Census-Income-Prediction-Final

May 1, 2022

1 Introduction and Background

One of the most pressing issues currently being faced in the United States today is poverty. While a common idea is that it involves a person or community lacking the necessary finances, resources, or even shelter. Even so, there is a far-reaching meaning when we closely look at what factors are associated with poverty. This led us to our next interest in how one's income is affected. Thus, in extension to our last project, we will be studying how attributes such as Race, Sex, Age, etc. may affect an individual's income. We aim to apply the statistical techniques and tools that we learned from this semester to build predictive models and discover which predictors are significant to this study. Additionally, we were interested in looking at Minnesota and Colorado individually, then compare how well our prediction models compare are the end of the project.

For the purpose of this study, an individual or personal income is reported based on the previous year of the person's wages, salaries, and any other type of money income received from an employer.

1.1 Importing Dataset

The data was extracted from the IPUMS website, a public database that provides numerous samples from around the world, including the American Community Survey (ACS) conducted by the U.S Census Bureau. This online database provided us with the option to select attributes that cater towards our study in regards to predicting an individual's income. This dataset will be the 2018 ACS (5-year estimate) and is observational since it has been collected to measure or survey an adult individual's income levels, demographics, employment rates, etc. for each state. The variables in the dataset included:

- **YEAR**: Year when the ACS was conducted
- **STATEFIP**: State FIPS Code
- **FAMSIZE**: Number of family members in household [**continuous**]
- **SEX**: Male or Female [**categorical**]
- **AGE**: Age in years (18 and older) [**continous**]
- **MARRNO**: Number of times the person was married [**continous**]
- **RACE**: Race (original version) [**categorical**]
- **YRSUSA1**: Years lived in the US [**continous**]
- **EDUC**: Educational attainment [**categorical**]

- **CLASSWKR**: Class of worker [**continous**]
- **UHRSWORK**: Usual hours worked per week [**continous**]
- **INCWAGE**: Individual's income and salary wage [**continous**]
- **POVERTY**: Poverty Status [**categorical**]

- **OCCSCORE**: Occupational Score [**continous**]
- **DIFFSENS**: Vision or hearing difficulty [**categorical**]

YEAR	STATEFIP	FAMSIZE	SEX	AGE	MARRNO	RACE	RACED	YRSUSA1	EDUC	EDUC
2018	8	1	1	22	0	1	100	0	6	63
2018	8	5	2	58	1	1	100	0	6	63
2018	8	5	1	58	1	1	100	0	11	115
2018	8	5	2	34	1	1	100	0	11	114
2018	8	5	1	39	1	1	100	0	11	114
2018	8	2	2	60	1	1	100	0	10	101

```
     YEAR         STATEFIP        FAMSIZE           SEX
Min.   :2018   Min.   : 8.00   Min.   : 1.000   Min.   :1.000
1st Qu.:2018   1st Qu.: 8.00   1st Qu.: 2.000   1st Qu.:1.000
Median :2018   Median :27.00   Median : 2.000   Median :2.000
Mean   :2018   Mean   :17.52   Mean   : 2.573   Mean   :1.504
3rd Qu.:2018   3rd Qu.:27.00   3rd Qu.: 3.000   3rd Qu.:2.000
Max.   :2018   Max.   :27.00   Max.   :17.000   Max.   :2.000
     AGE            MARRNO            RACE           RACED
Min.   :18.00   Min.   :0.0000   Min.   :1.000   Min.   :100.0
1st Qu.:34.00   1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:100.0
Median :50.00   Median :1.0000   Median :1.000   Median :100.0
Mean   :49.63   Mean   :0.9672   Mean   :1.408   Mean   :142.2
3rd Qu.:63.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:100.0
Max.   :95.00   Max.   :3.0000   Max.   :9.000   Max.   :990.0
   YRSUSA1           EDUC            EDUCD          CLASSWKR
Min.   : 0.00   Min.   : 0.000   Min.   :  2.00   Min.   :0.00
1st Qu.: 0.00   1st Qu.: 6.000   1st Qu.: 63.00   1st Qu.:1.00
Median : 0.00   Median : 7.000   Median : 71.00   Median :2.00
Mean   : 1.95   Mean   : 7.637   Mean   : 78.65   Mean   :1.47
3rd Qu.: 0.00   3rd Qu.:10.000   3rd Qu.:101.00   3rd Qu.:2.00
Max.   :89.00   Max.   :11.000   Max.   :116.00   Max.   :2.00
  CLASSWKRD        UHRSWORK         INCWAGE          POVERTY        OCCSCORE
Min.   : 0.00   Min.   : 0.00   Min.   :     0   Min.   :  0   Min.   : 0.00
1st Qu.:13.00   1st Qu.: 0.00   1st Qu.:     0   1st Qu.:197   1st Qu.:12.00
Median :22.00   Median :36.00   Median : 15980   Median :372   Median :24.00
Mean   :17.09   Mean   :27.11   Mean   : 33712   Mean   :334   Mean   :22.29
3rd Qu.:22.00   3rd Qu.:40.00   3rd Qu.: 49170   3rd Qu.:501   3rd Qu.:33.00
Max.   :29.00   Max.   :99.00   Max.   :508091   Max.   :501   Max.   :80.00
   DIFFSENS
Min.   :1.00
1st Qu.:1.00
Median :1.00
Mean   :1.07
3rd Qu.:1.00
Max.   :2.00
```

Here, we sought to build predictive models and formulate a hypothesis concerning the potential predictors that affect an individual income or wage (**INCWAGE**) for Colorado and Minnesota.

Additionally, IPUMS provided a Data Documentation Initiative (DDI) file. In short, this helped us relabel some values that were assigned to descriptive labels (i.e, the variable **SEX**, Male and Female code values are 1 and 2 respectively) and provided us with further details regarding variables that are included in the dataset.

2 Method and Results

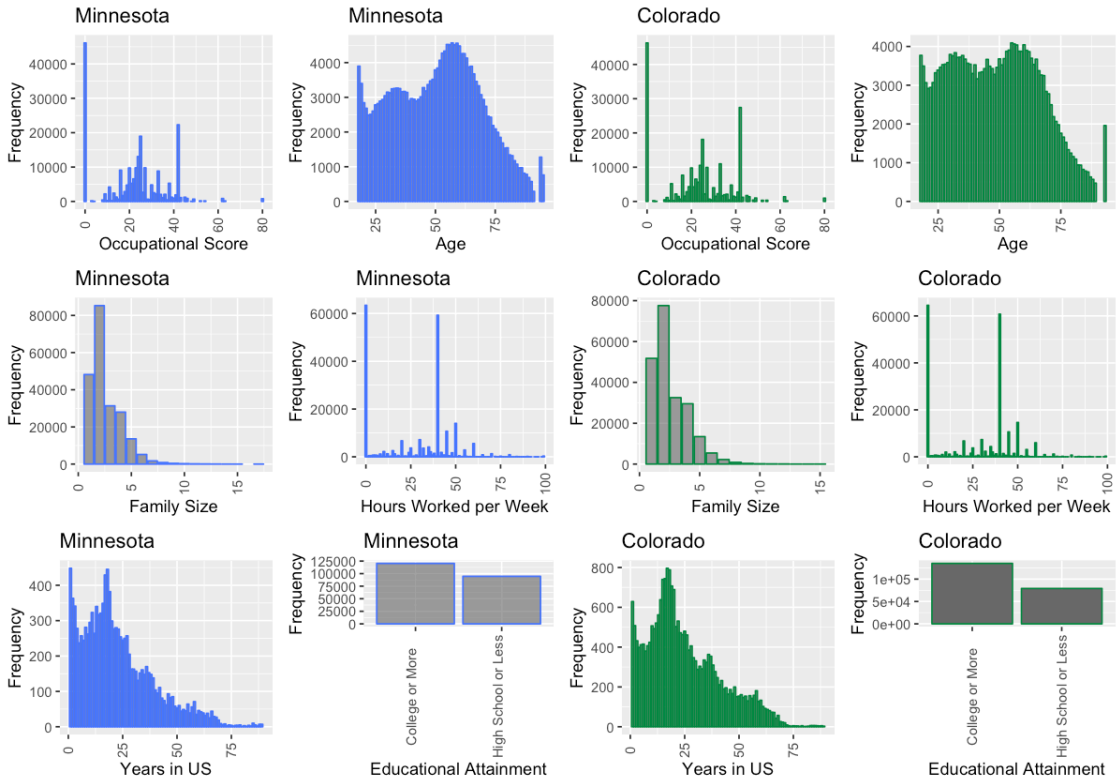
After importing the dataset, much of the project was spent cleaning and manipulating it. The process consisted of merging datasets, fixing structural errors, removing irrelevant data, and amending variables in order to ensure the dataset is consistent. The unfiltered dataset had raw numerical values that corresponded with a category, while the cleaned dataset had category descriptive values. However, due to complications, Anaconda environment did not support the **ipumsr** package. To work around this issue, we utilized R Studio and read_ipums_ddi() function to relabel the values within the original dataset to be more detailed. Additionally, we have sorted the data to only adult individuals, ages 18 and older, this is to reduce the 0 values with respect to income.

