Project Part B

Rushabh Khara

Introduction

The U.S. Commission on Civil Rights investigated claims of insurance redlining in Chicago. Initial data from the Illinois Department of Insurance highlighted policy actions by ZIP code from December 1977 to February 1978. This data represented over 70% of homeowner policies in Chicago. FAIR plan policies, often chosen by those denied standard insurance, were also considered. The Chicago Police provided 1975 theft data, emphasizing that insurers often use past years' data for decisions. Fire data for the same year was sourced from the Chicago Fire Department. Both datasets were organized by ZIP code. Lastly, the US Census Bureau offered demographic and residential data for Chicago's ZIP codes. To normalize differences, thefts were calculated as incidents per 1,000 residents, with similar adjustments for fire and insurance data. The goal of the following analysis is to explore the extent to which racial composition and age of housing affected underwriting practices after controlling for factors like fire, theft, and income.

Exploratory Data Analysis and Data Manipulation

Table 1: Summary Table

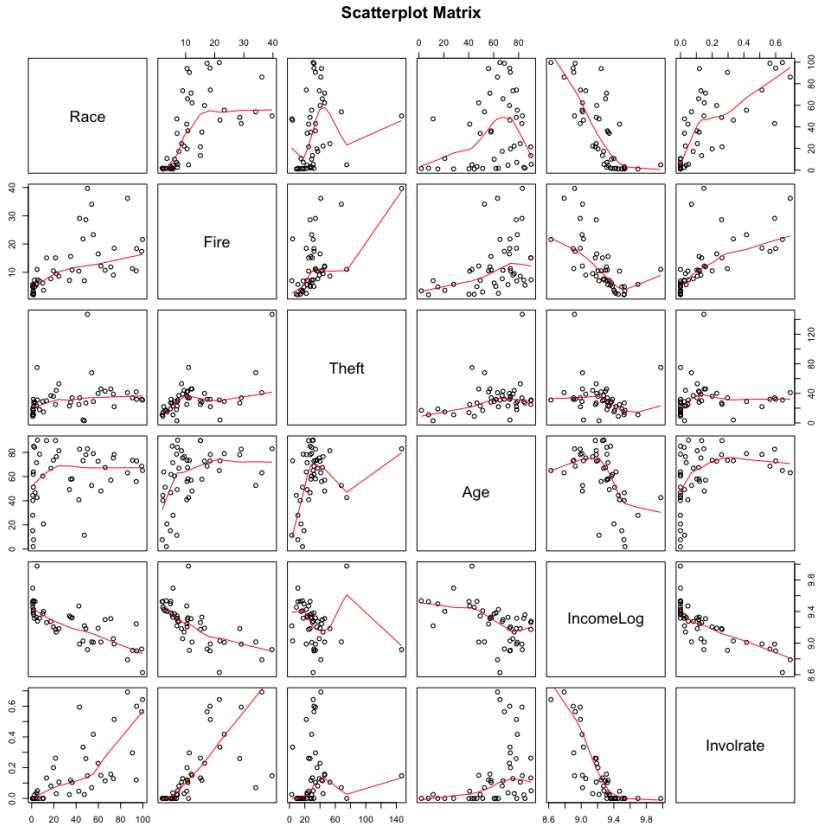
Zip	Race	Fire	Theft	Age	Volun	Invol	Income
Min. :60607	Min. : 1.00	Min. : 2.00	Min. : 3.00	Min. : 2.00	Min. : 0.50	Min. :0.0000	Min. : 5583
1st Qu.:60618	1st Qu.: 3.75	1st Qu.: 5.65	1st Qu.: 22.00	1st Qu.:48.60	1st Qu.: 3.10	1st Qu.:0.0000	1st Qu.: 8447
Median :60630	Median :24.50	Median :10.40	Median : 29.00	Median :65.00	Median : 5.90	Median :0.4000	Median :10694
Mean :60631	Mean :34.99	Mean :12.28	Mean : 32.36	Mean :60.33	Mean : 6.53	Mean :0.6149	Mean :10696
3rd Qu.:60642	3rd Qu.:57.65	3rd Qu.:16.05	3rd Qu.: 38.00	3rd Qu.:77.30	3rd Qu.: 9.65	3rd Qu.:0.9000	3rd Qu.:11989
Max. :60657	Max. :99.70	Max. :39.70	Max. :147.00	Max. :90.10	Max. :14.30	Max. :2.2000	Max. :21480

