Economic mobility across generations in USA

Final Exam for Data-based statistical decision model

Kyung Yoon Lee

Sep 07 2018

Data description.

We will look at economic mobility across generations in the 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, we don't have that individual-level data, but we do have aggregate statistics about economic mobility for several hundred communities, containing most of the American population, and covariate information about those communities. We are interested in predicting economic mobility from the characteristics of communities.

```
dat <- read.csv("mobility.csv")
attach(dat)
```

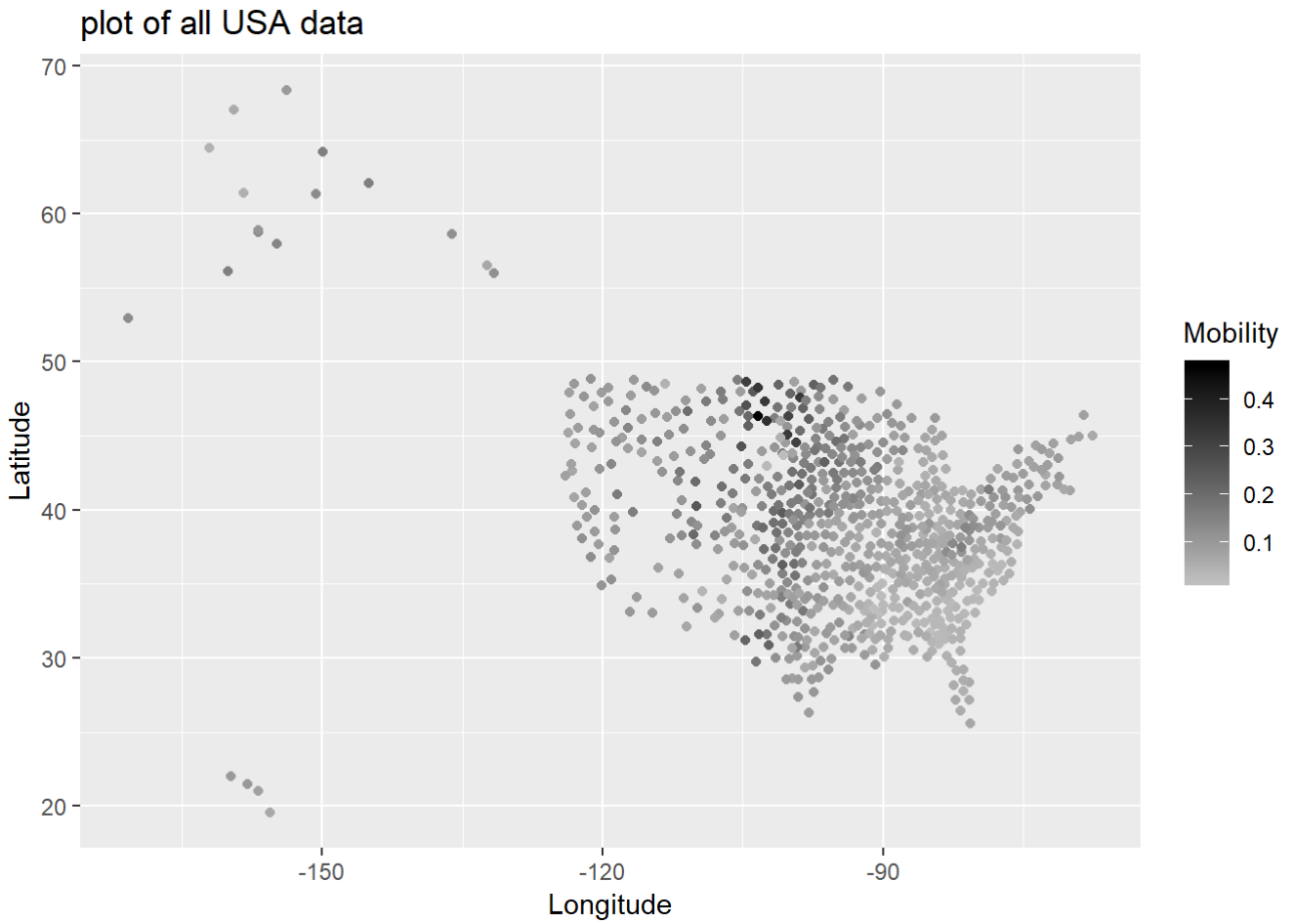
```
library(ggplot2)
library(gridExtra)
library(maps)
library(mapdata)
library(dplyr)
library(car)
library(rosm)
library(prettymapr)
library(boot)
library(data.table)
library(splines)
```

1. A map of mobility

a. Make a plot where the x and y coordinates are longitude and latitude, and mobility is indicated by color (possibly grey scale), by a third coordinate, or some other suitable device. Make sure your map is legible. Describe the geographic pattern in words.

Geographicly, north central and south central USA seems to have high Mobility. Also some points with high mobility points can be detected in Alaska. However, it is difficult to get some idea with this plot since there are too many points and color which seem similar. Thus, i will further discuss about this plot with question number 1-b.

```
p <- ggplot(dat, aes(Longitude,Latitude, color = Mobility)) + geom_point() + scale_color_gradient(low="grey", high="black") +
  labs(title = "plot of all USA data")
p
```



since it was difficult to interpret the location by the plot I have made above, I have used rosm package to point out the locations into the real-scale map. This is only to get the idea of the exact location.

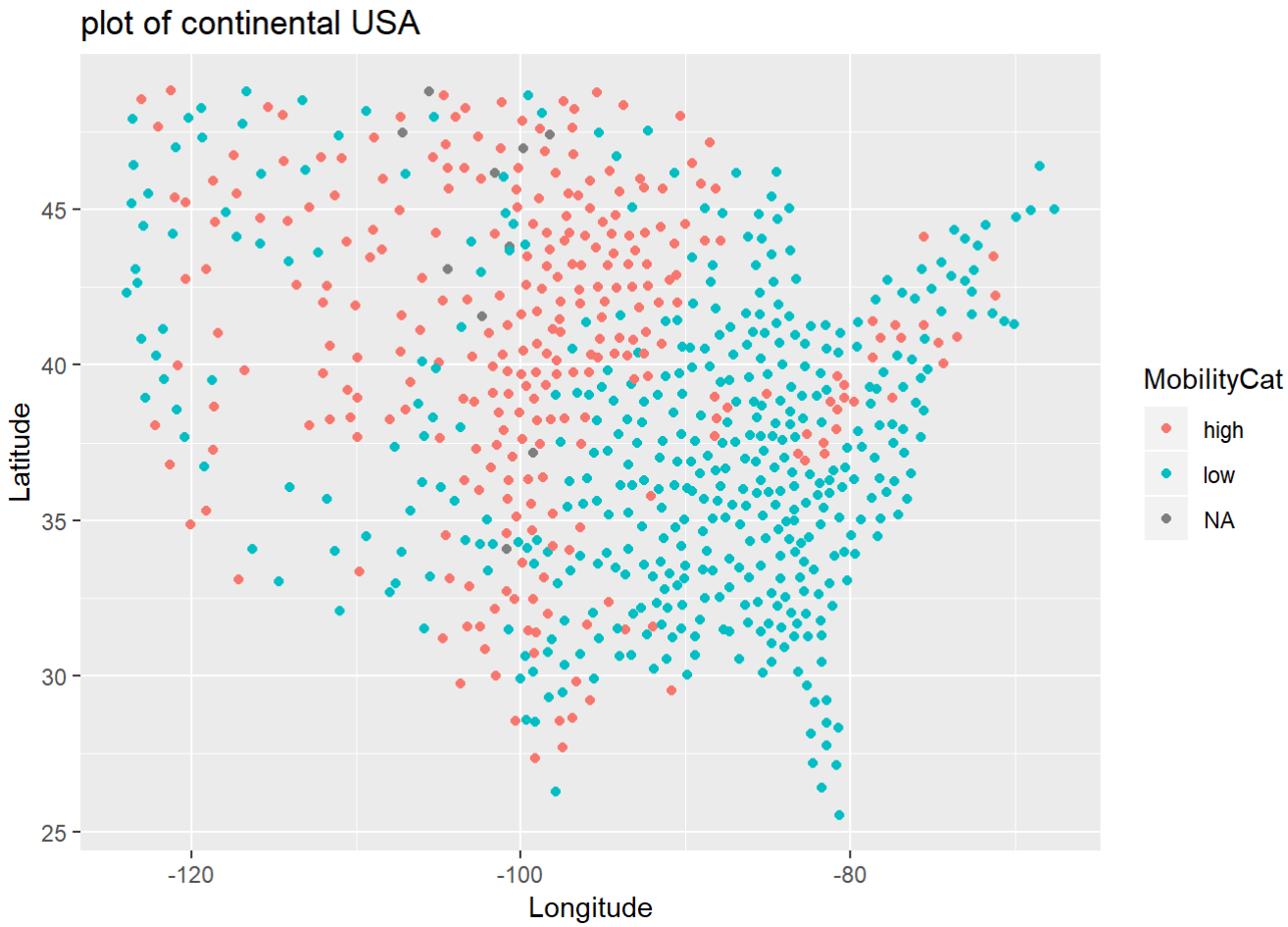
inference : https://github.com/paleolimbot/rosm (https://github.com/paleolimbot/rosm)

```
map1 <- makebbox(70, -60, 15, -175)
prettymap({
  bmaps.plot(map1)
  osm.points(Longitude, Latitude, pch=1, cex=0.6, col = 'white')})
```

b. Discretizing the Mobility values may enhance visualizing. Create a new variable, called MobilityCat with values high if Mobility > 0.1, and low otherwise. Make a plot where the x and y coordinates are longitude and latitude, and the categorized mobility (i.e. MobilityCat) is indicated by color. This time, filter your observations so that only the continental part of USA is visible (that is, remove data corresponding to Alaska and Hawaii). Has the geographic pattern become clearer?

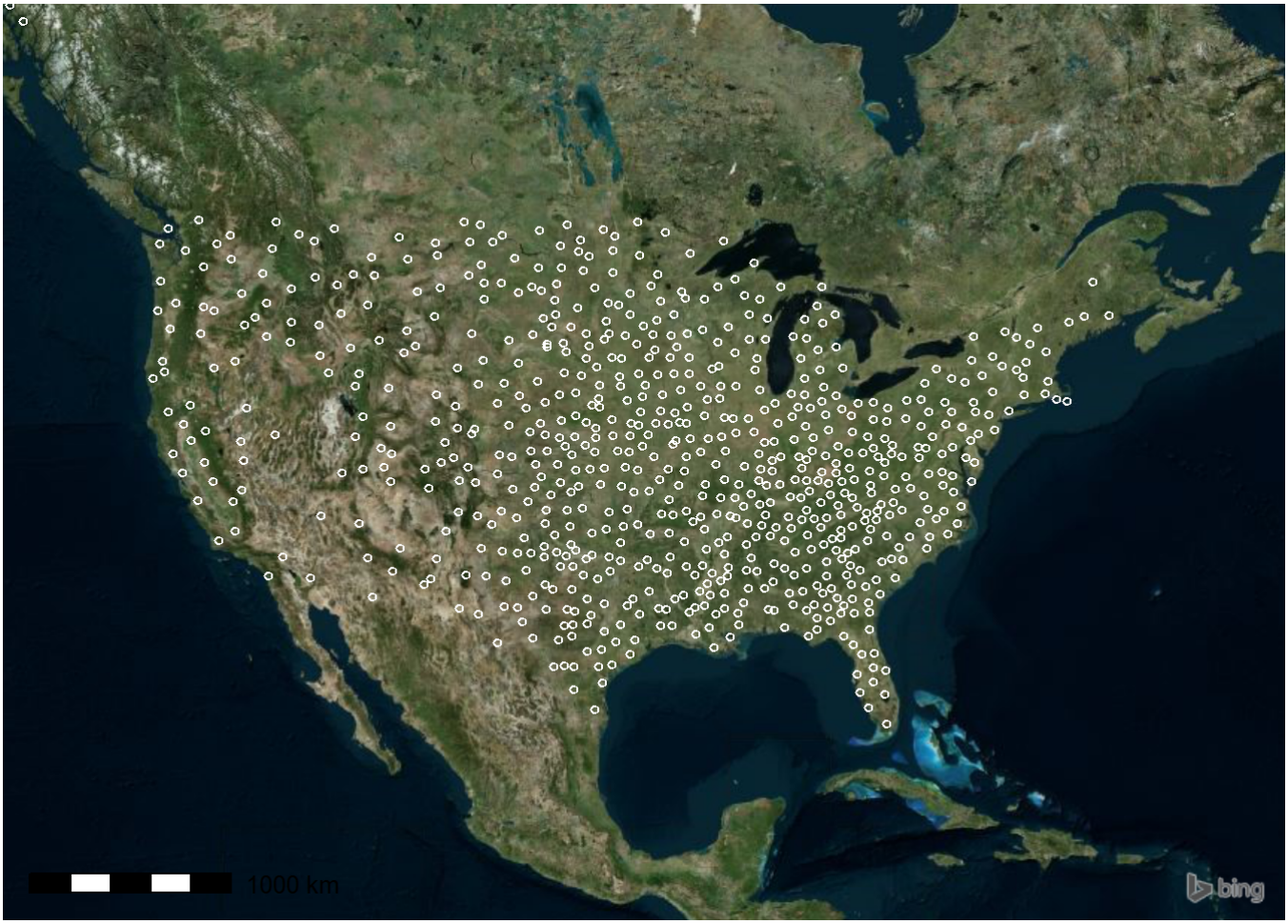
By discretizing the Mobility values, as it can be seen in the plot below, I was able to enhance visibility by making a new category variable. By using filter() function, I have zoomed into only continental part of USA, excluding Alaska and Hawaii region in my plot. By making this plot, i was able to conclude that north central and south central USA has high mobility ratio. Comparing east and west, it was obvious that west USA has hight mobility ratio than east. Most of the cities located in east USA seemed as a tomb for social level escalation.

```
dat_map <- dat %>% mutate(Mobility, MobilityCat = ifelse(Mobility > 0.1, "high", "low"))
dat_map <- dat_map %>% filter(Longitude > -150, Latitude < 50)
ggplot(dat_map, aes(x = dat_map$Longitude, y = dat_map$Latitude, col = dat_map$MobilityCat)) + geom_po
int() +
  labs(x = "Longitude", y = "Latitude", colour = "MobilityCat", title = "plot of continental USA")
```



Again, i have made a map with plots overlapped to see if plots are on the continental USA correctly.

```
map2 <- makebbox(50, -60, 25, -130)
prettymap({
  bmaps.plot(map2)
  osm.points(Longitude, Latitude, pch=1, cex=0.6, col = 'white')})
```



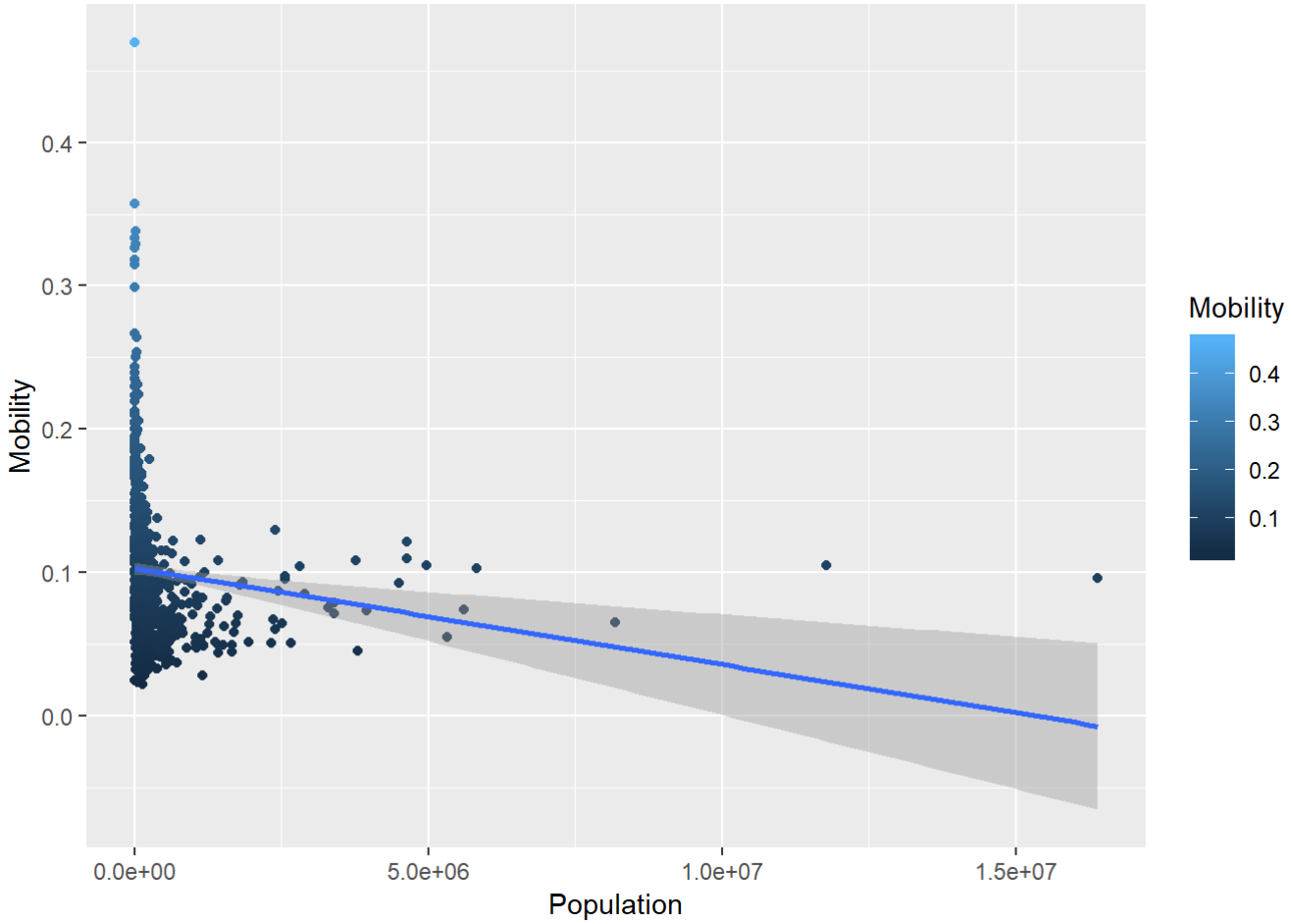
2. A bunch of simple regression model

Make scatter plots of mobility against each of the following variables. Include on each plot a line for the simple or univariate regression, and give a table of the regression coefficients. Carefully explain the interpretation of each coefficient. Do any of the results seem odd?

a. Population

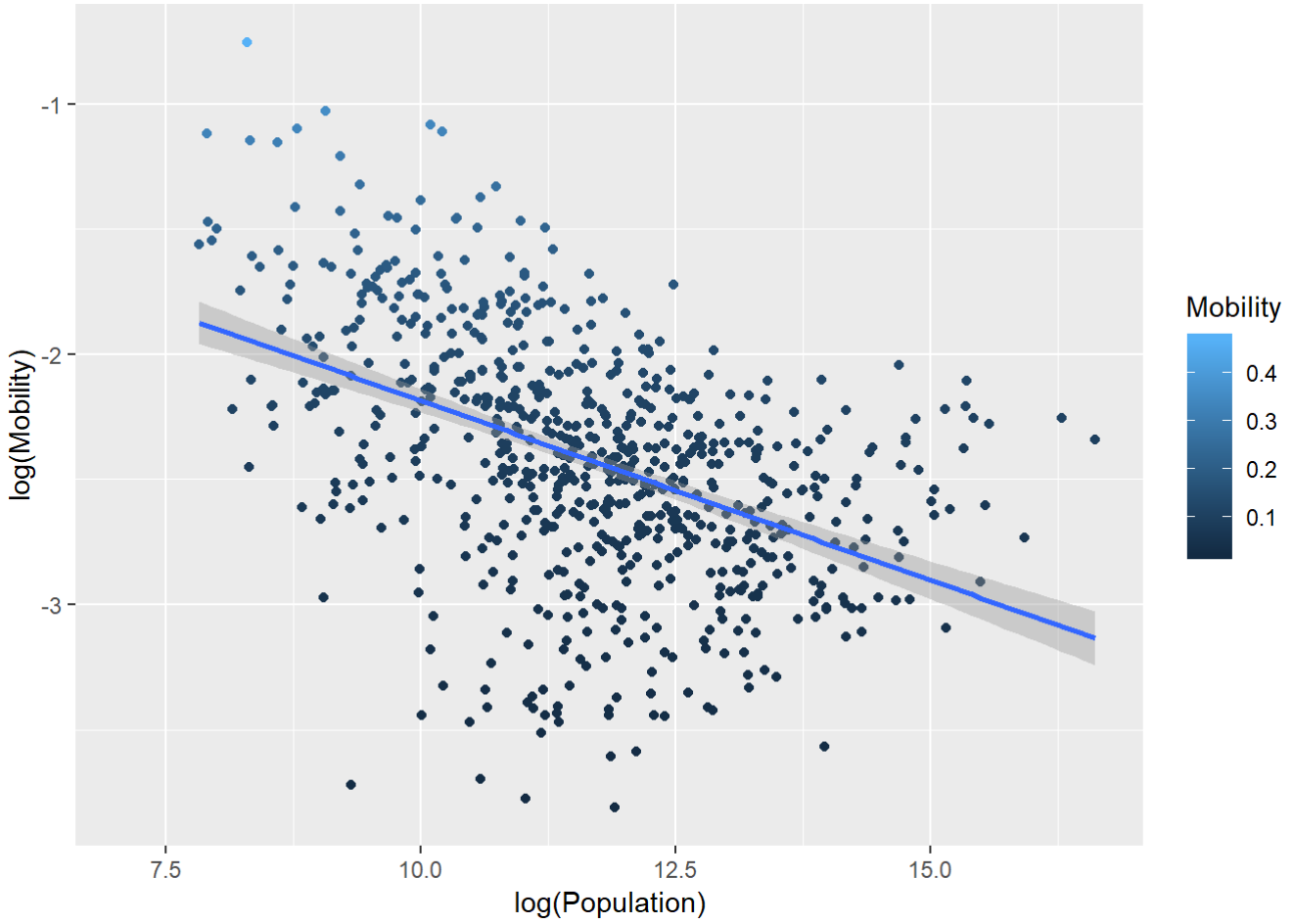
It was difficult to look at the relationship between mobility and population since most of the data was located in the left side of the scatter plot. Thus, to make this plot more visible, I have used log() function to spread out the data and look at the relationship.

```
ggplot(dat_map, aes(Population,Mobility, color = Mobility)) + geom_point() + geom_smooth(method = lm)
```



By doing so, I was able to see that there is a negative linear relathionship between these two varaibles. As population gets larger, mobility gets smaller.

```
ggplot(dat_map, aes(log(Population),log(Mobility), color = Mobility)) + geom_point() + geom_smooth(met
hod = lm)
```



As it was seen in the plot, the summary of linear model shows that population has negative relationship with Mobility by -6.733e-09. Population seems as a relevent variable since it's p-value is small. However, since its R square value is only about 0.016, i should see another variables as well in predicting mobility.

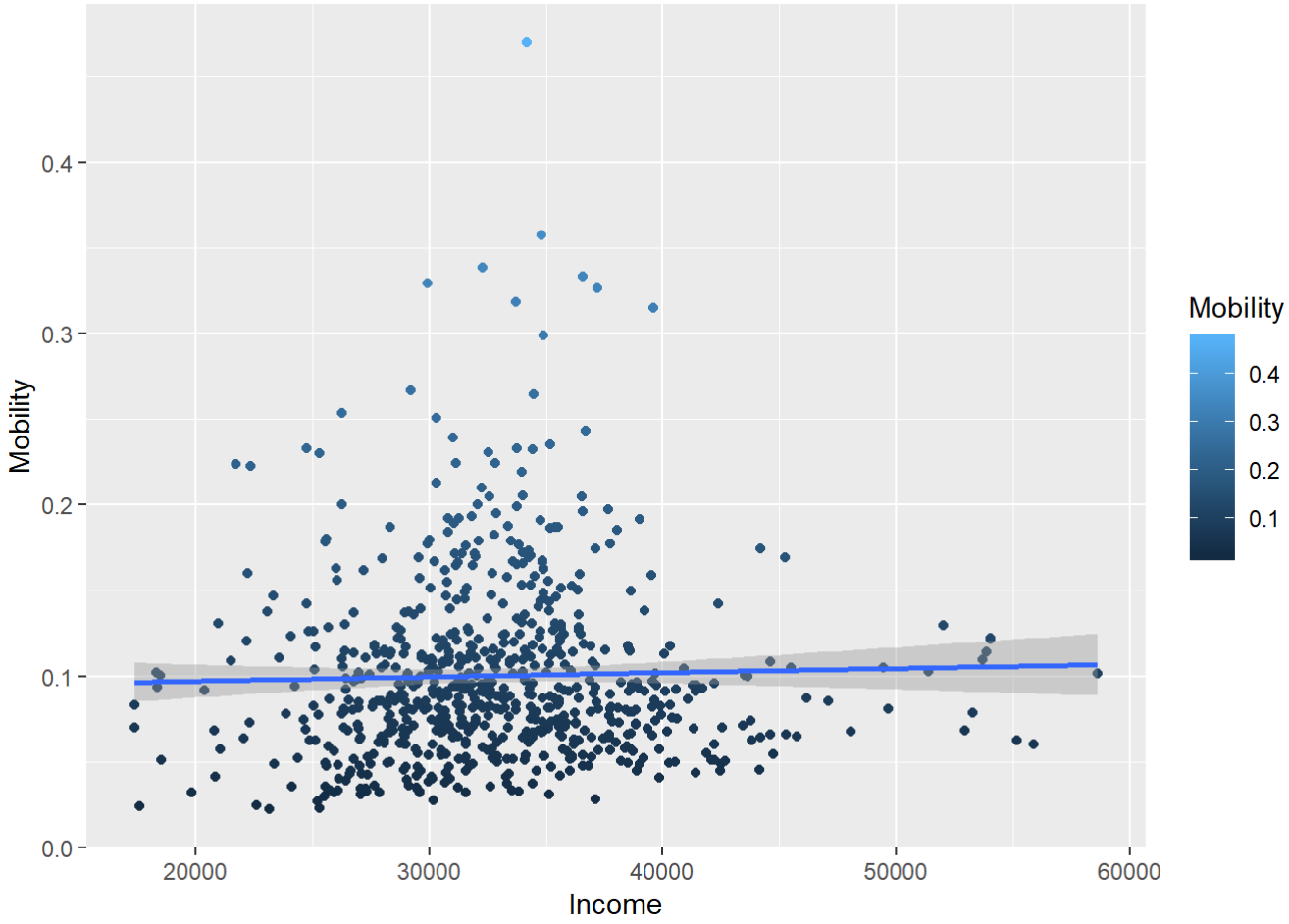
```
a <- lm(Mobility ~ Population, data = dat_map)
summary(a)
```

```
##
## Call:
## lm(formula = Mobility ~ Population, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07987 -0.03368 -0.01038  0.01846  0.36676
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.030e-01  2.099e-03  49.056  < 2e-16 ***
## Population  -6.733e-09  1.849e-09  -3.642 0.000291 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05255 on 710 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.01834,    Adjusted R-squared:  0.01695
## F-statistic: 13.26 on 1 and 710 DF,  p-value: 0.0002906
```

b. Mean household income per capita

The relationship between income and mobility cannot be explaned in linear model. It seems as there is no relationship between these two variables since the line is flat.

```
ggplot(dat_map, aes(Income,Mobility, color = Mobility)) + geom_point() + geom_smooth(method = lm)
```



By taking a look at summary, i was able to conclude that my prediction was correct since p-value is high enough to say there is no relationship.

