> **ECO483 Winter 2025** Health and Economic Inequality, *University of Toronto*
>
> Prof. Michael Stepner

In-Class Empirical Lab

Now that you've completed the *At-Home Setup* assignment, you have a template for how to write code in a way that anyone can run it. Even an automated script can run it and reproduce your results without any manual interactions.

For the in-class lab, we're now going to dive deeper into data analysis. This assignment will ask you to manipulate datasets, generate new variables, run OLS regressions, and generate figures.

During this assignment:

- 🛜 You are allowed to get help from the internet.
- 🤖 You are allowed to get help from AI tools.
- 🙋 You are allowed to get help from the professor and the TA.
- ⛔ You are **not allowed** to copy and paste code from your classmates or code that you do not understand provided by an AI.
  - Your code may be compared to other students' code.
  - If asked, you must be able to verbally explain what your code is doing and how you developed it.
- 📝 You are **required** to respond to the *Empirical Lab: Reflection on AI* survey on Quercus within 24 hours after the empirical lab ends. (You can complete the survey during the in-class lab or after class.)
  - It is an academic offense to be dishonest in your description of how you used AI tools to assist with your work. To be clear: in this class, you are allowed to use AI tools so long as you accurately describe how they were used.
  - If you did not use AI, you can check a box to say so and skip the remaining questions, but **you must still complete the survey**.
  - Late submissions will reduce your grade on the empirical lab.

🛠️ Advice and Troubleshooting

I recommend completing this assignment using the same development environment you used for the *At-Home Setup*. (Although you are not required to.)

1. Edit and run your code interactively in Github Codespaces.
2. Then commit your code and check your results by following the instructions in this guide.

Useful References

The same example code from the *At-Home Setup* is still available for your reference in this repository. There is an `example_notebook` and `example_script` for Python, R and Stata (so 6 examples in total). You can also refer back to the code you submitted in the *At-Home Setup* assignment, if anything there is helpful to you.

For more detailed guidance and troubleshooting advice for common issues, click the link corresponding to your development environment in the table below:

	Python	R	Stata
☁️ **Online**	Guide	Guide	Guide

➡️ Submitting your work

- **You must save your code in Github, in a file named `submission`** located in the root directory of the repository (not in a subfolder).
    - This will be `submission.ipynb` if you're submitting a Jupyter Notebook.
    - This will be `submission.Rmd` if you're submitting an R Markdown document.
    - This will be `submission.py` , `submission.r` or `submission.do` if you're submitting a script, depending on the programming language you chose.
- Your code must output all your results to the `results` folder.
    - Your numerical estimates must be output to a YAML file named `results/lab.yaml` . Each time a question asks you to output a number, it is marked with (🔢 → `name_of_estimate`).
    - Your figures must be output to images saved as `.png` files in the `results` folder. Each time a question asks you to output a figure, it is marked with (📊 → `results/filename_of_figure.png`)
    - **These files in the `results` folder will not be committed to the Github repository.** You will NOT see them in the Github website within the `results` folder.
        - They will be visible in your Github Codespace after running your code.
        - When you submit your *code* by committing it to Github, your *results* will be automatically recalculated by running your code. You can check your results to see if your code ran successfully and if you got the right answers.

Tasks to Complete

These tasks are approximately ordered in increasing difficulty. They get harder as you go.

- I recommend solving the questions roughly in order and seeking resources to help you answer them if you're not sure how.
- Don't worry if you can't finish every question.

If you're waiting a few minutes at the start of the lab for your Github Codespace to load, I recommend carefully reading all the questions. You can start to think about how to answer them while you wait.

1. Load Table 6 from the Health Inequality Project's Data Tables.

    - You can either load it as a local filed stored in the repository or directly from the web.

        - To load it as a local file, I recommend using the relative path: `data/raw/health_ineq_online_table_6.csv` . (See examples of how to use relative paths in the example code.)
        - To load it directly from the web, I recommend using this URL: `https://github.com/UofT-Econ-DataAnalytics/files/releases/download/files/health_ineq_online_table_6.csv`

    - If you click through to the Health Inequality Project's data page, you can open the README for Table 6 to see a description of the data.

    - This dataset contains estimates of the life expectancy of American men and women in 2001 to 2014, separately for each "Commuting Zone" and each income quartile. Commuting zones are a geographic area, similar in scope to a city—but covering the entire country.

2. **[1 point]** Within these geographic areas, let's examine how strongly the life expectancy of different demographic groups is correlated with one another.

