A Predictive Analysis of Income in the United States

Kaitlyn Hohmeier

STT 450, Fall 2020

Advisor: Dr. Cuixian Chen

Abstract: The factors that affect income levels in the United States are numerous and complex. Over the course of this project, we examined four general categories of explanatory variables—gender, race, type of employment, and poverty levels—and analyzed their significance in contributing to incomes in a particular region. To accomplish this, we utilized data from the 2015 American Community Survey, conducted by the U.S. Census Bureau, which divides the United States into regions called census tracts. We studied a number of regression models, in particular non-linear models, and classifiers, such as logistic regression, linear and quadratic discriminant analysis, bagging, and boosting. We also implemented 5-fold cross validation both for regression and classification procedures to help us develop our models. Our findings are summarized and demonstrated with a variety of graphical and numerical analyses and aids. Ultimately, we wish to obtain regression and classification models to effectively predict and analyze income levels in the United States. We did in fact develop some successful models based on our dataset that could effectively predict income and income levels in different census tracts.

Keywords: income, income levels, census data, poverty levels, regression modeling, classification analysis, predictive analysis.

1. Introduction

Our goal is to examine and measure multiple demographic factors that influence income level. We wish to formulate a model that can accurately predict income level based on characteristics such as gender, race and ethnicity, type of employment, region of the United States,

and poverty levels in the region. By examining a variety of socioeconomic variables, we will see what factors relate most strongly to economic status and income level and provide a holistic view of income levels in the United States.

Variable Name	*Description*	*Variable Type*
Income	median household income (the response variable)	Numerical
State	State, Washington, D.C., or Puerto Rico	Categorical
County	county or county equivalent, with each county divided into multiple census districts	Categorical
Men	number of men in the census tract	Numerical
Women	number of women in the census tract	Numerical
Hispanic	percentage of the population of the census tract that is Hispanic/Latino	Numerical
White	percentage of the population of the census tract that is white/Caucasian	Numerical
Black	percentage of the population of the census tract that is black/African-American	Numerical
Native	percentage of the population of the census tract that is Native American/Native Alaskan	Numerical
Asian	percentage of the population of the census tract that is Asian	Numerical
Pacific	percentage of the population of the census tract that is Native Hawaiian or Pacific Islander	Numerical
IncPerCap	income per capita	Numerical
Poverty	percentage of the population of the census tract under the national poverty level	Numerical
ChildPov	percentage of the children in the census tract under the national poverty level	Numerical
Professional	percentage of the population of the census tract employed in management, business, science, and arts	Numerical
Service	percentage of the population of the census tract employed in service jobs	Numerical
Office	percentage of the population of the census tract employed in sales and office jobs	Numerical
Construction	percentage of the population of the census tract employed in natural resources, construction, and maintenance	Numerical
Production	percentage of the population employed of the census tract in production, transportation, and material movement	Numerical

Table 1.1. List of variables to be included in this research project.

The dataset to be used for this research was originally obtained from US Census data via the 2015 American Community Survey 5-year estimates and subsequently compiled and organized by a machine learning engineer who posted the data set to Kaggle (reference [1]). It contains comprehensive income and demographic information on every county in every state in the United States. Furthermore, each county is split into additional regions called "census tracts," which are determined by the U.S. Census Bureau. The original sample size of this dataset, prior to any attempts to clean up the data, was 74,000 observations. After performing some initial exploratory data analysis, including omitting any rows with missing data, a new sample size of 72,727 complete observations was obtained. In total, the dataset contains 37 variables; however, for the purposes of this project, not all of the variables will be used, as we wish to examine only those variables that are most likely to significantly affect income level.

The above table (Table 1.1) provides descriptions for the variables chosen initially to focus on for exploratory data analysis. Once we proceed to model selection in the next section, not all of these variables may be used in our modeling.

2. Exploratory Data Analysis

Because the goal of this project is to create a statistical model that predicts income levels based on a number of explanatory variables, this project will involve supervised statistical learning. For this data set, the response variable will be income—a numerical variable—and it is included in this dataset. The research performed in this project is regression predictive modeling, since we seek to predict a numerical response variable.

Numerical and graphical analyses were performed on the numerical and categorical variables. These graphs and results are documented in the following sections.

2.1 Numerical Analysis and Graphical Analysis of Numerical Variables

In this section, we will describe the numerical and graphical analyses performed on the numerical variables in this data set.

	N		Mean	SD		Min	Q1	Median	Q3	Max
Income	72727		57259.14	28663.97		2611	37716.0	51106.0	70147.0	248750.0
IncPerCap	72727		28520.77	14831.78		1188	19192.0	25372.0	33901.0	254204.0
Poverty	72727		16.89	12.95		0	7.3	13.4	23.1	98.6
ChildPov	72727		22.47	19.11		0	7.0	17.8	33.7	100.0
Professional	72727		34.80	14.91		0	24.1	32.6	43.8	100.0
Service	72727		19.08	8.14		0	13.5	17.9	23.5	74.1
Office	72727		23.93	5.72		0	20.2	23.7	27.5	74.4
Construction	72727		9.31	5.94		0	5.0	8.4	12.5	69.3
Production	72727		12.88	7.57		0	7.1	11.8	17.5	60.0

Table 2.1. First part of five-number summaries for numerical variables in Income dataset.

	N		Mean	SD		Min	Q1	Median	Q3	Max
Men	72727		2153.86	1050.08		16	1434.0	2002.0	2685.5	27962.0
Women	72727		2229.70	1072.69		25	1490.0	2085.0	2788.0	27250.0
Hispanic	72727		16.87	22.94		0	2.4	7.0	20.5	100.0
White	72727		62.06	30.68		0	39.4	71.4	88.4	100.0
Black	72727		13.24	21.75		0	0.7	3.7	14.3	100.0
Native	72727		0.72	4.46		0	0.0	0.0	0.4	100.0
Asian	72727		4.59	8.79		0	0.2	1.4	4.8	91.3
Pacific	72727		0.14	1.03		0	0.0	0.0	0.0	84.7

Table 2.2. Second part of five-number summaries for numerical variables in Income dataset.

As the first step in exploratory data analysis on the numerical variables, we examined the five-number summaries for the numerical variables we are most likely to use in our modeling. These results are depicted in Table 2.1 and 2.2. From this numerical summary, we can see that a number of the variables appear to be skewed (for example, income appears highly right-skewed). When we proceed to model selection in the next section after completing exploratory data analysis on all of the variables, we may have to consider a transformation, such as a log() transformation, on some of these variables, in order to account for this skewedness.

Next, we looked at the correlation coefficients between the numerical variables.

	Income	IncPerCap	Poverty	ChildPov	Professional	Service	Office	Construction	Production	Men	Women
Income	1	0.835569	-0.70358	-0.66388	0.7334454	-0.58811	-0.06758	-0.3329961	-0.499943	0.176304	0.166493
IncPerCap	0.835569	1	-0.61277	-0.58988	0.8052115	-0.58667	-0.10701	-0.4111344	-0.551712	0.025686	0.038203
Poverty	-0.70358	-0.61277	1	0.898781	-0.5493317	0.59245	-0.02795	0.17110565	0.3319685	-0.15366	-0.14457
ChildPov	-0.66388	-0.58988	0.898781	1	-0.5741526	0.558358	-0.01406	0.21123591	0.375497	-0.14983	-0.13763
Professional	0.733445	0.805212	-0.54933	-0.57415	1	-0.65858	-0.1858	-0.547989	-0.691084	0.067048	0.083962
Service	-0.58811	-0.58667	0.59245	0.558358	-0.6585836	1	-0.09361	0.1113866	0.2056465	-0.11078	-0.10476
Office	-0.06758	-0.10701	-0.02795	-0.01406	-0.1857966	-0.09361	1	-0.1865026	-0.142384	0.04269	0.094632
Construction	-0.333	-0.41113	0.171106	0.211236	-0.547989	0.111387	-0.1865	1	0.3153237	0.021618	-0.04779
Production	-0.49994	-0.55171	0.331969	0.375497	-0.6910842	0.205647	-0.14238	0.31532372	1	-0.06226	-0.08677
Men	0.176304	0.025686	-0.15366	-0.14983	0.0670482	-0.11078	0.04269	0.02161834	-0.062258	1	0.934995
Women	0.166493	0.038203	-0.14457	-0.13763	0.0839622	-0.10476	0.094632	-0.047787	-0.086767	0.934995	1
Hispanic	-0.22805	-0.30878	0.350066	0.323598	-0.3356928	0.290102	0.007189	0.26913928	0.1326557	0.116731	0.098213
White	0.314785	0.381268	-0.52974	-0.49859	0.3514853	-0.47473	-0.02964	-0.0408594	-0.12759	-0.0181	-0.03875
Black	-0.31053	-0.28388	0.411073	0.413108	-0.2473963	0.37842	0.056579	-0.1309832	0.1406315	-0.13479	-0.08263
Native	-0.07208	-0.07525	0.08693	0.071659	-0.0433884	0.052787	-0.04635	0.07203366	0.0071985	-0.02984	-0.03821
Asian	0.282894	0.208129	-0.12345	-0.15717	0.2674567	-0.08115	-0.03727	-0.2405928	-0.222471	0.100555	0.09886
Pacific	0.007838	-0.02641	0.00801	0.002886	-0.0347876	0.054136	0.021407	0.00603	-0.010594	0.029251	0.020499

Table 2.3. A subset of the correlation coefficients obtained for each variable.

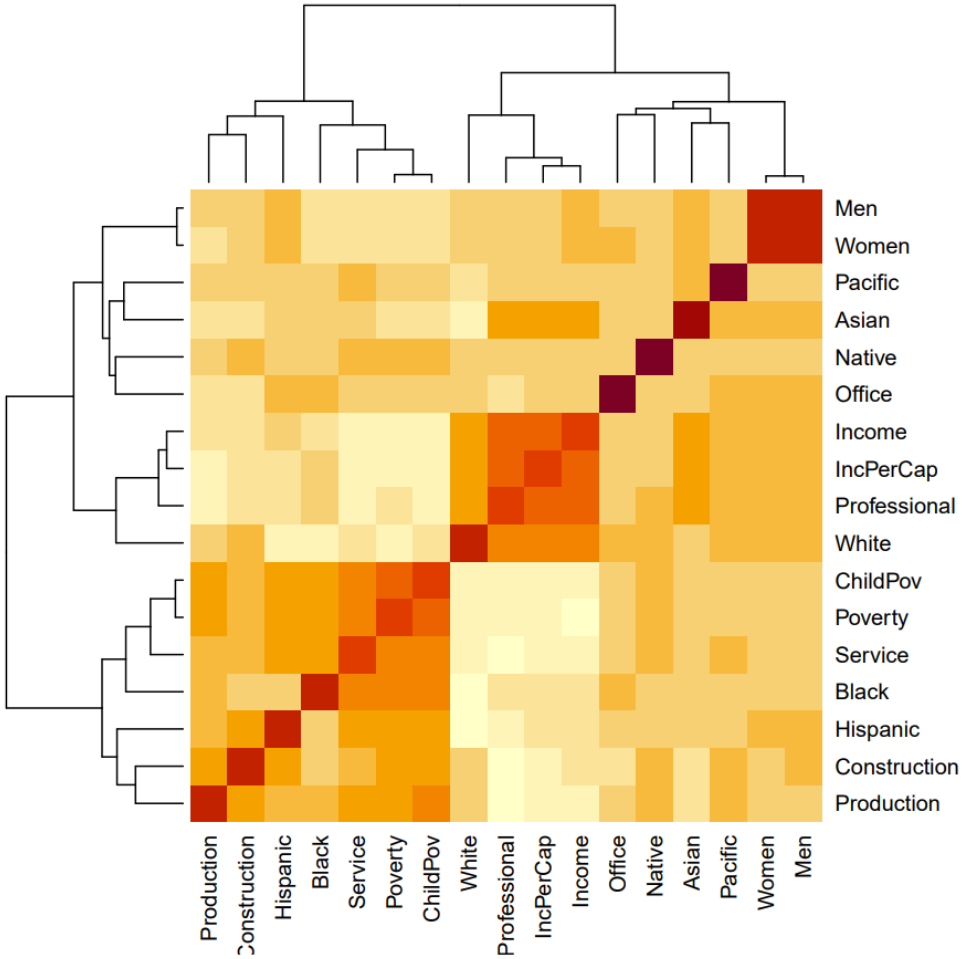

Figure 2.1. Heat Map of Numerical Variables.

Our last step in the numerical analysis was to create a correlation coefficient matrix for the variables, which subsequently lead us to our first graphical analysis of the numerical variables via a heat map of the correlation coefficients for each variable (shown in Figure 2.1). From the correlation coefficients and the heat map, a number of interesting relationships among the variables emerged. A subset of some the most noteworthy correlation coefficients are presented in Table 2.3; the omitted correlation coefficients from Table 2.3 are the coefficients between the race and ethnicity variables, which do not have any meaning or significance in this data analysis.

Concentrating first on the relationship between income and race and ethnicity, there appeared to be some weak correlations between the percentage of each racial group in a census tract and the median income within that census tract. The strongest of these associations could be seen within the Caucasian and African American groups. There was a weak to moderate positive relationship ($r = 0.3148$) between median income and the percentage of Caucasians in a census tract. This correlation coefficient increased to $r = 0.3813$ when looking at income per capita. There was a weak to moderate negative relationship ($r = -0.3105$) between median income and the percentage of African-Americans in a census tract. When examining income per capita, an additional weak to moderate negative relationship between income per capita and percentage of Hispanics in a census tract emerged.

Looking next at the relationship between gender and income, the correlation coefficient for median income versus percentage of men in a census tract was $r = 0.1763$, and similarly for women, $r = 0.1665$. This suggests that there did not appear to be any strong correlation between gender and income.

Type of employment also appeared to have a strong relationship to income. Each type of employment is explained and described in Table 1. For example, the correlation coefficient for the

relationship between median income and the percentage of individuals employed in a professional field was r = 0.7334, which indicated a strong positive relationship. The coefficient increased to r = 0.8052 when looking at income per capita. On the other hand, service employment appear to have a moderately strong negative relationship to median income (r = -0.5881), while production employment had a moderate negative relationship to median income (r = -0.4999), and construction employment had a weak to moderate negative relationship to median income (r = -0.3329). Multicollinearity is a potential concern, because there may be a relationship between type of employment and some other characteristic, such as race. For example, there is a positive weak to moderate relationship between the percentage of African-Americans in a census tract and the percentage of individuals employed in a service industry (r = 0.3784), while there is a negative moderate relationship between employment in a service industry and the percentage of Caucasians in a census tract (r = -0.4747). The occupation and race variables will need to be investigated further to determine if multicollinearity is playing a role.

Perhaps unsurprisingly, a strong negative relationship between median income and the percentage of individuals in the census tract living in poverty was indicated in the correlation coefficient, as well as the strong negative relationship between median income and the percentage of children living in poverty. A closer examination also yielded several noteworthy relationships in the data. First, there was a moderate negative relationship between percentage of Caucasians living in a census tract and percentage of individuals living in poverty—that is, a higher percentage of Caucasians in a census tract correlates to a lower percentage of individuals in poverty. On the other hand, certain minority groups, in particular African-Americans, have a positive correlation with poverty, meaning that a higher percentage of African-Americans in a census tract correlates to a higher percentage of individuals in poverty in the census tract.

In the next stage of graphical analysis, we will examine boxplots of each numerical variable.

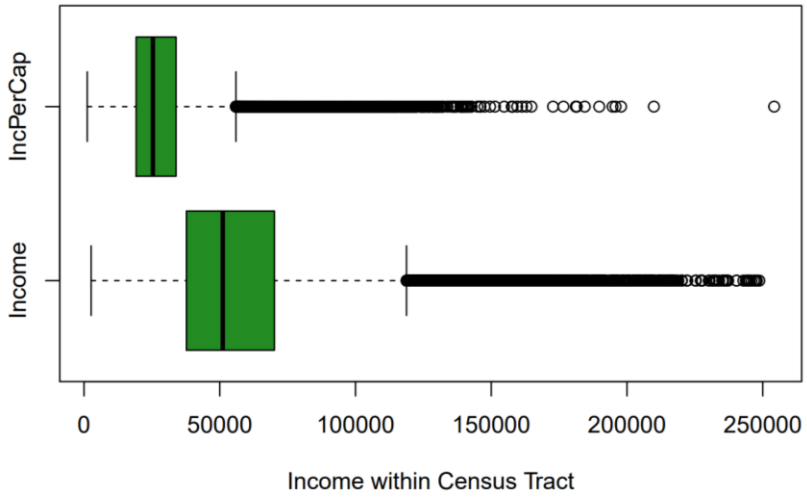

Figure 2.2. Boxplots of income and income per capita within each census tract.

Figure 2.2 depicts boxplots of median income and income per capita within each census tract. From this, we can see that the median income within each census tracts is around fifty thousand dollars ($50,000), a result also obtained from the numerical analysis performed earlier. This suggests that the median U.S. income is around fifty thousand dollars ($50,000). There is significant variation in median income and income per capita in the United States, as can be seen from the spread of the data. The outliers in the plots indicate that there are many highly affluent areas in the U.S. where median income and income per capita is much higher than that of most of the United States. This suggests that the data is right-skewed, as observed earlier in the numerical analysis of these variables.

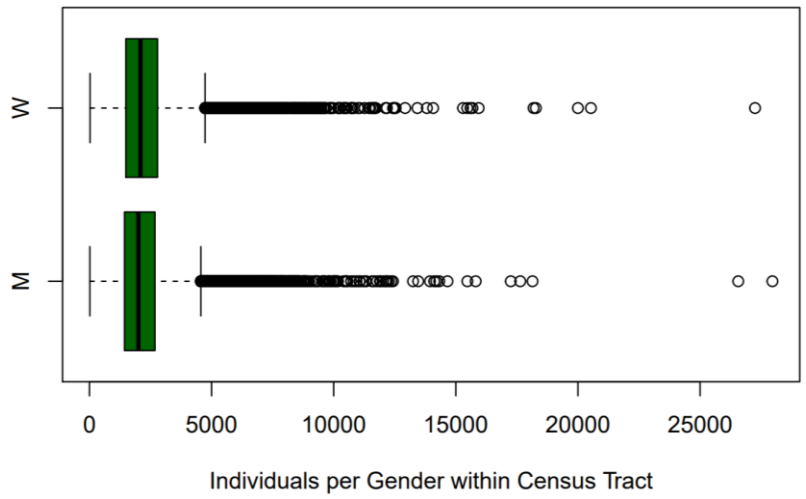

Figure 2.3. Boxplots of number of individuals per gender within each census tract.

The boxplots shown in Figure 2.3 depict the number of men (labeled "M") and women (labeled "W") within each census tract. The median numbers of men and women in each census tract is roughly equal, though there appear to be slightly more women than men, on average, within each census tract. This result suggests that in the U.S. population as a whole, there appear to be slightly more women than men in the population. The data appears to be right-skewed.

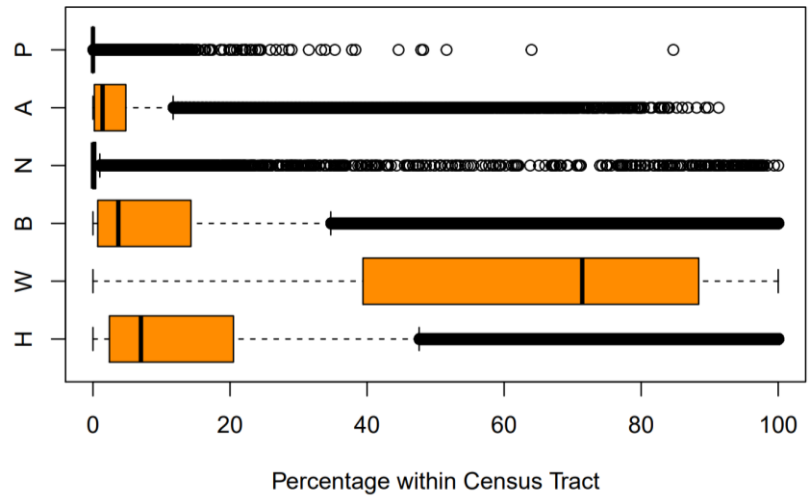

Figure 2.4. Boxplots of percentage of each ethnic/racial group within each census tract.

Next, we examine boxplots of the racial and ethnic breakdown of the census tracts, from which we can draw some conclusions about the racial demographics of the United States (see

Figure 2.4). From these boxplots, we can see that Caucasians (labeled “W” on the boxplot) are overwhelmingly the predominant racial demographic within each census tract, and hence in the U.S. population as a whole. The largest minority group is Hispanics (labeled “H”), followed by African-Americans (labeled “B”) and Asians (labeled “A”). Native people groups (labeled “N” in the boxplot) and Pacific Islander (labeled “P”) form the smallest racial demographics in the United States. The outliers in the data indicate that although whites are the largest racial population in the U.S., there are many census tracts and regions where overall minority populations are the majority racial demographic. In addition, the population percentages for the five minority groups appear to be right-skewed, while the distribution for the population percentage of Caucasians looks slightly left-skewed.

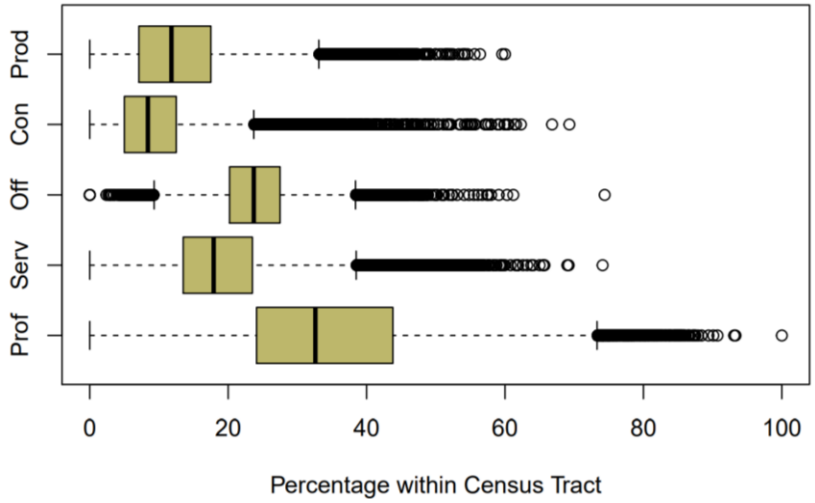

Figure 2.5. Boxplots of percentage of each employment classification within each census tract.

From Figure 2.5, we can see that professional employment (labeled “Prof” in the boxplot) has the highest median percentage of all the other employment types, while the construction industry (labeled “Con”) has the smallest percentage employed. The medians of the percentage of individuals employed in the service industry (“Serv”), the production industry (“Prod”), and office employment (“Off”) fall between these other two categories. We can see that there are a number

of outliers in the boxplots. For the production, construction, service, and professional industries, we see that these distributions may be right-skewed, though the distribution for the professional industry appears to be the closest to a Normal distribution of these four. The office industry has outliers on both sides of the lines for its boxplot, though this distribution appears to be right-skewed as well.

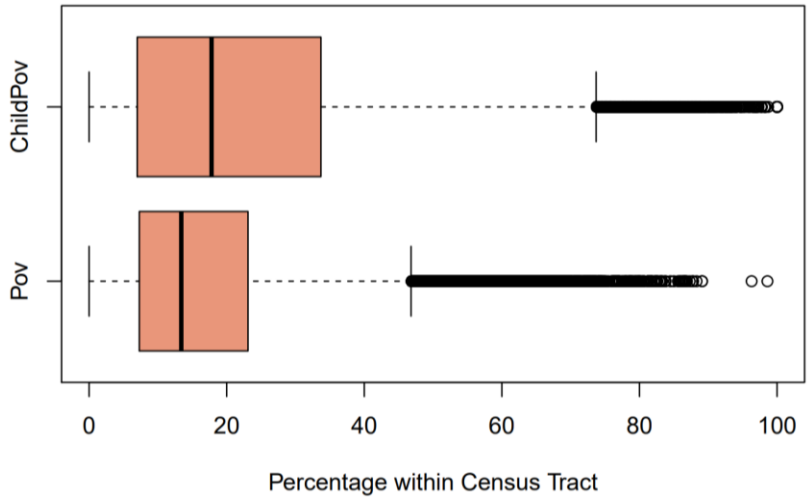

Figure 2.6. Boxplots of the percentages of poverty and child poverty within each census tract.

