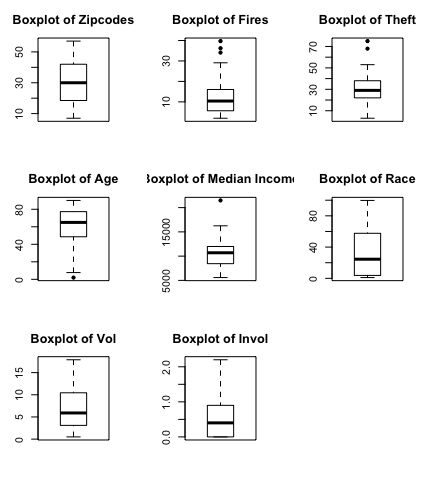
Dan DeFronzo, Kevin Falk, Max Abrams, Raj Rai

**Overview of Project:**

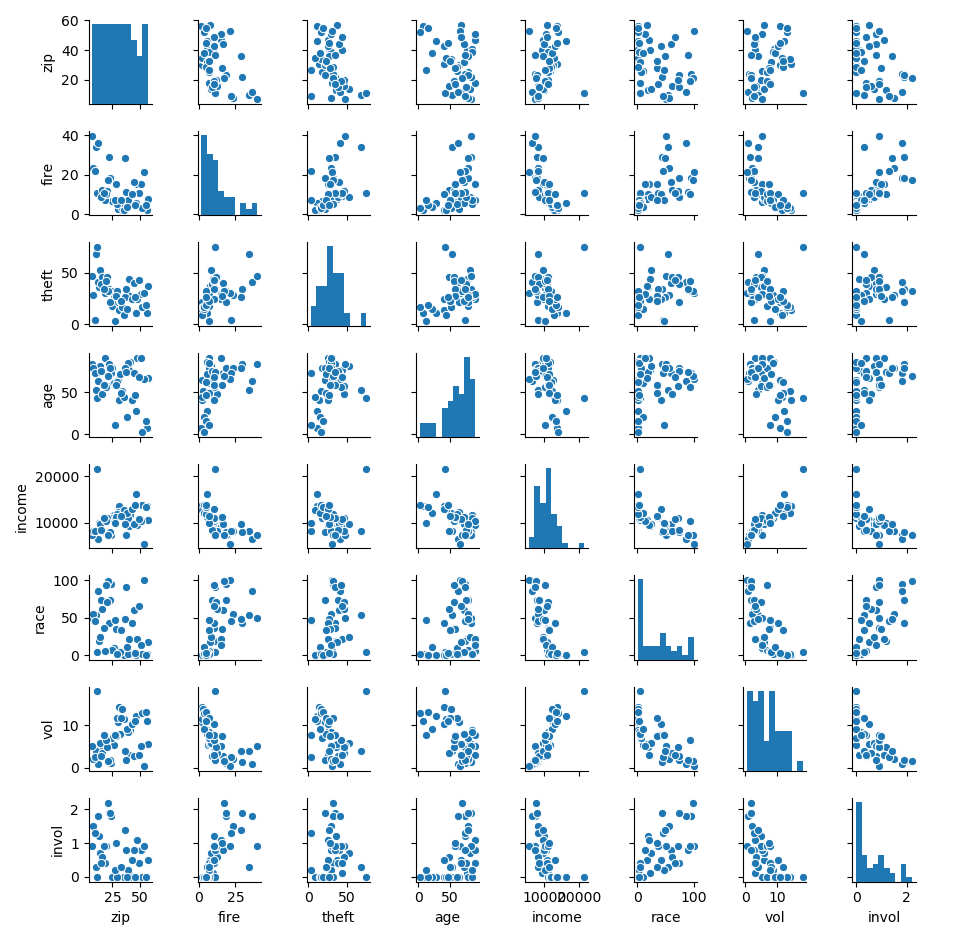
We are analyzing data given to us by the Illinois Department of Insurance, Chicago Police and Fire Departments, and the US Census Bureau. Our goal is to be able to draw any conclusions with regards to if insurance companies are redlining, or refusing to insure certain neighborhoods within Chicago. We have been given 8 total variables; zip code, race, fire, theft, age, vol, invol, and income in order to be able to draw some conclusions. From these variables, we see that we have seven qualitative variables and one categorical variable being the zip code. Out of these variables, we have chosen our response variable to be invol, or Involuntary Market Activity Variable. This is the best measure of people who wanted insurance, but were denied it and thus given a FAIR plan policy. This variable may be misrepresented by the people who have given up on finding insurance and may not want it at all, but the variable is still a good measure. The rest of the variables are all being used as our predictor variables. We will analyze which variables have the strongest correlation and thus be able to see which variables have the greatest impact on insurance companies redlining certain communities.

