1 Data Project: Relationship between Degree Earned and Family Income

Mauro Taraborelli – *mauro@maurotaraborelli.com*

1.1 Introduction

The project studies the relationship between the highest degree earned by United States residents and their family income in constant dollars.

Access to education and its funding is the subject of many discussions on social mobility and redistribution of income. The study explores data from a long running social survey to verify one of the main topic of these discussions: if family income is related to education level.

1.2 Data

The study uses General Social Survey (GSS) data for the year 2012.

1.2.1 General Social Survey (GSS)

The General Social Survey (GSS) has been monitoring change in American society since 1972. Until 1994, it was conducted almost annually. Since 1994, the GSS has been conducted in even numbered years.

The vast majority of GSS data is obtained in face-to-face interviews, computer-assisted personal interviewing (since 2002), and by telephone.

The interviewees are individuals English and Spanish speaking persons 18 years of age or older, living in the United States. They are selected from metropolitan and rural areas. Multiple level of stratification for region, race, age, income and sex was employed to guarantee a random sample.

The target sample size is of 1500 observations every year until 1994. Since 1994 the GSS has been administered to two samples in even-numbered years, each with a target sample size of 1500.

For more information please refer to GSS FAQ [1].

1.2.2 Study characteristics

The data come from a survey and not from an experiment, so the study can be characterized as **observational**: it can establish only correlation between the variables examined and not causation.

However, GSS data are random samples taken from US residents, so the study's findings could be **generalized** to the entire US residents population.

1.2.3 Variables

The study uses two variables chosen from those collected by GSS survey:

- **Highest degree**: it is a ordinal categorical variable collected with the question "Did you ever get a high school diploma or a GED certificate?" – its levels are *Less than high school*, *High school*, *Associate/Junior College*, *Bachelor's*, *Graduate*
- **Family income in constant US dollars**: it is a continuous numerical variable and it measure the inflation-adjusted family income

1.3 Exploratory data analysis

1.3.1 Data preparation

Before exploring the data, we create a dataset with the two variables of interest for the year 2012 and we give them a clearer label. We obtain a sample size of 1,974 observations.

```
# Filter Highest.Degree and Family.Income.Constant.USD for 2012
study_data = gss[gss$year == 2012, c(12,27)]
# Relabel the two variables
colnames(study_data) <- c("Highest.Degree","Family.Income.Constant.USD")
# Count the observations
nrow(study_data)
```

```
## [1] 1974
```

1.3.2 Summary statistics

```
# Three graphs in a row
par(mfrow = c(1, 3))
# Barplot of Highest.Degree
par(mar=c(6,4,3,2))
barplot(table(study_data$Highest.Degree), las=2, main="Highest Degree")
# Histogram of Family.Income.Constant.USD
par(mar=c(6,1,3,2))
hist(study_data$Family.Income.Constant.USD, main="Family Income in constant USD", xlab="USD")
# Boxplot of Family.Income.Constant.USD by Highest.Degree
par(mar=c(6,2,3,4))
boxplot(study_data$Family.Income.Constant.USD ~ study_data$Highest.Degree, las= 2, main="Family Income in cons
```

