The Catholic University of America
School of Engineering – Department of Electrical Engineering and Computer Sciences

Career Path after College Degree in Maryland in 2013

Project 1

CSC530 – Introduction to Data Analysis

By: Vy Bui and Duong Le

Advisor: Dr. Le He

Fall 2014

I. Introduction

This report focuses on the statistical modeling and analysis results associated with the study of career after college degree. One of the most popular question of high school students may find themselves asking, "Why is it important to go to college?" The answer may become different in some particular way but one of our best answer is to have more opportunity. It serves as the gateway to students' future endeavor. In this report, earnings is be considered as a primary unit of information about the career pathways of college students after completing their degree. There are many reports from The United States Census Bureau about education attainment and earnings. This report serves as an addition information about the earnings and majors of college degree (bachelors or higher). This report also documents the inference techniques used during the subsequent statistical analyses.

The remainder of this report organized as follows. Section II gives the details about how we obtained the public data. Section III introduces the background review about the state-of-the-art techniques for this problem or similar problems. Section IV presents the implementation approach. Section V provides results analysis. Section VI concludes about the lessons and the future work of this study.

II. Data Collection

1. American Community Survey

The raw data was collected in the American Community Survey (ACS) in 2013. There are many subjects in the person record including age, citizenship, ability to speak English, hours worked, health insurance, occupation, place of work, etc. In this report, we only take into account the records of their highest degree (a bachelor's degree, master's degree, professional degree, or doctorate degree) and the majors of their bachelor's degree. The data included the annual earnings in the year of 2013 of Maryland residents who were employed full-time.

2. Assumptions and Limitations of ACS Data[1]

ACS collects information about respondents only one time, we don't know about their job history or what changes they might make in their future careers. Thus we assume that a person will work in the same occupation, full-time without changing their level of education and the economy will stay the same without inflation. In reality, the economy is constantly changing. Education and experiential requirements for an occupation may change.

ACS does not collect data on professional training and certifications that may impact the way respondents answer the survey questions.

ACS does not collect data on field of degrees other than the bachelor's degree. For example, a person may major in any subject and then go on to get Master's in Business Administration (MBA), Masters of Teaching (MAT) or law degree (JD), a medical degree (MD) or a doctorate (PhD) which often lead to certain types of professions. These subject areas of the advanced degrees are likely to impact earnings more than the bachelor's field. Additionally, one could go on to an academic master's or doctorate degree in an unrelated area or in their minor area and then go on to an occupation seemingly unrelated to either degree, the data don't allow us to know this kind of detail.

III. Background Review

In this course, we learn about the statistical analysis, which applies the confidence interval, hypothesis tests, and linear/nonlinear regression techniques to approximate the model. In this section, we briefly review what learn in Statistics. The details of the techniques and why we used them is presented Section IV.

1. **Confidence Interval**

   Confidence interval for population mean gives indication of how accurately the sample mean estimate the population mean. A 95% confidence interval is defined as an interval calculated in such a way that if a large number of samples were drawn from a population and the interval calculated for each of these samples, 95% of the intervals will contain the true population mean value.

2. **Hypothesis Tests**

   A hypothesis test uses a sample to test hypothesis about the population from which a sample is drawn. This helps us make decisions or draw conclusions about the population. A hypothesis test has the following components:

   Null Hypothesis H_0: is a hypothesis about the population from which a sample or samples are drawn. It is usually a hypothesis about the value of an unknown parameter such as the population mean or variance.

   e.g. H_0: The population mean is equal to five. The null hypothesis is adopted unless proven false.

   Alternative Hypothesis $H_1$is the hypothesis that will be accepted if there is enough evidence to reject the null hypothesis.

3. **Linear/Nonlinear Regression**

   A general linear model is used to predict the value of a continuous variable (known as response variable) from one or more explanatory variables. A general linear model takes the form:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

   Where y is the response variable, $x_n$are the explanatory variables, β_n are coefficients to be estimated and ε represents the random error. The random errors are assumed to be independent, to follow a normal distribution with a mean of zero, and to have the same variance for all values of the explanatory variables. Simple linear regression, multiple linear regression, polynomial regression, analysis of variance, two-way analysis of variance, analysis of covariance, and experimental design models are all types of general linear model.

