Relationship between the level of education and the age when the first child is born

Introduction:

In this research I would like to investigate the following question: is there a relationship between the level of education of a person and the age when (s)he decides to have the first child.

It is often told that modern people tend to have families and children later. Relying on a limited sample of people I know, I can not jump to such a conclusion. Therefore, I think that researching this particular question on a big dataset would be both interesting and valuable.

Data:

In this research I am using the modified dataset (all missing values have been recoded as “NA”) from General Social Survey (GSS). GSS has been monitoring societal change and studying the growing complexity of American society since 1972. 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.

Units of observation (cases) in this study are a set of answers by respondents to interview's questions (the set of each respondent's answers has a numerical identification) and a year when this respondent was surveyed (numerical 1972-2012).

I'm going to use 2 variables:

degree: highest degree achieved (categorical variable with set of values {Lt High School, High School, Junior College, Bachelor, Graduate})

agekdbrn: age when the 1st child is born (numerical with values from the interval [9,65])

My study is observational since it is not an experemint. In an observational study, researchers collect data in a way that does not directly interfere with how the data arise, which is our case (merely observe). In an experiment, on the other hand, researchers randomly assign subjects to various treatments, and can therefore establish causal connections between explanatory and response variables (not our case).

Population of interest is residents of the United States. There are some possible sources of bias:

Convenience sample: individuals who are easily accessible are more likely to be included in the survey

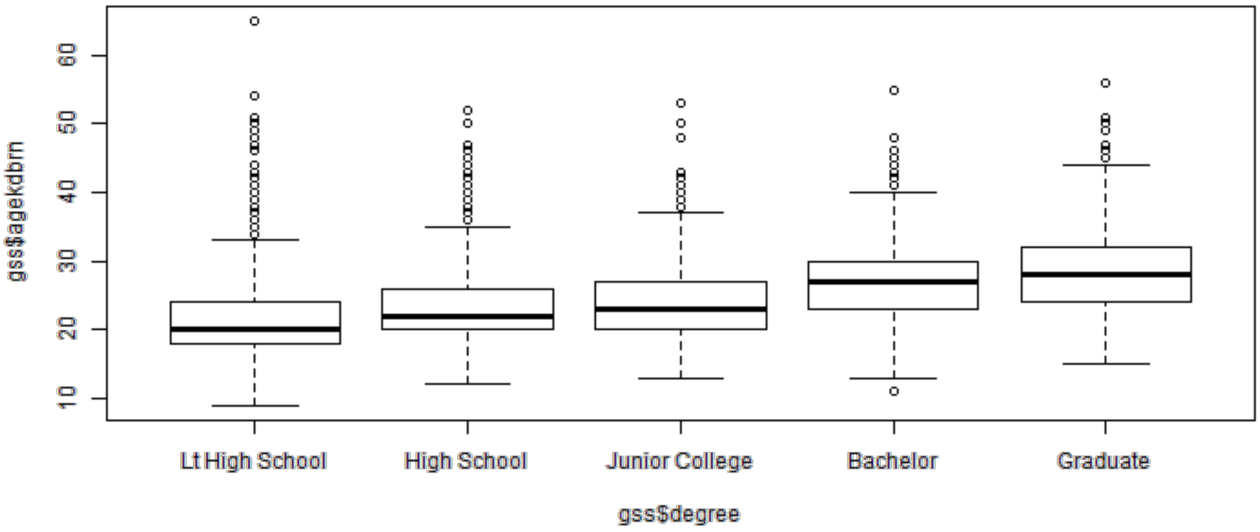
Non-response: if only a non-random fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population

Voluntary response: occurs when the sample consist of people who volunteer to respond because they have strong opinions on the issue. As long as all theses 3 possible sources of bias are avoided the data looks reliable and it could be generalized to the US population.

Based on an observational study (which we have), we can only establish an association, in other words correlation, between the explanatory and response variables (NOT a causal connection, since we do not have an experiment). There may be other variables that we do not control in this study that contribute to the observed difference.

