EDUCATION LEVEL AND FATHER'S EDUCATION LEVEL

Introduction:

Is there a relationship between one's education level and the level of education of his own father?

The relationship between the education level of parents and children has been the subject of many research studies. I want to investigate this association because I think if you understand better how education level of parents influence children's education and behavior, it is possible to identify better the causes of inequality across generations, how to deal this problem and reverse negative trends.

Data:

The dataset used is the data collected by the “General Social Survey” (GSS). GSS is a sociological survey on demographic characteristics and attitudes of residents of the United States.

[Excerpted from http://www.norc.org/Research/Projects/Pages/general-social-survey.aspx]

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop cross-national models of human society; and to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting. GSS questions cover a diverse range of issues including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior.

The cases are the residents of the United States interviewed in the GSS survey.

The two variables are:

1. paeduc_cat - Father's degree
2. educ - Highest year of school completed

The first variable is categorical and the second variable is numerical discrete. The variable paeduc_cat is not part of the original data set but it is obtained by transforming the original numerical variable paeduc (Highest year of school completed, father) into a categorical variable for convenience (Less than Middle School if paeduc< 8, Middle school if 8<=paeduc< 12, High School if 12<=paeduc<16 and Bachelor or Graduate if paeduc >= 16).

The type of study is observational because the data were collected in a way that does not directly interfere with how the data arise (a survey).

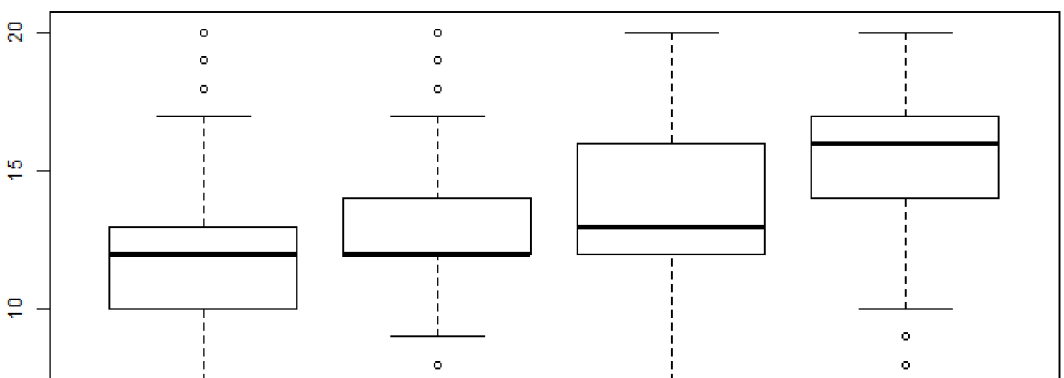
The population of interest is all residents of the United States. The findings from the analysis can be generalized to that population because the sample of residents in the survey is selected randomly. There is not convenience or voluntary bias because the sample is random. There could be a non-response bias but it is likely that the respondents are a random fraction of the sample so they are still representative of the initial random sample.

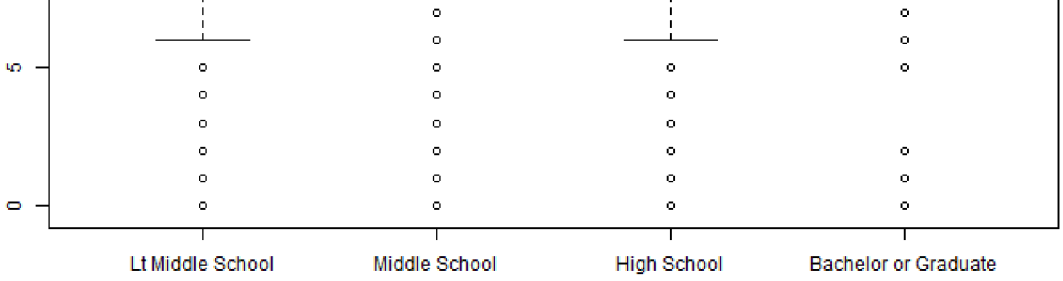
These data cannot be used to establish causal links between the variables because the type of study is observational. In this case it is only possible to establish an association between variables.