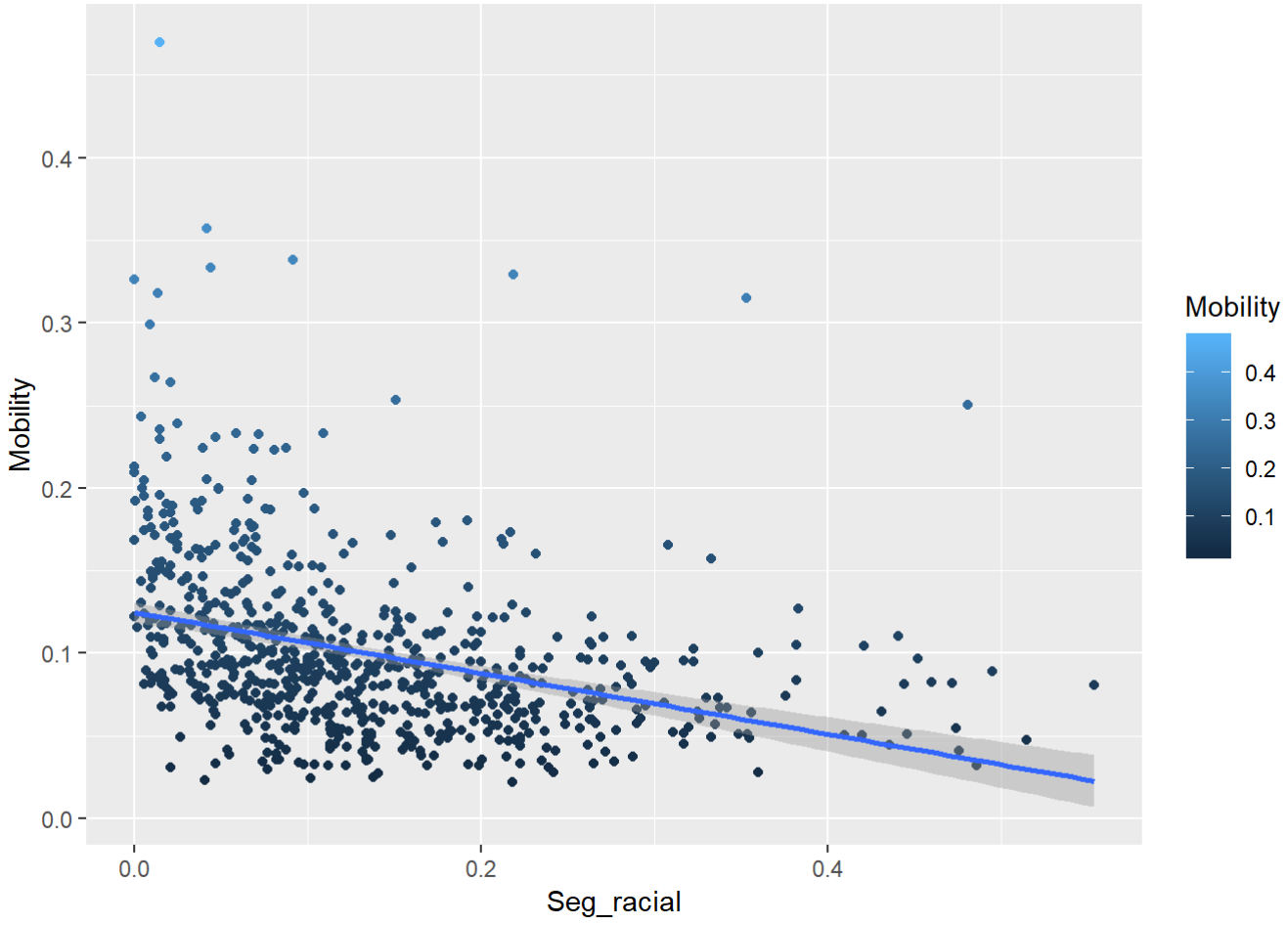
```
b <- lm(Mobility ~ Income, data = dat_map)
summary(b)
```

```
##
## Call:
## lm(formula = Mobility ~ Income, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07583 -0.03477 -0.01086  0.01784  0.36904
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.219e-02  1.158e-02   7.963 6.68e-15 ***
## Income      2.477e-07  3.475e-07   0.713    0.476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05302 on 710 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.0007149,  Adjusted R-squared:  -0.0006926
## F-statistic: 0.5079 on 1 and 710 DF,  p-value: 0.4763
```

c. Racial segregation

With racial segregation, Mobility seems to have negative relationship. As the measure of residential segregation by race increases, the moblity gets lower. Perhaps this is because the residential segregation by race means inequal chance of education by race.

```
ggplot(dat_map, aes(Seg_racial,Mobility, color = Mobility)) + geom_point() + geom_smooth(method = lm)
```



With summary function, I was able to conlcude that seg_racial has negative relationship with Mobility. P-value was small enough to conclude that this is meaningful data in explaining Mobility.

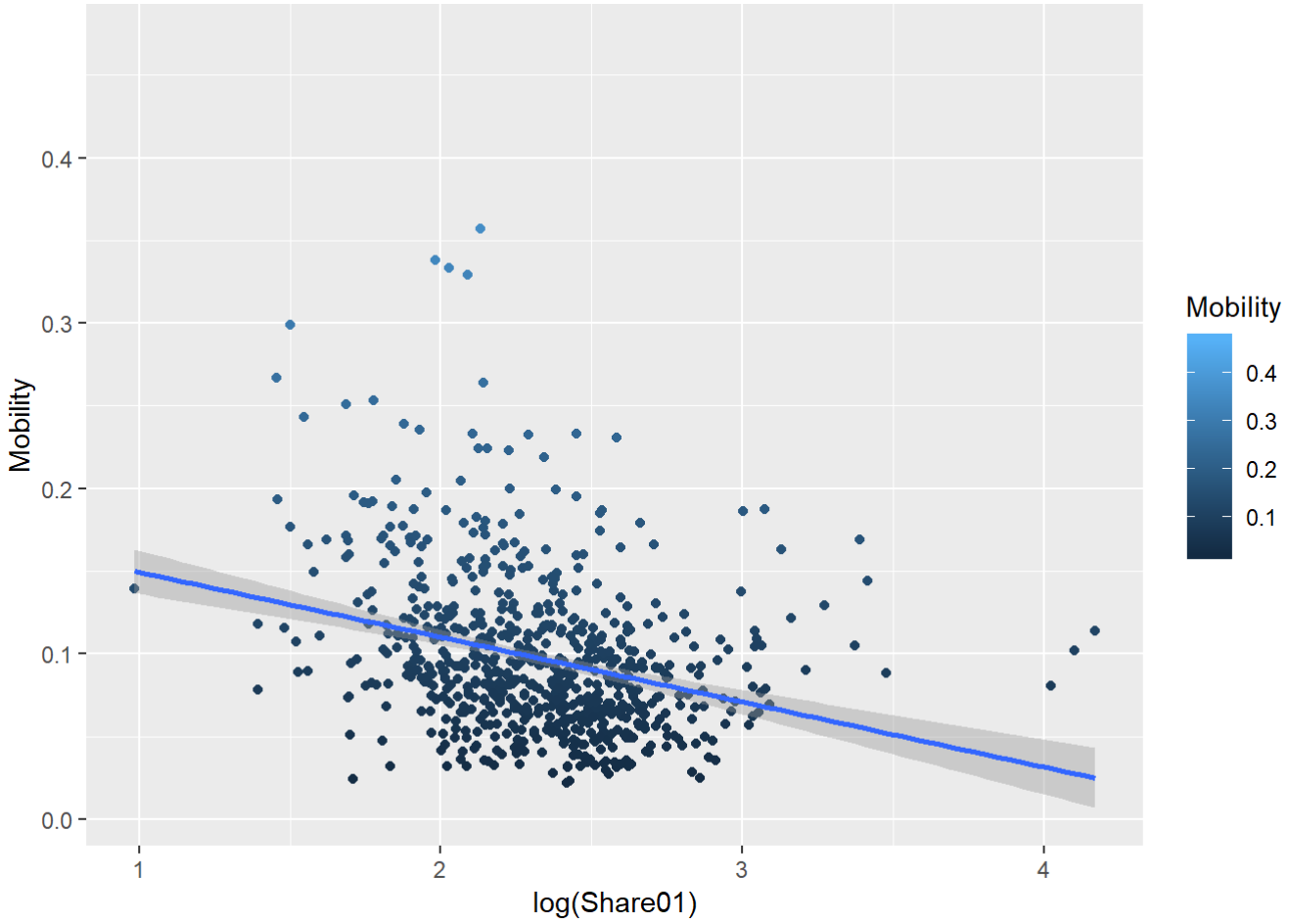
```
c <- lm(Mobility ~ Seg_racial, data = dat_map)
summary(c)
```

```
##
## Call:
## lm(formula = Mobility ~ Seg_racial, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09433 -0.03188 -0.00850  0.01969  0.34762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.124843   0.003098  40.300   <2e-16 ***
## Seg_racial  -0.184333   0.018603  -9.909   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04971 on 710 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.1215, Adjusted R-squared:  0.1203
## F-statistic: 98.19 on 1 and 710 DF,  p-value: < 2.2e-16
```

d. Income share of the top 1%

With this plot, it seems as if share income of 1% rich gets higher, the mobility rate decreases. This plot was suprising for me since with question number 2-b, i have concluded that income has no relationship with Mobility. However, this plot is saying that somehow income has a relationship with mobility.

```
ggplot(dat_map, aes(log(Share01),Mobility, color = Mobility)) + geom_point() + geom_smooth(method = l
m)
```



Summary of lm model tells me that share01 has small p-value, thus it can be used in explaining mobility. Again, it raises question why only share01 has relationship but not income variable?

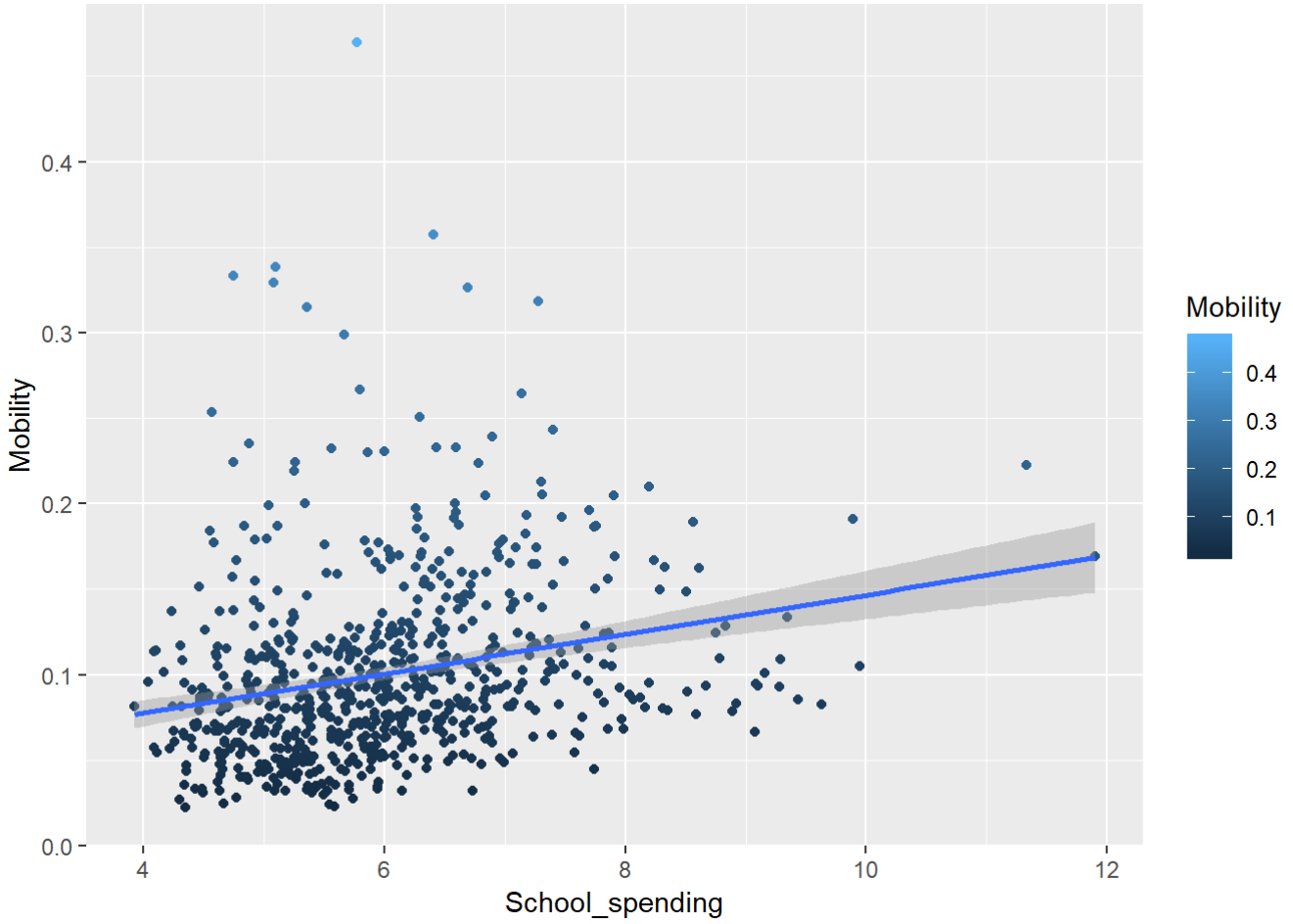
```
d <- lm(Mobility ~ Share01, data = dat_map)
summary(d)
```

```
##
## Call:
## lm(formula = Mobility ~ Share01, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08271 -0.03130 -0.01041  0.01944  0.25528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1166345  0.0043007  27.120  < 2e-16 ***
## Share01     -0.0017511  0.0003572  -4.903 1.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04748 on 691 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.03362,    Adjusted R-squared:  0.03222
## F-statistic: 24.04 on 1 and 691 DF,  p-value: 1.178e-06
```

e. Mean school expenditures per pupil

If the average spending per people of public school gets higher, mobilty rate also increases. Easy to comprehend since it means students from low income household can also be supported with high public education quality , raising the possbility of social level change.

```
ggplot(dat_map, aes(School_spending,Mobility, color = Mobility)) + geom_point() + geom_smooth(method = lm)
```



P-value of school spending variable is very small, implying that this is meaningful varaible.

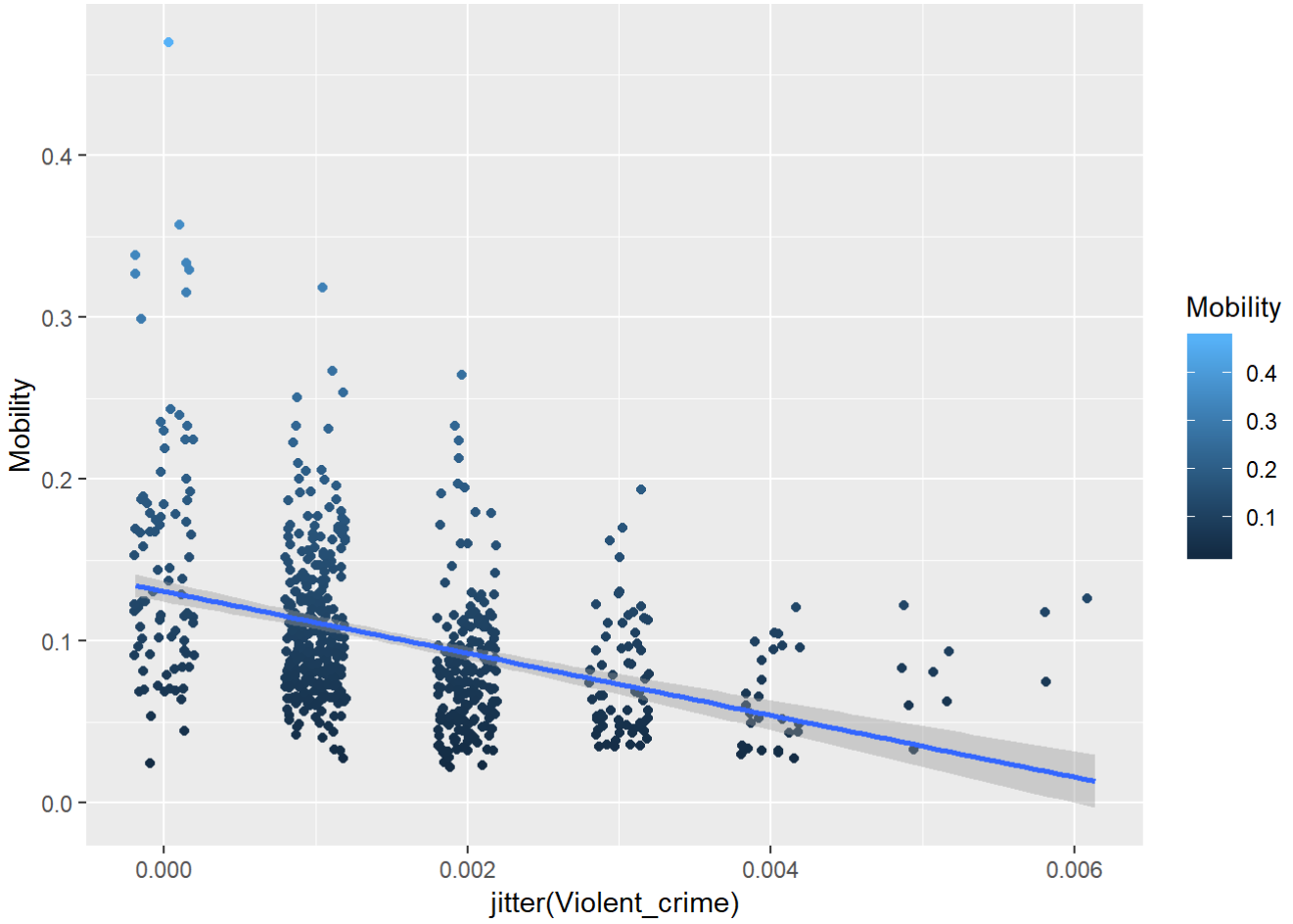
```
e <- lm(Mobility ~ School_spending, data = dat_map)
summary(e)
```

```
##
## Call:
## lm(formula = Mobility ~ School_spending, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07706 -0.03412 -0.01029  0.02101  0.37174
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.031470   0.010515   2.993  0.00286 **
## School_spending 0.011519   0.001729   6.663 5.41e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05154 on 707 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.05908,    Adjusted R-squared:  0.05775
## F-statistic: 44.39 on 1 and 707 DF,  p-value: 5.413e-11
```

f. Violent crime rate

Since violent crime had discrete value around 0.000 ~ 0.006, it was difficult to see how data was spread out. Thus i have used jitter() option to check how much data points are located in what violent crime rate. By doing so, i was able to see that most of the cities have violent rate around 0.001 ~ 0.002.

```
ggplot(dat_map, aes(jitter(Violent_crime),Mobility, color = Mobility)) + geom_point() + geom_smooth(me
thod = lm)
```



By using summary function, i was suprised that estimate value of violent_crime is very high compared to other variables. However, if we consider that one point increase in violent crime is 0.001, we cannot conlcude that change in violent crime rate will have huge impact on Mobility.

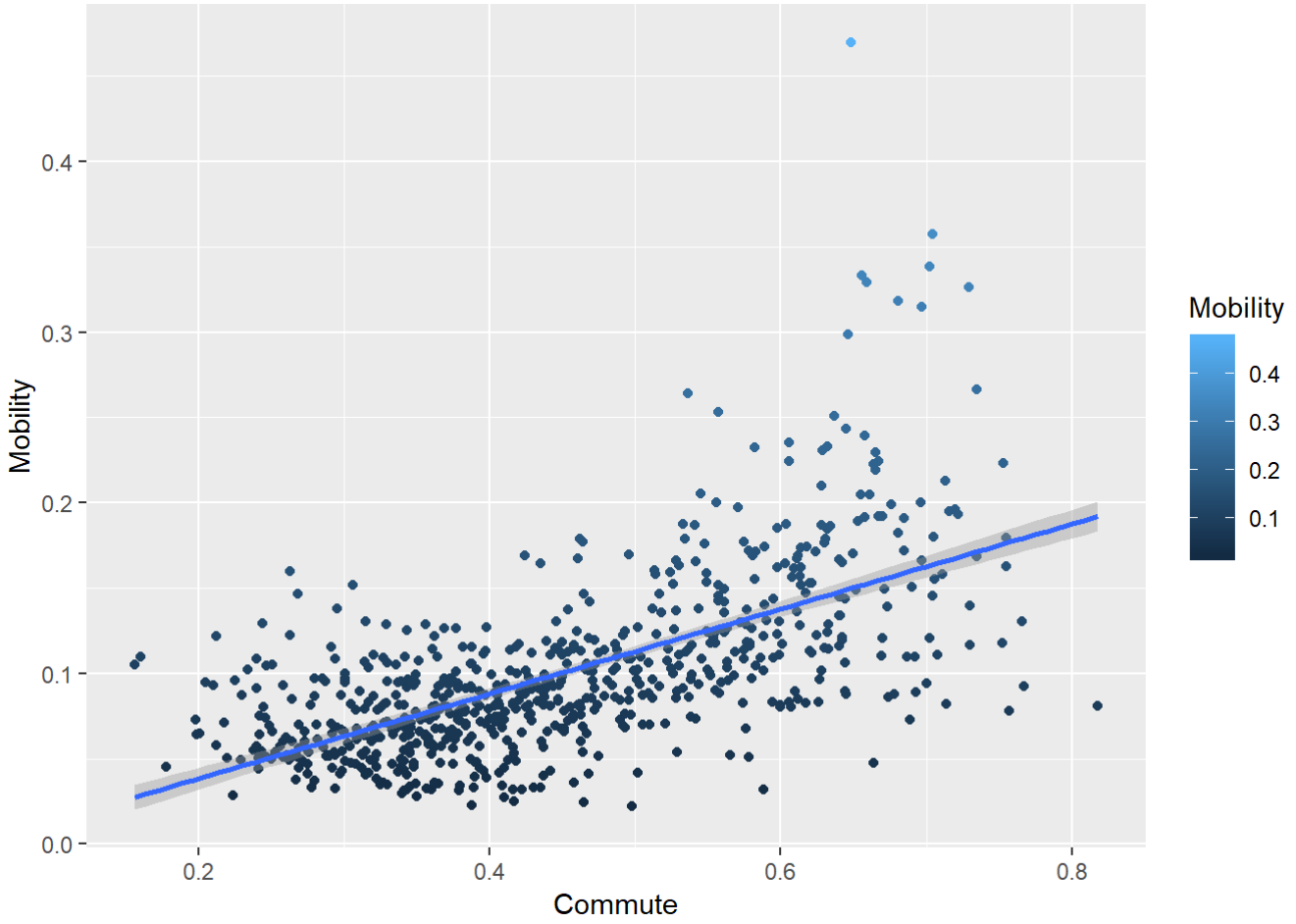
```
f <- lm(Mobility ~ Violent_crime, data = dat_map)
summary(f)
```

```
##
## Call:
## lm(formula = Mobility ~ Violent_crime, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10679 -0.03345 -0.01056  0.02208  0.33865
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.131046   0.003317   39.50   <2e-16 ***
## Violent_crime -19.333361   1.777514  -10.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04953 on 683 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.1476, Adjusted R-squared:  0.1464
## F-statistic: 118.3 on 1 and 683 DF,  p-value: < 2.2e-16
```

g. Fraction of workers with short commutes.

It seems commute has strong linear relationship with Mobility. However, it is difficult to understand the relationship between commute and mobiliy with my common sense.

```
ggplot(dat_map, aes(Commute,Mobility, color = Mobility)) + geom_point() + geom_smooth(method = lm)
```



p-value of commute is small enough to conlcude that it is a meaningful variable. Again, i will have to look on other variables as well since i cannot get the idea how commute can have positive relationship with mobility.

```
f <- lm(Mobility ~ Commute, data = dat_map)
summary(f)
```

```
##
## Call:
## lm(formula = Mobility ~ Commute, data = dat_map)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11092 -0.02301 -0.00470  0.01718  0.31996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.011365   0.005382  -2.112   0.0351 *
## Commute      0.248615   0.011479  21.659   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04116 on 710 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.397
## F-statistic: 469.1 on 1 and 710 DF,  p-value: < 2.2e-16
```

3 All things considered

Run a linear regression of mobility against all appropriate covariates.

I have decided to exclude factor variables. They are ID, Name and State.

```
fit.all3 <- lm(Mobility ~ . -ID -Name -State, data = dat)
```

a. Report all regression coefficients and their standard errors; you may use either a table or a figure as you prefer.