Upon examining the `insure` dataset summary, a observations is evident. The scale of the Income variable substantially surpasses other variables in the dataset. Furthermore, it's common to apply a natural logarithm transformation to `Income` variables to mitigate potential skewness, ensuring that exceptionally high values don't unduly influence the results.

It is also important to note that in order identify insurability of a neighbourhood, we need a comprehensive variable that would include information regarding voluntary as well as involuntary policies in order to make an informed decision. It is not advisable to use `Invol` as a sole response variable a `Invol` of value 0 doesn't mean that neighbourhood would be redlining free. It could be that the residents, due to economic constraints, might not be in a position to afford insurance despite the need. It's possible that the residents are not aware of the insurance options available to them or the benefits of such policies. This lack of awareness could lead to a low number of policies, both voluntary and involuntary. Sometimes, external incentives or government programs might cover some risks, reducing the perceived need for private insurance. It is just to say that there can be multitudes of reasons, making the `Invol` a volatile variable and more prone to external factors. We introduce a new and more robust variable `Involrate`. The measure is a proportion that indicates the share of involuntary market activity relative to the total insurance activity in the neighborhood. This metric provides a standardized way to gauge the reliance on the involuntary market.

$$Involrate = \frac{Invol}{Invol + Volun}$$

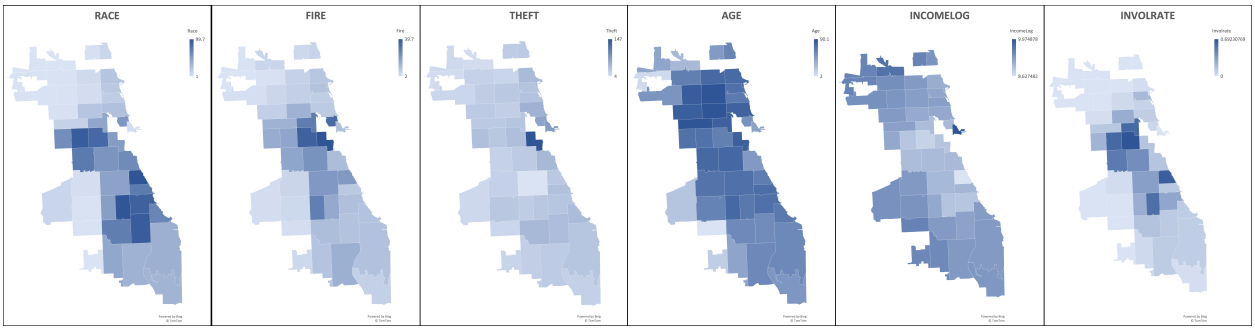
It is not to say that the new introduced variable has no pitfalls. In sparsely populated areas, even a few involuntary policies can drastically affect the Involrate, making it seem disproportionately high. Hence, this variable should only be used as a device to better understand the situation rather than taking it as a definitive solution. If decision-makers rely too heavily on `Involrate` without considering other contextual and qualitative data, it can lead to a narrow perspective on insurability.



