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.

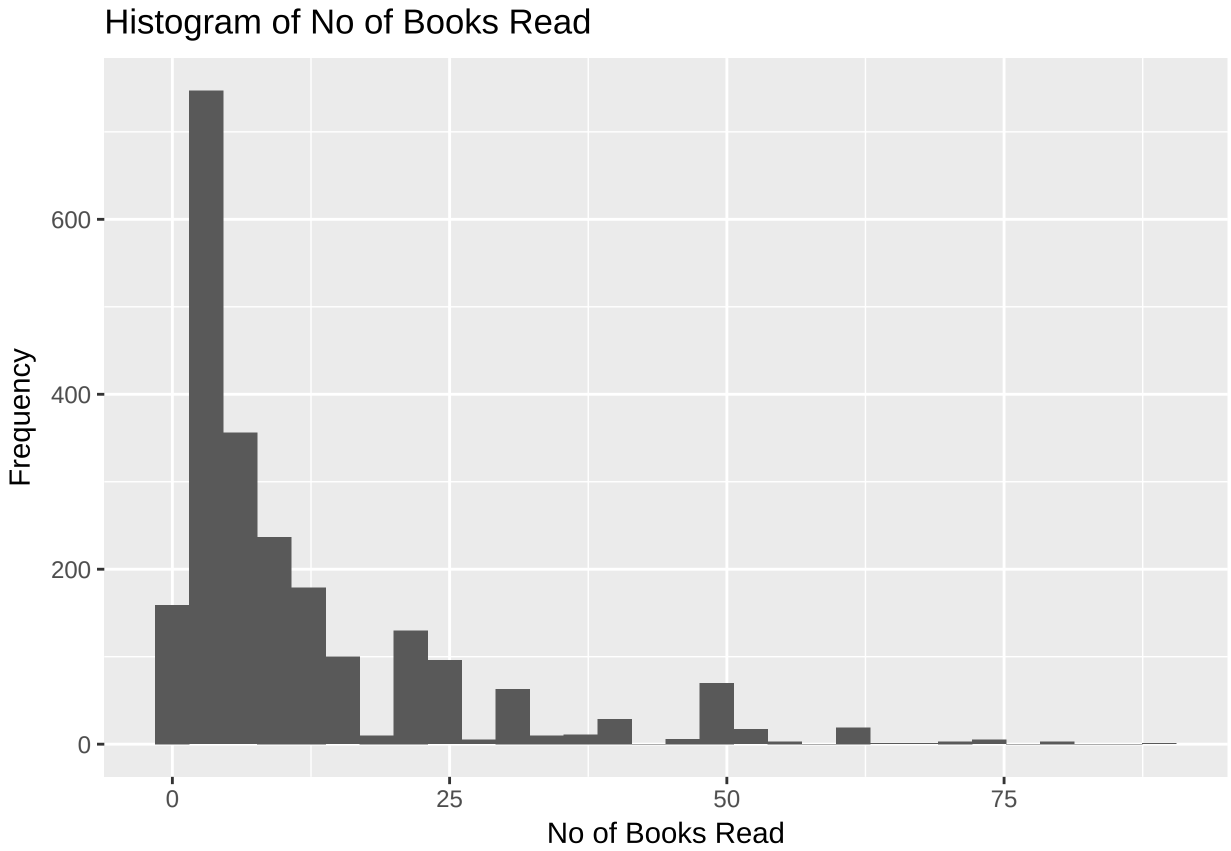
Part 3 - Exploratory data analysis:

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.

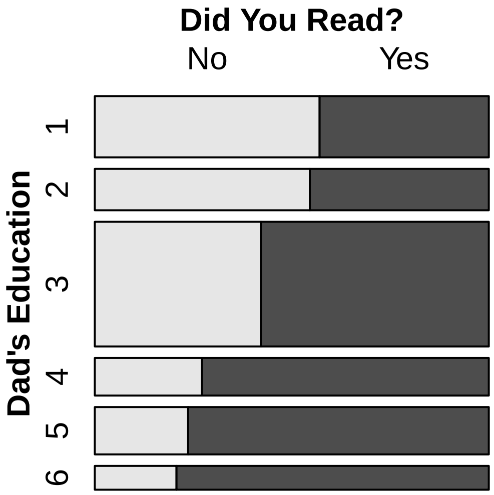
Did you read?	No of Observations
No	1547
Yes	2261

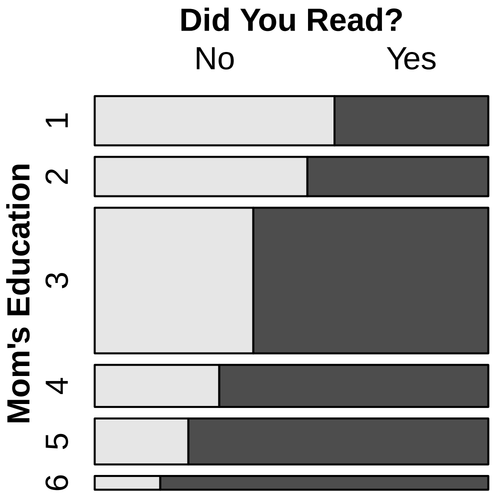


Key descriptive statistics for the number of books read is as follows:

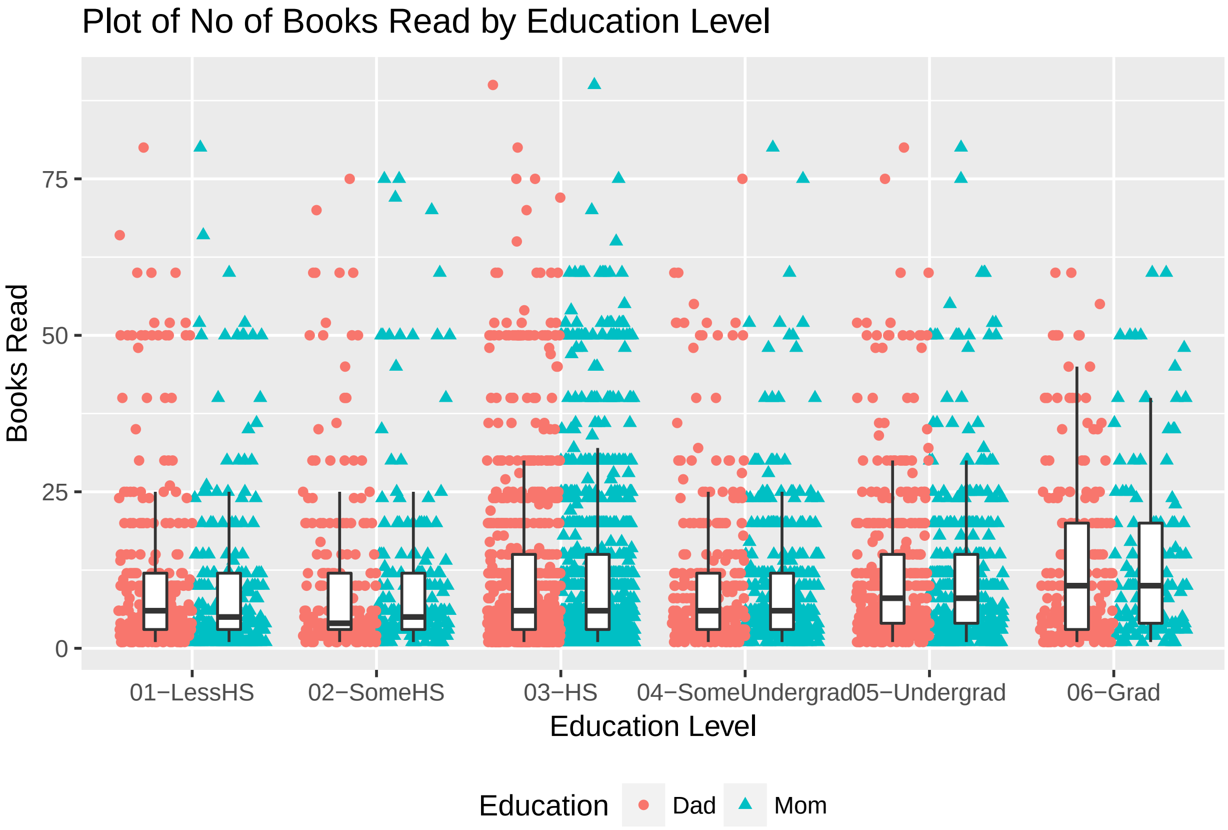
```
##       vars    n  mean    sd median trimmed  mad min max range skew
## Books    1 2261 11.63 13.59      6    8.67 5.93   1  90    89 2.14
##       kurtosis   se
## Books     4.74 0.29
```

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.



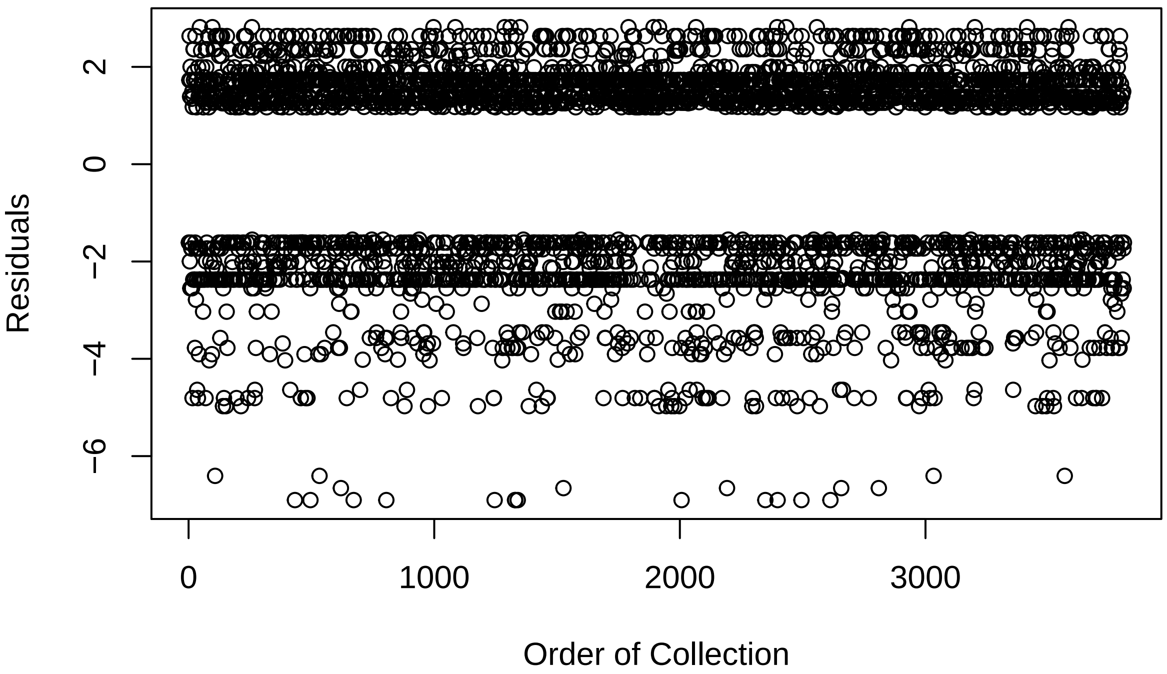
Part 4 - Inference:

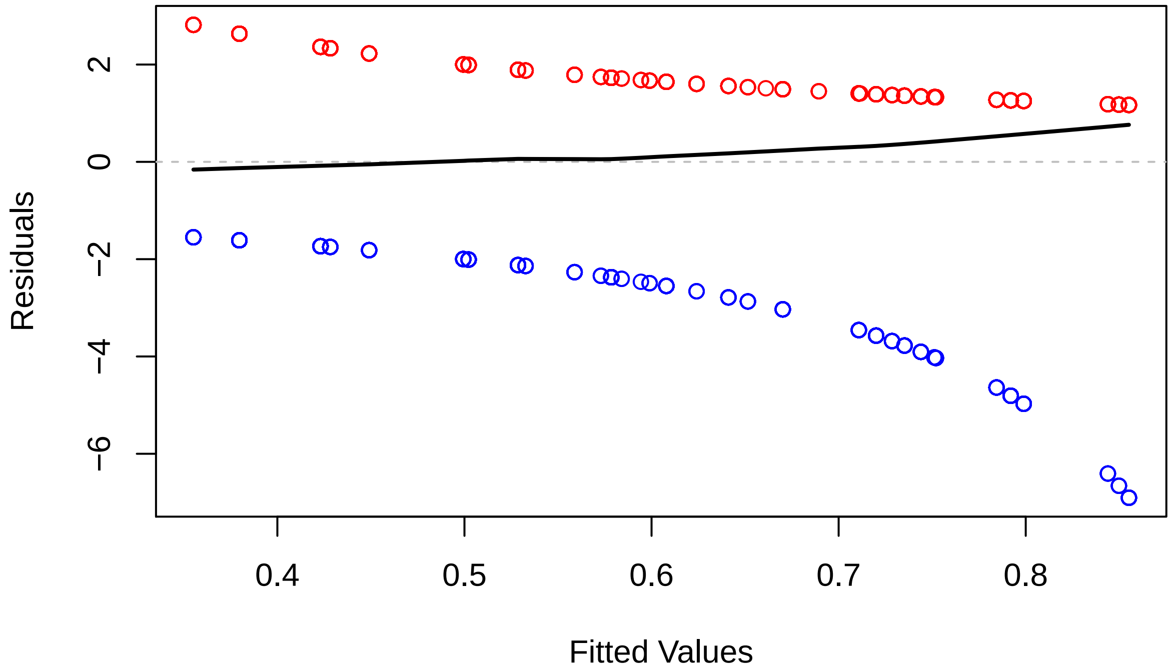
The relationship between the dichotomous dependent variable `Read` is modelled against two independent `DadEducation` and `MomEducation` using logistic regression. Depedent variable and its outcomes are **independent**. One area of concern is that two explanatory variables are not necessarily independent of each other. It is quite possible that individuals seek partners of similar education level.

```
lr <- glm(Read ~ MomEducation + DadEducation, data = books, family = "binomial")
summary(lr)
```

```
##
## Call:
## glm(formula = Read ~ MomEducation + DadEducation, family = "binomial",
##     data = books)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -1.966   -1.181    0.683    1.046    1.439
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.49084    0.09042  -5.428 5.69e-08 ***
## MomEducation02-SomeHS           0.28632    0.15253   1.877   0.0605 .
## MomEducation03-HS             0.60541    0.12907   4.690 2.73e-06 ***
## MomEducation04-SomeUndergrad  0.72736    0.16088   4.521 6.15e-06 ***
## MomEducation05-Undergrad      0.99766    0.17013   5.864 4.51e-09 ***
## MomEducation06-Grad           1.39379    0.26105   5.339 9.33e-08 ***
## DadEducation02-SomeHS        -0.10571    0.14316  -0.738   0.4603    
## DadEducation03-HS             0.20196    0.12072   1.673   0.0943 .  
## DadEducation04-SomeUndergrad  0.78462    0.15852   4.950 7.43e-07 ***
## DadEducation05-Undergrad      0.83010    0.15827   5.245 1.56e-07 ***
## DadEducation06-Grad           0.87265    0.20000   4.363 1.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 5144.3  on 3807  degrees of freedom
## Residual deviance: 4830.1  on 3797  degrees of freedom
## AIC: 4852.1
## 
## Number of Fisher Scoring iterations: 4
```

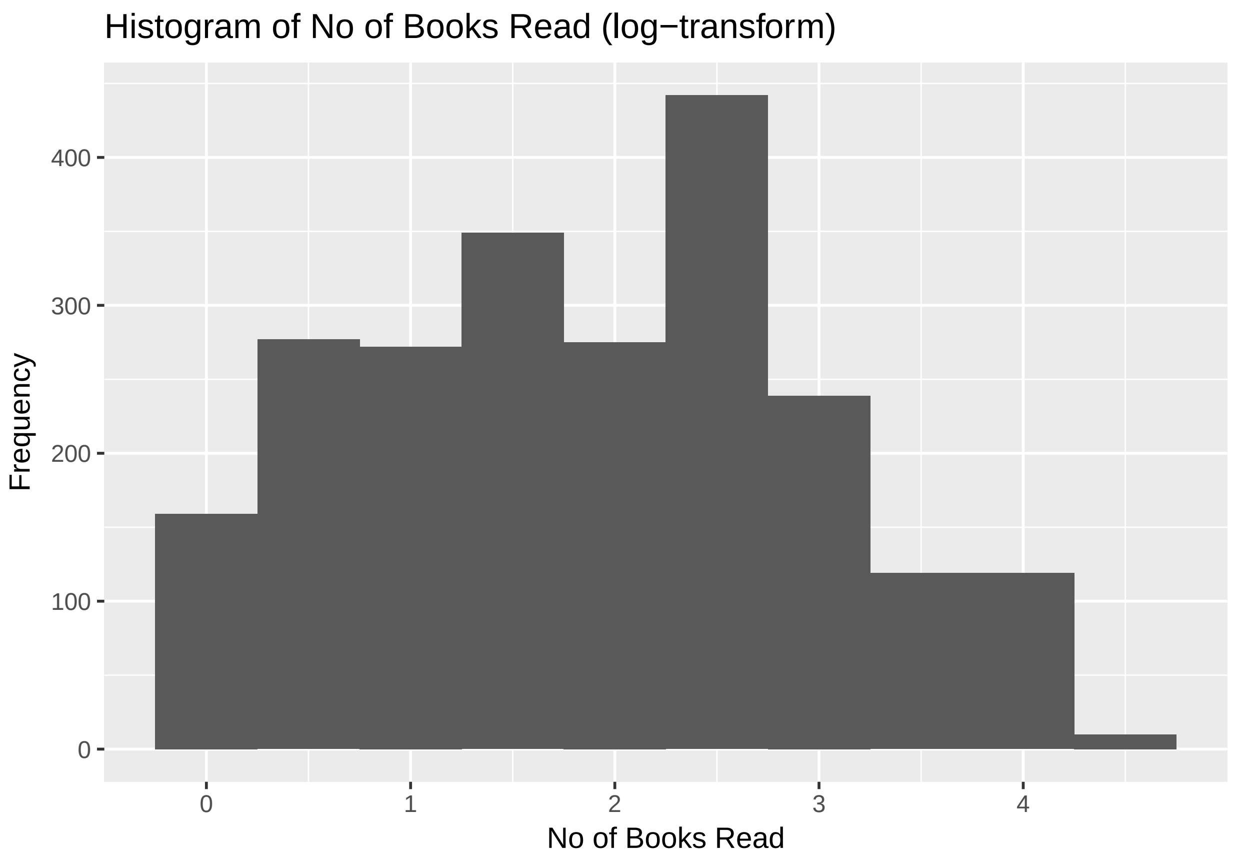
Graphs below illustrate residuals to demonstrate model fit.

