PROJECT UTS: GAM

Anak Agung Ayu Diva Shanty Darmawan – 01112190018

Untuk project UTS kali ini, akan dilakukan simulasi Generalized Additive Model (GAM) menggunakan salesData2 yaitu data penjualan rumah yang telah diberikan pada awal semester genap.

 # Data Cleanup

Data cleanup disini berisi perintah yang akan menyatakan ulang data yang digunakan dan memeriksa serta menghapus nilai yang kosong atau missing values.

```
rm(list = ls())
library(readr)
library(ggplot2)
data = read.csv('salesData2.csv')
set.seed(1)
summary(data)
```

```
       ID             City          Province
 Min.   :    1   Min.   :20.00   Min.   :1.000
 1st Qu.: 7486   1st Qu.:54.00   1st Qu.:2.000
 Median :14994   Median :54.00   Median :2.000
 Mean   :14999   Mean   :56.02   Mean   :2.022
 3rd Qu.:22502   3rd Qu.:65.00   3rd Qu.:2.000
 Max.   :30014   Max.   :98.00   Max.   :3.000
    SALEDT              Price             SQFT
 Length:29209       Min.   :  1200   Min.   :1.000e+00
 Class :character   1st Qu.:133000   1st Qu.:6.159e+03
 Mode  :character   Median :202000   Median :1.184e+04
                    Mean   :221522   Mean   :2.353e+05
                    3rd Qu.:287900   3rd Qu.:4.126e+04
                    Max.   :760000   Max.   :2.123e+09
     RMTOT            RMBED            BATH
 Min.   : 1.000   Min.   : 1.000   Min.   :1.000
 1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.:1.000
 Median : 3.000   Median : 3.000   Median :1.000
 Mean   : 3.227   Mean   : 3.186   Mean   :1.619
 3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.:2.000
 Max.   :30.000   Max.   :12.000   Max.   :7.000
    FLR1AREA         BSMT          HEATSYS
 Min.   : 192   Min.   :1.000   Min.   :1.000
 1st Qu.: 748   1st Qu.:6.000   1st Qu.:1.000
 Median :1001   Median :7.000   Median :2.000
 Mean   :1067   Mean   :5.958   Mean   :2.443
 3rd Qu.:1268   3rd Qu.:7.000   3rd Qu.:3.000
 Max.   :5366   Max.   :7.000   Max.   :6.000
     ATTIC             SFLA         GRADE
 Min.   :0.0000   Min.   : 192   Length:29209
 1st Qu.:1.0000   1st Qu.:1200   Class :character
 Median :1.0000   Median :1604   Mode  :character
```

```
Mean   :0.8829   Mean   :1686
3rd Qu.:1.0000   3rd Qu.:2048
Max.   :1.0000   Max.   :8580
  SALE_YEAR        SALE_MONTH        EFF_AGE
Min.   :-5.000   Min.   : 1.000   Min.   :  0.00
1st Qu.:-4.000   1st Qu.: 5.000   1st Qu.: 16.00
Median :-2.000   Median : 7.000   Median : 29.00
Mean   :-2.757   Mean   : 7.018   Mean   : 52.23
3rd Qu.:-1.000   3rd Qu.: 9.000   3rd Qu.: 40.00
Max.   :-1.000   Max.   :12.000   Max.   :468.00
   Style
Length:29209
Class :character
Mode  :character
```

Sebelum menyatakan kembali data, kita gunakan perintah `rm(list = ls())` untuk menghapus environment data dari data-data projek sebelumnya. Data ini disimpan dalam format CSV sehingga perlu diekstrak menggunakan fungsi read.csv. Di sini, kita gunakan data sebagai nama variabel `data` untuk memanggil data kita selanjutnya. `Set.seed(1)` digunakan dalam projek ini untuk mendapatkan nilai yang sama setiap kali di run.

 # Data Exploration

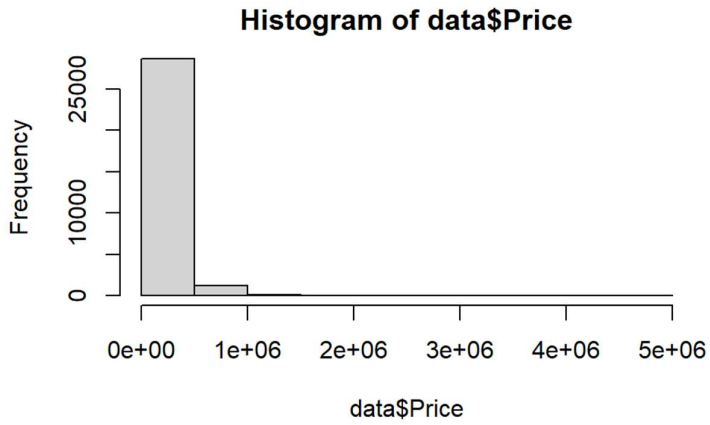
Melalui Data Exploration, kita dapat mengetahui variabel-variabel mana saja yang akan digunakan untuk mensimulasikan GAM.

Dependent Variable

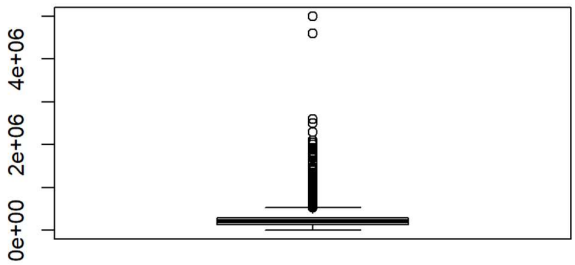
Dependent Variable adalah variable yang akan menjadi variable yang kita lihat untuk mengukur apakah berhasil modelnya atau tidak. Anggaplah dependent variable seperti efeknya dari model yang ingin kita simulasikan. Variable yang kita gunakan sebagai dependent variable adalah variable Price.

```
summary(data$Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1200  133000  204500  228956  290000 5000000
```

```
hist(data$Price) #datanya skewed ke kiri
```

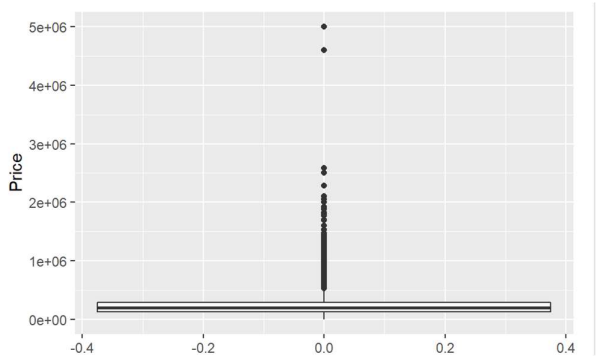



```
boxplot(data$Price)
```



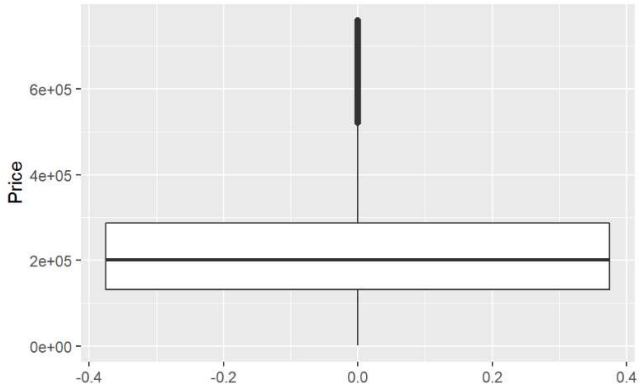
Dari histogram dan boxplot dapat dilihat bahwa datanya strongly skewed, dengan banyak data berada di bagian kiri apabila dilihat dari histogram.

```
ggplot(data) + geom_boxplot(aes(y = Price))
```



Dari histogram di atas, dapat dilihat bahwa terdapat outliers di dependent variable. Hal ini dapat menjadi alasan mengapa kenapa data distribusinya strongly skewed. Karena outliers ini memiliki kemungkinan mempengaruhi model yang akan dibuat, maka harus dihapus terlebih dahulu melalui fungsi dibawah ini.

```
boxplot.stats(data$Price, coef = 3)$out
out = boxplot.stats(data$Price, coef = 3)$out
out_key = which(data$Price %in% c(out))
data = data[-out_key, ]
ggplot(data) + geom_boxplot(aes(y = Price))
```



Bisa dilihat histogram diatas bahwa data menjadi terlihat lebih baik dan lebih tidak strongly skewed. Setelah menganalisa dependent variable, kita dapat melanjutkan proses dengan menganalisa independent variable.

Independent Variable

Apabila dependent variable adalah efek, maka independent variable adalah penyebab dari efek tersebut. Independent variable digunakan sebagai variable yang akan mempengaruhi bagaimana hasil model yang akan dihasilkan. Independent variable yang digunakan pada kali ini adalah: City, Province, SQFT, RMTOT, RMBED, BATH, FLR1AREA, BSMT, ATTIC, SFLA, GRADE, EFF_AGE, Style.

```
table(data$GRADE)
    A     B     C     D     E 
   45  2499 23213  3114   866 

table(data$City)
   20    21    23    24    25    27    28    30    31 
  717    27   204    58    71   362   129    86    11 
   32    33    43    44    45    47    48    49    54 
   56  2878   865   372    51   550   254    28 13857 
   59    60    62    64    65    66    67    69    70 
  743   479   124   304    97   196   195  1639   105 
   71    72    74    75    76    77    78    80    81 
  278   183   400   831   331   127    48   539   385 
   82    83    84    85    87    88    89    91    92 
  104   154    59   121   190    45   326    22   122 
   93    95    96    97    98 
   77   171   191   339   236 

table(data$Province)
    1     2     3 
 5001 19060  5676 

 table(data$BSMT)
    1     2     3     4     5     6     7 
 1598  2558   909   219  1905   741 21807 

table(data$ATTIC)
    0     1 
 3603 26134 

table(data$Style)
    B     H     R     S     U 
 3623  1920   654  1081 22459 

summary(data$SQFT)
     Min.   1st Qu.    Median      Mean   3rd Qu. 
0.000e+00 6.005e+03 1.140e+04 2.311e+05 4.030e+04 
     Max. 
2.123e+09 
summary(data$RMTOT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   3.000   3.226   4.000  30.000 
```

```
summary(data$RMBED)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.184   4.000  12.000

summary(data$BATH)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.621   2.000   7.000

summary(data$FLR1AREA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    192     748    1002    1069    1270    5366

summary(data$SFLA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    192    1200    1604    1687    2050    8580

summary(data$EFF_AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   16.00   29.00   52.07   40.00  468.00

data = data[data$SQFT != 0,]
```

Perintah di atas berfungsi untuk menghapus data-data yang memiliki nilai 0 dalam variable SQFT.

