Determination of Key Factors Influencing Insurance Claim Activity INVOLACT in Chicago

For the Bachelor of Science Honours Degree in
Financial Mathematics and Industrial Statistics

By

A.K.A.N Vidyarathna - SC/2021/12454

Supervisor :
Dr. W.A.R. De Mel
Department of Mathematics
University of Ruhuna
Matara

2024.07.12

Declaration

I, A.K.A.N. Vidyarathna, declare that the presented project report titled "Determinationof key factors Influencing insurance claim activity INVOLACT" is uniquely prepared.by me based on the group project carried out under the supervision of Dr. W.A.R. DeMel, Department of Mathematics, Faculty of Science, University of Ruhuna, as a partialfulfillment of the requirements of the level II Case Study course unit, MIS 2231 of theBachelor of Science Honours Degree in Financial Mathematics and Industrial Statisticsin the Department of Mathematics, Faculty of Science, University of Ruhuna, Sri Lanka.It has not been submitted to any other institution or study program by me for any otherpurpose.

It has not been submitted to any other institution or study program by me for any other purpose.

Signature : ....
Date : ....

Supervisor's Recommendation

I certify that this study was carried out by A.K.A.N Vidyarathna under my supervision

Signature : ....
Date : ....

Department of Mathematics
Faculty of Science
University of Ruhuna

Abstract

In this case study, 47 geographical areas of Chicago were examined for the determinants of insurance claim activity. The factors that were considered were racial composition, the fire and theft rate, housing age, and income level. As for the regression model, the study considers a multiple linear regression technique. The results of the study indicate higher categories for percentages of minority people and theft rates are associated with increases in INVOLACT. These disparities underline very important socio-economic inequality, confirming on the possible continuous practice of discrimination in the insurance sector. The study provides relevant insides to policyholders and insurance companies to strategize for insurance service provision that is accessible on equal terms. This concise overview helps potential readers grasp the research problem, justification, methodology, and key finding together with their importance hence allowing the reader to make a decision of the relevance of the full article to their interests.

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

Acknowledgement

In preparation for our case study, I had to seek the help and guidance of some respected individuals who deserve our deepest gratitude. First of all, my 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 are especially thankful to our course coordinator, Dr. A.W.L. Pubudu Thilan, who gave us many more valuable instructions to fulfill our case study successfully, and Ms. R.M.V. Lakmini for her guidance in the regression analysis part. As well, I would like to thank all the supervisors on the supervisory board for their great support of 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-II Course Unit. Finally, I am thankful to everyone, especially my group members, who played a part, big or small, in making this research a success.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the study

Insurance redlining is a discriminatory practice where insurance companies deny coverage, charge higher premiums, or offer limited coverage based on factors such as the racial composition, socioeconomic status, or geographic location of the insured individuals or properties. In a study of insurance availability in Chicago, the U.S. Commission on Civil Rights attempted to examine charges by several community organizations that insurance companies were redlining their neighborhoods.
ie. canceling policies or refusing to insure or renew.

There are several factors that affect the redlining of insurance, such as:

- Demographic Factors: Race, Ethnicity, Income Level, Education Level
- Geographic Location
- Historical Factors
- Risk Assessment Regulatory Environment
- Industry Practices
- Public Perception and Bias

Using that data, in our research, we basically consider race, fire rate, theft rate, age of housing, INVOLACT, income, and side as the factors that affect the insurance redlining. For insurance, many states have a "FAIR" plan available for (and limited to) those who cannot obtain insurance in the regular market. So an area with a high number of FAIR plan policies is an area where it is hard to get insurance in the regular market. We will use the name “INVOLACT” for new FAIR plan policy updates. As per the per the above description, we can say that insurance redlining is limiting the services that insurance companies provide to policyholders by considering those factors. Therefore, we have to reduce those limitations and expose their facilities to more people. We will be mainly focusing on this factor and how significantly it will affect the insurance claim activity.

1.2 Problem Statement

As graduates studying actuarial science, it is good that we know about the insurance redlining. Although the fair housing act officially outlawed the practice of redlining in Chicago, despite being illegal, redlining and discriminatory housing practices still persist to this day, resulting in the denial or restriction of insurance services to residents or businesses in those areas. Based on those facts, this case study investigates how the factors of race, fire, theft, age, income and side affect the INVOLACT. This is a result of variation of many factors that affect the INVOLACT. Therefore, there is a need to examine and analyze the relationship between these variables and INVOLACT to limit the redlining process.

1.3 Research Questions

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

1.4 Objectives

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. In addition to that, studying the individual contributions of these variables to INVOLACT, identifying whether they have positive or negative impact on INVOLACT, providing empirical evidences and insights to formulate effective strategies can also be taken as the objectives of this study. We are planning to achieve our goal using the multiple regression model and then the 'Backward Elimination' method to select the best regression equation. The mentioned regression model will be taking using the R software.

1.5 Hypothesis

Hypothesis 1

- Null hypothesis (H_0) : There is no linear relationship between the dependent variable and independent variables in Chicago. $(B_i = 0)$

- Alternative hypothesis (H_1) : There is a linear relationship between the dependent variable and independent variables in Chicago. $(B_i \neq 0)$

1.6 Significance of the Study

This study has significant practical applications across various domains. Firstly, it enables policymakers to design targeted interventions that address disparities, fostering inclusivity in crucial areas such as education and workforce participation.

Also, organizations can refine their practices by implementing informed diversity and inclusion strategies, leading to more equitable environments and opportunities for all members. Additionally, communities can benefit from insights into how demographic factors shape involvement, guiding initiatives that promote cohesion and ensure equitable access to resources and opportunities.

Finally, research in this area contributes to advancing knowledge and understanding societal dynamics, informing future studies and policy discussions on social justice and equity. Ultimately, by addressing disparities and promoting inclusivity, this research contributes to building more equitable and cohesive societies.

Chapter 2

Literature Review

Insurance redlining refers to the practice of denying or limiting insurance coverage to certain geographical areas, often based on the racial or ethnic composition of the neighborhood. Several factors contribute to variations in INVOLACT, including socio-economic variables, demographic characteristics, and environmental risks. The FAIR plan is designed to provide insurance to those unable to secure it in the voluntary market, often targeting high-risk areas. This literature review aims to explore existing research on the factors influencing insurance claim activity, with a focus on race, fire incidents, theft rates, and income levels.

The racial composition of a neighborhood has been consistently identified as a critical factor in insurance redlining. Historically, minority neighborhoods have faced discrimination through practices such as redlining, where these areas were marked as high-risk and subsequently denied insurance coverage or offered it at higher premiums. This legacy continues to influence present-day insurance practices, leading to greater reliance on involuntary market programs in predominantly minority communities. Research by Squires [1997] highlights how minority-dominated neighborhoods often face higher insurance premiums or outright denial of coverage, irrespective of the actual risk levels. The article Squires [2003] further states 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 play a significant role in influencing insurance claim activity, particularly in areas prone to wildfires or urban blight. 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. Moreover, the severity of fire damage can impact insurance affordability and availability, further affecting INVOLACT in affected communities. High rates of theft and property crime contribute to increased insurance claim activity, reflecting the need for coverage against losses due to burglary, vandalism, and other criminal activities. Fire and theft rates are also critical factors affecting insurance policies. Higher incidents of fires and thefts in a neighborhood typically lead to increased insurance costs and policy cancellations. Insurance companies often justify redlining by citing higher risks associated with these factors. In Chicago, data from the 1970s show that neighborhoods with more fires and thefts faced more challenges in obtaining insurance, further marginalizing these communities.

The age of the houses is also a crucial factor that affects the insurance redlining. The article Klein and Grace [2001] explains that Community and consumer advocates argue that insurers' underwriting practices and pricing systems unfairly discriminate against neighborhoods with older and lower-value housing, leading to higher rates and reduced availability of coverage.

Income levels serve as a crucial determinant of insurance claim activity. Bostic and Surette [2001] examined how lower median incomes in minority neighborhoods often lead to higher premiums and less favorable terms. 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. Conversely, affluent communities may exhibit lower INVOLACT due to greater investment in risk mitigation measures and higher insurance coverage limits.

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. The use of quantitative methods allows researchers to collect and analyze numerical data related. This approach identify the relationships between new FAIR plan policies (INVOLACT) 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. The mentioned regression model will be taking using the R software.

3.1.1 Justifying the research approach

Objective Measurement and Analysis : Quantitative research accurately measures variables like race, fire rate, theft, age, income, and side using numerical data. This method eliminates biases and ensures findings are based on empirical evidence, eliminating subjective interpretations.

Statistical Validation : Multiple regression models and the 'Backward Elimination' method in R software, validate findings by identifying statistically significant relationships between variables and INVOLACT.

Generalizability: The aim of the study is likely to draw generalizable conclusions about the relationships between the dependent and independent variables. Larger sample sizes enhances generalizability, allowing for broader conclusions about factors influencing engagement levels across a wider population, making insights more applicable for effective strategies.

Identification of Patterns and Trends: Multiple regression analysis, can identify patterns and trends in data, revealing complex interactions between variables like income and housing age, which is crucial for understanding INVOLACT's multifaceted nature.

Precision and Control: Quantitative research offers precise measurement and control over variables through structured data collection instruments and statistical controls, ensuring accurate and directly attributable findings to the variables of interest.

Empirical Evidence: Quantitative research uses numerical data to provide empirical evidence, supporting conclusions and recommendations, making them more convincing and actionable for policymakers and stakeholders.

Efficiency of data: Using statistical software like R, is efficient and reproducible, ensuring consistency in findings and verification, thereby enhancing the robustness and credibility of a study.

Backward Elimination Method: The backward elimination method is a systematic method used in quantitative research to select models by iteratively removing the least significant variables, ensuring the final model only includes factors impacting INVOLACT, aligning with data-driven decision-making.

