Chicago Insurance Redlining

An Analysis of Institutionalized Racism’s Impact on Low-Income Communities

MA576 Final Project

Pratibha Pattabhiraman

April 24, 2016

I. Introduction

The data set I chose for this report was found in the faraway library package under ‘chredlin’, and it stood out to me in particular because of its apt relevance in our country’s current racial climate. The concept of redlining is not foreign, nor is it specific to this time period in Chicago. Redlining has been happening all over the country in the later half of the 20th century, more commonly in major cities, and even in Boston. The goal of redlining has most often been a tool for politicians and other officials to lay the groundwork for “legalized racism.” By keeping folks of color out of white neighborhoods and opportunities, either directly or by increasing prices, cities can create visible lines between different communities. This was extremely harmful to minorities back in the 70s, and still carries out into the 21st century with repercussions for those communities today. It’s important to realize that racism isn’t always interpersonal or blatant. More often times, it’s institutional and subtle.

Many cities are still divided into the “good” or “bad” areas, terms and spaces known colloquially to local residents only. The majority of white people still live in the “good” or more wealthy areas, and the majority of black/Hispanic people still live in the “bad” or more impoverished areas. This split is a direct result of redlining. By keeping the low-income communities in a separate part of the city, the government is able to clearly define where its money should be going and who it should be benefitting, and 9 times out of 10, that money never reaches the communities that need it the most. This lack of funding and support at an institutional level forces families to send their children to public schools with minimal funding, forces children to receive poor education, and creates an environment where high education is nearly impossible. In addition, the added stress of racism and living in poverty creates health issues for communities of color, but access to quality healthcare and funds to see proper primary care physicians is nearly nonexistent. In spite of all this, most people try and juggle multiple minimum wage jobs and struggle to simultaneously spend time with their children at home. All of this can take a toll on these people, and not getting insurance because of the situation they’ve been placed into can cause even more stress. But then a multitude of stereotypes get painted onto people of color, typically from upper-class white folk, about their laziness or inability to reach higher educations, but where you live and what kind of community you’re a part of has a great deal to do with your future success. And that’s what institutions need to realize.

That is also why this particular dataset is so interesting to me. I’ve never really thought about redlining in terms of insurance, but it makes sense in the context of the time period and who companies would trust to live responsibly. Poor communities are often known for high crime rates, older housing developments, and high racial compositions. As a result, insurance companies redlined them in terms of access to good coverage because they knew that it would be more expensive to insure a black family living in poor conditions than it would be to insure a white family living comfortably in a rich community. While these actions seem logical, and even economical, they are products of racism and are should be taken into account when discussing the racial disparities that exist in our country today.

My estimate for what this data set will reveal in terms of racism and wealth disparity shouldn’t be very surprising. As a preliminary analysis, I would like to explore if it is indeed true and significant that white communities have lower amounts of crime, and have greater incomes than communities of color. And then to jump off that, if it is correct, I would expect that zip codes with higher minority populations, lower incomes, and higher crime rates are expected to receive little to no insurance. On the other hand, areas with little to no minority population, higher incomes, and lower crime rates are expected to receive greater insurance coverage. It would also be interesting to see what the age of the buildings in each zip code tell about the people living there, because to me, it would seem logical that older housing developments would be cheaper and more affordable for people with lower incomes, which tends to be people of color. Following this frame of logic, newer housing developments should be tied to communities with higher levels of income, most often afforded to white people. It will be interesting to see what the data has to say about these claims, but sociologically it makes sense and is very real for a lot of people of color trying to get the same treatment as their white counterparts to this day.

II. Methods

This study, which focuses on insurance availability in Chicago, was used by the U.S. Commission on Civil Rights to try and resolve charges made by various community organizations that claimed insurance companies were redlining their neighborhoods. This means that the companies were either canceling, or failing to renew, insurances. So the Illinois Department of Insurance provided the total number of cancellations, non-renewals, new policies, and renewals of homeowners by zip code from December 1977 through February 1978 in order be transparent against any charges coming their way.