The scatterplot matrix reveals the following insights:

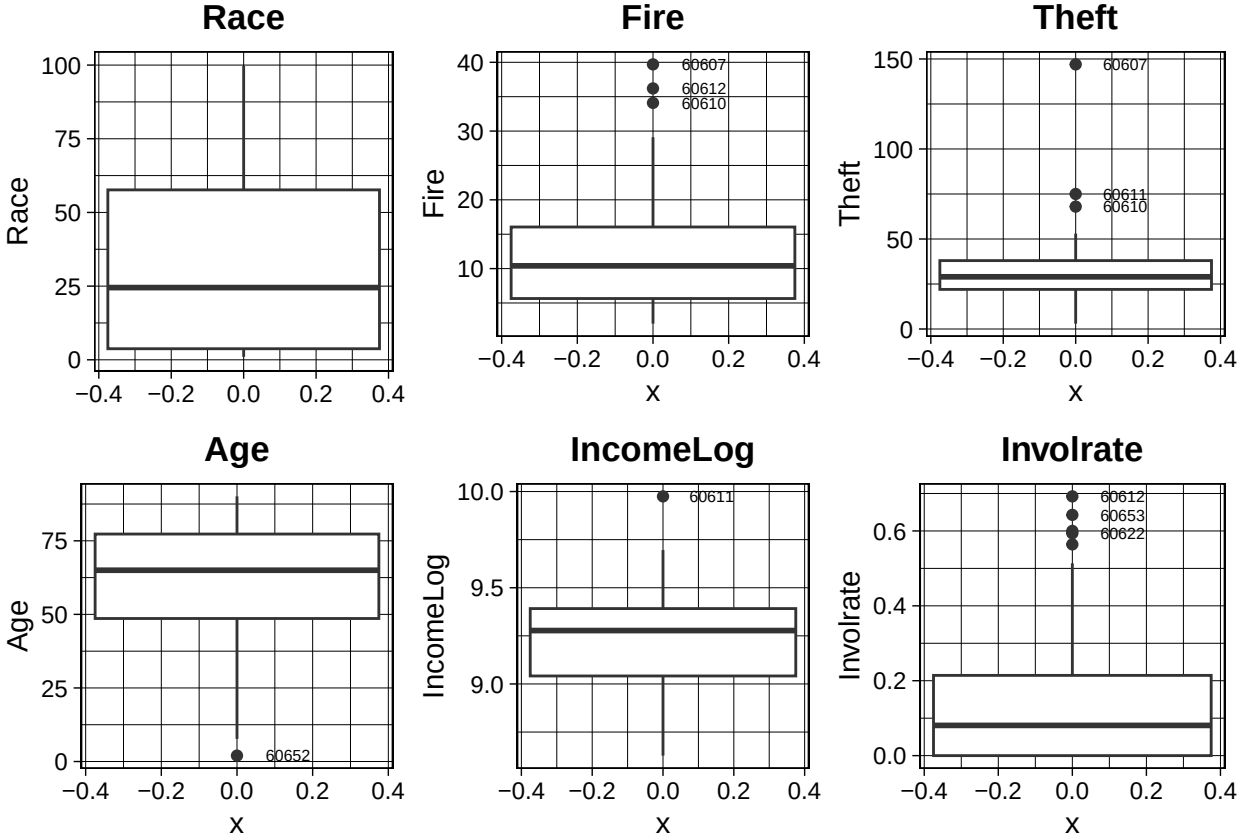
1) A strong positive correlation between `Race` and `Involrate`, suggesting racial factors may influence insurance involrate.
2) `Fire` incidents correlate positively with `Theft` and `Age`, and negatively with `IncomeLog`, implying neighborhoods with more fires might have higher thefts and older houses, but lower income levels.
3) `Age` negatively correlates with `IncomeLog`, suggesting older neighborhoods might have lower incomes.
4) `IncomeLog` has a pronounced negative relationship with `Involrate` , indicating areas with low incomes might get their insurances rejected more often.
5) The logarithmic transformation of our Income data not only reduces its scale but also yields a more normalized distribution.

To enhance our data visualization and facilitate a more robust interpretation, we can project the values onto a map of Chicago as heatmap.



