Introduction to econometrics
Econ 4811 – Spring 2018

Problem Set 7

Due: *Beginning of class, May 2, 2018*

Question 1

Evidence shows that many undergraduate students have bad eating habits. Caffeine, candy, and fast food are essential parts of their diets. We hypothesize that the ability to learn is partially related to a student's physical health, and so we decide to evaluate if eating better foods will improve a student's academic performance. We decide to design a randomized control trial (RCT).

We take all non-senior students in Principals of Economics and randomly assign some of them to receive a week's worth of fresh fruit and vegetables at their apartment/dorm/home each week for the rest of their time at CU Denver. For each of these students, once they graduate we collect information from the CU Denver administration on their high school GPA, gender, college GPA, and if they are an ICB student or not. Also, we have the students go to a health clinic as seniors to record their blood pressure and weight.

Our main model is: $GPA_i = \beta_0 + \beta_1 FV_i + u_i$

where GPA is the student's final college GPA and FV is an indicator equal to 1 if the student was assigned and ate the delivered fruits and vegetables.

a) Why do you need to go through the trouble of a running a RCT in this case?
Why can't we just ask students about their diets and then run a regression like:
$GPA_i = \beta_0 + \beta_1 DietQuality + u_i$? Explain the general problem and give a specific example.

b) Let's assume you have full compliance and you know that randomization worked perfectly. Does your estimate of the impact of the treatment on GPA give you the causal relationship between a healthier diet and academic performance as measured by GPA? If yes, give two key reasons that you can make this statement. If no, give an example of why you may not be able to make this statement.

c) Let's say you find out compliance is not perfect in the treatment group. Specifically, you know that some students don't like the taste of fruits and vegetables and so won't eat them. Other than these people, compliance is perfect. Can you still get the causal effect of receiving and eating fruits and vegetables on GPA using $GPA_i = \beta_0 + \beta_1 FV_i + u_i$? Explain why or why not. Remember, FV=1 only if the people are assigned to the treatment and eat the fruits and vegetables and 0 otherwise.

d) Provide one reasonable and plausible example of why you might have selective attrition in this study?

e) Given your concerns about attrition, you want to check to see if your randomization is working. Given the data mentioned at the beginning of this question, what regression(s) would

you run to test if FV is random. Explain the intuition of the test, give the specific model(s) you would run, and what result from those models would make you feel better that randomization worked.

f) You find out that your randomization did not work. To account for this you decide to calculate the D-i-D estimate. You collect the GPA before treatment started for all the students in your sample. You find the following:
For students in the FV=1 group:
GPA before treatment: 3.5
GPA after treatment: 3.0

For students in the FV=0 group:
GPA before treatment: 4.0
GPA after treatment: 3.0

i.) What is the estimated effect of FV on GPA only using post-treatment data?
ii.) What is the D-i-D estimate of the effect FV on GPA?
iii.) What is the intuition for why these estimates are the same or different?

g) When using this model what are you assuming the GPA would have been in Treatment if they had not received the fruits and vegetables? How is this assumption called?

h) How would you test this assumption?

Question 2

Many studies have attempted to identify the causal effect of education on wages. The relationship between wages and education has been studied across the globe, for different population subgroups, age groups, students who attend elite colleges, those that don't and every possibility in between.

(a) Using the data in **ps7q2.dta**, estimate the model

$$\text{lnwage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{age}_i + u_i \qquad [1]$$

where $\text{lnwage}_i = \log(\text{wage})$
educ_i=years of education
age_i=age in years

by the method of OLS. Interpret your results.

(b) What condition needs to be satisfied in order to interpret $\hat{\beta}_1$ as the causal effect of education on wages? Be precise. Does this seem like a plausible assumption? Why or why not?

(c) It has been suggested that the cost of tuition at the local college is an indicator of the price of higher education and, therefore, is likely to affect the probability a student attends and completes college. It has also been claimed that tuition costs at the local college should not affect the wages a person earns. This suggests that the tuition costs at the local college is a valid instrument for education in [1]. Discuss.

