Please be concise. Upload your answers in moodle before 23:59 on May 25. The solutions should include a .pdf or .doc document with the answers to each one of the questions and the log files or do files of the exercises done using stata. We should be able to replicate your results following your log or do file. If you have any question, do not hesitate to send me an email: teresa.molina@novasbe.pt

1. (12 points) **Panel Data**

In this exercise you are asked to set up a model for estimating the effect of alcohol consumption on labor market outcomes. The data set **alcohol_main.dta** comes from the National Longitudinal Survey of Youth (NLSY) and includes information on labor market outcomes, alcohol consumption, and assorted demographics for individuals in 1989 and 1994 waves. The data are restricted to young adults who are between the ages of 24 and 32 in 1989 (and hence 29-37 in 1994). Each individual has a unique identifier (variable named id) and the year is indicated by the variable named year.

While I have defined the basic outline of the question for you, you need to refine it. There is no single "right answer". What is important is to think about **what question do you want to answer with your model**, and properly interpret the results in that context. There is n single right answer for what dependent variable or key explanatory variables to choose:

1. Select the labor market outcome you are interested on: currently employed, annual hours worked, annual income from wages, average hourly wage.
2. Select the kind of alcohol consumption to investigate: any consumption, any heavy consumption, amount consumed, frequency of consumption. You may have to define some of these variables. For example, what does "heavy" mean, some studies classified as alcohol abusers female (male) respondents reporting six or more drinks at one time on average for three (five) occasions during the last 30 days.

Some of these variables are already available in the database, but for most of them you will need to do some computation.

(a) (1 point) State the research question, define the labor market outcome and the measure of alcohol consumption you will analyze and very briefly motivate your choice. (*Max: 700 characters.*)

(b) (2 points) Describe the data (*Max: 900 characters.*)

   1. State whether it is a balanced or unbalanced panel, and what is the retention rate.
   2. Describe the distribution of your outcome of interest and of alcohol consumption.

(c) (2.5 points) To answer your research question you decide to exploit the panel data nature of the NLSY and estimate a model controlling for individual-specific heterogeneity. (*Max: 2,500 characters without including formulas.*)
   Describe in detail the model you will use to estimate the relation between your labor market outcome and the measure of alcohol consumption (including control variables). Write down the **equation** you will estimate, be careful with the **notation** you use, and state any **assumptions** needed to consistently estimate the parameter of interest. Think about the set of **control variables** you want to include in the model. *Hint: Some of these variables are available in the regression, however, you may want to transform some of them, or you may want to construct new variables. Carefully, review the list of demographic variables in the appendix.* Think about the **standard errors** of your model.

(d) (1 point) Instead of using individual fixed effects, you could have used random effects estimation. Under which assumptions would you prefer the random effects model? and why? (*Max: 900 characters*)

(e) (2 points) Estimate your model, report the estimates and standard errors and interpret the results. Focus only on the relation between alcohol consumption and the labor market outcome. (*Max: 900 characters, without including the table with the results.*)

(f) (1.5 points) Modify your model to analyze whether alcohol consumption has a differential effect on the labor market outcome according to the gender of the respondent. Present your model, report and discuss your results. (*Max: 1200 characters, without including the table with the results and the model.*)

(g) (2 points) Discuss whether your model has a causal interpretation. Explain, with examples, what are the main threats to the identification strategy? (*Max: 1,200 characters*)

2. (8 points) **Determining causal effects using propensity scores**

In order to evaluate the causal relationship between alcoholism and labor market you are asked to apply matching methods for 1994. The data set **alcohol_PSM1994.dta** has been prepared for conducting this exercise. The goal of this exercise is to estimate the propensity score (PS) of being an alcohol abuser, use the PS to construct a counterfactual group of non-abusers and estimate the effect of being an alcohol abuser on average hourly wage (`hrwage`). The variable `alcoholic` takes value equal to 1 for those identified as alcohol abusers, and 0 otherwise.

(a) (1 point) Plot the density curves of hourly wage for the group of alcohol abusers and for the group of non-alcohol abusers. Discuss the graph.(*Max: 700 characters*) (*Hint: the graph should be meaningful, if it is very skewed drop the outliers, but comment on the existence of outliers*)

(b) (1.5 points) Provide sample means for the group of alcohol abusers and for the group of non-alcohol abusers. How similar are the two groups? Test whether these differences are statistically significant, and discuss the implications of your findings for the identification strategy. (*Max: 900 characters*)

(c) (2 points) Select variables on which to match: especially those related to treatment receipt (e.g., alochol use) and the outcomes. Run a model to estimate a propensity score as the predicted probability of being an alcohol abuser. Present your results in a table and on one plot overlay the density curves of the propensity scores for alcoholic abusers and non-abusers. Discuss your results and the quality of your model based on the predicted propensity score and the common support. (*Max: 2,500 characters*)

(d) (2 points) Use propensity score matching to estimate the effect of being an alcoholic abuser on hourly wages. Discuss the matching method, and why do you think it is the best matching method for this particular case. Discuss your results, and whether your estimates could be interpreted as causal. (*Max: 1200 characters, without counting the tables.*)

(e) (1.5 points) Use propensity score weighting to estimate the effect of being an alcoholic abuser on hourly wages. Discuss the results. (*Max: 900 characters, without counting the tables.*)