Exploratory data analysis:

```
plot(gss$agekdbrn ~ gss$degree)
```



In this plot we can see the tendancy: people with higher educational level have kids later.

```
by(gss$agekdbrn, gss$degree, summary)
```

```
## gss$degree: Lt High School
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       9      18      20      21      24      65    8926 
## -------------------------------------------------------- 
## gss$degree: High School
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##      12      20      22      23      26      52   19669 
## -------------------------------------------------------- 
## gss$degree: Junior College
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##      13      20      23      24      27      53    1709 
## -------------------------------------------------------- 
## gss$degree: Bachelor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##      11      23      27      27      30      55    5367 
## -------------------------------------------------------- 
## gss$degree: Graduate
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    15.0    24.0    28.0    28.2    32.0    56.0    2430
```

Comparing the age when the first child is born with educational level we can see increase (or rare invariance) of Median, Mean and both quartiles while education level increases. Boxplot supports my assumption that there is a tendency that more educated people prefer to have kids later.

Inference:

In order to perform a test to prove my hypothesis that there is correlation between educational level of a person and the time when (s)he decides to have a baby, I divide our sample into 5 groups according to their educational levels. So my data is presented as one numerical and one categorical variable (with 5 levels). It means that I only can perform a hypothesis test. ANOVA and pairwise tests are suitable for this situation.

At first, I have to make a null hypothesis which always states that there is nothing going on and that on average ages when the first child is born are the same for all educational levels: H0: Mu1=..=Mu5. Alternative hypothesis: Ha: at least one pair of means (as a measure of average value) are different from each other.

To perform such test we have to check conditions.

1.Independence: a) within group independence: This means that the sample observations must be independent of each other in every group.

```
data = na.omit(subset(gss, select = c(agekdbrn, degree)))
a = subset(data, degree == "Lt High School")
b = subset(data, degree == "High School")
c = subset(data, degree == "Junior College")
d = subset(data, degree == "Bachelor")
e = subset(data, degree == "Graduate")
dim(a)
```

```
## [1] 2896    2
```

```
dim(b)
```

```
## [1] 9618    2
```

```
dim(c)
```

```
## [1] 1361    2
```

```
dim(d)
```

```
## [1] 2635    2
```

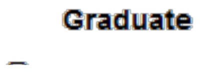
```
dim(e)
```

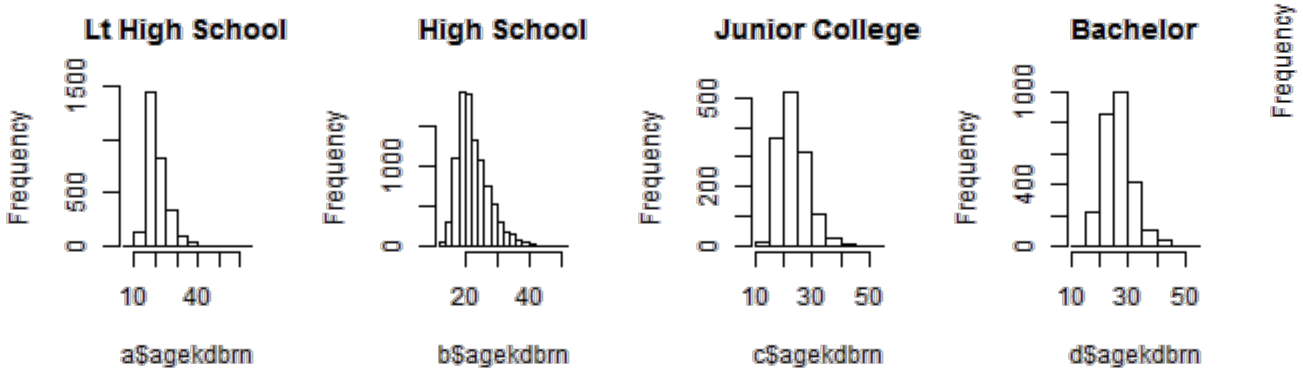
```
## [1] 1440    2
```

Since we have a random sample of people, and as numbers of people in each group are less than 10% of their respective population (as shown above) we can conclude that this condition is met.

