What Makes People Earn More Money*

An Analysis of the Income and Influencing Factors of Young adults in the United States

Zifeng Zhu

27 April 2022

Abstract

As income inequaility of individuals arises as a more significant issue in all countries across the world nowadays, it is important for us to understand the leading factors of this social problem. This paper is going to explore how gender, race, education level and other factors influence the income of individuals in the United States in 1994. After our data analysis, we found that men generally had higher income than women, and non-black, non-hispanic americans earned more money. In addition, citizens with higher education degrees had a higher income level than others. We learned that factors including gender and race can lead to a huge gap between people's income, however, today the influence of gender and race on income level is gradually decreasing.

Contents

*Code and data are available at: https://github.com/zhuzifen/sta304_final_paper

1 Introduction

The author Ram pointed out that the income inequality of individuals in the United States has become a hot spot in contemporary society, and high income inequality among states is likely to reduce the economic growth rate of the United States (Ramet al., 2015). Income equality is important. Typically, factors that affect an individual's income include the individual's state, age, education, and hours worked. The purpose of studying personal income and its influencing factors is to possibly point out ways to narrow the gap of personal income to achieve income balance.

Our report focuses on the analysis of the 1994 Longitudinal Survey of American Youth (NLSY) wave data. The data provides details on youth's education, ethnicity, employment and income. The raw data has 37 variables. This report explores a variety of key factors related to youth income in the United States. First, we extracted only some of the research variables of interest including year of birth, ethnicity, gender, type of residence, highest degree, total income over the past calendar years, and total hours per week. In this report the analysis tool we are using is R (R Core Team 2020)

In addition to our introduction, this report has a data [2] section below, which visually shows the relationship between American income level, gender, race, type of residence and education in 1994. The following method [**??**] section analyzes the above problem with unpaired two-sample t-test, one-way ANOVA and linear regression with dummy variables. Next, the result [**??**] section summarizes the findings of this report. The discussion [5] at the end discusses the findings of this article from a global perspective, expanding on it relationship to the world. It also describes the weakness of the data we used in this report, as well as methods for future improvements.

2 Data

We willing using 1994 wave of NLYSY1979, a survey data of young adults in the United States. The data comes from https://www.nlsinfo.org/content/cohorts/nlsy79/other-documentation/codebook-supplement/nlsy79-attachment-4-fields-study#business. The extracted data of interest are described as follows:

Variable.Names	Type	Description
YEAR_OF_BIRTH	discrete	year of birth
RACE	categorical	race
GENDER	categorical	sex
URBAN_RURAL_	categorical	Whether lives in an urban or rural area
EDU_DEGREE	categorical	The highest degree R has completed
INCOME_	categorical	income
HOURS_WORKED	categorical	Total number of hours per week from all jobs reported that year

For missing values in the dataset, imputation or deletion of missing values is used for processing. After obtaining a clean dataset, first perform an exploratory analysis of the data, including generating additional variable metrics for all categorical variables, obtaining descriptive statistics for the entire sample, obtaining summary statistics grouped by category, and using bar charts, histograms, etc. Graphs and scatter plots to visualize data.

2.1 Data Collection

This data is from the NLSY which can be found on the NLSY website where NLSY means the National Longitudinal Survey of Youth. This report is about the United States of America. The data was collected

Table 2: Income by Gender statistic

GENDER	N	Mean	Sd
FEMALE	3349	18239.95	14920.75
MALE	3670	27416.62	20294.18

Table 3: Income by Race statistic

RACE	N	Mean	Sd
BLACK	1981	18791.37	15124.91
HISPANIC	1292	21404.86	16139.06
NON-BLACK, NON-HISPANIC	3746	25847.25	20317.65

in 1994 in USA. This survey contains 12686 observations, 6283 of them are females, and 6403 are males. A total of 5908 households had been selected for this survey. The participants were born between 1957 and 1964. At the time of first interview, respondents' ages ranged from 14 to 22. The respondents were 53 to 62 at the time of their 2018 interviews (the most recent survey year). Interviews were conducted annually from 1979 to 1994 and on a biennial basis thereafter. This dataset is unique and it takes a long time to collect.