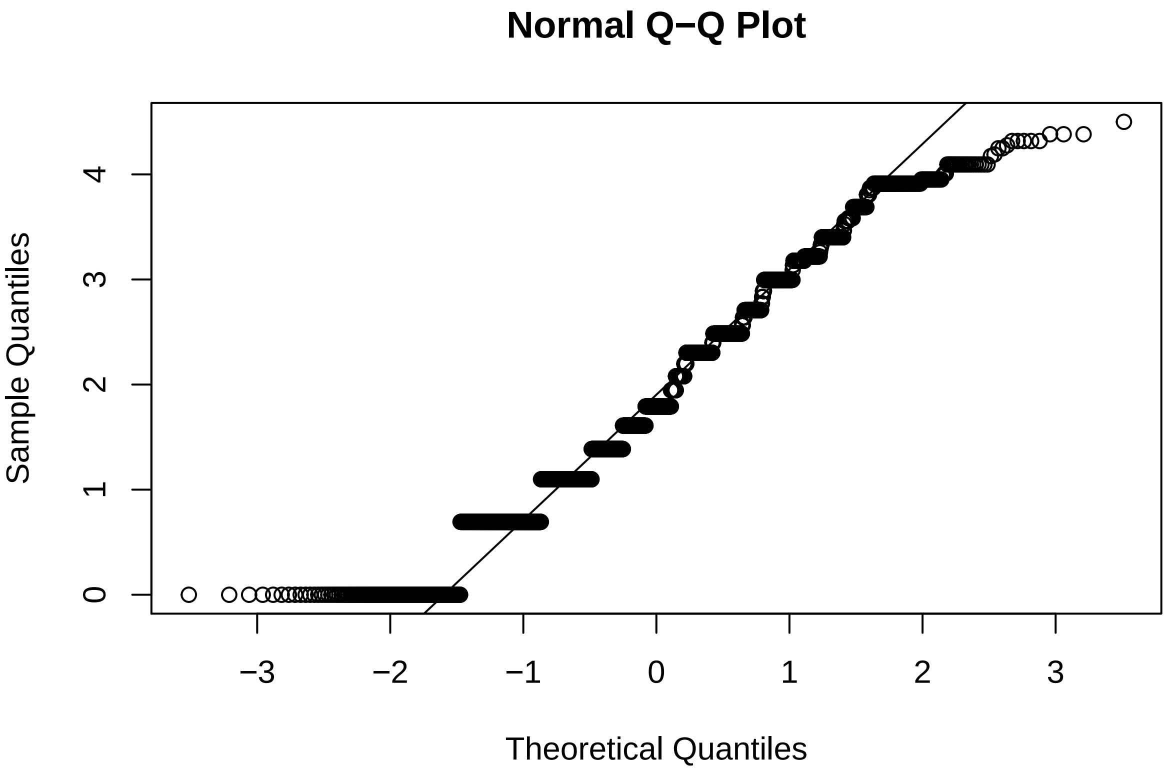




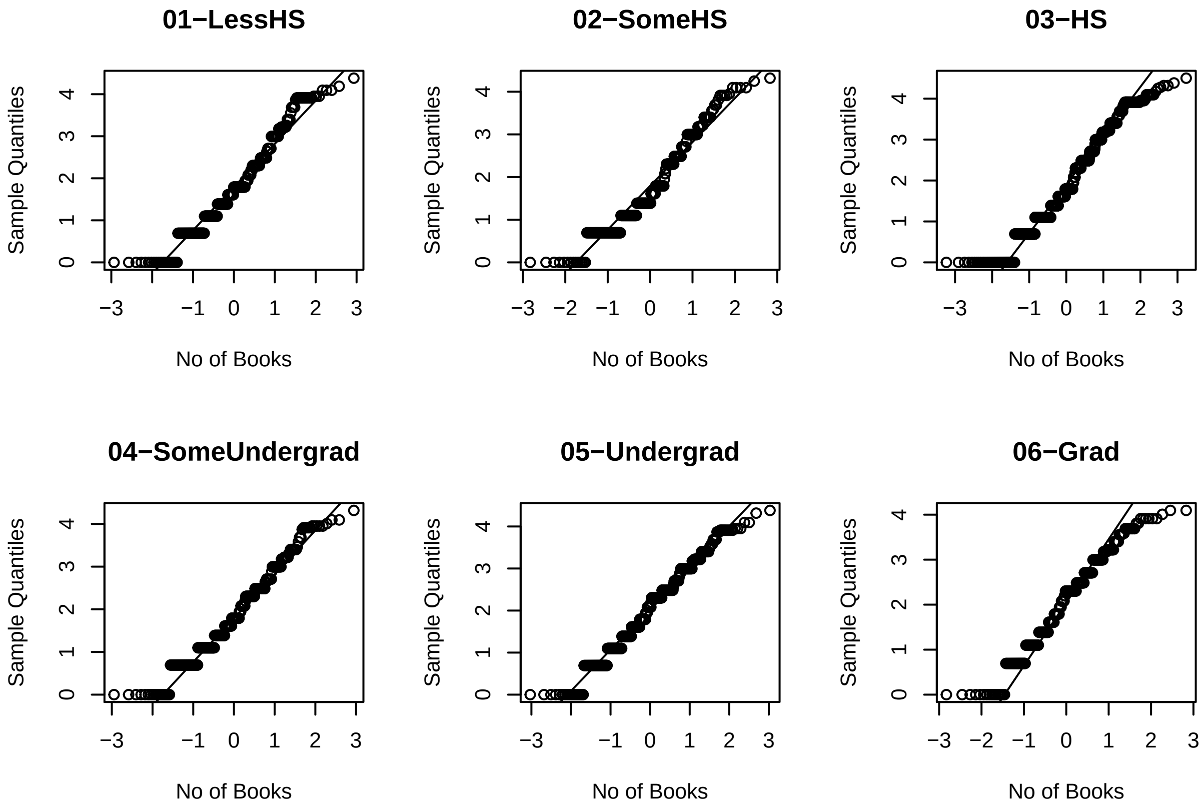
Looking at logistic regression summary table, it appears that lower education levels - *Less than 9th Grade* and *Some High School* - bear no significant relationship with the reading flag. However, higher education levels - from High School to College - are significant in predicting whether an individual is reading books recreationally.

