THE JOURNEY FROM WILD TO TEXTBOOK DATA: A CASE STUDY FROM THE NATIONAL LONGITUDINAL SURVEY OF YOUTH

Dewi Amaliah
Department of Econometrics and Business Statistics
Monash University
Clayton, VIC 3800
dama0007@student.monash.edu

Dianne Cook
Department of Econometrics and Business Statistics
Monash University
Clayton, VIC 3800
dicook@monash.edu

Emi Tanaka
Department of Econometrics and Business Statistics
Monash University
Clayton, VIC 3800
emi.tanaka@monash.edu

Kate Hyde
Department of Econometrics and Business Statistics
Monash University
Clayton, VIC 3800

Nicholas Tierney
Telethon Kids Institute
Nedlands, WA 6009

January 14, 2022

Abstract

The National Longitudinal Survey of Youth (NLSY79) is a prominent open data source that has been important for educational purposes and multidisciplinary research on longitudinal data. Subsets of this data can be found in numerous textbooks and research articles. However, the steps and decisions taken to get from the raw data to the textbook data are never clearly articulated. This article describes our journey when trying to re-create a textbook example data set from the original database, with the goal being to refresh the textbook data more regularly. Thus, this paper demonstrates the process – extracting, tidying, cleaning, and exploring with documentation – to make refreshed data available for education or research. Three new data sets and the code to produce them are provided in an accompanying open source R package, called [CENSORED]. As a result of this process, some recommendations are also made for the NLSY79 curators for incorporating data quality checks and providing more convenient samples of the data to potential users.

Keywords Data cleaning; Data tidying; Reproducible workflow; Longitudinal data; NLSY79; Initial data analysis;

1   Introduction

Statistics and data science education relies on cleaned and simplified data, suitably called textbook data, for clear examples about how to apply different techniques. An example of this the wages data made public by Singer and Willett (2003) which is commonly used to teach generalized linear models, including hierarchical, mixed effects and multilevel. The freely available data contains records hourly wages of a sample of high school dropouts, from 1979-1994, along with demographic variables including education and race, taken from the National Longitudinal Survey of Youth (NLSY79) (Bureau of Labor Statistics, U.S. Department of Labor 2021).

This textbook data was used as an example in Ozlem Ilk's PhD thesis supervised by author Cook (Ilk 2004), and was developed into a case study for use in teaching exploratory data analysis at Iowa State University. The story from modeling the data (and as reported by Singer and Willett) is that wages increase with length of time in the workforce, higher level of education leads to higher wages, and that race makes a difference, on average. The story from an exploratory analysis is very different: the individual experience varies a lot from the overall average. Some individuals experience a decline in wages the longer they are in the workforce, and most experience volatility in their wages.

This disparity between the average and the individual is a part of statistics that is given less attention. It is this disparity which makes this particular data set very important as a textbook example. Textbook data sets have staying power if it has a challenge that has never been completely solved. The iris data (XXX REF) is a prime example, it has withstood the test of time because the three species cannot be perfectly classified, and so it continues to challenge researchers and instructors to do better in the analysis. (A side note: the iris data is best replaced today with the penguins data (XXX REF), which has similar qualities, is new and does not suffer from a connection with eugenics (Stodel 2020).) The wages data is in this class of textbook data, because it presents this challenge for exploratory data analysis: how can we better summarize and explain the individual experience?

For the field of statistics, and data science by association, it is increasingly important to reach the individual. One might describe this as a divergence of purpose, statistics for public policy or statistics for the public. The two are not the same. As the world becomes more electronically connected combating misinformation and mitigating conspiracy theories require that statistics address the individual. For example, with the wages data, the message for the individual is that you are more than your demographic. The majority of people tracked are not like the average, the average is an anomaly. If you have a bad experience, that your wages have declined over time, you are not alone, there are others like you, and more than you think.

As a textbook data though, the dates are tired. The most recent year in the data is 1994, 10 years prior to when Singer and Willett (2003) was published. Teachers of statistics need use contemporary data sets to show how techniques are relevant to today's students. Using tired old textbook data sets sets up a misconception that the field is not current. The wages data is extracted from NSLY79, one of the best examples of open data, which is constantly being updated. It should be possible to continuously refresh the textbook data from the data repository. This is our story of the journey from open (Open Knowledge Foundation 2021) but wild data to textbook data, it's not entirely pretty.

Singer and Willett (2003) used the wages and other variables of high school dropouts from the NLSY79 data as an example data set to illustrate longitudinal data modeling of wages on workforce experience, with covariates education and race. This data has been playing an important role in research in various disciplines, including but not limited to economics, sociology, education, public policy, and public health for more than a quarter of the century (Pergamit et al. 2001). In addition, this is considered a carefully designed longitudinal survey with high retention rates, making it suitable for life course research (Pergamit et al. 2001; Cooksey 2017). According to Cooksey (2017), thousands of articles, and hundreds of book chapters and monographs have utilized this data. Moreover, the NLSY79 is considered the most widely used and most important cohort in the survey data (Pergamit et al. 2001). Our aim is to refresh this textbook data and append it with data from 1994 through to the latest data reported in 2018, a purpose that is consistent with Grimshaw (2015)'s statistics education goal of embracing authentic data experiences. Here, we investigate the process of getting from the raw NLSY79 data to a textbook data set as similar as possible to that provided by Singer and Willett (2003). We should also note that race is a variable in the original data set, and for compatibility it is also provided with the refreshed data, for the *purposes of studying racism, not race* (Fullilove 1998).

This paper demonstrates the steps of cleaning data, and documents the process, as recommended by Huebner, Vach, and Cessie (2016). They emphasize that making the data cleaning process accountable and transparent is imperative and essential for the integrity of downstream statistical analyses and model building. Data

cleaning can be considered to be a part of what is called "initial data analysis" (IDA) (Chatfield 1985). In IDA one would also explore the data, especially to check if the data is consistent with assumptions required for modeling. This is also related to exploratory data analysis (EDA), coined by Tukey (1977) with a focus on learning from data. EDA can be considered to encompass IDA. Dasu and Johnson (2003) say that data cleaning and exploration, without naming it as IDA, is a difficult task and consumes 80% of the data mining task.

Our process of cleaning builds heavily on the `tidyverse` approach (Wickham et al. 2019). The data is first organised into "tidy data" (Wickham 2014) and then further wrangled using the data pipeline and split-apply-combine approach (Wickham 2011). (Tidy data shouldn't be confused with "tame data" which Kim, Ismay, and Chunn (2018) coined to refer to textbook data sets suitable for teaching, particularly teaching statistics.) The resulting (tame) data is provided in a new R package called `yowie` which includes the code so that the process is reproducible, and could be used to further refresh the data as new records are made available in the NLSY79 database.