It's important to mention that zip codes `60627` and `60635` are now defunctional and thus excluded from our mapping. However, this exclusion is statistically negligible due to the ample sample size available for analysis.

The heatmap confirms our initial observations, effectively highlighting the varying degrees of correlation between `Involrate` and other variables through color intensities. It reveals a strong positive correlation with `Race` and `Fire`, a almost no correlation with `Theft` and `Age`, and a high negative correlation with `IncomeLog`. We also observe a notable negative correlation between `Race` and `IncomeLog`, suggesting that lower `IncomeLog` may be the factor in insurance cancellations than `Race`. However, the simultaneous negative correlation of both `Involrate` and `Race` and other variables with `IncomeLog` complicates the attribution of causality, making it challenging to isolate the effects of these variables. As long as insurance companies utilize income to deny insurance without discriminating, it is legal to reject insurance based on low income.



Notably, zip codes `60607`, `60610`, `60611`, and `60612` recur as outliers across various boxplots. It would be good practice to delve deeper into these and other outlying observations to extract more granular insights. We would focus on outliers in `Fire`, `Theft`, `IncomeLog`, and `Involrate` and try to examine `Race` and `Age` values of the outlying observations in order identify how these factors relate.

Table 2: Outlier Table

Zip	Race	Fire	Theft	Age	IncomeLog	Involrate
60610	54.0	34.1	68	52.6	9.015663	0.0697674
60611	4.9	11.0	75	42.6	9.974877	0.0000000
60622	43.1	29.1	34	82.7	8.986572	0.5937500
60612	86.2	36.2	41	63.1	8.789508	0.6923077
60607	50.2	39.7	147	83.0	8.917177	0.1475410
60653	99.7	21.6	31	65.0	8.627482	0.6428571
60621	98.9	17.4	32	68.6	8.925321	0.5641026
60652	1.4	3.4	17	2.0	9.535463	0.0000000

Upon examining the summary, it's evident that outliers are predominantly associated with observations having a significant minority composition. Notably, outliers related to `Involrate` exhibit a high minority `Race` composition, with the exception of 60622. This may suggest the potential incorporation of `Race` in underwriting processes. Let's delve deeper into the primary outliers that repeatedly emerged.

1) ZIP code `60610` exhibits a minority `Race` composition near the 3rd quantile, exceptionally high incidents of both `Fire` and `Theft`, a lower `IncomeLog`, and a `Involrate` below the median.

2) ZIP code `60611` is characterized by a lower minority `Race` composition, a heightened theft rate, the highest `IncomeLog`, and an a `Involrate` of 0.

3) ZIP code `60607` possesses a higher minority `Race` composition, the peak values in both `Theft` and `Fire` incidents, a diminished `IncomeLog`, and a comparatively low `Involrate`.

4) ZIP code `60612` possesses extremely high minority `Race` composition, high `Fire` and moderate `Theft` rate, a below median `IncomeLog` and the highest `Involrate`.

