STAT 429: Linear Models and Their Applications

Using MLR to Determine Factors Affecting Total Educational Attainment

December 10, 2021

Author *UCID*

[redacted] [redacted]

Daniel Girvitz

[redacted]

Redacted 2022-06-19

1 Introduction

In the United States of America, community colleges offer a two-year alternative to bachelor's degrees called "associate degrees". Having completed such a degree at a community college, a student may transfer to another college where he can complete a bachelor's degree, if he so chooses; not all students choose to do so.

In April 1995, Cecilia Elena Rouse published an article in the Journal of Business & Economic Statistics entitled "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment". Ms. Rouse attempted to determine if "community colleges may provide a place in higher education for those who otherwise might not attend college either for financial reasons or because they were not yet ready to attend four-year college" or if "local community colleges may also draw away students who might otherwise have attended a four-year college". The former had been previously termed *the democratization effect*, the latter – *the diversion effect*. Sociological reasons aside, which we are not interested in, Ms. Rouse "found" that a closer local community college will increase total years of schooling slightly and not likely change the likelihood of receiving a bachelor's degree".

Initially our goal of the project was to determine if we could predict the education level that a student would obtain based on socioeconomic factors however, due to further research we realized that our initial approach could not be analyzed through MLR. The scope was too advanced to truly grasp the understanding of our data set so we decided to change the response variable to composite test score which is graded at the end of the 12th education level (aka grade 12). This variable would make much more sense to use because score is continuous and education level is categorical and in MLR models you cannot use categorical data as your response variable. We also realized that the composite test score is heavily dependent with the individual's education level which is why we chose to solely analyze the composite test score as we are able to get a better understanding of the data.

2 Analysis

We chose to use MLR (multiple linear regression) initially to analyze the data obtained in order to find the effects that each socioeconomic element has on an individual's composite test score. The elements that we found has the greater effects on an individual's test score and education are gender, unemployment, family income, ethnicity, parent's education, and wage.

The data frame we use includes 4,739 observations on 14 variables. Every section in this report includes the necessary information to understand the analysis. Further details are located in the appendix.

First we will explore the data to better understand the data set by representing and splitting the individual variables visually in several different ways. After this, we will confirm the LINE assumption and then will create an MLR model to better analyze the data and variables.

2.1 Education Level vs Composite Test Score/Gender

By inspection, we found that there was no significant statistical difference between the genders of male and female. When we had represented this data in the form of box plots, both for education levels as well as composite test scores, we were unable to solely trust our eyes for analyzing the difference. Thus were turned to math in order to hopefully uncover the hidden information.

First, we tried the Bartlett and Levene's test on the data, which concluded that for both the educational level and composite test scores, we failed to reject the null hypothesis of the equal variance assumption. Next, we built a SLR (Simple Linear Regression) model with one categorical predictor, and had failed to reject the equality of means on total education attainment. This is also known as a gender treatment effect. However, since the MLR model tells us something different, we decided to dig deeper into the data.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.00	12.00	13.00	13.83	16.00	18.00

Figure 1: Summary of Education Level MALE

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.00	12.00	13.00	13.79	16.00	18.00

Figure 2: Summary of Education Level FEMALE

As one can see, all the measures are equal between the two genders except for the mean, whose difference is statistically insignificant. This treatment effect can be confirmed by a linear model of education.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28.95	44.77	52.02	51.66	58.81	71.36

Figure 3: Summary of Composite Test Score MALE

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
30.91	43.39	50.23	50.26	56.87	72.81

Figure 4: Summary of Composite Test Score FEMALE

With a linear model of the student's composite test score, with the factor of gender, we find that there is indeed a significant statistical difference between treatment effects. This difference has a p value of 3.27e-08. According to the model, males average 1.4017 points higher than females.

As one can see, the confidence intervals for each education level increases until it gets to level 18. For both male and female, the scores seem drop between levels 15 and 16. For each confidence intervals however, males have a slightly higher interval than females.

Basically, as composite test scores increases, we can expect a higher level of education.

If one looks at the data collected for 18 years of education, there is a dip in this data. An inference about why this anomaly occurs is that it is a small sample size for this level. The graphs that would represent this showed that the highest count for education level 18 is 9 individuals, where as education level 14 has multiple columns with counts above 40. Therefore, the small sample size could be affecting the confidence intervals for education level 18 which could explain the discrepancy.

Ed 12	Ed 13	Ed 14	Ed 15	Ed 16	Ed 17	Ed 18
(46.5, 46.9)	(50.3, 50.8)	(52.1, 52.5)	(54.9, 55.3)	(57.2, 57.6)	(58.7, 59.1)	(56.0, 56.4)

Figure 5: Confidence Intervals: Education Level vs Composite Test Scores MALE

Ed 12	Ed 13	Ed 14	Ed 15	Ed 16	Ed 17	Ed 18
(46.4, 46.8)	(48.4, 48.8)	(49.9, 50.3)	(52.3, 52.7)	(55.1, 55.5)	(57.4, 57.8)	(54.1, 54.7)

Figure 6: Confidence Intervals: Education Level vs Composite Test Scores FEMALE

When comparing education levels by composite test scores, we see that as the composite test scores increase so does the education level. This occurs all the way up until the 18th education level where we can see that the test scores drop back down for both males and females. However, the general trend is that as the composite test scores increases, so does the educational level. A possible reason for this outlying trend could be that there is a very small sample size for education level 18, along with the combination of a large portion of people with low test scores.

2.2 Education Level vs Composite Test Score/Unemployment

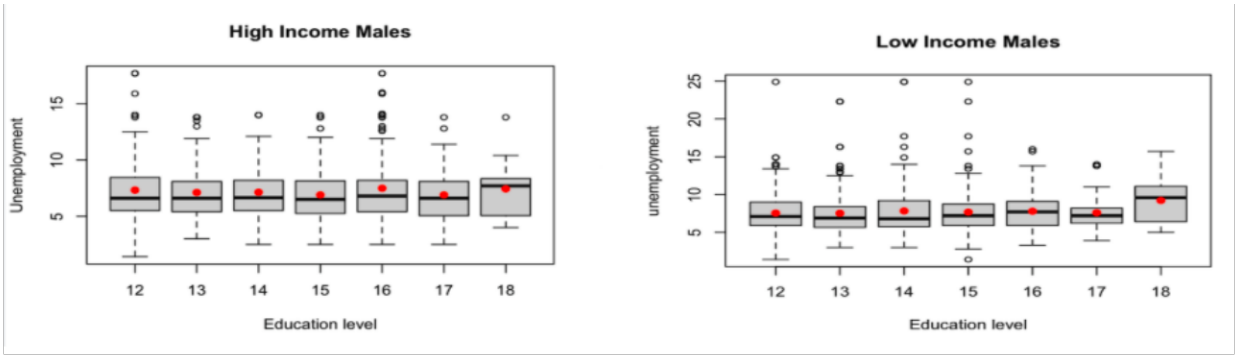

Figure 7: Education Level vs Unemployment with High and Low Family Income MALE

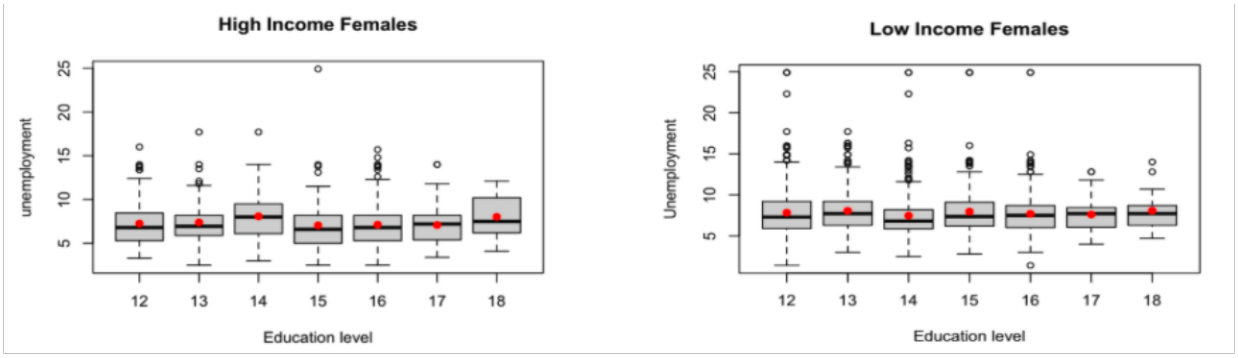

Figure 8: Education Level vs Unemployment with High and Low Family Income FEMALE

A linear model of education versus unemployment tells us that the country unemployment rate in 1980 does not have a significant statistical influence on educational attainment. We had conducted a two-tail test and obtained a p-value of 0.503 which is greater than our α value of 0.05. Therefore, we fail to reject the null hypothesis and that $\beta_1 = 0$.

When we fixed family income, and chose either "low" or "high" income, we found that the country unemployment rate in 1980 still did not have a significant statistical influence on educational attainment. From another statistical test, we found that we also failed to reject the null hypothesis.

This conclusion is also the same when looking at the female data for the same fixed elements.

2.3 Education Level vs Composite Test Score/Family Income

2.3.1 MALE Education vs Composite Test Score

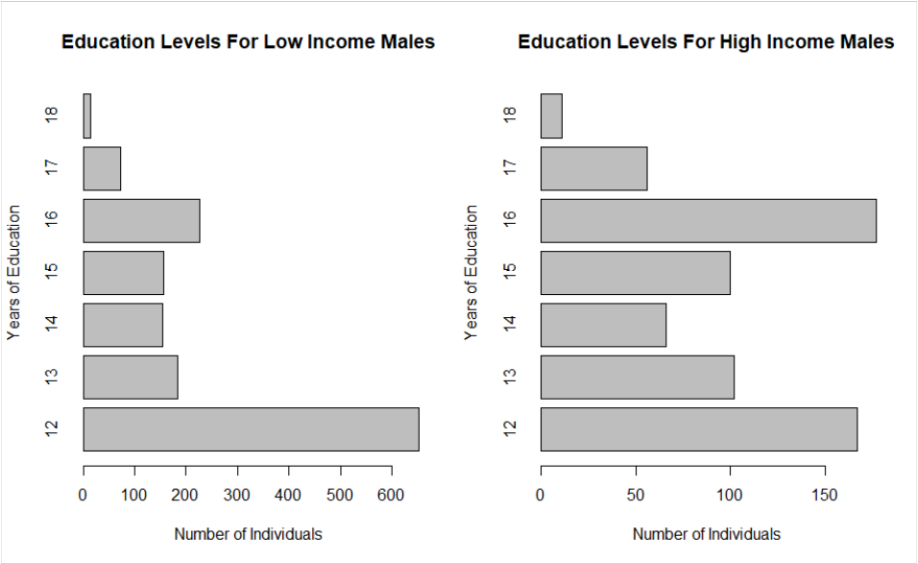

Figure 9: Bar Plot of Number of Individuals vs Years of Education (High and Low Income) MALE

We started by getting a general look on the number of individuals in each education level. Generally speaking, for low income males, most of the occurrences are in the 12 years of education category. For high income males however, we have a significantly smaller sample size. The occurrences are more spread out, however we do see some similarity in numbers between 12 yrs and 18 yrs.