3.2 Conceptual Model

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

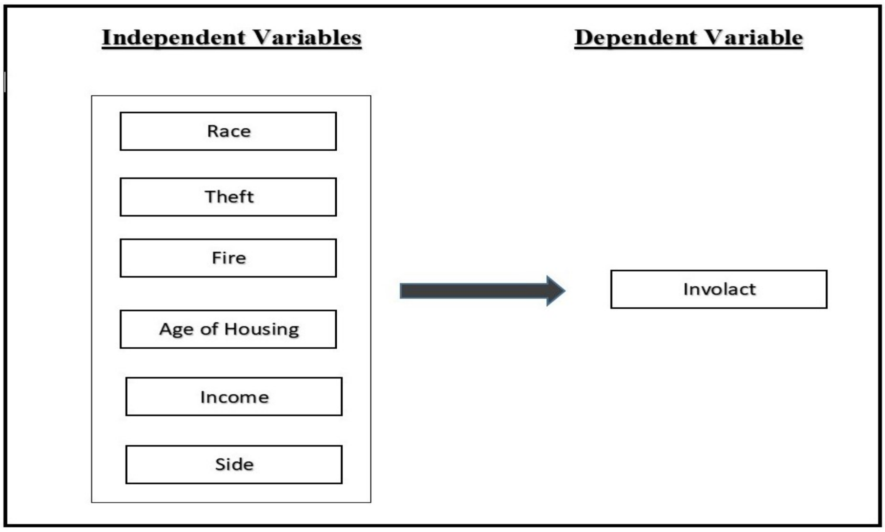


Figure 3.1: *Conceptual Model*

3.3 Research Design

Under the quantitative approach, correlational research is used as the research design for this study. 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 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, and Age of housing). So, we use multiple linear regression and approach it using the basic concepts of simple linear regression and ordinary least squares method.

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. The purpose of regression analysis is to determine the existence, the kind, or the extend of a relationship (in the form of a mathematical equation) between variables. Then we can use that formula to predict values for the dependent variable when only the independent variables are known.

- **Simple Linear Regression :** Simple linear regression is a regression model that estimates the relationship between one independent variable and dependent variable. If we let y represent dependent variable and x represent independent variable, then the equation of a straight line relating these two variables is:

  $$\mathbf{y} = \beta_0 + \beta_1 x$$

  where β_0 is the intercept and β_1 is the slope. Now the data points do not fall exactly on a straight line, so the above equation should be modified to account for this. Let the difference between the observed value of y and the straight line ($\beta_0 + \beta_1 x$) be an error ε.

  $$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

  Above equation is called as linear regression model.

- **Least Square Method :** Then we can use the functions below for finding least square line.

  $$\hat{\beta_1} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.2}$$

  $$\hat{\beta_0} = \bar{y} + \hat{\beta_0}\ \bar{x} \tag{3.3}$$

3.3.1 Multiple Linear Regression

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.

There are four assumptions in multiple linear regression as Normality of errors, Linearity, Multicolinearity and Homoscedasticity. First of all we should find they are satisfying or not. By checking for these assumptions and addressing any violations, we can ensure that our multiple linear regression model produces reliable results, valid inferences, and generalizable insights.

In general, the response y may be related to k regressor or predictor variables.
The model :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{3.4}$$

is called a multiple linear regression model with k regressors.

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

$\beta_k x_k$ = the regression coefficient of the last independent variable

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

A matrix formulation of the multiple regression model

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses. Here, we review basic matrix algebra, as well as learn some of the more important multiple regression formulas in matrix form.
As always, let's start with the simple case first. Consider the following simple linear regression function:

$$\mathrm{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i = 1, \ldots, n$$

If we actually let $i = 1, \ldots, n$, we see that we obtain n equations:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned} \tag{3.5}$$

Well, that's a pretty inefficient way of writing it all out! As you can see, there is a pattern that emerges. By taking advantage of this pattern, we can instead formulate the above simple linear regression function in matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{1\mathrm{n}} \\ 1 & x_{21} & ... & x_{2\mathrm{n}} \\ \vdots & \vdots & & \\ 1 & x_{\mathrm{n}1} & ... & x_{\mathrm{nn}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{3.6}$$

That is, instead of writing out the n equations, using matrix notation, our simple linear regression function reduces to a short and simple statement:

$$Y = X\beta + \epsilon \tag{3.7}$$

where:

- Y is the $n \times 1$ vector of observed values.
- X is the $n \times (p+1)$ matrix of predictors (including a column of ones for the intercept).
- β is the $(p+1) \times 1$ vector of regression coefficients.
- ϵ is the $n \times 1$ vector of errors.

To find the least squares estimator for β, we solve the normal equations:

$\mathrm{X}^T X\beta = X^T Y$

The solution to this equation is:

$$\beta = (X^T X)^{-1} X^T Y \tag{3.8}$$

This formula is used to calculate the estimated coefficients in a multiple linear regression model. Then, the best regression equation should be selected related to the equation of this linear combination. This selection process aims to reduce the set of predictor variables to find the most important predictor variable or variables while maintaining a good explanation of the data.

There are mainly three methods for this process,

1. All possible regressions
   All possible regressions goes beyond stepwise regression and literally tests all possible subsets of the set of potential independent variables.

2. Forward selection method
   The forward selection approach starts with no variables and adds each new variable incrementally, testing for statistical significance.

3. Backward elimination
   The backward elimination method begins with a full model and then removes the least statistically significant variables one at a time.

Among these methods, this is study is done by using the backward elimination method because it aligns with the objective of studying the individual contributions of variables to INVOLACT. It help simplifying the model and also this method ensures than all remaining predictor variables in the best regression equation are statistically significant and draw meaningful conclusions on this study.

In addition to these, by selecting forward selection method or all possible regression may increase the risk of over fitting. Therefore, it can affect to occur the errors.

Chapter 4

Data

4.1 Data Set

This dataset is about the Race, Fire rate, Theft rate, Age of housing, INVOLACT, Income and Side of the Chicago. The R code used to read this dataset 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 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.

Table 4.1: *Independent and Dependent Variables*

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

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.

Table 4.2: *Data Dictionary Table*

Variable	**Abbreviation**	**Measurement**	**Variable** Type	**Missing Data Indicators**
New FAIR plan policies	INVOLACT	Percentage	Numeric	N/A
Minority percentage	Race	Percentage	Numeric	N/A
Fire rate	Fire	Percentage	Numeric	N/A
Theft rate	Theft	Theft per 1000 population	Numeric Numeric	N/A
Age of housing	Age	Percentage	Numeric	N/A
Median family income	Income	Median family income in $1000	Numeric Numeric	N/A
Side	Side	n:North , s:South	Categorical	N/A

4.4 Preparation for analysis

4.4.1 Missing values and Null values

Missing data is important because, depending on the type, it can sometimes bias our results. This means our results may not be generalizable outside of the study because the data come from an unrepresentative sample.

But the following R output shows that our data set has no missing values and Null values.

4.4.2 Assigning Numerical Values for Categorical Variables

Standard linear regression models require numerical inputs for both the dependent and independent variables. Categorical variables, on the other hand, represent distinct groups or qualities. An example, we have one categorical variable as ‘side’. They don’t inherently have numerical values. To incorporate categorical variables into regression models, we need to convert them into a format suitable for numerical analysis. This process is called encoding. We use dummy coding method for encoding.

Dummy Coding : This is the most common technique. It creates a new binary variable (0 or 1) for each category of the original categorical variable. In here categorical variable with two levels (e.g., north=n, south=s) would be converted into one dummy variables. One variable would represent ‘n’ (0 for n) and the other would represent ‘s’ (1 for s).

The R code used to assigning numerical values is in the appendix.

We will not use information on the south side against the north until later. In summary section, according to the reduced model we will find North or South will be significantly affected or not.

4.4.3 Some Adjustment of variables

The question of which variables should be included in a regression to adjust their effect is challenging. Statistically, it can be done, but the important question is whether it should be done at all. For example, adjusting for factors like job type and length of service can reduce or even disappear the gender difference in income. If the effect of adjusting for income differences were to remove the race effect, this would pose a non-statistical question. Log(income) is used due to skewness and because income is better considered on a multiplicative scale.In other words, \$ 1,000 is worth a lot more to a poor person than a millionaire because \$ 1,000 is a much greater fraction of the poor person’s wealth.
Then our full model like,

4.4.4 Checking Assumptions

Before leaping to any conclusions, we should check the model assumptions. Understanding and verifying these assumptions is necessary for accurate model interpretation and prediction.

1. **Normality of errors**
   The residual values are normally distributed. This can be checked by either using a normal probability plot or a histogram. The figure 4.1 and figure 4.2 shows that the residual values are normally distributed.

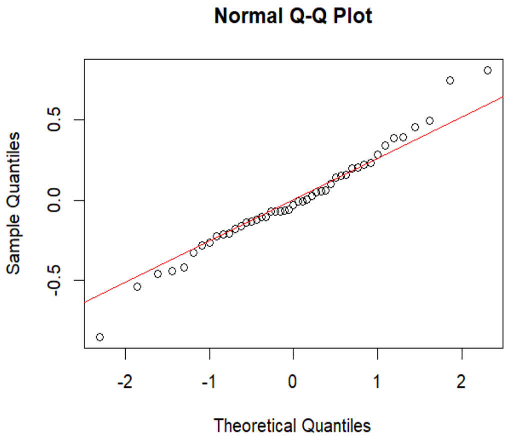


Figure 4.1: *Q-Q Plot with outliers*

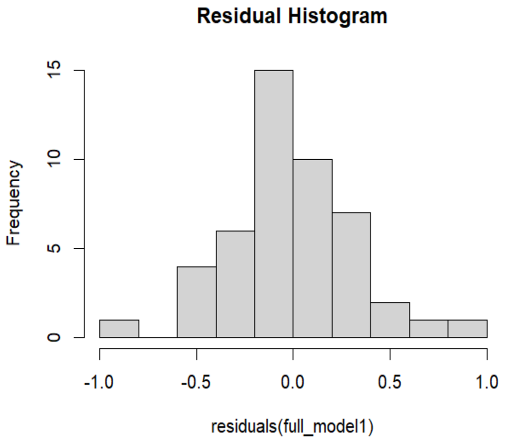


Figure 4.2: *Residual Histogram*

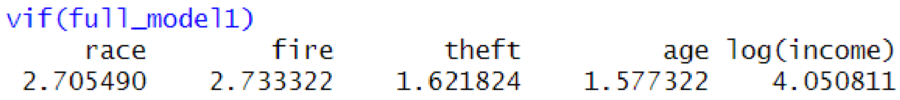
2. **Linearity**

   There must be a linear relationship between the dependent and the independent variables. This can be illustrated by scatterplots showing a linear or curvilinear relationship. The figure 5.3 in exploratory data analysis section indecates that there is a linear relationship between the dependent and the independent variables.

3. **Multicolinearity**

   Multicollinearity is another assumption, meaning that the independent variables are not highly correlated with each other. Multicollinearity makes it difficult to identify which variables better explain the dependent variable. This assumption is verified by computing a matrix of Pearson's bivariate correlations among all the independent variables. If there is no collinearity in the data, then all the values should be less than 0.8. The figure 5.4 shows that the independent variables are not highly correlated with each other.