(d) Estimate [1] using the method of instrumental variables with

(i) the tuition at the local college at the time the respondent was 17
(ii) the tuition at the local college at the time the respondent was 18
(iii) the tuition at the local college at the time the respondent was 17 and at age 18.

From a theoretical perspective, which of the three sets of instruments is most appealing? Why? Are there other controls that you think belong in the model for the exclusion assumption to be plausible? What are they? Explain.

For each set of instruments, estimate the first stage regression and model [1] by the method of instrumental variables. What is meant by weak instruments? How can you test for weak instruments? Discuss the results of the three sets of IV estimates with this in mind. Compare your estimated returns to education with the estimates in (a)

(e) An alternative school of thought argues that the key impediment to going to (and completing) college is parental resources. This school argues that the education of a person's father (or mother) is a good instrument for one's own education. Discuss whether this instrument is likely to satisfy the exclusion assumption.

Estimate [1] using father's education as an instrument. Discuss the first stage and IV estimates. What do you conclude?

(f) What is meant by an over-identified model? Give an intuitive explanation of how you might test whether instruments satisfy the exclusion restriction in a model that is overidentified. Explain why you cannot do this in an exactly identified model.

(g) Estimate the first stage regression using tuition at the local college at the time the respondent was 17, at the time the respondent was age 18 and father's education as instruments.

$$\text{educ}_i = \gamma_0 + \gamma_1 \text{ tuit17}_i + \gamma_2 \text{ tuit18}_i + \gamma_3 \text{ fatheduc}_i + \gamma_{\gamma 4} \text{ age}_i + \text{v}_i \qquad [2]$$

Calculate the predicted value of education, educ_hat, and include that predicted value in [1] in place of education. Estimate the same model using ivregress. In what ways do the two sets of estimates differ? Which has a larger estimated standard error for $\hat{\beta}_1$? Why?

Test whether the instruments in [2] are "weak". Test whether the instruments satisfy the exclusion restriction. What do you conclude?

(h) Calculate the residual, vhat, from [2] and estimate [3] by the method of least squares.

$$\text{lnwage}_i = \delta_0 + \delta_1 \text{ educ}_i + \delta_2 \text{ vhat}_i + \delta_3 \text{ age}_i + \omega_i \qquad [3]$$

How are $\hat{\delta}_1$ and $\hat{\delta}_3$ related to your IV estimates of [1] in (g)?

(i) A third school of thought argues that education is a marker for cognitive achievement and it is cognitive skills that are rewarded in the labor market. The respondents in this study completed a cognitive assessment and the scores have been converted to a z score, cog. Like education, it is

potentially correlated with other, unobserved characteristics that affect wages and so is properly treated as an endogenous covariate. Include both cog and educ in [1] and re-estimate the model using all 3 instruments. Test whether cognitive achievement and education are jointly significant in the IV regression. Discuss the results.

Question 3

The IV intuition police

Many studies have focused on the economic returns to education. In these studies they usually see if an extra year of education has an effect on a person's income Income=beta_0+beta_1 YrsSchool +u_i. For all these studies, a major concern is the fact that inherent intelligence should enter as a determinant of earnings, but that it is close to impossible to measure and therefore represents an omitted variable. Assume that the coefficient on years of education is the parameter of interest.

a) Given that education and income is correlated with inherent intelligence, the OLS estimator for the returns to education could be biased. In what direction are you biasing your estimate of the returns to education by omitting inherent intelligence from the regression?

To overcome this problem, various authors have used instrumental variables estimation techniques. For each of the potential instruments listed below, discuss whether you believe the proposed instrument for education is valid (remember what the 2 conditions are for instrument validity).

These have all been legitimately attempted by some researcher at some point.

b) The individual's postal zip code used as dummy variables to indicate the area an individual lives

c) The individual's SAT score.

d) Years of education for the individual's mother or father

e) If a good affordable college was built and opened close to where the person lived when they turned 16

f) Whether the individual was randomly assigned to receive a tutor

g) The number of siblings the individual has.