Conducting the Levene test for equality of variances on our sample for low and high income males, we reject the null hypothesis on both since our p-values are significantly less than our α value of 0.05. Thus this confirms the non-homogeneity of the variances in our data.

$$\text{Levene Test (low income Males)} \rightarrow \text{ p-value} = 0.00586 < 0.05, \text{ reject } H_0$$

$$\text{Levene Test (high income Males)} \rightarrow \text{ p-value} = 0.00150 < 0.05, \text{ reject } H_0$$

Looking now at boxplots for low income males, we can see that the tests scores are generally rising as the education level increases. However in the 18th education level category, we see the test scores decrease. This could be down to several factors however the sample size as said before is significantly smaller that the other levels, so this would be ultimately classified as an outlier.

For high income males we see the same general trend. However, between the 12th and 17th education level groups, we see the greatest change in the box plot values as there is no overlap. Considering they are both high income, we can see some reasons for the difference upon further inspection using other fixed factors.

2.3.2 FEMALE Education vs Composite Test Score

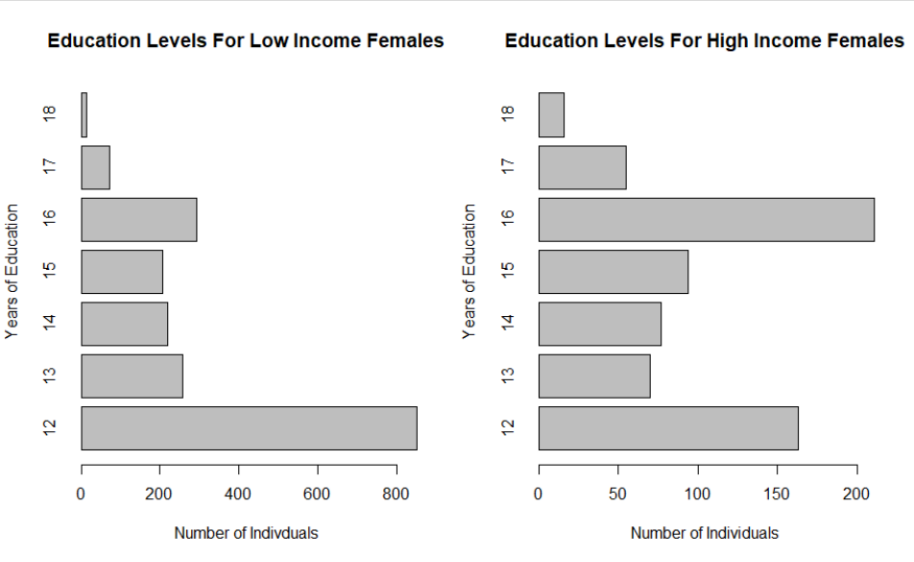

Figure 10: Bar Plot of Number of Individuals vs Years of Education (High and Low Income) FEMALE

Generally speaking, females tend to have the same trend as the males for this comparison. We see that similarity again in the bar plot for 12th and 16th education level for high income females.

The Levene test revealed that we actually fail to reject the null hypothesis for this test of equal variances for low income females as our p value was greater than our α of 0.05. However, for high income females, we reject the null.

$$\text{Levene Test (low income Females)} \rightarrow \text{ p-value} = 0.389 > 0.05, \text{ FTR } H_0$$
$$\text{Levene Test (high income Females)} \rightarrow \text{ p-value} = 0.00867 < 0.05, \text{ reject } H_0$$

For the box plots that were created to represent this data, we see generally the same trend as we did for males, with some boxes overlapping with the medians. Most of the difference between the data occurs however at the 12th and 17th education level.

2.4 Education Level vs Composite Test Score/Ethnicity

2.4.1 MALE Education vs Composite Test Score

When looking at education level vs composite test score for males and fixing income and ethnicity variables we see that there is for the most part a positive linear relationship between them. which means the higher the test score, the higher level of education. However, we do see a drop off at education level 18. Students at education level 18 tend to have a lower test score than students at education level 17.

When we do a hypothesis test to test if the means of the different groups are the same, we get a p-value of less than 0.001 for all the graphs, which tells us that the means for all the groups are significantly different. This is true for all the groups. For high income Hispanics, there seems to be not as many students at the education level 18. However, other ethnic groups, have more individuals at high income than for low income. Of all the different male graphs, this graphs tends to have the most difference in variance from group to group.

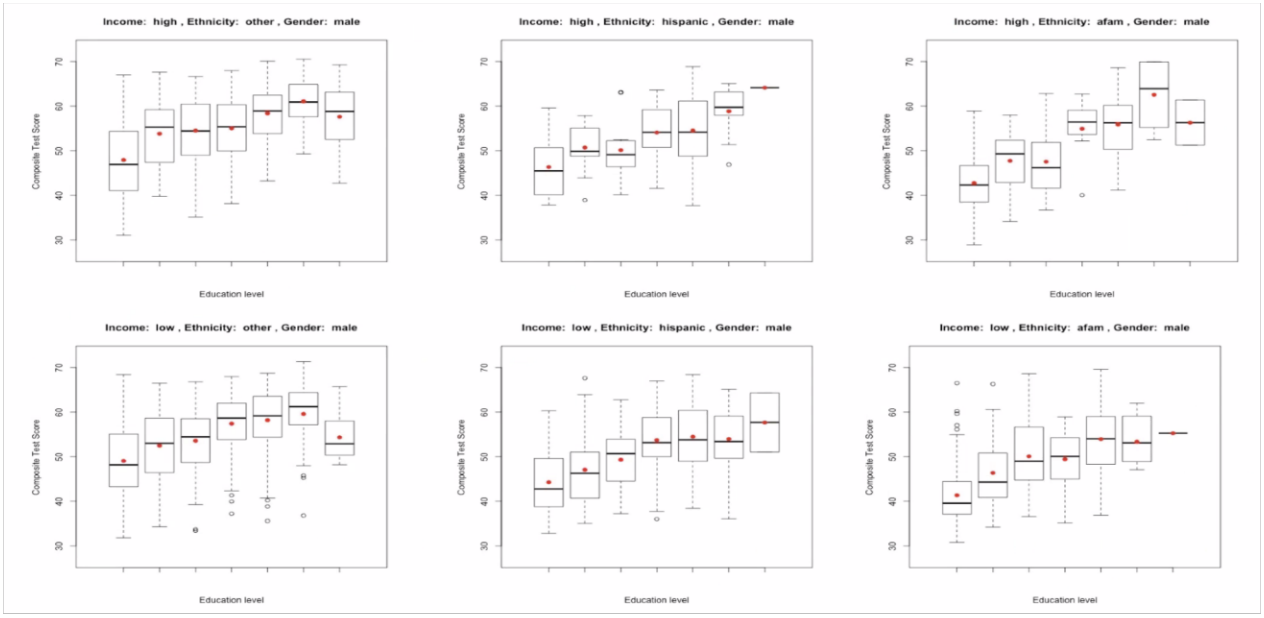

Figure 11: Education Level vs Composite Test Score with Fixed Family Income and Varying Ethnicity MALE

Looking at low income "other" males, there is a clear linear relationship between the box plots, however, this group has the biggest drop off at the 18th education level in comparison to the other groups. This means that students at higher education level tend to do worse on their test score. This group of students also had the most outliers. Even though students scored low on their test scores, they went further into their education. The variance is about the same for all groups in this graph as well.
Lastly, for low income African American males, there is also a significant amount of outliers. However, on the contrary to low income, other ethnicity groups, students did well on their test scores, and they tend to not go far in their education. We also see there was only one student who was documented that went and finished masters.

2.4.2 FEMALE Education vs Composite Test Score

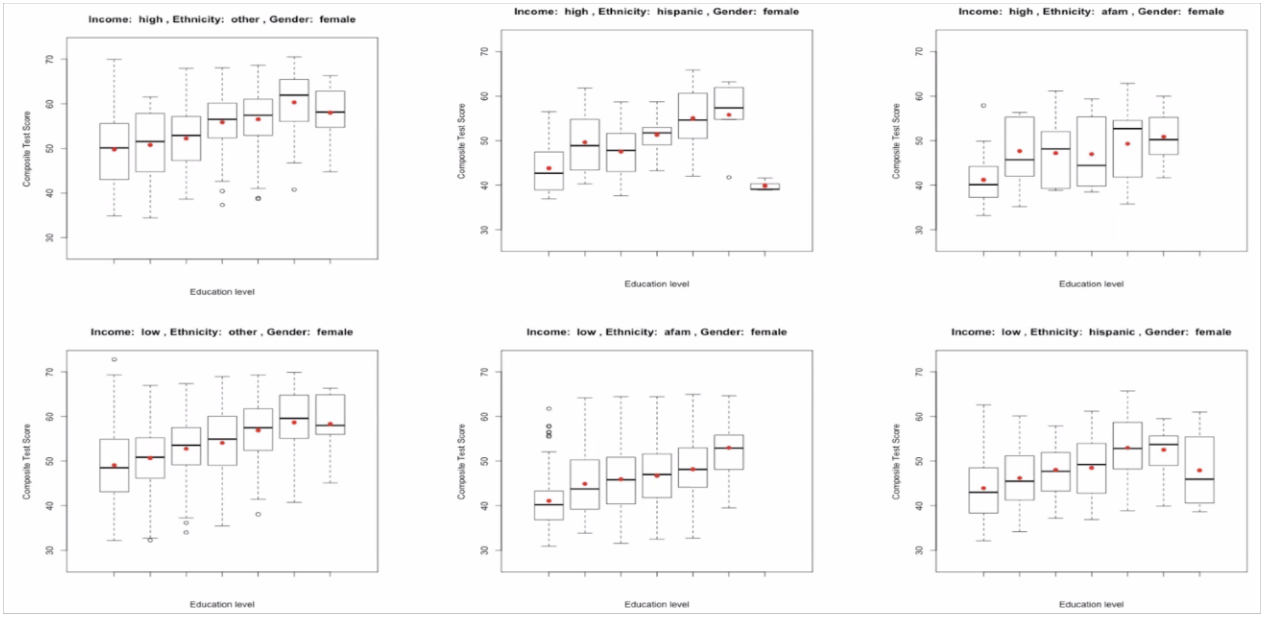

Figure 12: Education Level vs Composite Test Score with Fixed Family Income and Varying Ethnicity FEMALE

When looking at education level vs composite test score for females and fixing income and ethnicity variables we once again see a positive linear relationship between them.

