Final project for Analysis of Large Scale Data
2025

For the final project, you will replicate results from an existing (published) article and extend the analysis in at least one significant way. Students will proceed in two steps:

1. **Submit a 1-page proposal** identifying the article to replicate and the proposed extension (details below). This is due Nov 5, 2025, and it will be worth 5 of the 35 total points.

2. **Present their replication on the last day of class** (Dec 10, 2025). All files for the final project must be submitted no later than 1 hour before the start of the last class. The presentation should discuss your efforts to replicate the article, how you chose to extend the analysis, and any discrepancies you found. Be sure to show both your results and the original article results side by side in the same table so one doesn't need to have the original article in hand.

A few notes:

- The project points are divided as follows:
    - 1-page proposal: 5 points
    - Presentation delivery (self-graded): 10 points
    - Content of the presentation, code, and documentation of your process: 20 points
- Everything in your slides should be presented professionally, including tables and graphs. We will have a lecture on presenting graphs nicely.
- Many or most of you will not be able to replicate the article results. This is OK and expected. What I'm looking for is what you do to try to replicate. If something doesn't work, then try something else. Keep track of the things you try and what you learn. Your presentation is about what you did to try to replicate the article, how you chose to extend the analysis, and what you learned.
- Your presentation (10 of the 35 points) will be self-graded. You will provide a brief assessment of your strengths and areas for growth. See syllabus for more information.

Please read this document carefully. You will lose points for not following instructions

Submit all materials electronically via courseworks. The presentation must be in PDF format; do not submit a PPT document. In addition to the presentation, also provide a PDF of the original article and all do-files created. You do not need to submit the data sets used, but must have them prepared to share in case I ask.

You will work in pairs. Clearly label all files submitted via courseworks as follows. Combine your uni with your partner's uni in alphabetical order followed by a brief description of the file. For example, if my partner is Dolly Parton (dp1001), submitted files will be labeled as follows:

dp1001_kld2128 _proposal.pdf → 1 page proposal
dp1001_kld2128_article.pdf → article used for replication
1st submission (Nov 5)

dp1001_kld2128_final.pdf → final presentation
dp1001_kld2128_proc.do → do-file to process data
dp1001_kld2128_analysis.do → do-file to perform statistical analyses
2nd submission (Dec 10)

The project should roughly include the following steps:

Step 1. Choose an article to replicate, preferably on a topic you are interested in.

- As a general rule, remember that you are going to be replicating this article. Does it look replicable to you? Choose one that uses an easily (publicly) accessible data set, clearly describes the methodology, and uses analytical techniques you are familiar with.
- **There must be at least one regression in the paper.**
- Be careful when choosing articles that merge multiple data sets, use non-public versions of data sets, or perform advanced statistical analyses. It is okay to choose such articles, but they may be overly burdensome for you to work with.
- Read the article carefully, making sure you understand the steps they've taken. If it says they estimate 'cox proportional hazards models' and you don't know what that is, find another article. In general, when they say something like 'account for complex survey design' this simply means they use sample weights (which we cover in the course).
- Do not worry if the original authors used something other than Stata (it doesn't matter). Also, don't worry if they used a public data set that we did not directly cover in class (e.g., YRBS, PULSE, or CPS) -- this is OK, as long as you are comfortable learning about that data set on your own.
- **Tip:** Some of the articles assigned as optional readings, or other articles I bring up in the lecture slides, might be good candidates for replication. If you are stuck, try starting with the slides and syllabus to find ideas.

Step 2. Write a 1-page proposal for your project (due Nov 5, 2025).

- **On Nov 5,** you will turn in a 1-page document that:
  - summarizes the article you are replicating;
  - explains how you plan extend the analyses;
  - makes an argument for why this is a good article for this project;
  - and lists potential challenges you expect to face or limitations you might run into.
- **You do not need to have started any analyses yet.** I just want to know that you have read the article thoroughly, thought about why it is a good fit for the project, and are prepared for any potential challenges.
- **Please also include a PDF of the original article.**
- I will quickly scan the original article you propose, but it is your responsibility to make sure it is appropriate for your project and explain this in your 1-page proposal.
- It is not permissible to change articles after we have agreed upon one (i.e., after Nov 5). Choose your article carefully.

Step 3. Access the same data they used, and clean and process the data exactly as they did to the best of your ability.

- Compare your summary statistics to theirs to see how well you can replicate their variable definitions and study sample size.
- Note any discrepancies, and try to correct them to the best of your ability.
- If discrepancies still exist, describe the steps you have taken to eliminate them or identify the source of the discrepancies.

Step 4. Using the processed data, perform the same statistical analyses they performed.

- You do not need to replicate all results – only the main regression results.
- Are your results the same, very similar, or substantively different?
- Are there any variable definitions or sample selection differences from the previous step that may be driving the differences?
- Go back to Step 3 if necessary to fix any discrepancies. Document your process.

Step 5. Extend the analysis in an important way.

- For example:
  - Perhaps they omitted a variable with many missing values, but you feel it's an important confounder, so you impute the missing values instead.
  - Perhaps you hypothesize that the estimates may differ by race/ethnicity of the individuals, so you perform the analysis stratifying by race/ethnicity.
  - Perhaps you decide to add additional years of data to see if the results hold over time.
  - Perhaps you decide to modify the inclusion/exclusion criteria of the study because you disagree with the original authors' decision.
- **<u>These are obviously only general suggestions that need to be tailored to the article you choose.</u>**
- You need to have clear logic driving your choice. The extension should be meaningful and interesting – not a change just for the sake of making a change. You will need to communicate this logic in your presentation.

Step 6. Prepare a presentation about your process and findings (7 minutes total).

- Provide a brief overview of the study you will replicate: why it's of interest (1-2 bullet points at most), what data sources were used, what methodology was used, and what the main findings were. Be very brief (~1-2 slides, preferably 1). Your presentation should mostly be about your results, not theirs.
- Show us your results compared to the original results. Were you able to replicate their summary statistics and regression results? Why or why not? What steps did you take to try to fix any discrepancies?
- Tell us about how you chose to extend the analysis and what you found when you did that. Why did you choose to do this extension or make this change? What new conclusions did you make?
- Throughout the presentation, the focus should be on **<u>your process of attempting to replicate the paper and what you learned</u>**. We care less about the actual findings (e.g., "smoking is associated with lung cancer") than your process of working with the data, resolving discrepancies, producing interesting results, and making decisions when handling complex data and analyses.