Economics 4161/8803: Machine Learning for Economics

Georgia Institute of Technology

Spring 2025

Project Description

Instructions

- The purpose of the assignment is to have you demonstrate that you can apply the techniques learned in the course to an important real-world setting in order to obtain answers to economic questions of interest.
- This document contains an outline of the projects topic and the expected form of the final report.
- You will work in groups. You should form groups of 4 or 5 people as soon as possible, and fill out the following Google form with your group information by February 14: https://forms.gle/Ni5VyjpTvoycRoAY7
- Your group will submit a final report and relevant code by 5 PM ET by **May 1, 2025**. You should submit only one report for each group.
- The project incorporates strategies from some of the major topics in the course (i.e. data visualization, supervised learning, unsupervised learning) in a way that is appropriate to answer the question of interest.
- While technical skills in implementing ML techniques are important, the best projects will also motivate carefully the rationale for the analysis, and provide a compelling explanation of the results in the form of a well written report. The second half of this document provides specific details of how the report should be formatted. I will also go through this in the lecture for further clarification.
- While we are using `R` as a programming language for problem sets during the course, you may use any statistical software to produce the analysis in your report. However, I will not be providing support in all languages only use a language other than `R` if you already know it well, and will not need support in conducting the analysis.

Topic : The Effects of 401(k) Participation on Wealth

Over the years US has introduced several tax deferred retirement savings plans in order to increase individual savings for retirements. One of the most popular plans are an individual retirement accounts (IRAs) and 401(k) plans. Both of these plans allow an individual to deduct part of their taxable income towards a retirement account and allows them to invest the assets within the plan. The employers may choose to match a certain percentage of an employees contribution, making it an attractive policy. Because 401(k) plans are provided by employers, only workers in firms offering plans are eligible for participation, whereas participation in IRAs is open to everyone (Chernozhukov and Hansen 2004).

- The goal of this project is to study the effect of participating in 401(k) on individual wealth, using the 1991 SIPP dataset described above.
- This project uses dataset is drawn from the 1991 Survey of Income and Program Participation (SIPP) and consists of 9,915 observations. The observational units are household reference persons aged 25 - 64 and spouse if present. Households are included in the sample if at least one person is employed and no one is self employed.
- The data set was analysed in Chernozhukov and Hansen (2004). Please refer to the introduction (Section I) and data (Section III) for more details on the setting and data.
  - Problem Set 1 will further familiarise you with the dataset and the setting. You may use elements of this analysis (e.g. code, figures, tables) in your final report.
  - Using this data and the tools learned in the course, you will construct an algorithmic model to predict wealth, based on various individual characteristics, and check its accuracy against the actual wealth variables. You may choose to use one or all the three wealth variables as your outcomes of interest.
  - You should use this model to estimate the **treatment effect** of participation in a 401(k) plan.
    * Note that there is no single technique that is unambiguously best suited for this task.

    - However, you should only choose one technique and develop your model using that technique.
  - One of the key conceptual tasks in this exercise is to construct a model that uses variables or individual characteristics that affect wealth and are observed in the data. While ultimately you have to choose how to do this in your paper, you might want to think:
    - How should you measure wealth? The dataset includes three candidates - net financial assets, net non-401(k) assets, total wealth. You may choose one measure or all three.
    - Participating in a 401(k) plan may be endogenous (think about selection issues). Can your model account for this endogeneity? In other words, is this endogeneity a cause of concern at all for your model?
    - How should you divide the training and test samples? Alternatively, what aspects of the prediction exercise should you keep in mind while making that decision?

- Some other important aspects that you may want to address are:
  - Analyzing the test data, what are the overall effects of participation in 401(k) plan? Is there any interesting heterogeneity based on income or education level?
  - How does your model perform? What are its strengths and weaknesses? Which features of the outcome variable does it predict well, and which ones does it miss?

- Based on the model you develop, you will then estimate the treatment effect of 401(k) participation. Through out the course, we will discuss examples from the economics literature that exploit ML techniques to estimate treatment effects. You can use these methods as your references while conducting your own analysis.