For high income "other" females, there is a noticeable steady linear relationship, with increasing means, and equal variances. When looking at the outliers, we see that students that obtained low values on there test scores tended to go far into their

education level.
For high income Hispanic females we see students have a very low test score at the 18th education level.
For high income African American females, there are no females at the 18th education level. With the increase of means, there is a higher variance in the middle of college years.

Looking at low income Hispanic females, we see a significance amount of outliers at the 17th education level.
Lastly, low income African American females have a substantial amount of outliers at the education level 12. Even though the students did well on their test scores, they did not pursue post secondary.

2.5 Education Level vs Composite Test Score/Parent Education

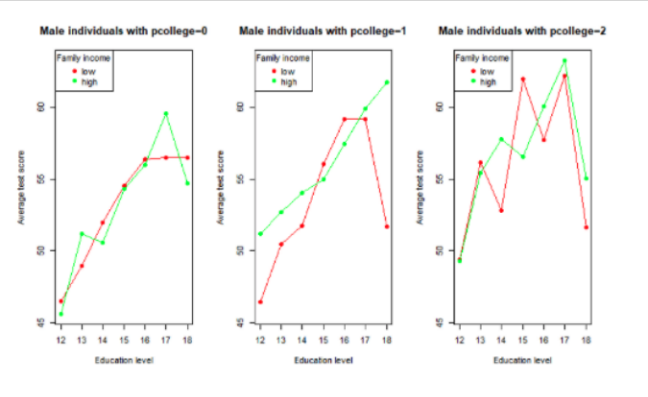

Figure 13: Education Level vs Composite Test Score with Fixed Family Income and Varying PCollege MALE

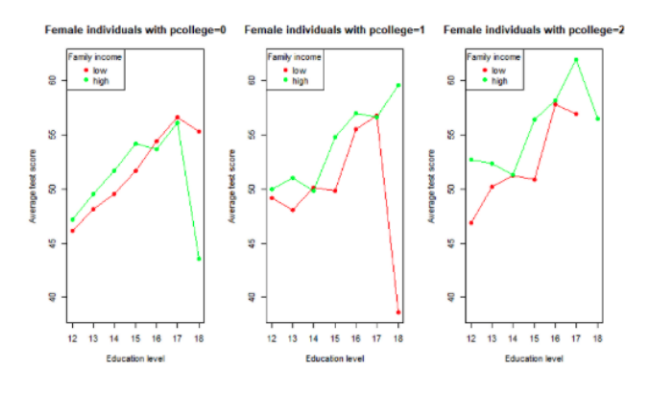

Figure 14: Education Level vs Composite Test Score with Fixed Family Income and Varying PCollege FEMALE

By inspection, we found that males from high-income families generally have higher scores than males from low-income families. From additional testing, we found that the average composite test score is indeed lower among males from low income families compared to males from high-income families. Among the entirety of the male population, it was confirmed by the attained p-value = 1.871e-13 that we reject the null hypothesis of equality of means and treatment effects since it is significantly smaller than the α value of 0.05.

Also by inspection, each additional college-educated parent tends to raise the individual's mean score. Average composite test scores increases with each additional college-educated parent among entire male population and among both of low and high income.

We found that females from high-income families have higher scores than females from low-income families. Each additional college-educated parent tends to raises mean score of the students.

2.6 Education Level vs Composite Test Score/Wage

For high income individuals, the intervals for their wage with respect to their education level remain relatively the same. However, we see that individuals with high income and 18 years of education have the lowest wage. From looking at the data, low income individuals by far have the largest confidence interval for the 18th education level.

For low income individuals, as their wage increases, we see a slight increase in education level but nothing worth analyzing. However for high income individuals, there is no significant fluctuation.

One can also see by a boxplot representation that there is a variation in wage for high income individuals at the 18th education level. This has a larger variation in wage than low income individuals in the 18th education level. This allowed us to come to the conclusion that wage has a higher impact on low income individuals.

When looking at education level vs wage, with a fixed income of high or low, we see that there is no drastic difference between education level and wage for high income individuals. However, for low income individuals, we see that the 18th education level has a much higher confidence interval for an individual's wage than any other education level.

Ed 12	Ed 13	Ed 14	Ed 15	Ed 16	Ed 17	Ed 18
(9.57, 9.65)	(9.61, 9.69)	(9.72, 9.81)	(9.63, 9.70)	(9.67, 9.75)	(9.48, 9.55)	(9.23, 9.32)

Figure 15: Confidence Intervals: Education Level vs Wage (dollar/hr) MALE

Ed 12	Ed 13	Ed 14	Ed 15	Ed 16	Ed 17	Ed 18
(9.40, 9.48)	(9.33, 9.41)	(9.38, 9.46)	(9.43, 9.49)	(9.42, 9.50)	(9.44, 9.52)	(9.75, 9.83)

Figure 16: Confidence Intervals: Education Level vs Wage(dollar/hr) FEMALE

3 MLR Models

Before we start doing analysis using MLR, we want to make sure that the LINE assumptions are met. To do this, we used the following graphs to confirm the assumptions.

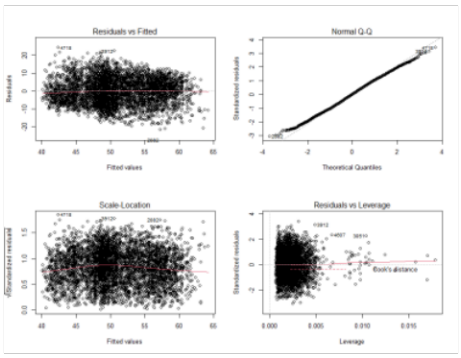

Figure 17: LINE Assumption Diagrams

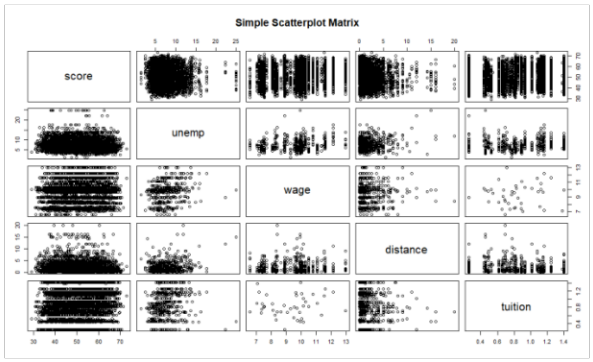

Figure 18: Scatterplot Matrix Diagram

From the diagrams, we can see that the first assumption of linearity is met because of the QQ-plot. Looking at the QQ-plot we find that the fit is adequate therefore the model satisfies linearity. By looking at the residuals vs fitted graph, we can confirm homoscedasticity. This meets the equal variance assumption. By the diagrams (17 and 18), we can assume that all of the LINE assumptions are met.

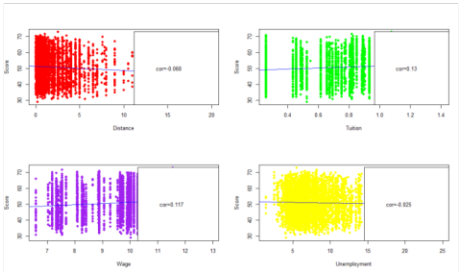

Figure 19: MLR Residuals Model

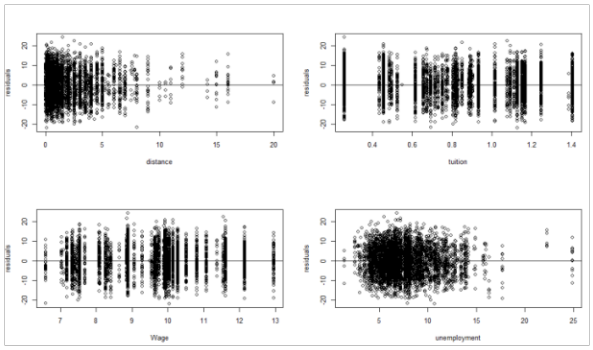

Figure 20: MLR Residuals Model

First, we constructed an MLR model with “education” as the response variable and all other variables as predictors. Some variables, when it was necessary, we converted to factors. We get the following summary after we created the model in R:

```
> summary(fit.education)

Call:
lm(formula = education ~ factor(gender) + factor(ethnicity) +
    score + factor(fcollege) + factor(mcollege) + factor(home) +
    factor(urban) + unemp + wage + distance + tuition + factor(income) +
    factor(region), data = edata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2586 -1.1251 -0.2188  1.1335  5.1428

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.862924   0.217133  45.423  < 2e-16 ***
factor(gender)male    -0.129014   0.044891  -2.874 0.004072 **
factor(ethnicity)hispanic  0.009194   0.078904   0.117 0.907247
factor(ethnicity)other    -0.308261   0.067611  -4.559 5.26e-06 ***
score                  0.089834   0.002847  31.558  < 2e-16 ***
factor(fcollege)yes    0.513740   0.064308   8.611  < 2e-16 ***
factor(mcollege)yes    0.383678   0.072338   5.304 1.18e-07 ***
factor(home)yes        0.145662   0.059265   2.458 0.014015 *
factor(urban)yes       0.040734   0.056521   0.721 0.471138
unemp                  0.030046   0.009023   3.330 0.000876 ***
wage                  -0.037013   0.018299  -2.023 0.043162 *
distance              -0.036449   0.010822  -3.368 0.000763 ***
tuition               -0.209261   0.089600  -2.335 0.019559 *
factor(income)low     -0.381584   0.053728  -7.102 1.41e-12 ***
factor(region)west    -0.188441   0.069813  -2.699 0.006975 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.528 on 4724 degrees of freedom
Multiple R-squared:  0.273,    Adjusted R-squared:  0.2708
F-statistic: 126.7 on 14 and 4724 DF,  p-value: < 2.2e-16
```

