

**Chicago Insurance Redlining**

*(Regression Analysis Report)*

*By*

**Malik Hassan Qayyum**

*Supervisor*

**Dr. Ellie Small**

Table of Contents

[Background 3](#_Toc26640371)

[Data Source 4](#_Toc26640372)

[Variables 4](#_Toc26640373)

[Goal 5](#_Toc26640374)

[Initial Data Analysis 5](#_Toc26640375)

[Summary 5](#_Toc26640376)

[Scaling Income variable 5](#_Toc26640377)

[Assumptions 7](#_Toc26640378)

[Checking the linear Structure of the model 7](#_Toc26640379)

[Checking Normality 10](#_Toc26640380)

[Checking Error Variance 11](#_Toc26640381)

[Checking Collinearity 11](#_Toc26640382)

[Checking Unusual Observations 15](#_Toc26640383)

[Checking Leverage Points 15](#_Toc26640384)

[Checking Outliers 17](#_Toc26640385)

[Checking Influential Observations 20](#_Toc26640386)

[Transformations 21](#_Toc26640387)

[Power Transformation 21](#_Toc26640388)

[Polynomials 24](#_Toc26640389)

[Evaluating the model 25](#_Toc26640390)

[Train and Test data 25](#_Toc26640391)

[Training the model 26](#_Toc26640392)

[Testing the model 26](#_Toc26640393)

[References 28](#_Toc26640394)

# Background

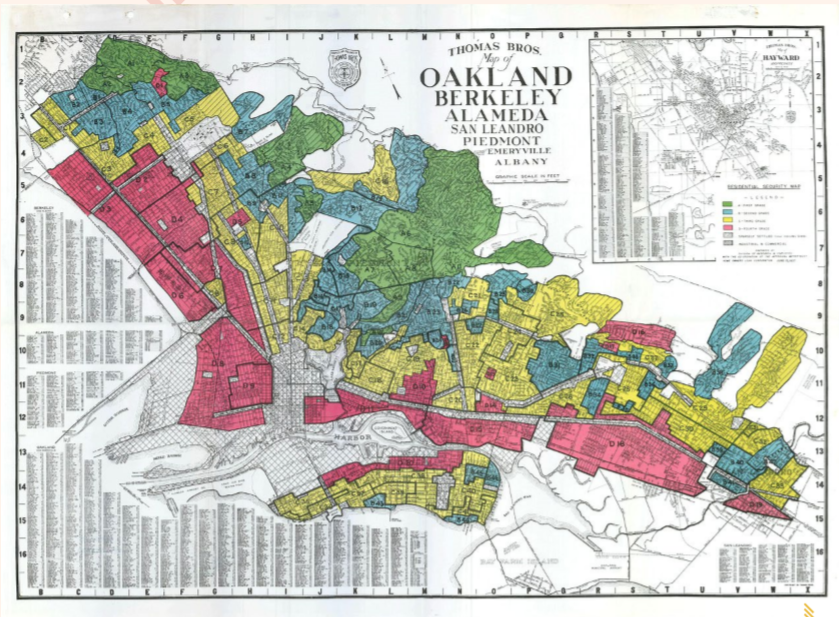
The term redlining originates from the 1930s practice of color-coding maps of cities based on different neighborhoods’ eligibility to receive a loan or mortgage. The lowest ranked neighborhoods were often literally lined in red and were almost always a community of color or other marginalized identity.

Redlining began in 1935 when the Home Owner’s Loan Corporation began producing maps of virtually every major city upon request of the Federal Home Loan Bank Board.

Neighborhoods were color coded based on their desirability, from “A - First Grade” to “D - Fourth Grade.” Most often the “D” ranking neighborhoods were black communities, or other communities of minorities, while the “A” ranking neighborhoods were affluent white suburbs.

The maps were used by both public and private banks and loan offices to directly discriminate and refuse loans to residents of the “D” neighborhoods.

The Fair Housing Act of 1968 made discrimination during the process of selling a house illegal, yet redlining was not effectively outlawed until 1977. The Home Mortgage Disclosure Act of 1975 required transparency thus making redlining unfeasible, and was followed by the Community Reinvestment Act of 1977 that finally prohibited it



A red lined map of Oakland, California, created by Home Owner’s Loan Corporation.

## Data Source

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, i.e. canceling policies or refusing to insure or renew.

First the Illinois Department of Insurance provided the number of cancellations, non-renewals, new policies, and renewals of homeowners and residential ﬁre insurance policies by ZIP code for the months of December 1977 through February 1978. The companies that provided this information account for more than 70% of the homeowner’s insurance policies written in the City of Chicago. The department also supplied the number of FAIR plan policies written a renewed in Chicago by zip code for the months of December 1977 through May 1978. Since most FAIR plan policyholders secure such coverage only after they have been rejected by the voluntary market, rather than as a result of a preference for that type of insurance, the distribution of FAIR plan policies is another measure of insurance availability in the voluntary market.

Secondly, the Chicago Police Department provided crime data, by beat, on all thefts for the year 1975. Most Insurance companies claim to base their underwriting activities on loss data from the preceding years, i.e. a 2-3-year lag seems reasonable for analysis purposes. the Chicago Fire Department provided similar data on ﬁres occurring during 1975. These ﬁre and theft data were organized by zip code.

Finally, the US Bureau of the census supplied data on racial composition, income and age and value of residential units for each ZIP code in Chicago. To adjust for these differences in the populations size associated with different ZIP code areas, the theft data were expressed as incidents per 1,000 population and the ﬁre and insurance data as incidents per 100 housing units.

## Variables

Following are the variables of the data source.

**race racial:** composition in percent minority

**fire:** ﬁres per 100 housing units

**theft:** theft per 1000 population

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

**volact:** new homeowner policies plus renewals minus cancellations and non-renewals per 100 housing units

**involact:** new FAIR plan policies and renewals per 100 housing units

**income:** median family income

## Goal

To compute the effect of different parameters on insurance redlining in 1975, in which race has been a dominant contributor. To Creating a Linear model for the involuntary market activity variable (the number getting FAIR plan insurance) based on the other parameters. Hence, we can compare the parameters who effects the redlining most in the past vs the one’s which are affecting it now. This regression analysis will give a comparison matric to the policy maker to measure the changes of insurance redlining now and then.

# Initial Data Analysis

## Summary

summary(chicago)

## 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

## volact involact income

## Min. : 0.50 Min. :0.0000 Min. : 5583

## 1st Qu.: 3.10 1st Qu.:0.0000 1st Qu.: 8447

## Median : 5.90 Median :0.4000 Median :10694

## Mean : 6.53 Mean :0.6149 Mean :10696

## 3rd Qu.: 9.65 3rd Qu.:0.9000 3rd Qu.:11989

## Max. :14.30 Max. :2.2000 Max. :21480

Income has much bigger numbers then other parameters. It would have greater weight in the regression model. Hence, to avoid this we standardize the income variable.

## Scaling Income variable

ch = chicago

ch$income = ch$income/1000

summary(ch)

## 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

## volact involact income

## Min. : 0.50 Min. :0.0000 Min. : 5.583

## 1st Qu.: 3.10 1st Qu.:0.0000 1st Qu.: 8.447

## Median : 5.90 Median :0.4000 Median :10.694

## Mean : 6.53 Mean :0.6149 Mean :10.696

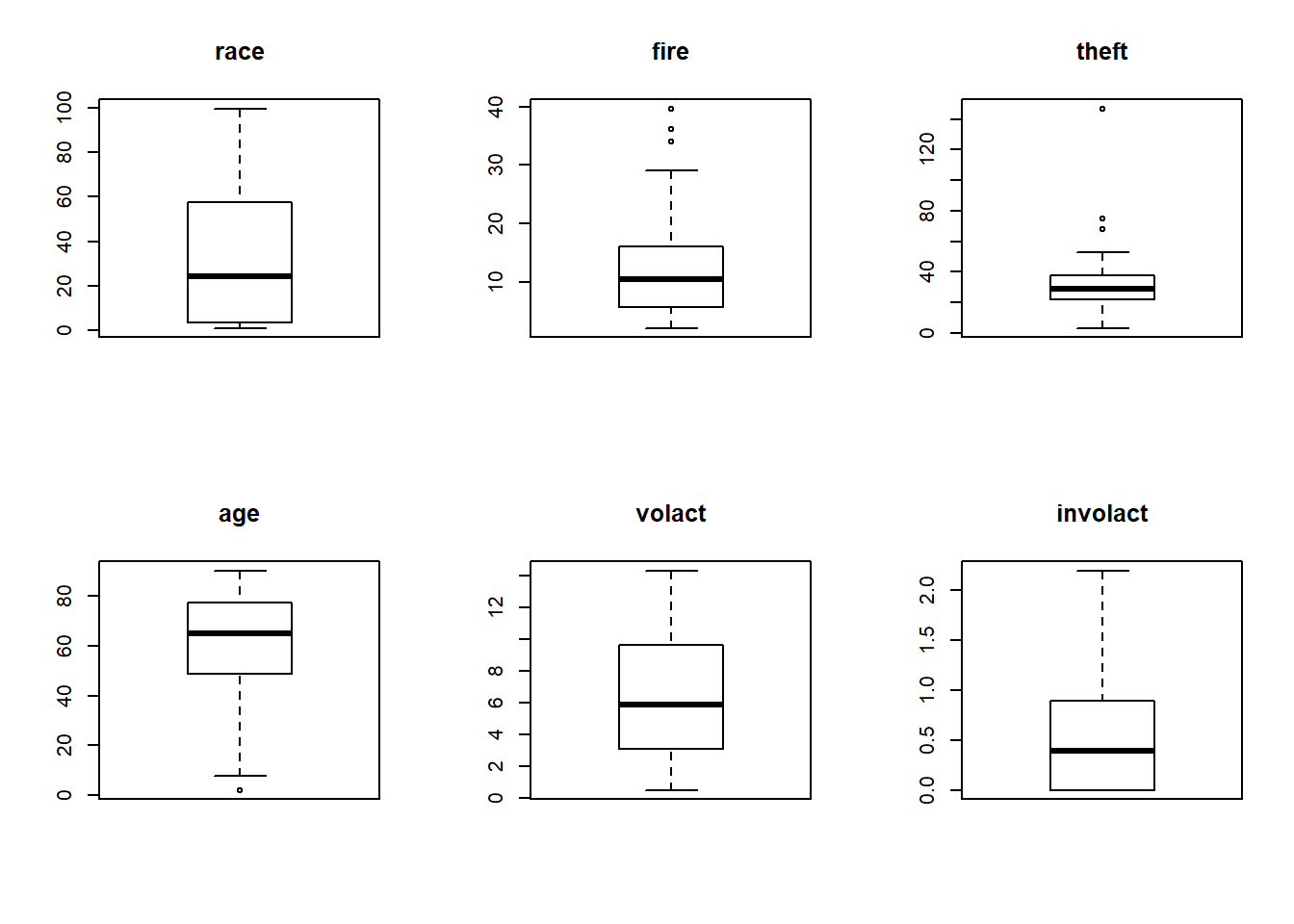
## 3rd Qu.: 9.65 3rd Qu.:0.9000 3rd Qu.:11.989

## Max. :14.30 Max. :2.2000 Max. :21.480

par(mfrow=c(2,3))

**for**(i **in** 1:6)

boxplot(chicago[,i],main=names(chicago)[i])



Boxplots show some unusual observations, that we are later going to deal with.