    - Create a scatterplot showing life expectancy for American men in the bottom income quartile (`le_agg_q1_m`) on the y-axis, and life expectancy for women in the bottom income quartile (`le_agg_q1_f`) on the x-axis.

        - Output this graph as an image using the PNG file format: (📊 → `results/scatter_bottom_quartile_men_vs_women.png`)

    - Create a scatterplot showing life expectancy for American women in the top income quartile (`le_agg_q4_f`) on the y-axis, and life expectancy for women in the bottom income quartile (`le_agg_q1_f`) on the x-axis.

        - Output this graph as an image using the PNG file format: (📊 → `results/scatter_top_vs_bottom_quartile_women.png`)

3. **[2 points]** We'll now calculate some numbers that quantify the relationships we plotted, by running OLS regressions.

    - Looking at the scale on the axes of the first scatterplot, we can see that women live longer on average than men. Comparing women to men in the bottom income quartile, and computing an unweighted mean over the commuting zones in this sample...how many years longer do women live than men? (🔢 → `le_diff_women_men`)

    - It is clear from the first scatterplot that life expectancy for men and women in the bottom income quartile living in the same geographic area are strongly correlated. Regress `le_agg_q1_m` on `le_agg_q1_f` . What is the share of variation in life expectancy for low-income men that can be explained by the life expectancy of low-income women in the same area? In other words, what is the R-squared of this regression? (🔢 → `r2_low_income_men`)

    - The second scatterplot already showed that life expectancy of low-income women does not strongly predict the life expectancy of high-income women in the same area. So let's add more predictors: we'll use the life expectancy of high-income men and low-income men as well as low-income women.
      Regress `le_agg_q4_f` on `le_agg_q4_m`, `le_agg_q1_m` and `le_agg_q1_f`. What is the R-squared of this regression? (🔢 → `r2_high_income_women_multiple_predictors`)

    There are two main reasons why this last regression does not explain a large share of the geographic variation in the life expectancy of high-income women. First, people with high incomes have more similar life expectancies across areas—life expectancy is more sensitive to geographic location for people with low incomes. But a second reason is that life expectancy estimates for people with higher incomes are noisier—rich people die less often, so we have less precise life expectancy estimates for them.

    - To reduce the influence of noise, let's restrict our attention to areas with larger populations: specifically, populations greater than 150,000 people (`pop2000 > 150000`). Run the same OLS regression of `le_agg_q4_f` on `le_agg_q4_m`, `le_agg_q1_m` and `le_agg_q1_f` after applying this sample restriction. What is the R-squared of this regression? (🔢 → `r2_high_income_women_multiple_predictors_large_population`)

4. Let's now examine trends over time in life expectancy. Load Table 9 from the Health Inequality Project's Data Tables.

    - You can either load it as a local filed stored in the repository or directly from the web.

        - To load it as a local file, I recommend using the relative path: `data/raw/health_ineq_online_table_9.csv`. (See examples of how to use relative paths in the example code.)
        - To load it directly from the web, I recommend using this URL: `https://github.com/UofT-Econ-DataAnalytics/files/releases/download/files/health_ineq_online_table_9.csv`

    - This dataset is similar to Table 6, but contains separate observations for life expectancy for each year from 2001 to 2014.

5. **[1 point]** We'll measure the trends in life expectancy growth over time, comparing women in the bottom income quartile to those in the top income quartile.

    - How much did average life expectancy for women in the bottom income quartile increase per year? Regress `le_agg_q1_f` on `year`, and output the slope of this OLS regression. (🔢 → `le_trend_low_income_women`)

    - How much did average life expectancy for women in the top income quartile increase per year? Regress `le_agg_q4_f` on `year`, and output the slope of this OLS regression. (🔢 → `le_trend_high_income_women`)

    - You ought to find that life expectancy was growing roughly twice as fast for women in the top income quartile than the bottom income quartile.

6. **[1 point]** Economists are often interested in whether trends are driving convergence or divergence in an outcome: are the groups that started further behind catching up? In the previous question, we observed *divergence* since high-income women started with higher life expectancies and had faster-growing life expectancies. In this question, we'll examine whether the life expectancies of low-income women are converging across areas.

   - Calculate the median life expectancy of low-income women in 2001, which is the first year in our sample. In other words, we want to calculate the median of `le_agg_q1_f` in `year==2001`.