   Another method to detect multicollinearity is to calculate the variance inflation factor (VIF) for each independent variable, and we can get all of VIF values are less than 5. It can be says that variables are moderately correlated.



```
vif(full_model1)
     race      fire     theft       age log(income)
 2.705490  2.733322  1.621824  1.577322    4.050811
```

Figure 4.3: *VIF Values*

4. **Homoscedasticity**

   The homoscedasticity assumes that the variance of the residual errors is similar across the value of each independent variable. One way of checking that is through a plot of the predicted values against the standardized residual values to see if the points are equally distributed across all the values of the independent variables. In Figure 4.4 we can see points are equally distributed.

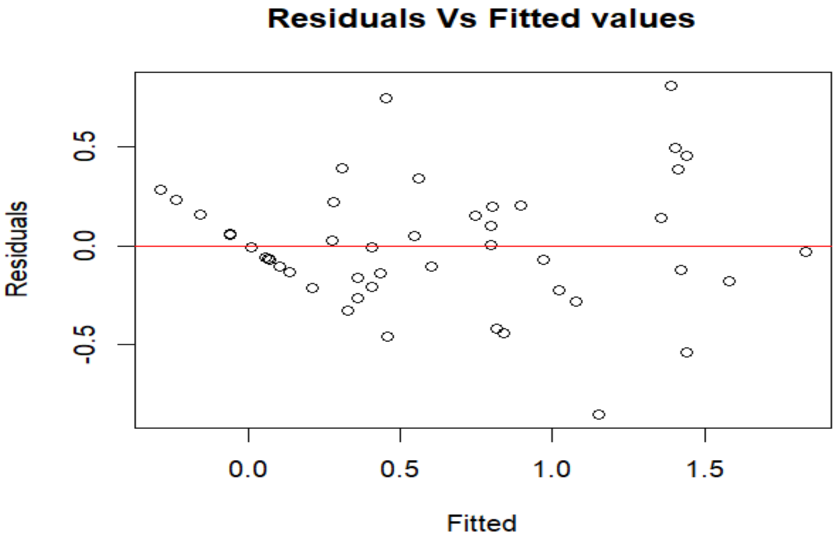


Figure 4.4: *Plot of Predicted Values vs. Standardized Residuals*

4.4.5 Outliers

Now let's look at influence, what happens if points are excluded? We plot the leaveout-one difference in for Cook distances. If the data contains large residuals (outliers) and/or high leverage, these can distort the result of a regression and its accuracy. Cook's distance gives a measure of the effect of deleting a given observation. Large Cook's distances are generally considered to be worth closer examination in the analysis.

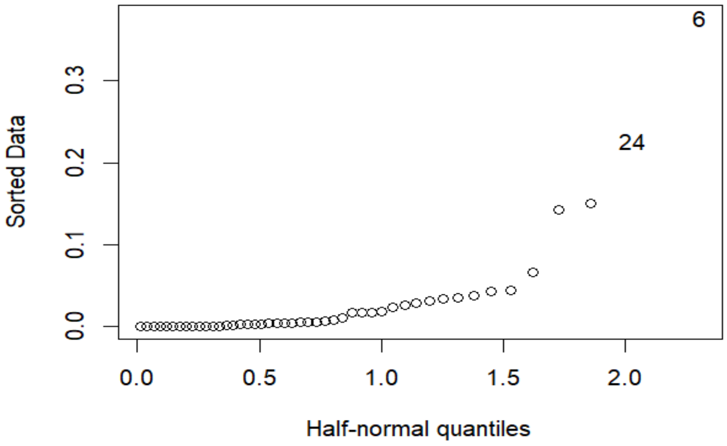
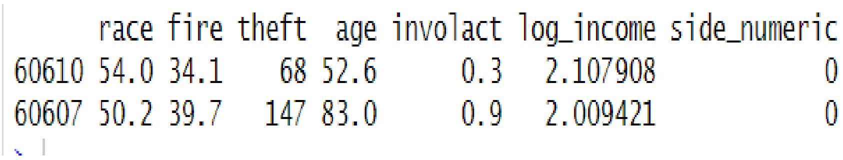


Figure 4.5: *A half-normal plot of the Cook distances*

See Figure 4.5 where cases 6 and 24 stick out. It is worth looking at other leaveout one coefficient plots also. We check the jackknife residuals for outliers. Jackknife residuals are most sensitive to outlier detection and are superior in terms of revealing other problems with the data. For that reason, most diagnostics rely upon the use of jackknife residuals. Here we get jackknife residual range as -3.184960 and 2.792884. So, the smallest studentized residual is -3.184960 and the largest is 2.792884.

Let's take a look at the two cases in 6 row and 24 row. The 6 row represent the ZIP code No:60610 and 24 represent ZIP code No:60607. These are high theft and fire zip codes.



```
      race fire theft  age involact log_income side_numeric
60610 54.0 34.1    68 52.6      0.3   2.107908            0
60607 50.2 39.7   147 83.0      0.9   2.009421            0
```

Figure 4.6: *Outlier include rows*

Since what we are doing here is to get an idea of the data to take legal action against redlining, I decided to remove this data as an outlier and do further calculations to get the correct information in the data.

4.4.6 Transformation Variables

We now look for transformations. We try some partial residual plots as seen in figure 4.7 :

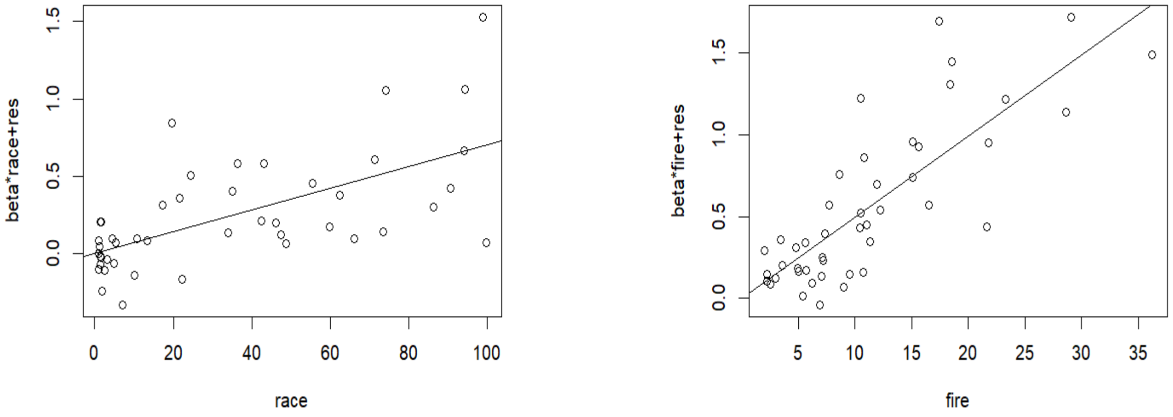


Figure 4.7: *Partial residualplots for race and fire.*

These plots indicate no need to transform. The race variable transformation would have been challenging to interpret, but other partial residual plots and polynomial predictor transformations were not worthwhile. The response transformation due to the zeros in the response which would have limited possibilities and made interpretation more difficult. The decision to avoid such transformations is based on the lack of worthwhile alternatives.

Chapter 5

Results

5.1 Exploratory Data Analysis

Table 5.1: *Descriptive Statistics Table*

Variable	Minimum	Q1	Median	Mean	Q3	Maximum	Total Observation
Race	1.00	3.75	24.5	34.99	57.65	99.70	45
Fire	2.00	5.65	10.40	12.28	16.05	39.70	45
Theft	3.00	22.00	29.00	32.36	38.00	147.00	45
Age	2.00	48.60	65.00	60.33	77.3	90.10	45
Involact	0.0000	0.0000	0.4000	0.6149	0.9000	2.2000	45
Log Income	1.720	2.134	2.370	2.339	2.484	3.067	45
Numeric Side	0.0000	0.0000	0.0000	0.4681	1.0000	1.0000	45

The above 5.1 Table illustrates the summary statistics of each variable. According to this table, the total count of observations is 45 and there are no missing values and duplicate values. We see that there is a wide range in the race variable, with some zip codes almost entirely minority or non-minority. This is good for our analysis since it will reduce the variation in the regression coefficient for race, allowing us to assess this effect more accurately. If all the zip codes were homogeneous, we would never be able to discover an effect from these aggregated data. We also note some skewness in the theft and income variables. We got zero value for INVOLACT (response variable). The number of new FAIR Plan Policy renewals per 100 housing units may be zero.

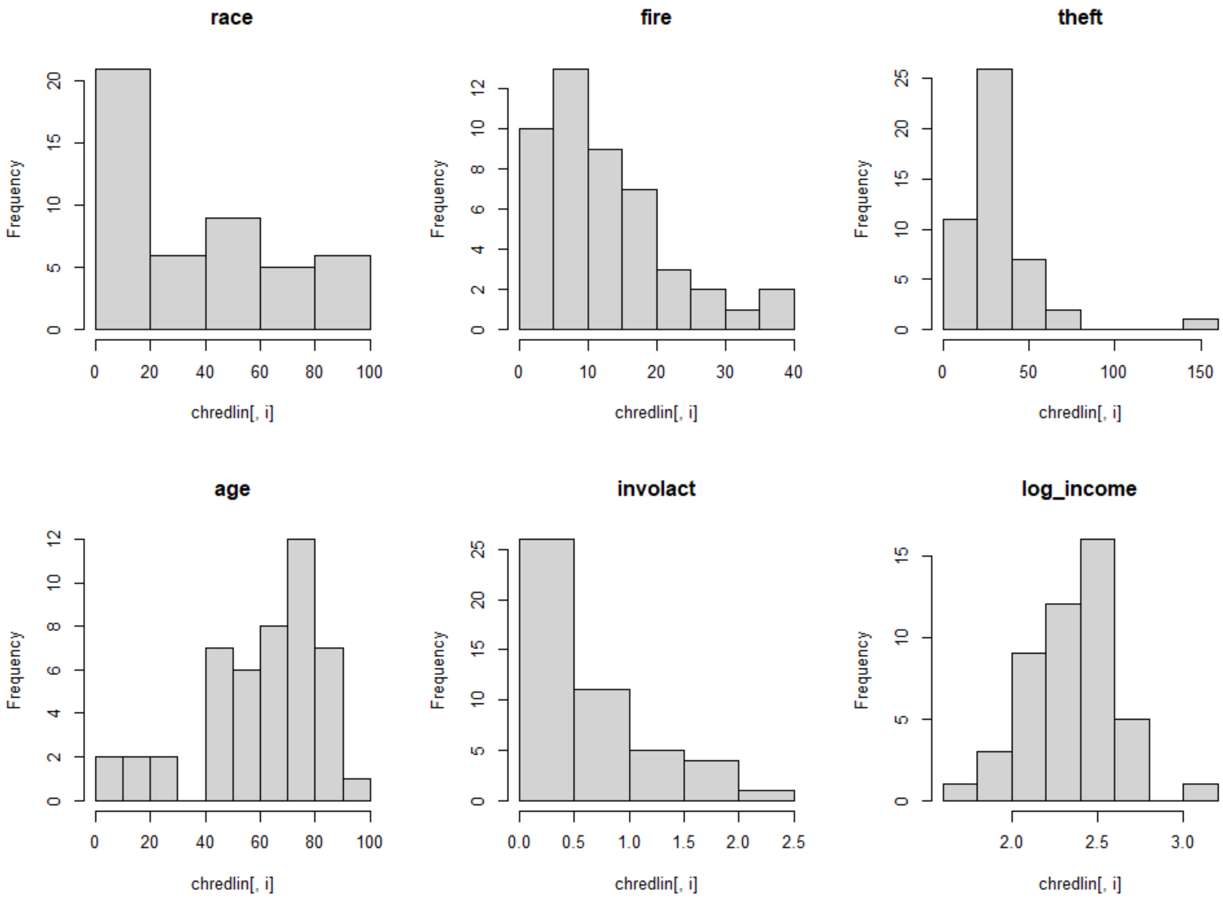


Figure 5.1: *Histograms of All Variables*

In the above figure 5.1 is shown that histogram of all variables. We see that there is a wide range in the race variable. INVOLACT, fire and theft values are positively skewed because the majority of the data are on the left side of the mean. The figure shows that the Age values are negatively skewed because the majority of data points are in the right side of the mean. We can see that the log-income values like bell shape.