2.2 Descriptive statistic

There are slightly more men than women in the data in table 2. For average earnings, men are significantly more than women by nearly $10,000.

The average income gap for young people of different races is shown in table 3. Among them, the average income of black people is obviously relatively low.

The income of American youth is different whether they live in the city or in the countryside in table 4. Young people living in cities have higher incomes, which may be related to higher consumption levels in cities relative to rural areas.

The average earnings of young people with different highest education levels vary widely (table 5). And obviously the higher the education or professional degree, the higher the average income.

2.3 Visualizations

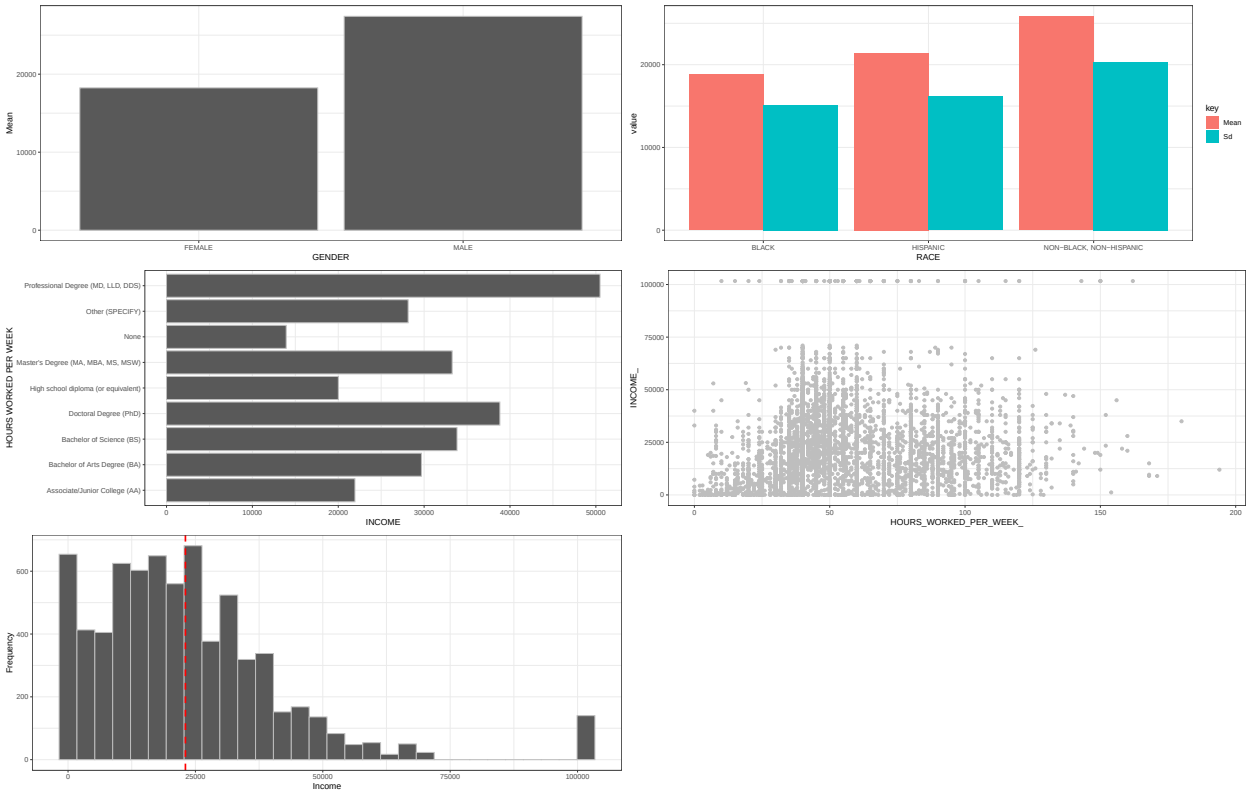
From the grid arrangement diagram it can be concluded that:
(1) The average income of men is higher than that of women.
(2) Blacks have the lowest mean and standard deviation of income.
(3) Professional degrees (MD, LLD, DDS) have the highest income, followed by PhD. The minimum is no education. The higher the education, the higher the average income.
(4) There is no obvious linear relationship between income and total hours worked per week.
(5) Income's distribution is right-tailed. In further analysis, the data need to be log-transformed to adjust their distribution to a normal distribution.

