Final Project

Prof. Chouldechova

Assigned: November 28, 2020

Due: 11:59PM ET Monday, December 14, 2020

Important note

> Regardless of your grading basis, in order to pass the course you must achieve a score of at least 50% (at least 21/40) on the final project.

Project outline

Your task for this project is to write a report that examines **one** of the following two questions, each of which probes the question of income inequality between groups:

> 1. Sex-related differences: Is there a significant difference in income between **men and women**? Does the difference vary depending on other factors (e.g., education, marital status, criminal history, drug use, childhood household factors, profession, etc.)?

> 2. Race-related differences: Is there a significant difference in income across **racial groups**? Does the difference vary depending on other factors (e.g., education, marital status, criminal history, drug use, childhood household factors, profession, etc.)? Note: for this problem you may find it easiest to focus on differences between just two groups. E.g., Black and non-Black, non-Hispanic.

To address this problem you will use the NLSY97 (National Longitudinal Survey of Youth, 1997 cohort) data set. The NLSY97 data set contains survey responses on thousands of individuals who have been surveyed every one or two years starting in 1997.

Base data set: To get you started, I've pulled together 90 variables from the broader data set. This base data set is posted on the course website, along with accompanying info files that tell you about the variables. There's also a starter script that gets you started with some basic variable renaming (though you'll want to go further than what's shown there).

A natural outcome variable for the data is `TOTAL INCOME FROM WAGES AND SALARY IN PAST CALENDAR YEAR` (2017 survey question). This variable gets renamed to `YINC-1700_2017` by the starter script provided. Note that this quantity is **truncated** aka **topcoded**, meaning that you do not get to see the actual incomes for the top 2% of earners. For the top 2% of earners, the income variable is set to the average income among the 2% of earners. The implication of this topcoding is something you'll want to discuss as part of your analysis.

> You are **not** expected to use all 90+ variables in your analysis. It suffices to choose a total of 8-14 variables (income, sex/race, + 6-12 others) and to perform a thorough analysis using just those variables.

Collaboration policy

This is a **group project**. While you may undertake the project independently, you may also work in teams of 2-3 students.

While you are not prohibited from discussing the project outside of your group, everything that your team submits must be your team's own work product. You are not permitted to share project code or text with members of other teams.

All student are expected to comply with the CMU policy on academic integrity. This policy can be found online at http://www.cmu.edu/academic-integrity/ (http://www.cmu.edu/academic-integrity/).

Any submitted project that is deemed by the instructor to be in violation of the collaboration policy or academic integrity policy will receive a score of 0. Since a passing score on the project is necessary for passing the class, anyone deemed to be in violation of the policies will automatically fail the class.

Project requirements

Your end-product for the project will be an R Markdown report that covers the following.

1. Data processing and summarization (12 points)

You should begin by describing the data you have available. You will want to display tabular summaries of means and proportions where appropriate. Since the main question is one of gender differences, you may want your tabular summaries to also break things down by gender.

Your score for this section will be based on the following criteria:

- Meaningful variable names, factor variables, and factor level names
- Insightful graphical and tabular summaries of the data
- Proper labelling of figure axes and table columns

- Discussion of the graphical and tabular summaries.

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year college student trying to learn more about income inequality between men and women.

2. Methodology (10 points)

In this section you should provide an overview of the approach you took to exploring and analyzing the data. This is where you tell the story of how you got to your main findings. It's too tedious to carefully format plots and tables for every approach you tried, so you can also use this section as a place to explain the various types of analyses that you tried.

You should address *at least* the following questions:

- How did you deal with missing values? What impact does your approach have on the interpretation or generalizability of the resulting analysis?
- How did you deal with topcoded variables? What impact does your approach have on the interpretation or generalizability of the resulting analysis?
- Did you produce any tables or plots that you thought would reveal interesting trends but didn't?
- What relationships did you investigate that don't appear in your findings section?
- What's the analysis that you finally settled on? What income and gender related factors do you investigate in the final analysis?

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year college student trying to learn more about income inequality between men and women.

3. Findings (13 points)

In this section you give a careful presentation of your main findings concerning the main problem of race/sex-related income (in)equality. You should provide, where appropriate:

- Tabular summaries (with carefully labelled column headers)
- Graphical summaries (with carefully labelled axes, titles, and legends)
- Regression output + interpretation of output + interpretation of coefficients
- Assessments of statistical significance (output of tests, models, and corresponding p-values)

As part of your analysis you must run a regression model. When running regressions, you should discuss whether the standard diagnostic plots indicate issues with the model (trends in residuals, variance issues, outliers, etc.). You will not receive full credit for your regression unless you clearly display and discuss the diagnostic plots.

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year college student trying to learn more about income inequality between men and women.

4. Discussion (5 points)

In this section you should summarize your main conclusions. You should also discuss potential limitations of your analysis and findings. Are there potential confounders that you didn't control for? Are the models you fit believable?

You should also address the following question: **How much confidence do you have in your analysis?** Do you believe your conclusions? Are you confident enough in your analysis and findings to present them to policy makers? (*You will not be deducted points for saying that you are unsure of your analysis. This is just something I want you to reflect upon.*)

Submitting data files

You may assume that the person grading your report will have the file called `nlsy97_Nov2020.csv` in their working directory. Any other files that you need in order for your Rmd file to knit should be sumitted along with your report. You will be responsible for submitting:

- The **Rmd** file that generates your analysis
- The resulting **html** file produced by knitting