Selanjutnya, akan kita lihat korelasi diantara dependent variable dengan masing-masing independent variable.

```
cor(data$Price, data$SQFT)
      -0.01093608
```

Dapat dilihat bahwa, hubungan antara `Price` dan `SQFT` menghasilkan nilai yang negatif yang artinya variable `SQFT` tidak terlalu signifikan bagi variable `Price`.

```
cor(data$Price, data$FLR1AREA)
      0.4345601
```

Dapat dilihat bahwa, hubungan antara `Price` dan `FLR1AREA` menghasilkan nilai yang cukup besar yang artinya variable `FLR1AREA` memiliki korelasi yang moderate bagi variable `Price`.

```
cor(data$Price, data$SFLA)
      0.7104939
```

Dapat dilihat bahwa, hubungan antara `Price` dan `SFLA` menghasilkan nilai yang besar yang artinya variable `SFLA` memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$RMTOT)
      0.2652995
```

Dapat dilihat bahwa, hubungan antara `Price` dan `RMTOT` menghasilkan nilai kecil yang artinya variable `RMTOT` memiliki korelasi yang lemah bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$RMBED)
      0.280041
```

Dapat dilihat bahwa, hubungan antara `Price` dan `RMBED` menghasilkan nilai kecil yang artinya variable `RMBED` memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$EFF_AGE)
      -0.3595609
```

Dapat dilihat bahwa, hubungan antara `Price` dan `EFF_AGE` menghasilkan nilai yang negatif yang artinya variable `EFF_AGE` tidak memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$City)
      -0.02829117
```

Dapat dilihat bahwa, hubungan antara `Price` dan `City` menghasilkan nilai yang negatif yang artinya variable `City` tidak memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$Province)
      -0.07577558
```

Dapat dilihat bahwa, hubungan antara `Price` dan `Province` menghasilkan nilai yang negatif yang artinya variable `Province` tidak memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$BATH)
      0.5961186
```

Dapat dilihat bahwa, hubungan antara `Price` dan `BATH` menghasilkan nilai yang cukup besar yang artinya variable `BATH` memiliki korelasi moderate bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$BSMT)
      0.2269471
```

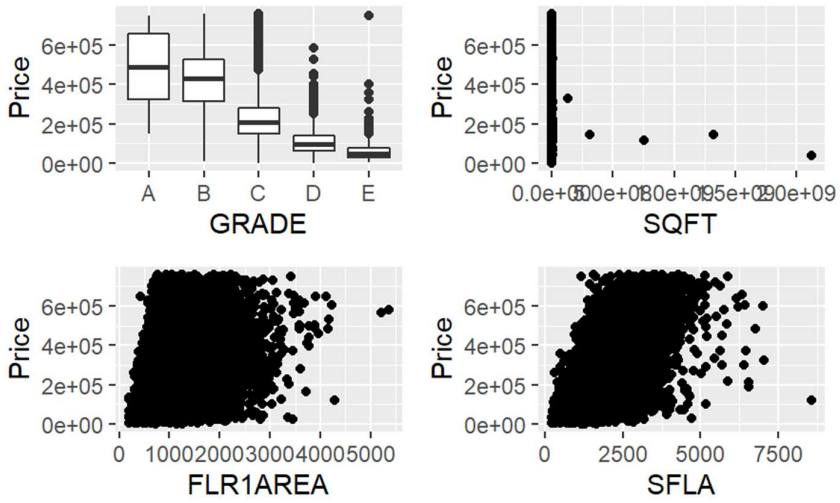
Dapat dilihat bahwa, hubungan antara `Price` dan `BSMT` menghasilkan nilai kecil yang artinya variable `BSMT` memiliki korelasi yang lemah bagi variable `Price` di model yang akan dibuat.

```
cor(data$Price, data$ATTIC)
      -0.2861129
```

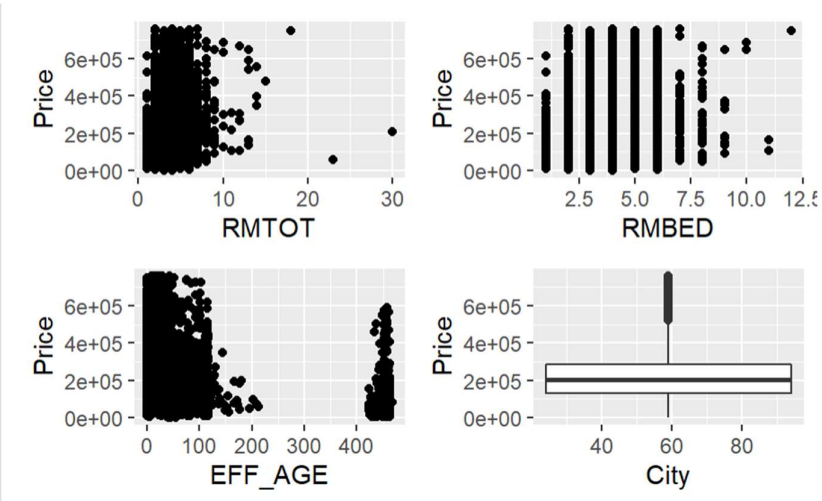
Dapat dilihat bahwa, hubungan antara `Price` dan `ATTIC` menghasilkan nilai yang negatif yang artinya variable `ATTIC` tidak memiliki korelasi yang signikan bagi variable `Price` di model yang akan dibuat.

Sekarang, akan dibandingkan korelasi yang telah didapatkan di atas dengan boxplot dari masing-masing independent variable.

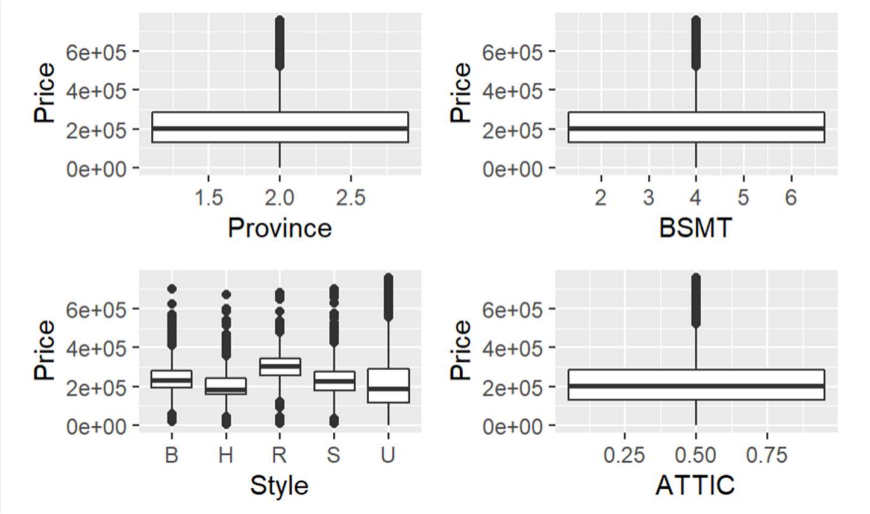
```
library(ggpubr)
ggarrange(ggplot(data) + geom_boxplot(aes(GRADE, Price)),
          ggplot(data) + geom_point(aes(SQFT, Price)),
          ggplot(data) + geom_point(aes(FLR1AREA, Price)),
          ggplot(data) + geom_point(aes(SFLA, Price)))
```




```
ggarrange(ggplot(data) + geom_point(aes(RMTOT, Price)),
          ggplot(data) + geom_point(aes(RMBED, Price)),
          ggplot(data) + geom_point(aes(EFF_AGE, Price),
          ggplot(data) + geom_boxplot(aes(City, Price)))
```



```
ggarrange(ggplot(data) + geom_boxplot(aes(Province, Price)),
          ggplot(data) + geom_boxplot(aes(BSMT, Price)),
          ggplot(data) + geom_boxplot(aes(Style, Price)),
          ggplot(data) + geom_boxplot(aes(ATTIC, Price)))
```



Dari plot-plot yang sudah dibuat, kita mendapatkan kesimpulan yang sama seperti pada correlation coefficients test pada hubungan dependent variable dan masing-masing independent variable. Seperti dapat dilihat pada plot `SFLA`, terlihat dapat hubungan yang signifikan antara variable `SFLA` dengan variable `Price`. Juga dapat dilihat dari plot `SQFT`, terlihat bahwa variable `SQFT` tidak memiliki hubungan yang signifikan dengan `Price`. Ditambah lagi, kita dapat melihat harga rata-rata rumah menurun dengan variable `GRADE` dan harga rata-rata tidak beraturan dengan variable `Style`.

 # Splitting Data

Pada section ini, kita akan membagi data ke dalam train set dan test set. Pembagian data ini dilakukan dengan menggunakan fungsi `CreateDataPartition`. Pembagian data ini akan dilakukan berdasarkan variable `GRADE` sebagai salah satu independent data yang categorical di model GAM. Data akan dibagi mengikuti 80:20 ratio, dengan 80% train set dan 20% test set.

```
library(lattice)
library(caret)
datap = createDataPartition(data$GRADE, p = 0.8, list = FALSE)

train.sd = data[datap,]
test.sd = data[-datap,]

rbind("Train Set" = table(train.sd$GRADE),
      "Test Set"= table(test.sd$GRADE))
                A    B     C    D   E
      Train Set 36 1960 18245 2451 677
      Test Set   9  489  4561  612 169
```

Dengan melakukan pengecekan pada fungsi `rbind`, dapat kita lihat bahwa di dalam train set dan test set terdapat variable `GRADE` dengan proporsi yang telah ditetapkan.

```
rbind("Train Set" = table(train.sd$BSMT),
      "Test Set"= table(test.sd$BSMT))
                  1    2   3   4    5   6     7
      Train Set 1251 1973 705 167 1479 588 17206
      Test Set   300  523 177  49  392 148  4251
```

Selain itu juga terdapat proporsi yang pas diantara train set dan test set untuk variable BSMT sehingga kita dapat melanjutkan pembuatan model GAM dengan data yang telah dibagi ini.

 # Creating GAM Model

Generalized Additivie Model (GAM) adalah model linier umum dimana variable response linier bergantung secara linier pada fungsi mulus yang tidak diketahui dari beberapa variable prediktor.

```
library(foreach)
library(splines)
library(gam)
mod1 = gam(log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) + s(SFLA) +
s(EFF_AGE) + City + Province + s(RMTOT) + s(RMBED) + s(BATH) + BSMT +
ATTIC + Style, data = train.sd)
summary(mod1)
```

```
Call: gam(formula = log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) +
s(SFLA) +
    s(EFF_AGE) + City + Province + s(RMTOT) + s(RMBED) + s(BATH)
+
    BSMT + ATTIC + Style, data = train.sd)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-4.84250 -0.18279  0.01597  0.19346  2.43975

(Dispersion Parameter for gaussian family taken to be 0.1516)

    Null Deviance: 10936.51 on 23368 degrees of freedom
Residual Deviance: 3537.156 on 23327.3 degrees of freedom
AIC: 22281.03

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
               Df Sum Sq Mean Sq    F value
GRADE           4 2763.5  690.88  4556.3329
s(SQFT)         1    0.9    0.92     6.0386
s(FLR1AREA)     1  553.9  553.94  3653.2069
s(SFLA)         1 1073.2 1073.24  7077.9662
s(EFF_AGE)      1 1536.4 1536.38 10132.3469
City            1    1.1    1.07     7.0322
Province        1   26.9   26.87   177.2374
s(RMTOT)        1    3.2    3.19    21.0212
s(RMBED)        1    0.0    0.03     0.1754
s(BATH)         1   20.6   20.60   135.8793
BSMT            1   41.5   41.50   273.7113
ATTIC           1   33.1   33.08   218.1883
Style           4   19.2    4.81    31.7092
Residuals   23327 3537.2    0.15
               Pr(>F)
GRADE       < 2.2e-16 ***
s(SQFT)      0.014004 *
s(FLR1AREA) < 2.2e-16 ***
s(SFLA)     < 2.2e-16 ***
s(EFF_AGE)  < 2.2e-16 ***
City         0.008011 **
```

