Project of STAT6026

2023-10-31

Part A(page 1-5)

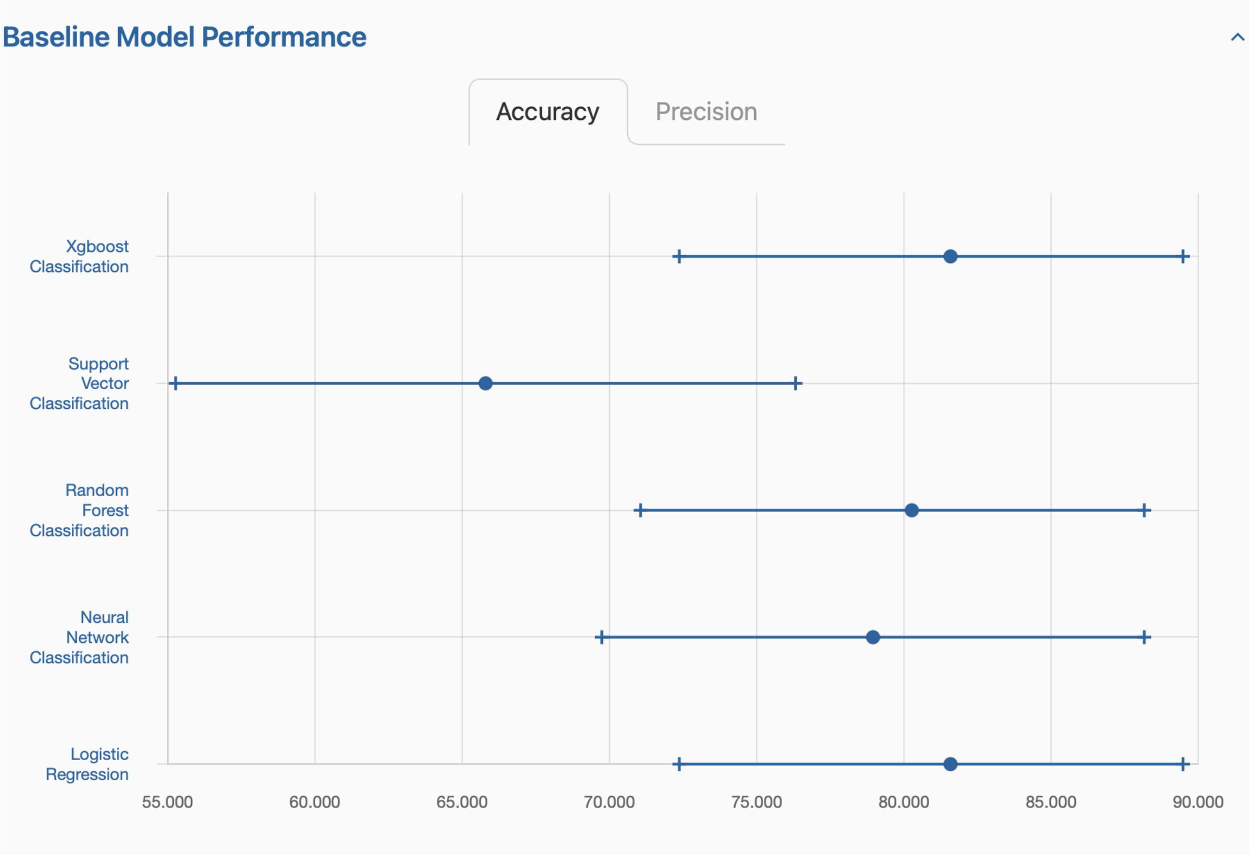


Figure 1: source:https://archive.ics.uci.edu/dataset/45/heart+disease

This graphic is from webpage of Heart Disease Datasets which is one of the most popular datasets in UC Irvine Machine Learning Repository. The graphic shows the accuracy of each machine learning method.The point in the middle shows the average performance of each ML methods while the 2 boundaries show the best and the worst performance.The graphic is breif , direct and easy for people to compare each methods.We can see Logistic Regression and Xgboost Classification are likely to perform better than other methods. And in the case Support Vector Classification is likely to perform worst.

The advantages of the graphic are showing the average performance directly and showing random effects in testing machine learning. If the graphic could sort the methods from top to bottom by average accurancy, the comparison may be more direct.

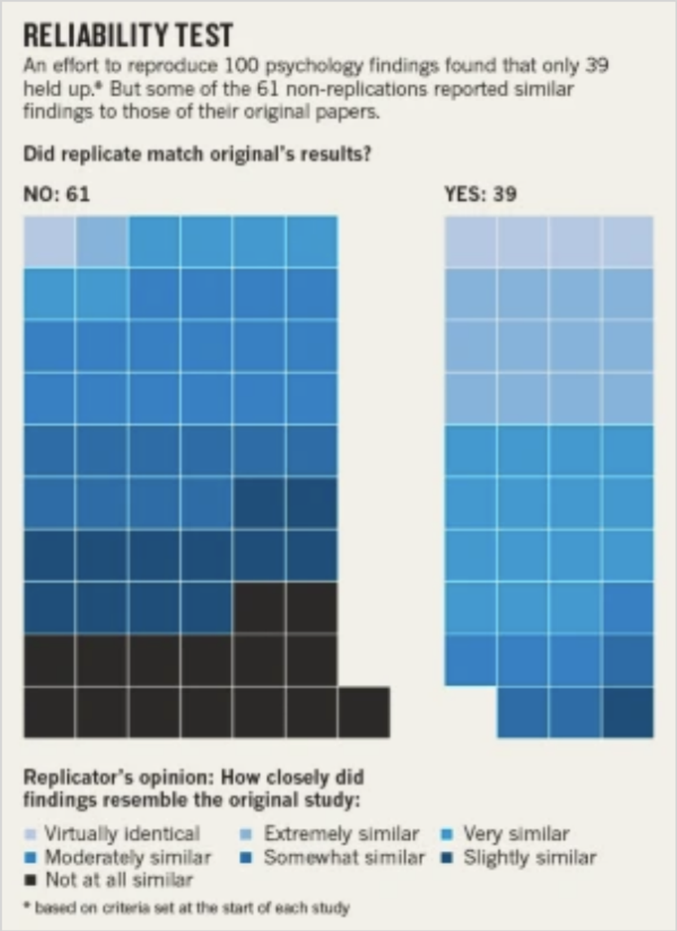


Figure 2: source:https://www.nature.com/articles/nature.2015.18248

The picture directly shows the ratio of successful replication results which strongly suggest that many psychology studies should be doubted because the high rate of unsuccessful reproducing.Besides, it also show the similarity between original paper and reproducing results by color from light to dark(which always mean negative) which is also very direct and brief.We can see “not at all similar” accounts for 15 in 100 which may also imply a crisis in psychology study. However, although some work may be reproduced unsuccessfully ,the findings share some similarity. Besides,the we can clearly see the different color pattern between “Yes” and “No” which reflect the fact that “Yes” in replication means more similar and vice vesra.

