Determination of key Factors Influencing Insurance Claim Activity INVOLACT

For the Bachelor of Science Honours Degree in

Financial Mathematics and Industrial Statistics

By

J.M.K.L Jayasinghe SC/2021/12456

Supervisor :

Dr. W.A.R. De Mel

Department of Mathematics

University of Ruhuna

Matara

2024.07.12

Abstract

In this case study, investigate the impact of racial composition, fire and theft rates, age of housing and income on insurance redlining in Chicago in 47 zip codes. Here we are going to analyze this relationship using regression approach in multiple linear regression model. The objective of researching the effects of factors such as race, fire, theft, age, income and side on Involact is to understand how these factors influence individuals' levels of engagement in various activities and contexts. So in this research mainly focused to understand the factors influence to insurance claim activity Involact.

Key Words : *Involact, Race, Theft, Fire, Income, Age of housing, Chicago, Multiple linear regression*

Acknowledgement

In preparation for this case study, we had to take the help and guidance of some respected persons, who deserve our deepest gratitude. First of all, our sincere gratitude goes to our supervisor Dr. W.A.R. De Mel, Department of Mathematics, Faculty of Science, and the University of Ruhuna for giving us good guidelines for the case study through numerous consultations.

We would thankful to all the supervisors in the supervisory board for their great support towards us. And also,we would be thankful to the Head of the Department and all the other staff members of the Department of Mathematics for their assistance during the Case Study I Course Unit. Specially thankful to Ms. R.M.V Lakmini for guidance of Regression analysis part.

Finally, we are thankful to especially our group members, everyone who played a part, big or small, to make this research a success.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of the study

Insurance redlining is a discriminatory practice wherein insurance companies limit or deny services to residents of certain areas based on racial or socioeconomic characteristics. This phenomenon has significant implications for fairness and equity in insurance practices. The Chredlin dataset, provided by Juan G. Faraway, offers a valuable resource for examining these issues, as it contains detailed information on various factors that might influence insurance claim activity. The dataset includes variables such as Race, Fire, Theft, Age, Income, and Side, alongside Involact, which measures the involvement or activity in insurance claims. By analyzing this data using a multiple linear regression model, we aim to understand the underlying patterns and relationships that contribute to insurance redlining. This study focuses on identifying the factors that significantly impact Involact, thereby shedding light on the broader dynamics of insurance redlining. Through this analysis, we hope to uncover the extent to which racial and socioeconomic factors influence insurance claim activity and contribute to discriminatory practices within the industry.

1.2 Research Quection

1. How the racial composition in percentage of minority effect to the new FAIR plan policies and renewals per 100 housing units
2. How the fires per 100 housing units effect to the new FAIR plan policies and renewals per 100 housing units
3. How the thefts per 1000 population effect to the new FAIR plan policies and renewals per 100 housing units
4. How the percentage of housing units built before 1939 effect to the new FAIR plan policies and renewals per 100 housing units
5. How the median family income in thousands of dollars effect to the new FAIR plan policies and renewals per 100 housing units
6. How the north or south side of Chicago effect to the new FAIR plan policies and renewals per 100 housing units

1.3 Objectives

The primary objective of this research is to obtain the best-fitted line using a multiple linear regression model to analyze the Chredlin dataset. Specifically, the study aims to:

1. Investigate the relationship between Involact and the other factors in the dataset, including Race, Fire, Theft, Age, Income, and Side.
2. Identify the significant predictors of insurance claim activity (Involact).
3. Interpret the results to understand how each factor influences insurance redlining practices.

1.4 Significance of the Study

This study is significant as it addresses the critical issue of insurance redlining, a practice that perpetuates inequality and discrimination within the insurance industry. By analyzing the factors that influence insurance claim activity (Involact), this research provides valuable insights into how discriminatory practices can be identified and mitigated. Understanding the impact of variables such as Race, Fire, Theft, Age, Income, and Side on insurance claims can help policymakers and stakeholders develop more equitable insurance policies.

Additionally, the findings can assist insurance companies in creating strategies that are based on legitimate risk factors rather than discriminatory criteria, promoting fairness and inclusivity. This study also contributes to the broader field of social justice by highlighting the importance of addressing systemic discrimination in financial services. The methodological approach of using multiple linear regression and backward elimination offers a rigorous framework for analyzing complex datasets, making the results robust and reliable. Overall, this research aims to foster a more equitable insurance industry and inform future policies and practices that combat discrimination.

Chapter 2

Literature Review

Insurance claim activity is a multifaceted phenomenon influenced by a variety of factors. This section explores existing research on key determinants such as race, fire incidents, theft rates, and income levels. These factors are critical in understanding the dynamics of insurance practices and potential disparities, especially in the context of redlining.

Race :
The role of race in insurance claim activity has been extensively studied, highlighting persistent disparities in insurance practices. Research by Squires and Velez (1987) found that minority neighborhoods often face higher insurance premiums and reduced access to insurance services, even after accounting for other risk factors. This racial disparity is a manifestation of systemic biases that persist within the insurance industry. The Chredlin dataset and other studies confirm that race remains a significant predictor of insurance claim activity, indicating that discriminatory practices continue to affect minority communities. These findings underscore the need for regulatory measures to ensure fair treatment across racial groups.A study by **?** demonstrates that minority neighborhoods are still more likely to be underserved by voluntary insurance markets, necessitating the use of involuntary market solutions like FAIR plans.

Fire Incidents :
Fire incidents are a critical factor influencing insurance claim activity, as they represent a significant risk to property. Studies such as those by Kleindorfer and Kunreuther (1999) have shown that areas with higher incidences of fires tend to have higher insurance claims. This is because insurance companies price their policies based on the perceived risk of fire damage. The frequency and severity of fire incidents can vary widely based on geographic location, building materials, and community fire prevention measures. The Chredlin dataset provides valuable insights into how fire risks are distributed across different neighborhoods and how these risks influence insurance claims. Effective fire pre-

vention strategies and community education are essential in reducing fire-related claims and ensuring fair insurance practices.Research by Ansfield [2021] demonstrated a strong correlation between the frequency of fire incidents and elevated INVOLACT, as insurers adjust premiums and coverage options in response to heightened fire risks.

Theft Rates :
Theft rates are another significant determinant of insurance claim activity. Areas with higher crime rates, particularly theft, tend to see more frequent insurance claims as policyholders seek compensation for stolen property. Research by Skipper and Kwon (2007) indicates that insurance companies often adjust premiums and coverage terms based on local crime statistics. This means that neighborhoods with high theft rates may face higher premiums or more stringent coverage conditions. The Chredlin dataset helps illustrate the impact of theft rates on insurance claim activity, revealing patterns that can inform both insurers and policymakers. Strategies to reduce theft, such as improved security measures and community policing, can help mitigate these risks and promote equitable insurance practices.