Modelling

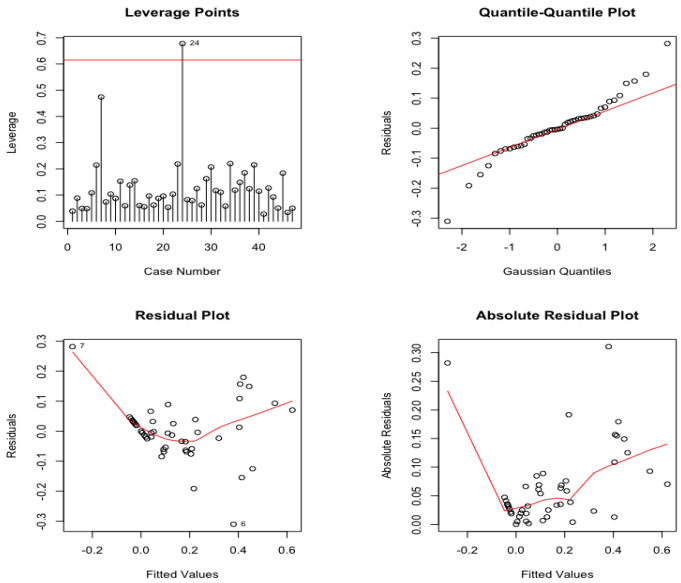
Concluding our exploratory data analysis and implementing the required transformations, we should now transition to constructing a robust linear regression model to assess insurability. We utilise robust linear regression model to reduce the influence of outliers on our model. It is important to take into account the sequencing of variables when fitting the model. Primarily, variables such as `Fire`, `Theft`, and `Income` should be considered before delving into factors like `Race` and `Age`. This sequence reflects the belief that insurance companies predicate their decisions on established legal frameworks. Subsequently, `Race` and `Area` are introduced as auxiliary variables that might explain any residual variability within the fitted model. However, the ordering won't affect the coefficients of the model, it is more about how we want to interpret the model. First, we start by fitting a full model.

Table 3: Summary of Robust Linear Model

Estimate	Value	Std. Error	t value
(Intercept)	2.7790601516053	0.900857166019377	3.08490652728573
Fire	0.0121339033583535	0.00204730214158228	5.92677705547444
Theft	-0.003729241662672	0.000658109039169348	-5.66660149111305
IncomeLog	-0.293721083828627	0.0934474279090759	-3.14316927068786
Race	0.00146704176933507	0.000581427630316287	2.52317174630491
Age	0.000187461266703543	0.000640852922927247	0.292518392281445

$$Involrate = 2.7790 + 0.0121 \times Fire - 0.0037 \times Theft - 0.2937 \times IncomeLog + 0.0014 \times Race + 0.0001 \times Age$$

Upon reviewing the model's summary output, it is evident that all variables, except `Age`, hold statistical significance. Significance of `Race` in the model indicates insurance company utilising `Race` for redlining. However, no conclusions can be made before examining the diagnostic plots.



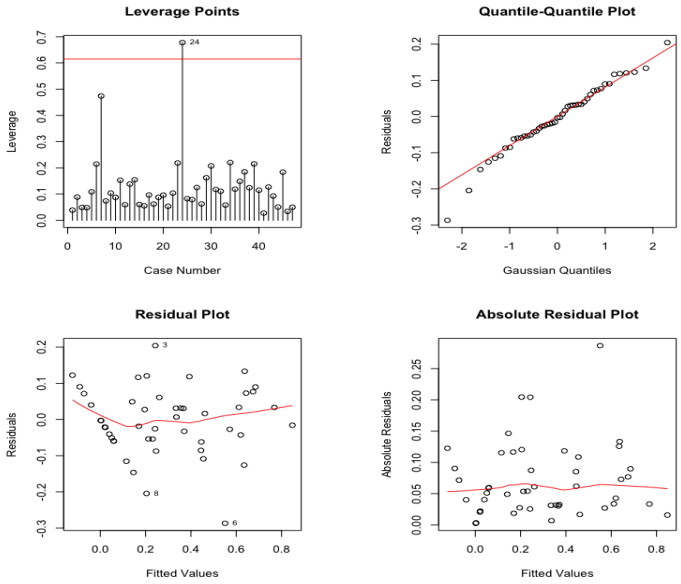
The residual plots exhibit a funnel-shaped pattern, signaling an issue with heteroscedasticity. This is further evidenced by the QQ-plot, where the points diverge from the expected line at both ends. Given that `Involrate` includes zero values, a logarithmic transformation isn't feasible. Therefore, we opt for a square root transformation of `Involrate` to mitigate the heteroscedasticity.

