STAT 420: Data Analysis Project

A survey of economic mobility across generations in contemporary USA

Submitted By: Pragya Mishra (pragyam3)

Submitted On: 08/01/2019

Introduction

A survey of economic mobility across generations in contemporary USA.

The data come from a large study, based on tax records, which allowed researchers to link the income of adults to the income of their parents several decades previously. For privacy reasons, I don't have that individual-level data, but I do have aggregate statistics about economic mobility for several hundred communities, containing most of the American population, and covariate information about those communities, containing most of the American population, and covariate information about those communities. I am interested in predicting economic mobility from the characteristics of communities.

Inspiration

Inspired by homework assignment from CMU class 36-401 (*Modern Regression*) and 36-402 (*Undergraduate Advanced Data Analysis*).

Motivation

1.Hands on experince with real life datasets. 2.Practise with all techniques learnt in STAT420. 3.Discover how applied statistics can help us answer socio-economic questions.

Dataset

A snippet. (Only first few columns).

```
# Data
mobility <- read.csv("mobility.csv")
mobility$Urban = as.factor(mobility$Urban)
mobilityData = mobility[complete.cases(mobility),]
attach(mobilityData)
knitr::kable(head(mobility)[,1:10])
```

ID	Name	Mobility	State	Population	Urban	Black	Seg_racial	Seg_income	Seg_poverty
100	Johnson City	0.0622	TN	576081	1	0.021	0.090	0.035	0.030
200	Morristown	0.0537	TN	227816	1	0.020	0.093	0.026	0.028
301	Middlesborough	0.0726	TN	66708	0	0.015	0.064	0.024	0.015
302	Knoxville	0.0563	TN	727600	1	0.056	0.210	0.092	0.084
401	Winston-Salem	0.0448	NC	493180	1	0.174	0.262	0.072	0.061
402	Martinsville	0.0518	VA	92753	0	0.224	0.137	0.024	0.015

Description

The data file `mobility.csv` has information on 741 communities. The variable I want to predict is economic mobility; the rest are predictor variables or covariates.

1. Mobility: The probability that a child born in 1980???1982 into the lowest quin- tile (20%) of household income will be in the top quintile at age 30. Individuals are assigned to the community they grew up in, not the one they were in as adults.
2. Population in 2000.
3. Is the community primarily urban or rural?
4. Black: percentage of individuals who marked black (and nothing else) on cen- sus forms.
5. Racial segregation: a measure of residential segregation by race.
6. Income segregation: Similarly but for income.
7. Segregation of poverty: Specifically a measure of residential segregation for those in the bottom quarter of the national income distribution.
8. Segregation of affluence: Residential segregation for those in the top qarter.
9. Commute: Fraction of workers with a commute of less than 15 minutes.
10. Mean income: Average income per capita in 2000.
11. Gini: A measure of income inequality, which would be 0 if all incomes were perfectly equal, and tends towards 100 as all the income is concentrated among the richest individuals (see Wikipedia, s.v. ??? Gini coefficient???).
12. Share 1%: Share of the total income of a community going to its richest 1%.
13. Gini bottom 99%: Gini coefficient among the lower 99% of that community.
14. Fraction middle class: Fraction of parents whose income is between the na- tional 25th and 75th percentiles.
15. Local tax rate: Fraction of all income going to local taxes.
16. Local government spending: per capita.
17. Progressivity: Measure of how much state income tax rates increase with in- come.

18. EITC: Measure of how much the state contributed to the Earned Income Tax Credit (a sort of negative income tax for very low-paid wage earners).
19. School expenditures: Average spending per pupil in public schools.
20. Student/teacher ratio: Number of students in public schools divided by num- ber of teachers.
21. Test scores: Residuals from a linear regression of mean math and English test scores on household income per capita.
22. Highschooldropoutrate:Also,residualsfromalinearregressionofthedropout rate on per-capita income.
23. Colleges per capita
24. College tuition: in-state, for full-time students
25. College graduation rate: Again, residuals from a linear regression of the actual graduation rate on household income per capita.
26. Labor force participation: Fraction of adults in the workforce.
27. Manufacturing: Fraction of workers in manufacturing.
28. Chinese imports: Growth rate in imports from China per worker between 1990 and 2000.
29. Teenage labor: fraction of those age 14???16 who were in the labor force.
30. Migration in: Migration into the community from elsewhere, as a fraction of 2000 population.
31. Migration out: Ditto for migration into other communities.
32. Foreign: fraction of residents born outside the US.
33. Social capital: Index combining voter turnout, participation in the census, and participation in community organizations.
34. Religious: Share of the population claiming to belong to an organized religious body.
35. Violent crime: Arrests per person per year for violent crimes.
36. Singlemotherhood:Numberofsinglefemalehouseholdswithchildrendivided by the total number of households with children.
37. Divorced: Fraction of adults who are divorced.
38. Married: Ditto.
39. Longitude: Geographic coordinate for the center of the community
40. Latitude: Ditto
41. ID: A numerical code, identifying the community.
42. Name: the name of principal city or town.
43. State: the state of the principal city or town of the community.

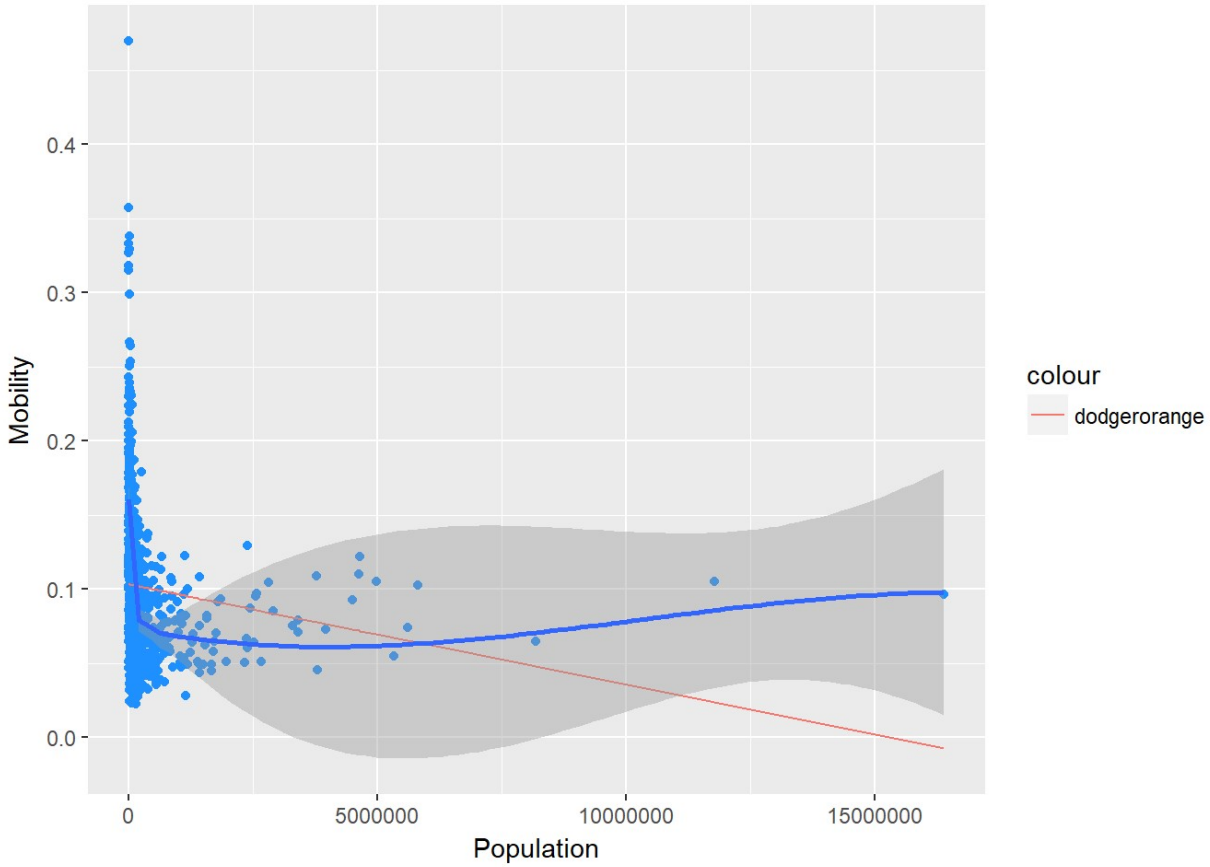
Methods

Exploratory Data Analysis

Scatterplots

Population

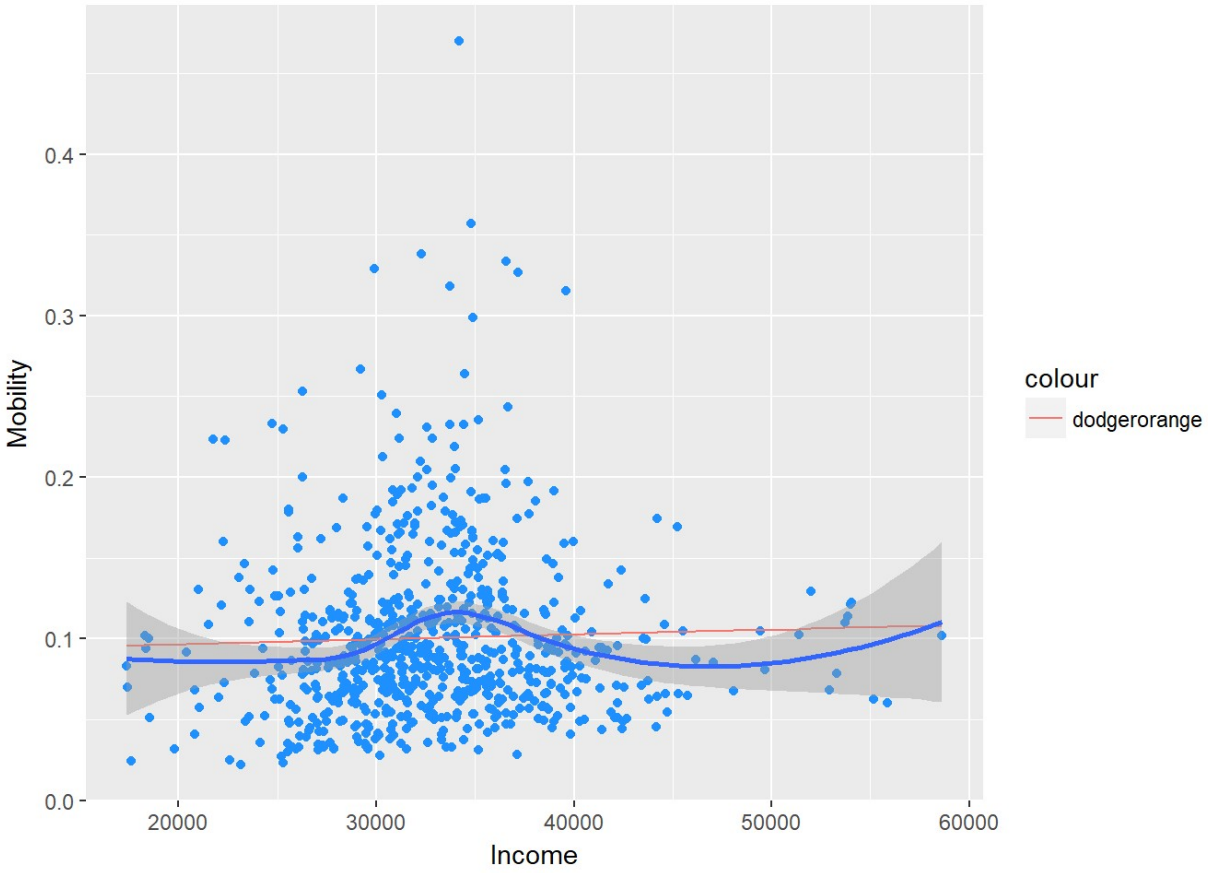
```
popMobilityData = as.data.frame(cbind(mobility$Population,mobility$Mobility))
popMobilityData = popMobilityData[complete.cases(popMobilityData),]
colnames(popMobilityData) = c("Population","Mobility")
pred.Population <- predict(lm(Mobility ~ Population, data = popMobilityData))


p1 <- ggplot(popMobilityData, aes(x = Population, y = Mobility))

p1 + geom_point(col="dodgerblue") + geom_line(aes(y = pred.Population,col="dodgerorang
e"))+ geom_smooth()
```



There's relatively much higher variance in Mobility of regions with less population and they are also regions of highest Mobility.

Mean household income per capita

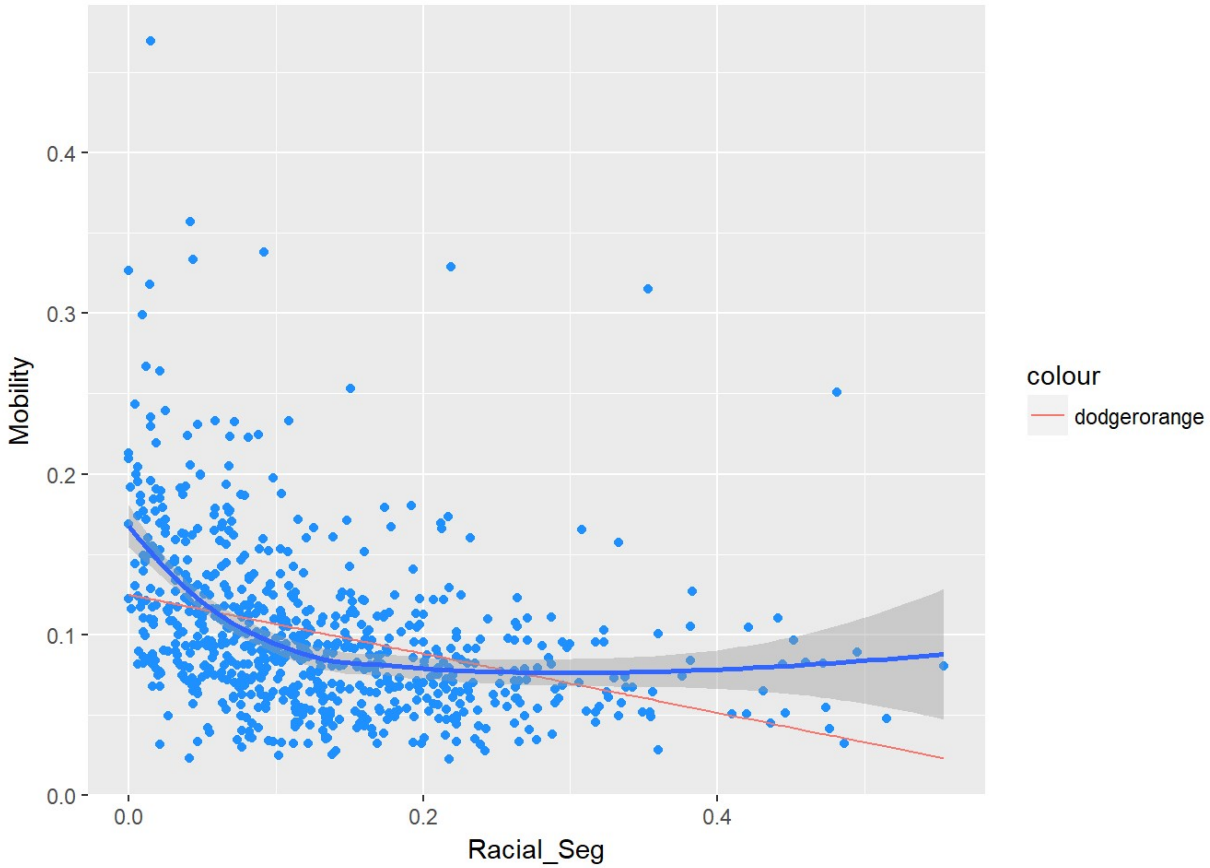
```
incomeData = as.data.frame(cbind(mobility$Income,mobility$Mobility))
incomeData = incomeData[complete.cases(incomeData),]
colnames(incomeData) = c("Income","Mobility")
pred.Inc <- predict(lm(Mobility ~ Income, data = incomeData))


p1 <- ggplot(incomeData, aes(x = Income, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Inc,col="dodgerorange"))+geom_smooth()
```



Not a clear relationship; middle income groups have highest range of Mobility.

Racial segregation

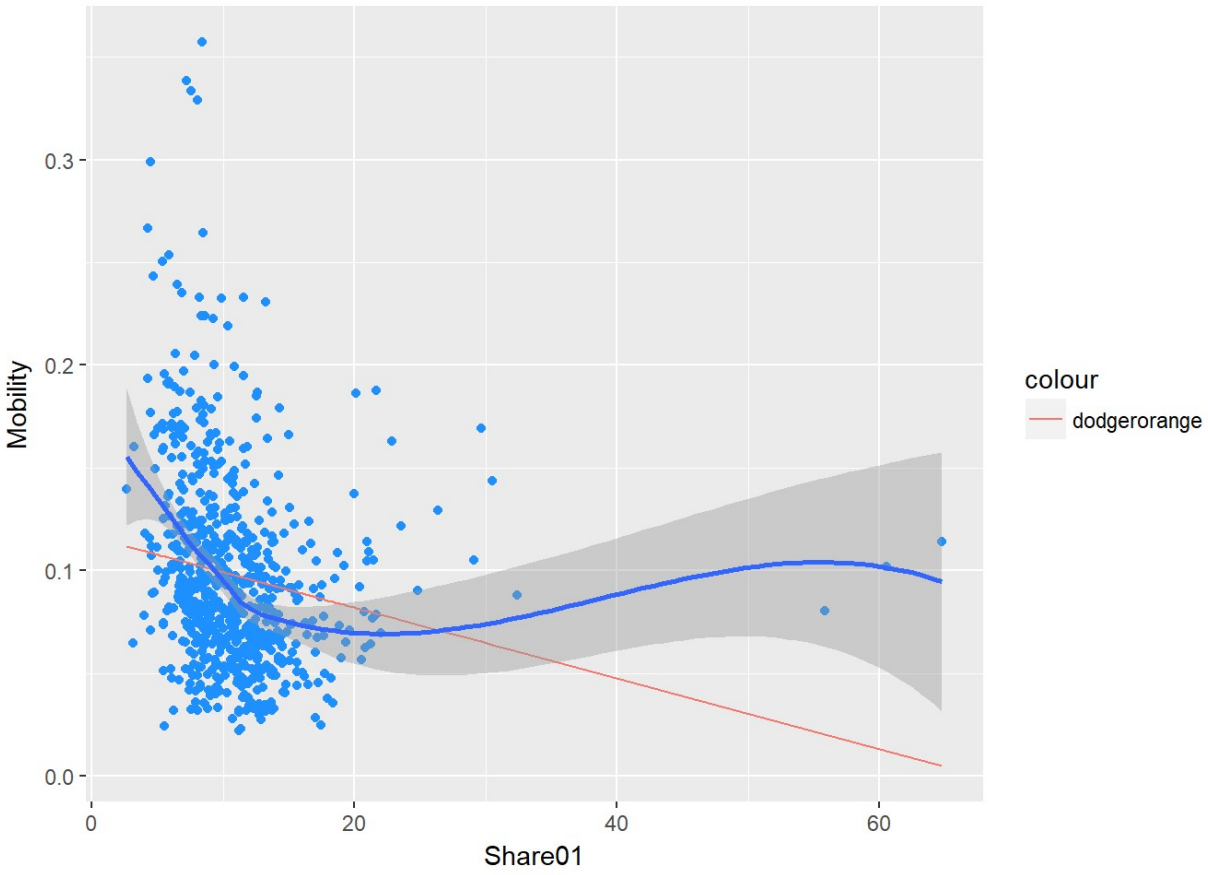
```
raceData = as.data.frame(cbind(mobility$Seg_racial,mobility$Mobility))
raceData = raceData[complete.cases(raceData),]
colnames(raceData) = c("Racial_Seg","Mobility")
pred.Race <- predict(lm(Mobility ~ Racial_Seg, data = raceData))


p1 <- ggplot(raceData, aes(x = Racial_Seg, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Race,col="dodgerorange"))+geom_smooth()
```



Areas with less racial segregation see higher range of Mobility.