Income Levels :
Income levels play a crucial role in influencing insurance claim activity. Low-income neighborhoods often face greater financial instability, making them more vulnerable to the impacts of loss and damage. Studies like those by Harrington and Niehaus (2003) have shown that lower-income individuals are less likely to have sufficient insurance coverage, leading to higher claim activity when losses occur. Moreover, income disparities can exacerbate issues of underinsurance and non-insurance, as lower-income households may struggle to afford adequate coverage. The Chredlin dataset highlights the correlation between income levels and insurance claim activity, emphasizing the need for policies that ensure affordable insurance options for all income groups. Addressing income-related disparities in insurance coverage is critical for promoting financial resilience and equity. Research by Li et al. [2022] found that lower-income neighborhoods tend to have higher INVOLACT, driven by limited access to preventive measures, higher vulnerability to environmental risks, and reliance on insurance as a safety net.

Geographic Side :
Geographic location, including whether a neighborhood is on the North or South side of a city, affects insurance engagement through varying regional risks, such as natural disasters and crime rates. Urban areas, particularly those historically subjected to redlining and with higher crime rates, show higher reliance on residual market insurance. Grace et al. [2004] found that high-risk locations, such as those prone to natural disasters or with high crime rates, have significant engagement with involuntary market solutions due

to the unwillingness of standard insurers to cover these risks.

Chapter 3

Materials and Method

3.1 Research Approach

The overall research approach for this study is quantitative research approach. This approach identify the relationships between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes. This likely involves collecting data on insurance policies, racial demographics, crime rates, age of housing and income levels within each zip code, and then using statistical methods to analyze and interpret these relationships. Quantitative method such as linear regression model, correlation analysis will be used to detect these relationship patterns.

3.2 Conceptual Model

The below figure illustrates the general conceptual model for this case study.

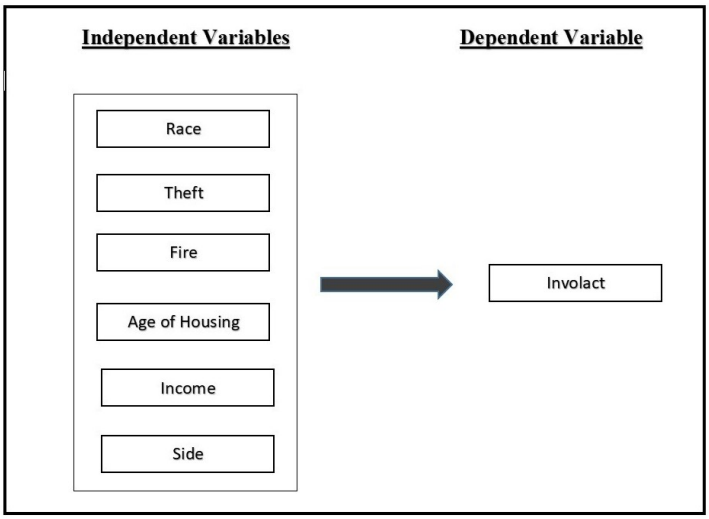

Figure 3.1: *Conceptual Model*

3.3 Research Design

The research design is a correlational research design. Correlational research is research designed to discover relationships among variables and to allow the prediction of future events from present knowledge. In our case study, our aim to explore relationship of insurance redlining (Involact) in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes.

Since we have single dependent variable (Involact) and several independent variables (Race, Theft, Fire, Income, Age of Housing). So, we use multiple linear regression as the statistical tool in modelling this relationship. Multiple linear regression is a statistical model that uses a straight line to estimate the relationship between a quantitative dependent variable and two or more independent variables.

The formula for a multiple linear regression is:

$$\mathbf{y} = \beta_0 + \beta_1 X_1 + \beta_2 X2 + ... + \beta_n X_n + \epsilon$$

Where,

y = the predicted value of the dependent variable

β_0 = the y-intercept (value of y when all other parameters are set to 0)

$\beta_1 X_1$ = the regression coefficient β_1 of the first independent variable X_1 (The effect that increasing the value of the independent variable has on the predicted y value)

... = do the same for however many independent variables we are testing

$\beta_1 X_1$ = the regression coefficient of the last independent variable

ϵ = model error (How much variation there is in our estimate of y)

Chapter 4

Data

4.1 Dataset

This data set is about the new FAIR plan policies and renewal (INVOLACT), racial consumption, fire and theft rates, age of housing and income in 47 zip codes.

The R code used to read this data set is in the appendix.

4.2 Metadata

The source of these data used for the case study has been obtained by the "Linear Models with R" written by Julian J. Faraway. Faraway [2005]

4.3 Data Dictionary

4.3.1 Variable Description

1. INVOLACT:
   Involuntary Market Assistance (INVOLACT) is a program that offers Fair Access to Insurance Requirements (FAIR) policies for high-risk or uninsurable properties. The number of FAIR policies issued or renewed per 100 housing units indicates the demand for alternative insurance options.

2. Race :
   In the context of insurance data, race can be an independent variable used to analyze disparities or patterns in insurance coverage, pricing, or access to INVOLACT. The race is a racial composition in percentage of minority.

3. Fire :
   The "fire fires per 100 housing units" metric is crucial for insurance companies to assess fire-related claims and premiums, as higher rates indicate higher fire damage risks, potentially influencing insurance pricing and policy availability.

4. Theft :
   In insurance datasets measures theft incidents per 1000 individuals, influencing insurance coverage and premiums. Higher rates indicate higher property loss or damage.

5. Age :
   Insurance data reveals that older housing units constructed before 1939 are more vulnerable to damage and require unique insurance considerations, impacting coverage options and prices. This age plays a significant role in INVOLACT.

6. Income :
   Income significantly influences insurance coverage, with higher income levels providing more comprehensive policies and lower-income individuals facing financial constraints, affecting insurers' risk assessment and premium pricing. In this income is the median income family in thousands of dollars.

7. Side :
   Normally, in this express North or South side in Chicago.

Dependent Variable	**Independent Variable**
INVOLACT	Race
	Fire
	Theft
	Age
	Income
	Side

Table 4.1: *Independent and Dependent Variables*

There are both categorical and numerical variables here. The independent variable Side, is a categorical variable. The other variables are numerical, and most of which are continuous. The dependent variable INVOLACT and independent variables are continuous and only Theft is discrete.

Variable	**Variable Type**	**Measurement Units**
INVOLACT	Numeric	Percentage
Race	Numeric	Percentage
Fire	Numeric	Percentage
Theft	Numeric	Theft per 1000 population
Age	Numeric	Percentage
Income	Numeric	Median family income in $1000
Side	Factor	n-: North , s-: South

Table 4.2: *Data Dictionary Table*

Chapter 5

Results

5.1 Exploratory Data Analysis

Variable	Minimum	Q1	Median	Mean	Q3	Maximum	Total Observation
Race	1	3.75	24.5	34.99	57.65	99.7	47
Fire	2	5.65	10.4	12.28	16.05	39.7	47
Theft	3	22	29	32.36	38	147	47
Age	2	48.6	65	60.33	77.3	90.1	47
Involact	0	0	0.4	0.6149	0.9	0.2	47
Income	5.583	8.4447	10.694	10.696	11.989	21.48	47
Side	0.0000	0.0000	0.0000	0.4681	1	1	47

Table 5.1: *Descriptive Statistics Table*

The above Table illustrates the summary statistics of each variable. According to this table, the total count of observations is 47 and there are no missing values and duplicate values.

