OLS Regression Model

Brian Lin

August 12, 2018

Contents

Library Used

```
library(readxl, warn.conflicts = FALSE, quietly = TRUE)
library(stringr, warn.conflicts = FALSE, quietly = TRUE)
library(dplyr, warn.conflicts = FALSE, quietly = TRUE)
library(readr, warn.conflicts = FALSE, quietly = TRUE)
library(randomForestSRC, warn.conflicts = FALSE, quietly = TRUE)
library(effects, warn.conflicts = FALSE, quietly = TRUE)
library(gridExtra, warn.conflicts = FALSE, quietly = TRUE)
library(ggRandomForests, warn.conflicts = FALSE, quietly = TRUE)
```

Guiding Question: How well could we predict the median student debt for each school using the College Accessibility variables? Is there any interactions between these variables?

Use Random ForestSRC to Find the Varialbe Interaction

```
#The variable that are in our interests are selected.
#Use RandomForestSRC to check for the potential interaction.

College <- College %>%
  select("INSTNM","COSTT4_A","MD_EARN_WNE_P10","AVGFACSAL","TUITIONFEE_IN","TUITIONFEE_OUT","PREDDEG","

for(i in 1:ncol(College)){
College <- College[College[,c(i)] != "PrivacySuppressed" & College[,c(i)] != "NULL" ,]
}

College[,2:8] <- lapply(College[,2:8],as.integer)
College$PREDDEG <- as.factor(College$PREDDEG)

summary(College)
```

```
##     INSTNM             COSTT4_A     MD_EARN_WNE_P10    AVGFACSAL
##  Length:3122        Min.   : 5886   Min.   : 15200   Min.   : 1366
##  Class :character   1st Qu.:15185   1st Qu.: 29600   1st Qu.: 5235
##  Mode  :character   Median :24254   Median : 35700   Median : 6394
##                     Mean   :26812   Mean   : 37135   Mean   : 6695
##                     3rd Qu.:34241   3rd Qu.: 42600   3rd Qu.: 7847
##                     Max.   :67572   Max.   :122600   Max.   :21354
##  TUITIONFEE_IN   TUITIONFEE_OUT  PREDDEG  MD_INC_DEBT_MDN
##  Min.   : 1036   Min.   : 1154   1: 365   Min.   : 2200
##  1st Qu.: 5504   1st Qu.:10985   2:1062   1st Qu.: 8726
##  Median :13124   Median :16788   3:1694   Median :14698
##  Mean   :15948   Mean   :19096   4:   1   Mean   :14004
##  3rd Qu.:23191   3rd Qu.:25433            3rd Qu.:18714
##  Max.   :53000   Max.   :53000            Max.   :33481
```

```
typeof(College$MD_INC_DEBT_MDN)
```