```
Province    < 2.2e-16 ***
s(RMTOT)    4.566e-06 ***
s(RMBED)     0.675327
s(BATH)     < 2.2e-16 ***
BSMT        < 2.2e-16 ***
ATTIC       < 2.2e-16 ***
Style       < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
(Intercept)
GRADE
s(SQFT)         3.7  12.95 6.560e-10 ***
s(FLR1AREA)     3.0  17.02 4.867e-11 ***
s(SFLA)         3.0 297.29 < 2.2e-16 ***
s(EFF_AGE)      3.0 450.16 < 2.2e-16 ***
City
Province
s(RMTOT)        3.0   1.44  0.227913
s(RMBED)        3.0   5.37  0.001075 **
s(BATH)         3.0   1.29  0.275913
BSMT
ATTIC
Style
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
preplot.gam1 = preplot(mod1)
pdf(file = "~/nty/interactive mod1.pdf")
for (i in 1:length(preplot.gam1)){
  plot(preplot.gam1[[i]])
}
dev.off()
      null device
          1
```

Dengan fungsi GAM ini, dapat kita lihat model dengan variabel-variabel yang telah dinyatakan dan bagaimana efek dari independent variable kepada dependent variable. Model ini memiliki nilai AIC 22281.03.

Apabila di analisa lebih, model ini memiliki nilai AIC yang cukup besar sehingga model ini belum dapat dinilai akurat. Artinya juga bahwa ada salah satu atau beberapa dari independent variable yang kurang cocok dan tidak memberikan pengaruh pada modelnya. Sehingga harus dicoba lagi untuk mencari model yang lebih akurat.

```
mod2 = gam(log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) + s(SFLA) +
s(EFF_AGE) + City + Province + s(RMTOT) + s(BATH) + BSMT + ATTIC +
Style, data = train.sd)
summary(mod2)
```

```
Call: gam(formula = log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) +
s(SFLA) +
    s(EFF_AGE) + City + Province + s(RMBED) + s(BATH) + BSMT +
    ATTIC + Style, data = train.sd)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-4.84250 -0.18256  0.01609  0.19355  2.44139

(Dispersion Parameter for gaussian family taken to be 0.1516)

    Null Deviance: 10936.51 on 23368 degrees of freedom
Residual Deviance: 3537.867 on 23331.3 degrees of freedom
AIC: 22277.73

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
               Df Sum Sq Mean Sq    F value
GRADE           4 2762.2  690.55  4553.9654
s(SQFT)         1    0.9    0.91     6.0277
s(FLR1AREA)     1  553.9  553.92  3652.9752
s(SFLA)         1 1072.2 1072.22  7070.9888
s(EFF_AGE)      1 1536.8 1536.83 10134.9983
City            1    1.1    1.07     7.0880
Province        1   26.9   26.86   177.1382
s(RMBED)        1    2.4    2.43    16.0490
s(BATH)         1   20.4   20.40   134.5589
BSMT            1   41.7   41.72   275.1349
ATTIC           1   33.5   33.48   220.7799
Style           4   19.3    4.82    31.8174
Residuals   23331 3537.9    0.15
               Pr(>F)
GRADE       < 2.2e-16 ***
s(SQFT)      0.014090 *
s(FLR1AREA) < 2.2e-16 ***
s(SFLA)     < 2.2e-16 ***
s(EFF_AGE)  < 2.2e-16 ***
City         0.007766 **
Province    < 2.2e-16 ***
s(RMBED)    6.192e-05 ***
s(BATH)     < 2.2e-16 ***
BSMT        < 2.2e-16 ***
ATTIC       < 2.2e-16 ***
Style       < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
```

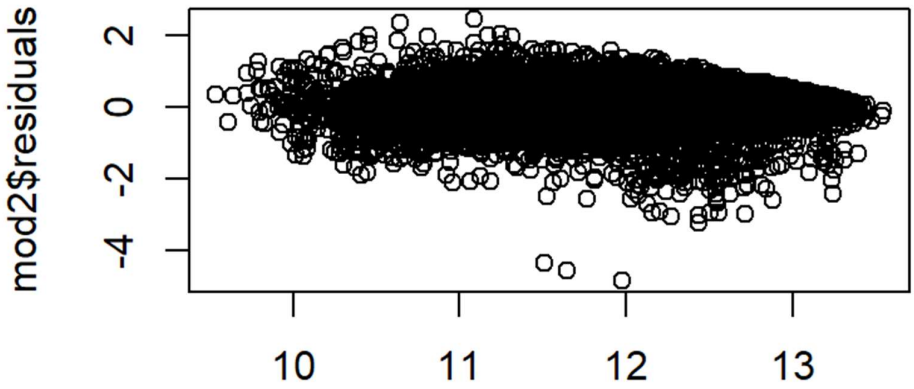
```
(Intercept)
GRADE
s(SQFT)         3.7  12.85 7.930e-10 ***
s(FLR1AREA)     3.0  17.04 4.725e-11 ***
s(SFLA)         3.0 297.62 < 2.2e-16 ***
s(EFF_AGE)      3.0 450.08 < 2.2e-16 ***
City
Province
s(RMBED)        3.0   7.11 9.018e-05 ***
s(BATH)         3.0   1.33    0.2638
BSMT
ATTIC
Style
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
preplot.gam2 = preplot(mod2)
pdf(file = "~/nty/interactive mod2.pdf")
for (i in 1:length(preplot.gam2)){
  plot(preplot.gam2[[i]])
}
dev.off()
```

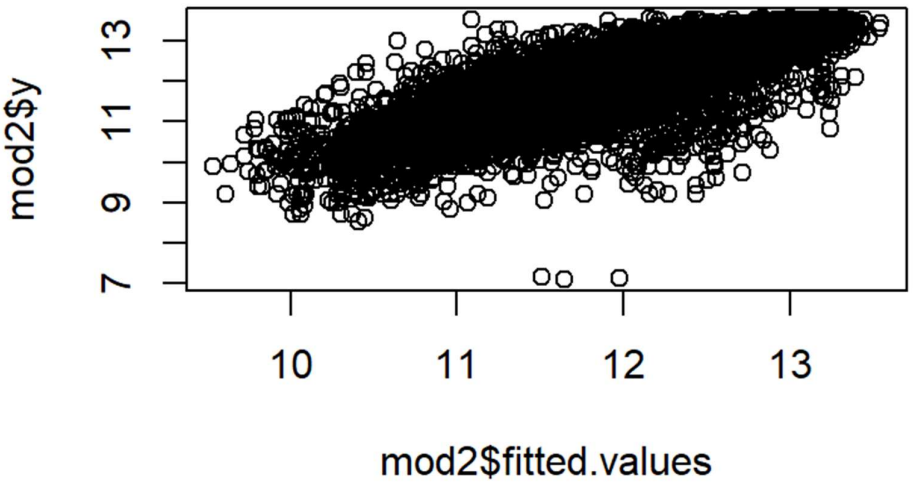
```
null device
          1
```

Dengan mod2 ini dapat dilihat bahwa kita mendapatkan nilai AIC yang lebih kecil dengan membuang independent variabel `RMTOT` yaitu 2227.73. Sehingga kita akan melanjutkan proses simulasi model menggunakan mod2 ini.

```
plot(mod2$fitted.values, mod2$residuals)
```

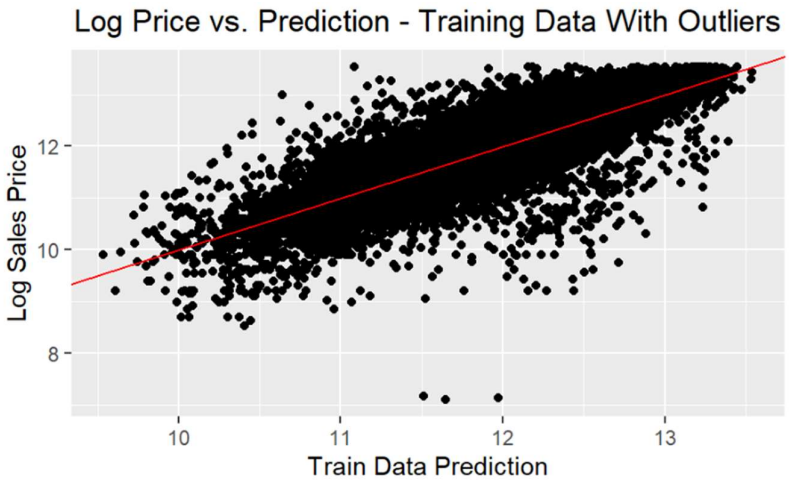



```
plot(mod2$fitted.values, mod2$y)
```



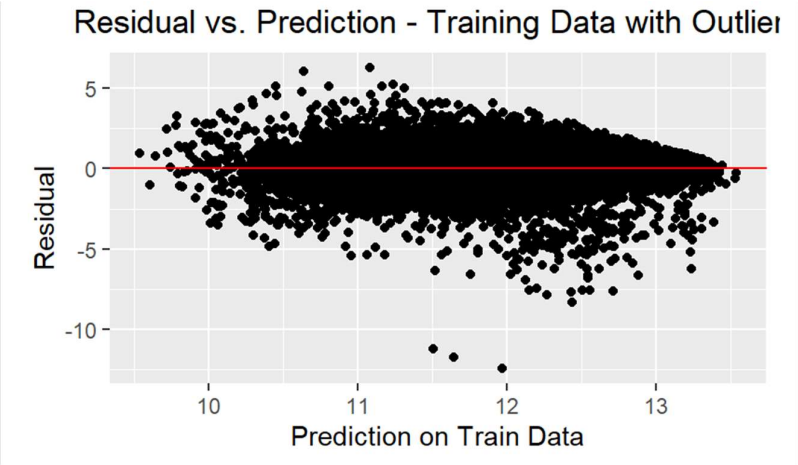

```
mod2$aic
```
```
22277.73
```

```
ggplot() +
  geom_point(aes(x = mod2$fitted.values, y = log(train.sd$Price))) +
  geom_abline(aes(intercept = 0, slope = 1), colour = "red") +
  ggtitle("Log Price vs. Prediction - Training Data With Outliers") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Train Data Prediction", y = "Log Sales Price")
```




```
residu = data.frame(x = rstandard(mod2))
pred = mod2$fitted.values

ggplot() +
  geom_point(aes(x = mod2$fitted.values, y = residu$x)) +
  geom_abline(aes(intercept = 0, slope = 0), colour = "red") +
  ggtitle("Residual vs. Prediction - Training Data with Outliers") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Prediction on Train Data", y = "Residual")
```



Dari 2 plot di atas, dapat dilihat bahwa data yang telah di prediksi oleh mod2 berada di garis regresi atau mendekati garis regresinya. Hal ini menunjukkan bahwa prediksi model mirip dengan nilai aslinya. Tetapi dari plot dapat dilihat masih ada beberapa outliers yang ada di dalam plot ini sehingga harus dibuang terlebih dahulu agar mendapatkan prediksi model yang lebih akurat.

