ECON 4225 Fall 2025

Homework 1

Deadline. Solutions have to be submitted to Canvas **before the beginning of class (10:15AM) on October 14**. You must submit **two separate files**: (1) A PDF document containing your answers to the questions, including all tables and figures you produce. (2) A separate file with the codes to generate your results.

Grading. 80% of your grade for the homework is based on the approach taken and the numerical results. The remaining 20% is awarded for a clear presentation of your results, i.e. clearly label all figures and tables and discuss all findings.

Late Submissions. Submissions received after the deadline without a valid excuse will receive a deduction of points from the final score, 10 percentage points for submissions up to 24 hours after the deadline, 20 percentage points up to 48 hours after the deadline. Submissions later than 48 hours after the deadline receive a score of zero.

Collaboration. I encourage you to discuss the homework with your classmates and to work on it together. However, each of you has to submit their own solution. I.e. everyone has to write up their answers independently, you must not submit a joint solution or the same solution / same codes.

Data. This homework assignment uses the dataset *PSID_clean.csv*, which you should download from Canvas before you start. The file *data_codebook.pdf*, also available on Canvas, provides a description for each of the variables in the dataset. The dataset is constructed from the 2019 wave of the Panel Study of Income Dynamics (PSID). For the purpose of this homework you can assume that the data covers the entire population, i.e. you do not have to apply any sample weights during your analysis. Whenever you are asked to compare your results to the Survey of Consumer Finances (SCF), you will find the relevant information from the SCF in the lecture slides. I.e. you do not need to compute any results from the SCF yourself.

Part 1: Overview of the Dataset

We begin by looking at some summary statistics of the dataset. This helps us to understand better what we are working with. Please answer the following questions:

1. How many households are in the dataset? How many variables are there?
2. What is the average age of the household head in the dataset? What is the highest and lowest age?
3. What is the average income of households? How does this compare to the average income in the Survey of Consumer Finances (SCF)? What could be a reason why the income in the PSID is different from the SCF?
4. What is the highest and lowest household income in the dataset? Why can the lowest household income be negative?
5. What is the average age of the spouse? [Hint: To answer this question, make sure to restrict your dataset to households where a spouse is present.]
6. Plot a histogram of the number of members per household. What is the share of households with 5 or more members? What is the share of households with only one member? Compare it to the histogram from the lecture slides.

Part 2: Income Distribution

Next we look at the distribution of household income. Answer the following questions:

1. Plot the histogram of household income, using a bin width of $5,000. Make sure to comment on your findings.
2. Plot the Lorenz curve of household income and comment on your findings.
3. Compute the coefficient of variation and compare it to the SCF.
4. Compute income adjusted for the number of household heads, i.e. for all households with a spouse present divide total income by two. Plot the Lorenz curve of this adjusted income in the same graph as the Lorenz curve for income. What do you conclude from this comparision?
5. Compute the 30th, 50th, 90th, and 99th percentile of income. Compute the 90-30, 90-50, 30-10 and 99-50 ratios. Make sure to interpret the numbers that you find. Compare your results to the SCF.
6. Compute the share of income received by each quintile of the income distribution. Compute also the average income of each quintile. Again, interpret your findings and compare them to the SCF results from the lecture.
7. Compute the average income and share of income received for the top 1% of the income distribution. How does this compare to the SCF?

Part 3: Labor Income

We conclude with the distribution of earnings. Please answer the following questions:

1. Create a variable that is the total labor income (earnings) of the household. Compute the average earnings and share of total earnings for each quintile of the earnings distribution. [Hint: Total labor earnings is the sum of the labor earnings of both spouses.]

2. Compare the distribution of household earnings to the distribution of income. Compare the share of each quintile of the respective distributions as well as the coefficient of variation. Plot the Lorenz curve of labor earnings and of total household income in the same graph. What do you conclude on the relative inequality in earnings and total income?

3. Create a variable for the share of labor earnings in total household income. What is the average share of labor earnings in the dataset?

4. Compute the average share of labor earnings by quintile of the income distribution. Compute also the share of labor earnings for the top 1% of households. Compare both to the SCF. What could be a reason why your results differ from the SCF?

5. For this subquestion and the next subquestion, restrict the sample to households in which the head has positive weeks worked and positive labor earnings. Compute the weekly wage of the household head. Compute the variance of the log of earnings, the log of wages, and the log of weeks worked of the household head. What do you conclude for the relative contribution of hours worked and wages to inequality in labor earnings? Is there any caveat to this way of computing the contribution of hours worked to earnings inequality? [Hint: Remember that Labor earnings = weekly wage $\times$ weeks worked.]

6. Run a regression of log wages of the household head on age, age^2, education and occupation of the head. Compute the residuals from this regression. Compute the variance of the residuals. What share of inequality in wages do age, education, and occupation explain together? [Hint: Make sure to classify education and occupation as categorical variables before running the regression.]