Exploratory data analysis:

The side-by-side boxplot below compares the highest year of school completed across the different father's education level. The plot could suggest a positive association between the two variables: the more a father has studied the more likely the son/daughter studies. In this association the explanatory variable is the father's degree (paeduc_cat) and the response variable is the highest year of school completed (educ).

```
boxplot(gss$educ ~ gss$paeduc_cat)
```





To complete the previous plot, this is the summary statistics per group

```
by(gss$educ, gss$paeduc_cat, summary)
```

```
## gss$paeduc_cat: Lt Middle School
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    10.0    12.0    11.4    13.0    20.0      14
## ------------------------------------------------------------
## gss$paeduc_cat: Middle School
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    12.0    12.0    12.7    14.0    20.0      15
## ------------------------------------------------------------
## gss$paeduc_cat: High School
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    12.0    13.0    13.8    16.0    20.0      17
## ------------------------------------------------------------
## gss$paeduc_cat: Bachelor or Graduate
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    14.0    16.0    15.5    17.0    20.0       9
```

Inference:

The hypotheses for the inference are:

- H0 - Null hypothesis: The mean highest year of school completed is the same across all the different education levels of the father.
- HA - Alternative hypothesis: At least one pairs of means are different from each other.

The method chosen to perform the hypothesis test is ANOVA with pairwise tests because the explanatory variable is categorical with more than 2 levels (paeduc_cat) and the response variable is numerical.

ANOVA (analysis of variance) is used to compare means across several groups with a single hypothesis test. The test statistics used by ANOVA is F statistics that is the ratio between the between-groups variability and within-group variability. A large F value corresponds to large between-groups variability relative to within-group variability and consequently to strong evidence against the null hypothesis. A p-value is calculated from the upper tail of the F distribution. If the p-value is lower than the significance level, we can reject the null hypothesis. To identify which groups differ in means, a pairwise t-test has to be used for each different pair of groups with a modified significance level.

The conditions to be satisfied for ANOVA are:

1. Independence within groups and between groups
2. Approximate normality (of the response variable)
3. Equal variance

The sample of residents in the survey is selected randomly so data are a simple random sample. In addition, the sample size of each group represents less than 10% of the respective population (see the summary table below for the counts of each group).

```
table(gss$paeduc_cat)
```

```
##
##     Lt Middle School        Middle School          High School
##                 7100                 10499                 15219
## Bachelor or Graduate
##                 6197
```

There is also independence between groups because the groups are not paired since individuals of each group are different and independent from individuals of other groups. So the first condition is satisfied.

To check the second condition, normal probability plots for each one of the groups will be used.

