LPO 8852: Regression II
Vanderbilt University
Take-Home Final Exam
December 9, 2021

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 your own 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 assumptions explicitly. Email your completed exam to sean.corcoran@vanderbilt.edu before 11:59 p.m. on **Tuesday December 14**. Good luck!

Question 1. Proponents of school choice argue that public schools will improve with greater competition from charter and private schools. The theory is that, in the absence of competition, traditional public schools have weak incentives to perform at their highest levels. Competition raises the risk of losing students (and funding), and public schools respond by increasing effort, using resources more efficiently, or prioritizing different outcomes. Empirically estimating the causal effects of competition on student outcomes and school performance is challenging, however. (**30 points**)

(a) Author 1 approached this problem by using a cross-sectional regression model of the following type:

$$y_{is} = \beta_0 + \beta_1 CharterComp_s + \gamma' X_i + \delta' W_s + u_{is}$$

in which y_{is} is an outcome for student i in traditional public school s, X_i is a vector of student-level covariates, and W_s is a vector of school-level characteristics. $CharterComp_s$ is a measure of charter school competition in the vicinity of school s; for example, this measure could be the number of charter schools within 1 mile of school s, or the percent of all same-grade-level students within 1 mile of school s enrolled in charter schools. The student outcomes (y_{is}) may include, for example, test scores, attendance, on-time graduation, or behavioral infractions. The coefficient of interest is β_1, the relationship between student outcomes at traditional schools and the extent of local charter school competition. Note there are multiple student observations per school, and the schools in the data are located within the same large urban district.

What key assumption(s) must hold in order to interpret β_1 as the causal effect of charter competition in Author 1's model? If there is suspected bias, in what direction do you believe it goes? Carefully explain. (**5 points**)

(b) Author 2 modified the approach of Author 1 by incorporating multiple years of data and estimating the following model with school fixed effects:

$$y_{ist} = \beta_0 + \beta_1 CharterComp_{st} + \gamma' X_{it} + \delta' W_{st} + \theta_s + \eta_t + u_{ist}$$

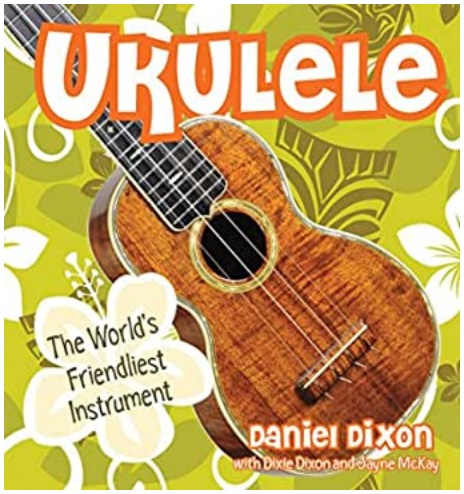
In addition to the terms listed in part (a), this model includes a school fixed effect (θ_s) and year dummies (η_t). Note the $CharterComp_{st}$ measure is now time varying—it increases as charter schools open near school s and/or gain market share. How does this approach improve upon that of Author 1, if at all? (**5 points**)

(c) What key assumption(s) must hold in order to interpret β_1 as the causal effect of charter competition in Author 2's model? What do you think the biggest threat(s) to these assumptions are in this context? (**5 points**)