Income share of the top 1%

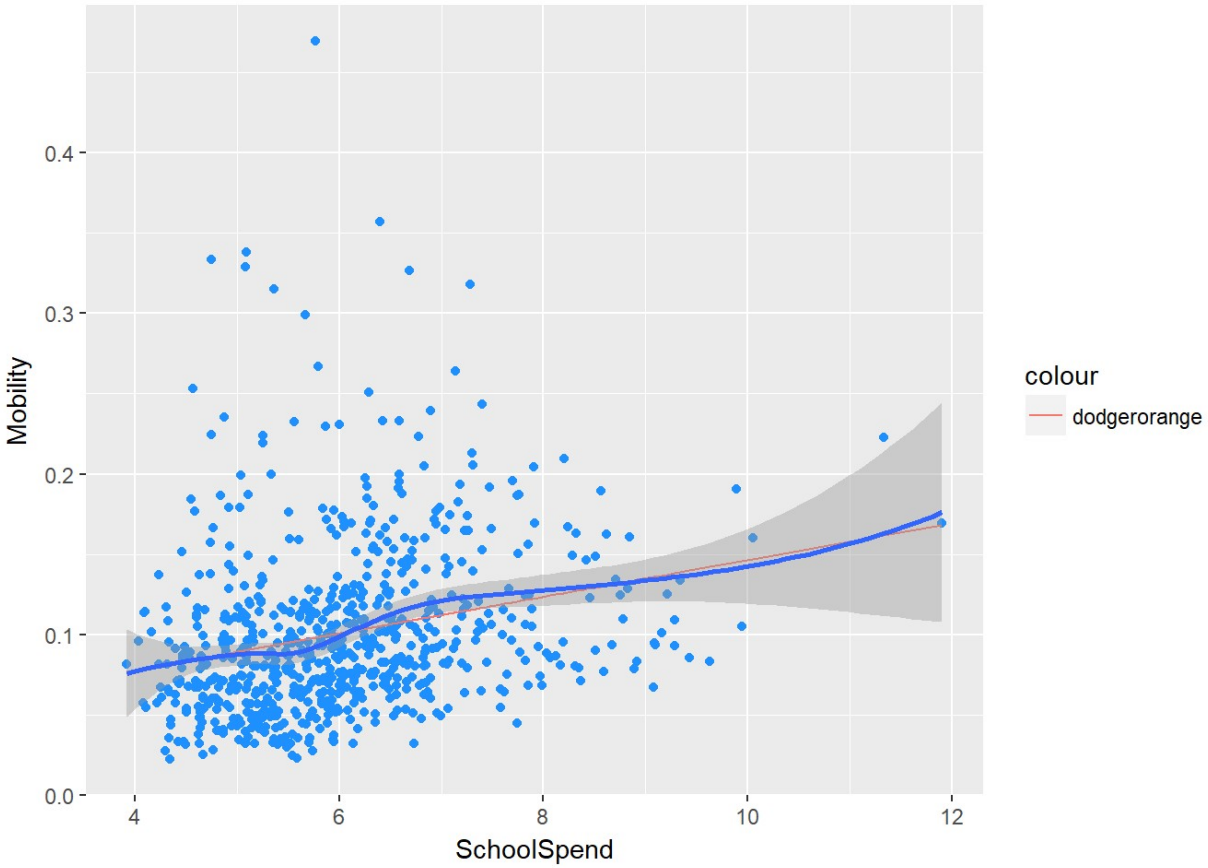
```
incomeData = as.data.frame(cbind(mobility$Share01,mobility$Mobility))
incomeData = incomeData[complete.cases(incomeData),]   # complete.cases: Return a logi
cal vector indicating which cases are complete, i.e., have no missing values.
colnames(incomeData) = c("Share01","Mobility")
pred.Inc <- predict(lm(Mobility ~ Share01, data = incomeData))


p1 <- ggplot(incomeData, aes(x = Share01, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Inc,col="dodgerorange"))+geom_smooth()
```



An interesting relationship. Areas with less share in top 1% see a higher range of Mobility.

Mean school expenditures per pupil

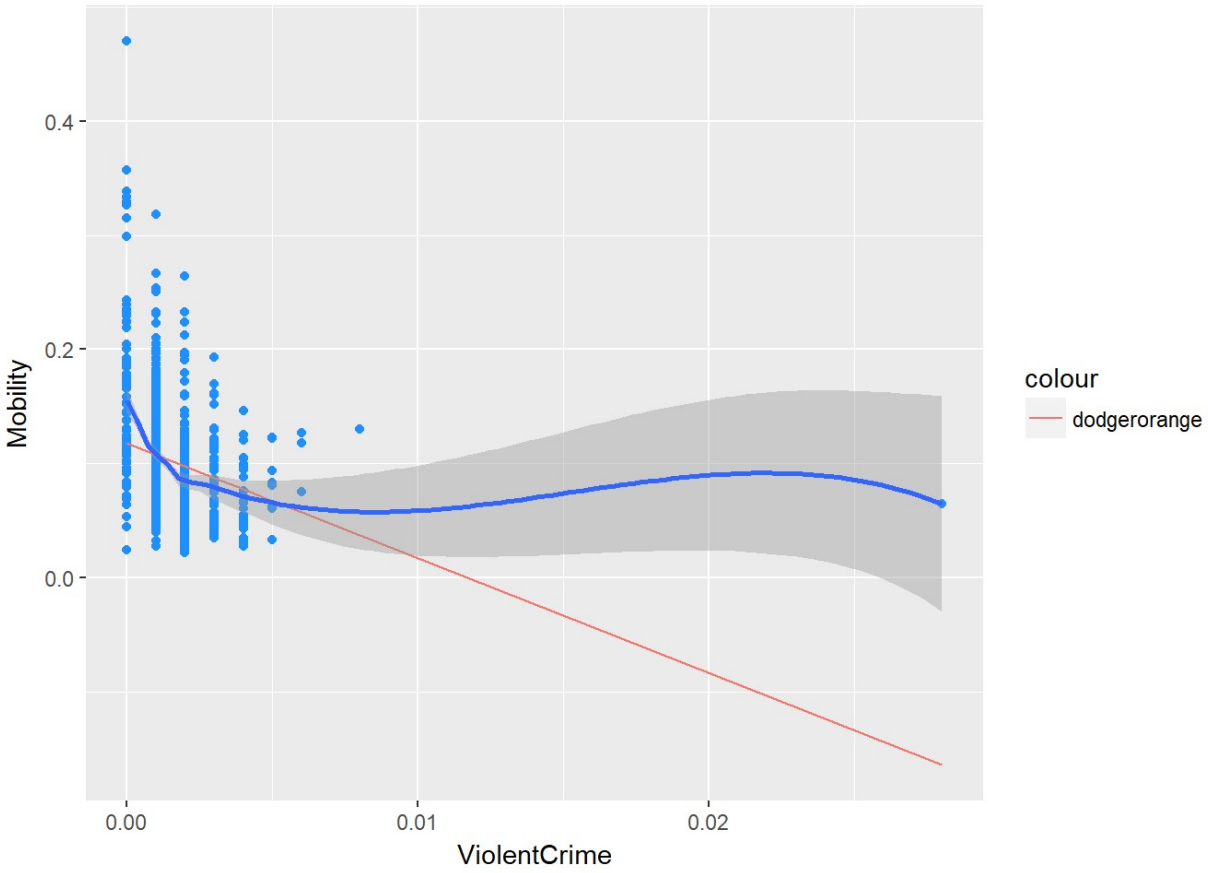
```
schoolData = as.data.frame(cbind(mobility$School_spending,mobility$Mobility))
schoolData = schoolData[complete.cases(schoolData),]
colnames(schoolData) = c("SchoolSpend","Mobility")
pred.Sch <- predict(lm(Mobility ~ SchoolSpend, data = schoolData))


p1 <- ggplot(schoolData, aes(x = SchoolSpend, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Sch,col="dodgerorange"))+geom_smooth()
```



More expenditure, more Mobility, intuitive.

Violent crime rate

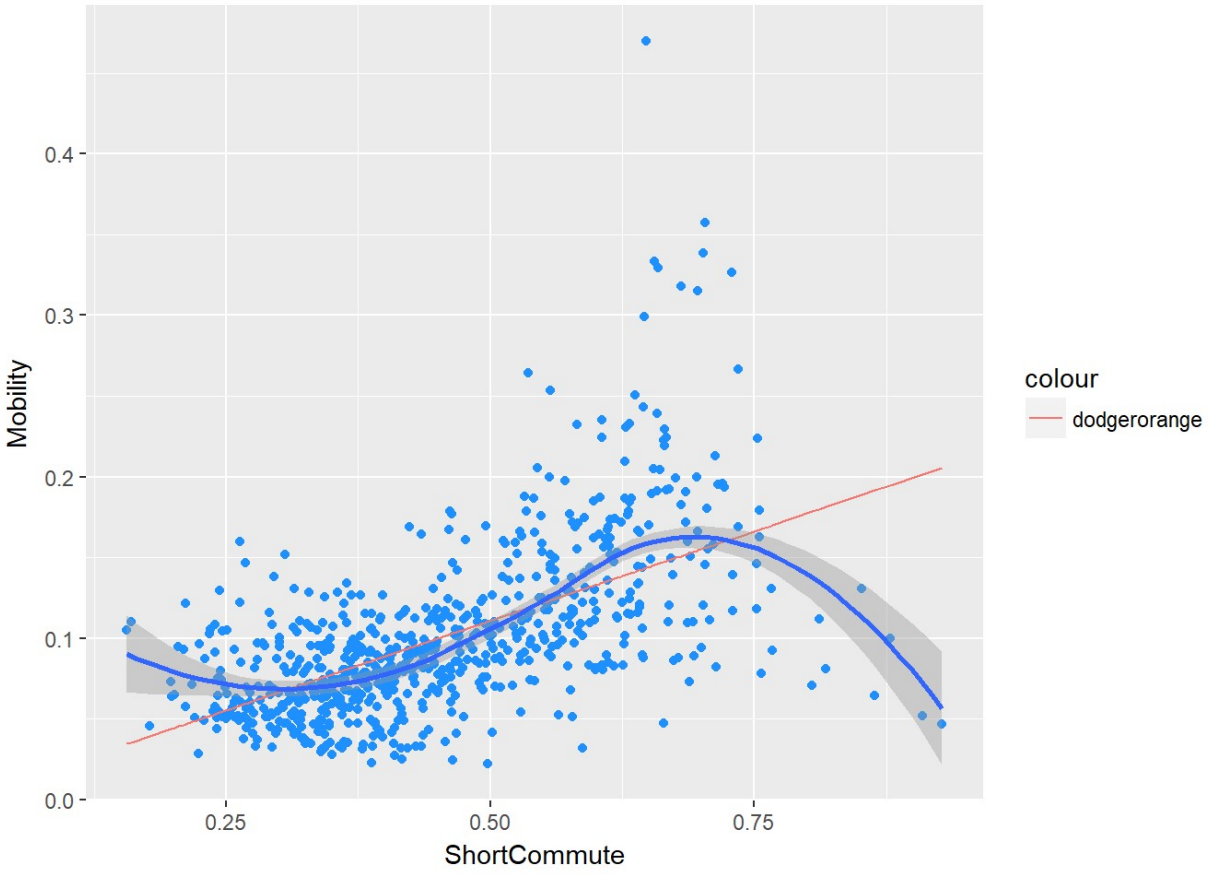
```
crimeData = as.data.frame(cbind(mobility$Violent_crime,mobility$Mobility))
crimeData = crimeData[complete.cases(crimeData),]
colnames(crimeData) = c("ViolentCrime","Mobility")
pred.Sch <- predict(lm(Mobility ~ ViolentCrime, data = crimeData))


p1 <- ggplot(crimeData, aes(x = ViolentCrime, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Sch,col="dodgerorange"))+geom_smooth()
```



Less crime, more Mobility; intuitive.

Fraction of workers with short commutes.

```
commuteData = as.data.frame(cbind(mobility$Commute,mobility$Mobility))
commuteData = commuteData[complete.cases(commuteData),]
colnames(commuteData) = c("ShortCommute","Mobility")
pred.Comm <- predict(lm(Mobility ~ ShortCommute, data = commuteData))


p1 <- ggplot(commuteData, aes(x = ShortCommute, y = Mobility))

p1 + geom_point(col="dodgerblue") +
  geom_line(aes(y = pred.Comm,col="dodgerorange"))+geom_smooth()
```



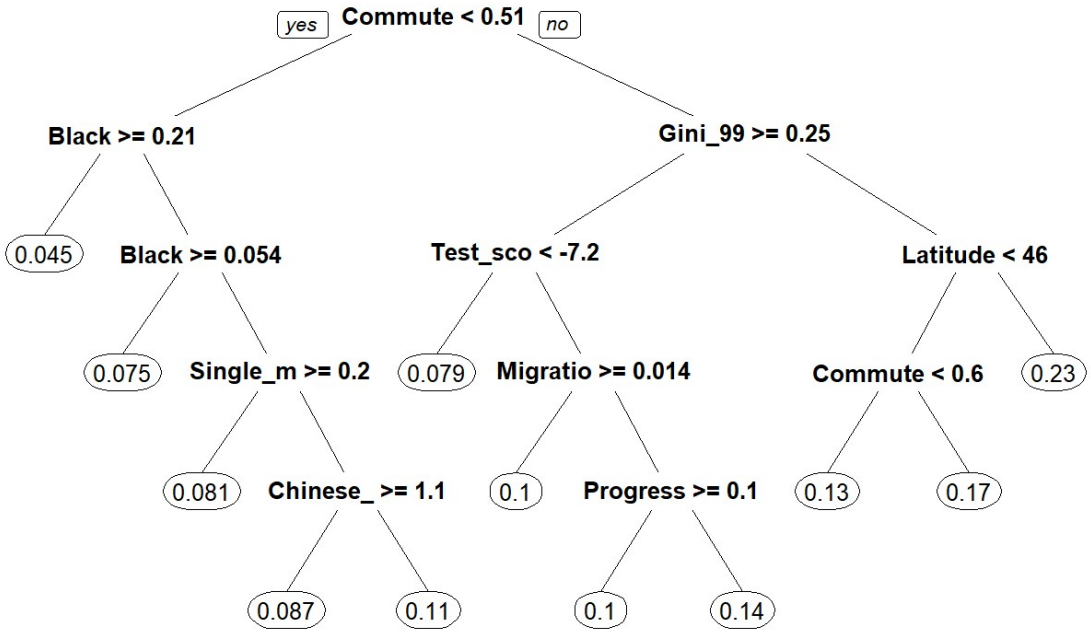
Nearby jobs, more Mobility :)

Note: All of these individual predictors aren't considered in isolation; so the observed variations as they appear on the plots might not really be because of the predictor considered.

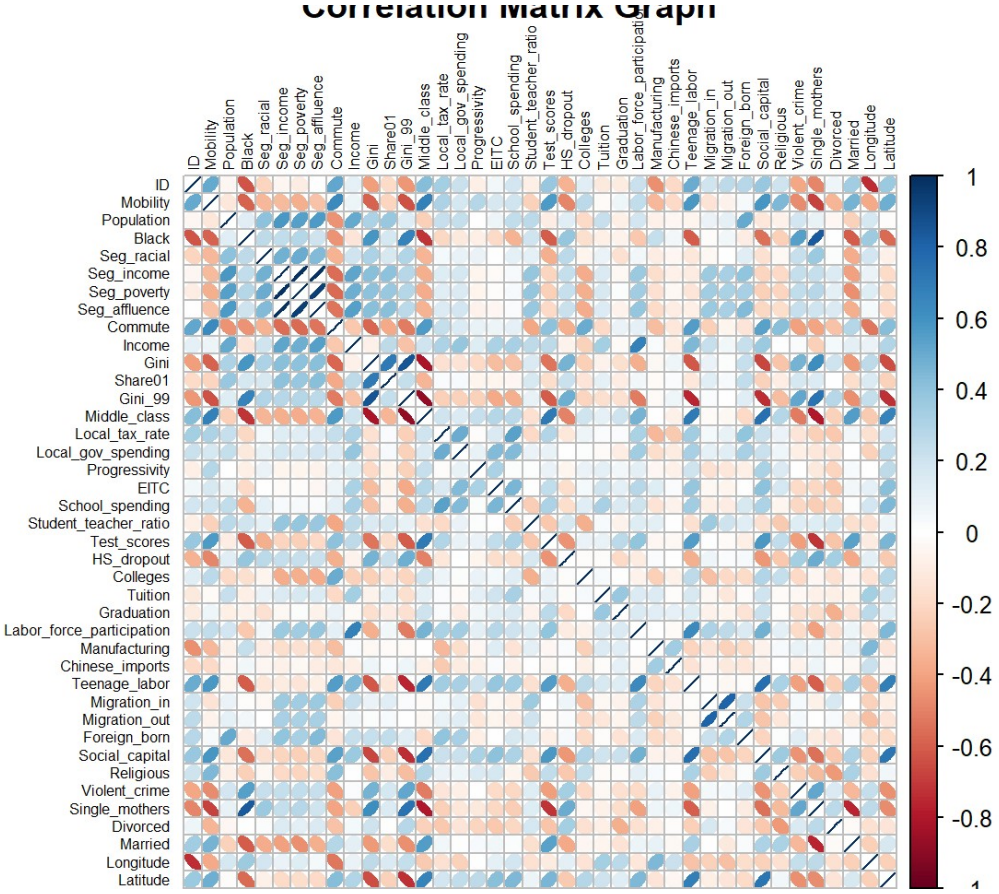
Model Selection

0. Tree model, Correlations

```
# Dropping unique IDs and Names;
# and also States (too many levels.)

drops <- c("Name", "ID","State")
dataset = mobilityData[ , !(names(mobilityData) %in% drops)]

# Lets see interactions in a tree model
form <- as.formula(Mobility ~ .)
model <- rpart(form,data=dataset)
prp(model)
```




```
# Correlations
data = na.omit(mobility)
#round(cor(data[sapply(data,is.numeric)], use="pairwise.complete.obs"),2)
corrplot(cor(data[sapply(data,is.numeric)]),method ="ellipse",
       title =" Correlation Matrix Graph",tl.cex = .5,tl.pos ="lt",tl.col ="black" )
```



Most important explanatory variable is `Commute` ; and the threshold value seperating low and high values of `Commute` is 0.511 . The fact that both limbs are branched means that other variables explain a significant amount of the variation in `Mobility` levels for values of `Commute` .

High levels of correlation amongst multiple predictors is also confirmed by the `corrplot` .

