Kyle Kazemini

U1127157

Math 3080-001

January 30, 2020

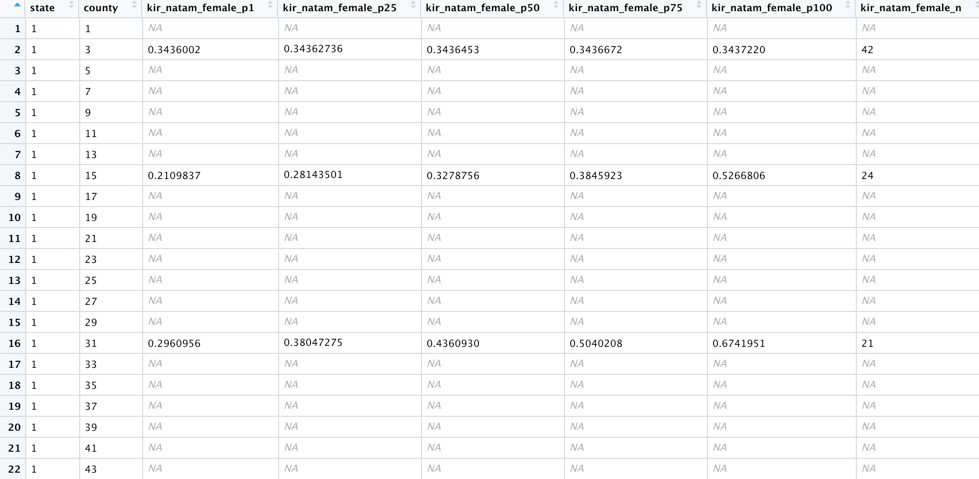
**Economic Mobility**

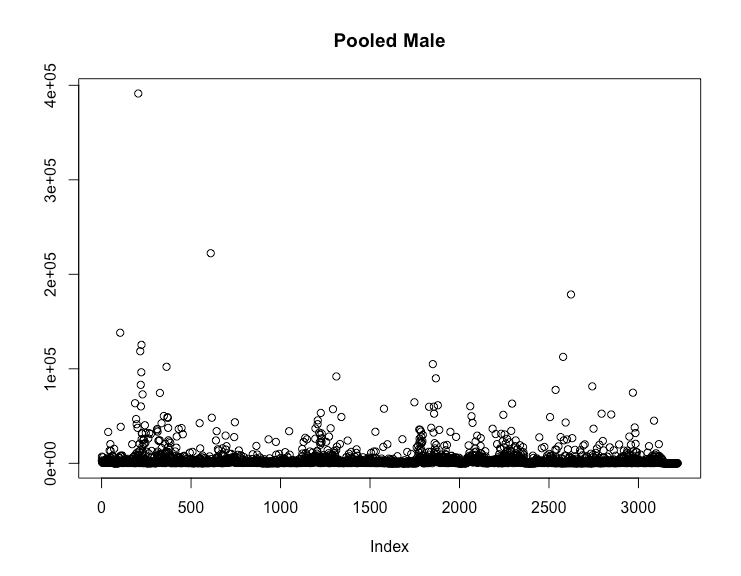
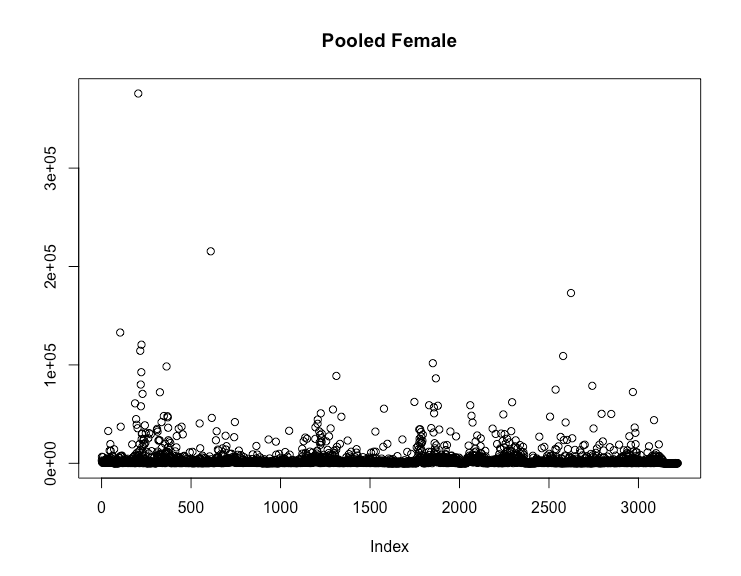
**Introduction**