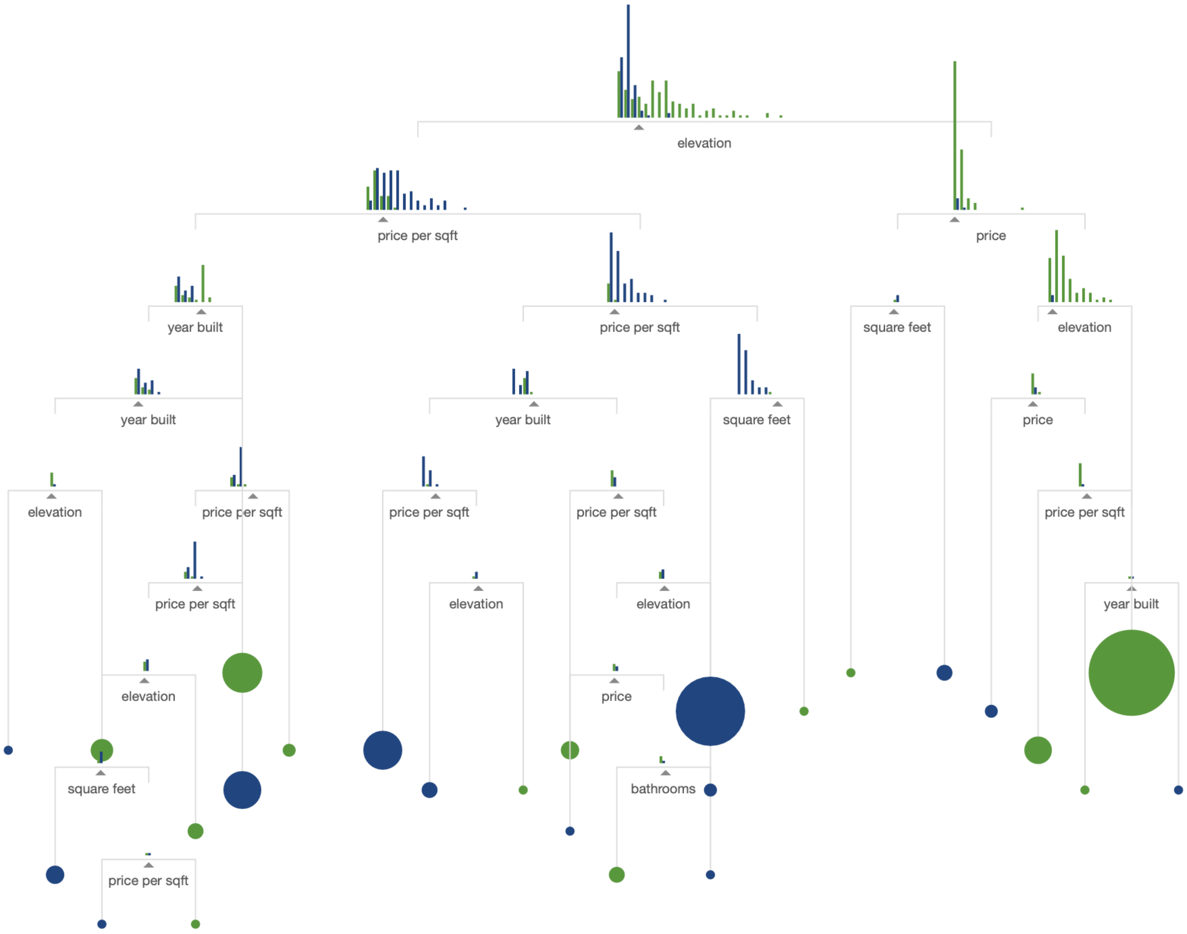


Figure 3: sourse:http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

The graphic show how tree decision method which is one of the classical machine learning method works. The 2 colored histogram on each nodes clearly shows how it split a data subset into 2 partition depending on point on x(different covariates) axis.

The original graphic is dynamic which really help ML beginners learn the principle of tree decision quickly.

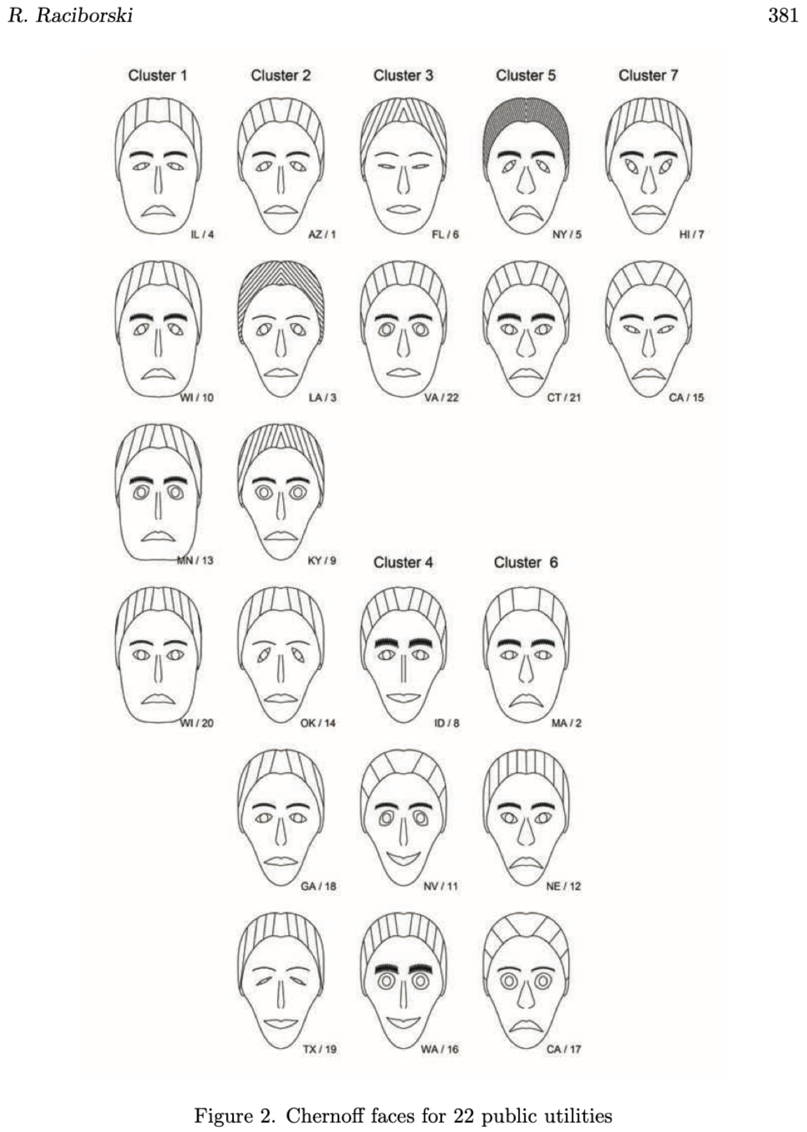

R. Raciborski 381

Figure 2. Chernoff faces for 22 public utilities

Figure 4: source:https://journals.sagepub.com/doi/pdf/10.1177/1536867X0900900302

This is also an interesting graphic which use chernoff face to show different clusters. We can also find companies in the same cluster share what in similarity quickly because people are sensitive to face features .