To polish our datasets, we worked with both the unfiltered and cleaned the dataset to impute correct values and corresponding descriptions. We noticed that there were variables whose values that represented NA's and too many descriptive elements for what we needed. The mismatched values were then removed by replacing the zero's with NA's so that the linear model do not interpret these zero values as actual input. Then we reduced the amount of description for the Race and Education variables due to its substantial amount of redundant factors.

FAMSIZE	SEX	AGE	MARRNO	RACE	YRSUSA1	EDUC	CLASSWKR	UHRS
3	Male	37	1	White	NA	College or More	Works for wages	40
3	Female	37	1	White	NA	College or More	Works for wages	50
2	Female	25	0	White	NA	College or More	Works for wages	36
5	Female	43	1	White	NA	College or More	Works for wages	32
5	Male	42	1	White	NA	College or More	Works for wages	50
5	Female	18	0	White	NA	High School or Less	N/A	0

2.1 Exploratory Analysis

Following the data cleaning process, we performed simple graphical summaries of the dataset to explore each variable, where we color-coordinated the states with blue and green to Minnesota and Colorado respectively. This is to help distinguish between the plots, to remove any confusion, we kept the colors consistent with the corresponding states. Furthermore, these graphical summaries helped us identify any new noticeable irregularities of the dataset. Note that **OCCSCORE** has an unusual amount of 0 values, this was a similar irregularity in that we found during the cleaning process. Thus, we replaced these values with NA to help with our analysis process later.



Above, we printed histograms corresponding to each numeric variable - **income, occupational score, age, family size, hours worked per week, years in the U.S.**, and **educational attainment** - to explore distribution of each variable. **Educational attainment** was reduced from 10 to 2 categories for simplicity.

Looking at Minnesota's data none of the graphs have a normal distribution, especially when we focus on **income, occupational score, and hours worked per week,**. Most income values are in the zero range, however, there is a slight hump there on after. Furthermore, a lot of adult individuals have an income of 0, this may be due to some of the individuals were around the age of 18-20 could be that they are college students, none of them have started to work, and so on. Another idea is that many citizens left out income information for privacy reasons, however, there could be numerous reasons why there were incomplete observations. A few outliers existed at $500,000$ which meant that the wealthiest individuals (the 1% in America). **Occupational score** follows a random histogram with most of its survey at 0 and between 15 to 43. The numeric values that were graphed on the x-axes represented an occupation, therefore we took note that many people between 15 and 43 have similar jobs. A few outliers also resided at the values 80 which could indicate a higher paying occupation. **Hours worked per week** has 2 large pillars in there graph, 0 and 40. This seems logical, because most people work 40 hours per week. The 0th pillar looked to represents individuals with no jobs a very small fraction of outliers work roughly 100 hours per week (which do, in fact exist).

Minnesota's **Age** histogram resembled a bimodal distribution, **Family Size** had a skewed distribution, and **Years in the U.S.** had a skewed bimodal distribution. Looking at the **Age** histogram most individuals were around the ages 50 to 70 and a slightly smaller amount around 18 to 35.

This seemed odd since we hypothesized that most working individuals would be around 25 to 40. The **Family Size** histogram skewed around a family size of 2 then decreased down to a family of 6. We see outliers at a family size of 15, wow! **Years in the U.S.** had two large distribution points at 0 and 20. This can mean that many people in 2018 just moved to Minnesota and a large portion of the population had lived in Minnesota for 20 years.

Observing Colorado's histograms we see similar distributions to Minnesota. Colorado's histrograms are on the same variable details as Minnesota's therefore we will do more comparing than explanation. If the reader is intersted in more details about the variables please visit Ipums. **Individual income** has a similar distribution to Minnesota's **individual income** histogram, however, Colorado has a smaller peak than Minnesota. Colorado also has more outliers at $500,000$ than Minnesota's which was interesting. **Occupation score** had a similar range as Minnesota's score, but Colorado had more individuals with an occupation score of 40. **Age** is distribution is much different than Minnesota's. Most people's age in Minnesota was around 50 to 70 with a smaller group around 18 to 40. Notice Colorado's age group has similar peak at 25 to 40 and 50 to 70. We could see a greater amount of younger people work in Colorado than Minnesota. **Family Size** and **Hours worked per week** was almost the same as Minnesota's histogram so we wont comment more on that than what is mentioned above. **Years in the U.S.** had a smaller amount of people who have 0 years in the U.S. than Minnesota's.

Our last two graphs within **Educational Attainment** presented the number of individuals who would graduate or attended college vs the individuals who graduated or attended high school. Colorado and Minnesota had more individuals that graudate or attend college, however, Colorado had more individuals who graduated or attended high school than Minnesota.