Below, you can see all regression coefficients and their standard errors. I can see bunch of variable which have high p-value.

```
summary(fit.all3)
```

```
##
## Call:
## lm(formula = Mobility ~ . - ID - Name - State, data = dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079757 -0.010815 -0.001319  0.009626  0.130681
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.766e-01  7.221e-02   2.446 0.014902 *
## Population                  1.561e-09  2.172e-09   0.719 0.472790
## Urban                       1.568e-03  3.515e-03   0.446 0.655716
## Black                       8.856e-02  2.540e-02   3.487 0.000546 ***
## Seg_racial                 -4.837e-02  1.654e-02  -2.924 0.003664 **
## Seg_income                  1.064e+00  8.313e-01   1.280 0.201160
## Seg_poverty                -8.582e-01  4.471e-01  -1.920 0.055648 .
## Seg_affluence              -3.178e-01  4.163e-01  -0.763 0.445690
## Commute                     7.548e-02  2.551e-02   2.959 0.003285 **
## Income                      3.059e-07  5.984e-07   0.511 0.609546
## Gini                        2.929e+00  2.888e+00   1.014 0.311106
## Share01                    -2.937e-02  2.889e-02  -1.017 0.309980
## Gini_99                    -3.033e+00  2.888e+00  -1.050 0.294341
## Middle_class                8.649e-02  4.265e-02   2.028 0.043240 *
## Local_tax_rate              1.329e-01  2.379e-01   0.559 0.576667
## Local_gov_spending          9.929e-07  2.761e-06   0.360 0.719344
## Progressivity               5.602e-03  1.119e-03   5.005 8.58e-07 ***
## EITC                       -5.897e-04  4.092e-04  -1.441 0.150404
## School_spending            -1.286e-03  2.066e-03  -0.622 0.533997
## Student_teacher_ratio      -5.020e-04  1.021e-03  -0.492 0.623294
## Test_scores                 4.603e-04  2.758e-04   1.669 0.095920 .
## HS_dropout                 -1.918e-01  7.679e-02  -2.498 0.012926 *
## Colleges                   -1.053e-01  7.219e-02  -1.458 0.145586
## Tuition                    -3.329e-08  4.002e-07  -0.083 0.933746
## Graduation                 -1.386e-02  1.264e-02  -1.096 0.273738
## Labor_force_participation -6.895e-02  4.756e-02  -1.450 0.148010
## Manufacturing              -1.727e-01  2.528e-02  -6.831 3.39e-11 ***
## Chinese_imports            -8.122e-04  6.989e-04  -1.162 0.245929
## Teenage_labor              -2.125e+00  1.928e+00  -1.103 0.270884
## Migration_in               -8.819e-02  2.763e-01  -0.319 0.749778
## Migration_out              -5.249e-01  3.380e-01  -1.553 0.121224
## Foreign_born                1.071e-01  4.983e-02   2.148 0.032313 *
## Social_capital             -2.021e-03  2.430e-03  -0.832 0.406137
## Religious                   6.082e-02  1.157e-02   5.256 2.46e-07 ***
## Violent_crime              -3.194e+00  1.481e+00  -2.156 0.031700 *
## Single_mothers             -3.469e-01  8.331e-02  -4.164 3.87e-05 ***
## Divorced                    7.964e-02  1.417e-01   0.562 0.574460
## Married                    -8.914e-02  6.706e-02  -1.329 0.184548
## Longitude                   1.129e-04  2.049e-04   0.551 0.581773
## Latitude                    1.424e-03  5.312e-04   2.680 0.007682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0224 on 378 degrees of freedom
##   (323 observations deleted due to missingness)
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.7418
## F-statistic: 31.72 on 39 and 378 DF,  p-value: < 2.2e-16
```

b. Explain why the ID variable must be excluded

Since Id has no meaning at all and is a factor variable, it can be excluded.

c. Explain which other variables, if any, you excluded from the regression, and why. (If you think they can all be used, explain why.) [For this question, do not use any automated variable selection, and try to keep as many variables as possible.]

If i only consider p-value, i can exclude population, urban, black and so on. There are so many variables that can be excluded if i only consider p-value. However, i would like to reject this kind of approach since i believe some of the variable included in excluding list has some meaning, as we have discussed in question number 2. Rather than that, i have decided to exclude data by VIF values and single linear model p-value of each variable. Thus, i would love to answer this qestion together with 3-e.

d. Compare the coefficients you found in problem 2 to the coefficients for the same variables in this regression. Are they significantly different? Have any changed sign?

Below, you can see a table with order of population, income, seg_racial, share01, school_spending, violent_crime and commute with question number 2 and 3. Coefficient values should be different since we are using different linear model in Q2 and Q3. For each of them,

Population : changed sign and changed significantly.

Income : no big change

Seg_racial : decreased relationship, however hard to say significant

share01 : increased relationship

school_spending : changed sign

violent_crime : did not change sign, but changed significantly

commute : decreased significantly

```
question2 <- c(-6.733e-09, 2.477e-07, -0.184333, -0.0017511, 0.011519, -19.333361, 0.248615)
question3 <- c(5.722e-10, 3.362e-07, -6.261e-02, -1.643e-02, -1.557e-03, -3.350e+00, 4.415e-02)
question4 <- rbind(question2,question3)
question4 <- as.data.frame(question4)
colnames(question4) <- c("Population", "Income", "Seg_racial","Share01", "School_spending", "Violent_c
rime", "Commute" )
question4
```

```
##           Population    Income Seg_racial    Share01 School_spending
## question2 -6.733e-09 2.477e-07  -0.184333 -0.0017511        0.011519
## question3  5.722e-10 3.362e-07  -0.062610 -0.0164300       -0.001557
##           Violent_crime  Commute
## question2     -19.33336 0.248615
## question3      -3.35000 0.044150
```

e. Take a look at the variation inflation factor for each variable. Report those variables with VIF greater than 10. Do you suspect a (nearly) multicollinearity? If so, give a reason for, andsuggest a way to avoid it.

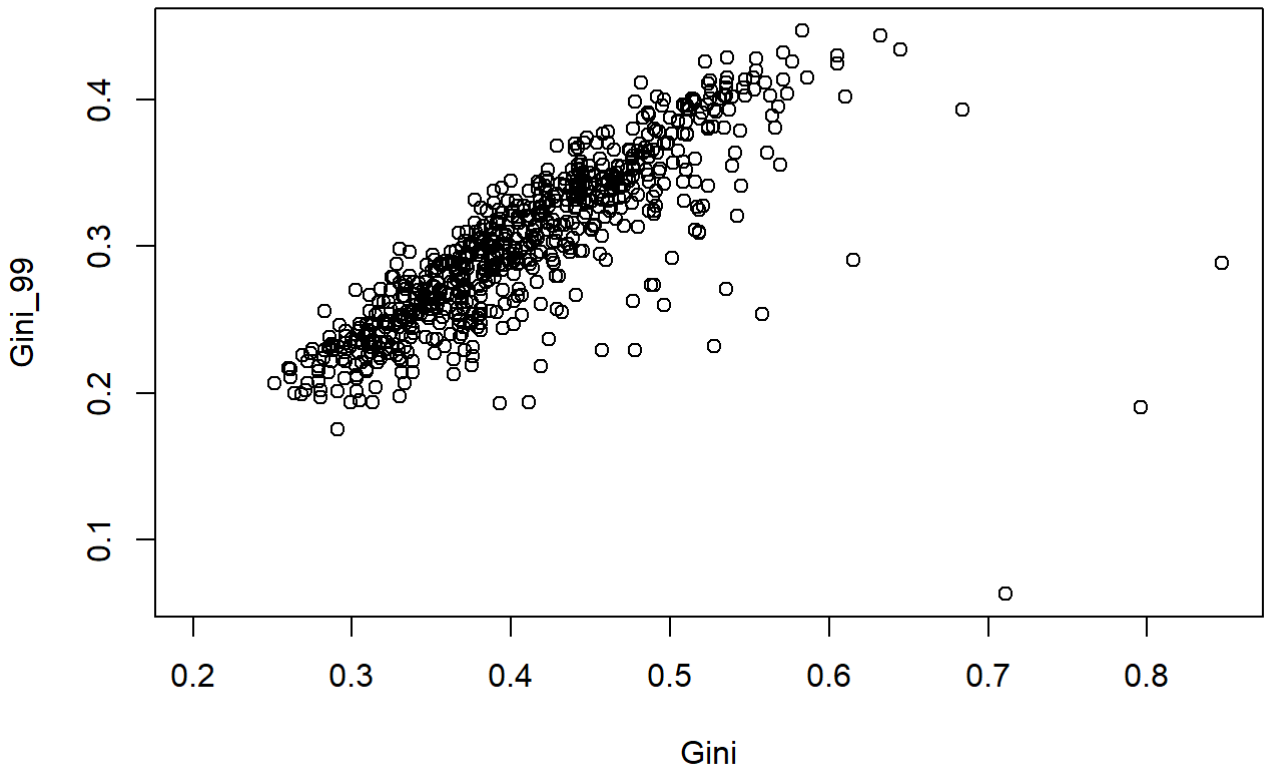
I can see several variables with high VIF. From now on i would like to take a look one by one.

```
vif(fit.all3)
```

```
##                Population                     Urban
##                  3.320686                  2.567701
##                     Black                Seg_racial
##                 10.527764                  2.134496
##                Seg_income               Seg_poverty
##                525.197773                130.147684
##            Seg_affluence                   Commute
##                161.727935                  8.003940
##                    Income                      Gini
##                  7.376884              41784.423509
##                   Share01                   Gini_99
##              11726.020668              22476.048718
##              Middle_class            Local_tax_rate
##                  9.164688                  2.569307
##       Local_gov_spending              Progressivity
##                  2.245337                  1.481558
##                      EITC            School_spending
##                  2.234908                  3.946428
##    Student_teacher_ratio                Test_scores
##                  2.691779                  3.591586
##                HS_dropout                   Colleges
##                  1.857290                  1.865790
##                   Tuition                Graduation
##                  1.835598                  2.282166
## Labor_force_participation              Manufacturing
##                  5.947274                  2.745890
##           Chinese_imports               Teenage_labor
##                  1.365216                  6.087540
##              Migration_in              Migration_out
##                  5.423010                  5.010401
##              Foreign_born             Social_capital
##                  3.417968                  6.739760
##                 Religious              Violent_crime
##                  2.294771                  1.806381
##            Single_mothers                  Divorced
##                16.377285                  4.012991
##                   Married                 Longitude
##                  6.407447                  4.274026
##                  Latitude
##                  5.946518
```

first of all, I thought that Gini and Gini99 will have high relationship among each other since the concept of the variables are almost similar. Moreover, by using lm() and summary() function, i was able to see linear relationship. Thus i have decided to exclude Gini_99. The reason why i have excluded Gini_99 rather than Gini is because i thought saving Gini_99 and share01 will somehow give me some inspiration in later analysis.

```
plot(Gini, Gini_99)
```



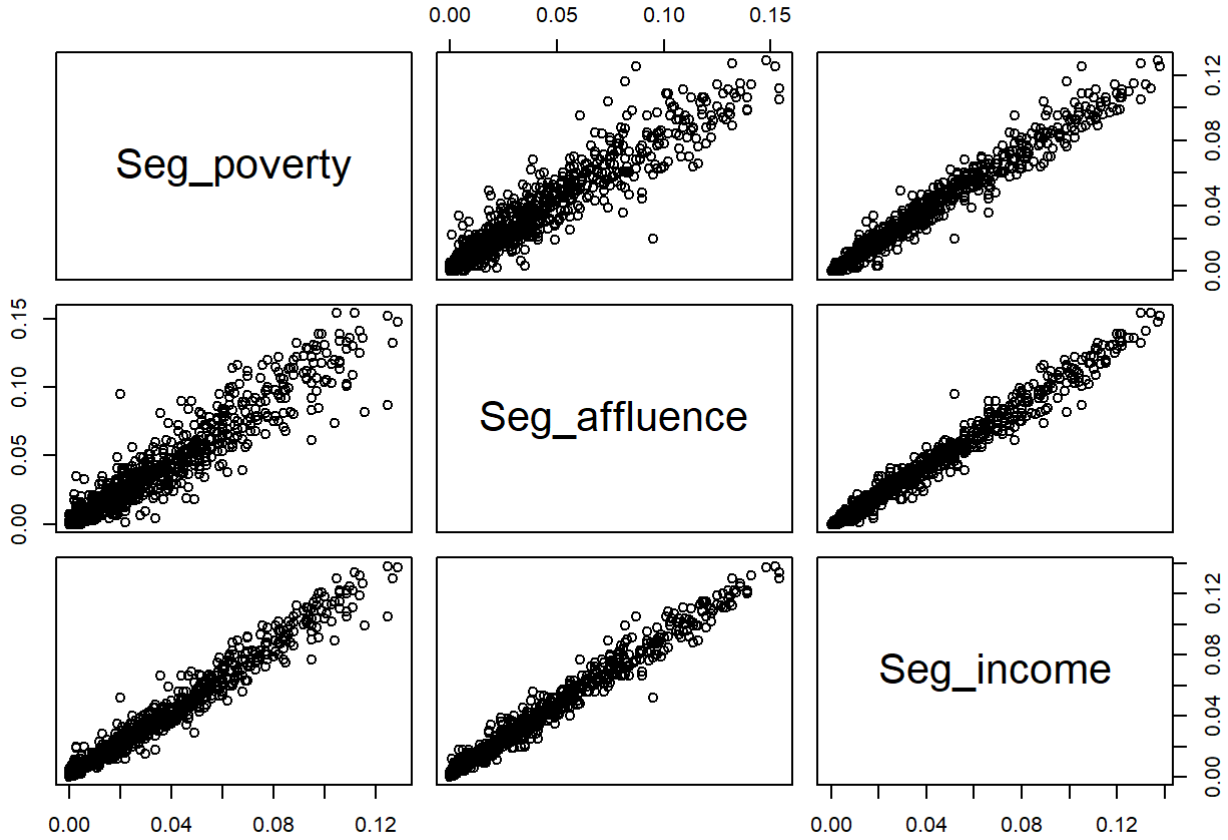

```
summary(lm(Gini ~ Gini_99, data = dat))
```

```
##
## Call:
## lm(formula = Gini ~ Gini_99, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07710 -0.02662 -0.00661  0.01451  0.56240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07956    0.01036   7.682 5.23e-14 ***
## Gini_99      1.09584    0.03381  32.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05025 on 707 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.5978, Adjusted R-squared:  0.5972
## F-statistic:  1051 on 1 and 707 DF,  p-value: < 2.2e-16
```

Secondly, i have decided to take a look on seg_poverty, seg_affluence and seg_income. According to variable explanation, seg_income is a concpet including seg_poverty and seg_affluence. This is the reason why seg_income and seg_poverty has linear relationship and seg_affluence as well. Thus, i decided to exclude seg_poverty and seg_affluence since it can be explained by seg_income. Nonetheless, i though excluding seg_poverty and seg_affluence would be better than excluding seg_income since those two varialbe have

common social level in their explanation. Seg_povery is ranged from low to middlehigh while seg_affluence is ranged from middlelow to high. Thus range between middelow to middlehigh is counted twice if i decide to use these two variables.

```
pairs(~ Seg_poverty + Seg_affluence + Seg_income)
```



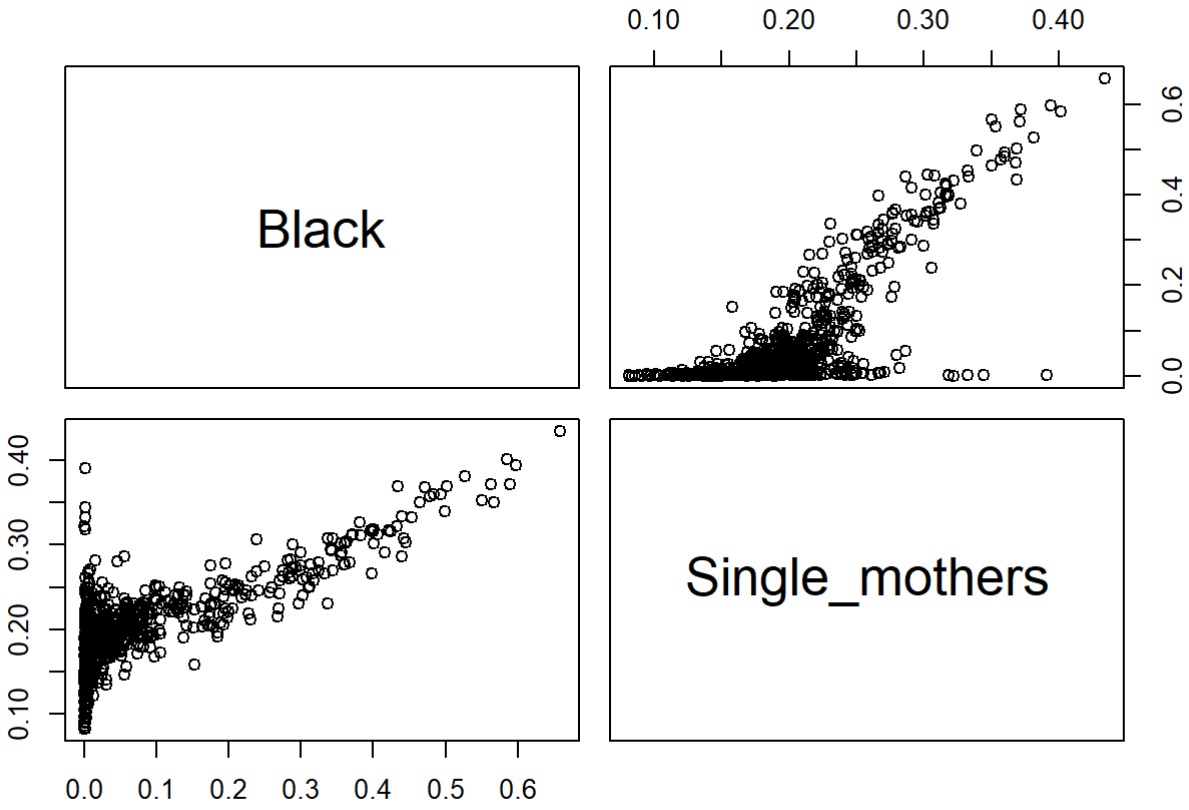
After excluding Gini, seg_affluence and sef_poverty, i was able to see that extremely high VIF values have disappeared. However, i was able to see some variables which still have VIF higher than 10. Those are Black and single_mothers.

```
fit.all32 <- lm(Mobility ~ . -ID -Name -State - Gini - Seg_poverty -Seg_affluence, data = dat)
vif(fit.all32)
```

```
##              Population                   Urban
##                3.282520                2.560158
##                   Black              Seg_racial
##               10.519553                2.062987
##              Seg_income                 Commute
##                5.855364                7.966747
##                  Income                 Share01
##                7.363764                1.529608
##                 Gini_99            Middle_class
##                7.522621                8.968121
##          Local_tax_rate       Local_gov_spending
##                2.536954                2.231926
##           Progressivity                    EITC
##                1.478441                2.201991
##          School_spending    Student_teacher_ratio
##                3.912414                2.686520
##             Test_scores               HS_dropout
##                3.495506                1.847984
##                Colleges                  Tuition
##                1.860555                1.824632
##              Graduation Labor_force_participation
##                2.189736                5.827826
##           Manufacturing          Chinese_imports
##                2.723406                1.325295
##           Teenage_labor             Migration_in
##                5.977678                5.375782
##           Migration_out             Foreign_born
##                4.939827                3.305565
##          Social_capital                Religious
##                6.593842                2.279542
##           Violent_crime           Single_mothers
##                1.792154               16.051577
##                Divorced                 Married
##                3.940653                6.343653
##               Longitude                Latitude
##                4.232717                5.893256
```

With the plots and summary below, I can see there is some relationship between Black and single_mothers. However, i decided not to exclude any of these variables since i believe i will need them in predicting mobility. (In fact, these two variables were shown as meanigful variables in later analysis with low p-value. Please note that there might be some collinearity.)

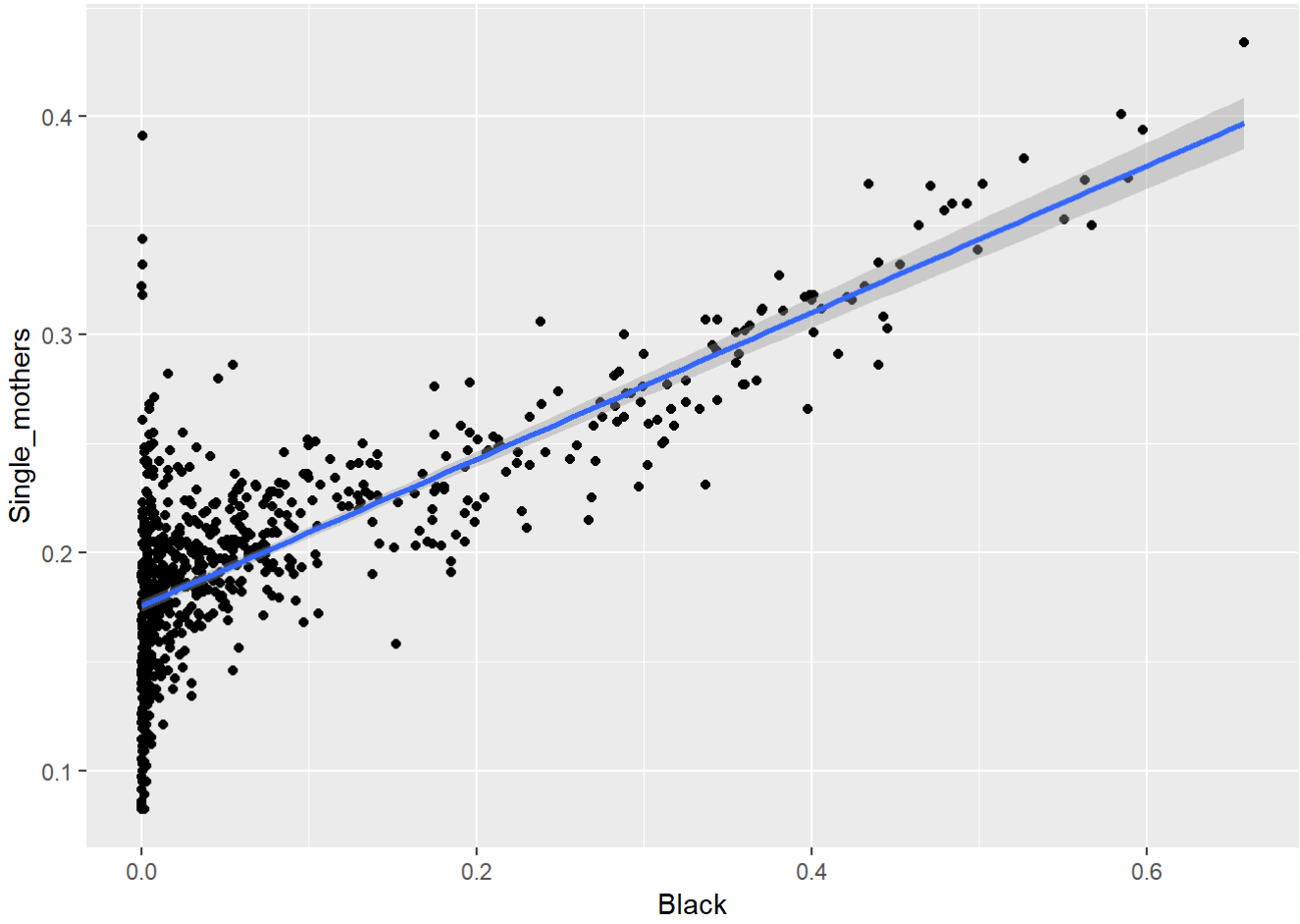
```
index <- vif(fit.all32)[vif(fit.all32)>10]
pairs(dat[,names(index)])
```




```
summary(lm(Black ~ Single_mothers, data = dat))
```

```
##
## Call:
## lm(formula = Black ~ Single_mothers, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41542 -0.04435 -0.00831  0.04485  0.22393
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.28309    0.01116  -25.38   <2e-16 ***
## Single_mothers  1.78904    0.05348   33.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07714 on 739 degrees of freedom
## Multiple R-squared:  0.6022, Adjusted R-squared:  0.6017
## F-statistic:  1119 on 1 and 739 DF,  p-value: < 2.2e-16
```

```
ggplot(dat, aes(Black,Single_mothers)) + geom_point() + geom_smooth(method = lm)
```



4. Please in my front yard

a. Inspect the missingness pattern in variables Colleges, Tuition and Graduation. [Note: NA is a missing value.] How many observations have no measurements for these variables?

Colleges have 157, Tuition have 161 and Graduation have 160 NA missing values.