This paper is structured in the following way. Section 2 describes the NLSY79 data source. Section 3 presents the steps of cleaning the data, including getting and tidying the data from the NLSY79 and IDA to find and repair anomalies. Our final subset is compared to the old textbook subset in Section 4. Finally, Section 5 summarizes the contribution and makes recommendations for the NLSY79 data curators.

2 The NLSY79

2.1 Database

The NLSY79 is a longitudinal survey administered by the U.S Bureau of Labor Statistics that follows the lives of a sample of American youth born between 1957-1964 (Bureau of Labor Statistics, U.S. Department of Labor 2021). The cohort originally included 12,686 respondents aged 14-22 when first interviewed in 1979. It comprised of Blacks, Hispanics, economically disadvantaged non-Black non-Hispanics, and youth in the military. In 1984 and 1990, two sub-samples were dropped from the interview; the dropped subjects were the 1,079 members of the military sample and 1,643 members of the economically disadvantaged non-Black non-Hispanics, respectively. Hence, 9,964 respondents remain in the eligible samples. The surveys were conducted annually from 1979 to 1994 and biennially thereafter. Data are currently available from Round 1 (1979 survey year) to Round 28 (2018 survey year).

Although the main focus area of the NLSY is labor and employment, the NLSY also covers several other topics, including education; training and achievement; household, geography, and contextual variables; dating, marriage, cohabitation; sexual activity, pregnancy, and fertility; children; income, assets and program participation; health; attitudes and expectations; and crime and substance use.

There are two ways to conduct the interview of the NLSY79, which are face-to-face or by telephone interviews. In recent survey years, more than 90 percent of respondents were interviewed by telephone (Cooksey 2017).

2.2 Target data

The NLSY79 data used in Singer and Willett (2003) contains the longitudinal measurements on yearly mean hourly wages with years of workforce experience, and demographic variables education and race, from 1979 through to 1994. In addition, the cohort is restricted to male high-school dropouts who first participated in the study at age 14-17 years. Thus the target data set is to collect the same variables for the extended time frame of 1994 through to 2018, the most recent year reported.

3 Data cleaning

M. van der Loo and de Jonge (2018) describe the notion of a "statistical value chain" where the production stages of the data cleaning process are earmarked as raw data (data as arrived to the desk of an analyst), input data (data organised with correct type and identified variables) and valid data (data that faithfully represent the variables). In this section, we outline the steps to download the raw data (Section 3.1) and then tidy the raw data into input data, specifically for the demographic variables (Section 3.1.1) and the employment variables (Section 3.1.2), so that the resulting input data can be used downstream for validating the data as described in Section 3.2.

Navigating the data source

🗄 NLSY79 (`https://www.nlsinfo.org/investigator/pages/search?s=NLSY79`)

✔ The CASEID will be always be selected.

✔ The 3 recommended demographic variable (sample ID, race and sex) were selected.

For the remaining variables, we went to the "Variable Search" tab and select variables as follows

　▷ Education, Training and Achievement Scores

　　▷ Education ▷ Summary measures ▷ All schools ▷ By year ▷ Highest grade completed

　　　✔ All 80 variables in Highest grade completed were selected.

　▷ Employment

　　▷ Summary measures ▷ By job

　　　▷ Hours worked

　　　　✔ All 442 primary variables in Hours worked were selected.

　　　▷ Hourly wages

　　　　✔ All 156 variables in Hourly wages were selected.

　▷ Household, Geography and Contextual Variables

　　▷ Context ▷ Summary measures ▷ Basic demographics ▷ Date of birth

　　　✔ All 4 variables in Date of birth were selected.

☁ To download all 686 variables selected, we then navigate to the tab "Save / Download" then select the tab "Advanced Download". We select the R Source code and Comma-delimited datafile of selected variables with Reference Number as column headers. We name the filename "NLSY79" and press the download button. There are also options to get control or dictionary files for SAS, SPSS and STATA.

Figure 1: The above documents the steps taken to select variables of interest and download the raw data.

3.1 Getting the data

The NLSY79 data contains a large number of variables but for our aim, the scope required is limited to demographic profiles and wages data. More specifically, we went to the NLSY79 database website at `https://www.nlsinfo.org/content/cohorts/nlsy79/get-data`, clicked on the direct link to NSLY79 data and navigated as described in Figure 1.

The downloaded data set comes as a zip file, containing the following set of files:

- `NLSY79.csv`: Comma Separated Value format of the response data,
- `NLSY79.dat`: .dat format of the response data,
- `NLSY79.NLSY79`: Tagset of variables that can be uploaded to the website to recreate the data set, and
- `NLSY79.R`: R script for reading the data into R and converting the variables' names and label into something more sensible.

We alter only the file path in `NLSY79.R` and run the script without any other alteration. This results in an initial processing of the raw data into two data sets, `categories_qnames` (where the observations are stored in categorical/interval values) and `new_data_qnames` (the observations are stored in integer form).

The raw data, `new_data_qnames`, is organised such that each row corresponds to an individual. As respondents can have multiple jobs at specific years, the column names, such as `HRP1_1979`, `HRP2_1979`, `HRP1_1980` and `HRP2_1980`, contain the information about the job number up to 5 (`HRP1` = job 1, `HRP2` = job 2) and the year. The raw data consequently has a large number of columns (686 to be specific). The values in the cell under the variables that begin with `HRP` correspond to the hourly wage in dollars. A glimpse of this data output is shown below.

```
#> 'data.frame':    12686 obs. of  686 variables:
#>  $ CASEID_1979                : int  1 2 3 4 5 6 7 8 9 10 ...
#>  $ HRP1_1979                  : int  328 385 365 NA 310 NA NA NA 214 NA ...
```

```
#>  $ HRP2_1979                    : int  NA NA NA NA 375 NA NA NA NA NA ...
#>  $ HRP3_1979                    : int  NA NA 275 NA NA NA NA NA NA NA ...
#>  $ HRP4_1979                    : int  NA NA NA NA NA 250 NA NA NA NA ...
#>  $ HRP5_1979                    : int  NA NA NA NA NA NA NA NA NA NA ...
#>  $ HRP1_1980                    : int  NA 457 397 NA 333 275 300 394 200 318 ...
#>  $ HRP2_1980                    : int  NA NA 367 NA NA NA NA NA NA NA ...
#>  $ HRP3_1980                    : int  NA NA 380 NA NA NA 290 NA NA NA ...
#>  $ HRP4_1980                    : int  NA NA NA NA NA NA NA NA NA NA ...
#>   [list output truncated]
```

According to Wickham (2014), tidy data sets comply with three rules: (i) each variable forms a column, (ii) each observation forms a row, and (iii) each type of observational unit forms a table. The raw data does not comply with these rules, thus we re-arrange and wrangle the data into tidy data form, columns corresponding to individual ID, year, job number, wage in dollars and the demographic variables. This is done using the `tidyverse` suite of packages (Wickham et al. 2019), `tidyr` (Wickham 2020) to pivot the data into long form, `dplyr` (Wickham et al. 2020), and `stringr` (Wickham 2019) for creating new variables, and levels of factors by text wrangling. The long form of the data makes it possible to do these data transformations efficiently, and it is an intermediate step towards the final target data. The code for tidying the data are demonstrated at `CENSORED/articles/raw-to-input-data.html` but also described in the subsequent subsections.