```
bin = which(abs(residu) > 3)

if(length(bin)>0){
  train.outliers = train.sd
  train.outliers$outliers = 0
  train.outliers$outliers[bin] = 1
  train.outliers$pred = mod2$fitted.values
  train.outliers$pred.dollar = exp(train.outliers$pred)
  train.sd.2 = train.sd[-bin,]
} else{
  train.sd.2 = train.sd
}
```

Dalam fungsi di atas, semua nilai yang lebih besar daripada 3 (> 3) akan dibuang. Outliers ini yang terdapat di dalam data `train.sd` dibuang sehingga menghasilkan `train.sd.2` yang merupakan data baru tanpa outliers.

```
mod2.final = gam(log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) + s(SFLA)
+ s(EFF_AGE) + City + Province + s(RMTOT) + s(BATH) + BSMT + ATTIC +
Style, data = train.sd.2)
summary(mod2.final)
```

```
Call: gam(formula = log(Price) ~ GRADE + s(SQFT) + s(FLR1AREA) +
s(SFLA) +
    s(EFF_AGE) + City + Province + s(RMBED) + s(BATH) + BSMT +
    ATTIC + Style, data = train.sd.2)
Deviance Residuals:
      Min        1Q    Median        3Q
-1.245831 -0.182650  0.002615  0.171643
      Max
 1.208485

(Dispersion Parameter for gaussian family taken to be 0.1082)

    Null Deviance: 9819.411 on 23008 degrees of freedom
Residual Deviance: 2485.925 on 22971.31 degrees of freedom
AIC: 14173.54

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
               Df  Sum Sq Mean Sq    F value
GRADE           4 2730.87  682.72  6308.6724
s(SQFT)         1    0.99    0.99     9.1656
s(FLR1AREA)     1  600.10  600.10  5545.2205
s(SFLA)         1 1012.60 1012.60  9356.9350
s(EFF_AGE)      1 1534.77 1534.77 14182.1138
City            1    0.31    0.31     2.8836
Province        1   21.27   21.27   196.5354
s(RMBED)        1    2.40    2.40    22.2004
s(BATH)         1   25.44   25.44   235.0574
BSMT            1   24.68   24.68   228.0746
ATTIC           1    3.70    3.70    34.1754
Style           4   19.00    4.75    43.8940
Residuals   22971 2485.92    0.11
               Pr(>F)
GRADE       < 2.2e-16 ***
s(SQFT)      0.002469 **
s(FLR1AREA) < 2.2e-16 ***
s(SFLA)     < 2.2e-16 ***
s(EFF_AGE)  < 2.2e-16 ***
City         0.089499 .
Province    < 2.2e-16 ***
s(RMBED)    2.471e-06 ***
s(BATH)     < 2.2e-16 ***
BSMT        < 2.2e-16 ***
ATTIC       5.104e-09 ***
Style       < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
(Intercept)
GRADE
s(SQFT)         3.7  10.73 3.469e-08 ***
s(FLR1AREA)     3.0  14.42 2.220e-09 ***
s(SFLA)         3.0 358.73 < 2.2e-16 ***
s(EFF_AGE)      3.0 614.86 < 2.2e-16 ***
City
Province
s(RMBED)        3.0   8.11 2.155e-05 ***
s(BATH)         3.0   4.82  0.002356 **
BSMT
ATTIC
Style
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

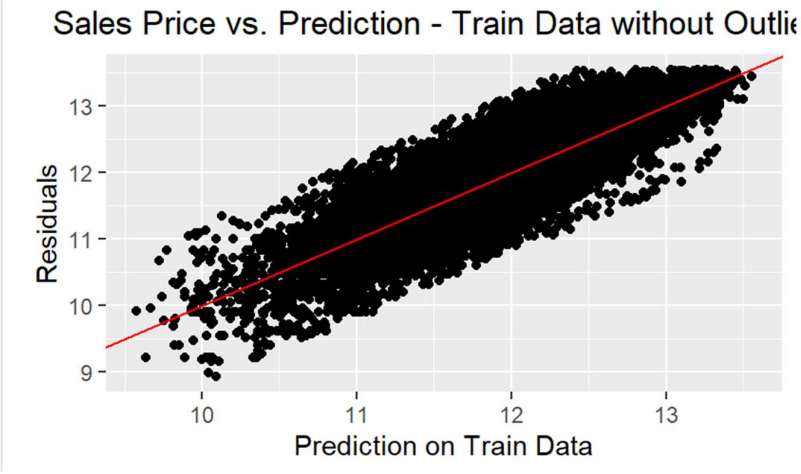
```
preplot.gam2.final = preplot(mod2.final)
pdf(file = "~/nty/interactive mod2.final.pdf")
for (i in 1:length(preplot.gam2.final)){
  plot(preplot.gam2.final[[i]])
}
dev.off()
```

```
null device
          1
```

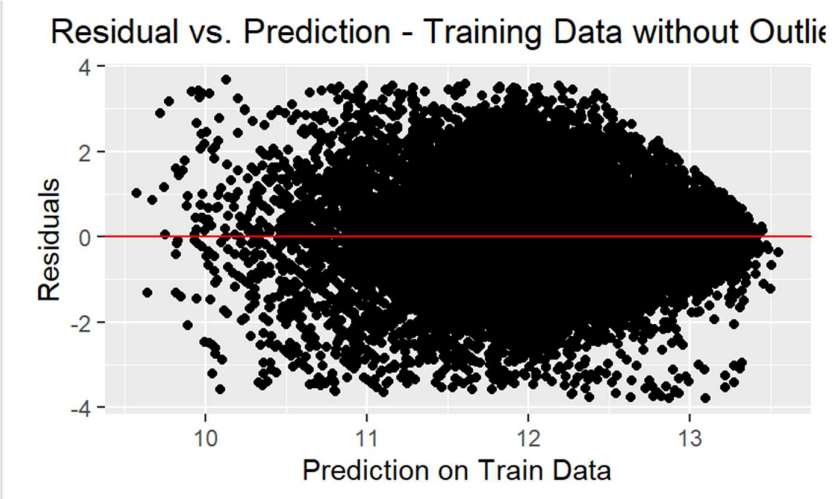
Dengan fungsi GAM telah dibuat model yang lebih akurat dengan menggunakan data train.sd.2 yang sudah tidak ada outliers. Dari mod2.final, didapatkan nilai AIC 14173.54. Dari nilai AIC ini dapat dilihat bahwa terdapat penurunan nilai yang signifikan dengan dibuangnya outliers yang ada. Prediksi model yang lebih baik akan kita gunakan dalam proses simulasi model selanjutnya.

```
residu.final = data.frame(x = rstandard(mod2.final))
pred = mod2.final$fitted.values
```

```
ggplot() +
  geom_point(aes(x = mod2.final$fitted.values, y =
log(train.sd.2$Price))) +
  geom_abline(aes(intercept = 0, slope = 1), colour = "red") +
  ggtitle("Sales Price vs. Prediction - Train Data without Outliers")
+
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Prediction on Train Data", y = "Residuals")
```




```
ggplot() +
  geom_point(aes(x = mod2.final$fitted.values, y = residu.final$x)) +
  geom_abline(aes(intercept = 0, slope = 0), colour = "red") +
  ggtitle("Residual vs. Prediction - Training Data without Outliers")
+
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Prediction on Train Data", y = "Residuals")
```



Dari 2 plot di atas, dapat dilihat bahwa prediksi model ini lebih mendekati lagi dengan nilai aslinya. Dapat dilihat juga bahwa residunya telah berkurang secara signifikan. Dapat dilihat juga dari plot kedua bahwa residu terletak tidak di satu spesifik tempat saja, melainkan random dan tersebar di seluruh model.

 # Cheking for Multicollinearity and Error

```
library(car)
vif(mod2.final)
                      GVIF Df GVIF^(1/(2*Df))
GRADE             1.745730  4        1.072129
s(SQFT)           1.000965  1        1.000482
s(FLR1AREA)       1.754752  1        1.324670
s(SFLA)           3.022615  1        1.738567
s(EFF_AGE)        1.156553  1        1.075431
City              1.145566  1        1.070311
Province          1.167127  1        1.080337
s(RMBED)          1.382873  1        1.175956
s(BATH)           2.031623  1        1.425350
BSMT              1.307211  1        1.143333
ATTIC             1.192308  1        1.091929
Style             1.310142  4        1.034344
```

Dapat dilihat bahwa nilai VIF dari independent variable yang ada di dalam mod2 ini memiliki nilai yang lebih kecil dari 5 maka artinya semua independent variable ini memiliki korelasi sedang sehingga kemungkinan terdapat hubungan diantara independent variable juga kecil. Maka model yang dihasilkan cukup akurat. Selanjutnya akan kita run error.analysis pada data `train.sd.2`.

```
comp = function(pred, obs){
  n = length(obs)
  rsq = cor(pred, obs)^2
  mse = sum((pred - obs)^2)/ n
  semse = sd((pred - obs)^2) / sqrt(n)
  rmse = sqrt(mse)
  se = sd(pred - obs) / sqrt(n)
  mae = sum(abs(pred - obs)) / n
  mape = sum(abs(pred-obs)/obs)/n*100

return(list("n" = n, "R2" = rsq, "MSE" = mse, "SEMSE" = semse, "RMSE"
= rmse, "SE" = se, "MAE" = mae, "MAPE" = mape))
}
```

```
comp(mod2.final$fitted.values, mod2.final$y)
      $n
      [1] 23009

      $R2
      [1] 0.7468458

      $MSE
      [1] 0.1080414

      $SEMSE
      [1] 0.00127332

      $RMSE
      [1] 0.3286965

      $SE
      [1] 0.002166984

      $MAE
      [1] 0.2431503

      $MAPE
      [1] 2.038631
```