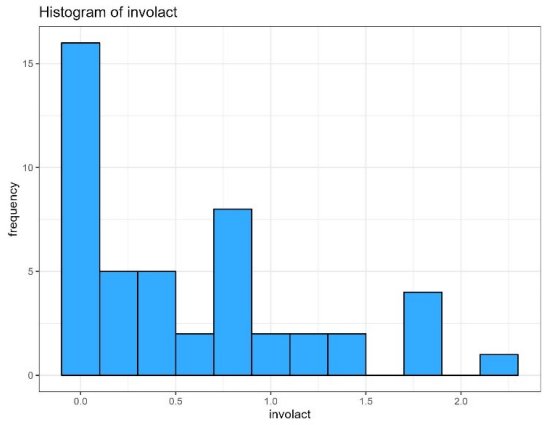

Figure 5.1: *Histogram of Involact*

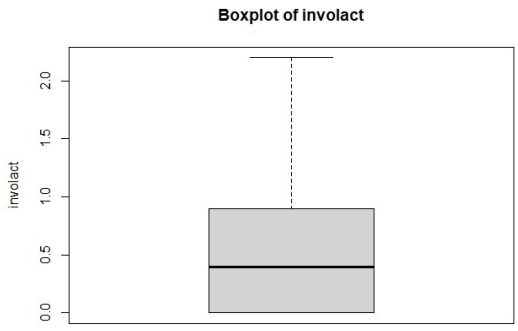

Figure 5.2: *Boxplot of Involact*

Figure 5.1 shows that the Involact values are positively skewed because the majority of the data are on the left side of the mean. According to the Boxplot, we can say that there are no outliers in Involact.

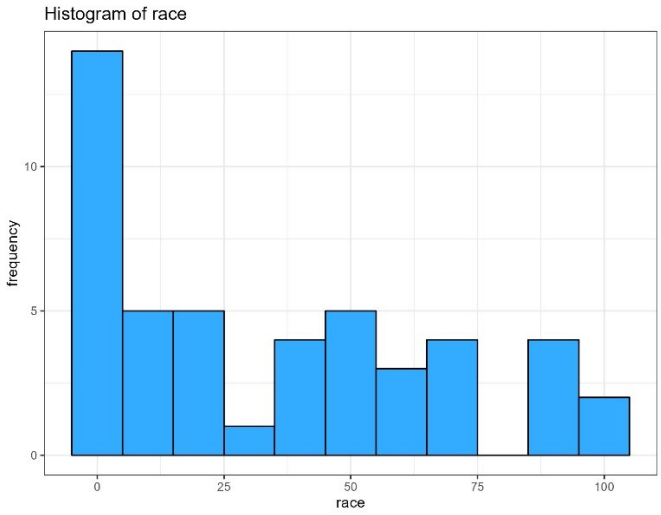

Figure 5.3: *Histogram of Race*

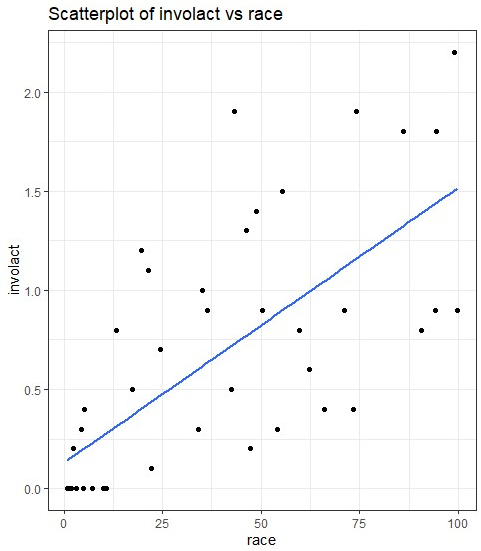

Figure 5.4: *Scatterplot of Race vs Involact*

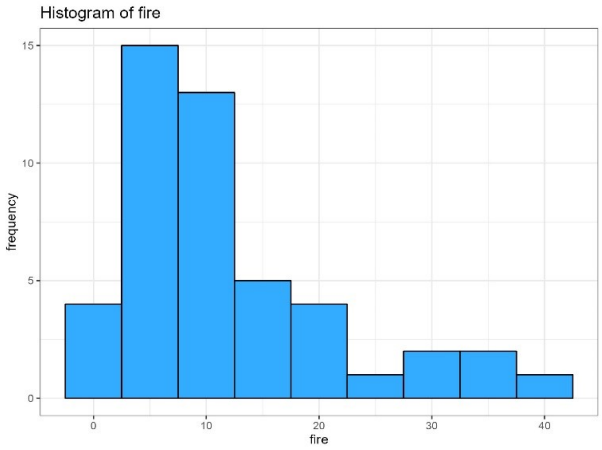

Figure 5.5: *Histogram of Fire*

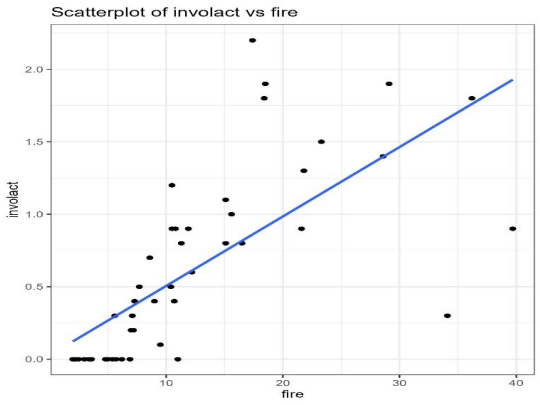

Figure 5.6: *Scatterplot of Fire vs Involact*

In the above graph 5.3, It is shown that the histogram of "Race" and figure 5.5 is shown that the histogram of Fire. These distributions are also positively skewed.

The figure 5.4 illustrates the relationship between the Involact and the race. It has a positive relationship. Same as the scatterplot of Fire vs Involact.