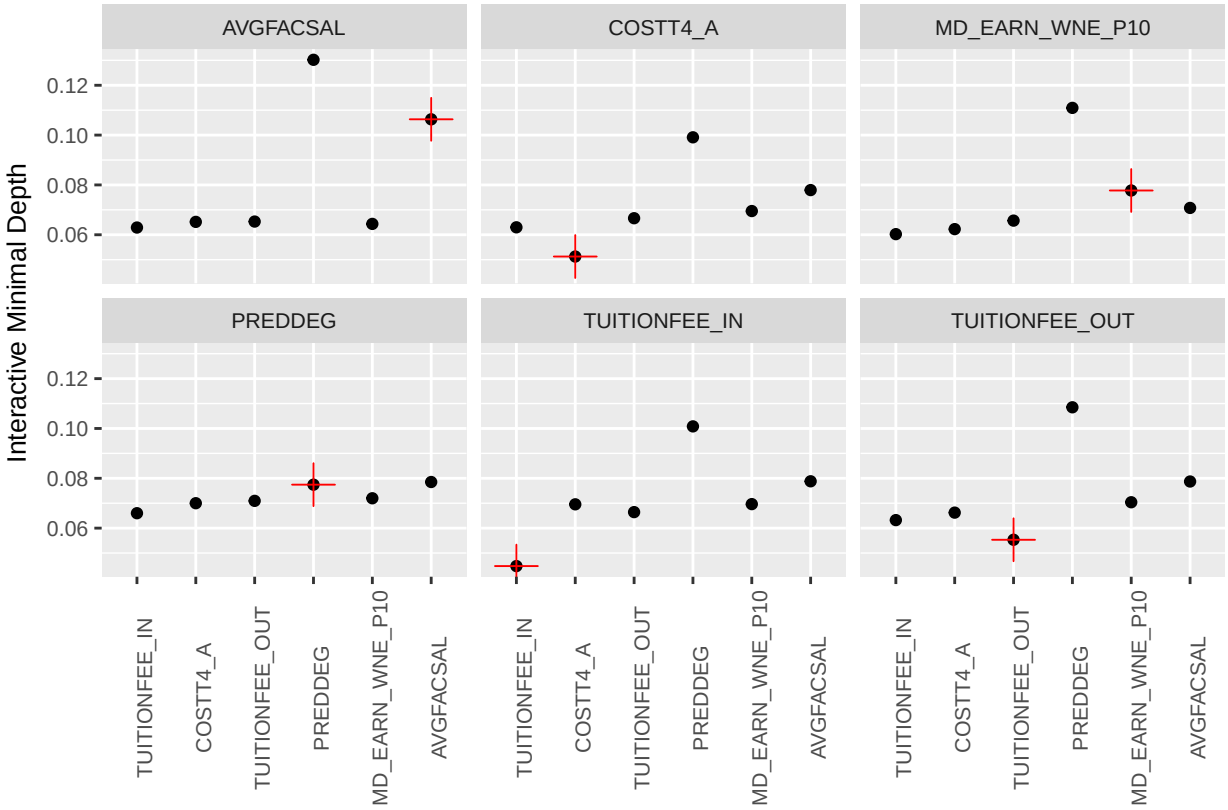
```
## [1] "integer"
```

```
Subset <- College[,-c(1)]

rfsrc_m1 <- rfsrc(MD_INC_DEBT_MDN~ ., data = as.data.frame(Subset))
```

```
max_var <- max.subtree(rfsrc_m1, conservative = TRUE)
Top <- max_var$topvars

gg_int <- gg_interaction(find.interaction(rfsrc_m1,
                                          xvar.names = Top),
                                          sorted = FALSE,
                                          verbose = FALSE)
```

```
##
##                              Method: maxsubtree
##                      No. of variables: 6
##   Variables sorted by minimal depth?: TRUE
##
##                  TUITIONFEE_IN COSTT4_A TUITIONFEE_OUT PREDDEG
## TUITIONFEE_IN             0.04     0.07           0.07    0.10
## COSTT4_A                  0.06     0.05           0.07    0.10
## TUITIONFEE_OUT            0.06     0.07           0.06    0.11
## PREDDEG                   0.07     0.07           0.07    0.08
## MD_EARN_WNE_P10           0.06     