Economic mobility is an individual’s ability to improve their economic status - usually measured by an increase in income. Some common ways to advance to a higher economic status (called vertical economic mobility) are job changes, marriage, and inheritance. Economic mobility can be a change between parents and children, called “intergenerational,” or over the course of one individual’s lifetime, called “intragenerational.” Since the 1970’s, the United States has been thought to have low vertical economic mobility compared to other first-world countries. This seems to be especially true for low income areas in the United States. This belief even comes through in popular culture, with many songs and movies stating that the only way to make it out of low-income neighborhoods is to become a professional athlete or musician. The goal of this project is to analyze how well these beliefs hold up. This project will focus less on children whose parents are already wealthy and more on children who grow up in low income areas. The main goal is to analyze how difficult it is to achieve intergenerational economic mobility in the United States.

**The Data**

This data was taken from Opportunity Insights. This site has publicly available datasets that are part of their research studies. They are made public for use by other researchers, and the general public, for use in other projects. The dataset is titled “All Outcomes by County, Race, Gender and Parental Income Percentile.” I chose this dataset because it relates parental income to location, as well as other demographics.

At first glance, the dataset seems to have a sparse structure. This is probably due to the fact that it’s such a large set of data. It has dozens of fields for each state, and each state has multiple counties. Some states have over fifty counties. There are so many fields because the dataset covers multiple areas for each race, gender, and income level, broken down into percentiles for each of these fields. Every field won’t be listed here because there are so many. Important fields will be analyzed in the next section. Here’s a small screenshot of the beginning of the dataset and a couple of graphs that show pooled success statistics for all races, split into males and females. For the remainder of the project, details about statistics and procedures will be discussed. Overarching patterns and their interpretations will be explained in the conclusion. 



