Econometrics 2

Central European University

**Final Assignment**

Due Friday 30 April 2020, 15.30

This assignment is based on the paper by Holden, Kristian L. "Buy the book? Evidence on the effect of textbook funding on school-level achievement." *American Economic Journal: Applied Economics* (2016): 100-127.

Read the paper and the data\_readme.pdf for a description of the variables used in the dataset.

For this assignment you should submit a .pdf (or word) document and a separate do-file.

Download paper and data using the following link:

<https://www.aeaweb.org/articles?id=10.1257/app.20150112&within%5Btitle%5D=on&within%5Babstract%5D=on&within%5Bauthor%5D=on&journal=4&q=buy+the+book&from=j>

Question 1: Errors in Variables

Use CA\_schools\_ms data

1. Set the seed to a number of your own choice. Why do we set the seed? Create a normally distributed random variable using the following command: *gen me=rnormal(0,2)*.
2. Standardize the api\_rank and readingscore by subtracting the mean and dividing by the standard deviation. In your opinion, why do we use sometimes the standardized version of a variable? Use the standardized version of the variables for the reminder of the exercise.
3. Regress Reading score on API rank, what is the R-square? Regress API rank on Reading score, what is the R square? Describe, why do you think this is the case. What does the R-square say about causal inference?
4. gen api\_error= api\_rank +me. What is this variable saying? Now regress readingscore on API rank. What is the coefficient on API rank? Regress readingscore on api\_error. Does the coefficient on api\_error coincide with the one on api\_rank? Why? Can you test formally for the equality of these coefficients?
5. Now use the following commands:

sum api\_rank, d

gen api\_rank\_el=api\_rank+me if api\_rank>`r(mean)'

replace api\_rank\_el=api\_rank if api\_rank<=`r(mean)'

What are we doing by running these lines of code?

1. Regress readingscore on api\_rank\_el. Does the coefficient on api\_rank\_el coincide with the coefficient from the regression of readinscore on api\_rank? Does it coincide with the coefficient you get from regressing readingscore on api\_error? Why do you think is the case?
2. What happens if you instead generate a normally distributed random variable with mean 0.5 and variance =2 and repeat 1.d)? What happens if you instead generate a normally distributed random variable with mean 0 and variance =3 and repeat 1.d)? Discuss.
3. Assume now that the measurement error is in the dependent variable. Where the new dependent variable measured with error will be readingscore\_error=readingscore+me. Run the specification using readingscore\_error as the dependent variable. From the lecture notes, we know that when measurement error in the dependent variable, the error is iid and the covariance between the measurement error and the regressor is zero, the OLS estimator is unbiased but the error variance is large. Do your estimates reflect this? Why? Why not?

Question 2: Replication Regression Discontinuity

1. Summarize the main findings of the paper. Is this a novel research question? What are the main endogeneity concerns when studying the effect of books on performance? What is the identification strategy that the author proposes? How does this identification strategy allow to causally identify the effects of textbooks on performance? What type of regression discontinuity is this? Why?
2. Use the codes already available at the AEJ website and replicate Figure 1, Figure 2, Figure 3 (Panel A), and Table 5. Comment on the codes already provided by the author and submit a do-file with only relevant parts of the code. (Non-mandatory: Can you provide a different coding alternative?) **Bonus 2 points:** Automate the generation of Table 5 (Hint: use outreg2 or estout)
3. What are the main threats to the identification strategy? If you would have any data how would you test for potential threats to identification?