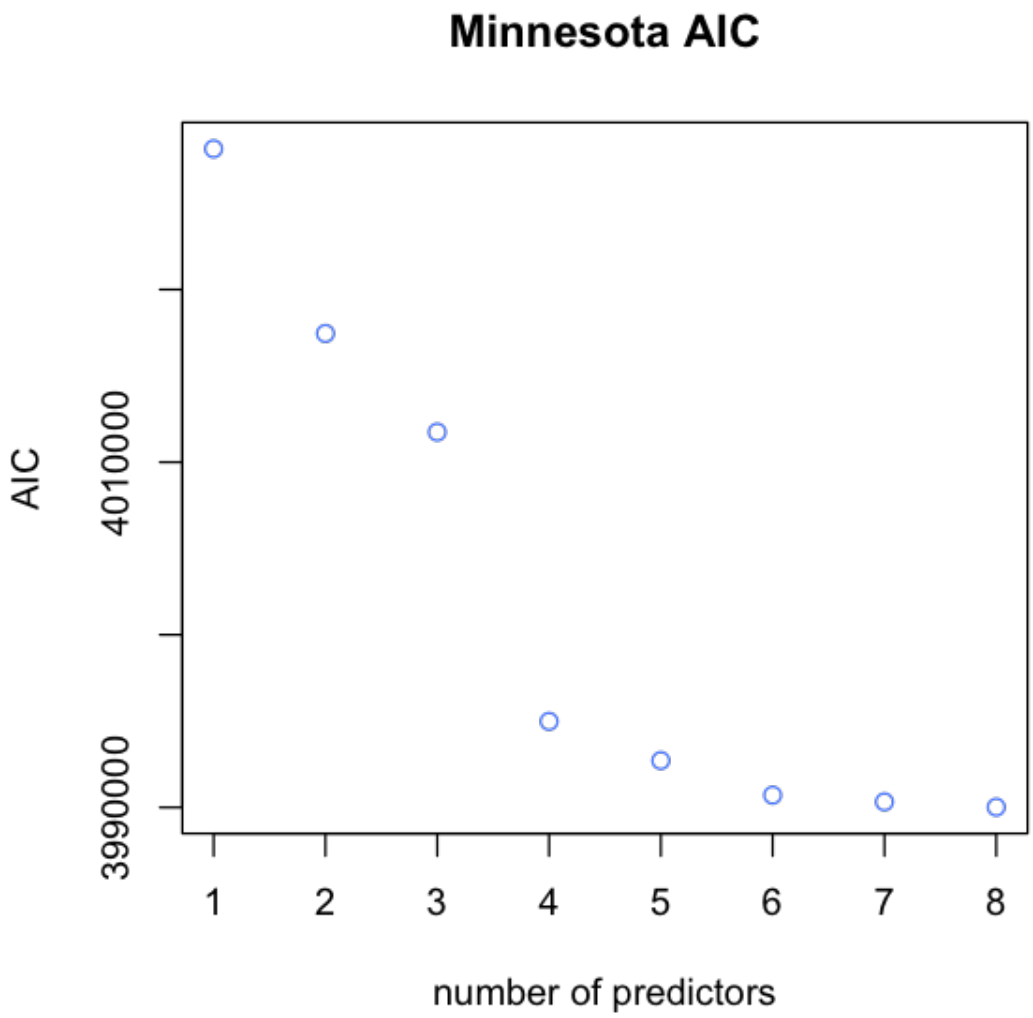
2.2 AIC Model Selection

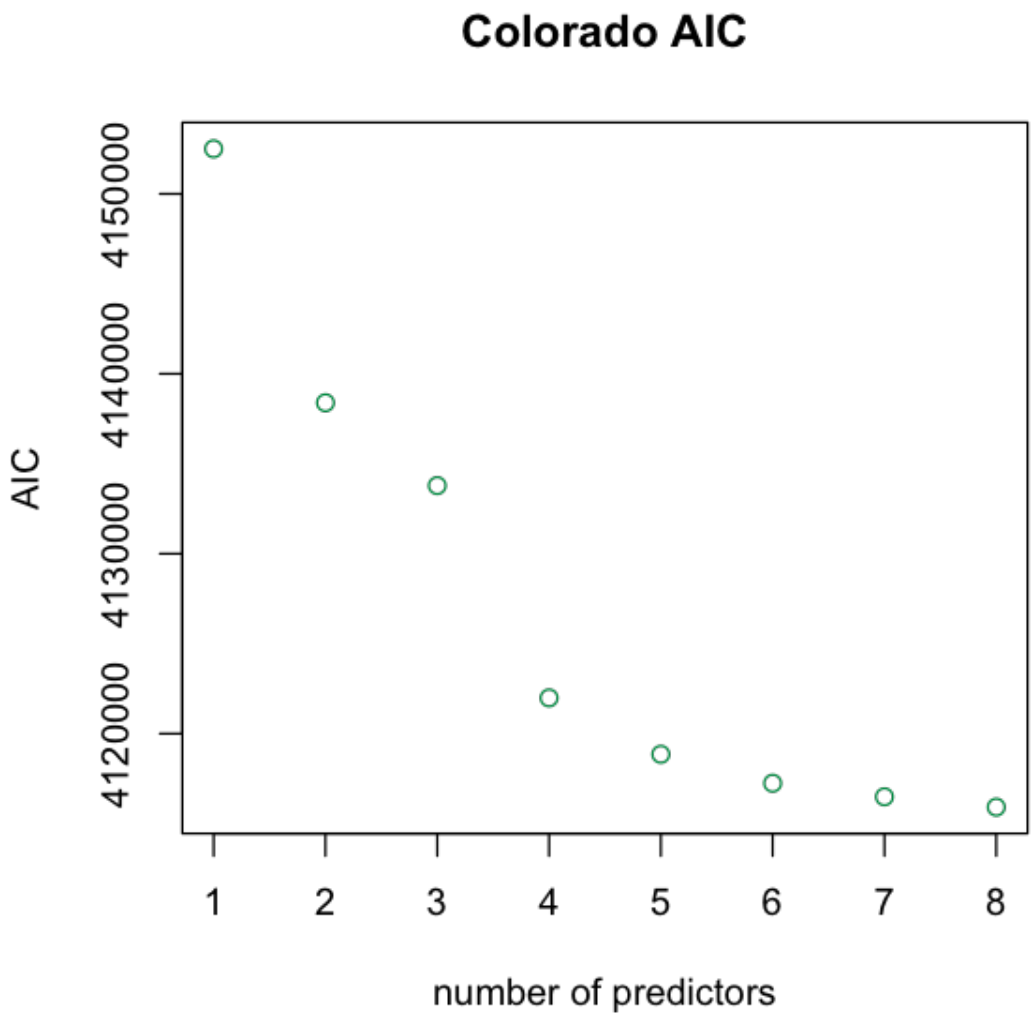
`Table of best models (in terms of RSS) for the Minnesota dataset:`

(Intercept)	FAMSIZE	SEXMale	AGE	MARRNO	RACEAsian American or Pacific Islander	RACEI
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE

`Table of best models (in terms of RSS) for the Colorado dataset:`

(Intercept)	FAMSIZE	SEXMale	AGE	MARRNO	RACEAsian American or Pacific Islander	RACEI
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE


Minnesota AIC
AIC
4010000
3990000
1
2
3
4
5
6
7
8
number of predictors



Next, we used AIC to help us with the model selection process in terms of RSS (size $k = 1, 2, \cdots, 8$). The table and plots below would help us to determine the the "best" model to move forward with.

The `regsubset$which` table and AIC plot helped us decide which predictors to use. In the AIC plot for both Colorado and Minnesota, the most ideal AIC for both was 8. However, we wanted a definite answer that 8 was the best number of predictors. Thus, we concluded with the best three AIC values. We decided to select 6th, 7th, and 8th for our predictors.

The 6 predictors were **Age, Education, Class of Work, Hours worked, and Occupational score**. **Class of work** was interpreted as two predictors in `R`, self-employed and works for wages. The 7th predictor added **SEX** and the 8th predictor added **Marriage**. We decided to add **race** to the 8th predictors because we believed that race is a huge factor in someone's income. Hence, our 8th predictor model was actually 9.

```
[1] "Anova test for Minnesota:"
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
215015	3.665492e+14	NA	NA	NA	NA
215014	3.656511e+14	1	898090625983	528.1047	1.010242e-116

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
215014	3.656511e+14	NA	NA	NA	NA
215013	3.647418e+14	1	909300128691	536.0267	1.928465e-118

```
[1] "Anova test for Colorado:"
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
214117	4.474257e+14	NA	NA	NA	NA
214116	4.456595e+14	1	1.766195e+12	848.5642	3.457256e-186

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
214116	4.456595e+14	NA	NA	NA	NA
214115	4.446904e+14	1	969105699696	466.617	2.25252e-103

To test which linear model is best we utilized the F-test with significant level of $\alpha = 0.05$. For Colorado and Minnesota we compared the linear models with 6, and 7 predictors.

Our first hypotheses are:

$$H_0 : Y_i = \beta_0 + \beta_{age}\,(age) + \beta_{educ}\,(educ) + \beta_{classwkr}\,(classwkr) + \beta_{occscore}\,(occscore) + \varepsilon_i$$

with

$$H_1 : \text{SEX should be included in the model.}$$

We found that the p-values was less than α hence reject the reduced model, 6 predictors, and compare the next following models, 7 to 9.

The second hypotheses:

$$H_0 : \text{Model with 7 predictors.}$$

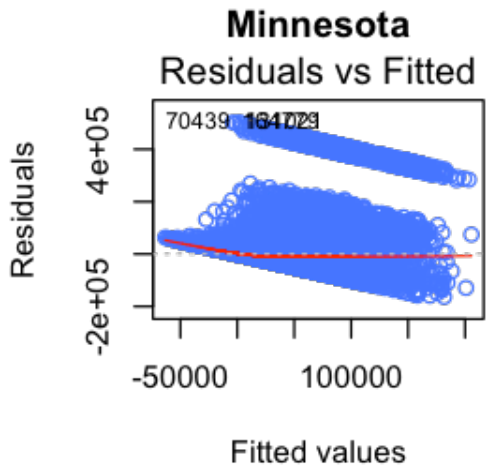
with

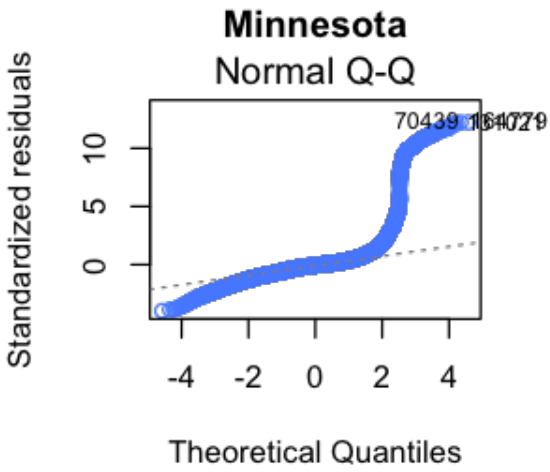
$$H_1 : \text{SEX or MARRNO (or both) should be included in the model.}$$

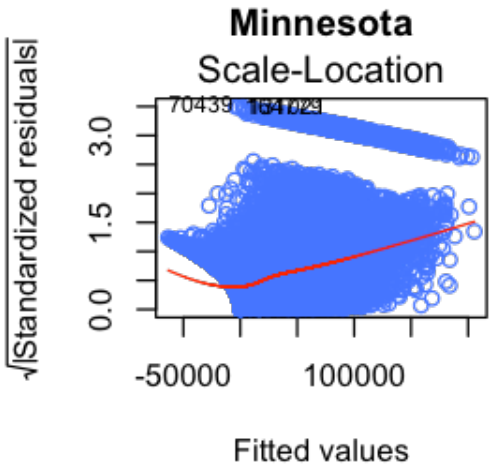
In both states the p-value was less than α therefore we concluded with 9 predictors.