0.06           0.07    0.11
## AVGFACSAL                 0.06     0.07           0.07    0.13
##                  MD_EARN_WNE_P10 AVGFACSAL
## TUITIONFEE_IN               0.07      0.08
## COSTT4_A                    0.07      0.08
## TUITIONFEE_OUT              0.07      0.08
## PREDDEG                     0.07      0.08
## MD_EARN_WNE_P10             0.08      0.07
## AVGFACSAL                   0.06      0.11
```

```
plot(gg_int)
```



The lower the interactive minimal depth suggests interactions between variables.

Interactions between variable were indentified.

Model Selecting

We use OLS regression to predict median student debt for each institution.

Two model were considered:

1: OLS regression without interaction terms 2: OLS regression with interaction terms

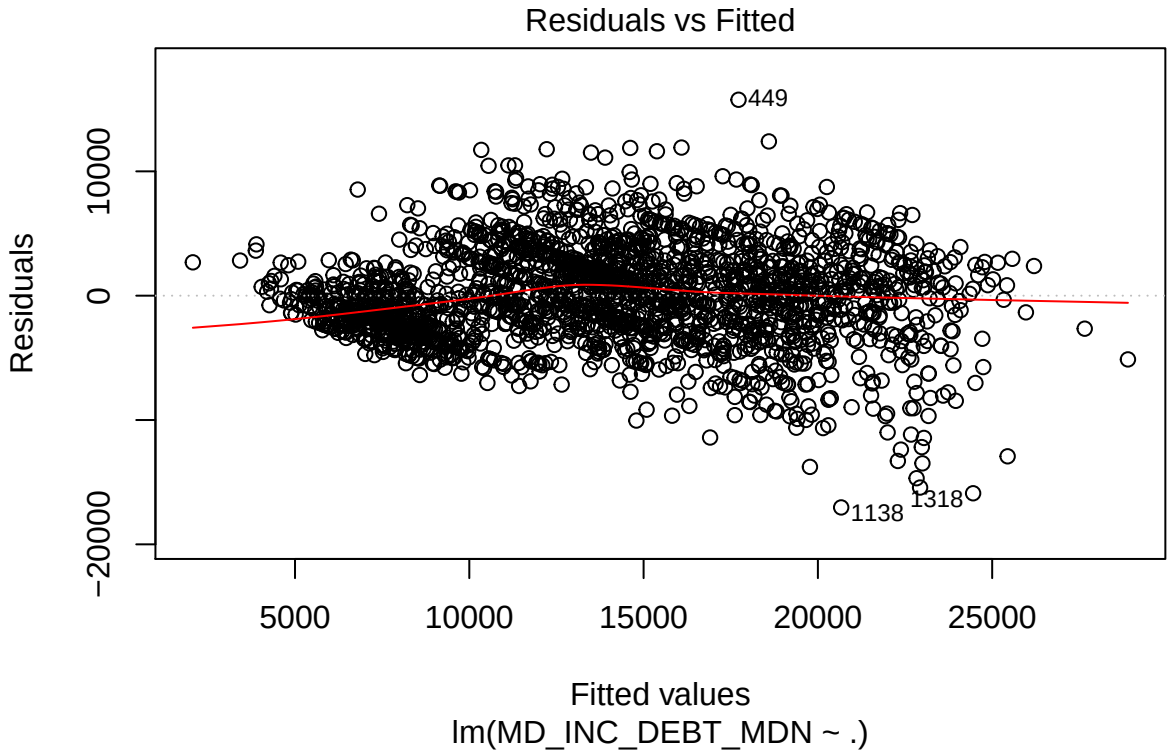
```
set.seed(3122)

Index <- sample(1:3122, 2185, replace = FALSE)

Train <- Subset[Index,]
Test <-  Subset[-Index,]

lm_m1 <- lm(MD_INC_DEBT_MDN~ . , data = Train)
summary(lm_m1)
```