- Finally, you should replicate the results in Panel A in Table 3 of Chernozhukov and Hansen (2004). For the replication exercise, do not worry about replicating the standard errors. Focus only on the coefficient estimates.
  - This replication exercise involves estimating a set of OLS and 2SLS regressions on the dataset. These regressions will serve as 'traditional' econometric counterparts of your analysis.

- You should provide a careful comparison of the results from your analysis to the results from the replication exercise.
- You can use interpretation and explanations from the original paper as a reference, however, all explanations in your report should be your own words.

References

Chernozhukov, Victor, and Christian Hansen. 2004. The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and Statistics* 86 (3): 735–751.

Final Report

- The report should contain the following sections:

    1. **Abstract** summarizing the question and results (no more than 150 words)
    2. **Introduction**: A 1-1.5 page summary of the project including data, methods, and main results.
    3. **Data**: A section summarizing the data set, providing relevant summary statistics, graphs, etc.
    4. **Methods**: A section on the context, including a brief description of the policy, methodology, and the results. Explain clearly the techniques that you used and why you used those techniques.
    5. **Results and Replication Exercise**: Discuss the results from your analysis, and relate them to economic intuition. This should be followed by a short section on the results of the replication exercise. Provide a discussion comparing your results with that from the replication exercise.
    6. **Conclusion**: Broad summary of the results *including a discussion of the caveats in interpreting the results.*
    7. **References** (if you cite any papers or books, or any other reference).
    8. Do **not** include printouts of your code, but attach to your submission the programs necessary for replication

- **The report should be around 10 - 12 pages long (not including the appendix)**
- Follow all of the style guidelines that appear at the end of this document
- Only include material that is relevant to the argument in your paper. Some students have a tendency to include as much course material as possible in the hopes of showing every fact learned from lecture. You should attempt to build a tight argument in the paper, without extraneous material. Remember, brevity is the key.
- Similarly, when deciding the methods you want to use, more sophisticated does not always mean best. Part of demonstrating knowledge is being able to choose the best tool available to answer the research question, even if it is not the most sophisticated. You should not aim to include every technique you have learned in the course, instead choose the ones appropriate for answering the question.

Evaluation Guidelines

- Successful projects will fully develop the assigned topic, including the counterfactual prediction, interpretation of results and economic intuition.
- You should work hard on both form and substance: the report should be clear and in line with the standards of professional writing for economists. See the style guidelines below for hints, and compare the style and formatting of your paper to the references in the syllabus.
- While precision and technical expertise is necessary in choosing and estimating your model, and producing predictions with it, I particularly emphasize that mature understanding of what you are doing, and why, is the hallmark of a successful project. Thus, when evaluating a project that employs the most sophisticated tools, but with limited understanding of the underlying data and poor interpretation of the results, versus a project that uses basic tools, but demonstrates full understanding of the data, technique and results, I will almost surely give a higher score to the latter.
- I will give substantial weight to the quality of the discussion of the results that you obtain: what are their limitations and strengths? Why? Are there other avenues of analysis that, although not pursued in the current report, may yield interesting results? For example, this can be motivated in light of the data you have.
- Successful projects always result from the full involvement of everyone in the group. Specialization of different individuals to different tasks is efficient and necessary, but everyone should also be on board with the overall direction of the project, and contribute their views and expertise where needed. Lack of cohesion or free-riding within the group results in projects that read like collection of disconnected pieces, which would result in low scores. While, I encourage you to come up with a dynamic that suits best for your group, I am happy to advice on how to navigate work-flow for a successful project.
    - Additionally each student will submit a short write-up describing their contribution to the group project. The grade for this write-up will depend on the quality of that contribution as reflected in the final report.

Style Guidelines

In reading the papers, most people pay considerable attention to style (correct spelling and grammar, clear exposition, good organization). Reports that are difficult to read routinely get ignored, even if they contain good ideas. Thus, it will pay to develop the habit of working hard to craft a clear explanation of your ideas.

Many of the following suggestions are standard good practice. Others are matters of taste.

1. In writing up a research report, one should have an audience in mind. I suggest that you take the audience to be your fellow students in this course.

2. Include a cover page with the following information: Title; date; your name; your e-mail address; the word Abstract; an abstract of 150 words or fewer. If you have acknowledgments to make (thanking a fellow student for helpful comments, for examples), put these on the bottom of the cover page. The text of the paper begins on the next page.