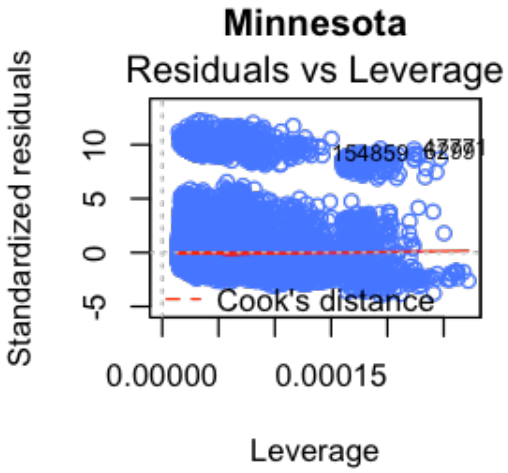
```
Call:
lm(formula = INCWAGE ~ SEX + AGE + MARRNO + EDUC + CLASSWKR +
    UHRSWORK + OCCSCORE, data = mn)

Residuals:
    Min      1Q  Median      3Q     Max
-163669  -15422   -2560    7377  503409
```

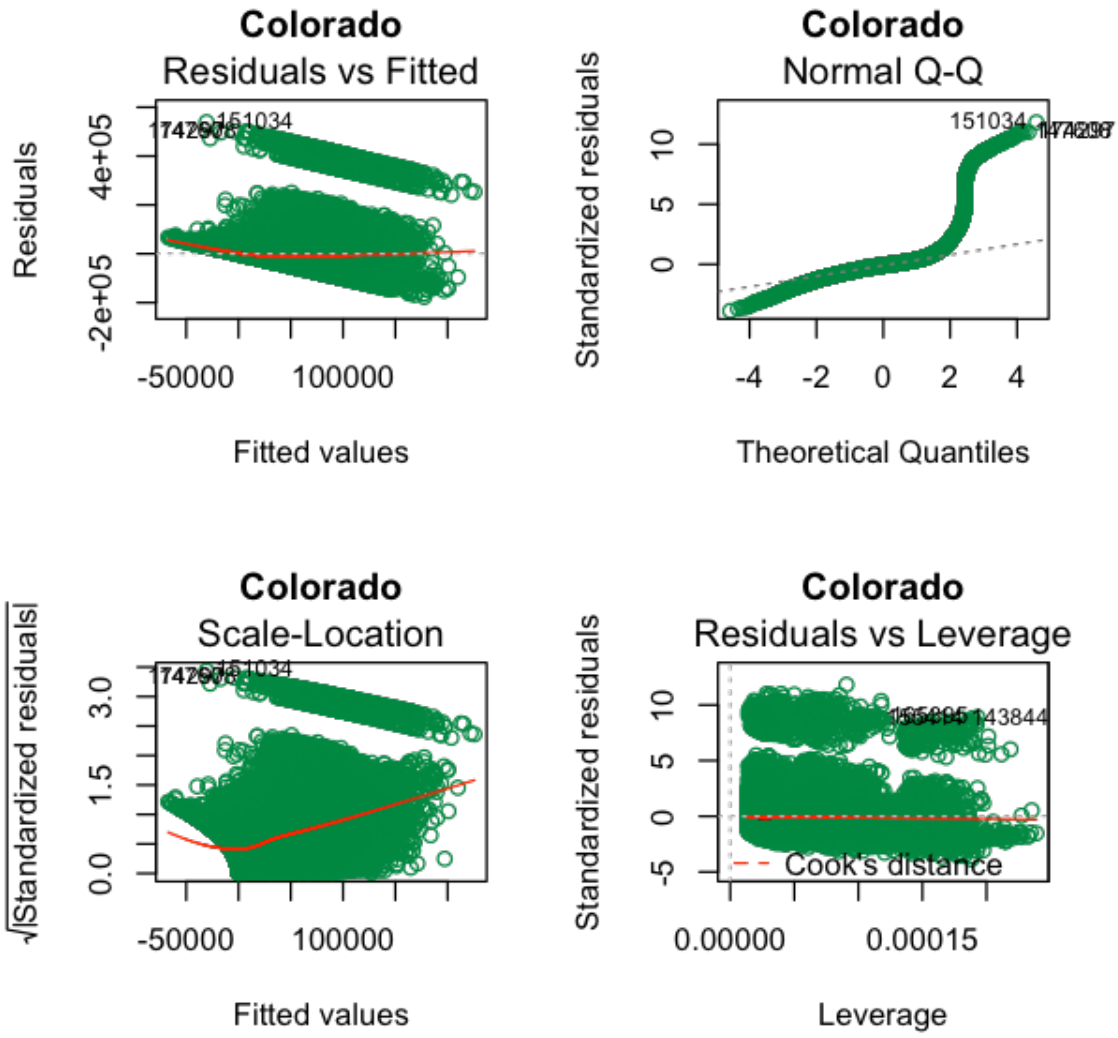
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -9626.157    450.983  -21.34   <2e-16 ***
SEXMale                   4553.458    183.959   24.75   <2e-16 ***
AGE                        143.003      6.401   22.34   <2e-16 ***
MARRNO                    3493.393    150.888   23.15   <2e-16 ***
EDUCHigh School or Less  -9375.328    188.532  -49.73   <2e-16 ***
CLASSWKRSelf-employed   -60589.087    475.741 -127.36   <2e-16 ***
CLASSWKRWorks for wages -36307.390    402.289  -90.25   <2e-16 ***
UHRSWORK                  1108.137      6.065  182.72   <2e-16 ***
OCCSCORE                  1557.655     10.248  152.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41190 on 215013 degrees of freedom
Multiple R-squared:  0.3705,        Adjusted R-squared:  0.3705
F-statistic: 1.582e+04 on 8 and 215013 DF,  p-value: < 2.2e-16



Call:
lm(formula = INCWAGE ~ SEX + AGE + MARRNO + EDUC + CLASSWKR +
    UHRSWORK + OCCSCORE, data = co)

Residuals:
    Min      1Q  Median      3Q     Max
-176638  -18078   -3621    9102  538657