Figure 21: MLR Model for Education

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            37.99079    0.85409  44.481  < 2e-16 ***
factor(gender)male      1.09370    0.20748   5.271 1.41e-07 ***
factor(ethnicity)hispanic  2.73681    0.36347   7.530 6.06e-14 ***
factor(ethnicity)other     6.36572    0.30103  21.146  < 2e-16 ***
factor(education)13        2.98083    0.33302   8.951  < 2e-16 ***
factor(education)14        4.31825    0.35559  12.144  < 2e-16 ***
factor(education)15        6.28375    0.34800  18.057  < 2e-16 ***
factor(education)16        8.20055    0.30096  27.248  < 2e-16 ***
factor(education)17       10.17527    0.48423  21.013  < 2e-16 ***
factor(education)18        7.12428    0.95912   7.428 1.30e-13 ***
factor(fcollege)yes        1.29603    0.29951   4.327 1.54e-05 ***
factor(mcollege)yes        1.01925    0.33573   3.036  0.00241 **
factor(home)yes            0.67536    0.27441   2.461  0.01389 *
factor(urban)yes          -0.36422    0.26169  -1.392  0.16406
unemp                     -0.07478    0.04184  -1.787  0.07397 .
wage                       0.19416    0.08472   2.292  0.02197 *
distance                  -0.11372    0.05017  -2.267  0.02344 *
tuition                    2.18730    0.41418   5.281 1.34e-07 ***
factor(income)low          0.18951    0.25023   0.757  0.44888
factor(region)west         0.60056    0.32390   1.854  0.06378 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.072 on 4719 degrees of freedom
Multiple R-squared:  0.3421,    Adjusted R-squared:  0.3395
F-statistic: 129.2 on 19 and 4719 DF,  p-value: < 2.2e-16
```

Figure 22: MLR Model for Composite Test Score

As you can see, our R^2 value for the education model is very small and would not be a good standpoint to analyze. Since our R^2 value is small, not much of our data can be used to explain the variation in y. This can be because our response variable is categorical and to address this issue we will instead use the score variable as our response variable instead. When we now create a linear model with the score as our response variable and education as one of the predictors, we obtain the following model:

With this, more of the variation in our composite test score can be explained by the variation in our predictors since our R^2 value has increased. In order to see an increase the R^2 value, we are going to perform a backwards elimination to further see what variables can be removed that are not needed in our model.

When we performed the backwards elimination test, we found that urban, wage, unemployment and income can be removed as they had the highest p values. After eliminating those elements, we produced a model that was deemed as the most effective model to this point. When we now fit the linear model with the remaining variables, we now produced the following summary:

```
Call:
lm(formula = score ~ gender + ethnicity + fcollege + mcollege +
    home + distance + tuition + education + region)

Residuals:
     Min       1Q   Median       3Q      Max
-21.6773  -5.1026  -0.0029   5.0415  24.1345

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       16.22430    0.93593  17.335  < 2e-16 ***
gendermale         1.11036    0.20777   5.344 9.51e-08 ***
ethnicityhispanic  2.70340    0.36102   7.488 8.29e-14 ***
ethnicityother     6.43673    0.29377  21.910  < 2e-16 ***
fcollegeyes        1.30823    0.29182   4.483 7.53e-06 ***
mcollegeyes        1.06769    0.33520   3.185  0.00146 **
homeyes            0.73947    0.27340   2.705  0.00686 **
distance          -0.12240    0.04578  -2.674  0.00753 **
tuition            2.35172    0.38652   6.084 1.26e-09 ***
education          1.92923    0.06105  31.603  < 2e-16 ***
regionwest         0.73865    0.32129   2.299  0.02155 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.1 on 4728 degrees of freedom
Multiple R-squared:  0.3356,    Adjusted R-squared:  0.3342
F-statistic: 238.9 on 10 and 4728 DF,  p-value: < 2.2e-16
```

Figure 23: MLR Model for Composite Test Score Final

As we can see, the p values for all of the variables is less than 0.05 which is our alpha. This can be confirmed by looking at the AIC values for each of the variables.

To determine which model to use to analyze the data, we will compare the two models using the F test. Our null hypothesis is that the two models are the same, and our alternative hypothesis is that they are significantly different. Once we run the F test, we obtained a p value that is slightly greater (0.05515) than 0.05 (our alpha), and thus fail to reject our null hypothesis and conclude to use the reduced model.

We will now try different transformations to see if we can create a better model. The first transformation that we tried is the log transformation. We found that the R^2 decreased to 33.91%. This showed that it did decrease but not significantly. Next we tried the inverse transformation, which had an R^2 value of 33.2%. Lastly we tried the square root transformation and obtained an R^2 value of 34.3%. Therefore, since there were no drastic changes, we chose to stick with the original reduced MLR model to analyze the data.

4 Conclusion

The most optimal model that we found for our data set was the reduced MLR model which can be represented as the following:

$$score = 39.11 + 1.1(gender = MALE) + 2.76(ethnicity = HISPANIC) + 6.51(ethnicity = OTHER) + 1.28(fcollege = YES) + 1.03(mcollege = YES) + 0.71(home = YES) - 0.12(distance) + 2.32(tuition) + 2.95(education = 13) + 4.30(education = 14) + 6.26(education = 15) + 8.16(education = 16) + 10.14(education = 17) + 7.003(education = 18) + 0.66(region = WEST)$$

We determined that MLR is unfortunately not the strongest model to use due to the fact that our R^2 value is well below 50% regardless of the changes we had applied to our data and analysis. We can also conclude that MLR is not the best approach to use to test this hypothesis. We believe that if we were able to alter our model and choose a different model, we use a multinomial model. We believe that this would be a better model to represent the data that we collected since majority of the variables are categorical in our data set. We can also infer that the reason as to why our R^2 value is so low is because there are variety of external factors that affect an individual's "choice" in education. Unfortunately are unable to account for those external factor (motivation, career goals, mental health...etc). In spite of the trends we uncovered, the significance of these external factors, all unstudied, translates mathematically into variance, which may have negatively impacted our R^2.

The original MLR returns an R^2 value of 0.27, so we determined that it does not help explain the data due to high variability. After that we put composite test score as the response variable instead and it increased the R^2 value to 0.35 approximately, making it the highest value we were able to obtain. Since education level and composite test score are highly correlated, we can make inference of a student's education level through the composite test score.

Prediction intervals, even when obtained by fixing some predictors, are too large to be of any practical use, due to the high variability of the data. This we have tested, for example, with model education score, for males who come from low-income families. Furthermore, we know roughly 44 percent of males from low-income families belong to education level 12, and a low-income male is more likely to belong to the 12th education level rather than to any other education level; and this we have determined is useful information.

By testing, we found the mean composite test score (46.59179) for males from low-income families who belong to the 12th education level. The expected educational attainment for this score is 13.20617 and the PI (predication interval) is (10.170206, 16.24214). The expected value and PI are rigorously-defined notions. However, the PI is unfortunately too wide to be of any use for the education levels and does not fall within out data scope. In conclusion, we determined that MLR is not a good approach to this data set however, multinomial would be a logical replacement.

Appendix

Reference(s):
Rouse, C.E. (1995). Democratization or Diversion? The Effect of Community Colleges on Educational Attainment. Journal of Business & Economic Statistics, 12, 217–224.

Link to Data Set:
https://vincentarelbundock.github.io/Rdatasets/csv/AER/CollegeDistance.csv
https://vincentarelbundock.github.io/Rdatasets/doc/AER/CollegeDistance.html

R code for figures used:
https://drive.google.com/drive/folders/1g2wYV4ak-3TbRgJCsyqPJZmOe6VDl9-D?usp=sharing

Figure 1-4 & 17-23:

```
eData<-read.csv("CollegeDistance.csv")
eData


eData$gender= factor(eData$gender)
eData$ethnicity= factor(eData$ethnicity)
eData$fcollege= factor(eData$fcollege)
eData$mcollege= factor(eData$mcollege)
eData$home= factor(eData$home)
eData$urban= factor(eData$urban)
eData$income= factor(eData$income)
eData$region= factor(eData$region)

str(eData)
head(eData)
View(eData)
class(eData)
typeof(eData)
summary(eData)
colnames(eData)


gender=eData$gender
ethnicity=eData$ethnicity
score=eData$score
home=eData$home
urban=eData$urban
unemp=eData$unemp
wage=eData$wage
distance=eData$distance
tuition=eData$tuition
education=eData$education
income=eData$income
region=eData$region
fcollege = eData$fcollege
mcollege = eData$mcollege


#score
fit.score <- lm(score~gender+ethnicity+fcollege+mcollege+home+urban+unemp+wage+distance+
tuition+factor(education)+income+region, data =eData)
summary(fit.score)
#anova(fit.score)


#unemp
fit.unemp <- lm(unemp~gender+ethnicity+fcollege+mcollege+home+urban+score+wage+distance+
```

```
tuition+education+income+region)
summary(fit.unemp)
anova(fit.unemp)

#wage
fit.wage <- lm(wage~gender+ethnicity+fcollege+mcollege+home+urban+score+unemp+distance+
tuition+education+income+region)
summary(fit.wage)
anova(fit.wage)

#distance
fit.distance <- lm(distance~gender+ethnicity+fcollege+mcollege+home+urban+score+unemp+
wage+tuition+education+income+region)
summary(fit.distance)
anova(fit.distance)



#Best way is to fit score as Y
summary(fit.score)
anova(fit.score)

fit.score$coefficients
fit.score$df.residual
vcov(fit.score)

#plot
par(mfrow = c(2,2))
plot(fit.score)


#diagnostic plot


par(mfrow=c(2,2))
plot(eData$distance, fit.removed_wage$residuals,xlab="distance", ylab="residuals")
abline(h=0)

plot(eData$tuition, fit.removed_wage$residuals,xlab="tuition", ylab="residuals")
abline(h=0)

plot(eData$wage, fit.removed_wage$residuals,xlab="Wage", ylab="residuals")
abline(h=0)


plot(eData$unemp, fit.removed_wage$residuals,xlab="unemployment", ylab="residuals")
abline(h=0)

plot(fit.removed_wage$fitted.values, fit.removed_wage$residuals,xlab="fitted",
ylab="residuals")
abline(h=0)

#Holds LINE

#A scatter plot is not a good choice for categorical variables,
#so it wouldn't really make sense to "add" those variables to
#this scatter matrix.