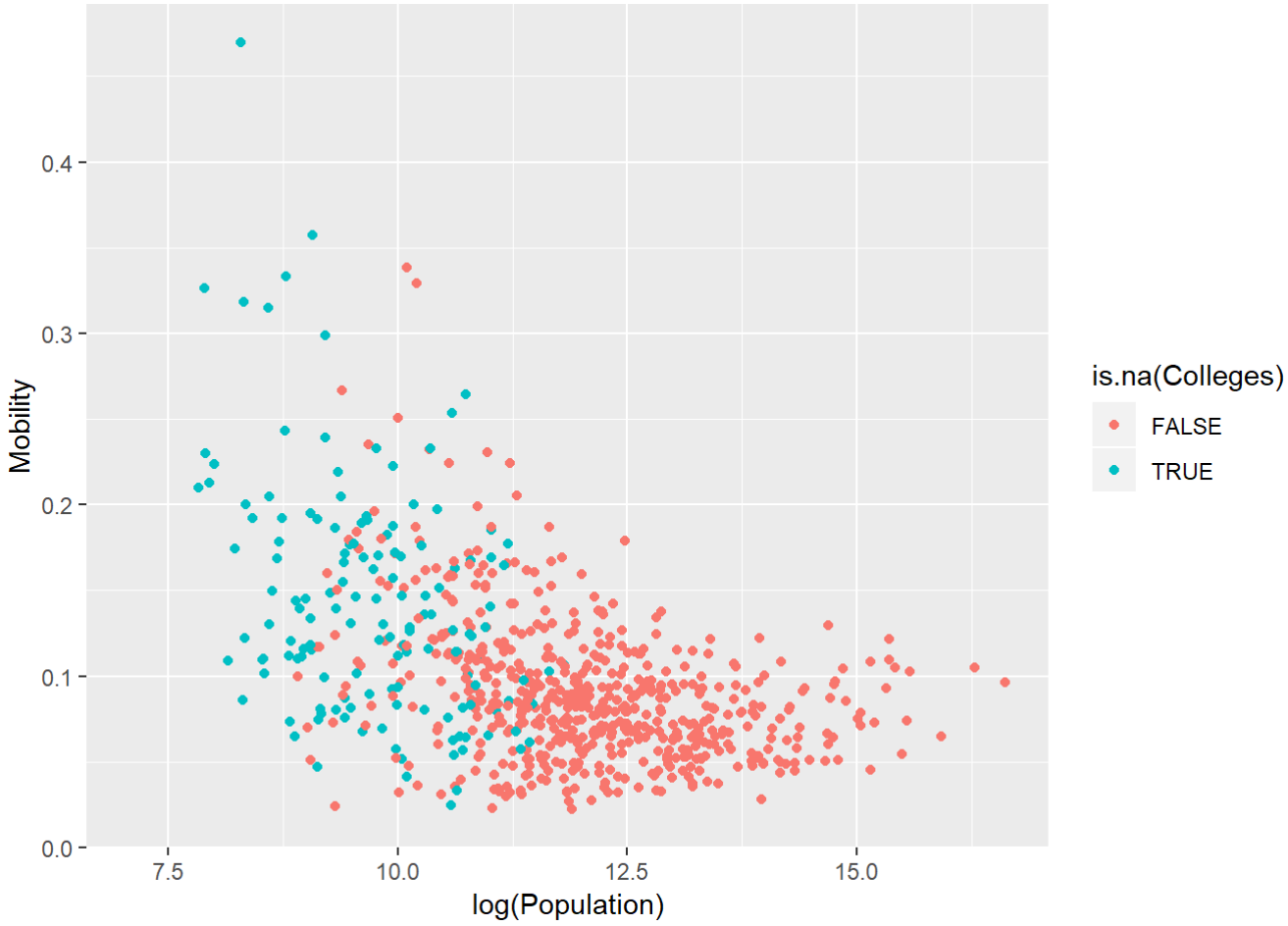
```
Howmuchna <- c(sum(is.na(dat$Colleges)),sum(is.na(dat$Tuition)),sum(is.na(dat$Graduation)))
Howmuchna
```

```
## [1] 157 161 160
```

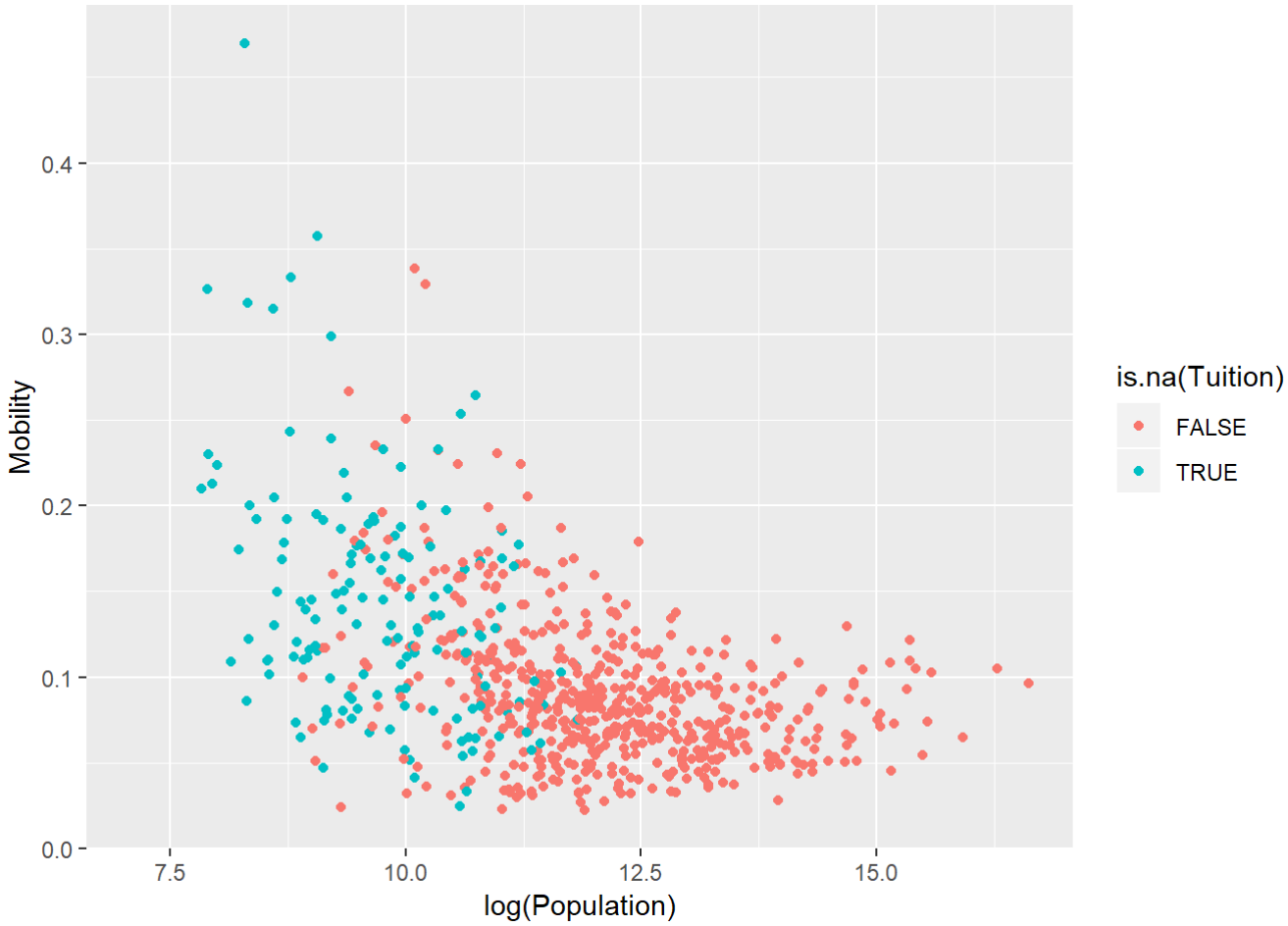
b. Did the missing values happen at random? To answer this, plot a scatter of Mobility and Population (choose a suitable scale for Population), and inspect which data points have missing values in (all of, or some of) variables Colleges, Tuition and Graduation.

According to three plots i have drawn below, i can conclude that missing values did not happen at random. (I have differentiated each of the plots by using color option by colleges, tuituon and graduation.) Rather than that, it seems like cities with low population tend to have missing values. This could be because of lack of population, there were some difficulties in getting statistic results. For other pluasible explanation, NA could be deemed as 0. I would like to discuss how i am going to deal with NA values in later question.

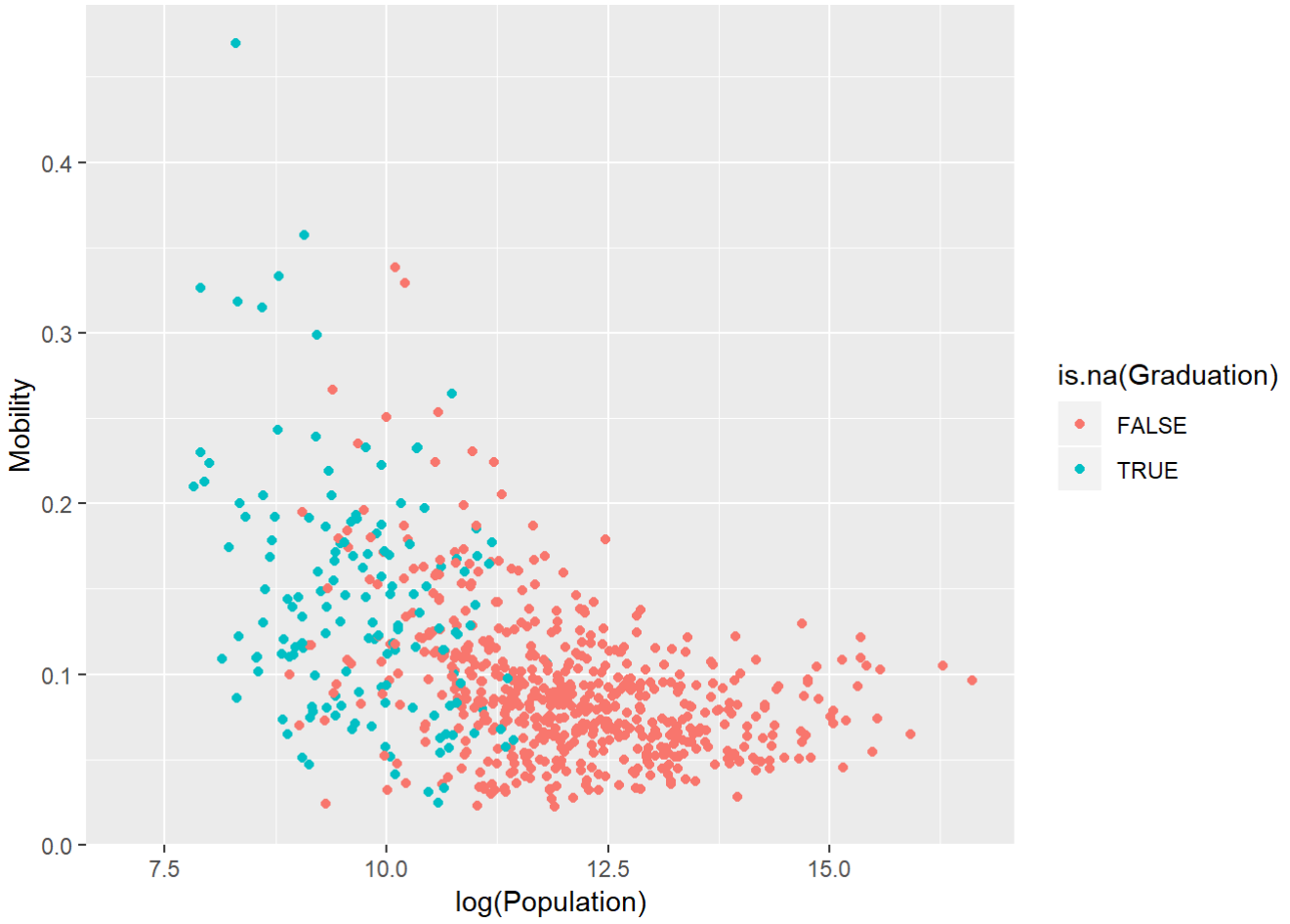
```
ggplot(mapping = aes(log(Population), Mobility, color = is.na(Colleges))) + geom_point()
```




```
ggplot(mapping = aes(log(Population), Mobility, color = is.na(Tuition))) + geom_point()
```



```
ggplot(mapping = aes(log(Population), Mobility, color = is.na(Graduation))) + geom_point()
```



[This is what i have tried in order to replace NA, However i am not going to use this idea afterall since Q4-C requires me to replace NA to 0]

As i have concluded that NA has some pattern, meaning that most of the NA values come frome places where population is smaller then median, i thought replacing all NA values into 0 is not reasonable. Thus, i wanted to replace NA values into mean values of each variables with values uwith population smaller than median. Therefore, first i have made a new data with population under median and calculated mean of each variable, college, tuition and graduation, After that, i have replaced NA with mean values of those variables.

```
dat4 <- dat %>% filter(dat$Population < median(dat$Population))
dat4c <- na.omit(dat4$Colleges)
dat4t <- na.omit(dat4$Tuition)
dat4g <- na.omit(dat4$Graduation)
dat44 <- dat %>% mutate(Colleges = ifelse(is.na(Colleges),mean(dat4c),Colleges)) %>% mutate(Tuition =
ifelse(is.na(Tuition),mean(dat4t),Tuition)) %>% mutate(Graduation = ifelse(is.na(Graduation),mean(dat4
g),Graduation))
```

with the dataframe below, you can see what diffrence happens if i change NA into mean values but not 0. However, since question below asks me to change NA into 0, i will follow the instruction first.

```
q4ba <- c(mean(na.omit(dat$Colleges)), mean(dat4c), mean(dat44$Colleges))
q4bb <- c(mean(na.omit(dat$Tuition)), mean(dat4t), mean(dat44$Tuition))
q4bc <- c(mean(na.omit(dat$Graduation)),mean(dat4g), mean(dat44$Graduation))
q4bd <- rbind(q4ba,q4bb,q4bc)
q4bd <- as.data.frame(q4bd)
colnames(q4bd) <- c("Pop all", "Pop under median", "Pop all after NA transefer")
q4bd
```

```
##            Pop all Pop under median Pop all after NA transefer
## q4ba  2.311473e-02     3.600465e-02               2.584579e-02
## q4bb  4.354729e+03     3.563901e+03               4.182903e+03
## q4bc -1.032702e-05    -7.056872e-03              -1.531848e-03
```

c. Create a new variable, called HE, whose value is TRUE if there is a higher education institution in the community, is FALSE if not. Replace all NA values in variables Colleges, Tuition and Graduation with 0

I have replaced all of the NA values in colleges, tuition and graduation into 0.

```
dat44c <- dat %>% mutate(Colleges = ifelse(is.na(Colleges),0,Colleges)) %>% mutate(Tuition =  ifelse(i
s.na(Tuition),0,Tuition)) %>% mutate(Graduation = ifelse(is.na(Graduation),0,Graduation))
Howmuchna2 <- c(sum(is.na(dat44c$Colleges)),sum(is.na(dat44c$Tuition)), sum(is.na(dat44c$Graduation)))
Howmuchna2
```

```
## [1] 0 0 0
```

I have created new variable called HE, whose value is True if there is a higher education and False if not. You can see in below table that places with no college have been assigned to False. (Insdie the code, i have considered not only colleges but also tuition and graduation.)

```
dat44ca <- dat44c %>% mutate(HE = ifelse(dat44c$Colleges == 0 | dat44c$Graduation == 0 | dat44c$Tuitio
n == 0, 'FALSE', 'TRUE'))
head(dat44ca[dat44ca$Colleges == 0,c(26,44)])
```

```
##    Colleges    HE
## 9         0 FALSE
## 16        0 FALSE
## 40        0 FALSE
## 43        0 FALSE
## 44        0 FALSE
## 54        0 FALSE
```

5. All things considered, again.

Fit a linear regression model, incorporating your findings in problems 3 and 4. If you have removed, created, or modified variables, explain. Report all regression coefficients and their standard errors. Use this model for all problems below.

Before excluding some variables in question number 5 to make a final lm model, i would like to show what i have done in question number 3 and 4. In question number 3, i have decided to exclude Gini, Seg_poverty and seg_affluence in order to control VIF rather than using PCA. In question number 4, i haved replaced NA in colleges, tuition and graduation into 0. Below is the baseline before i decide to exclude some other variables.

```
fit.all5a <- lm(Mobility ~ . -ID -Name -State - Gini_99 - Seg_poverty -Seg_affluence, data = dat44ca)
summary(fit.all5a)
```

```
##
## Call:
## lm(formula = Mobility ~ . - ID - Name - State - Gini_99 - Seg_poverty -
##     Seg_affluence, data = dat44ca)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.076451 -0.013637 -0.002004  0.010444  0.134967
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 7.134e-02  6.755e-02   1.056 0.291462
## Population                  5.827e-10  2.350e-09   0.248 0.804273
## Urban                      -5.147e-04  3.856e-03  -0.133 0.893858
## Black                       6.739e-02  2.628e-02   2.564 0.010653 *
## Seg_racial                 -6.790e-02  1.677e-02  -4.049 5.98e-05 ***
## Seg_income                 -1.284e-02  9.299e-02  -0.138 0.890215
## Commute                     4.436e-02  2.556e-02   1.736 0.083210 .
## Income                      3.271e-07  5.615e-07   0.583 0.560455
## Gini                       -5.095e-02  5.125e-02  -0.994 0.320682
## Share01                     6.439e-04  6.167e-04   1.044 0.296991
## Middle_class                1.214e-01  3.991e-02   3.042 0.002478 **
## Local_tax_rate              6.926e-01  2.095e-01   3.306 0.001015 **
## Local_gov_spending          7.722e-07  1.855e-06   0.416 0.677375
## Progressivity               7.786e-03  1.157e-03   6.730 4.76e-11 ***
## EITC                       -5.747e-04  4.140e-04  -1.388 0.165672
## School_spending            -1.464e-03  1.886e-03  -0.776 0.438072
## Student_teacher_ratio      -6.217e-05  1.006e-03  -0.062 0.950744
## Test_scores                -5.120e-05  2.607e-04  -0.196 0.844376
## HS_dropout                 -2.062e-01  7.856e-02  -2.624 0.008958 **
## Colleges                   -3.283e-02  7.199e-02  -0.456 0.648594
## Tuition                     1.321e-07  4.323e-07   0.306 0.760017
## Graduation                 -1.534e-02  1.307e-02  -1.173 0.241259
## Labor_force_participation -4.744e-02  4.549e-02  -1.043 0.297541
## Manufacturing              -1.500e-01  2.511e-02  -5.974 4.47e-09 ***
## Chinese_imports            -7.875e-04  7.581e-04  -1.039 0.299423
## Teenage_labor              -2.488e+00  1.999e+00  -1.245 0.213709
## Migration_in               -4.783e-01  2.910e-01  -1.643 0.100948
## Migration_out              -5.134e-02  3.471e-01  -0.148 0.882471
## Foreign_born                5.487e-02  4.805e-02   1.142 0.254016
## Social_capital             -5.404e-03  2.199e-03  -2.457 0.014340 *
## Religious                   5.617e-02  1.159e-02   4.847 1.69e-06 ***
## Violent_crime              -3.436e+00  1.527e+00  -2.250 0.024898 *
## Single_mothers             -3.493e-01  8.016e-02  -4.358 1.60e-05 ***
## Divorced                   -7.352e-02  1.384e-01  -0.531 0.595518
## Married                    -2.958e-02  6.242e-02  -0.474 0.635755
## Longitude                  -5.559e-05  2.074e-04  -0.268 0.788794
## Latitude                    1.844e-03  5.256e-04   3.509 0.000492 ***
## HETRUE                      8.111e-04  4.209e-03   0.193 0.847270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02682 on 490 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.7466, Adjusted R-squared:  0.7275
## F-statistic: 39.03 on 37 and 490 DF,  p-value: < 2.2e-16
```

[below are the variables i decided to exclude.]

longitude, latitude : i would like to exclude these variables since it contains only location information but does not have relationship with what defines mobility rate.

share01 : i decided to exclude share01 since Gini can be used as a replacement.

chinese imports : i decided to exclude chinese imports since i cannot find relationship between mobility with chinese imports. I doubt there will have any kind of relationship.

income : according to single linear model, p-value was big enough to conlcude that income is not valid variable. If it is not relative variable in single linear model, mostly we can say it is not relative variable in multi regression model as well. Thus decided to exclude.

colleges : as well as income, colleges showed high p-value in single linear model.

graduation : same reason with income and colleges.

Foreign_born : same as graduation.

```
summary(lm(Mobility ~ Income, data = dat44ca ))
```

```
##
## Call:
## lm(formula = Mobility ~ Income, data = dat44ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07531 -0.03433 -0.01083  0.01870  0.36887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.025e-02  1.133e-02   7.962 6.52e-15 ***
## Income      3.094e-07  3.395e-07   0.911    0.362
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05266 on 727 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.001141,   Adjusted R-squared:  -0.0002325
## F-statistic: 0.8308 on 1 and 727 DF,  p-value: 0.3623
```

```
summary(lm(Mobility ~ Colleges, data = dat44ca ))
```

```
##
## Call:
## lm(formula = Mobility ~ Colleges, data = dat44ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07845 -0.03451 -0.01042  0.01930  0.36876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.100940   0.002578  39.160   <2e-16 ***
## Colleges    -0.027899   0.090951  -0.307    0.759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05269 on 727 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.0001294,  Adjusted R-squared:  -0.001246
## F-statistic: 0.0941 on 1 and 727 DF,  p-value: 0.7591
```

```
summary(lm(Mobility ~ Graduation, data = dat44ca ))
```

```
##
## Call:
## lm(formula = Mobility ~ Graduation, data = dat44ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07806 -0.03428 -0.01142  0.01938  0.36927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10042    0.00195  51.509   <2e-16 ***
## Graduation   0.01886    0.01606   1.174    0.241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05264 on 727 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.001894,   Adjusted R-squared:  0.0005206
## F-statistic: 1.379 on 1 and 727 DF,  p-value: 0.2406
```

```
summary(lm(Mobility ~ Foreign_born, data = dat44ca ))
```

```
##
## Call:
## lm(formula = Mobility ~ Foreign_born, data = dat44ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07883 -0.03405 -0.01084  0.01889  0.36872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.101034   0.002525  40.007   <2e-16 ***
## Foreign_born -0.014742   0.038738  -0.381    0.704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05269 on 727 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.0001992,  Adjusted R-squared:  -0.001176
## F-statistic: 0.1448 on 1 and 727 DF,  p-value: 0.7037
```

In conclusion, my final model would be : fit.all5b <- lm(Mobility ~ . -ID -Name -State -Gini -Seg_poverty -Seg_affluence -Longitude -Latitude -Share01 -Chinese_imports -Income -Colleges -Graduation -Foreign_born, data = dat44ca)

```
fit.all5b <- lm(Mobility ~ . -ID -Name -State -Gini -Seg_poverty -Seg_affluence -Longitude -Latitude -Share01 -Chinese_imports -Income -Colleges -Graduation -Foreign_born, data = dat44ca)
summary(fit.all5b)
```

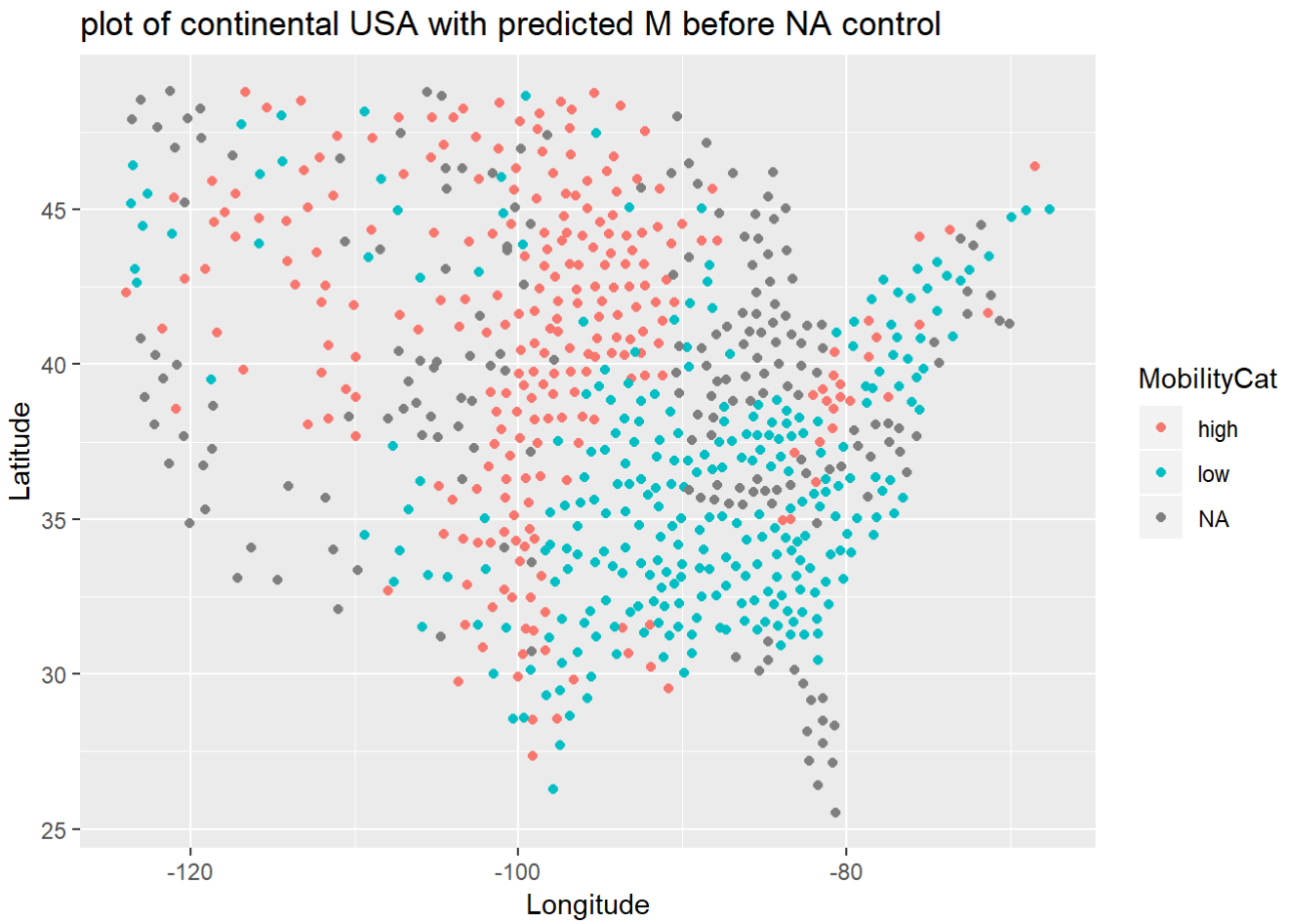
```
##
## Call:
## lm(formula = Mobility ~ . - ID - Name - State - Gini - Seg_poverty -
##     Seg_affluence - Longitude - Latitude - Share01 - Chinese_imports -
##     Income - Colleges - Graduation - Foreign_born, data = dat44ca)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06928 -0.01390 -0.00197  0.01111  0.14330
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.822e-02  6.296e-02   1.560 0.119346
## Population                  3.200e-09  2.043e-09   1.567 0.117833
## Urban                      -2.374e-04  3.866e-03  -0.061 0.951056
## Black                       4.835e-02  2.263e-02   2.137 0.033101 *
## Seg_racial                 -5.990e-02  1.619e-02  -3.700 0.000239 ***
## Seg_income                 -4.448e-02  8.771e-02  -0.507 0.612311
## Commute                     4.925e-02  2.242e-02   2.196 0.028527 *
## Gini_99                    -9.364e-02  5.005e-02  -1.871 0.061946 .
## Middle_class                1.179e-01  3.811e-02   3.094 0.002083 **
## Local_tax_rate              6.826e-01  2.014e-01   3.389 0.000757 ***
## Local_gov_spending          1.187e-06  1.823e-06   0.651 0.515347
## Progressivity               8.342e-03  1.152e-03   7.242 1.69e-12 ***
## EITC                       -4.439e-04  4.145e-04  -1.071 0.284699
## School_spending            -5.108e-04  1.790e-03  -0.285 0.775490
## Student_teacher_ratio       1.585e-03  8.754e-04   1.810 0.070858 .
## Test_scores                -3.412e-05  2.528e-04  -0.135 0.892687
## HS_dropout                 -1.891e-01  7.648e-02  -2.472 0.013764 *
## Tuition                     1.961e-08  3.958e-07   0.050 0.960511
## Labor_force_participation -1.769e-02  4.297e-02  -0.412 0.680755
## Manufacturing              -1.680e-01  2.367e-02  -7.099 4.36e-12 ***
## Teenage_labor              -1.159e+00  1.889e+00  -0.613 0.539853
## Migration_in               -5.386e-01  2.810e-01  -1.917 0.055865 .
## Migration_out              -5.254e-02  3.347e-01  -0.157 0.875329
## Social_capital             -3.861e-03  1.978e-03  -1.952 0.051473 .
## Religious                   4.507e-02  1.066e-02   4.228 2.81e-05 ***
## Violent_crime              -4.129e+00  1.496e+00  -2.759 0.006003 **
## Single_mothers             -3.056e-01  7.881e-02  -3.877 0.000120 ***
## Divorced                   -7.305e-02  1.147e-01  -0.637 0.524461
## Married                     4.131e-03  5.502e-02   0.075 0.940179
## HETRUE                      3.247e-04  3.783e-03   0.086 0.931640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02711 on 498 degrees of freedom
##   (213 observations deleted due to missingness)
## Multiple R-squared:  0.7368, Adjusted R-squared:  0.7215
## F-statistic: 48.08 on 29 and 498 DF,  p-value: < 2.2e-16
```

6. Make a map of predicted mobility.

Make a map of the model's predicted mobility. How does it compare, qualitatively, to the map of actual mobility?