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -14045.350    472.901  -29.70   <2e-16 ***
SEXMale                    6238.918    203.022   30.73   <2e-16 ***
AGE                         201.855      7.074   28.54   <2e-16 ***
MARRNO                     3259.486    150.893   21.60   <2e-16 ***
EDUCHigh School or Less   -9856.518    212.883  -46.30   <2e-16 ***
CLASSWKRSelf-employed    -60073.431    522.444 -114.98   <2e-16 ***
CLASSWKRWorks for wages  -38802.663    438.651  -88.46   <2e-16 ***
UHRSWORK                   1255.505      6.669  188.25   <2e-16 ***
OCCSCORE                   1552.693     10.654  145.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45570 on 214115 degrees of freedom
Multiple R-squared:  0.3702,        Adjusted R-squared:  0.3702
F-statistic: 1.573e+04 on 8 and 214115 DF,  p-value: < 2.2e-16
```


Minnesota
Residuals vs Fitted
Residuals
Fitted values
-2e+05
4e+05
-50000
100000


Minnesota
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-4
-2
0
2
4
0
5
10


Minnesota
Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.5
3.0
-50000
100000


Minnesota
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
-5
0
5
10
0.00000
0.00015



The summary of both 9 predictor models we noticed that most predictors reject the null hypothes meaning they have an association to an individuals income. We found it surprising that African americans and other races did not have a significant p-values. Is known that most poor communities are black and other races so this was odd to see that the model did not find it significant. After further analysis, the reason the p-value for African Americans and other races was not significant was that the model associates any predictor that ***increases*** an individual income not decrease. This could be why White has a significant p-value since there was an association that affected the individual's income.

The F-test for both models was below our significant level $\alpha = 0.05$, therefore, we needed one or more of the predictors included in the model.

Notice that the β_0 for Minnesota was larger than Colorado's β_0. Both β_0 estimates were negative, which suggested that individuals had a decreasing income, however, Colorado had a larger

decreasing income than Minnesota's. The β_0's for Minnesota and Colorado were -5415.4094 and -22812.300, respectively. We also observed that if an individual was a male, the amount of income increase the most adjusting for the other predictors in the model. On the other hand, if an individual is self-employed their income decrease the most when all other predictors are fixed. These observations were for both states.

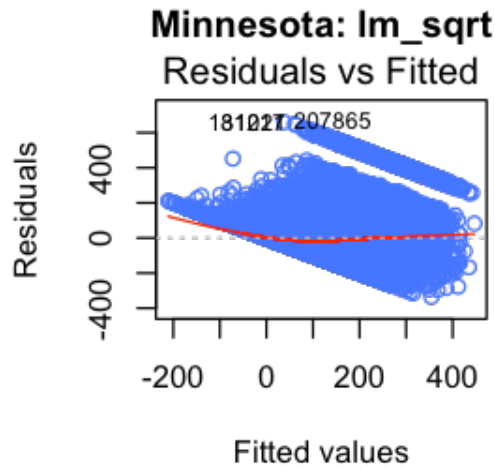
Based on the Residuals vs. Fitted, Normal Q-Q, and Residual vs. Leverage plots, there were noticeable violations of our assumptions. First, the Residual vs Fitted plot of Colorado and Minnesota appeared to have non-constant variance, therefore, this violated our variance assumption. Both graphs had a negative trend gap between $600,000$ to $300,000$ and a negative trend around 0.

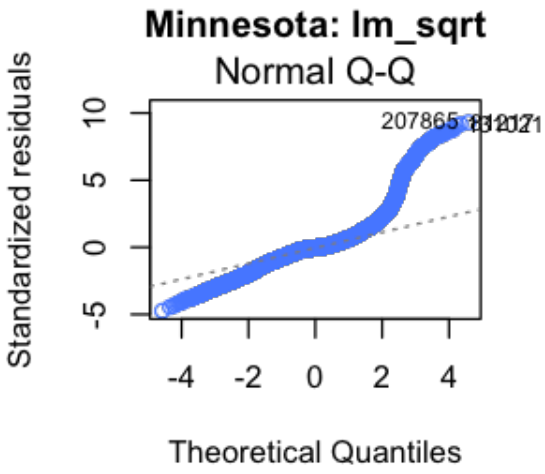
Furthermore, the Colorado and Minnesota Normal Q-Q plot seemed to violate our normality assumption. In both graphs, there was a snake structure stray away from the linear line. Around the middle of the linear line, the points seemed to stabilize but as the quantile decreased, the points veered away from the linear line. We believe that this was most likely caused from the violation of linearity assumption. We won't focus on the Scale-Location plot, however, note that it is a similar type of plot to the Residuals vs Fitted.

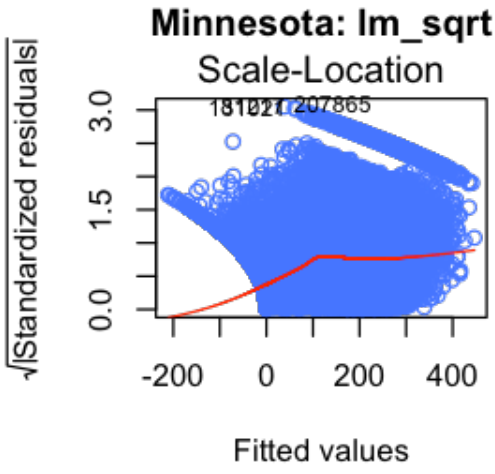
The Residuals vs. Leverage in Minnesota has more scatter compared to Colorado's Residual vs Leverage. Yet, both plots showed no potential influential points or any points that crossed Cook's distance. There were a few outliers in both graphs and that could be explained from the exploratory analysis done above. Nevertheless, there were no points that we had considered influential points.

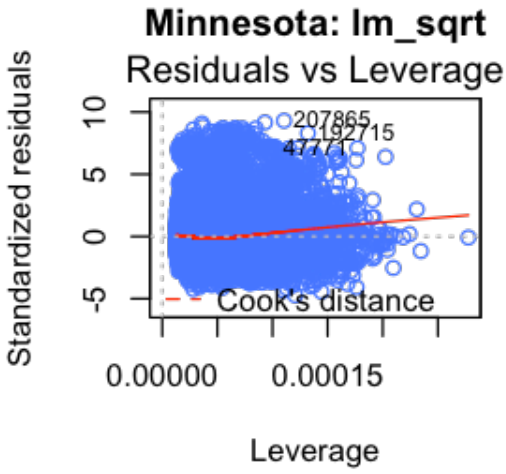
2.3 Transformations

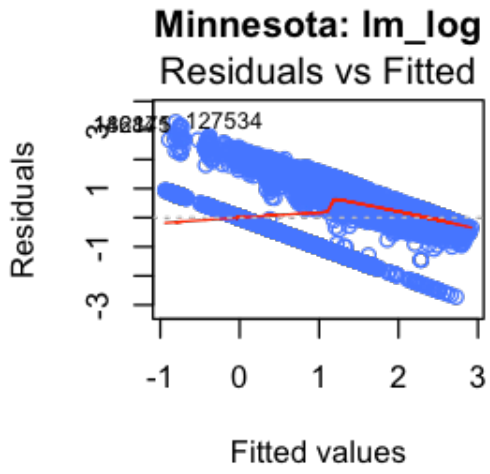
In hopes to fix these issues we applied the square-root, logorithm, and inverse transformations to fix the nonconstant variance, linearity violation, and nonnormality. Below we showed plots corresponding to each transformation and the change the transformation has done.

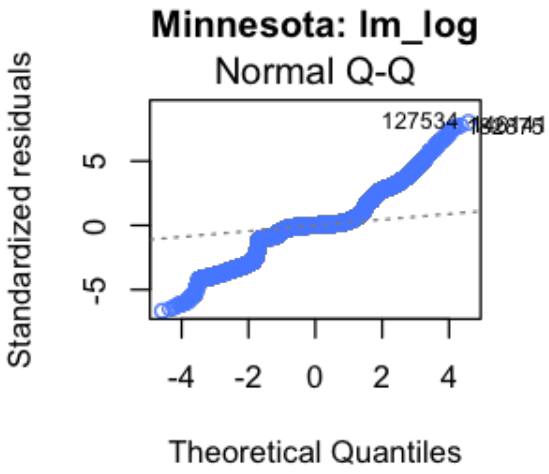

Minnesota: lm_sqrt
Residuals vs Fitted
Residuals
Fitted values
-200
0
200
400
-400
0
400