It may cost a little time to describe these features with raw data.However, This graph may really be useful for people to see the hidden structure of data , how well cluster works and a good and quick approach to show high dimensional(more than 2D) data as well.

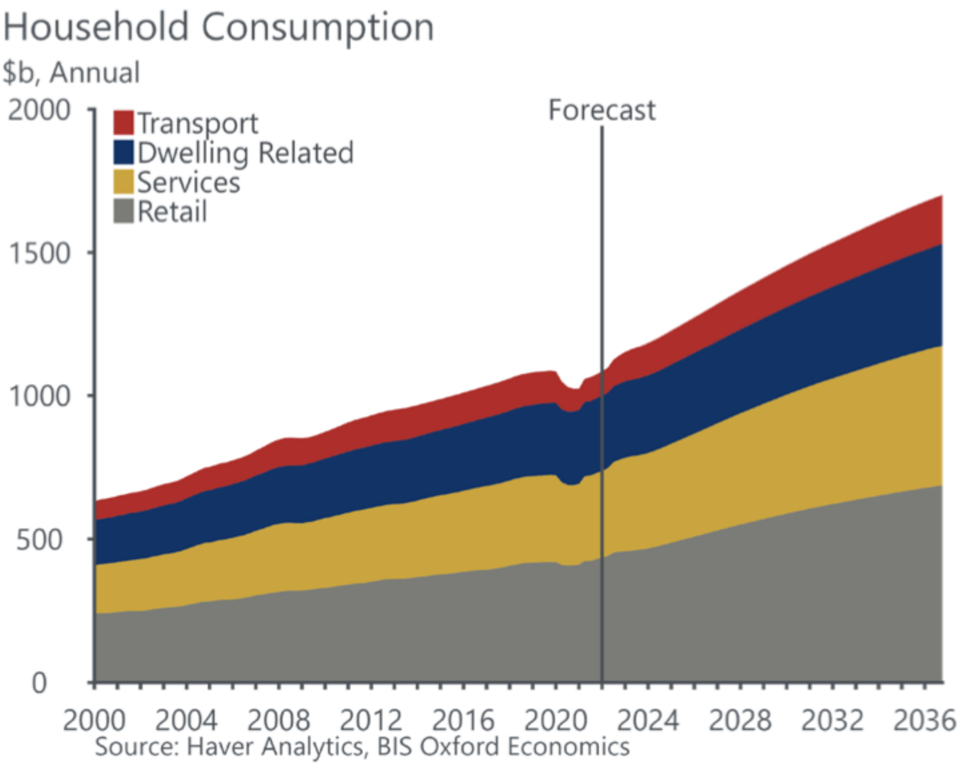


Figure 5: source:https://www.portofmelbourne.com/wp-content/uploads/BISOE-Port-of-Melbourne-Container-Forecasts-and-Sensitivities-20221223.pdf

It shows the household consumption in different part and total in one graphic.Besides ,it also show the change and the rate of total and each cosumption change.The data on 2020 which showed an abnormal down because of the pandemic breaking out.

The author wanted to show the increasing trend of the household consumption therefore he or she plot the forecast.It may be clear to see the total growth and retail growth but a little harder to get the growth of other part due to the graphic type. But since the retail accounts most ,this may be reasonable. And we may compare the growth or growth rate by slope.

Part B (page 6-13)

Introduction

This part focus on the analysis on insurance availability around 1978 in Chicago.The data has full description in GDA project description. The original variables include Zip,Race,Fire,Theft,Age,Volun, Invol and Income.

Zip which represents different neighborhood is the identity of each neighborhood.Race,Fire,Theft,Age and Income may be our important covariates especially Race and Age since the objectives of the study are to explore the extent to which racial composition and age of housing affect underwriting practices.

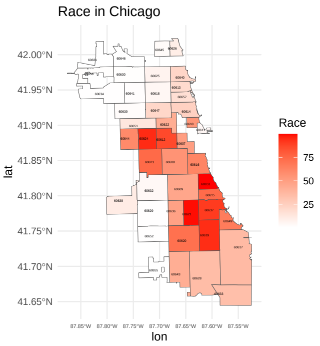
Volun represents the a rate of "accepted" requests to purchase insurance while Invol is likely to play the opposite role meaning rate of "rejected".Since these 2 variables are rate, to some degree,they can show the rate of acceptance and rejection regardless of different size of neighborhood . However,the ratio between $Invol$ and $Volun + Invol$ may be more reasonable to represent the extent of rejection.So here I produce a variable VIRatio which is $\frac{Invol}{Volun+Invol}$ as response variable of the first regression model. Besides ,insurance was probably not compulsory so we can see the $max(\frac{Volun}{100})$ is around 0.1 while $max(\frac{Invol}{100})$ is around 0.01 which may also suggest us to consider Volun as a response variable.In the second regression model I take Volun as response variable.

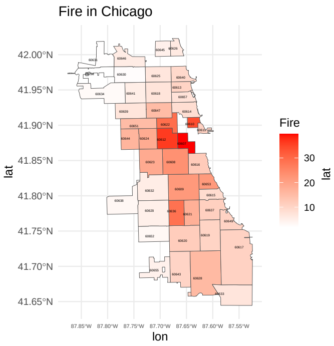
Let's first have a general picture of the data.