For individuals who do read books, the following analysis checks if there is statistically significant variations in mean values across various education levels. Based on the histogram above in Part 3, the distribution of number of books read is heavily right-skewed and harldy normal. In order to perform the analysis, the data is log-transformed.

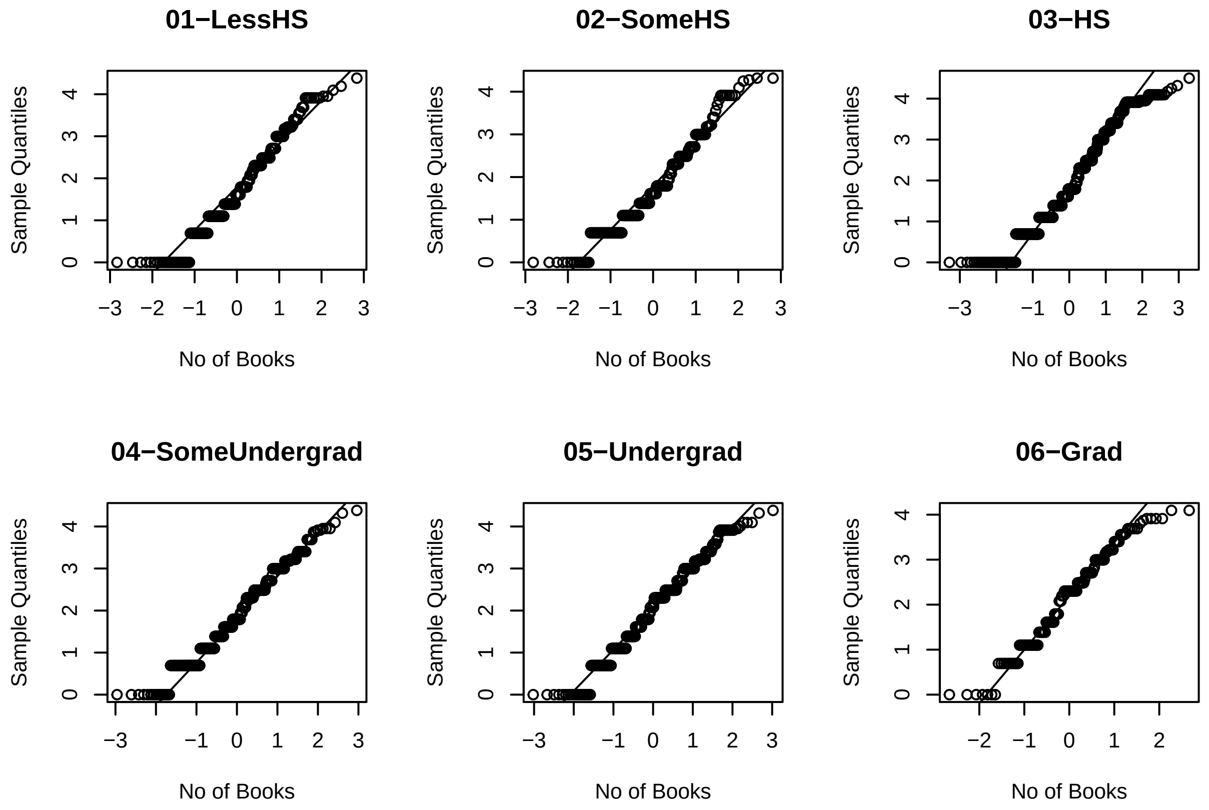

Histogram of No of Books Read (log−transform)
Frequency
400
300
200
100
0
0
1
2
3
4
No of Books Read