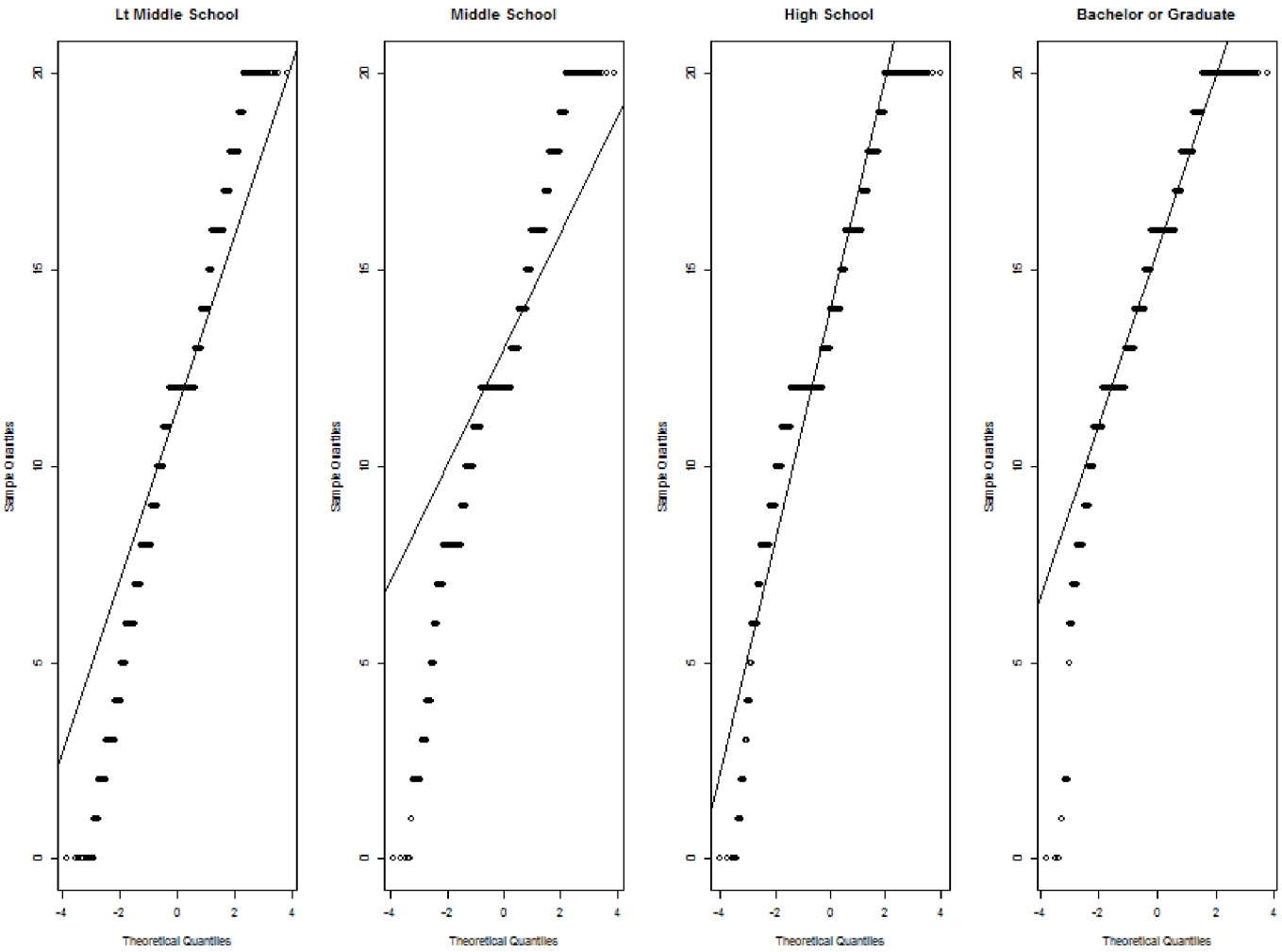
```
windowpar <- par(mfrow = c(1, 4))
qqnorm(gss$educ[gss$paeduc_cat == "Lt Middle School"], main = "Lt Middle School")
qqline(gss$educ[gss$paeduc_cat == "Lt Middle School"])

qqnorm(gss$educ[gss$paeduc_cat == "Middle School"], main = "Middle School")
```

```
qqline(gss$educ[gss$paeduc_cat == "Middle School"])

qqnorm(gss$educ[gss$paeduc_cat == "High School"], main = "High School")
qqline(gss$educ[gss$paeduc_cat == "High School"])

qqnorm(gss$educ[gss$paeduc_cat == "Bachelor or Graduate"], main = "Bachelor or Graduate")
qqline(gss$educ[gss$paeduc_cat == "Bachelor or Graduate"])
```




```
par(windowpar)
```

As you can see from the plots above, the groups, especially “Middle School” and “Bachelor or Graduate”, seem to diverge a little from normality but it is possible to relax the normality condition and consider the second condition met since the sample size of the groups is large.