# Assumptions

## Checking the linear Structure of the model

ch= data.frame(ch)

lmod\_full <- lm(involact~., ch)

lmod\_full$rank

## [1] 7

summary(lmod\_full)

##

## Call:

## lm(formula = involact ~ ., data = ch)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.84296 -0.14613 -0.01007 0.18386 0.81235

##

## Coefficients:

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

## (Intercept) -0.486201 0.602048 -0.808 0.424109

## race 0.008527 0.002863 2.978 0.004911 \*\*

## fire 0.037780 0.008982 4.206 0.000142 \*\*\*

## theft -0.010160 0.002908 -3.494 0.001178 \*\*

## age 0.007615 0.003330 2.287 0.027582 \*

## volact -0.010180 0.027734 -0.367 0.715519

## income 0.025685 0.032199 0.798 0.429759

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

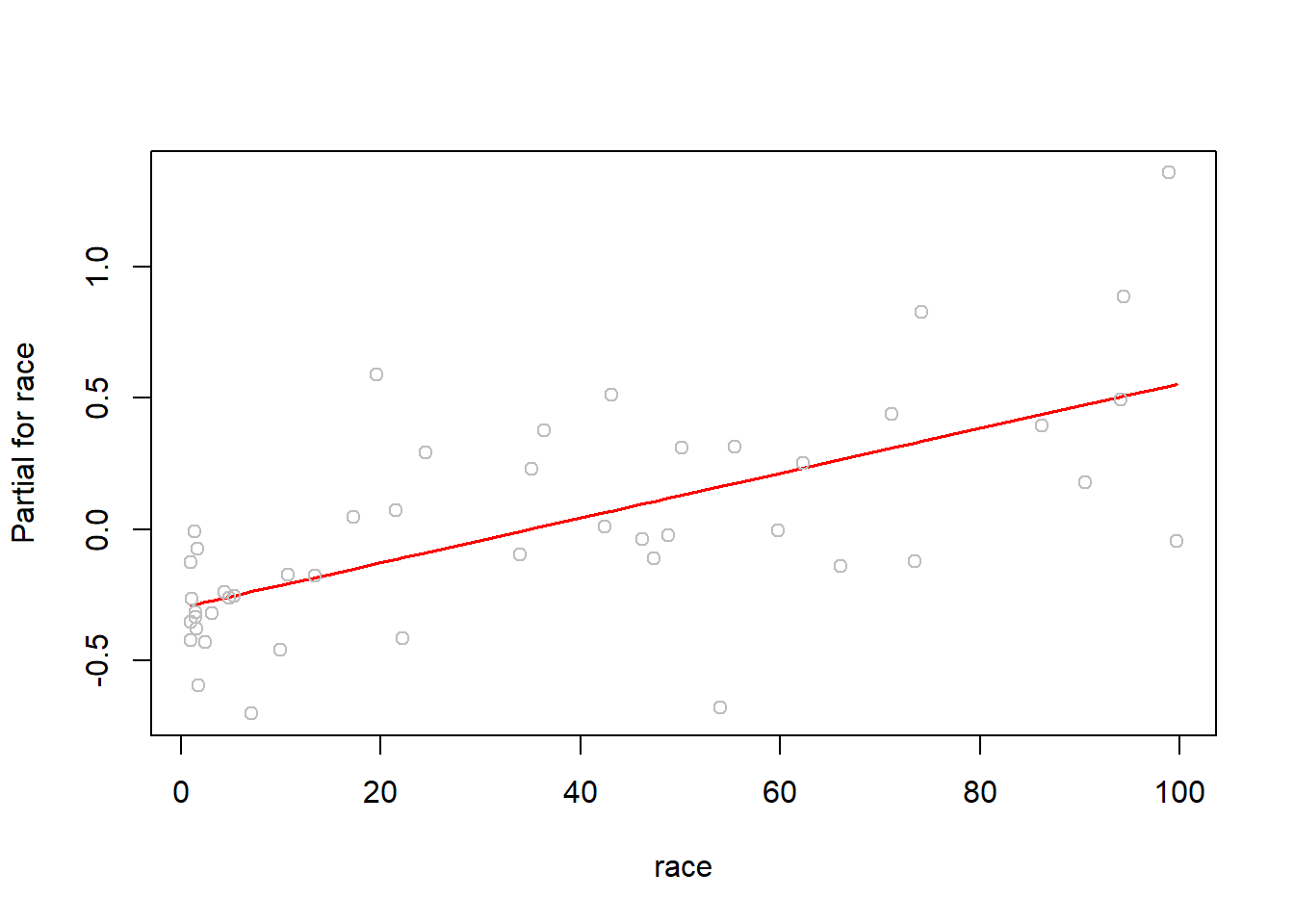
##

## Residual standard error: 0.3387 on 40 degrees of freedom

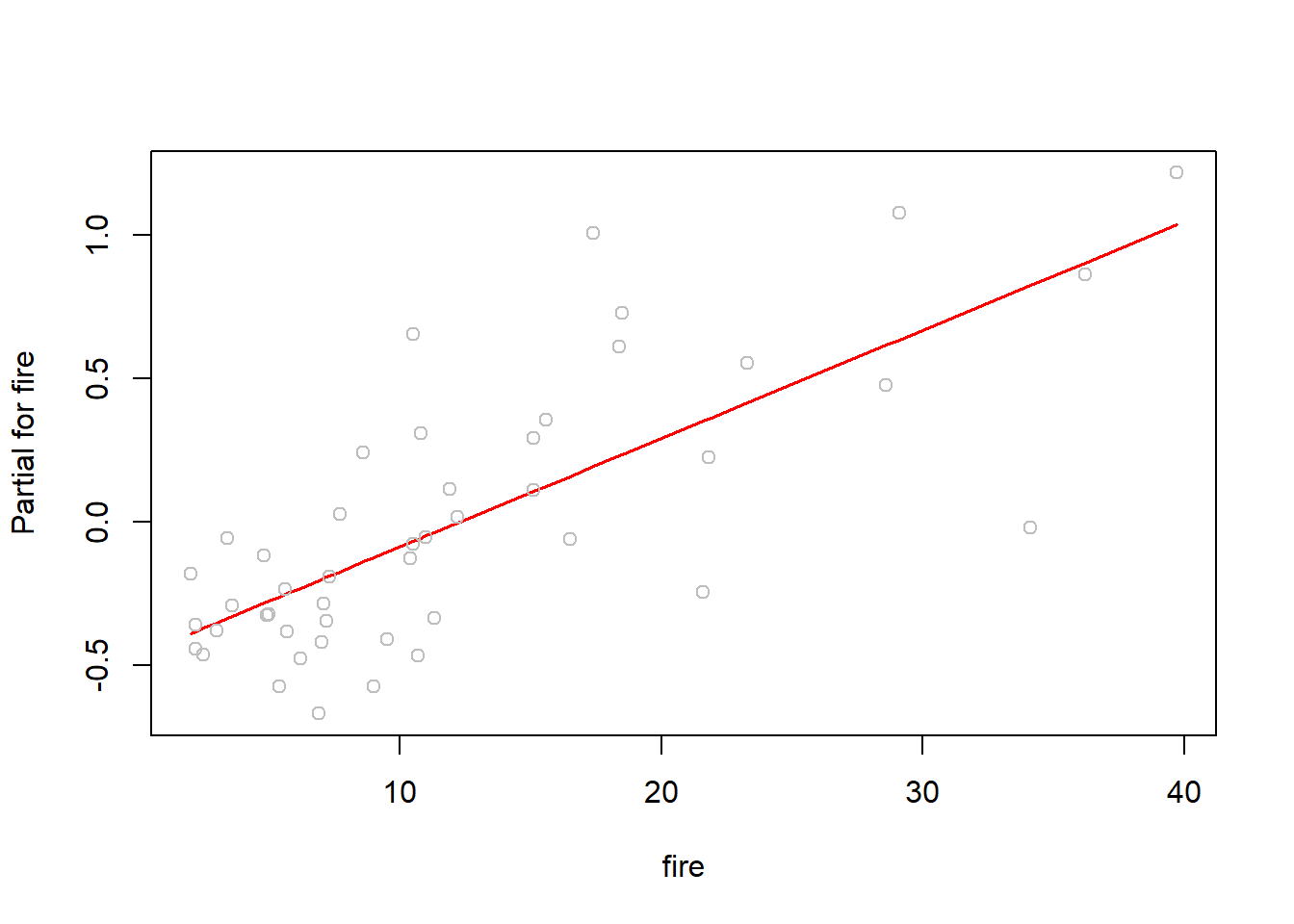
## Multiple R-squared: 0.7517, Adjusted R-squared: 0.7144