The overall percentage of individuals in poverty in each census tract and the percentage of children in poverty in each census tract are depicted in the boxplot in Figure 2.6. Here, we see that both distributions look right-skewed. Furthermore, the median percentage of children in poverty is greater than the median percentage of individuals in general who are in poverty.

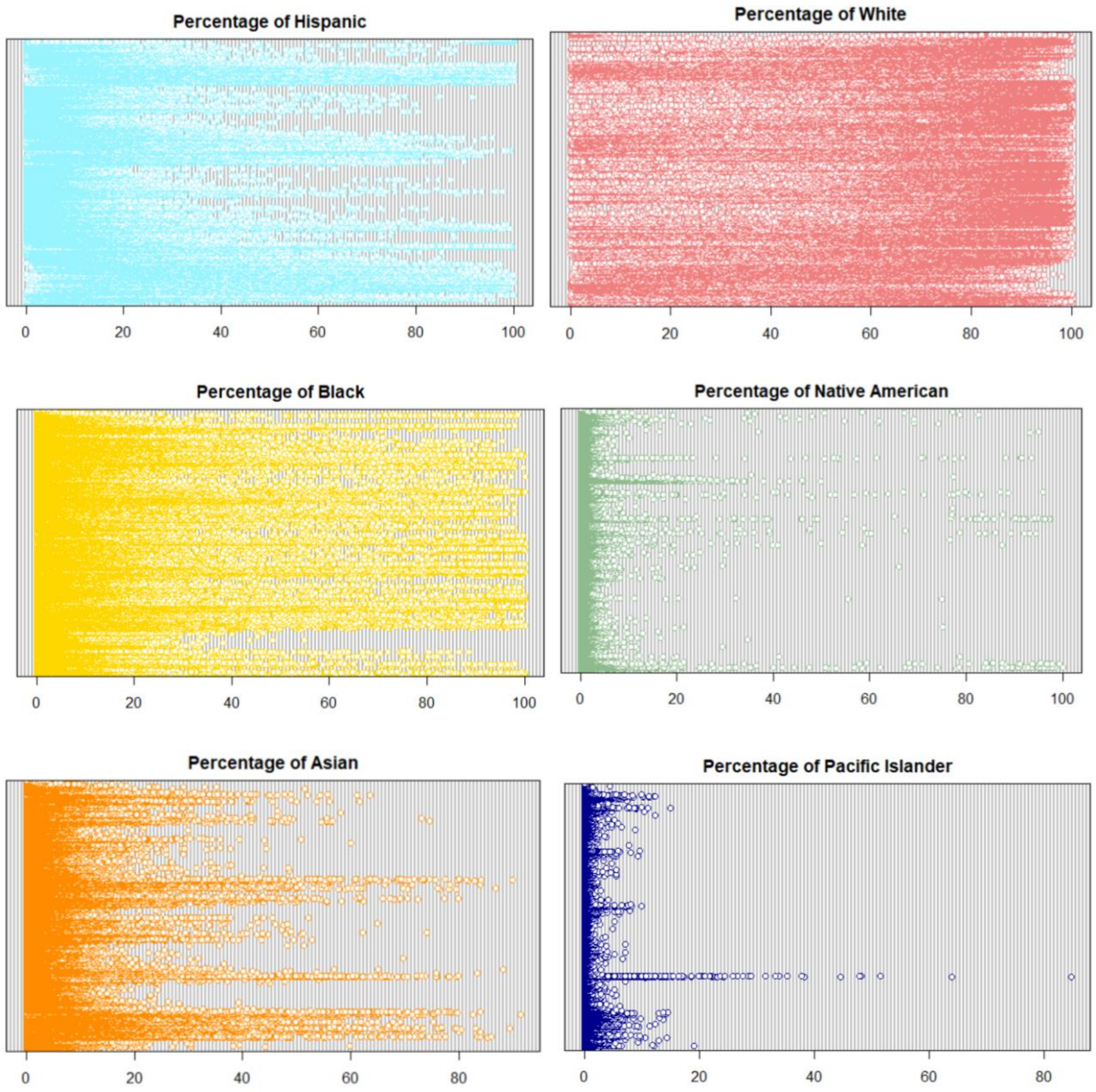

Figure 2.7. Scatterplots of percentage of each racial and ethnic demographic within each census tract.

To further graphically explored the racial demographic distributions, several scatterplots were created for each racial group. These plots are shown in Figure 2.7. The results of these scatterplots match the results obtained via the boxplots described earlier and show more nuance and variation in the percentages by showing a point for each census tract because the plots can

show regional variations in the populations of each demographic. Each point represents the percentage of that particular racial group in that census tract, and the darker portions of the scatterplot indicate a heavier concentration of points in that area. Hence, the plots show that Caucasians are the majority racial demographic group in the United States, although there are still regional differences in the percentages of each racial demographic group.

2.2 Numerical Analysis and Graphical Analysis of Categorical Variables

In this section, we will describe the numerical and graphical analyses performed on the graphical variables in this data set.

First, we performed numerical analysis on the categorical variables. Due to the large number of counties (thus creating a large number of census tracts), we were unable to perform analysis directly on the counties and census tracts, as the large number of observations would make contingency tables unfeasible and unwieldy. However, we were able to look at the number of census districts within each state, and so we focused our numerical and graphical analyses on that variable. We also performed graphical analysis on the counties by creating a small subset of the larger data set, so we were able to draw some conclusions about the County variable.

	Frequency	Percent	Cum. percent
California	7933	10.9	10.9
Texas	5197	7.1	18.1
New York	4783	6.6	24.6
Florida	4109	5.6	30.3
Pennsylvania	3178	4.4	34.7
Illinois	3106	4.3	38.9
Ohio	2930	4.0	42.9
Michigan	2727	3.7	46.7
North Carolina	2160	3.0	49.7
New Jersey	1983	2.7	52.4
Georgia	1949	2.7	55.1
Virginia	1861	2.6	57.6
Indiana	1499	2.1	59.7
Arizona	1480	2.0	61.7
Tennessee	1471	2.0	63.8
Massachusetts	1453	2.0	65.8
Washington	1442	2.0	67.7
Wisconsin	1386	1.9	69.6
Missouri	1384	1.9	71.5
Maryland	1377	1.9	73.4
Minnesota	1331	1.8	75.3
Colorado	1228	1.7	77.0
Alabama	1172	1.6	78.6
Louisiana	1122	1.5	80.1
Kentucky	1103	1.5	81.6
South Carolina	1080	1.5	83.1
Oklahoma	1037	1.4	84.5
Puerto Rico	874	1.2	85.7
Oregon	825	1.1	86.9
Iowa	821	1.1	88.0
Connecticut	817	1.1	89.1
Kansas	759	1.0	90.2
Arkansas	683	0.9	91.1
Nevada	671	0.9	92.0
Mississippi	654	0.9	92.9
Utah	582	0.8	93.7
Nebraska	528	0.7	94.5
New Mexico	498	0.7	95.1
West Virginia	484	0.7	95.8
Maine	351	0.5	96.3
Hawaii	309	0.4	96.7
Idaho	297	0.4	97.1
New Hampshire	292	0.4	97.5
Montana	268	0.4	97.9
Rhode Island	240	0.3	98.2
South Dakota	222	0.3	98.5
Delaware	213	0.3	98.8
North Dakota	205	0.3	99.1
Vermont	183	0.3	99.4
District of Columbia	175	0.2	99.6
Alaska	164	0.2	99.8
Wyoming	131	0.2	100.0
Total	72727	100.0	100.0

Table 2.4. Frequency table of each state, with number of census tracts in each state.

The table above (Table 2.4) is the frequency table for the State variable. This table shows the number of census tracts in each state, arranged in descending order. Hence, the state with the most census tracts (California) is listed first, while the state with the smallest number of census tracts (Wyoming) is listed last. The table also shows the percentage of the total number of census tracts that each state possesses, as well as the cumulative percentage, which is obtained from adding the percentage from each state cumulatively. Each county is divided into any number of census tracts, with more populous counties having more counties and census tracts, while counties with a smaller population have fewer counties and tracts. Hence, a state with a large number of census tracts is very like to be more populous than a state with a smaller number of such tracts. From this, we can conclude that the three largest states are California, Texas, and New York, while the smallest state (based on population) is Wyoming. We can also see that over fifty percent (50%) of the U.S. population appears to be concentrated in ten states, based on the cumulative percentage of census tracts. These states are California, Texas, New York, Florida, Pennsylvania, Illinois, Ohio, Michigan, North Carolina, and New Jersey.

From this frequency table, we also generated a sideways frequency bar chart (shown in Figure 2.8), with the same descending order as the frequency table. This bar graph graphically illustrates the results obtained in Table 2.4.

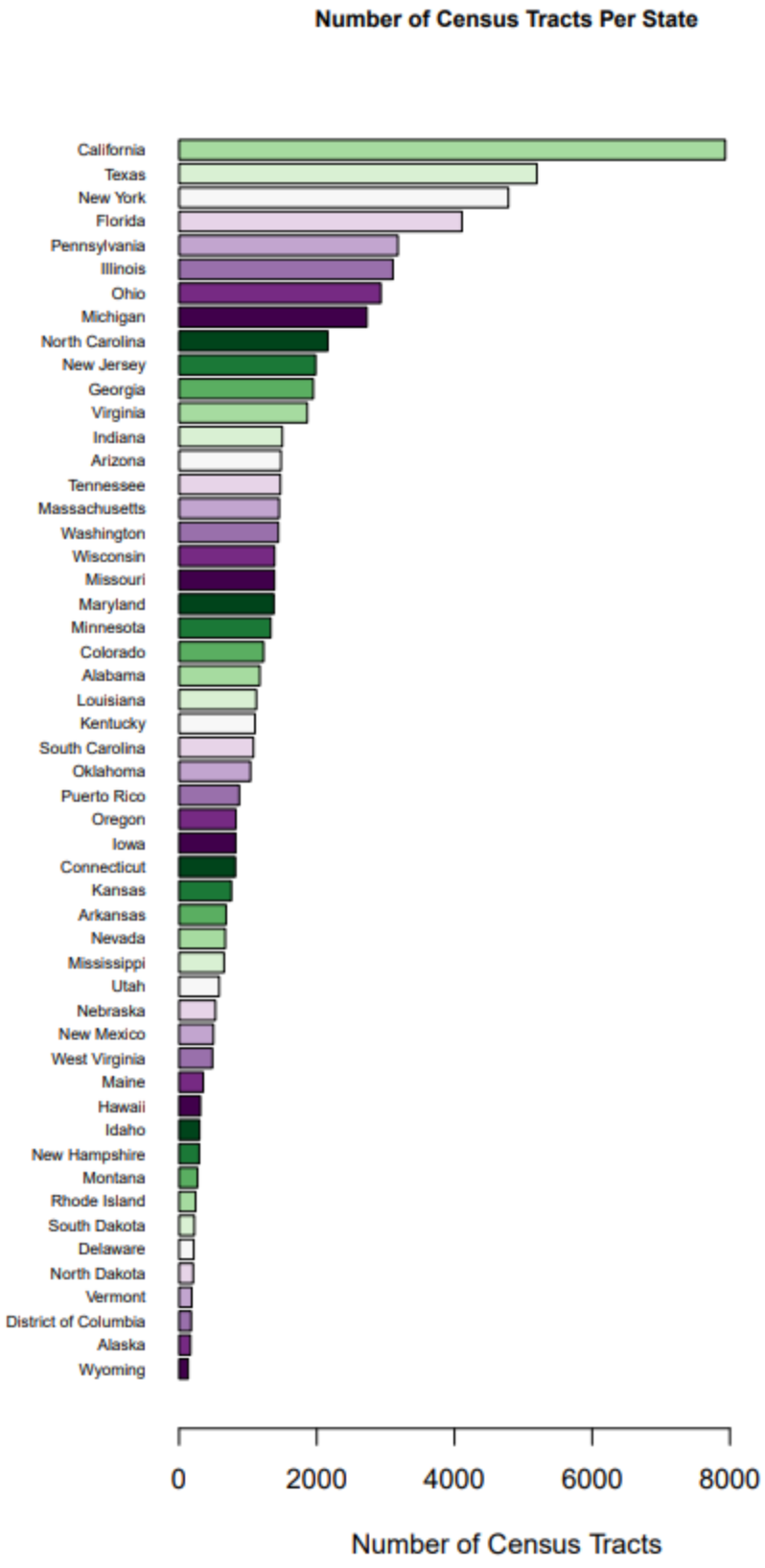

Figure 2.8. Sideways frequency bar chart of number of census districts per state and territory.

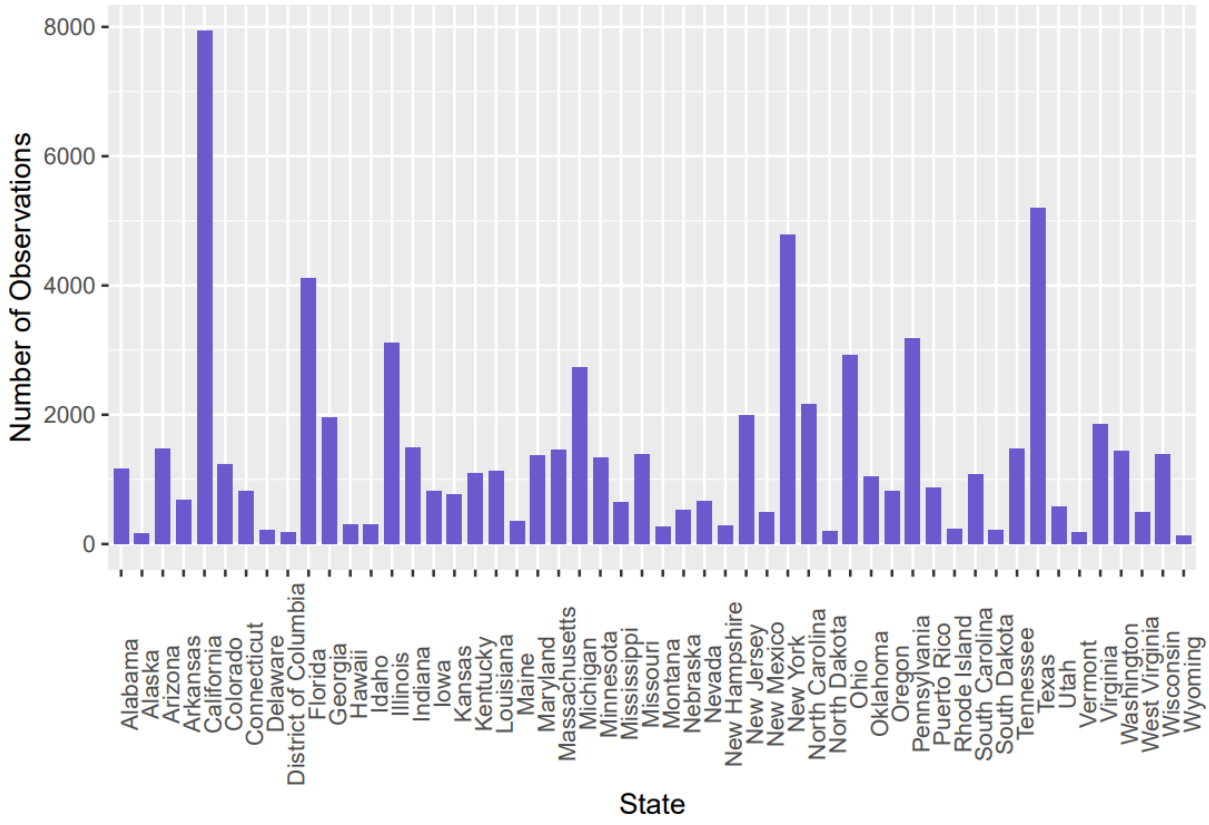

Figure 2.9. Frequency bar chart for state and territory, showing number of census tracts within each state or territory.

Aside from the sideways frequency bar chart in Figure 2.8., created in conjunction with the frequency table in Table 2.4, an alphabetically arranged frequency bar chart for all fifty states, Puerto Rico, and the District of Columbia was created and is shown in Figure 2.9. This chart shows the number of census tracts within each state and territory. Larger and more populous states have more census tracts, since there is more land to divide and more people to classify into each tract. The states with the largest numbers of census tracts are California, Texas, and New York, which indicates that these states are highly populous and are likely the most populous states in the U.S. The state with the least number of census tracts appears to be Wyoming, which suggests that Wyoming has the smallest population in any state or territory in the U.S.

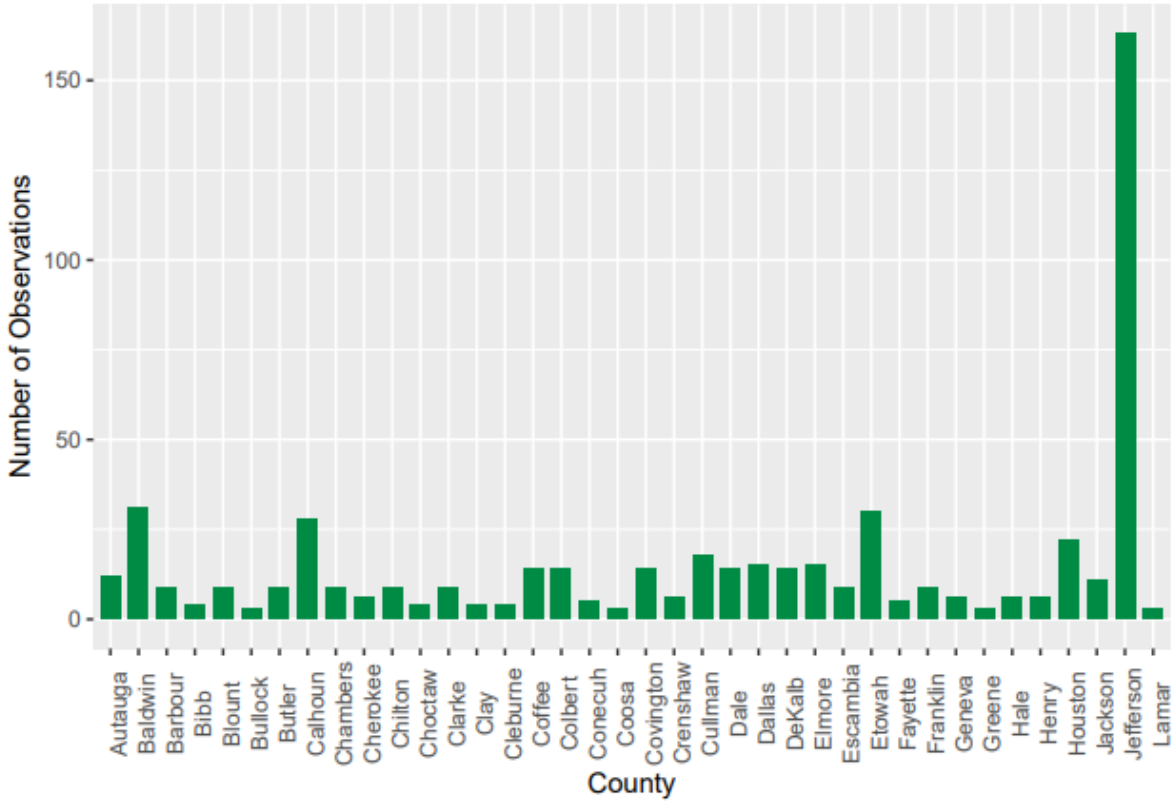

Figure 2.10. Frequency bar chart for counties, using the first 555 observations in the data.

Because of the large number of counties in the United States, creating a frequency bar chart that contained every county was unfeasible. Hence, only a subset of the data was examined, and a bar chart was created based on that subset. The results shown in Figure 2.10 are for several of the counties in the state of Alabama. The number of observations refers to the number of census tracts into which each county is divided, with larger and more populous countries containing more divisions (more census tracts) than the less populous counties. In this subset, we can see that Jefferson county has the most census tracts. This suggests that it is a very populated area, possibly the county that contains the state's capitol. We can draw similar conclusions from the other counties in the data set: a large number of census tracts within a county indicates a more populous area.

3. Modeling and Model Selection

In this section, we will discuss the results of our modeling and model selection. We wish to determine which numerical variables will provide the best predictions for income level and which model will best describe the data and offer the most predictive power. Our approach to developing the models was to group the models into four categories. First, we modeled income against gender. For our second model, we looked at income against race and ethnicity. With the third model, we looked at income against type of employment, and lastly, for our fourth model, we examined income against levels of poverty and child poverty. For each of these groups, due to the skewed nature of the data (as described in the exploratory data analysis in the previous section), we initially performed a log transformation on the response and predictor variables examined for each model. Then, we implemented linear and polynomial regression analysis.

3.1 Log Transformation Models

Model	R-Squared	Adj. R-Squared
Gender: log1p(Men) + log1p(Women)	0.05375	0.05373
Race: log1p(Hispanic) + log1p(White)+ log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific)	0.3873	0.3873
Type of employment: log1p(Professional) + log1p(Service) + log1p(Office) + log1p(Construction) + log1p(Production)	0.542	0.5419
Poverty levels: log1p(Poverty) + log1p(ChildPov)	0.729	0.729

Table 3.1. Logarithmic models for income against each set of predictors, along with R-squared and adjusted R-squared values.

Because we knew from our exploratory data analysis that we would be working with skewed data, we first started our modeling with log transformations on the data. The models, along with their R-squared and adjusted R-squared values, are shown in Table 3.1 above. We also performed an ANOVA analysis of each of these models, shown in Table 3.2 below.

Model	DF	RSS	F Statistic	P-Value
Income versus Gender	72724	16135.8	2066	< 2.2e-16
Income versus Race	72720	10447.7	7662	< 2.2e-16
Income versus Type of Employment	72721	7810.4	1.721e+04	< 2.2e-16
Income versus Poverty Levels	72724	4620.6	9.783e+04	< 2.2e-16

Table 3.2. ANOVA Table of each log model.