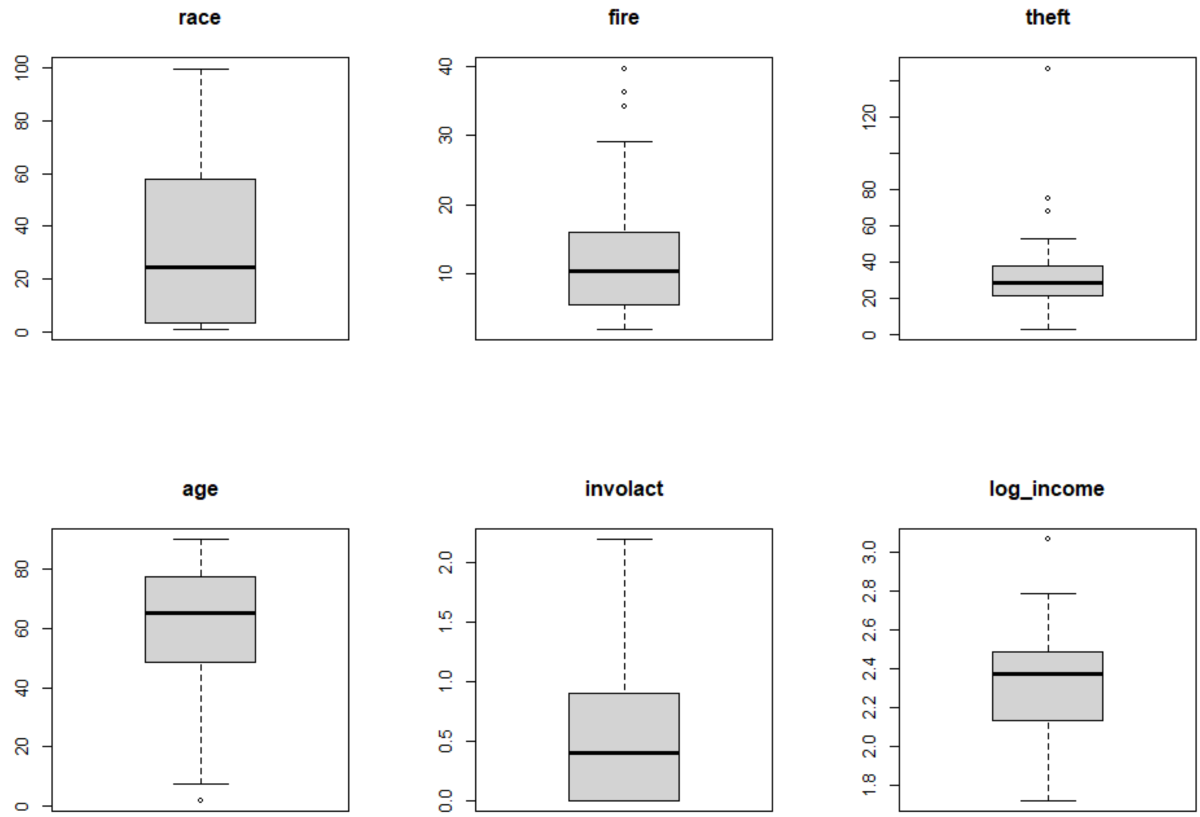


Figure 5.2: *Boxplot of all variables*

The boxplot drawn for each variable is shown in the above figure 5.2. When taken separately, outliers are shown in the variables fire, theft, age, and log-income. Fire rates, theft rates and income may be higher. Also, the age of housing may decrease due to new housing construction. The values we see as outliers may not actually be so. It is best to remove outliers only when we have a good reason to do so. Some outliers represent natural variations in the population and should be left as they are in our data set. These are called true outliers. That's why we used cook-distance to remove outliers.

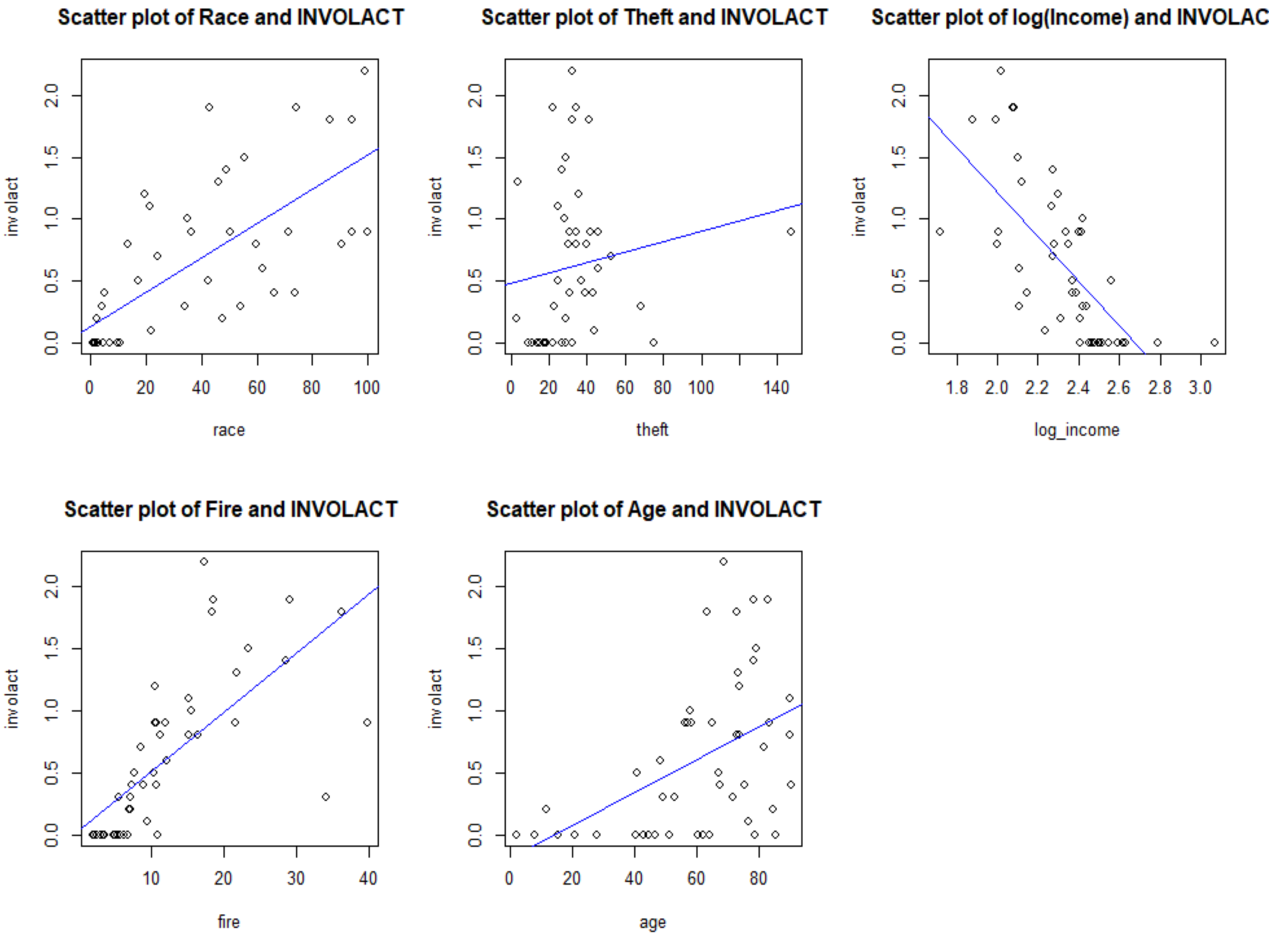


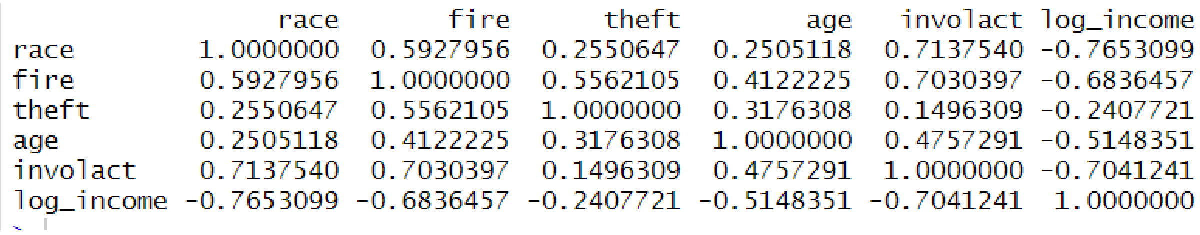
Figure 5.3: *Scatter Plots with all variables*

Figure 5.3 shows the relationship between other variables and INVOLACT. We can show that there is a negative relationship between INVOLACT and log-income. All of other variables have positive relationship with INVOLACT.

5.2 Quantitative Analysis

As mentioned earlier, there are no missing values and null values in this dataset. In here we use the model with outlier for this analysis.