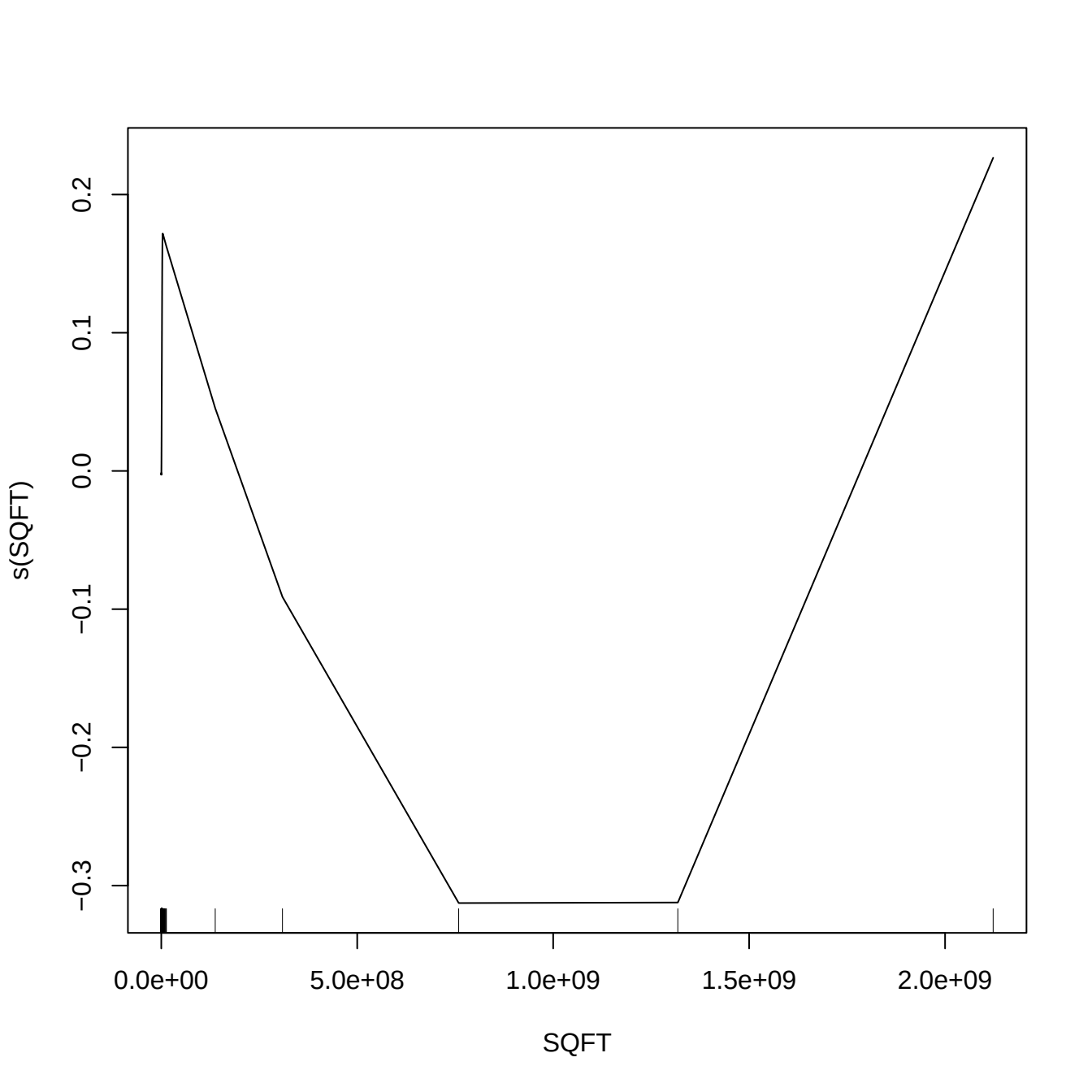
Dari hasil error analysis pada train.sd.2, kita dapatkan nilai RMSE yang kecil yaitu 0.33 maka dapat dinilai bahwa prediksi model ini akurat. Ditambah dengan MAPE sebesar 2.04% yang melengkapi kesimpulan kita bahwa prediksi model ini akurat. Selanjutnya akan dilakukan error anaylisis pada data `test.sd`.

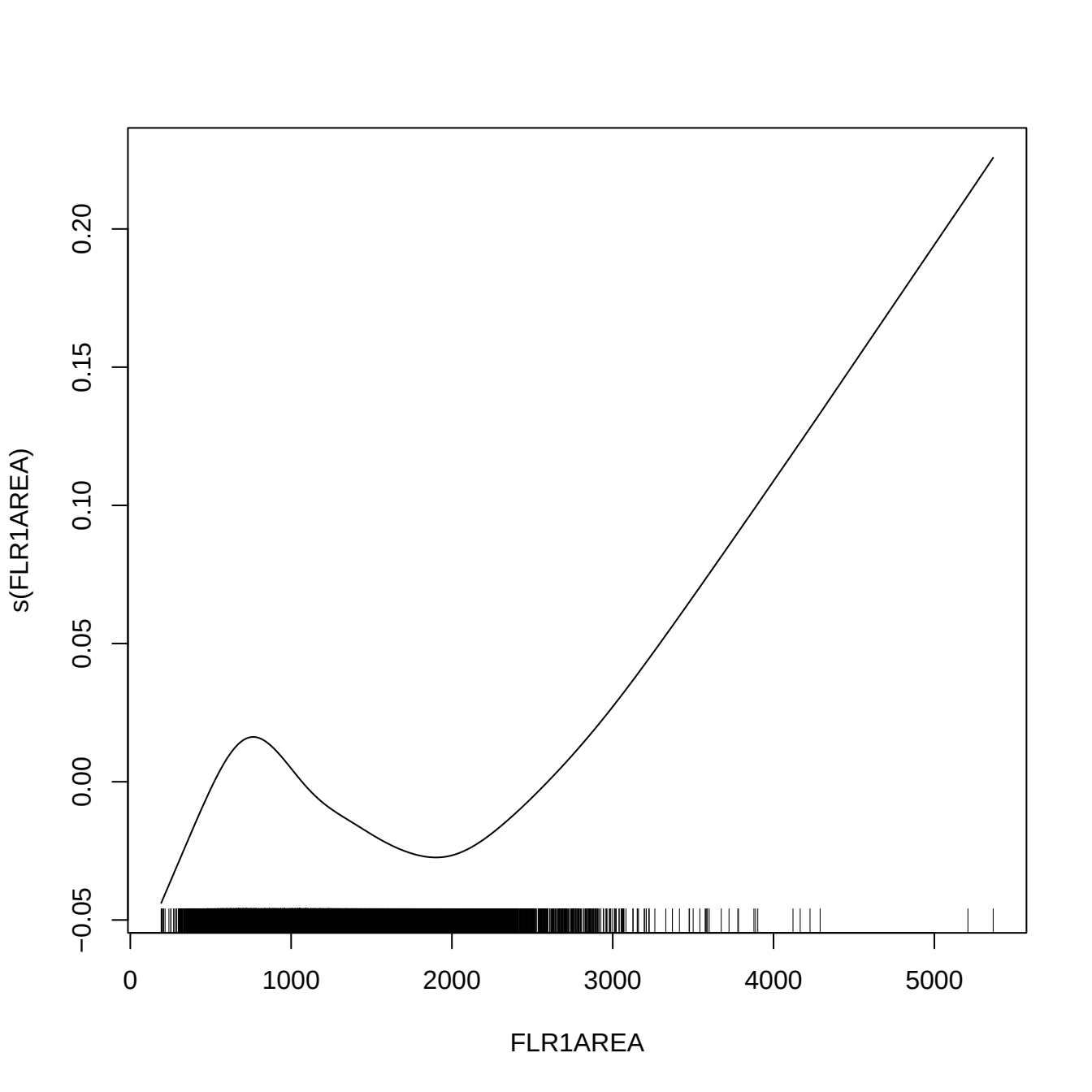
```
test.sd$prediction = predict(mod2.final , newdata = test.sd)
summary(test.sd$prediction)
        Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
         9.66   11.85   12.19   12.12   12.48   13.49
```