## F-statistic: 20.18 on 6 and 40 DF, p-value: 1.072e-10

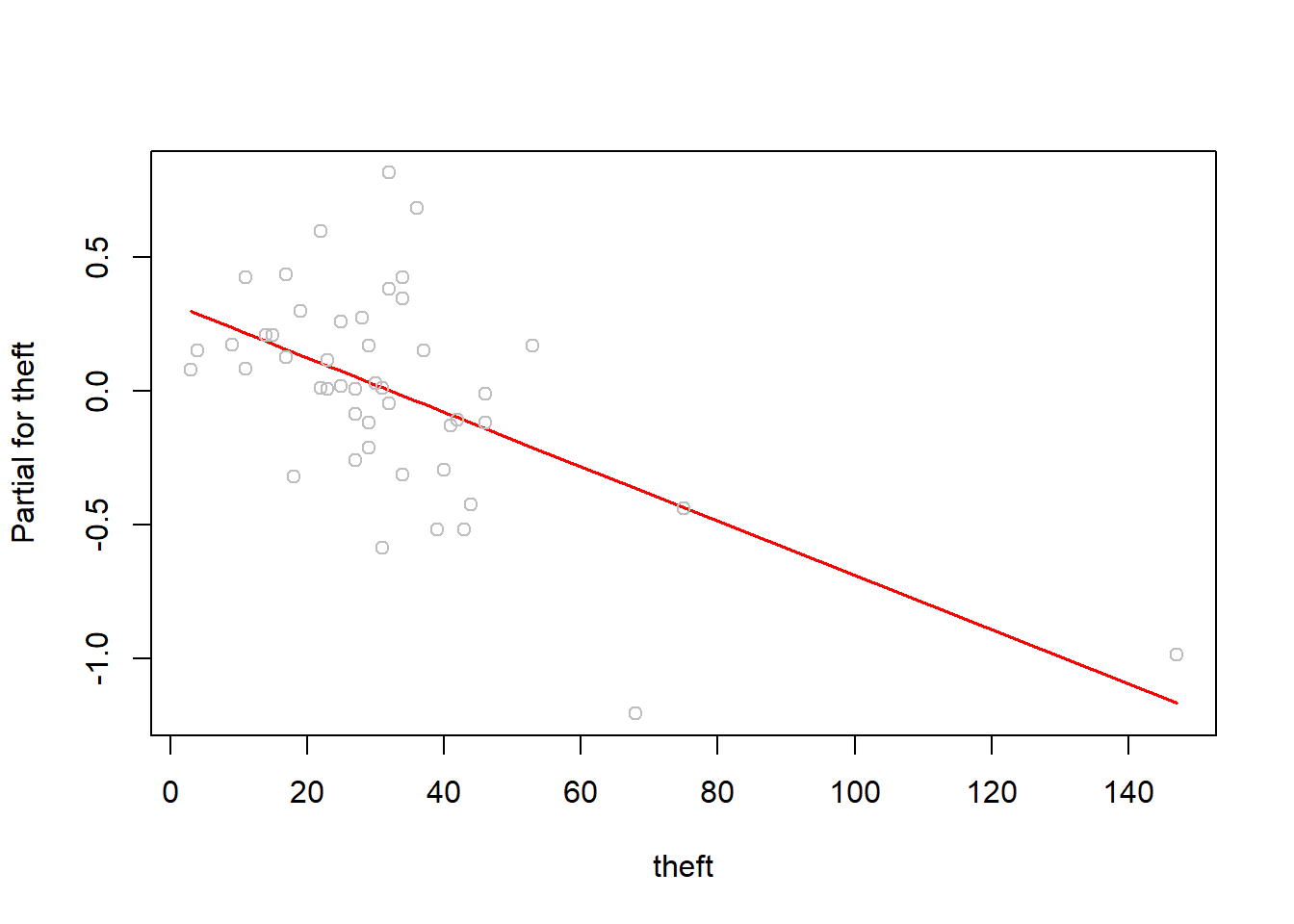
termplot(lmod\_full, partial.resid = T, terms=1)



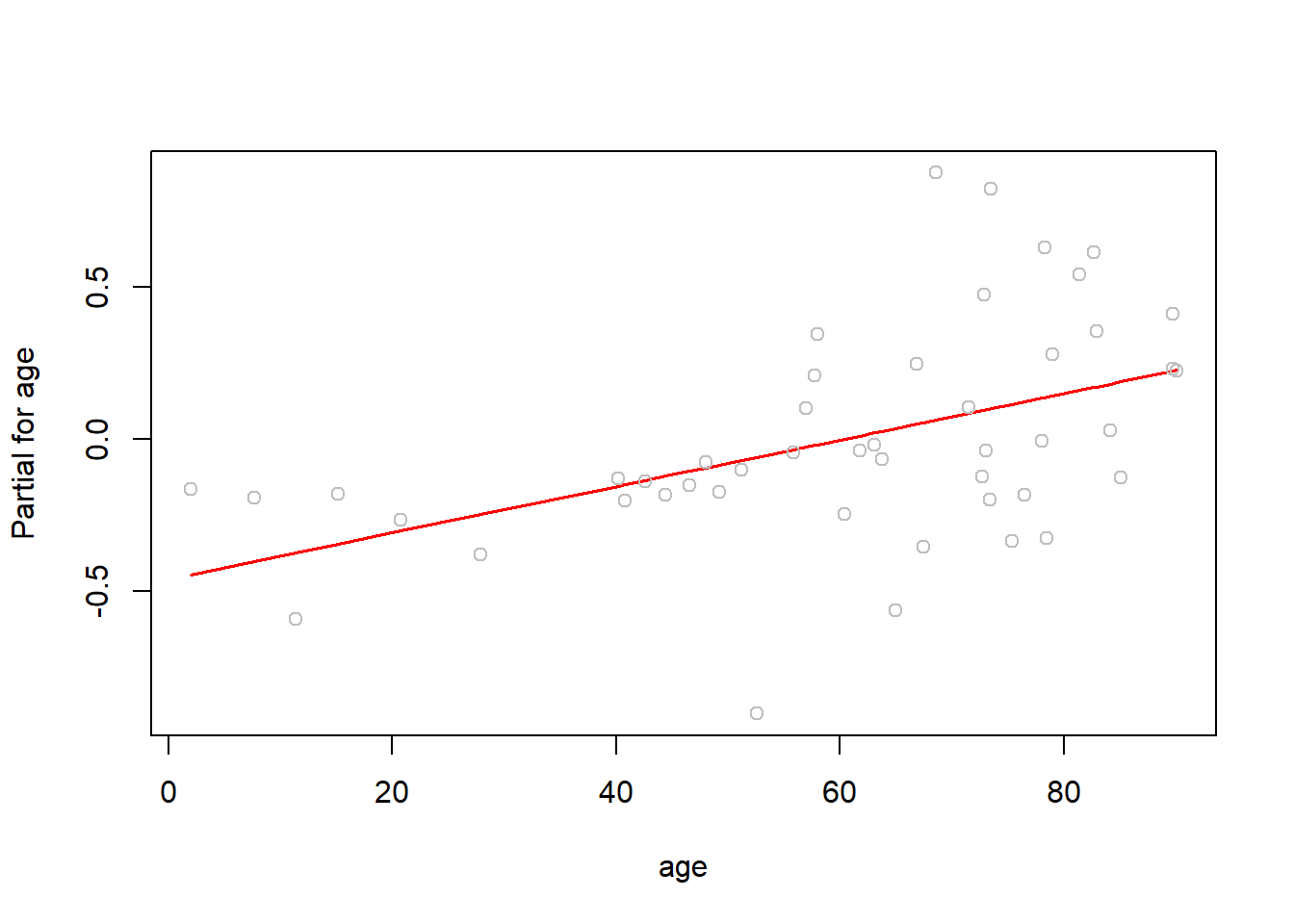
termplot(lmod\_full, partial.resid = T, terms=2)



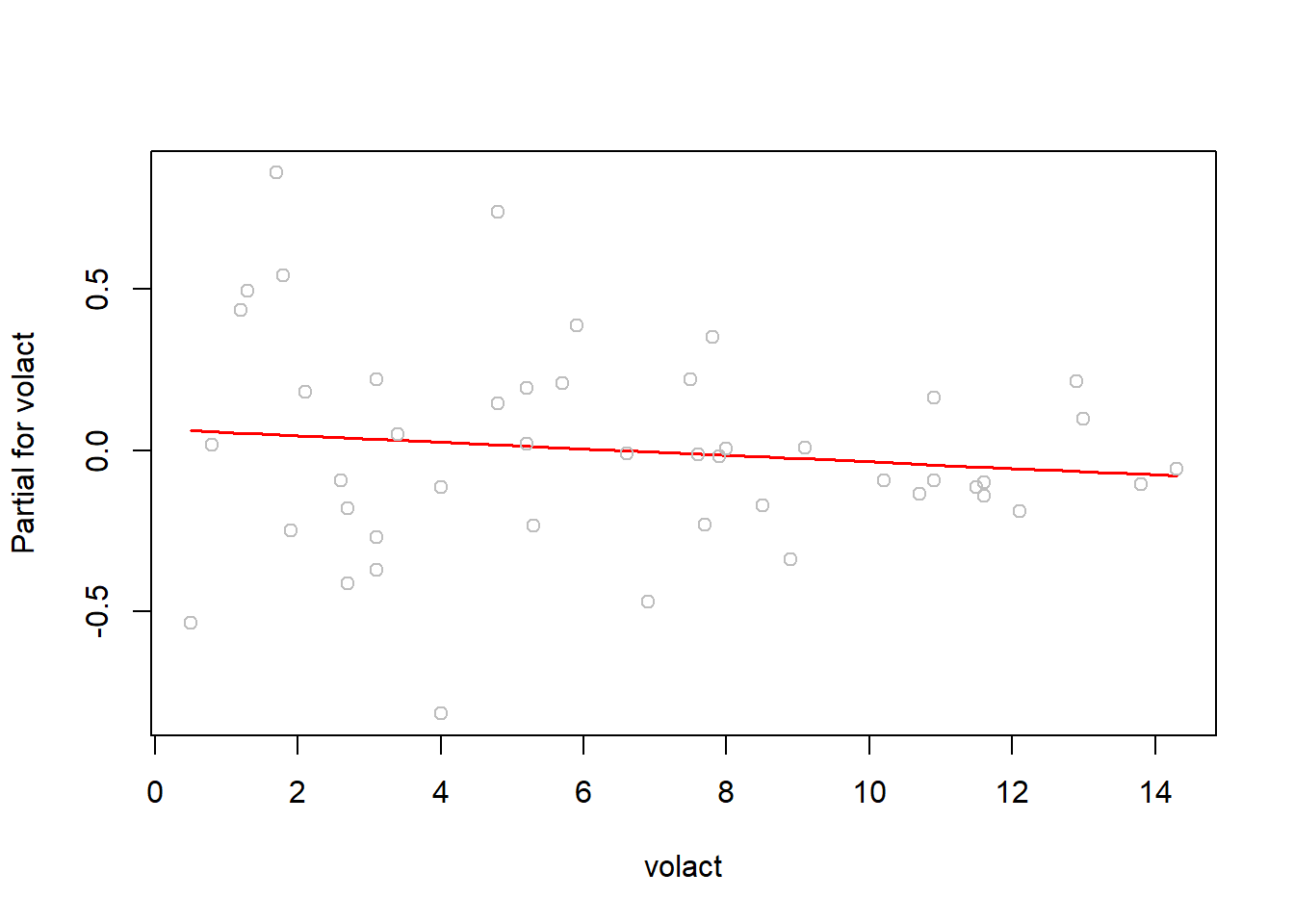
termplot(lmod\_full, partial.resid = T, terms=3)