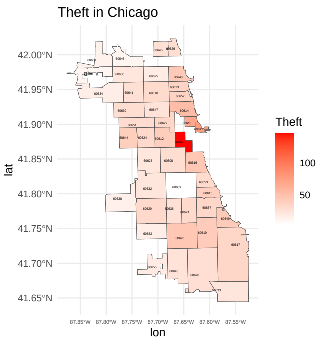
General picture

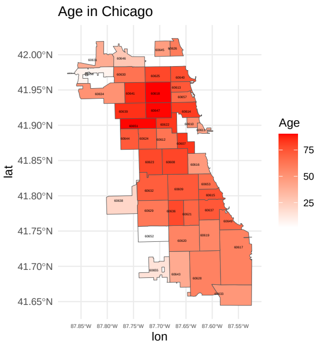
I download Chicago with zip code from Google to interpret the general view of data on location.The 8 graphics below respectively show Race,Fire,Theft, Age,Volun,Invol,Income and VIRatio on the map of Chicago .We may find that the VIRatio in the center part of the Chicago is more likely to be lower.Besides, it seems that the graphics of Race , Fire ,Invol and VIRatio share some pattern (deep red and light red seem to have similar areas) which suggests high positive correlation between variables .

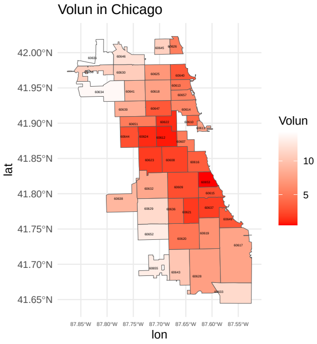
Finally , we also need to pay attention to the areas where $Invol$ and $VIRatio$ are zero.These areas are all in almost white color meaning very low in Race(almost 0) but show similar Fire,Theft and Age compared to areas around or near them.Meanwhile, these 0-Invol areas also show different house age which may suggests the relationship between "reject" and house age is weak. These features may suggest the race discrimination may happen in making the desicison of "accept" or "reject".

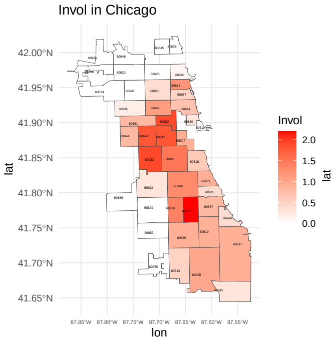


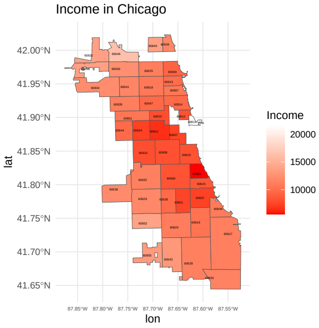


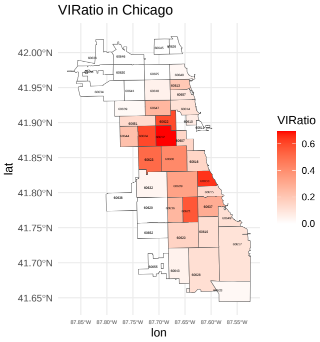








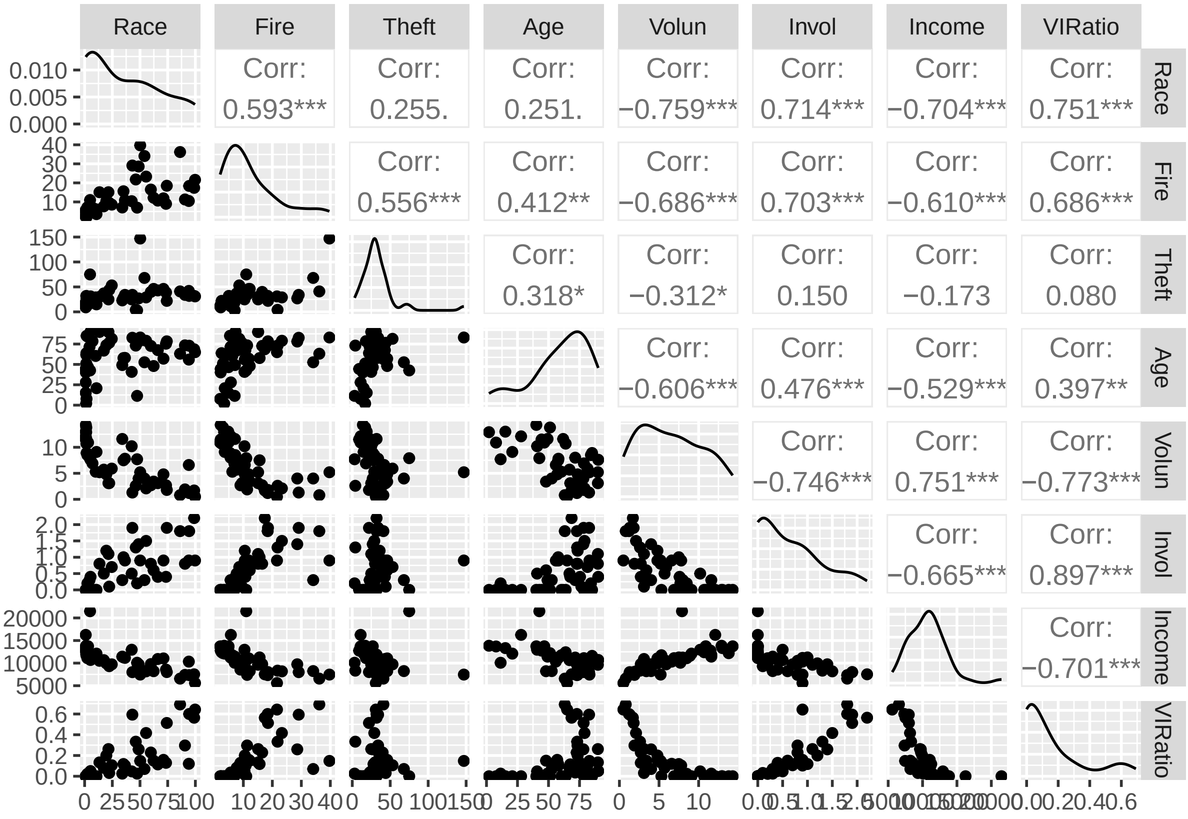




Matrix of correlation and scatter plot

Take a further look on the correlation between each variables. From the matrix of plots below, it seems that Race really show high correlation with VIRation which is corresponding to our finding in general picture's pattern.We may also find Race show a strong linear relationship with Volun and Invol which may also suggest affect the underwriting practice to a large extent. We may found fire and Income show high correlation with IVRatio while Theft show low correlation with IVRatio.

Besides, we can also easily learn the correlation between covariates such as Race ,Fire and income.The high correlation and strong linear pattern of the scatter plot between Race and Income is also corresponding to the pattern of deeper red closer to center if we review the general picture.When it comes to plot and correlation of theft and VIRatio, we may find the 2 variables seem to have no correlation.