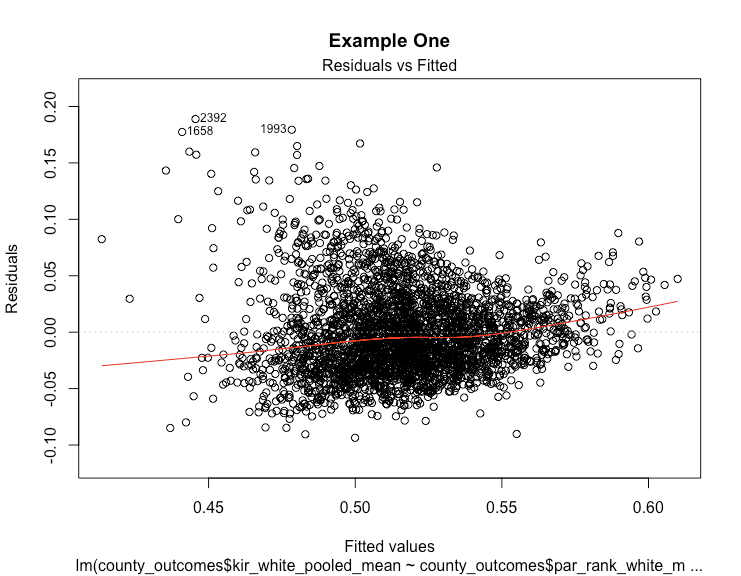
**Variables & Hypothesis**

One of the important variables that we’re concerned with is “kir”. This variable represents the mean percentile rank (relative to other children born in the same year) in the national distribution of individual income (i.e. just own earnings) measured as mean earnings in 2014-2015 for the baseline sample. The reason we’re interested in this variable is because we’re trying to find a correlation between parent’s income and child’s income/success. We’re also interested in “proginc” which represents the fraction of children who receive public assistance income, “kfr” which represents the mean percentile rank (relative to other children born in the same year) in the national distribution of household income (i.e. own earnings and spouse’s earnings) measured as mean earnings in 2014-2015 for the baseline sample, and “staycz” which represents the fraction of children who live in one of their childhood commuting zones in adulthood. The variable used to predict child income/success is “par\_rank\_[race]\_[gender]\_mean”. This variable represents the mean household income rank for parents of children of race [race] and gender [gender]. Parents are ranked relative to other parents with children in the same birth cohort.

**Model**

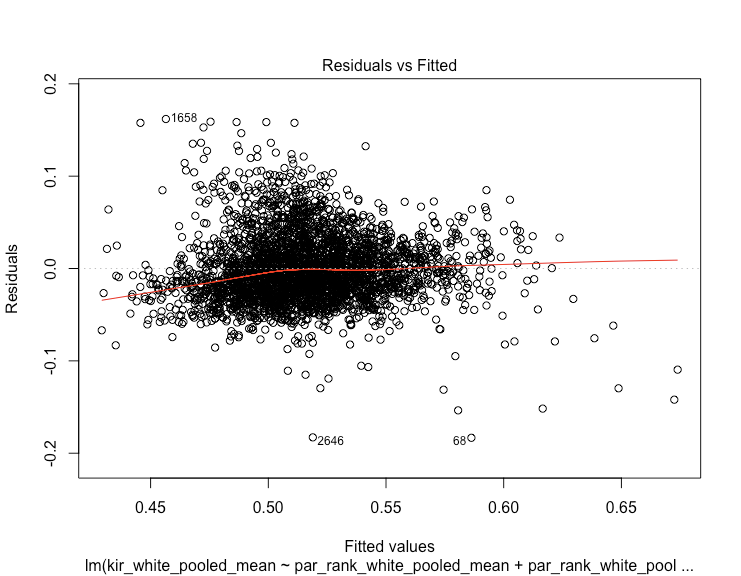
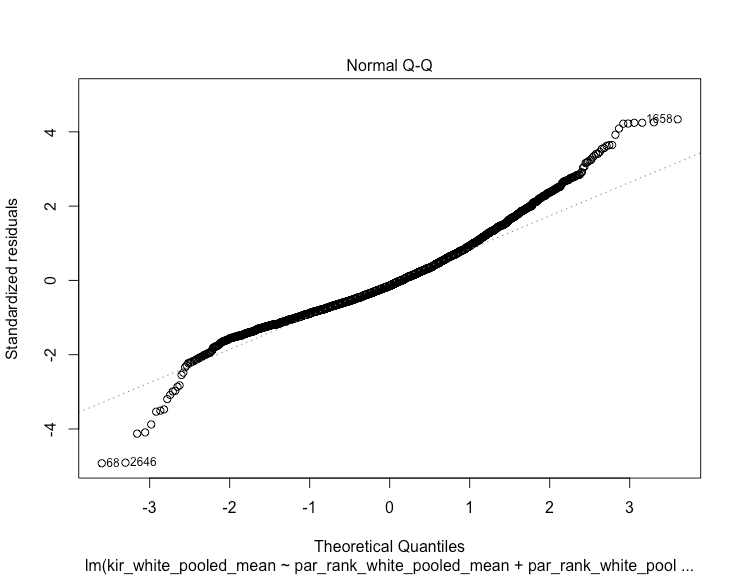
A simple linear regression model is of the form where and are parameters, and is a random, normally distributed error term with E() = 0 and V() = . The derivation of these values is important, but not relevant to this project’s goal. As discussed in the previous section, the “kir” variable, as well as a few others, will be modeled by “par\_rank\_[race]\_[gender]\_mean.” The values of the parameters and error term will be calculated using software. The code will be shown at the end of this project.