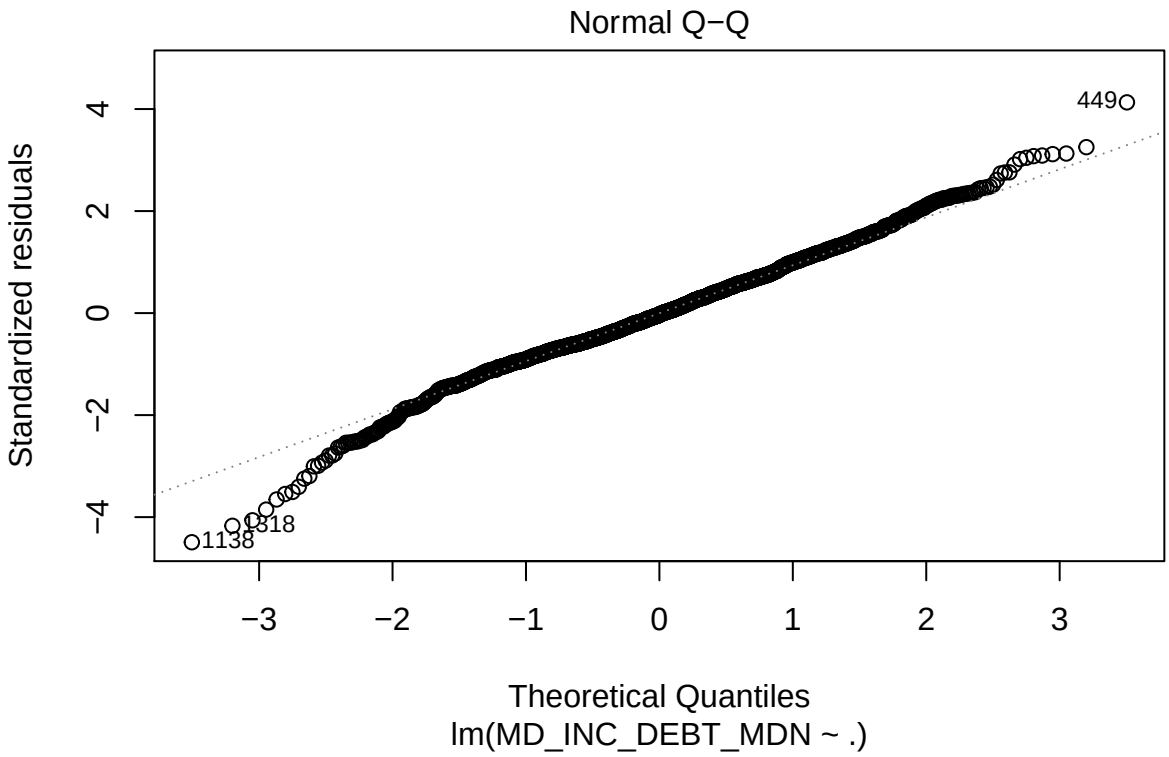
```
##
## Call:
## lm(formula = MD_INC_DEBT_MDN ~ ., data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -17030.8  -2428.1   -113.9   2411.3  15757.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.953e+03  4.515e+02  10.971  < 2e-16 ***
## COSTT4_A         1.854e-01  3.015e-02   6.151 9.15e-10 ***
## MD_EARN_WNE_P10  1.232e-01  1.184e-02  10.411  < 2e-16 ***
## AVGFACSAL       -8.294e-01  5.736e-02 -14.460  < 2e-16 ***
## TUITIONFEE_IN   -7.994e-02  3.595e-02  -2.224   0.0263 *
## TUITIONFEE_OUT   1.674e-01  2.520e-02   6.641 3.92e-11 ***
## PREDDEG2         1.254e+03  2.828e+02   4.433 9.76e-06 ***
## PREDDEG3         4.765e+03  3.072e+02  15.510  < 2e-16 ***
## PREDDEG4        -3.239e+03  3.945e+03  -0.821   0.4118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3820 on 2176 degrees of freedom
## Multiple R-squared:  0.6269, Adjusted R-squared:  0.6256
## F-statistic: 457.1 on 8 and 2176 DF,  p-value: < 2.2e-16
```