New <- data.frame(unemp,wage,distance,tuition ,resp=score)
pairs(~score+unemp+wage+distance+tuition,data=New,main="Simple Scatterplot Matrix")
```

```
#variable selection
fit.score
summary(fit.score)
step(fit.score, data=eData, direction="backward", k=2)

fit.removed_income =lm(score~gender+ethnicity+fcollege+mcollege+home+urban+unemp+wage+
distance+tuition+factor(education)+region,)
summary(fit.removed_income)
step(fit.removed_income, data=eData, direction="backward", k=2)



#remove urban and unemp since p-value high
fit.removed_urb_unemp =lm(score~gender+ethnicity+fcollege+mcollege+home+wage+distance+
tuition+factor(education)+region,)
summary(fit.removed_urb_unemp)
step(fit.removed_urb_unemp, data=eData, direction="backward", k=2)


#remove wage
fit.removed_wage =lm(score~gender+ethnicity+fcollege+mcollege+home+distance+tuition+
factor(education)+region)
summary(fit.removed_wage)

step(fit.removed_wage, data=eData, direction="backward", k=2)



#Trying some removal

anova(fit.score, fit.removed_wage)


#try transformation


#log
log_fit =lm(log(score)~gender+ethnicity+fcollege+mcollege+home+distance+tuition+
factor(education)+region)
summary(log_fit)
plot(log_fit)

#inverse
inverse_fit =lm(1/score~gender+ethnicity+fcollege+mcollege+home+distance+tuition+
factor(education)+region)
summary(inverse_fit)


#sqrt

sqrt_fit =lm(sqrt(score)~gender+ethnicity+fcollege+mcollege+home+distance+tuition+
factor(education)+region)
summary(sqrt_fit)
plot(sqrt_fit)
step(log_fit, data=eData, direction="backward", k=2)
```

```
#poly
poly_fit =lm(score~gender+ethnicity+fcollege+mcollege+home+distance+distance^2+tuition+
factor(education)+region)
summary(poly_fit)
plot(poly_fit)


#multicollinearity
#New is the correlation matrix of new variables
cor(New)
library(car)
vif(fit.removed_wage)

#latest model to describe the fit would be

# correlation coefficient
cor(score, distance)  #-0.06797927  #distance goes up score goes down
cor(score, tuition)   #0.1298585   #tuition goes up score goes up


par(mfrow=c(2,2))

plot(score~distance, data=eData, pch=19, col="red", xlab="Distance", ylab="Score")
abline(lm(score~distance,data=eData), col="blue")
legend("bottomright", paste("cor=", round(cor(score,distance),3), sep="") )


plot(score~tuition, data=eData, pch=19, col="Green", xlab="Tuition", ylab="Score")
abline(lm(score~tuition,data=eData), col="blue")
legend("bottomright", paste("cor=", round(cor(score,tuition),3), sep="") )


plot(score~wage, data=eData, pch=19, col="purple", xlab="Wage", ylab="Score")
abline(lm(score~wage,data=eData), col="blue")
legend("bottomright", paste("cor=", round(cor(score,wage),3), sep="") )



plot(score~unemp, data=eData, pch=19, col="Yellow", xlab="Unemployment", ylab="Score")
abline(lm(score~unemp,data=eData), col="blue")
legend("bottomright", paste("cor=", round(cor(score,unemp),3), sep="") )



Best.fit = fit.removed_wage
Best.fit
summary(Best.fit)
confint(Best.fit, prediction = "interval", level =0.95)



#Not the best coz R squared is small
#solutions

#Try different Transformations such as Multinomial (out of scope)
#Change Data set or find better predictors affecting score
```

Figure 5 & Figure 6:

```
ed12_m=eData$score[eData$education==12 & eData$gender=="male"]
ed13_m=eData$score[eData$education==13 & eData$gender=="male"]
```

```
ed14_m=eData$score[eData$education==14 & eData$gender=="male"]
ed15_m=eData$score[eData$education==15 & eData$gender=="male"]
ed16_m=eData$score[eData$education==16 & eData$gender=="male"]
ed17_m=eData$score[eData$education==17 & eData$gender=="male"]
ed18_m=eData$score[eData$education==18 & eData$gender=="male"]

mean_ed12_m = mean(ed12_m)
mean_ed13_m = mean(ed13_m)
mean_ed14_m = mean(ed14_m)
mean_ed15_m = mean(ed15_m)
mean_ed16_m = mean(ed16_m)
mean_ed17_m = mean(ed17_m)
mean_ed18_m = mean(ed18_m)
mean_ed_m=c(mean_ed12_m, mean_ed13_m, mean_ed14_m, mean_ed15_m, mean_ed16_m,
mean_ed17_m, mean_ed18_m)

var_ed12_m = var(ed12_m)
var_ed13_m = var(ed13_m)
var_ed14_m = var(ed14_m)
var_ed15_m = var(ed15_m)
var_ed16_m = var(ed16_m)
var_ed17_m = var(ed17_m)
var_ed18_m = var(ed18_m)
var_ed_m=c(var_ed12_m, var_ed13_m, var_ed14_m, var_ed15_m, var_ed16_m,
var_ed17_m, var_ed18_m)

sd_ed_m=sqrt(var_ed_m)

ed12_f = (eData$score[eData$education==12 & eData$gender=="female"])
ed13_f = (eData$score[eData$education==13 & eData$gender=="female"])
ed14_f = (eData$score[eData$education==14 & eData$gender=="female"])
ed15_f = (eData$score[eData$education==15 & eData$gender=="female"])
ed16_f = (eData$score[eData$education==16 & eData$gender=="female"])
ed17_f = (eData$score[eData$education==17 & eData$gender=="female"])
ed18_f = (eData$score[eData$education==18 & eData$gender=="female"])

mean_ed12_f = mean(ed12_f)
mean_ed13_f = mean(ed13_f)
mean_ed14_f = mean(ed14_f)
mean_ed15_f = mean(ed15_f)
mean_ed16_f = mean(ed16_f)
mean_ed17_f = mean(ed17_f)
mean_ed18_f = mean(ed18_f)
mean_ed_f=c(mean_ed12_f, mean_ed13_f, mean_ed14_f, mean_ed15_f, mean_ed16_f,
mean_ed17_f, mean_ed18_f)

var_ed12_f = var(ed12_f)
var_ed13_f = var(ed13_f)
var_ed14_f = var(ed14_f)
var_ed15_f = var(ed15_f)
var_ed16_f = var(ed16_f)
var_ed17_f = var(ed17_f)
var_ed18_f = var(ed18_f)
var_ed_f=c(var_ed12_f, var_ed13_f, var_ed14_f, var_ed15_f, var_ed16_f,
var_ed17_f, var_ed18_f)

sd_ed_f=sqrt(var_ed_f)

q1=qt(0.975,4738)*sd_ed_m/sqrt(4739)
mean_ed_m-q1
```

```
mean_ed_m+q1

q2=qt(0.975,4738)*sd_ed_f/sqrt(4739)
mean_ed_f-q2
mean_ed_f+q2
```

Figure 7 & 8:

```
#### Education level vs unemployment
#### Comparing "high" and "low" family income

eData<-read.csv("CollegeDistance.csv")

head(eData)
str(eData)
View(eData)
class(eData)
typeof(eData)
summary(eData)
colnames(eData)
eData

##### "High" family income
#### MALES ONLY

ed12_high_male=eData$unemp[eData$education==12 & eData$income=="high" & eData$gender=="male"]
ed13_high_male=eData$unemp[eData$education==13 & eData$income=="high" & eData$gender=="male"]
ed14_high_male=eData$unemp[eData$education==14 & eData$income=="high" & eData$gender=="male"]
ed15_high_male=eData$unemp[eData$education==15 & eData$income=="high" & eData$gender=="male"]
ed16_high_male=eData$unemp[eData$education==16 & eData$income=="high" & eData$gender=="male"]
ed17_high_male=eData$unemp[eData$education==17 & eData$income=="high" & eData$gender=="male"]
ed18_high_male=eData$unemp[eData$education==18 & eData$income=="high" & eData$gender=="male"]

#means: "High" family income- MALES only
mean_ed12_high_male=mean(ed12_high_male)
mean_ed13_high_male=mean(ed13_high_male)
mean_ed14_high_male=mean(ed14_high_male)
mean_ed15_high_male=mean(ed15_high_male)
mean_ed16_high_male=mean(ed16_high_male)
mean_ed17_high_male=mean(ed17_high_male)
mean_ed18_high_male=mean(ed18_high_male)
mean_ed_high_male=c(mean_ed12_high_male,mean_ed13_high_male,mean_ed14_high_male,mean_ed15_high_m
mean_ed16_high_male,mean_ed17_high_male,mean_ed18_high_male)
mean_ed_high_male


##### "High" family income
#### FEMALES ONLY

ed12_high_female=eData$unemp[eData$education==12 & eData$income=="high" &
eData$gender=="female"]
ed13_high_female=eData$unemp[eData$education==13 & eData$income=="high" &
eData$gender=="female"]
ed14_high_female=eData$unemp[eData$education==14 & eData$income=="high" &
eData$gender=="female"]
ed15_high_female=eData$unemp[eData$education==15 & eData$income=="high" &
eData$gender=="female"]
ed16_high_female=eData$unemp[eData$education==16 & eData$income=="high" &
eData$gender=="female"]
ed17_high_female=eData$unemp[eData$education==17 & eData$income=="high" &
eData$gender=="female"]
```

```
ed18_high_female=eData$unemp[eData$education==18 & eData$income=="high" &
eData$gender=="female"]

#Means: "High" family income - FEMALEs only
mean_ed12_high_female=mean(ed12_high_female)
mean_ed13_high_female=mean(ed13_high_female)
mean_ed14_high_female=mean(ed14_high_female)
mean_ed15_high_female=mean(ed15_high_female)
mean_ed16_high_female=mean(ed16_high_female)
mean_ed17_high_female=mean(ed17_high_female)
mean_ed18_high_female=mean(ed18_high_female)
mean_ed_high_female=c(mean_ed12_high_female,mean_ed13_high_female,mean_ed14_high_female,
mean_ed15_high_female, mean_ed16_high_female,mean_ed17_high_female,mean_ed18_high_female)
mean_ed_high_female