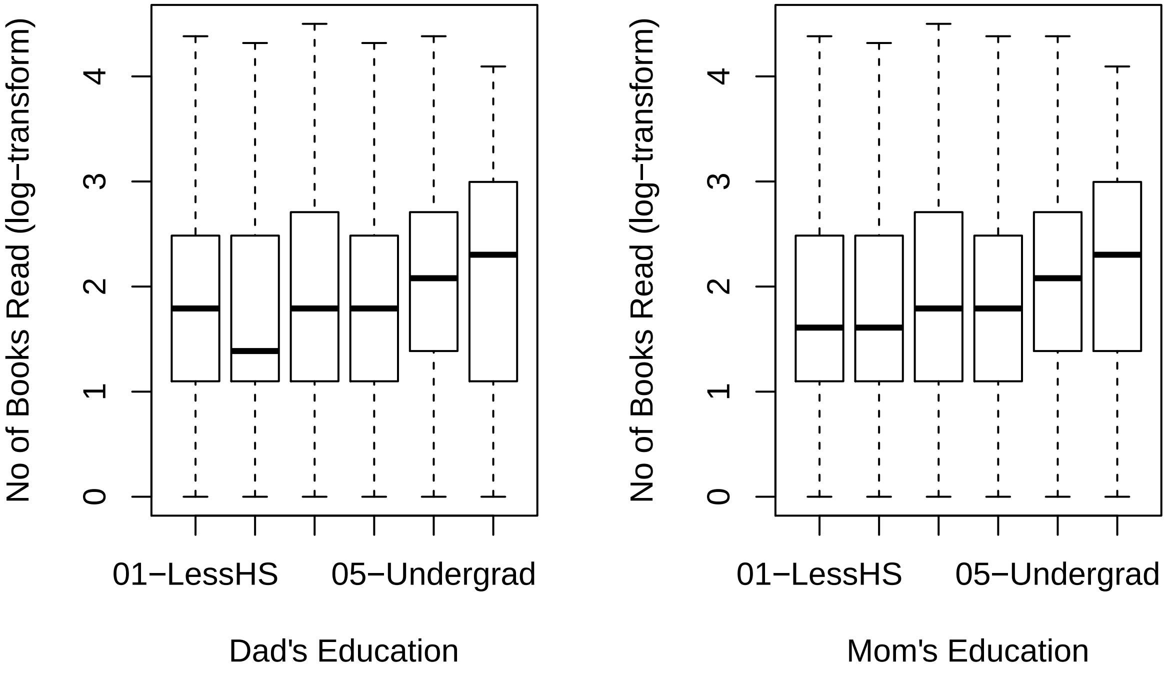

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0
1
2
3
4
−3
−2
−1
0
1
2
3

Below normal probability plots show log-transformed data per each `DadEducation` level.



Below normal probability plots show log-transformed data per each `MomEducation` level.





Based on the histogram and normal probability plots for log-transformed books data, it is reasonable to state that the distribution within the sample and within each group is **nearly normal** (with perhaps some issues towards the tails). Observations within each group and between groups are **independent**. The boxplots show that variability between groups is very similar. With these assumptions it is appropriate to use **ANOVA**.

H_0: The mean number of books read is the same for all levels of **dad's education**.

H_A: At least one mean is different.

```
summary(aov(log(b$Books) ~ b$DadEducation))
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## b$DadEducation     5     23   4.608    4.08 0.00109 **
## Residuals       2255   2547   1.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the significance level of 0.01, we reject the null hypothesis. Variation in mean number of books read is significantly different between some groups.

H_0: The mean number of books read is the same for all levels of **mom's education**.

H_A: At least one mean is different.

```
summary(aov(log(b$Books) ~ b$MomEducation))
```

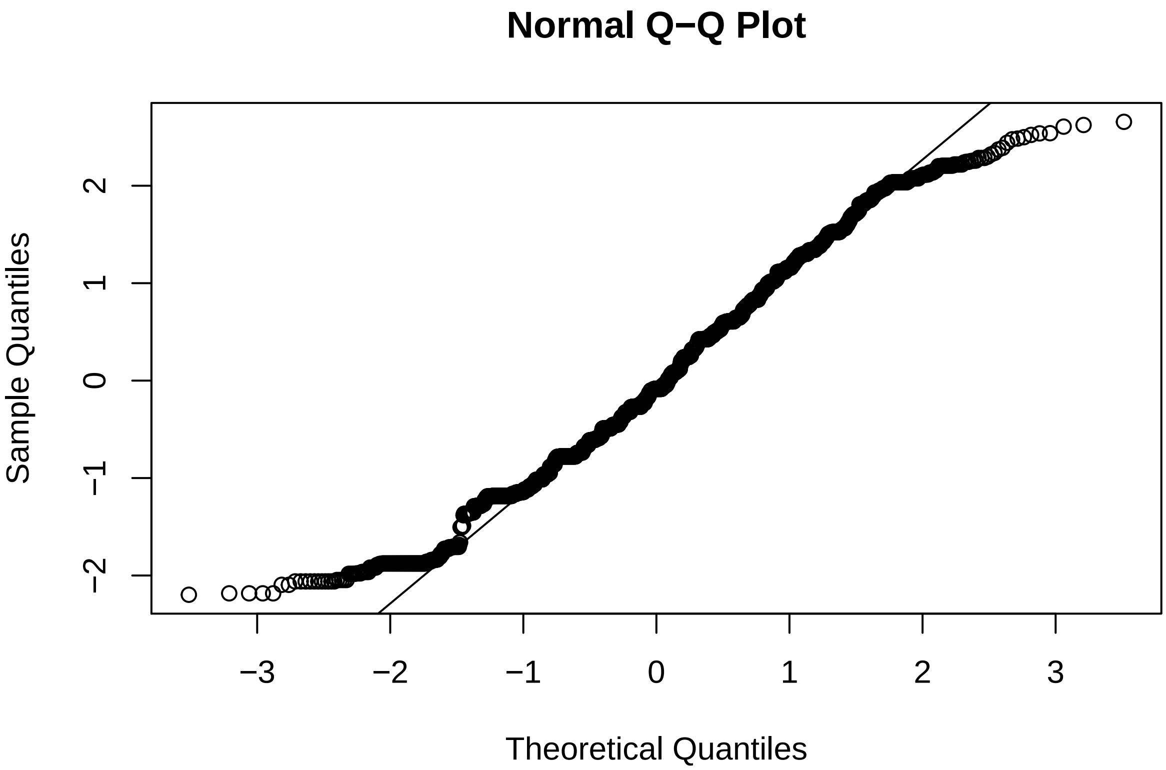
```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## b$MomEducation     5   27.9   5.589   4.958 0.000161 ***
## Residuals       2255 2541.9   1.127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