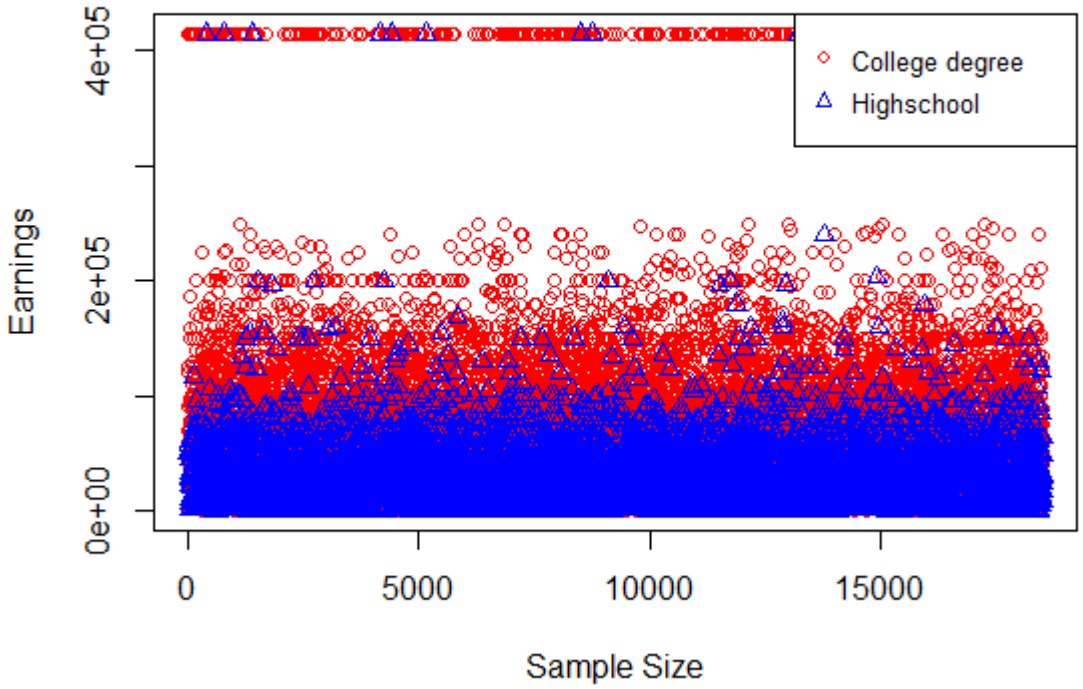
IV. Implementation approach & Results analysis

In this study, we use R [2] [3] to analyze data, create statistical plots, performs hypothesis tests and build regression models. R is open-source software and free to install under the GNU general public license. R has over 5000 add-on packages covering a broad range of statistical techniques.

A sample data size includes 57701 person records for all education attainments (Grade 1 – 11, 12th grade with no diploma, regular high school diploma, GED or alternative credential, 1 or more years of college credits with no degree, associate's degree, bachelor's degree and higher). We analyze those person records have education attainment beyond high school. Thus among 57701 person records, there are 12452 people having bachelor's degree or higher and 6070 people having high school diploma.

Figure 1 visualizes the earnings of people with college degree vs. those with high school diploma in Maryland in 2013. We can see that the higher of their degree, the more income a person can get. There are some people with high school diploma earn enormously (up to $415000), we consider these as outliners. One question may arise whether or not to exclude or include these outliners in the datasets, our approach is to look more into the person records such as: occupation, age/ work experience. We found 326 people having the wages or salary income equal $415000 in 2013. Among that, there are 11 people holding different positions: Construction manager, Sales representatives, wholesale and manufacturing, Medical and health services mangers, Chief executives and legislators, Maintenance and repair workers, general, supervisors of retail sales workers, etc only have high school diploma. Because ACS individual records do not include work experience so we take age into account. In 11 people holding high school degree, there are 9 people over 40 years old. We interpret these special 9 people as people having a great deal of experiences or people that work in the same organization for a long time, the salary/position they currently hold can be seen as the results of long-life working. Only 2 people have the maximum salary are under 40 years old. One is farmers, ranchers, and other agricultural managers 27 years old and another is funeral service manager and postmasters and mail superintendents 37 years old. As we mentioned in part II, the respondents do not get any professional training while answering the survey question, from this point of view we may want to exclude the 2 people from the datasets. However since the sample size are large, including the records of these 2 people will not impact much on the results of the data so we keep them.