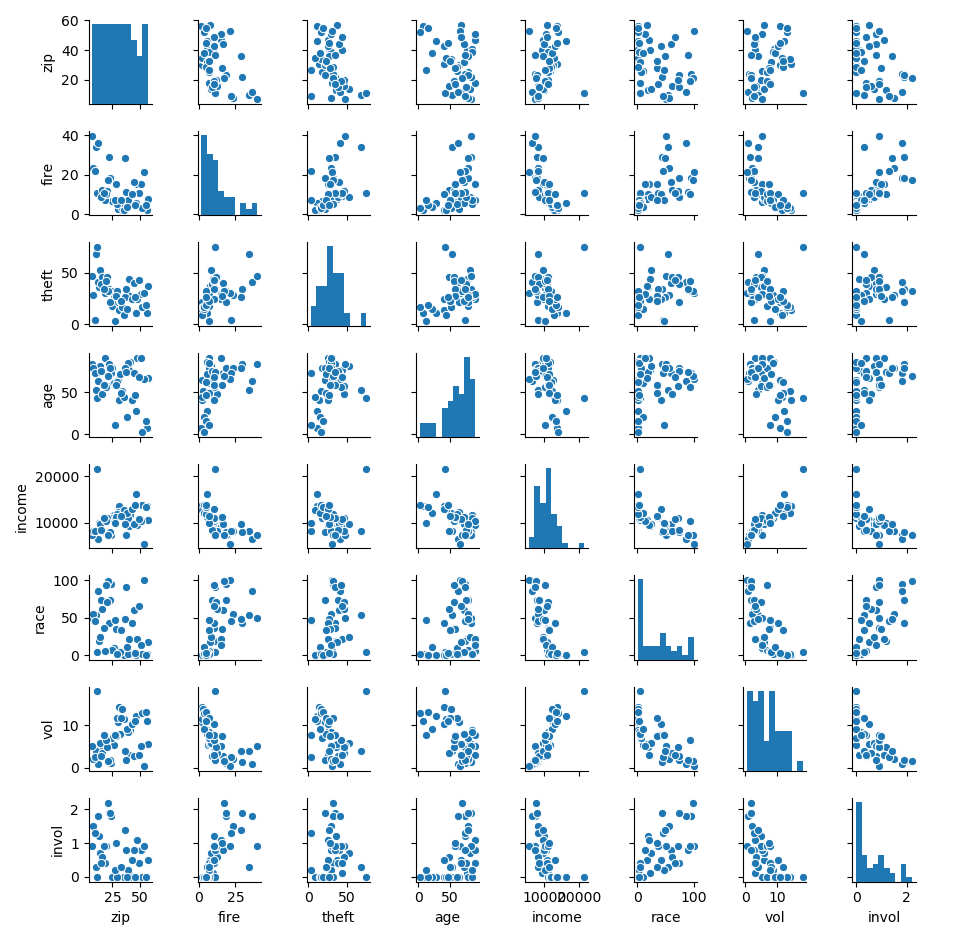
**Box Plots of all Variables:**

Our first step in this experiment is analyzing the data given to us for any outliers or unusual cases. We will do this easily by making boxplots for each variable provided.



Looking at these boxplots, it is clear that we have a total of 7 extreme points in our predictors within the given data. Three of these are seen in the “Boxplot of Fires”, two in the “Boxplot of Theft”, one in the “Boxplot of Age”, and one in the “Boxplot of Median Income”. These extreme points may affect our model but more analysis must be done to determine that.