1. Linear and quadratic relationships.

```
# 1st model
model1 = lm(Mobility~Population+Black+Urban+Seg_racial+Seg_income+Seg_poverty+Seg_affl
uence+Commute+Income+Gini+Share01+Gini_99+Middle_class+Local_tax_rate+Local_gov_spendi
ng+Progressivity+EITC+School_spending+Student_teacher_ratio+Test_scores+HS_dropout+Col
leges+Tuition+Graduation+Labor_force_participation+Manufacturing+Chinese_imports+Teena
ge_labor+Migration_in+Migration_out+Foreign_born+Social_capital+Religious+Violent_crim
e+Single_mothers+Divorced+Married+Longitude+Latitude+I(Population^2)+I(Black^2)+I(Seg_
racial^2)+I(Seg_income^2)+I(Seg_poverty^2)+I(Seg_affluence^2)+I(Commute^2)+I(Income^
2)+I(Gini^2)+I(Share01^2)+I(Gini_99^2)+I(Middle_class^2)+I(Local_tax_rate^2)+I(Local_g
ov_spending^2)+I(Progressivity^2)+I(EITC^2)+I(School_spending^2)+I(Student_teacher_rat
io^2)+I(Test_scores^2)+I(HS_dropout^2)+I(Colleges^2)+I(Tuition^2)+I(Graduation^2)+I(La
bor_force_participation^2)+I(Manufacturing^2)+I(Chinese_imports^2)+I(Teenage_labor^
2)+I(Migration_in^2)+I(Migration_out^2)+I(Foreign_born^2)+I(Social_capital^2)+I(Religi
ous^2)+I(Violent_crime^2)+I(Single_mothers^2)+I(Divorced^2)+I(Married^2)+I(Longitude^
2)+I(Latitude^2),data=dataset)

# Model 2
model2 <- step(model1,trace=0)
#summary(model2)
# Insignificant
summary(model2)$coefficients[summary(model2)$coefficients[ ,4] >= 0.05,]
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## Seg_racial         0.0600466 0.04075521   1.473  0.14148
## Teenage_labor     -2.5709355 1.69086973  -1.520  0.12921
## I(Seg_poverty^2)   1.9582161 1.31631876   1.488  0.13767
## I(EITC^2)          0.0001159 0.00007046   1.645  0.10069
## I(Migration_in^2) -4.3214634 2.58413335  -1.672  0.09528
```

```
# Model 3, Removed insignificant from previous
model3 <- update(model2,~.-I(Seg_poverty^2)-Teenage_labor-Seg_racial-I(EITC^2)-I(Migra
tion_in^2))
#summary(model3)

# Lets tag this model for compariosn later on
firstModel = model3
```

2. All 2-Way interactions compared. I have $700+$ 2-Way interactions; so I fit them randomly shuffled and distributed in 7 models; since I should only estimate around 100 predictors in one model; for given dataset size (400).

```
## 7 models now
model4 = lm(Mobility ~ .+School_spending:Chinese_imports+Chinese_imports:Foreign_born+Mi
  ddle_class:Teenage_labor+Commute:Migration_in+Labor_force_participation:Single_mothers
  +Gini:Middle_class+Test_scores:Migration_in+Seg_racial:Graduation+Population:Longitude
  +Commute:Graduation+Local_tax_rate:HS_dropout+Gini_99:Violent_crime+Migration_in:Marri
  ed+Colleges:Migration_in+Student_teacher_ratio:Migration_out+Manufacturing:Latitude+Bl
  ack:Migration_in+Seg_affluence:Graduation+Black:Test_scores+Local_gov_spending:College
  s+Gini:Married+Black:Tuition+Seg_affluence:Progressivity+Black:Middle_class+EITC:Chine
  se_imports+Seg_affluence:Migration_in+Graduation:Longitude+Test_scores:Social_capital
  +Seg_poverty:Graduation+Colleges:Single_mothers+Chinese_imports:Religious+Income:Share
  01+Population:Test_scores+Seg_affluence:Tuition+Local_gov_spending:Longitude+Income:Ma
  rried+Black:Social_capital+School_spending:Divorced+EITC:Violent_crime+Progressivity:C
  hinese_imports+Seg_racial:Student_teacher_ratio+Violent_crime:Married+Commute:Income+M
  igration_out:Violent_crime+Seg_affluence:Income+Colleges:Manufacturing+Seg_income:HS_d
  ropout+Test_scores:Married+Colleges:Divorced+Seg_poverty:Gini_99+Share01:Latitude+Seg_
  racial:Chinese_imports+Income:Foreign_born+Middle_class:School_spending+Labor_force_pa
  rticipation:Longitude+Progressivity:Violent_crime+Migration_in:Violent_crime+Seg_racia
  l:Migration_in+Black:Share01+Population:Social_capital+Seg_poverty:EITC+Black:Student_
  teacher_ratio+Student_teacher_ratio:Longitude+School_spending:Religious+Labor_force_pa
  rticipation:Chinese_imports+School_spending:Longitude+Gini:Share01+Migration_out:Relig
  ious+Local_tax_rate:Teenage_labor+Seg_affluence:Migration_out+Income:Latitude+Manufact
  uring:Single_mothers+Local_gov_spending:School_spending+School_spending:Violent_crime
  +Seg_poverty:Religious+Gini_99:Tuition+Student_teacher_ratio:HS_dropout+Seg_racial:Man
  ufacturing+Manufacturing:Violent_crime+Seg_affluence:Single_mothers+Seg_racial:Local_g
  ov_spending+Gini_99:Migration_in+EITC:Graduation+Population:Foreign_born+Gini_99:Local
  _gov_spending+EITC:Test_scores+Gini_99:Single_mothers+Seg_racial:Seg_affluence+Populat
  ion:Income+Local_tax_rate:Graduation+Social_capital:Violent_crime+Seg_affluence:Marrie
  d+Seg_poverty:Middle_class+Commute:Local_tax_rate+HS_dropout:Latitude+Local_gov_spendi
  ng:Manufacturing+Seg_poverty:Single_mothers+Population:Single_mothers+Population:Latit
  ude+Migration_in:Divorced,data=dataset)
```

```
model5 = lm(Mobility~.+Share01:Violent_crime+Black:EITC+Colleges:Latitude+Tuition:Migrat
  ion_out+Social_capital:Married+Chinese_imports:Longitude+Seg_affluence:Chinese_imports
  +Teenage_labor:Migration_out+Gini:Divorced+Population:Tuition+HS_dropout:Graduation+Ch
  inese_imports:Divorced+Gini:Test_scores+Colleges:Social_capital+Labor_force_participat
  ion:Divorced+Social_capital:Religious+Seg_income:Commute+Seg_affluence:Test_scores+Man
  ufacturing:Social_capital+School_spending:Latitude+Student_teacher_ratio:Single_mother
  s+Black:Gini+Seg_poverty:Violent_crime+Foreign_born:Longitude+Local_gov_spending:Marri
  ed+Progressivity:Single_mothers+Manufacturing:Chinese_imports+Black:Commute+Black:Gini
  _99+EITC:Manufacturing+Progressivity:Labor_force_participation+Migration_out:Latitude
  +Seg_affluence:Commute+Commute:Manufacturing+Seg_income:Migration_out+EITC:Foreign_bor
  n+Migration_in:Single_mothers+Foreign_born:Married+School_spending:Married+Test_score
  s:Manufacturing+Test_scores:Graduation+Migration_out:Social_capital+Local_gov_spendin
  g:Divorced+Local_gov_spending:Graduation+Seg_poverty:Labor_force_participation+Test_sc
  ores:Single_mothers+Commute:Tuition+Local_gov_spending:Migration_in+Income:Local_gov_s
  pending+Gini_99:Progressivity+Population:Seg_affluence+Test_scores:Violent_crime+Blac
  k:Teenage_labor+Seg_racial:Labor_force_participation+Progressivity:Student_teacher_rat
  io+Seg_poverty:HS_dropout+Gini_99:Labor_force_participation+Tuition:Divorced+Local_tax
  _rate:Colleges+EITC:HS_dropout+Gini:Student_teacher_ratio+Local_gov_spending:Violent_c
  rime+Colleges:Graduation+School_spending:Tuition+Local_gov_spending:Student_teacher_ra
  tio+Black:Single_mothers+Teenage_labor:Violent_crime+Migration_out:Foreign_born+Seg_in
  come:Violent_crime+Share01:Teenage_labor+Black:Labor_force_participation+Student_teach
  er_ratio:Violent_crime+Colleges:Migration_out+Income:EITC+Seg_affluence:Local_tax_rate
  +Share01:Social_capital+Seg_racial:Latitude+Colleges:Foreign_born+Gini:Chinese_imports
  +Gini_99:Student_teacher_ratio+Seg_racial:HS_dropout+Gini:Local_gov_spending+Commute:S
  ingle_mothers+Seg_affluence:Manufacturing+Seg_affluence:Social_capital+Seg_racial:Fore
  ign_born+Violent_crime:Single_mothers+Seg_racial:Tuition+Progressivity:Longitude+Seg_a
  ffluence:Foreign_born+Labor_force_participation:Violent_crime+Manufacturing:Divorced+P
  rogressivity:Religious+Share01:EITC+Local_tax_rate:Violent_crime+Middle_class:HS_dropo
  ut+Local_tax_rate:Divorced+Local_gov_spending:Tuition+Seg_income:Share01+Tuition:Latit
  ude,data=dataset)
```

```
model6 = lm(Mobility~.+Commute:Chinese_imports+Migration_out:Longitude+Seg_racial:EITC+S
  hare01:Labor_force_participation+Population:Manufacturing+Commute:Progressivity+Incom
  e:Labor_force_participation+Progressivity:Tuition+Local_tax_rate:Migration_out+Seg_rac
  ial:Progressivity+Middle_class:Student_teacher_ratio+EITC:Student_teacher_ratio+Incom
  e:Religious+Local_gov_spending:HS_dropout+School_spending:Foreign_born+Graduation:Chin
  ese_imports+Seg_racial:Gini+Gini:Foreign_born+Seg_racial:Share01+Seg_poverty:Migration
  _in+Student_teacher_ratio:Religious+Seg_income:Gini+HS_dropout:Colleges+Progressivity:
  Married+Share01:Foreign_born+Seg_income:Labor_force_participation+Middle_class:Manufac
  turing+HS_dropout:Divorced+Labor_force_participation:Social_capital+Test_scores:Religi
  ous+Income:Chinese_imports+Seg_income:Middle_class+Graduation:Labor_force_participatio
  n+Migration_in:Migration_out+Gini:Gini_99+Population:Local_tax_rate+EITC:School_spendi
  ng+Seg_poverty:Student_teacher_ratio+Income:Gini+EITC:Religious+Local_tax_rate:Longitu
  de+Local_tax_rate:Religious+HS_dropout:Tuition+Student_teacher_ratio:Test_scores+Gini:
  Longitude+Share01:Religious+Middle_class:Chinese_imports+Tuition:Violent_crime+Seg_pov
  erty:Foreign_born+Seg_poverty:Migration_out+Foreign_born:Latitude+Population:Labor_for
  ce_participation+Test_scores:Chinese_imports+Gini:Migration_out+Gini_99:Foreign_born+G
  ini:Labor_force_participation+Income:Test_scores+Middle_class:Tuition+Seg_income:Progr
  essivity+Gini:Migration_in+Graduation:Social_capital+Tuition:Foreign_born+Seg_racial:I
  ncome+Population:Local_gov_spending+Commute:HS_dropout+Graduation:Teenage_labor+Middle
  _class:Graduation+Gini_99:Married+Longitude:Latitude+Migration_in:Social_capital+Seg_i
  ncome:Gini_99+Gini:HS_dropout+Graduation:Manufacturing+Tuition:Social_capital+Seg_pove
  rty:Colleges+School_spending:Manufacturing+Seg_poverty:Longitude+Income:Tuition+School
  _spending:Graduation+School_spending:Test_scores+Local_tax_rate:Single_mothers+Seg_aff
  luence:EITC+Gini_99:Test_scores+Gini:Progressivity+Social_capital:Longitude+Commute:St
  udent_teacher_ratio+Progressivity:Divorced+Colleges:Chinese_imports+Seg_income:Local_g
  ov_spending+Income:Manufacturing+Seg_affluence:Teenage_labor+HS_dropout:Foreign_born+R
  eligious:Latitude+Divorced:Latitude+Single_mothers:Married+Seg_poverty:Commute+School_
  spending:Teenage_labor+Teenage_labor:Religious+Seg_poverty:Share01+Local_tax_rate:Tuit
  ion,data=dataset)
```

```
model7 = lm(Mobility~.+Seg_affluence:Latitude+Seg_racial:Single_mothers+Commute:Migratio
  n_out+Share01:Local_gov_spending+EITC:Social_capital+Test_scores:Latitude+Seg_affluenc
  e:Gini+Black:Seg_poverty+Student_teacher_ratio:Graduation+Local_gov_spending:Religious
  +Commute:Foreign_born+Labor_force_participation:Married+Progressivity:Social_capital+P
  rogressivity:Migration_out+Seg_income:Latitude+Single_mothers:Latitude+Seg_poverty:Loc
  al_gov_spending+Seg_affluence:HS_dropout+Religious:Single_mothers+Seg_income:School_sp
  ending+Labor_force_participation:Migration_out+Black:Seg_affluence+Gini_99:Colleges+Po
  pulation:Religious+Graduation:Latitude+HS_dropout:Migration_in+Graduation:Foreign_born
  +Local_tax_rate:Labor_force_participation+EITC:Married+Income:Middle_class+Student_tea
  cher_ratio:Divorced+Black:Colleges+Foreign_born:Divorced+HS_dropout:Violent_crime+Seg_
  racial:Test_scores+Social_capital:Single_mothers+Local_gov_spending:Test_scores+Violen
  t_crime:Latitude+Population:Migration_out+Black:Seg_income+Commute:Labor_force_partici
  pation+Seg_racial:School_spending+Gini_99:Social_capital+Local_tax_rate:Progressivity
  +Commute:Religious+Income:Migration_in+School_spending:Colleges+Tuition:Longitude+Stud
  ent_teacher_ratio:Manufacturing+Local_tax_rate:Social_capital+Population:Middle_class
  +Student_teacher_ratio:Chinese_imports+Labor_force_participation:Migration_in+Gini:Rel
  igious+Seg_racial:Married+Share01:Local_tax_rate+Population:Graduation+Commute:Divorce
  d+Seg_racial:Migration_out+Seg_poverty:Income+Population:Black+Middle_class:Divorced+P
  opulation:Divorced+Seg_poverty:Manufacturing+Seg_poverty:Teenage_labor+Tuition:Single_
  mothers+Seg_affluence:Share01+Migration_in:Latitude+Local_gov_spending:Single_mothers
  +HS_dropout:Manufacturing+Share01:Migration_in+Middle_class:Religious+Share01:Tuition
  +Share01:Test_scores+Seg_poverty:Chinese_imports+Local_gov_spending:EITC+Seg_income:In
  come+Population:Married+Divorced:Married+School_spending:Migration_out+Black:Married+P
  opulation:Violent_crime+Tuition:Married+Migration_in:Religious+Share01:School_spending
  +Student_teacher_ratio:Foreign_born+HS_dropout:Migration_out+Tuition:Graduation+Incom
  e:Single_mothers+Colleges:Teenage_labor+Seg_poverty:Divorced+Graduation:Migration_out
  +Seg_affluence:Local_gov_spending+Income:Colleges+Gini_99:Local_tax_rate+Commute:Colle
  ges+Income:HS_dropout+Middle_class:Foreign_born+Income:Graduation+Seg_poverty:Married,
  data=dataset)
```

```
model8 = lm(Mobility~.+Student_teacher_ratio:Labor_force_participation+Share01:Single_mo
  thers+School_spending:Migration_in+Seg_affluence:Religious+Seg_poverty:School_spending
  +Religious:Longitude+Graduation:Married+Population:Commute+Seg_income:Divorced+Gini:Gr
  aduation+Black:Chinese_imports+Local_tax_rate:Foreign_born+Seg_racial:Religious+Studen
  t_teacher_ratio:Social_capital+Population:Teenage_labor+Commute:School_spending+Gini:E
  ITC+Local_tax_rate:Manufacturing+HS_dropout:Single_mothers+Local_tax_rate:Chinese_impo
  rts+Religious:Divorced+Migration_out:Married+Middle_class:Labor_force_participation+Pr
  ogressivity:Latitude+Labor_force_participation:Religious+Gini_99:Latitude+Seg_racial:C
  ommute+Population:Seg_racial+Progressivity:School_spending+Local_tax_rate:Local_gov_sp
  ending+Commute:Teenage_labor+Income:Divorced+Seg_income:Single_mothers+Gini:Violent_cr
  ime+Test_scores:Teenage_labor+Seg_racial:Colleges+Local_tax_rate:EITC+Local_tax_rate:T
  est_scores+Income:Progressivity+Income:Gini_99+Population:Seg_income+Local_tax_rate:La
  titude+Share01:Chinese_imports+Seg_income:Religious+Test_scores:Foreign_born+Income:St
  udent_teacher_ratio+Teenage_labor:Latitude+Gini:Colleges+Commute:Test_scores+Middle_cl
  ass:Test_scores+Seg_income:Test_scores+Single_mothers:Longitude+Seg_income:Seg_affluen
  ce+Seg_affluence:Divorced+Seg_racial:Longitude+Commute:Gini_99+Seg_affluence:Middle_cl
  ass+Seg_poverty:Local_tax_rate+Tuition:Labor_force_participation+Colleges:Violent_crim
  e+Share01:Graduation+Local_gov_spending:Progressivity+Commute:Share01+Tuition:Migratio
  n_in+Manufacturing:Married+Gini_99:Chinese_imports+Gini_99:HS_dropout+Gini_99:Longitud
  e+Population:Gini_99+Graduation:Religious+Income:Social_capital+Labor_force_participat
  ion:Foreign_born+School_spending:Student_teacher_ratio+Gini:Tuition+Black:Migration_ou
  t+HS_dropout:Longitude+Graduation:Divorced+Progressivity:Migration_in+Progressivity:HS
  _dropout+Gini_99:Divorced+Seg_racial:Gini_99+EITC:Single_mothers+Seg_income:Longitude
  +Population:Progressivity+Population:Gini+Gini_99:Teenage_labor+EITC:Migration_out+Bla
  ck:Religious+Middle_class:Local_gov_spending+Share01:Progressivity+Seg_poverty:Progres
  sivity+Local_tax_rate:Student_teacher_ratio+Progressivity:Manufacturing+Seg_affluence:
  Labor_force_participation+Middle_class:Longitude+Black:Divorced+School_spending:Single
  _mothers+Manufacturing:Teenage_labor+HS_dropout:Labor_force_participation+Commute:Midd
  le_class+Gini:Teenage_labor,data=dataset)
```

```
model9 = lm(Mobility~.+Religious:Married+School_spending:Social_capital+Progressivity:EI
  TC+Share01:Divorced+Gini_99:Manufacturing+Seg_income:Tuition+Commute:EITC+Seg_income:C
  hinese_imports+Married:Longitude+Middle_class:Married+Seg_affluence:Violent_crime+Divo
  rced:Longitude+Seg_poverty:Gini+EITC:Teenage_labor+Manufacturing:Foreign_born+HS_dropo
  ut:Chinese_imports+Test_scores:Tuition+Test_scores:Migration_out+Seg_racial:Seg_povert
  y+EITC:Migration_in+Black:School_spending+Black:Progressivity+Colleges:Married+Gini:Si
  ngle_mothers+Teenage_labor:Single_mothers+Black:Graduation+Student_teacher_ratio:Colle
  ges+Test_scores:Colleges+Gini_99:Graduation+Seg_racial:Divorced+Population:HS_dropout
  +Chinese_imports:Migration_in+Seg_affluence:School_spending+Tuition:Chinese_imports+Se
  g_racial:Social_capital+Manufacturing:Migration_in+Chinese_imports:Teenage_labor+Colle
  ges:Longitude+Labor_force_participation:Teenage_labor+Population:Seg_poverty+Migration
  _out:Divorced+Manufacturing:Migration_out+HS_dropout:Teenage_labor+Seg_affluence:Gini_
  99+Progressivity:Colleges+Foreign_born:Single_mothers+Middle_class:EITC+Foreign_born:S
  ocial_capital+Population:Colleges+Test_scores:Divorced+Student_teacher_ratio:Migration
  _in+Share01:Student_teacher_ratio+Local_gov_spending:Chinese_imports+Violent_crime:Lon
  gitude+Seg_income:Local_tax_rate+Colleges:Labor_force_participation+Foreign_born:Viole
  nt_crime+Teenage_labor:Married+Commute:Gini+School_spending:Labor_force_participation
  +Commute:Longitude+Seg_income:Married+Black:Latitude+Seg_income:Colleges+Population:St
  udent_teacher_ratio+Manufacturing:Religious+Progressivity:Test_scores+Black:Income+Tee
  nage_labor:Migration_in+Chinese_imports:Married+Colleges:Tuition+Chinese_imports:Viole
  nt_crime+Black:Foreign_born+Graduation:Single_mothers+Seg_poverty:Test_scores+Seg_raci
  al:Seg_income+Share01:HS_dropout+Seg_racial:Violent_crime+Colleges:Religious+Tuition:M
  anufacturing+Teenage_labor:Longitude+Seg_income:Graduation+Labor_force_participation:L
  atitude+Black:Local_gov_spending+Teenage_labor:Divorced+Population:Share01+Gini:School
  _spending+Middle_class:Latitude+Commute:Latitude+Student_teacher_ratio:Married+Black:S
  eg_racial+Migration_in:Longitude+EITC:Longitude+Local_gov_spending:Foreign_born+Share0
  1:Longitude+Share01:Migration_out+Student_teacher_ratio:Tuition+Seg_poverty:Tuition+Vi
  olent_crime:Divorced+Income:Longitude+Black:Local_tax_rate,data=dataset)
```

```
model10 = lm(Mobility~.+Test_scores:Longitude+Gini:Latitude+EITC:Latitude+Gini:Manufactu
  ring+Student_teacher_ratio:Latitude+EITC:Tuition+Population:Chinese_imports+Gini_99:EI
  TC+Migration_in:Foreign_born+Foreign_born:Religious+Chinese_imports:Social_capital+Mar
  ried:Latitude+Gini:Local_tax_rate+Local_gov_spending:Teenage_labor+Gini_99:School_spen
  ding+Gini_99:Migration_out+Manufacturing:Longitude+Social_capital:Divorced+Middle_clas
  s:Progressivity+Progressivity:Teenage_labor+Test_scores:Labor_force_participation+Loca
  l_gov_spending:Migration_out+Seg_income:Student_teacher_ratio+Seg_racial:Local_tax_rat
  e+Seg_income:Foreign_born+Income:Migration_out+Income:Violent_crime+Population:School_
  spending+Seg_racial:Middle_class+Religious:Violent_crime+Black:Violent_crime+Income:Te
  enage_labor+Graduation:Migration_in+Seg_income:Manufacturing+Share01:Middle_class+Shar
  e01:Manufacturing+Teenage_labor:Social_capital+Seg_poverty:Seg_affluence+Local_tax_rat
  e:Migration_in+Share01:Gini_99+Seg_income:Migration_in+Population:Migration_in+Migrati
  on_out:Single_mothers+Commute:Violent_crime+Social_capital:Latitude+HS_dropout:Social_
  capital+Local_gov_spending:Social_capital+Share01:Colleges+Income:Local_tax_rate+EITC:
  Labor_force_participation+Local_tax_rate:Married+Test_scores:HS_dropout+Teenage_labor:
  Foreign_born+Progressivity:Foreign_born+Gini_99:Middle_class+Progressivity:Graduation
  +Tuition:Teenage_labor+Chinese_imports:Single_mothers+Student_teacher_ratio:Teenage_la
  bor+Chinese_imports:Migration_out+Single_mothers:Divorced+Labor_force_participation:Ma
  nufacturing+Black:Longitude+Middle_class:Violent_crime+Tuition:Religious+Middle_class:
  Social_capital+Graduation:Violent_crime+Seg_racial:Teenage_labor+Middle_class:Single_m
  others+Seg_income:EITC+Black:Manufacturing+Seg_poverty:Latitude+HS_dropout:Married+Com
  mute:Married+Commute:Local_gov_spending+Local_gov_spending:Latitude+Seg_affluence:Long
  itude+Local_gov_spending:Labor_force_participation+Seg_affluence:Student_teacher_ratio
  +HS_dropout:Religious+Seg_poverty:Social_capital+EITC:Divorced+Middle_class:Local_tax_
  rate+Chinese_imports:Latitude+Gini_99:Religious+School_spending:HS_dropout+Middle_clas
  s:Migration_in+Share01:Married+Seg_income:Teenage_labor+Black:HS_dropout+Gini:Social_c
  apital+Seg_affluence:Colleges+Commute:Social_capital+Population:EITC+Middle_class:Coll
  eges+Seg_income:Social_capital+EITC:Colleges+Local_tax_rate:School_spending+Income:Sch
  ool_spending+Middle_class:Migration_out+Seg_income:Seg_poverty,data=dataset)



# Only significant ones
```

```
model11 = lm(Mobility~.+Middle_class:Teenage_labor+Gini_99:Violent_crime+Black:Test_scor
  es+Seg_affluence:Progressivity+Seg_affluence:Migration_in+Labor_force_participation:Lo
  ngitude+Seg_poverty:EITC+Student_teacher_ratio:Longitude+School_spending:Religious+Loc
  al_tax_rate:Teenage_labor+Seg_poverty:Middle_class+Colleges:Latitude+Tuition:Migration
  _out+Gini:Divorced+Seg_poverty:HS_dropout+Local_tax_rate:Colleges+Seg_affluence:Local_
  tax_rate+Seg_racial:HS_dropout+Gini:Local_gov_spending+Progressivity:Religious+Middle_
  class:HS_dropout+Tuition:Latitude+Commute:Progressivity+Seg_racial:Progressivity+Stude
  nt_teacher_ratio:Religious+Middle_class:Manufacturing+HS_dropout:Divorced+Test_scores:
  Religious+Seg_income:Middle_class+Migration_in:Migration_out+Foreign_born:Latitude+Mig
  ration_in:Social_capital+School_spending:Test_scores+Gini_99:Test_scores+Commute:Stude
  nt_teacher_ratio+Commute:Migration_out+Test_scores:Latitude+Commute:Foreign_born+Gini_
  99:Colleges+Commute:Labor_force_participation+Local_tax_rate:Progressivity+Commute:Rel
  igious+Local_tax_rate:Social_capital+Seg_racial:Married+Share01:Test_scores+Income:Sin
  gle_mothers+Income:Colleges+Middle_class:Foreign_born+Seg_poverty:Married+HS_dropout:S
  ingle_mothers+Religious:Divorced+Gini_99:Latitude+Progressivity:School_spending+Seg_ra
  cial:Colleges+Commute:Test_scores+Gini_99:HS_dropout+Income:Social_capital+Gini_99:Div
  orced+Population:Progressivity+Seg_poverty:Progressivity+Black:Divorced+Commute:Middle
  _class+Gini_99:Manufacturing+HS_dropout:Chinese_imports+Labor_force_participation:Teen
  age_labor+Progressivity:Colleges+Foreign_born:Social_capital+Seg_racial:Seg_income+Sha
  re01:HS_dropout+Middle_class:Latitude+Test_scores:Longitude+Gini:Latitude+Seg_income:M
  anufacturing+Test_scores:HS_dropout+Progressivity:Graduation+Middle_class:Violent_crim
  e+Middle_class:Single_mothers+HS_dropout:Religious+Black:HS_dropout+Gini:Social_capita
  l+Population:EITC+EITC:Colleges,data=dataset)

# See if I am not estimating too many parameters
nrow(dataset)/3 >length(coef(model11))
```

```
## [1] TRUE
```