Below is the plot of predicted Mobility with bunch of NA Values on it. Qualitatively, i cannot say this is a good plot since there are too many NA values which prohibits one to have discrete interpretation. Thus, i would like to start taking care of those NA values.

```
dat6 <- dat44ca[,-c(1,2,4,10,11,13,14,15,26,28,31,35)]
fit.all6 <- lm(Mobility ~ . -Longitude -Latitude, data = dat6)
test_fit = predict(fit.all6, newdata = dat6)
dat6 = dat6 %>% mutate(test_fit) %>% arrange(desc(test_fit))
dat_map6 <- dat6 %>% mutate(Mobility, MobilityCat = ifelse(test_fit > 0.1, "high", "low"))
dat_map6 <- dat_map6 %>% filter(Longitude > -150, Latitude < 50)
ggplot(dat_map6, aes(x = dat_map6$Longitude, y = dat_map6$Latitude, col = dat_map6$MobilityCat)) + geo
m_point() +
  labs(x = "Longitude", y = "Latitude", colour = "MobilityCat", title = "plot of continental USA with
 predicted M before NA control")
```



First of all, i tried to take a look on if NA Values spotted in USA has a certain patter. I tried to have a look if NA happens in areas with small population density. However, by drawing the plot below, i was able to conclude there is no such relationship with population and NA. Rather than that, NA was spotted regardless of population. Therefore, i decided to replace NA into mean of each variable.

```
ggplot(mapping = aes(log(dat_map6$Population), dat_map6$Mobility, color = is.na(dat_map6$MobilityCa
t))) + geom_point() +
  labs(x = "Population", y = "Mobility")
```

Below, by using colsums() function, i tried to have a look in what column has the most NA value. Since NAs in diffrent variables make total 213 NA in test_fit value, i would love to replace NA to mean of each column variable.

```
colSums(is.na(dat6))
```

```
##                  Mobility                Population
##                        12                         0
##                     Urban                     Black
##                         0                         0
##               Seg_racial                Seg_income
##                         0                         0
##                   Commute                  Gini_99
##                         0                        32
##              Middle_class             Local_tax_rate
##                        32                         1
##       Local_gov_spending               Progressivity
##                         2                         0
##                      EITC             School_spending
##                         0                        10
##     Student_teacher_ratio               Test_scores
##                        30                        36
##               HS_dropout                    Tuition
##                       148                         0
## Labor_force_participation              Manufacturing
##                         0                         0
##             Teenage_labor               Migration_in
##                        32                        17
##             Migration_out             Social_capital
##                        17                        19
##                 Religious             Violent_crime
##                         0                        27
##            Single_mothers                  Divorced
##                         0                         0
##                   Married                 Longitude
##                         0                         0
##                  Latitude                        HE
##                         0                         0
##                  test_fit
##                       213
```

```
dat6a <- dat44ca[,-c(1,2,4,10,11,13,14,15,26,28,31,35)]
dat6b <- na.omit(dat6a$Gini_99)
dat6c <- na.omit(dat6a$Middle_class)
dat6d <- na.omit(dat6a$School_spending)
dat6e <- na.omit(dat6a$Student_teacher_ratio)
dat6f <- na.omit(dat6a$Test_scores)
dat6g <- na.omit(dat6a$HS_dropout)
dat6h <- na.omit(dat6a$Teenage_labor)
dat6i <- na.omit(dat6a$Migration_in)
dat6j <- na.omit(dat6a$Migration_out)
dat6k <- na.omit(dat6a$Social_capital)
dat6l <- na.omit(dat6a$Violent_crime)
dat6a <- dat6a %>% mutate(Gini_99 = ifelse(is.na(Gini_99),mean(dat6b), Gini_99)) %>%
  mutate(Middle_class = ifelse(is.na(Middle_class),mean(dat6c), Middle_class)) %>%
  mutate(School_spending = ifelse(is.na(School_spending),mean(dat6d), School_spending)) %>%
  mutate(Student_teacher_ratio = ifelse(is.na(Student_teacher_ratio),mean(dat6e), Student_teacher_rati
o)) %>%
  mutate(Test_scores = ifelse(is.na(Test_scores),mean(dat6f), Test_scores)) %>%
  mutate(HS_dropout = ifelse(is.na(HS_dropout),mean(dat6g), HS_dropout)) %>%
  mutate(Teenage_labor = ifelse(is.na(Teenage_labor),mean(dat6h), Teenage_labor)) %>%
  mutate(Migration_in = ifelse(is.na(Migration_in),mean(dat6i), Migration_in)) %>%
  mutate(Migration_out = ifelse(is.na(Migration_out),mean(dat6j), Migration_out)) %>%
  mutate(Social_capital = ifelse(is.na(Social_capital),mean(dat6k), Social_capital)) %>%
  mutate(Violent_crime = ifelse(is.na(Violent_crime),mean(dat6l), Violent_crime))
```

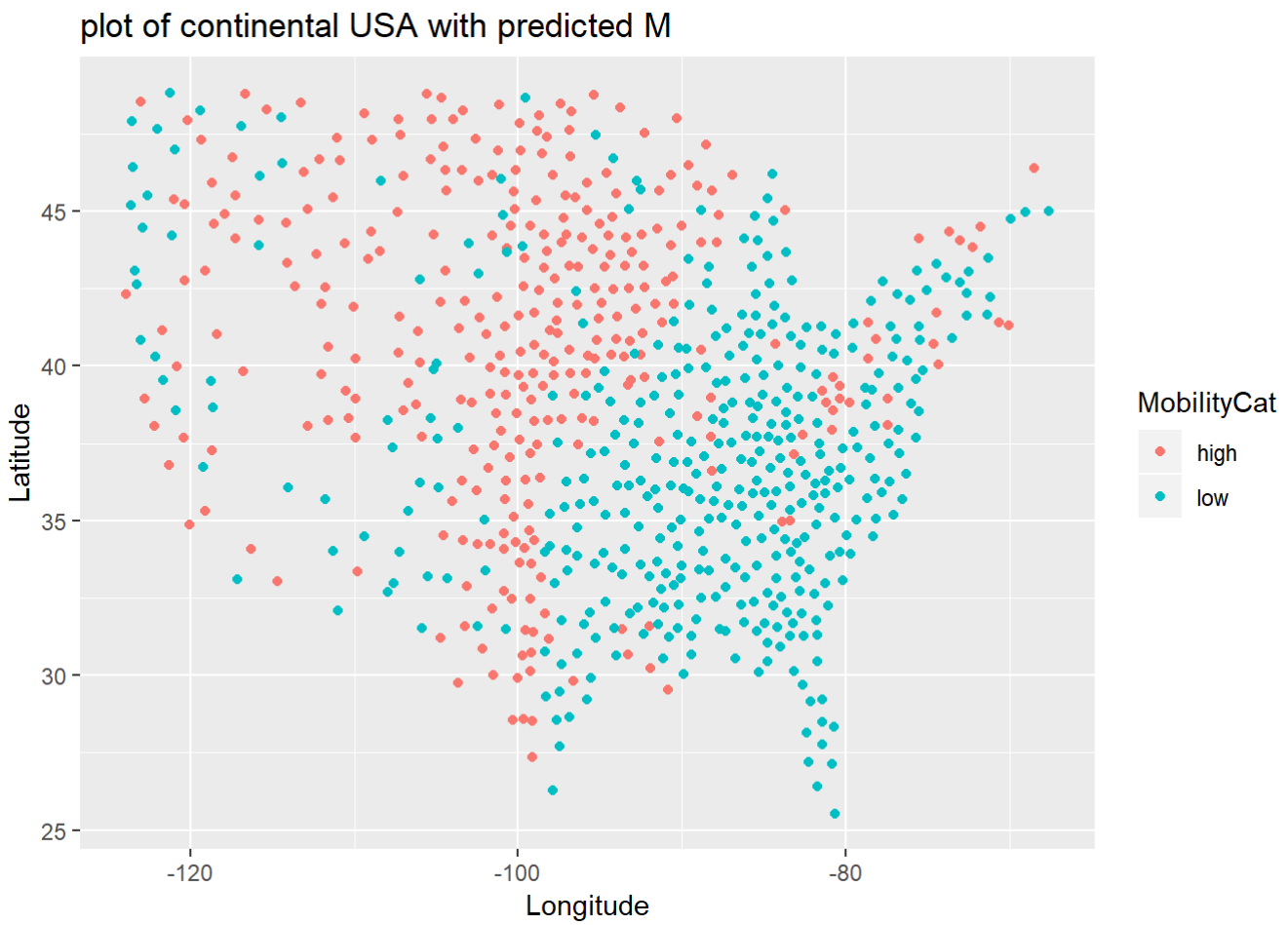
NA inside each variable has been replaced to mean of each variables. This was able because i found out that there is no relationship between NA and population.

```
colSums(is.na(dat6a))
```

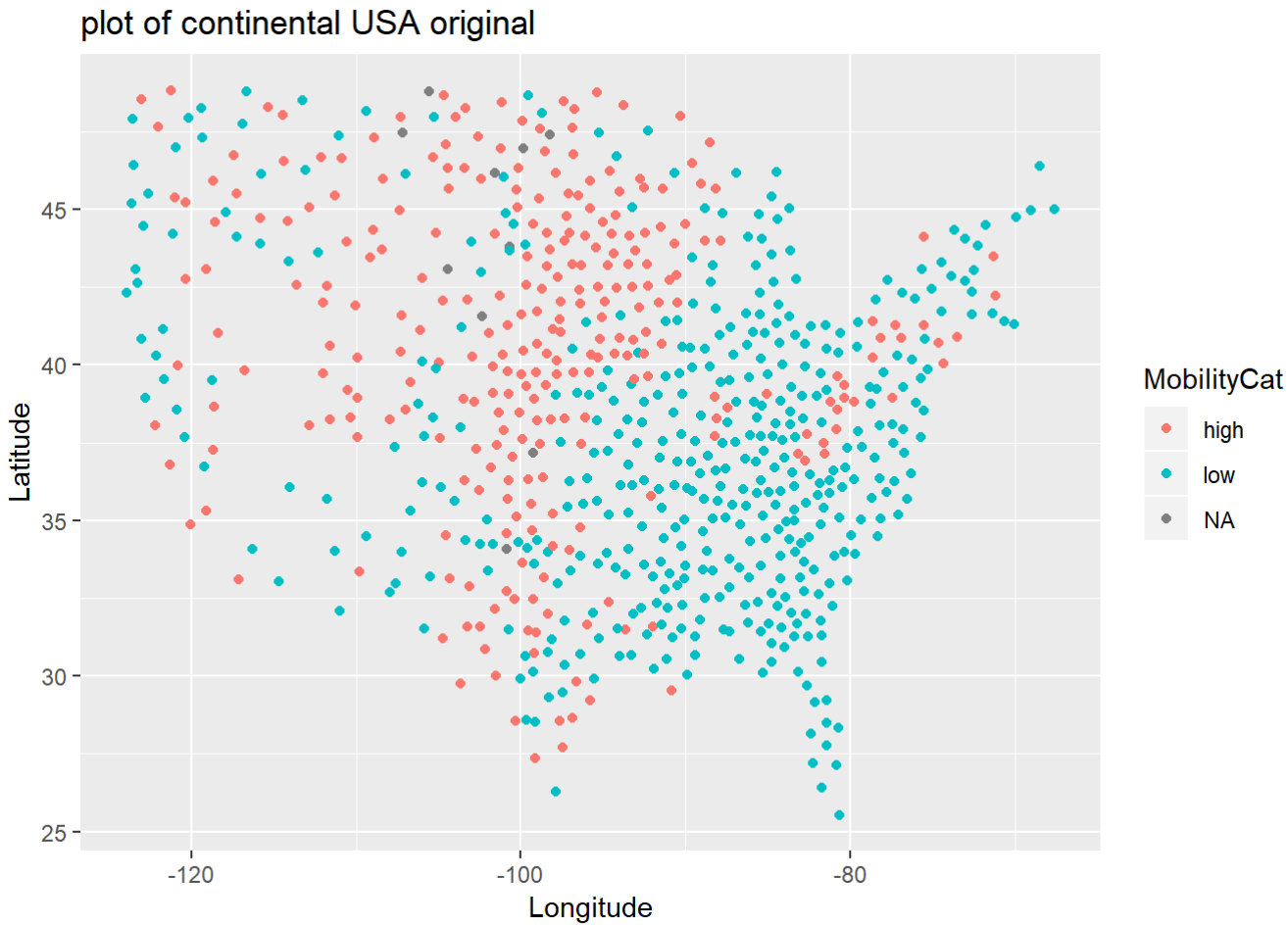
```
##                  Mobility                Population
##                        12                         0
##                     Urban                     Black
##                         0                         0
##                Seg_racial                Seg_income
##                         0                         0
##                   Commute                   Gini_99
##                         0                         0
##              Middle_class             Local_tax_rate
##                         0                         1
##        Local_gov_spending              Progressivity
##                         2                         0
##                      EITC            School_spending
##                         0                         0
##     Student_teacher_ratio               Test_scores
##                         0                         0
##                HS_dropout                   Tuition
##                         0                         0
## Labor_force_participation              Manufacturing
##                         0                         0
##              Teenage_labor               Migration_in
##                         0                         0
##              Migration_out             Social_capital
##                         0                         0
##                  Religious              Violent_crime
##                         0                         0
##             Single_mothers                  Divorced
##                         0                         0
##                    Married                 Longitude
##                         0                         0
##                   Latitude                        HE
##                         0                         0
```

With new data with only few NA, i have drawn the predicted mobility plot once again. As you can see, now there is literally no NA values in the plot. Thus, to compare its quality with actualy Mobility plot, i can see several spots that have changed its color with mobility prediction, but still majority of them matches with original mobility. In conclusion, i would love to say the prediction quality is not bad at all.

```
fit.all6a <- lm(Mobility ~ . -Longitude -Latitude, data = dat6a)
test_fita = predict(fit.all6a, newdata = dat6a)
dat6b = dat6a %>% mutate(test_fita) %>% arrange(desc(test_fita))
dat_map6a <- dat6b %>% mutate(MobilityCat = ifelse(test_fita > 0.1, "high", "low"))
dat_map6a <- dat_map6a %>% filter(Longitude > -150, Latitude < 50)
ggplot(dat_map6a, aes(x = dat_map6a$Longitude, y = dat_map6a$Latitude, col = dat_map6a$MobilityCat)) +
 geom_point() +
  labs(x = "Longitude", y = "Latitude", colour = "MobilityCat", title = "plot of continental USA with
 predicted M")
```




```
dat_map <- dat %>% mutate(Mobility, MobilityCat = ifelse(Mobility > 0.1, "high", "low"))
dat_map <- dat_map %>% filter(Longitude > -150, Latitude < 50)
ggplot(dat_map, aes(x = dat_map$Longitude, y = dat_map$Latitude, col = dat_map$MobilityCat)) + geom_po
int() +
  labs(x = "Longitude", y = "Latitude", colour = "MobilityCat", title = "plot of continental USA origi
nal")
```



7 Just because I was there

Find Pittsburgh in the data set. For this question, assume that the model (you have fitted just before) is well-fitted.

```
dat7 <- dat44ca[dat44ca$Name == 'Pittsburgh',]
dat7
```

```
##         ID       Name   Mobility State Population Urban Black Seg_racial
## 229 16300 Pittsburgh 0.09514869    PA    2561364     1 0.075       0.323
##     Seg_income Seg_poverty Seg_affluence Commute Income Gini Share01
## 229      0.081       0.069         0.089   0.287  38686 0.45   13.848
##     Gini_99 Middle_class Local_tax_rate Local_gov_spending Progressivity
## 229   0.312        0.533          0.025                2244             0
##     EITC School_spending Student_teacher_ratio Test_scores HS_dropout
## 229    0           8.199                  18.1       5.035      -0.01
##     Colleges Tuition Graduation Labor_force_participation Manufacturing
## 229    0.022    8789      0.092                     0.599         0.124
##     Chinese_imports Teenage_labor Migration_in Migration_out Foreign_born
## 229           0.556         0.004        0.007         0.011        0.025
##     Social_capital Religious Violent_crime Single_mothers Divorced Married
## 229          0.604     0.651         0.002          0.209    0.082   0.545
##     Longitude Latitude   HE
## 229 -79.57613 40.56791 TRUE
```

a. What its actual mobility? What is its predicted mobility, according to the model?

Actual mobility of Pitssburgh is 0.09514869 while predicted value is 0.09426714.

```
na.omit(dat6b[dat6b$Mobility == 0.09514869,c(1,33)])
```

```
##       Mobility  test_fita
## 378 0.09514869 0.09426714
```

b. Holding all else fixed, what is the predicted mobility if the violent crime rate is doubled? If it is halved?

If the crime rate is halved, the predicted mobility rises to 0.09871708, while original was 0.09426714. If the crime rate is doubled, the predicted mobility dropes to 0.08536725

```
Pits = dat6b %>% filter(dat6b$Mobility == 0.09514869 )
Pits$Violent_crime <- Pits$Violent_crime/2
pitsa <- predict(fit.all6a, newdata = Pits)
Pits = dat6b %>% filter(dat6b$Mobility == 0.09514869 )
Pits$Violent_crime <- Pits$Violent_crime
pitsb <- predict(fit.all6a, newdata = Pits)
Pits = dat6b %>% filter(dat6b$Mobility == 0.09514869 )
Pits$Violent_crime <- Pits$Violent_crime*2
pitsc<- predict(fit.all6a, newdata = Pits)
pitsall <- cbind(pitsa, pitsb, pitsc)
pitsall <- as.data.frame(pitsall)
colnames(pitsall) <- c("Crime /2 ", "Crime original", "Crime *2")
pitsall
```

```
##    Crime /2  Crime original   Crime *2
## 1 0.09871708     0.09426714 0.08536725
```

c. Provide a 95% confidence interval for the expected mobility at Pittsburgh.

Expected confidence interval would be 0.0841 ~ 0.1043.

```
Pits = dat6b %>% filter(dat6b$Mobility == 0.09514869 )
predict(fit.all6a, newdata = Pits, interval = "confidence")
```

```
##          fit        lwr       upr
## 1 0.09426714 0.08416592 0.1043684
```

d. Provide a 95% prediction interval for the expected mobility at Pittsburgh. Explain the difference.

Prediction interval would be ranged 0.0363 ~ 0.15220

```
Pits = dat6b %>% filter(dat6b$Mobility == 0.09514869 )
predict(fit.all6a, newdata = Pits, interval = "prediction")
```

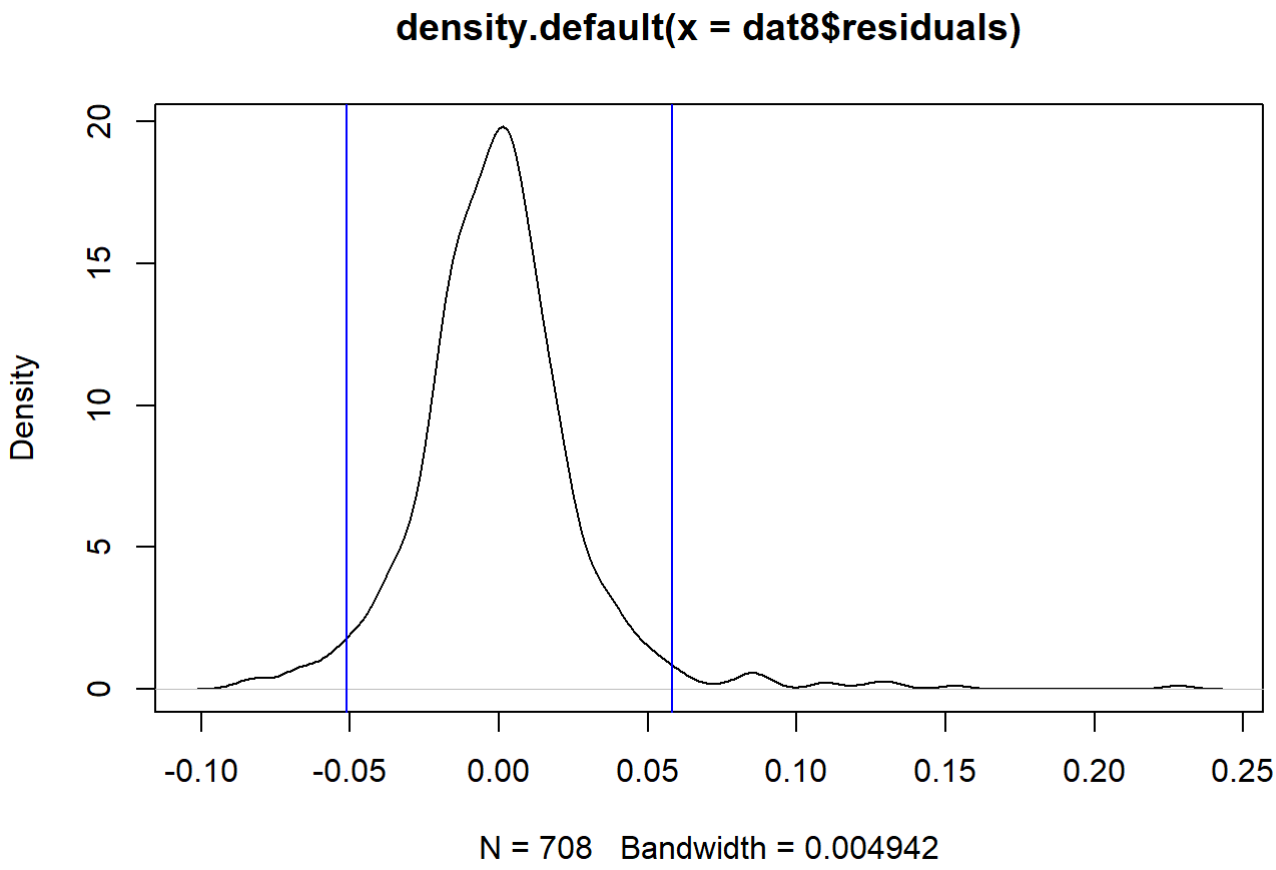
```
##          fit        lwr       upr
## 1 0.09426714 0.03632637 0.1522079
```

8. After making proper allowances

a. Make a map of the model's residuals.

