DATA 606 Data Project

Ilya Kats

May 2017

Part 1 - Introduction:

The project was conceived in March 2017 after President Trump's release of budget proposal for fiscal year 2018. The Budget proposes to eliminate federal funding to a number of independent agencies, including the National Endowment for the Arts. The NEA was founded in 1965 and is "dedicated to supporting excellence in the arts, both new and established; bringing the arts to all Americans; and providing leadership in arts education." I personally feel that it is very short-sighted to reduce support of the arts and wanted to pay a little respect to the NEA by choosing the data set and research question that relates to the arts and to the NEA.

The NEA tracks how Americans engage with the arts through the Survey of Public Participation in the Arts. The survey was conducted six times from 1982 to 2012, which is the last year the data is available for. The survey covered five broad areas: arts attendance, reading literary works, arts consumption through electronic media, arts creation and performance, and arts learning. For this research I have decided to look into whether individual's participation in the arts can be predicted by parents' education level. Due to the limited nature of this report, I have chosen to concetrate on just one area. Specifically, **is parents' education level predictive of individual's reading of literary works?**

The analysis is split into two parts. The first part analyzes whether parents' education level is predictive of reading or not reading any works. The second parts looks if it is predictive of the number of works read.

Part 2 - Data:

The NEA survey was administered in July 2012 as a supplement to the U.S. Census Bureau's Current Population Survey (CPS), and therefore is nationally representative. The 2012 SPPA included two core components: a questionnaire used in previous years to ask about arts attendance; and a new, experimental module on arts attendance. In addition, the survey included five modules designed to capture other types of arts participation as well as participation in other leisure activities. Respondents were randomly assigned to either of the survey's core questionnaires, and then were randomly assigned to two of the remaining five SPPA modules. Most SPPA questions address arts participation that occurred in the 12-month period prior to the survey's completion. The total sample size of the 2012 SPPA was 35,735 U.S. adults, ages 18 and over, of which 31.5 percent were represented by proxy respondents. The 2012 SPPA had a household response rate of 74.8 percent.

The survey materials, including collected data is available online: https://www.arts.gov/publications/additional-materials-related-to-2012-sppa. For the project data was downloaded in STATA format from the NEA site.

The following questions were selected for analysis:

- E11a: What is the highest degree or level of school your Father completed?
- E11b: And, what is the highest degree or level of school your Mother completed?
- C1Q13a: With the exception of books required for work or school, did you read any books during the last 12 months?
- C1Q13b: If Q13b is Yes, then about how many did you read during the last 12 months?

The following options are available for the questions related to education level:

- Less than 9th grade
- Some high school

- High school graduate (or GED)
- Some college
- College graduate (BA, AB, BS)
- Advanced or graduate degree (Masters, Professional, Doctoral)

Only observations containing valid entries for father's and monther's education level were selected for analysis.

Observations with the following answers for Q13b were also eliminated - 98 (representing *Don't Know*), 99 (representing *Refused to Answer*). Finally, observations with 100 in the answer were eliminated. These outlier observations were not explained in the survey documentation, but they lie outside of survey instructions to record the number between 1 and 97.

This is **an observational study**. As such this analysis cannot be used to show causation. It can only be used to demonstrate dependency of some variables.

The data set contains 3,808 observations. Each observation is an individual responding to the survey.

There are two **explanatory variables**. They are the highest degree or level of school completed by father and the highest degree or level of school completed by mother. Both are categorical. There are two **response variables**. They are whether an individual read any books or not (categorical) and number of books read in the last 12 months (numerical).

The study is representative of the adult population of the United States. Individuals where selected at random from the population. The analysis only considers observations directly about individual taking the survey (some questions in the survey are related to spouses). Since the selection was random and the sample is less than 10% of the population, it is reasonable to assume that the observations are **independent**. Results can be generalized to the entire adult population of the United States. However, generalization to other groups - such as kids and teenagers or other countries - is not appropriate.

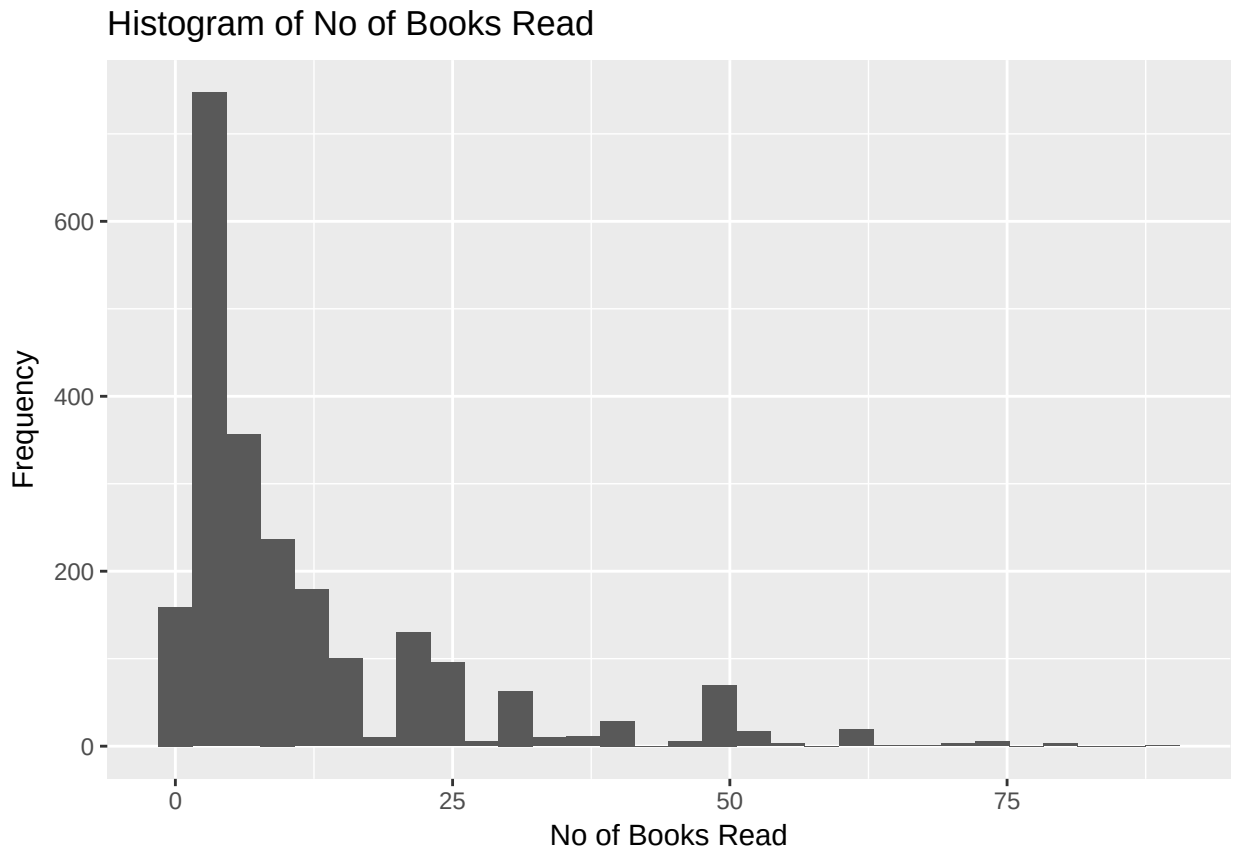
Part 3 - Exploratory data analysis:

Number of observations per each option under `DadEducation` and `MomEducation` is listed in the table below.

	Less than HS	Some Hs	HS	Some College	Undergraduate	Graduate
Dad's Education	694	469	1413	426	536	270
Mom's Education	558	446	1650	477	521	156

Number of observations per each option under `Read` is listed in the table below.

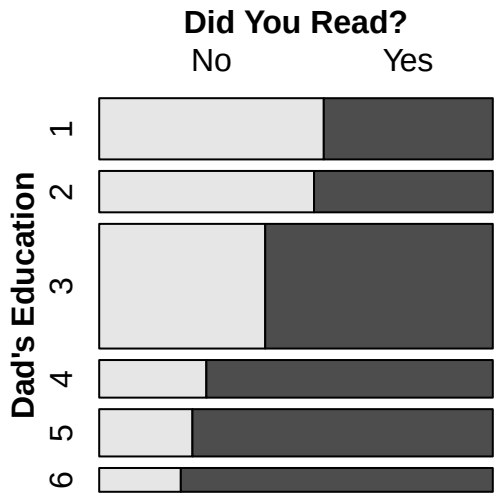
Did you read?	No of Observations
No	1547
Yes	2261

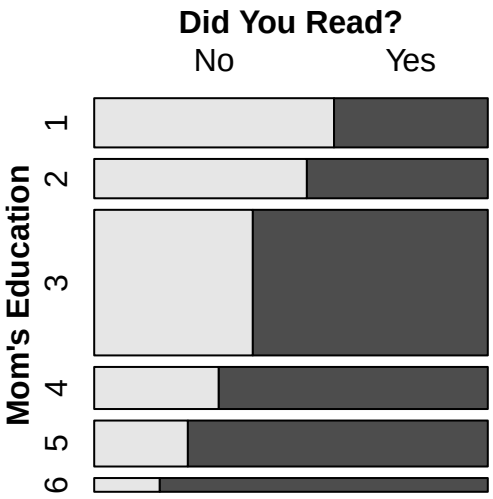


Key descriptive statistics for the number of books read is as follows:

```
##       vars    n  mean    sd median trimmed  mad min max range skew
## Books    1 2261 11.63 13.59      6    8.67 5.93   1  90    89 2.14
##       kurtosis   se
## Books     4.74 0.29
```

Mosaic plots below illustrate the breakdown between reading and not reading books per dad's and mom's education level. Visually it appears that higher parents' education level increases probability of reading books in adult life.





The plot below illustrates the spread of number of books read separated by parents' education level. The relationship (if any) is far less obvious from this representation.

Part 4 - Inference:

Part 5 - Conclusion:

References:

NEA Office of Research & Analysis, *2012 SPPA Questionnaire.*

Triplett, T. (October 2014). *2012 SPPA Public-Use Data File User's Guide.* Statistical Methods Group. Urban Institute.

Appendix A (Notes):

Appendix B (Code):

The code below was used to load original data and transform it for analysis.

```
library(foreign)
library(dplyr)
library(psych)
library(ggplot2)

# Load data.
sspa <- read.dta("C:\\Temp\\GitHub\\CUNY-DATA606\\Project\\sppa2012_public_stata.dta")

# Select relevant variables.
arts <- sspa %>%
  select(DadEducation = PEE11A, MomEducation = PEE11B,
         ConcertsFlag = PEC1Q3A, ConcertsNo = PTC1Q3B,
         BooksFlag = PEC1Q13A, BooksNo = PTC1Q13B) %>%
  filter(!is.na(DadEducation) & !is.na(MomEducation) & (!is.na(ConcertsFlag) | !is.na(BooksFlag)))

arts$BooksNo[arts$BooksFlag == "no"] <- 0
arts$ConcertsNo[arts$ConcertsFlag == "no"] <- 0

# Drop unused factor levels.
arts$DadEducation <- as.factor(as.character(arts$DadEducation))
arts$MomEducation <- as.factor(as.character(arts$MomEducation))

# Save for future use and analysis
saveRDS(arts, "arts.rds")
```