Regression models

Now, let us try regression modeling to have a deeper view on the relationship among several variables. Our response variable may be one of Volun , Invol and VIRatio. VIRatio is a product combining Volun and Invo and reflect the close ratio between rejection and all insurance requests.So we may first try

$$VIRatio_i = \beta_0 + \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Theft_i + \beta_4 Age_i + \beta_5 Income_i + \epsilon_i$$

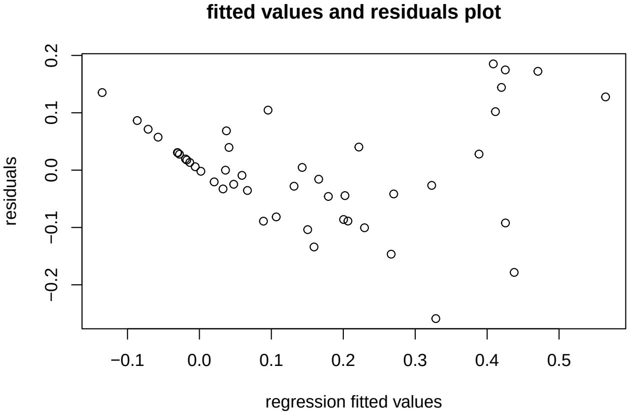
and

$$\epsilon_i \sim iid\ Norm(0, \sigma^2)$$

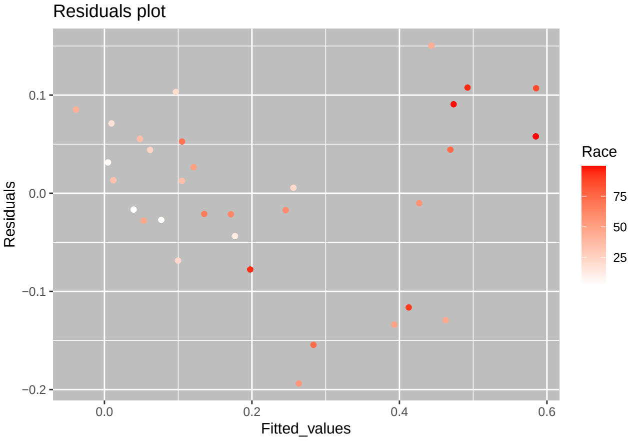
From the output we get

$$V I \hat{R}atio_i = \frac{-698.9749 + 29.64232 Race_i + 119.0461 Fire_i - 36.59007 Theft_i + 15.70569 Age_i - 0.001532509 Income_i}{10^4}$$

The output of linear regression shows that the coefficient $\beta_1, \beta_2, \beta_3$ are significant and indicating that in the model the coefficient of Race,Fire and Theft are important and should not be zero.However, when we plot the residuals ,we may find that the spread increases when the fitted value increases which may suggest transform the VIRatio to log(VIRaion).Besides, we also found a interesting pattern where low fitted values have a strong linear relationship with residuals. The strange points mainly come from the true value of $VIRatio_i$ which is 0.So the strange pattern of residuals can be explianed because all these points have the same VIRatio but different covariates values and the regression model is a linear predictor.



Refitting on data with $VIRatio \neq 0$

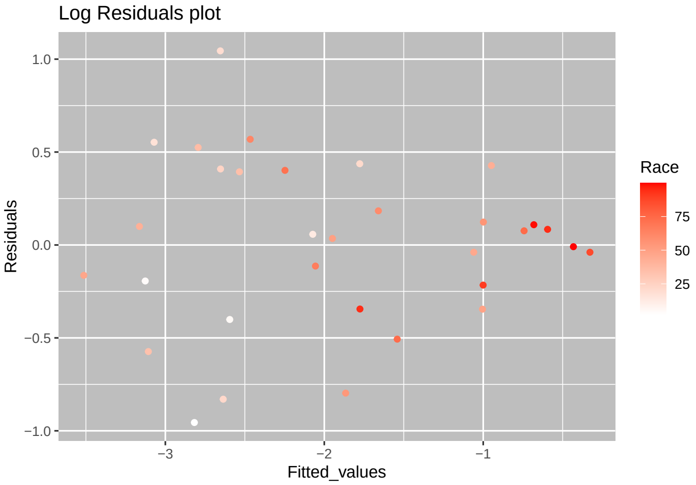


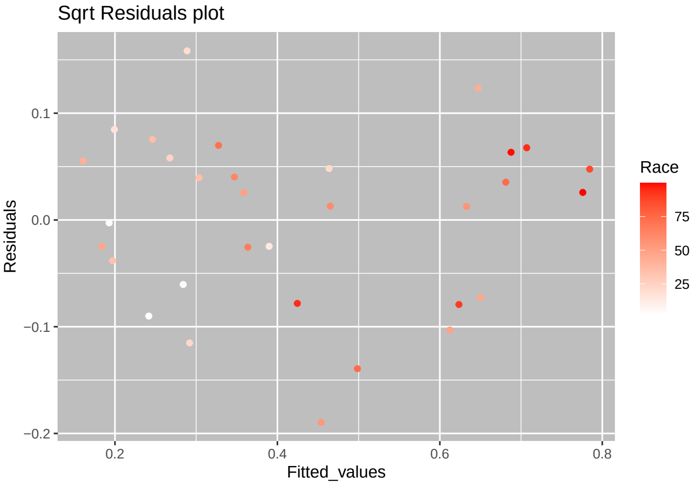
Let's first remove these $VIRatio_i = 0$ points ,fit the rest date with the same model again.The figure above shows the spread of residuals increase again when fitted values increase. So we apply a log or square-root transformation to $VIRatio_i$.Then fit the following model:

$$log(VIRatio_i) = \beta_0 + \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Theft_i + \beta_4 Age_i + \beta_5 Income_i + \epsilon_i \quad (1)$$

$$\sqrt{VIRatio_i} = \beta_0 + \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Theft_i + \beta_4 Age_i + \beta_5 Income_i + \epsilon_i \quad (2)$$

After refitting,from the plot below we may find the residuals pattern of fitted model (2) perform better than fitted model (1).So we choose the model (2).





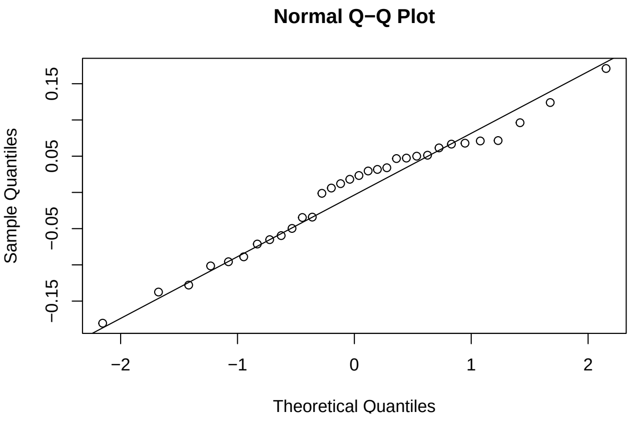
Then we may want to see the coefficients' significance to know whether we can reduce numbers of covariates to make model easier.