From the ANOVA analysis in R, model 2 (income versus race) was selected as the best model, but it had a relatively low R-squared and adjusted R-squared (0.3873, as shown in Table 3.1). The models for income versus type of employment and income versus poverty levels had much higher adjusted R-squared values, and based on this information, we conclude that these two models were the best ones since they explain most of the variation in income. Of these two, we observed from the residual plot for the poverty model that there was still a curved pattern in the residuals even after applying the log transformation (see Figure 3.4), which suggested that there was still some underlying issues with the logarithmic model. Hence, we concluded that the income versus type of employment logarithmic model was our best model. The full model would be

$$log(Income) = \beta_1 log(Professional) + \beta_2 log(Service) + \beta_3 log(Office)$$
$$+ \beta_4 log(Construction) + \beta_5 log(Production) + \varepsilon$$

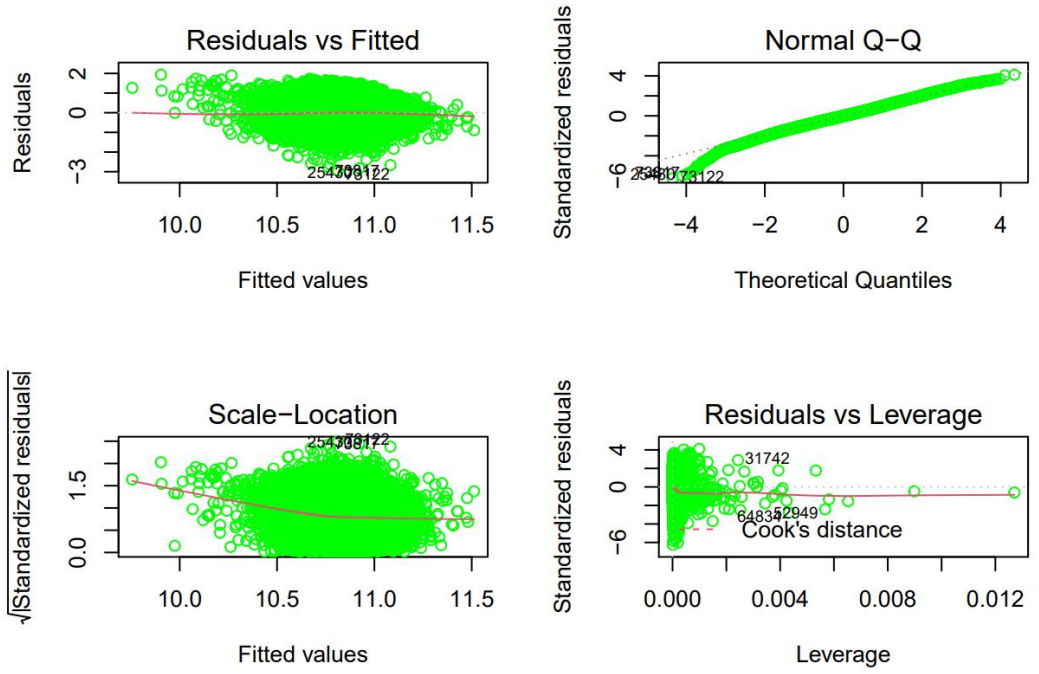

Figure 3.1. Residual, Normal Q-Q, and leverage plots for the income versus gender log model.

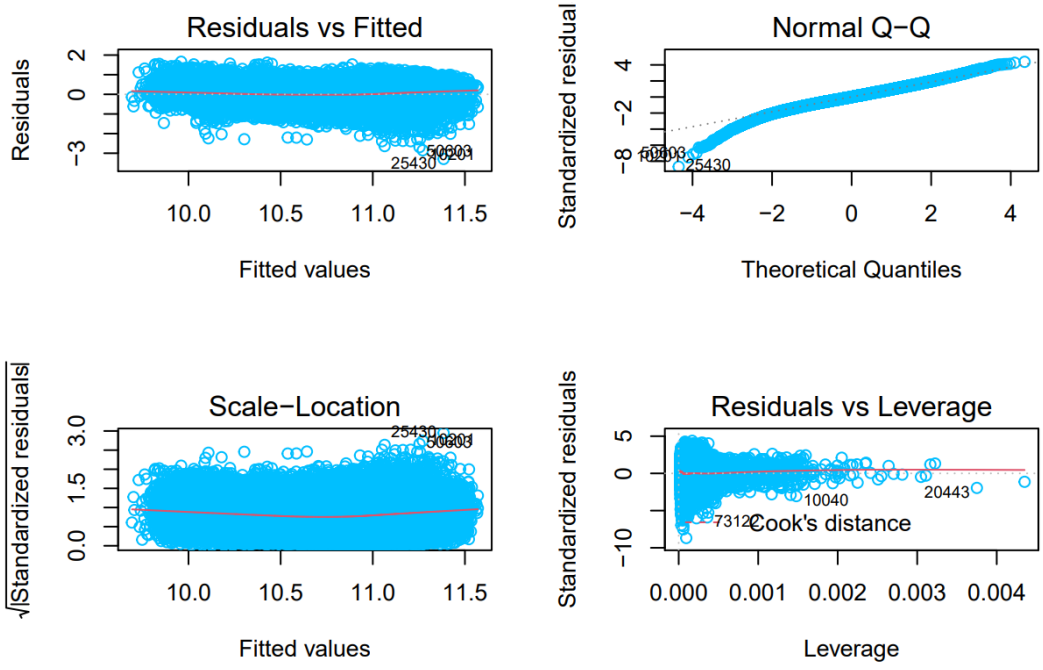

Figure 3.2. Residual, Normal Q-Q, and leverage plots for the income versus race log model.

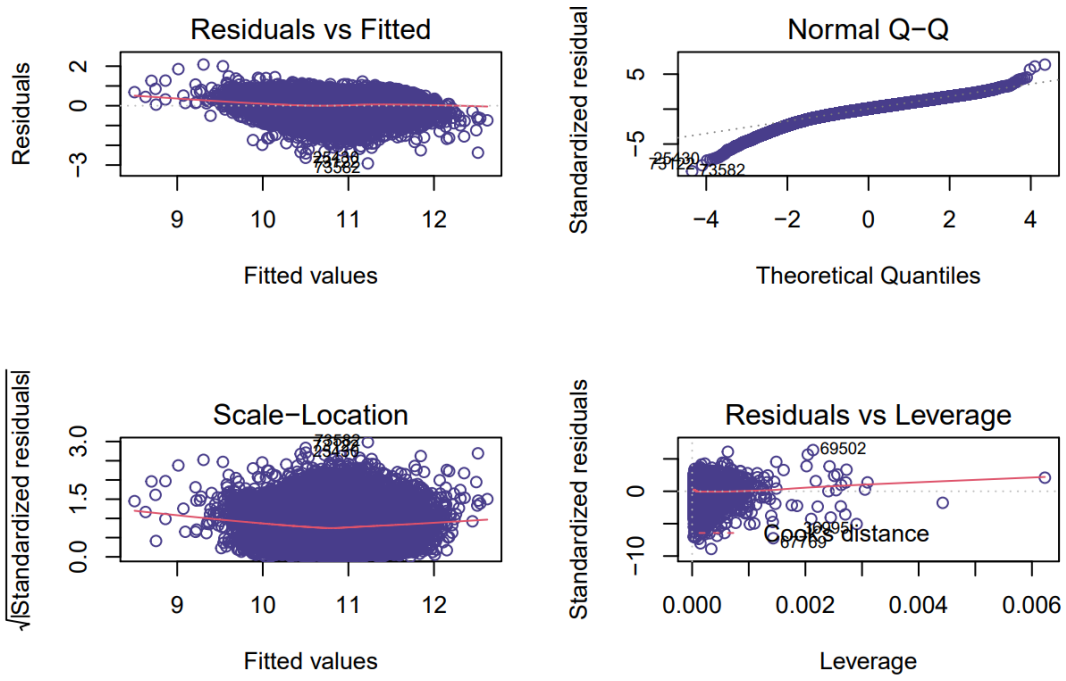

Figure 3.3. Residual, Normal Q-Q, and leverage plots for the income versus type of employment log model.

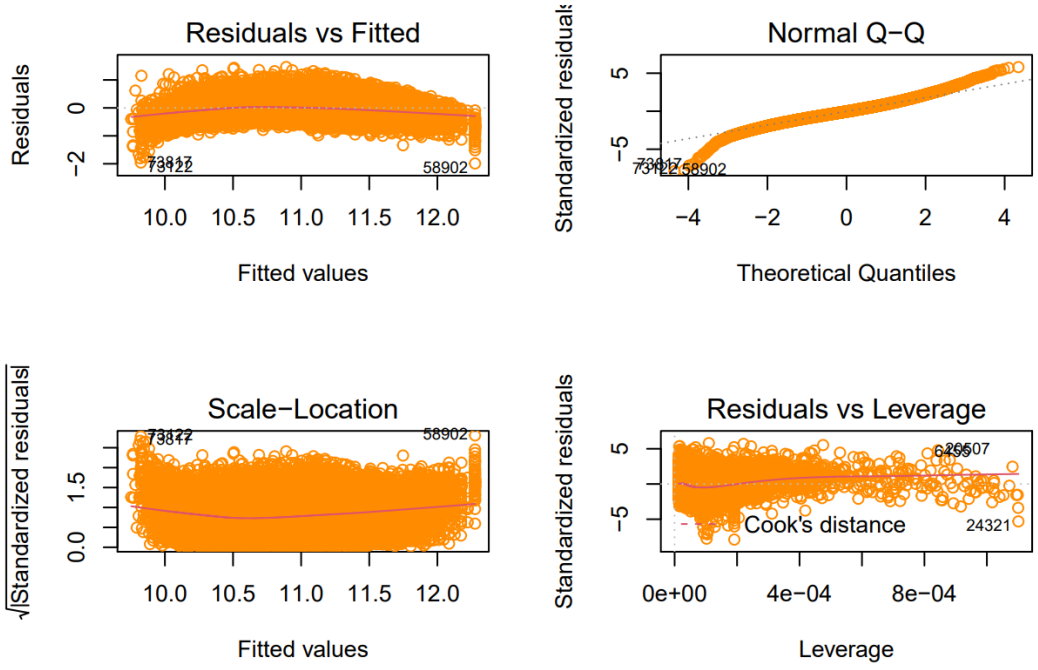

Figure 3.4. Residual, Normal Q-Q, and leverage plots for the income versus poverty levels log model.

In Figures 3.1, 3.2, 3.3, and 3.4, we can see the residual, normal Q-Q, and leverage plots for each model. As can been seen from the plots, the residual plot has an approximately random scattering of the points (with the exception of the residual plot of the poverty model, as described earlier). The Normal Q-Q plots show some heavy tails but on the whole do seem to indicate normality. The leverage plots also show a few potential high leverage observations. We decided to keep these observations in the model, since we determined that due to the size of the data set removal would be rather difficult, and we were also concerned that we would lose valuable information about the relationship between income and the predictors if we removed the high-leverage observations from the data.

3.2 Linear Regression Model

To verify that linear models would be an inappropriate fit for the data, we performed linear regression analysis on each of the four groups of predictors versus income as described in section 3.1. The results of these models are shown in Table 3.3.

Model	R-Squared	Adj. R-Squared
Gender: Men + Women	0.0311	0.03108
Race: Hispanic + White + Black + Native + Asian + Pacific	0.2425	0.2425
Type of Employment: Professional + Service + Office + Construction + Production	0.5613	0.5612
Poverty Levels: Poverty + ChildPov	0.5002	0.5002

Table 3.3. Linear models for income against each set of predictors, along with R-squared and adjusted R-squared values.

Although models 3 and 4 appear to have a rather high adjusted R-squared value, a closer examination of the models' residual, normal Q-Q, and leverage plots showed that the conditions for linearity appear to be violated, and hence the linear models are not a good fit for the data, as suspected based on the highly skewed nature of the data. These plots are shown in Figures 3.5, 3.6, 3.7, and 3.8. The residual plots show a distinct pattern and also indicate that homoscedasticity is violated. The normal Q-Q plots show a strong curved pattern and heavy tails, which suggest that the linear models are inappropriate because normality is violated.

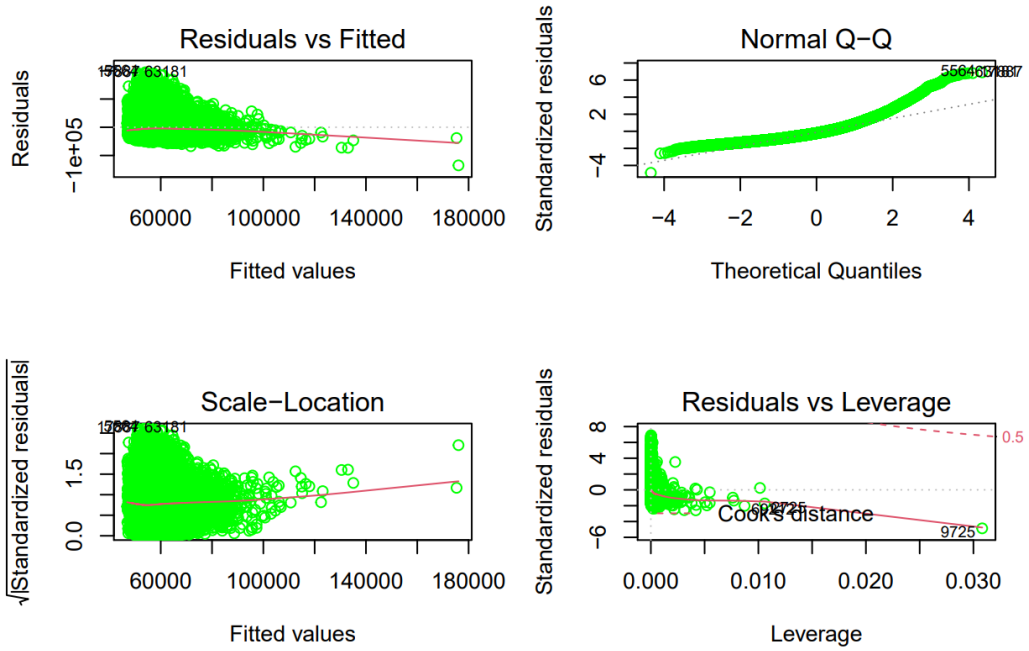

Figure 3.5. Residual, Normal Q-Q, and leverage plots for the income versus gender linear model.

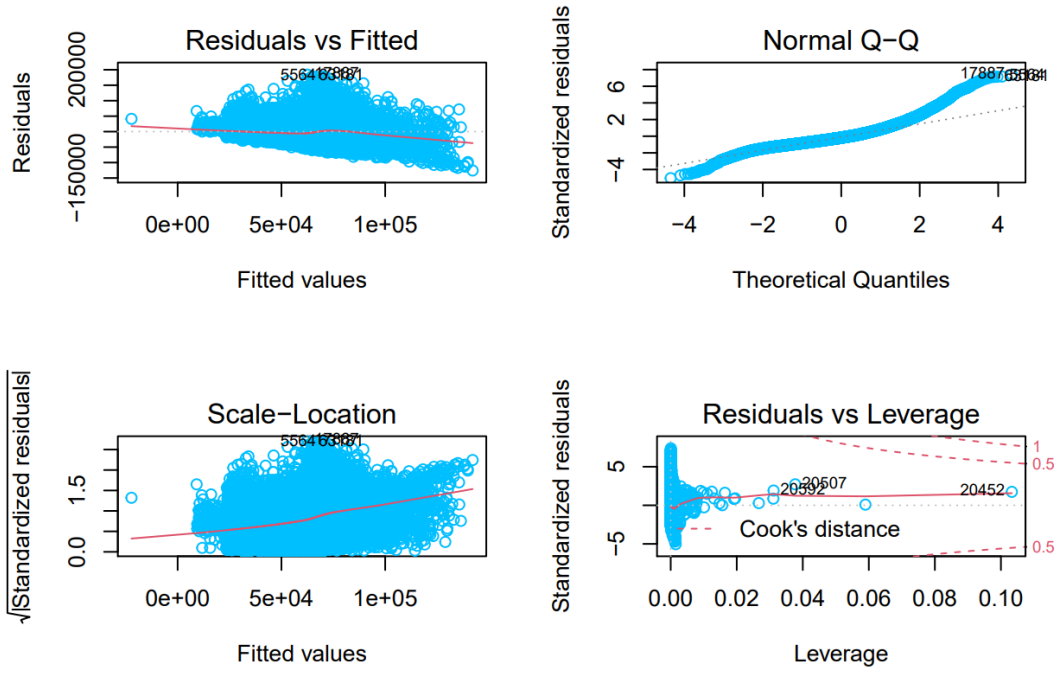

Figure 3.6. Residual, Normal Q-Q, and leverage plots for the income versus race linear model.

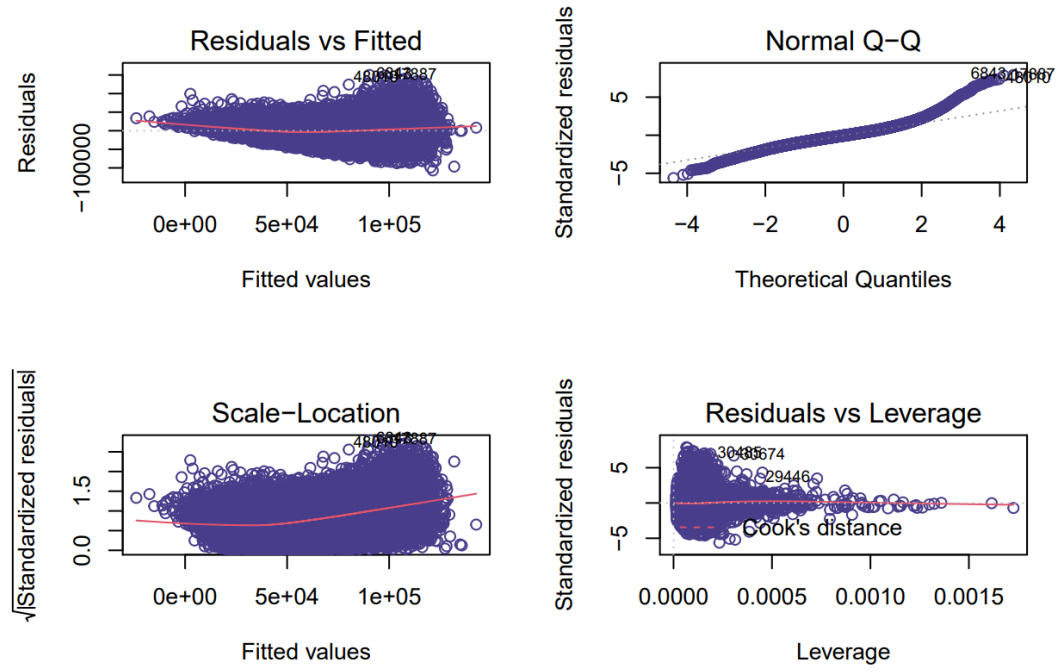

Figure 3.7. Residual, Normal Q-Q, and leverage plots for the income versus type of employment linear model.

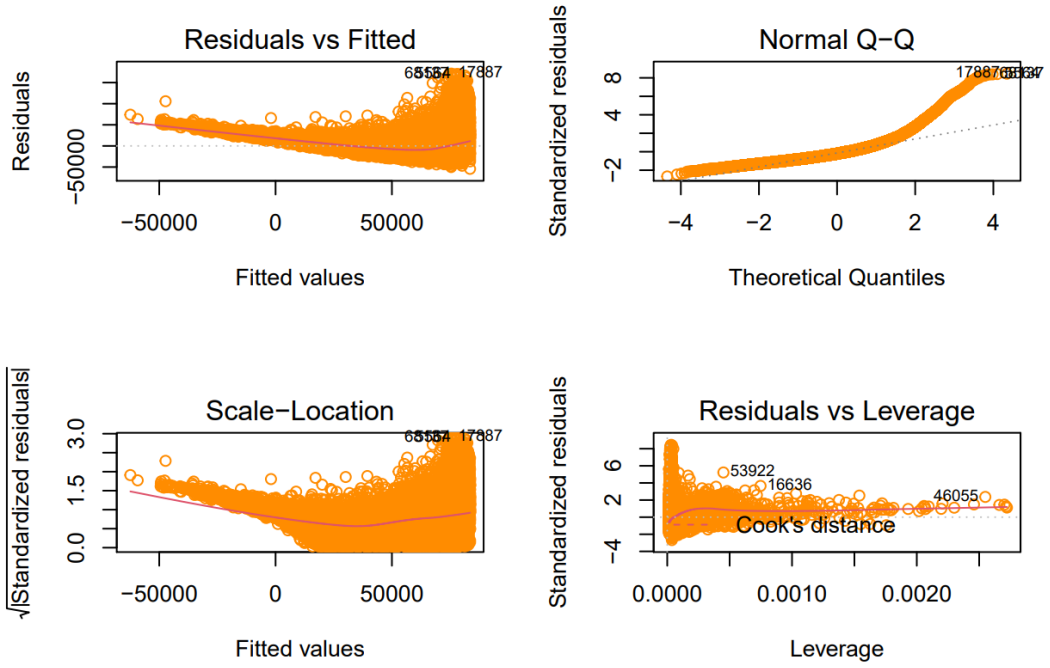

Figure 3.8. Residual, Normal Q-Q, and leverage plots for the income versus poverty levels linear model.

An ANOVA analysis of each of these four linear models indicated that the income versus race and the income versus poverty models were the most significant, but since the linear models are clearly inappropriate for this data, we chose not to use any of these models. We also tried linear models with interaction terms (see Table 3.4), but we did not see much improvement in the plots for the residuals, normal Q-Q, and leverage points (see Figures 3.9, 3.10, 3.11, and 3.12), indicating that the conditions for a linear model are violated. The problems that we had observed in the previous set of plots were still apparent here, such as patterns in the residual plots, violation of normality, and lack of homoscedasticity. Hence, we concluded that the linear models with the interaction terms were also not good fits for the data, so we rejected all of these models as well.

Model	R-Squared	Adj. R-Squared
Gender: Men * Women	0.03243	0.03239
Race: Hispanic * White * Black * Native * Asian * Pacific	0.3546	0.354
Type of employment: Professional * Service * Office * Construction * Production	0.6049	0.6048
Poverty levels: Poverty * ChildPov	0.5828	0.5828

Table 3.4. Linear models with interaction terms for income against each set of predictors, along with R-squared and adjusted R-squared values.

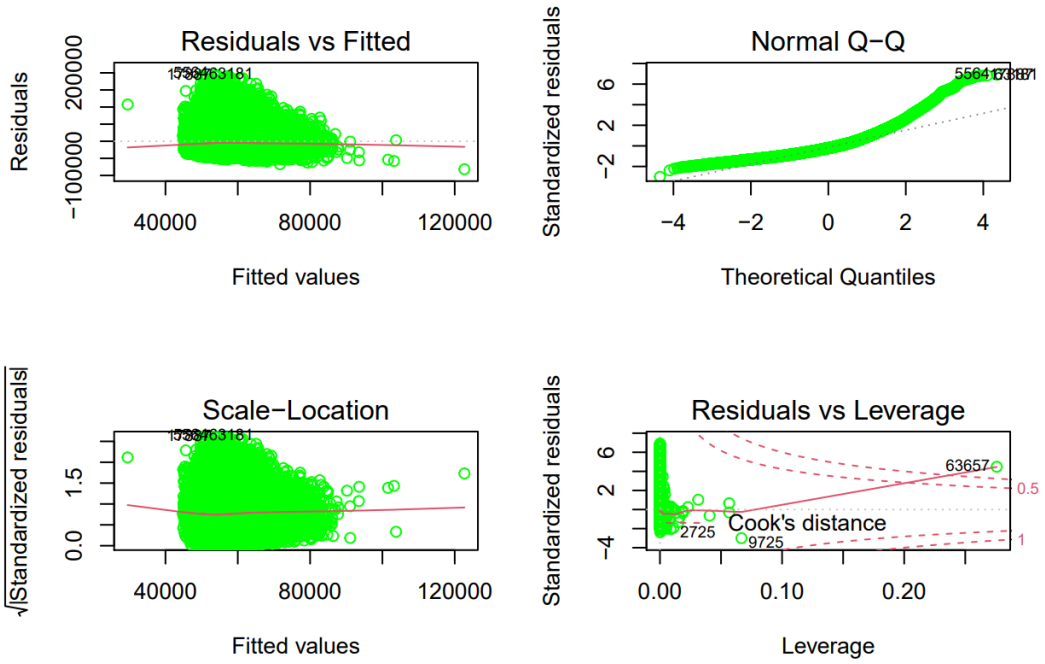

Figure 3.9. Residual, Normal Q-Q, and leverage plots for the income versus gender linear model with interaction terms.

