Paper Replication

Eco 231 - Undergraduate Econometrics

Prof. Carolina Caetano

For the replication assignment, you will have to reproduce the results reported by Ronald DeBeaumont in "Occupational differences in the wage penalty for obese women." Specifically, you will have to replicate Table 2, 3, and 4 appearing in the paper. Use the NLSY79_ECO231W.dta dataset for the replication.

Note several things regarding the data and variables:

- The file NLSY79_ECO231W.dta contains variables extracted from the NLSY79 survey for year 1990. Each observation is identified by the survey respondent ID (variable CASEID_1979). You will have to clean the dataset and prepare the variables before you can begin the replication. As in the homework, make sure you recode variables before making use of them.
- You can identify variables using the codebook PDF available on the course website. Figure 1 provides an example of a codebook entry. The variable name appears in the top left corner, in square brackets. Since Stata cannot use tilde (∼), hyphens, or dots in variable names, these are replaced by an underscore (e.g., QES−52.01 appears as QES_52_01 in the dataset). For more information on NLSY79 codebook structure, see https://www.nlsinfo.org/content/cohorts/nlsy79/using-and-understanding-the-data/nlsy79-documentation.
- Alternatively, you can use the interactive codebook available online at https://www.nlsinfo.org/investigator/pages/search.jsp (choose the NLSY79 (1979-2014) study). You can either browse or search for specific variables under the tab "Variable Search." Note that not all variables appearing online exist in the dataset supplied to you.
- In NLSY79, missing values (i.e., item non-response) are recorded by four categories: −1 (refusal), −2 (don't know), −3 (invalid skip), and −5 (non-interview). Make sure that you properly handle such invalid data values. Unlike these negative numbers, −4 (valid skip) is assigned to respondents when NLSY79 does not ask that question to those respondents. For instance, some questions might apply to only females, a certain age range, or only wage-salary workers (not self-employed). Check "Universe" in the codebook to identify such skip patterns.
- Some survey questions have only been asked once. These are variables that are not expected to change over time (e.g. gender, race, date of birth), in which case survey answers are expanded over the entire survey period. For example, if CASEID_1979 = 1 reported as female in 1979, she

will appear in the data as answering "female" in every survey year, including year 1990 in which we focus on (though only being asked in 1979).

- In the paper, all relevant employment data (including wage, tenure, occupation, collective bargaining, etc.) refer to the current or most recent job. This job is called, "CPS job" in the NLSY dataset. In many cases, employment information of the CPS job can be found in the same dataset provided to you. However, this is not the case for a couple of variables that we are interested in, such as Tenure and Collective Bargaining.

- In order to construct Tenure of the CPS job, we make use of how the NLSY survey collects employment information. In the dataset, you will notice that tenure information is collected across 5 different jobs in 1990 survey. Among job #1 $\sim$ #5, one of these has to be currently or most recently held job (the CPS job). Now, we look at the variables QES$-$52.01 through QES$-$52.05 which allow us to identify job number that corresponds to the CPS job. For instance, if QES$-$52.01 has a value 1, it means that job #1 is the current or most recent job to that respondent. For that person, tenure information of the CPS job is exactly the same as tenure information of the job #1, which can be found at TENURE1_1990. Do the same step for collective bargaining as well. For more detailed information of such linking process, read https://www.nlsinfo.org/content/cohorts/nlsy79/topical-guide/employment/jobs-employers/page/0/1.

- One more comment: while working with the data, you may be in a situation where total number of observations in your dataset differs from one reported in the paper. If this happens, some of your regression results may not be the same as the values presented in the paper. In this case, you might as well do the followings:

  - Check whether there are any mistakes in cleaning the dataset and correct whenever necessary. This helps you get the sample size reasonably close enough (e.g., your dataset has 200-300 more/less observations than what's reported in the paper).
  - Reproduce Table 2 before conducting regression analysis. If your results in the "mean" column are fairly close to the summary statistics in the paper, you can move on.

  Once you narrow the discrepancy, most of your estimates will look quite similar (although a few might look different). In practice, this is one of common procedures that researchers follow when it comes to the replication.

Figure 1: Codebook Entry Example