3.1.1 Tidying demographic variables

In our final target data, we wish to include the demographic variables with variable names specified in brackets: gender (`gender`), race (`race`), age (`age_1979`), highest grade completed (`hgc`), highest grade completed in terms of years, e.g. 9th grade = 9, 3rd year college = 15, (`hgc_i`) and the corresponding year this grade was completed (`yr_hgc`).

For gender and race, we only rename the column names. It is worth noting that using these two variable needs special attention. Gender, as reported in the data, only has two categories, which is recognised today as inadequate. Gender is not binary. Further, race is as reported in the database. When doing analysis with this variable, one should keep in mind that the purpose is to study racism rather than race.

The `new_data_qnames` contains the variables `Q1-3_A~Y_1979` and `Q1-3_A~Y_1981` which records two versions of the birth year of the respondent; this is also the case for the record of birth month (`Q1-3_A~M_1979` and `Q1-3_A~M_1981`). The record contains two versions of birth year and birth month as the survey recorded this in 1979 and 1981. We checked for consistency between the two versions and found no discrepancy where the responses were recorded in both 1979 and 1981. The age was then calculated using the birth year.

The next step is tidying to obtain `hgc` and `yr_hgc`. The highest grade completed are recorded in `new_data_qnames` as variables beginning with `Q3-4` and `HGC` with suffix of the year it was recorded. In addition the variables beginning with `HGCREV` contain the revised data. We choose to use this revised May data because it seemed to have less missing and presumably has been checked. However, there is no revised May data for 2012, 2014, 2016, and 2018, thus, we use the ordinary May data for these years.

The `hgc` is measured and could be updated in each period of the survey. We chose to only retain the highest grade completed for each individual and derived the year when they completed it (`yr_hgc`) by finding the minimum year with the highest completed grade.

Finally, we get all of the demographic profiles of the NLSY79 cohort. We then save this data as `demog_nlsy79`.

3.1.2 Tidying employment variables

Our target variables for the employment are to obtain respondent's mean hourly wage (`mean_hourly_wage`), the number of jobs (`number_of_jobs`) and the total hours of work per week (`total_hours`) for each survey year. As the data only reports up to 5 jobs for each respondent, the maximum number of jobs is capped at 5.

From 1979 to 1987, `new_data_qnames` only contains one version of hours worked per week for each job (in the variables with names starting with `QES-52A`). From 1988 onward, we selected the total hours worked per week, including hours working from home (`QES-52D`). However, in 1993, this variable was missing for the first and last job so we selected to use `QES-52A` instead. In addition, 2008 only had jobs 1-4 for the `QES-52D` variable, so we use only these.

The hourly wages are in the variables beginning with `HRP` in `new_data_qnames`. As a respondent may have multiple jobs, the `mean_hourly_wage` is computed as a weighted average of the hourly wage for each job with the number of hours worked for each job as weights (provided that the information on number of hours is available); if number of hours worked for any job is missing, then the `mean_hourly_wage` is computed as a simple average of all available hourly wages. Prior to computing the mean hourly wage, we undertook a number of steps to treat unusual observations as described below:

- If the hourly rate is recorded as 0, we set wage as missing.
- If the total hours of worked for the corresponding job is greater than 84 hours, we set the wage and hour worked as missing.

The number of jobs (`number_of_jobs`) for each respondent per year is computed from the number of non-missing values of hourly wage *and* hours worked. If either the hourly wage or hours worked is missing, we do not tally this.

```
#> # A tibble: 10 x 6
#>       id  year mean_hourly_wage total_hours number_of_jobs is_wm
#>    <int> <dbl>            <dbl>       <int>          <dbl> <lgl>
#> 1      1  1979             3.28          38              1 FALSE
#> 2      1  1981             3.61          NA              1 FALSE
#> 3      2  1979             3.85          35              1 FALSE
#> 4      2  1980             4.57          NA              1 FALSE
#> 5      2  1981             5.14          NA              1 FALSE
#> 6      2  1982             5.71          35              1 FALSE
#> 7      2  1983             5.71          NA              1 FALSE
#> 8      2  1984             5.14          NA              1 FALSE
#> 9      2  1985             7.71          NA              1 FALSE
#> 10     2  1986             7.69          NA              1 FALSE
```

The employment and demographic variables are then joined. These data are further filtered to the cohort who completed education up to 12th grade and participated in at least five rounds in the survey. We save the resultant wage data on this cohort as `wages`.

3.2 Initial data analysis

According to Huebner, Vach, and Cessie (2016), initial data analysis (IDA) is the step of inspecting and screening the data after being collected to ensure that the data is clean, valid, and ready to be deployed in the later formal statistical analysis. Moreover, Chatfield (1985) argues that the two main objectives of IDA are data description, which is to assess the structure and the quality of the data, and model formulation without any formal statistical inference.

In this paper, we conduct an IDA or a preliminary data analysis to assess the consistency of the data with the cohort information that the NLSY provides. In addition, we also aim to find the anomaly in the wages values using this approach. We mainly use graphical summaries to do the IDA using `ggplot2` (Wickham 2016) and `brolgar` (Tierney, Cook, and Prvan 2020).

As stated previously, the respondents' ages ranged from 12 to 22 when first interviewed in 1979. Hence, we validate whether all of the respondents were in this range. Additionally, the NLSY also provides the number of the survey cohort by their gender (6,403 males and 6,283 females) and race (7,510 Non-Black/Non-Hispanic; 3,174 Black; 2,002 Hispanic). To validate this, we used the `demog_nlsy79`, i.e., the data with the survey years 1979 sample. Tables 1 and 2 suggest that the demographic data we had is consistent with the sample information in the database.

In the next step, we explore the mean hourly wage data. In this case, we only explore the wages data of respondents that have the highest completed grade of up to 12th grade. We employ visualization techniques to perform the IDA as described next.

We randomly take 20 samples from the data and plot them, as shown in Figure 2. It shows that these respondents have a lot of variability in wages, for example, the IDs in panel numbers 5, 7, and 11. The plot also implies that the samples have a different pattern of mean hourly wages. Some have had flat wages for years but had a sudden increase in one particular year, then it gone down again, while the others experienced an upsurge in their wage, for instance, the IDs in panel 9. However, when checking the summary plots (Figure

Table 1: The frequency table of the age at the start of the survey (?CHECK) in the full NSLY79 data

Age	Number of individuals
15	1,265
16	1,550
17	1,600
18	1,530
19	1,662
20	1,722
21	1,677
22	1,680

Table 2: The contingency table for gender and race for the full NLSY79 data.