##### "low" family income
#### MALES ONLY

ed12_low_male=eData$unemp[eData$education==12 & eData$income=="low" & eData$gender=="male"]
ed13_low_male=eData$unemp[eData$education==13 & eData$income=="low" & eData$gender=="male"]
ed14_low_male=eData$unemp[eData$education==14 & eData$income=="low" & eData$gender=="male"]
ed15_low_male=eData$unemp[eData$education==15 & eData$income=="low" & eData$gender=="male"]
ed16_low_male=eData$unemp[eData$education==16 & eData$income=="low" & eData$gender=="male"]
ed17_low_male=eData$unemp[eData$education==17 & eData$income=="low" & eData$gender=="male"]
ed18_low_male=eData$unemp[eData$education==18 & eData$income=="low" & eData$gender=="male"]

#Means: "Low" fmaily income- MALES only
mean_ed12_low_male=mean(ed12_low_male)
mean_ed13_low_male=mean(ed13_low_male)
mean_ed14_low_male=mean(ed14_low_male)
mean_ed15_low_male=mean(ed15_low_male)
mean_ed16_low_male=mean(ed16_low_male)
mean_ed17_low_male=mean(ed17_low_male)
mean_ed18_low_male=mean(ed18_low_male)
mean_ed_low_male=c(mean_ed12_low_male,mean_ed13_low_male,mean_ed14_low_male,mean_ed15_low_male,
mean_ed16_low_male, mean_ed17_low_male,mean_ed18_low_male)
mean_ed_low_male


##### "low" family income
#### FEMALES ONLY

ed12_low_female=eData$unemp[eData$education==12 & eData$income=="low" & eData$gender=="female"]
ed13_low_female=eData$unemp[eData$education==13 & eData$income=="low" & eData$gender=="female"]
ed14_low_female=eData$unemp[eData$education==14 & eData$income=="low" & eData$gender=="female"]
ed15_low_female=eData$unemp[eData$education==15 & eData$income=="low" & eData$gender=="female"]
ed16_low_female=eData$unemp[eData$education==16 & eData$income=="low" & eData$gender=="female"]
ed17_low_female=eData$unemp[eData$education==17 & eData$income=="low" & eData$gender=="female"]
ed18_low_female=eData$unemp[eData$education==18 & eData$income=="low" & eData$gender=="female"]

#Means: "Low" family income- FEMALES only
mean_ed12_low_female=mean(ed12_low_female)
mean_ed13_low_female=mean(ed13_low_female)
mean_ed14_low_female=mean(ed14_low_female)
mean_ed15_low_female=mean(ed15_low_female)
mean_ed16_low_female=mean(ed16_low_female)
mean_ed17_low_female=mean(ed17_low_female)
mean_ed18_low_female=mean(ed18_low_female)
mean_ed_low_female=c(mean_ed12_low_female,mean_ed13_low_female,mean_ed14_low_female,
```

```
mean_ed15_low_female, mean_ed16_low_female, mean_ed17_low_female, mean_ed18_low_female)
mean_ed_low_female

#### Boxplot: Family income "high" males only
par(mfrow=c(2,2))
boxplot(ed12_high_male, ed13_high_male, ed14_high_male, ed15_high_male, ed16_high_male,
ed17_high_male, ed18_high_male,
        xlab="Education level",
        ylab="Unemployment",
        main="High Income Males",
        names=c(12:18),
        vertical = TRUE)
points(x=c(mean_ed_high_male),pch=19, col="red")


#### Boxplot: Family income "high" females only
boxplot(ed12_high_female, ed13_high_female, ed14_high_female, ed15_high_female,
ed16_high_female, ed17_high_female, ed18_high_female,
        xlab="Education level",
        ylab="unemployment",
        main="High Income Females",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_high_female),pch=19, col="red")
### Boxplot: Family income "low" males only
boxplot(ed12_low_male, ed13_low_male, ed14_low_male, ed15_low_male, ed16_low_male,
ed17_low_male, ed18_low_male,
        xlab="Education level",
        ylab="unemployment",
        main="Low Income Males",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_low_male),pch=19, col="red")
### Boxplot: Family income "low" females only
boxplot(ed12_low_female, ed13_low_female, ed14_low_female, ed15_low_female,
ed16_low_female, ed17_low_female, ed18_low_female,
        xlab="Education level",
        ylab="Unemployment",
        main="Low Income Females",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_low_female),pch=19, col="red")

##


#######

#### 6 linear models: 3 per gender

### MALES ONLY

## educ ~ unemployment: fit1
summary(eData$unemp[eData$gender=="male"])
fit1 <- lm(education[eData$gender=="male"]~unemp[eData$gender=="male"],data=eData)
summary(fit1)
anova(fit1)

## educ ~ unemployment, fixing income==low: fit2
fit2 <- lm(education[eData$gender=="male" & eData$income=="low"]~unemp
```

```
[eData$gender=="male" & eData$income=="low"],data=eData)
summary(fit2)
anova(fit2)

## educ ~ unemployment, fixing income==high: fit3
fit3 <- lm(education[eData$gender=="male" & eData$income=="high"]~unemp
[eData$gender=="male" & eData$income=="high"],data=eData)
summary(fit3)
anova(fit3)

### FEMALES ONLY

## educ ~ unemployment: fit4
summary(eData$unemp[eData$gender=="female"])
fit4 <- lm(education[eData$gender=="female"]~unemp[eData$gender=="female"],data=eData)
summary(fit4)
anova(fit4)

## educ ~ unemployment, fixing income==low: fit5
fit5 <- lm(education[eData$gender=="female" & eData$income=="low"]~unemp
[eData$gender=="female" & eData$income=="low"],data=eData)
summary(fit5)
anova(fit5)

## educ ~ unemployment, fixing income==high: fit6
fit6 <- lm(education[eData$gender=="female" & eData$income=="high"]~unemp
[eData$gender=="female" & eData$income=="high"],data=eData)
summary(fit6)
anova(fit6)
```

Figure 9 :

```
#### Base year composite test score vs education level
#### family income low & high
#### MALES ONLY

## Libraries and Data Sets
library(car)
library(ggplot2)
eData<-read.csv("CollegeDistance.csv")

eData
head(eData)
str(eData)
View(eData)
class(eData)
typeof(eData)
summary(eData)
colnames(eData)

### Section 1: Prepare Data

## low income/male
y_low_male=eData$score[eData$income=="low" & eData$gender=="male"]
x_low_male=eData$education[eData$income=="low" & eData$gender=="male"]

fit_low_male=lm(y_low_male~factor(x_low_male))
fit_low_male

summary(fit_low_male)
```

```
## Add Bartlett and Levene test
bartlett_test_low_male = bartlett.test(y_low_male~factor(x_low_male), data=eData)
levene_test_low_male = leveneTest(y_low_male~factor(x_low_male), data=eData)
bartlett_test_low_male
levene_test_low_male

## high income/male
y_high_male=eData$score[eData$income=="high" & eData$gender=="male"]
x_high_male=eData$education[eData$income=="high" & eData$gender=="male"]

fit_high_male=lm(y_high_male~factor(x_high_male))
fit_high_male

summary(fit_high_male)

## Add Bartlett and Levene test
bartlett_test_high_male = bartlett.test(y_high_male~factor(x_high_male), data = eData)
levene_test_high_male = leveneTest(y_high_male~factor(x_high_male), data = eData)
bartlett_test_high_male
levene_test_high_male

## Section 2: Bar Plot
# Organizing the plots
par(mfrow=c(2,2))

## Bar Plot: education levels for low income males
barplot_low_male = table(x_low_male)
barplot(barplot_low_male,
        horiz = TRUE,
        main = "Education Levels For Low Income Males",
        ylab = "Years of Education",
        xlab = "Number of Individuals")

## Bar Plot: education levels for high income males
barplot_high_male = table(x_high_male)
barplot(barplot_high_male,
        horiz = TRUE,
        main = "Education Levels For High Income Males",
        ylab = "Years of Education",
        xlab = "Number of Individuals")

## Section 3: Comp Scores vs Education Level Boxplots

# Scores: Male/Low income
score12_low_male = eData$score[eData$education == 12 & eData$income == "low"
                    & eData$gender == "male"]
score13_low_male = eData$score[eData$education == 13 & eData$income == "low"
                    & eData$gender == "male"]
score14_low_male = eData$score[eData$education == 14 & eData$income == "low"
                    & eData$gender == "male"]
score15_low_male = eData$score[eData$education == 15 & eData$income == "low"
                    & eData$gender == "male"]
score16_low_male = eData$score[eData$education == 16 & eData$income == "low"
                    & eData$gender == "male"]
score17_low_male = eData$score[eData$education == 17 & eData$income == "low"
                    & eData$gender == "male"]
score18_low_male = eData$score[eData$education == 18 & eData$income == "low"
                    & eData$gender == "male"]
```

```
# Means: Male/Low income
mean_score12_low_male = mean(score12_low_male)
mean_score13_low_male = mean(score13_low_male)
mean_score14_low_male = mean(score14_low_male)
mean_score15_low_male = mean(score15_low_male)
mean_score16_low_male = mean(score16_low_male)
mean_score17_low_male = mean(score17_low_male)
mean_score18_low_male = mean(score18_low_male)
mean_score_all_low_male = c(mean_score12_low_male,
                            mean_score13_low_male,
                            mean_score14_low_male,
                            mean_score15_low_male,
                            mean_score16_low_male,
                            mean_score17_low_male,
                            mean_score18_low_male)

# Scores: Male/High Income
score12_high_male = eData$score[eData$education == 12 & eData$income == "high"
                    & eData$gender == "male"]
score13_high_male = eData$score[eData$education == 13 & eData$income == "high"
                    & eData$gender == "male"]
score14_high_male = eData$score[eData$education == 14 & eData$income == "high"
                    & eData$gender == "male"]
score15_high_male = eData$score[eData$education == 15 & eData$income == "high"
                    & eData$gender == "male"]
score16_high_male = eData$score[eData$education == 16 & eData$income == "high"
                    & eData$gender == "male"]
score17_high_male = eData$score[eData$education == 17 & eData$income == "high"
                    & eData$gender == "male"]
score18_high_male = eData$score[eData$education == 18 & eData$income == "high"
                    & eData$gender == "male"]

# Means: Male/High Income
mean_score12_high_male = mean(score12_high_male)
mean_score13_high_male = mean(score13_high_male)
mean_score14_high_male = mean(score14_high_male)
mean_score15_high_male = mean(score15_high_male)
mean_score16_high_male = mean(score16_high_male)
mean_score17_high_male = mean(score17_high_male)
mean_score18_high_male = mean(score18_high_male)
mean_score_all_high_male = c(mean_score12_high_male,
                            mean_score13_high_male,
                            mean_score14_high_male,
                            mean_score15_high_male,
                            mean_score16_high_male,
                            mean_score17_high_male,
                            mean_score18_high_male)