Table 4: Income by Place statistic

URBAN_RURAL_	N	Mean	Sd
0: RURAL	1363	19599.31	15712.97
1: URBAN	5656	23866.82	19028.74

Table 5: Income by Place statistic

EDU_DEGREE	N	Mean	Sd
Associate/Junior College (AA)	778	21942.26	15773.41
Bachelor of Arts Degree (BA)	413	29684.31	21175.39
Bachelor of Science (BS)	724	33829.62	22232.34
Doctoral Degree (PhD)	42	38822.12	27125.78
High school diploma (or equivalent)	3719	20007.28	15235.22
Master's Degree (MA, MBA, MS, MSW)	460	33262.98	23165.49
None	668	13935.84	11999.38
Other (SPECIFY)	162	28135.13	22838.31
Professional Degree (MD, LLD, DDS)	53	50483.64	32020.13



3 Methods

The methods had been used here are unpaired two-sample t-test, one-way ANOVA and linear regression with dummy variables, etc.

4 Results

In this section, I will explore the analysis that interests me by dividing it into three questions. Hypothesis testing and regression analysis methods are mainly used.

4.1 Question 1

Is there a significant difference between American Youth adult income of male and female?

Here we use Two Sample t-test.

4.1.1 Hypothesis

Test if (mean(male income)-mean(female income)) or the difference of the income by gender is different from zero. Null Hypothesis: mean(male income) − mean(income) = 0 Alternative Hypothesis: mean(income) − mean(income) ≠ 0

4.1.2 Output

The test output is calculated as follows.

```
##
##  Welch Two Sample t-test
##
## data:  male$INCOME_ and female$INCOME_
## t = 21.708, df = 6719.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8347.988 10005.345
## sample estimates:
## mean of x mean of y
##  27416.62  18239.95
```

4.1.3 Statistical Interpretation of the Results

Assuming that the significance level is selected to be 0.05. p-value is the observed probability of the null hypothesis to happen.

Due to p-value $< 2.2e-16$, which smaller than the level of significance (0.05). We reject the null hypothesis, which shows that at the significance level of 0.05, American Youth average income from the 1994 wave in different genders There is a significant difference between the average income of American Youth men and women.

4.2 Question 2

Is there a significant difference between American Youth adult income by race?

Here, I want to compare income differences between different races, which is a similar problem to comparing income differences between different genders. But since race falls into 3 categories, one-way ANOVA is used here.

Null Hypothesis: There are no differences in variable values between groups of samples.
Alternative Hypothesis: At least one group of samples has significantly different values of the variable.

4.2.1 Hypothesis

Null Hypothesis: mean(BLACK income) = mean(HISPANIC income) = mean(NON-BLACK, NON-HISPANIC income), the average incomes of American Youth of different races are all equal.

Alternative Hypothesis: mean (BLACK income), mean (HISPANIC income), mean (NON-BLACK, NON-HISPANIC income are not all equal, and the average incomes of American Youth of different races are not all equal.

4.2.2 Output

The anova output is calculated as follows.

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## RACE           2 6.873e+10 3.437e+10   103.3 <2e-16 ***
## Residuals   7016 2.335e+12 3.328e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