The validity of this model will be tested, as well as verified using the coefficient of determination. This value is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model.

A trial and error process is useful for determining the validity of different models. Different races, genders, percentiles, and statistics were used to find the strength of these relationships. For example, the coefficient of determination for income rank of white males modeled by their parent’s income was 0.211. This value suggests that there’s a relationship between the two variables. This graph shows that there is little relationship:

Since there are so many variables in this dataset, a technique for picking the best ones for linear regression is extremely useful. The “leaps” library will be used for this selection process.

Using functions in this library, information can be generated that shows the usefulness of each variable. These functions also give the N best variables in the model, based on the impact they have on the dependent variable.

Based on these tools, the best two variables for making predictions about children's success are mean parent rank and standard error of parent rank. These two variables represent the parent’s income rank relative to other parents. Here is a graph showing the relationships, as well as a Q-Q plot to demonstrate normality of the distribution. The graph shows that there are a lot of outliers and noise in the data, but this is expected.

The model is now established. The R-squared value for the model is 0.375 and the adjusted R-squared value is 0.3746. These values show that this is a valid and useful model for predicting child income and success. As stated earlier, a moderate R-squared value is expected because this is a very large dataset. A very high R-squared isn’t likely for even a moderately large dataset because in the natural world, there’s always a certain amount of variance. This is a very large dataset considering the scope of this project.

**Rejection**

Before starting any modeling, it’s important to establish a region of rejection. Doing this before starting any testing eliminates bias and ensures that the analysis is as objective as possible. The model utility test for simple linear regression will be used in this analysis. The test will be explained in the next section. The rejection region is:

where represents the confidence level for the test and n is the number of data points. n - 2 is the degrees of freedom for the test. The t on the left-hand side of the inequality is the test statistic for this hypothesis. It’s compared with the value from the t-distribution in order to test for statistical significance. If the value from the test statistic is greater than the one from the distribution, is rejected and the simple linear regression model is statistically significant.

**Test**

Now that a region of rejection has been established, testing can be conducted. As stated earlier, the model utility test will be used. Note: all assumptions made by this test in general are also made in this use case. Here are the null and alternative hypotheses:

When the null hypothesis is true, independent of x. In other words, knowledge of x gives no information about the value of the dependent variable. When the null hypothesis is rejected, the simple linear regression model gives information about the relationship between the two variables. The model can then be used for further inferences and calculations.

The test statistic used for conducting the model utility test is shown below:

Since the value of is 0, the test statistic becomes:

The test is now ready to be conducted. Using software, a value from a t distribution is generated using alpha level = 0.95 and n - 2 degrees of freedom. This value is 0.8289082. The test statistic is also calculated using the formula above, and this value is 17.39. The standard error for is very small, so the t statistic is quite large. As stated earlier, the rejection region for the test is anything greater than the distribution value. Therefore, the null hypothesis is rejected and this test has shown that the regression model gives valuable information about child success. This test has shown that parent income rank is a good indicator of child income and success.

**Conclusion**

The purpose of this project was to analyze economic mobility in the United States. As stated in the introduction, economic mobility is an individual’s ability to improve their economic status. A simple linear regression model, as well as some other basic statistics, were used to analyze the relationship between parent income and child income/success. The regression model didn’t explain an extremely high percentage of the variation of the data, but this was expected. This is a large dataset, and it’s unlikely that a model can completely explain the data’s variation. That being said, the R squared and adjusted R squared values for this model show that it’s a good model for predicting economic mobility. Also, the model utility test indicates that parent income gives valuable information about child success/income. Therefore, this project has successfully shown that there is a relationship between parent income and child income. This suggests that economic mobility in the United States is difficult to achieve because children whose parents have high income tend to also have high income, and the same holds for low income parents and children.