#Organizing the plots
#par(mfrow = c(1,2))
# Boxplot: Male/Low Income
boxplot(
        score12_low_male, score13_low_male, score14_low_male, score15_low_male,
        score16_low_male, score17_low_male, score18_low_male,
        xlab = "Education Level",
        ylab = "Test Score",
        main = "Low Income Males",
        names = c(12:18),
        vertical = TRUE
```

```
)
# Marking all the mean points
points(x = c(mean_score_all_low_male), pch = 20, col = "red")

# Boxplot: Male/High Income
boxplot(
        score12_high_male, score13_high_male, score14_high_male, score15_high_male,
        score16_high_male, score17_high_male, score18_high_male,
        xlab = "Education Level",
        ylab = "Test Score",
        main = "High Income Males",
        names = c(12:18),
        vertical = TRUE
)
# Marking all the mean points
points(x = c(mean_score_all_high_male), pch = 20, col = "red")
```

Figure 10:

```
#### Base year composite test score vs education level
#### family income low & high
#### FEMALES ONLY

## Libraries and Data Sets
#library(yarrr)
library(car)
library(ggplot2)
eData<-read.csv("CollegeDistance.csv")

eData
head(eData)
str(eData)
View(eData)
class(eData)
typeof(eData)
summary(eData)
colnames(eData)

### Section 1: Prepare Data

## low income/female
y_low_female=eData$score[eData$income=="low" & eData$gender=="female"]
x_low_female=eData$education[eData$income=="low" & eData$gender=="female"]

fit_low_female=lm(y_low_female~factor(x_low_female))
fit_low_female

summary(fit_low_female)

## Add Bartlett and Levene test
bartlett_test_low_female = bartlett.test(y_low_female~factor(x_low_female), data=eData)
levene_test_low_female = leveneTest(y_low_female~factor(x_low_female), data=eData)
bartlett_test_low_female
levene_test_low_female

## high income/female
y_high_female=eData$score[eData$income=="high" & eData$gender=="female"]
x_high_female=eData$education[eData$income=="high" & eData$gender=="female"]

fit_high_female=lm(y_high_female~factor(x_high_female))
```

```
fit_high_female

summary(fit_high_female)

## Add Bartlett and Levene test
bartlett_test_high_female = bartlett.test(y_high_female~factor(x_high_female), data = eData)
levene_test_high_female = leveneTest(y_high_female~factor(x_high_female), data = eData)
bartlett_test_high_female
levene_test_high_female

## Section 2: Bar Plot
# Organizing the Plots
par(mfrow=c(2,2))

## Bar Plot: education levels for low income females
barplot_low_female = table(x_low_female)
barplot(barplot_low_female,
        horiz = TRUE,
        main = "Education Levels For Low Income Females",
        ylab = "Years of Education",
        xlab = "Number of Indivduals")

## Bar Plot: education levels for high income females
barplot_high_female = table(x_high_female)
barplot(barplot_high_female,
        horiz = TRUE,
        main = "Education Levels For High Income Females",
        ylab = "Years of Education",
        xlab = "Number of Individuals")

## Section 3: Comp Scores vs Education Level Boxplots

# Scores: Female/Low Income
score12_low_female = eData$score[eData$education == 12 & eData$income == "low"
                     & eData$gender == "female"]
score13_low_female = eData$score[eData$education == 13 & eData$income == "low"
                     & eData$gender == "female"]
score14_low_female = eData$score[eData$education == 14 & eData$income == "low"
                     & eData$gender == "female"]
score15_low_female = eData$score[eData$education == 15 & eData$income == "low"
                     & eData$gender == "female"]
score16_low_female = eData$score[eData$education == 16 & eData$income == "low"
                     & eData$gender == "female"]
score17_low_female = eData$score[eData$education == 17 & eData$income == "low"
                     & eData$gender == "female"]
score18_low_female = eData$score[eData$education == 18 & eData$income == "low"
                     & eData$gender == "female"]

# Means: Female/Low income
mean_score12_low_female = mean(score12_low_female)
mean_score13_low_female = mean(score13_low_female)
mean_score14_low_female = mean(score14_low_female)
mean_score15_low_female = mean(score15_low_female)
mean_score16_low_female = mean(score16_low_female)
mean_score17_low_female = mean(score17_low_female)
mean_score18_low_female = mean(score18_low_female)
mean_score_all_low_female = c(mean_score12_low_female,
                              mean_score13_low_female,
                              mean_score14_low_female,
                              mean_score15_low_female,
```

```
                            mean_score16_low_female,
                            mean_score17_low_female,
                            mean_score18_low_female)

# Scores: Female/High Income
score12_high_female = eData$score[eData$education == 12 & eData$income == "high"
                    & eData$gender == "female"]
score13_high_female = eData$score[eData$education == 13 & eData$income == "high"
                    & eData$gender == "female"]
score14_high_female = eData$score[eData$education == 14 & eData$income == "high"
                    & eData$gender == "female"]
score15_high_female = eData$score[eData$education == 15 & eData$income == "high"
                    & eData$gender == "female"]
score16_high_female = eData$score[eData$education == 16 & eData$income == "high"
                    & eData$gender == "female"]
score17_high_female = eData$score[eData$education == 17 & eData$income == "high"
                    & eData$gender == "female"]
score18_high_female = eData$score[eData$education == 18 & eData$income == "high"
                    & eData$gender == "female"]

# Means: Female/High Income
mean_score12_high_female = mean(score12_high_female)
mean_score13_high_female = mean(score13_high_female)
mean_score14_high_female = mean(score14_high_female)
mean_score15_high_female = mean(score15_high_female)
mean_score16_high_female = mean(score16_high_female)
mean_score17_high_female = mean(score17_high_female)
mean_score18_high_female = mean(score18_high_female)
mean_score_all_high_female = c(mean_score12_high_female,
                               mean_score13_high_female,
                               mean_score14_high_female,
                               mean_score15_high_female,
                               mean_score16_high_female,
                               mean_score17_high_female,
                               mean_score18_high_female)


#Organizing the plots
#par(mfrow = c(1,2))
# Boxplot: female/Low Income
boxplot(
        score12_low_female, score13_low_female, score14_low_female, score15_low_female,
        score16_low_female, score17_low_female, score18_low_female,
        xlab = "Education Level",
        ylab = "Test Score",
        main = "Low Income Females",
        names = c(12:18),
        vertical = TRUE
)
# Marking all the mean points
points(x = c(mean_score_all_low_female), pch = 20, col = "red")

# Boxplot: Female/High Income
boxplot(
        score12_high_female, score13_high_female, score14_high_female, score15_high_female,
        score16_high_female, score17_high_female, score18_high_female,
        xlab = "Education Level",
        ylab = "Test Score",
        main = "High Income Females",
        names = c(12:18),
```

```
        vertical = TRUE
)
# Marking all the mean points
points(x = c(mean_score_all_high_female), pch = 20, col = "red")
```

Figure 11 & Figure 12:

```
eData<-read.csv("CollegeDistance.csv")
for (income in c("low", "high")) {
  for (ethnicity in c("afam", "hispanic", "other")){
    for (gender in c("female", "male")){
      eData_score=eData$score[eData$income==income & eData$ethnicity==ethnicity &
      eData$gender==gender]
      eData_education=eData$education[eData$income==income & eData$ethnicity==ethnicity &
      eData$gender==gender]

      score_education12=eData$score[eData$education==12 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education13=eData$score[eData$education==13 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education14=eData$score[eData$education==14 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education15=eData$score[eData$education==15 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education16=eData$score[eData$education==16 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education17=eData$score[eData$education==17 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      score_education18=eData$score[eData$education==18 & eData$income==income &
      eData$ethnicity==ethnicity & eData$gender==gender]
      mean_score = c(mean(score_education12), mean(score_education13),
      mean(score_education14), mean(score_education15), mean(score_education16),
      mean(score_education17), mean(score_education18))

      graph_name = paste("Income: ", income, ", Ethnicity: ", ethnicity, ", Gender: ",
      gender, sep = " ")
      pdf(paste(graph_name, ".pdf"))
      boxplot (score_education12,
               score_education13,
               score_education14,
               score_education15,
               score_education16,
               score_education17,
               score_education18,
               xlim = c(0, 8),
               ylim = c(27, 73),
               xlab="Education level",
               ylab="Composite Test Score",
               main=paste(graph_name)
      )
      points(mean_score,pch=19, col="red")
      dev.off()
    }
  }
}
```

Figure 13:

```
eData<-read.csv("CollegeDistance.csv")

head(eData)
```

```
str(eData)
View(eData)
class(eData)
typeof(eData)
summary(eData)
colnames(eData)
eData

ed12_high_male=eData$unemp[eData$education==12 & eData$income=="high" & eData$gender=="male"]
ed13_high_male=eData$unemp[eData$education==13 & eData$income=="high" & eData$gender=="male"]
ed14_high_male=eData$unemp[eData$education==14 & eData$income=="high" & eData$gender=="male"]
ed15_high_male=eData$unemp[eData$education==15 & eData$income=="high" & eData$gender=="male"]
ed16_high_male=eData$unemp[eData$education==16 & eData$income=="high" & eData$gender=="male"]
ed17_high_male=eData$unemp[eData$education==17 & eData$income=="high" & eData$gender=="male"]
ed18_high_male=eData$unemp[eData$education==18 & eData$income=="high" & eData$gender=="male"]


ed12_low_male=eData$unemp[eData$education==12 & eData$income=="low" & eData$gender=="male"]
ed13_low_male=eData$unemp[eData$education==13 & eData$income=="low" & eData$gender=="male"]
ed14_low_male=eData$unemp[eData$education==14 & eData$income=="low" & eData$gender=="male"]
ed15_low_male=eData$unemp[eData$education==15 & eData$income=="low" & eData$gender=="male"]
ed16_low_male=eData$unemp[eData$education==16 & eData$income=="low" & eData$gender=="male"]
ed17_low_male=eData$unemp[eData$education==17 & eData$income=="low" & eData$gender=="male"]
ed18_low_male=eData$unemp[eData$education==18 & eData$income=="low" & eData$gender=="male"]


#means: "High" family income- MALES only
mean_ed12_high_male=mean(ed12_high_male)
mean_ed13_high_male=mean(ed13_high_male)
mean_ed14_high_male=mean(ed14_high_male)
mean_ed15_high_male=mean(ed15_high_male)
mean_ed16_high_male=mean(ed16_high_male)
mean_ed17_high_male=mean(ed17_high_male)
mean_ed18_high_male=mean(ed18_high_male)
mean_ed_high_male=c(mean_ed12_high_male,mean_ed13_high_male,mean_ed14_high_male,
mean_ed15_high_male,mean_ed16_high_male,mean_ed17_high_male,mean_ed18_high_male)
mean_ed_high_male