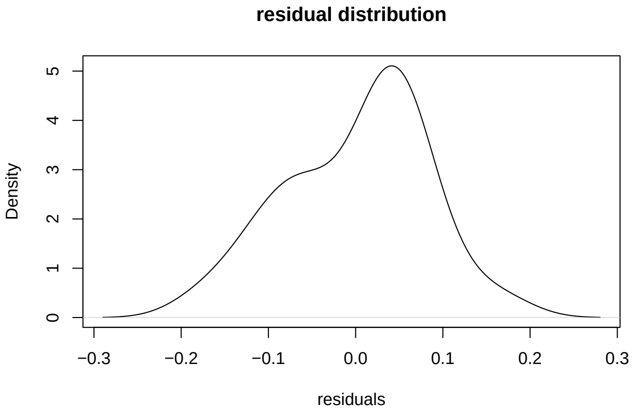
The outputs suggest that Intercept can not be rejected to be zero.So we may have a try to get a new reduced model:

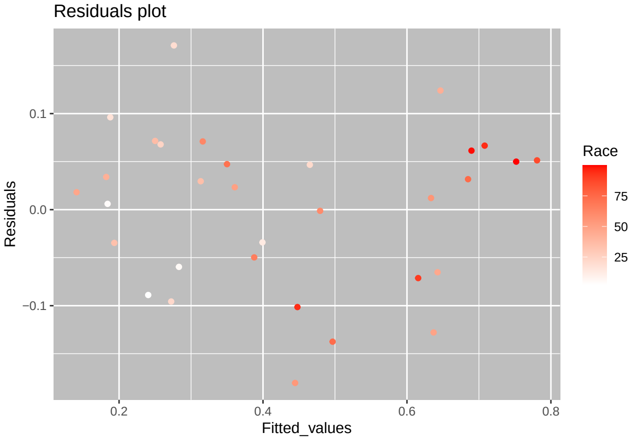
$$\sqrt{VIRatio_i} = \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Theft_i + \beta_4 Age_i + \beta_5 Income_i + \epsilon_i \quad (3)$$

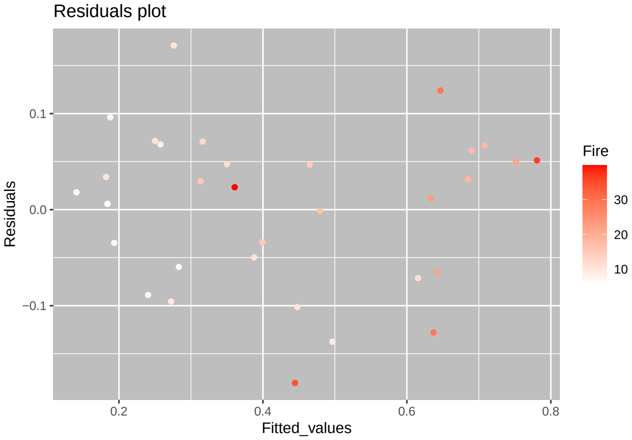
The output seems show that all coefficients are significant in this reduced model 3 so we can't reject each coefficients equals to zero.Meanwhile, the R-squared is around 0.95 which may imply really good fitted result. To see more residuals plot to see whether the residuals have a good pattern and how well they fits iid normal distribution assumption, we usually apply a QQ plot.The qqnorm seems to perform neither good nor bad because the point is not so near the qqline near zero while the overall performance seems not bad.We may also take a look at the residuals density plot which also explain the QQ plot results.The reasons for the not such good results of residuals could be small volume of data.

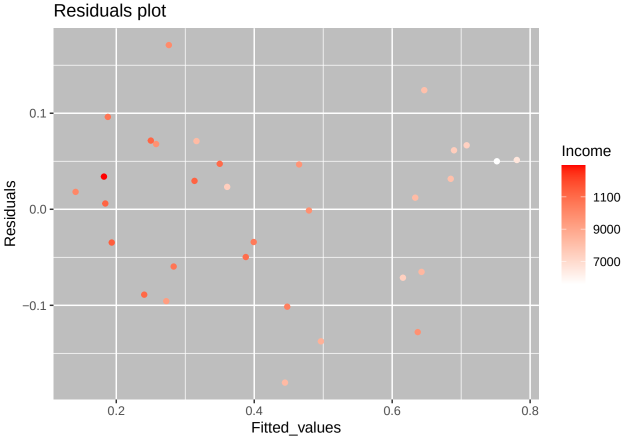
$$\sqrt{\widehat{VIRatio_i}} = 0.003967 Race_i + 0.010082 Fire_i - 0.003817 Theft_i + 0.005378 Age_i + -0.000017 Income_i \quad (4)$$