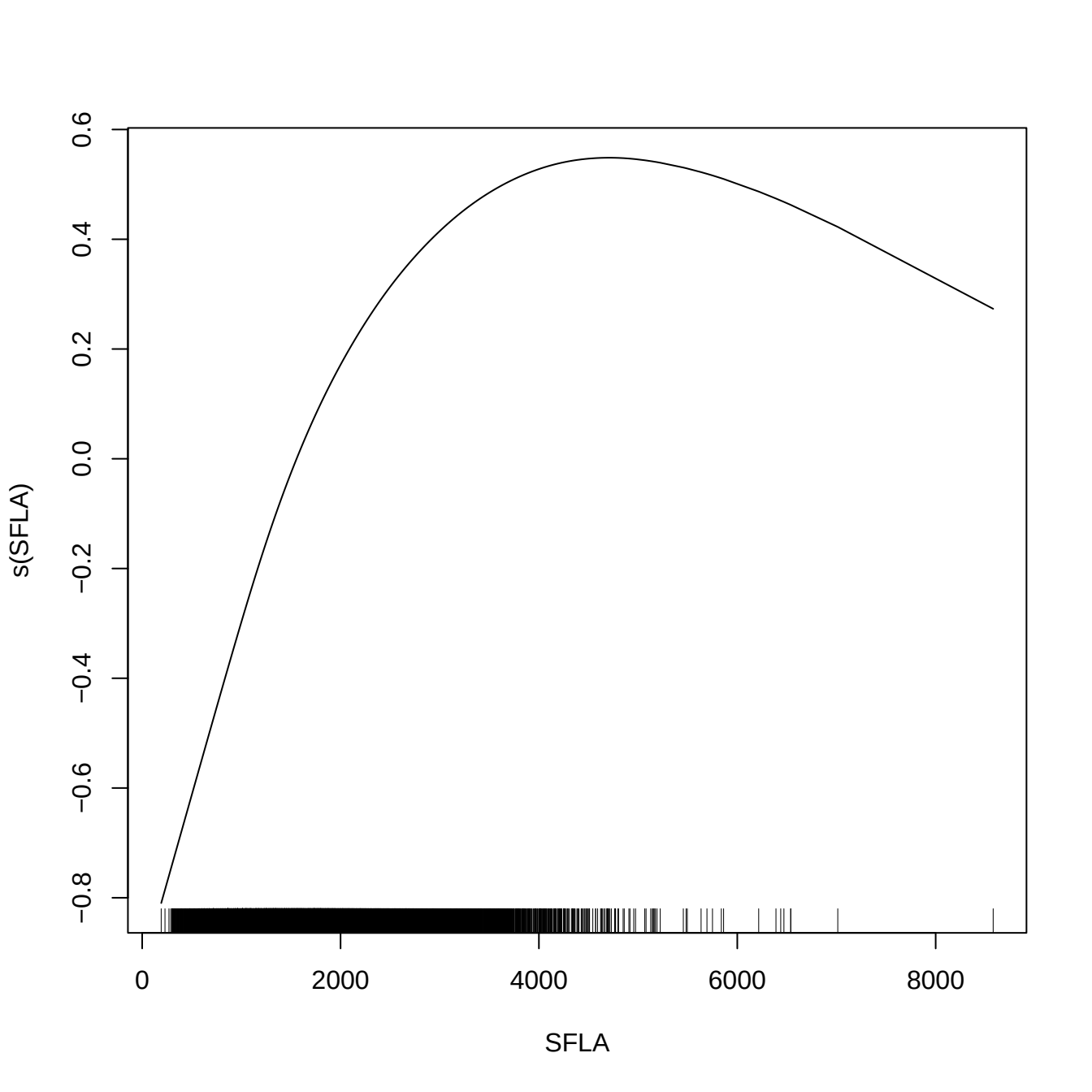
 # Kesimpulan

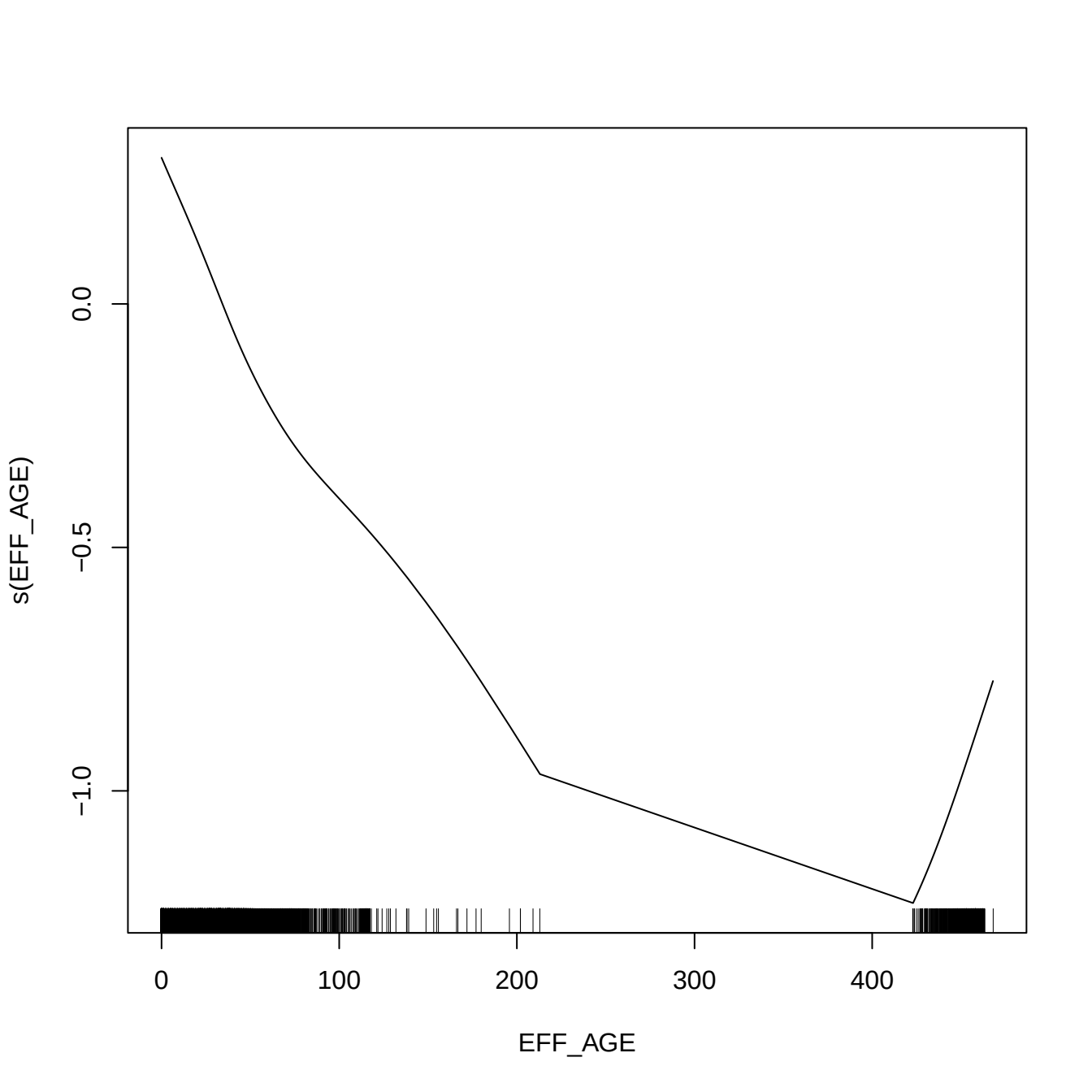
Pada projek UTS kali ini, kami gunakan GAM untuk mensimulasikan model yang dapat merepresentasikan efek dari independent variabel pada dependent variable. Independent variable yang digunakan kali ini adalah City, Province, SQFT, RMBED, BATH, FLR1AREA, BSMT, ATTIC, SFLA, GRADE, EFF_AGE, Style. Dependent variable yang digunakan adalah Price.

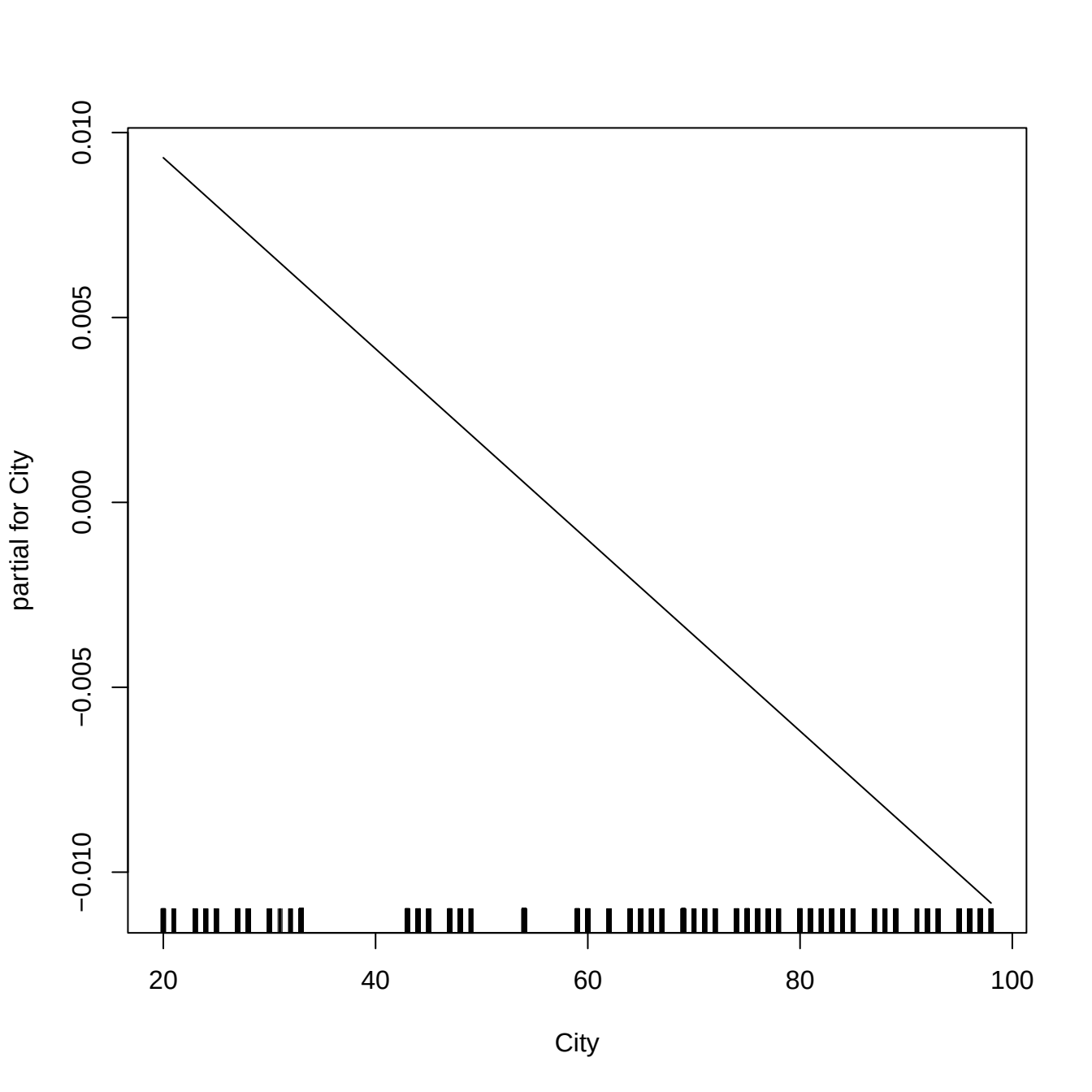
Model yang dihasilkan kali ini dinilai cukup akurat dengan AIC model final yang lebih kecil yaitu 14173.54 dibandingkan model dengan semua independent variable digunakan. Selain itu, juga karena RMSE yang dihasilkan kecil yaitu 0.33 dan MAPE sebesar 2.04% melengkapi kesimpulan bahwa model ini akurat. R-squared yang dihasilkan 0.7468% yang artinya model ini dapat menjelaskan 74.68% dari seluruh datanya.

