**Data:**

The following dataset is based on the study by Yakovlev (2018) on the alcohol consumption habits of Russian men.[[1]](#footnote-1) This study investigates both the determinants and the consequences of alcohol abuse. The dataset **Exam\_Data.dta** to be used here, contains information on a sample of Russian men, aged between 18 and 65 years old, over two decades (1994-2014). It includes information on individual’s age, income, education, marital status, drinking and smoking behavior, among other details. The table below provides a brief description of the variables available.

The dataset **Exam\_Data\_y2010.dta** provides similar information but for the **year 2010 only**.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| idind | unique longitudinal person ID |
| curwrk | 1 if currently employed, 0 otherwise |
| stroke | 1 if ever diagnosed with a stroke, 0 if never |
| smokes | 1 if the individual smokes, 0 otherwise |
| hdrinker | 1 if individual belongs to top 25% by alcohol intake, 0 otherwise |
| age | individual's age (in years) |
| year | year of reference |
| logvodka | vodka consumption (log) |
| logprice\_vodka\_real | price of vodka in logs and real prices |
| married | 1 if married or lives with spouse, 0 otherwise |
| logfamily\_income | household income (log) |
| logincome | individual income (log) |
| college | 1 if has college degree, 0 otherwise |
| age\_2 | age squared |
| died | 1 if dies before next round, 0 if survives |
| muslim | muslim religion |
| hdrinker50 | 1 if individual belongs to top 50% by alcohol intake, 0 otherwise |
| y2011 | 1 if year >= 2011, 0 if year < 2011 |
| logvodka\_next | vodka consumption in the subsequent year (log) |

**Before we do the 2019 exam question, let’s ‘feel’ STATA.**

1. Open a new folder, and download from Canvas the following files: Exam\_Data.dta Exam\_Data\_y2010.dta and Intro\_Session\_AER2022.do.
2. Open Intro\_Session\_AER2022.do.
3. Change directory in line 1 to you folder name

Let’s have a look at our database, and understand our variables.

1. A little exercise:
   1. We have the age variable. Let’s create a dummy variable for all individuals above the age of 40.
   2. Add a label to the variable.
2. Assign 1 to all individuals (in age\_40), also for those under 40.

**Question 1 A and B (Predicting heavy drinking)**

The dataset includes the variable *hdrinker*, which is a dummy variable equal to 1 for heavy drinkers (those who belong to the yearly top quartile by total alcohol intake), and 0 for all the other men in the sample. In this question, you will analyze some determinants of heavy drinking among Russian men.

1. Many presume that heavy drinkers are less educated on average. What is the percentage of college graduates in the subsamples of heavy drinkers versus others? And what is the pairwise correlation between heavy drinking (*hdrinker*) and college education (*college*)? Do you confirm the aforementioned assumption based on your results?
2. Using a t-test, formally test the null hypothesis that heavy drinkers are equally represented among college and non-college graduates. What is the alternative hypothesis you are testing against? What do you conclude based on your results?

1. Yakovlev, E. (2018), “Demand for alcohol consumption in Russia and its implication for mortality”, *American Economic Journal: Applied Economics*, 10(1): 106-149. [↑](#footnote-ref-1)