Gender	Race			Total
	Hispanic	Black	Non-Black, Non-Hispanic	
Male	1,000 (15.62%)	1,613 (25.19%)	3,790 (59.19%)	6,403 (100.00%)
Female	1,002 (15.95%)	1,561 (24.84%)	3,720 (59.21%)	6,283 (100.00%)
Total	2,002 (15.78%)	3,174 (25.02%)	7,510 (59.20%)	12,686 (100.00%)

3), we found that some observations had exceptionally high wages. Some of them, for example respondents in Figure 3 C had experienced unusual wages only in certain year.

The anomalies are also found in the total hours of work, where some observations reported as having worked for 420 hours a week in total. According to Pergamit et al. (2001), one of the flaws of the NLSY79 employment data is that the NLSY79 collects the information of the working hours since the last interview. Thus, it might be challenging for the respondents to track the within-job hours' changes between survey years, especially for the respondents with fluctuating working hours or seasonal jobs. It even has been more challenging since 1994, where the respondents had to recall two year periods. This shortcoming might also contribute to the fluctuation of one's wages data.

3.2.1 Replacing extreme values

As part of the IDA, which is the model formulation, we build a robust linear regression model to treat the extreme values in the data. Robust linear regression yields an estimation robust to the influence of noise or contamination (Koller 2016). It also aims to detect the contamination by weighting each observation based on how "well-behaved" they are, known as robustness weight. Observations with lower robustness weight are suggested as an outlier by this method (Koller 2016).

Since we work with longitudinal data, we build the model for each ID instead of the overall data. The robust mixed model could be the best model to be employed in this case. However, this method is too computationally and memory expensive, especially for a large data set, like the NLSY79 data. Therefore, the model for each ID is built utilizing the `nest` and `map` function from `tidyr` (Wickham 2020) and `purrr` (Henry and Wickham 2020), respectively. The full code for this is shown at `CENSORED/articles/input-to-valid-data.html` but also described in detail next.

We build the model using the `rlm` function from `MASS` package (Venables and Ripley 2002). We set the `mean_hourly_wage` and `year` as the dependent and predictor, respectively. Furthermore, we use M-Estimation with Huber weighting, where the observation with a slight residual gets a weight of 1, while the larger the residual, the smaller the weight (less than 1) (UCLA: Statistical Consulting Group 2021). However, the challenging part of detecting the anomaly using the robustness weight is determining the weight threshold in which the observations are considered outliers. Moreover, it should be noted that not all the outliers are due to an error. Instead, it might be that one had reasonably increasing or decreasing wages in a particular period.

To minimize the risk of mistakenly identifying an outlier as an "erroneous outlier", we simulate some thresholds and study how they affect the data. We find that 0.12 is the most reasonable value to be the threshold to
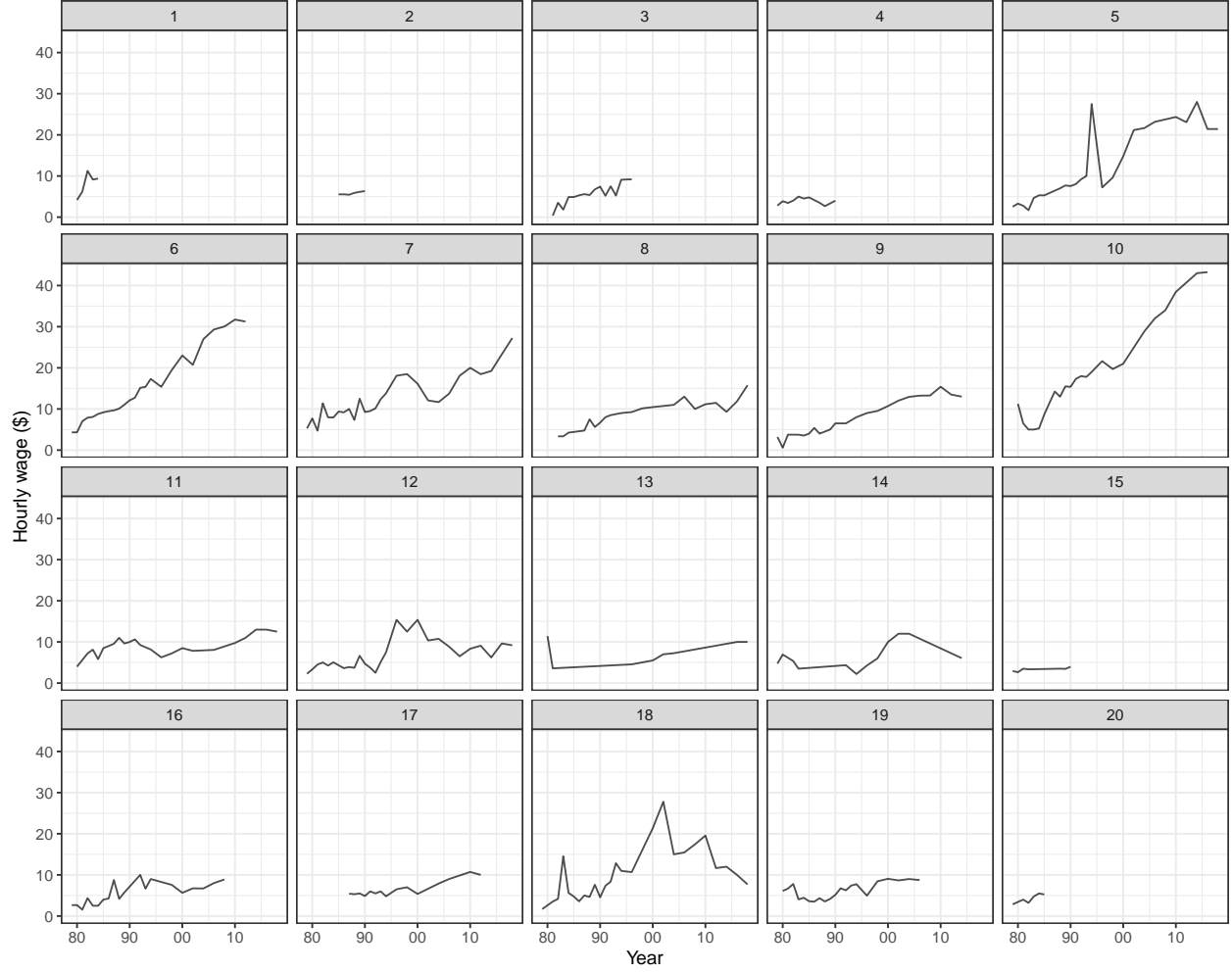
Figure 2: Longitudinal profiles of wages for a random sample of 20 individuals in the refreshed data. Most, but not all, individuals here experienced increasing wages over time, and several have experienced considerable fluctuation in wages. Some individuals are only measured for a short period.
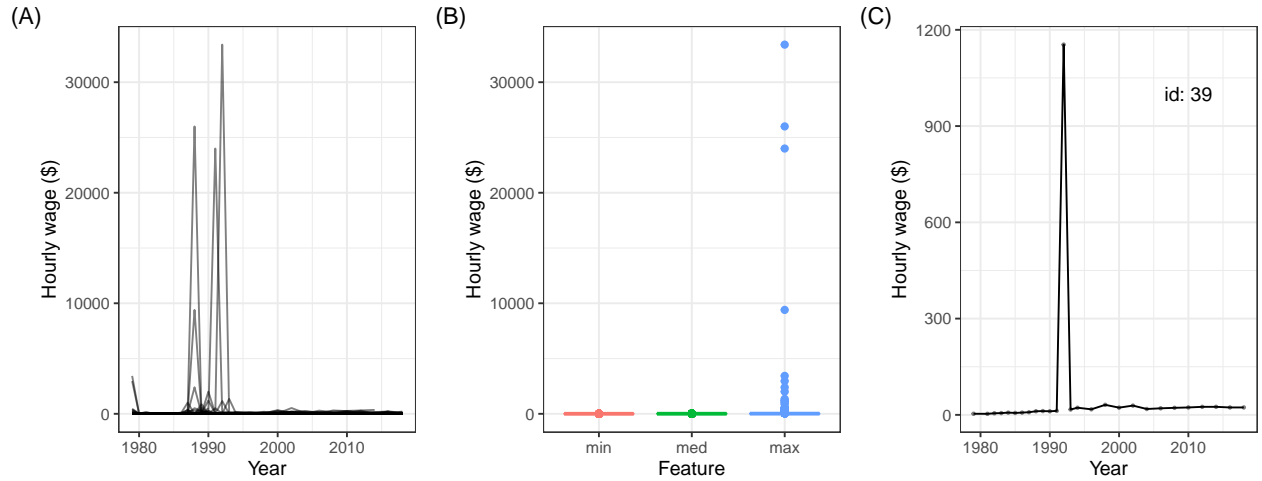

