**ITAM**

**Ciencia de Datos para Economistas**

**Fall 2017**

**Midterm Exam**

**PART I – TRUE OR FALSE OR UNCERTAIN (x points each)**

1. (3 points): With data on individual workers, you regress daily earnings on years of education, years of education squared, and a constant. You obtain the following sample regression function:

With this information, it is possible to estimate the increase in earnings associated with a one-year increase in education?

***A: False, it is impossible to tell without more information.***

2. (3 points): With estimated coefficients from a probit regression, it is possible to learn the statistical significance and sign only.

***A: True, magnitude it´s not possible in probit regression.***

3. (3 points): Mexico's government asks you to evaluate the impact of a cash transfer program on poverty outcomes. Households with less than 5 acres of land receive cash transfers, while those with more than 5 acres do not. The government only has data on poverty outcomes one year after the intervention. In this context, to evaluate this program, Difference-in-difference is preferred to Regression discontinuity and Fixed effects.

***A: False, regression discontinuity it`s the best instrument in this case.***

4. (3 points): Massachusetts runs advertisements on state-wide television and radio stations to promote safe driving and you are asked to evaluate this campaign. Legislators then provide you with county-level data on accident rates for the year before and the year after the ads ran. The legislators ask you to run a county fixed effects model to determine the impact of the program. It is valid to explain that you cannot run a fixed effects model since since you would need more than two years of data?

***A: False, you have to explain that you cannot run a fixed effects model since all counties were equally affected by the program.***

5. (5 points): Because omitted variables cause bias, it is always recommended to include all available explanatory variables in a regression.

***A: False, first of all, omitted variables do not always cause bias. Omitting a variable only causes bias when the omitted variable is correlated with the included variables. Second, there are reasons that some variables should not be included in a regression. One reason is that some variable may be an intermediate variable between the important explanatory variable and the dependent variable.***

6.-(i-iii) We have a classic example using race (black=1) and gender (female=1) dummies:

Log(wage) = β0 + β1Female + β2Black + β12(female ∗ Black) + e

Then we have:

1. (5 points): β0 represents the expected log wage of a black male (Female=0, Black=0).

**A:False, it represents log wage for non-black male**

1. (5 points): (ii) Expected log wage for non-black female is β0 + β1 and for black female is β0 +β1 +β2 +β3.

**A: True.**

1. (5 points): (iii) β1 represents the wage difference between non-black males and non-black females and β2 represents the wage difference between white females and black females.

**A: True.**

**PART II**

1. (10 points): Whether an individual has a college degree is clearly not random. To identify the causal impact of a college degree on wages, researchers decide to use father’s college degree status as an instrument for an individual's college degree status. Assuming there is a strong correlation between fathers' and children's college degree statuses, would this instrumental variable approach successfully estimate the causal effect of college education on wages? Explain. [4 sentences]

***No. This instrument is almost certainly not exogenous and thus not valid. Having a college-educated father may affect a child's wages through many channels other than the child's own college status. For example, college-educated fathers may have extensive social networks that help their children find better jobs regardless of whether the children are college-educated.***

1. (10 points): Some researchers tried to estimate the impact of a higher minimum wage on employment. New Jersey had increased its minimum wage and Pennsylvania had not. They ran the following regression

*Employment* = β0 + β1*NJ* + β2*After* + β3 *NJ\*After* + u

and found that the interaction coefficient was positive, suggesting that the minimum wage increase had actually increased employment in fast food restaurants.

Critic 1 of this study argues that the estimate is biased due to the fact that New Jersey’s workforce is less educated (and thus more likely to work in low paid jobs) than Pennsylvania’s workforce. Critic 2 argues that the estimate is biased due to the fact that Pennsylvania experienced a worse recession than New Jersey around the time the minimum wage law was passed.

Does either of these critics have a valid argument? Explain. [5 sentences]

***Critic 1 does not have a valid argument. The difference in education levels between nj and pa don't change over time (at least in this short time period). The nj dummy thus successfully controls for any such difference between the two states that is constant over time.***

***Critic 2 does have a valid argument. This difference-in-difference regression does not control for other factors that affect nj and pa in ways that change differentially over time. We thus may be confusing the impact of the law change with the impact of a differential economic shock.***

1. (10 points): A standard “money demand” function used by macroeconomist has the form:

*ln(m)* = β0 + β1*ln(GDP)* + β2*R* + u

where m is the quantity of (real) money, GDP is the value of (real) gross domestic product, and R is the value of the nominal interest rate measured in percent per year. Suppose that β1  = 1.0 and β2 = -0.02. What will happen to m if the interest rate increase from 4% to 5%?

***A 2% increase in GDP means that ln(GDP) increases by 0.02. The implied change in ln(m) is 1.0 × 0.02 = 0.02, which corresponds to a 2% increase in m. With R measured in percentage points, the increase in R is from 4.0 to 5.0 or 1.0 percentage point. This leads to a change in ln(m) of −0.02 × 1.0 = −0.02, which corresponds to a 2% fall in m.***

**PART III**

Now suppose you wanted to do an instrumental variables analysis of the effect of school quality on students’ test scores. Economist Bruce Sacerdote does precisely this in his paper “When the Saints Go Marching Out: Long-Term Outcomes for Student Evacuees from Hurricanes Katrina and Rita” (American Economic Journal: Applied Economics, 4(1): 109–35.).

The paper uses two major natural disasters (Hurricanes Katrina and Rita) as an instrument for school switching. In particular, for people living in the hardest hit areas, entire families were forced to evacuate and relocate to other neighboring states and regions (e.g. eastern Texas).