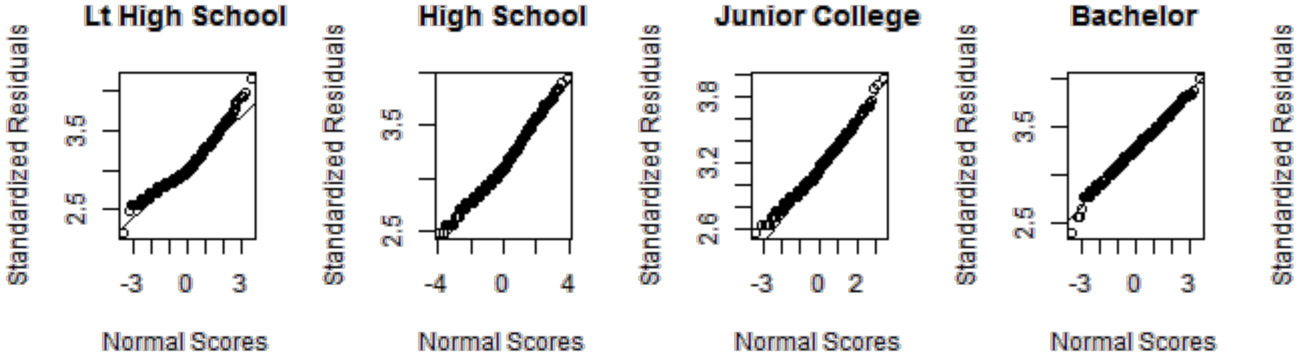
b) between group independence: This condition is also satisfied since every person could be included in only one group, so there is no pairing between the groups.

2.Approximately normal condition. The distribution of the response variable within each group should be approximately normal.





From the graphs above we can see that it is not the case and our distribution is skewed. Hence, we have to take logarithm to “normalize data.”



Now our data set is sutible and condition of “approximate normality” is met.

3.The last condition is the condition of constant variance, which says that variability should be consistent across groups. We can check this condition using side-by-side box plots, where it's going to be pretty easy to visually compare the variability across the groups, as well as by looking at the standard deviations for each group (log): 0.218 for “Lt High School”;

0.1994 for “High School”;

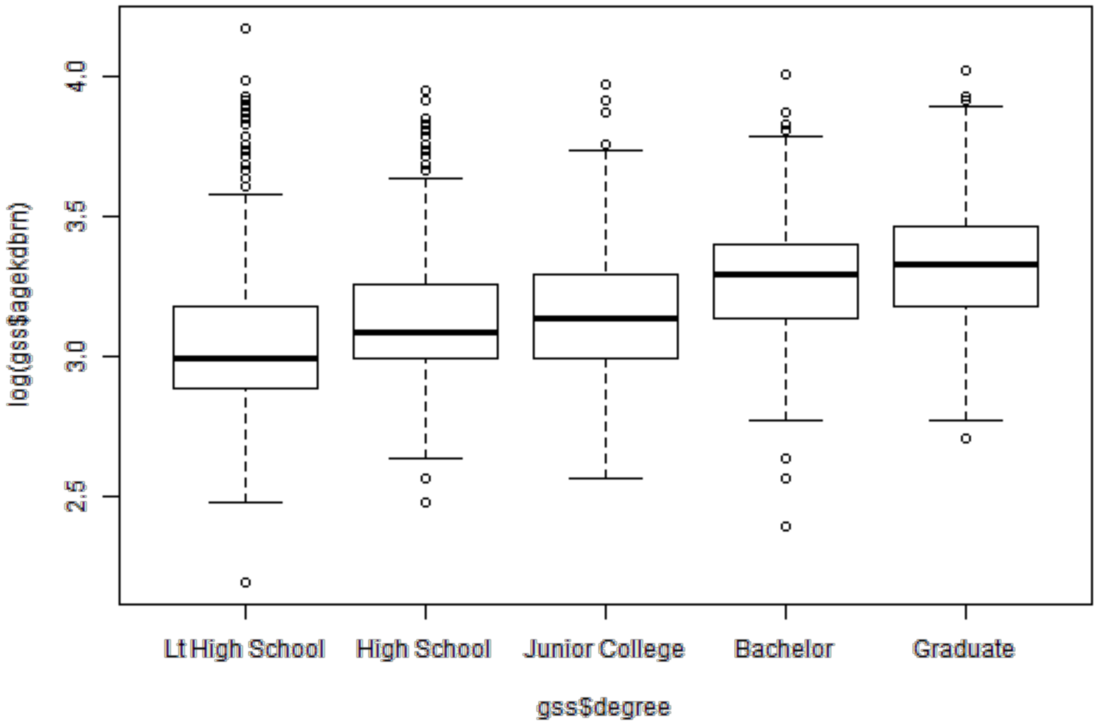
0.2052 for “Junior College”;