In addition to this information, the Chicago Police Department provided crime data on all thefts for the year 1975. Many insurance claim that they base their underwriting activities on loss data from precious years, so a 2-3 year lag is reasonable for analysis. The Fire Department in Chicago also provided similar data, but on fires from 1975. These are also organized by zip code.

Lastly, the US Bureau provided pertinent data on racial composition, income, and age of housing developments for each zip code in the Chicago area. In order to adjust for any differences in population size, the theft data was expressed as incidents per 1,000 people and the fire and insurance data were expressed as incidents per 100 housing units. Here’s a list of the independent and dependent variables before I go into the model we will using to illustrate the data:

**race** racial composition in percent of minorities

**fire** fires/100 housing units

**theft** theft/1000 population

**age** percent of housing units built before 1939

**income** median household income (in thousands)

**involact** new FAIR plan policies and renewals/100 housing units

**side** North or South side of Chicago

Since none of this data is categorical and all are continuous, it didn’t initially make sense to me that I should apply a generalized linear model. But I did apply one in the beginning anyway, using a Poisson regression, since the response variable is “policies and renewals per 100 housing units” and I didn’t want negative values to be possible as a response. I also decided to try a regular linear model later on, and use involact as the dependent variable. We’re using this value because it is the best way to measure the number of households getting denied insurance. However, I must mention that it is not a perfect value because some people who are denied insurance might give up and others may still not even be trying because they know the chances of receiving insurance in their situation might be slim to none.

When proceeding to fit a model, we must be careful not to jump to conclusions about correlation. Since we don’t have information on the racial composition of the individuals who are denied insurance, it’s difficult to make concrete conclusions about whether or not racism is the sole factor in the redlining. However, I do believe that in communities with higher minority compositions, the number of new insurance policies or renewals will be significantly lower that communities with a mainly white population, and it is not revolutionary to conclude that race plays an important part in that disparity. Also, race is a social construct that has been misused in these discussions of oppression. Researchers like to say that the true cause of issues like redlining, discrimination, etc., are not actually due to race, but in fact due to income, socioeconomic status, or crime rates. But race-based discrimination is what led to all these other disparities in the first place and overlooking race as merely a figment of society’s imagination would be irresponsible and can lead to misleading conclusions.

Having said that, we can run some initial analyses on the data, just to see some basic patterns between the variables. Using the summary() command, we can see that there’s a wide range in the race category, which is desirable because it gives us a fairly equal number of zip codes with high and low proportion of minorities. This reduces the variation in race, which allows us to assess the effect race has on involact more accurately. We are also able to see that income is skewed left and theft is skewed right, and that involact has a large number of zeroes. If this causes problems in the models later on, we can take steps to remedy the situation.

> summary(chredlin)

race fire theft age

Min. : 1.00 Min. : 2.00 Min. : 3.00 Min. : 2.00

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

Median :24.50 Median :10.40 Median : 29.00 Median :65.00

Mean :34.99 Mean :12.28 Mean : 32.36 Mean :60.33

3rd Qu.:57.65 3rd Qu.:16.05 3rd Qu.: 38.00 3rd Qu.:77.30

Max. :99.70 Max. :39.70 Max. :147.00 Max. :90.10

involact income side

Min. :0.0000 Min. : 5.583 n:25

1st Qu.:0.0000 1st Qu.: 8.447 s:22

Median :0.4000 Median :10.694

Mean :0.6149 Mean :10.696

3rd Qu.:0.9000 3rd Qu.:11.989

Max. :2.2000 Max. :21.480

> plot(chredlin)