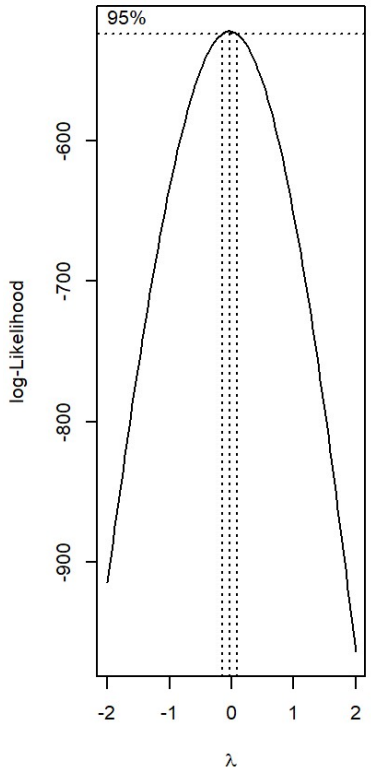
```
# Lets tag this model
secondModel = model11
```

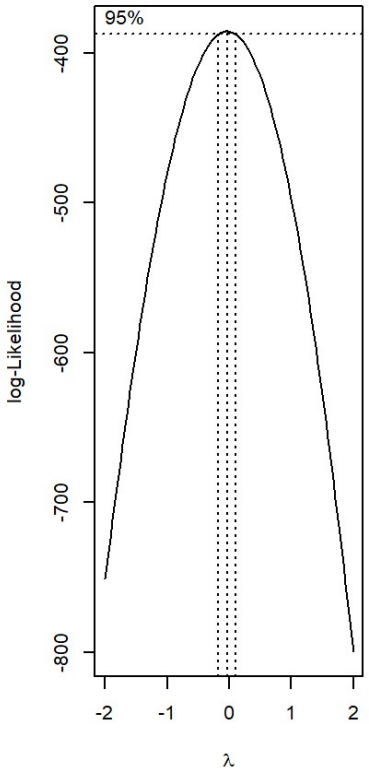
3. Final take at improvement

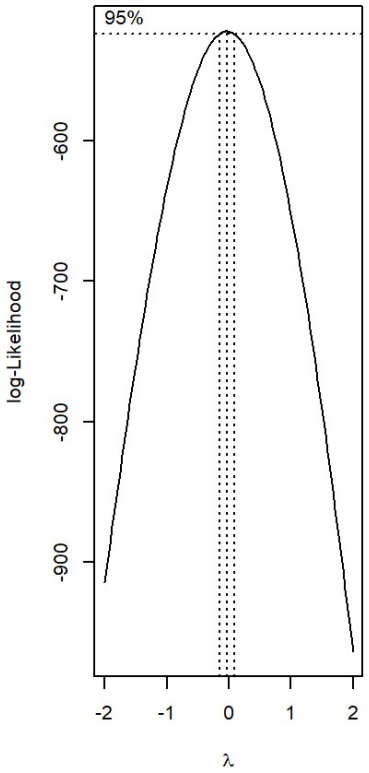
```
mod_both_aic = step(model3,model11,direction = "both",trace=0)

# Tag this too.
thirdModel = mod_both_aic
```

4. Transformations?

```
par(mfrow=c(1,3))
boxcox(firstModel, plotit = TRUE)
boxcox(secondModel, plotit = TRUE)
boxcox(thirdModel, plotit = TRUE)
```


95%
log-Likelihood
-600
-700
-800
-900
-2
-1
0
1
2
λ


95%
log-Likelihood
-400
-500
-600
-700
-800
-2
-1
0
1
2
λ


95%
log-Likelihood
-600
-700
-800
-900
-2
-1
0
1
2
λ

```
# Log Transforms
firstLog = lm(log(Mobility)~Seg_income+Seg_poverty+Commute+Income+Gini+Share01+Middle_cl
  ass+Progressivity+EITC+School_spending+HS_dropout+Colleges+Labor_force_participation+M
  anufacturing+Social_capital+Single_mothers+Longitude+Latitude+I(Seg_racial^2)+I(Commut
  e^2)+I(Gini_99^2)+I(Middle_class^2)+I(School_spending^2)+I(HS_dropout^2)+I(Social_capi
  tal^2)+I(Religious^2)+I(Longitude^2)+I(Latitude^2),data=dataset)
secondLog = lm(log(Mobility)~.+Middle_class:Teenage_labor+Gini_99:Violent_crime+Black:Te
  st_scores+Seg_affluence:Progressivity+Seg_affluence:Migration_in+Labor_force_participa
  tion:Longitude+Seg_poverty:EITC+Student_teacher_ratio:Longitude+School_spending:Religi
  ous+Local_tax_rate:Teenage_labor+Seg_poverty:Middle_class+Colleges:Latitude+Tuition:Mi
  gration_out+Gini:Divorced+Seg_poverty:HS_dropout+Local_tax_rate:Colleges+Seg_affluenc
  e:Local_tax_rate+Seg_racial:HS_dropout+Gini:Local_gov_spending+Progressivity:Religious
  +Middle_class:HS_dropout+Tuition:Latitude+Commute:Progressivity+Seg_racial:Progressivi
  ty+Student_teacher_ratio:Religious+Middle_class:Manufacturing+HS_dropout:Divorced+Test
  _scores:Religious+Seg_income:Middle_class+Migration_in:Migration_out+Foreign_born:Lati
  tude+Migration_in:Social_capital+School_spending:Test_scores+Gini_99:Test_scores+Commu
  te:Student_teacher_ratio+Commute:Migration_out+Test_scores:Latitude+Commute:Foreign_bo
  rn+Gini_99:Colleges+Commute:Labor_force_participation+Local_tax_rate:Progressivity+Com
  mute:Religious+Local_tax_rate:Social_capital+Seg_racial:Married+Share01:Test_scores+In
  come:Single_mothers+Income:Colleges+Middle_class:Foreign_born+Seg_poverty:Married+HS_d
  ropout:Single_mothers+Religious:Divorced+Gini_99:Latitude+Progressivity:School_spendin
  g+Seg_racial:Colleges+Commute:Test_scores+Gini_99:HS_dropout+Income:Social_capital+Gin
  i_99:Divorced+Population:Progressivity+Seg_poverty:Progressivity+Black:Divorced+Commut
  e:Middle_class+Gini_99:Manufacturing+HS_dropout:Chinese_imports+Labor_force_participat
  ion:Teenage_labor+Progressivity:Colleges+Foreign_born:Social_capital+Seg_racial:Seg_in
  come+Share01:HS_dropout+Middle_class:Latitude+Test_scores:Longitude+Gini:Latitude+Seg_
  income:Manufacturing+Test_scores:HS_dropout+Progressivity:Graduation+Middle_class:Viol
  ent_crime+Middle_class:Single_mothers+HS_dropout:Religious+Black:HS_dropout+Gini:Socia
  l_capital+Population:EITC+EITC:Colleges,data=dataset)
thirdLog = lm(log(Mobility)~Seg_income+Seg_poverty+Commute+Income+Gini+Share01+Middle_cl
  ass+Progressivity+EITC+School_spending+HS_dropout+Colleges+Labor_force_participation+M
  anufacturing+Social_capital+Single_mothers+Longitude+Latitude+I(Seg_racial^2)+
+I(Commute^2)+I(Gini_99^2)+I(Middle_class^2)+I(School_spending^2)+I(HS_dropout^2)+I(Soci
  al_capital^2)+I(Religious^2)+I(Longitude^2)+I(Latitude^2),data=dataset)
```

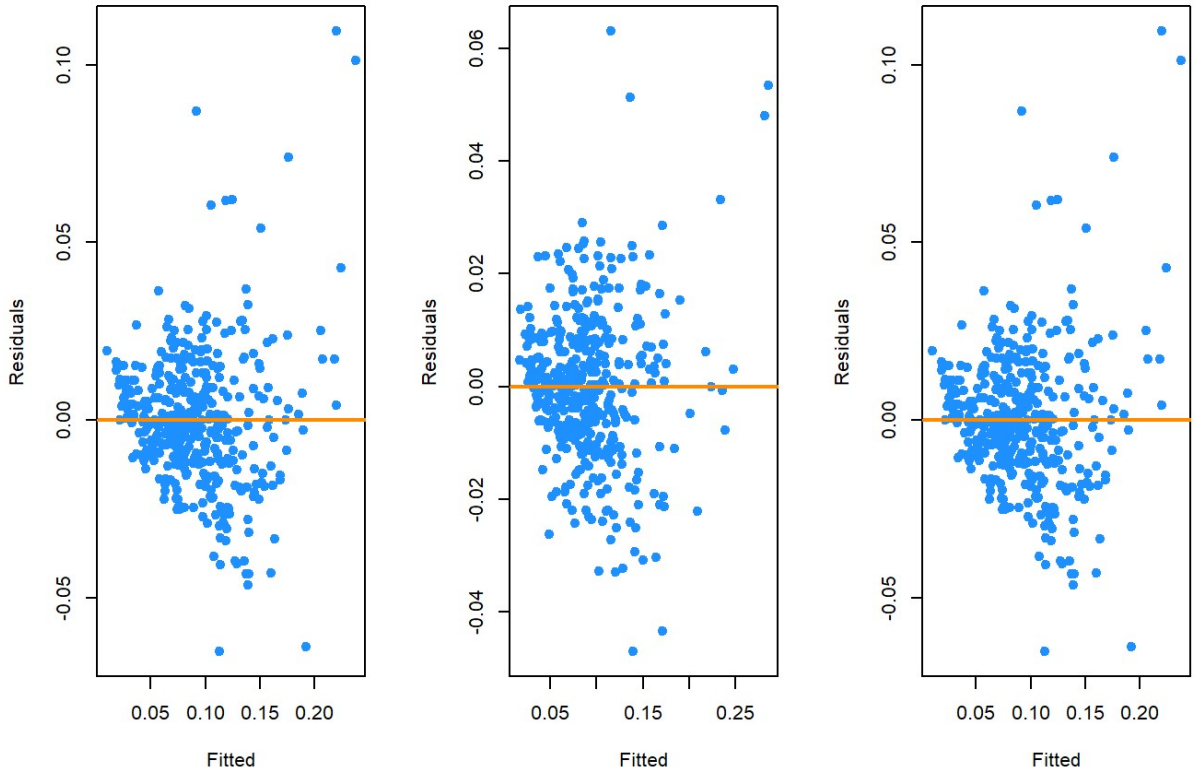
Model Diagnostics

I have Six models to dignose: `firstModel` , `secondModel` and `thirdModel` and their transformed `log` versions.

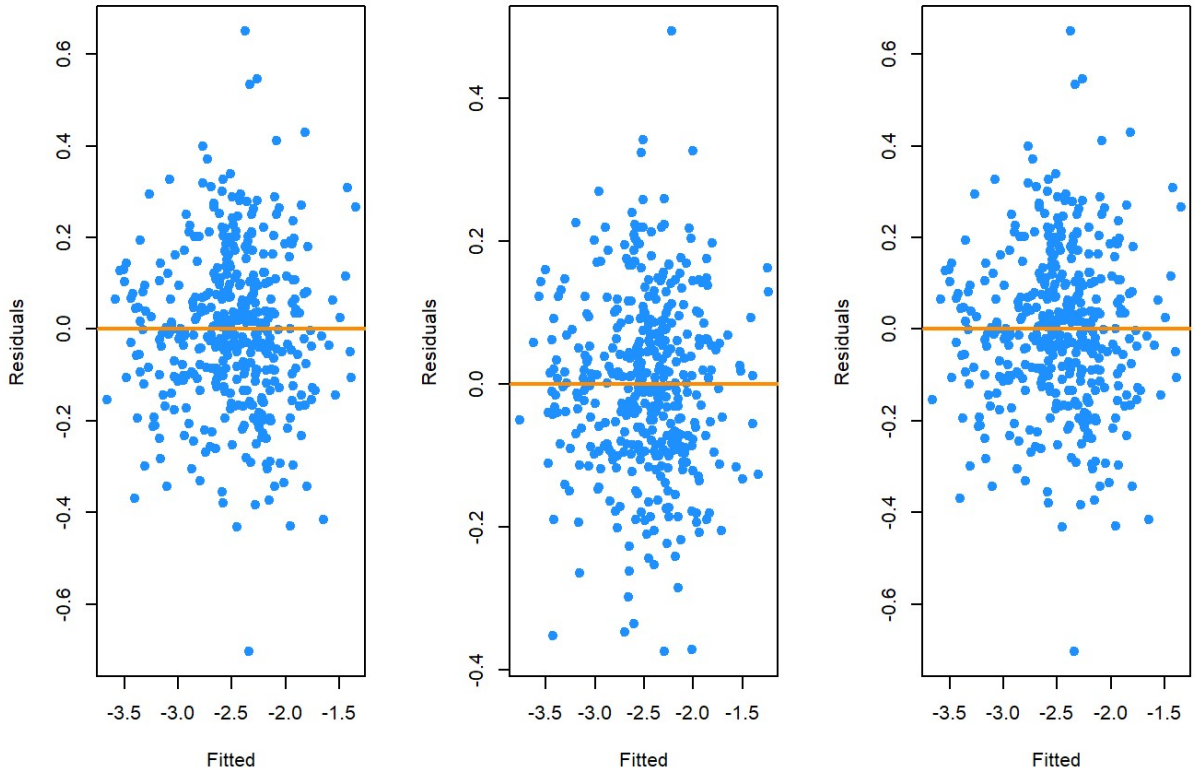
Model Assumptions

1. Linearity and Constant Variance

```
par(mfrow=c(1,3))
plot_fitted_resid(firstModel)
plot_fitted_resid(secondModel)
plot_fitted_resid(thirdModel)
```




```
par(mfrow=c(1,3))
plot_fitted_resid(firstLog)
plot_fitted_resid(secondLog)
plot_fitted_resid(thirdLog)
```




```
bptest(firstModel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  firstModel
## BP = 120, df = 28, p-value = 7e-13
```

```
bptest(secondModel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  secondModel
## BP = 170, df = 120, p-value = 0.002
```

```
bptest(thirdModel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  thirdModel
## BP = 120, df = 28, p-value = 7e-13
```

```
bptest(firstLog)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  firstLog
## BP = 70, df = 28, p-value = 2e-05
```

```
bptest(secondLog)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  secondLog
## BP = 130, df = 120, p-value = 0.3
```

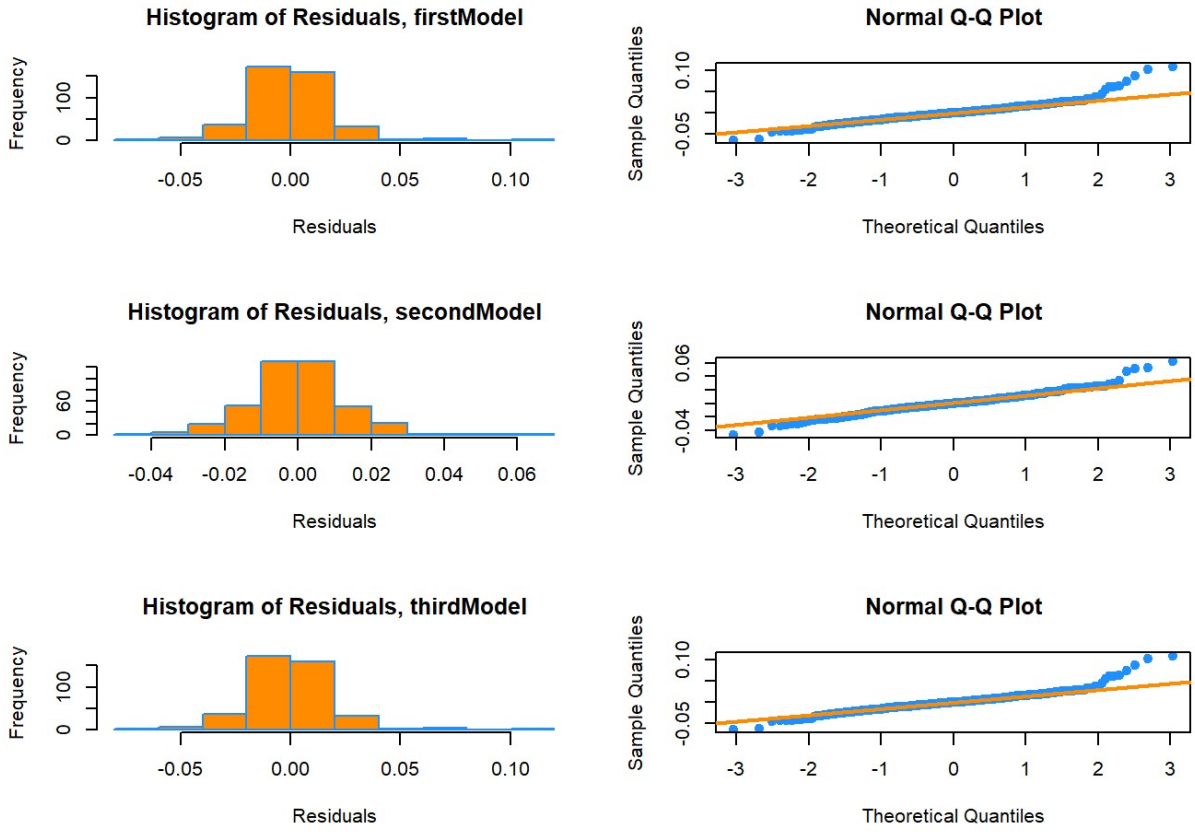
```
bptest(thirdLog)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  thirdLog
## BP = 70, df = 28, p-value = 2e-05
```

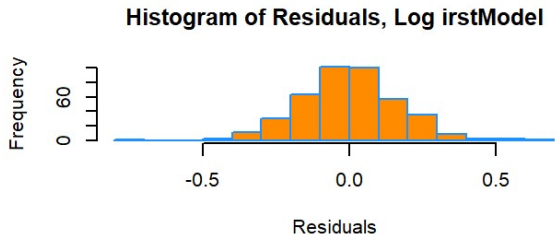
Based on these tests and plots, `secondLog` is the winner.