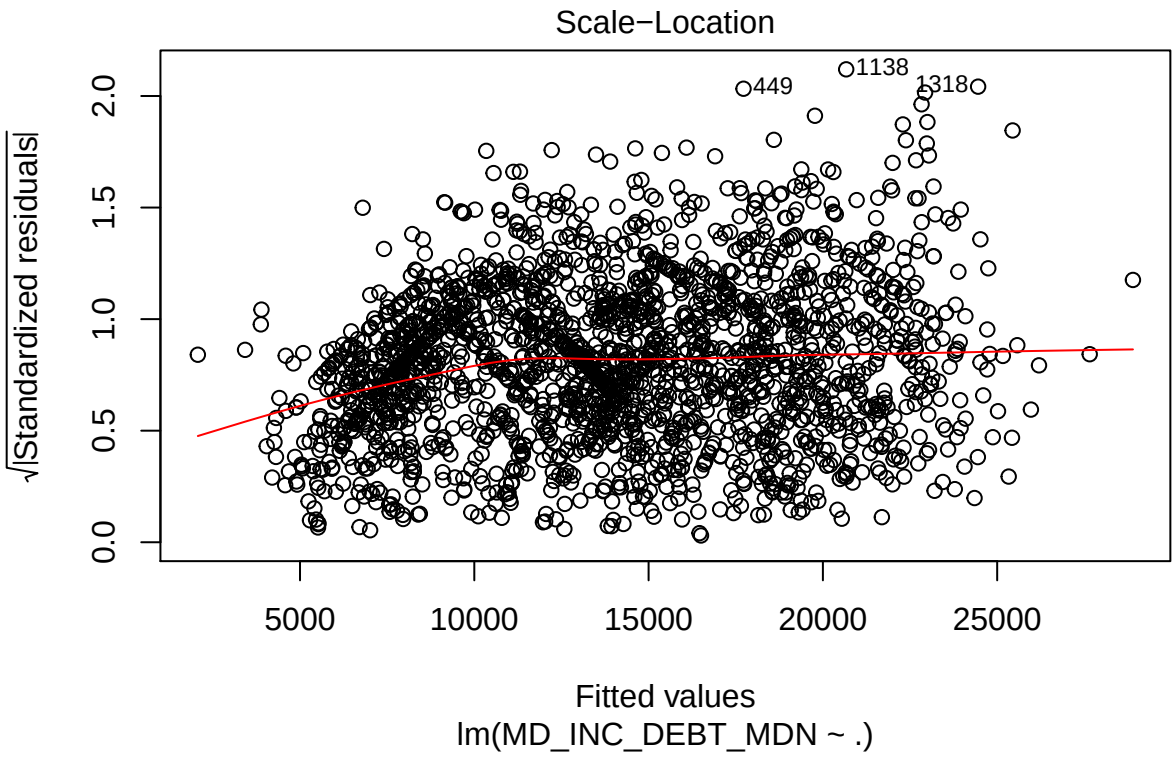
```
plot(lm_m1)
```

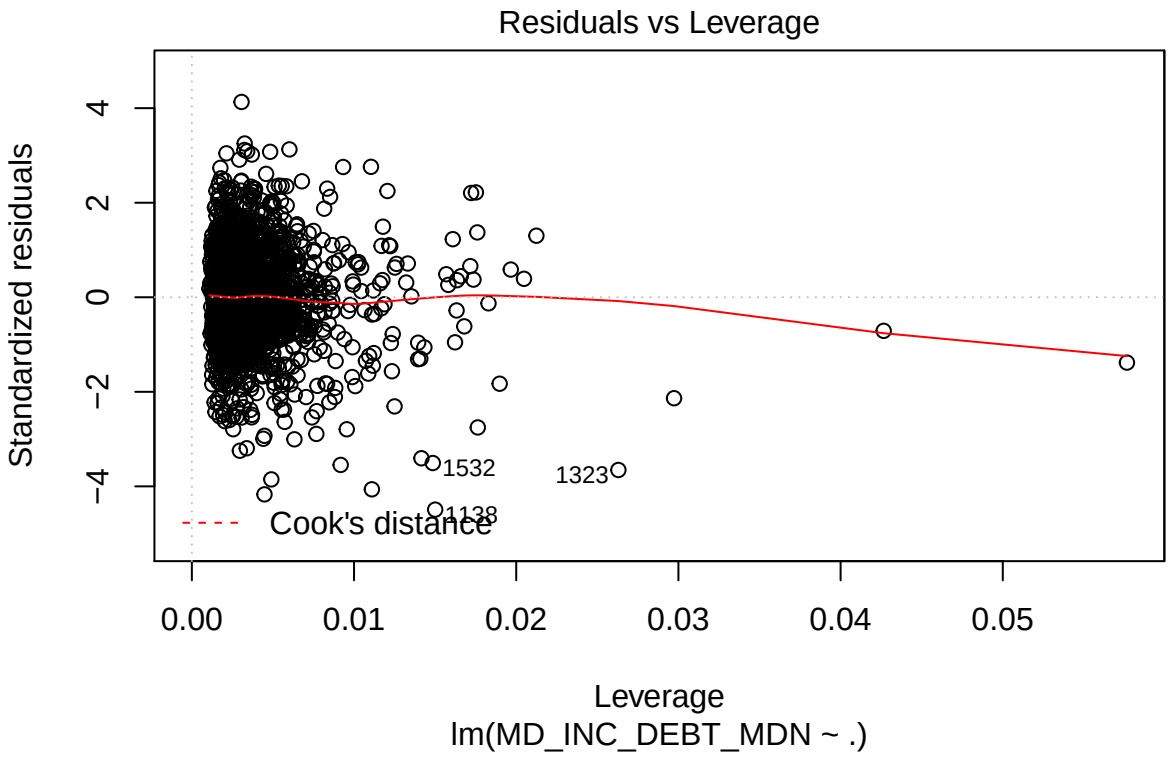
```
## Warning: not plotting observations with leverage one:
##   6
```






```
## Warning: not plotting observations with leverage one:
##   6
```


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
449
1138
1318
5000
10000
15000
20000
25000
Fitted values
lm(MD_INC_DEBT_MDN ~ .)




```
predict <- as.vector(predict(lm_m1,Test))
MSE1 <- mean((Test$MD_INC_DEBT_MDN - predict)^2)


lm_m2 <- lm(MD_INC_DEBT_MDN~ . + TUITIONFEE_IN*TUITIONFEE_OUT*COSTT4_A*MD_EARN_WNE_P10 , data = Train)
summary(lm_m2)
```