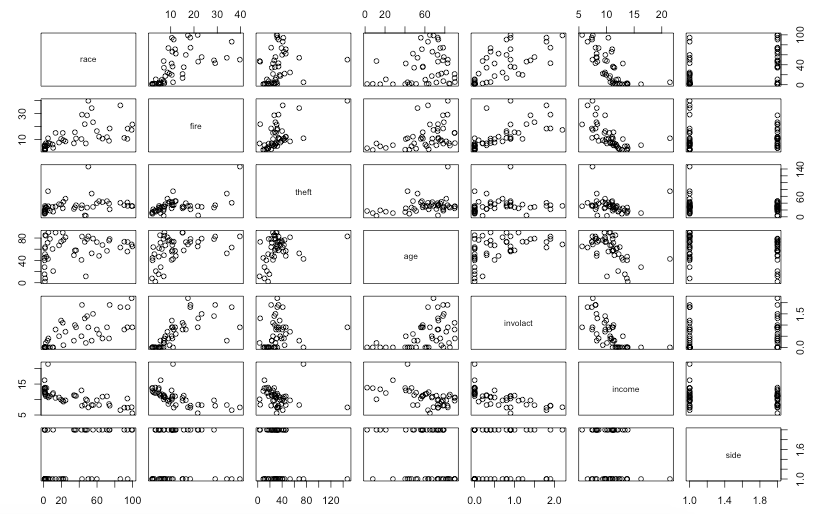
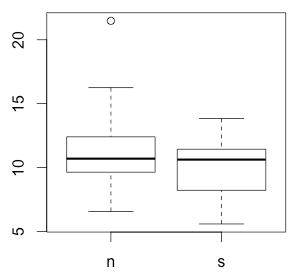
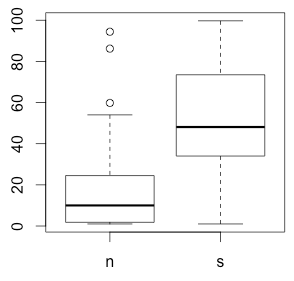
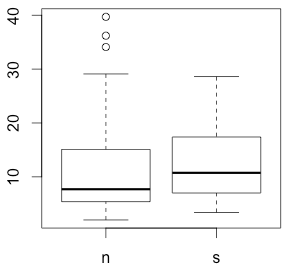
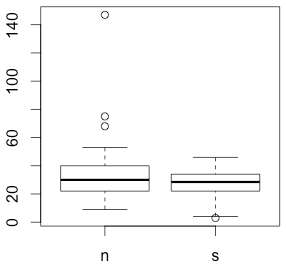


Figure 1: Scatterplot matrix of the data

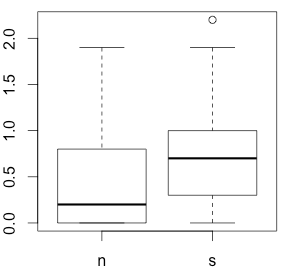
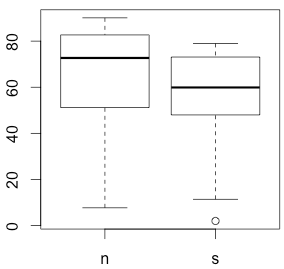
From the scatterplot matrix shown in Figure 1, we can see that there are positive correlations between race and theft, fire, age of housing developments, and involact. These are all not very surprising, and neither is the fact that income is negatively correlated with every aforementioned coefficient. Since it’s a little difficult to see the association between which side of the city you live on and the other variables, I decided to do separate boxplots for each relationship. From the boxplots below, we can gather that the South side of Chicago is highly associated with minority populations, while the North side seems to be less diverse. Apart from that, however, the rest of the variables seem to be fairly similar comparing North and South in terms of median. Income in the North side is shifted up slightly more, but doesn’t seem to be significant. In terms of involact, those living in the South side seem to be afforded greater amounts of new insurance and renewals. But this also isn’t significant, since we have more data recorded for the North side, and only a few for the South side. So even a couple high values for involact in the South side can make it seem like they receive better coverage. We won’t really know if this is truly significant or not until after we’ve run the models.



**Figure 2: Race Figure 3: Income**



**Figure 4: Theft Figure 5: Fire**



**Figure 6: Age Figure 7: Involact**

After examining this preliminary data and running a quick linear model of involact and race only, we can see that the percentage of minorities living in a particular zip code is indeed a significant predictor of whether or not that community will receive insurance.