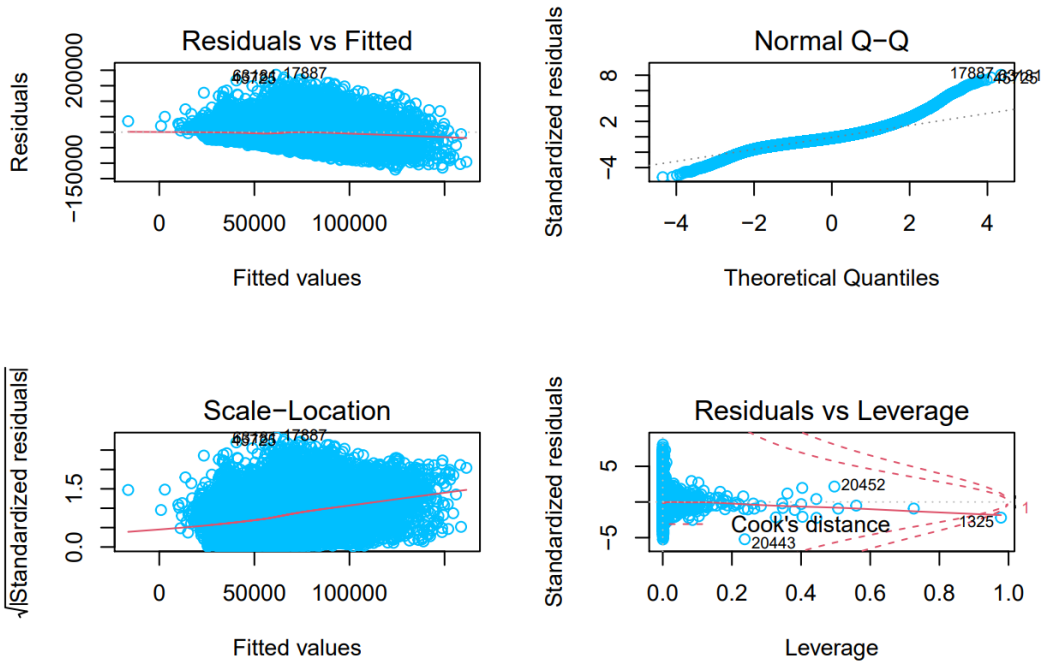

Figure 3.10. Residual, Normal Q-Q, and leverage plots for the income versus race linear model with interaction terms.

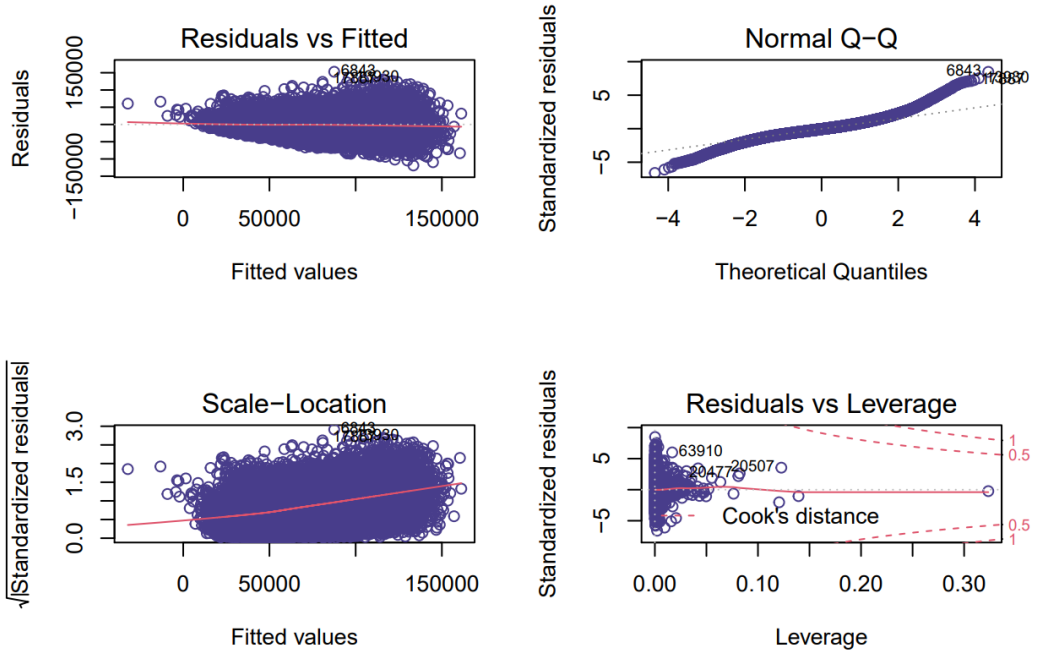

Figure 3.11. Residual, Normal Q-Q, and leverage plots for the income versus type of employment linear model.

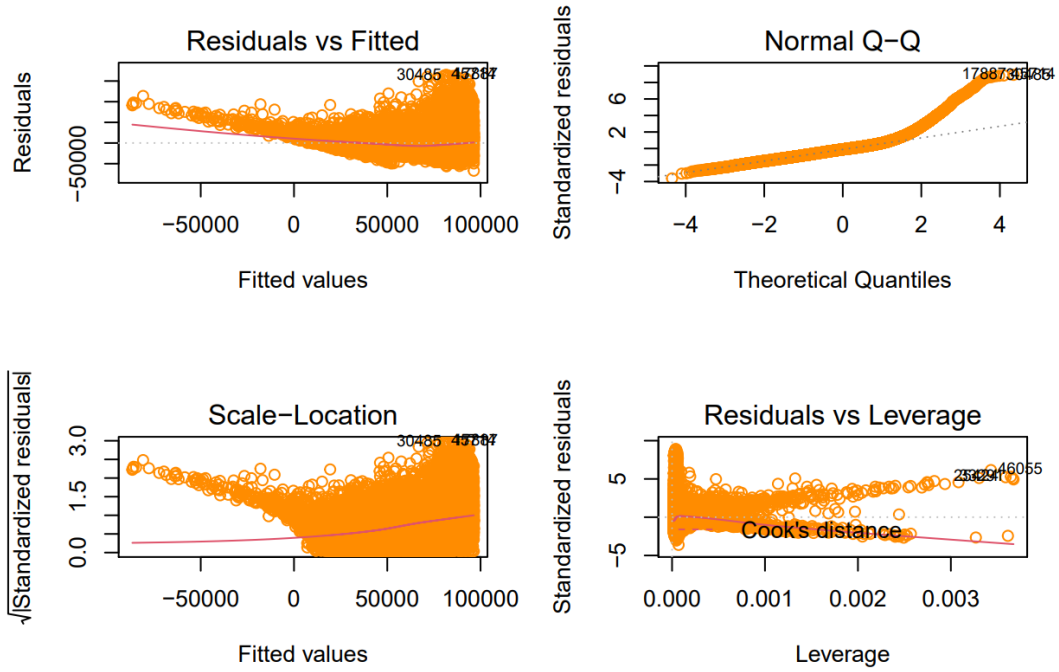

Figure 3.12. Residual, Normal Q-Q, and leverage plots for the income versus poverty levels linear model with interaction terms.

3.3 Polynomial Regression Model

Based on our results and observations in section 3.3, we decided that our income versus type of employment model was our best choice. Although the model of income versus poverty levels had a higher adjusted R^2 value, it is intuitive that poverty levels would affect median income, and the poverty levels log model still had some issues in its plots. Hence, we thought that the most new information could be gained from further analysis on the type of employment model. We performed further analysis with this set of variables by examining a polynomial model and another log model with quadratic terms. The polynomial model is as follows:

$$\begin{aligned} Income = {} & \beta_1 Professional + \beta_2 Service + \beta_3 Office + \beta_4 Construction + \beta_5 Production \\ & + \beta_6(Professional^2) + \beta_7(Service^2) + \beta_8(Office^2) + \beta_9(Construction^2) \\ & + \beta_{10}(Production^2) + \beta_{11}(Professional^3) + \beta_{12}(Service^3) + \beta_{12}(Office^3) \\ & + \beta_{13}(Construction^3) + \beta_{14}(Production^3) + \varepsilon \end{aligned}$$

The log model with quadratic terms is as follows:

$$log(Income) = \beta_1 log(Professional) + \beta_2 log(Service) + \beta_3 log(Office) + \beta_4 log(Construction) + \beta_5 log(Production) + \beta_6 (log1p(Professional)^2) + \beta_7 (log1p(Service)^2) + \beta_8 (log1p(Office)^2) + \beta_9 (log1p(Construction)^2) + \beta_{10} (log1p(Production)^2) + \varepsilon$$

Model	R-Squared	Adj. R-Squared
Polynomial: Professional + Service + Office + Construction + Production + I(Professional^2) + I(Service^2) + I(Office^2) + I(Construction^2) + I(Production^2) + I(Professional^3) + I(Service^3) + I(Office^3) + I(Construction^3) + I(Production^3)	0.5964	0.5963
Log with quadratic terms: log1p(Professional) + log1p(Service) + log1p(Office) + log1p(Construction) + log1p(Production) + I(log1p(Professional)^2) + I(log1p(Service)^2) + I(log1p(Office)^2) + I(log1p(Construction)^2) + I(log1p(Production)^2)	0.5659	0.5659

Table 3.5. Logarithmic models for income against each set of predictors, along with R-squared and adjusted R-squared values.

Each of these models, with its R-square and adjusted R-squared values, is shown in Table 3.5 below. The polynomial appears to be the better fit of the two, based on the adjusted R-squared values and the results of the residual, normal Q-Q, and leverage plots in Figures 3.13 and 3.14. However, the residual plot for the polynomial model may have some issues with non-homoscedasticity, and the tails on the log and quadratic model do not appear to be as heavy as on the polynomial model. On the other hand, the polynomial model does not appear to have as many high-leverage points as the log quadratic model. Ultimately, after comparing all of the regression

models described in Section 3 using the ANOVA table, the best fitting model is the third-degree polynomial model for Income versus employment type.

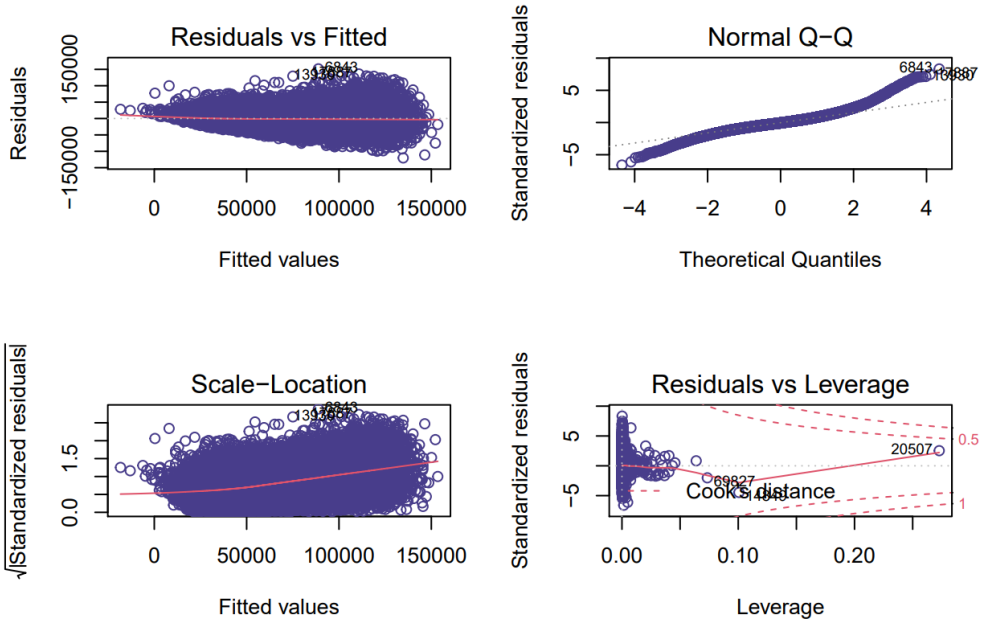

Figure 3.13. Residual, Normal Q-Q, and leverage plots for the income versus type of employment polynomial model.

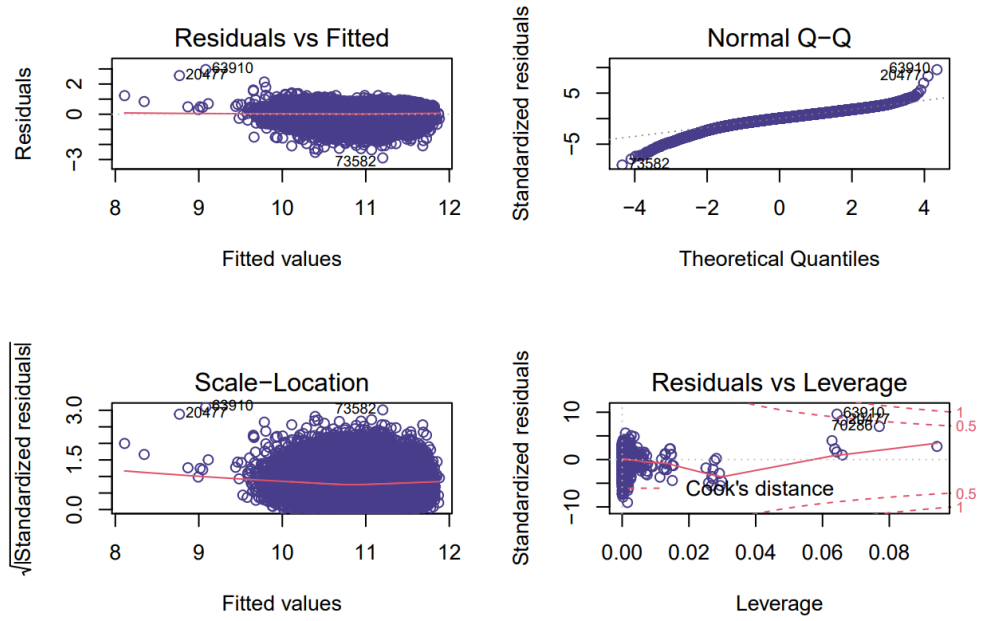

Figure 3.14. Residual, Normal Q-Q, and leverage plots for the income versus type of employment logarithmic model with quadratic terms.

4. Variable Selection

After designing different models by grouping the variables into four categories, as described in the previous section, we then performed variable selection using four different methods: best subset selection, forward stepwise selection, backward stepwise selection, and hybrid stepwise selection. The criteria examined for each method were residual sums of squares (RSS), adjusted R^2, Mallow's Cp, and BIC. Based on the number of variables selected from the values of adjusted R^2, Mallow's Cp, and BIC, we created three models per method using the variables chosen in each selection method. From our previous observations about the skewedness of the data, we decided that developing linear models from each variable selection method would be inappropriate; hence, we instead created log models with the variables chosen from each selection method.

4.1 Best Subset Selection

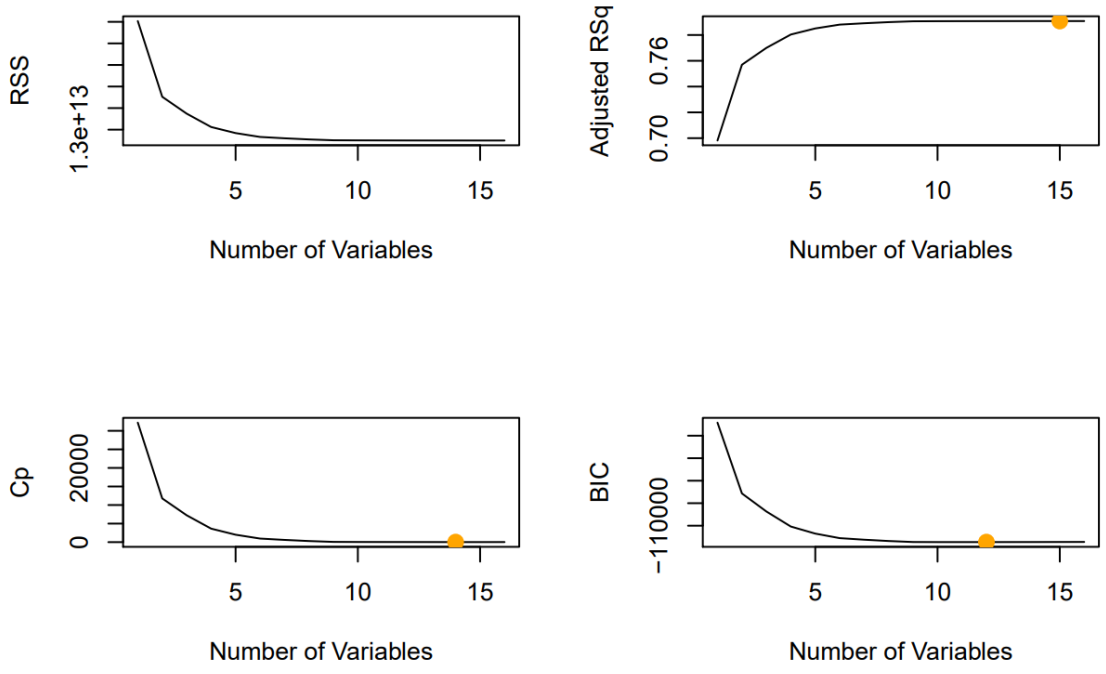

Figure 4.1. RSS, Adjusted R^2, Mallow's Cp, and BIC plots for best subset selection

The first variable selection method we performed was the best subset selection method. The plots shown in Figure 4.1 graphically illustrate the values of RSS, adjusted R^2, Mallow's Cp, and BIC obtained for each addition of a variable, with the maximum R^2 and minimum Mallow's Cp and BIC marked in orange. Hence, based on this output, we require a 15-variable model for to obtain maximum adjusted R^2, a 14-variable model for minimum Cp, and a 12-variable model for minimum BIC. These results, along with a determination of the variables selected via best subset selection, generated the models shown in Table 4.1.

Measure	**Number of Variables**	**Model Obtained**	**Adjusted R^2 of Model**
Adjusted R^2	15	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.8619
Mallow's Cp	14	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.8617
BIC	12	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(Professional) + log1p(Service) + log1p(Production)	0.8596

Table 4.1. Models obtained from maximum adjusted R^2, minimum Mallow's Cp, and minimum BIC for best subset selection

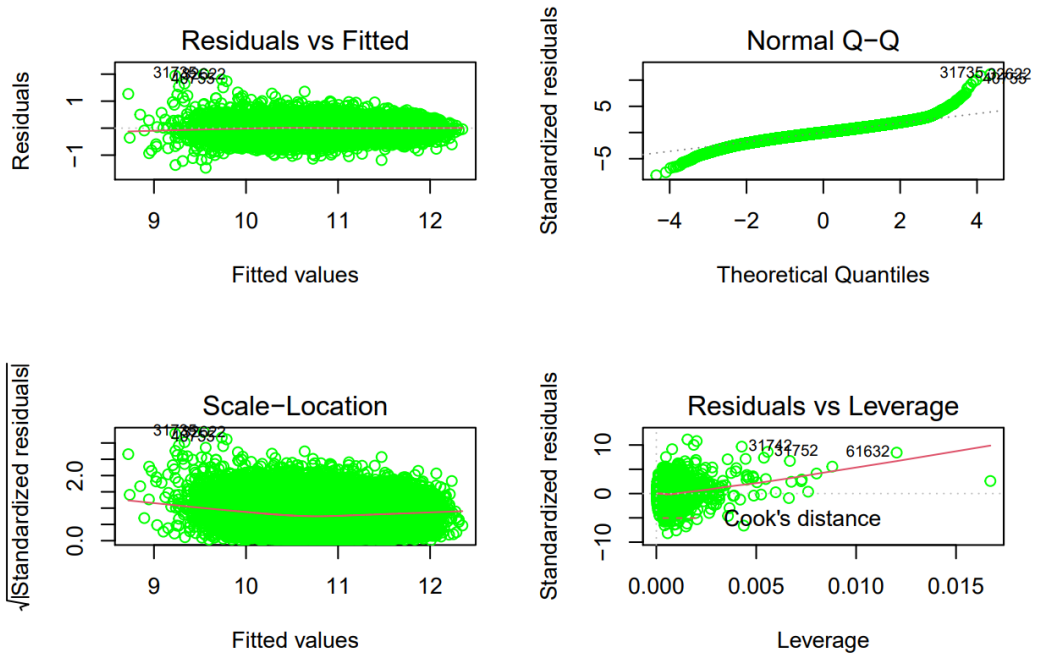

Figure 4.2. Residual, Normal Q-Q, and leverage plots for the model obtained from best subset selection based on maximum adjusted R^2.

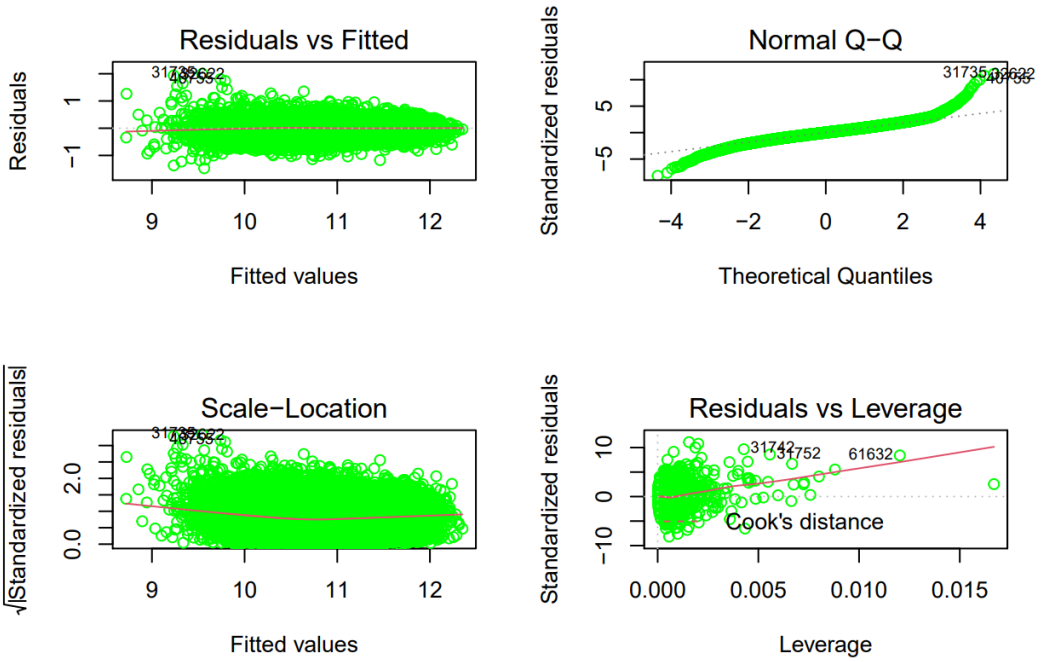

Figure 4.3. Residual, Normal Q-Q, and leverage plots for the model obtained from best subset selection based on minimum Mallow's Cp.

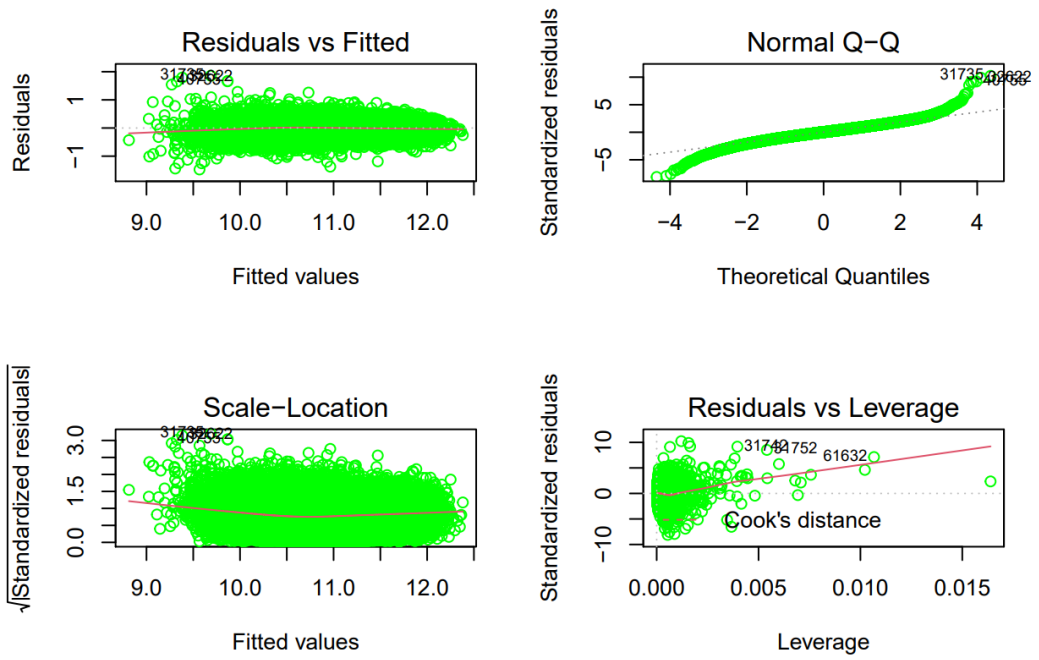

Figure 4.4. Residual, Normal Q-Q, and leverage plots for the model obtained from best subset selection based on minimum BIC.