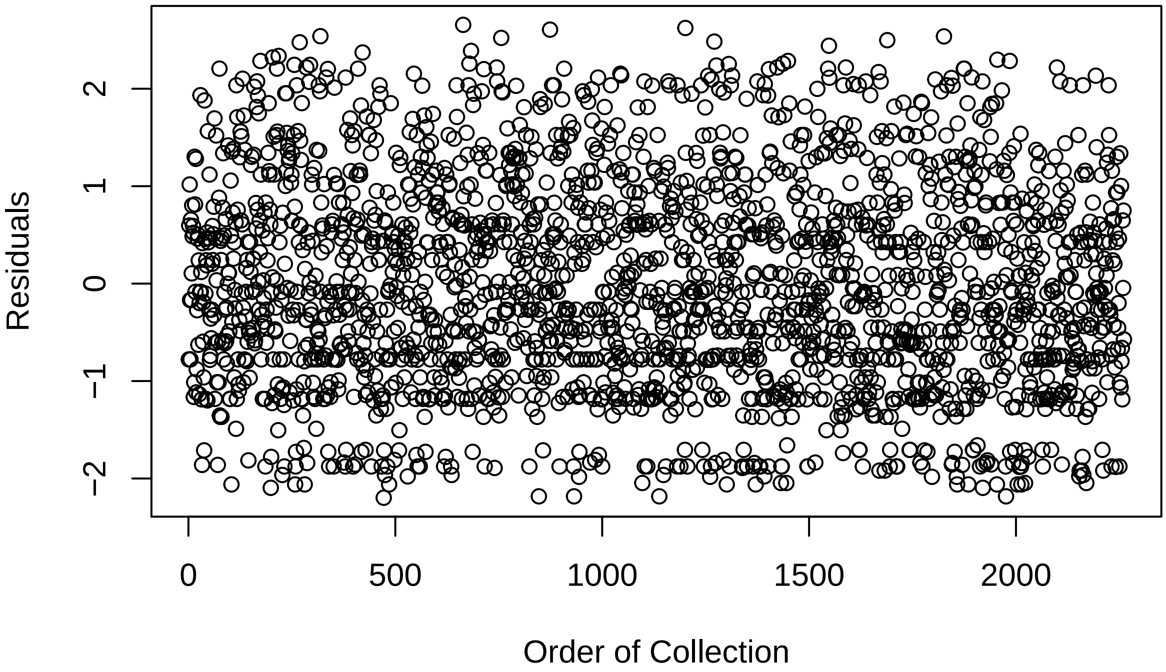
The p-value is close to 0, so we reject the null hypothesis. Variation in mean number of books read is significantly different between some groups.

Fitting a linear regression model to the log-transformed data (summary table below), we see that there is little statistically significant relationship between education levels and number of books read (with the exception of mom's higher education levels). Comparing linear regression model to ANOVA results, we see that even though there may be difference between some groups in mean number of books, there is no significant relationship when comparing all observations.

```
l <- lm(log(b$Books) ~ b$DadEducation + b$MomEducation)
summary(l)
```

```
##
## Call:
## lm(formula = log(b$Books) ~ b$DadEducation + b$MomEducation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19744 -0.77783 -0.08469  0.75839  2.65551
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.70601    0.07657  22.280  < 2e-16 ***
## b$DadEducation02-SomeHS       -0.04657    0.10389  -0.448  0.65398
## b$DadEducation03-HS            0.02050    0.08386   0.244  0.80687
## b$DadEducation04-SomeUndergrad -0.01937    0.09968  -0.194  0.84597
## b$DadEducation05-Undergrad     0.12235    0.09814   1.247  0.21263
## b$DadEducation06-Grad          0.10781    0.11374   0.948  0.34331
## b$MomEducation02-SomeHS        0.05182    0.11254   0.460  0.64526
## b$MomEducation03-HS            0.14993    0.09271   1.617  0.10599
## b$MomEducation04-SomeUndergrad  0.15443    0.10846   1.424  0.15462
## b$MomEducation05-Undergrad     0.23238    0.10881   2.136  0.03282 *
## b$MomEducation06-Grad          0.36907    0.13619   2.710  0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.062 on 2250 degrees of freedom
## Multiple R-squared:  0.01284,    Adjusted R-squared:  0.008451
## F-statistic: 2.926 on 10 and 2250 DF,  p-value: 0.001189
```


Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles
−3
−2
−1
0
1
2
3



Plots above demonstrate that residuals of the model are nearly normal and that there is constant variability. There may be linearity in the data (refer to Part 3) and observations are independent. Conditions for linear regression are satisfied.

Part 5 - Conclusion:

The analysis indicates that parents' education level is a significant predictor of whether an individual will read books recreationally in his or her adult life. Higher education level indicates higher probability of reading books. More surprisingly, parents' education level does not predict the number of books an individual reads. There are possibly other variables that predict how many books an individual reads a year. Further analysis can reveal those. This analysis only considered reading. It may be interesting to see if similar results can be observed in other areas, for example attendance of classical music concerts (and number of concerts attended).

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):

It was interesting to work on this assignment with real data and real results (provided my analysis is valid). I have tried to structure this report in a way that may be presented outside of class. This appendix (and appendix B with R code) includes some notes on assignment to show my work and opinions that I would not want included in the report.

I originally started looking at number of books. After some analysis I realized that I am not seeing any zeros, but there must be people who do not read. The research question is completely different with and without zeros. This data has a significant number of individuals that do not read. This skewed any analysis I tried to do on all data at once. The best solution I could come up with was to separate the research into two questions - whether an individual reads and, if yes, how many books. My research indicates that there are other ways to handle this type of situation. I used logistic regression for the first question and linear regression for the second. I believe there is a way to combine these models, but this is beyond my current abilities.

Quite honestly I found logistic regression challenging. I do feel it was a proper method to use in this situation, but I am still not 100% positive on how to properly test model fit and conditions.

I think presenting results of the project in class would be interesting. I would like to get my classmates' opinions on my research and see what they were up to. I would enjoy it more and find it more valuable than the exercises we presented over the course. However, creating this report was also very valuable, so I think the output of the assignment should be a report which is then presented to the class. Perhaps the last meetup can be everyone's presentations.

