LPO 8852: Regression II
Vanderbilt University
Midterm Exam
October 16, 2018

Name: _____

I agree to the terms of Vanderbilt University's honor code:

Signature: _____

Instructions: Read each question carefully and provide clear, concise responses. You may write directly on the exam. 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. You may not use your textbook or other materials for this exam. You have 80 minutes to complete the exam: 2:30 p.m. to 3:50 p.m. Good luck!

Question 1. In a 2017 paper, Anderson, Butcher, & Schanzenbach examined the effects of the federal No Child Left Behind (NCLB) accountability policy on student obesity. Their hypothesis was that schools facing accountability pressure reallocated time away from physical activity, such as recess and gym class, and toward reading and mathematics instruction, for which schools were held accountable under NCLB. To test this hypothesis, the authors analyzed a panel of school-level data from Arkansas which contained average test scores, obesity rates, and other student demographics. The unit of observation was a school-year, and the data spanned 1998 through 2010. (**30 points**)

(a) NCLB came into effect in 2002. Define $overwgt_{st}$ as the percent of students with a body mass index (BMI) above the threshold for overweight and $post_t = 1$ if the year is 2002 or later. Anderson et al. considered estimating the regression model below to estimate the effect of NCLB on the percent overweight. $pctpoor_{st}$ is the percent of students below the poverty line in school s in year t, $pctnw_{st}$ is the percent nonwhite, $mathprof_{st-1}$ is the percent proficient in math in the prior year, and $year_t$ is a linear time trend. Clearly explain what β represents in this case. (**5 points**)

$$overwgt_{st} = \alpha + \beta post_t + \gamma pctnw_{st} + \delta pctpoor_{st} + \phi mathprof_{st-1} + \omega year_t + u_{st}$$

(b) Under what conditions can we interpret the coefficient β as the causal effect of NCLB? Are these conditions likely to hold here? Briefly explain why or why not. (**5 points**)

(c) Anderson et al. recognized that not all schools faced the same accountability pressure under NCLB. Schools near the threshold of failing to make "adequate yearly progress" (AYP) as defined under NCLB were under strong pressure to improve math and reading scores while those away from this threshold did not experience the same pressure. Let $\tau_{st} = 1$ in the first year school s was at risk for failing to make AYP *and all years following* ($= 0$ otherwise). Write down a revised version of the regression model above that uses a difference-in-differences approach to estimate the effect of NCLB accountability pressure on the percent overweight. (Hint: only some schools were ever at risk during this period.) Carefully explain the how the key regression coefficient(s) here should be interpreted. (**5 points**)

(d) Under what conditions can we interpret the difference-in-differences coefficient in your model in part (c) as the causal effect of NCLB? Are these conditions likely to hold here? Briefly explain why or why not. (**5 points**)

(e) Table 2 below from Anderson et al. reports their DD estimate of the effect of NCLB pressure on the percent overweight. Carefully provide a written interpretation of the point estimate in column (3). (**5 points**)

Table 2. Effects of Accountability Pressures on School Rates of Overweight Students

	(1)	(2)	(3)	(4)
Pressured in past	1.033*** (0.351)	1.220*** (0.313)	1.205*** (0.286)	0.522*** (0.151)
Overweight rate (previous year)				0.608*** (0.0153)
Overall proficiency rate	NO	YES	YES	YES
Demographic controls	NO	NO	YES	YES
Observations	4,588	4,588	4,588	4,588
R^2	0.007	0.125	0.248	0.496

Notes: Pressured in past is defined as whether a school's minimum-scoring subgroup had a proficiency rate within 5 points of the AYP target for some year in the past. Overweight rate includes all weights above normal weight. Demographic controls are a quartic in percent nonwhite and a quartic in percent economically disadvantaged. Overall proficiency rate controls are a quartic in the standardized overall literature proficiency rate and a quartic in the standardized overall math proficiency rate. All models include an annual trend. Standard errors which are robust to heteroskedasticity and within-school correlation are in parentheses.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

(f) To provide support for the identifying assumptions of their DD regression model, Anderson et al. estimate the "event study" model shown below. This model includes separate dummy variables for years before and after a school faced accountability pressure. For instance, $D_{st,i} = 1$ if year t represents year i, relative to the time period in which school s faced accountability pressure (e.g., $D_{st,-2} = 1$ if t is two years prior to accountability pressure for school s; $D_{st,2} = 1$ if t is two years following accountability pressure. D always equals zero if a school never experienced accountability pressure). Briefly explain how the β_i coefficients should be interpreted. What kinds of estimates for the β_i would support a causal interpretation of the DD design? (**5 points**)

$$overwgt_{st} = \alpha + \sum_{i=-4}^{9} \beta_i D_{st,i} + \gamma pctnw_{st} + \delta pctpoor_{st} + \phi mathprof_{st-1} + \omega year_t + u_{st}$$

Question 2. You are interested in the expected wage premium from earning a GED for the typical high school dropout. You have collected a dataset representing nearly 50,000 high school dropouts in your state with labor market information (hourly wages, hours and weeks worked, work experience, etc.), GED completion status, and some demographic and academic indicators from high school (GPA, credits completed, race/ethnicity, gender, free or reduced price lunch status, English proficiency, etc.) (**18 points**)

(a) You begin by estimating the OLS regression model below, where $wage_i$ is individual i's hourly wage at age 25 and $GED_i = 1$ if the individual earned a GED by age 25 (= 0 otherwise). Does the population regression function 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**)

$$wage_i = \beta_0 + \beta_1 GED_i + u_i$$

(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. (**5 points**)

(c) To mitigate OVB you opt for a propensity score matching approach in which you match dropouts who have earned a GED by age 25 to similar dropouts who have not earned a GED by age 25. Matching is based on the individual's probability of having earned a GED by age 25. Your propensity score model includes a long list of covariates measured prior to age 18 and thus are not affected by the receipt of a GED. These include student demographics, credits accumulated in high school, GPA in completed courses, days absent, etc. Carefully explain the assumptions required for this method to provide plausibly causal estimates of the effect of earning a GED on wages. (**5 points**)

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

Question 3. You are interested in the effect of attending a charter school on middle school math and English achievement. You have access to a longitudinal student-level dataset that follows students from grade 4 to grade 8. This dataset contains annual test scores in math and English, relevant student demographics, and school ID variables (some of these schools are charter schools). (**19 points**)

(a) You have considered an OLS regression that relates a measure of student achievement to $charter_{it}$, a dummy variable equal to 1 if the student is enrolled in a charter school (= 0 otherwise) and a vector of student covariates ($\mathbf{X_{it}}$) that you believe may be related to charter school attendance and student achievement. Carefully explain how this model may suffer from omitted variables bias due to selection on unobservables. (**5 points**)

$$zscore_{it} = \alpha + \beta charter_{it} + \gamma' \mathbf{X_{it}} + u_{it}$$

(b) As an alternative to the regression model in part (a), you are considering a model with *student fixed effects* (δ_i). Under what conditions can this model be estimated? Under what conditions can we interpret the coefficient β as the causal effect of attending a charter school in this model? Carefully explain. (**6 points**)

(c) Are there any downsides to the regression specification described in part (b)? What are they? Briefly explain. (**5 points**)

(d) Finally, when estimating regression models like this one, most researchers allow for correlation in the error term within students over time. Briefly explain why this is important. (**3 points**)