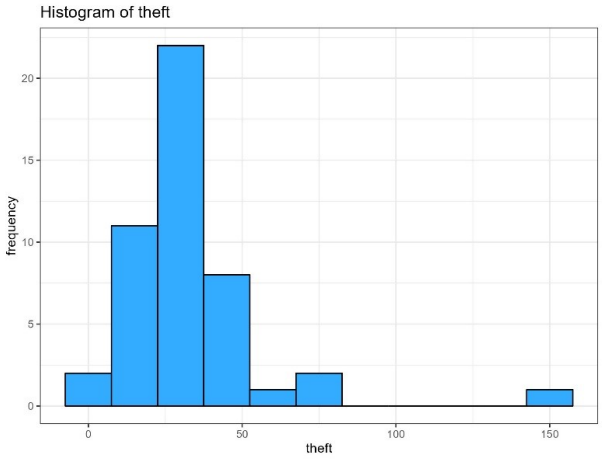

Figure 5.7: *Histogram of Theft*

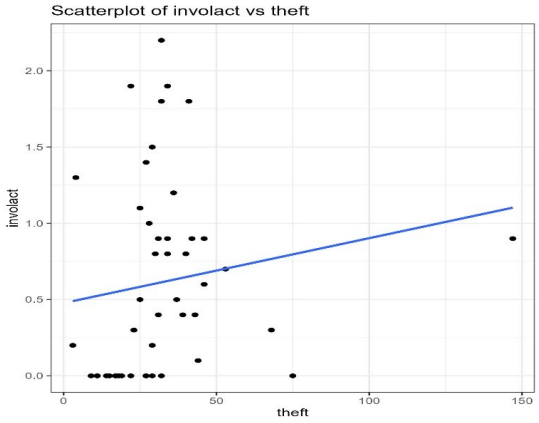

Figure 5.8: *Scatterplot of Theft vs Involact*

According to the figure 5.7 and figure 5.8, this data set is positively skewed because the majority of data points are in the left of the mean.

The scatter plot of Theft vs Involact shows that they have positive relationship. It means, if the theft incidents are increasing then the involact will go up.

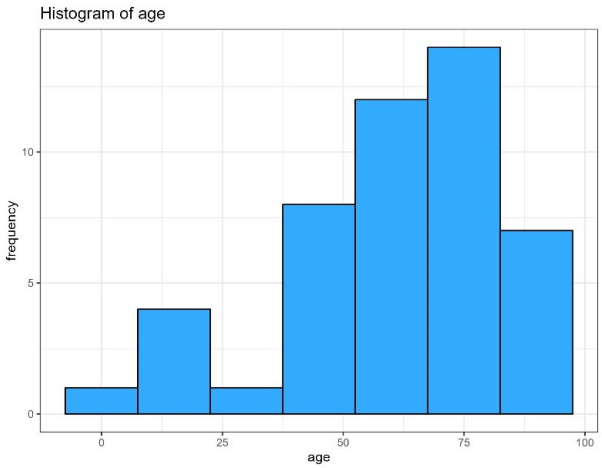

Figure 5.9: *Histogram of Age*

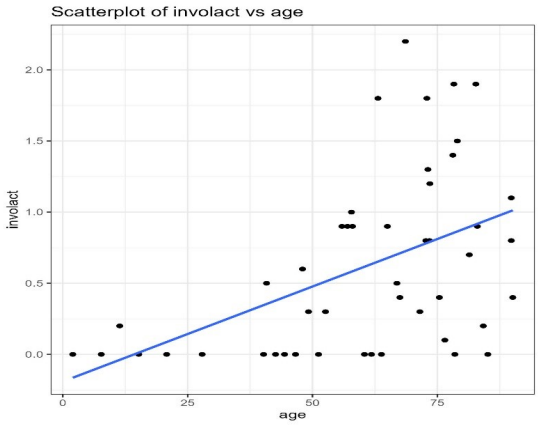

Figure 5.10: *Scatterplot of Age vs Involact*

The figure 5.9 shows that the Age values are negatively skewed because the majority of data points are in the right side of the mean.

The figure 5.10 illustrates the relationship between the age and involact.

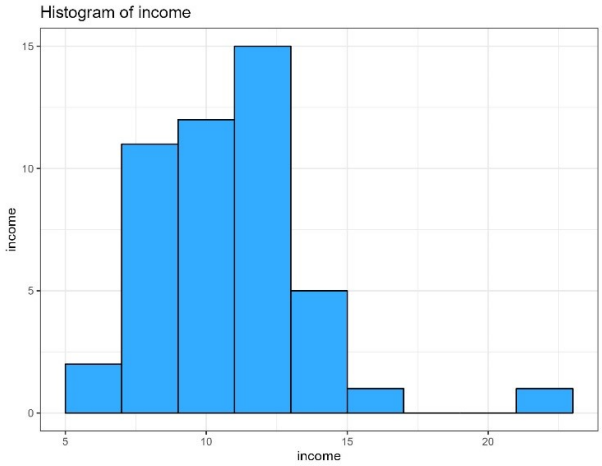

Figure 5.11: *Histogram of Income*

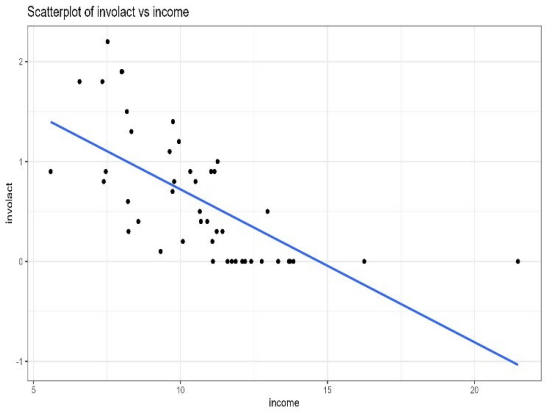

Figure 5.12: *Scatterplot of Income vs Involact*

According to the figure 5.11 and figure 5.12, this data set is positively skewed because the majority of data points are in the left if the mean.

The scatter plot of income vs involact shows that they have negative relationship. It means, when the income is decreasing then the involact will go up.

The below Boxplots show the dispersion of our independent variables of this data set.