Annex: Variables available

Labor market variables

- wgsal – total wage and salary income in the past calendar year, in dollars
- hrswrk – total number of hours worked in the past calendar year
- wkswrk – total number of weeks worked in the past calendar year
- wksue – total number of weeks spent unemployed in the past calendar year
- wksolf – total number of weeks spent out of the labor force in the past calendar year
- empst – a categorical variable indicating the individual's current employment status. It is defined as follows:
  1 = Employed
  2 = Unemployed
  3 = Out Of Labor Force
- numjob – total number of jobs the individual has had in their lifetime

Alcohol consumption variables

- drinkev – a dummy variable = 1 if the individual has ever had a drink, = 0 otherwise
- drnkmo – a dummy variable = 1 if the individual has had a drink in the last month, 0 otherwise
- drnk6m – a categorical variable indicating the number of times in the past month the individual has had 6 or more drinks in one sitting. It is defined as follows:
  0 = Never
  1 = Once
  2 = 2 Or 3 Times
  3 = 4 Or 5 Times
  4 = 6 Or 7 Times
  5 = 8 Or 9 Times
  6 = 10 Or More Times
- days – the number of days in the last month the individual has had at least 1 drink per day
- perday – the average number of drinks per day on a day when the individual drinks (this is 0 if the individual doesn't drink alcohol)
- gtint – a categorical variable that answers the question of whether the individual has ever drunk more than intended. It is defined as follows:
  0 = Don't Drink
  1 = Happened 3+ Times In Past Year
  2 = Happened 2 Times In Past Year
  3 = Happened 1 Time In Past Year
  4 = Happened In Lifetime Other Than Past Year
  5 = Never Happened

Demographic variables (which may or may not be useful for the analysis)

- sex – a categorical variable = 1 if the individual is a man and =2 if a woman
- race – a categorical variable = 1 if the individual is Hispanic, =2 if the individual is Black and =3 if the individual is White
- south14 – a dummy variable = 1 if the individual lived in the south when they were 14 years old
- wdad14 – a dummy variable =1 if the individual lived with their father when they were 14
- wmom14 – a dummy variable = 1 if the individual lived with their mother when they were 14
- dadwork – a dummy variable = 1 if the individual's father worked when they were 14. This is set to 0 if they didn't know, which often happens if they didn't live with dad, so this variable should always be used along with wdad14
- momwork – a dummy variable = 1 if the individual's mother worked when they were 14. This is set to 0 if they didn't know, which often happens if they didn't live with mom, so this variable should always be used along with wmom14
- dadhgc – the number of years of education the individual's father has. This is set to 0 if they didn't know, which often happens if they didn't live with dad, so this variable should always be used along with wdad14
- momhgc – the number of years of education the individual's mother has. This is set to 0 if they didn't know, which often happens if they didn't live with mom, so this variable should always be used along with wmom14
- numsib – the number of siblings the individual has
- hvsib – a dummy variable =1 if the individual has a sibling in the data set
- sibid1 – the value of the variable id for the individual's sibling in the data set. This is missing if there is no sibling in the data set
- religkid – a categorical variable reporting what religion the individual was at age 14. It is defined as follows:
  0 = None, No Religion
  1 = Protestant, unspecified
  2 = Baptist
  3 = Episcopalian
  4 = Lutheran
  5 = Methodist
  6 = Presbyterian
  7 = Roman Catholic
  8 = Jewish
  9 = Other
- relignow – a categorical variable reporting what religion the individual is now. It is defined the same as religkid
- afqtrev – the percentile in which the individual scored on an intelligence test given in 1979
- height – the individual's height, measured in inches

- weight – the individual's weight, measured in pounds
- health – a dummy variable = 1 if the individual has a health problem that limits the amount or kind of work that can be done
- higrad – the number of years of education the individual has completed
- numkid – the number of children the individual has
- urbrur – a dummy variable = 1 if the individual lives in an urban area
- famsz – the number of people in the individual's family (i.e. self, plus spouse, plus dependent children)
- faminc – net income for the family in the past year, measured in dollars
- povst – a dummy variable =1 if the individual's family was below the poverty line last year
- region – a categorical variable for the region the individual lives in. It is defined as follows:
  1 = Northeast
  2 = North Central
  3 = South
  4 = West
- urate – the unemployment rate for the local labor market of the individual. (note: this variable might look a little funny to you because it was created as the midpoint of a range so that the place the individual lives could not be identified)
- marst – a categorical variable for the individual's marital status. It is defined as follows:
  1 = never married
  2 = married with a spouse present
  3 = other
- delin – frequency count of the number of illegal activities, excluding underage alcohol consumption that the individual reported engaging in during 1980.
- depression – measure of depression, is the total score on a 7-item version of the Center for Epidemiological Studies Depression Scale (CES-D) measured in 1994.
- alchfirstage – age at which the individual reports they first had a drink.
- drugusage – a binary indicator that the individual had a positive response to at least one of two questions pertaining to frequency of marijuana or cocaine in their lifetime.

Important: The set of variables with prefix DMISS in the dataset alcohol_PS are binary indicator variables that equal one if the corresponding variable is missing and zero otherwise. You should include these variables along with the variable they refer to when estimating the propensity score. For example, if you want to use alchfirstage in your model, you should also include DMISS_alchfirstage.