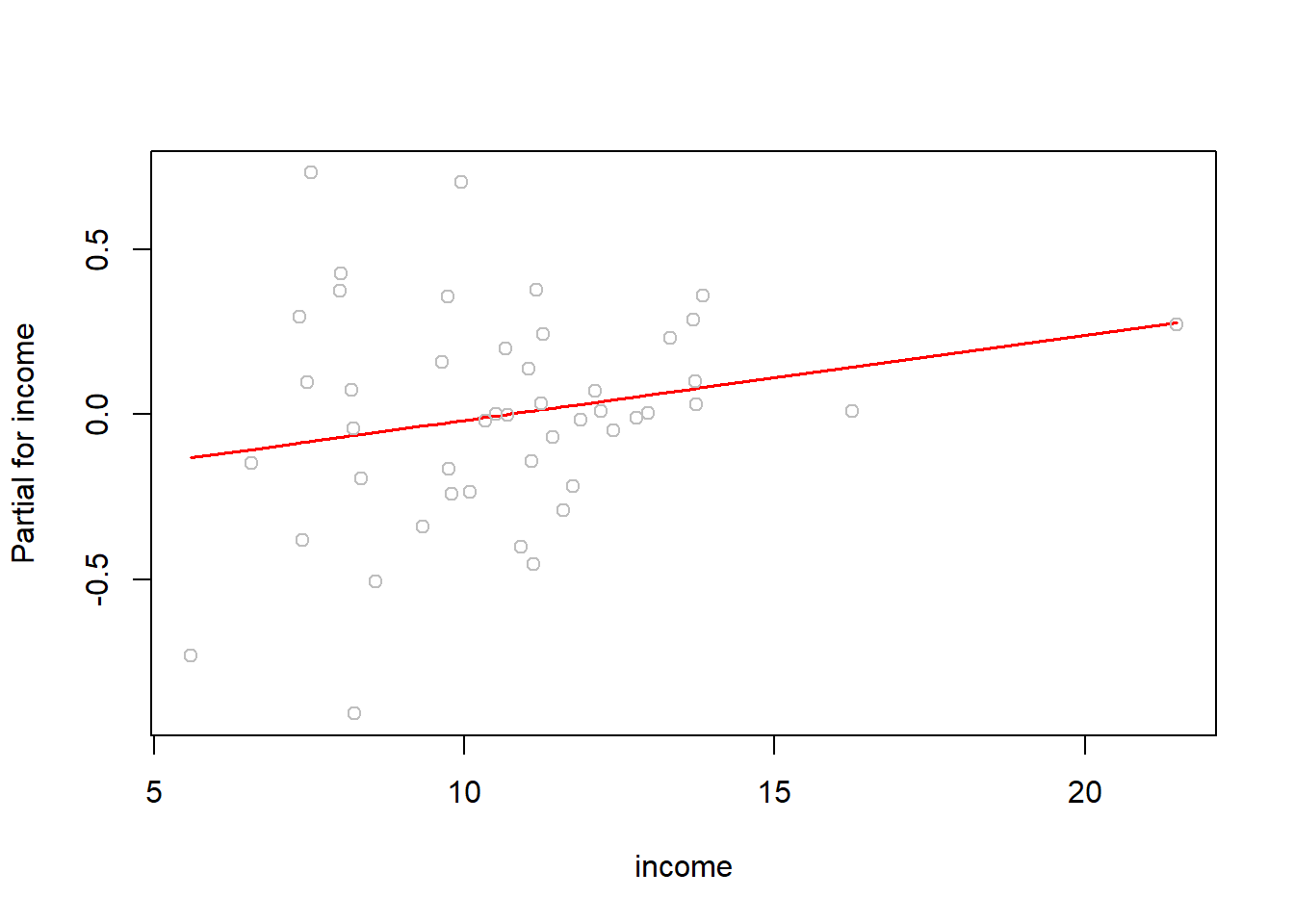
termplot(lmod\_full, partial.resid = T, terms=4)



termplot(lmod\_full, partial.resid = T, terms=5)



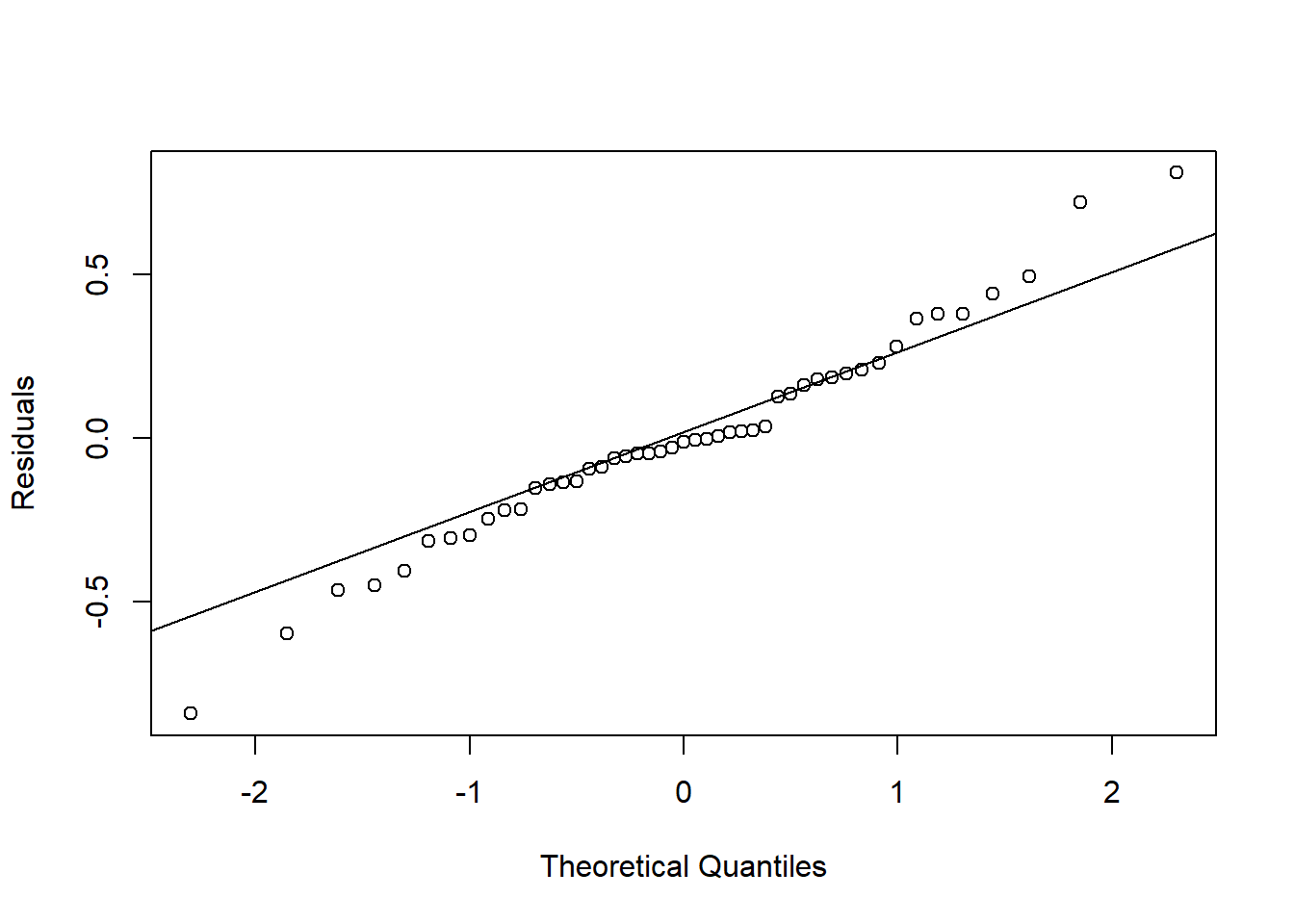
termplot(lmod\_full, partial.resid = T, terms=6)

  
Structure is linear as we checked for every predictor.

## Checking Normality

qqnorm(residuals(lmod\_full),ylab = "Residuals", main ="")

qqline(residuals(lmod\_full))

  
It looks we have fatter tails distribution.   
We use Shapiro Test for verification.