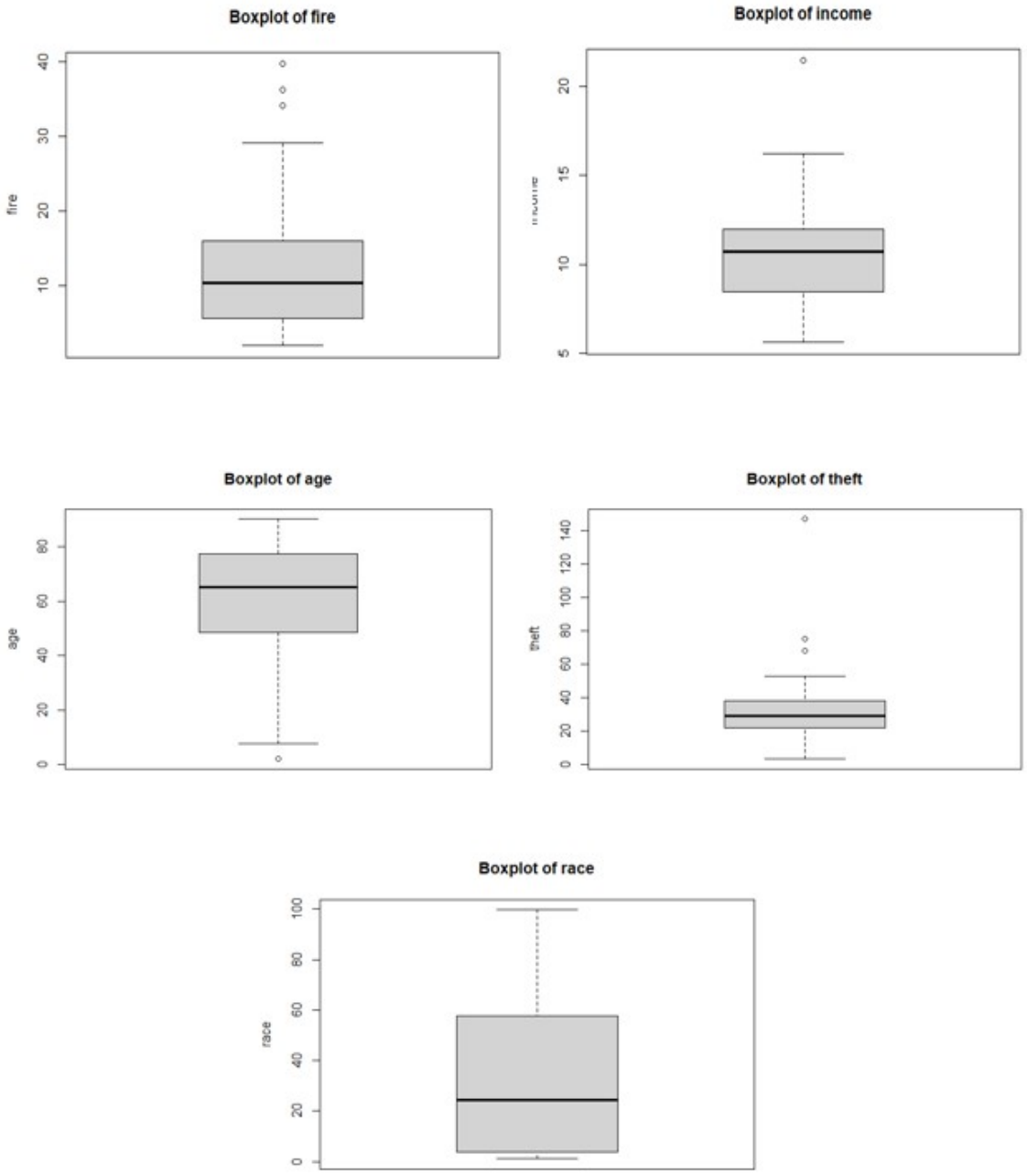

Figure 5.13: *Boxplot of Race, Fire, Theft, Age, Income*

According to these, we can say that fire, income, age and theft have the outliers and race has no outliers.

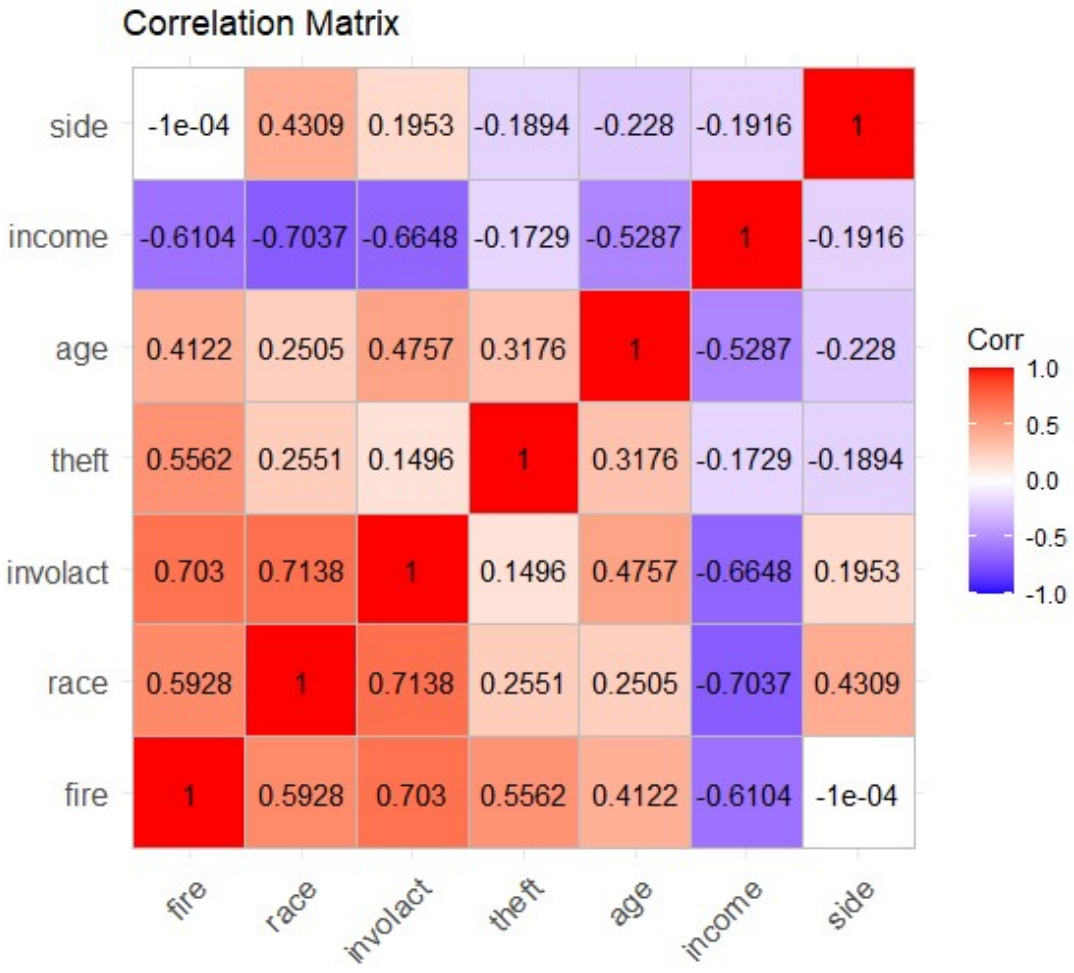

Figure 5.14: *correlation matrix*

Figure 5.14 illustrates the heat map of the variables. According to this figure, the highest correlation is between involact and race, which is scaled at 0.7138

5.2 Quantitative Analysis

According to the above Boxplots, we can see that there is a wide range in the race variable and The response involact has a large number of zeros.