**Scatter Plot of Y vs. all X’s:**

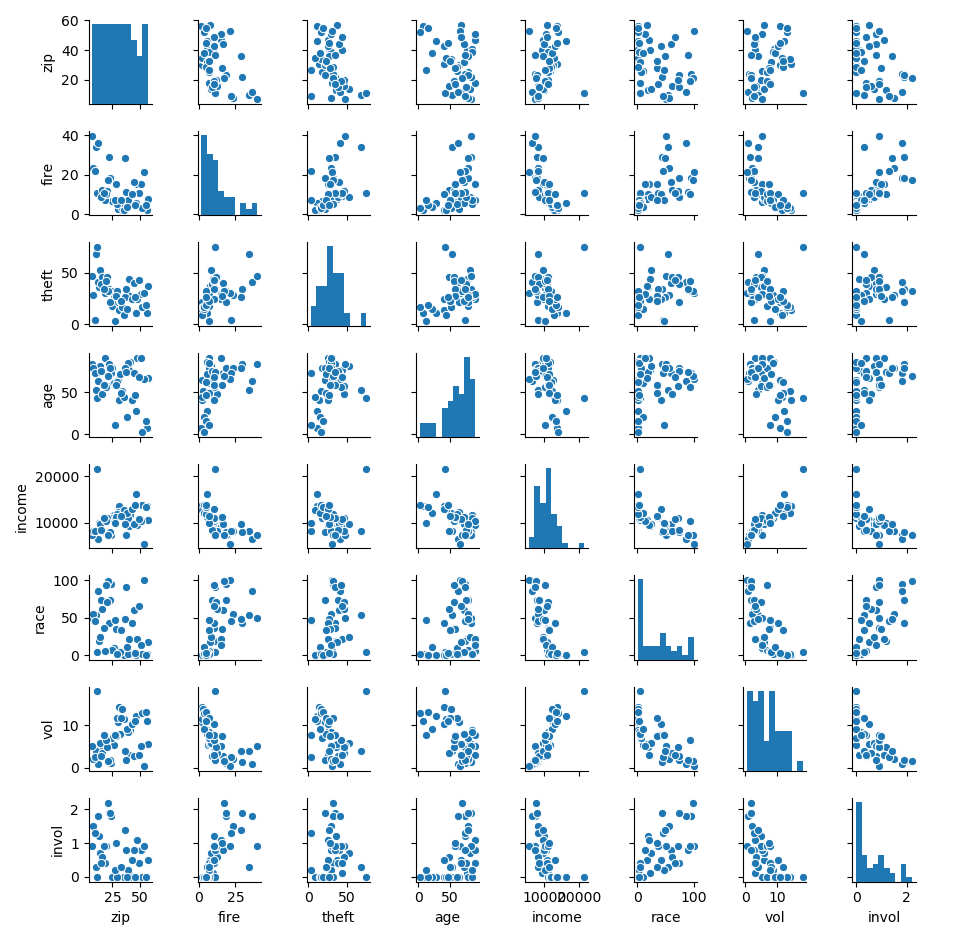


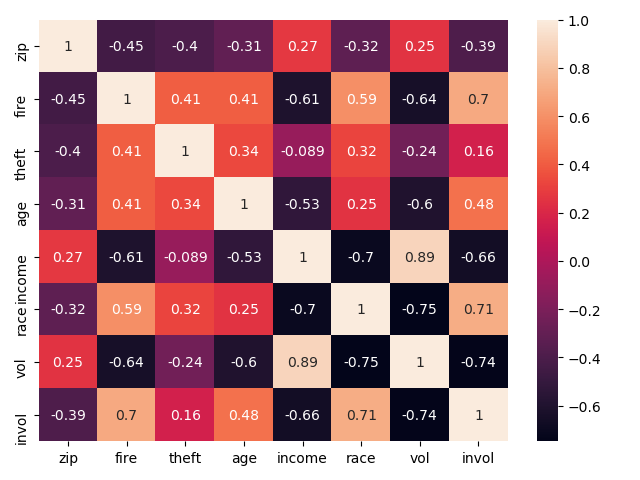
Based on the scatterplots, there appears to be correlations between the variable Invol and the following predictor variables: Fire, Age, Income, Race, and Vol. The predictor variables Zip and Theft do not have as obvious of a correlation with Invol. Zip and Theft have correlation coefficients of -0.39 and 0.16, with Invol, respectively. This is not entirely surprising regarding Zip, as it is not entirely a quantitative variable. While Zip is indeed composed of numbers, in context Zip could also be thought of as a qualitative predictor variable, since it describes an area rather than an amount. This could be a point for further background of the dataset; perhaps there is a gradient of the zipcodes and the area they represent, e.g. the higher the zip code number, the further east a residential area is.