*#Checking using Shapiro Test*

shapiro.test(residuals(lmod\_full))

##

## Shapiro-Wilk normality test

##

## data: residuals(lmod\_full)

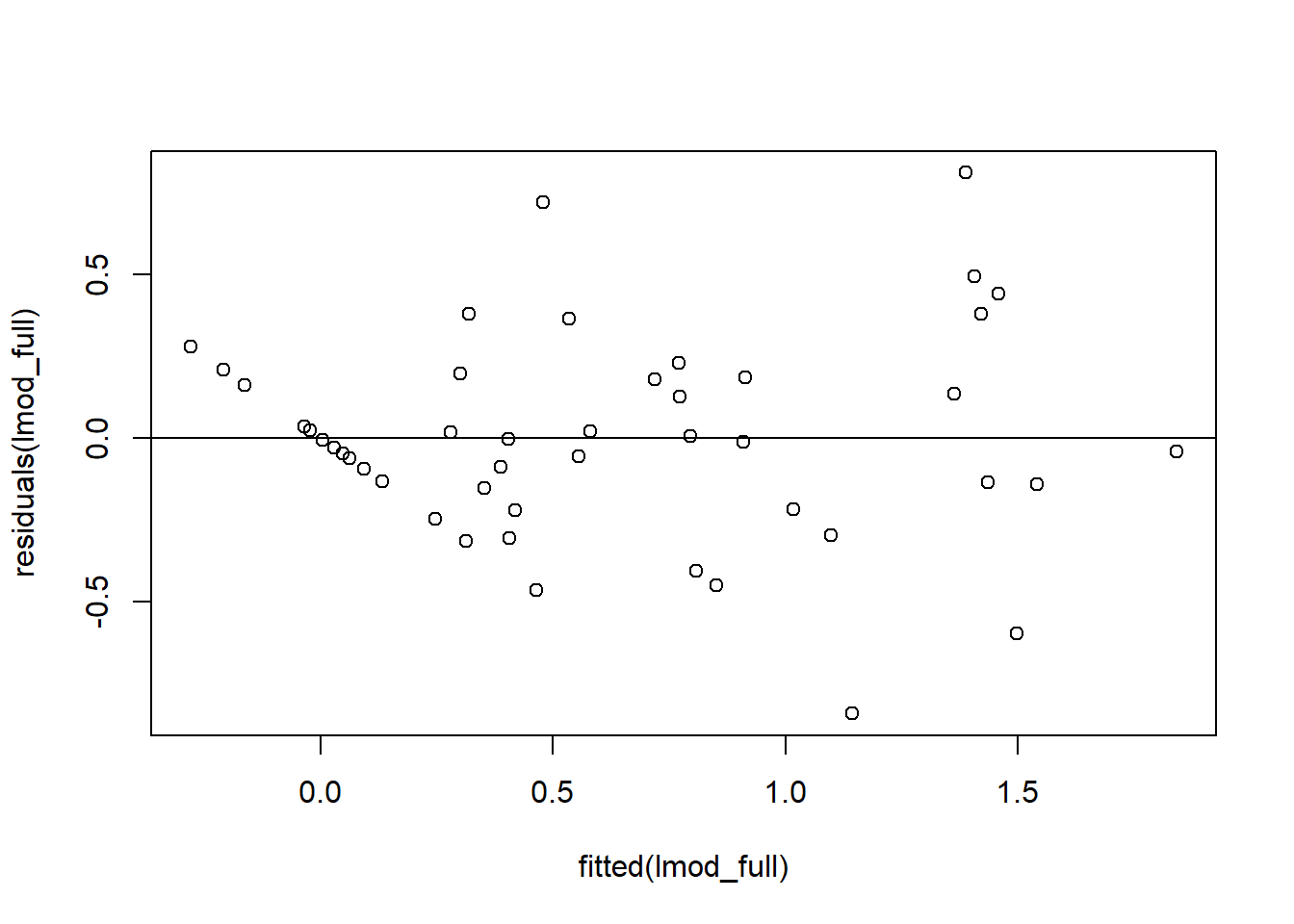
## W = 0.98095, p-value = 0.6317

P-value High, Accepting Null hypotheses. Dist. is Normal.

## Checking Error Variance

plot(fitted(lmod\_full),residuals(lmod\_full))

abline(h=0)

  
Looks constant variance with few anomalies.

## Checking Collinearity

X = model.matrix(lmod\_full)[,-1]

cor(X)

## race fire theft age volact income

## race 1.0000000 0.5927956 0.2550647 0.2505118 -0.7594196 -0.7037328

## fire 0.5927956 1.0000000 0.5562105 0.4122225 -0.6864766 -0.6104481

## theft 0.2550647 0.5562105 1.0000000 0.3176308 -0.3116183 -0.1729226

## age 0.2505118 0.4122225 0.3176308 1.0000000 -0.6057428 -0.5286695

## volact -0.7594196 -0.6864766 -0.3116183 -0.6057428 1.0000000 0.7509780

## income -0.7037328 -0.6104481 -0.1729226 -0.5286695 0.7509780 1.0000000

vif(X)

## race fire theft age volact income

## 3.491088 2.798840 1.684571 2.266203 4.851903 3.153110

Every Predictor is under 5. we are safe. (volact has relatively high correlation with other predictors  
Building model without volact:

lmod1\_without\_volcat = lm(involact ~ race + fire + theft + age + income, ch )

summary(lmod1\_without\_volcat)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age + income, data = ch)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.84428 -0.15804 -0.04093 0.18116 0.80828

##

## Coefficients:

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

## (Intercept) -0.608979 0.495260 -1.230 0.225851

## race 0.009133 0.002316 3.944 0.000307 \*\*\*

## fire 0.038817 0.008436 4.602 4e-05 \*\*\*

## theft -0.010298 0.002853 -3.610 0.000827 \*\*\*

## age 0.008271 0.002782 2.973 0.004914 \*\*

## income 0.024500 0.031697 0.773 0.443982

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3351 on 41 degrees of freedom

## Multiple R-squared: 0.7508, Adjusted R-squared: 0.7204

## F-statistic: 24.71 on 5 and 41 DF, p-value: 2.159e-11

Model without volcat performs better.  
  
Now, Checking this using anova().

anova(lmod1\_without\_volcat,lmod\_full)

## Analysis of Variance Table

##

## Model 1: involact ~ race + fire + theft + age + income

## Model 2: involact ~ race + fire + theft + age + volact + income

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 41 4.6047

## 2 40 4.5892 1 0.015457 0.1347 0.7155

H0: Beta(r) = 0   
H1: Beta(r) != 0  
  
High p-value accept Null Hypotheses.   
simple words: volcat is not significant.   
  
Checking Summary of the model

summary(lmod1\_without\_volcat)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age + income, data = ch)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.84428 -0.15804 -0.04093 0.18116 0.80828

##

## Coefficients:

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

## (Intercept) -0.608979 0.495260 -1.230 0.225851

## race 0.009133 0.002316 3.944 0.000307 \*\*\*

## fire 0.038817 0.008436 4.602 4e-05 \*\*\*

## theft -0.010298 0.002853 -3.610 0.000827 \*\*\*

## age 0.008271 0.002782 2.973 0.004914 \*\*

## income 0.024500 0.031697 0.773 0.443982

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3351 on 41 degrees of freedom

## Multiple R-squared: 0.7508, Adjusted R-squared: 0.7204

## F-statistic: 24.71 on 5 and 41 DF, p-value: 2.159e-11

P-value of income is high. means it insignificant.  
  
Let’s try removing income.

lmod2\_without\_volcat\_income = lm(involact ~ race + fire + theft + age, ch)

summary(lmod2\_without\_volcat\_income)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age, data = ch)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.87108 -0.14830 -0.01961 0.19968 0.81638

##

## Coefficients:

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

## (Intercept) -0.243118 0.145054 -1.676 0.101158

## race 0.008104 0.001886 4.297 0.000100 \*\*\*

## fire 0.036646 0.007916 4.629 3.51e-05 \*\*\*

## theft -0.009592 0.002690 -3.566 0.000921 \*\*\*

## age 0.007210 0.002408 2.994 0.004595 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3335 on 42 degrees of freedom

## Multiple R-squared: 0.7472, Adjusted R-squared: 0.7231

## F-statistic: 31.03 on 4 and 42 DF, p-value: 4.799e-12

Removing income does not make much of the difference  
  
Comparing models using anova():

anova(lmod2\_without\_volcat\_income,lmod1\_without\_volcat)

## Analysis of Variance Table

##

## Model 1: involact ~ race + fire + theft + age

## Model 2: involact ~ race + fire + theft + age + income

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 42 4.6718

## 2 41 4.6047 1 0.067101 0.5975 0.444

H0: Beta(r) = 0   
H1: Beta(r) != 0  
  
Significance level = 5%   
High p-value accept Null Hypothesis.   
simple words: income is not significant.

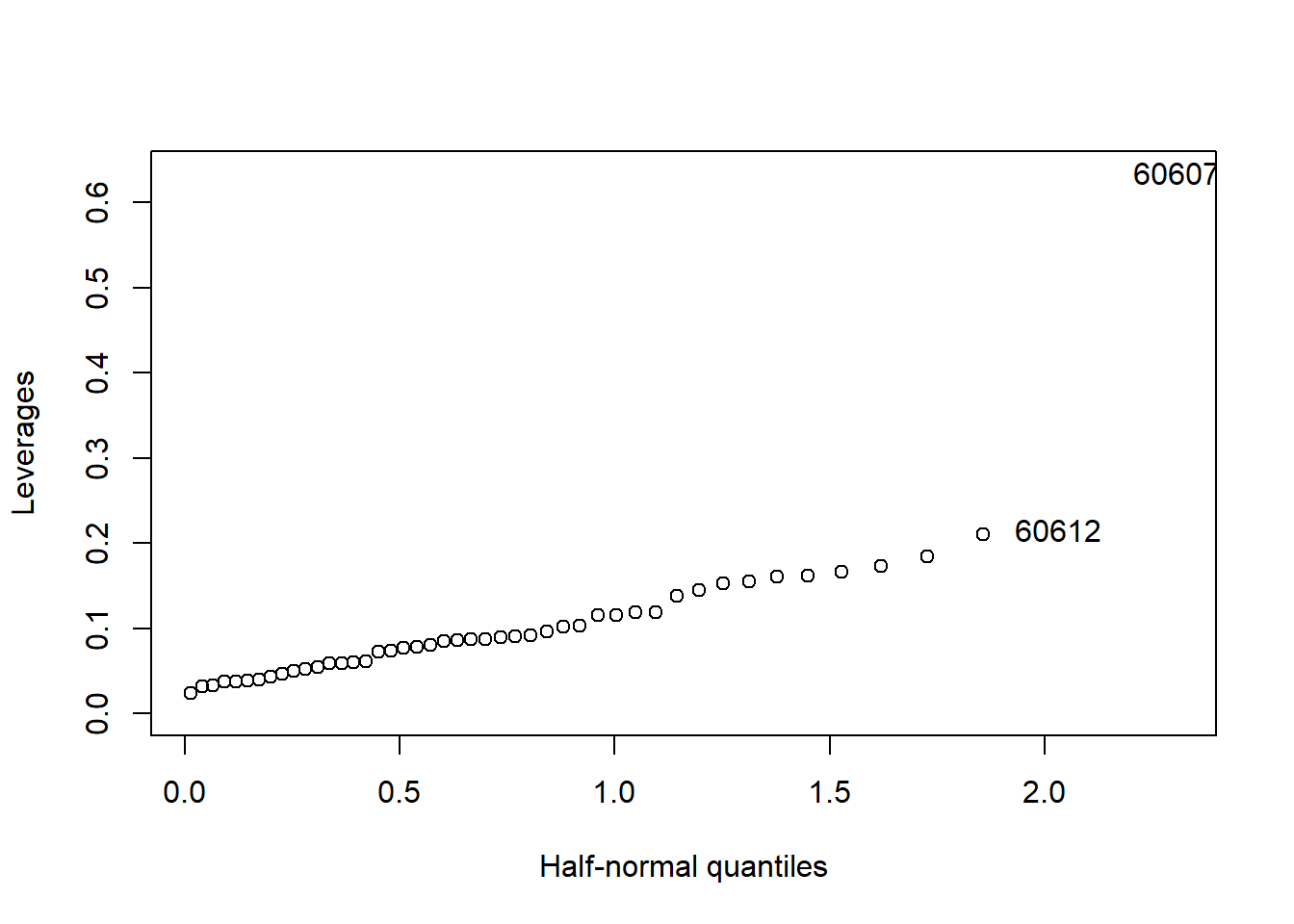
# Checking Unusual Observations

## Checking Leverage Points

zips = row.names(ch)