5.2.1 Correlation Analysis



```
                 race       fire      theft        age   involact log_income
race        1.0000000  0.5927956  0.2550647  0.2505118  0.7137540 -0.7653099
fire        0.5927956  1.0000000  0.5562105  0.4122225  0.7030397 -0.6836457
theft       0.2550647  0.5562105  1.0000000  0.3176308  0.1496309 -0.2407721
age         0.2505118  0.4122225  0.3176308  1.0000000  0.4757291 -0.5148351
involact    0.7137540  0.7030397  0.1496309  0.4757291  1.0000000 -0.7041241
log_income -0.7653099 -0.6836457 -0.2407721 -0.5148351 -0.7041241  1.0000000
```

Figure 5.4: *Correlation Matrix*

1. **Question 01 -** Relationship between race and INVOLACT :
   The correlation coefficient of 0.7137540 between racial composition in percentage of minority and new FAIR plan policies and renewals per 100 housing units (INVOLACT) implies that the relationship between these two variables is high positive. This suggests that when the racial composition in percentage of minority increasing, the number of FAIR plan policies to increase.

2. **Question 02 -** Relationship between fire and INVOLACT:
   The correlation coefficient of 0.7030397 between fire rate and new FAIR plan policies and renewals per 100 housing units (INVOLACT) implies that the relationship between these two variables is high positive. This suggests that when the fire rates increasing, the number of FAIR plan policies to increase.

3. **Question 03 -** Relationship between theft and INVOLACT:
   The correlation coefficient of 0.1496309 between theft rate and new FAIR plan policies and renewals per 100 housing units (INVOLACT) implies that the relationship between these two variables is weekly positive. This suggests that when the theft rates increasing, the number of FAIR plan policies to increase.

4. **Question 04 -** Relationship between age and INVOLACT:
   The correlation coefficient of 0.4757291 between age of housing and new FAIR plan policies and renewals per 100 housing units (INVOLACT) implies that the relationship between these two variables is moderate positive. This suggests that when the age of housing increasing, the number of FAIR plan policies to increase.

5. **Question 05 -** Relationship between income and INVOLACT:
   The correlation coefficient of -0.7041241 between log-income and INVOLACT. So we can say that median family income and new FAIR plan policies and renewals per 100 housing units (INVOLACT) implies that the relationship between these two variables is moderate moderately positive. This suggests that when the median family income increasing, the number of FAIR plan policies to decreasing.

5.2.2 Estimate Model Parameters

In this section, relationship between the dependent and independent variables are mathematical represented. Then, which independent variables are to be included is chosen to assess the model's fit. When fitting a multiple linear regression model, a researcher will likely include independent variables that are not important in predicting the dependent variable y. In the analysis we will try to eliminate these variable from the final equation. The objective in trying to find the "Best Equation" will be to find the simplest model that adequately fits the data. Our objective will be to find the equation with the least number of variables that still explain a percentage of variance in the dependent variable that is comparable to the percentage explained with all the variables in the equation.
As mentioned in multiple linear regression in research design part, we will use the backward elimination method for this.

Backward Elemination Method

The backward elimination method begins with a full model and then removes the least statistically significant variables one at a time. Here we use null hypothesis and alternative hypothesis and find out whether the null hypothesis is accepted or not.

The test statistic we use is the F partial value.

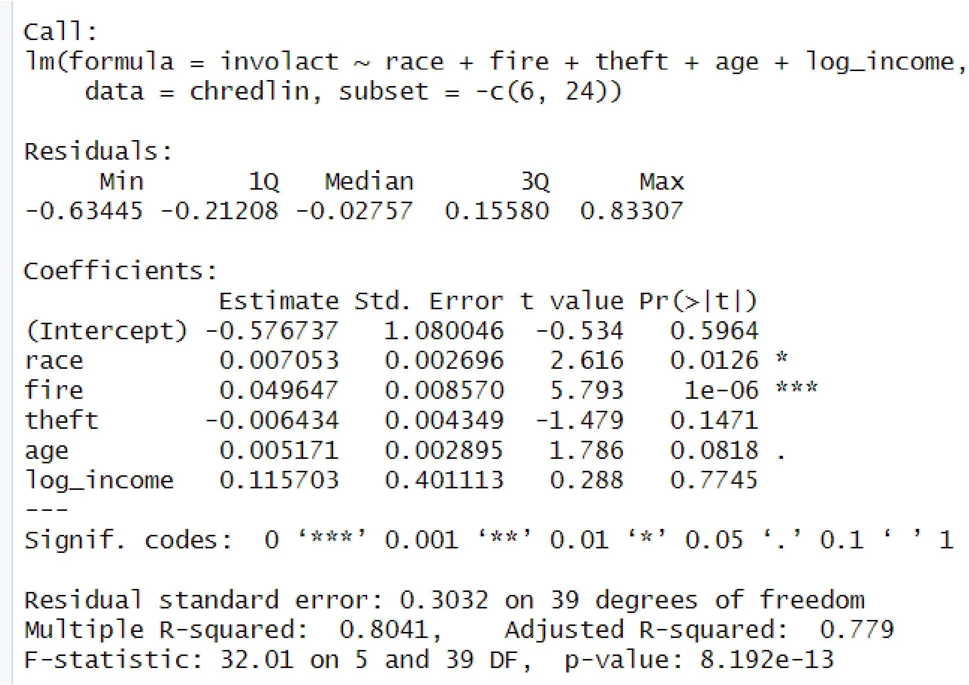
$$\textbf{Test Statistic} = \mathbf{F}_{\textbf{partial}} = \frac{[SS_{\textbf{Reg(Full)}} - SS_{\textbf{Reg(Reduced)}}]}{MS_{\textbf{Error(Full)}}}$$

In this our hypothesis are,

Null hypothesis (H_0) : Reduced model is suitable.
Alternative hypothesis (H_1) : Full model is needed.

Full Model :

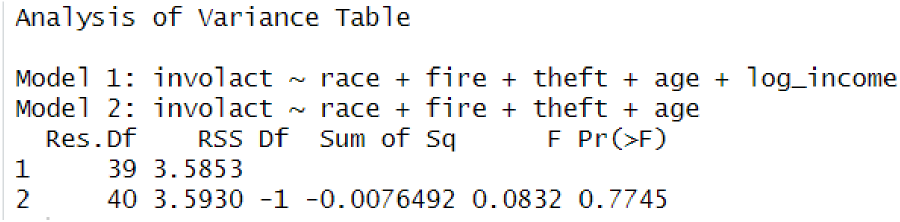
First we get linear model with all variables. Then we summarize the full model and find the variable with the highest p value.



```
Call:
lm(formula = involact ~ race + fire + theft + age + log_income,
    data = chredlin, subset = -c(6, 24))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63445 -0.21208 -0.02757  0.15580  0.83307 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(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    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3032 on 39 degrees of freedom
Multiple R-squared:  0.8041,	Adjusted R-squared:  0.779 
F-statistic: 32.01 on 5 and 39 DF,  p-value: 8.192e-13
```

Figure 5.5: *Summary Full Model*

Figure 5.5 illustrates that in full model R^2 is 80.41% and "log income" has highest p $= 0.7745 \succ 0.05$ value. Then we reduced that variable and get new model call as reduce model 1. After we use Nested ANOVA method and consider probability of F partial value is greater than or not 95% Confidence Interval.



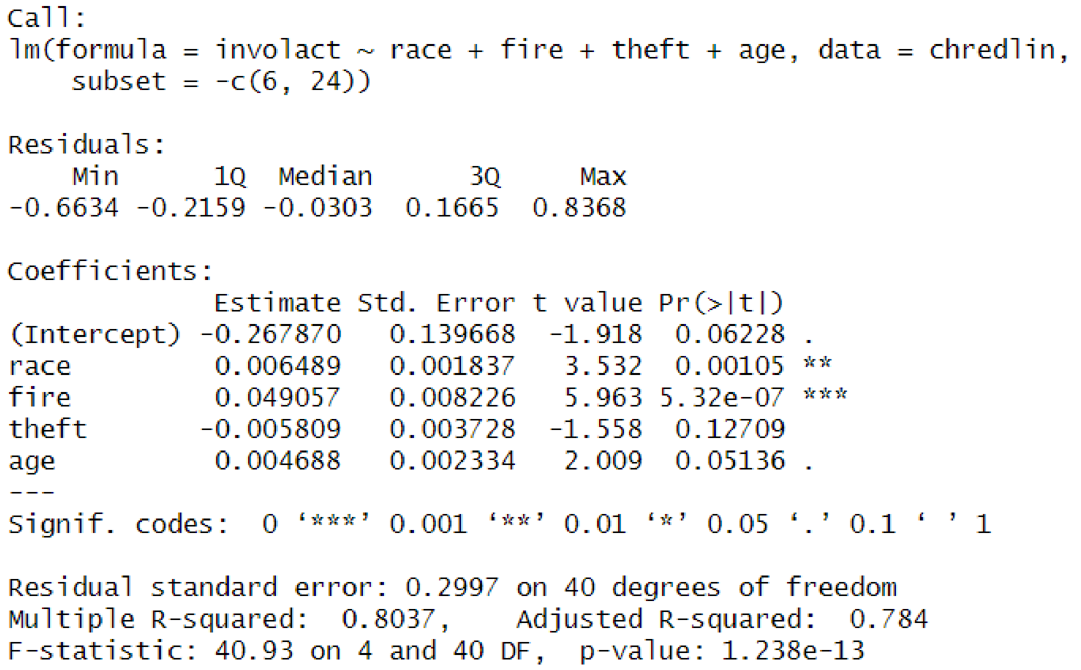
```
Analysis of Variance Table

Model 1: involact ~ race + fire + theft + age + log_income
Model 2: involact ~ race + fire + theft + age
  Res.Df    RSS Df  Sum of Sq      F Pr(>F)
1     39 3.5853                            
2     40 3.5930 -1 -0.0076492 0.0832 0.7745
```

Figure 5.6: *Nested ANOVA table for full and reduced model 1*

The figure 5.6 shows that F partial is 0.0832 and p value is $0.7745 \succ 0.05$. This implies that the p-value is not in the rejection region. So, there is not much evidence to reject the null hypothesis. So, we can say that the reduced model 1 is suitable.