Below is a plot of how residuals are distributed. I am going to make a map with residuals diveded to high and low for better visibility.

```
num <- as.numeric(names(fit.all6a$residuals))
dat8 <- dat6b[num,]
dat8$residuals <- fit.all6a$residuals
dat8 <- dat8 %>% filter(Longitude > -150, Latitude < 50)
plot(density(dat8$residuals))
abline(v = quantile(dat8$residuals,probs=c(0.025,0.975), type=5), col = 'blue')
```

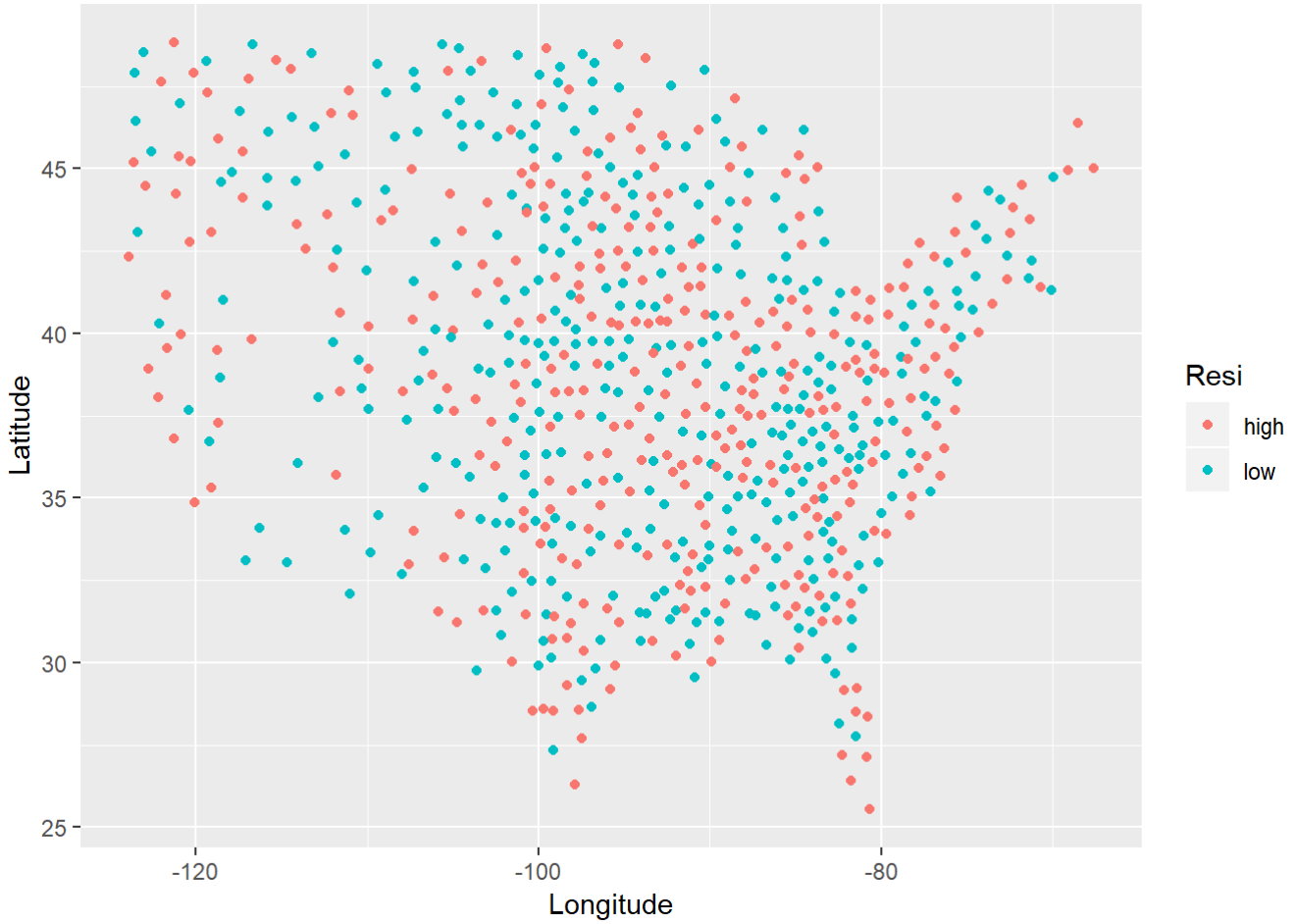



```
mean(dat8$residuals,probs=c(0.025,0.975), type=5)
```

```
## [1] -0.0001891521
```

I have made a new variable Resi to classify residuals into high and low. With this new variable resi, i was able to see in which location residual is high. However, i was not able to find any pattern within this map.

```
dat8 <- dat8 %>% mutate(Resi = ifelse(dat8$residuals > -0.0001891521, "high", "low"))
map_res <- ggplot(data = dat8, mapping = aes(x = Longitude, y = Latitude, col = Resi)) +
  geom_point()
map_res
```



b. What are the five communities with the largest positive residuals? The five with the most negative residuals? Provide the names of the communies.

The name of the highest residuals would be Bowman, Willston, Dickinson, Carrington, Lemmon. The name of the lowest residuals would be Shelby, Seymour, Littlefield, Nantucket and Bethel.

```
num <- as.numeric(names(fit.all6a$residuals))
dat8c <- dat[num,]
dat8c$residuals <- fit.all6a$residuals
high <- dat8c[,c(2,44)] %>% arrange(desc(residuals)) %>% head(5)
low <- dat8c[,c(2,44)] %>% arrange(residuals) %>% head(5)
cbind(high,low)
```

```
##         Name residuals        Name   residuals
## 1     Bowman 0.2280201      Shelby -0.08675708
## 2  Williston 0.1530334     Seymour -0.08254459
## 3  Dickinson 0.1393201 Littlefield -0.08139703
## 4 Carrington 0.1320135   Nantucket -0.07683132
## 5     Lemmon 0.1303660      Bethel -0.07012253
```

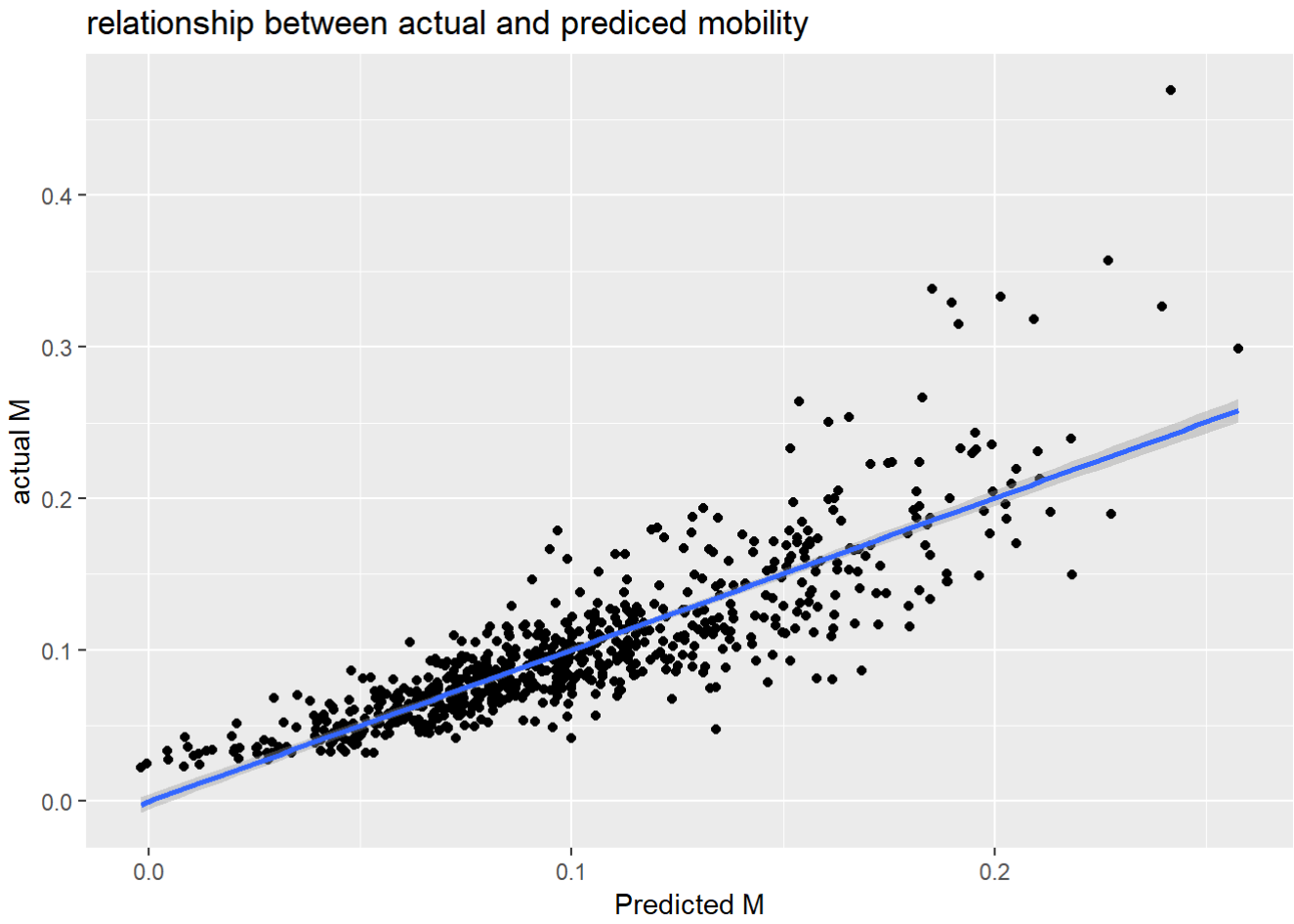
```
prettymap({
  bmaps.plot(map2)
  osm.points(dat8ca$Longitude, dat8ca$Latitude, pch=1, cex=0.6, col = 'white')
  osm.text(dat8ca$Longitude, dat8ca$Latitude, labels= dat8ca$Name, adj=c(-0.2, 0.5), cex=0.8, col = 'w
hite')})
```

9. Expectations and reality

a. Make a scatterplot of actual mobility against predicted mobility. Is the relationship linear? Should it be, is the model right? Is the relationship flat? Should it be, is the model right?

I believe the relationship should be linear. Being flat means that acutaly mobility and predicted mobility does not have any relationship. This implies that my prediction could be wrong. However, if the relationship is linear, which means if actual value grows predicted value also grows, then i can conclude that my prediction might be corret.

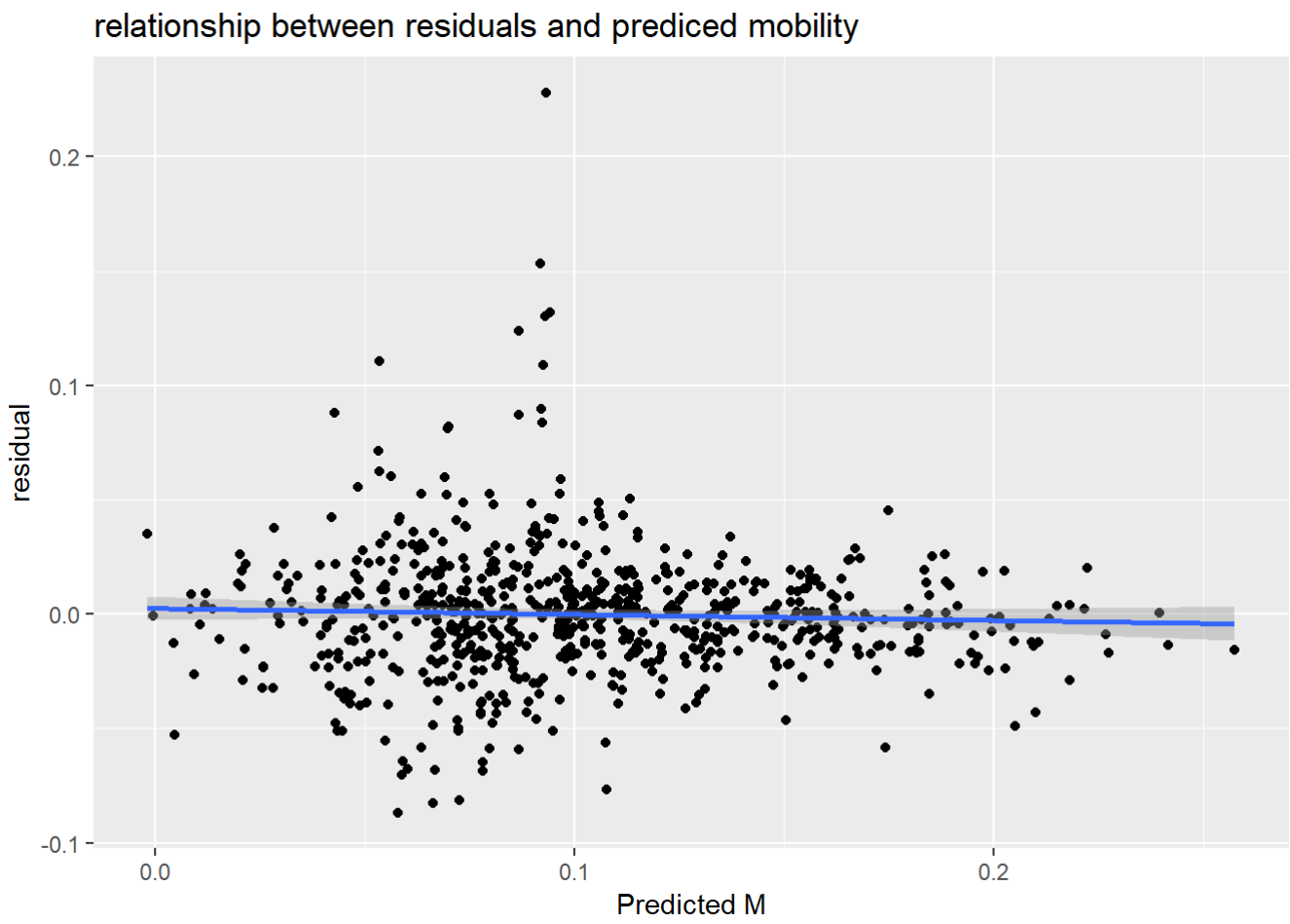
```
ggplot(data = dat_map6a, mapping = aes(x = dat_map6a$test_fit, y = dat_map6a$Mobility)) +
  geom_point() + geom_smooth(method = 'lm')  +
  labs(x = "Predicted M", y = "actual M", title = "relationship between actual and prediced mobility")
```



b. Make a scatterplot of the model's residuals against predicted mobility. Is the relationship linear? Should it be, is the model right? Is the relationship flat? Should it be, is the model right?

Relationship between residuals and predicted mobility is flat. This implies that there is no relationship between residuals and predicted mobility. Thus i can conclude that this model is right. (If there is a relationship between residual and predicted value, this implies that i am missing some variable in my prediction.)

```
ggplot(data = dat8, mapping = aes(x = dat8$test_fita, y = dat8$residuals)) +
  geom_point() + geom_smooth(method = 'lm')  +
  labs(x = "Predicted M", y = "residual", title = "relationship between residuals and prediced mobilit
y")
```



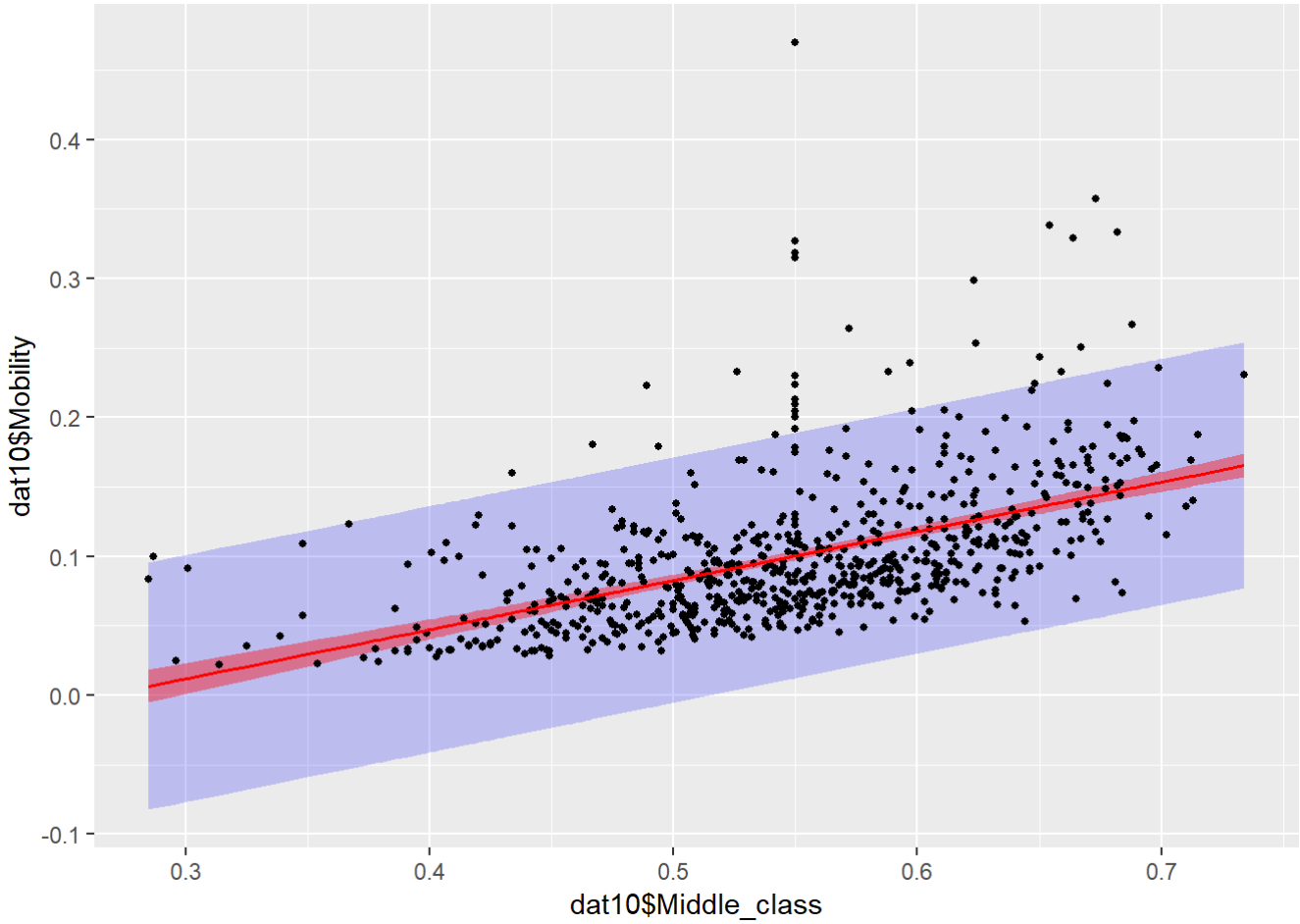
10. Cross-validation, bootstrap and smoothing

For this question, focus on predicting mobility by the fraction of middle class in the community.

a. Fit a simple linear regression model y = β0+β1x+ϵ to predict Mobility by Middle_class. Create a plot showing the data points, the fitted regression line, and 95% confidence and prediction intervals. For the intervals, assume that the error ϵ is i.i.d. N(0, σ2). (Is the assumption right?)

Seemeingly purple line is prediction interval and red line is confidence interval.

```
fit.mid <- lm(Mobility ~ Middle_class, data = dat6a)
pred_mid <- predict(fit.mid, dat6a, interval = "prediction")
colnames(pred_mid) <- c("fit_pred", "lwr_pred", "upr_pred")
dat10 <- cbind(dat6a, pred_mid)
conf_mid <- predict(fit.mid, dat6a, interval = "confidence")
colnames(conf_mid) <- c("fit_conf", "lwr_conf", "upr_conf")
dat10 <- cbind(dat10, conf_mid)
ggplot(dat10, aes(x = dat10$Middle_class)) +
  geom_ribbon(aes(ymin = dat10$lwr_pred, ymax = dat10$upr_pred),
              fill = "blue", alpha = 0.2) +
  geom_ribbon(aes(ymin = dat10$lwr_conf, ymax = dat10$upr_conf),
              fill = "red", alpha = 0.4) +
  geom_point(aes(y = dat10$Mobility), colour = "black", size = 1) +
  geom_line(aes(y = dat10$fit_pred), colour = "red", size = 0.7)
```



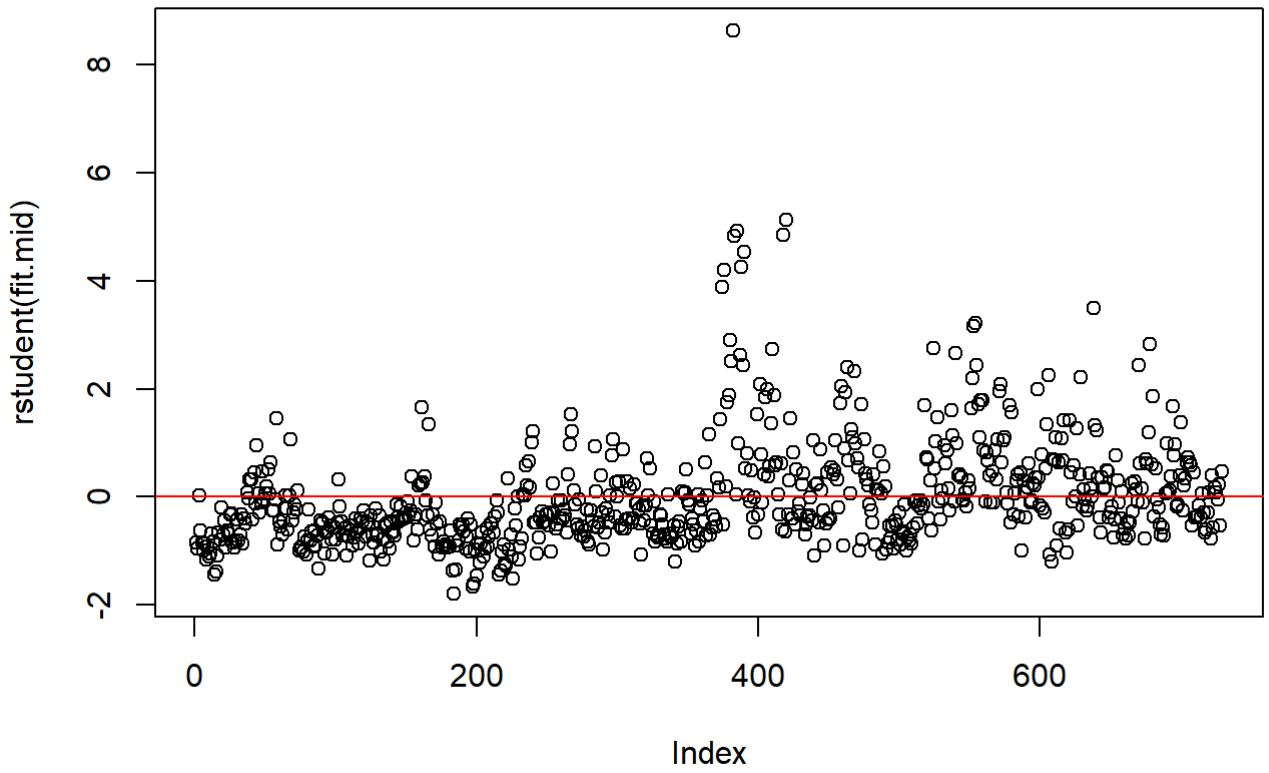
In order to use Regression Analysis, we have to look over standardized resuduals distribution. By taking a look over this distribution, we can know if our assumptions used in Regression Analysis are acceptable. To do so, i have made a plot with rstudent(linearmodel). By looking over this data, i was able to figure out that

Residuals must evenly distributed from Y = 0. However, i can say that below 200, most residuals are under 0 and around 400 ~ 600, residuals are over the 0 horizontal line. Thus, i cannot say that in every X, residuals are evenly distributed.

Most of the residuals should be within range of +-2. However, i was able to see a lot of outliers.

In conclusion, this plot shows no evidence that we can continue with our Regression Analysis. This is the basic reason why we have to move on smoothing spline in questions that will follow.

```
plot(rstudent(fit.mid))
abline(h=0, col ='red')
```



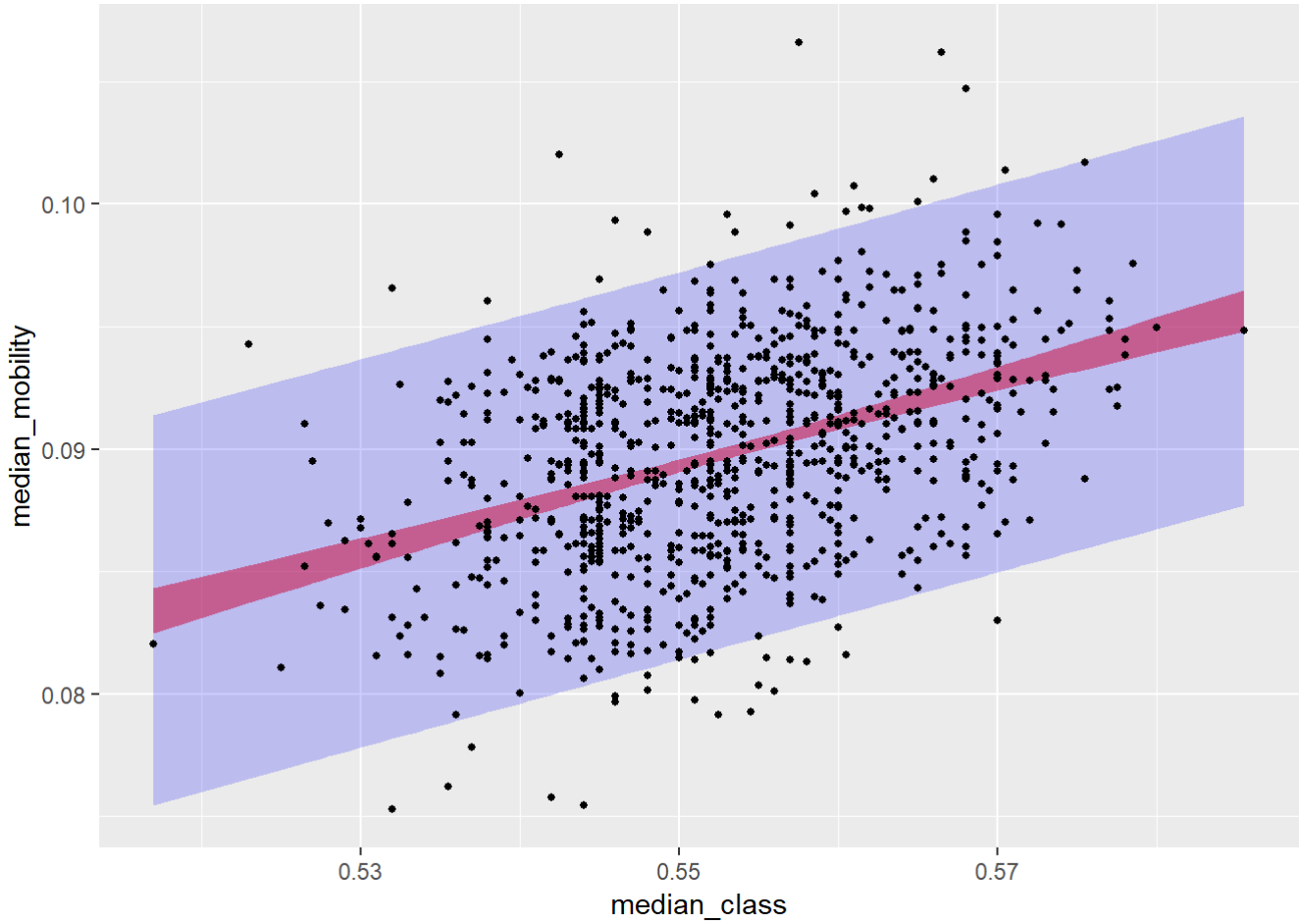
b. Use a resampling method to obtain 95% confidence interval. Can you build a 95% prediction interval?