	Estimate	**Std.Erroe**	**t value**	**Pr(value)**
(Intercept)	0.129218	0.096611	1.338	0.188
race	0.013882	0.002031	6.836	1.78e-08 ***

Table 5.2: *relationship between involact and race*

Homeowners with a higher percentage of minorities tend to get more default fair plan insurance. Insurance companies could argue that their decision to deny coverage in certain neighborhoods was based on the substantial fire-related losses they had experienced in those areas, and any resulting discriminatory impact was an unintended consequence of a legitimate business practice. Below figure shows Relationship between race and fire

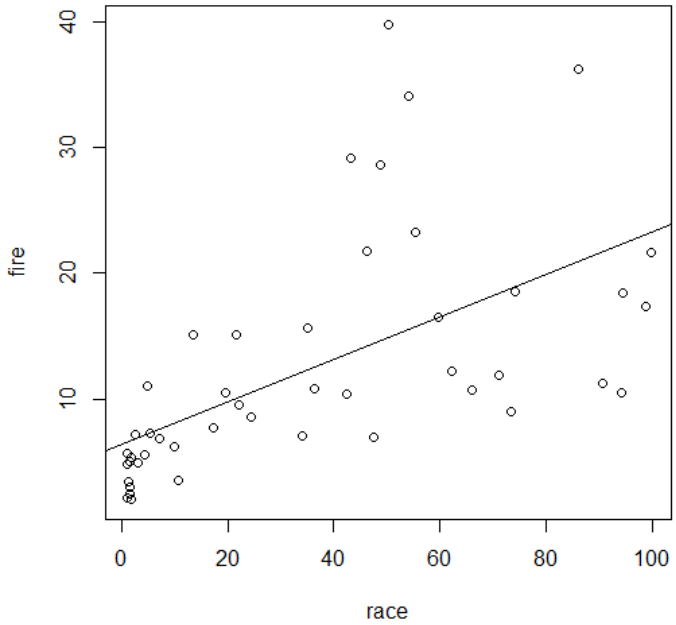

Figure 5.15: *Relationship between race vs fire*

In order to observe the effect of income differences, race variables were initially removed and income variables were included in the analysis instead.Due to the skewness in the income data, I choose to use the logarithm of income.

5.2.1 Correlation Analysis

	Estimate	Std.Erroe	t value	Pr(value)
(Intercept)	-1.185540	1.100255	-1.078	0.287550
race	0.009502	0.002490	3.817	0.000449 ***
fire	0.039856	0.008766	4.547	4.76e-05 ***
theft	-0.010295	0.002818	-3.653	0.000728 ***
age	0.008336	0.002744	3.038	0.004134 **
log(Income)	0.345762	0.400123	0.864	0.392540

Table 5.3: *Correlation Coefficient Table*

According to the details obtained above, R squared value of this Chredlin dataset is 0.7517. Thearefor, the model is effectively capturing the relationship between the independent variables and Involact. .

Regresion Equation:

Involact = -1.185540 + 0.009502 race + 0.039856 fire -0.010295 theft + 0.008336 age + 0.345762 log(income)

To ensure the validity and reliability of this model, the best way is checking model assumptions at first.

5.3 Check the Model Assumptions

5.3.1 First Assumption

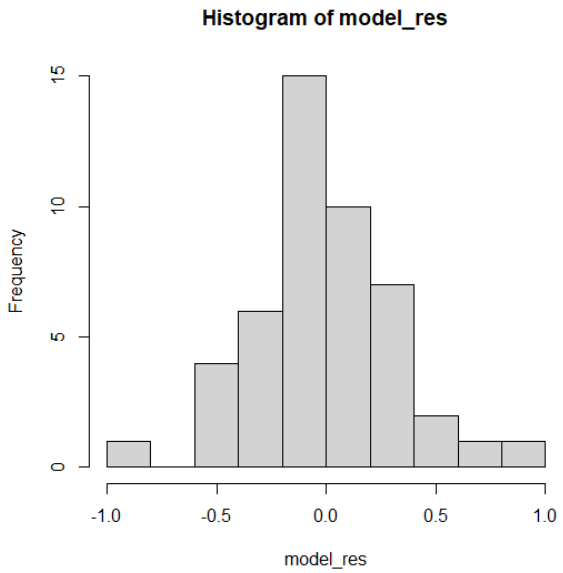

Figure 5.16: *Histogram of Model-residuals*

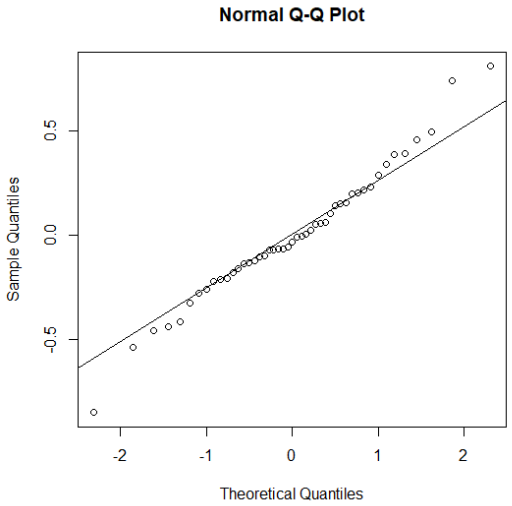

Figure 5.17: *Normal q-q plot*

The histogram of the model residuals is used to visualize the distribution of the residuals, which are the differences between the observed and predicted values from the regression model. By checking above histogram and QQ-plot, we can see that the shape of the histogram is symmetric and straight line on the q-q plot. Thus, we can say that the data is followed the first assumption.

5.3.2 Second Assumption

This assumption states that the relationship between the dependent variable and each of the independent variables is linear. In other words, the effect of the independent variables on the dependent variable is additive and proportional.

The above scatter plots which are plot each independent variable against to the dependent variables, are shown that there must be a linear relationship between the dependent variable and the independent variables. Therefore, we can see that the data we considered to this research is follow the second assumption for the multiple linear regression.

5.3.3 Third Assumption

*Third Assumption The third assumption of multiple linear regression is homoscedasticity, which means that the variance of the residuals (the errors) should be constant across all levels of the independent variables. This assumption ensures that the spread or "noise" around the predicted values is the same for all predicted values. When homoscedasticity is violated (i.e., heteroscedasticity), the precision of the coefficient estimates decreases, leading to inefficient estimates and biased standard errors, which affects hypothesis testing. To check for homoscedasticity, we can plot the residuals against the predicted values and look for a random scatter.