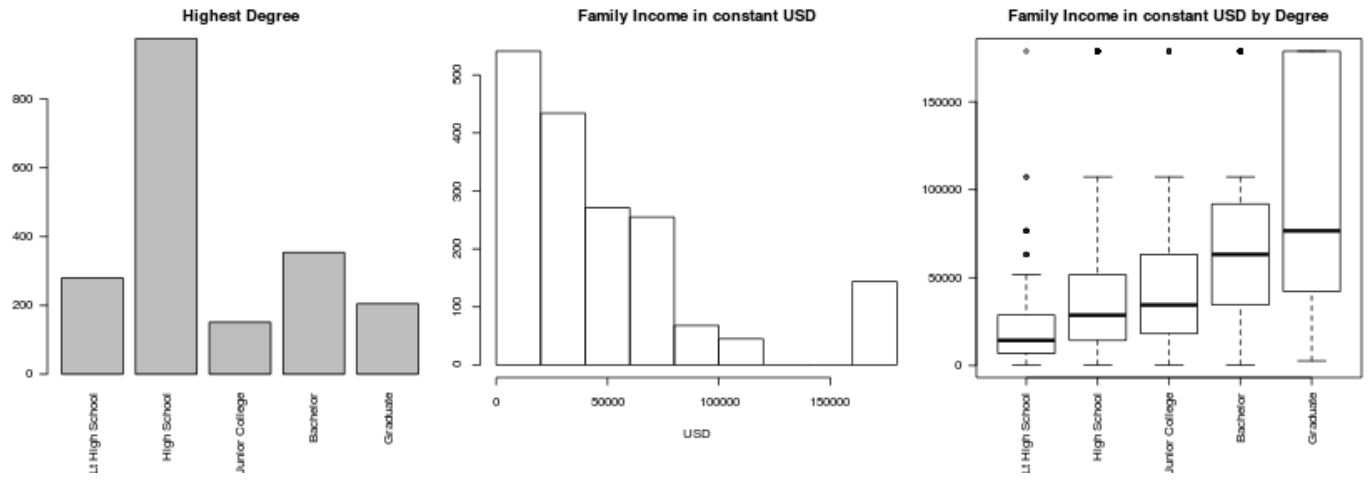

Figure 1: Summary statistics visualization

1.3.2.1 Highest degree Highest degree is a categorical variable. We summarize it with a contingency table, a frequency table and a barplot (see Figure 1).

```
# Contingency table for Highest.Degree
table(study_data$Highest.Degree)
```

```
##
## Lt High School    High School Junior College       Bachelor       Graduate
##            280            976            151            354            205
```

```
# Frequency table for Highest.Degree
prop.table(table(study_data$Highest.Degree))
```

```
##
## Lt High School    High School Junior College       Bachelor       Graduate
##        0.14242        0.49644        0.07681        0.18006        0.10427
```

We can see that high school as the highest degree has nearly 50% percent of observations.

1.3.2.2 Family income in constant USD Family income in constant USD is a continuous numerical variable. We summarize it with mean, range and quantiles, and with a histogram (see Figure 1).

```
# Mean, range and quartiles for Family.Income.Constant.USD
summary(study_data$Family.Income.Constant.USD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     383   16300   34500   48400   63200  179000     216
```

We can see that the distribution is right skewed and unimodal, with 50% of observations in the 16,300-63,200 USD (constant dollar) range, and there is a maximum value of 179,000 USD. There are some clear outliers in the upper quantiles of the distribution.

There are 216 observations with missing income values. Filtering them out brings the number of observations to 1752. The sample size remains significant for the study.

```
# Filter NAs
study_data = study_data[complete.cases(study_data),]
# Count observations after the filter
nrow(study_data)
```

```
## [1] 1752
```

```
# Contingency table for Highest.Degree after the filter
table(study_data$Highest.Degree)
```

```
##
## Lt High School    High School Junior College       Bachelor       Graduate
##            230            881            132            324            185
```

1.3.2.3 Relationship among family income and highest degree Finally, we explore the relationship among family income and highest degree (see Figure 1).

We can see that exists a positive association, but the wider interquantile range in the college groups and the presence of outliers in the high school and less than high school groups, means that such a relationship is not strong and that family income could be associated with other variables.

1.4 Inference

The study want to establish if there is a statistical significant difference between the mean family income in constant dollars of United States resident grouped by the highest degree earned.

In statistical inference terms, we test a null hypotheses (H_0) where the mean family income in constant dollars is equal for all the highest degree groups, and an alternative hypotheses (H_A) where at least one pair of means are different from each other.

$H_0 : \mu_{LHS} = \mu_{HS} = \mu_{JC} = \mu_B = \mu_G$

$H_A : \textit{the average income in constant dollar } (\mu_i) \textit{ varies across some (or all) groups}$

1.4.1 ANOVA and its conditions

We could test the hypotheses doing a pairwise comparison of means across many groups. But we could find a difference just by chance, even if there is no difference in the population.

So we first use a method called **analysis of variance** (ANOVA) [2] that uses a single hypotheses test to check whether the mean across many groups are equal. If we can reject the null hypotheses using ANOVA, then the results of pairwise comparison are more significant.

ANOVA uses F test statistic, which represents a standardized ratio of variability in the sample means relative to the variability within the group. The larger the observed variability in the sample means relative to the within group observations, the larger the F will be and the stronger the evidence against the null hypotheses.