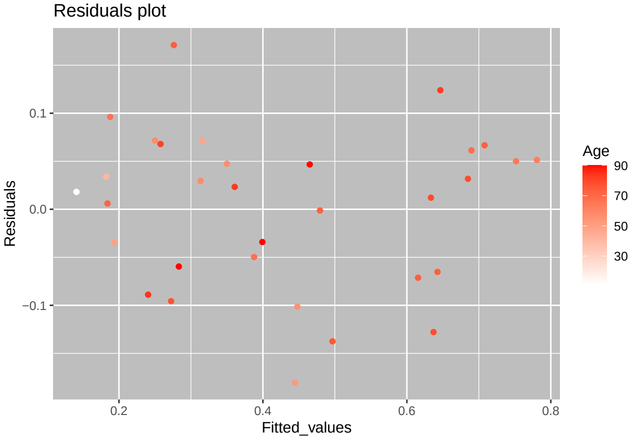


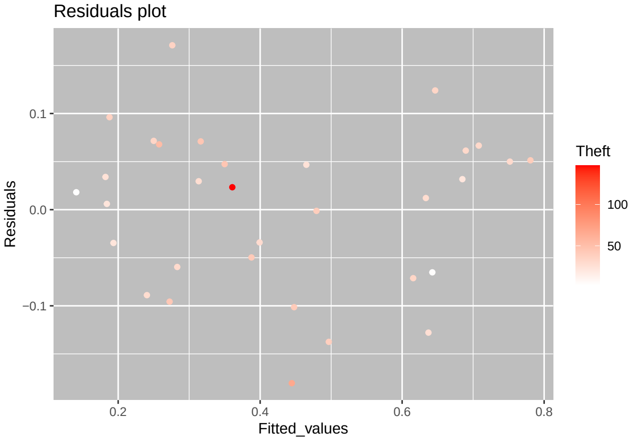



Residuals plot


Residuals plot


Residuals plot


Residuals plot


Residuals plot

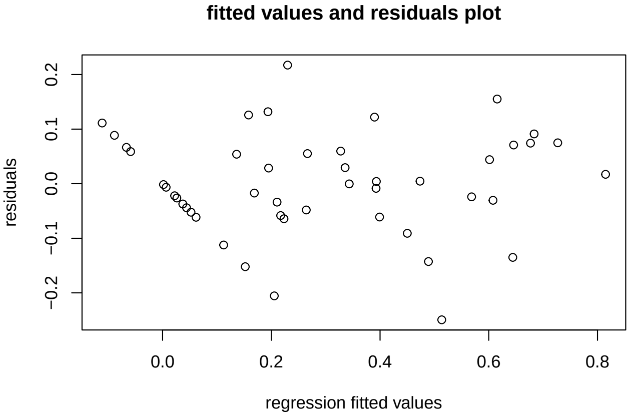
However ,the performance of this model 4 may be good enough to explain the data and residuals and show how Race and Age affect VIRatio.Meanwhile, from the colored residuals plot we ma also find the Race , Fire and Income play more important roles because the dark red and light red data points seem to be classified by fitted values.But the different colored points of Age and Theft are much more widely distributed on Fitted values which may suggest Age and Theft are weak factors.

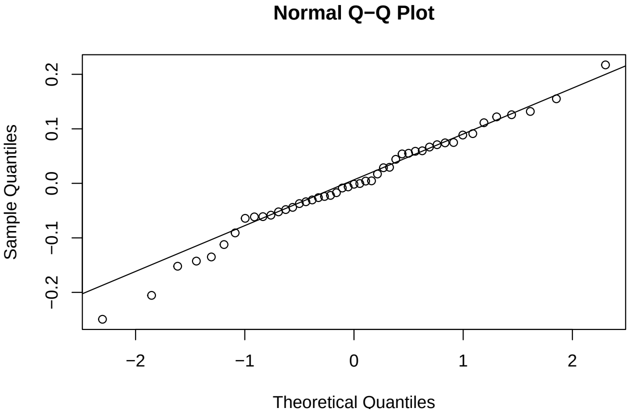
Fit model to standrdized data

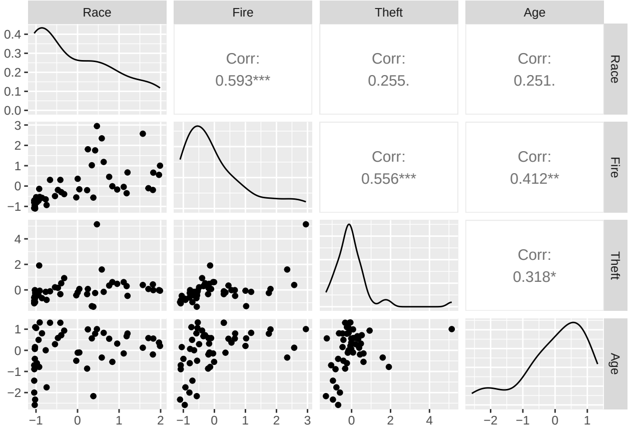
If we want to compare which variable may have the most strong influence, we may try to standardize the data to the same scale.So we standardize all covarites and then fit and select model on the whole data points.After serval practises on fitting and selections,we finally get the model below:

$$\sqrt{VI\hat{R}atio_i} = 0.29260143 + 0.14751409\hat{R}ace_i + 0.12114695\hat{F}ire_i - 0.07771471\hat{T}heft_i + 0.07341757\hat{A}ge_i \quad (5)$$

The qqplot seems to be good although the 0-Invol pattern exists.In the model (5) We can find Race is much more influential to “rejected”.Higher race composition means higher rate of “rejected”.So does Fire while Theft and Age have weak infulence.





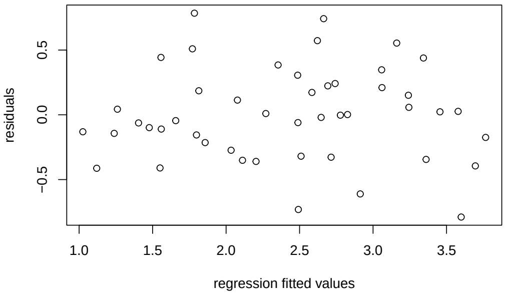


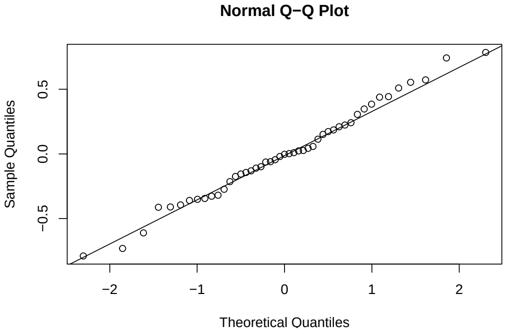
Finally, the Theft’s coefficients in 2 model (4) and (5) are negative which may be contradict to the reality.The reasons may be various such as reduced data ,correlation (the correlation matrix plot somehow show the Theft to balance the other data influence on fitting) and so on.Since Theft is weak ,it may not cause big problems.

Modeling on Volun as response variable

The above discussion we pay much attention on “reject” and let us view the “accept” which may not completely showed by “reject” because the problems with data eg. $Invol = 0$ or much more complex fact in reality. Similarly, our model is

$$\sqrt{Volun} = \beta_0 + \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Theft_i + \beta_4 Age_i + \beta_5 Income_i + \epsilon_i \tag{6}$$



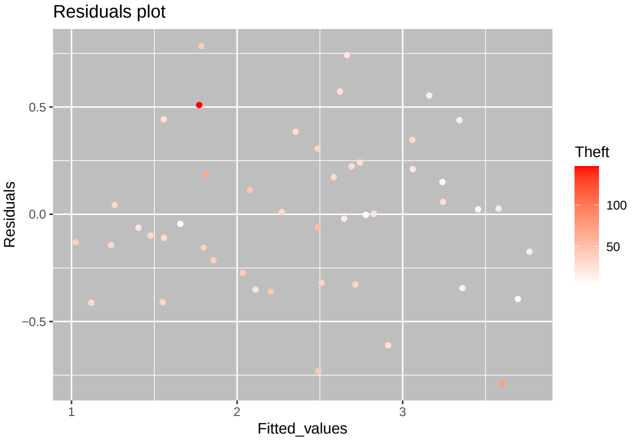


The fitted results of model 6 seems good from figures above.