0.1862 for “Bachelor”;

0.2007 for “Graduate”;

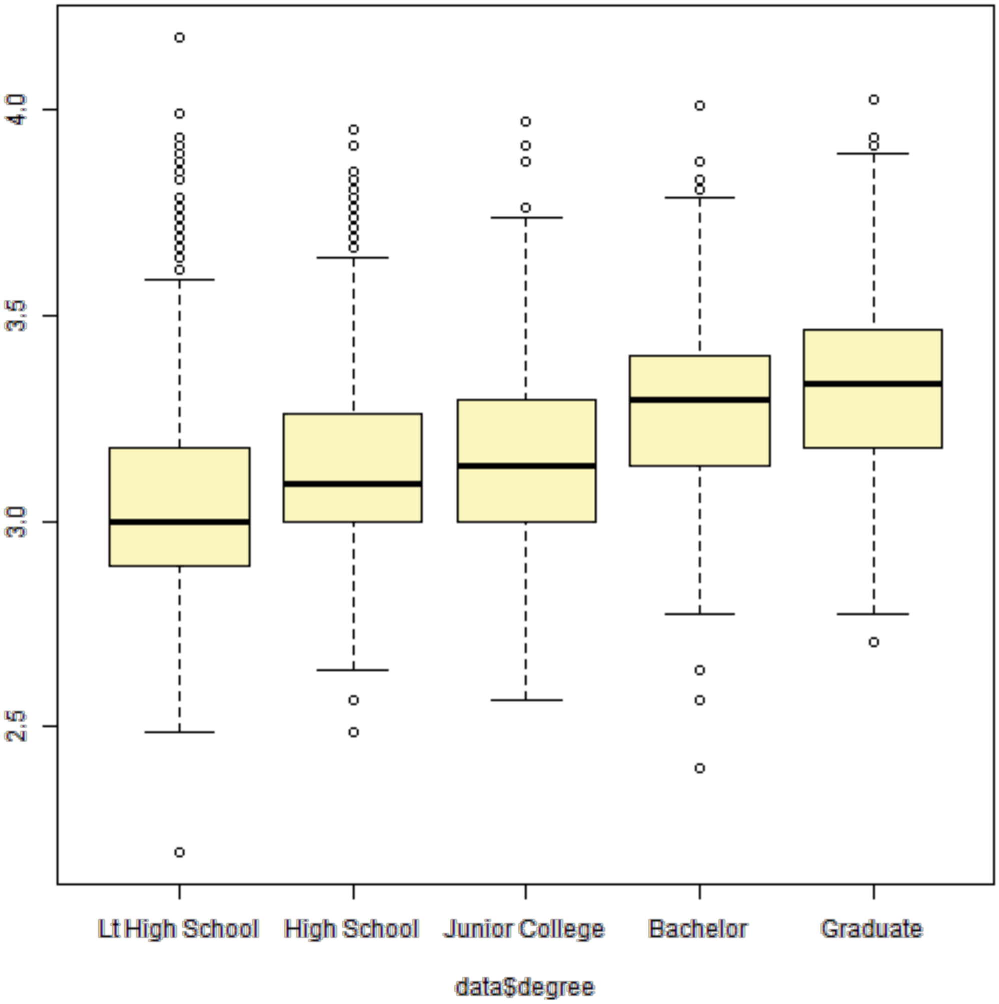
From this we can see that we have roughly equal variability across the groups and that our third condition is satisfied. But I have to mention that my previous hypothesis changed: H0: ln(Mu1)=..=ln(Mu5). Alternative hypothesis: Ha: at least one pair of means (as a measure of average value) are different from each other.

```
plot(log(gss$agekdbrn) ~ gss$degree)
```



```
load(url("http://bit.ly/dasi_gss_ws_cl"))
source("http://bit.ly/dasi_inference")
inference(y = log(data$agekdbrn), x = data$degree, est =
"mean", type = "ht",
    alternative = "greater", method = "theoretical", inf_plot
= FALSE, sum_stats = TRUE)
```

```
## Warning: package 'BHH2' was built under R version 3.0.3
## Response variable: numerical, Explanatory variable:
categorical
## ANOVA
## Summary statistics:
## n_Lt High School = 2896, mean_Lt High School = 3.034, sd_Lt
High School = 0.218
## n_High School = 9618, mean_High School = 3.116, sd_High
School = 0.1994
## n_Junior College = 1361, mean_Junior College = 3.155,
sd_Junior College = 0.2052
## n_Bachelor = 2635, mean_Bachelor = 3.275, sd_Bachelor =
0.1862
## n_Graduate = 1440, mean_Graduate = 3.319, sd_Graduate =
0.2007
```



```
## 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              4    131    32.8     810 <2e-16
## Residuals 17945    726     0.0
##
## Pairwise tests: t tests with pooled SD
##                 Lt High School High School Junior College
Bachelor
## High School                 0          NA
NA       NA
## Junior College              0           0
NA       NA
## Bachelor                    0           0
0       NA
## Graduate                    0           0
0        0
```