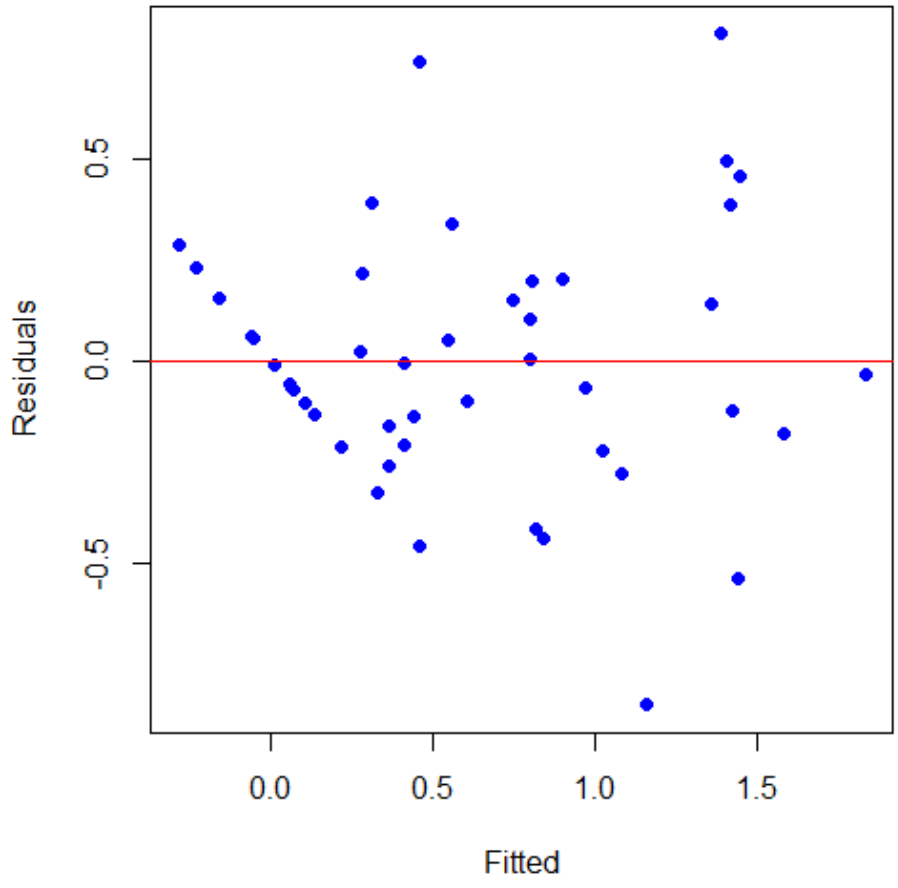

Figure 5.18: *Predicted values vs Standardized Residuals*

The plot of the Predicted values against to the Standardized residuals values, we can see that there is a straight line on the plot. That means the considered data were accept this assumption, Homoscedasticity

Let's examine the influence of individual points by analyzing what happens when they

are excluded. We'll plot the leave-one-out differences for theft and Cook's distances.

According to the below figure where cases 6 and 24 are notable. It's also beneficial to inspect other leave-one-out coefficient plots. Additionally, we should check the jackknife residuals to identify any outliers.

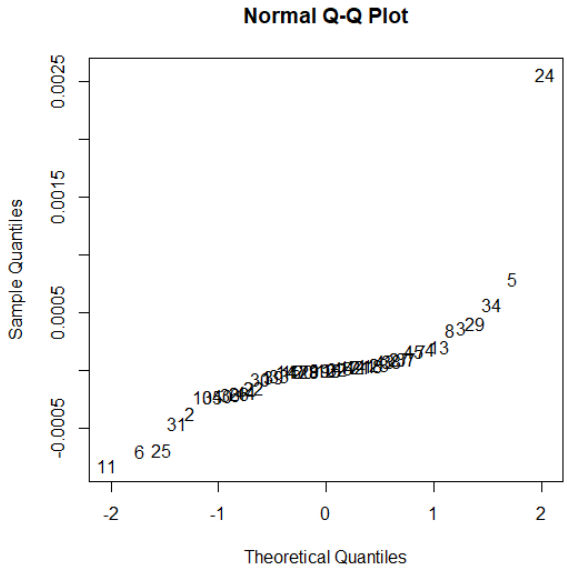

Figure 5.19: *jacknife_residualsplot*

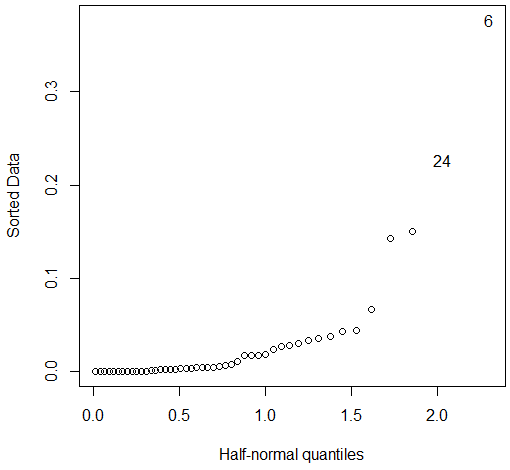

Figure 5.20: *cook_residualsplot*

Cook's distance is a measure used in regression analysis to identify influential data points. It assesses the impact of removing a particular observation on the estimated regression coefficients.

As shown in the cook statistic plot, the two cases 6 and 24 are particularly notable and those two cases can be identified as high theft and fire zip codes in the chredlin dataset and the model that can be obtained by removing those two cases can be shown as follows.

	Estimate	**Std.Erroe**	**t value**	**Pr(value)**
(Intercept)	-0.576737	1.080046	-0.534	0.5964
race	0.007053	0.002696	2.616	0.0126 *
fire	0.049647	0.008570	5.793	1e-06 ***
theft	-0.006434	0.004349	-1.479	0.1471
age	0.005171	0.002895	1.786	0.0818
log(income)	0.115703	0.401113	0.288	0.7745

Table 5.4: *Correlation Coefficient Table with exlude cases 6 and 24*

The analysis of the plots indicated that there was no need to transform the variables. Transforming the race variable would have complicated the interpretation, so it was

avoided. After examined other partial residual plots and experimented with polynomial terms for the predictors, but no transformations appeared beneficial. Also decided against transforming the response variable due to the presence of zeros, which would have restricted transformation options and made interpretation more difficult.

Next, we proceeded to variable selection. Our primary interest was not in selecting a single best model but in understanding the dependency of the response variable on the race variable. The x estimate of the race variable's effect was our main focus. However, collinearity with other variables might cause this estimate to vary substantially depending on the included variables.

To address this issue, we excluded two influential points from the analysis and ensured that the race variable was included in every model considered. This approach allowed us to consistently measure the effect of race, which is the primary predictor of interest in this model.

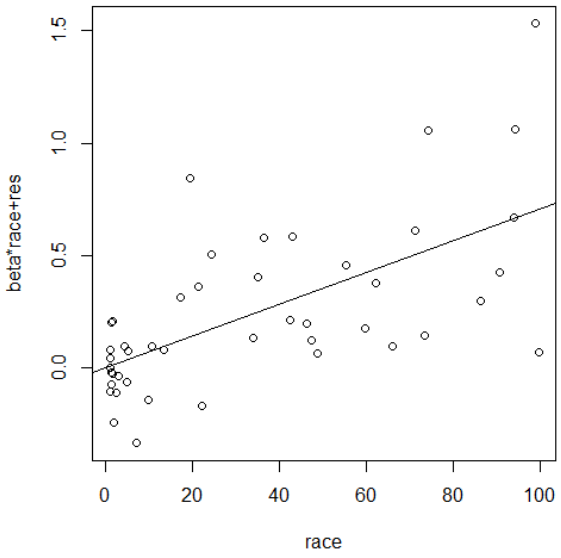

Figure 5.21: *partial residual plot for race*

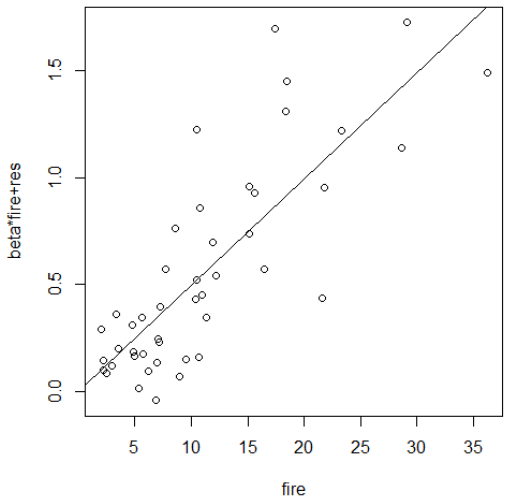

Figure 5.22: *partial residual plot for fire*

	Estimate	**Std.Erroe**	**t value**	**Pr(value)**
(Intercept)	-0.267870	0.139668	-1.918	0.06228
race	0.006489	0.001837	3.532	0.00105 **
fire	0.049057	0.008226	5.963	5.32e-07 ***
theft	-0.005809	0.003728	-1.558	0.12709
age	0.004688	0.002334	2.009	0.05136