Because schools in areas affected by these natural disasters were disproportionately low performing schools, many students affected were not only forced to switch schools, but also ended up switching to a *higher quality school*. Sacerdote finds initial declines in test scores (likely due to the shock of displacement), but “by the third and fourth years after the disaster, evacuees displaced [by these natural disasters] see a 0.18 standard deviation improvement in scores.” Sacerdote also finds that the improvements observed were “concentrated among students initially in the lowest quintiles of the test score distribution.”

1. (10 points): Define the outcome (Y), the endogenous regressor (X), and the instrument (Z) in the context of this study.

* ***Outcome (Y) – students’ test scores***
* ***Endogenous Regressor (X) – school quality***
* ***Instrument (Z) – natural disaster that forces a child to switch to a school of higher quality***

2. (5 points): Is this instrument relevant? Explain why or why not.

***Yes: it is correlated with switching to a higher quality school***

3. (10 points): Is this instrument exogenous? Explain why or why not.

***Yes: it was unexpected and the school changes made would almost certainly not have happened in the absence of a natural disaster. Thus there is no “reverse causality” (i.e. it is highly unlikely that students selected schools that would be hit by a natural disaster so as to get placed into better performing schools)***

4. (15 points): For what group of individuals is this study valid? List at least two groups for which this study is unlikely to yield valid estimates.

***This study is valid for individuals in low-performing schools in the American south. Specifically, it tells us the effect on students’ test scores when they are unexpectedly placed into better performing schools. It is unlikely to teach us about how test scores would change if students from high-performing schools were placed into schools of differing quality, nor is it likely to tell us about the effects one might observe in another country where there are fewer disparities in educational quality across regions.***

**Part IV**

In the paper “Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns,” authors Douglas Almond, Joseph J. Doyle Jr., Amanda E. Kowalski, and Heidi Williams use discontinuities in how very low birth weight babies are cared for to estimate the marginal returns to medical care for at-risk newborns. In particular, the “treatment” is that below a birth weight of 1500 grams, babies are given much more in-hospital care than similar babies weighing just above 1500 grams. The authors are therefore interested in the effect of additional in-hospital care on the outcomes of low birth weight babies.

The full paper has been posted and the abstract is as follows:

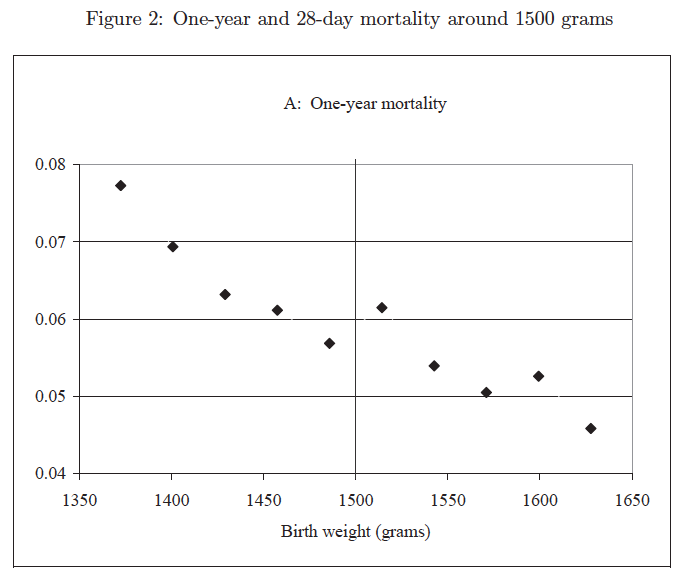
*We estimate marginal returns to medical care for at-risk newborns by comparing health outcomes and medical treatment provision on either side of common risk classifications, most notably the “very low birth weight” threshold at 1500 grams. First, using data on the census of US births in available years from 1983-2002, we find evidence that newborns with birth weights just below 1500 grams have lower one-year mortality rates than do newborns with birth weights just above this cutoff, even though mortality risk tends to decrease with birth weight. One-year mortality falls by approximately one percentage point as birth weight crosses 1500 grams from above, which is large relative to mean one-year mortality of 5.5% just above 1500 grams. Second, using hospital discharge records for births in five states in available years from 1991-2006, we find evidence that newborns with birth weights just below 1500 grams have discontinuously higher costs and frequencies of specific medical inputs. We estimate a $4,000 increase in hospital costs as birth weight approaches 1500 grams from above, relative to mean hospital costs of $40,000 just above 1500 grams. Taken together, these estimates suggest that the cost of saving a statistical life of a newborn with birth weight near 1500 grams is on the order of $550,000 in 2006 dollars.*

1. (15 points): What are the requirements to use regression discontinuity design ***in this case*** and are they likely to be met?

*From the class notes:*

* *Assignment to treatment must be based only on the cutoff score. –* ***likely to be met; we know that babies under 1500g are given more care as measured by cost.***
* *The assignment variable cannot be caused by treatment. –* ***likely to be met; birth weight cannot be influenced by care received ex-post.***
* *The assignment variable can even be totally unrelated to outcome and have no particular substantive meaning. –* ***not of use here***
* *The ideal assignment variable is a continuous variable. Such variables maximize the chance of correctly modeling the regression line for each group, which is crucial to the success of the RD design. –* ***met: weight is a continuous ariable***
* *Also important is that individuals cannot perfectly control the value of the assignment variable. In other words, we don't want individuals to be able to control whether they are treated or not. –* ***likely to be met; if we believe that doctors use a rule of thumb, then individuals cannot control treatment.***

2. (5 points): Panel 1 of figure 2 is below. Describe the visual evidence for using regression discontinuity design in this case.



***We see that there is a jump in one-year mortality as we cross the 1500g cutoff, which suggests that there is a discontinuity in treatment at the 1500g threshold.***