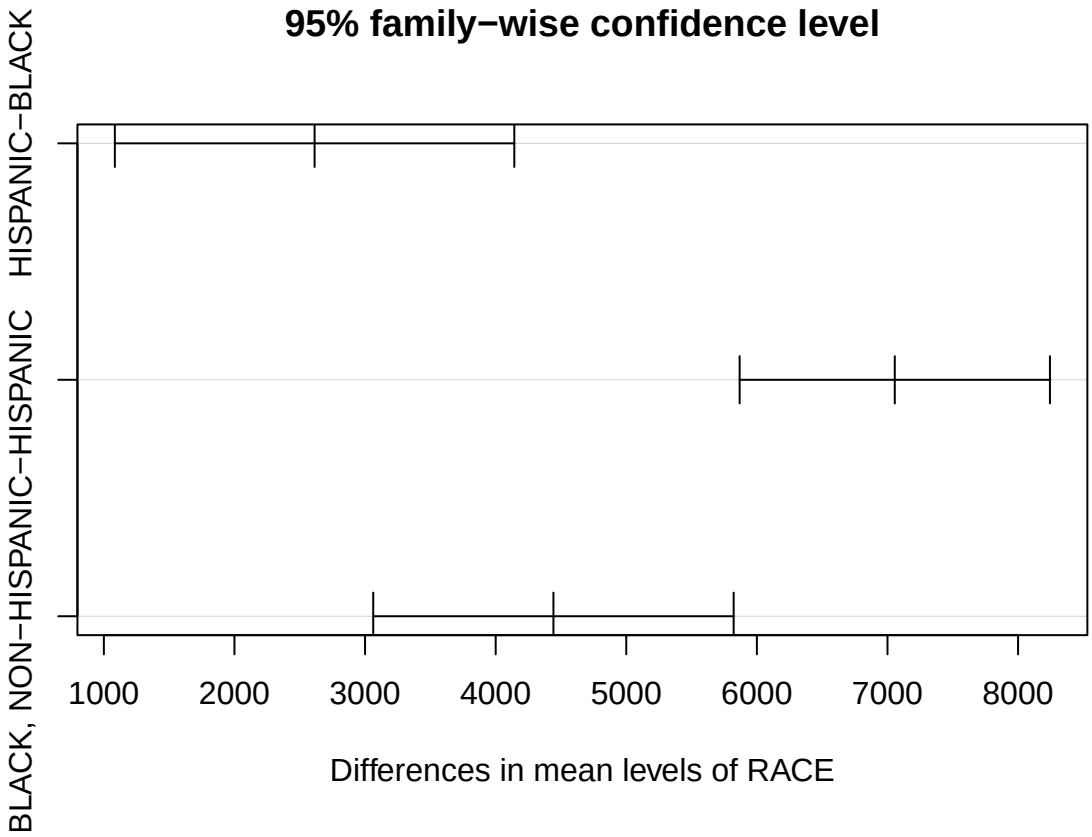
4.2.3 Statistical Interpretation of the Results

Due to p-value $< 2.2e-16$ which smaller than the level of significance (0.05). We reject the null hypothesis, which shows that at the significance level of 0.05, the average income of American Youth from the 1994 wave varies significantly between different races.

4.2.4 Multiple comparisons

The analysis of variance only tells us that the three groups are different, but it does not tell us which two groups have obvious differences. At this time, we need to use the TukeyHSD function to perform a multiple comparison analysis of the mean.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = INCOME_ ~ RACE, data = NLSY.data)
##
## $RACE
##                                       diff      lwr      upr     p adj
## HISPANIC-BLACK                    2613.496 1084.138 4142.855 0.0001843
## NON-BLACK, NON-HISPANIC-BLACK     7055.883 5867.799 8243.967 0.0000000
## NON-BLACK, NON-HISPANIC-HISPANIC 4442.387 3062.563 5822.211 0.0000000
```



It is observed that the differences in the pairwise comparisons between the three races are significant.

4.3 Question 3

Considering the education, how does Urban against Rural affect the income of American Youth adult?

Here I use the linear regression method and add the dummy variable "URBAN_RURAL" to the independent variables of the linear regression. For dummy variables, the value is 1 for Urban and 0 for Rural.

Linear regression with dummy variables to compare each metric to the underlying category. Here, the predictors of the linear regression are educational attainment (EDU) and the dummy variable we focus on "URBAN_RURAL," and the response variable is the income of American young adult(INCOME_).

The summary results of the model are as follows:

```
##
## Call:
## lm(formula = INCOME_ ~ EDU + URBAN_RURAL, data = NLSY.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43162 -11399  -2136   8146  87472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14180.9      516.3  27.468  < 2e-16 ***
## EDU            3218.0      119.6  26.907  < 2e-16 ***
```