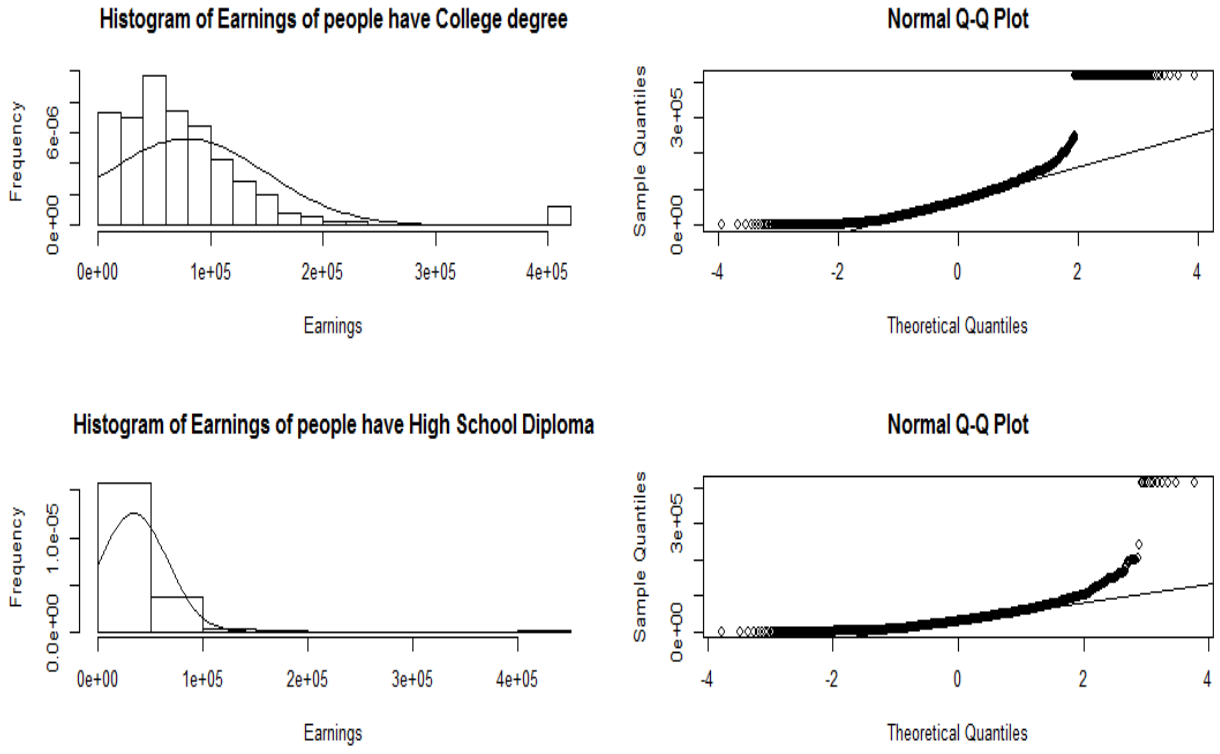
Figure 1. Education Attainment vs. Earnings



```
> summary(WCollege$WAGP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     20   37000   65000   78940  100000  415000
> summary(HighSchool$WAGP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     40   13000   29000   34200   48000  415000
```

The above R code shows that the in general, people have College degree tend to earn more than one with High school diploma. The histogram in Figure 1a shows that the earning of people having College degree is likely to sparse while histogram in Figure 1b indicate the earning of people having High school diploma have a tendency to concentrate to lower income. Figure 1c and 1d is a normal probability plot to determine whether a sample is drawn from a normal distribution. The way in which the data points fall around the straight line tells us something about the shape of the distribution relative to the normal distribution. These normal probability plots appear in Figure 1c and 1d allow us to conclude that the normality assumption is not justified. Apart from the outliers, data points curve from the above line to below the line and then back to above the line. We may conclude that data has a positive skew/ right-skewed.

Figure 1
(a) (c)
(b) (d)



The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger. In other words, no matter what the shape of the population distribution. This fact holds especially true for larger sample sizes over 30. All this is saying is that as we take more samples, especially large ones, our graph of

the sample means will look more like a normal distribution. In ACS datasets, we have sample size equal to 57701 observations so the sampling means of people income approaches a normal distribution. This is important because the following of this report use the methods that's true for normal distribution.

Confidence interval for population mean gives indication of how accurately the sample mean estimate the population mean. A 95% confidence interval is defined as an interval calculated in such a way that if a large number of samples were drawn from a population and the interval calculated for each of these samples, 95% of the intervals will contain the true population mean value. The simplest way to obtain a confidence interval for sample mean is with Student t's test. We run the test in R with the data which contains the records of income of people hold bachelor degree or higher. The output is:

```
One Sample t-test
data:  WCollege$WAGP
t = 123.9757, df = 12452, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 77694.42 80190.71
sample estimates:
mean of x
 78942.56
```

From the results, we can see that the mean income of people with college degree is $78942.56, with a 95% CI of $77694.42 and $80190.71.
Similarly with the data of people hold high school diploma.

```
One Sample t-test
data:  HighSchool$WAGP
t = 84.3468, df = 6069, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 33402.73 34992.34
sample estimates:
mean of x
 34197.54
```

From the results, we can see that the mean income of people with college degree is $34197.54, with a 95% CI of $33402.73 and $34992.34. This is another evidence saying that with higher degree, the more income a person can get.
Hypothesis testing is useful because it allow us to use data as evidence to support or refute a claim, and from there we can make decisions. In our experiments, we run one-tailed t-test to see whether the income of people with college degree is greater than $60000.

```
> t.test(WCollege$WAGP, mu=60000, conf.level = 0.95,
alternative="greater")
   One Sample t-test
```