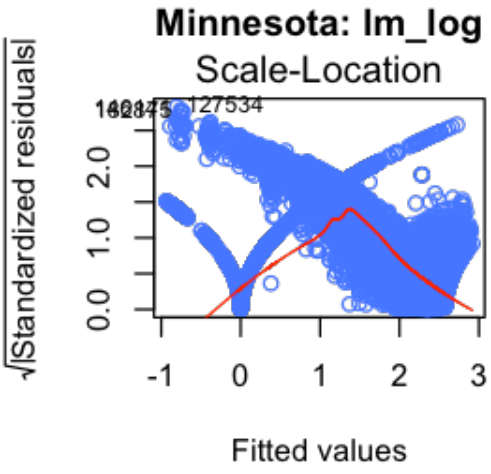

Minnesota: lm_sqrt
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-4
-2
0
2
4
-5
0
5
10

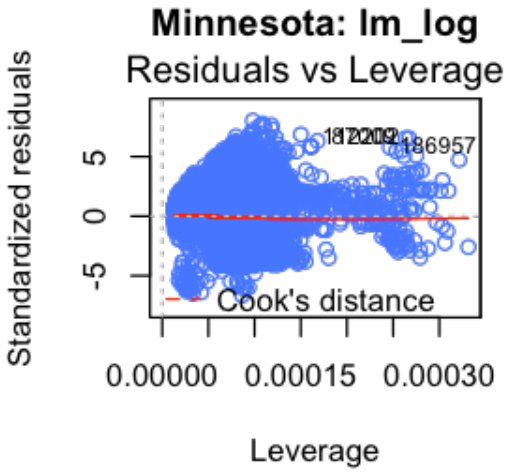

Minnesota: lm_sqrt
Scale-Location
√|Standardized residuals|
Fitted values
-200
0
200
400
0.0
1.5
3.0

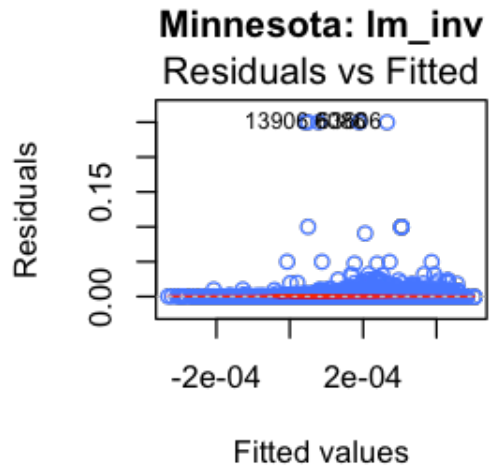

Minnesota: lm_sqrt
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.00000
0.00015
-5
0
5
10

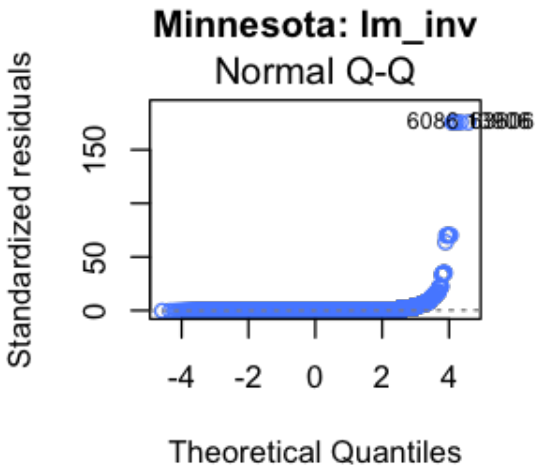

Minnesota: lm_log
Residuals vs Fitted
Residuals
Fitted values
127534

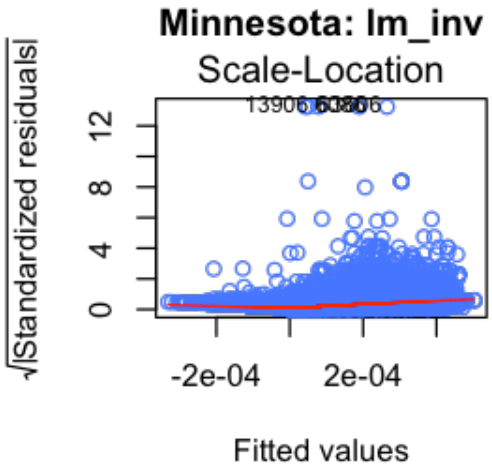

Minnesota: lm_log
Normal Q-Q
Standardized residuals
Theoretical Quantiles
127534

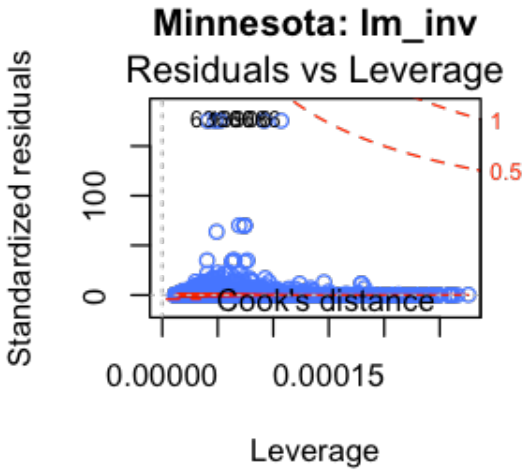

Minnesota: lm_log
Scale-Location
√|Standardized residuals|
Fitted values
127534

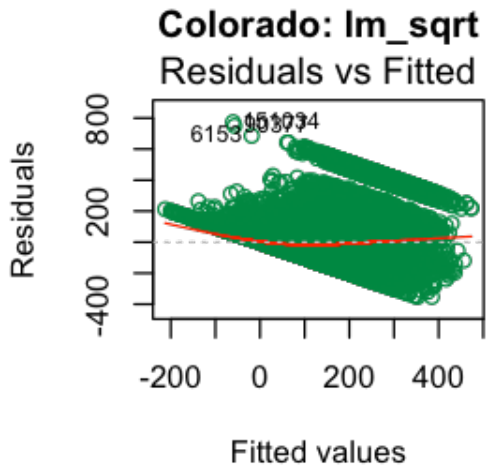

Minnesota: lm_log
Residuals vs Leverage
Standardized residuals
Cook's distance
186957
Leverage

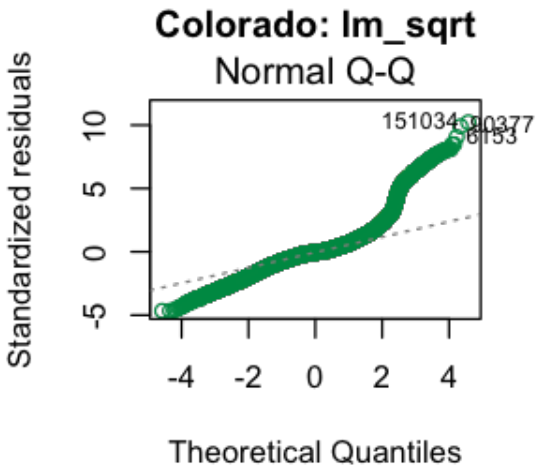

Minnesota: lm_inv
Residuals vs Fitted
Residuals
0.00
0.15
-2e-04
2e-04
Fitted values

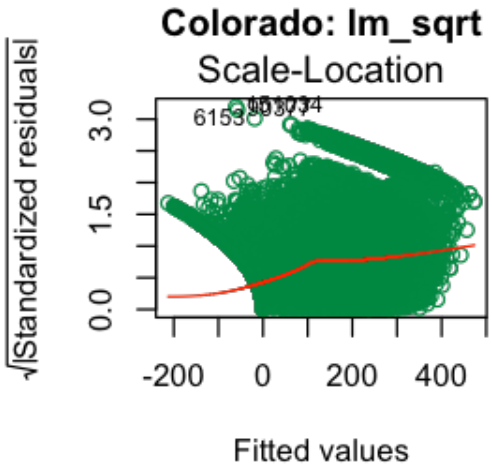

Minnesota: lm_inv
Normal Q-Q
Standardized residuals
0
50
150
-4
-2
0
2
4
Theoretical Quantiles

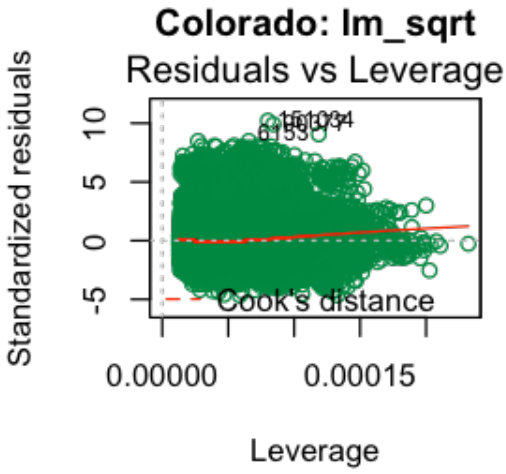

Minnesota: lm_inv
Scale-Location
√|Standardized residuals|
0
4
8
12
-2e-04
2e-04
Fitted values

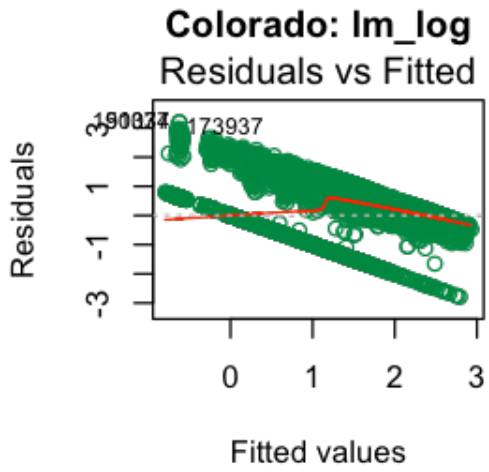

Minnesota: lm_inv
Residuals vs Leverage
Standardized residuals
0
100
1
0.5
Cook's distance
0.00000
0.00015
Leverage

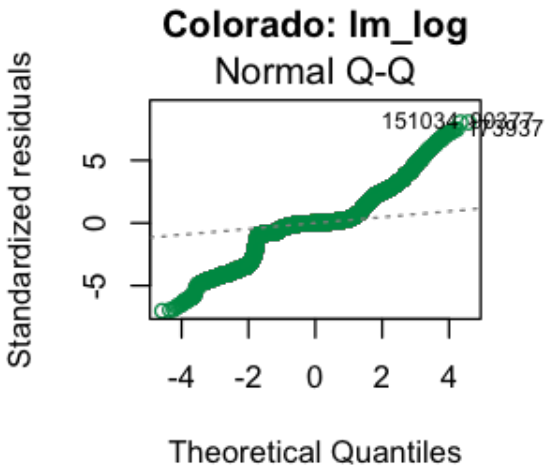

Colorado: lm_sqrt
Residuals vs Fitted
Residuals
Fitted values
-400
200
800
-200
0
200
400

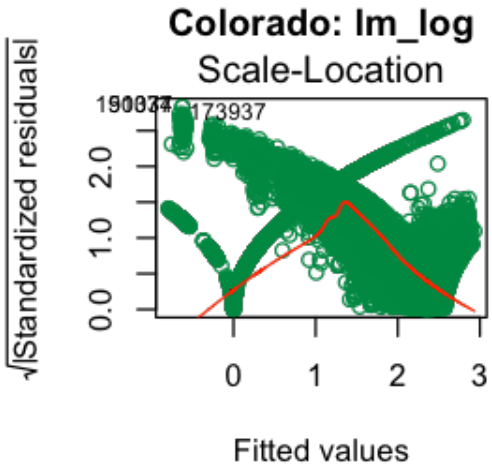

Colorado: lm_sqrt
Normal Q-Q
Standardized residuals
Theoretical Quantiles
151034
-5
0
5
10
-4
-2
0
2
4

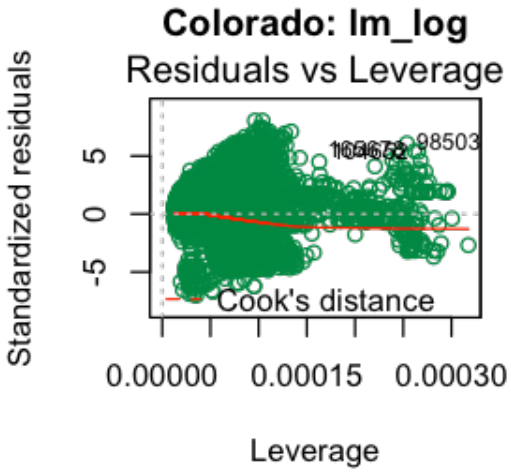

Colorado: lm_sqrt
Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.5
3.0
-200
0
200
400

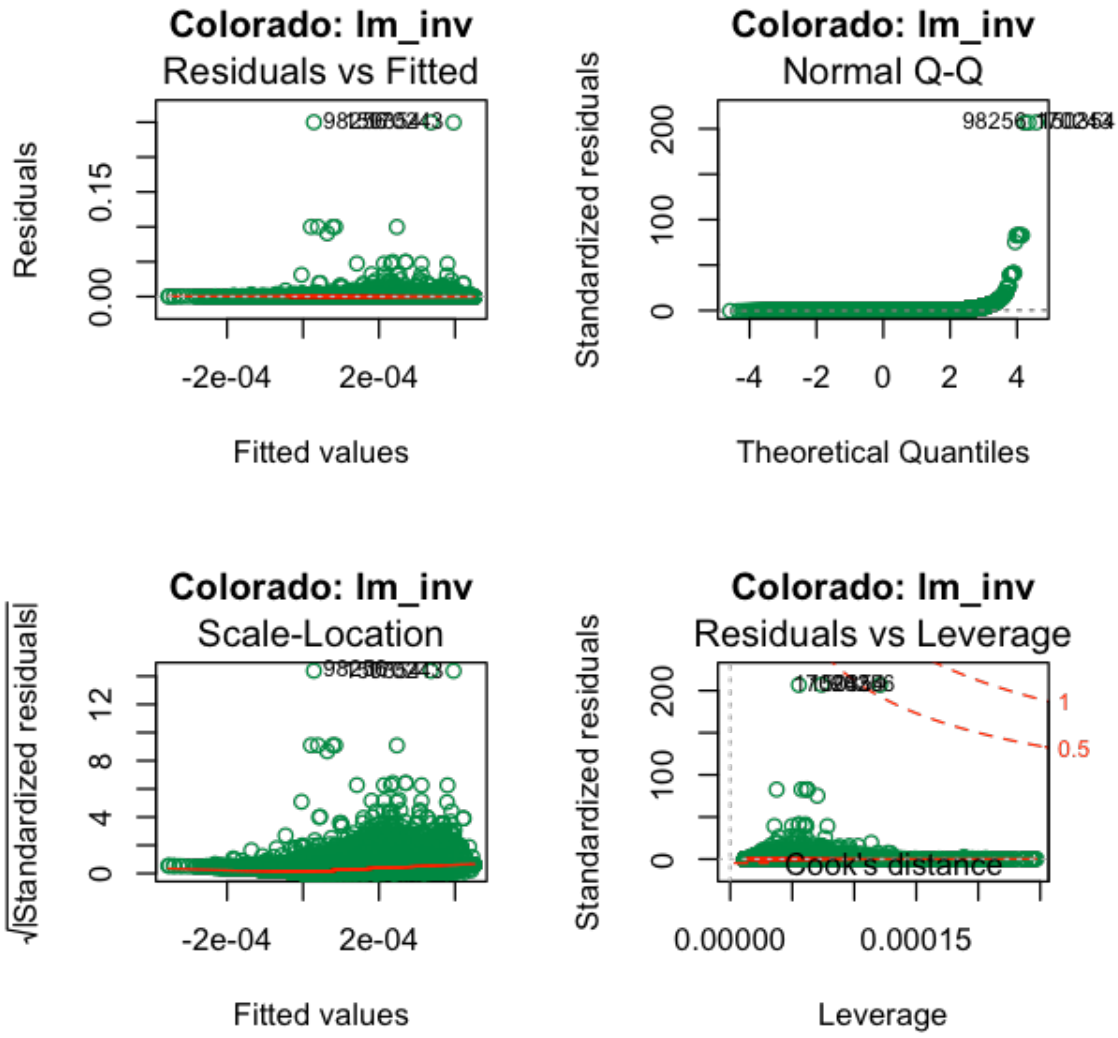

Colorado: lm_sqrt
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
-5
0
5
10
0.00000
0.00015


Colorado: lm_log
Residuals vs Fitted
Residuals
Fitted values
173937


Colorado: lm_log
Normal Q-Q
Standardized residuals
Theoretical Quantiles


Colorado: lm_log
Scale-Location
√|Standardized residuals|
Fitted values
173937


Colorado: lm_log
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



Unfortunately, the square root and logorithmic transformations faired no better than our original linear models. Both states still showed violations in variance and linearity. The Normal Q-Q has a random string structure that shows nonnormality and the Residuals vs Leverage shows no potential influential points or leverage points. Sadly, the first two transformation didn't improve our model, however, the inverse transformation was more promising.

First the Residuals vs Fitted in both states show more consitency around 0 with a few outliers above 12. Colorado's Residuals vs Fitted is more consistent in variance with the majority of the points across 0. Minnesota's Resiudal vs Fitted has a similar plot with points around 0. Both the plots show us that the points do not violate constanst variance, however, we still need to do further research to conclude to that decision.

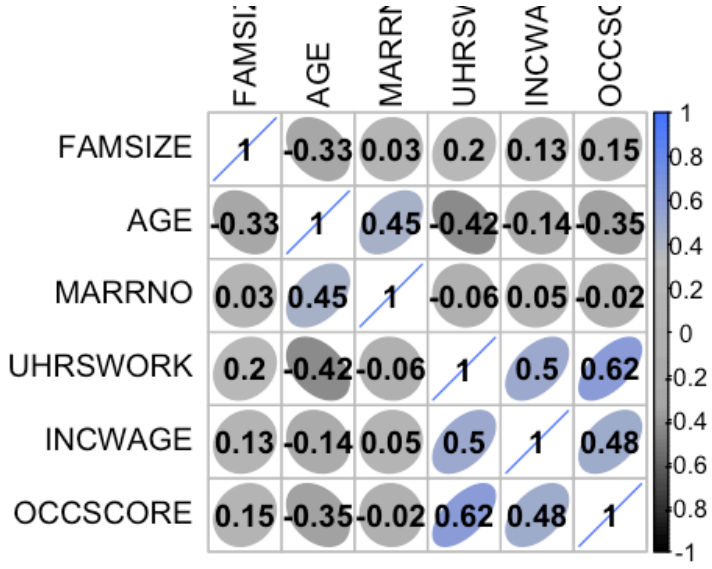
The Normal Q-Q plot for the inverse transformation is a flat line with all the points aligning on the line for Colorado and Minnesota. The points tail off in the end showing some nonnormality at the

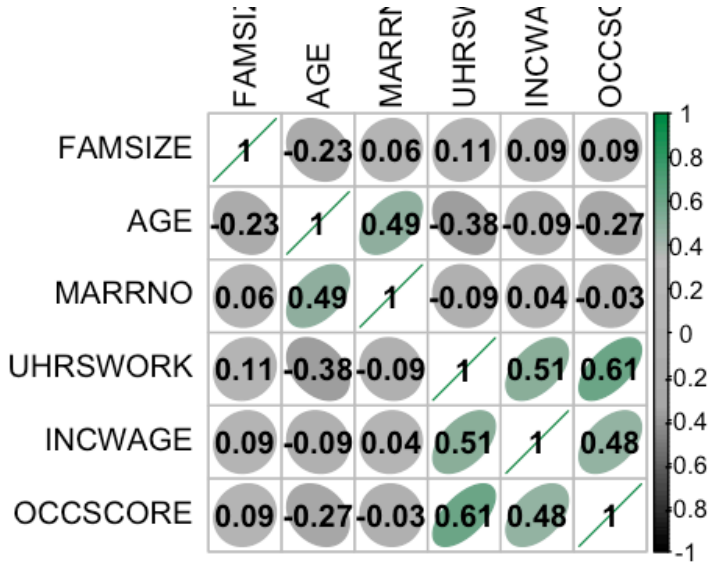
end of the points in both states. However, we can not conclude if it is or isn't violating linearity assumption becuase the linear line is along 0. There is no increase or decrease in the linear line. Hence, more research need to be made.

The Residual vs Leverage plot in both states finally show us some points that can possibly be potential influential points. Most of the points lie around 0 in Colorado and Minnesota, though, Colorado has a potential influential point and a handful of leverage pooints. Minnesota has no potential influential points, but a few leverage points. These graphs resembles more of our exploratory analysis where **occupational score, income, and hours worked** had more outliers than the other variables.

Overall, we decided to move forward with our selected linear model since the inverse model showed less issues.

3 Checking for Multicollinearity





We have plotted a correlation plot on numeric values, where nothing too alarming that sugget there is some multicollinearity here. Notice in both plots with respect to INCWAGE that UHRSWORK is approximately 0.50, OCCSCORE is approximately 0.48. There is noticeable negative correlations but nothing that is alarming.