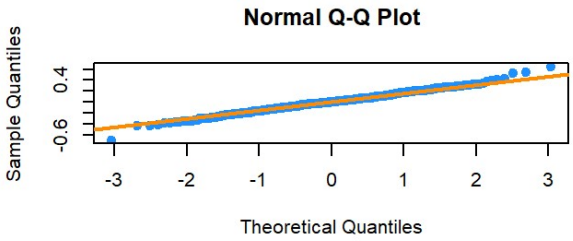
2. Normailty of errors

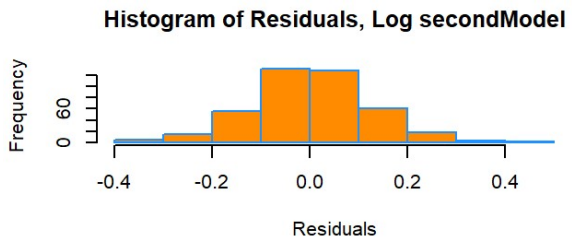
```
par(mfrow=c(3,2))
hist(resid(firstModel),xlab = "Residuals",,main = "Histogram of Residuals, firstModel",c
  ol = "darkorange",border = "dodgerblue")
plot_qq(firstModel)
hist(resid(secondModel),xlab = "Residuals",,main = "Histogram of Residuals, secondMode
  l",col = "darkorange",border = "dodgerblue")
plot_qq(secondModel)
hist(resid(thirdModel),xlab = "Residuals",,main = "Histogram of Residuals, thirdModel",c
  ol = "darkorange",border = "dodgerblue")
plot_qq(thirdModel)
```

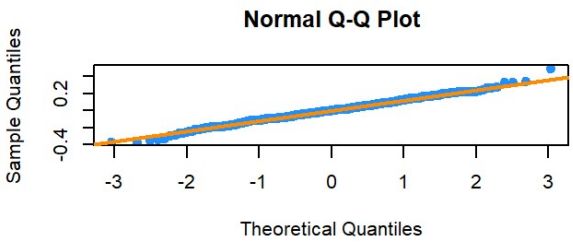


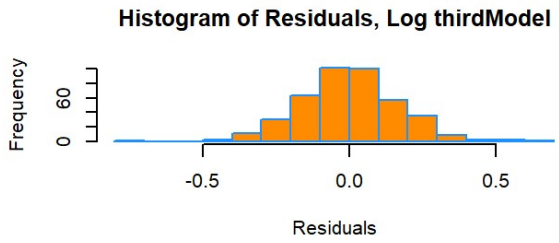

```
par(mfrow=c(3,2))
hist(resid(firstLog),xlab = "Residuals",,main = "Histogram of Residuals, Log irstModel",
  col = "darkorange",border = "dodgerblue")
plot_qq(firstLog)
hist(resid(secondLog),xlab = "Residuals",,main = "Histogram of Residuals, Log secondMode
  l",col = "darkorange",border = "dodgerblue")
plot_qq(secondLog)
hist(resid(thirdLog),xlab = "Residuals",,main = "Histogram of Residuals, Log thirdMode
  l",col = "darkorange",border = "dodgerblue")
plot_qq(thirdLog)
```

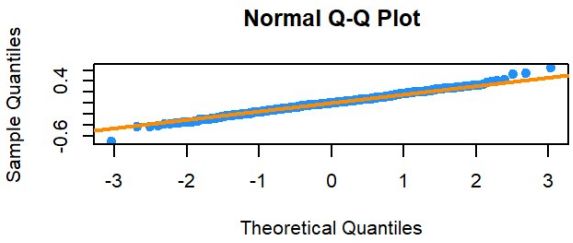













```
shapiro.test(resid(firstModel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(firstModel)
## W = 0.93, p-value = 1e-13
```

```
shapiro.test(resid(secondModel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(secondModel)
## W = 0.97, p-value = 3e-07
```

```
shapiro.test(resid(thirdModel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(thirdModel)
## W = 0.93, p-value = 1e-13
```

```
shapiro.test(resid(firstLog))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(firstLog)
## W = 0.99, p-value = 0.04
```

```
shapiro.test(resid(secondLog))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(secondLog)
## W = 0.99, p-value = 0.1
```

```
shapiro.test(resid(thirdLog))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(thirdLog)
## W = 0.99, p-value = 0.04
```

`secondLog` model is winner based on plots and tests.

Unsual Observations

Leverage:

```
# Leverage
length(hatvalues(firstModel)[hatvalues(firstModel) > 2 * mean(hatvalues(firstModel))])
```

```
## [1] 29
```

```
length(hatvalues(secondModel)[hatvalues(secondModel) > 2 * mean(hatvalues(secondMode
  l))])
```

```
## [1] 27
```

```
length(hatvalues(thirdModel)[hatvalues(thirdModel) > 2 * mean(hatvalues(thirdModel))])
```

```
## [1] 29
```

```
length(hatvalues(firstLog)[hatvalues(firstLog) > 2 * mean(hatvalues(firstLog))])
```

```
## [1] 29
```

```
length(hatvalues(secondLog)[hatvalues(secondLog) > 2 * mean(hatvalues(secondLog))])
```

```
## [1] 27
```

```
length(hatvalues(thirdLog)[hatvalues(thirdLog) > 2 * mean(hatvalues(thirdLog))])
```

```
## [1] 29
```

Outliers:

```
# Outliers
length(rstandard(firstModel)[abs(rstandard(firstModel)) > 2])
```

```
## [1] 21
```

```
length(rstandard(secondModel)[abs(rstandard(secondModel)) > 2])
```

```
## [1] 23
```

```
length(rstandard(thirdModel)[abs(rstandard(thirdModel)) > 2])
```

```
## [1] 21
```

```
length(rstandard(firstLog)[abs(rstandard(firstLog)) > 2])
```

```
## [1] 20
```

```
length(rstandard(secondLog)[abs(rstandard(secondLog)) > 2])
```

```
## [1] 17
```

```
length(rstandard(thirdLog)[abs(rstandard(thirdLog)) > 2])
```

```
## [1] 20
```

Influential:

```
# Influential
length(cooks.distance(firstModel)[cooks.distance(firstModel) > 4 / length(cooks.distance
  (firstModel))])
```

```
## [1] 30
```

```
length(cooks.distance(secondModel)[cooks.distance(secondModel) > 4 / length(cooks.distan
  ce(secondModel))])
```

```
## [1] 52
```

```
length(cooks.distance(thirdModel)[cooks.distance(thirdModel) > 4 / length(cooks.distance
  (thirdModel))])
```

```
## [1] 30
```

```
length(cooks.distance(firstLog)[cooks.distance(firstLog) > 4 / length(cooks.distance(fir
  stLog))])
```

```
## [1] 32
```

```
length(cooks.distance(secondLog)[cooks.distance(secondLog) > 4 / length(cooks.distance(s
  econdLog))])
```

```
## [1] 54
```

```
length(cooks.distance(thirdLog)[cooks.distance(thirdLog) > 4 / length(cooks.distance(thi
  rdLog))])
```

```
## [1] 32
```

Prsence of Outliers, Influential points and Leveraging points is expected in real life datasets.

Evaluations

##adj.r.squared

```
summary(firstModel)$adj.r.squared
```

```
## [1] 0.7947
```

```
summary(secondModel)$adj.r.squared
```

```
## [1] 0.8714
```

```
summary(thirdModel)$adj.r.squared
```

```
## [1] 0.7947
```

```
summary(firstLog)$adj.r.squared
```

```
## [1] 0.8571
```

```
summary(secondLog)$adj.r.squared
```

```
## [1] 0.9024
```

```
summary(thirdLog)$adj.r.squared
```

```
## [1] 0.8571
```

##calc_loocv_rmse

```
calc_loocv_rmse(firstModel)
```

```
## [1] 0.02121
```

```
calc_loocv_rmse(secondModel)
```

```
## [1] 0.02224
```

```
calc_loocv_rmse(thirdModel)
```

```
## [1] 0.02121
```

```
calc_loocv_rmse(firstLog)
```

```
## [1] 0.188
```

```
calc_loocv_rmse(secondLog)
```

```
## [1] 0.1954
```

```
calc_loocv_rmse(thirdLog)
```

```
## [1] 0.188
```

##VIF

```
# VIF
sum(vif(firstModel)>5)/length(coef(firstModel))
```

```
## [1] 0.5517
```

```
sum(vif(secondModel)>5)/length(coef(secondModel))
```

```
## [1] 0.9508
```

```
sum(vif(thirdModel)>5)/length(coef(thirdModel))
```

```
## [1] 0.5517
```

```
sum(vif(firstLog)>5)/length(coef(firstLog))
```

```
## [1] 0.5517
```

```
sum(vif(secondLog)>5)/length(coef(secondLog))
```

```
## [1] 0.9508
```

```
sum(vif(thirdLog)>5)/length(coef(thirdLog))
```

```
## [1] 0.5517
```

##AIC

```
#AIC
extractAIC(firstModel)
```

```
## [1]    29 -3244
```

```
extractAIC(secondModel)
```

```
## [1]   122 -3367
```

```
extractAIC(thirdModel)
```

```
## [1]    29 -3244
```

```
extractAIC(firstLog)
```

```
## [1]    29 -1420
```

```
extractAIC(secondLog)
```

```
## [1]   122 -1508
```

```
extractAIC(thirdLog)
```

```
## [1]    29 -1420
```

##BIC

```
#BIC
extractAIC(firstModel,k=log(nrow(dataset)))
```

```
## [1]    29 -3127
```

```
extractAIC(secondModel,k=log(nrow(dataset)))
```

```
## [1]   122 -2875
```

```
extractAIC(thirdModel,k=log(nrow(dataset)))
```

```
## [1]    29 -3127
```

```
extractAIC(firstLog,k=log(nrow(dataset)))
```

```
## [1]    29 -1303
```

```
extractAIC(secondLog,k=log(nrow(dataset)))
```

```
## [1]   122 -1015
```

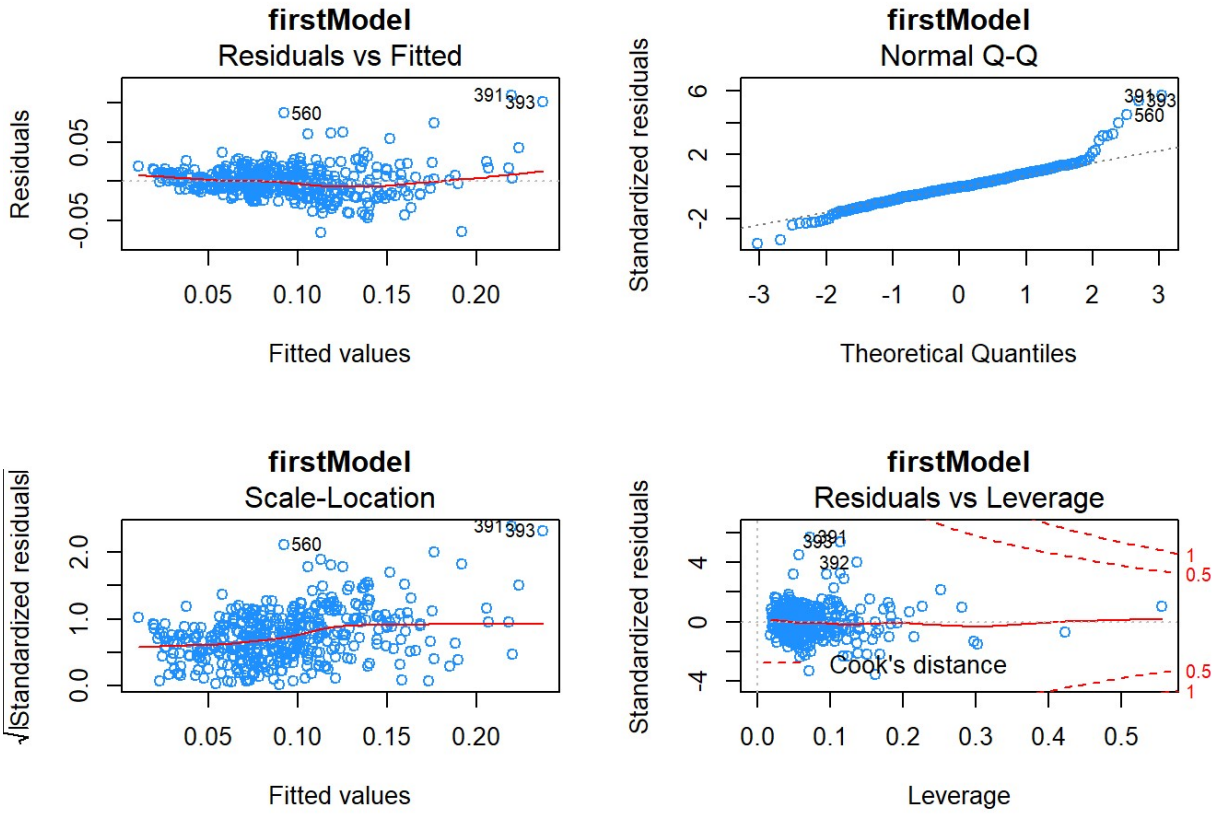
```
extractAIC(thirdLog,k=log(nrow(dataset)))
```

```
## [1]    29 -1303
```

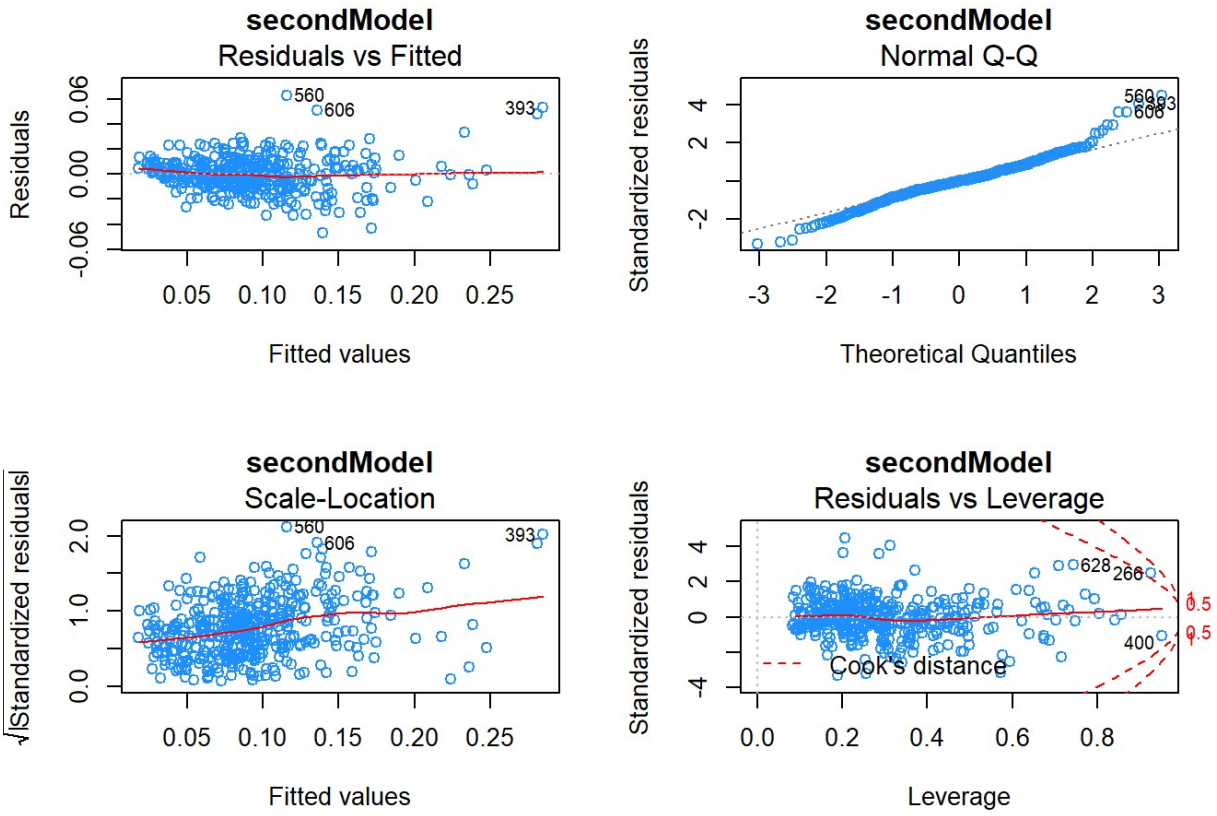
Depending upon our criteria, I have different winners. For `adj.r.squared` , its `secondLog` and for `LOOCV_RMSE` , its tie between `firstModel` and `thirdModel` . And even after considering 2-Way interactions; I still have multicollinearity issues.