It might be useful to transform, or center, the data in Theft vs. Invol. It seems that Invol is highest for the middle values of Theft. Invol is the number of new and renewed FAIR plan insurance policies. The background description of this dataset suggests that Invol is a good measure of being denied by most insurance companies. It would make sense that FAIR plan policies operate in accordance with the principle of affirmative action, and cater to more underprivileged areas, communities, and people.Those with low crime and theft rates would most likely not be denied by other insurance companies and would not need to get FAIR plan assistance. Those with extremely high crime and theft, while they would be denied by most insurance companies, would most likely also not be selected for FAIR plan policies as their crime may be too high. This would explain why FAIR plan policies (Invol) are issued mostly in the median values for Theft. It may be beneficial to treat Zip as a qualitative variable in the linear model, and center the values for Theft.

Transforming the data for age, by squaring it, increased the adjusted R2 of the overall model and decreased the p-value of the F-test. Therefore, we decided to square the values of the ages when supplying the data to fit the model. During this testing, we also determined that including ZIP data gave us no increase in accuracy while increasing the p-value of (decreasing confidence in) our model, so we decided to omit that variable.

**Pearson Correlation Between all Pairs of Variables:**





The Pearson Correlation is a very organized and effective way to display the relationships between each pair of variables that are used throughout this experiment. They are especially useful because they show us potential variables to be eliminated from the model, as well as any hard-to-see trends in the data. The highest absolute value correlation with respect to our response variable is (somewhat expectedly) vol, and following vol is fire. The lowest correlated variable is theft.

Amongst the predictors, income and vol seem to be highly correlated, and one of the two may need to be excluded. Neither ZIP nor theft have high correlation with any of the other variables (including the response variable), and could be possibly excluded later if we need to reduce the complexity of our model in order to make it more accurate. We also notice that our predictor variables income and vol are highly correlated so it could be beneficial to remove one of these variables.