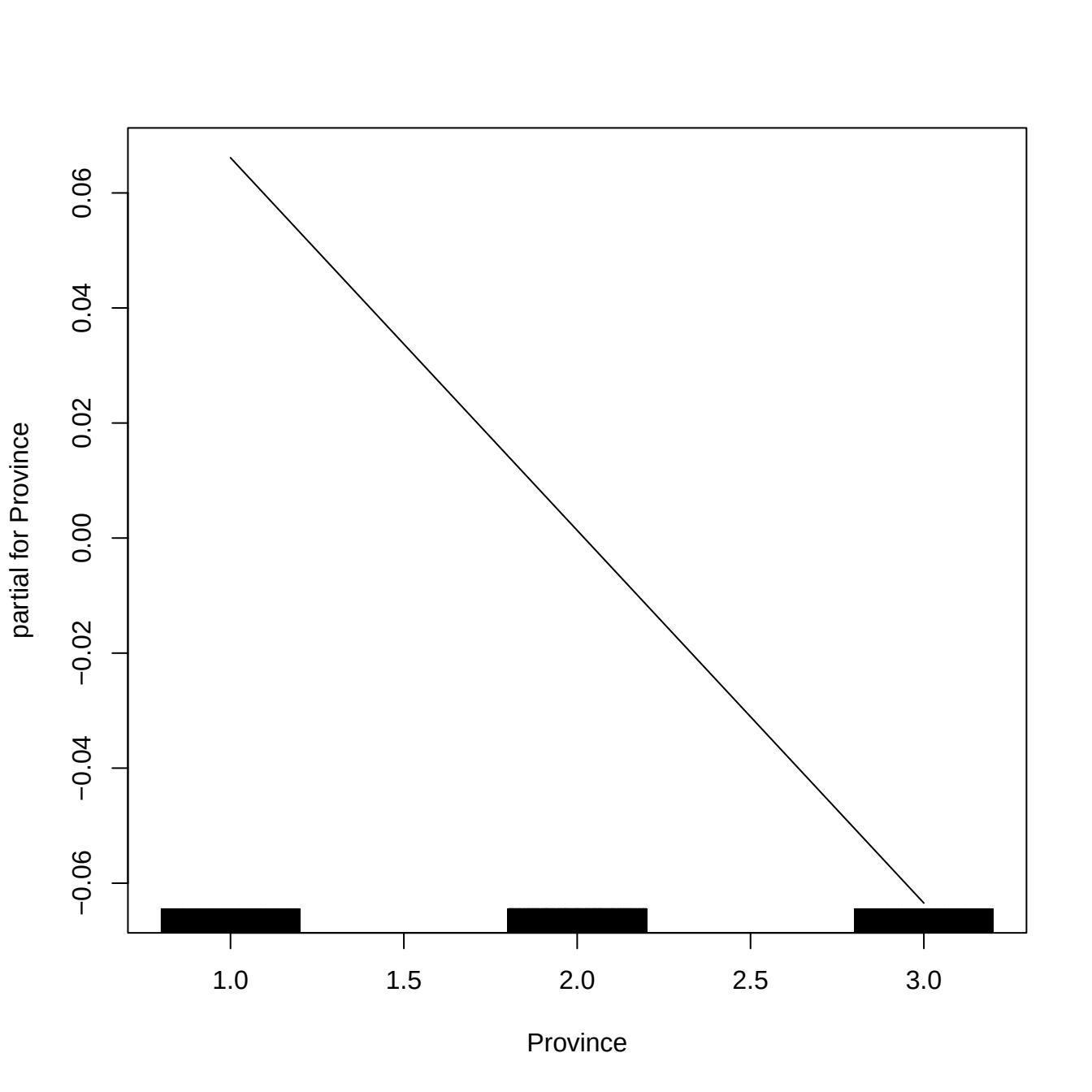

s(SQFT)
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
SQFT

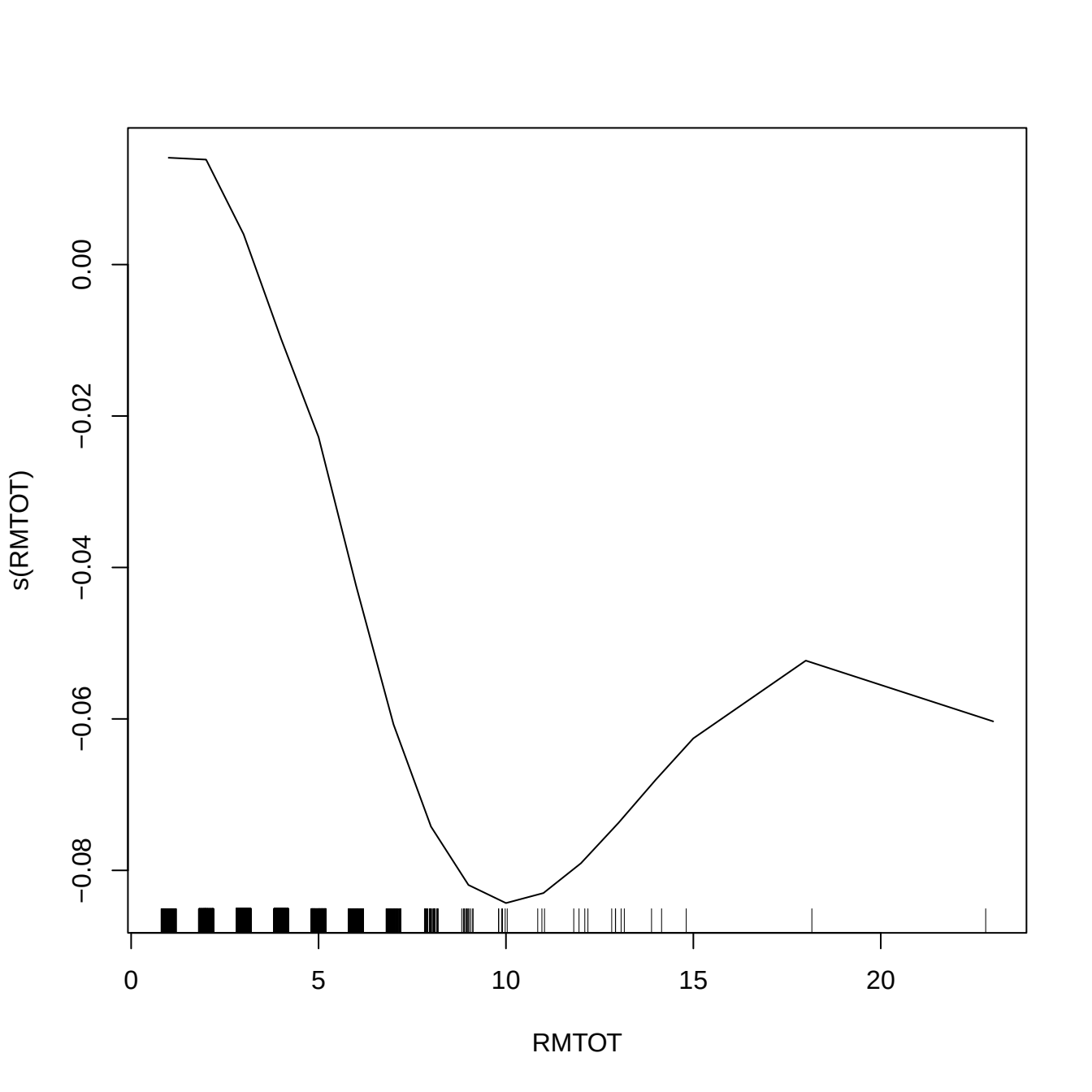

s(FLR1AREA)
0.20
0.15
0.10
0.05
0.00
−0.05
0
1000
2000
3000
4000
5000
FLR1AREA

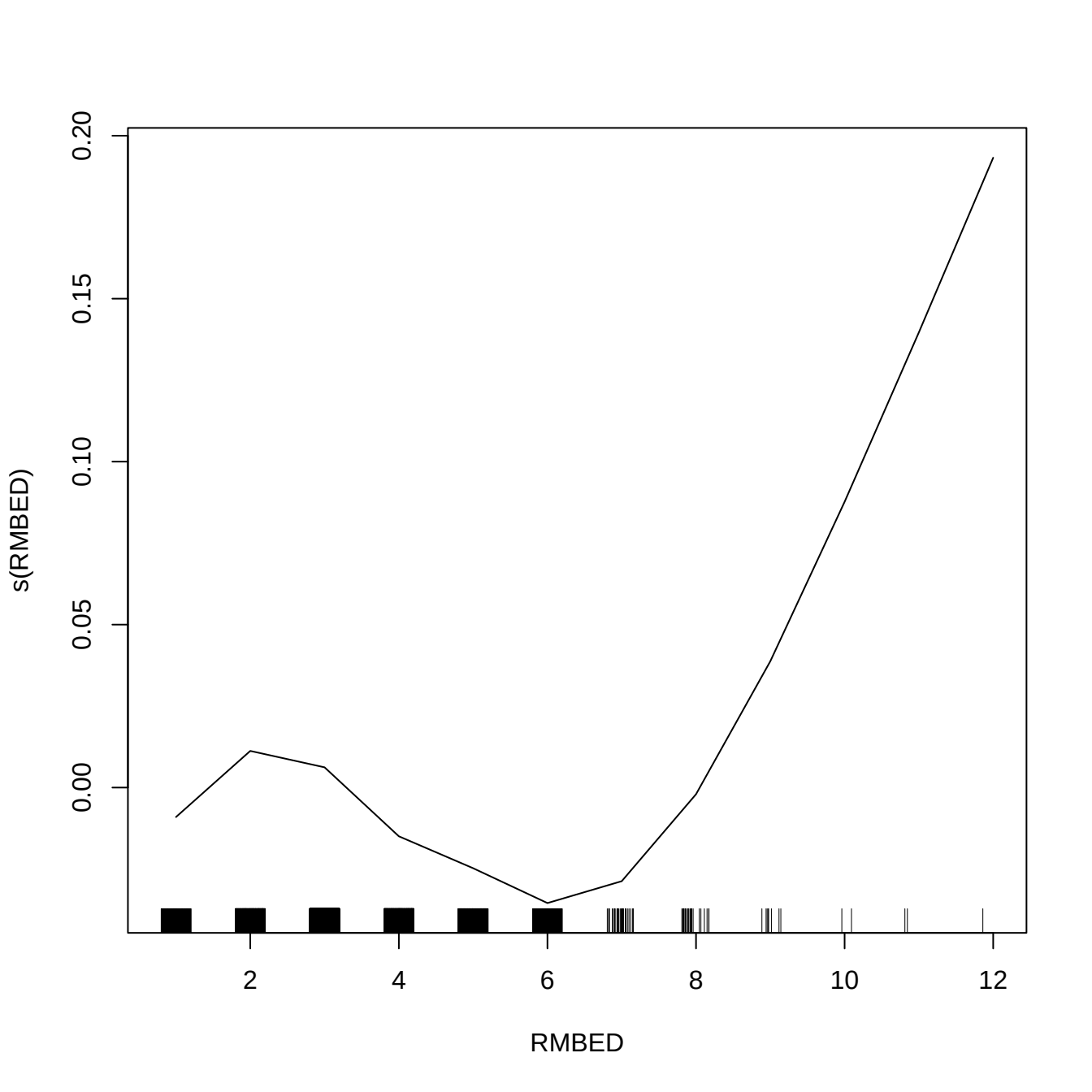

s(SFLA)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
2000
4000
6000
8000
SFLA

