Assignment 2

Income distribution (ECON 473)
Winter 2026

due Feb 26 at 23.59pm via *mycourses*

Data. Work with data from the Canadian Income Survey (CIS) and either the Survey of Labour and Income Dynamics (SLID) or the Survey of Consumer Finances (SCF). I will assign each student two years from two data sets. For the data assigned to you, see the separate file posted on *mycourses*.

You can access and download it from Odesi, at `http://odesi1.scholarsportal.info/webview/`. To do so, you need to be on the McGill network, either at the university or connected via VPN. Once on Odesi, I could download the dataset for each year. For import into Stata, it is convenient to download the comma separated version. (Note that the file is large.) In the SCF/SLID, use the files on individuals (not families).

Tasks. Use these data to perform an analysis in the spirit of HPV and AKK. You should be able to answer all questions from what we have covered in class and in the Stata sessions. Keep discussions brief. Create tables and figures that are easy to read.

This is an individual assignment. Upload your solution to *mycourses*. Hand in the code file you use for your analysis as well as a file containing outputs, like tables, figures, and replies to questions (submit a *pdf* version of this). As a default, I assume that you will use Stata. If you have access to a different statistical package and prefer to use that, you may do so. If you use Stata, **you need to provide your do file**, and I recommend that you also provide a log file. If you use different software, you need to provide your code for that.

Some hints: In principle, one should use sampling weights in such an analysis, so do so where possible. Be careful with missing data, which is coded in special ways (values like 99999996 or 999.6). Depending on the format in which you download the data, these may already be transformed into other missing value codes. Consult the Data Dictionary for details.

Operationally, it is easiest if you do your analysis data file by data file, or if you merge data from the same survey and do your analysis survey by survey. If you find it easier to plot in Excel, you can compute statistics in Stata and then generate a graph in Excel.

Questions. Answer each question for all the years/data sets assigned to you. (Two years from two data sets for each student.)

1. *Variables.* You will need to use or construct a variable for hourly wages, and use variables for demographics. In particular, you will need to use information on Total Income, After-tax income, Wages and Salaries, Hours worked, Education, Age, and Gender, as well as the survey weights. List the codes of these variables for your two data sets. Use variables for individuals, not families.

   Note: Since we will be focussing on wages in this assignment, you do not need to consider business income (differently from the first assignment).

2. *Building hourly wages.*

   (a) Download the consumer price index for the relevant years from Statistics Canada. Plot its evolution.

   (b) Compute hourly wages for each individual. Deflate the hourly wage using the CPI (use Statscan's annual CPI). To make wage statistics more informative for us, use 2020 as the base year.

   Hint: maybe the most straightforward way of doing this is to do it year by year in your do file, with one line of code per year.

3. *Sample selection.* Drop from your sample individuals with missing information, zero hours, or a wage below 50% of the minimum wage. Next, keep only individuals between the ages of 25 and 60 (inclusive), who work more than 260 hours. Report in a table like Table 1 of HPV how many observations you have (not using weights) at each stage. Use this selected sample for the remainder of the assignment.

   Hint: use the yearly minimum wages I provide on *mycourses* as an Excel file. In principle, you would need to be more precise here, since minimum wages are set at the provincial level in Canada. But to keep things simple, you can use the provided numbers for all individuals. (The file contains the unweighted average of the prevailing minimum wage in Alberta, BC, Ontario and Québec, in *current* dollars.)

   Hint 2: In some surveys, age is not available, only a variable for year of birth, in 5-year intervals. I ask you to use a sample of individuals aged 25 to 60. I suggest that if most individuals in an age group fall in the sample, you keep that age group. For example, those born in 1968 are 25 years old in 1993. You could then include the group of all of those born in 1965-1969.

4. *The evolution of inequality.*

   (a) Plot the time series of the log variance of total income, after-tax income, and wages and salaries (similar to HPV, F11). Plot, in a different graph, the time series of the Gini coefficient of these variables.

   (b) Plot, in a single graph, the time series of the log variances of annual wages, hourly wages, and annual hours, as well as the correlation of log hours and log wages. Repeat this for men, and for women.

   (c) Plot the time series of percentiles 5, 50 and 95 of the distribution of hourly wages.

   A "time series" here consists of values for the four years you are analyzing.

5. *Composition-adjusted wages.* Here, you will compute composition-adjusted wages, as in AKK.

   (a) Generate a variable `potential_experience` as age minus years of schooling minus 6. You may need to create a new variable *years of schooling* based on a categorical schooling variable for this.

   (b) Create a variable `group` distinguishing gender $\times$ education $\times$ potential experience groups. How many groups do you have?

Hint: use the `egen ... group` function.

Hint 2: You need to make sure that your groups are identical across your different data sources.

(c) Compute the share of the sample (using weights!) in each group in each year. (Show only the code for this.)

(d) Compute the mean wage in each group and year. (Show only the code for this.)

(e) For each year, compute the composition-corrected mean wage for (a) the entire population, (b) women, (c) men, (d) university graduates, and (e) high-school graduates, as discussed in the lecture on AKK. Report naive and composition-corrected wages for each of these five broad groups, for each of your years.

(f) Discuss how composition-adjusted wage changes over the sample period compare to naive ones for the five broad groups.

Hint: There are several ways of addressing this question. Probably the most elegant option is to use the `collapse` command in Stata, at the group level. This was not covered in the Stata intro. If you try this, be aware that `collapse` overwrites your data set in memory, so you need to make sure that you save it first. Another option is to report the group-level information using `tab` and `tab ..., sum`, copy it to Excel (note the *Copy table* function in Stata), and process it there.