ANOVA gives significant results if three conditions on the data are checked:

1. **indipendence**: *Are data indipendent?* – GSS data consist in a random sample from less than 10% of the population and so they could be considered independent
2. **approximately normal**: *Have data a normal distribution?* – normal probability plots for each group are shown in Figure 2, they visualize the difference among observations distribution and standard distribution; we can see that there is some deviation from normality in each group, especially in the upper quantiles

```
# Five graphs in a row
par(mfrow = c(1,5))
# Iterate on the groups and graph a QQ plot to test normality
degrees = c("Lt High School","High School","Junior College","Bachelor","Graduate")
for (i in 1:5) {
  qqnorm(study_data[study_data$Highest.Degree == degrees[i],]$Family.Income.Constant.USD, main=degrees[i])
  qqline(study_data[study_data$Highest.Degree == degrees[i],]$Family.Income.Constant.USD)
}
```

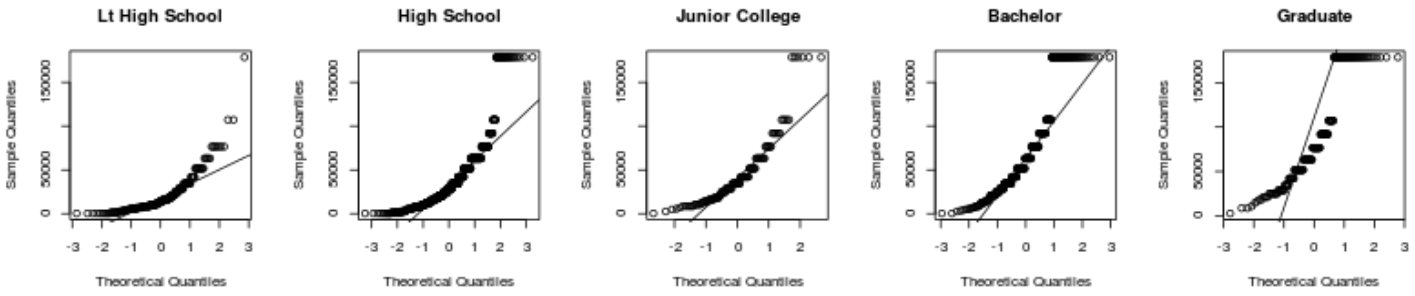

Figure 2: Check data normality in family income grouped by degree

3. **constant variance**: *Is the variability across group about equal?* – we can check the variability with the boxplot in Figure 1; we can see that the total range and the interquantile range of the groups are different, with the lowest variability in the Less than high school group and the highest variability in the Graduate group

The conditions on normality and constant variance are not fully respected. We use ANOVA in our hypotheses test, but we report the uncertainty in the results.

1.4.2 Computation

```
# ANOVA for the mean income in constanst dollars grouped by degree
anova(lm(Family.Income.Constant.USD ~ Highest.Degree, data=study_data))
```

```
## Analysis of Variance Table
##
## Response: Family.Income.Constant.USD
##                   Df   Sum Sq  Mean Sq F value Pr(>F)
## Highest.Degree    4 8.28e+11 2.07e+11     121 <2e-16 ***
## Residuals      1747 3.00e+12 1.72e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA reports a F test statistic of 121 and a p-value of approximately zero. This mean that the probability of observing a F value of 121 or higher, if the null hypotheses were true, is very low.

So we can reject the null hypotheses and we can say that the average income in constant dollar varies across some (or all) groups in a statistically significant way.

Since the null hypotheses has been rejected, we can do a pairwise comparison to find out which groups have different means.

For every possible pair of groups (10 pairs), we use a t test statistic to confirm the null hypotheses that the means of the two groups are equal ($H_0 : \mu_{Group1} = \mu_{Group2}$) or the alternative hypotheses that they are different ($H_A : \mu_{Group1} \neq \mu_{Group2}$). To avoid the increase of Type I error rate (rejecting a true null hypotheses), we apply a *Bonferroni* correction to the p-values which are multiplied by the number of comparison. With this correction, the difference of the means has to be bigger to reject the null hypotheses.