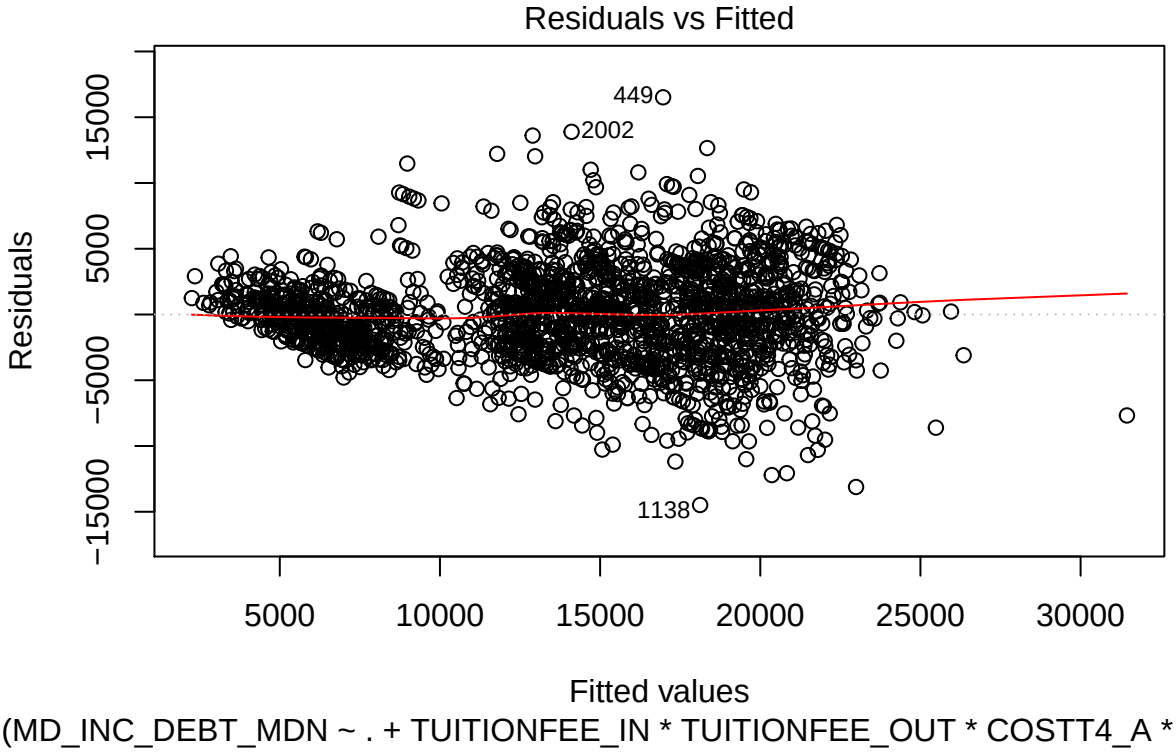
```
##
## Call:
## lm(formula = MD_INC_DEBT_MDN ~ . + TUITIONFEE_IN * TUITIONFEE_OUT *
##     COSTT4_A * MD_EARN_WNE_P10, data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14484.4  -2048.4     49.7   1896.6  16514.3
##
## Coefficients:
##                                                      Estimate
## (Intercept)                                        -3.574e+03
## COSTT4_A                                            1.781e-01
## MD_EARN_WNE_P10                                     1.442e-01
## AVGFACSAL                                          -2.077e-01
## TUITIONFEE_IN                                       2.615e+00
## TUITIONFEE_OUT                                     -9.389e-01
## PREDDEG2                                            8.363e+02
## PREDDEG3                                            2.828e+03
```

```
## PREDDEG4                                                -7.534e+03
## TUITIONFEE_IN:TUITIONFEE_OUT                            -5.671e-05
## COSTT4_A:TUITIONFEE_IN                                  -8.647e-05
## COSTT4_A:TUITIONFEE_OUT                                  6.886e-05
## MD_EARN_WNE_P10:TUITIONFEE_IN                           -4.883e-05
## MD_EARN_WNE_P10:TUITIONFEE_OUT                           2.277e-05
## COSTT4_A:MD_EARN_WNE_P10                                 5.783e-06
## COSTT4_A:TUITIONFEE_IN:TUITIONFEE_OUT                    8.141e-10
## MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT             1.479e-09
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN                   1.422e-09
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_OUT                 -1.576e-09
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT -1.554e-14
##                                                         Std. Error t value
## (Intercept)                                              2.689e+03  -1.329
## COSTT4_A                                                 1.983e-01   0.898
## MD_EARN_WNE_P10                                          8.124e-02   1.775
## AVGFACSAL                                                6.041e-02  -3.438
## TUITIONFEE_IN                                            3.550e-01   7.366
## TUITIONFEE_OUT                                           2.830e-01  -3.317
## PREDDEG2                                                 2.577e+02   3.245
## PREDDEG3                                                 3.035e+02   9.318
## PREDDEG4                                                 3.665e+03  -2.056
## TUITIONFEE_IN:TUITIONFEE_OUT                             1.318e-05  -4.302
## COSTT4_A:TUITIONFEE_IN                                   1.180e-05  -7.327
## COSTT4_A:TUITIONFEE_OUT                                  1.456e-05   4.731
## MD_EARN_WNE_P10:TUITIONFEE_IN                            9.244e-06  -5.282
## MD_EARN_WNE_P10:TUITIONFEE_OUT                           7.023e-06   3.242
## COSTT4_A:MD_EARN_WNE_P10                                 5.466e-06   1.058
## COSTT4_A:TUITIONFEE_IN:TUITIONFEE_OUT                    1.300e-10   6.262
## MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT             3.153e-10   4.691
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN                   2.719e-10   5.230
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_OUT                  3.247e-10  -4.855
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT    3.251e-15  -4.781
##                                                         Pr(>|t|)
## (Intercept)                                             0.183984
## COSTT4_A                                                0.369097
## MD_EARN_WNE_P10                                         0.076087 .
## AVGFACSAL                                               0.000597 ***
## TUITIONFEE_IN                                           2.48e-13 ***
## TUITIONFEE_OUT                                          0.000924 ***
## PREDDEG2                                                0.001193 **
## PREDDEG3                                                 < 2e-16 ***
## PREDDEG4                                                0.039903 *
## TUITIONFEE_IN:TUITIONFEE_OUT                            1.77e-05 ***
## COSTT4_A:TUITIONFEE_IN                                  3.31e-13 ***
## COSTT4_A:TUITIONFEE_OUT                                 2.38e-06 ***
## MD_EARN_WNE_P10:TUITIONFEE_IN                           1.40e-07 ***
## MD_EARN_WNE_P10:TUITIONFEE_OUT                          0.001203 **
## COSTT4_A:MD_EARN_WNE_P10                                0.290164
## COSTT4_A:TUITIONFEE_IN:TUITIONFEE_OUT                   4.56e-10 ***
## MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT            2.89e-06 ***
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN                  1.85e-07 ***
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_OUT                 1.29e-06 ***
## COSTT4_A:MD_EARN_WNE_P10:TUITIONFEE_IN:TUITIONFEE_OUT 1.86e-06 ***
```