hat\_vals = hatvalues(lmod2\_without\_volcat\_income)

halfnorm(hat\_vals,labs = zips, ylab = "Leverages")

  
Zip Code: 60607 seems to be high leverage point.  
  
Checking this observation:

row1 = which(rownames(ch) == 60607)

ch[row1,]

## race fire theft age volact involact income

## 60607 50.2 39.7 147 83 5.2 0.9 7.459

We can observe high theft in this observation  
  
See what’s happens if we remove this observation.

lmod3\_modified\_1 = lm(involact ~ race + fire + theft + age ,ch[-row1])

summary(lmod3\_modified\_1)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age, data = ch[-row1])

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.87108 -0.14830 -0.01961 0.19968 0.81638

##

## Coefficients:

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

## (Intercept) -0.243118 0.145054 -1.676 0.101158

## race 0.008104 0.001886 4.297 0.000100 \*\*\*

## fire 0.036646 0.007916 4.629 3.51e-05 \*\*\*

## theft -0.009592 0.002690 -3.566 0.000921 \*\*\*

## age 0.007210 0.002408 2.994 0.004595 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3335 on 42 degrees of freedom

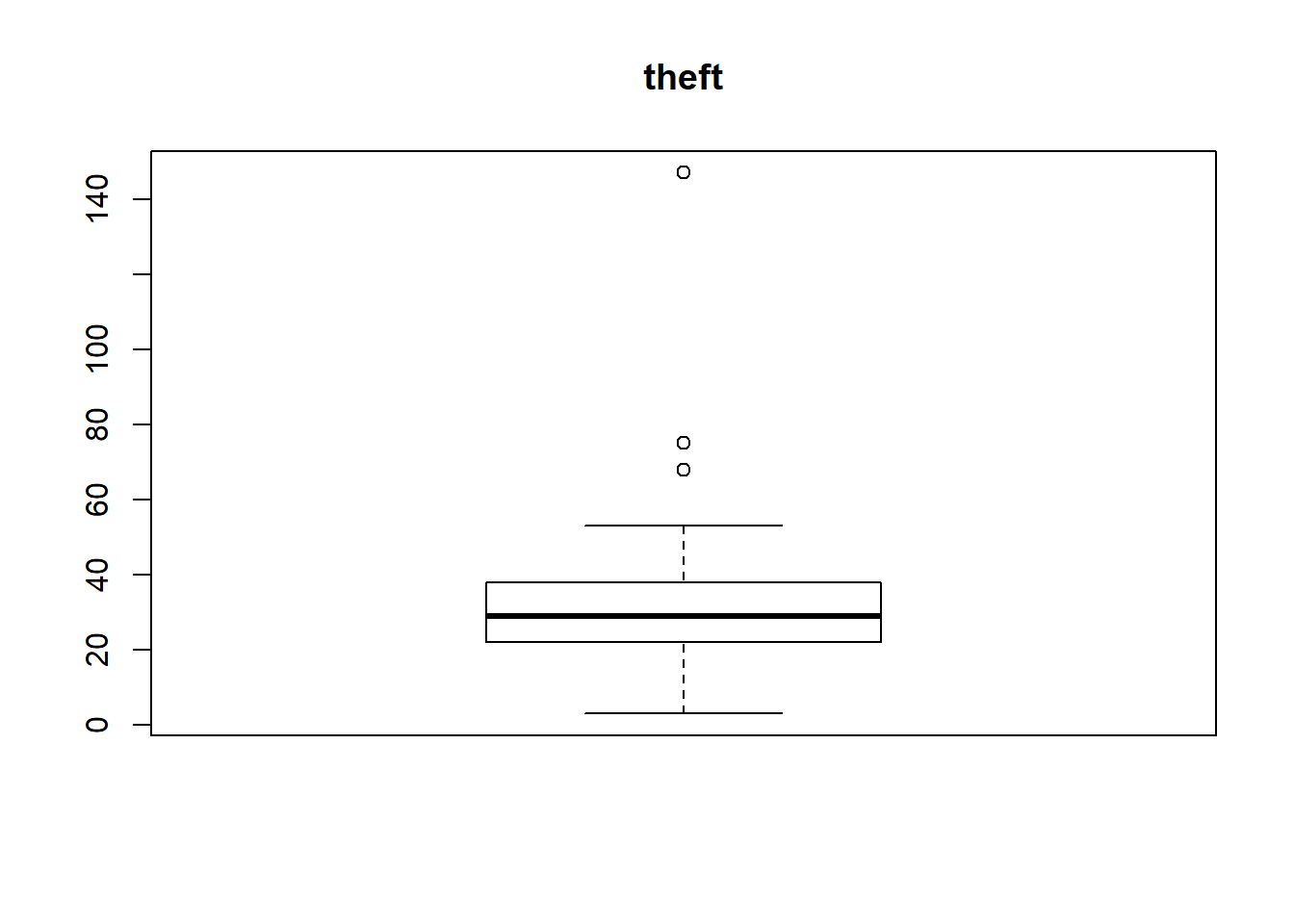
## Multiple R-squared: 0.7472, Adjusted R-squared: 0.7231

## F-statistic: 31.03 on 4 and 42 DF, p-value: 4.799e-12

No Effect. As it is not affecting the model let it be in the model but report the case.   
  
Further Investigation:

theft = 3

boxplot(ch[,theft],main=names(ch)[theft])

  
This observation theft value is far higher than other observation. it must be reported.

## Checking Outliers

sort(abs(residuals(lmod2\_without\_volcat\_income)))

## 60631 60619 60632 60616 60638 60618

## 0.001966931 0.004373166 0.007351602 0.013648913 0.018400932 0.019607322

## 60651 60643 60645 60612 60646 60635

## 0.021481962 0.036764563 0.038561267 0.043673115 0.069513114 0.079401224

## 60634 60629 60607 60636 60630 60609

## 0.085058456 0.090890292 0.093118206 0.104531778 0.110429605 0.118843923

## 60633 60655 60608 60620 60639 60644

## 0.126700197 0.131776158 0.148096306 0.160332787 0.169898411 0.186612598

## 60657 60656 60647 60611 60626 60628

## 0.193319018 0.206048996 0.207880464 0.212589033 0.222429899 0.238847285

## 60627 60652 60637 60640 60641 60614

## 0.250934003 0.255827264 0.308243446 0.314421019 0.323936345 0.350925819

## 60649 60617 60624 60615 60622 60625

## 0.358846125 0.360328508 0.385193257 0.451842984 0.457316717 0.460601200

## 60623 60653 60613 60621 60610

## 0.510361779 0.627657515 0.714891996 0.816376747 0.871077427

Following are the outlier observations

60653 60613 60621 60610   
0.990274659 1.127907196 1.288022823 1.374325778 

Now Let’s Try by removing them

*#Getting Outlier Rows*

rn = rownames(ch)

rows\_outliers = subset(ch, rn == 60610 | rn == 60621 | rn == 60613 | rn == 60653)

rows\_outliers

## race fire theft age volact involact income

## 60613 19.6 10.5 36 73.5 4.8 1.2 9.948

## 60610 54.0 34.1 68 52.6 4.0 0.3 8.231

## 60653 99.7 21.6 31 65.0 0.5 0.9 5.583

## 60621 98.9 17.4 32 68.6 1.7 2.2 7.520

*#buliding model with out them*

lmod2\_without\_volcat\_income\_outlier\_removed = lm(involact ~ race + fire + theft + age, data = ch[-c(60653,60613,60621,60610),])

summary(lmod2\_without\_volcat\_income)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age, data = ch)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.87108 -0.14830 -0.01961 0.19968 0.81638

##

## Coefficients:

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

## (Intercept) -0.243118 0.145054 -1.676 0.101158

## race 0.008104 0.001886 4.297 0.000100 \*\*\*

## fire 0.036646 0.007916 4.629 3.51e-05 \*\*\*

## theft -0.009592 0.002690 -3.566 0.000921 \*\*\*

## age 0.007210 0.002408 2.994 0.004595 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3335 on 42 degrees of freedom

## Multiple R-squared: 0.7472, Adjusted R-squared: 0.7231

## F-statistic: 31.03 on 4 and 42 DF, p-value: 4.799e-12

summary(lmod2\_without\_volcat\_income\_outlier\_removed)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age, data = ch[-c(60653,

## 60613, 60621, 60610), ])

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.87108 -0.14830 -0.01961 0.19968 0.81638

##

## Coefficients:

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

## (Intercept) -0.243118 0.145054 -1.676 0.101158

## race 0.008104 0.001886 4.297 0.000100 \*\*\*

## fire 0.036646 0.007916 4.629 3.51e-05 \*\*\*

## theft -0.009592 0.002690 -3.566 0.000921 \*\*\*

## age 0.007210 0.002408 2.994 0.004595 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3335 on 42 degrees of freedom