Coefficients:

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

(Intercept) 0.129218 0.096611 1.338 0.188

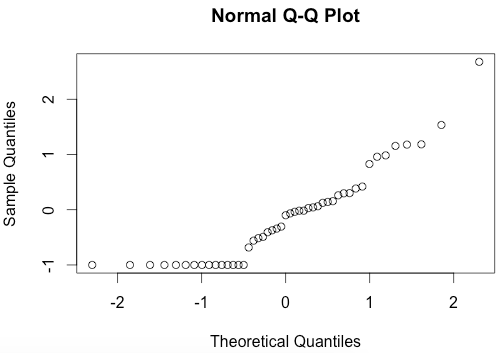
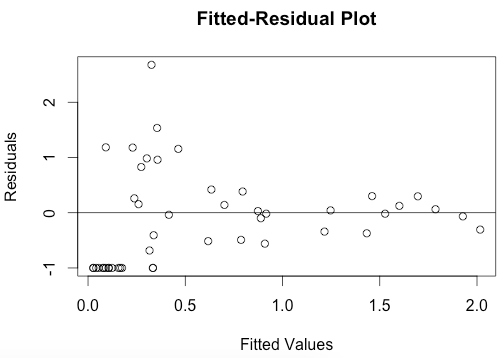
race 0.013882 0.002031 6.836 1.78e-08 \*\*\*

With this information, we can now run a series of models that will give us more concrete evidence on whether the other variables are also significant predictors. The initial model I’m going to run will be a generalized linear model using Poisson regression. I chose Poisson because this data set seemed like count data to me, and it illustrated the number of time a certain event occurred within a specific timeframe.

m1 <- glm(involact ~ race + age + log(income) + theft + fire +

side1, family = poisson)

I chose to use log(income) not only because this variable is skewed, and because income is better illustrated on a multiplicative scale. The only significant variable from this summary output ended up being age, but before I could make any hard conclusions, I had to check the model assumptions with a residual and Normal QQ plot.



These plots seem to suggest heteroscedasticity and non-normality. Heteroscedasticity means that the residuals get smaller as the predicted value increases which is not good, because we want constant variance in our model. There are many reasons for this but the two main ones are that the precision of the estimates will be lower without constant variance, and the p-values for the coefficients are based on satisfying the assumption of constant variance. So we must find a way to try and fix this. I re-built the model including m1$residuals at one of the predictors, which seemed to remedy the situation slightly but there were still the zero outlier variables hanging out in the bottom left side of the graph. I didn’t feel comfortable removing those values or transforming the dependent variable since the 0 values are important to the interpretation, so I must continue without using a GLM. Another reason for my decision is that our data set only has 47 observations, which is not nearly enough to claim that n is large, so fixing heteroscedasticity in a small sample such as this one is not easy or the best option. The linear model I’m starting off with is:

m2 <- lm(involact ~ race + age + log(income) + theft + fire)

III. Results

Coefficients:

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

(Intercept) -1.185540 1.100255 -1.078 0.287550

race 0.009502 0.002490 3.817 0.000449 \*\*\*

age 0.008336 0.002744 3.038 0.004134 \*\*

log(income) 0.345762 0.400123 0.864 0.392540

theft -0.010295 0.002818 -3.653 0.000728 \*\*\*

fire 0.039856 0.008766 4.547 4.76e-05 \*\*\*

From this output, we can see that race, age of housing developments, theft, and fire are all significant predictors for approved and renewed insurance. However, we must first check the assumptions of the model because we interpret further.

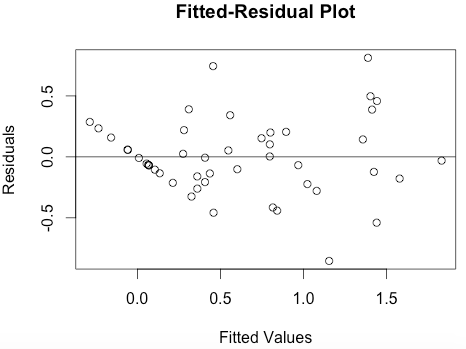


Figure 8: Residual-fitted values plot

Looking at the above plot, we don’t see anything out of the ordinary. The points seem to cluster randomly about the horizontal line at 0, and besides the diagonal line of points on the left side (which are due to the abnormal values of 0 for involact), the model passes this test.

