Longitudinal Data Cleaning - A Case of the NLSY79 Data

A Preprint

Dewi Amaliah *
Department of Econometric and Business Statistics
Monash University
Clayton, VIC 3168
dama0007@student.monash.edu

Dianne Cook
Department of Econometric and Business Statistics
Monash University
Clayton, VIC 3168
dicook@monash.edu

February 21, 2021

Abstract

Enter the text of your abstract here.

Keywords longitudinal data · data cleaning · data tidying · robust linear model

1 Introduction

2 The NLSY79

3 The NLSY79 Data Cleaning

3.1 Getting and Tyding the Data

The NLYS data is stored in a database and could be downloaded by variables. Several variables are available to be downloaded and could be browsed by index. For the wages data set, we only extracted these variables:

- Education, Training & Achievement Scores
  - Education -> Summary measures -> All schools -> By year -> Highest grade completed
    - Downloaded all of the 78 variables in Highest grade completed.
- Employment
  - Summary measures -> By job -> Hours worked and Hourly wages
    - Downloaded all of the 427 variables in Hours worked
    - Downloaded all of the 151 variables in Hourly wages

    Both hours worked and hourly wages are recorded by the job, up to five jobs for each id/subject.
- Household, Geography & Contextual Variables

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Optional.

  - Context -> Summary measures -> Basic demographics
    * Downloaded year and month of birth, race, and sex variables.

    There are two versions of the year and month of birth, i.e. 1979 and 1981 data. We downloaded these two versions.

The downloaded data set came in a csv (wages-high-school-demo.csv) and dat (wages-high-school-demo.dat) format. We only used the .dat format. It also came along with these files:

- wages-high-school-demo.NLSY79: This is the tagset of variables that can be uploaded to the web site to recreate the data set.
- wages-high-school-demo.R: This is an R script provided automatically by the database for reading the data into R and convert the variables' name and its label into something more sensible. We utilized this code to do an initial data tidying. It produced two data set, `categories_qnames` (the observations are stored in categorical/interval values) and `new_data_qnames` (the observations are stored in integer form). In this case, we only used the latter.

`new_data_qnames` is still untidy, where all of the variables for each year and each job being stored in one column, hence the data contains a huge number of columns (686 columns). Thus, the data should be tidied and wrangled first to extract the demographic and employment variables that we want to put in the final data set.

3.1.1 Tidying demographic variables

Date of birth, age in 1979, gender, race, highest grade completed (factor and integer), and the year when the highest grade completed are the variables that we want to use in the cleaned data set. There are two versions of date of birth variable, which are the 1979 version and the 1981 version. In this case, we only used the 1979 data. We also did a consistency check for the 1979 and 1981 data and flag the inconsistent observations.

3.2 Initial Data Analysis

3.3 Robust Linear Model for Influential Values Treatment

4 Conclusion