Appendix B (Code):

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

```r
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")
```

The code below was used to load transformed data and further adjust it for analysis.

```r
library(foreign)
library(tidyr)
library(dplyr)
library(psych)
library(ggplot2)
library(vcd)
library(gridExtra)

knitr::opts_chunk$set(echo = TRUE)

arts <- readRDS("C:\\Temp\\GitHub\\CUNY-DATA606\\Project\\arts.rds")

arts$DadEducation <- as.character(arts$DadEducation)
arts$DadEducation[arts$DadEducation == "High school graduate (or GED)"] <- "03-HS"
arts$DadEducation[arts$DadEducation == "Less than 9th grade"] <- "01-LessHS"
arts$DadEducation[arts$DadEducation == "College graduate (BA, AB, BS)"] <- "05-Undergrad"
arts$DadEducation[arts$DadEducation == "Some College"] <- "04-SomeUndergrad"
arts$DadEducation[arts$DadEducation == "Some high school"] <- "02-SomeHS"
arts$DadEducation[arts$DadEducation == "Advanced or graduate degree"] <- "06-Grad"
arts$DadEducation <- as.factor(arts$DadEducation)
```

```r
arts$MomEducation <- as.character(arts$MomEducation)
arts$MomEducation[arts$MomEducation == "High school graduate (or GED)"] <- "03-HS"
arts$MomEducation[arts$MomEducation == "Less than 9th grade"] <- "01-LessHS"
arts$MomEducation[arts$MomEducation == "College graduate (BA, AB, BS)"] <- "05-Undergrad"
arts$MomEducation[arts$MomEducation == "Some College"] <- "04-SomeUndergrad"
arts$MomEducation[arts$MomEducation == "Some high school"] <- "02-SomeHS"
arts$MomEducation[arts$MomEducation == "Advanced or graduate degree"] <- "06-Grad"
arts$MomEducation <- as.factor(arts$MomEducation)

books <- arts %>%
  filter(BooksNo >= 0 & BooksNo < 98) %>%
  select(DadEducation, MomEducation, Read = BooksFlag, Books = BooksNo)

books$Read <- as.character(books$Read)
books$Read[books$Read == "no"] <- "0"
books$Read[books$Read == "yes"] <- "1"
books$Read <- as.numeric(books$Read)
```

Code below was used to generate all plots.

```r
# Histogram of number of books read
ggplot(data = filter(books, Books > 0), aes(x = Books)) +
  geom_histogram() +
  xlab("No of Books Read") + ylab("Frequency") +
  ggtitle("Histogram of No of Books Read")

# Mosaic plot of reading flag by education level
m1 <- grid.grabExpr(mosaic(Read ~ DadEducation, data = books,
        labeling_args = list(set_varnames = c(Read = "Did You Read?",
                                              DadEducation = "Dad's Education")),
        set_labels = list(Read = c("No", "Yes"),
                          DadEducation = c("1", "2", "3", "4", "5", "6"))))
m2 <- grid.grabExpr(mosaic(Read ~ MomEducation, data = books,
        labeling_args = list(set_varnames = c(Read = "Did You Read?",
                                              MomEducation = "Mom's Education")),
        set_labels = list(Read = c("No", "Yes"),
                          MomEducation = c("1", "2", "3", "4", "5", "6"))))
grid.arrange(m1, m2, ncol=2)

# Scatterplot and boxplots of number of books by education level
books %>%
  filter(Books > 0) %>%
  select(Dad = DadEducation, Mom = MomEducation, Books) %>%
  gather(Education, Level, Dad:Mom) %>%
  ggplot(aes(x = Level, y = Books, shape = Education)) +
  geom_point(aes(color = Education),
             position = position_jitterdodge(dodge.width = 0.8)) +
  geom_boxplot(fill = "white",
               outlier.colour = NA,
               position = position_dodge(width=0.8),
               width = 0.25,
               show.legend = FALSE) +
  xlab("Education Level") + ylab("Books Read") +
  ggtitle("Plot of No of Books Read by Education Level") +
  theme(legend.position="bottom")

# Residuals plots for logistic regression
plot(lr$residuals, books$read,
     xlab = "Order of Collection", ylab = "Residuals")

plot(lr$fitted.values, lr$residuals, col=c("blue","red")[1+books$Read],
     xlab = "Fitted Values", ylab = "Residuals")
abline(h=0,lty=2,col="grey")
lines(lowess(lr$fitted.values,lr$residuals),col="black",lwd=2)

# Log-transform plots for number of books
b <- books %>%
  filter(Books > 0)

ggplot(data = b, aes(x = log(Books))) +
  geom_histogram(bins = 10) +
  xlab("No of Books Read") + ylab("Frequency") +
  ggtitle("Histogram of No of Books Read (log-transform)")

qqnorm(log(b$Books))
qqline(log(b$Books))

# Normal probability plots for education levels
op <- par(mfrow=c(2,3))
for(i in levels(b$DadEducation)){
  tmp <- with(b, log(Books[DadEducation == i]))
  qqnorm(tmp, xlab="No of Books",main = i)
  qqline(tmp)
}
par(op)

op <- par(mfrow=c(2,3))
for(i in levels(b$MomEducation)){
  tmp <- with(b, log(Books[MomEducation == i]))
  qqnorm(tmp, xlab="No of Books",main = i)
  qqline(tmp)
}
par(op)

# Boxplots for number of books by education level
op <- par(mfrow=c(1,2))
plot(log(b$Books) ~ b$DadEducation,
     xlab = "Dad's Education", ylab = "No of Books Read (log-transform)")
plot(log(b$Books) ~ b$MomEducation,
     xlab = "Mom's Education", ylab = "No of Books Read (log-transform)")
par(op)

# Residuals plots for linear regression
qqnorm(l$residuals)
qqline(l$residuals)

plot(l$residuals, xlab = "Order of Collection", ylab = "Residuals")
```