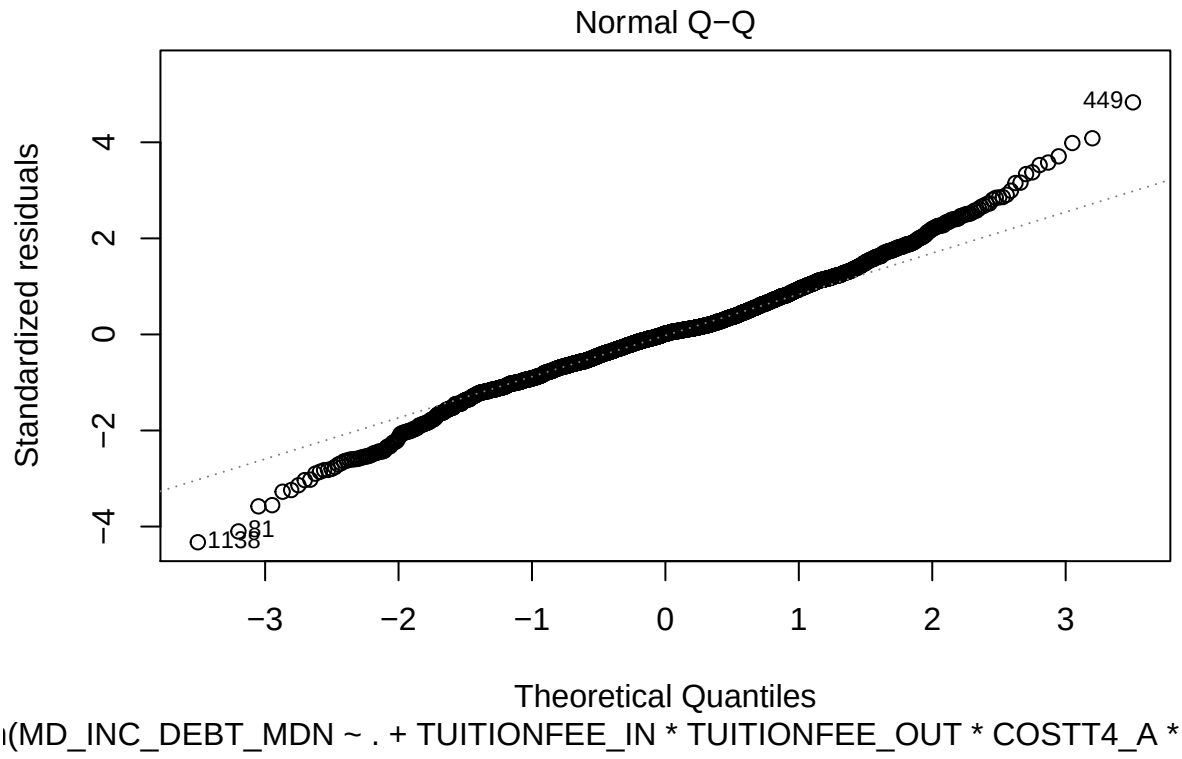
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3424 on 2165 degrees of freedom
## Multiple R-squared:  0.7018, Adjusted R-squared:  0.6991
## F-statistic: 268.1 on 19 and 2165 DF,  p-value: < 2.2e-16
```

