EKHM66 Fall 2023
Assignment 2: Experiments, instrumental variables, and regression discontinuity

Hand in no later than October 11th at 16:00 via the hand-in function at CANVAS. The exercise is to be performed in pairs. Use Stata to answer the questions and submit your answers as a PDF file with your names included.

In this exercise, we apply what we have learned about the experimental ideal, instrumental variables (IV) regression, and regression discontinuity.

Part I

We start with a dataset called *healthRAND.dta*. It is a random subsample (n=5000) from the dataset we worked with during Lab 2 (RAND Health Experiment). The dataset consists of a treatment variable ("ins"), an outcome variable ("health"), and a number of socio-economic and demographic independent variables.

The research question we are working with is: What is the causal effect of health insurance on health?

List of variables:

Name	Format	Description
health	Integer (0-5)	A self-reported health-score.
ins	Integer (0/1)	A dummy variable for having any kind of health insurance (1=insured).
female	Integer (0/1)	A dummy variable for gender (1=female).
non-white	Integer (0/1)	A dummy variable for ethnicity (1=non-white).
educ	Integer	A continuous variable for number of years of education.
income	Float	Yearly family income in USD.
family_id	Integer	A family identification variable.

Questions:

1. Load the dataset. Report output of differences in means for the dependent variable and one independent variable of your choice by treatment status. Describe and interpret the reported differences in means by treatment status.

2. Choose an independent variable and use a t-test to determine whether an observed difference in means between the treated and untreated is statistically significant. Report and comment on the difference in means and associated p-value.

3. Use regression to test whether the mean health score differs between treated and untreated. Report full regression output and comment on the results. Explain whether

the analysis answers the research question satisfactorily. Elaborate on whether the assignment to treatment has been randomized or not and how that affects answering the research question.

Part II

The dataset we work with is *ak91_subsample.dta*. It is a random subsample (n=200,000) of the Angrist and Krueger dataset we worked with in Lab 2. The dataset consists of a treatment variable ("schooling"), an outcome variable ("logwage"), and two variables for year and quarter of birth. Each individual appears in the data once.

The research question we are working with is: What is the causal effect of schooling on earnings?

List of variables:

Name	**Format**	**Description**
ln_wage	Float	Log of wage.
schooling	Float	Years of education.
yqob	Float	Year and quarter of birth.
qobi	Integer (0/1)	Quarter-of-birth instrument (1=born in 4th quarter of year).

Questions:

4. Explain in words why the instrumental variable method solves the confounding problem in this case, and why the provided instrument is appropriate to use.

5. Test the assumptions of the IV analysis in Stata (when appropriate), report output, and interpret. If a formal test is not appropriate, explain in words whether you think the assumption holds or not, and why.

6. Run a Two-Stage Least Squares regression of earnings on schooling. Report output from first and second stages and interpret.

Part III

The dataset we work with is *mlda.dta*. This RD setup exploits knowledge of the fact that the Minimum Legal Drinking Age (MLDA) in the U.S. is 21 years, combined with mortality rate data with high precision in the time dimension.

The research question we are working with is: What is the causal effect of alcohol consumption on young adult mortality?

List of variables:

Name	Description
agecell	Age (decimal years).
all	All-cause mortality (deaths/100,000).
homicide	Homicide mortality (deaths/100,000).
mva	Motor vehicle accident mortality (deaths/100,000).

Questions:

7. In this RD setting (discussed also in the course book), define the following:
   - Causal variable of interest
   - Running variable
   - At least one potential confounding factor.

8. Describe the confounding problem in this setting, and how this specific design can potentially create a causal estimate to answer the research question.

9. Present two scatterplots, showing the running variable on the x axis, and:
   - Motor vehicle accident mortality on the y axis
   - Homicide mortality on the y axis.

   Comment briefly on the graphs and what the reason behind differences between the two graphs may be.

10. Generate a treatment dummy that switches in accordance with the rule that we are exploiting. Run two RD regressions with the running variable and the RD dummy as independent variables, and:
    - Motor vehicle accident mortality as outcome
    - Homicide mortality as outcome.

    Report the regression results and comment on the magnitude and significance of the RD dummy coefficient in the two models.