Plots

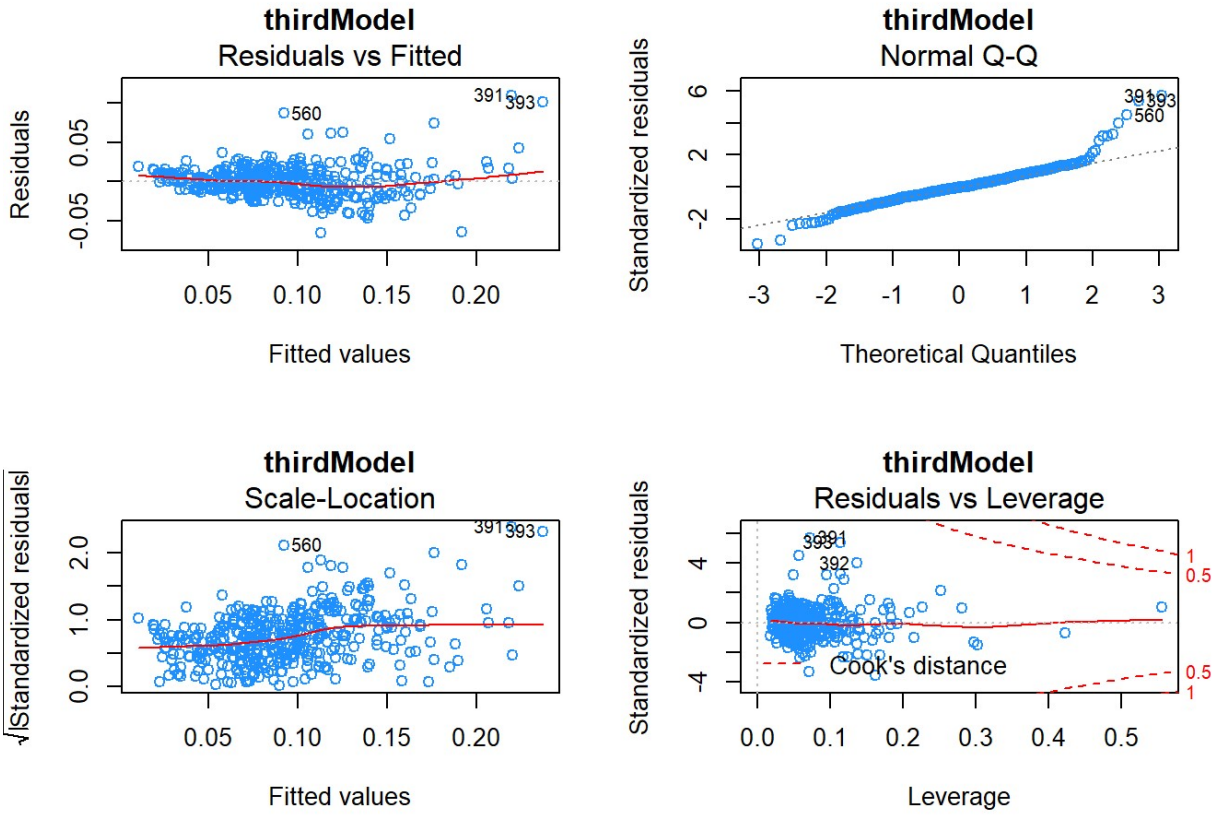
```
par(mfrow = c(2, 2))
plot(firstModel,main="firstModel",col='dodgerblue')
```



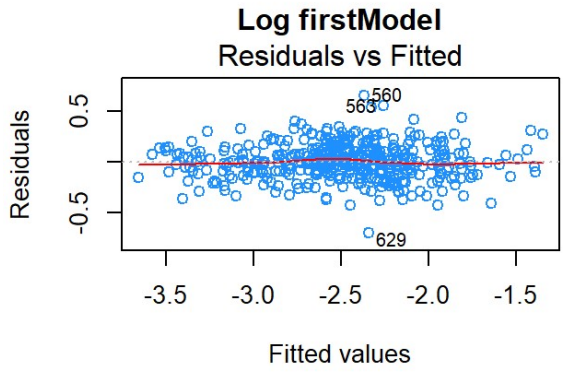

```
par(mfrow = c(2, 2))
plot(secondModel,main="secondModel",col='dodgerblue')
```

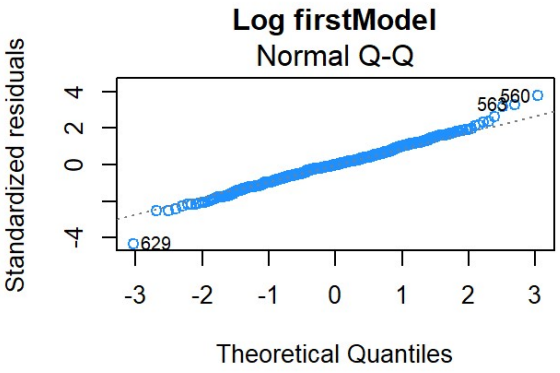


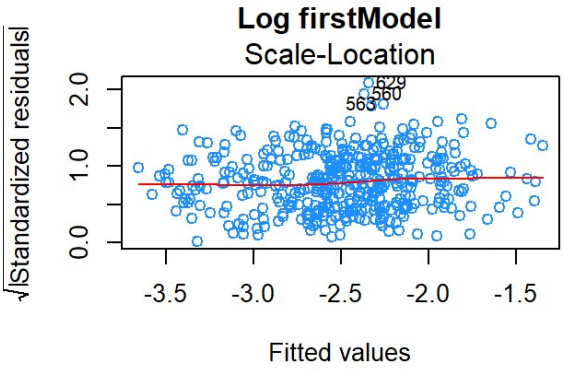

```
par(mfrow = c(2, 2))
plot(thirdModel,main="thirdModel",col='dodgerblue')
```

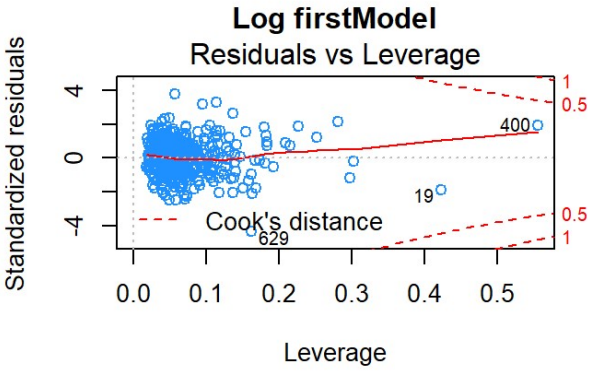



```
par(mfrow = c(2, 2))
plot(firstLog,main="Log firstModel",col='dodgerblue')
```

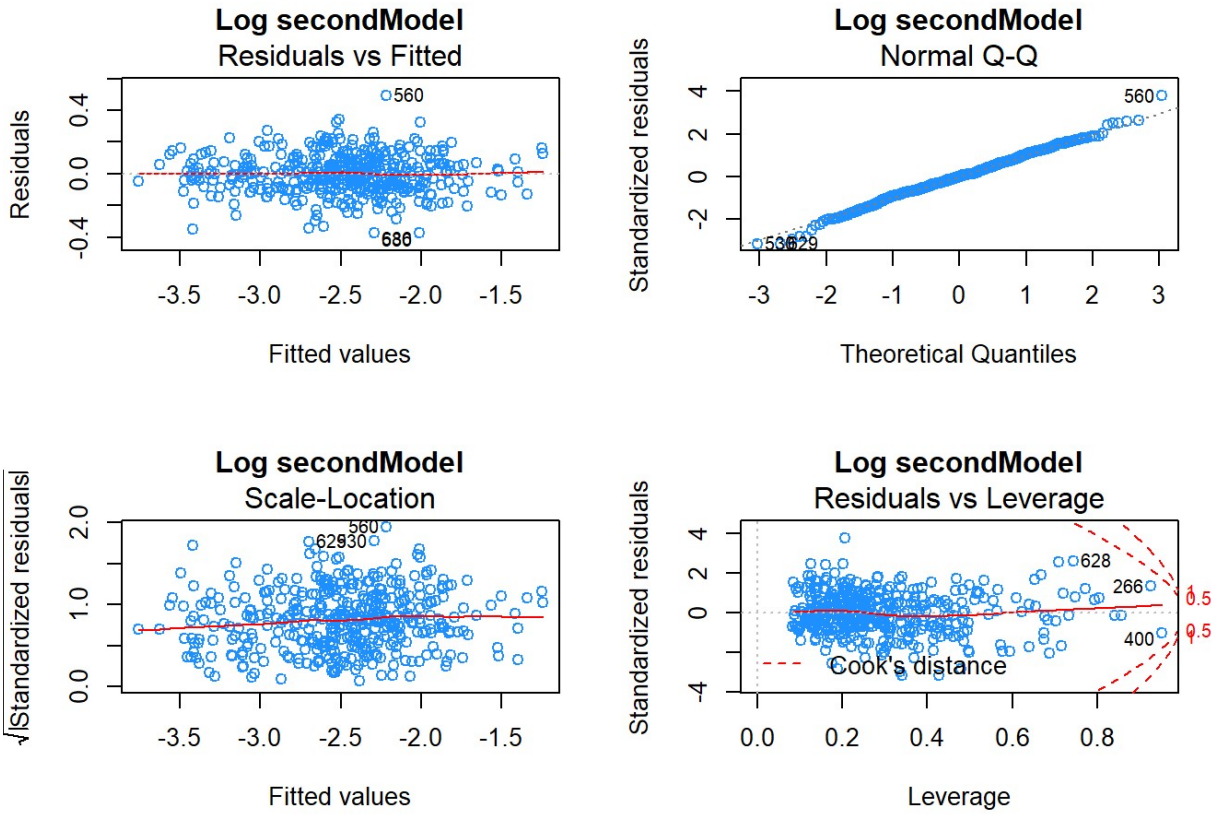




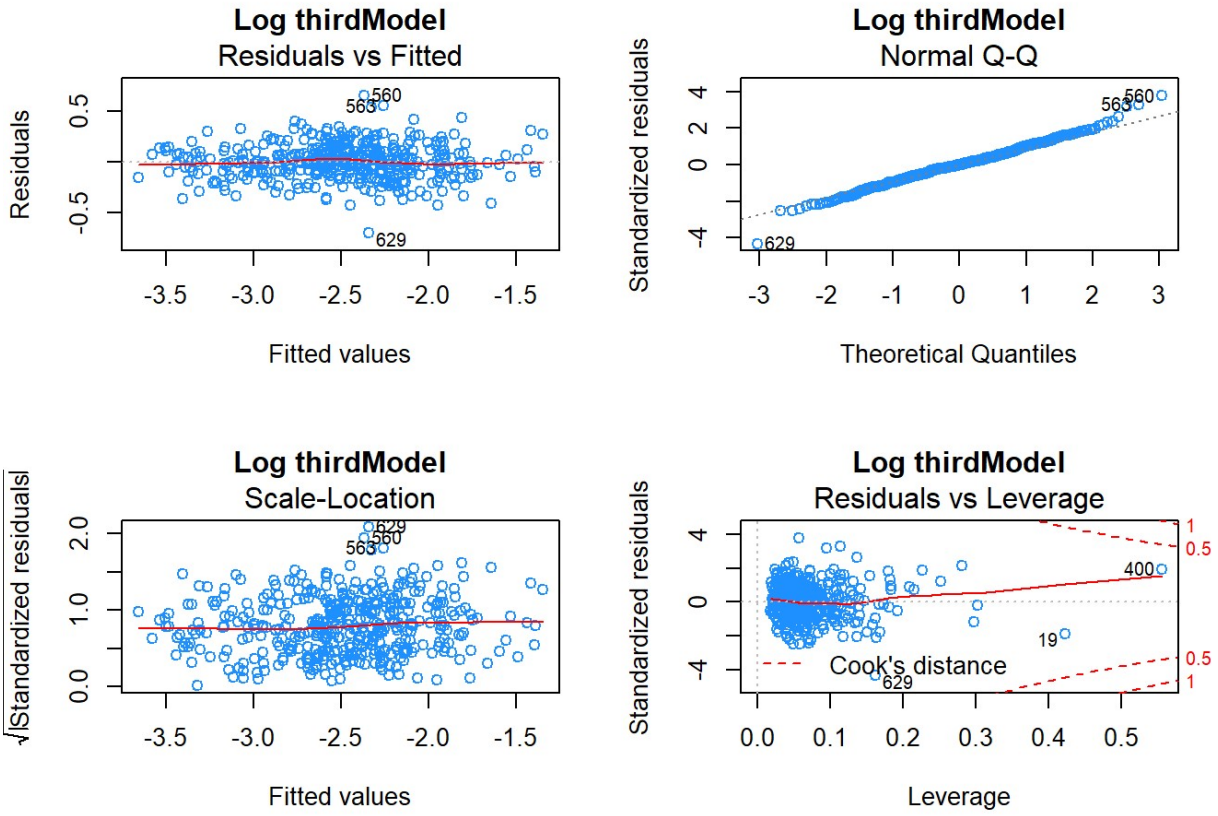





```
par(mfrow = c(2, 2))
plot(secondLog,main="Log secondModel",col='dodgerblue')
```




```
par(mfrow = c(2, 2))
plot(thirdLog,main="Log thirdModel",col='dodgerblue')
```



Results

I chose the `secondLog` model as our final one, even of its issues with multicollinearity, its still the best in terms of `Adjusted R2` value.

Lets see which predictors I finally have.

```
length(coef(secondLog))
```

```
## [1] 122
```

```
names(coef(secondLog))
```

```
##   [1] "(Intercept)"
##   [2] "Population"
##   [3] "Urban1"
##   [4] "Black"
##   [5] "Seg_racial"
##   [6] "Seg_income"
##   [7] "Seg_poverty"
##   [8] "Seg_affluence"
##   [9] "Commute"
##  [10] "Income"
##  [11] "Gini"
##  [12] "Share01"
##  [13] "Gini_99"
##  [14] "Middle_class"
##  [15] "Local_tax_rate"
##  [16] "Local_gov_spending"
##  [17] "Progressivity"
##  [18] "EITC"
##  [19] "School_spending"
##  [20] "Student_teacher_ratio"
##  [21] "Test_scores"
##  [22] "HS_dropout"
##  [23] "Colleges"
##  [24] "Tuition"
##  [25] "Graduation"
##  [26] "Labor_force_participation"
##  [27] "Manufacturing"
##  [28] "Chinese_imports"
##  [29] "Teenage_labor"
##  [30] "Migration_in"
##  [31] "Migration_out"
##  [32] "Foreign_born"
##  [33] "Social_capital"
##  [34] "Religious"
##  [35] "Violent_crime"
##  [36] "Single_mothers"
##  [37] "Divorced"
##  [38] "Married"
##  [39] "Longitude"
##  [40] "Latitude"
##  [41] "Middle_class:Teenage_labor"
##  [42] "Gini_99:Violent_crime"
##  [43] "Black:Test_scores"
##  [44] "Seg_affluence:Progressivity"
##  [45] "Seg_affluence:Migration_in"
##  [46] "Labor_force_participation:Longitude"
##  [47] "Seg_poverty:EITC"
##  [48] "Student_teacher_ratio:Longitude"
##  [49] "School_spending:Religious"
```

```
##  [50] "Local_tax_rate:Teenage_labor"
##  [51] "Seg_poverty:Middle_class"
##  [52] "Colleges:Latitude"
##  [53] "Tuition:Migration_out"
##  [54] "Gini:Divorced"
##  [55] "Seg_poverty:HS_dropout"
##  [56] "Local_tax_rate:Colleges"
##  [57] "Seg_affluence:Local_tax_rate"
##  [58] "Seg_racial:HS_dropout"
##  [59] "Gini:Local_gov_spending"
##  [60] "Progressivity:Religious"
##  [61] "Middle_class:HS_dropout"
##  [62] "Tuition:Latitude"
##  [63] "Commute:Progressivity"
##  [64] "Seg_racial:Progressivity"
##  [65] "Student_teacher_ratio:Religious"
##  [66] "Middle_class:Manufacturing"
##  [67] "HS_dropout:Divorced"
##  [68] "Test_scores:Religious"
##  [69] "Seg_income:Middle_class"
##  [70] "Migration_in:Migration_out"
##  [71] "Foreign_born:Latitude"
##  [72] "Migration_in:Social_capital"
##  [73] "School_spending:Test_scores"
##  [74] "Gini_99:Test_scores"
##  [75] "Commute:Student_teacher_ratio"
##  [76] "Commute:Migration_out"
##  [77] "Test_scores:Latitude"
##  [78] "Commute:Foreign_born"
##  [79] "Gini_99:Colleges"
##  [80] "Commute:Labor_force_participation"
##  [81] "Local_tax_rate:Progressivity"
##  [82] "Commute:Religious"
##  [83] "Local_tax_rate:Social_capital"
##  [84] "Seg_racial:Married"
##  [85] "Share01:Test_scores"
##  [86] "Income:Single_mothers"
##  [87] "Income:Colleges"
##  [88] "Middle_class:Foreign_born"
##  [89] "Seg_poverty:Married"
##  [90] "HS_dropout:Single_mothers"
##  [91] "Religious:Divorced"
##  [92] "Gini_99:Latitude"
##  [93] "Progressivity:School_spending"
##  [94] "Seg_racial:Colleges"
##  [95] "Commute:Test_scores"
##  [96] "Gini_99:HS_dropout"
##  [97] "Income:Social_capital"
##  [98] "Gini_99:Divorced"
```

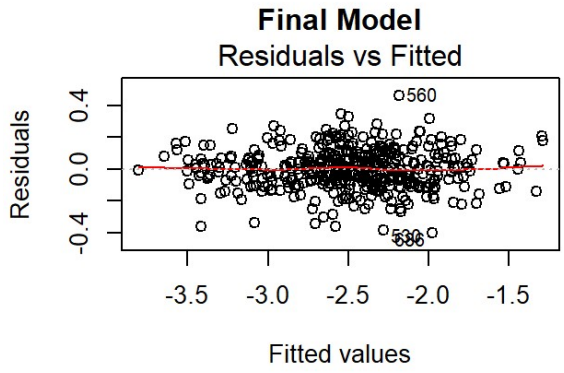
```
##  [99] "Population:Progressivity"
## [100] "Seg_poverty:Progressivity"
## [101] "Black:Divorced"
## [102] "Commute:Middle_class"
## [103] "Gini_99:Manufacturing"
## [104] "HS_dropout:Chinese_imports"
## [105] "Labor_force_participation:Teenage_labor"
## [106] "Progressivity:Colleges"
## [107] "Foreign_born:Social_capital"
## [108] "Seg_racial:Seg_income"
## [109] "Share01:HS_dropout"
## [110] "Middle_class:Latitude"
## [111] "Test_scores:Longitude"
## [112] "Gini:Latitude"
## [113] "Seg_income:Manufacturing"
## [114] "Test_scores:HS_dropout"
## [115] "Progressivity:Graduation"
## [116] "Middle_class:Violent_crime"
## [117] "Middle_class:Single_mothers"
## [118] "HS_dropout:Religious"
## [119] "Black:HS_dropout"
## [120] "Gini:Social_capital"
## [121] "Population:EITC"
## [122] "EITC:Colleges"
```

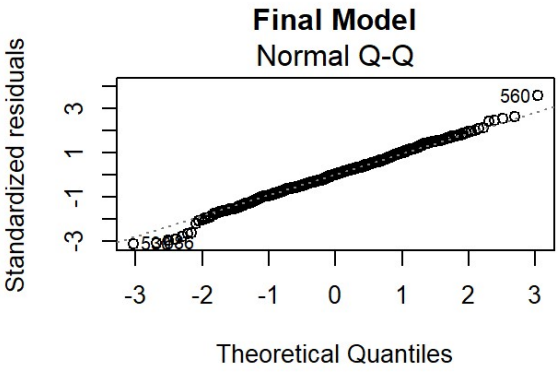
I found that its a highly interactive model. And looking at the significance level of each contributor I found that there's still hige scope for improvement. Lets do that one last time.

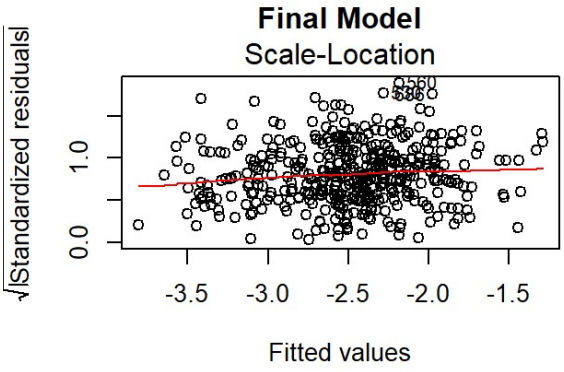
```
finalModel = step(secondLog,trace=0)
```

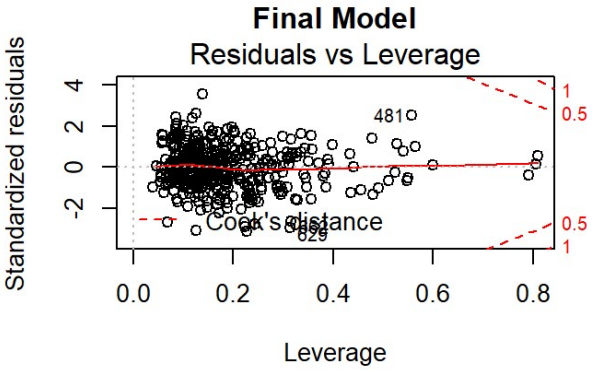
And see how it performs.

```
par(mfrow = c(2, 2))
plot(finalModel,main="Final Model")
```









##adj.r.squared

```
# Evaluation
summary(finalModel)$adj.r.squared
```

```
## [1] 0.9103
```

```
summary(finalModel)$adj.r.squared > summary(secondLog)$adj.r.squared
```

```
## [1] TRUE
```

##calc_loocv_rmse

```
calc_loocv_rmse(finalModel)
```

```
## [1] 0.1578
```

```
calc_loocv_rmse(finalModel)<calc_loocv_rmse(secondLog)
```

```
## [1] TRUE
```

##VIF

```
# VIF
sum(vif(finalModel)>5)/length(coef(finalModel))
```

```
## [1] 0.9221
```

```
sum(vif(finalModel)>5) < sum(vif(secondLog)>5)
```

```
## [1] TRUE
```

##AIC

```
#AIC
extractAIC(finalModel)
```

```
## [1]    77 -1574
```

```
extractAIC(secondLog)
```

```
## [1]   122 -1508
```

##BIC

```
#BIC
extractAIC(finalModel,k=log(nrow(dataset)))
```

```
## [1]    77 -1263
```

```
extractAIC(secondLog,k=log(nrow(dataset)))
```

```
## [1]   122 -1015
```

I found a better performing model.

Discussion

With the `finalModel` I have some observations.