Table 4: Summary of Robust Linear Model

Estimate	Value	Std. Error	t value
(Intercept)	1.26148708499378	1.20485700982761	1.04700149038787
Fire	0.013093585613007	0.00273817696030564	4.78186245915439
Theft	-0.00325189213310123	0.000880192020426298	-3.69452580531946
IncomeLog	-0.141738182430743	0.124981842642287	-1.1340701931913
Race	0.00379499881801145	0.000777633994065346	4.8801863691321
Age	0.00260497327068164	0.000857112720620574	3.03924234002216

$$\sqrt{Involrate} = 1.2614 + 0.0130 \times Fire - 0.0032 \times Theft - 0.1417 \times IncomeLog + 0.0037 \times Race + 0.0026 \times Age$$

Upon applying transformations to the dependent variable, the resultant model exhibits notable difference. Notably, the variable `Age` has emerged as statistically significant predictors along with existing significant variables, whereas `IncomeLog` has been rendered non-significant. To substantiate the robustness of this revised model, it is important to analyse the associated diagnostic plots.



The diagnostic plots can be interpreted in a structured manner as follows:

1) Residual Plots: The residuals' distribution displays a nearly uniform spread across the range, pointing to the assumption of homoscedasticity being met. The Lowess curve, while exhibiting a slight curvature, is within acceptable bounds, allowing us to proceed under the linearity assumption.

2) Normal Q-Q Plot: The data points' alignment closely with the reference line suggests the residuals are approximately normally distributed. A few discrepancies at the beginning can be considered statistically insignificant given the sample size, reinforcing our confidence in the normality assumption.

3) Leverage Plot: The plot identifies a single observation (24) as marginally exceeding the leverage threshold. Its minimal departure from the reference line implies it's not problematic from an influence perspective.

Having addressed and validated the primary assumptions for our regression model, we can confidently proceed with our analysis.

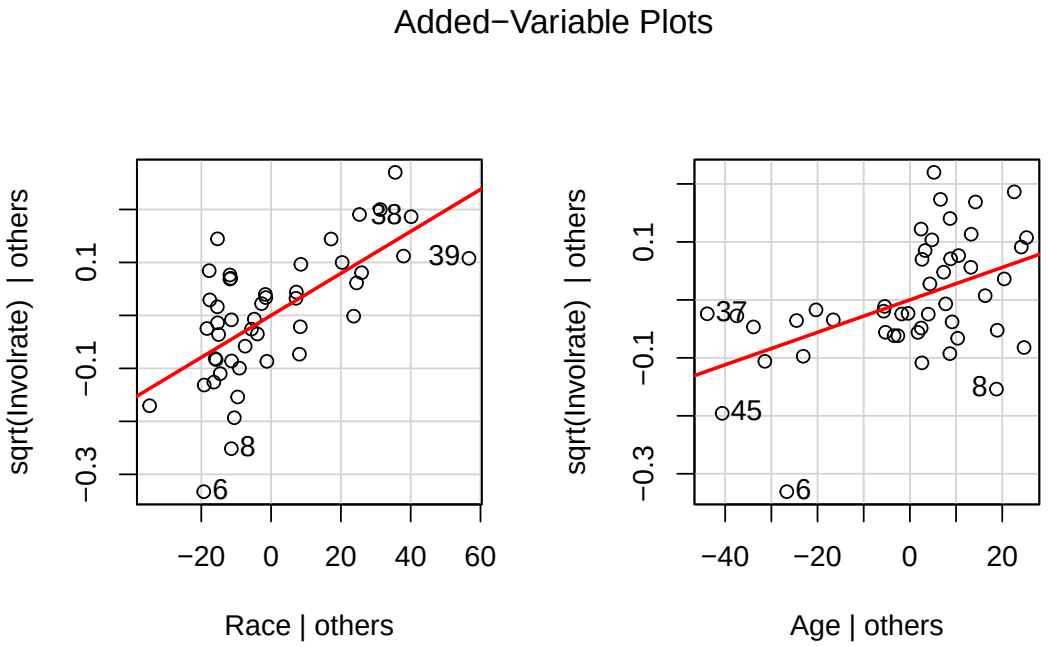
Table 5: VIF table

Fire	Theft	IncomeLog	Race	Age
2.733322	1.621824	4.050811	2.70549	1.577322

The Variance Inflation Factor (VIF) analysis indicates an absence of multicollinearity concerns within the model. This conclusion is supported by the application of a commonly accepted rule of thumb, which states that a VIF below 10 signifies a model is free from multicollinearity issues. This signifies that standard errors of `Race` and `Age` are not inflated by other variables.