#Means: "Low" fmaily income- MALES only
mean_ed12_low_male=mean(ed12_low_male)
mean_ed13_low_male=mean(ed13_low_male)
mean_ed14_low_male=mean(ed14_low_male)
mean_ed15_low_male=mean(ed15_low_male)
mean_ed16_low_male=mean(ed16_low_male)
mean_ed17_low_male=mean(ed17_low_male)
mean_ed18_low_male=mean(ed18_low_male)
mean_ed_low_male=c(mean_ed12_low_male,mean_ed13_low_male,mean_ed14_low_male,
mean_ed15_low_male,mean_ed16_low_male,mean_ed17_low_male,mean_ed18_low_male)
mean_ed_low_male

#### Boxplot: Family income "high" males only
par(mfrow=c(2,2))
boxplot(ed12_high_male,ed13_high_male,ed14_high_male,ed15_high_male,ed16_high_male,
ed17_high_male,ed18_high_male,
        xlab="Education level",
        ylab="Unemployment",
        main="High Income Males",
        names=c(12:18),
        vertical = TRUE)
points(x=c(mean_ed_high_male),pch=19, col="red")
```

```
### Boxplot: Family income "low" males only
boxplot(ed12_low_male,ed13_low_male,ed14_low_male,ed15_low_male,ed16_low_male,
ed17_low_male,ed18_low_male,
        xlab="Education level",
        ylab="unemployment",
        main="Low Income Males",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_low_male),pch=19, col="red")
```

Figure 14:

```
##### "High" family income
#### FEMALES ONLY

ed12_high_female=eData$unemp[eData$education==12 & eData$income=="high" &
eData$gender=="female"]
ed13_high_female=eData$unemp[eData$education==13 & eData$income=="high" &
eData$gender=="female"]
ed14_high_female=eData$unemp[eData$education==14 & eData$income=="high" &
eData$gender=="female"]
ed15_high_female=eData$unemp[eData$education==15 & eData$income=="high" &
eData$gender=="female"]
ed16_high_female=eData$unemp[eData$education==16 & eData$income=="high" &
eData$gender=="female"]
ed17_high_female=eData$unemp[eData$education==17 & eData$income=="high" &
eData$gender=="female"]
ed18_high_female=eData$unemp[eData$education==18 & eData$income=="high" &
eData$gender=="female"]

#Means: "High" family income - FEMALEs only
mean_ed12_high_female=mean(ed12_high_female)
mean_ed13_high_female=mean(ed13_high_female)
mean_ed14_high_female=mean(ed14_high_female)
mean_ed15_high_female=mean(ed15_high_female)
mean_ed16_high_female=mean(ed16_high_female)
mean_ed17_high_female=mean(ed17_high_female)
mean_ed18_high_female=mean(ed18_high_female)
mean_ed_high_female=c(mean_ed12_high_female,mean_ed13_high_female,mean_ed14_high_female,
mean_ed15_high_female,mean_ed16_high_female,mean_ed17_high_female,mean_ed18_high_female)
mean_ed_high_female

ed12_low_female=eData$unemp[eData$education==12 & eData$income=="low" &
eData$gender=="female"]
ed13_low_female=eData$unemp[eData$education==13 & eData$income=="low" &
eData$gender=="female"]
ed14_low_female=eData$unemp[eData$education==14 & eData$income=="low" &
eData$gender=="female"]
ed15_low_female=eData$unemp[eData$education==15 & eData$income=="low" &
eData$gender=="female"]
ed16_low_female=eData$unemp[eData$education==16 & eData$income=="low" &
eData$gender=="female"]
ed17_low_female=eData$unemp[eData$education==17 & eData$income=="low" &
eData$gender=="female"]
ed18_low_female=eData$unemp[eData$education==18 & eData$income=="low" &
eData$gender=="female"]

#Means: "Low" family income- FEMALES only
mean_ed12_low_female=mean(ed12_low_female)
mean_ed13_low_female=mean(ed13_low_female)
```

```
mean_ed14_low_female=mean(ed14_low_female)
mean_ed15_low_female=mean(ed15_low_female)
mean_ed16_low_female=mean(ed16_low_female)
mean_ed17_low_female=mean(ed17_low_female)
mean_ed18_low_female=mean(ed18_low_female)
mean_ed_low_female=c(mean_ed12_low_female,mean_ed13_low_female,mean_ed14_low_female,
mean_ed15_low_female,mean_ed16_low_female,mean_ed17_low_female,mean_ed18_low_female)
mean_ed_low_female

#### Boxplot: Family income "high" females only
boxplot(ed12_high_female,ed13_high_female,ed14_high_female,ed15_high_female,
ed16_high_female,ed17_high_female,ed18_high_female,
        xlab="Education level",
        ylab="unemployment",
        main="High Income Females",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_high_female),pch=19, col="red")

## Boxplot: Family income "low" females only
boxplot(ed12_low_female,ed13_low_female,ed14_low_female,ed15_low_female,
ed16_low_female,ed17_low_female,ed18_low_female,
        xlab="Education level",
        ylab="Unemployment",
        main="Low Income Females",
        names=c(12:18),
        vertical = T)
points(x=c(mean_ed_low_female),pch=19, col="red")

## educ ~ unemployment: fit1
summary(eData$unemp[eData$gender=="male"])
fit1 <- lm(education[eData$gender=="male"]~unemp[eData$gender=="male"],data=eData)
summary(fit1)
anova(fit1)

## educ ~ unemployment, fixing income==low: fit2
fit2 <- lm(education[eData$gender=="male" & eData$income=="low"]~unemp
[eData$gender=="male" & eData$income=="low"],data=eData)
summary(fit2)
anova(fit2)

## educ ~ unemployment, fixing income==high: fit3
fit3 <- lm(education[eData$gender=="male" & eData$income=="high"]~unemp
[eData$gender=="male" & eData$income=="high"],data=eData)
summary(fit3)
anova(fit3)

### FEMALES ONLY

## educ ~ unemployment: fit4
summary(eData$unemp[eData$gender=="female"])
fit4 <- lm(education[eData$gender=="female"]~unemp[eData$gender=="female"],data=eData)
summary(fit4)
anova(fit4)

## educ ~ unemployment, fixing income==low: fit5
fit5 <- lm(education[eData$gender=="female" & eData$income=="low"]~unemp
[eData$gender=="female" & eData$income=="low"],data=eData)
summary(fit5)
anova(fit5)
```

```
## educ ~ unemployment, fixing income==high: fit6
fit6 <- lm(education[eData$gender=="female" & eData$income=="high"]~unemp
[eData$gender=="female" & eData$income=="high"],data=eData)
summary(fit6)
anova(fit6)
#End of code for education level vs unemployment
```

Figure 15 & Figure 16:

```
ed12_h=eData$wage[eData$education==12 & eData$income=="high"]
ed13_h=eData$wage[eData$education==13 & eData$income=="high"]
ed14_h=eData$wage[eData$education==14 & eData$income=="high"]
ed15_h=eData$wage[eData$education==15 & eData$income=="high"]
ed16_h=eData$wage[eData$education==16 & eData$income=="high"]
ed17_h=eData$wage[eData$education==17 & eData$income=="high"]
ed18_h=eData$wage[eData$education==18 & eData$income=="high"]

ed12_l=eData$wage[eData$education==12 & eData$income=="low"]
ed13_l=eData$wage[eData$education==13 & eData$income=="low"]
ed14_l=eData$wage[eData$education==14 & eData$income=="low"]
ed15_l=eData$wage[eData$education==15 & eData$income=="low"]
ed16_l=eData$wage[eData$education==16 & eData$income=="low"]
ed17_l=eData$wage[eData$education==17 & eData$income=="low"]
ed18_l=eData$wage[eData$education==18 & eData$income=="low"]

mean_ed12_h = mean(ed12_h)
mean_ed13_h = mean(ed13_h)
mean_ed14_h = mean(ed14_h)
mean_ed15_h = mean(ed15_h)
mean_ed16_h = mean(ed16_h)
mean_ed17_h = mean(ed17_h)
mean_ed18_h = mean(ed18_h)
mean_ed_h=c(mean_ed12_h, mean_ed13_h, mean_ed14_h, mean_ed15_h,mean_ed16_h,
mean_ed17_h, mean_ed18_h)
mean_ed_h

var_ed12_h = var(ed12_h)
var_ed13_h = var(ed13_h)
var_ed14_h = var(ed14_h)
var_ed15_h = var(ed15_h)
var_ed16_h = var(ed16_h)
var_ed17_h = var(ed17_h)
var_ed18_h = var(ed18_h)
var_ed_h=c(var_ed12_h, var_ed13_h, var_ed14_h, var_ed15_h, var_ed16_h,
var_ed17_h, var_ed18_h)
var_ed_h

sd_ed_h=sqrt(var_ed_h)
sd_ed_h

mean_ed12_l = mean(ed12_l)
mean_ed13_l = mean(ed13_l)
mean_ed14_l = mean(ed14_l)
mean_ed15_l = mean(ed15_l)
mean_ed16_l = mean(ed16_l)
mean_ed17_l = mean(ed17_l)
mean_ed18_l = mean(ed18_l)
mean_ed_l=c(mean_ed12_l, mean_ed13_l, mean_ed14_l, mean_ed15_l, mean_ed16_l,
mean_ed17_l, mean_ed18_l)
mean_ed_l
```

```
var_ed12_l = var(ed12_l)
var_ed13_l = var(ed13_l)
var_ed14_l = var(ed14_l)
var_ed15_l = var(ed15_l)
var_ed16_l = var(ed16_l)
var_ed17_l = var(ed17_l)
var_ed18_l = var(ed18_l)
var_ed_l=c(var_ed12_l, var_ed13_l, var_ed14_l, var_ed15_l, var_ed16_l,
var_ed17_l, var_ed18_l)
var_ed_l

sd_ed_l=sqrt(var_ed_l)
sd_ed_l

q1=qt(0.975,4738)*sd_ed_h/sqrt(4739)
mean_ed_h-q1
mean_ed_h+q1

q2=qt(0.975,4738)*sd_ed_l/sqrt(4739)
mean_ed_l-q2
mean_ed_l+q2
```