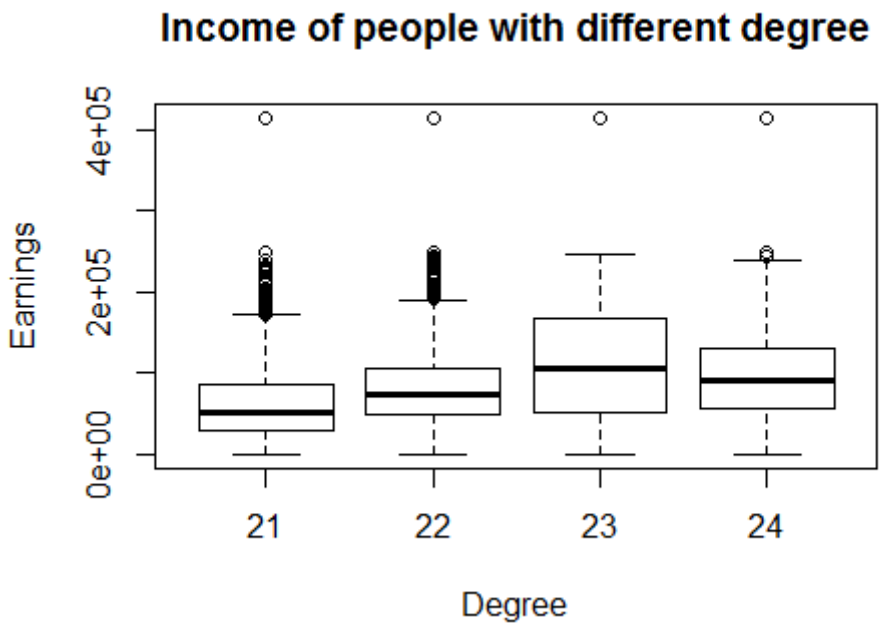
```
data:  WCollege$WAGP
t = 29.7484, df = 12452, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 60000
95 percent confidence interval:
 77895.11      Inf
sample estimates:
mean of x
 78942.56
```

From the output, we can see that the mean income of people with college degree is $78942.56. The one-side 95% CI tells us that the mean income is likely to be greater than $77895.11. The p-value of 2.2×10^{-16} tells us that if the mean income were $60000, the probability of selecting a sample with mean income greater than or equal this one would be very low percentage. Because the p-value is less than the significance level of 0.05, we can reject the null hypothesis that the mean income of people with college degree is equal to $60000.

```
> t.test(WCollege$WAGP, mu=40000, conf.level = 0.95,
alternative="less")
   One Sample t-test
data:  WCollege$WAGP
t = 61.1575, df = 12452, p-value = 1
alternative hypothesis: true mean is less than 40000
95 percent confidence interval:
     -Inf 79990.02
sample estimates:
mean of x
 78942.56
```

From the output, we can see that the mean income of people with college degree is $78942.56. The one-side 95% CI tells us that the mean income is likely to be greater than $79990.02.11. The p-value of 1 tells us that if the mean income were $40000, the probability of selecting a sample with mean income less than or equal this one would be very low percentage. Because the p-value is not less than the significance level of 0.05, we cannot reject the null hypothesis that the mean income of people with college degree is equal to $40000. This means that there is no evidence that the income of people with college degree is less $40000.

We can see from the boxplot below that there are some differences in the income between the groups of people holding different degrees. Again, we want to denote that the degree levels 21 to 24 are Bachelor, Mater, Professional degree beyond college degree and Docterate degree, respectively. We wish to perform an analysis of variance to determine whether these differences are statistically significant. We assume that the income is normally distributed and that the variance is the same for all 4 groups. A significance level of 0.05 is used.




```
Analysis of Variance Table
Response: WCollege$WAGP
              Df     Sum Sq    Mean Sq F value    Pr(>F)
SCHL           1 3.6148e+12 3.6148e+12  759.53 < 2.2e-16 ***
Residuals 12451 5.9258e+13 4.7593e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that the p-value of group effect is 2.2×10^{-16}. This means that if the effect of all 4 groups of different degree were the same, we would have very low percentage of seeing differences between groups as large or larger than this. As the p-value is less than the significance level of 0.05, we can reject the null hypothesis that the mean income is the same for all four groups., in favor of the alternative hypothesis that the mean income is different for at least one pair of groups.
This study is continued where we use pairwise pairwise t-tests to further investigate the differences between individual pairs of groups after performing an analysis of variance.

```
Pairwise comparisons using t tests with pooled SD
data:  WCollege$WAGP and WCollege$SCHL
   21      22      23
22 < 2e-16 -       -
23 < 2e-16 < 2e-16 -
24 < 2e-16 1.3e-11 < 2e-16
P value adjustment method: bonferroni
```

Before examining the results, we want to denote that the degree levels 21 to 24 are Bachelor, Mater, Professional degree beyond college degree and Docterate degree, respectively. The

output shows p-value for each of the comparisons. From the output, we can see that the comparison of Bachelor and Master is statistically significant. This means that the income for Bachelor is significantly different from the income of Master degree. Similar to other pairs test.

Model building helps us to understand the relationships between variables and to make predictions about the future observations. In this part, we build regression models and other models in the general linear model family to predict the income of a person by taking into account what degree they have and what is the major in their bachelor degree.