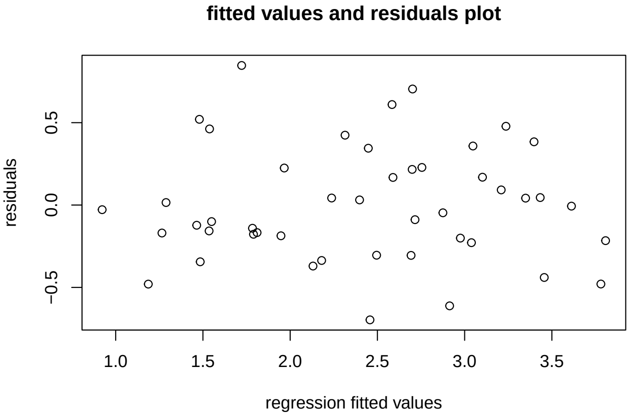
The fitted model is given by:

$$\sqrt{\hat{Volun}} = 3.386915 - 0.013038 Race_i - 0.025713 Fire_i + 0.005356 Theft_i + -0.011476 Age_i + 0.000030 Income_i \tag{7}$$

Similarly, we find a strange positive value as coefficent of Theft.The reason may be the extreme data point(deep red) as shown in the below figure cause overfitting.



Let's remove the point and refit model.After sereval practices on model fitting and selection , the R output results suggest us to remove income and Theft.



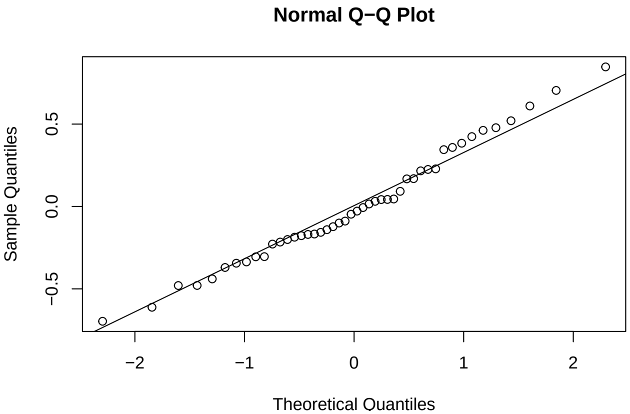
So the final model is given by:

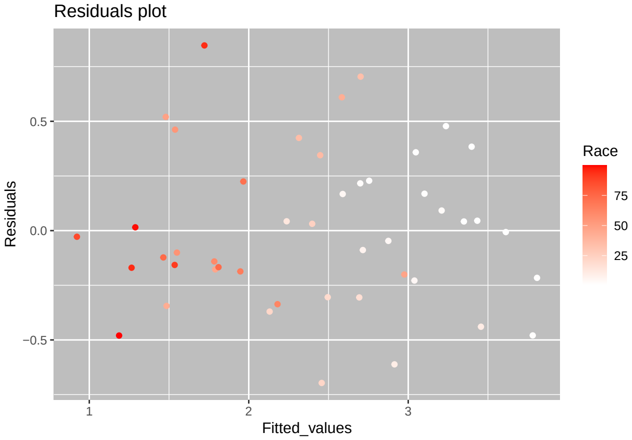
$$\sqrt{Volun} = \beta_0 + \beta_1 Race_i + \beta_2 Fire_i + \beta_3 Age_i + \epsilon_i \tag{8}$$

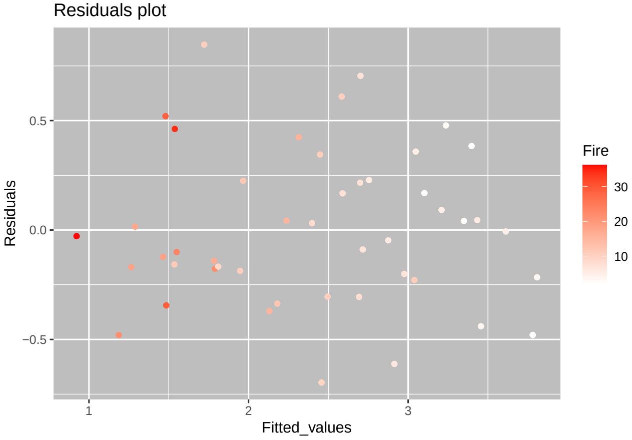
The fitted model is:

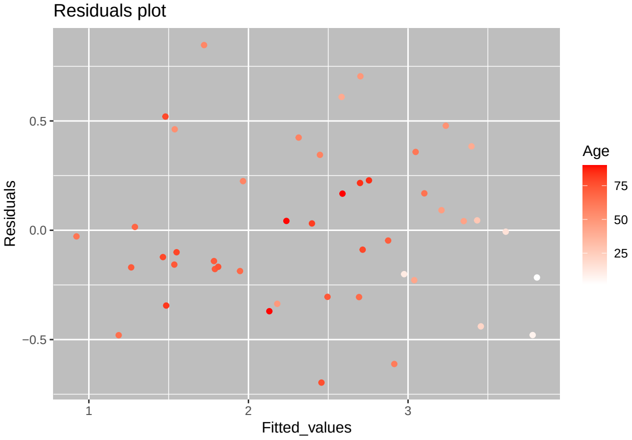
$$\sqrt{\hat{Volun}} = 3.95811 - 0.01318 Race_i - 0.03184 Fire_i - 0.01184 Age_i \tag{9}$$

The result of model (8) show that all three factors have negative impact on $\sqrt{Volun}$.The residuals show good patterns shown by QQ plot.We can clearly see the general color changes when fitted values increase from the plot below which also suggests a good fitted results.So to conclude , Race,Fire and Age affect the acceptance to a large extent.









Conclusion

From the overall analysis ,we may find racial composition affect the underwriting practice to a large extent while the age of house do not show such strong effect on “rejection”.However, when we model on Volun as response variable , we find both racial composition and age of housing affect the underwriting practices to an obvious large extent after controlling for factors like fire and others according to each coefficients meaning in regression model is showing how response variable change with the coveriate increase when keeping the other covariates unchanging.The racial composition plays an important role may imply the race discrimination exists in the insurance company while the age of house also may be taken into account of decision of insurance company.

To summarize,both of both racial composition and age of housing may affect the “accept” and “reject” significantly according to the 3 model (4) (5) and 9) when the other conditions keep the same.Meanwhile,racial composition is obviously influential while the influence age of house is shown much weaker.