I had some difficulties in understanding the question. Thus i would like to give two type of solutions within this question.

[First]

The first way is to resample mobility and middle_class and apply lm function. By using this method, i have drawn the confidence interval and prediction interval.

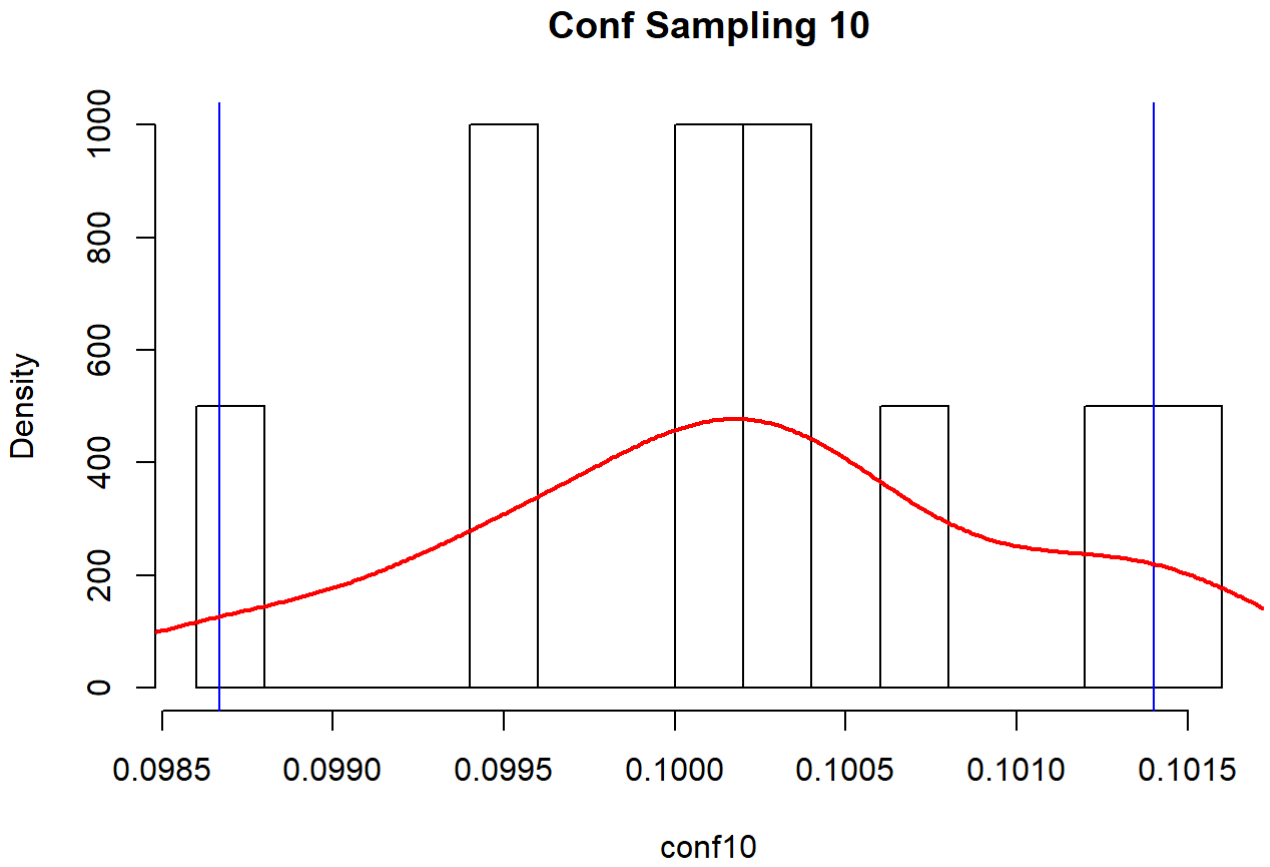
```
boot <- list()

for(i in 1:1000){
  boot[[i]] <- dat %>%
    sample_n(size = 100) %>%
    summarise(median_mobility = median(Mobility, na.rm = T), median_class = median(Middle_class, na.rm
 = T))
}

boot10 = data.table::rbindlist(boot)
model = lm(median_mobility ~ median_class, data = boot10)
con = predict(model, newdata = boot10, interval = 'confidence')
pre = predict(model, newdata = boot10, interval = 'prediction')

ggplot(boot10, aes(x = median_class)) +
  geom_ribbon(aes(ymin = con[,"lwr"], ymax = con[,"upr"]),
              fill = "red", alpha = 0.5) +
  geom_ribbon(aes(ymin = pre[,"lwr"], ymax = pre[,"upr"]),
              fill = "blue", alpha = 0.2) +
  geom_point(aes(y = median_mobility), colour = "black", size = 1)
```



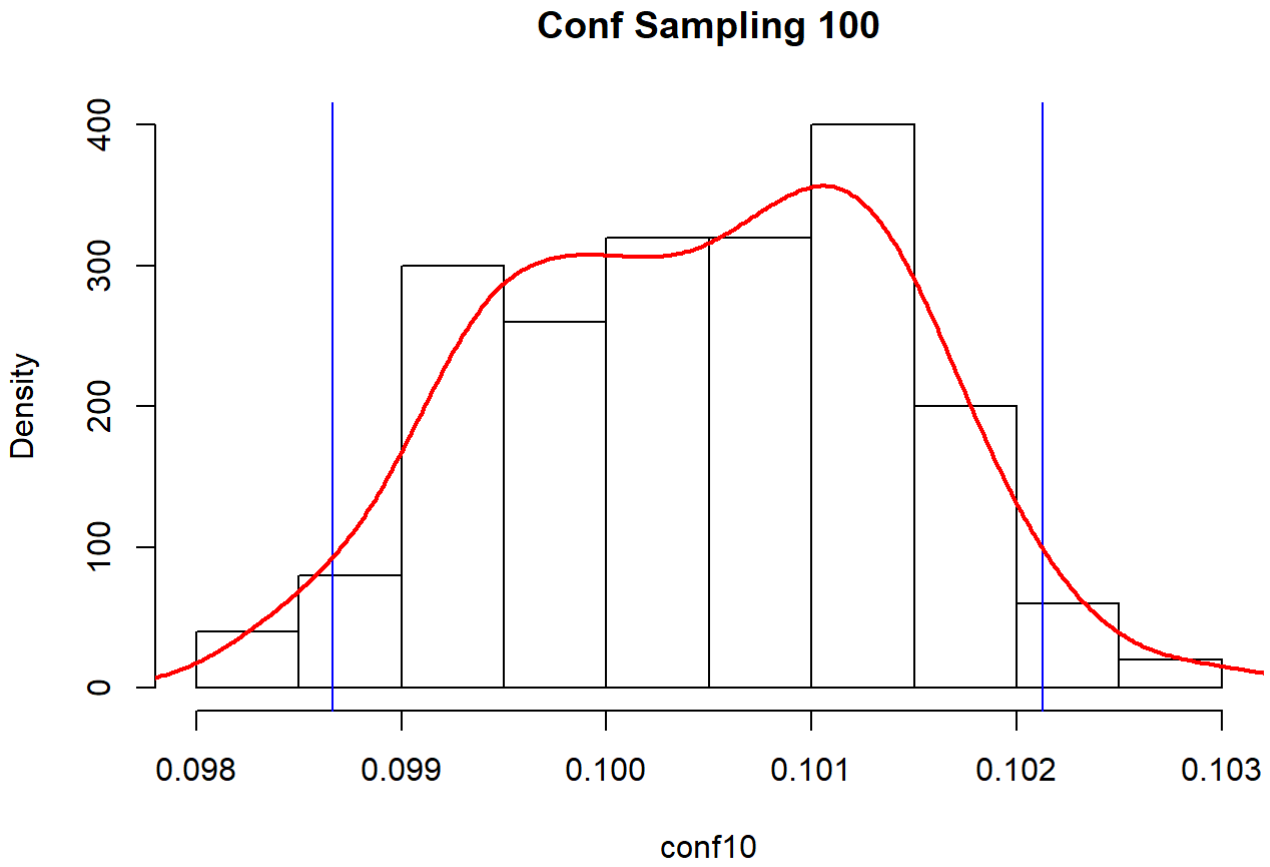
[Second]

Second way is to apply lm function into original data and export only fitted value with confidnce and predicted option in lm. Afterward, resample within those fitted values. Below are plots with resampling method with diffrent sampling n 10, 100, 1000 and 10000. You can see as number gets larger, it follows normal distribution. 95% confidence interval is printted as blue line.

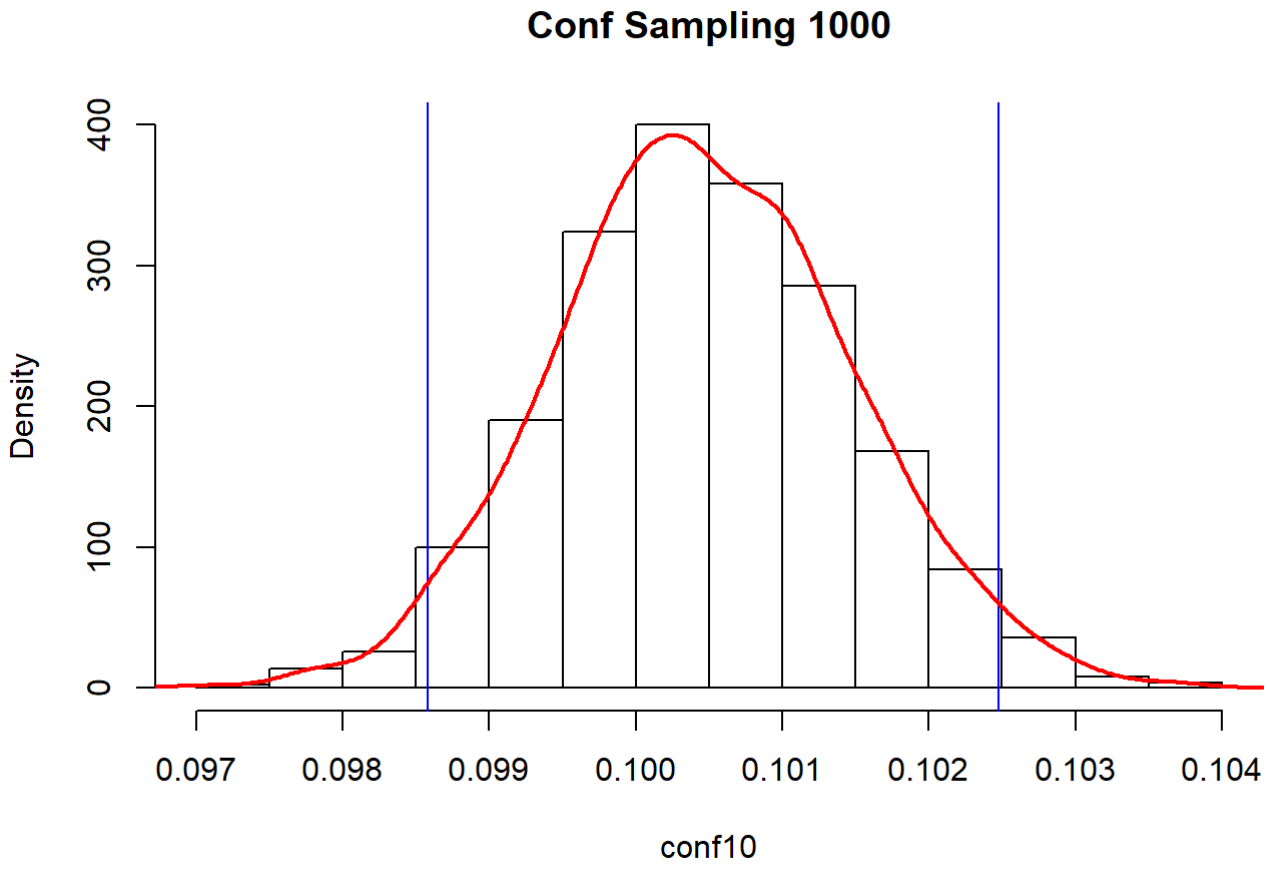
```
dat10b <- na.omit(dat10[,36])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-10
conf10 <- NULL
for(draw in 1:sampN) {
    conf10 = c(conf10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(conf10, prob=T, breaks =10, main = 'Conf Sampling 10')
x <- conf10
abline( v = quantile(conf10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(conf10),col="red", lwd=2)
```



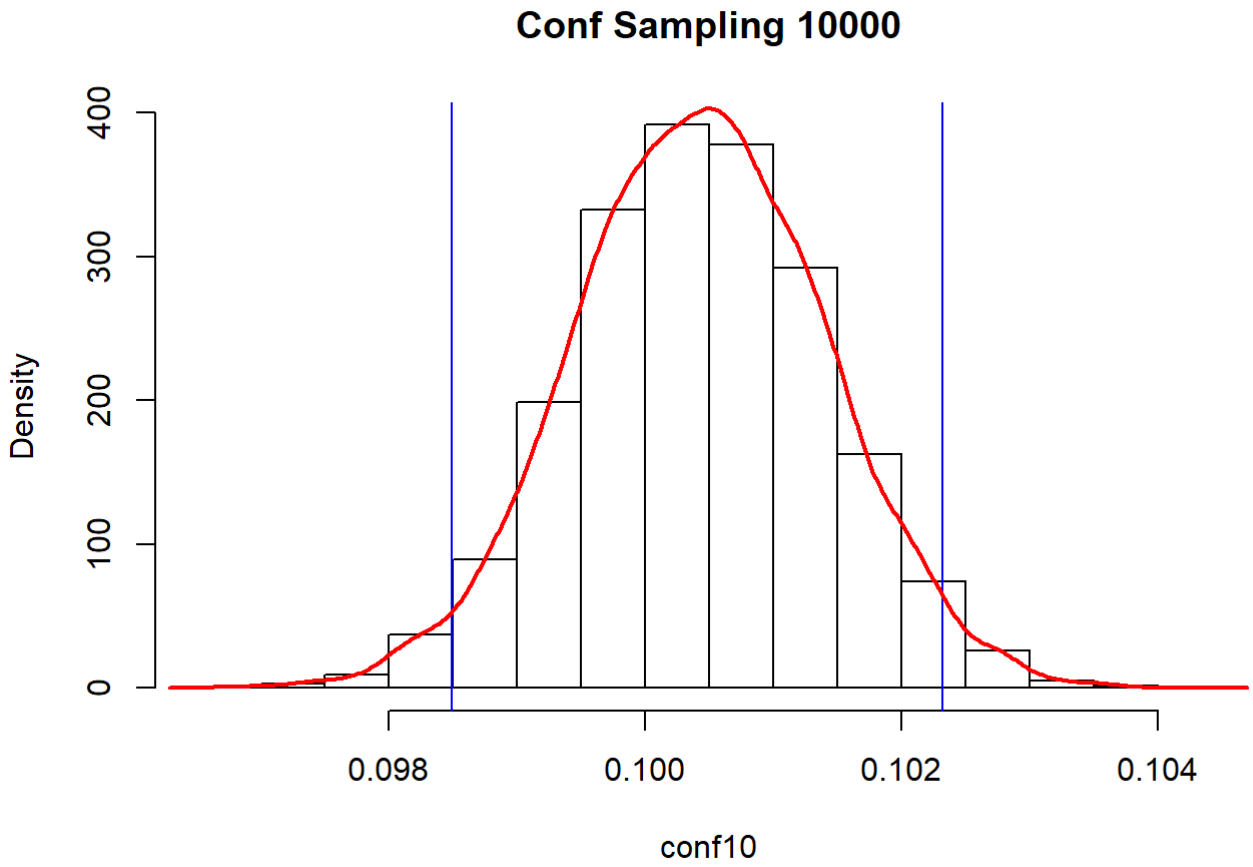

```
dat10b <- na.omit(dat10[,36])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-100
conf10 <- NULL
for(draw in 1:sampN) {
    conf10 = c(conf10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(conf10, prob=T, main = 'Conf Sampling 100')
x <- conf10
abline( v = quantile(conf10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(conf10),col="red", lwd=2)
```




```
dat10b <- na.omit(dat10[,36])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-1000
conf10 <- NULL
for(draw in 1:sampN) {
    conf10 = c(conf10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(conf10, prob=T, main = 'Conf Sampling 1000')
x <- conf10
abline( v = quantile(conf10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(conf10),col="red", lwd=2)
```

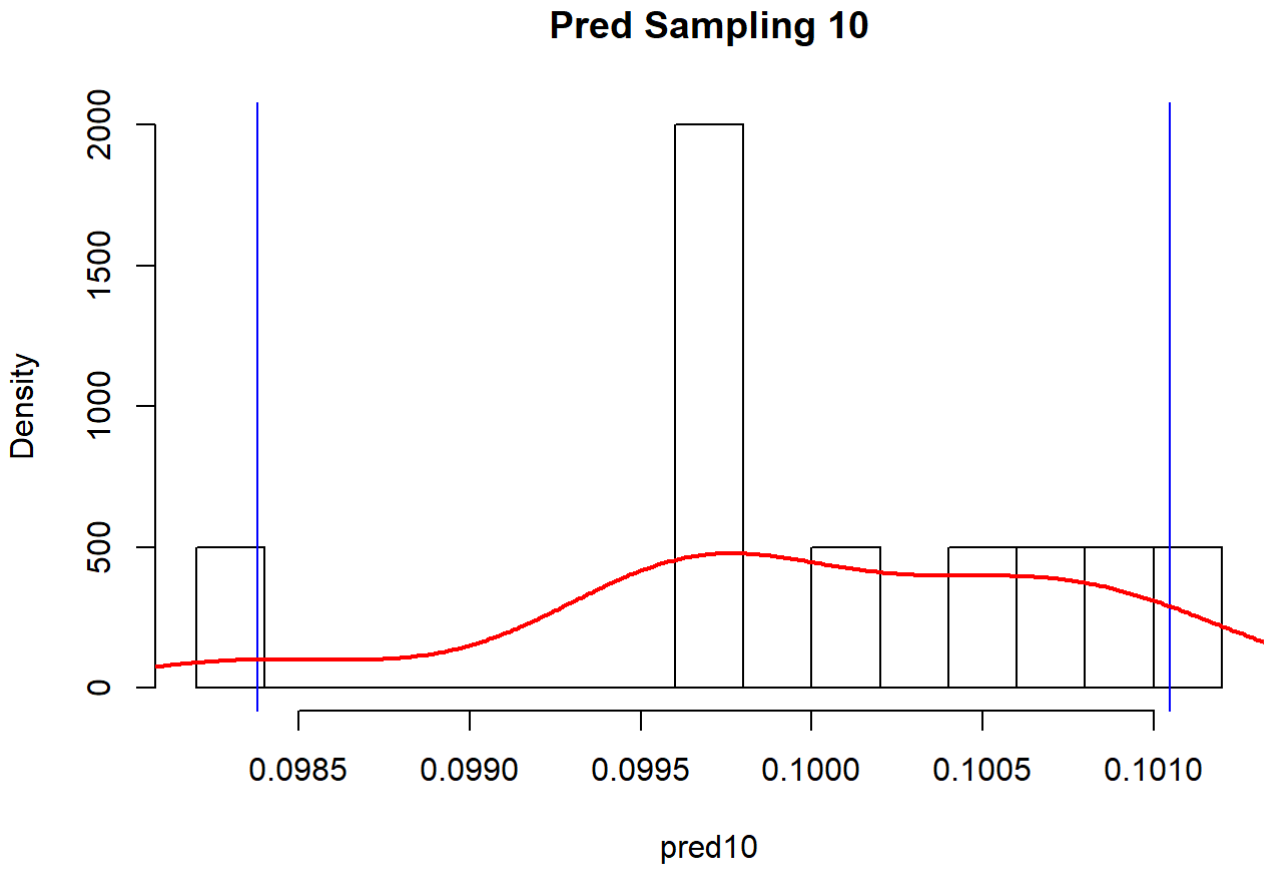



```
dat10b <- na.omit(dat10[,36])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-10000
conf10 <- NULL
for(draw in 1:sampN) {
    conf10 = c(conf10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(conf10, prob=T, main = 'Conf Sampling 10000')
x <- conf10
abline( v = quantile(conf10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(conf10),col="red", lwd=2)
```

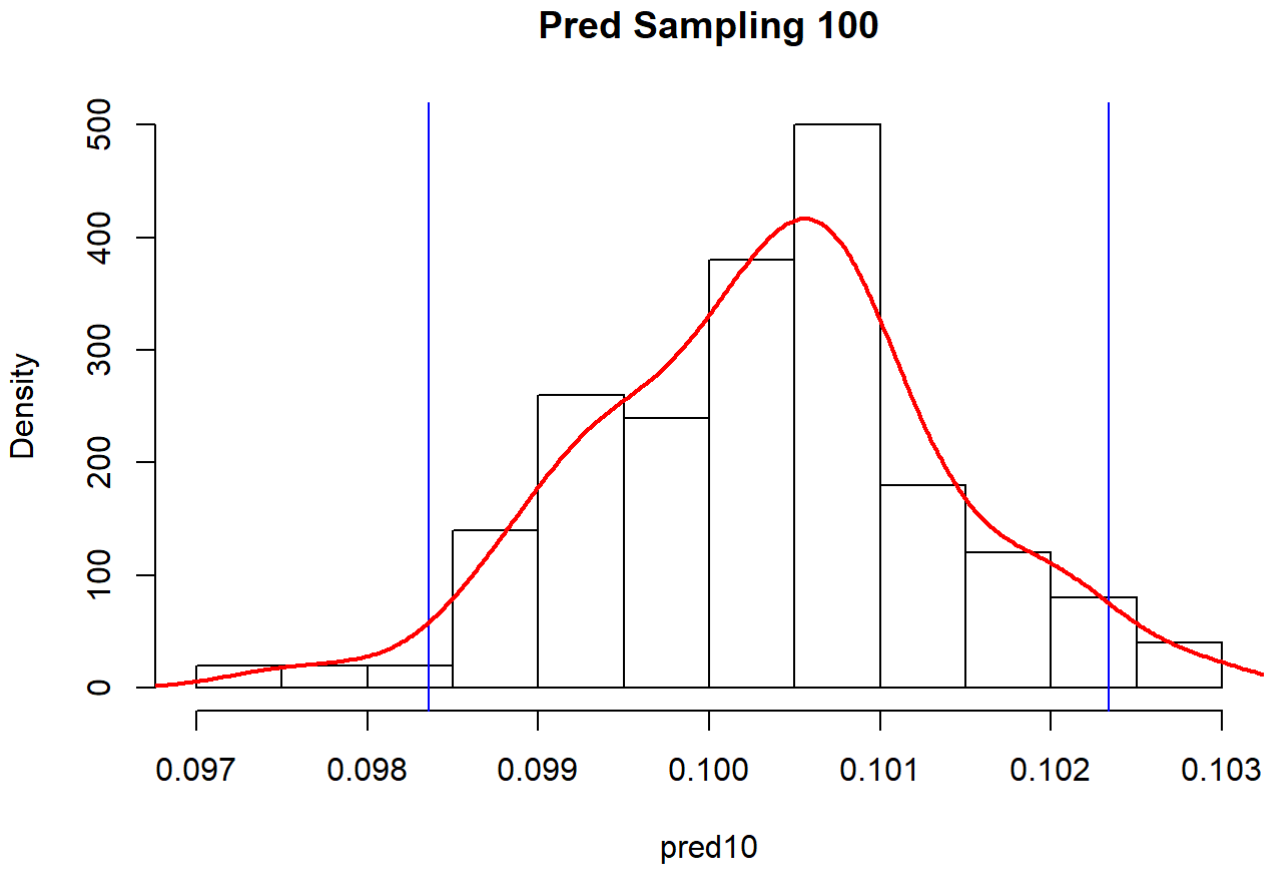


Below are plots with predicted value and 95% interval. As well as confidence interval, you can see it follows normality.