```
Call:
lm(formula = WCollege$WAGP ~ SCHL + FOD1P, data = WCollege)
Coefficients:
(Intercept)         SCHL        FOD1P
 -3.346e+05    1.890e+04    8.181e-01
```

From the output, we can see that formula has two coefficients that express the effect of degree (SCHL) and majors (FOD1P) on the income of a person (WAGP). So the model formula is:

$$Income = -3.346 \times 10^5 + SCHL + FOD1P$$

```
Call:
lm(formula = WCollege$WAGP ~ SCHL + FOD1P, data = WCollege)
Residuals:
    Min      1Q  Median      3Q     Max
-124085  -39313  -13301   22590  351598
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.346e+05  1.500e+04 -22.308   <2e-16 ***
SCHL         1.890e+04  6.842e+02  27.630   <2e-16 ***
FOD1P        8.181e-01  3.681e-01   2.222   0.0263 *  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 68980 on 12450 degrees of freedom
Multiple R-squared: 0.05787,    Adjusted R-squared: 0.05772
F-statistic: 382.4 on 2 and 12450 DF,  p-value: < 2.2e-16
```

The hypothesis tests for the model coefficients tell us that the intercept, degree and major are significantly different from 0. The p-value for the F-test is less than 0.05, which tells us that the model explains a significant amount of variation in the income of the data.
To view confidence interval for the coefficients estimates, we run the following code.

```
> confint(model, level=0.95)
                    2.5 %        97.5 %
(Intercept) -3.640294e+05 -3.052230e+05
SCHL         1.756200e+04  2.024409e+04
FOD1P        9.651589e-02  1.539775e+00
```

From the output, we can see that the 95% CI for the intercept is -3.640294×10^5 to $-3.052230\text{e} \times 10^5$. CI for SCHL and FOD1P can be read accordingly.
So we have the model that we want, we now want to use it to make predictions for new data. R has a convenient function for making predictions, called `predict`. Once our new data is arranged in a data frame, we run the following codes and observe the output:

```
> newdata<-data.frame(SCHL=c(21, 24), FOD1P=c(2102, 2102))
> newdata$predictions<-predict(model, newdata,
interval="confidence", level=0.95)
> newdata
  SCHL FOD1P predictions.fit predictions.lwr predictions.upr
1   21  2102        64057.54        61806.56        66308.51
2   24  2102       120766.68       117120.65       124412.71
```

From the output, we can see that for a person has bachelor in Computer Science has an income about $64057.54 while a person has a PhD in Computer Science earns $120766.68 which makes logic results.

V. Conclusion & Future work

In this study, we learn about R programming in Statistics. This practice helps us understand more about the theory we cover in the course. Finding real public data for our experiments is also a challenging task, it is because the ACS collected a large number of observations so they arrange their data title into particular codelists - ACSPUMS2013CodeLists.xls. We attached this document in the submission. It took us sometimes to see the whole picture of their data.

In the future, we can use the same methodology for the income of particular major and degree. For example, Figure 2 compare the wage/salary of bachelors major in Computer Science (CS) vs. those major in Education.

Fig 2a. Earnings of Bachelors-Computer Science

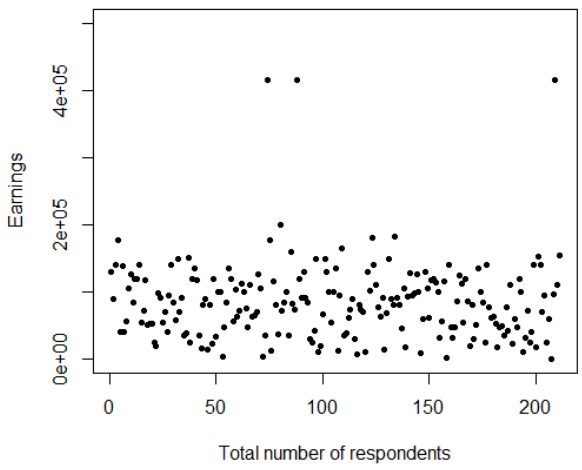
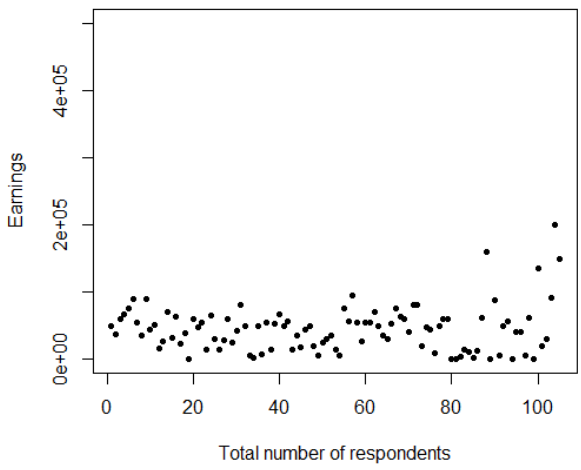
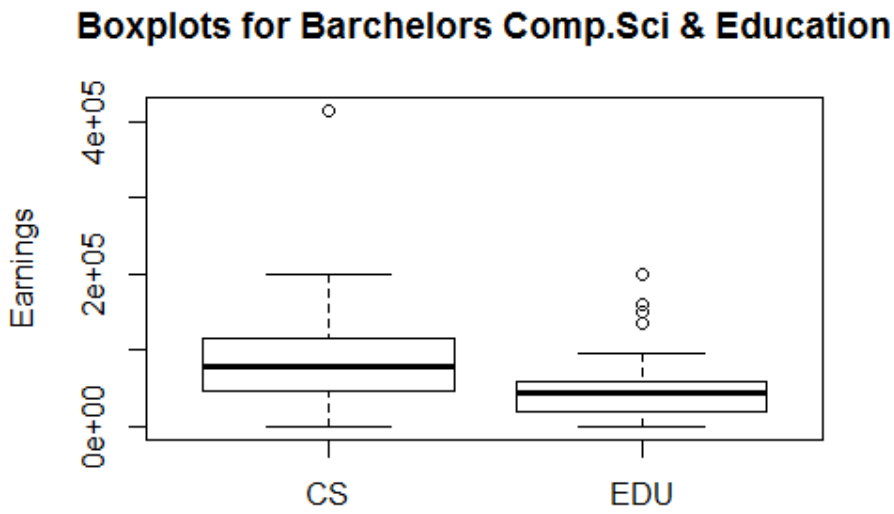


Fig 2b. Earnings of Bachelors-General Education



The following R codes produce a summary of the income of CS and Education Bachelors including the mean, median, range, and interquartile range.