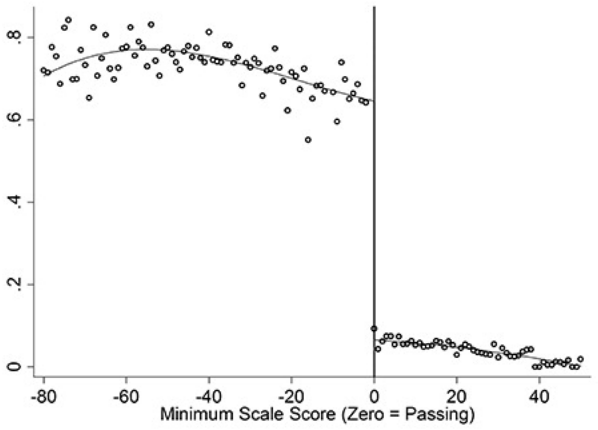
(d) In this school district, charter schools are responsible for finding their own space, and are more likely to locate in neighborhoods where space can be found. These spaces often include former schools, churches, strip malls, or industrial space. Author 3 uses an instrumental variable for $CharterComp_{st}$: the number of nearby buildings that are large enough to house a charter school. For example, if $CharterComp_{st}$ is the number of charter schools within a 1 mile radius of school s, the instrument z_{st} is the number of large buildings (say 30,000-60,000 square feet) within a 1 mile radius of school s. The estimating equation is otherwise the same as part (b), with the exception of the instrumental variable. How does this approach improve upon that of Authors 1-2, if at all? (**5 points**)

(e) What are the assumptions necessary for z_{st} to be a valid instrument for $CharterComp_{st}$ in this application? Do you think these assumptions are likely to hold in this context? Carefully explain why or why not. (**5 points**)

(f) Suppose that <u>both</u> Author 2 and Author 3's approaches are valid. That is, the necessary assumptions hold in both cases for interpreting β_1 as causal. At the same time, Authors 2 and 3 get substantively different estimates for β_1 using the same data. How could both approaches be valid yet yield different conclusions about β_1? (**5 points**)



Question 2. In many public colleges, incoming students are required to take a placement exam to assess whether they must enroll in a remedial education course in the subject (often math or English). Your local college administers such an exam and requires students below a certain threshold score to enroll in the remedial course. You are interested in whether these courses—which typically delay students' entry into non-remedial courses—make it less likely that students graduate in four years. You have collected data on recent placement test scores and remedial course enrollment (on the y-axis) and produced the following graph:



You have also collected subsequent graduation outcomes and other "pre-treatment" student variables (such as high school GPA and demographics) for the same students used in the graph above. (**40 points**)

(a) The fitted line above is a cubic function that is allowed to vary on either side of "0". Write down the regression equation that was used here, and be sure to define the variables you include. (**4 points**)

(b) Does it appear that the college strictly adhered to its rule requiring students scoring below the threshold to enroll in remedial courses? Briefly explain. (**3 points**)

(c) Carefully explain how you would use a regression discontinuity design to estimate the impact of mandated remedial education courses on students' graduation rates. For this part, write down the model you would use, and explain how it would be estimated and interpreted. (**10 points**)

(d) What are the key assumptions required for a causal interpretation of your model in part (c)? What might you do, if anything, to evaluate the plausibility of these assumptions? Carefully explain. (**10 points**)

(e) Define and apply the following terms to this problem: *average treatment effect, local average treatment effect (LATE), treatment-on-the-treated (TOT), and intent-to-treat (ITT).* That is, explain what each refers to in this specific case. Does your model in part (c) estimate the average treatment effect of remedial courses? Explain why or why not. (**8 points**)

(f) State whether the following statement is true or false. If false, explain why. (**5 points**)

If there is evidence of manipulation in the running variable in a regression discontinuity design, the treatment effect estimator must be biased.

Question 3. Suppose a researcher is interested in the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

She suspects that x_2 is endogenous and wants to use another variable z as an instrument. For each of the following, indicate whether the statement is true or false. If the statement is false, provide a brief explanation of what is wrong with it. (Make clear that you know why the statement is false). (**14 points - 2 each**)

(a) To be a valid instrument, z must be correlated with x_2.

(b) To be a valid instrument, z must be uncorrelated with y.

(c) If the researcher includes z in the equation above and estimates it by OLS, a statistically significant coefficient on z means z is not a valid instrument.

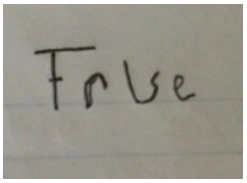
(d) In the first stage of 2SLS, x_2 is regressed on x_1 and z.