The third condition is not satisfied because the groups have different variability and the sample sizes are different as you can see from the initial side-by-side box plot.

To summarize, the results may be unreliable due to the equal variance condition is not met.

To perform inference, the function inference from the website http://bit.ly/dasi_inference will be called with the parameter set for the ANOVA method:

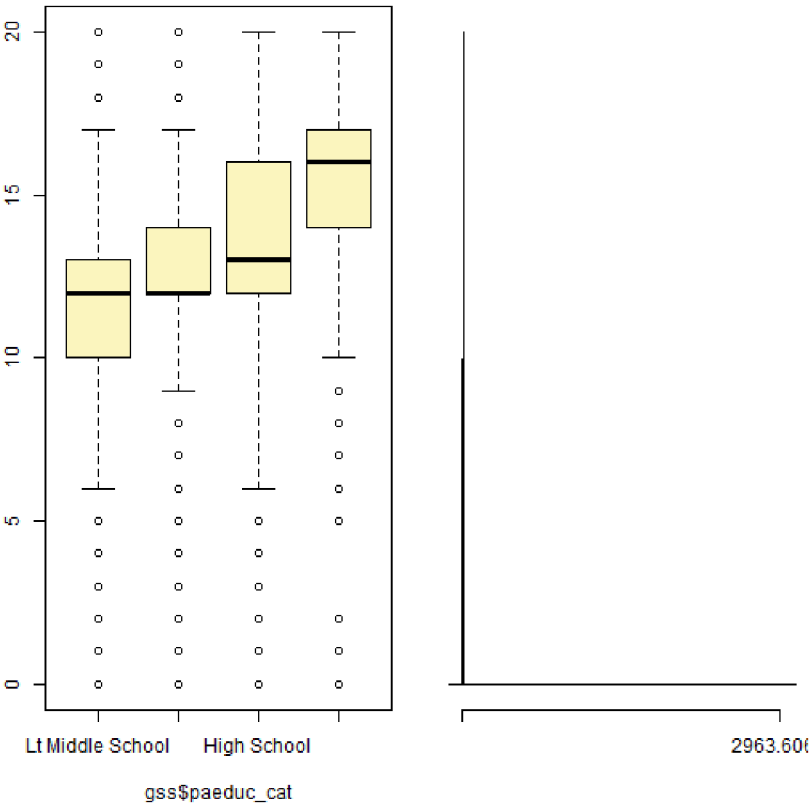
```
source("http://bit.ly/dasi_inference")
inference(y = gss$educ, x = gss$paeduc_cat, est = "mean", type = "ht", alternative =
"greater",
    method = "theoretical")
```

```
## Warning: package 'BHH2' was built under R version 3.0.3
```

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_Lt Middle School = 7086, mean_Lt Middle School = 11.42, sd_Lt Middle School = 3.195
## n_Middle School = 10484, mean_Middle School = 12.68, sd_Middle School = 2.615
## n_High School = 15202, mean_High School = 13.82, sd_High School = 2.448
## n_Bachelor or Graduate = 6188, mean_Bachelor or Graduate = 15.46, sd_Bachelor or
Graduate = 2.427
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
```