**Simple Linear Model of each X vs. Y:**

Zipcode Model - Zipcodes vs Invol : **Y = 1.13480 - .01699X**

Fire Model - Fire vs Invol: **Y = .02671 + .04790X**

Theft Model - Theft vs Invol: **Y = .401011 + .007074X**

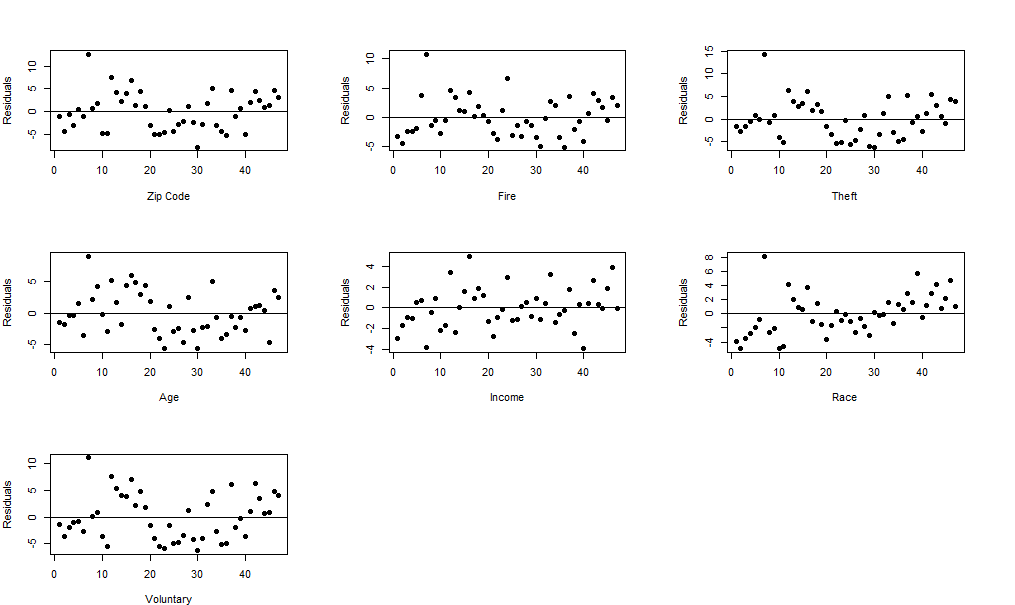
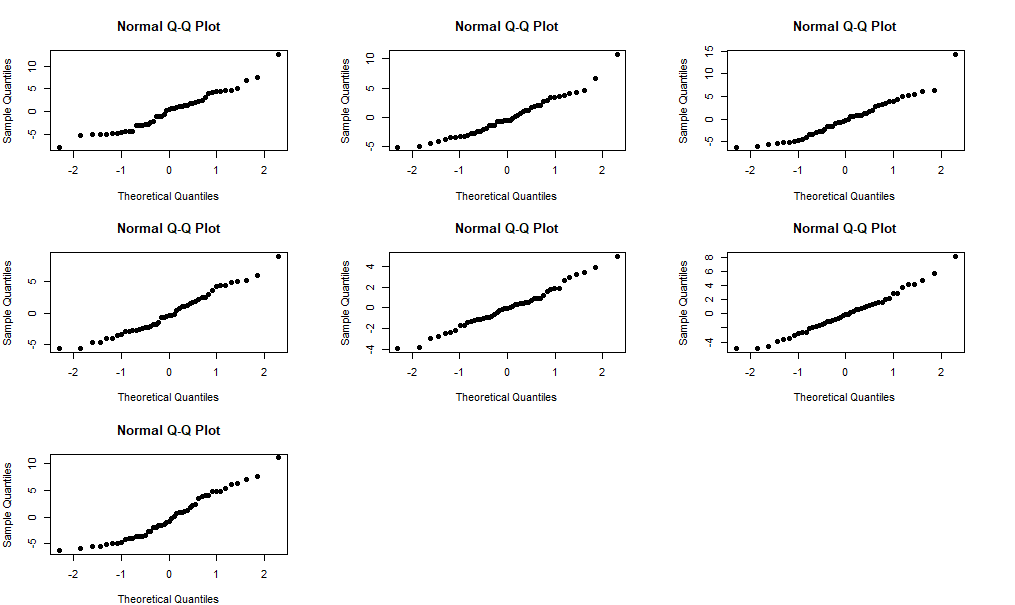
Age Model - Age vs Invol: **Y = -0.19089 + .01336X**

Income Model - Income vs Invol: **Y = 2.251364 - .000153X**

Race Model - Race vs Invol: **Y = .12922 + .01388X**

Vol Model - Vol vs Invol: **Y = 1.3493 - .1089X**

**Residuals vs Predictor Variable:**



Looking at the residuals of our models we can determine whether some of the assumptions of our linear models are correct. By calculating the expected value of the residuals in R, we see that the mean of the residuals all equal or are close to 0. This satisfies one assumption of our model. By looking at the plots of the residuals vs X variables and the distribution of residuals, ignoring outliers, we see that the variance of the error terms are constant which satisfies another assumption of our model.

**Descriptive Statistics**

The descriptive statistics for each of these variables was easily found using R.

Using the function, summary() in R, this displays the five number summary of each variable in the data set as well as the means for those variables. This gives us a better representation of what numbers and ranges the variables could possibly have.

> summary(insurance)

**zip fire theft age income race**

Min. : 7.0 Min. : 2.00 Min. : 3.00 Min. : 2.00 Min. : 5583 Min. : 1.00

1st Qu.:18.5 1st Qu.: 5.65 1st Qu.:22.00 1st Qu.:48.60 1st Qu.: 8447 1st Qu.: 3.75

Median :30.0 Median :10.40 Median :29.00 Median :65.00 Median :10694 Median :24.50

Mean :30.6 Mean :12.28 Mean :30.23 Mean :60.33 Mean :10696 Mean :34.99 3rd Qu.:42.0 3rd Qu.:16.05 3rd Qu.:38.00 3rd Qu.:77.30 3rd Qu.:11989 3rd Qu.:57.65

Max. :57.0 Max. :39.70 Max. :75.00 Max. :90.10 Max. :21480 Max. :99.70

**vol invol**

Min. : 0.500 Min. :0.0000

1st Qu.: 3.100 1st Qu.:0.0000

Median : 5.900 Median :0.4000

Mean : 6.743 Mean :0.6149

3rd Qu.:10.450 3rd Qu.:0.9000

Max. :17.900 Max. :2.2000

We also found out the standard deviation and variance of each variable in the set using the functions sd() and var() from R. These are a great indicator of how volatile and spread out the data is within that set for each one of our variables.