Reduced model 1:

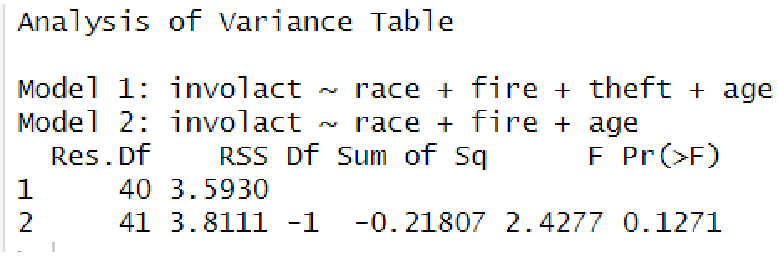
Secondly we get summary of reduced model 1 and find the variable with the highest p value.



```
Call:
lm(formula = involact ~ race + fire + theft + age, data = chredlin,
    subset = -c(6, 24))

Residuals:
    Min      1Q  Median      3Q     Max
-0.6634 -0.2159 -0.0303  0.1665  0.8368

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(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 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2997 on 40 degrees of freedom
Multiple R-squared:  0.8037,	Adjusted R-squared:  0.784
F-statistic: 40.93 on 4 and 40 DF,  p-value: 1.238e-13
```

Figure 5.7: *Summary Reduced Model 1*

Figure 5.7 illustrates that in reduced model 1 R^2 is 80.37% and “Theft” has highest p $= 0.12709 \succ 0.05$ value. Then we reduced that variable and get new model call as reduce model 2. After we use Nested ANOVA method and consider probability of F partial value is greater than or not 95% Confidence Interval.



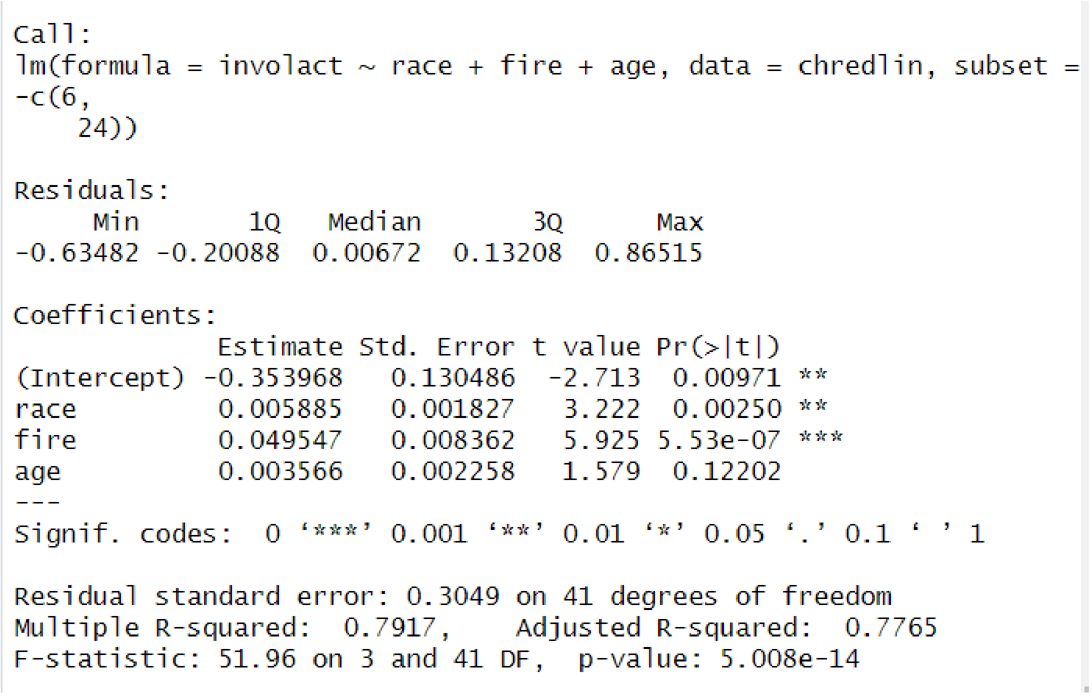
```
Analysis of Variance Table

Model 1: involact ~ race + fire + theft + age
Model 2: involact ~ race + fire + age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     40 3.5930
2     41 3.8111 -1  -0.21807 2.4277 0.1271
```

Figure 5.8: *Nested ANOVA table for reduce model 1 and reduced model 2*

The figure 5.8 shows that F partial is 2.4277 and p value is $0.1271 \succ 0.05$. This implies that the p-value is not in the rejection region. So, there is not much evidence to reject the null hypothesis. So, we can say that the reduced model 2 is suitable.

Reduced model 2:

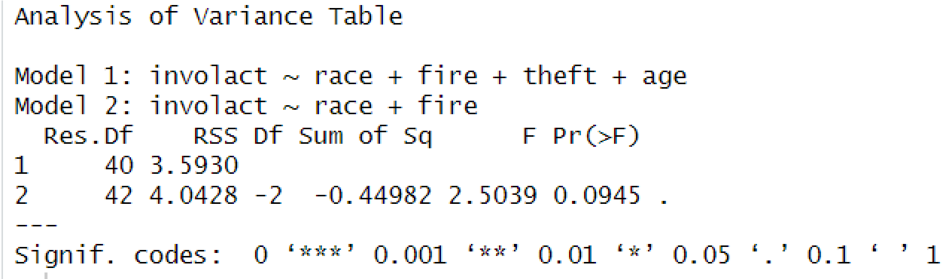
After we get summary of reduced model 2 and find the variable with the highest p value.



```
Call:
lm(formula = involact ~ race + fire + age, data = chredlin, subset =
-c(6,
    24))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63482 -0.20088  0.00672  0.13208  0.86515 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.353968   0.130486  -2.713  0.00971 ** 
race         0.005885   0.001827   3.222  0.00250 ** 
fire         0.049547   0.008362   5.925 5.53e-07 ***
age          0.003566   0.002258   1.579  0.12202    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3049 on 41 degrees of freedom
Multiple R-squared:  0.7917,	Adjusted R-squared:  0.7765 
F-statistic: 51.96 on 3 and 41 DF,  p-value: 5.008e-14
```

Figure 5.9: *Summary Reduced Model 2*

Figure 5.9 illustrates that in reduced model 1 R^2 is 79.17% and “Age” has highest p = 0.12202 $\succ$ 0.05 value. Then we reduced that variable and get new model call as reduce model 3. After we use Nested ANOVA method and consider probability of F partial value is greater than or not 95% Confidence Interval.



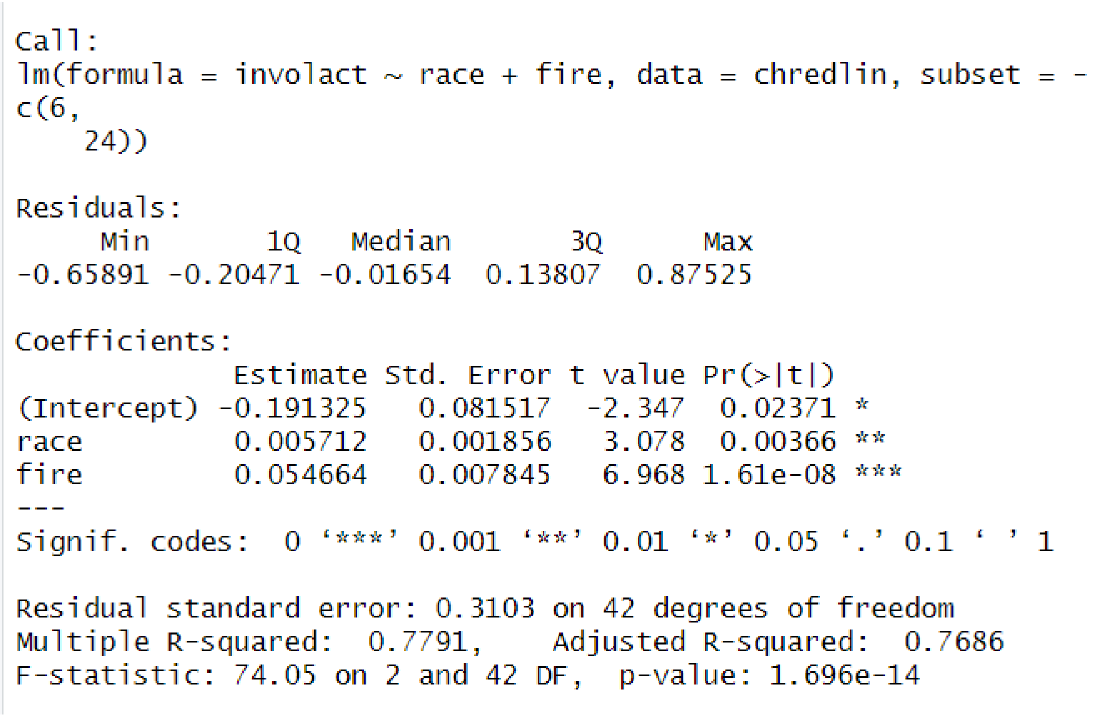
```
Analysis of Variance Table

Model 1: involact ~ race + fire + theft + age
Model 2: involact ~ race + fire
  Res.Df    RSS Df Sum of Sq      F Pr(>F)  
1     40 3.5930                             
2     42 4.0428 -2  -0.44982 2.5039 0.0945 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.10: *Nested ANOVA table for reduce model 2 and reduced model 3*

The figure 5.10 shows that F partial is 2.5039 and p value is 0.0945 $\succ$ 0.05. This implies that the p-value is not in the rejection region. So, there is not much evidence to reject the null hypothesis. So, we can say that the reduced model 3 is suitable.

Reduced model 4:

After we get summary of reduced model 3 and find the variable with the highest p value.