Figure 3: Summary plots to check the cleaned data reveal more cleaning is necessary: longitudinal profiles of wages for all individuals 1979-2018 (A), boxplots of minimum, median, and maximum wages of each individual (B), and one individual (id=39) with an unusual wage relative to their years of data (C). Some values of hourly wages are unbelievable, and some individuals have extemely unusual wages in some years.
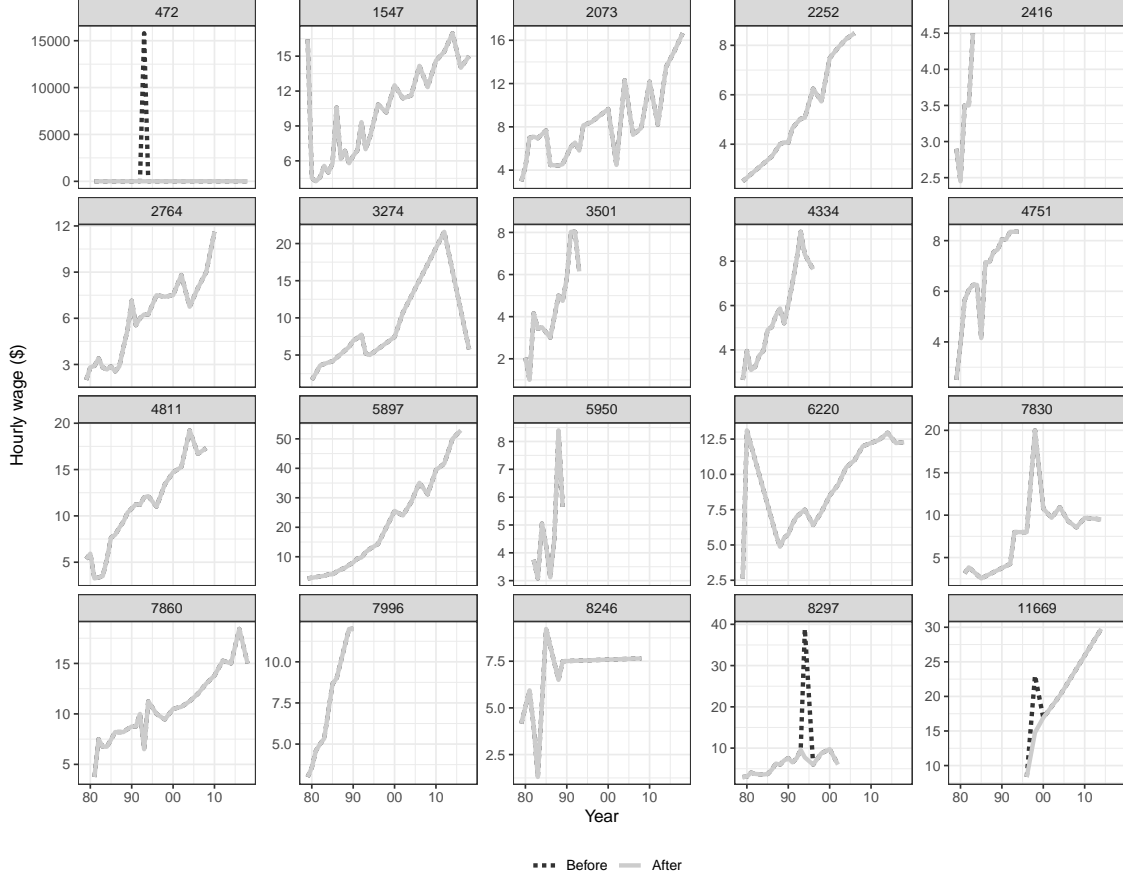
Figure 4: Comparison between the original (black dots) and the corrected (solid grey) mean hourly wage for a sample of individuals. A robust linear model prediction was used to correct mean hourly wages value. We can see that some extreme spikes, corresponding to implausible wages, have been replaced with values more similar to wages in neighboring years, but otherwise the profiles are not changed. Some spikes might remain when wage vaues are plausible.

minimize that drawback's risk because it still captures the sensible spikes in the data. In other words, we keep maintaining the natural variability of the wages while minimizing anomalies because of the error in the data recording. After deciding the threshold, we impute the observations whose weights are less than 0.12 with the models' predicted value. We then flag those observations in a new variable called `is_pred`.

Figure 4 shows the mean hourly wage before and after the extreme values are replaced. It implies that the fluctuation can still be observed in the data after the treatment. However, the large spikes, which are considered "erroneous outliers", are already eliminated from the data. Hence, the model produces a data set with a more reasonable degree of fluctuation.