|  |  |  |
| --- | --- | --- |
| Variable | Standard Dev. | Variance |
| Zipcode | 14.59103 | 212.8982 |
| Fires | 9.302266 | 86.53215 |
| Thefts | 14.53364 | 211.2266 |
| Age | 22.57496 | 509.629 |
| Income | 2754.198 | 7585607 |
| Race | 32.58761 | 1061.953 |
| Vol | 4.296092 | 18.45641 |
| Invol | 0.6338216 | .4017299 |

An anova test was also ran on each of the simple linear regression models listed above,

The following information is R output for the anova tables of the linear models listed above.

**Zipcode Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Fire | 1 | 2.8279 | 2.82791 | 8.1305 | 0.006549 \*\* |
| Residuals | 45 | 15.651 | 0.34781 |  |  |

**Fire Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Fire | 1 | 9.1338 | 9.1338 | 43.979 | 3.591e-08\*\*\* |
| Residuals | 45 | 9.3458 | 0.2077 |  |  |

**Theft Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Theft | 1 | 0.4863 | 0.48626 | 1.2161 | 0.276 |
| Residuals | 45 | 17.9933 | 0.39985 |  |  |

**Age Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Age | 1 | 4.1823 | 4.1823 | 13.163 | 0.0007258 |
| Residuals | 45 | 14.2973 | 0.3177 |  |  |

**Income Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Income | 1 | 8.1684 | 8.1684 | 35.648 | 3.452e-07\*\*\* |
| Residuals | 45 | 10.3112 | 0.2291 |  |  |

**Race Model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Race | 1 | 9.4143 | 9.4143 | 46.733 | 1.784e-08 \*\*\* |
| Residuals | 45 | 9.0653 | 0.2015 |  |  |

**Voluntary Model:**

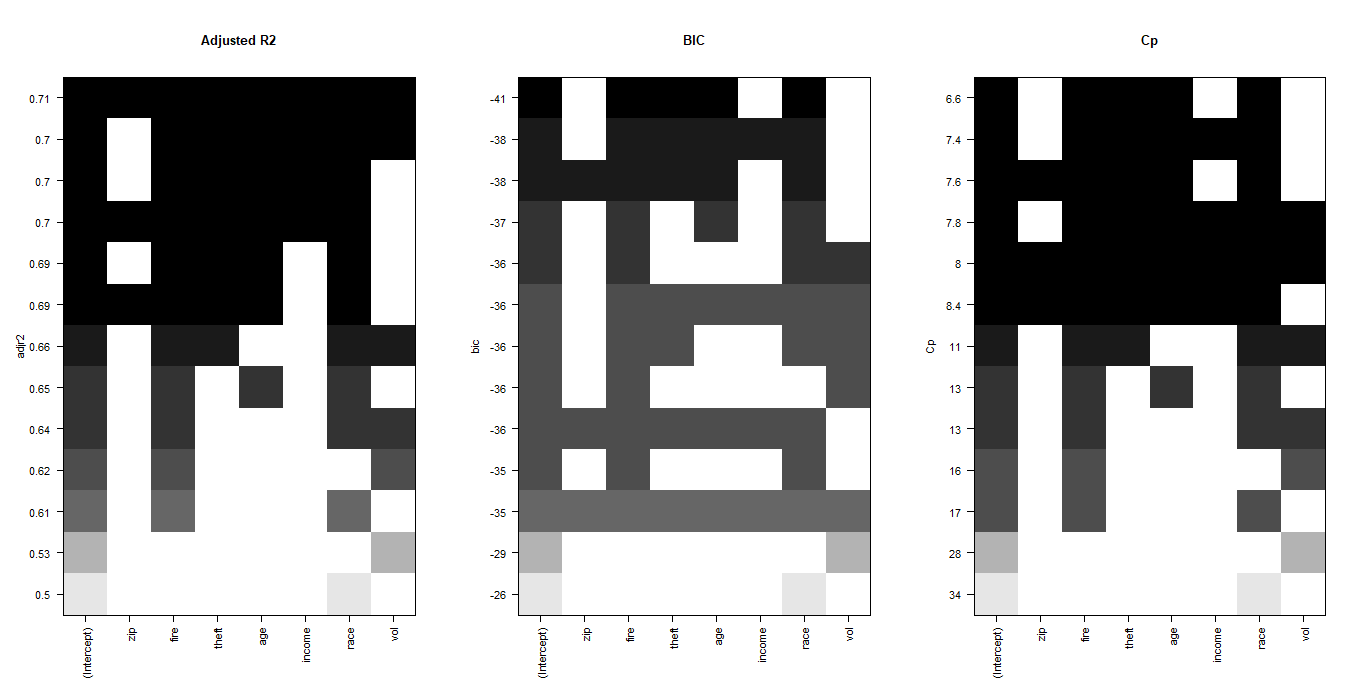
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Vol | 1 | 10.0715 | 10.0715 | 53.903 | 3.18e-09 \*\*\* |
| Residuals | 45 | 8.4081 | 0.1868 |  |  |