Table 5.5: *Correlation Coefficient Table without log(income)*

According to the above table, there are lower significance values for age and theft and

higher significance values for race and fire. Therefore, when choosing our best model, we have to remove the variables of age and theft from the system. And race and fire variables show a positive relationship to Involuct.
By eliminating cases 6 and 4, the resulting model is race, fire, and log(income), because the significance of the race variable is low, so the model without the log(income) variable fits relatively well.

	Estimate	**Std.Erroe**	**t value**	**Pr(value)**
(Intercept)	-0.191325	0.081517	-2.347	0.02371 *
race	0.005712	0.001856	3.078	0.00366 **
fire	0.054664	0.007845	6.968	1.61e-08 *

Table 5.6: *Correlation Coefficient Table for Best Model*

Residual Standard Error (0.31): On average, the actual values of Involact deviate from the predicted values by about 0.31 units.

Multiple R-Squared (0.779): About 77.9 percent of the variance in Involact is explained by the model. This indicates a strong relationship between the independent variables (race and fire) and the dependent variable (Involact).

After the calculations, regression equation of the new model is, :

$$\textbf{Involact = -0.191325 + 0.005712 race + 0.054664 fire}$$

5.4 Discussion and Conclusion

5.4.1 Discussion

The multiple linear regression analysis of the chredlin dataset reveals significant relationships between the percentage of minority population (race), the incidence of fires (fire), and the demand for Involuntary Market Assistance (INVOLACT) policies. However, several factors warrant careful consideration when interpreting these results.

Ecological Correlations

One primary concern is the aggregate nature of the data. The Involact variable is an aggregate measure, not a perfect indicator of individuals being denied insurance. This raises the issue of ecological correlations, where relationships observed at the aggregate level may not accurately reflect individual-level associations. We have implicitly assumed that the probability of a minority homeowner obtaining a FAIR plan is constant across zip codes after adjusting for other covariates. This assumption is unlikely to be entirely true. If the probability varies systematically, our conclusions may be off the mark. To address this, obtaining individual-level data would be beneficial for a more precise analysis.

Effect Size and Practical Implications

While the model coefficients for race and fire are statistically significant, the practical significance of these effects is relatively small. The largest observed value of Involact is only 2.2 percent, and most values are much lower. Even in the worst-case scenario, the number of affected individuals is small, suggesting that while disparities exist, the overall impact on the population is limited.

Potential Latent Variables

There is always the possibility of a latent variable driving the observed relationships. For instance, a variable related to the insurance business that we have not included in the model could be the true cause of the observed patterns. This possibility casts a shadow of doubt on our conclusions. An expert with firsthand knowledge of the insurance industry might propose such a variable, highlighting the need for domain-specific insights.

Data Aggregation

The degree of data aggregation can influence the significance of predictors. By fitting separate models to different subsets of the data (e.g., north and south sides of Chicago), we observe variations in the significance of predictors. For example, race is significant in the north but not in the south. This suggests that the relationship between predictors and Involact may vary across different areas, and aggregating data without considering

regional differences might lead to misleading conclusions..

North vs. South Analysis

	Estimate	Std.Erroe	t value	Pr(value)
(Intercept)	-0.255828	0.159601	-1.603	0.125446
race	0.005322	0.002728	1.951	0.065930 .
fire	0.059969	0.012614	4.754	0.000138 ***

Table 5.7: *Model for the south of Chicago*

On the south side, fire is a highly significant predictor, while race is marginally significant. The model explains approximately 72.63 percentage of the variance in Involact.

	Estimate	Std.Erroe	t value	Pr(value)
(Intercept)	0.049954	0.119353	0.419	0.67961
race	0.015869	0.004845	3.275	0.00346 **
fire	0.008281	0.011900	0.696	0.49377

Table 5.8: *Model for the north of Chicago*

On the north side, race is a significant predictor, while fire is not. The model explains approximately 64.45 percentage of the variance in Involact.

5.4.2 Conclusion

The multiple linear regression analysis highlights significant relationships between the percentage of minority population, fire incidents, and the demand for INVOLACT policies. However, several reservations must be acknowledged:

1. **Ecological Correlations:** Aggregate data may not accurately reflect individual-level associations. Obtaining individual-level data is crucial for more precise conclusions.

2. **Effect Size:** The practical significance of the effects is relatively small, indicating that while disparities exist, the overall impact is limited.

3. **Latent Variables:** Potential unobserved variables could be the true drivers of the observed relationships, introducing uncertainty in our conclusions.

4. **Data Aggregation:** The degree of data aggregation affects the significance of predictors. Regional differences should be considered to avoid misleading conclusions.

The results suggest a need for policy interventions to address insurance disparities, especially in minority communities and high-risk areas. However, the limitations of the study must be carefully considered, and further research with individual-level data is necessary to validate and extend these findings. Presenting these results in a non-technical manner to stakeholders, particularly in adversarial settings, can be challenging, and care must be taken to communicate the inherent uncertainties and subtleties of the analysis.

Bibliography

B. Ansfield. The crisis of insurance and the insuring of the crisis: Riot reinsurance and redlining in the aftermath of the 1960s uprisings. *Journal of American History*, 107(4): 899–921, 2021.

J. J. Faraway. *Linear Models with R.* CHAPMAN HALL/CRC, Boca Raton London NewYork Washington, D.C., 2005.

M. F. Grace, R. W. Klein, and P. R. Kleindorfer. Homeowners insurance with bundled catastrophe coverage. *Journal of Risk and Insurance*, 71(3):351–379, 2004.

D. Li, G. D. Newman, B. Wilson, Y. Zhang, and R. D. Brown. Modeling the relationships between historical redlining, urban heat, and heat-related emergency department visits: an examination of 11 texas cities. *Environment and Planning B: Urban Analytics and City Science*, 49(3):933–952, 2022.

Appendix A

Appendix

A.1 To read the data set

Click here to read the Insurance data set
chredlin

A.2 R code file

The data set can be read as using this R code.
Click here to see the Insurance Data R cose