```
# Pairwise t test for the mean income in constanst dollars grouped by degree
# With Bonferroni correction
pairwise.t.test(study_data$Family.Income.Constant.USD, study_data$Highest.Degree,  p.adj="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  study_data$Family.Income.Constant.USD and study_data$Highest.Degree
##
##                Lt High School High School Junior College Bachelor
## High School    1.4e-06        -           -              -
## Junior College 3.2e-07        0.2140      -              -
## Bachelor       < 2e-16        < 2e-16     2.3e-10        -
## Graduate       < 2e-16        < 2e-16     < 2e-16        0.0011
##
## P value adjustment method: bonferroni
```

We can see that for nine group pairs the p-value is lower than the significance level of 0.05 and so the null hypotheses are rejected: the difference of the means of these nine groups is statistically significant.

The null hypotheses is not rejected for the pair High school–Junior college. The difference of the means of this pair is not statistically significant and it is due to chance.

1.5 Conclusion

The study establishes a positive correlation among the highest degree earned by United States residents and their family income in constant dollars.

We used data from the 2012 edition of General Social Survey (GSS) so that we could generalize our results to the entire United States residents population. We grouped the family income in constants dollars by the highest degree earned by the interviewees (less than high school, high school, junior college, bachelor's and graduate), and by visually exploring the data we noticed a positive correlation among the two variables.

We tested our hypotheses with ANOVA and pair comparisons and we find out that the mean incomes of the groups are significantly different from one another. The only exception being among high school degree and junior college degree: it seems that junior college degree were not better, in term of income, than a high school one.

However, these results could not be considered definitive.

We noticed in the data exploration a wide range of income and the presence of outliers in many of the groups. This mean that other variables could be strongly correlated with income. Moreover some of the conditions for the statistical inference methods used were not fully respected, and so we have to be cautious in interpreting the results.

Future research could address these shortcomings by analyzing the interaction of other variables and by using more sophisticated statistical techniques.

It could be even interesting repeating the study for other years covered by GSS survey (1972–2012) and compare the results.

1.6 References

1.6.1 Data reference

General Social Survey Cumulative File, 1972-2012 Coursera Extract. Modified for Data Analysis and Statistical Inference course (Duke University).

R dataset could be downloaded at http://bit.ly/dasi_gss_data.

Original data:

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

1.6.2 Other references

1. General Social Survey (GSS) FAQ. URL: http://publicdata.norc.org:41000/gssbeta/faqs.html. Accessed 03/30/2014
2. Comparing many means with ANOVA. In Diez M David, Barr D Christopher, Çetinkaya-Rundel Mine (2012), *OpenIntro Statistics, Second Edition*, URL: http://www.openintro.org/stat/textbook.php.

1.7 Appendix

Example of the data used in the study.

```
# First 50 observations
head(study_data, n=50L)
```

```
##       Highest.Degree Family.Income.Constant.USD
## 55088       Bachelor                     178712
## 55089    High School                     178712
## 55090    High School                      91920
```

```
## 55091    High School                    107240
## 55092       Bachelor                     42130
## 55093       Bachelor                     21065
## 55094 Junior College                     24895
## 55095 Lt High School                      4213
## 55096 Lt High School                       383
## 55097       Bachelor                     24895
## 55098 Lt High School                       383
## 55099       Bachelor                     42130
## 55100    High School                      6894
## 55101    High School                     18193
## 55103    High School                     42130
## 55104    High School                     42130
## 55105    High School                     34470
## 55106    High School                     51705
## 55107    High School                     18193
## 55109 Lt High School                     34470
## 55110 Junior College                     76600
## 55111 Junior College                    107240
## 55112    High School                     91920
## 55114    High School                    178712
## 55116       Bachelor                     34470
## 55117    High School                      6894
## 55118       Bachelor                     63195
## 55119 Junior College                     42130
## 55120       Graduate                    178712
## 55121       Bachelor                     51705
## 55122    High School                     51705
## 55123       Graduate                     51705
## 55124    High School                     76600
## 55125    High School                     34470
## 55126       Bachelor                     24895
## 55127    High School                     34470
## 55128       Graduate                     76600
## 55129    High School                     91920
## 55130       Graduate                    178712
## 55131       Bachelor                    178712
## 55132       Bachelor                    178712
## 55133       Bachelor                    178712
## 55134       Bachelor                    178712
## 55135       Graduate                    178712
## 55136       Bachelor                     51705
## 55137       Graduate                    178712
## 55138       Bachelor                     51705
## 55139       Bachelor                    178712
## 55140       Graduate                    178712
## 55141 Junior College                    107240
```

1.8 Copyright