```
## URBAN_RURAL1   3237.4      530.9   6.098 1.13e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17550 on 7016 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.1008
## F-statistic: 394.5 on 2 and 7016 DF,  p-value: < 2.2e-16
```

The equation of the fitted model is as follows:

$$Income = 14180.9 + 3218\text{Eductation} + 3237.4\text{URBAN}$$

The above results indicate that the income of American youth living in Urban has increased by 3,532.34 relative to living in Rural.

5 Discussion

5.1 Conclusions

This report analyzes the income of American youth in 1994 and its influencing factors.

First, the data set was processed and the key variables were extracted, and the statistical summary tables and histograms, bar graphs, and scatter plots were visualized to explore the relationship between income and its influencing variables. Conclusion: The average income of men is higher than that of women. Blacks have the lowest average income and standard deviation, followed by HISPANIC, and non-blacks and non-Asian people have the highest incomes. Professional degrees (MD, LLD, DDS) have the highest income, followed by doctoral degrees.

Then in the main analysis part, I asked 3 questions to find answers using hypothesis testing and linear regression analysis. First, use a one-sample t-test to test whether the average income of American youth during the 1994 wave was equal to 25,000. The conclusion is that at the significance level of 0.01, the average income of young Americans in the 1994 wave was not equal to 25,000. Whether it is greater than or less than 25,000, we need to conduct a one-tailed test to determine, this is the direction of future exploration. Then use an unpaired two-sample t test to compare the income differences of American youths of different genders. The conclusion is that at the significance level of 0.01, there is a significant difference between the average income of young men in the United States and the average income of women. Then use one-way analysis of variance to compare the income differences of American youths of different races. The conclusion is that the pairwise comparisons between the three races are all significant. Then use simple linear regression analysis to analyze how the degree of education affects American Youth's income. The conclusion is that with the addition of one unit of the educational degree variable indicators, income will increase by 3,431.97 US dollars. Finally, we used linear regression with dummy variable to analyze how Urban against Rural affects the income of American Youth. The conclusion is that the income of American youth living in Urban has increased by 3,532.34 relative to living in Rural.

5.2 Connection to the world

First of all, this report reflects the traditional society's idea that men should be working and women stay at home as a housewife. One important reason is that men are much more educated than women, and we find relevant data showing that in 1989-1990 government statistics showed that schools continued to enroll more men than women. In primary education, only 45% of students are girls. The proportion of women in school drops to 33% at the secondary level, 27% at technical colleges and 19% at university. In the society of

the 19th century, the traditional idea in that social were that men were educated and then earned money to support their families, and women were responsible for housework and reproduction.(La Verle Berry 1994)

In recent years, many countries have made considerable progress in closing the gender gap in education and income level. More than two-thirds of countries have achieved gender parity, but there are also some that have not fully achieved it, especially in Africa, the Middle East and South Asia, where girls are more likely than boys to be disadvantaged.(unicef 2020) However, with the progress of the times, the ratio of male and female enrollment has gradually balanced, indicating that people's ideas in the old society are gradually changing, and the status of men and women in society is gradually keep in a balance. Even in many countries, there are more women with higher education than men, which is a good thing in today's society.

At the same time, the report reflects that the income level of different races has significant difference back in 1994. However, nowadays the world becomes more diverse, and people from different origin and culture are more equally welcomed and competitive in the job market.

5.3 Weakness and the next steps

The set of data we analyses in the report also have some weakness. There are two main point we could improve in our further report.

First of all, the data we analyses in this report is the summarized data, not the original data set. Which means all the data we see is they have been cleaned and filtered. Although it makes our work more efficient, we cannot see the details about the original data, which may cause some small deviations in the data we analyze. If in future analysis, except for the cleaned data, it is better to use the original data set as a reference to see what details need to be modified.

Secondly, this set of data was firstly collected in 1979. From the above analysis, we can see that many Americans did not have much education when writing this questionnaire, so it they are likely to have misunderstandings or dyslexia when completing the questionnaire, which also leads to our analysis results may be biased. This is an unavoidable disadvantage due to the data collected several decades ago.

5.4 Suggestion based on the result

Based on our analysis above, there are some suggestions for people all over the world. We encourage people of all ages in every country receive education. First of all, receiving education can change people's thinking and it can also avoid some misunderstandings of children in an early age. This will also be good for the society, which will make the whole society more harmonious. Also, being educated can create more employment opportunities. A good educational background will help people to find a better job, is the most direct way to improve the life. Moreover, receiving education can also help a country to improve the economy. Higher educational background can help people get better job and been well paid. It increase the GDP of a country and decrease the poverty rate at the same time.

Appendix

A Additional details

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to collect data on fertility, family planning, and maternal and child health.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The USA Statistical Service created the dataset.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The funded the creation of dataset is United Nations Fund for Population Activities (UNFPA), the United States Agency for International Development (USAID), and other donor agencies to implement a number of population-related activities.