`Commute` which I considered to be the most important predictor of economic mobility, is not significant considering the interactions.

```
coef(summary(finalModel))['Commute',]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##    -0.6405     1.3084    -0.4896     0.6247
```

The significant predictors.

```
sum(summary(finalModel)$coefficients[ ,4] < 0.05)/length(coef(finalModel))
```

```
## [1] 0.5844
```

```
summary(finalModel)$coefficients[summary(finalModel)$coefficients[ ,4] < 0.05,]
```

```
##                                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                            -3.312e+00  1.500e+00  -2.208 2.793e-02
## Population                              4.545e-08  2.185e-08   2.080 3.826e-02
## Black                                   2.777e+00  6.935e-01   4.005 7.612e-05
## Middle_class                            1.435e+00  4.040e-01   3.552 4.361e-04
## Local_tax_rate                          1.702e+01  5.767e+00   2.952 3.377e-03
## EITC                                   -1.071e-02  4.171e-03  -2.568 1.065e-02
## School_spending                         1.531e-01  3.823e-02   4.006 7.574e-05
## Student_teacher_ratio                   1.304e-01  4.720e-02   2.763 6.041e-03
## Test_scores                            -1.068e-01  3.071e-02  -3.477 5.727e-04
## Colleges                               -2.526e+00  6.032e-01  -4.188 3.581e-05
## Tuition                                -9.327e-05  2.275e-05  -4.100 5.164e-05
## Labor_force_participation              -7.322e+00  9.446e-01  -7.751 1.061e-13
## Manufacturing                          -1.688e+00  2.429e-01  -6.949 1.873e-11
## Teenage_labor                          -3.716e+02  9.945e+01  -3.737 2.184e-04
## Migration_out                           1.395e+01  6.814e+00   2.047 4.141e-02
## Violent_crime                           1.499e+02  6.849e+01   2.189 2.928e-02
## Single_mothers                         -3.749e+00  5.973e-01  -6.277 1.048e-09
## Divorced                               -1.393e+01  5.162e+00  -2.698 7.314e-03
## Latitude                                7.468e-02  1.766e-02   4.227 3.039e-05
## School_spending:Religious              -2.404e-01  6.394e-02  -3.759 2.007e-04
## Local_tax_rate:Teenage_labor           -4.140e+03  1.120e+03  -3.695 2.557e-04
## Seg_affluence:Local_tax_rate            1.362e+02  4.271e+01   3.189 1.558e-03
## Tuition:Latitude                        2.301e-06  5.729e-07   4.016 7.272e-05
## Seg_racial:Progressivity                1.759e-01  7.721e-02   2.279 2.330e-02
## Commute:Student_teacher_ratio          -1.312e-01  4.842e-02  -2.709 7.091e-03
## Commute:Migration_out                  -4.116e+01  1.351e+01  -3.048 2.486e-03
## Commute:Foreign_born                    6.662e+00  2.728e+00   2.442 1.510e-02
## Commute:Labor_force_participation       6.383e+00  1.907e+00   3.347 9.067e-04
## HS_dropout:Single_mothers               8.832e+01  1.336e+01   6.613 1.459e-10
## Religious:Divorced                      2.049e+01  4.339e+00   4.723 3.407e-06
## Gini_99:Latitude                       -2.732e-01  5.424e-02  -5.036 7.724e-07
## Commute:Test_scores                     3.769e-02  1.555e-02   2.423 1.592e-02
## Gini_99:HS_dropout                     -5.202e+01  1.358e+01  -3.832 1.514e-04
## Income:Social_capital                  -1.012e-05  2.103e-06  -4.811 2.254e-06
## Gini_99:Divorced                        2.881e+01  1.421e+01   2.028 4.330e-02
## Seg_poverty:Progressivity              -5.987e-01  2.756e-01  -2.172 3.053e-02
## Black:Divorced                         -3.485e+01  6.738e+00  -5.172 3.959e-07
## Labor_force_participation:Teenage_labor  7.265e+02  1.663e+02   4.368 1.665e-05
## Foreign_born:Social_capital            -8.313e-01  3.074e-01  -2.705 7.183e-03
## Seg_income:Manufacturing                1.503e+01  4.886e+00   3.077 2.262e-03
## Progressivity:Graduation               -1.639e-01  6.309e-02  -2.599 9.763e-03
## Black:HS_dropout                       -1.694e+01  5.706e+00  -2.968 3.205e-03
## Gini:Social_capital                     4.541e-01  1.558e-01   2.914 3.800e-03
## Population:EITC                         4.534e-09  2.301e-09   1.971 4.957e-02
## EITC:Colleges                           1.673e-01  8.013e-02   2.088 3.751e-02
```

Proportion of variation in `Mobility` explained by chosen predictors.

```
summary(finalModel)$r.squared
```

```
## [1] 0.9266
```

Really a better model than `secondLog`

```
analysis = anova(finalModel,secondLog,test="F")
analysis$`Pr(>F)`[2]
```

```
## [1] 0.9999
```

YES!! I fail to reject the NULL hypothesis that the smaller model is as good as the bigger one.

Appendix

Model Summaries
secondLog

```
summary(secondLog)
```

```
##
## Call:
## lm(formula = log(Mobility) ~ . + Middle_class:Teenage_labor +
##     Gini_99:Violent_crime + Black:Test_scores + Seg_affluence:Progressivity +
##     Seg_affluence:Migration_in + Labor_force_participation:Longitude +
##     Seg_poverty:EITC + Student_teacher_ratio:Longitude + School_spending:Religious +
##     Local_tax_rate:Teenage_labor + Seg_poverty:Middle_class +
##     Colleges:Latitude + Tuition:Migration_out + Gini:Divorced +
##     Seg_poverty:HS_dropout + Local_tax_rate:Colleges + Seg_affluence:Local_tax_rate +
##     Seg_racial:HS_dropout + Gini:Local_gov_spending + Progressivity:Religious +
##     Middle_class:HS_dropout + Tuition:Latitude + Commute:Progressivity +
##     Seg_racial:Progressivity + Student_teacher_ratio:Religious +
##     Middle_class:Manufacturing + HS_dropout:Divorced + Test_scores:Religious +
##     Seg_income:Middle_class + Migration_in:Migration_out + Foreign_born:Latitude +
##     Migration_in:Social_capital + School_spending:Test_scores +
##     Gini_99:Test_scores + Commute:Student_teacher_ratio + Commute:Migration_out +
##     Test_scores:Latitude + Commute:Foreign_born + Gini_99:Colleges +
##     Commute:Labor_force_participation + Local_tax_rate:Progressivity +
##     Commute:Religious + Local_tax_rate:Social_capital + Seg_racial:Married +
##     Share01:Test_scores + Income:Single_mothers + Income:Colleges +
##     Middle_class:Foreign_born + Seg_poverty:Married + HS_dropout:Single_mothers +
##     Religious:Divorced + Gini_99:Latitude + Progressivity:School_spending +
##     Seg_racial:Colleges + Commute:Test_scores + Gini_99:HS_dropout +
##     Income:Social_capital + Gini_99:Divorced + Population:Progressivity +
##     Seg_poverty:Progressivity + Black:Divorced + Commute:Middle_class +
##     Gini_99:Manufacturing + HS_dropout:Chinese_imports + Labor_force_participation:Te
enage_labor +
##     Progressivity:Colleges + Foreign_born:Social_capital + Seg_racial:Seg_income +
##     Share01:HS_dropout + Middle_class:Latitude + Test_scores:Longitude +
##     Gini:Latitude + Seg_income:Manufacturing + Test_scores:HS_dropout +
##     Progressivity:Graduation + Middle_class:Violent_crime + Middle_class:Single_mothe
rs +
##     HS_dropout:Religious + Black:HS_dropout + Gini:Social_capital +
##     Population:EITC + EITC:Colleges, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3740 -0.0838  0.0008  0.0769  0.4933
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           8.57e-01   3.79e+00    0.23  0.82116
## Population                            4.21e-08   3.09e-08    1.36  0.17414
## Urban1                               -3.21e-03   2.71e-02   -0.12  0.90592
## Black                                 2.57e+00   8.48e-01    3.04  0.00261
## Seg_racial                           -3.22e+00   2.05e+00   -1.57  0.11690
## Seg_income                           -2.60e+00   1.68e+01   -0.15  0.87717
## Seg_poverty                           8.62e+00   1.52e+01    0.57  0.57097
## Seg_affluence                         3.24e-01   3.63e+00    0.09  0.92901
```

```
## Commute                                  7.02e-01   2.20e+00    0.32  0.74984
## Income                                   2.18e-05   1.78e-05    1.22  0.22216
## Gini                                    -3.13e+01   2.21e+01   -1.42  0.15768
## Share01                                  3.07e-01   2.19e-01    1.40  0.16204
## Gini_99                                  3.43e+01   2.21e+01    1.55  0.12159
## Middle_class                            -2.57e+00   3.94e+00   -0.65  0.51374
## Local_tax_rate                           1.04e+01   9.04e+00    1.15  0.24980
## Local_gov_spending                       1.78e-04   1.13e-04    1.57  0.11764
## Progressivity                            9.83e-02   8.73e-02    1.13  0.26098
## EITC                                    -1.08e-02   8.05e-03   -1.34  0.17978
## School_spending                          1.39e-01   4.61e-02    3.01  0.00285
## Student_teacher_ratio                    1.24e-01   6.06e-02    2.04  0.04181
## Test_scores                             -7.98e-02   4.01e-02   -1.99  0.04780
## HS_dropout                              -1.03e+01   1.87e+01   -0.55  0.58235
## Colleges                                -8.78e+00   1.39e+01   -0.63  0.52758
## Tuition                                 -8.02e-05   2.64e-05   -3.03  0.00264
## Graduation                               7.40e-02   9.97e-02    0.74  0.45845
## Labor_force_participation               -8.63e+00   2.18e+00   -3.96 0.000094
## Manufacturing                            5.47e-02   3.51e+00    0.02  0.98758
## Chinese_imports                         -1.27e-05   6.13e-03    0.00  0.99834
## Teenage_labor                           -4.27e+02   1.41e+02   -3.03  0.00264
## Migration_in                            -4.28e-01   3.79e+00   -0.11  0.91022
## Migration_out                            1.58e+01   8.06e+00    1.97  0.05024
## Foreign_born                            -8.78e-01   5.09e+00   -0.17  0.86302
## Social_capital                           1.97e-01   1.54e-01    1.28  0.20195
## Religious                               -1.83e-01   1.31e+00   -0.14  0.88918
## Violent_crime                            1.07e+02   2.38e+02    0.45  0.65352
## Single_mothers                          -3.28e+00   3.16e+00   -1.04  0.30032
## Divorced                                -1.63e+01   6.89e+00   -2.37  0.01847
## Married                                 -2.83e-01   7.49e-01   -0.38  0.70525
## Longitude                                1.83e-04   1.85e-02    0.01  0.99210
## Latitude                                -1.33e-02   7.73e-02   -0.17  0.86335
## Middle_class:Teenage_labor               4.63e+01   2.19e+02    0.21  0.83281
## Gini_99:Violent_crime                    1.75e+00   3.57e+02    0.00  0.99610
## Black:Test_scores                        1.23e-02   1.80e-02    0.68  0.49744
## Seg_affluence:Progressivity              2.43e-01   1.04e+00    0.23  0.81507
## Seg_affluence:Migration_in              -5.35e+01   4.47e+01   -1.20  0.23163
## Labor_force_participation:Longitude     -1.62e-02   2.67e-02   -0.61  0.54426
## Seg_poverty:EITC                         8.07e-02   1.30e-01    0.62  0.53447
## Student_teacher_ratio:Longitude          6.26e-04   6.23e-04    1.00  0.31599
## School_spending:Religious               -1.99e-01   8.09e-02   -2.45  0.01470
## Local_tax_rate:Teenage_labor            -2.71e+03   1.67e+03   -1.62  0.10564
## Seg_poverty:Middle_class                -2.20e+00   2.84e+01   -0.08  0.93838
## Colleges:Latitude                        1.25e-01   1.68e-01    0.74  0.45755
## Tuition:Migration_out                   -3.16e-04   4.09e-04   -0.77  0.43949
## Gini:Divorced                           -7.74e+00   1.93e+01   -0.40  0.68880
## Seg_poverty:HS_dropout                   1.01e+01   2.43e+01    0.42  0.67828
## Local_tax_rate:Colleges                 -1.10e+01   9.41e+01   -0.12  0.90706
## Seg_affluence:Local_tax_rate             1.09e+02   4.91e+01    2.22  0.02687
```

```
## Seg_racial:HS_dropout                      1.91e+00   6.98e+00    0.27  0.78403
## Gini:Local_gov_spending                   -4.04e-04   2.73e-04   -1.48  0.13913
## Progressivity:Religious                    8.17e-02   5.70e-02    1.43  0.15278
## Middle_class:HS_dropout                    2.37e+00   1.99e+01    0.12  0.90537
## Tuition:Latitude                           2.07e-06   6.62e-07    3.13  0.00190
## Commute:Progressivity                      3.73e-02   1.25e-01    0.30  0.76557
## Seg_racial:Progressivity                   2.38e-01   9.95e-02    2.40  0.01718
## Student_teacher_ratio:Religious            4.35e-03   4.54e-02    0.10  0.92373
## Middle_class:Manufacturing                -8.69e-01   3.99e+00   -0.22  0.82774
## HS_dropout:Divorced                        6.20e+01   4.62e+01    1.34  0.18067
## Test_scores:Religious                      1.64e-02   1.36e-02    1.20  0.23114
## Seg_income:Middle_class                   -2.26e+00   2.66e+01   -0.08  0.93251
## Migration_in:Migration_out                 2.07e+02   1.24e+02    1.67  0.09506
## Foreign_born:Latitude                     -4.90e-02   1.06e-01   -0.46  0.64548
## Migration_in:Social_capital                2.02e+00   1.43e+00    1.41  0.15819
## School_spending:Test_scores                1.42e-03   2.07e-03    0.68  0.49396
## Gini_99:Test_scores                        3.21e-02   5.43e-02    0.59  0.55431
## Commute:Student_teacher_ratio             -1.37e-01   6.28e-02   -2.18  0.02999
## Commute:Migration_out                     -4.56e+01   1.63e+01   -2.79  0.00557
## Test_scores:Latitude                       5.07e-04   5.42e-04    0.93  0.35100
## Commute:Foreign_born                       5.48e+00   3.61e+00    1.52  0.12952
## Gini_99:Colleges                           3.68e+00   2.07e+01    0.18  0.85942
## Commute:Labor_force_participation          5.85e+00   2.96e+00    1.97  0.04925
## Local_tax_rate:Progressivity              -1.53e+00   2.01e+00   -0.76  0.44544
## Commute:Religious                         -2.58e-01   8.31e-01   -0.31  0.75660
## Local_tax_rate:Social_capital             -7.79e-01   2.15e+00   -0.36  0.71710
## Seg_racial:Married                         5.15e+00   3.38e+00    1.52  0.12887
## Share01:Test_scores                       -7.00e-04   5.26e-04   -1.33  0.18422
## Income:Single_mothers                     -7.28e-05   8.41e-05   -0.87  0.38709
## Income:Colleges                            3.30e-05   1.51e-04    0.22  0.82722
## Middle_class:Foreign_born                  1.13e+00   6.94e+00    0.16  0.87089
## Seg_poverty:Married                       -1.15e+01   1.36e+01   -0.85  0.39870
## HS_dropout:Single_mothers                  8.54e+01   2.44e+01    3.50  0.00055
## Religious:Divorced                         1.90e+01   5.18e+00    3.68  0.00028
## Gini_99:Latitude                          -2.26e-01   1.16e-01   -1.95  0.05199
## Progressivity:School_spending             -1.59e-02   1.13e-02   -1.41  0.15956
## Seg_racial:Colleges                       -5.43e+00   6.26e+00   -0.87  0.38621
## Commute:Test_scores                        2.18e-02   2.37e-02    0.92  0.35869
## Gini_99:HS_dropout                        -5.53e+01   2.34e+01   -2.36  0.01892
## Income:Social_capital                     -1.20e-05   2.98e-06   -4.01 0.000077
## Gini_99:Divorced                           4.70e+01   2.72e+01    1.73  0.08493
## Population:Progressivity                   1.67e-08   1.67e-08    1.00  0.31867
## Seg_poverty:Progressivity                 -1.06e+00   1.25e+00   -0.84  0.39899
## Black:Divorced                            -3.11e+01   8.07e+00   -3.86  0.00014
## Commute:Middle_class                      -1.03e+00   2.90e+00   -0.35  0.72302
## Gini_99:Manufacturing                     -3.78e+00   4.77e+00   -0.79  0.42814
## HS_dropout:Chinese_imports                 3.03e-01   2.98e-01    1.02  0.30974
## Labor_force_participation:Teenage_labor    7.24e+02   2.07e+02    3.50  0.00054
## Progressivity:Colleges                     1.02e-01   7.05e-01    0.15  0.88460
```

```
## Foreign_born:Social_capital             -6.89e-01   5.69e-01   -1.21  0.22665
## Seg_racial:Seg_income                   -2.32e+00   4.87e+00   -0.48  0.63453
## Share01:HS_dropout                       1.69e-01   1.60e-01    1.05  0.29310
## Middle_class:Latitude                    9.04e-02   8.97e-02    1.01  0.31422
## Test_scores:Longitude                   -3.10e-04   2.13e-04   -1.46  0.14575
## Gini:Latitude                            5.75e-02   8.05e-02    0.71  0.47545
## Seg_income:Manufacturing                 1.11e+01   6.25e+00    1.77  0.07767
## Test_scores:HS_dropout                   1.85e-03   1.25e-01    0.01  0.98824
## Progressivity:Graduation                -1.35e-01   7.29e-02   -1.84  0.06613
## Middle_class:Violent_crime              -1.58e+02   2.58e+02   -0.61  0.54093
## Middle_class:Single_mothers              3.62e+00   4.73e+00    0.76  0.44500
## HS_dropout:Religious                    -1.85e+00   4.81e+00   -0.39  0.70042
## Black:HS_dropout                        -1.73e+01   8.17e+00   -2.12  0.03524
## Gini:Social_capital                      4.90e-01   2.35e-01    2.08  0.03806
## Population:EITC                          3.38e-09   3.79e-09    0.89  0.37414
## EITC:Colleges                            1.03e-01   1.55e-01    0.66  0.50894
##
## (Intercept)
## Population
## Urban1
## Black                                   **
## Seg_racial
## Seg_income
## Seg_poverty
## Seg_affluence
## Commute
## Income
## Gini
## Share01
## Gini_99
## Middle_class
## Local_tax_rate
## Local_gov_spending
## Progressivity
## EITC
## School_spending                         **
## Student_teacher_ratio                   *
## Test_scores                             *
## HS_dropout
## Colleges
## Tuition                                 **
## Graduation
## Labor_force_participation               ***
## Manufacturing
## Chinese_imports
## Teenage_labor                           **
## Migration_in
## Migration_out                           .
## Foreign_born
```

```
## Social_capital
## Religious
## Violent_crime
## Single_mothers
## Divorced                                *
## Married
## Longitude
## Latitude
## Middle_class:Teenage_labor
## Gini_99:Violent_crime
## Black:Test_scores
## Seg_affluence:Progressivity
## Seg_affluence:Migration_in
## Labor_force_participation:Longitude
## Seg_poverty:EITC
## Student_teacher_ratio:Longitude
## School_spending:Religious               *
## Local_tax_rate:Teenage_labor
## Seg_poverty:Middle_class
## Colleges:Latitude
## Tuition:Migration_out
## Gini:Divorced
## Seg_poverty:HS_dropout
## Local_tax_rate:Colleges
## Seg_affluence:Local_tax_rate           *
## Seg_racial:HS_dropout
## Gini:Local_gov_spending
## Progressivity:Religious
## Middle_class:HS_dropout
## Tuition:Latitude                        **
## Commute:Progressivity
## Seg_racial:Progressivity                *
## Student_teacher_ratio:Religious
## Middle_class:Manufacturing
## HS_dropout:Divorced
## Test_scores:Religious
## Seg_income:Middle_class
## Migration_in:Migration_out              .
## Foreign_born:Latitude
## Migration_in:Social_capital
## School_spending:Test_scores
## Gini_99:Test_scores
## Commute:Student_teacher_ratio           *
## Commute:Migration_out                   **
## Test_scores:Latitude
## Commute:Foreign_born
## Gini_99:Colleges
## Commute:Labor_force_participation       *
## Local_tax_rate:Progressivity
```

```
## Commute:Religious
## Local_tax_rate:Social_capital
## Seg_racial:Married
## Share01:Test_scores
## Income:Single_mothers
## Income:Colleges
## Middle_class:Foreign_born
## Seg_poverty:Married
## HS_dropout:Single_mothers               ***
## Religious:Divorced                      ***
## Gini_99:Latitude                        .
## Progressivity:School_spending
## Seg_racial:Colleges
## Commute:Test_scores
## Gini_99:HS_dropout                      *
## Income:Social_capital                   ***
## Gini_99:Divorced                        .
## Population:Progressivity
## Seg_poverty:Progressivity
## Black:Divorced                          ***
## Commute:Middle_class
## Gini_99:Manufacturing
## HS_dropout:Chinese_imports
## Labor_force_participation:Teenage_labor ***
## Progressivity:Colleges
## Foreign_born:Social_capital
## Seg_racial:Seg_income
## Share01:HS_dropout
## Middle_class:Latitude
## Test_scores:Longitude
## Gini:Latitude
## Seg_income:Manufacturing                .
## Test_scores:HS_dropout
## Progressivity:Graduation                .
## Middle_class:Violent_crime
## Middle_class:Single_mothers
## HS_dropout:Religious
## Black:HS_dropout                        *
## Gini:Social_capital                     *
## Population:EITC
## EITC:Colleges
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.146 on 296 degrees of freedom
## Multiple R-squared:  0.931,  Adjusted R-squared:  0.902
## F-statistic: 32.9 on 121 and 296 DF,  p-value: <2e-16
```