Though our model presented evidence that each predictor was necessary we wanted to make sure that there was no multicollinearity. However, we would assume that variables like **hours worked per work vs occupational score, age vs marriage, and age vs hours worked per week** would show some correlation.

3.1 Minnesota

`VIF for the Minnesota dataset is:`

	GVIF	Df	GVIF^(1/(2*Df))
SEX	1.072358	1	1.035547
AGE	1.785806	1	1.336341
MARRNO	1.304837	1	1.142295
EDUC	1.110222	1	1.053671
CLASSWKR	3.546338	2	1.372287
UHRSWORK	2.033330	1	1.425949
OCCSCORE	2.841562	1	1.685693

`The condition number for the Minnesota dataset is:`

291.427390215319

`Correlation Model Matrix for the Minnesota dataset is:`

	SEXMale	AGE	MARRNO	EDUCHigh School or Less	CLASS
SEXMale	1.000000000	-0.03672824	-0.059084758	0.053602325	0.1231
AGE	-0.036728239	1.00000000	0.448833756	0.140047723	0.0405
MARRNO	-0.059084758	0.44883376	1.000000000	-0.009269013	0.0621
EDUCHigh School or Less	0.053602325	0.14004772	-0.009269013	1.000000000	-0.004
CLASSWKRSelf-employed	0.123137889	0.04051634	0.062142192	-0.004130447	1.0000
CLASSWKRWorks for wages	-0.003484196	-0.47078507	-0.138822334	-0.170480644	-0.492
UHRSWORK	0.174927503	-0.42109493	-0.061212850	-0.193432496	0.1628
OCCSCORE	0.144203411	-0.34897610	-0.024518222	-0.283487525	0.1468

It was questionable that the VIF for each predictor was less than 5, however, the condition number was very high. The condition number suggested that there was evidence of multicollinearity. Based on the correlation table, notice that some variables were partially correlated such as class of work vs age, age vs married, hours worked per week vs work wage and more. Despite that, most of the correlation values were equal or below |0.50|, hence correlation was not as evident. This was shocking since one would expect strong correlation between these variables.

3.2 Colorado

`VIF for the Colorado dataset is:`

	GVIF	Df	GVIF^(1/(2*Df))
SEX	1.062217	1	1.030639
AGE	1.712289	1	1.308545
MARRNO	1.350457	1	1.162092
EDUC	1.088235	1	1.043185
CLASSWKR	3.419890	2	1.359888
UHRSWORK	2.021369	1	1.421748
OCCSCORE	2.805418	1	1.674938

`The condition number for the Colorado dataset is:`

282.154587229977

`Correlation Model Matrix for the Colorado dataset is:`

	SEXMale	AGE	MARRNO	EDUCHigh School or Less	CLASS
SEXMale	1.00000000	-0.050745503	-0.06751557	0.025823138	0.06575
AGE	-0.05074550	1.000000000	0.48961612	0.007703764	0.06291
MARRNO	-0.06751557	0.489616117	1.00000000	-0.017197877	0.07357
EDUCHigh School or Less	0.02582314	0.007703764	-0.01719788	1.000000000	-0.0482
CLASSWKRSelf-employed	0.06575163	0.062913495	0.07357839	-0.048263930	1.00000
CLASSWKRWorks for wages	0.04968261	-0.429024011	-0.15066917	-0.108146545	-0.4894
UHRSWORK	0.18846830	-0.378788267	-0.08972016	-0.153898856	0.09716
OCCSCORE	0.15770840	-0.269854734	-0.02681930	-0.260625883	0.17754

Similar to Minnesota's analysis, the VIF for each predictor was below 5 and the condition number was significantly high. Next, the correlation table had the same predictors being correlated as in Minnesota with their correlation value to be less than or equal to |0.50|.

To convince ourselves, we decided to check how the VIF and condition number would change if we removed the predictor with the highest VIF. Below is our process of reducing multicollinearity.

3.3 Removing Predictor with the Highest VIF

From our table above we decided to remove **CLASSWKR** from our linear model because it had the highest VIF. Below was we see the output of our decision.

3.4 Minnesota

`VIF after removing CLASSWKR for Minnesota dataset is:`

SEX 1.05332594319139 **AGE** 1.60465721589475 **MARRNO** 1.30405303915097 **EDUC** 1.10238780597687 **UHRSWORK** 1.80196215116246 **OCCSCORE** 1.74213854233254

`The updated condition number for the Minnesota dataset is:`

176.418442090851

The VIF for the remaining predictors decreased to roughly under 2, however, the important value was the conditioning number. The condition number decreased by more than 100 which meant that removing class of work reduced some multicollinearity.

3.5 Colorado

`VIF after removing CLASSWKR for Colorado dataset is:`

SEX 1.05224820406581 **AGE** 1.5677527686008 **MARRNO** 1.34901968610419 **EDUC** 1.08374638042191 **UHRSWORK** 1.7785206560782 **OCCSCORE** 1.70899020018738

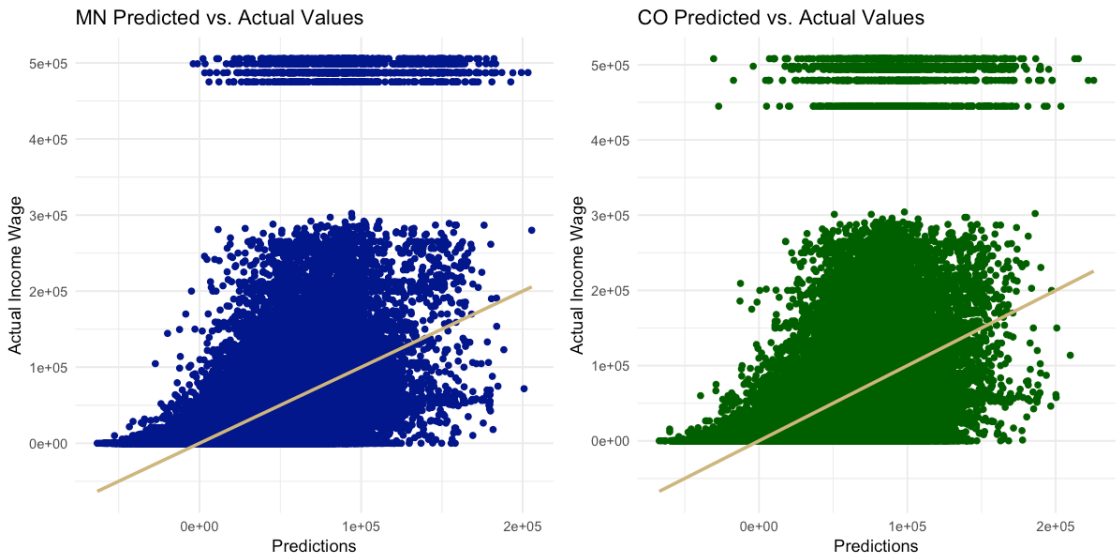
`The updated condition number for the Colorado dataset is:`

167.28610121578