```
> summary(BachelorCS$WAGP)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
80   47500   80000   85210  115500  415000
> summary(BachelorEdu$WAGP)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
520   18800   44000   45140   60000  200000
```



The comparative boxplots show that people have Bachelors in CS tend to have higher income than Education. There are one outlier among the data for CS major, the occupation of this person is in in the category of Chief executives and Legislators. He/she earns $415000 in 2013 which is reasonable. There is also some Education major people's income that barely qualifies as an outlier. The boxplot for CS major is a bit taller, and the lower whisker a bit longer, than that for Education major people. We conclude that apart from the outliers, the spread in income is larger for CS major people and is much larger when the outliers are considered.

This report only do experiments on datasets of people in Maryland in 2013. We can investigate more in other states in the US and from different years. That will benefit us to see the overall picture of career path of people with different education attainment. This also help the young people who haven't decided where to go on with their pathways to see the projection of older people in specified career path.

VI. Appendix

1. ACS Data details

ACS denote the educational attainment levels as follows:

bb .N/A (less than 3 years old)
01 .No schooling completed
02 .Nursery school, preschool
03 .Kindergarten
04-14 .Grade 1 – Grade 11
15 .12th grade - no diploma

16 .Regular high school diploma
17 .GED or alternative credential
18 .Some college, but less than 1 year
19 .1 or more years of college credit, no degree
20 .Associate's degree
21 .Bachelor's degree
22 .Master's degree
23 .Professional degree beyond a bachelor's degree
24 .Doctorate degree

2. Our R codes

```
# Read datasets into R
Datasets<-read.csv("C:/Users/vbui/Google Drive/CUA Fall
2014/CSC530/Project1/ss10pmd1.csv")
# Extract the data into subset included People with College
and Highschool degree
WCollege<-subset(Datasets, Datasets$SCHL>=21 & Datasets$WAGP >
0)
HighSchool<-subset(Datasets, Datasets$SCHL==16 &
Datasets$WAGP > 0)

# Summary: mean, median, range and interquartiles range. How
many missing values a variable has.
summary(WCollege$WAGP)
summary(HighSchool$WAGP)
# Calculate the mean, standard deviation and variance
Mean.WCollege<-mean(WCollege$WAGP)
Mean.WCollege
Mean.HighSchool<-mean(HighSchool$WAGP)
Mean.HighSchool
Std.WCollege<-sd(WCollege$WAGP)
Std.WCollege
Std.HighSchool<-sd(HighSchool$WAGP)
Std.HighSchool
Variance.WCollege<-var(WCollege$WAGP)
Variance.WCollege
Variance.HighSchool<-var(HighSchool$WAGP)
Variance.HighSchool

# A hypothesis test that can help to determine whether a
sample has been drawn from a normal distribution
# The null hypothesis for the test is that the sample is drawn
from a normal distribution
shapiro.test(Datasets$WAGP[3:5000])
shapiro.test(HighSchool$WAGP[3:5000])
# Conclude: Not follow normal distribution.

# Confidence Interval for a sample mean with Student's T-Tests
```

```
t.test(WCollege$WAGP, conf.level = 0.95)
t.test(WCollege$WAGP, mu=40000, conf.level = 0.95,
alternative="greater")
t.test(WCollege$WAGP, mu=40000, conf.level = 0.95,
alternative="less")
t.test(HighSchool$WAGP, conf.level = 0.95)
t.test(HighSchool$WAGP, mu=30000, conf.level = 0.95,
alternative="greater")
t.test(HighSchool$WAGP, conf.level = 0.95, alternative="less")

boxplot(WCollege$WAGP~SCHL, WCollege,main="Income of people
with different degree", xlab="Degree",ylab="Earnings")
# Analysis of Variance
EarningsANOVA<-aov(WAGP~SCHL, WCollege)
anova(EarningsANOVA)
coef(EarningsANOVA)
TukeyHSD(EarningsANOVA)
pairwise.t.test(WCollege$WAGP, WCollege$SCHL,
p.adj="bonferroni")
bartlett.test(WAGP~SCHL, WCollege)
var.test(WAGP~SCHL, WCollege)
# Prediction Interval
predict(lm(WCollege$WAGP~1), interval="prediction",
level=0.90)[1, ]
predict(lm(HighSchool$WAGP~1), interval="prediction",
level=0.95)[1, ]
# ?? Meaning of Prediction Interval??