Isolated Effects

In the scatterplot matrices, we analyzed the bivariate associations involving `Involrate` and other covariates. However, these matrices do not provide insights into the partial effect of the predictor variable, controlling for other variables in the model. To address this, we employ added-variable plots to explore the adjusted effect.



The added-variable plot indicates that `Race` and `Age` have a significant positive partial correlation with `Involrate`, after adjusting for potential covariates. This provides statistical support to the claim that private insurance companies had been employing non-permissible metrics, specifically `Race` composition and `Age` of the house, as determinants in their insurance approval decisions.

Location

To enhance the utility of location in our analytical model, we employ the `chredlin` dataset from the `faraway` package, which encompasses the same data with zip codes distinctly classified as either north side or south side of Chicago. Evaluating the robustness of our model by focusing solely on one geographic area could provide valuable insights into its performance.

Table 6: Summary of North Side Model

Estimate	Value	Std. Error	t value
(Intercept)	-1.14814531900679	1.96183792298426	-0.585239639602993
Fire	0.0123431475135117	0.00498060421091598	2.47824299840153
Theft	-0.00351742423386782	0.00128584394750041	-2.73549853441036
IncomeLog	0.103664805150166	0.201319450240799	0.5149269234849
Race	0.00604052434641658	0.00174083546943251	3.4698996272093
Age	0.00413131777461464	0.00150082523089062	2.75269744243507

Table 7: Summary of South Side Model

Estimate	Value	Std. Error	t value
(Intercept)	2.76117163222901	1.27627631691809	2.16345911588847
Fire	0.0156763668148406	0.00352396375519339	4.44850398694873
Theft	-0.000879526092746236	0.00188230812311192	-0.467259362028446
IncomeLog	-0.301270260178217	0.134353757376739	-2.24236572210951
Race	0.00233842765142307	0.000952932686353845	2.45392742311156
Age	0.00144583270506393	0.00109788458061426	1.31692595978986

As we can understand from the summary outputs of model above, our model generalises well in the north side, but performs differently in the south side. `Age` and `Theft` turn out to be non significant contributor, whereas `IncomeLog` is significant. This suggests that our model doesn't generalise well and should only be used at macroscopic scale. It is also possible due to the small sample size resulting by subsetting the data, the model fit doesn't have enough information. However, in both the models, `Race` appears to be a significant factor which means it was used as an identifying factor when deciding insurability of a neighbourhood.

Final Model

$$\sqrt{Involrate} = 1.2614 + 0.0130 \times Fire - 0.0032 \times Theft - 0.1417 \times IncomeLog + 0.0037 \times Race + 0.0026 \times Age$$

Conclusion

Based on the regression analysis, it can be statistically inferred that underwriting practices were significantly influenced by the variables `Race` and `Age` in addition to `Fire` and `Theft`, potentially indicating discriminatory practices in insurance underwriting. The variable `IncomeLog`, however, did not emerge as a statistically significant predictor, suggesting its limited role in the decision-making process. Furthermore, when the model was applied to a specific geographic location (north side or south side), the covariates exhibited distinct behaviors, implying that the model's applicability may be more appropriate at a macroscopic level rather than a localized one. This observation underscores the possibility that the current model may not fully capture the intricacies of the issue at hand. In light of these findings, it is recommended that policymakers consider these statistical insights to formulate more equitable policies and regulations that actively prevent redlining, particularly those practices rooted in racial discrimination.