Performing this test we get p-value < 2e-16, which corresponds to the probability that the differences between the groups variabilities are greater or equal to what we observe if, in fact, the means of all groups are equal. Since p-value < significance level (0.05) we can reject Ho and say that at least one log(mean) is different. Pairwise tests show that all p-values for different pairs are insignificant(->0) and are less than modified significance level (0.05/10=0.005), which states that neither pair of log(mean) are equal.

Conclusion:

Despite the fact that we were working with logarithm, we can generalize our conclusions to means themselves. So ANOVA test states that my initial guess was right and that the difference between average ages when people have their first child correlate with their level of education. Furthermore, this association is strong between all education level groups. Unfortunately, I do not see the way to perform an experiment to find out if there is a causal relationship. Anyhow, I got an interesting result which could be used to provide better social programs for young families. Maybe it could be helpful to consider some of the so needed reforms (For example in the USA parental leave is extremely short, approximately 3 months.) Anyway, this was a fun research with great results which could be a starting point for a bunch of new studies) Thank you for your time!

References:

1. 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 (https://d396qusza40orc.cloudfront.net

/statistics%2Fproject%2Fgss1.html)

2. https://d396qusza40orc.cloudfront.net/statistics/lab_resources/Rcommands.pdf

3. http://www.rstudio.com/ide/docs/authoring/using_markdown?version=0.98.501&mode=desktop

4. https://class.coursera.org/statistics-001

Appendix:

```
head(data, 55)
```

```
##       agekdbrn         degree
## 29389       21    High School
## 29391       25 Lt High School
## 29392       23 Lt High School
## 29395       17 Lt High School
## 29396       23    High School
## 29397       17    High School
## 29398       17    High School
## 29401       29       Bachelor
## 29402       32       Bachelor
## 29406       34       Bachelor
## 29410       35       Graduate
## 29415       32    High School
## 29416       14 Lt High School
## 29418       23       Bachelor
## 29419       20    High School
## 29421       20    High School
## 29423       27       Bachelor
## 29429       38       Bachelor
## 29431       20       Graduate
## 29432       29    High School
## 29435       18    High School
## 29436       25       Bachelor
## 29437       19 Lt High School
## 29438       19    High School
## 29439       18    High School
## 29444       26       Bachelor
## 29446       27    High School
## 29448       23 Lt High School
## 29449       24    High School
## 29451       20    High School
## 29452       24    High School
## 29458       21 Lt High School
## 29459       19 Lt High School
## 29460       21    High School
## 29461       32 Lt High School
## 29462       24    High School
## 29466       18       Bachelor
## 29467       32       Graduate
## 29469       24       Graduate
## 29474       26    High School
## 29475       28    High School
## 29476       29    High School
## 29478       30 Lt High School
## 29479       31    High School
## 29480       21 Lt High School
## 29482       32       Graduate
## 29484       29       Graduate
## 29487       18    High School
## 29490       18 Junior College
## 29492       23    High School
## 29494       32 Lt High School
## 29495       25    High School
## 29500       27       Graduate
## 29504       26 Lt High School
## 29505       28 Lt High School
```