```
plot(lm_m2)
```

```
## Warning: not plotting observations with leverage one:
##   6
```

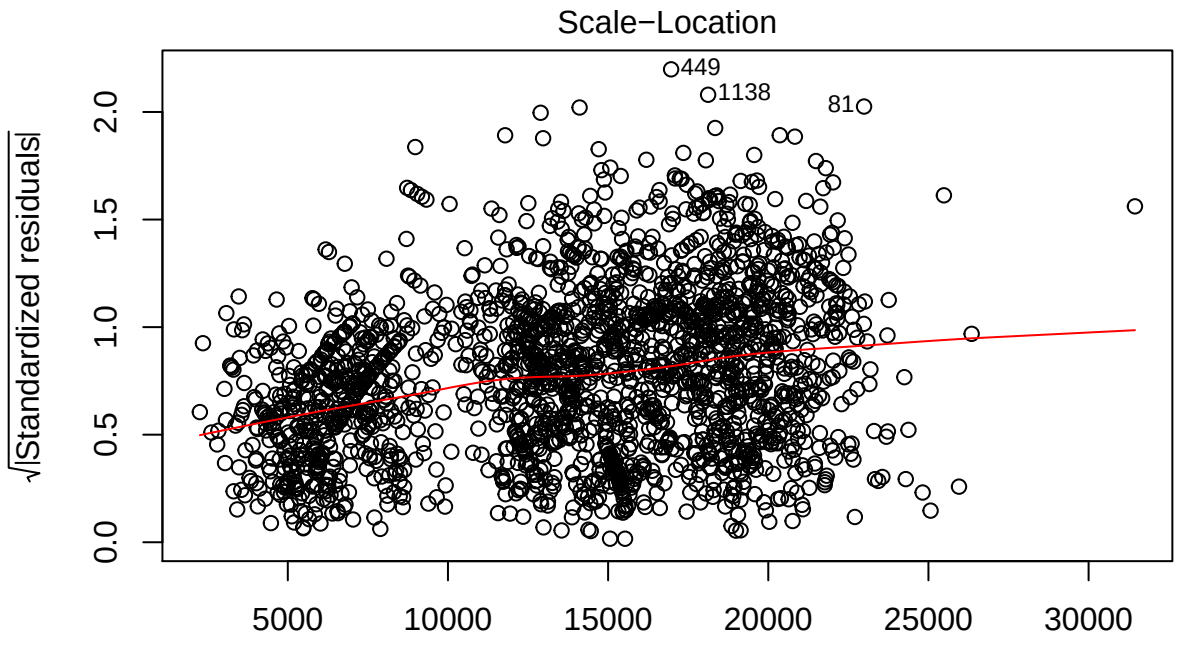


(MD_INC_DEBT_MDN ~ . + TUITIONFEE_IN * TUITIONFEE_OUT * COSTT4_A * MD_

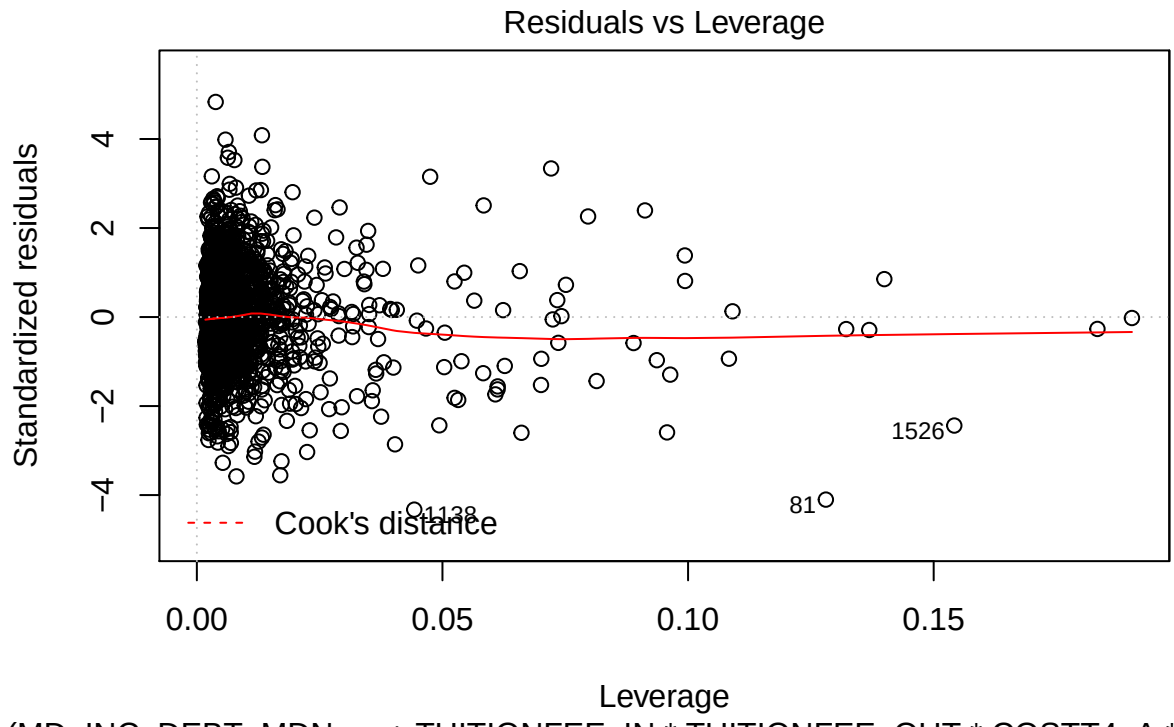


ı(MD_INC_DEBT_MDN ~ . + TUITIONFEE_IN * TUITIONFEE_OUT * COSTT4_A * MD_

```
## Warning: not plotting observations with leverage one:
##   6
```



ı(MD_INC_DEBT_MDN ~ . + TUITIONFEE_IN * TUITIONFEE_OUT * COSTT4_A * MD_



l(MD_INC_DEBT_MDN ~ . + TUITIONFEE_IN * TUITIONFEE_OUT * COSTT4_A * MD_

```
predict <- as.vector(predict(lm_m2,Test))
MSE2 <- mean((Test$MD_INC_DEBT_MDN - predict)^2)
```

Result

The result consist with the RandomForestSRC model result.

The OLS regression model with interactions has lower testing MSE.

Hence, we were able to demotraste that the interaction exist between the College Accessibility variables.

It has a MSE of 10919424 for predicting the median student debt for each school.