# Bootstrap goes here
library(boot)
# function to obtain R-Squared from the data
rsq <- function(formula, data, indices) {
  d <- data[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(coef(fit))
}
# bootstrapping with 1000 replications
results <- boot(data=WCollege, statistic=rsq, R=1000,
formula=WCollege$WAGP~SCHL+FOD1P)
# view results
results
plot(results)
# get 95% confidence interval
boot.ci(results, conf = 0.95, type=c("basic"))
```

```
# Frequency tables: summarize a categorical variable by
displaying the number of observations belonging to each
category
# The below command shows how many people get a specified
degree (>=21 for bachelors and higher) vs. their income
table(Datasets$WAGP, useNA="ifany")

# Creating plots
# This plot tells Education Attainment vs. Income
plot(WCollege$WAGP[1:500], pch=1, col=2, main="Figure 1.
Education Attainment vs. Earnings", xlab="Sample Size",
ylab="Earnings")
par(new=T)
plot(HighSchool$WAGP[1:500], pch=2, col=4,
axes=F,xlab="",ylab="")
par(new=T)
legend("topright", legend=c("College degree", "Highschool"),
pch=c(1,2), col=c(2,4), cex=0.8)

# Find records of people with maximum income
Max_College_HS<-subset(Datasets, (Datasets$SCHL>=21 |
Datasets$SCHL==16)& Datasets$WAGP==415000)
Max_College_HS<-subset(Datasets, Datasets$SCHL==16 &
Datasets$WAGP==415000 & Datasets$AGEP>50)
Max_College_HS<-subset(Datasets, Datasets$SCHL==16 &
Datasets$WAGP==415000 & Datasets$AGEP<50)
HighSchool_minus2<-subset(HighSchool, !HighSchool$row.names ==
27632)
Max_College_HS<-subset(HighSchool_minus2,
HighSchool_minus2$WAGP==415000 & HighSchool_minus2$AGEP<50)
# \n 0-15: Less than High School diploma\n16: Regular High
School Diploma\n17-20: With High School diploma but no college
degree\n21-24: bachelor's degree and beyond
# This plot tells income of Bachelors, Masters, Professional
degree beyond bachelor and Docterate
plot(WCollege)
plot(WCollege$SCHL, WCollege$WAGP, pch=20, main="Fig 1.
Education Attainment vs. Earnings", xlab="Education
Attainment", ylab="Earnings")
# Earnings of All majors Bachelors
Bachelor<-subset(Datasets, Datasets$SCHL==21 & Datasets$WAGP >
0)
plot(Bachelor$WAGP, pch=20, main="Fig 2. Earnings of
Bachelors, All majors", xlab="Total number of respondents",
ylab="Earnings")
# Earnings of Computer Science Bachelors
par(mfrow=c(1,4))
```

```
BachelorCS<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==2102)
plot(BachelorCS$WAGP, pch=20, main="Fig 3a. Earnings of
Bachelors, Computer Science", xlab="Total number of
respondents", ylab="Earnings")
BachelorPsycho<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==5200)
plot(BachelorPsycho$WAGP, pch=20, main="Fig 3b. Earnings of
Bachelors, Psychology", xlab="Total number of respondents",
ylab="Earnings")
BachelorHistory<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==6402)
plot(BachelorHistory$WAGP, pch=20, main="Fig 3c. Earnings of
Bachelors, History", xlab="Total number of respondents",
ylab="Earnings")
BachelorBussiness<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==6200)
plot(BachelorBussiness$WAGP, pch=20, main="Fig 3d. Earnings of
Bachelors, Bussiness", xlab="Total number of respondents",
ylab="Earnings")

par(mfrow=c(1,2))
BachelorCS<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==2102)
plot(BachelorCS$WAGP, pch=20, ylim=c(1,500000), main="Fig 2a.
Earnings of Bachelors-Computer Science", xlab="Total number of
respondents", ylab="Earnings")
BachelorEdu<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==2300)
plot(BachelorEdu$WAGP, pch=20, ylim=c(1,500000), main="Fig 2b.
Earnings of Bachelors-General Education", xlab="Total number
of respondents", ylab="Earnings")
summary(BachelorCS$WAGP)
summary(BachelorEdu$WAGP)
BachelorCS_EDU<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & (Datasets$FOD1P==2102 |
Datasets$FOD1P==2300))
boxplot(BachelorCS_EDU$WAGP~FOD1P, BachelorCS_EDU, main =
"Boxplots for Barchelors Comp.Sci & Education", xaxt = "n",
xlab = "", ylab = "Earnings")
axis(1, at=1:2, labels=c("CS", "EDU"))

#Plot for differnt degree of the same majors
#CS<-subset(Datasets, (Datasets$SCHL==21 | Datasets$SCHL==22 |
Datasets$SCHL==23) & Datasets$WAGP > 0 & Datasets$FOD1P==2102)
CS.Bachelors<-subset(Datasets, Datasets$SCHL==21 &
Datasets$WAGP > 0 & Datasets$FOD1P==2102)
```