As can be seen in Table 4.1, the adjusted R^2 values for these models are very high, with the model obtained from the maximum adjusted R^2 measure narrowly having the highest adjusted R^2 value (0.8619). This suggests that that this model is the best one obtained via best subset selection; however, we must also verify this claim graphically. Hence, after creating each of these models, we then performed graphical analysis on each model to verify that the models graphically fit the data. The residual, normal Q-Q, and leverage plots for each model are shown in Figures 4.2, 4.3, and 4.4. The graphs for teach model look nearly identical, which makes sense because the adjusted R^2 values for all the models are very close. The normal Q-Q plots appear to have some heavy tails, although the residual plots do not appear to have any significant patterns. There are also a few high-leverage points, as can be seen in the leverage plot. In keeping with our earlier decision, we decided to keep these observations in the model, since we determined that due to the size of the data set removal would be rather difficult, and we were also concerned that we would

lose valuable information about the relationship between income and the predictors if we removed the high-leverage observations from the data.

From the results of these plots and the adjusted R^2 values shown in Table 4.1, the best model appears to be the one obtained from maximum adjusted R^2, though as noted previously, the model based on minimum Mallow's Cp has a nearly identical adjusted R^2 value, and so could also be considered a good choice.

4.2 Forward Stepwise Selection

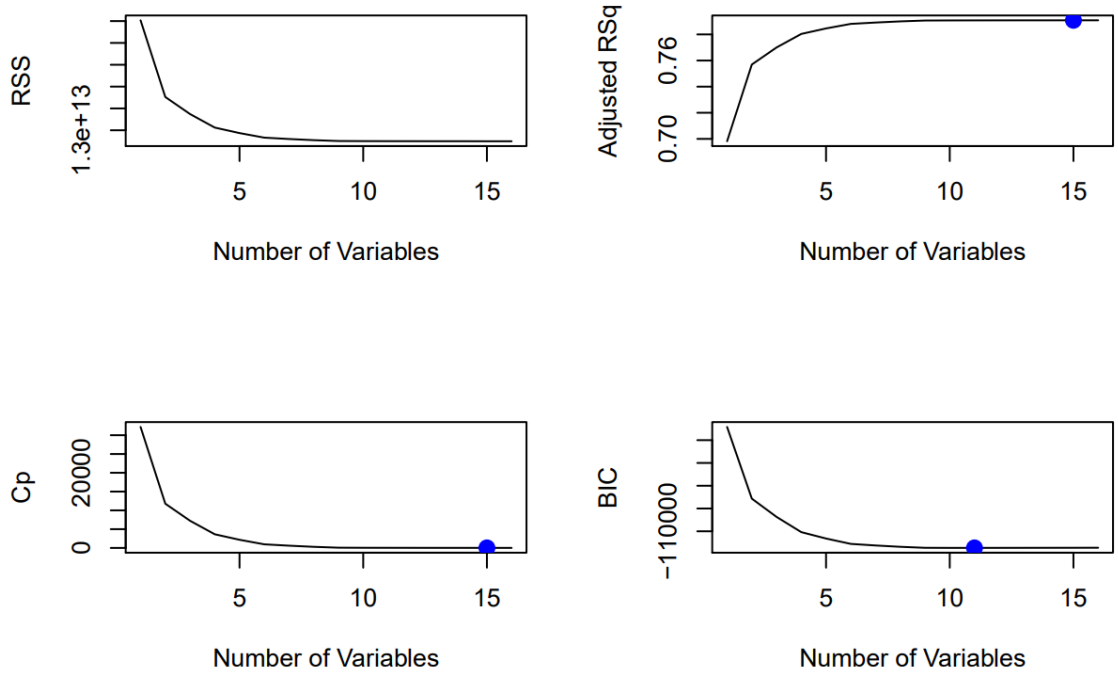

Figure 4.5. RSS, Adjusted R^2, Mallow's Cp, and BIC plots for forward stepwise selection

Next, we performed variable selection with the forward stepwise selection method. The plots shown in Figure 4.5 graphically illustrate the values of RSS, adjusted R^2, Mallow's Cp, and BIC obtained for each addition of a variable, with the maximum R^2 and minimum Mallow's Cp and BIC marked in blue. Hence, based on this output, we require a 15-variable model for to obtain maximum adjusted R^2, a 15-variable model for minimum Mallow's Cp, and an 11-variable model for minimum BIC. These results, along with a determination of the variables selected via forward

stepwise selection, generated the models shown in Table 4.2. The models for maximum adjusted R^2 and minimum Mallow's Cp are identical.

Measure	**Number of Variables**	**Model Obtained**	**Adjusted R^2 of Model**
Adjusted R^2 and Mallow's Cp	15	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Service) + log1p(Office) + log1p(Construction)	0.864
BIC	11	log1p(Men) + log1p(Women) + log1p(White) + log1p(Black) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(Professional) + log1p(Service) + log1p(Construction)	0.8588

Table 4.2. Models obtained from maximum adjusted R^2, minimum Mallow's Cp, and minimum BIC for forward stepwise selection

From Table 4.2, we can see that the best model obtained, based on maximum adjusted R^2 and minimum Mallow's Cp, has a higher adjusted R^2 value than the best model obtained in the previous section from best subset selection. Both models obtained here also utilize different variables than the models found in the previous section. For example, the 15-variable model shown in Table 4.2 uses the "Service" variable, which is absent from the 15-variable model in Table 4.1. Hence, the two variable selection methods yielded several slightly different models. From these two models, the highest adjusted R^2 value is 0.864, associated with the 15-variable maximum adjusted R^2 and minimum Mallow's Cp variable selection criteria. This suggests that this is the

best model obtained from forward stepwise selection, and this model also has the highest overall adjusted R^2 of all models obtained thus far, which indicates that this is the best model so far.

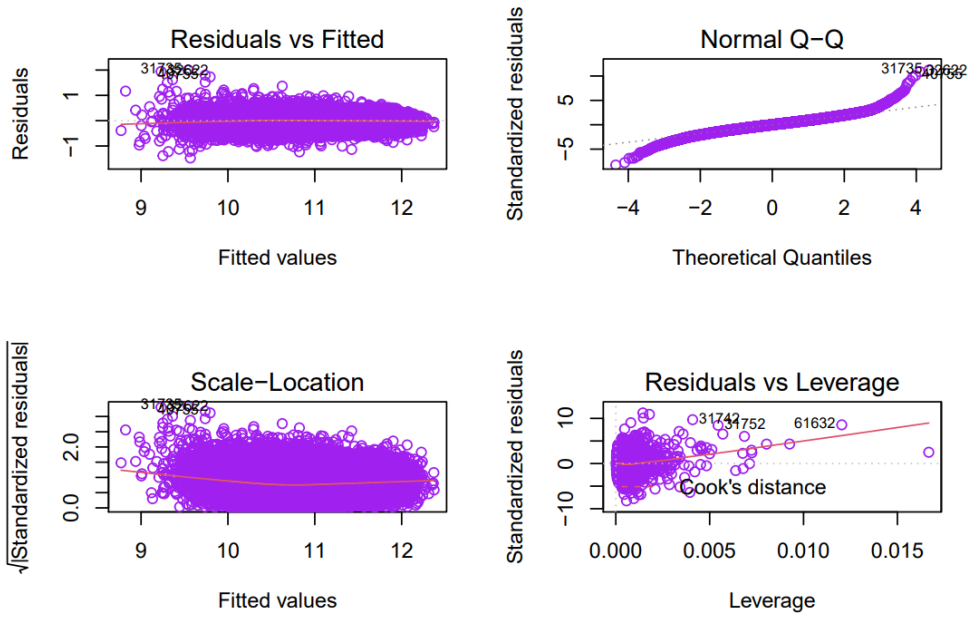

Figure 4.6. Residual, Normal Q-Q, and leverage plots for the model obtained from best subset selection based on maximum adjusted R^2 and minimum Mallow's Cp.

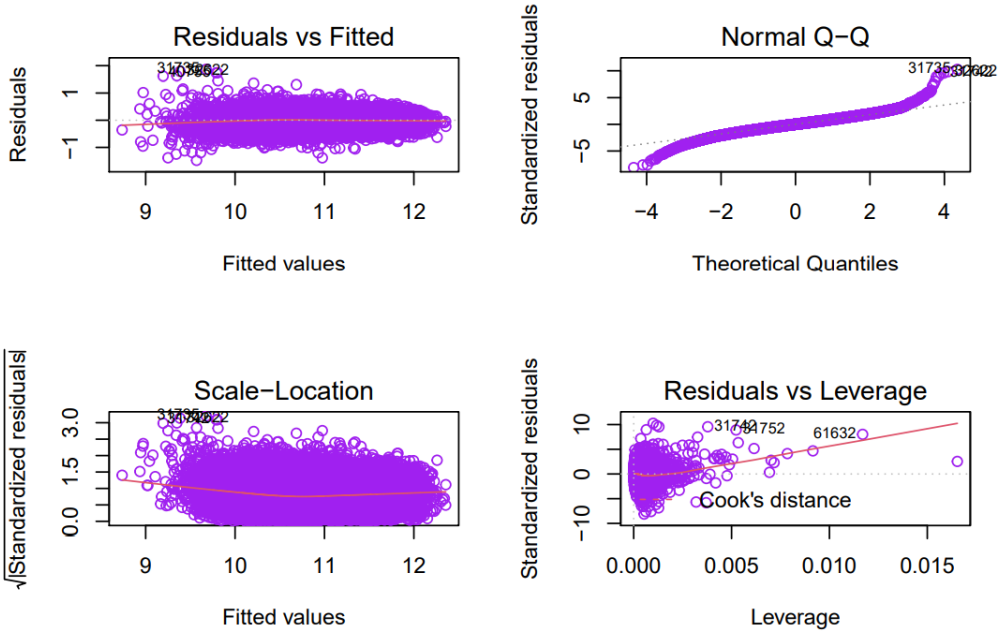

Figure 4.7. Residual, Normal Q-Q, and leverage plots for the model obtained from best subset selection based on minimum BIC.

To verify that these models are good fits for the data, we can also look at the graphical summaries of these models, which are shown in Figures 4.6 and 4.7. From the residual, normal Q-Q, and leverage plots for each model, the residuals for both models appear to be randomly scattered and do not have a significant pattern. The normal Q-Q plots have some heavy tails, and there are also a few high-leverage points. Based on the adjusted R^2 squared values and the residual, normal Q-Q, and leverage plots for all the models analyzed in best subset selection and forward stepwise selection, the 15-variable model obtained in this section appears to be the best model so far. However, we did observe that the variable "Native" was no longer statistically significant in this logarithmic model. In a linear model, this variable would have been considered statistically significant, but utilizing a logarithmic model likely caused this variable to no longer be significant.

4.3 Backward Stepwise Selection

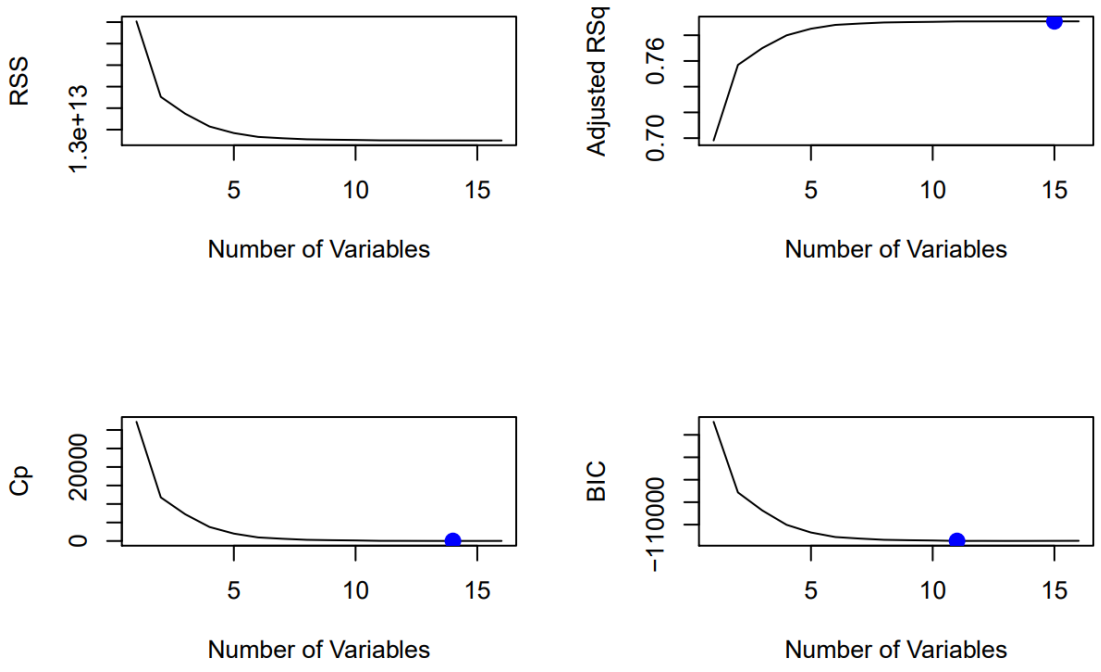

Figure 4.8. RSS, Adjusted R^2, Mallow's Cp, and BIC plots for backward stepwise selection

Next, we performed variable selection with the forward stepwise selection method. The plots shown in Figure 4.8 graphically illustrate the values of RSS, adjusted R^2, Mallow's Cp, and

BIC obtained for each addition of a variable, with the maximum R^2 and minimum Mallow's Cp and BIC marked in blue. Hence, based on this output, we require a 15-variable model for to obtain maximum adjusted R^2, a 14-variable model for minimum Mallow's Cp, and an 11-variable model for minimum BIC. These results, along with a determination of the variables selected via backward stepwise selection, generated the models shown in Table 4.3.

Measure	**Number of Variables**	**Model Obtained**	**Adjusted R^2 of Model**
Adjusted R^2	15	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.8619
Mallow's Cp	14	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.8617
BIC	11	log1p(Men) + log1p(Women) + log1p(White) + log1p(Black) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.857

Table 4.3. Models obtained from maximum adjusted R^2, minimum Mallow's Cp, and minimum BIC for backward stepwise selection

From Table 4.3, it can been seen that the models obtained from maximum adjusted R^2 and minimum Mallow's Cp are the same models obtained from these criteria in best subset selection. The minimum BIC model, with 11 variables, has the same number of variables as the minimum BIC model obtained via forward stepwise selection, but the variables selected for each model are different. The adjusted R^2 value obtained for the minimum BIC model from best subset selection is also higher than the adjusted R^2 value for the minimum BIC model obtained here, from backward stepwise selection.

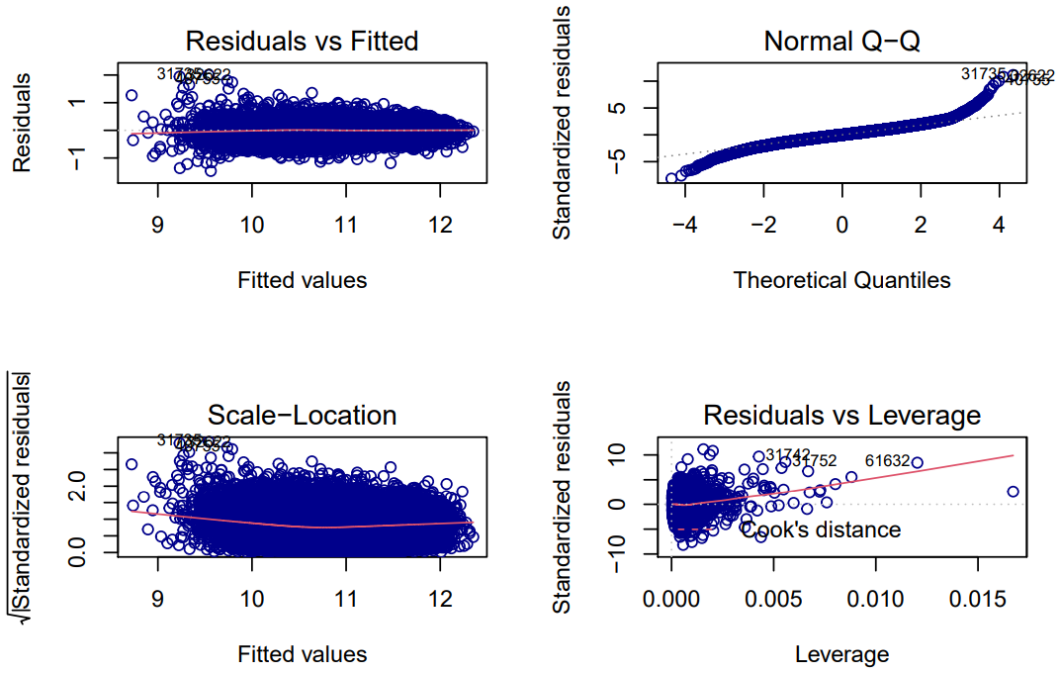

Figure 4.9. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on maximum adjusted R^2.

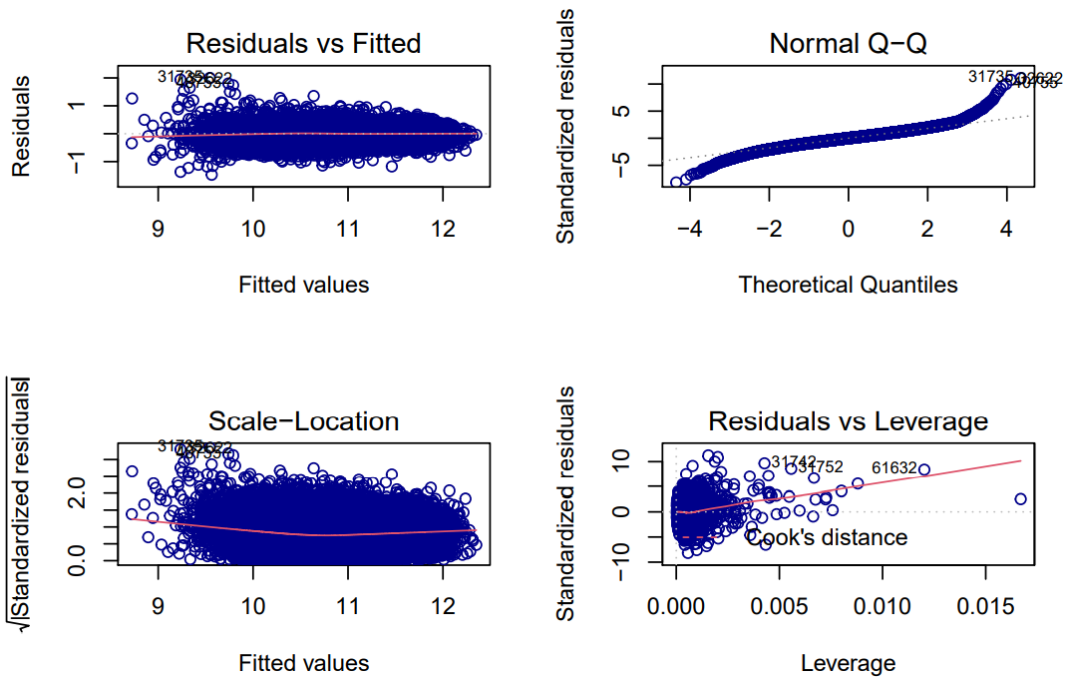

Figure 4.10. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on minimum Mallow's Cp.

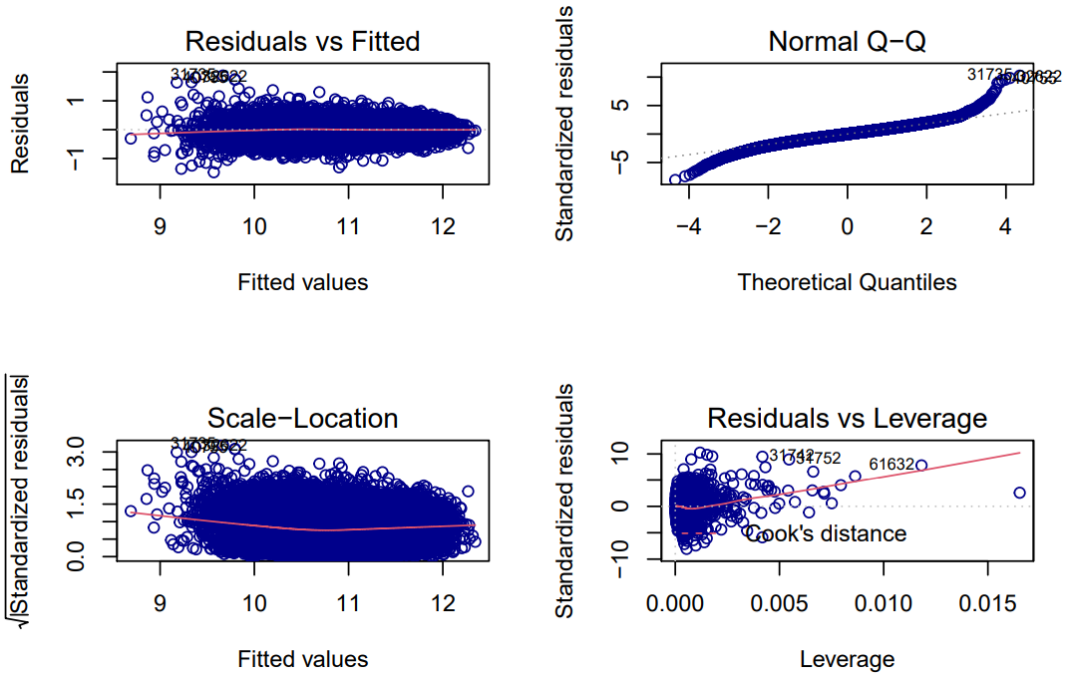

Figure 4.11. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on minimum BIC.

From graphical analysis of the models (see Figures 4.9, 4.10, and 4.11), we can see that the residuals appear randomly scattered, and there does not appear to be any apparent pattern. The slight curvature in the scale-location plot for this model looks smaller than in the scale-location plots for the previous models. The normal Q-Q plot has some heavy tails, and the leverage plot indicates that there are a few high-leverage observations or outliers. Based on our observations and analyses of all models so far, the maximum adjusted R^2-minimum Mallow's Cp model obtained in section 4.2 appears to be the best model obtained so far.