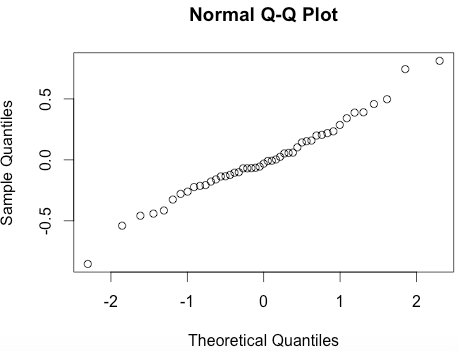


Figure 9: Normal QQ plot

Looking at the Normal QQ plot above, we can see a fairly linear line, with insignificant skewness at either end. The model also passes this test. One last factor we must examine are points of inference. Looking at the preliminary data analysis and quantiles for each of the variables, I noticed that both theft and fire had abnormally large maximum values. So I looked into the data and saw that entry #24 was the cause for these two values. In addition, the medians for both variables were much lower than the means which is sufficient evidence that there may be outliers present in the data. Removing this value from the data set gave a new summary for fire and theft, but the fire median was still 2 values lower than the mean so I removed another data point, which turned out to be entry #23. Just to be sure, I also removed entry #6. I wanted to see if the same variables remain significant after running this new model:

m3 <- lm(involact ~ race + age + log(income) + theft + fire, chredlin3)

The summary was as follows:

Coefficients:

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

(Intercept) -0.269600 1.097688 -0.246 0.8073

race 0.006086 0.002778 2.190 0.0347 \*

age 0.003757 0.003075 1.222 0.2293

log(income) -0.003348 0.408531 -0.008 0.9935

theft -0.005188 0.004421 -1.173 0.2479

fire 0.055488 0.009644 5.754 1.24e-06 \*\*\*

Age and theft are no longer significant, and fire remains the most significant factor.

The issue of variable transformation doesn’t seem important in this model since the response has a great deal of 0 values and interpretation would seem difficult. So we are able to fit the best model, which I believe to be:

m4 <- lm(involact ~ race + age + theft + fire, chredlin3)

Coefficients:

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

(Intercept) -0.278523 0.138566 -2.010 0.05138 .

race 0.006102 0.001843 3.312 0.00201 \*\*

age 0.003773 0.002412 1.564 0.12578

theft -0.005207 0.003720 -1.400 0.16951

fire 0.055496 0.009474 5.858 8.16e-07 \*\*\*

---

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

Residual standard error: 0.2968 on 39 degrees of freedom

Multiple R-squared: 0.7962, Adjusted R-squared: 0.7753

F-statistic: 38.1 on 4 and 39 DF, p-value: 5.584e-13

This model has a R2 value of .7962, which is equivalent to the m3 R2, meaning that the loss of income didn’t have a negative impact on the variability in involact explained by the predictors.

IV. Conclusion

Since we found that fire was the most significant predictor, with the highest t-value, insurance companies would be correct in saying that the reason why they were denying service in certain neighborhoods was due to a great number of fire-related incidents, and that any allegations of discrimination was merely a byproduct of their valid business decisions. Simply put, they would be able to fully deny racism and malpractice, but communities of color are not ignorant. They can see through the denial and are able to recognize that years of slavery, segregation, and housing redlining by institutions are the main reasons why they live in the communities that they do. I graphed a plot that compared age of housing with race, and found that communities with the highest percentage of minorities (over 60%) live in neighborhoods where over 58% of the housing developments were built before 1939. These kinds of infrastructures are not conducive to health or safety, and it would make sense that the greatest number of fire-related incidents occur in these environments. To see if this was true, I also ran a correlation test between fire and age and found that with a correlation coefficient of .412 and a p-value of .003987, this relationship was highly significant.

For insurance companies to claim that they are denying particular communities insurance based on the number of fires they’ve sustained in the past couple of years is just a way to mask the true problem and place blame where it shouldn’t fall. The quality of education we receive, our public health, and the types of privileges we are afforded by the government and other institutions are all byproducts of the racial disparity that most people believe to be a thing of the past. While blatant racism may not exist today, we must still grapple with its consequences and find ways to remedy the injustices against people of color in our country, instead of turning a blind eye and accepting redlining as the norm.