Further, Figure 5 A shows that after eliminating the extreme values, the highest value has decreased to be around $350. The spikes are still observed but not as extreme as the original data set. In Figure 5 B, we plot the three features of mean hourly wages, namely the minimum, median, and maximum value. We still can see some extreme values in maximum wages, but consider it as a natural variability of the data. In Figure 5 C, we can see that after imputing the extreme value in ID=39, we can see how the the wages change over the years more clearly.

Finally, we save the imputed data and set the appropriate data type for the variables.

3.3 Recap

Figure 6 summarizes the steps taken to go from raw to input to valid data (M. van der Loo and de Jonge 2018) to create a refreshed wages data set.
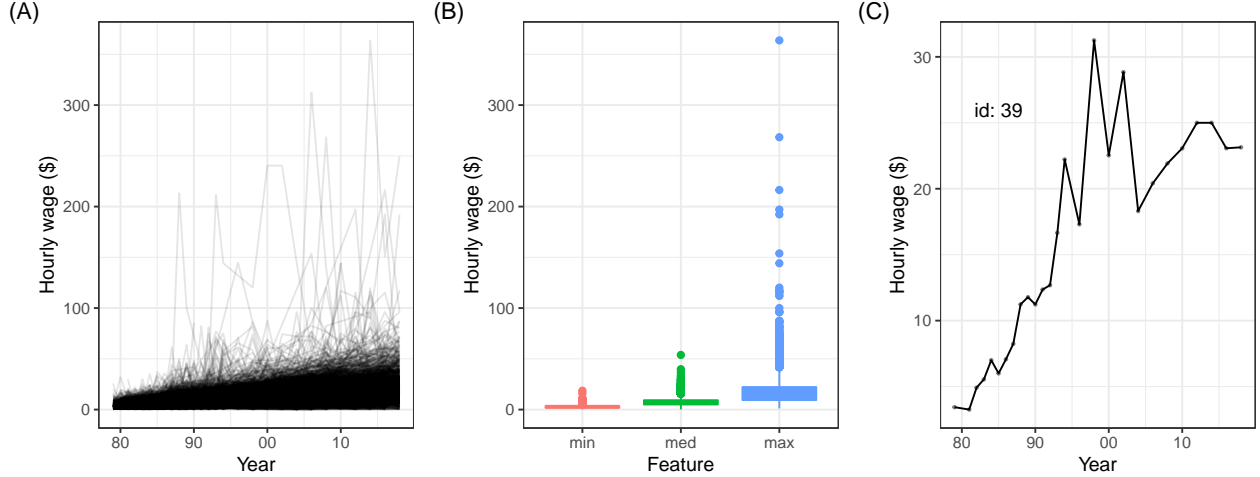
9

Figure 5: Re-make of the summary plots of the fully processed data suggest that it is now in a reasonable state: longitudinal profiles of wages for all individuals 1979-2018 (A), boxplots of minimum, median, and maximum wages of each individual (B), and one individual with an unusual wage relative to their years of data (C).

4 Comparison of refreshed with the original data

Now is the time to see how close we have come to the original textbook data. The original set contains wages of high school dropouts (Singer and Willett 2003) from 1979 through to 1994. The refreshed data set for comparison is also wages on high school drop outs from 1979 to 2018. The original set is available in the R package `brolgar` and the refreshed data is available in the R package `[CENSORED]`.

There are two aspects of the original data that make a direct comparison difficult. The time variable provided is "experience in the workforce". It is not clear how this is calculated. One would expect that there is a record of the day the individual first started a job, and this is used to adjust the year of collection. We have not been able to find this information in the database. Secondly, wages were inflation-adjusted to 1990 prices, which is not done for the refreshed data.

The treatment of unlikely wages differ in the refreshed data. In the original data by Singer and Willett (2003), wages greater than \$75 are set to be missing. However, in recent context, this value is too low to be set as the maximum threshold. We opted to use the weights from a robust linear regression to determine what should be set as missing value as described in Section 3.2.1.

Figure 7 shows a comparison of the two sets. (A direct ID matching is not possible.) There are 888 individuals in the original and 1,188 individuals in the refreshed data. This suggests some individuals were removed in the original data. In addition, in the original set the racial breakdown was 246 Black, 204 Hispanic and 438 White participants, while in the refreshed data there are 346 Black, 219 Hispanic, and 623 White participants, so the proportions have not been exactly reproduced. On education, in the refreshed data there are very few individuals with less than a 12th grade education, which is different from the original. In the original data there were 366 individuals with up to 8th grade and 522 9th-12th grade, as compared to 967 with 12th grade in the refreshed data.

5 Summary

This paper has described the stages to take a particular open data set and make it a textbook data set, ready for the classroom or research. In the first stage, we showed the steps performed to get the data from the NLSY79 database. The data format needed conversion to tidy format, and this was described. After that, an initial data analysis was conducted to investigate and screen the quality of the data. We found and fixed the anomalous observations in wages using a robust linear regression model. Finally a comparison is made between the original and refreshed data sets.