4.4 Hybrid Selection

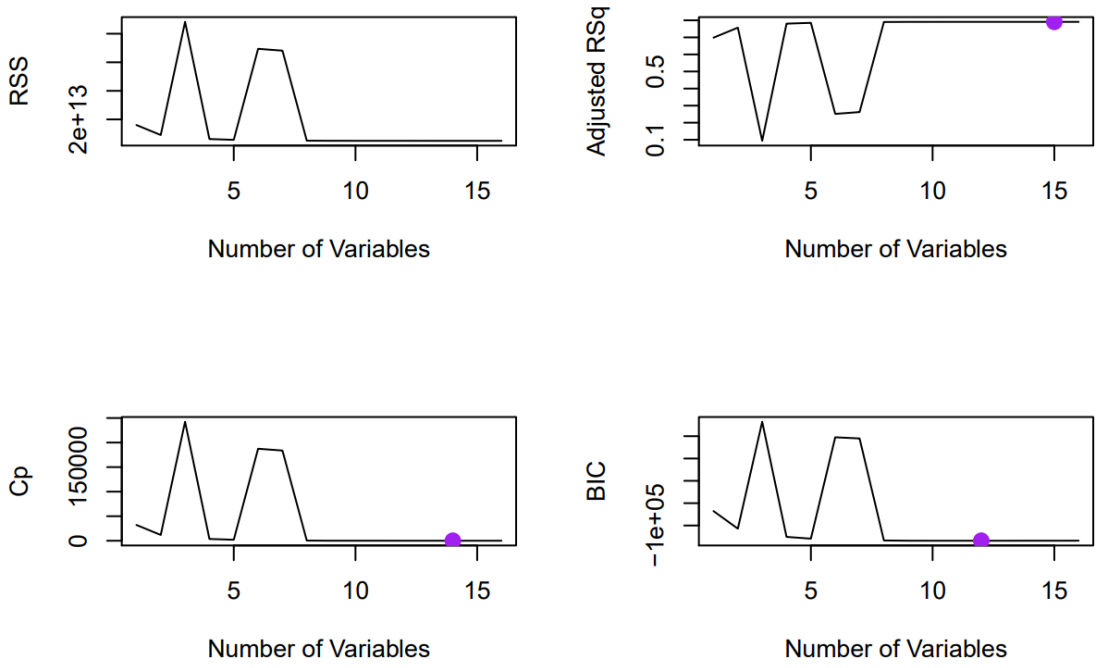

Figure 4.12. RSS, Adjusted R^2, Mallow's Cp, and BIC plots for hybrid selection

Lastly, we performed hybrid selection to choose variables and models. This approach utilizes a combination of forward and backward stepwise variable selection. The plots shown in Figure 4.12 graphically illustrate the values of RSS, adjusted R^2, Mallow's Cp, and BIC obtained for each addition of a variable, with the maximum R^2 and minimum Mallow's Cp and BIC marked in purple. Hence, based on this output, we require a 15-variable model for to obtain maximum

adjusted R^2, a 14-variable model for minimum Cp, and a 12-variable model for minimum BIC. These results, along with a determination of the variables selected via best subset selection, generated the models shown in Table 4.4.

Measure	**Number of Variables**	**Model Obtained**	**Adjusted R^2 of Model**
Adjusted R^2	15	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(Professional) + log1p(Service) + log1p(Construction) + log1p(Production)	0.8613
Mallow's Cp	14	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Office) + log1p(Construction) + log1p(Production)	0.8617
BIC	12	log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(IncPerCap) + log1p(Poverty) + log1p(Professional) + log1p(Service) + log1p(Construction) + log1p(Production	0.861

Table 4.4. Models obtained from maximum adjusted R^2, minimum Mallow's Cp, and minimum BIC for hybrid selection

Table 4.4 summarizes each of the three models obtained via hybrid selection. Interestingly, the minimum Mallow's Cp model had the highest adjusted R^2. This is likely because the logarithmic model for highest adjusted R^2 has two non-statistically significant variables (Black and Native). However, the adjusted R^2 values for all three models were ultimately very close

(0.8613, 0.8617, and 0.861, respectively). The maximum adjusted R^2 model and the minimum BIC model also used a few different variables than the models obtained in the previous sections.

Figures 4.13, 4.14, and 4.15 show the residual, normal Q-Q, and leverage plots for each of these three models. They look essentially the same as the other plots previously examined and analyzed in the previous three sections, except that the normal Q-Q plots seem to have heavier tails on the right side of the plot. The residuals look randomly scattered, with no apparent significant patterns, and there are also a few potential high leverage points. The noticeably heavier tails in the normal Q-Q plots is concerning, which suggests that these models, despite their high adjusted R^2 values, may not be the best fits for this data set.

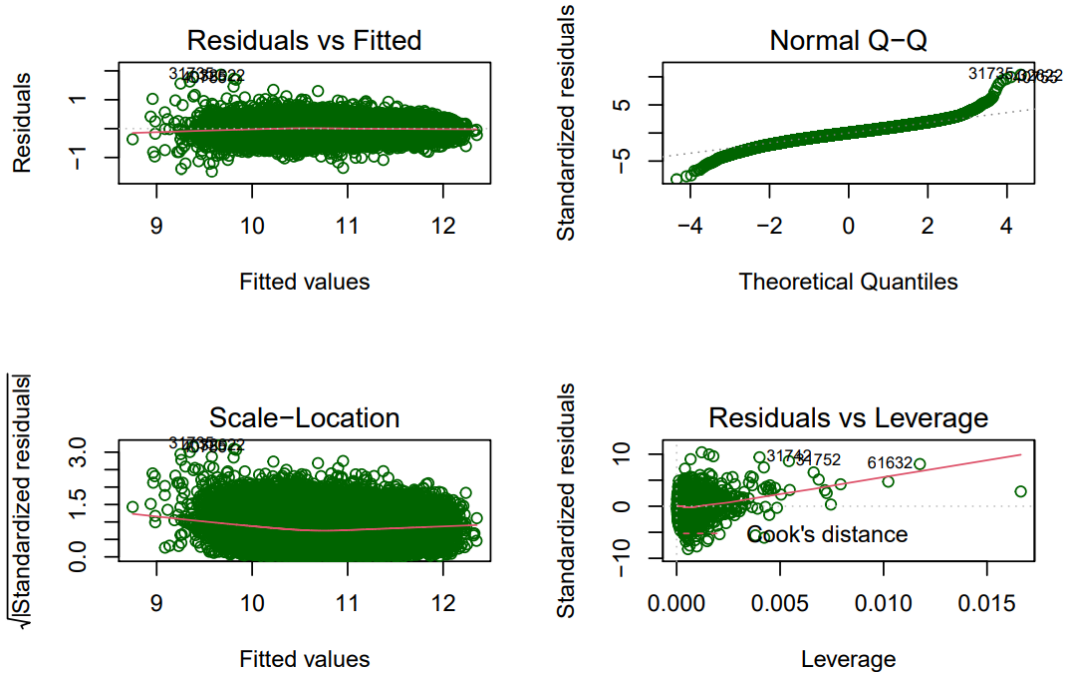

Figure 4.13. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on maximum adjusted R^2

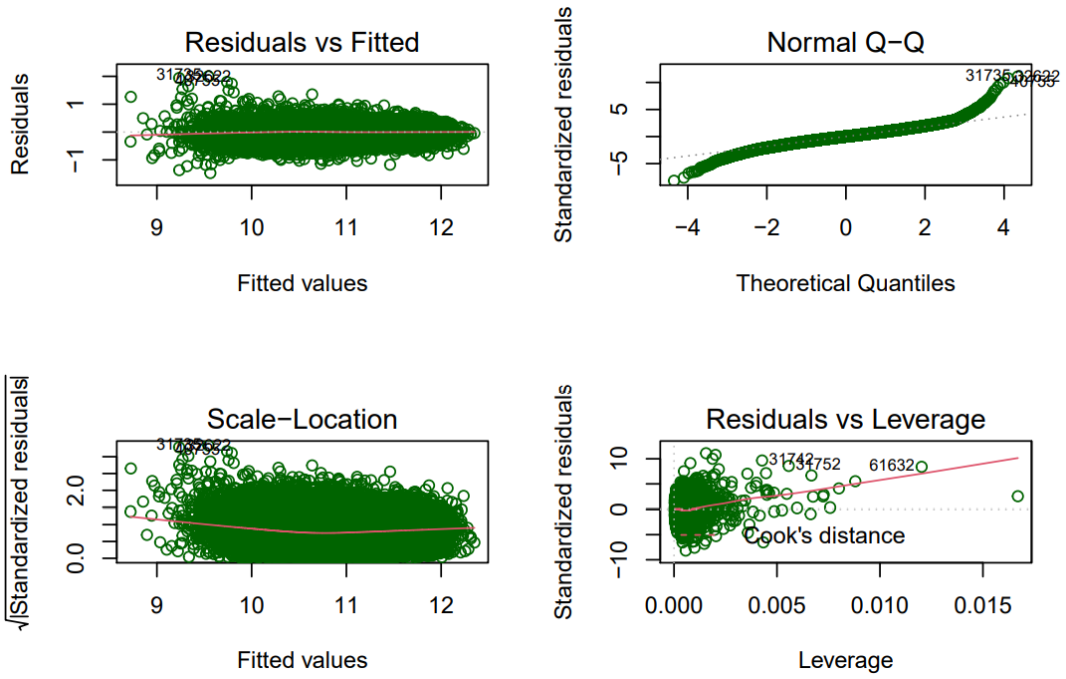

Figure 4.14. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on minimum Mallow's Cp

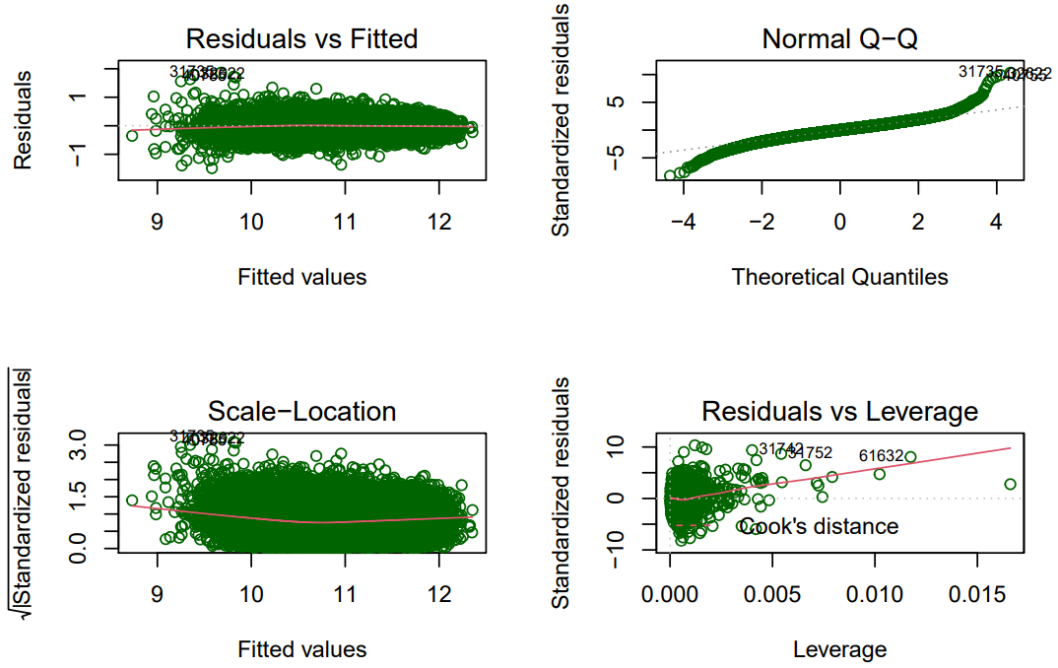

Figure 4.15. Residual, Normal Q-Q, and leverage plots for the model obtained from backward stepwise selection based on minimum BIC

4.5 Best Model Overall

From the variable selection methods in section 4 and from our model designs in section 3, we ultimately concluded that the best model obtained overall was the maximum adjusted R^2-minimum Mallow's Cp model obtained from forward stepwise selection, as it appears to be the best model of all models that we have examined. It has the highest adjusted R^2 and has the best-looking residual, normal Q-Q, and leverage plots. Although the normal Q-Q plots have some heavy tails and there are a few high-leverage plots or potential outliers, this model best represents and models our data set with the numerical variables that we have considered in this report.

5. Classification

In this section, we will perform classification analysis on our dataset. For this analysis, we will be focusing on the IncPerCap variable, which provides income per capita (that is, income per individual), rather the Income variable used in the previous sections (which provides median household income). The reasons for this decision will be discussed below. The particular classifiers examined are logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and KNN (with the value of k ranging from 1 to 10). This analysis ultimately contains two parts. First, we built models with all observations and then predicted outcomes for all observations—that is, we trained on the full dataset and then tested on the full dataset. Then, in the second part, we created a training set from the first 50% of all observations in the data and then used the remaining 50% as the testing set. We then compared the results of each of these training and testing methods to see how our models performed in these situations. Ultimately, our measures will include accuracy, sensitivity, specificity, and running time for each classification method.

Sensitivity is defined as the probability that a test will indicate a "positive" result among those who have that positive result, and specificity is the probability that a test will indicate a

"negative" result among individuals are in fact "negative" (see reference [3]). For the purposes of this report, a "positive" state will be the "high" income category, while a "negative" state will be defined as the "low" income category.

In order to perform classification, we required a binary factor variable as a response. As our data contained two categorical variables (State and County), but did not contain any binary factor categorical variables, we decided to create our own from our dataset. We decided that the best way to create such a variable was to divide our dataset into two groups: one group being at or below the federal poverty line (which we called "low") and the other group being above the poverty line (which we called "high"), where "low" and "high" reference low and high socioeconomic status. This would provide the binary response variable that we desired, as well as provide some additional information about our data.

However, precisely where the poverty line falls depends on a number of factors, including size of household and region, and there is disagreement even among different U.S. government agencies where that line is located when considering all of these factors. We ultimately located and utilized a poverty level classification scheme from 2015 provided by the U.S. Census Bureau (see reference [2]), which set the weighted average poverty level threshold for an individual at $12,082, regardless of region. We chose to use the poverty level for an individual because we did not have information in our data set about size of household, but we did have income per capita (that is, income per individual) in our data set. Hence, for this reason, in this section we will utilize income per capita to establish the binary income levels variable, rather than using median household income as we have done in the previous sections. The information on our new binary variables is shown below in Table 5.1.

Variable Name	*Description*	*Variable Type*
Income_Levels	Binary variable with two levels: "high" and "low," as determined by the poverty level income threshold	Factor
Income_Level_Binary	Binary variable with two levels: 0 (corresponding to "low") and 1 (corresponding to "high")	Categorical

Table 5.1. Summary table showing variable name, description, and type for the two newly created binary variables using the methodology described above.

5.1 Logistic Regression

The first method of classification performed was logistic regression. The logistic model, also known as a logit model, takes the following general form:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

The model for this data set was initially constructed by training on the full dataset and testing on the full dataset; for the second part of the classification analysis, the model was trained on the first 50% of all observations in the data and then used the remaining 50% as the testing set. First, what variables to include in the logistic regression model had to be determined, based on the variable selections for the models examined in the previous sections. This investigation lead to the model shown in Table 5.2, where all the variables included are statistically significant for the full data set. Unfortunately, the "Service" and Construction" variables were no longer significant when the model was trained on 50% of the data and tested on the remaining 50%, but for consistency these two variables were included in the classification procedure. The accuracy, sensitivity, specificity, and running time for the model were also determined and are recorded in Table 5.2. For brevity and to avoid redundancy, the confusion matrix from which the sensitivity and specificity values were obtained is omitted from this report.

Logistic regression model (Full data, 50% data)	**Accuracy**		**Sensitivity**		**Specificity**		**Procedure Run Time (seconds)**	
	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
Income_Level_Binary ~ Men + Women + Hispanic + Black + Native + Poverty + ChildPov + Professional + Service + Construction	97.26%	97.15%	99.08%	98.99%	65.31%	67.49%	1.21	0.42

Table 5.2. Logistic regression model, accuracy, sensitivity, and specificity obtained from training on the full data set and testing on the full data set. The percentages shown are rounded from R output.

From this output, it can be seen that both versions of the logistic regression have very high accuracy and sensitivity, and specificity is moderately high. We can conclude, then, that the logistic regression model in both cases has very high prediction accuracy, and furthermore, based on our earlier definitions of sensitivity and specificity, we can see that the model has a very strong ability to predict the census tracts that fall within the "high" income classification based on income per capita. Its ability to predict the census tracts that fall within the "low" income classification is not as powerful, but is nonetheless still quite high. Between the full model and the 50% data model, we can see that the prediction accuracy and the sensitivity are essentially equal, but there is an over 2% increase in sensitivity. The produce also ran much quicker for the 50% model than for the full data model. This suggests that training the model on 50% of the data and testing on the

remaining 50% of the data has several advantages over the other method of modeling and would likely be the best choice for developing the logistic model.

5.2 Linear Discriminant Analysis (LDA)

Next, linear discriminant analysis (LDA) was performed on the data set. Because, from the exploratory data analysis, we can see that this data is highly skewed, it seems likely that LDA may not model the data well due to violations of key assumptions. For example, LDA assumes that each class predictor has a normal distribution with the same variance, but as we can see from the graphs in Figures 5.1 and 5.2, the distribution of the class called group 0 (which under the definition described earlier is the "low" income group) appears to be normally distributed, but the distribution of the class called group 1 (the "high" income group) looks left-skewed. Figure 5.1 depicts the LDA results from performing this procedure on the full dataset, while the results in Figure 5.2 are from LDA as performed on the training set of the first 50% of the data.

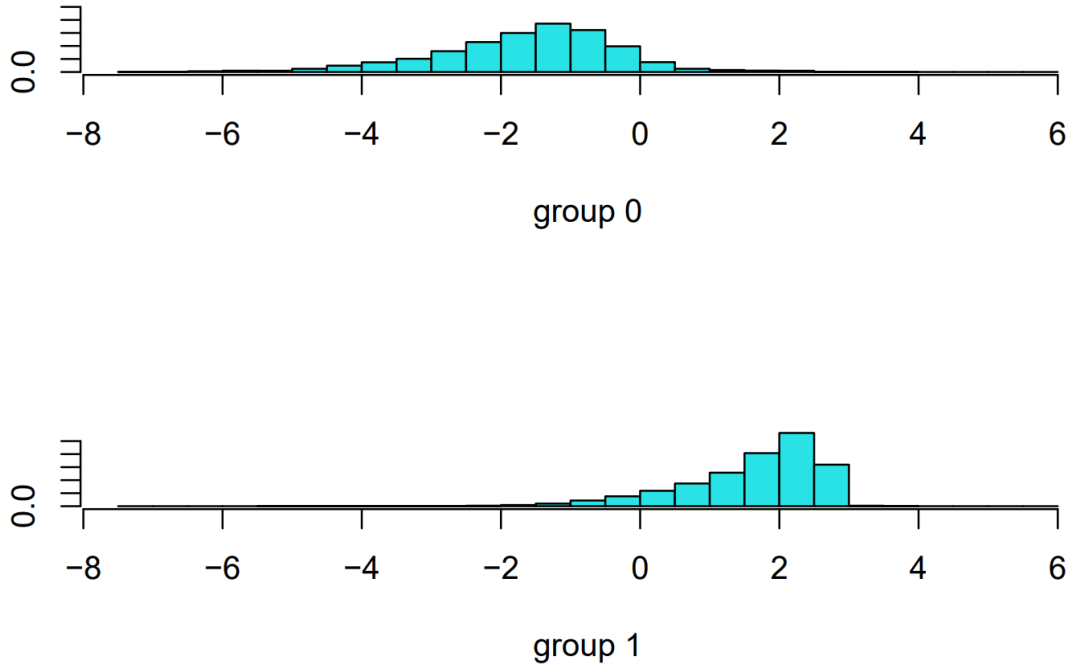

Figure 5.1. Graphs of the distributions of each class: group 0 ("low" income group) and group 1 ("high" income group) as obtained from LDA run on the full data set.

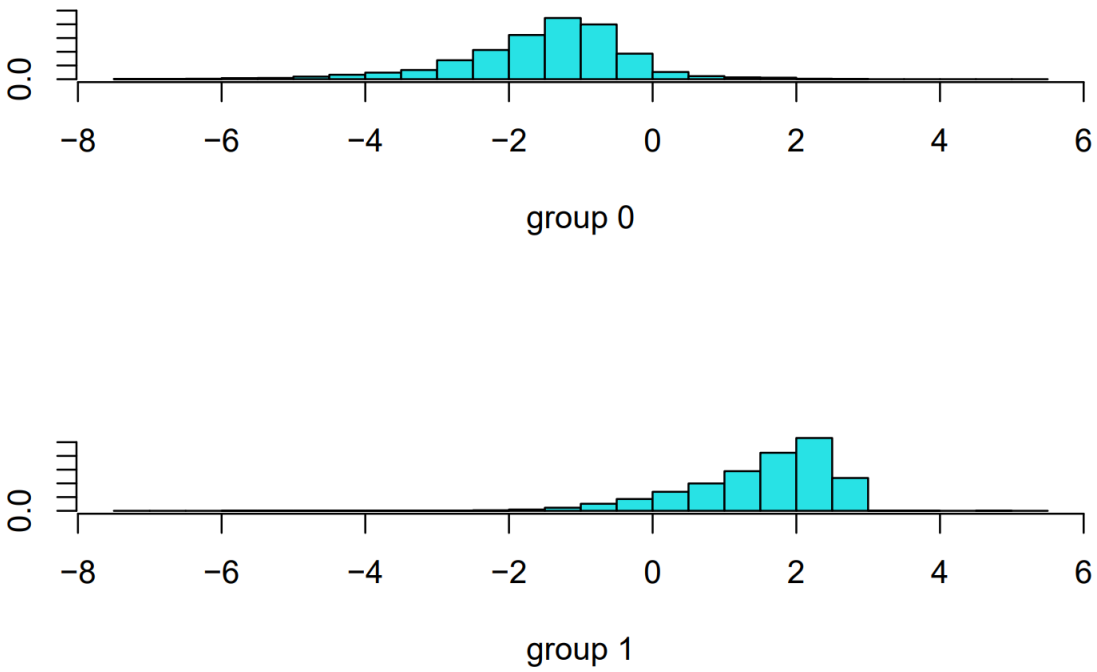

Figure 5.2. Graphs of the distributions of each class: group 0 (“low” income group) and group 1 (“high” income group) as obtained from LDA run on the data training set of the first 50% of observations.

Despite these violations of assumptions, we chose to proceed with LDA in order to obtain a comprehensive classification analysis for our data. As with logistic regression, we performed LDA on both the full data set and on training and testing sets that comprised 50% each of the data. Also, in order to remain consistent, the same model from logistic regression was used in LDA. The results of the LDA are shown in Table 5.3. The accuracy, sensitivity, specificity, and running time for the model were also determined and are recorded in Table 5.3. For brevity and to avoid redundancy, the confusion matrix from which the sensitivity and specificity values were obtained is omitted from this report.

Accuracy		Sensitivity		Specificity		Procedure Run Time (seconds)	
Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
96.35%	96.41%	97.80%	97.51%	70.96%	78.64%	0.51	0.29

Table 5.3. Accuracy, sensitivity, specificity, and procedure run time for LDA on the full dataset and on the training and testing sets. The percentages shown are rounded from R output.

Interestingly, despite the previously-described violation of an assumption for LDA, the results in Table 5.3 show that LDA still performed very well, with high accuracy and sensitivity and moderately high specificity. Its running time was also much quicker than that of logistic regression; even when running with the full data set, the procedure needed only around half a second to complete. This procedure's specificity was also much higher than that of logistic regression, and there was a large increase in specificity (about 8%) when the classification procedure was utilized with the 50% training data and 50% testing data.

5.3 Quadratic Discriminant Analysis (QDA)

Based on the violations of assumptions observed when performing LDA, it seems possible that quadratic discriminant analysis (QDA) may provide a better predictive analysis of this data set. Performing QDA on both the full data set and on training and testing sets that comprised 50% each of the data yielded the results shown below in Table 5.4. Again, the same model from logistic

regression and LDA was used here to keep results consistent and allow for comparisons among the different classifiers.

Accuracy		**Sensitivity**		**Specificity**		**Procedure Run Time (seconds)**	
Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
94.92%	95.09%	95.98%	96.46%	76.33%	73.11%	0.41	0.18

Table 5.4. Accuracy, sensitivity, specificity, and procedure run time for QDA on the full dataset and on the training and testing sets. The percentages shown are rounded from R output.