**Code**

#'Math 3080 Term Project

#'Spring 2020

#'Kyle Kazemini

#'February 25, 2020

#First, import and view the data, as well as some useful libraries

library(leaps)

library(stats)

setwd("/Users/kylekazemini/Downloads")

county\_outcomes <- read.csv("county\_outcomes.csv")

#Check if it's in the format we want

is.data.frame(county\_outcomes)

View(county\_outcomes)

l <- sapply(county\_outcomes, function(x) is.factor(x))

m <- county\_outcomes[, l]

ifelse(sapply(m, function(x) length(levels(x)) == 1, "DROP”, “NODROP"))

#Select variables using leaps library

reg <- regsubsets(as.matrix(response) ~ as.factor(m), data = county\_outcomes, nbest = 2, really.big = TRUE)

plot(reg, scale = "adjr2")

summary(reg)

#Establish a model

linearModel <- lm(kir\_white\_pooled\_mean ~ par\_rank\_white\_pooled\_mean + par\_rank\_white\_pooled\_mean\_se, data = county\_outcomes)

selectedModel <- step(linearModel)

summary(selectedModel)

#Graph the model

plot(selectedModel)

#T statistic

t <- summary(selectedModel)$t.value

t

#T value

pt(0.95, df = 3217)

#Other plots

plot(county\_outcomes$kir\_pooled\_female\_n, main = "Pooled Female", ylab = "")

plot(county\_outcomes$kir\_pooled\_male\_n, main = "Pooled Male", ylab = "")

############################################################################

#Practice Code

mean(county\_outcomes$kir\_natam\_female\_p1, na.rm = TRUE)

response <- county\_outcomes['kir\_white\_pooled\_mean']

predictors <- county\_outcomes[, !names(county\_outcomes) %in% "kir\_white\_pooled\_mean"]

predictors2 <- na.omit(predictors)

fit <- lm(county\_outcomes$kir\_white\_pooled\_mean ~ county\_outcomes$par\_rank\_white\_male\_mean)

print(fit)

summary(fit)$r.squared

summary(fit)$adj.r.squared

modelSelection <- lm(county\_outcomes$jail\_pooled\_female\_p100 ~ . - 1, data = county\_outcomes)

model1 <- glm(county\_outcomes$kir\_natam\_female\_p100 ~ . - 1, data = county\_outcomes)

scatter.smooth(county\_outcomes$par\_rank\_white\_male\_mean, county\_outcomes$kir\_white\_pooled\_mean)

**Sources**

“Data Library.” *Data | Opportunity Insights*, opportunityinsights.org/data/.

Devore, Jay L. *Probability and Statistics for Engineering and the Sciences*. Nelson, 2018.

Hays, William L. *Statistics*. Wadsworth/Thomson Learning, 2008.

Kassambara, and Sfd. “Best Subsets Regression Essentials in R.” *STHDA*, 11 Mar. 2018, [www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/](http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/).

Prabhakaran, Selva. “Eval(ez\_write\_tag([[728,90],'r\_statistics\_co-Box-3','ezslot\_4',109,'0','0']));Linear Regression.” *Linear Regression With R*, r-statistics.co/Linear-Regression.html.

Prabhakaran, Selva. “Eval(ez\_write\_tag([[728,90],'r\_statistics\_co-Box-3','ezslot\_4',109,'0','0']));Model Selection Approaches.” *Model Selection*, r-statistics.co/Model-Selection-in-R.html.

REnthusiastREnthusiast 1, et al. “Error in Contrasts When Defining a Linear Model in R.” *Stack Overflow*, 1 May 1963, stackoverflow.com/questions/18171246/error-in-contrasts-when-defining-a-linear-model-in-r.

Rio, Eric del. “Attempt at Multiple Regression in R.” *Medium*, Human Systems Data, 29 Mar. 2017, medium.com/humansystemsdata/attempt-at-multiple-regression-in-r-ff9a110a1405.

“Socioeconomic Mobility in the United States.” *Wikipedia*, Wikimedia Foundation, 22 Jan. 2020, en.wikipedia.org/wiki/Socioeconomic\_mobility\_in\_the\_United\_States.