4. *Any other comments?*
    - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The instances represent USA. The types are: Western, Central, Greater Accra, Volta, Eastern, Ashanti, Brong Ahafo, Northern, Upper West, and Upper East, corresponding to different regions of the country.

2. *How many instances are there in total (of each type, if appropriate)?*

- There are no data show the number of instances.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The dataset does contain all possible instances.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

- Each instance consists of 5 variables: gender, age, residence, region and education level.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

- No.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- No.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- There are no relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- There are no recommended data splits.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- There are no errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

- There is no confidential data, and the dataset is available to public.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

- In Table 2.4, the word "Father dead" and "Mother dead" might be offensive because the word "dead" might remind people who lost their parents in a negative way.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- The survey was designed to obtain completed interviews of 4,500 women age 15-49. In addition, all males age 15-59 in every third selected household were interviewed, to obtain a target of 1,500 men.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- It is not possible to identify individuals in any way.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- Sensitive columns may include but not limited to: "Father dead," "Mother dead" and "Both dead."

16. *Any other comments?*

- None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- These data were acquired from survey responses,, the data were validated since if there are any mistakes workers will come back and get them again.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Manual human curation had been used.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The sample was randomly selected from households.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- There are a lot of workers that had involved in data collection, they got paid and trained by the government.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- Household listing began in August and lasted for about two months.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Not sure.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- I collect the data from the DHS project website.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- The individuals voluntarily participated in the interview with data collectors. Also, the notice of data collection is not provided.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- The individuals consented to the collection and use of their data, however, the exact language to which the individuals consented is not available.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- A mechanism to revoke consent was not provided.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- An analysis of the potential impact of the dataset and its use on data subjects was not conducted.

12. *Any other comments?*

- None. **Preprocessing/cleaning/labeling**

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- The data was obtained in PDF format originally. A summary table on page 12 from the survey PDF was converted to a useable data frame in R using the library pdftools.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

- The raw data obtained from the PDF is saved in inputs/data/raw_data.csv

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- R software is available at https://www.R-project.org/

4. *Any other comments?*

- The library used to convert data from PDF is available at https://docs.ropensci.org/pdftools/

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

- The dataset has not been used for other tasks yet.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

- http://dhsprogram.com/pubs/pdf/FR106/FR106.pdf

3. *What (other) tasks could the dataset be used for?*

- The dataset can be used for analyzing educational level among all male and female age groups in USA in 1994.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- The cleaning process is very specific to consider how this table was formatted, in which case it might not work on other tables.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- The dataset would not be appropriate for purposes other than studying the educational level in 1994 USA.

6. *Any other comments?*

- None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

- No, the dataset is available to public and being used for personal uses only.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset will be distributed using Github.

3. *When will the dataset be distributed?*

- The dataset wwill be distributed in April 2022.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- The dataset will be released under MIT license.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- There are no restrictions.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- No such controls are applicable.

7. *Any other comments?*

- None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- USA Statistical Service hosting the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- Can be contacted via github.

3. *Is there an erratum? If so, please provide a link or other access point.*

- There is no currently available erratum.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- There is no plan to update the dataset.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- There are no applicable limits as people in USA voluntarily participate.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

- The older versions will not be supported. Dataset consumers could check the updates through github commit history.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- Currently there is no mechanism for accepting other contributions.

8. *Any other comments?*

- None.

B Reference

Ram, Rati. Real and Nominal Interstate Income inequality in the United States: Further Evidence[J]. United States. International Advances in Economic Research. 2015,21.1: 131-132.

La Verle Berry, ed. 1994. *Ghana: A Country Study. Washington: GPO for the Library of Congress.*

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

unicef. 2020. *Gender and Education.*