Comparing the results of QDA, shown above in Table 5.4, with the results from LDA, QDA has a slightly lower accuracy, sensitivity, and specificity than LDA, whiling also having a marginally faster run time than LDA. However, we must remember that since certain assumptions for LDA were violated, despite its high accuracy, sensitivity, and specificity, LDA is not a good fit for the data. Overall, QDA performs very well for this data, both when training and testing on the full dataset and when using 50% of the data for training and 50% for testing. It has high accuracy and sensitivity, and its specificity is also high.

5.4 K-Nearest Neighbors (KNN)

The last classification procedure used is K-nearest neighbors (KNN), with k values ranging from 1 to 7. Although KNN was performed for k-values from 1 to 7, for brevity only the results from 3 to 7 are included in Table 5.5.

KNN k-value	Accuracy		Sensitivity		Specificity		Procedure Run Time	
	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
k = 3	97.24%	94.64%	99.49%	98.94%	57.75%	25.67%	1.81 min	25.19 sec
k = 4	96.68%	94.48%	99.34%	98.97%	50.04%	22.58%	2.07 min	30.09 sec
k = 5	96.59%	94.60%	99.58%	99.35%	44.24%	18.41%	1.76 min	17.02 sec
k = 6	96.31%	94.53%	99.53%	99.41%	39.91%	16.25%	2.03 min	17.20 sec
k = 7	96.20%	94.54%	99.65%	99.55%	35.71%	14.24%	2.12 min	17.42 sec

Table 5.5. Accuracy, sensitivity, specificity, and procedure run time for KNN, k from 3 to 7 on the full dataset and on the training and testing sets. The percentages shown are rounded from R output.

The two most obvious issues with the KNN procedure are the rather low specificity, which worsened as the value of k increased, and the somewhat lengthy running times. Based on the earlier definition of specificity, this means that the KNN classification procedure does not work well when trying to predict whether a census tract falls into the "low" income category based on income per capita. Although KNN's specificity was moderately high for small k values when testing and training on the full data set, the sensitivity worsened when training on the first 50% of the data and testing on the last 50% of the data, with its worst specificity performance at k = 7, with a specificity of 14.24%. However, its very high sensitivity indicates that it does perform well on census tracts in the "high" income category. The running time is the other issue with the KNN classification procedure; although training on 50% of the data did help to cut back significantly on the procedure's run time, its run times were still much longer than the other three procedures examined in the procedures, which could run as fast as fractions of a second.

5.5 Summary and Conclusions

	Accuracy		Sensitivity		Specificity		Procedure Run Time	
	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
Logistic Regression	97.26%	97.15%	99.08%	98.99%	65.31%	67.49%	1.21 sec	0.42 sec
LDA	96.35%	96.41%	97.80%	97.51%	70.96%	78.64%	0.51 sec	0.29 sec
QDA	94.92%	95.09%	95.98%	96.46%	76.33%	73.11%	0.41 sec	0.18 sec
k = 3	97.24%	94.64%	99.49%	98.94%	57.75%	25.67%	1.81 min	25.19 sec
k = 4	96.68%	94.48%	99.34%	98.97%	50.04%	22.58%	2.07 min	30.09 sec
k = 5	96.59%	94.60%	99.58%	99.35%	44.24%	18.41%	1.76 min	17.02 sec
k = 6	96.31%	94.53%	99.53%	99.41%	39.91%	16.25%	2.03 min	17.20 sec
k = 7	96.20%	94.54%	99.65%	99.55%	35.71%	14.24%	2.12 min	17.42 sec

Table 5.6. Summary table of results from logistic regression, LDA, QDA, and KNN, provided in one table for comparison.

After performing four classification methods—logistic regression, LDA, QDA, and KNN—the results of all of these classifiers were compiled in a single table, Table 5.6 (above), for comparison. These results are also summarized graphically in Figure 5.3 and Figure 5.4, with side-to-side boxplots. Figure 5.3 depicts the accuracy, sensitivity, and specificity of each classifier utilizing the full dataset, while Figure 5.4 illustrates the accuracy, sensitivity, and specificity of each classifier when the first 50% of the dataset is used as the training set and the remaining 50% of the dataset is used as the testing set. For conciseness, the only k-value for KNN that is included in the side-to-side bar plots is k = 3.

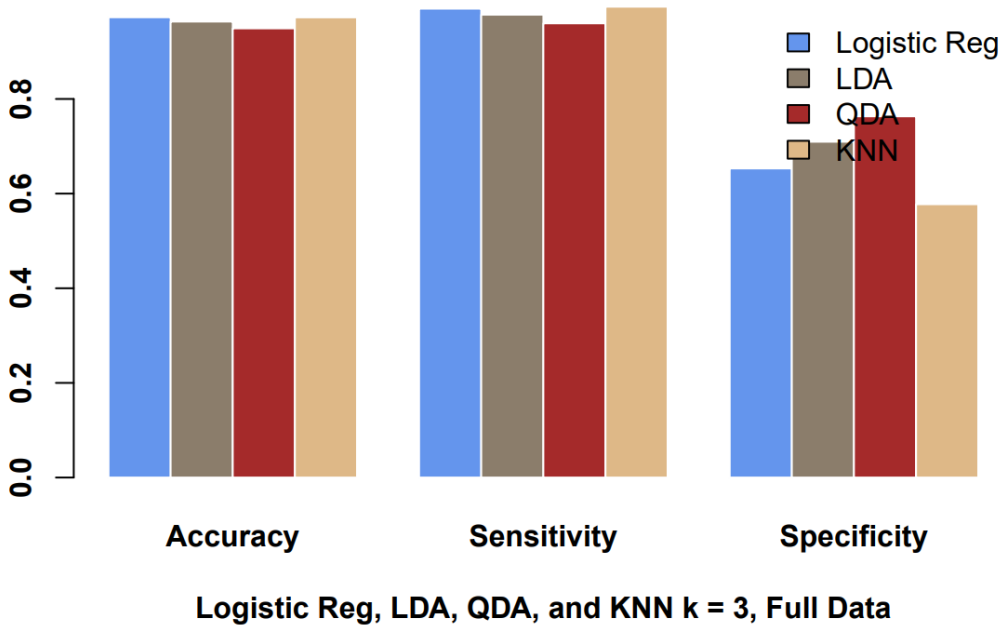

Figure 5.3. Side to side bar chart of accuracy, sensitivity, and specificity of logistic regression, LDA, QDA, and KNN (k = 3), with full data.

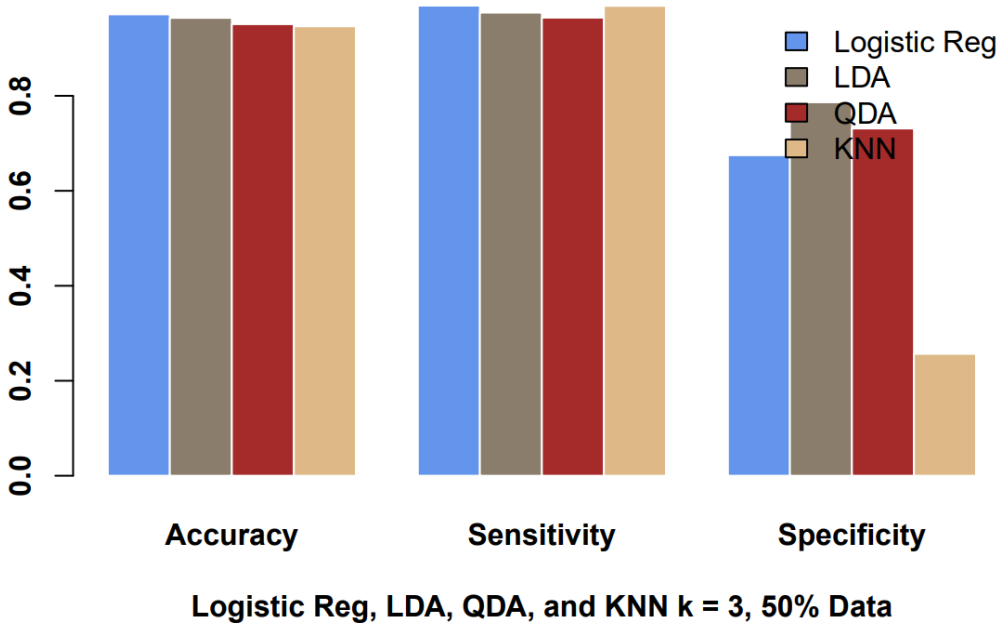

Figure 5.4. Side to side bar chart of accuracy, sensitivity, and specificity of logistic regression, LDA, QDA, and KNN (k = 3), with 50% training and 50% testing data.

Based on the results obtained in Table 5.6 and Figures 5.3 and 5.4, the classifier with the highest overall accuracy and sensitivity was the logistic regression model analyzed in section 5.1. Although this classifier did not have the highest specificity—LDA performed on 50% of the data had the highest specificity—and it was not the quickest classification method—QDA performed on 50% of the data had the quickest runtime—the accuracy of the logistic regression model and its very high sensitivity offered the best results of all the classifiers. Its runtime was also not that much slower than QDA and was still extremely fast. The runtime was especially improved when logistic regression was run on 50% of the data, as it ran to completion after a mere half a second. Hence, the logistic regression model run on 50% of the data was the best classification method for this data set, since it had the highest accuracy and the highest sensitivity.

6. Resampling

For this section, classification will be performed using 5-fold cross validation with eight classifiers: logistic regression, LDA, QDA, KNN, a decision tree, bagging, random forest, and boosting. As in the previous section, the prediction accuracies, specificities, sensitivities, and running times of all methods will be summarized and compared, as well the standard errors for all classifiers. Both numerical and graphical comparisons will be utilized for these eight classifiers. For the categorical response variable, the binary income levels variable whose creation was described in section 5 will again be utilized here for classification with 5-fold cross validation.

In addition, regression analysis will also be re-performed using KNN, a decision tree, bagging, random forest, and boosting, since these methods can not only be used for classification but also for regression. Since regression requires a numerical response variable, for this portion of the analysis we will return to using the Income variable, which as described in the exploratory data

analysis in section 2 provides the median income for a census tract. We will then compare the results from these regression methods to the results obtained in parts 3 and 4 of this report.

6.1. Classification with All Eight Classifiers

First, 5-fold cross validation was performed with all eight classifiers described in the introduction to this section. For consistency, the same variables used in section 5 to build the models were also utilized here. Also in keeping with the precedent used in section 5, 50% of the data was used as a training set, while the remaining 50% of the data was the test set.

KNN k value	Accuracy
k = 1	94.24%
k = 2	94.05%
k = 3	94.83%
k = 4	94.77%
k = 5	94.90%
k = 6	94.86%
k = 7	94.88%
k = 8	94.85%
k = 9	94.82%
k = 10	94.79%

Table 6.1. Numerical summary of accuracy measurements for KNN, k from 1 to 10.

For KNN, some additional preparatory work had to be done before its performance could be compared to the other classifiers. Because we wished to know the best k value for KNN, a procedure was run in R in order to make this determination, testing values of k from 1 to 10. The best k value was ascertained by performing 5-fold cross validation for each k value and finding

the overall accuracy for each value of k. The accuracy results are shown above in Table 6.1, as well as summarized graphically in Figure 6.1 below. Based on the numerical and graphical summaries, it can be seen that the k value with the highest overall accuracy was k = 5. Hence, using this value of k, another KNN procedure was run with 5-fold cross validation in order to record accuracy, standard error, sensitivity, specificity, and run time, to be compared with the other classifiers.

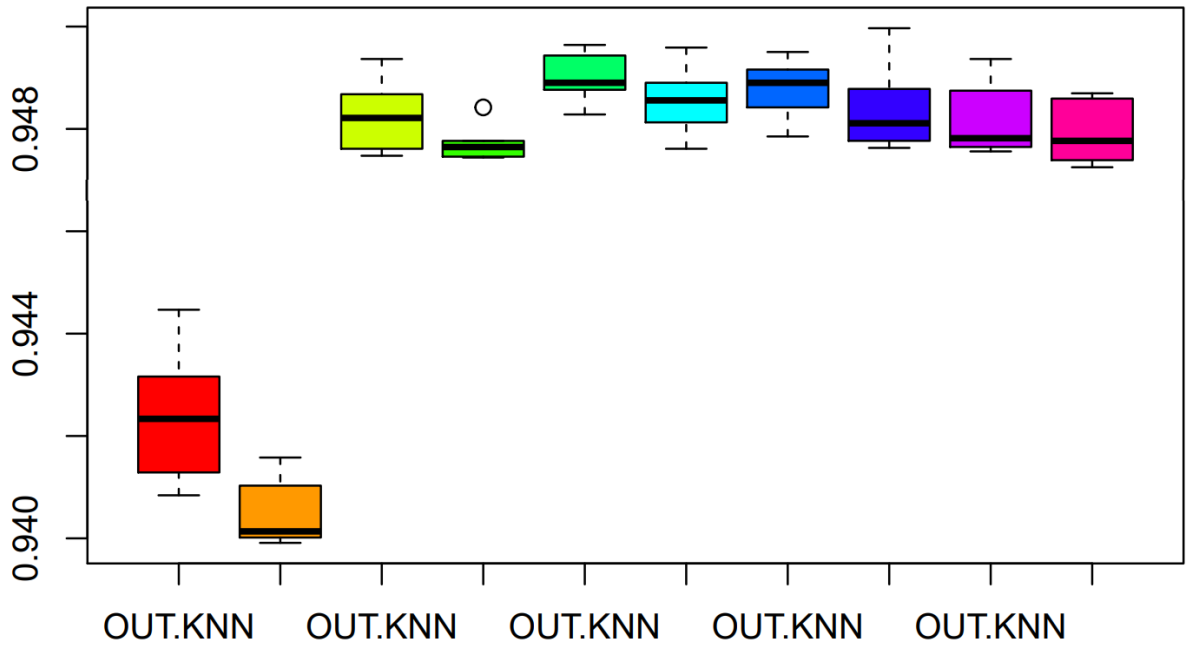

Figure 6.1. Side-to-side boxplots of accuracy measurements of KNN k values from 1 to 10 (also numerically recorded in Table 6.1).

Five-fold cross validation was also performed on the other seven classifiers, with overall prediction accuracy, standard error, specificity, sensitivity, and running time calculated for all of them. These results are summarized below in Table 6.2. For brevity's sake, the only k value recorded for KNN is k = 5, since per the previous analysis it was found to be the best k value with the highest overall accuracy.

	Accuracy	Standard Error	Sensitivity	Specificity	Procedure Run Time
Logistic Regression	97.23%	0.00146	99.07%	64.93%	2.83 sec
LDA	96.33%	0.00071	97.79%	70.68%	1.50 sec
QDA	94.92%	0.00077	95.98%	76.48%	1.17 sec
KNN k = 5	94.90%	0.00054	99.44%	15.33%	28.93 sec
Decision Tree	96.66%	0.00163	98.52%	64.11%	2.40 sec
Bagging	97.25%	0.00026	98.78%	70.57%	5.62 min
Random Forest	96.77%	0.00156	98.74%	62.13%	4.38 min
Boosting	96.77%	0.00156	98.74%	62.13%	4.81 min

Table 6.2. Summary table of results from logistic regression, LDA, QDA, KNN, decision tree, bagging, random forest, and boosting, provided in one table for comparison.

Based on the results shown in Table 6.2, it can be seen that logistic regression had the highest accuracy of all eight classifiers, as also obtained during the classification analysis in section 5. Logistic regression also had the highest sensitivity, at 99.07%, also as obtained in section 5. QDA performed with the highest specificity at about 76%. These results can also be seen in Figure 6.2 below, which shows the accuracy, sensitivity, and specificity of the classifiers. We can also see from the figure that the accuracy and sensitivity values for the eight classifiers were always high and very close together, though logistic regression outperformed the other classifiers in both areas. Ultimately, though, the average accuracy for all the classifiers examined was very high;

even the lowest accuracy, obtained from QDA, was about 95%. There tends to be more variability and more lackluster performance when considering specificity, and KNN with k = 5 had the lowest specificity at around 15%.

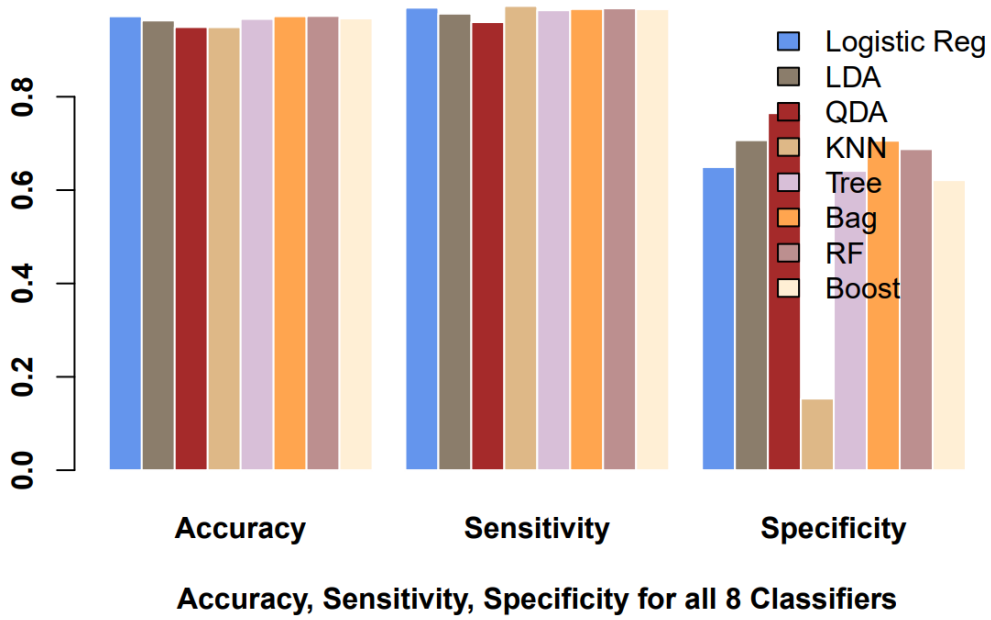

Figure 6.2. Side-to-side bar graph of accuracy, sensitivity, and specificity for all eight classifiers.

The standard error values for all of the classifiers were very low, though the classifier with the smallest standard error was bagging, while the decision tree had the highest standard error. This can also be seen in Figure 6.3, which shows a bar plot of the standard errors for the eight classifiers. LDA, QDA, and KNN also tended to have small standard errors compared to the other classifiers. However, although bagging had the lowest standard error, it also had the longest runtime at around 7 minutes. Random forest and boosting also had lengthy runtimes at around five and a half minutes; the long runtimes for these three classifiers likely stemmed from the large dataset used for this project. The lengthy running times for these three classifiers is especially apparent in Figure 6.4, which graphically depicts the run times for all eight classifiers in seconds. The run times for logistic regression, LDA, QDA, and the decision tree are superior to those obtained for KNN, bagging, random forest, and boosting.

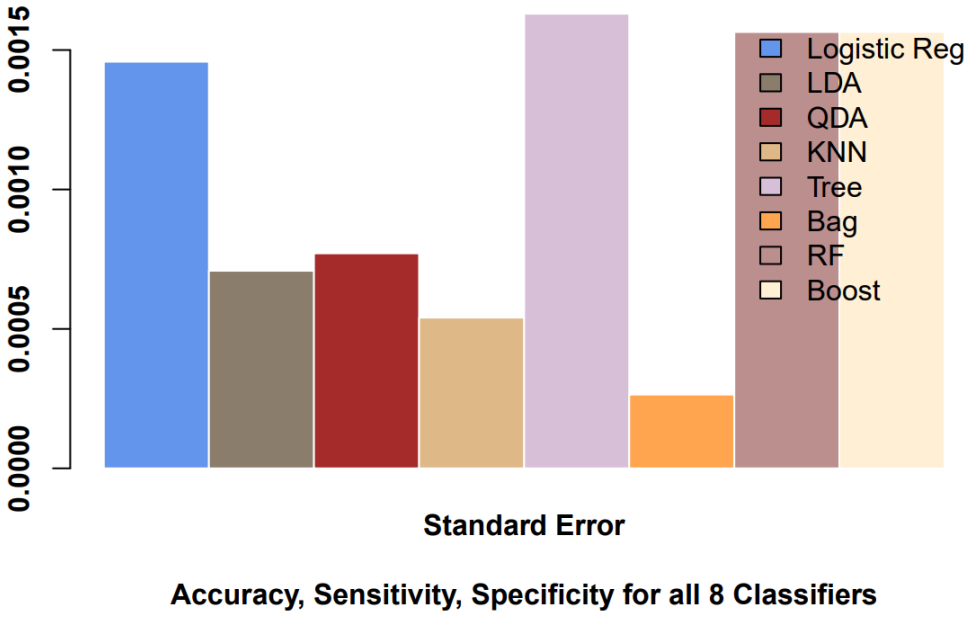

Figure 6.3. Bar plot of the standard error measurements obtained for each classifier.

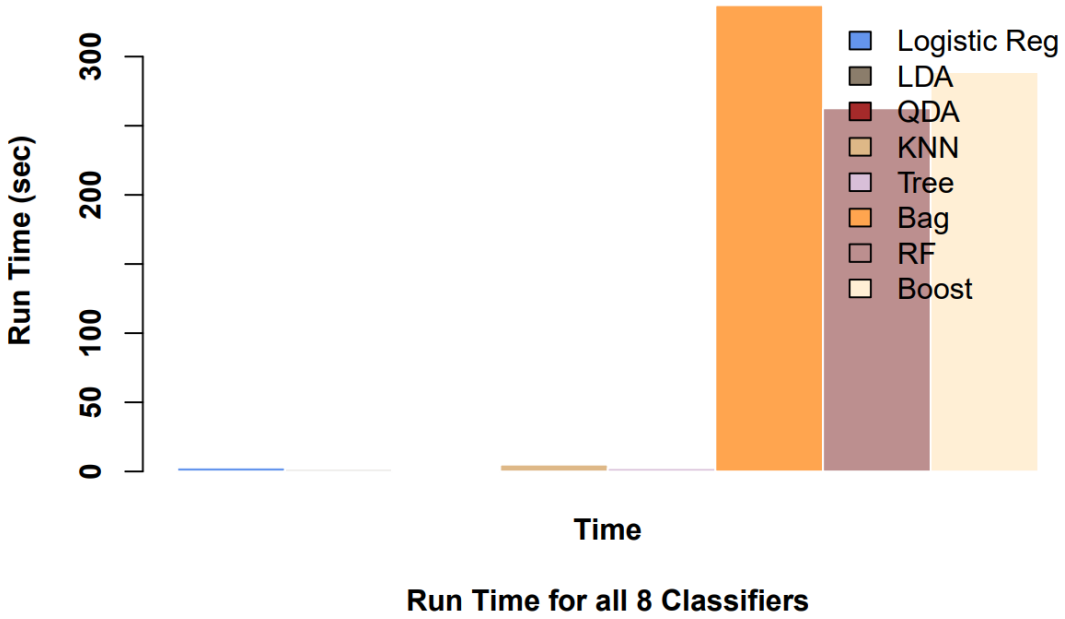

Figure 6.4. Side-to-side bar plot of running times (in seconds) for each classifier.

To further compare the accuracy of the eight classifiers, a side-to-side box plot was created to summarize and compare the accuracy results. This graph is shown in Figure 6.4 and indicates

the same conclusions as before. Logistic regression has the highest overall accuracy, and its high sensitivity and efficient performance suggest that this classifier is the best choice for the dataset.