   - Create an indicator variable which is equal to 1 if the Commuting Zone had above-median life expectancy for low-income women in 2001, and equal to 0 otherwise. Note that the dataset has an observation for every commuting zone in every year, but this indicator variable should be the same for every year in any given commuting zone.

   - In commuting zones where life expectancy in 2001 was below median, how much did average life expectancy for women in the bottom income quartile increase per year? Restricting to this sample, regress `le_agg_q1_f` on `year` and output the slope of this OLS regression. (🔢 → `le_trend_low_income_women_below_median`)

   - In commuting zones where life expectancy in 2001 was above median, how much did average life expectancy for women in the bottom income quartile increase per year? Restricting to this sample, regress `le_agg_q1_f` on `year` and output the slope of this OLS regression. (🔢 → `le_trend_low_income_women_above_median`)

   - You ought to find that life expectancy was growing roughly twice as fast for women in the bottom income quartile living in areas with below median life expectancy in 2001. We therefore see *convergence* in the geographic distribution of life expectancy for low-income women.

7. We'll now return to examining the average life expectancy in each commuting zone during 2001 to 2014. But we'll combine that information with information on the characteristics of those commuting zones.

   - Join/merge the data from Table 6 and Table 10 in the Health Inequality Project's Data Tables.

   - Both data tables have the variable `cz` as a unique identifier, which can be used to join them together. Note that you'll find there are 595 Commuting Zones that have both life expectancy and characteristics data, but a further 146 Commuting Zones only have data on their characteristics without any data on life expectancies.

   - Both data tables are available either as a local filed stored in the repository or directly from the web.

     - For example, to load table 10 as a local file, I recommend using the relative path: `data/raw/health_ineq_online_table_10.csv`. (See examples of how to use relative paths in the example code.)
     - To load table 10 directly from the web, I recommend using this URL: `https://github.com/UofT-Econ-DataAnalytics/files/releases/download/files/health_ineq_online_table_10.csv`

8. **[1.5 points]** It's time to run some regressions describing the characteristics of places with high life expectancy for women with incomes in the bottom income quartile. These are "naïve regressions" in the terminology of Deryugina & Molitor's (2021) *Journal of Economic Perspectives* paper from the Class 3 readings.

    - We'll focus our attention on three variables: the percentage of people without health insurance in 2010 (`puninsured2010`), the extent of local income inequality as measured by the Gini coefficient (`gini99`) and the amount of local government expenditures (`subcty_exp_pc`).

    - One complication in measuring how each of these variables is associated with life expectancy is that each is measured in different units: percentages, Gini index and dollars. A common strategy is to re-normalize each variable in standard deviation units, so that we can ask "how much does life expectancy change when this variable increases by 1 standard deviation?" To implement this strategy, create three new variables: one for each variable, but measuring the Z score instead of its raw value.

        - To calculate the Z score, you must subtract the mean and divide this difference by the standard deviation. Therefore, an observation with a Z score equal to 1 has a raw value equal to the mean plus 1 standard deviation.

    - On average, how does the average life expectancy of women in the bottom income quartile change if the percentage of people without health insurance increases by 1 standard deviation? (🔢 → `uninsured_1sd_low_income_women`) Note that the direction of this effect is opposite from what our intuition would tell us: areas with higher uninsurance rates have higher life expectancies.

    - On average, how does the average life expectancy of women in the bottom income quartile change if the Gini coefficient increases by 1 standard deviation? (🔢 → `gini_1sd_low_income_women`)

    - On average, how does the average life expectancy of women in the bottom income quartile change if local government expenditures increase by 1 standard deviation? (🔢 → `expenditure_1sd_low_income_women`)

9. **[0.5 points]** We'll plot the data underlying these "naïve regressions" on a single scatterplot.

    - Restrict your sample to areas with a population greater than 600,000 (`pop2000 > 600000`). This roughly corresponds to the 100 commuting zones with the largest populations. This restriction will reduce the number of dots on our scatterplot, and also focus our attention on areas with more precise estimates of life expectancy—i.e. less noisy data.

    - After performing this sample restriction, create a scatterplot with life expectancy for low-income women on the x-axis (`le_agg_q1_f`). On the y-axis, plot in three separate colours the Z score values for the percentage of people without health insurance, the Gini coefficient and the local government expenditures. (📊 → `results/naive_regressions_low_income_women.png`)

    - Just looking at the scatterplot, you can see that none of these relationships are impressively strong or predictive.