EKHM66 Fall 2023
Assignment 1: Regression

Hand in no later than September 20nd 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 regarding exploring a dataset using Stata and about Ordinary Least Squares (OLS) regression and non-experimental data.

Part I

We start with a dataset called *healthNHIS.dta*. It is a random subsample (n=5000) from the dataset we worked with during Lab 1 (NHIS Health Survey). 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 (=tables) 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. (1 point)

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. (1 point)

3. Use regression to test whether the mean health score differs between treated and untreated. Report full regression output and comment on the results. (1 point)

4. Is there a way to satisfactorily answer the research question using the current data? Why, or why not? (1 point)

Part II

The dataset we work with is *earnings.dta*. It is a random subsample (n=2500) of the PSID dataset we worked with in Lab 1. The dataset consists of a treatment variable ("college"), an outcome variable ("logwage"), and other socio-economic and demographic independent variables. The data are a panel of individuals over time.

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

List of variables:

Name	Format	Description
logwage	Float	Log of wage.
college	Integer (0/1)	A dummy variable for having a college degree (1=degree).
female	Integer (0/1)	A dummy variable for gender (1=female).
non-white	Integer (0/1)	A dummy variable for ethnicity (1=non-white).
occupation	Integer (0/1)	A dummy variable for occupation (1=blue-collar).
industry	Integer (0/1)	A dummy variable for industry (1=manufacturing).
id	Integer	A person identification variable.

Questions:

5. Use an appropriate method in Stata to look for evidence of selection into treatment based on gender and ethnicity. Report output and interpret. (1 point)

6. Run a bivariate regression model (model 1) of college degree on earnings. Report output and interpret. Explain whether this analysis answers the research question or not and why. (1 point)

7. Which of the available independent variables would you choose as control variables in a multivariate regression model, and why? (1 point)

8. Run your chosen model (model 2) and store the estimates. Then, add one or more variables that you consider a "bad control", run and store the model (model 3). Report the models side by side and compare. Explain what changed and why. (1 point)

9. Considering model 2. It is reasonable to think that an individual's ability would affect the level of income. If omitting the variable "ability" in the regression, will that bias the estimate of college? If yes, in what direction? (1 point)

10. Explain shortly, verbally and with equations, why random assignment to treatment solves the selection problem. (1 point)