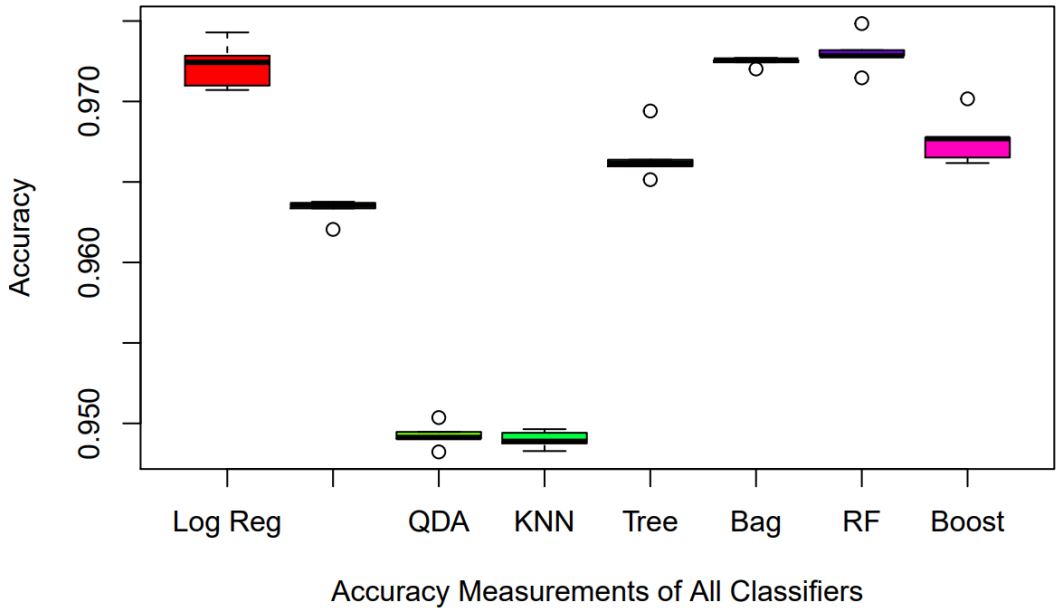

Figure 6.5. Side-to-side boxplot of accuracy measurements for each classifier.

6.2. Regression Revisited: KNN, Decision Tree, Bagging, Random Forest, Boosting

In section 6.1, we performed additional classification analysis by performing 5-fold cross validation with eight classifiers. In this section, we will revisit regression analysis. As KNN, the decision tree, bagging, random forest, and boosting can be utilized for regression as well as for classification, we will be using these five methods to perform additional regression analysis. For this reason, the numerical Income variable, which reports the median income in a census tract (see section 2), will be used as the response variable for the regression models. In addition, for consistency with how the classification models have been implemented, the same variables will be used for all models, except that the Income variable is being used as the response variable.

Since in part 2 of this report it was determined that linear models were not good fits for the data due to the skewedness of the data, we will be comparing the regression results obtained here with the two best models found from our previous regression modeling: the polynomial type of

employment model found in part 3 and the non-linear, logarithmic model created from maximum adjusted R^2 and minimum Mallow's Cp found with forward stepwise selection in part 4. Comparisons will be made based on procedure runtime and mean squared error (MSE) of the models. No linear models will be considered in this analysis.

For KNN, so that this procedure could be compared with the other regression procedures, the best k was sought. In this case, since we are working with regression, we want to find the k value that minimizes MSE. From an examination of Table 6.3 and Figure 6.6, it can be seen that the lowest MSE was obtained with k = 6 (though ultimately all MSE values obtained were very high). Hence, we treated this as our "best k," and we implemented regression with KNN separately using k = 6 in order to record run time and MSE for comparison with the other regression procedures.

KNN k value	**MSE**
k = 1	614745161
k = 2	478104948
k = 3	439377429
k = 4	424743221
k = 5	419887722
k = 6	419355153
k = 7	420700824
k = 8	423656520
k = 9	427315207
k = 10	430987732

Table 6.3. Numerical summary of MSE measurements for KNN, k from 1 to 10.

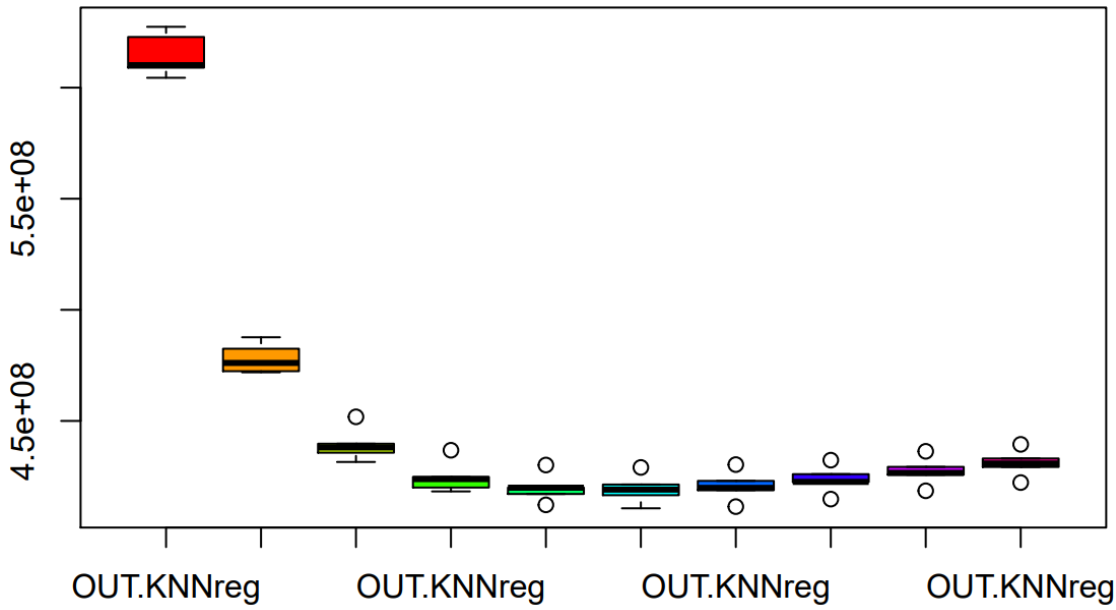

Figure 6.6. Boxplot of MSE measurements obtained from KNN for values of k from 1 to 10.

Procedure	Mean Squared Error (MSE)	Run Time
KNN, k = 6	419355153	1.43 sec
Decision Tree	237339549	1.51 sec
Bagging	159856095	3.59 hours
Random Forest	157407190	2.05 hours
Boosting	159844672	6.54 min

Table 6.4. Numerical summary of MSE and run time for KNN, decision tree, bagging, random forest, and boosting regression procedures.

Once the preparatory work was completed for KNN, the other regression procedures were implemented using 5-fold cross validation, and the run time for each procedure and the average MSE values were recorded for the five folds. The results are shown above in Table 6.4. Perhaps the most startling results are the astronomical run times required for bagging and random forest, at about three and a half hours and nearly two hours, respectively. The computing power required

to implement these procedures was the most challenging aspect of this analysis, because the large size of our data set caused these procedures to be highly expensive and time-consuming operations. This indicates that although random forest and bagging had the two lowest MSE values, their run times are prohibitive, making them highly inefficient regression methods for this data set. From Table 6.4, discounting bagging and random forest, it seems that the decision tree regression method offers the best alternative. Although it did not have the lowest MSE, it had the quickest run time and still had the third-lowest MSE. Ultimately, however, all the procedures considered here had extremely high MSE values. The graphical representations of these results can be seen in Figures 6.7 and 6.8 below. In particular, Figure 6.7, showing the run time in minutes for the five regression procedures, shows how the run time for bagging and random forest dwarfs the run times of the other procedures.

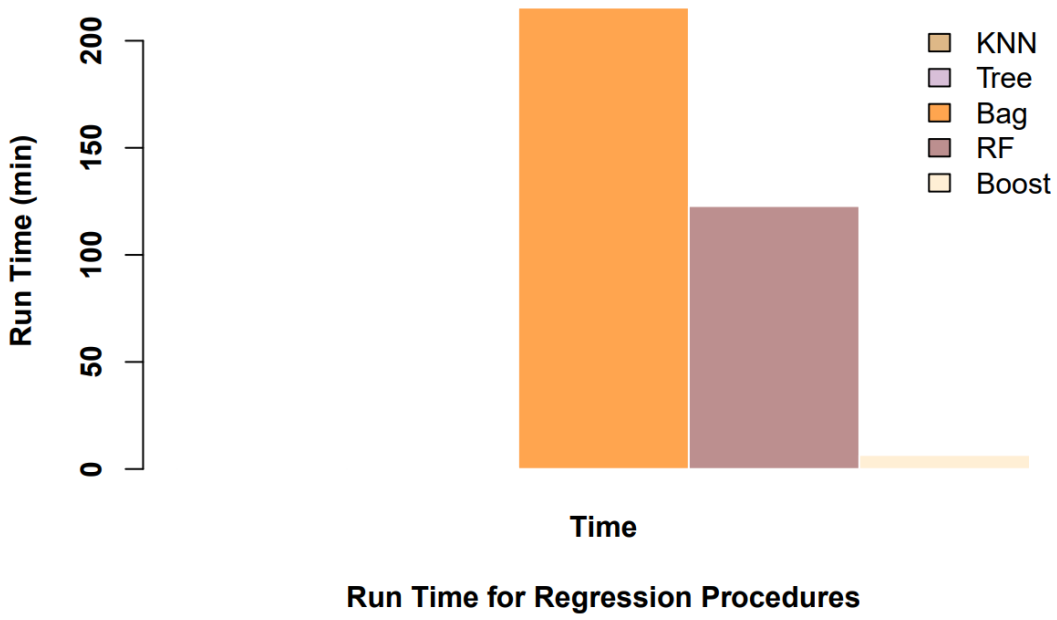

Figure 6.7. Side-to-side bar graph of run time for KNN, decision tree, bagging, random forest, and boosting for regression.

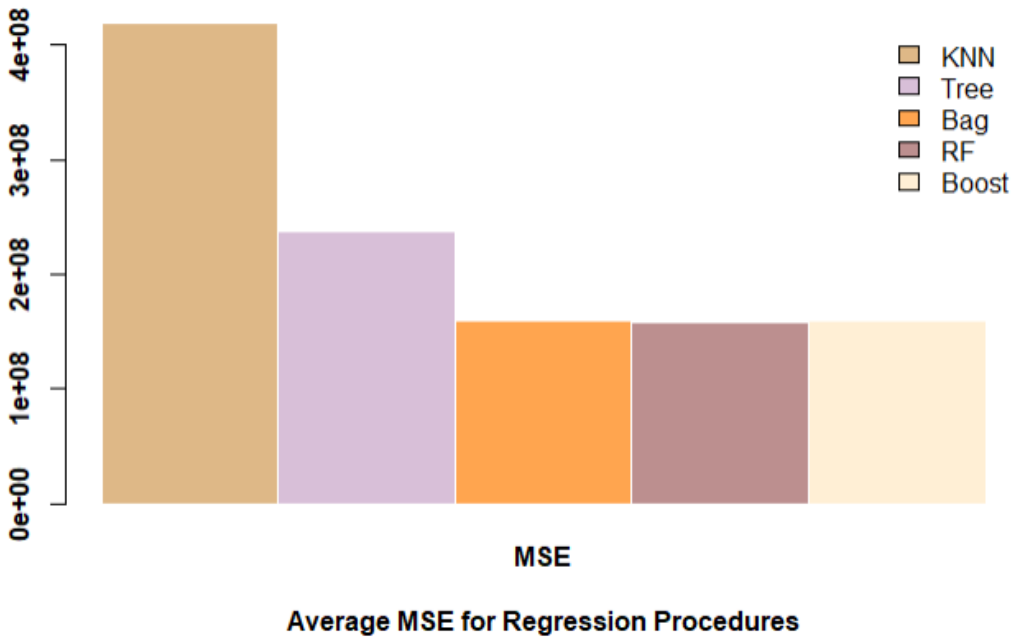

Figure 6.8. Side-to-side bar graph of average MSE (averages across the five folds from 5-fold cross validation) for KNN (k=6), decision tree, bagging, random forest, and boosting.

Figure 6.8, above, illustrates that bagging, random forest, and boosting had the lowest MSE values, but since these procedures had lengthy runtimes, it can be seen that the decision tree is the next best regression alternative for this data set.

6.3 MSE and Run Time Comparisons with Polynomial and Non-Linear Modeling

In this section, we will compare the regression techniques from KNN, the decision tree, bagging, random forest, and boosting with the third-degree type of employment polynomial model from part 3 and the non-linear logarithmic model from part 4 of this report. These were found to be the two best models based our preliminary modeling and our variable selection methods. For reference, these two models are reproduced below as Table 6.5. The analysis and comparison of these regression methods will be based on MSE and run time, so to obtain the MSE and run time for the two regression procedures shown in Table 6.5, these models were run again in R. This generated the results for MSE and run time shown in Table 6.6.

Model	R-Squared	Adj. R-Squared
Polynomial: Professional + Service + Office + Construction + Production + I(Professional^2) + I(Service^2) + I(Office^2) + I(Construction^2) + I(Production^2) + I(Professional^3) + I(Service^3) + I(Office^3) + I(Construction^3) + I(Production^3)	0.5964	0.5963
Non-linear logarithmic: log1p(Men) + log1p(Women) + log1p(Hispanic) + log1p(White) + log1p(Black) + log1p(Native) + log1p(Asian) + log1p(Pacific) + log1p(IncPerCap) + log1p(Poverty) + log1p(ChildPov) + log1p(Professional) + log1p(Service) + log1p(Office) + log1p(Construction)	0.8641	0.864

Table 6.5. Reproduction of two models considered in previous regression analysis.

Procedure	Mean Squared Error (MSE)	Run Time
Polynomial	331612785	0.074 sec
Non-Linear	4098956405	0.081 sec
KNN, k = 6	419355153	1.43 sec
Decision Tree	237339549	1.51 sec
Bagging	159856095	3.59 hours
Random Forest	157407190	2.05 hours
Boosting	159844672	6.54 min

Table 6.6. Numerical summary of MSE and run time for polynomial model, non-linear model, KNN, decision tree, bagging, random forest, and boosting regression procedures.

When all of these models are compared, it can be seen that the decision tree regression technique offers the lowest MSE. The polynomial modeling procedure was the quickest, with a

runtime of less than one second, but the MSE of the decision tree is much lower, and its run time is not that much slower at around two seconds. We can again see that bagging and random forest are highly inefficient regression techniques for this data set, and so we will discount them. In fact, since run time comparisons among the polynomial, logarithmic, and the other more efficient regression models is too difficult to depict graphically with bagging and random forest included in the bar chart (see Figure 6.9 below), an additional bar chart was created with bagging, random forest, and boosting omitted (see Figure 6.10). As can be seen from Figure 6.10, the decision tree, logarithmic, and polynomial regression models had the fastest run times. Since the decision tree regression model also had the lowest MSE (when not considering bagging and random forest), it seems that this regression procedure is the best one for our data set.

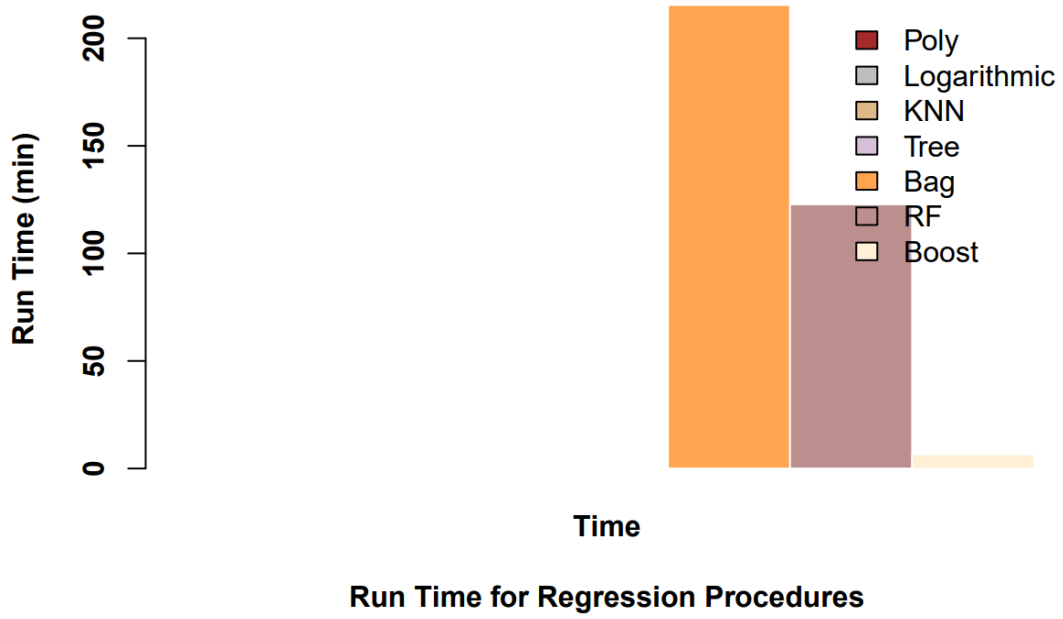

Figure 6.9. Bar plot of run times for all seven regression procedures. Notice that the only visible bars are for bagging, random forest, and boosting due to their lengthy run times.

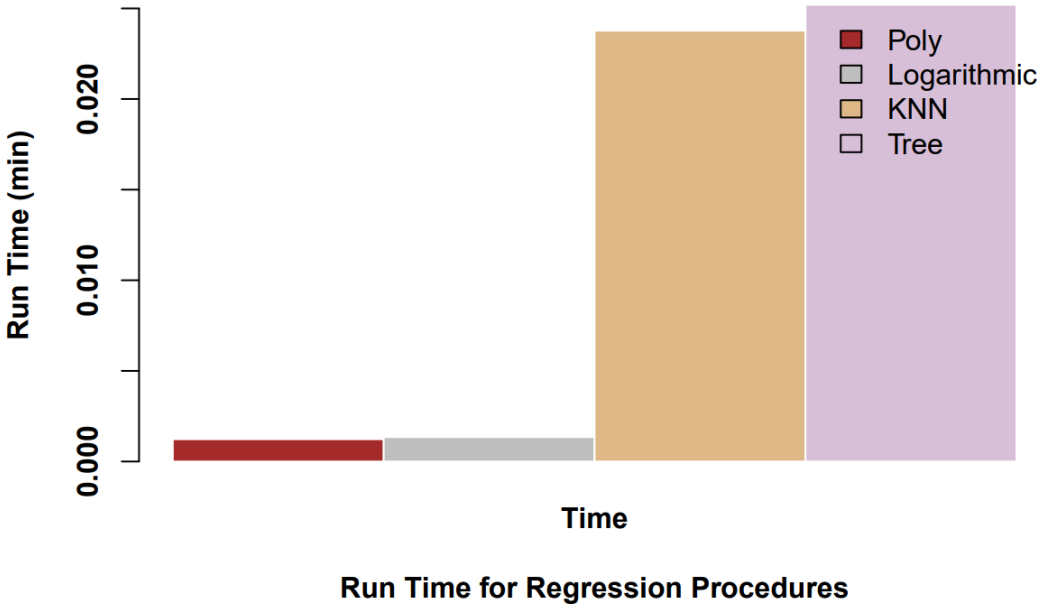

Figure 6.10. Bar graph of run times for the polynomial, logarithmic, KNN, and decision tree regression procedures.

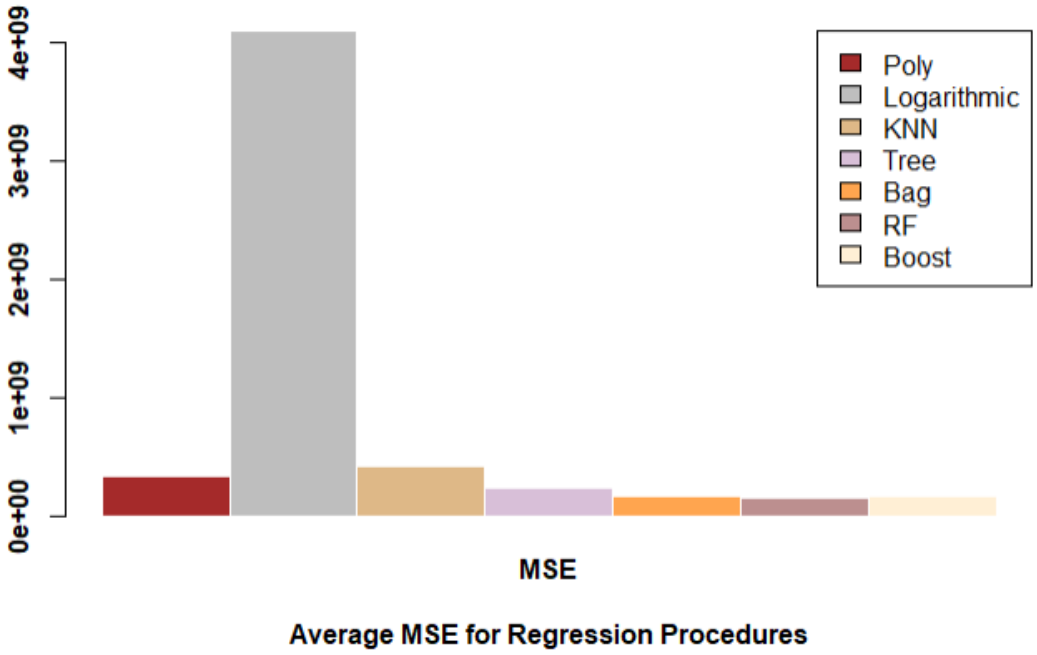

Figure 6.11. Bar graph of average MSE (averages across the five folds from 5-fold cross validation) for polynomial model, logarithmic model, KNN (k=6), decision tree, bagging, random forest, and boosting.

6.4 Summary and Conclusions

From the classification analysis performed with five-fold cross validation, logistic regression is the best classifier for this data set. It has the highest accuracy and sensitivity, has relatively high specificity, and has a quick run time. A potential problem encountered during classification analysis was some lackluster specificity values. This means that our models had difficult predicting which census tracts fell into the “low” income classification. To improve the specificity of the logistic regression model, a higher probability threshold might improve logistic regression’s specificity and thus improve its prediction accuracy for census tracts in the “low” income classification. The large size of the data set makes bagging, random forest, and boosting inefficient classifiers in this situation. The same is also true of regression, where the run time significantly worsened for bagging and random forest. For regression, it was found that the decision tree offered the best regression model for the data set, based on its run time and MSE, and it was run time that provided the largest difficulty in this section of our analysis. The other potential issue found with this regression analysis is the rather large MSE values, which suggests that underfitting occurred. Future investigation should take this result into account to improve modeling and variable selection. Ultimately, we found that the decision tree offered the best regression technique for our data set, since it had the lowest MSE (when bagging and random were removed from consideration) and is an efficient procedure.

7. Conclusions and Discussion

Through much analysis and testing for many different models—both regression and classification—we ultimately found regression and classification models that were effective in predicting income. For regression, we ultimately found that a decision tree offered the best predictive tool for income, and with classification, our logistic regression model was highly

successful and accurate at predicting the income level of a census tract. Our results indicated that our models do need some refinements but are overall accurate at predicting income and income levels. Ultimately, our best models included a variety of different variables, such as gender, race, type of employment, and poverty levels, which shows that not any one factor can be used to predict income. Rather, a variety of factors can contribute to one's income and to the income and poverty levels in an area as a whole. Future research in this area should include some investigation of other socioeconomic factors not included in this data set, such as percentage of single parents in a census tract (which from other research is strongly correlated to poverty and low incomes) and education levels (such as percentage of individuals with a college degree, percentage of individuals who are skilled tradespeople, and percentage of individuals who never graduate high school) in a census tract. These factors may reveal additional insights in how to predict income and income levels in the United States.

8. References

[1] M. (2019). [Income data from U.S. Census statistics, 2015.]. Unpublished raw data. https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2015_county_data.csv

[2] U.S. Census Bureau. [Poverty Thresholds for 2015 by Size of Family and Number of Related Children Under 18 Years]. (2020, August 21). Unpublished raw data. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html.

[3] Pennsylvania State University. [Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive value]. https://online.stat.psu.edu/stat507/lesson/10/10.3.