(e) Omitting x_1 from the first-stage equation can bias the estimated coefficients in the second stage.

(f) The researcher obtains the Wu-Hausman test statistic using the Stata command `estat endog`. A significant statistic implies rejection of the null hypothesis that x_2 is endogenous.

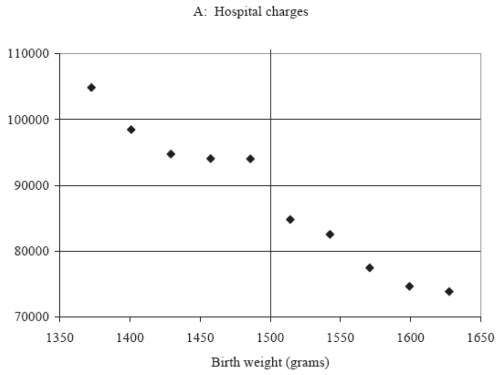
(g) If z is a weak instrument, significance tests of β_2 will reject the null hypothesis too frequently.

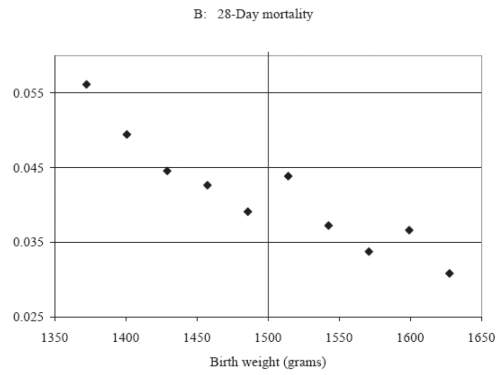
Figure 1: Frlse



Question 4. Consider the following regression discontinuity model constructed to estimate the impact of greater health care spending on the mortality of at-risk infants. The design exploits the fact that many hospitals have rules requiring greater care for newborns with especially low birthweight. For example, in some hospitals, newborns with very low weight (those $< 1500g$) are sent directly to the neonatal intensive care units (NICU) or to hospital wards with greater nurse supervision. The design here hopes to leverage the difference in health care at $1500g$ to estimate the benefits of greater health care spending on newborn outcomes. (**16 points**)

You have data on a large sample of children with low birth weights (1350 to 1650 g). The outcome of interest is 28-day mortality ($y = 1$ if the child dies within 28 days of birth). The key explanatory variable is hospital spending in dollars on the newborn (x). The two figures below show the relationship between birthweight and hospital spending (Figure A) and between birthweight and 28-day mortality (Figure B).





The table below reports coefficient estimates (and standard errors) for the following two regressions, where $D_i = 1$ if $BW_i < 1500$:

$$x_i = \alpha_0 + \alpha_1 BW_i + \alpha_2 D_i + u_i$$
$$y_i = \gamma_0 + \gamma_1 BW_i + \gamma_2 D_i + v_i$$

	(X) Hospital Charges in dollars	(Y) The newborn died within 28 days
Constant	260,250	0.168
	(23,000)	(0.021)
BW (in grams)	-115	-0.000083
	(15.1)	(0.00002)
D (BW<1500 grams)	7670	-0.0228
	(2300)	(0.003)

(a) Qualitatively, what do the figures suggest is the relationship between greater health care spending and outcomes for low weight newborns? Briefly explain. **(4 points)**

(b) Using the results reported in the table above, calculate the fuzzy RD estimate of the effect of increased hospital spending on 28-day infant mortality. Interpret this coefficient: what is your predicted change in 28-day mortality rate if spending on low birthweight newborns increases by $10,000? **(6 points)**

(c) What assumptions must be correct in order for the estimate in part (b) to be a consistent estimate of the causal impact of greater health care spending on newborn mortality? **(4 points)**

(d) Suppose a public health advocate uses the results in part (b) to argue for more health care spending for newborns in general. Given what you know of RD, what word of caution would you have for this person? **(2 points)**