## Multiple R-squared: 0.7472, Adjusted R-squared: 0.7231

## F-statistic: 31.03 on 4 and 42 DF, p-value: 4.799e-12

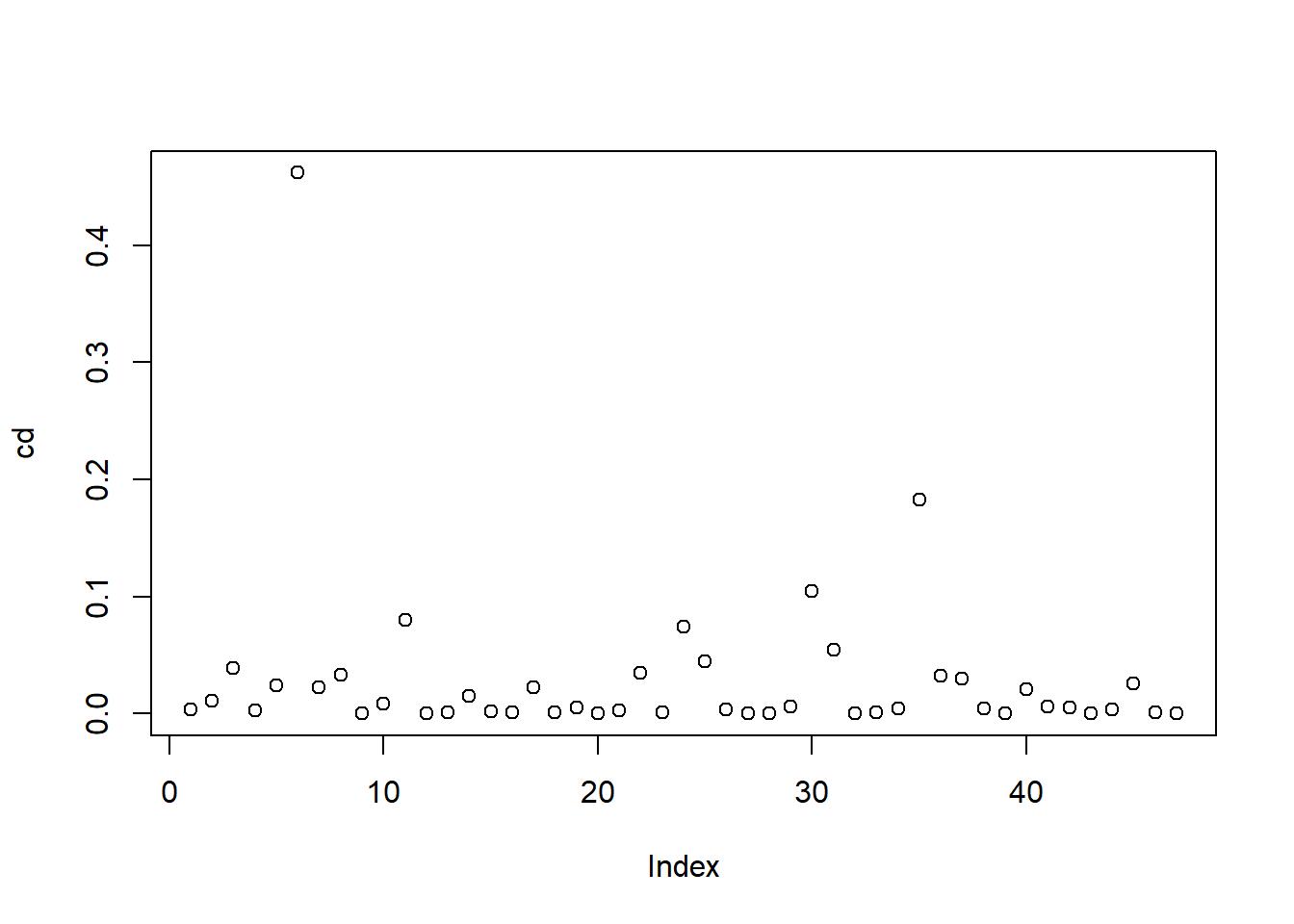
Does not make any difference in the result. Hence, we let them in the model.

## Checking Influential Observations

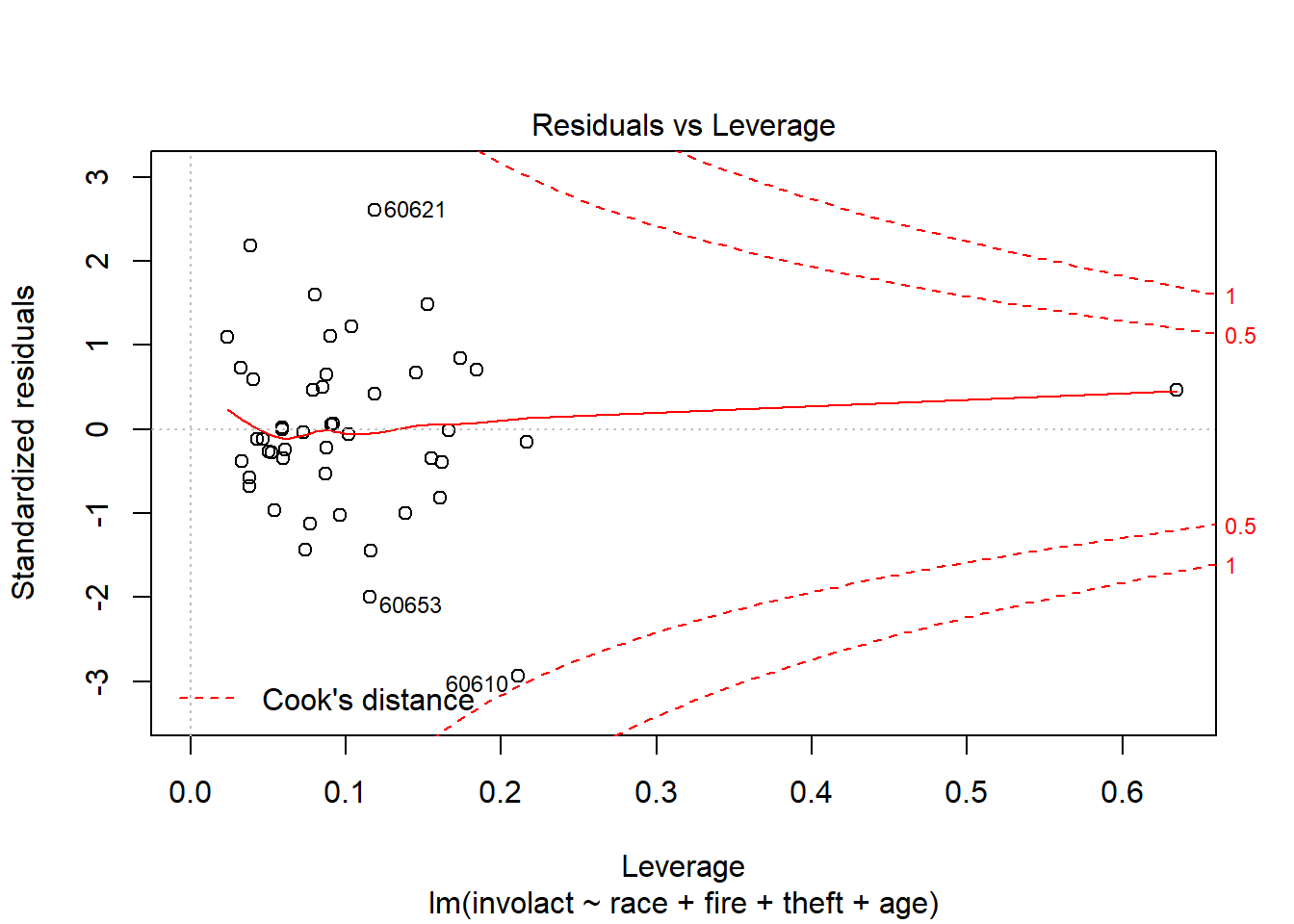
cd = cooks.distance(lmod2\_without\_volcat\_income)

plot(cd)

abline(h=0.5)

  
2nd method

plot(lmod2\_without\_volcat\_income)

  
No point is Over 0.5.  
No Influential Observations.

# Transformations

## Power Transformation

*# Reponses do not have strictly positive number*

unique(ch$involact)

## [1] 0.0 0.1 1.2 0.5 0.7 0.3 0.4 1.1 1.9 0.2 0.8 1.8 0.9 1.5 0.6 1.3 1.4 2.2 1.0

which(ch$involact == 0.0)

## [1] 1 7 8 12 13 14 15 16 17 18 32 33 37 42 47

*#scale response by adding* 10^-100

ch2 = ch

ch2$involact = ch$involact + (10^-100)

*#creating new model with scaled response*

lmod4\_without\_volcat\_income\_resposeScaledPositive = lm(involact ~ race + fire + theft + age, ch2)