```
## Analysis of Variance Table
##
## Response: y
##              Df Sum Sq Mean Sq F value Pr(>F)
## x             3  61969   20656    2964 <2e-16
## Residuals 38956 271524       7
##
## Pairwise tests: t tests with pooled SD
##                      Lt Middle School Middle School High School
## Middle School                       0             NA          NA
## High School                         0              0          NA
## Bachelor or Graduate                0              0           0
```



The p-value calculated can be interpreted as a conditional probability, the probability of at least as large a ratio between the “between” and “within” group variabilities if in fact the means of all groups are equal.

Since the p-value calculated by ANOVA is lower than the significance level 0.05, we reject the null hypothesis and conclude that the data provide convincing evidence that at least one pair of means are different. Since all the p-values of the pairwise tests from the ANOVA output are lower than the modified significance level (0.05/3 = 0.0167), we can conclude that there is strong evidence that all means are different.

Because we have used ANOVA, it is not possible to apply other methods like confidence interval and so there is nothing to compare.

Conclusion:

The result of ANOVA allows the conclusion that there is an association between USA residents' education level and their father's education level (under the condition that the results may be unreliable due to the equal variance not met). If a father has studied more, it is more likely that children have a better education level. Because this is only an association, it is not possible to conclude that there is causality between the two education levels but it would be interesting to verify it and include in the study other variables like mother's education level. Another possible extension of the study would be the research of other factors that explain the variability of USA residents' education level like family income and race and the comparison with data from other countries.

References:

Data citation:

Smith, Tom W., Michael Hout, and Peter V. Marsden. General Social Survey, 1972-2012 [Cumulative File]. ICPSR34802-v1. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut /Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2013-09-11. doi:10.3886/ICPSR34802.v1

Persistent URL: http://doi.org/10.3886/ICPSR34802.v1

The data set used in this study is an extract of GSS provided by the Coursera DASI team who have removed missing values from the responses and created factor variables when appropriate to facilitate analysis.

The codebook https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html contains the list of variables, their possible values and the original survey questions.

Appendix:

Extract of the data set

```
gss[1:65, c("paeduc", "paeduc_cat", "educ")]
```

```
##    paeduc           paeduc_cat educ
## 1      10        Middle School   16
## 2       8        Middle School   10
## 3       8        Middle School   12
## 4      16 Bachelor or Graduate   17
## 5       8        Middle School   12
## 6      18 Bachelor or Graduate   14
## 7      16 Bachelor or Graduate   13
## 8      16 Bachelor or Graduate   16
## 9      12          High School   12
## 10     10        Middle School   12
## 11     12          High School   13
## 12     NA                  <NA>    6
## 13      5     Lt Middle School    9
## 14     NA                  <NA>    8
## 15     NA                  <NA>    9
## 16     NA                  <NA>   14
## 17      8        Middle School   14
## 18     NA                  <NA>    8
## 19      8        Middle School   17
## 20     12          High School   14
## 21     NA                  <NA>   12
## 22      8        Middle School   11
## 23      3     Lt Middle School   13
## 24      8        Middle School   12
## 25      8        Middle School   16
## 26      7     Lt Middle School   12
## 27     12          High School   12
## 28     12          High School   12
## 29     12          High School   12
## 30      8        Middle School    9
## 31      6     Lt Middle School    6
## 32     NA                  <NA>   14
## 33     12          High School   16
## 34     NA                  <NA>    8
## 35      4     Lt Middle School   12
## 36     12          High School   14
## 37     NA                  <NA>   13
## 38     NA                  <NA>   10
## 39     NA                  <NA>    6
## 40      3     Lt Middle School   12
## 41      5     Lt Middle School    8
## 42      8        Middle School   14
## 43     NA                  <NA>    8
## 44     NA                  <NA>    8
## 45      8        Middle School    8
## 46     12          High School   12
## 47      8        Middle School   14
## 48     NA                  <NA>   12
## 49      8        Middle School    9
## 50      8        Middle School    8
## 51     14          High School   10
## 52     10        Middle School   14
## 53      3     Lt Middle School    9
## 54     13          High School   12
## 55     NA                  <NA>   10
## 56     NA                  <NA>   12
## 57     12          High School   14
## 58     12          High School   12
## 59     NA                  <NA>   12
## 60     NA                  <NA>   11
## 61     NA                  <NA>   12
## 62     NA                  <NA>    9
## 63      8        Middle School   12
## 64      6     Lt Middle School   11
## 65      8        Middle School   12
```