Introduction to Statistical Methods

Replication Project Instructions

Instructor: Siwei Cheng
Department of Sociology
New York University

Project Overview

Replication is an important part of social science research. It provides a hands-on method to train your quantitative analysis skills. In this final course project, we ask everyone to replicate the basic descriptive and regression results from a research article. This part includes downloading and cleaning the data, generating descriptive tables and figures, estimating regression models, and summarizing these results into publication-ready tables.

We require you to submit a package containing **three files**. The first is a LaTeX-formatted PDF document that includes your own replication tables and figures. The second is a project report (no longer than 2-page double-spaced) . The file may contain tables and figures and should be compiled in LaTeX. The third is a file containing your computational codes.

Note that this is an individual assignment. You should work alone and try really hard to do things on your own. Getting stuck and working your way out of a roadblock is the most important experience of advanced empirical work. But if you get stuck, we encourage you to ask each other for hints or discuss the problem together.

Replicating the Descriptive Statistics and Regression Models in a Published Paper

We ask you to replicate descriptive and regression analysis in a published paper. It contains the following steps:

1. Downloading the article.

    - Mandel, Hadas, and Moshe Semyonov. "Going back in time? Gender differences in trends and sources of the racial pay gap, 1970 to 2010." *American Sociological Review* 81, no. 5 (2016): 1039-1068.
    - You should download the online appendix available on the ASR website.

- Read the article.

2. Obtaining the raw data.

   - Browse the Integrated Public Use Microdata Series (IPUMS) website:

     https://www.ipums.org/

   - IPUMS provides census and survey data from around the world integrated across time and space. IPUMS integration and documentation makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community context. Data and services available free of charge.
   - The data sets used in this paper include the decennial years between 1970 and 2010. Refer to the data description on page 1045 of the paper: *"Data for 1980 and 2000 were derived from the 5 percent census samples; data for 1970 and 1990 were derived from the 1 percent census sample; and data for 2010 were derived from the American Community Survey (ACS). Because the analysis focuses on comparisons between blacks and whites, we excluded all other racial/ethnic groups (i.e., Hispanics, Asians, and other races) from the analysis."*

3. Cleaning the data

   - You should follow the authors' descriptions in the "Data Source and Variables" section to set up the variables for analysis. (Hint: see Table A1a and A1b for the list of variables).
   - After setting up the variables, you should also restrict your sample according to the authors' description. For example, the analysis focused on blacks and whites and excluded other racial/ethnic groups (i.e., Hispanics, Asians, and other races), so you should set up your sample accordingly. (Hint: check your sample sizes with Table A1a and A1b. Make sure to include a "year" variable to indicate the year in which an observation comes from.)
   - Note that your earnings variables should adjust for inflation using the CPI99 variable on the IPUMS website:

     https://cps.ipums.org/cps/cpi99.shtml

   - For the occupational categories, see the csv document in the course website for the crosswalk between *occ1990* and the two-digit (82 categories) coding scheme. Your computational codes should include commands for loading this crosswalk table into R and recoding occupation into the two-digit scheme.

- You will also need to construct Index of Dissimilarity in R, and create this index to capture the percentage of either blacks or whites that would have to change occupation to reach equal occupational distributions.

4. Replicating tables and figure for descriptive statistics.

   - You should replicate the descriptive statistics presented in Table A1a and A1b in the appendix. The numbers in the tables may not be exactly the same as those in the published table, but they should be quite similar.
   - You should also make a similar plot as Figure 1 to describe the trends in average weekly earnings by gender and race.
   - After creating the descriptive tables and figure, interpret the findings yourself, and then compare your own interpretation with the discussion in the article. This is to make sure you understand the descriptives in the sample statistics, and it is also possible that you come up with even better interpretations than the authors. This will be a good opportunity for learning and practice.

5. Replicating results from regression models.

   - Replicate the linear regression models for blacks and whites separately. The authors' purpose of running these models is to generate coefficients for conducting the Oaxaca-Blinder decomposition. However, in this part of your project, we only require you to focus on the regression models.
   - You should set the models as described by the authors and replicate the coefficients reported in their Table A2a and A2b.
   - Format your tables in LaTeX.

6. Write a brief project report (2 double-spaced pages). Use this brief project report to discuss any issues that arise as you try to replicate the findings, any of the authors' data decisions that you don't agree with, and any future extensions that you think are necessary.

Some Additional Notes on the Replication Project

- Sample definition: Due to the complexity and ambiguity of the 1970 sample, we are excluding year 1970 from the replication exercise. Use the 5 percent samples for 1980 and 2000, the 1 percent metro sample for 1990, and the ACS data for 2010.

- A key part of real-life data analysis is making data decisions. This replication project is an opportunity for you to start thinking seriously about data decisions and data strategies. These include, for example:

    1. Pay attention to whether the variable coding is consistent or different across years. For example, some variables (e.g. weeks worked in the last year) are coded as continuous variables in some years but in intervals in other years. In this case, you may consider using the mid-point of the intervals as proxy for the numeric values of these variables.
    2. Pay attention to how missing values are coded in the data.
    3. Think carefully about how inflation adjustment for the wage variable should be conducted (hint: it's not complicated.). If you are unclear, use Google.
    4. Whenever you create a new variable, check the descriptives, plot a histogram or bar chart, tabulate it with the existing variable(s) based on which it is created, count the number of non-missing values, etc.

- It is OK if your final sample sizes and descriptive statistics do not exactly match the tables in the published article. But if they are very different, you should be able to speculate some explanations in your report. If you don't like any of the authors' decisions, you can also write about that in your report. You should also feel free and encouraged to compare your results with your classmates and discuss where these differences arise.

- **Bottom line:** The course and lab instructors will NOT be providing detailed instructions on how to set up the data. We will only be answering questions if you think there is a problem with the data or analysis. In fact, when it comes to real life research, there is no one there to give us definitive solutions to data analysis! We want to make the process of thinking through and discussing these data decisions an important part of your data analysis training. Enjoy!