summary(lmod2\_without\_volcat\_income)$r.squared

## [1] 0.7471912

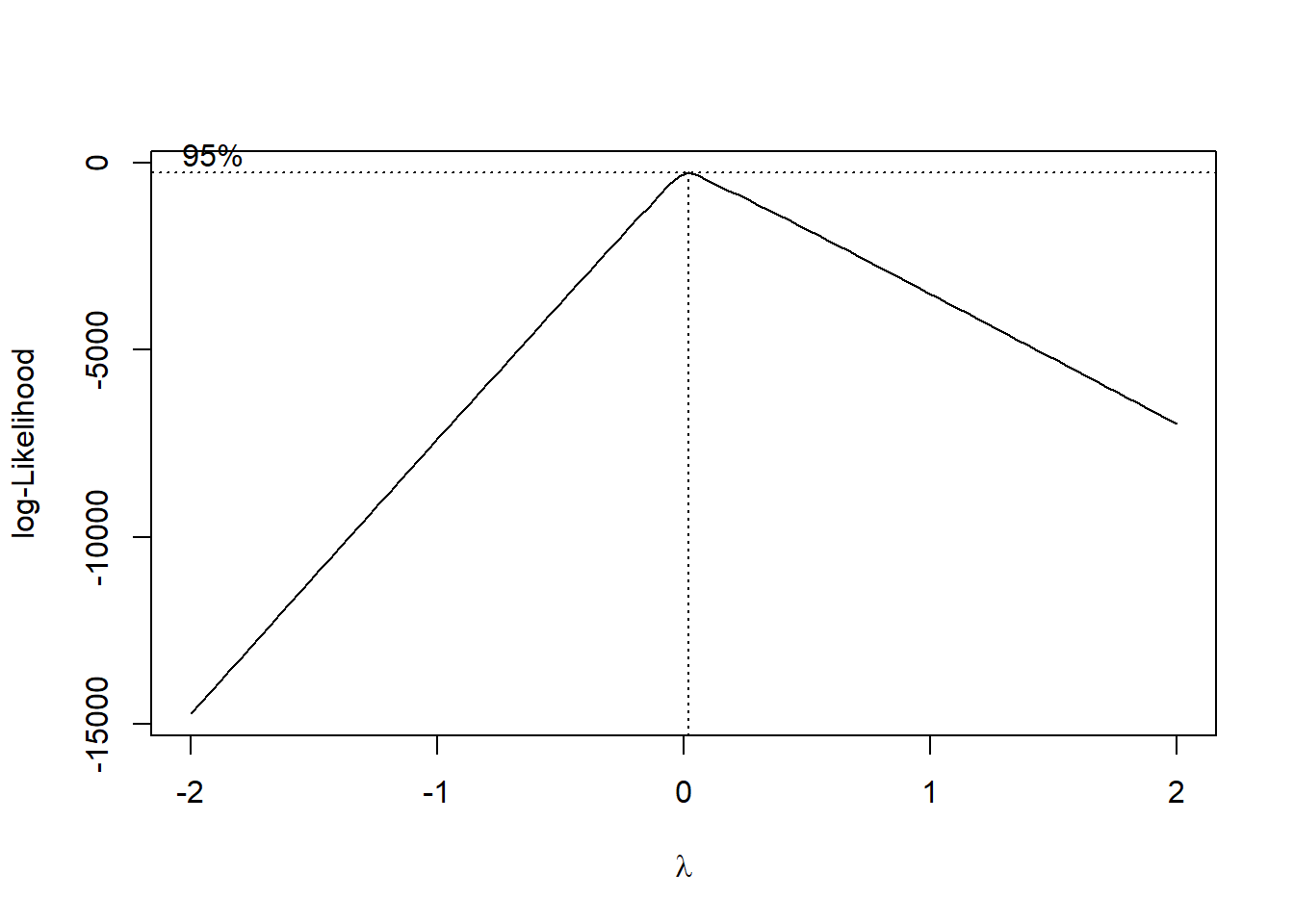
summary(lmod4\_without\_volcat\_income\_resposeScaledPositive)$r.squared

## [1] 0.7471912

*#doesnt not makes much of the difference*

*#Ploting boxcox*

bc = boxcox(lmod4\_without\_volcat\_income\_resposeScaledPositive, plotit=T)

  
Unable to interpret the diagram. Let’s try the transform directly if the model works better then fine else revert.

bc$x[which.max(bc$y)]

## [1] 0.02020202

Best possible Power transformation is ^0.0202

Applying Transformation:

lmod5\_without\_volcat\_income\_resposeScaledPositive\_powerT = lm(involact^0.0202 ~ race + fire + theft + age, ch2)

summary(lmod4\_without\_volcat\_income\_resposeScaledPositive)

##

## Call:

## lm(formula = involact ~ race + fire + theft + age, data = ch2)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.87108 -0.14830 -0.01961 0.19968 0.81638

##

## Coefficients:

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

## (Intercept) -0.243118 0.145054 -1.676 0.101158

## race 0.008104 0.001886 4.297 0.000100 \*\*\*

## fire 0.036646 0.007916 4.629 3.51e-05 \*\*\*

## theft -0.009592 0.002690 -3.566 0.000921 \*\*\*

## age 0.007210 0.002408 2.994 0.004595 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3335 on 42 degrees of freedom

## Multiple R-squared: 0.7472, Adjusted R-squared: 0.7231

## F-statistic: 31.03 on 4 and 42 DF, p-value: 4.799e-12

summary(lmod5\_without\_volcat\_income\_resposeScaledPositive\_powerT)

##

## Call:

## lm(formula = involact^0.0202 ~ race + fire + theft + age, data = ch2)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.5646 -0.2323 -0.0291 0.2607 0.5683

##

## Coefficients:

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

## (Intercept) -0.0787021 0.1312383 -0.600 0.55194

## race 0.0072822 0.0017063 4.268 0.00011 \*\*\*

## fire 0.0076946 0.0071624 1.074 0.28882

## theft -0.0006683 0.0024339 -0.275 0.78499

## age 0.0071348 0.0021785 3.275 0.00212 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3018 on 42 degrees of freedom

## Multiple R-squared: 0.6134, Adjusted R-squared: 0.5765

## F-statistic: 16.66 on 4 and 42 DF, p-value: 2.992e-08

Before Transformation R2 was 0.7472  
After Transformation R2 is 0.6134  
Model works better without Transformation. Hence continuing with model 2.

## Polynomials

Let’s try polynomial on “theft”:

*#Previous model*

summary(lmod2\_without\_volcat\_income)$r.squared

## [1] 0.7471912

summary(lmod2\_without\_volcat\_income)$adj.r.squared

## [1] 0.7231142

*#model with poly 2*

lmod7\_without\_volcat\_income\_poly2 = lm(involact ~ race + fire + poly(theft,2) + age, ch2)

summary(lmod7\_without\_volcat\_income\_poly2)$r.squared

## [1] 0.7472763

summary(lmod7\_without\_volcat\_income\_poly2)$adj.r.squared

## [1] 0.7164564

*#model with poly 3*

lmod8\_without\_volcat\_income\_poly3 = lm(involact ~ race + fire + poly(theft,3) + age, ch2)

summary(lmod8\_without\_volcat\_income\_poly3)$r.squared

## [1] 0.76032

summary(lmod8\_without\_volcat\_income\_poly3)$adj.r.squared

## [1] 0.724368

*#model with poly 4*

lmod9\_without\_volcat\_income\_poly4 = lm(involact ~ race + fire + poly(theft,4) + age, ch2)

summary(lmod9\_without\_volcat\_income\_poly4)$r.squared

## [1] 0.7606418

summary(lmod9\_without\_volcat\_income\_poly4)$adj.r.squared

## [1] 0.71768

3rd polynomial of theft makes model better. Hence, continue with model 8.

# Evaluating the model

## Train and Test data

*##Test\_data*

*#Selecting Few random rows from data*

test\_data = ch2[sample(nrow(ch2), 5), ]

options(scipen = 999) *#disabling Scintific notation*

round(test\_data,2)

## race fire theft age volact involact income

## 60630 1.6 2.5 22 63.8 10.7 0.0 12.40

## 60609 46.2 21.8 4 73.1 2.6 1.3 8.33

## 60643 42.5 10.4 25 40.8 10.2 0.5 12.96

## 60607 50.2 39.7 147 83.0 5.2 0.9 7.46

## 60639 2.5 7.2 29 84.2 8.5 0.2 11.08

*##Train\_data*

rows= as.numeric(row.names(test\_data))

train\_data = ch2[-rows,]

head(round(train\_data,2))

## race fire theft age volact involact income

## 60626 10.0 6.2 29 60.4 5.3 0.0 11.74

## 60640 22.2 9.5 44 76.5 3.1 0.1 9.32

## 60613 19.6 10.5 36 73.5 4.8 1.2 9.95

## 60657 17.3 7.7 37 66.9 5.7 0.5 10.66

## 60614 24.5 8.6 53 81.4 5.9 0.7 9.73

## 60610 54.0 34.1 68 52.6 4.0 0.3 8.23

## Training the model

*#Training the model*

lmod\_final = lm(involact ~ race + fire + poly(theft,3) + age, data = train\_data)

summary(lmod\_final)

##

## Call:

## lm(formula = involact ~ race + fire + poly(theft, 3) + age, data = train\_data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.71209 -0.16182 -0.01792 0.17004 0.79694

##

## Coefficients:

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

## (Intercept) -0.450491 0.174390 -2.583 0.013550 \*

## race 0.007628 0.002036 3.747 0.000564 \*\*\*

## fire 0.039080 0.008425 4.639 0.0000372 \*\*\*

## poly(theft, 3)1 -1.415081 0.406476 -3.481 0.001221 \*\*

## poly(theft, 3)2 -0.074267 0.380385 -0.195 0.846193

## poly(theft, 3)3 0.562725 0.381402 1.475 0.147931

## age 0.005282 0.002882 1.833 0.074258 .

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3328 on 40 degrees of freedom

## Multiple R-squared: 0.7603, Adjusted R-squared: 0.7244

## F-statistic: 21.15 on 6 and 40 DF, p-value: 0.00000000005386

## Testing the model

*#Testing the model*

test\_data\_with\_removed\_varibles = test\_data[,-c(5,6,7)]

predict(lmod\_final,test\_data\_with\_removed\_varibles )

## 60630 60609 60643 60607 60639

## 0.1252968 1.2089165 0.6096953 0.9044523 0.3797168

round(test\_data,2)

## race fire theft age volact involact income

## 60630 1.6 2.5 22 63.8 10.7 0.0 12.40

## 60609 46.2 21.8 4 73.1 2.6 1.3 8.33

## 60643 42.5 10.4 25 40.8 10.2 0.5 12.96

## 60607 50.2 39.7 147 83.0 5.2 0.9 7.46

## 60639 2.5 7.2 29 84.2 8.5 0.2 11.08

Predicted values are almost same as the actual values. Hence, Our Model Performs Well.

# References

Jackson, Kenneth T. Crabgrass Frontier: The Suburbanization of the United States. New York: Oxford UP, 1985. Print.

Squires, Gregory D. “Racial Profiling, Insurance Style: Insurance Redlining and the Uneven Development of Metropolitan Areas.” Journal of Urban Affairs 25.4 (2003): 391-410. Print.

Badger, Emily. “Redlining: Still a thing.” The Washington Post. WP Company, 28 May 2015. Web. 19 Dec. 2016.

Tootell, G. M. B. “Redlining in Boston: Do Mortgage Lenders Discriminate Against Neighborhoods?” The Quarterly Journal of Economics 111.4 (1996): 1049-079. Print.