We can use an Anova table to determine whether a relationship in our linear model exists. First, we test our null hypothesis that the regression coefficients equal 0 with an F test. The p value of this test determines whether we can reject our null hypothesis and conclude that there exists a relationship between the variables. If we have a high F-value, and the p value is > 0.05, then we know that there is sufficient evidence to reject our hypothesis. In every test, except for our theft model, we can determine that there exists a linear relationship between the predictor variables and our response variable, invol. Our theft model’s hypothesis does not have sufficient evidence to reject which means there could possibly be no relationship between our theft variable and our response variable giving us further reason to disregard theft as a predictor variable.

**Variable Selection:**

Below is pictured several information criteria for variable selection in the model; adjusted R-squared (adjr2), Schwarz' Bayesian Criterion (BIC), and Mallows’ Criterion (Cp), for each variable in the insurance dataset. Models with large adjusted R-squared values and fewer variables are generally preferred. The va

lues of BIC and Cp are not meaningful, but rather the comparison across all models of these criteria.

****

Based on all three plots, dropping the Zip variable seems to make sense. Based on the adjusted R-squared plot, the model with fewest variables and highest R-squared value includes Fire, Theft, Race, and Vol, as the linear model with these predictors captures 66% of the data. The same model can be confirmed from the Cp plot.

**Variance Inflation Factor (VIF)**

VIF is a formal method of detecting the presence of multicollinearity among predictor variables. Indicating that including every predictor in the model may not be a good idea, since some predictors may vary or correlate with other predictors, causing the model to be overfit, resulting in a weak correlation with the data.

VIF is calculated for a model in which a predictor variable is regressed against all other predictor variables. For example, in a dataset with predictors , , and , VIF would be calculated for each of the following linear models: , , and . Generally a VIF which is greater than 5 indicates multicollinearity; that predictor is correlated with other predictors and it might be a good idea to omit it from the regression model. The VIF values for all predictors in the Insurance dataset are shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| VIF Zip | VIF Fire | VIF Theft | VIF Age | VIF Income | VIF Race | VIF Vol |
| 1.443627 | 2.210049 | 1.733334 | 2.234269 | 6.026122 | 3.315207 | 7.618781 |

Income and Vol both have VIF values greater than 5, so based on VIF, dropping Income and Vol may be a good control measure to limit multicollinearity.

However, upon further examination it appears that Vol is intercorrelated to Income, but income is not intercorrelated to Vol. It might not be necessary to drop both of these predictors from the model. Below is the VIF of income when it it is regressed against all predictors *except* for Vol:

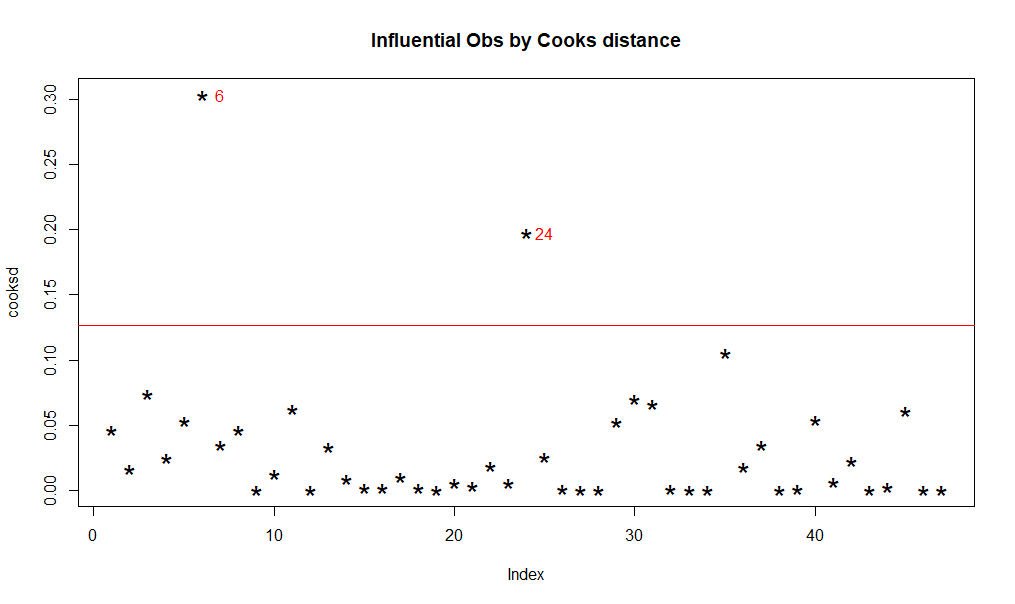
lm(income ~ zip + fire + theft + age + race)

VIF(income) = 3.671594

Since removing Vol from the measure of multicollinearity reduces the VIF, we only need to drop Vol. Additionally, the adjusted R2 of a reduced model which excludes vol and includes income is higher than one which excludes income and includes vol. Therefore, we decided to remove vol from the model, and keep income as a predictor.

Zip has the lowest VIF in the above table, thus is most likely not correlated to other predictors. Therefore it might seem like it should be kept in the model. However the reason it is not multicorrelated is most likely because Zip is categorical in nature. The zipcodes are not a numerical measurement as much as they are a qualitative index of *where* the numerical observations in the dataset came from. Any correlation that zipcodes may have with Invol would not really be meaningful, as the zipcode numbers themselves don’t increase or decrease with any particular geographical pattern.

**Cooks Distance:**



Looking at the Cook's distance of our model, we can see that there are two points in the data that are influential. However, a high influence of a point does not necessarily mean that the point should be excluded. Because this data was collected freely and fairly by credible sources, we are going to include these outliers in our model. The extreme values in the data could be due to the result of numerous external factors taking place within the given communities.

**The Model:**

After the analysis of each of our predictor variables, we have decided that due to the strong collinearity between income and vol, both having a VIF > 5, one of them would be insignificant in the model and should possibly be removed from the data. We also see that when we run an F test on the multiple regression models including every variable, the F-statistic is lower than when they are excluded from the model. This gives us further reason to not reject the null hypothesis and conclude there is not a relationship between some of the variables and our response variable. Surprisingly, the adjusted R-squared value for the full model with zipcode and vol is barely greater than the adjusted R-squared value in the reduced model including variables; fire, theft, age, income, and race. However, we have decided that the most effective model to use is the one excluding the variables zip and vol, and with a squaring transformation on our age predictor. We included the transformed variable because it increased the F-statistic and lowered the p-value associated with the test. Primarily because vol is strongly correlated with another predictor (income) and the zip code is a qualitative variable that has no correlation with any of the variables. Both variables contribute minimal accuracy to our model while decreasing our confidence in the results. The model we have decided on that is a best fit for this data is identified below.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.940e-01 5.093e-01 -1.363 0.180409

fire 2.972e-02 7.863e-03 3.779 0.000501 \*\*\*

theft -1.273e-02 4.397e-03 -2.896 0.006034 \*\*

newage 9.716e-05 2.977e-05 3.264 0.002223 \*\*

race 1.221e-02 2.673e-03 4.567 4.46e-05 \*\*\*

income 4.673e-05 3.623e-05 1.290 0.204247

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3455 on 41 degrees of freedom

Multiple R-squared: 0.7352, Adjusted R-squared: 0.7029

F-statistic: 22.76 on 5 and 41 DF, p-value: 7.308e-11

Y = -0.06940 + 0.02972 - 0.01273 - 0.00009716 + 0.01221 + 0.00004673

= fire, = theft, = age(squared), = race,  = income