finalModel

```
summary(finalModel)
```

```
##
## Call:
## lm(formula = log(Mobility) ~ Population + Black + Seg_racial +
##     Seg_income + Seg_poverty + Seg_affluence + Commute + Income +
##     Gini + Share01 + Gini_99 + Middle_class + Local_tax_rate +
##     Local_gov_spending + Progressivity + EITC + School_spending +
##     Student_teacher_ratio + Test_scores + HS_dropout + Colleges +
##     Tuition + Graduation + Labor_force_participation + Manufacturing +
##     Teenage_labor + Migration_in + Migration_out + Foreign_born +
##     Social_capital + Religious + Violent_crime + Single_mothers +
##     Divorced + Married + Longitude + Latitude + Student_teacher_ratio:Longitude +
##     School_spending:Religious + Local_tax_rate:Teenage_labor +
##     Seg_affluence:Local_tax_rate + Gini:Local_gov_spending +
##     Progressivity:Religious + Tuition:Latitude + Seg_racial:Progressivity +
##     HS_dropout:Divorced + Test_scores:Religious + Migration_in:Migration_out +
##     Migration_in:Social_capital + School_spending:Test_scores +
##     Gini_99:Test_scores + Commute:Student_teacher_ratio + Commute:Migration_out +
##     Test_scores:Latitude + Commute:Foreign_born + Commute:Labor_force_participation +
##     Seg_racial:Married + HS_dropout:Single_mothers + Religious:Divorced +
##     Gini_99:Latitude + Commute:Test_scores + Gini_99:HS_dropout +
##     Income:Social_capital + Gini_99:Divorced + Seg_poverty:Progressivity +
##     Black:Divorced + Labor_force_participation:Teenage_labor +
##     Foreign_born:Social_capital + Test_scores:Longitude + Seg_income:Manufacturing +
##     Progressivity:Graduation + Middle_class:Violent_crime + Black:HS_dropout +
##     Gini:Social_capital + Population:EITC + EITC:Colleges, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4022 -0.0782  0.0012  0.0781  0.4631
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -3.31e+00   1.50e+00   -2.21  0.02793
## Population                            4.54e-08   2.18e-08    2.08  0.03826
## Black                                 2.78e+00   6.93e-01    4.01  7.6e-05
## Seg_racial                           -2.45e+00   1.28e+00   -1.91  0.05650
## Seg_income                           -3.84e+00   5.72e+00   -0.67  0.50242
## Seg_poverty                           7.57e-01   3.02e+00    0.25  0.80221
## Seg_affluence                        -2.06e+00   3.01e+00   -0.68  0.49414
## Commute                              -6.41e-01   1.31e+00   -0.49  0.62475
## Income                                5.83e-06   4.45e-06    1.31  0.19144
## Gini                                 -2.87e+01   1.93e+01   -1.49  0.13785
## Share01                               2.96e-01   1.93e-01    1.53  0.12649
## Gini_99                               3.65e+01   1.95e+01    1.87  0.06165
## Middle_class                          1.43e+00   4.04e-01    3.55  0.00044
## Local_tax_rate                        1.70e+01   5.77e+00    2.95  0.00338
## Local_gov_spending                    1.61e-04   9.43e-05    1.71  0.08887
## Progressivity                         2.14e-03   3.01e-02    0.07  0.94331
## EITC                                 -1.07e-02   4.17e-03   -2.57  0.01065
```

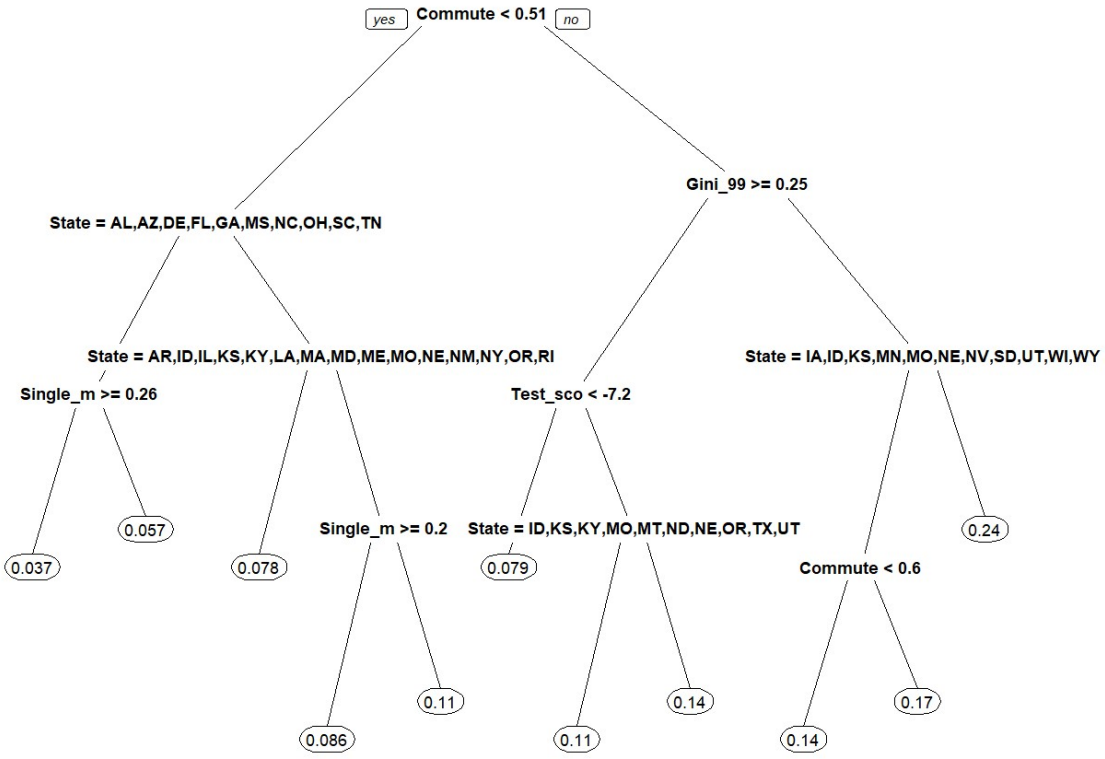
```
## School_spending                          1.53e-01   3.82e-02    4.01  7.6e-05
## Student_teacher_ratio                    1.30e-01   4.72e-02    2.76  0.00604
## Test_scores                             -1.07e-01   3.07e-02   -3.48  0.00057
## HS_dropout                              -9.68e+00   5.07e+00   -1.91  0.05707
## Colleges                                -2.53e+00   6.03e-01   -4.19  3.6e-05
## Tuition                                 -9.33e-05   2.27e-05   -4.10  5.2e-05
## Graduation                               6.86e-02   8.90e-02    0.77  0.44107
## Labor_force_participation               -7.32e+00   9.45e-01   -7.75  1.1e-13
## Manufacturing                           -1.69e+00   2.43e-01   -6.95  1.9e-11
## Teenage_labor                           -3.72e+02   9.94e+01   -3.74  0.00022
## Migration_in                            -2.75e+00   2.99e+00   -0.92  0.35819
## Migration_out                            1.40e+01   6.81e+00    2.05  0.04141
## Foreign_born                            -2.47e+00   1.43e+00   -1.73  0.08414
## Social_capital                           1.45e-01   1.07e-01    1.36  0.17620
## Religious                               -1.07e-01   5.73e-01   -0.19  0.85123
## Violent_crime                            1.50e+02   6.85e+01    2.19  0.02928
## Single_mothers                          -3.75e+00   5.97e-01   -6.28  1.0e-09
## Divorced                                -1.39e+01   5.16e+00   -2.70  0.00731
## Married                                 -4.14e-01   5.37e-01   -0.77  0.44138
## Longitude                               -1.08e-02   9.03e-03   -1.20  0.23236
## Latitude                                 7.47e-02   1.77e-02    4.23  3.0e-05
## Student_teacher_ratio:Longitude          6.73e-04   4.98e-04    1.35  0.17742
## School_spending:Religious               -2.40e-01   6.39e-02   -3.76  0.00020
## Local_tax_rate:Teenage_labor            -4.14e+03   1.12e+03   -3.70  0.00026
## Seg_affluence:Local_tax_rate             1.36e+02   4.27e+01    3.19  0.00156
## Gini:Local_gov_spending                 -3.85e-04   2.25e-04   -1.71  0.08869
## Progressivity:Religious                  8.60e-02   4.54e-02    1.90  0.05879
## Tuition:Latitude                         2.30e-06   5.73e-07    4.02  7.3e-05
## Seg_racial:Progressivity                 1.76e-01   7.72e-02    2.28  0.02330
## HS_dropout:Divorced                      7.10e+01   3.82e+01    1.86  0.06408
## Test_scores:Religious                    1.91e-02   9.72e-03    1.96  0.05029
## Migration_in:Migration_out               1.83e+02   1.05e+02    1.74  0.08199
## Migration_in:Social_capital              1.73e+00   1.19e+00    1.45  0.14759
## School_spending:Test_scores              2.38e-03   1.65e-03    1.44  0.15031
## Gini_99:Test_scores                      5.35e-02   3.94e-02    1.36  0.17550
## Commute:Student_teacher_ratio           -1.31e-01   4.84e-02   -2.71  0.00709
## Commute:Migration_out                   -4.12e+01   1.35e+01   -3.05  0.00249
## Test_scores:Latitude                     6.71e-04   4.25e-04    1.58  0.11516
## Commute:Foreign_born                     6.66e+00   2.73e+00    2.44  0.01510
## Commute:Labor_force_participation        6.38e+00   1.91e+00    3.35  0.00091
## Seg_racial:Married                       3.57e+00   2.21e+00    1.62  0.10617
## HS_dropout:Single_mothers                8.83e+01   1.34e+01    6.61  1.5e-10
## Religious:Divorced                       2.05e+01   4.34e+00    4.72  3.4e-06
## Gini_99:Latitude                        -2.73e-01   5.42e-02   -5.04  7.7e-07
## Commute:Test_scores                      3.77e-02   1.56e-02    2.42  0.01592
## Gini_99:HS_dropout                      -5.20e+01   1.36e+01   -3.83  0.00015
## Income:Social_capital                   -1.01e-05   2.10e-06   -4.81  2.3e-06
## Gini_99:Divorced                         2.88e+01   1.42e+01    2.03  0.04330
## Seg_poverty:Progressivity               -5.99e-01   2.76e-01   -2.17  0.03053
```

```
## Black:Divorced                            -3.48e+01   6.74e+00   -5.17  4.0e-07
## Labor_force_participation:Teenage_labor    7.27e+02   1.66e+02    4.37  1.7e-05
## Foreign_born:Social_capital               -8.31e-01   3.07e-01   -2.70  0.00718
## Test_scores:Longitude                     -2.53e-04   1.73e-04   -1.46  0.14390
## Seg_income:Manufacturing                   1.50e+01   4.89e+00    3.08  0.00226
## Progressivity:Graduation                  -1.64e-01   6.31e-02   -2.60  0.00976
## Middle_class:Violent_crime                -2.36e+02   1.30e+02   -1.82  0.06989
## Black:HS_dropout                          -1.69e+01   5.71e+00   -2.97  0.00321
## Gini:Social_capital                        4.54e-01   1.56e-01    2.91  0.00380
## Population:EITC                            4.53e-09   2.30e-09    1.97  0.04957
## EITC:Colleges                              1.67e-01   8.01e-02    2.09  0.03751
##
## (Intercept)                               *
## Population                                *
## Black                                     ***
## Seg_racial                                .
## Seg_income
## Seg_poverty
## Seg_affluence
## Commute
## Income
## Gini
## Share01
## Gini_99                                   .
## Middle_class                              ***
## Local_tax_rate                            **
## Local_gov_spending                        .
## Progressivity
## EITC                                      *
## School_spending                           ***
## Student_teacher_ratio                     **
## Test_scores                               ***
## HS_dropout                                .
## Colleges                                  ***
## Tuition                                   ***
## Graduation
## Labor_force_participation                 ***
## Manufacturing                             ***
## Teenage_labor                             ***
## Migration_in
## Migration_out                             *
## Foreign_born                              .
## Social_capital
## Religious
## Violent_crime                             *
## Single_mothers                            ***
## Divorced                                  **
## Married
## Longitude
```

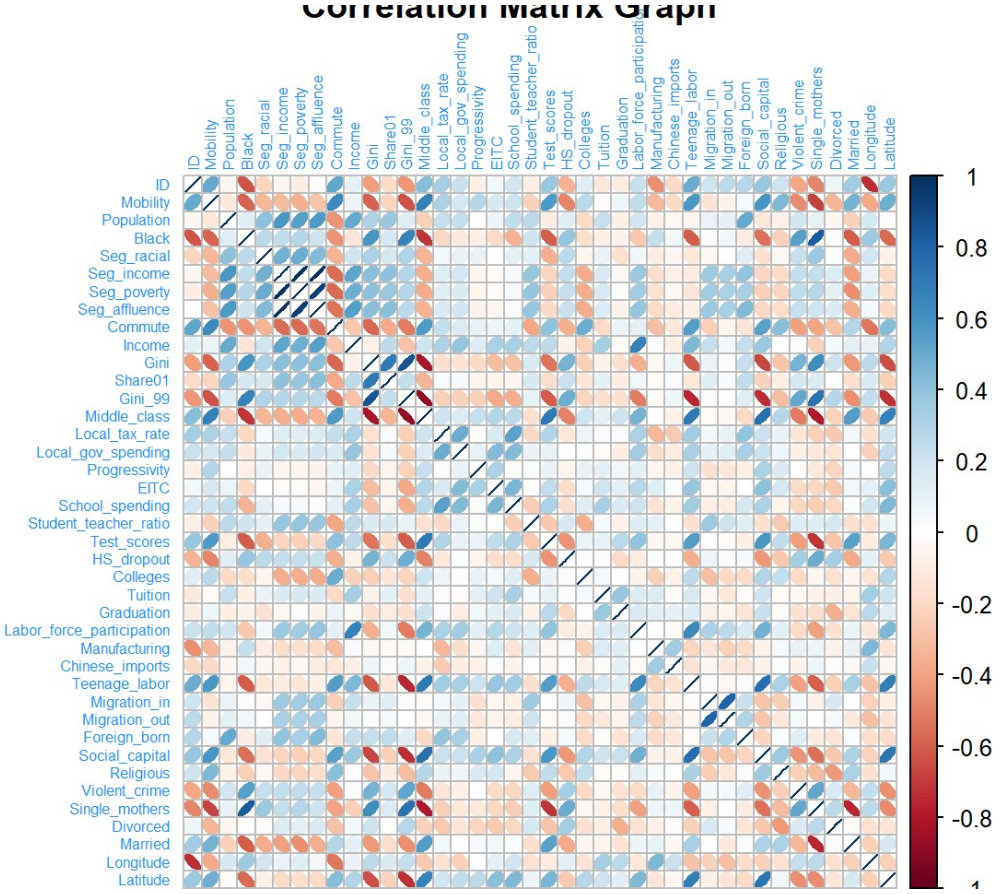
```
## Latitude                                    ***
## Student_teacher_ratio:Longitude
## School_spending:Religious                   ***
## Local_tax_rate:Teenage_labor                ***
## Seg_affluence:Local_tax_rate                **
## Gini:Local_gov_spending                     .
## Progressivity:Religious                     .
## Tuition:Latitude                            ***
## Seg_racial:Progressivity                    *
## HS_dropout:Divorced                         .
## Test_scores:Religious                       .
## Migration_in:Migration_out                  .
## Migration_in:Social_capital
## School_spending:Test_scores
## Gini_99:Test_scores
## Commute:Student_teacher_ratio               **
## Commute:Migration_out                       **
## Test_scores:Latitude
## Commute:Foreign_born                        *
## Commute:Labor_force_participation           ***
## Seg_racial:Married
## HS_dropout:Single_mothers                   ***
## Religious:Divorced                          ***
## Gini_99:Latitude                            ***
## Commute:Test_scores                         *
## Gini_99:HS_dropout                          ***
## Income:Social_capital                       ***
## Gini_99:Divorced                            *
## Seg_poverty:Progressivity                   *
## Black:Divorced                              ***
## Labor_force_participation:Teenage_labor     ***
## Foreign_born:Social_capital                 **
## Test_scores:Longitude
## Seg_income:Manufacturing                    **
## Progressivity:Graduation                    **
## Middle_class:Violent_crime                  .
## Black:HS_dropout                            **
## Gini:Social_capital                         **
## Population:EITC                             *
## EITC:Colleges                               *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.14 on 341 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.91
## F-statistic: 56.7 on 76 and 341 DF,  p-value: <2e-16
```

A model with `States`

```
# Dropping unique IDs and Names; and also States (too many levels.)

drops <- c("Name", "ID")
dataset = mobilityData[ , !(names(mobilityData) %in% drops)]

# Lets see interactions in a tree model
form <- as.formula(Mobility ~ .)
model <- rpart(form,data=dataset)
prp(model)
```




```
# Correlations
data = na.omit(mobility)
#round(cor(data[sapply(data,is.numeric)], use="pairwise.complete.obs"),2)
corrplot(cor(data[sapply(data,is.numeric)]),method ="ellipse",
       title =" Correlation Matrix Graph",tl.cex = .5,tl.pos ="lt",tl.col ="dodgerblu
e" )
```




```
  # State model
   modelState = lm(Mobility~Population+Black+Urban+State+Seg_racial+Seg_income+Seg_pover
  ty+Seg_affluence+Commute+Income+Gini+Share01+Gini_99+Middle_class+Local_tax_rate+Local
  _gov_spending+Progressivity+EITC+School_spending+Student_teacher_ratio+Test_scores+HS_
  dropout+Colleges+Tuition+Graduation+Labor_force_participation+Manufacturing+Chinese_im
  ports+Teenage_labor+Migration_in+Migration_out+Foreign_born+Social_capital+Religious+V
  iolent_crime+Single_mothers+Divorced+Married+Longitude+Latitude+I(Population^2)+I(Blac
  k^2)+I(Seg_racial^2)+I(Seg_income^2)+I(Seg_poverty^2)+I(Seg_affluence^2)+I(Commute^
  2)+I(Income^2)+I(Gini^2)+I(Share01^2)+I(Gini_99^2)+I(Middle_class^2)+I(Local_tax_rate^
  2)+I(Local_gov_spending^2)+I(Progressivity^2)+I(EITC^2)+I(School_spending^2)+I(Student
  _teacher_ratio^2)+I(Test_scores^2)+I(HS_dropout^2)+I(Colleges^2)+I(Tuition^2)+I(Gradua
  tion^2)+I(Labor_force_participation^2)+I(Manufacturing^2)+I(Chinese_imports^2)+I(Teena
  ge_labor^2)+I(Migration_in^2)+I(Migration_out^2)+I(Foreign_born^2)+I(Social_capital^
  2)+I(Religious^2)+I(Violent_crime^2)+I(Single_mothers^2)+I(Divorced^2)+I(Married^2)+I
  (Longitude^2)+I(Latitude^2),data=dataset)

# Significant ones
 summary(modelState)$coefficients[summary(modelState)$coefficients[ ,4] < 0.05,]
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.9469495 0.46561787   2.034 0.042854
## StateID         -0.0656679 0.02511683  -2.614 0.009386
## StateOR         -0.0659371 0.02949013  -2.236 0.026091
## Seg_poverty     -1.3854550 0.66473818  -2.084 0.037984
## Commute         -0.2491659 0.10790298  -2.309 0.021611
## Latitude        -0.0153889 0.00714860  -2.153 0.032135
## I(Seg_racial^2) -0.3031704 0.10762885  -2.817 0.005171
## I(Commute^2)     0.3239426 0.10399229   3.115 0.002016
## I(HS_dropout^2)  4.7448992 2.17346628   2.183 0.029801
## I(Latitude^2)    0.0002159 0.00009572   2.255 0.024840
```

Only Two states are significant; thus I am okay dropping them in the beginning.