The data cleaning process is documented and the code has been made available. These provide the opportunity to again refresh the textbook data as updated data is published in the NLSY79 database. Determining the
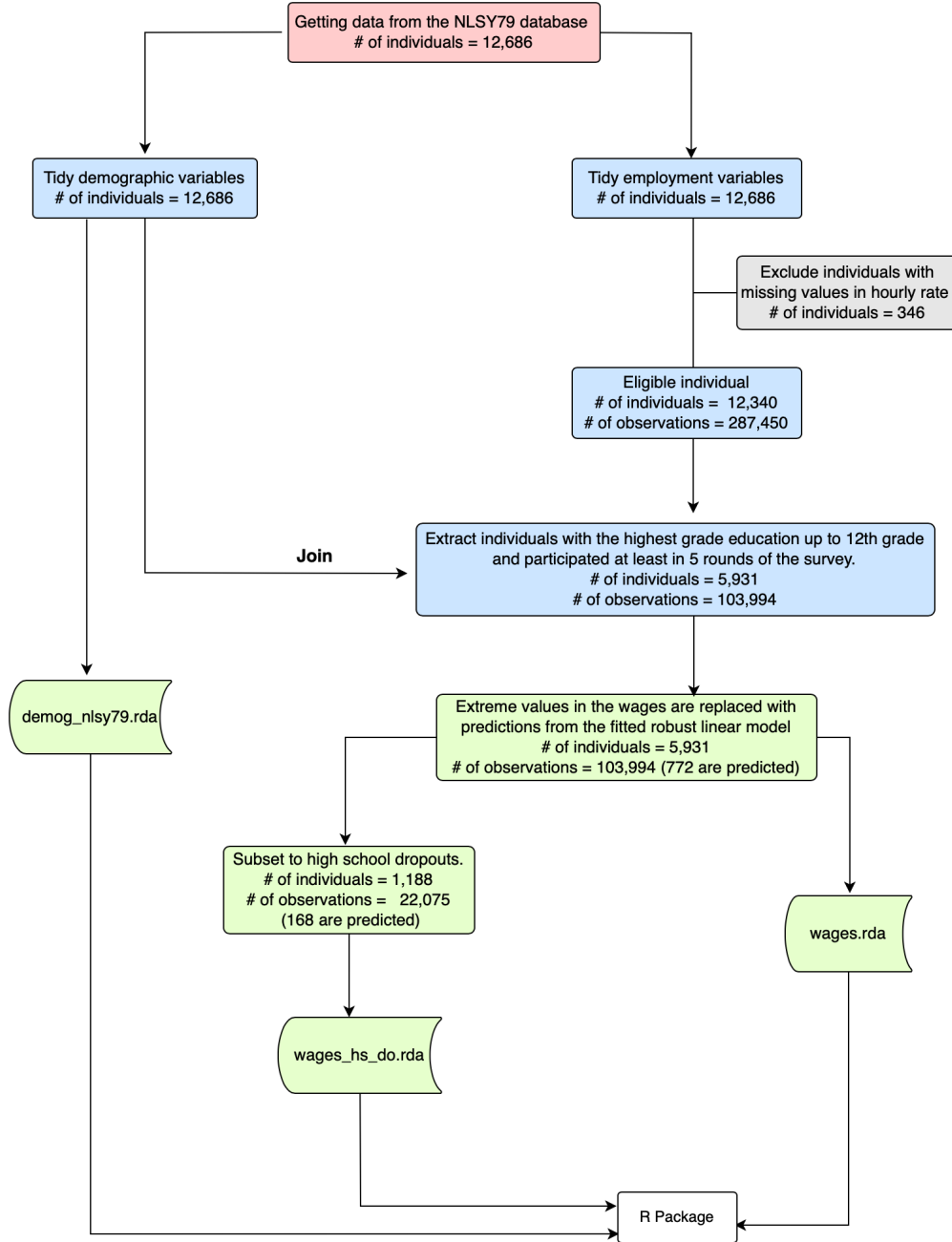
Figure 6: The stages of data cleaning from the raw data to get three datasets contained in `[CENSORED]`. "#
of individuals" means the number of respondents included in each stage, while "# of observations" means the
number of rows in the data. The color represents the stage of data cleaning in statistical value chain (M. P. J.
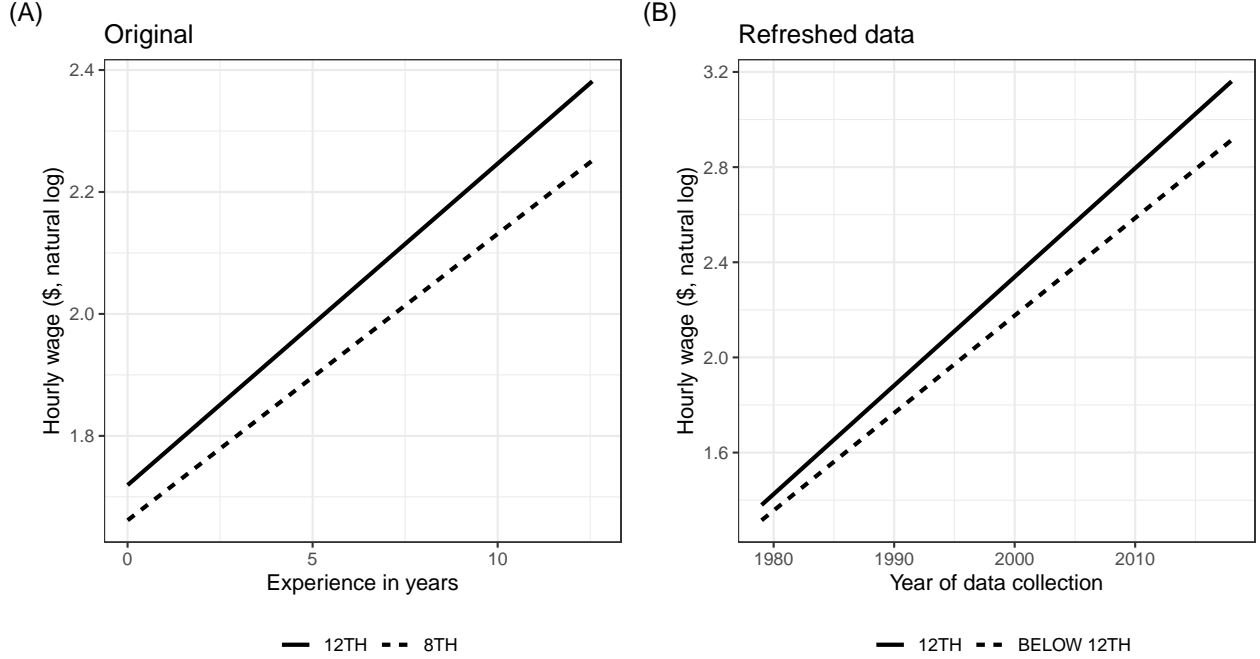van der Loo and de Jonge 2021). Pink, blue, and green represent the raw, input, and valid data, respectively.

Figure 7: Comparison of original textbook example (A) with refreshed data (B). The original data was inflation-adjusted to 1990 prices and the individual's time of collection was converted to a length of experience in the workforce, which makes it difficult to precisely compare the two sets.

appropriate robustness weight to threshold the anomalous observations is documented, and a `shiny` (Chang et al. 2020) app is provided to assist. The current subset is made available in a new R package, called `[CENSORED]`.

Various difficulties were encountered in trying to refresh the data, which include:

- determining which records should be downloaded from the database.
- there are many errors in the data, e.g. hourly wages greater than \$30,000 per hour.
- there is no explicit variable in the database recording high school dropout, which means we needed to compare date of 12th grade with the individual's age.
- calculating experience in the workforce would require knowing the time of first job within the first year the individual was recorded.

Ultimately, the refreshed data is reasonably similar to the original, but unsatisfactorily far from it. The last step required would be to inflation-adjust wages, but this is better to do with each wave of new data added, so that it is relative to the last date in the data.

Some readers may disagree with our decisions made to produce the refreshed textbook data and may have better insight than us in producing a more appropriate textbook data. We do not assert that we have produced the best textbook data, but rather we describe our journey to provide a reasonable textbook data set. All code and documentation are provided for transparency, and future updates of the `[CENSORED]` package may contain additional variables, or filters of the full set, if it is deemed important.

Finally, for the data providers we recommend that a validation system with clear rules is added on data entry, and that alternative output formats, such as a tidy format would help users make better use of the resource. The problem with many of the wages records is that there are implausible values, or confusion on how to record wages for multiple jobs. These values can be validated with simple checks at data entry. Providing an open data resource also is accompanied with the responsibility that the data, especially data that is as valuable as this, is reliable. Users need to be able to trust the data.