3. Your paper should be divided into sections, to help guide the reader.

4. Number the pages. No plastic covers or binders, please.

5. In the introduction, present an overview of your paper and summarize your findings. In the conclusion, give suggestions for future research.

6. Be explicit about your data set. State the sample size. Describe any data cleaning or aggregation you did for the analysis. Clearly explain the rationale of doing so as well. You will want to include a plot and/or a table with basic statistics (means, standard deviations) of the data. Remember, this is the data you use for the analysis.

7. Number the equations.

8. Tables:

   (a) Number the tables, and on each include a descriptive header (Means and Standard Deviations of Data, or Variance Decompositions, for example).

   (b) Include the tables within the text, in the appropriate place.

(c) Tables should not run over page boundaries, unless they are too long to fit on a single page. That is, if you include a table in the text, you should insure that you place it so that it does not run from one page to the next.

(d) Make every effort to make each table self-contained, even though this will require you to redundantly present information that is also stated in the body of the paper itself. This is now the standard in the profession and you should look at a paper published recently to see how much detail is included in tables.

i. In notes at the bottom of each table, define the symbols that are in the table, or give a precise reference to where the definition may be found. It is not adequate to simply state "definitions are in the paper or see section 2 of the paper for definitions". Instead say something like Variable definitions: y =log per capita income in 1992 dollars, r =interest rate on 3 month Treasury bills (end of quarter), and so on. Alternatively, for many of you it might be best to include a table that defines the symbols, and in subsequent tables say see Table x for variable definitions where x is the number of the table that defines the symbols. (You will also present such information in the text itself.)

ii. In tables that present regression results, include a note that describes the estimation technique ("The probit was estimated by maximum likelihood, assuming normality," for example.). You will also present such information in the text itself.

iii. If a given set of variables appears in more than one tableas is often the case there is no need to repeat the variable definitions. Instead one of the notes to (say) Table 2 can say "Variable definitions are given in notes to Table 1."

iv. Be sure to include the name of the dependent/outcome variable somewhere in the table.

v. When possible use words to describe the variables in your model. For example, if years of schooling is a regressor in your model write out "Years of schooling" not "Y_RSCH" if that is the name of the variable in your statistical software.

9. Figures:

(a) Number the figures, and on each include a descriptive header ("Parental Income versus SAT Score," for example).

(b) Include the figures within the text, in the appropriate place.

(c) Figures should not run over page boundaries, and must always fit on a single page. That is, if you include a figure in the text, you should insure that you place it so that it does not run from one page to the next.

10. Reporting of estimates:

(a) Do not report more than 3 or 4 digits. Example: report 0.412, not 0.4117678.

(b) Avoid long strings of zeroes at the beginning of a number. You can always retroactively rescale variables and coefficients.

(c) For regression models, report standard errors, not t-statistics. Standard errors belong in parentheses under the coefficients. Example:

$$\text{Report} \quad \begin{matrix} 0.412 \\ (0.146) \end{matrix} \quad \text{not} \quad \begin{matrix} 0.412 \\ 0.146 \end{matrix}$$

11. References:

(a) All references cited in the paper should be listed in a bibliography at the end of the paper. Cite these in the text as Walter (1995).

(b) When you reference a specific result, such as a point estimate of a parameter, or a theorem that establishes a particular claim, give the page number, such as Walter (1995, p361). When you reference a general result, for example noting other papers that have studied topics similar to yours, no page number is needed.

(c) For the reference section at the end, you can use MLA or Chicago Manual of Style. Refer to the style guides here for MLA and here for Chicago. Note you can use Google Scholar to generate references in either of these styles.

12. Computer code: You do not need to include you programs in the paper. I should be able to figure out what you did without seeing it explicitly.

13. It is a violation of scholarly ethics to repeat a passage, even a sentence, from another source without putting the passage in quotes and citing the source: the usual

publication details in case of printed matter, the URL and date in the case of web-only material. This rule applies even when you are describing dreary facts: if you repeat a description from another paper of how data were collected, or the steps in computing an estimate, you must put the passage in quotation marks and cite the original source.