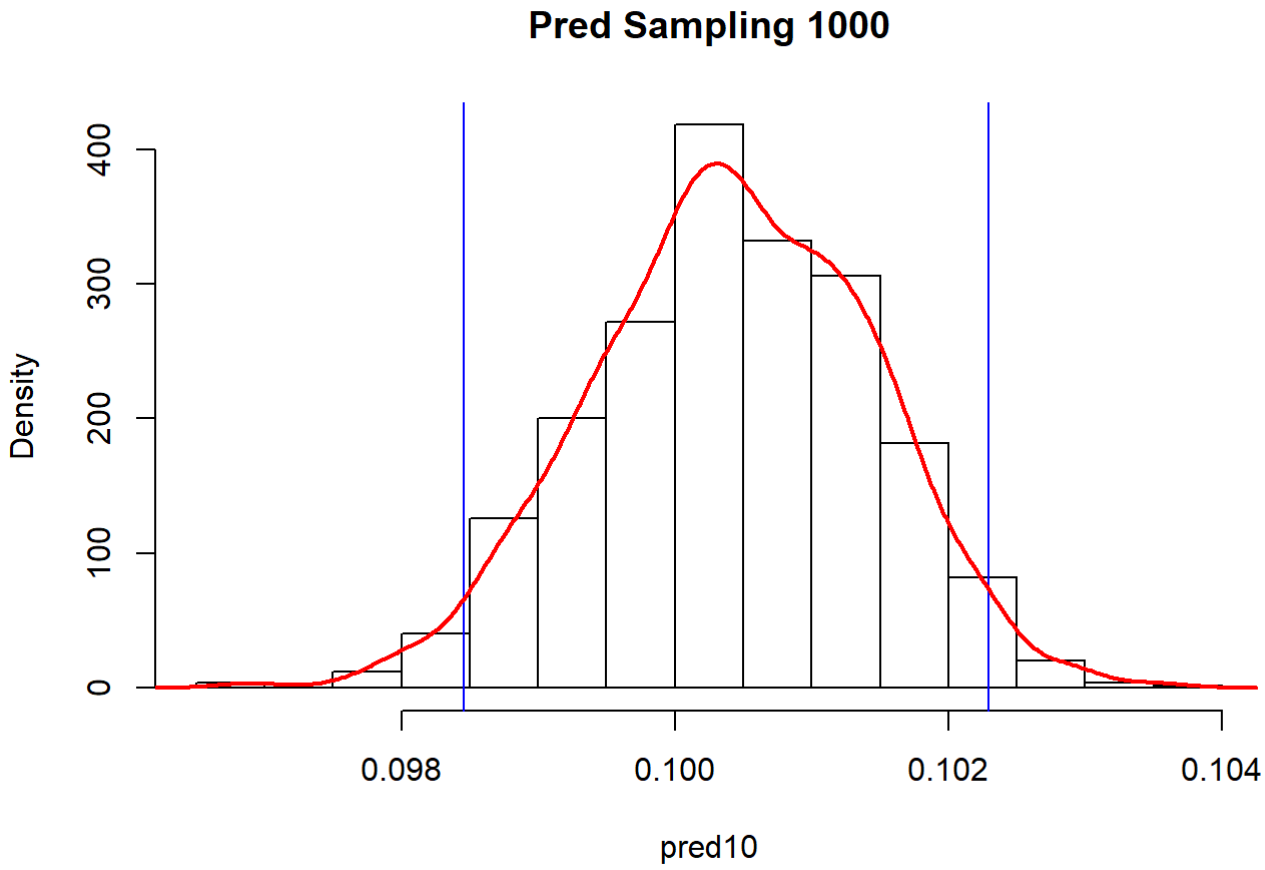
```
dat10b <- na.omit(dat10[,33])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-10
pred10 <- NULL
for(draw in 1:sampN) {
    pred10 = c(pred10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(pred10, prob=T, breaks = 10 ,main = 'Pred Sampling 10')
x <- pred10
abline( v = quantile(pred10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(pred10),col="red", lwd=2)
```



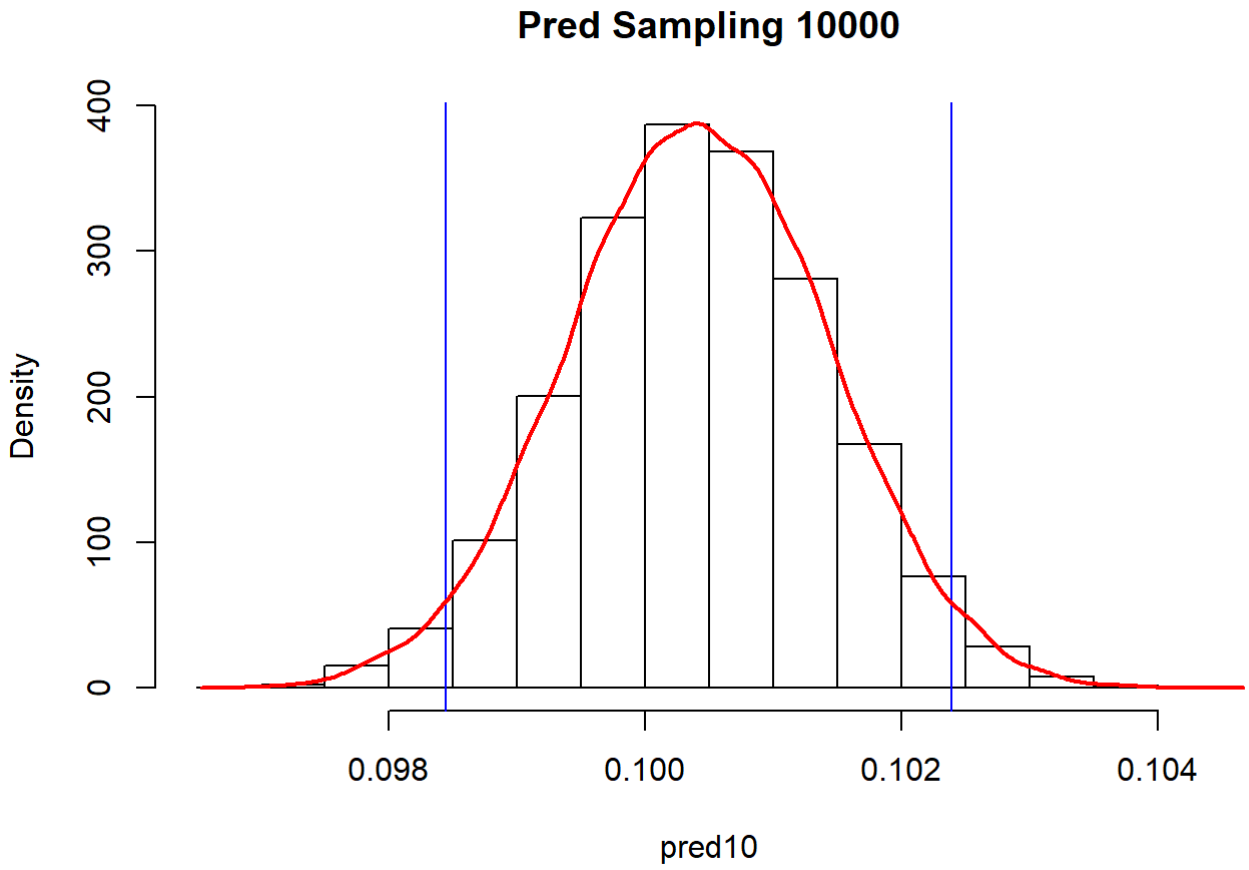

```
dat10b <- na.omit(dat10[,33])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-100
pred10 <- NULL
for(draw in 1:sampN) {
    pred10 = c(pred10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(pred10, prob=T, breaks = 10 ,main = 'Pred Sampling 100')
x <- pred10
abline( v = quantile(pred10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(pred10),col="red", lwd=2)
```




```
dat10b <- na.omit(dat10[,33])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-1000
pred10 <- NULL
for(draw in 1:sampN) {
    pred10 = c(pred10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(pred10, prob=T ,main = 'Pred Sampling 1000')
x <- pred10
abline( v = quantile(pred10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(pred10),col="red", lwd=2)
```




```
dat10b <- na.omit(dat10[,33])
dat10b <- as.data.frame(dat10b)
n <- length(dat10b$dat10b)
sampN<-10000
pred10 <- NULL
for(draw in 1:sampN) {
    pred10 = c(pred10, mean(dat10b$dat10b[sample(n, size=n, replace=T)]))
}
hist(pred10, prob=T ,main = 'Pred Sampling 10000')
x <- pred10
abline( v = quantile(pred10,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(pred10),col="red", lwd=2)
```



c. Use a smoothing spline to do a nonparametric regression of Mobility on Middle_class. Use cross-validation to choose the degree of flexibility. Then use a resampling method to obtain 95% confidence interval. Create a plot showing the data points, the spline and the confidence interval.

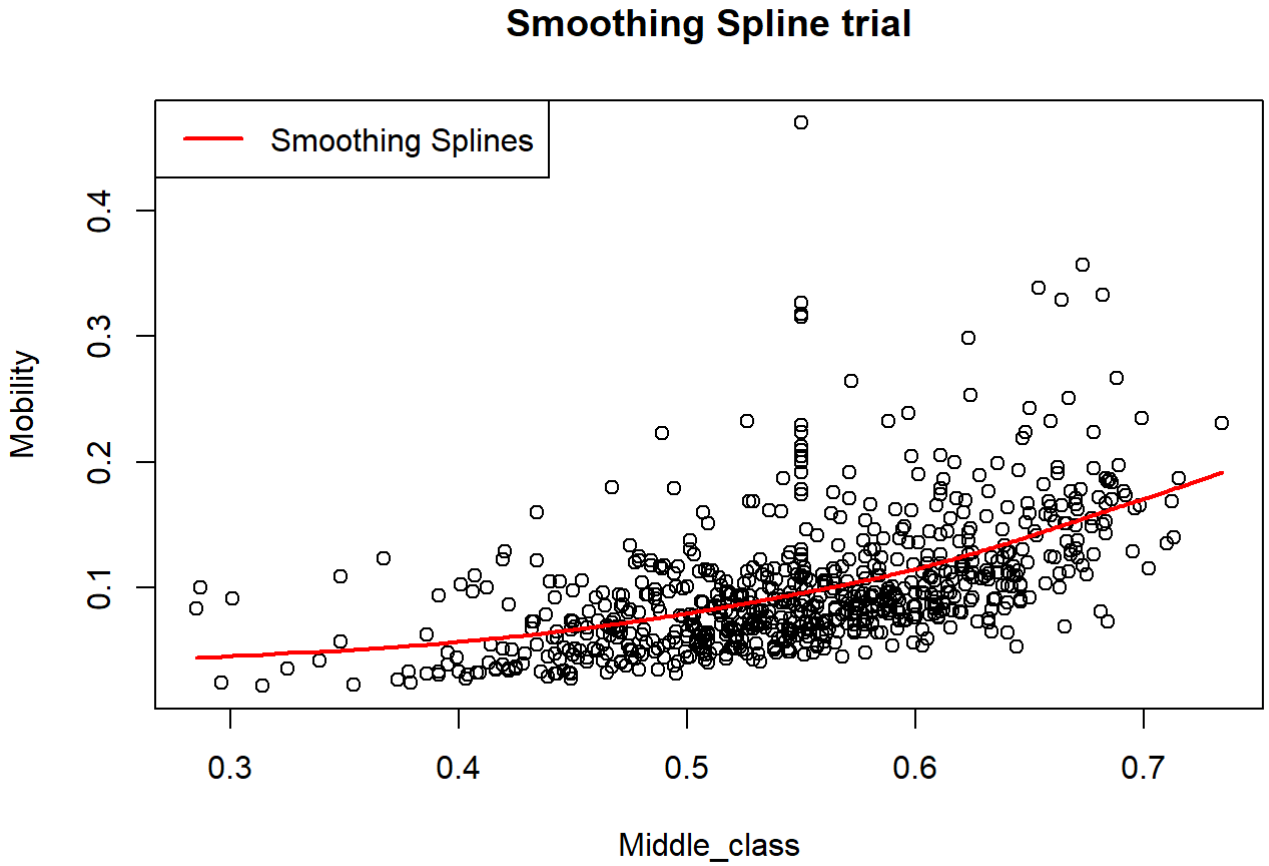
Below is the original trial with smoothing spline to do a nonparametric regression. In my original trial, i was able to see that df was setted as 3.75. (later on, i found out by boostrap that appropriate df would be around 5 ~ 6)

```
dat10c <- data.frame()
dat10ca <- dat6a[,c(1,9)]
dat10ca <- na.omit(dat10ca)
dat10c <- dat10ca %>% mutate(middle = dat10ca$Middle_class) %>% mutate(mobi = dat10ca$Mobility)
dat10c <- dat10c[,c(3,4)]
```

```
smooth10 <- smooth.spline(x=dat10c[,1],y=dat10c[,2],cv=TRUE)
smooth10$df
```

```
## [1] 3.751332
```

```
fit10c = lm(dat10c$middle~dat10c$mobi,data=dat10c)
pred10c = predict(fit10c , newdata = dat10c, se=T)
plot(dat10c$middle, dat10c$mobi, main = "Smoothing Spline trial", xlab = "Middle_class", ylab = "Mobil
ity")
lines(smooth10 ,col="red",lwd=2)
legend("topleft",("Smoothing Splines"),col="red",lwd=2)
```



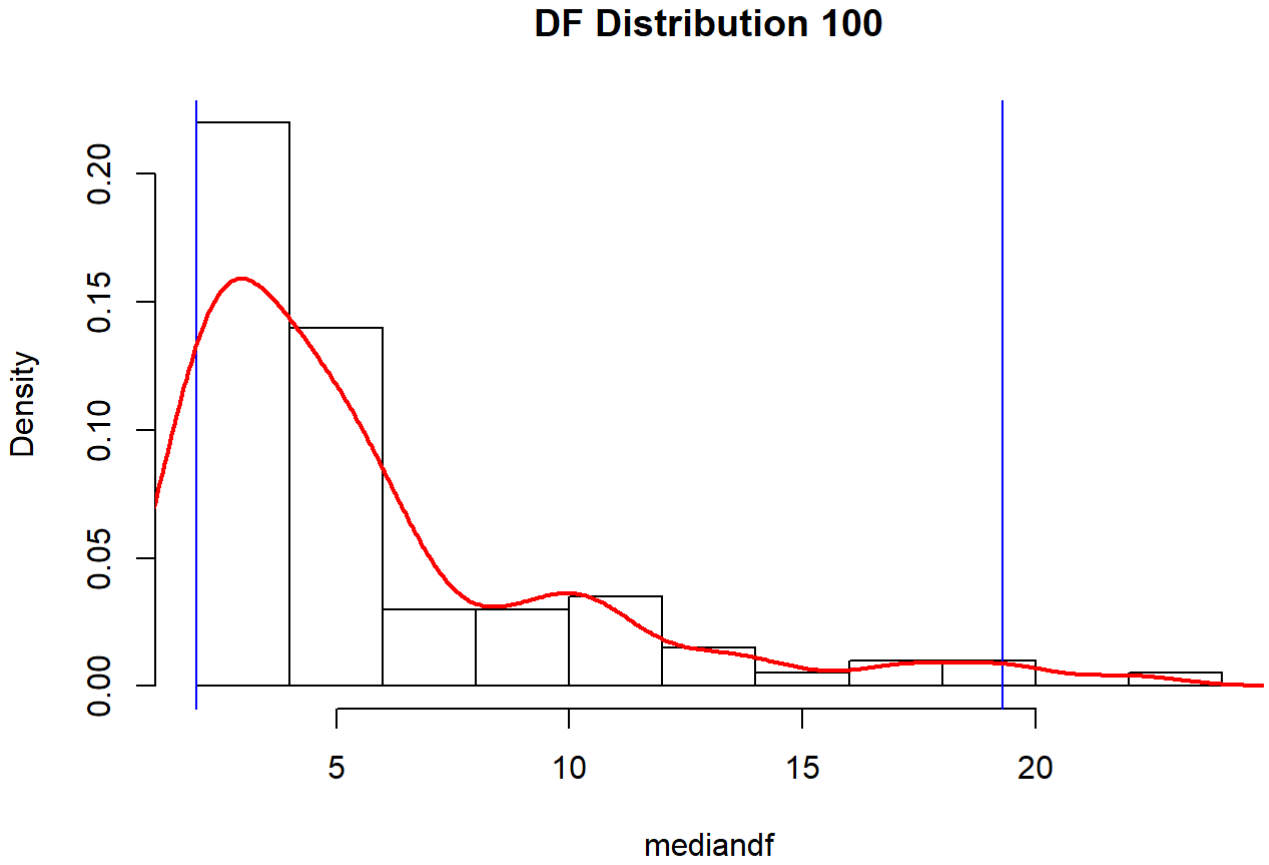
Below, i am going to use boostrap function to find out what DF is appropriate. i have sampled 1000 times the data and calculated df from each of them. By doing so, i will get 1 degree of freedom. Again i repeated getting this degree of freedom process for 100 times, thus i can get 100 degree of freedom. By doing so, i was able to darw the histrogram you can see below. Usually, as i have repeated this process, i was able to get df value around 5 ~ 6. Thus, i have decided to use 5 as my df.

```
mediandf <- list()

for (a in 1:100) {
  middleboot <- list()
  for(i in 1:1000){
    middleboot[[i]] <- dat10c %>% sample_n(size = 100) %>% summarise(meanmid = median(middle), meanmob
 = median(mobi))}
  boot10c = data.table::rbindlist(middleboot)
  model = smooth.spline(x=unlist(boot10c[,1]),y=unlist(boot10c[,2]),cv=TRUE)
  mediandf[[a]]<-model$df
}
```

```
mediandf <- unlist(mediandf)
mean(mediandf)
```

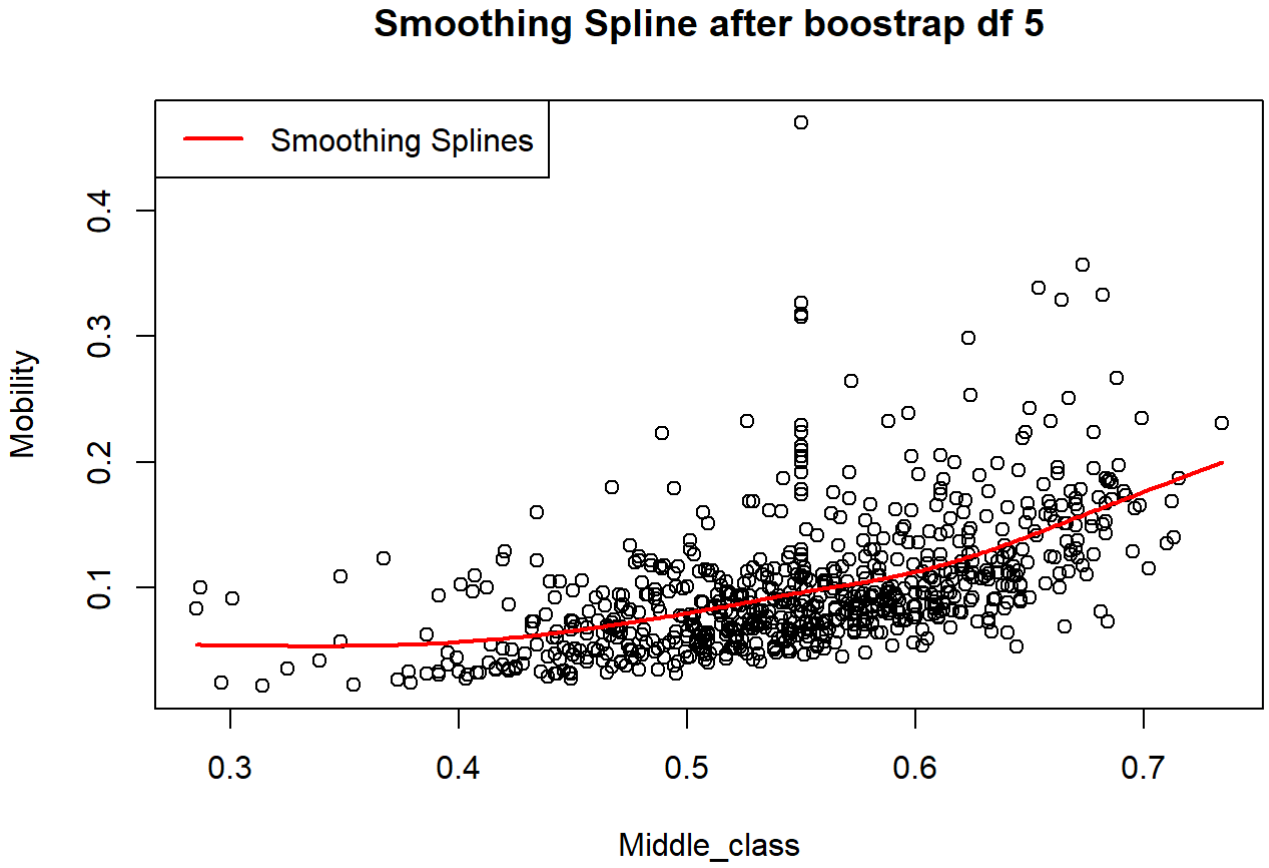
```
## [1] 5.693789
```

```
hist(mediandf, prob=T, breaks =10, main = 'DF Distribution 100')
abline( v = quantile(mediandf,probs=c(0.025,0.975), type=5), col = 'blue')
lines(density(mediandf),col="red", lwd=2)
```



This is a plot with df = 5 as i have decided by boostrap results above.

```
smooth10a <- smooth.spline(x=dat10c[,1],y=dat10c[,2],cv=TRUE, df = 5)
plot(dat10c$middle, dat10c$mobi, main = "Smoothing Spline after boostrap df 5", xlab = "Middle_class",
 ylab = "Mobility")
lines(smooth10a ,col="red",lwd=2)
legend("topleft",("Smoothing Splines"),col="red",lwd=2)
```

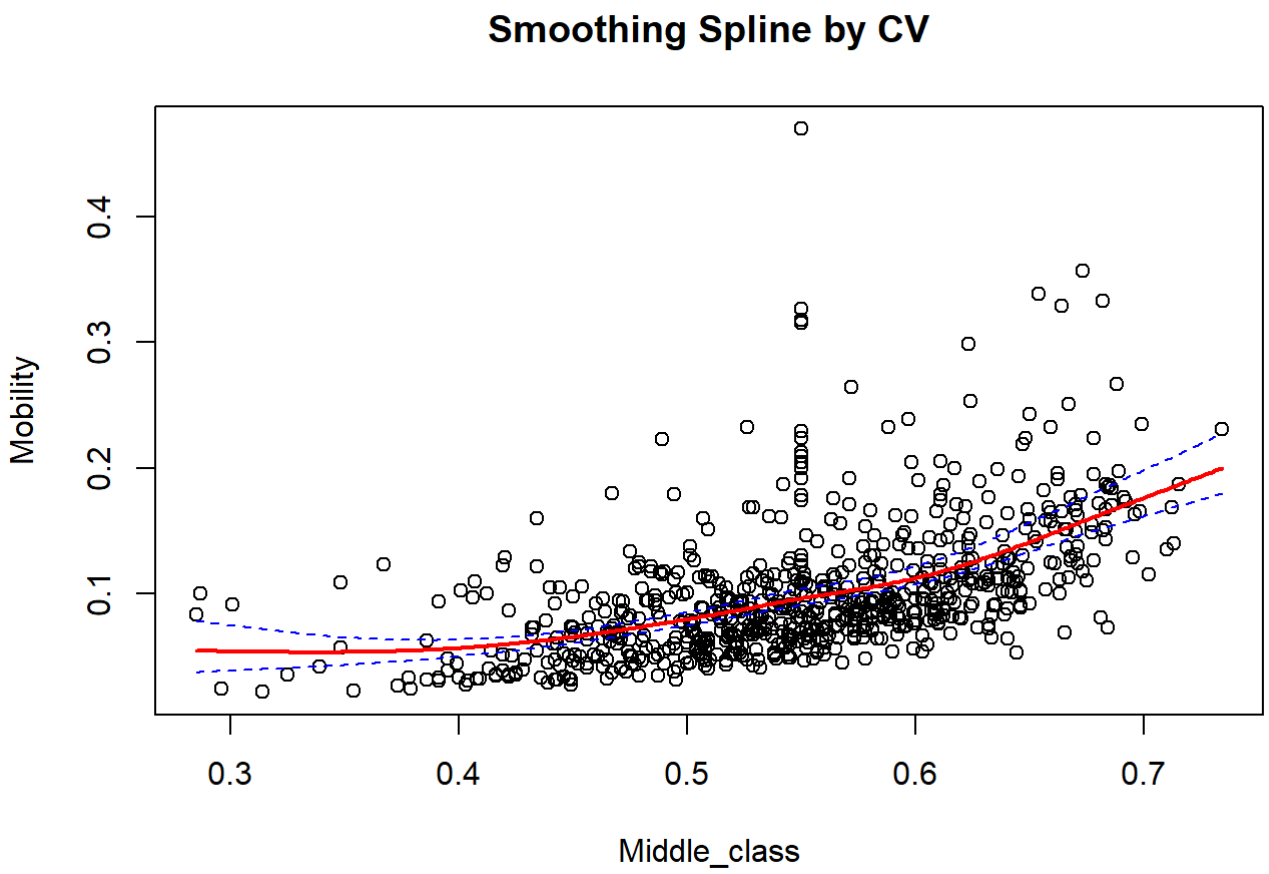


Next, I have included degree of freedom as 5 which i have found above inside the functions below. After, in order to draw a smoothing spline and its confidence interval, i had to calculate each points lower and higher value. For example, the formula for higher confidence level would be 2*spline.main - apply(spline.boots,1,quantile,probs=alpha/2).

Inference : https://stackoverflow.com/questions/23852505/how-to-get-confidence-interval-for-smooth-spline (https://stackoverflow.com/questions/23852505/how-to-get-confidence-interval-for-smooth-spline)

```
resampler <- function(data) {
  n <- nrow(data)
  resample.rows <- sample(1:n,size=n,replace=TRUE)
  return(data[resample.rows,])
}

spline.estimator <- function(data,m=300) {
  fit <- smooth.spline(x=data[,1],y=data[,2],cv=TRUE, df = 5)
  eval.grid <- seq(from=min(data[,1]),to=max(data[,1]),length.out=m)
  return(predict(fit,x=eval.grid)$y) # We only want the predicted values
}

spline.cis <- function(data,B,alpha=0.05,m=300) {
  spline.main <- spline.estimator(data,m=m)
  spline.boots <- replicate(B,spline.estimator(resampler(data),m=m))
  cis.lower <- 2*spline.main - apply(spline.boots,1,quantile,probs=1-alpha/2)
  cis.upper <- 2*spline.main - apply(spline.boots,1,quantile,probs=alpha/2)
  return(list(main.curve=spline.main,lower.ci=cis.lower,upper.ci=cis.upper,
              x=seq(from=min(data[,1]),to=max(data[,1]),length.out=m)))
}
```

```
sp.cis <- spline.cis(dat10c, B=1000,alpha=0.05)
plot(dat10c[,1],dat10c[,2], main = "Smoothing Spline by CV",
     xlab = "Middle_class", ylab = "Mobility")
lines(x=sp.cis$x,y=sp.cis$main.curve, col = "red", lwd=2)
lines(x=sp.cis$x, y=sp.cis$lower.ci, lty=2, col = "blue")
lines(x=sp.cis$x,y=sp.cis$upper.ci, lty=2, col = "blue")
```



d. Test whether smoothing is needed here.

shapiro.test is used to check whether (x) follows normality.

H0 : follows normality

H1 : does not follow normality

With shapiro.test, we can see that p-value is under 0.05. Thus i can conclude we should reject H0 and follow H1, which means linear model and residual does not follow normality. Thus we are not allowed to use linear regression in the data set. Thus we should consider using another model, which would be smoothing spline.

```
shapiro.test(fit.mid$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.mid$residuals
## W = 0.82355, p-value < 2.2e-16
```

Moreover, if we look at the Q-Q plot, we can see that bunch of the data is located outside the line. Thus, it is hard to say that this linear model and its residual follows normality.

```
plot(fit.mid, which = 2)
```