```
Call:
lm(formula = involact ~ race + fire, data = chredlin, subset = -
c(6,
    24))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65891 -0.20471 -0.01654  0.13807  0.87525 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3103 on 42 degrees of freedom
Multiple R-squared:  0.7791,	Adjusted R-squared:  0.7686 
F-statistic: 74.05 on 2 and 42 DF,  p-value: 1.696e-14
```

Figure 5.11: *Summary Reduced Model 3*

According to Figure 5.11, we can see that R^2 value is 77.91% and all variables interpret p-value below the significant intervals. All the independent variables having p-values below the significance level suggests that the full model explains a substantial portion of the variance in the dependent variable, and all the included variables are statistically significant contributors to the model.

For this study, the best regression equation is,

INVOLACT = - 0.191325 + 0.005712.Race + 0.054664.Fire

In this model, the R^2 value in is 77.91%. This indicates that a significant portion of variability in INVOLACT is explained by this model. The adjusted R^2 value of this model is 76.86%. The difference between the R^2 value and adjusted R^2 value is very small (about 0.0105) than the other steps. This means the addition of predictor variables do not improve the model significantly. In other words, it means that this is the best regression model for this dataset.

5.3 Assess Model Fit

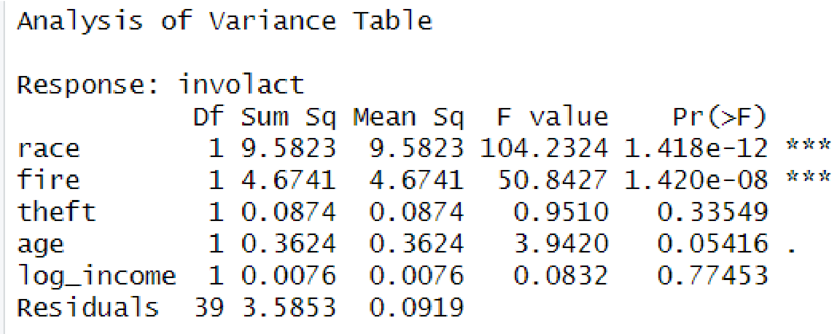
Here are the regression equations of two models.

Full Model :

INVOLACT = - 0.576737 + 0.007053.Race + 0.049647.Fire - 0.006434.Theft + 0.005171.Age + 0.115703.log(Income)

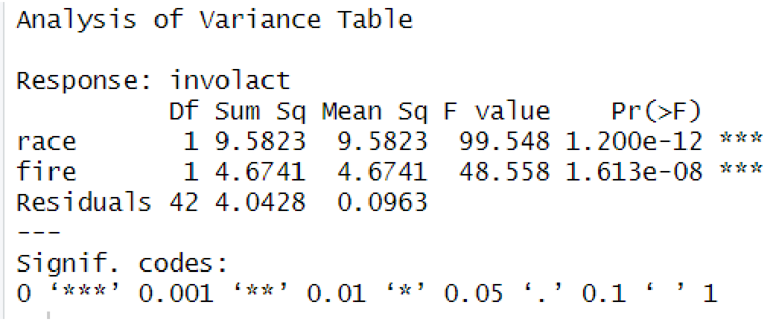
Reduced Model :

INVOLACT = - 0.191325 + 0.005712.Race + 0.054664.Fire



```
Analysis of Variance Table

Response: involact
           Df Sum Sq Mean Sq  F value    Pr(>F)    
race        1 9.5823  9.5823 104.2324 1.418e-12 ***
fire        1 4.6741  4.6741  50.8427 1.420e-08 ***
theft       1 0.0874  0.0874   0.9510   0.33549    
age         1 0.3624  0.3624   3.9420   0.05416 .  
log_income  1 0.0076  0.0076   0.0832   0.77453    
Residuals  39 3.5853  0.0919                       
```

Figure 5.12: *ANOVA table of full model*



```
Analysis of Variance Table

Response: involact
          Df Sum Sq Mean Sq F value    Pr(>F)    
race       1 9.5823  9.5823  99.548 1.200e-12 ***
fire       1 4.6741  4.6741  48.558 1.613e-08 ***
Residuals 42 4.0428  0.0963                      
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.13: *ANOVA table of reduced model*

1. **Hypothesis 1**

The full model is considered for checking this hypothesis.

- $H_0 : \beta_1 = 0 \ \ and \ \ H_1 : \beta_1 \neq 0$
  The p-value of race is 0.0126 and it is less than 0.05. That means the value is in the rejection region. So, the null hypothesis is rejected. Therefore, $\beta_1 \neq 0$. This suggests that there is a linear relationship between race and INVOLACT. According to the reduced model, $\beta_1 = 0.005712$.

- $H_0 : \beta_2 = 0 \ \ and \ \ H_1 : \beta_2 \neq 0$
  The p-value of fire is 1×10^{-06} and it is less than 0.05. That means the value is in the rejection region. So, the null hypothesis is rejected. Therefore, $\beta_2 \neq 0$. This suggests that there is a linear relationship between fire and INVOLACT. According to the reduced model, $\beta_2 = 0.054664$.

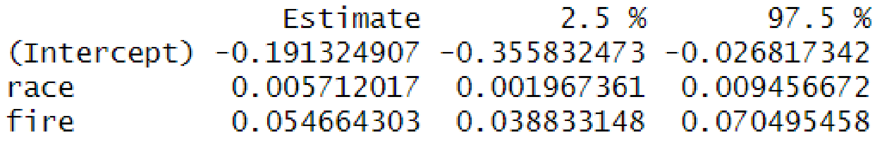
- $H_0 : \beta_3 = 0 \ \ and \ \ H_1 : \beta_3 \neq 0$
  The p-value of theft is 0.1471 and it is greater than 0.05. So, there is not much evidence to reject the null hypothesis. Therefore, $\beta_3 = 0$. This suggests that there is no a linear relationship between side numeric and INVOLACT. According to the reduced model, $\beta_3 = 0$.

- $H_0 : \beta_4 = 0 \ \ and \ \ H_1 : \beta_4 \neq 0$
  The p-value of age is 0.0818 and it is greater than 0.05. So, there is not much evidence to reject the null hypothesis. Therefore, $\beta_4 = 0$. This suggests that there is no a linear relationship between age and INVOLACT. According to the reduced model, $\beta_4 = 0$.

- $H_0 : \beta_5 = 0 \ \ and \ \ H_1 : \beta_5 \neq 0$
  The p-value of log-income is 0.7745 and it is greater than 0.05. So, there is not much evidence to reject the null hypothesis. Therefore, $\beta_5 = 0$. This suggests that there is no a linear relationship between log-income and INVOLACT. According to the reduced model, $\beta_5 = 0$.

Confidence Interval For Reduced Model

A 95% confidence interval for β_i has two equivalent definitions: The interval is the set of values for which a hypothesis test to the level of 5% cannot be rejected. The interval has a probability of 95% to contain the true value of β_i



```
               Estimate        2.5 %       97.5 %
(Intercept) -0.191324907 -0.355832473 -0.026817342
race         0.005712017  0.001967361  0.009456672
fire         0.054664303  0.038833148  0.070495458
```

Figure 5.14: *Confidence Interval for Reduced Model*

Intercept:
The 95% confidence interval for the intercept is from approximately -0.355832 to -0.026817. This means that we are 95% confident that the true intercept of the population regression line falls within this range.

Race:
The 95% confidence interval for the coefficient of race is from approximately 0.001967 to 0.009457. This indicates that we are 95% confident that the true effect of racial composition in percentage of minority between 0.001967 and 0.009457. Since the interval does not include 0, it suggests that racial composition in percentage of minority has a statistically significant positive effect on INVOLACT.

Fire:
The 95% confidence interval for the coefficient of fire is from approximately 0.038833 to 0.009456. This means that we are 95% confident that the true effect of fire rate is between 0.038833 and 0.009456. Since the interval does not include 0, it suggests that fire rate has a statistically significant positive effect on INVOLACT.

Finally, the suitable model we obtained can be written as follows,

INVOLACT = - 0.191325 + 0.005712.Race + 0.054664.Fire

Where,

β_0 = -0.191325
This is the y-intercept (value of y when all other parameters are set to 0)

β_1 = 0.005712
This is the slope and indicates that percentage of minority(race) must increase by 0.005712 to increase one additional INVOLACT value.

β_2 = 0.054664
This is the slope and indicates that Fire rate must increase by 0.054664 to increase one additional INVOLACT value.

Chapter 6

Discussion And Conclusion

6.1 Model Interpretation

We obtain a multiple linear regression model in this analysis to understand the relationships between new FAIR planning policies (INVOLACT) and various predictors as racial composition, fire and theft rates, age of housing and median family income in Chicago's 47 ZIP codes.

After satisfying by assumptions, it is possible to use multiple linear regression for the data set we selected. Our regression model which starts with a complete model, removed the least significant variables based on their p-value using backward elimination method. Refined. The initial full model showed a R^2 value of 80.41%, indicating that the variables could explain approximately 80% of the variance in INVOLACT. However, "log-income" had a high p-value (0.7745), suggesting that it was not a significant predictor. This was cause to the creation of reduced model 1, excluding the "log-income". The reduced model shows a R^2 value of 80.37%. A new model was created as reduced model 2 excluding "theft" due to showing the highest p-value (0.12709) with "theft". The reduced model 2 had a R^2 value of 79.17%. Then we created reduced model3 excluding "age" because reduced 2 had highest p-value (0.12202) with "age". The reduced model 3 had a R^2 value of 77.91% and the p-values of all remaining variables were below the significance threshold, indicating their importance in predicting INVOLACT.
Nested ANOVA tests comparing the full and reduced models showed that the p-value for F_{partial} was greater than 0.05 at each step (0.1917 and 1.6587), meaning there was no significant difference between two models. This supports the conclusion that reduced model 3 is the most efficient and effective for predicting INVOLACT.

Finally, the suitable model we obtained can be written as follows,
INVOLACT = - 0.191325 + 0.005712.Race + 0.054664.Fire

6.2 Link Back To The Question

In our problem statement aimed to determine the factors influencing the number of new FAIR planning policies in Chicago. The final model shows that racial composition and fire rate are significant predictors of INVOLACT, while variables such as theft rate, age of housing and median family income do not effect significantly to the model. This finding suggests that neighborhoods with racial compositions and fire rates are more likely have higher levels of new FAIR plan policies.

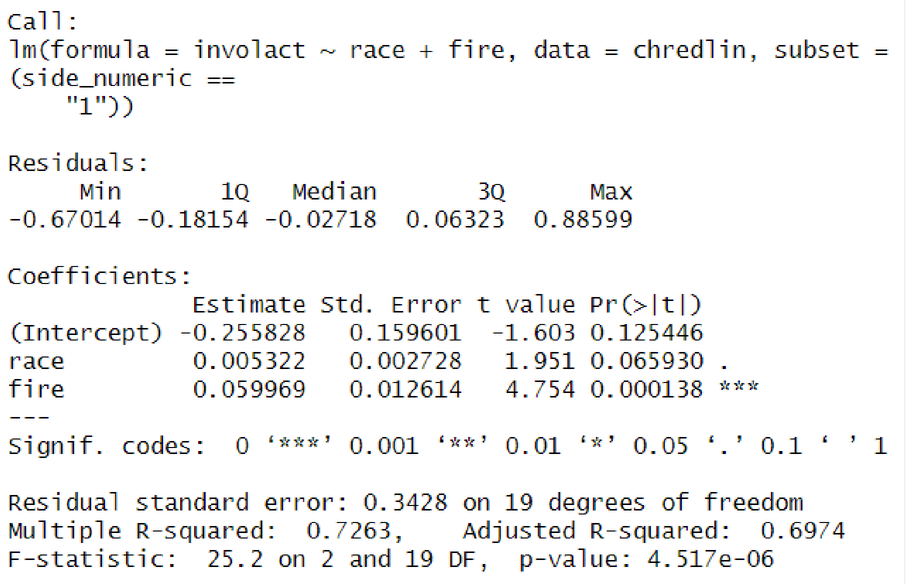
6.3 Compare with other related findings

The current literature that describes urban socio-economic dynamics slightly supports the findings of this study. Many times through studies, it has been shown that those with high percentages of minorities and in high-theft neighborhoods often face a higher rate in insurance claims, partly as a result of systemic issues like bad infrastructure and services. For instance, earlier research into urban crime and socio-economic factors has tended to find lower incomes and higher minority populations illustrating higher instances of crime, which can in turn result in higher insurance claims. The current study therefore underlines such patterns by showing that insurance claims activity is similarly influenced by socio-economic and demographic variables.