```
CS.Masters<-subset(Datasets, Datasets$SCHL==22 &
Datasets$WAGP > 0 & Datasets$FOD1P==2102)
CS.PhDs<-subset(Datasets, Datasets$SCHL==24 & Datasets$WAGP >
0 & Datasets$FOD1P==2102)
plot(CS.Bachelors$WAGP, col=2, xlim=c(1,400),xlab="Samples
Size",ylab="Earnings",main="Earnings of CS, All degrees")
par(new=T)
plot(CS.Masters$WAGP, col=4, axes=F,xlab="",ylab="")
par(new=T)
plot(CS.PhDs$WAGP, pch=10, axes=F,xlab="",ylab="")
par(new=T)
legend(locator(1), legend=c("Bachelor", "Master", "PhD"),
col=par("red", "blue", "black"), pch=c(1,1,10))

# Histogram
hist(Datasets$WAGP)
hist(WCollege$WAGP, main = "Histogram of Earnings of people
have College degree", xlab = "Earnings", ylab = "Frequency")
hist(HighSchool$WAGP, main = "Histogram of Earnings of people
have High School Diploma", xlab = "Earnings", ylab =
"Frequency")

# Fit a normal distribution curve to the data
hist(Datasets$WAGP,freq=F)
curve(dnorm(x, mean(Datasets$WAGP), sd(Datasets$WAGP)), add=T)

par(mfrow=c(2,2))
hist(WCollege$WAGP,freq=F, main = "Histogram of Earnings of
people have College degree", xlab = "Earnings", ylab =
"Frequency")
curve(dnorm(x, mean(WCollege$WAGP), sd(WCollege$WAGP)), add=T)
qqnorm(WCollege$WAGP)
qqline(WCollege$WAGP)
# Conclusion: Data points curve from above the line to below
the line and then back to above the line. Data has positive
skew/ right-skewed
hist(HighSchool$WAGP,freq=F, main = "Histogram of Earnings of
people have High School Diploma", xlab = "Earnings", ylab =
"Frequency")
curve(dnorm(x, mean(HighSchool$WAGP), sd(HighSchool$WAGP)),
add=T)
qqnorm(HighSchool$WAGP)
qqline(HighSchool$WAGP)

# A normal probability plot to determine whether a sample is
drawn from a normal distribution
qqnorm(Datasets$WAGP)
```

```
qqline(Datasets$WAGP)

# Conclusion: A sample is not drawn from a normal distribution

# Stem-and-Leaf Plots
stem(Datasets$WAGP)
stem(WCollege$WAGP)
stem(HighSchool$WAGP)

# Bar Charts
barplot(table(Datasets$WAGP, Datasets$SCHL, useNA="ifany"))
barplot(table(WCollege$WAGP, WCollege$SCHL, useNA="ifany"))

# Boxplot
boxplot(Datasets$WAGP~SCHL, Datasets)
boxplot(WCollege$WAGP~SCHL, WCollege)
boxplot(HighSchool$WAGP~SCHL, HighSchool)

#CI & Hypothesis
# Kruskal-Wallis Test
kruskal.test(WCollege$WAGP~SCHL, WCollege)
pairwise.wilcox.test(WCollege$WAGP, WCollege$SCHL)

# Wilcoxon Rank-Sum test
wilcox.test(Bussiness$Earnings, mu=60000, alternative="less",
conf.level=0.95)

# Build a model to predict a person's income from their degree
and their majors
#model<-lm(Datasets$WAGP~SCHL+FOD1P, Datasets)
model<-lm(WCollege$WAGP~SCHL+FOD1P, WCollege)
formula(model)
summary(model)
coef(model)
confint(model, level=0.95)
residuals(model)
WCollege$resids<-rstudent(model)
WCollege$fittedvals<-fitted(model)
# Create residual plots for a model object

plot(resids~WAGP, WCollege)
plot(resids~SCHL, WCollege)

# Leverage
hatvalues(model)
# Plot of the residuals against the leverage
#Cook's Distance
```

```
cooks.distance(model)
par(mfrow=c(3,2))
hist(WCollege$resids)
plot(model, which=2)
plot(model, which=1)
plot(resids~WAGP, WCollege)
plot(resids~SCHL, WCollege)

plot(model, which=4)
plot(model, which=5)
plot(model, which=6)

newdata<-data.frame(SCHL=c(21, 21, 24, 24), FOD1P=c(2300,
2102, 2300, 2102))
newdata$predictions<-predict(model, newdata,
interval="confidence", level=0.95)
newdata
```

VII. References

[1] PUMS documentation of American Community Survey

(http://www.census.gov/acs/www/data_documentation/pums_documentation/, assessed on October 2014)

[2] S. Stowell, "Using R for Statistics", Apress 1 edition. June 23, 2014
[3] Online Course about R programming

(http://www.lynda.com/R-tutorials/Using-single-mean-Hypothesis-test-confidence-interval/142447/149950-4.html, assessed on October 2014)