Similarly with Colorado, the VIF of the remaining predictors decreased to be less than 2 and, again, the conditioning number decreased by more than 100. However, we tried removing the next highest VIF value associated with UHRSWORK, the VIF seemed to increased. Thus, we decided to only leave out **CLASSWKR**

3.6 Prediction Plot

From all our analysis, statistical inferences, linear regression model assumption, predictor selection, and VIF and conditioning number checks we decided to move forward with our linear model and check how well it predicted. The predictors we concluded with were **sex, age, married, education, hours worked per week, and occupation score**. We decided not to remove **class of work** since its initial VIF was below 5. We also removed **race** because of the misinterpration in our earlier linear model. We ran the model on the full dataset and not a test dataset to verify its completeness.



Minnesota's and Colorado's prediction plots had a similar resemblance with many points around the central region of the linear plot. The plots displayed our prediction vs the actual income value. Our linear model crossed through a collection of points, therefore calculating most of the values correctly. What made the dataset particularly difficult to analyze was simply the sheer amount of data and the intricate nature of analyzing human behavior. Creating a linear model in the sociosciences can be complicated due to the amount of convulated data there is. Assumptions have huge role in social sciences, because statistically we must be aware of our limitations of our data and understand the right assumption to be conditioned on. Socially, we cant assume some one's income based on race and occupation. The extent to which assumptions are falsifiable when you are conditioning on variables affected by race, family size, and occupation is very large. Our linear model could not pick-up on every association or correlation, however, we were confident that this prediction plot had done a fair job. Also, there was a lot of underlying and missing data which making our analysis incomplete.

Continuing on to our analysis of the plot, we could see that both plots hd a large amount of outlier points near $500,000$, however, it was not enough to skew our linear model. The reason for this could be that there was so much data between the values of 0 to $300,000$ that the amount of outliers were miniscule compared to the amount of data there was in the center.

3.7 Conclusion

For this study, we wanted to be able to predict an individual's income by applying the statistical techniques and tools that we have learned from this course. Based on our analysis, the results of our predictions didn't turn out as we had hoped. Assuming that the model was correct, we found that **sex, age, married, education, hours worked per week, and occupation score** were significant in both states. This was especially interesting; before performing regression, we assumed that race and the class of workers would be significant. As we stated earlier, this could a consequence of underlying and missing data possibly impacting our prediction model at the end. Additionally, we noticed many similarities between the two states as we performed our analysis, which came to a surprise. Our original speculation was that there would be minor differences

in predictors that would affect one's income. This is because there are minor differences in an individual's income, being that the median income reported for Minnesota is $\$38,881$ compared to Colorado's, $\$39,545$ [2].

Overall, we learned that data in the social sciences can be difficult to work with and that a robust statistical technique may be needed to be done in order to run a more in-depth and accurate analysis of this study. Another issue that we ran into is our dataset was very large, when running through our code, it took some time to compile and often crashed our program. Although the results were not what we have anticipated, this project gave us the opportunity to put building predictive models into practice and apply these techniques and tools to real-world datasets.

If we had the option to extend this study, we could amend our response to be binary in order to run logistic regression. For the future research, we'd suggest predicting an adult individual who makes more than a fixed income, i.e creating a new variable that includes a binary response, which could be the threshold we set our income to be.

4 Sources

[1] IPUMS USA. "U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH." Retrieved from https://usa.ipums.org/usa/.

[2] U.S Census Bureau. "QuickFacts." Retrieved from https://www.census.gov/quickfacts/.