6 Acknowledgements

We would like to thank Aarathy Babu for the insight and discussion during the writing of this paper.

The entire analysis is conducted using `R` (R Core Team 2020) in RStudio IDE using these packages: `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2020), `readr` (Wickham and Hester 2020), `tidyr` (Wickham 2020), `stringr` (Wickham 2019), `purrr` (Henry and Wickham 2020), `brolgar` (Tierney, Cook, and Prvan 2020), `patchwork` (Pedersen 2020), `kableExtra` (Zhu 2019), `MASS` (Venables and Ripley 2002), `janitor` (Firke 2020), and `tsibble` (Wang, Cook, and Hyndman 2020). The paper was generated using `knitr` (Xie 2014) and `rmarkdown` (Xie, Dervieux, and Riederer 2020).

7 Supplementary Materials

- **Codes**: R script to reproduce data tidying and cleaning is available in this page.
- **R Package**:`[CENSORED]` is a data container R package that contains 3 datasets, namely the high school mean hourly wage data, high school dropouts mean hourly wage data, and demographic data of the NLSY79 cohort. This package can be accessed here.
- **shiny app**: An interactive `shiny` web app to visualise the effect of selecting different weight threshold for substituting the wages data to its predicted value from a fit of the robust linear regression model. This app can be accessed here with the source code provided here.

8 Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the supplementary materials.

References

Bureau of Labor Statistics, U.S. Department of Labor. 2021. "National Longitudinal Survey of Youth 1979 Cohort, 1979-2016 (Rounds 1-28)." Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. *shiny: Web Application Framework for R.* `https://CRAN.R-project.org/package=shiny`.

Chatfield, C. 1985. "The Initial Examination of Data." *Journal of the Royal Statistical Society. Series A. General* 148 (3): 214–53.

Cooksey, Elizabeth C. 2017. "Using the National Longitudinal Surveys of Youth (NLSY) to Conduct Life Course Analyses." In *Handbook of Life Course Health Development*, edited by Richard M. Lerner Neal Halfon Christoper B. Forrest, 561–77. Cham: Springer. https://doi.org/`https://doi.org/10.1007/978-3-319-47143-3_23`.

Dasu, Tamraparni, and Theodore Johnson. 2003. *Exploratory Data Mining and Data Cleaning.* Wiley Series in Probability and Statistics. Hoboken: WILEY.

Firke, Sam. 2020. *janitor: Simple Tools for Examining and Cleaning Dirty Data.* `https://CRAN.R-project.org/package=janitor`.

Fullilove, M. T. 1998. "Comment: Abandoning "Race" as a Variable in Public Health Research–an Idea Whose Time Has Come." *American Journal of Public Health* 88 (9): 1297–98.

Grimshaw, Scott D. 2015. "A Framework for Infusing Authentic Data Experiences Within Statistics Courses." *The American Statistician* 69 (4): 307–14. `https://doi.org/10.1080/00031305.2015.1081106`.

Henry, Lionel, and Hadley Wickham. 2020. *purrr: Functional Programming Tools.* `https://CRAN.R-project.org/package=purrr`.

Huebner, Marianne, Werner Vach, and Saskia le Cessie. 2016. "A Systematic Approach to Initial Data Analysis Is Good Research Practice." *The Journal of Thoracic and Cardiovascular Surgery* 151 (1): 25–27.

Ilk, Ozlem. 2004. "Exploratory Multivariate Longitudinal Data Analysis and Models for Multivariate Longitudinal Binary Data." PhD thesis, Iowa State University. `https://doi.org/10.31274/rtd-180813-11012`.

Kim, A. Y, C. Ismay, and J. Chunn. 2018. "The Fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses." *Technology Innovations in Statistics Education* 11 (1). `https://doi.org/10.5070/T511103589`.

Koller, Manuel. 2016. "robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models." *Journal of Statistical Software* 75 (6): 1–24.

Open Knowledge Foundation. 2021. "Open Definition. Defining Open in Open Data, Open Content, and Open Knowledge." 2021. `http://opendefinition.org/od/2.1/en/`.

Pedersen, Thomas Lin. 2020. *patchwork: The Composer of Plots.* `https://CRAN.R-project.org/package=patchwork`.

Pergamit, Michael R., Charles R. Pierret, Donna S. Rothstein, and Jonathan R. Veum. 2001. "Data Watch: The National Longitudinal Surveys." *The Journal of Economic Perspectives* 15 (2): 239–53.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Singer, Judith D, and John B Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford u.a: Oxford Univ. Pr.

Stodel, Megan. 2020. "Stop Using Iris." `https://www.meganstodel.com/posts/no-to-iris/`.

Tierney, Nicholas, Di Cook, and Tania Prvan. 2020. *brolgar: BRowse Over Longitudinal data Graphically and Analytically in R.* `https://github.com/njtierney/brolgar`.

Tukey, John W. (John Wilder). 1977. *Exploratory Data Analysis.* Addison-Wesley Series in Behavioral Science. Reading, Mass.: Addison-Wesley Pub. Co.

UCLA: Statistical Consulting Group. 2021. "Robust Regression | R Data Analysis Examples." February 2021. `https://stats.idre.ucla.edu/r/dae/robust-regression/`.

van der Loo, Mark P. J., and Edwin de Jonge. 2021. "Data Validation Infrastructure for R." *Journal of Statistical Software* 97 (10): 1–31. `https://doi.org/10.18637/jss.v097.i10`.

van der Loo, Mark, and Edwin de Jonge. 2018. *Statistical Data Cleaning with Applications in r.*

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s.* Fourth. New York: Springer. `http://www.stats.ox.ac.uk/pub/MASS4`.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics* 29 (3): 466–78. `https://doi.org/10.1080/10618600.2019.1695624`.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software, Articles* 40 (1): 1–29. `https://doi.org/10.18637/jss.v040.i01`.

———. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23.

———. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. `https://ggplot2.tidyverse.org`.

———. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations.* `https://CRAN.R-project.org/package=stringr`.

———. 2020. *tidyr: Tidy Messy Data.* `https://CRAN.R-project.org/package=tidyr`.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. `https://doi.org/10.21105/joss.01686`.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation.* `https://CRAN.R-project.org/package=dplyr`.

Wickham, Hadley, and Jim Hester. 2020. *readr: Read Rectangular Text Data.* `https://CRAN.R-project.org/package=readr`.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. `http://www.crcpress.com/product/isbn/9781466561595`.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook.* Boca Raton, Florida: Chapman; Hall/CRC. `https://bookdown.org/yihui/rmarkdown-cookbook`.

Zhu, Hao. 2019. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* `https://CRAN.R-project.org/package=kableExtra`.