6.4 Conclusion

This study identified key predictors of Chicago's new FAIR planning policies using a strong multiple linear regression approach. The final model emphasizes the significant roles of racial composition and fire rates in determining INVOLACT levels.

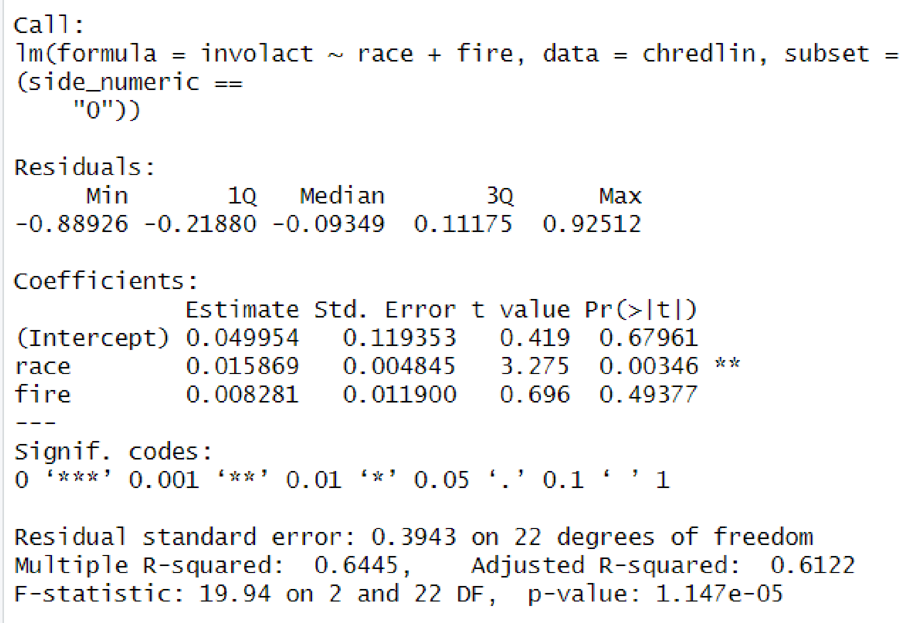
The response to insurance denied to minority homeowners is less than perfect and raises ecological correlations. It is unlikely that the probability of a minority homeowner obtaining a FAIR plan is constant across zip codes. If truth is variation about a constant, then the conclusions may be way off the mark. If a study is to be efficient, individual-level data are required in order for latent variables that might be the real reason for observed dependencies to be identified. Firsthand insurance business knowledge may introduce such a latent variable and render conclusions dubious. Another problem is developmentally appropriate data aggregation. Fit the model to the south of Chicago:



```
Call:
lm(formula = involact ~ race + fire, data = chredlin, subset =
(side_numeric ==
    "1"))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67014 -0.18154 -0.02718  0.06323  0.88599 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3428 on 19 degrees of freedom
Multiple R-squared:  0.7263,	Adjusted R-squared:  0.6974 
F-statistic:  25.2 on 2 and 19 DF,  p-value: 4.517e-06
```

Figure 6.1: *Reduced model with side==s*

and now to the north:



```
Call:
lm(formula = involact ~ race + fire, data = chredlin, subset =
(side_numeric ==
    "0"))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.88926 -0.21880 -0.09349  0.11175  0.92512 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(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   
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3943 on 22 degrees of freedom
Multiple R-squared:  0.6445,	Adjusted R-squared:  0.6122 
F-statistic: 19.94 on 2 and 22 DF,  p-value: 1.147e-05
```

Figure 6.2: *Reduced model with side==n*

We see that race is significant in the north, but not in the south. fire is significant in the south, but not in the north. By continuing to subsetting, one can make any predictor insignificant. On the other hand it is folly to aggregate all data without regard as to whether it is reasonable to do so. Clearly a judgement has to be made and this can be a point of contention in legal cases.

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.

R. W. Bostic and B. J. Surette. Have the doors opened wider? trends in homeownership rates by race and income. *The Journal of Real Estate Finance and Economics*, 23: 411–434, 2001.

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.

R. W. Klein and M. F. Grace. Urban homeowners insurance markets in texas: A search for redlining. *Journal of Risk and Insurance*, pages 581–613, 2001.

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.

G. D. Squires. *Insurance redlining: Disinvestment, reinvestment, and the evolving role of financial institutions*. The Urban Insitute, 1997.

G. D. Squires. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410, 2003.

Ansfield (2021), Bostic and Surette (2001), Grace et al. (2004), Klein and Grace (2001), Li et al. (2022), Faraway (2005), Squires (1997), Squires (2003)

Appendix A

Appendix

A.1 To Read Data Set

Clich Here to See the Insurance Data Set

A.2 R Code File

Clich Here to See the Insurance Data R Code

```
library(ggplot2)
library(dplyr)
library(faraway)
library(car)

data(”chredlin”)
View(chredlin)

summary(chredlin)

# Check for null values in the entire data frame
all-null-values ← is.null(chredlin)
all-null-values

# count total null values
sum(is.null(chredlin))

# Check for missing values in the entire data frame
```

```
all-missing-values ← is.na(chredlin)
all-missing-values

# count total missing values
sum(is.na(chredlin))

#Assigning numerical values for non-numeric values
side-num ← c(“n”=0,“s”=1)
side-numeric ← side-num[chredlin$side]

# Some adjustment - log(income)
correlation-matrix ← cor(chredlin[, c(“race”,“fire”, “theft”, “age”, “involact”, “income”)])
print(correlation-matrix)
cor(chredlinrace, chredlinincome)

plot( race ˜  income, chredlin)
abline(lm(race ˜  income, chredlin))
# race and income have high correlation, so we get log(income) values

full-model1 ← lm(involact ˜  race + fire + theft + age + log(income), chredlin)

summary(full-model1)
plot(full-model1)

log-income ← log(chredlin$income)
chredlin$ log-income ← log-income
chredlin$ side-numeric ← side-numeric
View(chredlin)

chredlin ← chredlin[,c(-6,-7)]
View(chredlin)
summary(chredlin)

# all plots
pairs(chredlin)

# Check the normality
qqnorm(residuals(full-model1))
qqline(residuals(full-model1), col=“red”,lwd=1)
```

```
hist(residuals(full-model1), main="Residual Histogram" )

#Linearity - scatter plots

# Multicolinearity

correlation-matrix ← cor(chredlin[, c("race","fire", "theft", "age", "involact", "log-income")])
print(correlation-matrix)

#VIF (less than 10)
vif(full-model1)

# Homoscedasticity

plot(fitted(full-model1), residuals(full-model1), xlab = "Fitted", ylab = "Residuals",
main = "Residuals Vs Fitted values")
abline(h=0, , col="red",lwd=1)

# Cook distances - Checking for outliers
cook-model ← influence(full-model1)
cook-model
qqnorm(cook-model$coef[,4])

halfnorm(cooks.distance(full-model1))

# Check jackknife residuals for outlier
range(rstudent(full-model1))

# 2 Outliers
chredlin[c(6,24),]

# Now our new full model
full-model2 ← lm (involact ˜  race + fire + theft + age + log-income, chredlin, subset=-
c(6, 24))
summary(full-model2)
plot(full-model2)

# Compute correlation matrix after romove outlier
correlation-matrix ← cor(chredlin[, c("race","fire", "theft", "age", "involact", "log-income")])
```

```
print(correlation-matrix)

# Parcial residual plots race & fire
prplot(full-model2,1)
prplot(full-model2,2)

#Quantitative summary
summary(chredlin,subset=-c(6, 24))

Graphical Summary (Strip Plots, Histograms, Boxplots)
par(mfrow=c(2,3))
for (i in 1:6) stripchart (chredlin [,i], main=names(chredlin) [i], vertical=TRUE, method=“jitter”)
#OR
for(i in 1:6) hist(chredlin[,i],main=names(chredlin)[i])
for(i in 1:6) boxplot(chredlin[,i],main=names(chredlin)[i])
pairs(chredlin)

#Scatter plots

plot(involact ˜ race, chredlin, main=“Scatter plot of Race and INVOLACT”)
abline(lm(involact ˜ race, chredlin), col=“blue”,lwd=1)

plot(involact ˜ fire, chredlin, main=“Scatter plot of Fire and INVOLACT”)
abline(lm(involact ˜ fire, chredlin), col=“blue”,lwd=1)

plot(involact ˜ theft, chredlin, main=“Scatter plot of Theft and INVOLACT”)
abline(lm(involact ˜ theft, chredlin), col=“blue”,lwd=1)

plot(involact ˜ age, chredlin, main=“Scatter plot of Age and INVOLACT”)
abline(lm(involact ˜ age, chredlin), col=“blue”,lwd=1)

plot(involact ˜ log-income, chredlin, main=“Scatter plot of log(Income) and INVOLACT”)
abline(lm(involact
textasciitilde log-income, chredlin), col=“blue”,lwd=1)

# Backward Elemination

full-model ← lm (involact ˜ race + fire + theft + age + log-income, chredlin, subset=-
c(6, 24))
```

```
summary(full-model)

# Reduce 1
reduce-model1 ← lm(involact ˜ race + fire + theft + age , chredlin, subset=-c(6, 24) )

anova(full-model,reduce-model1)
summary(reduce-model1)

# Reduce 2
reduce-model2 ← lm(involact ˜ race + fire + age, chredlin, subset=-c(6, 24) )

anova(reduce-model1,reduce-model2)
summary(reduce-model2)

# Reduce 3
reduce-model3 ← lm(involact ˜ race + fire, chredlin, subset=-c(6, 24) )

anova(reduce-model1,reduce-model3)
summary(reduce-model3)

estimate-model ← reduce-model3
anova(full-model)
anova(estimate-model)
anova(full-model,estimate-model)

#AIC (less AIC)
AIC(full-model)
AIC(estimate-model)

Confint(estimate-model)

#Now we sea how to affect side data for our model

full-model-south ← lm (involact ˜ race+fire, chredlin , subset=(side-numeric == “1”))
summary (full-model-south)

full-model-north ← lm (involact ˜ race+fire, chredlin , subset=(side-numeric == “0”))
summary (full-model-north)
```