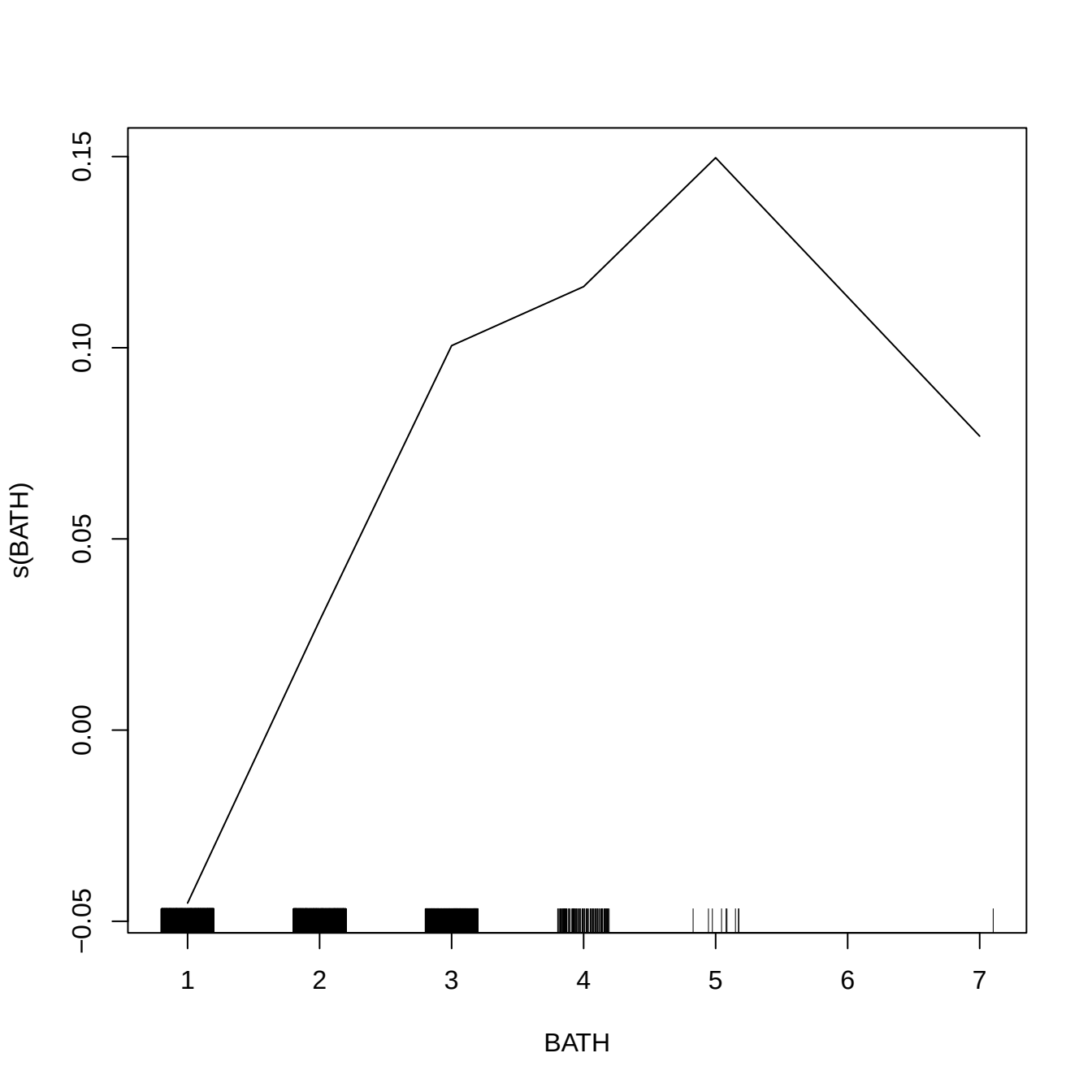

s(EFF_AGE)
0.0
−0.5
−1.0
0
100
200
300
400
EFF_AGE

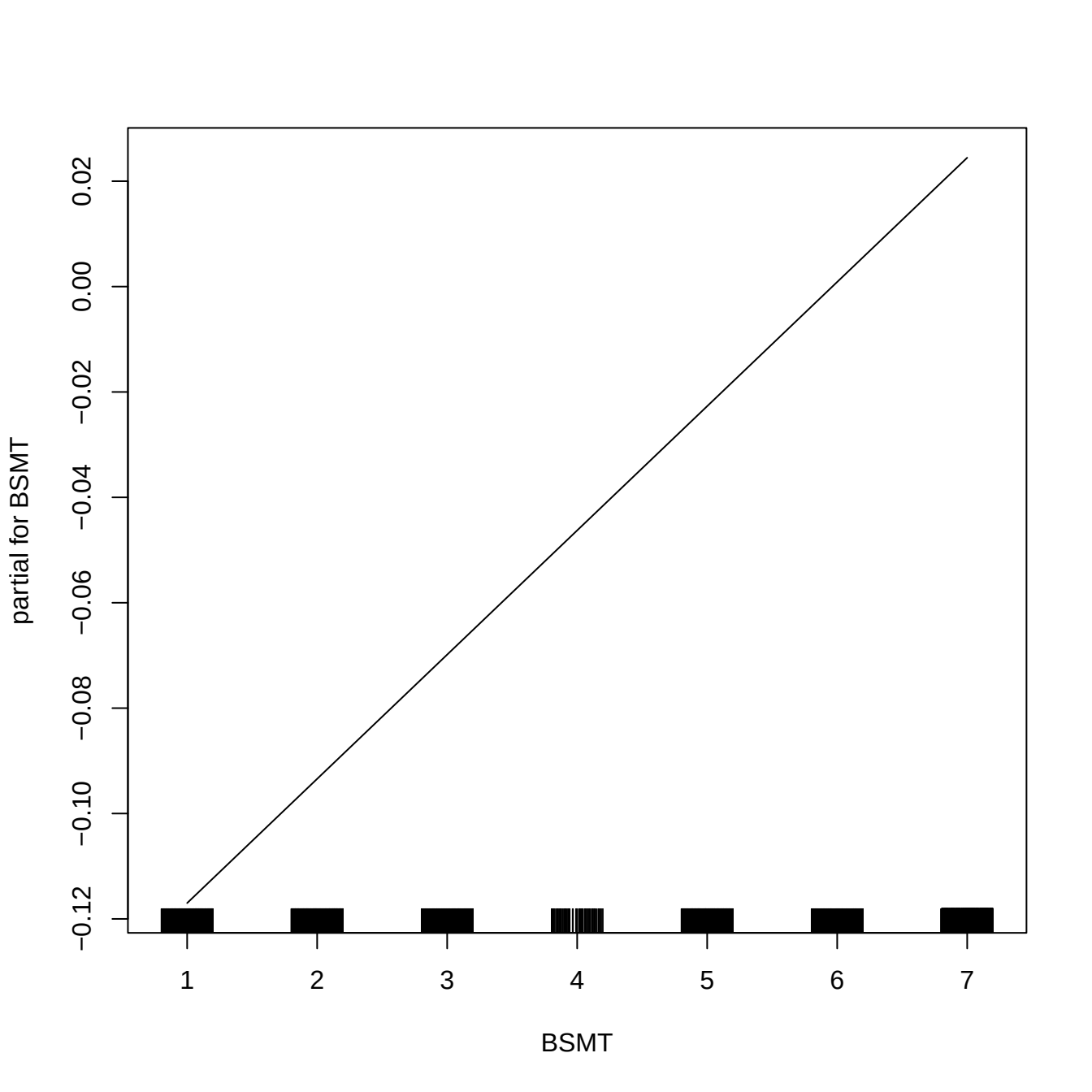

0.010
0.005
0.000
−0.005
−0.010
partial for City
20
40
60
80
100
City

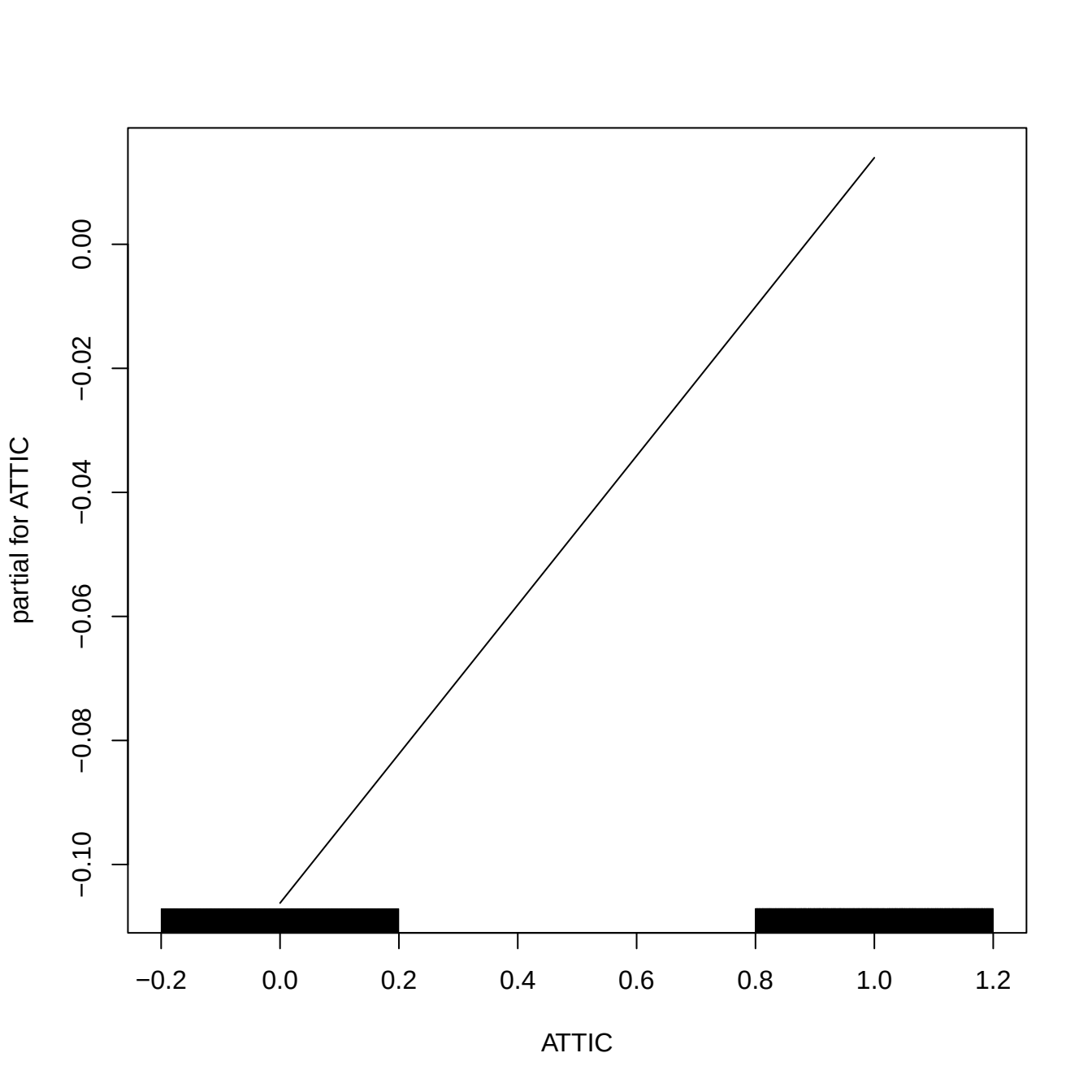

partial for Province
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
1.0
1.5
2.0
2.5
3.0
Province

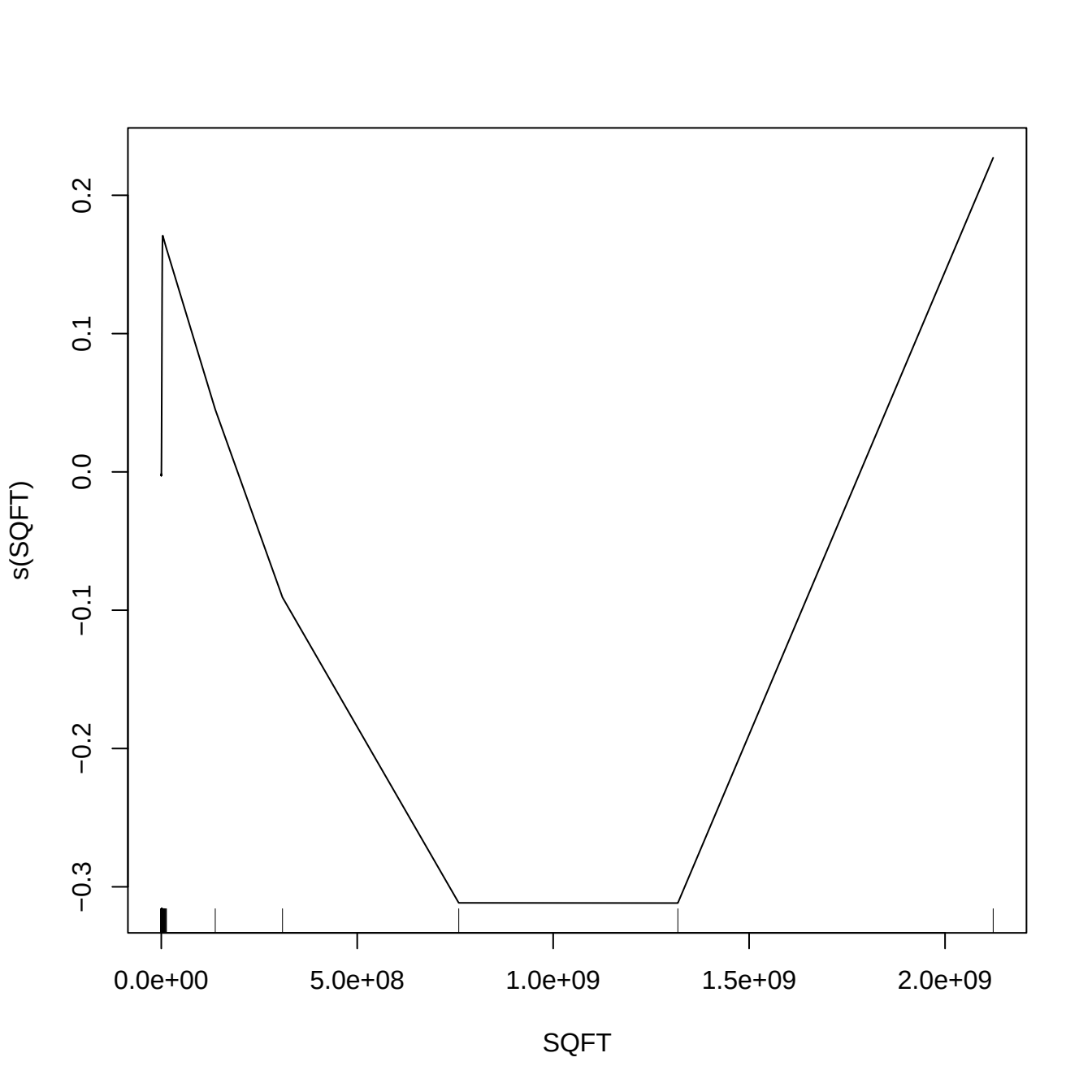

s(RMTOT)
0.00
−0.02
−0.04
−0.06
−0.08
0
5
10
15
20
RMTOT

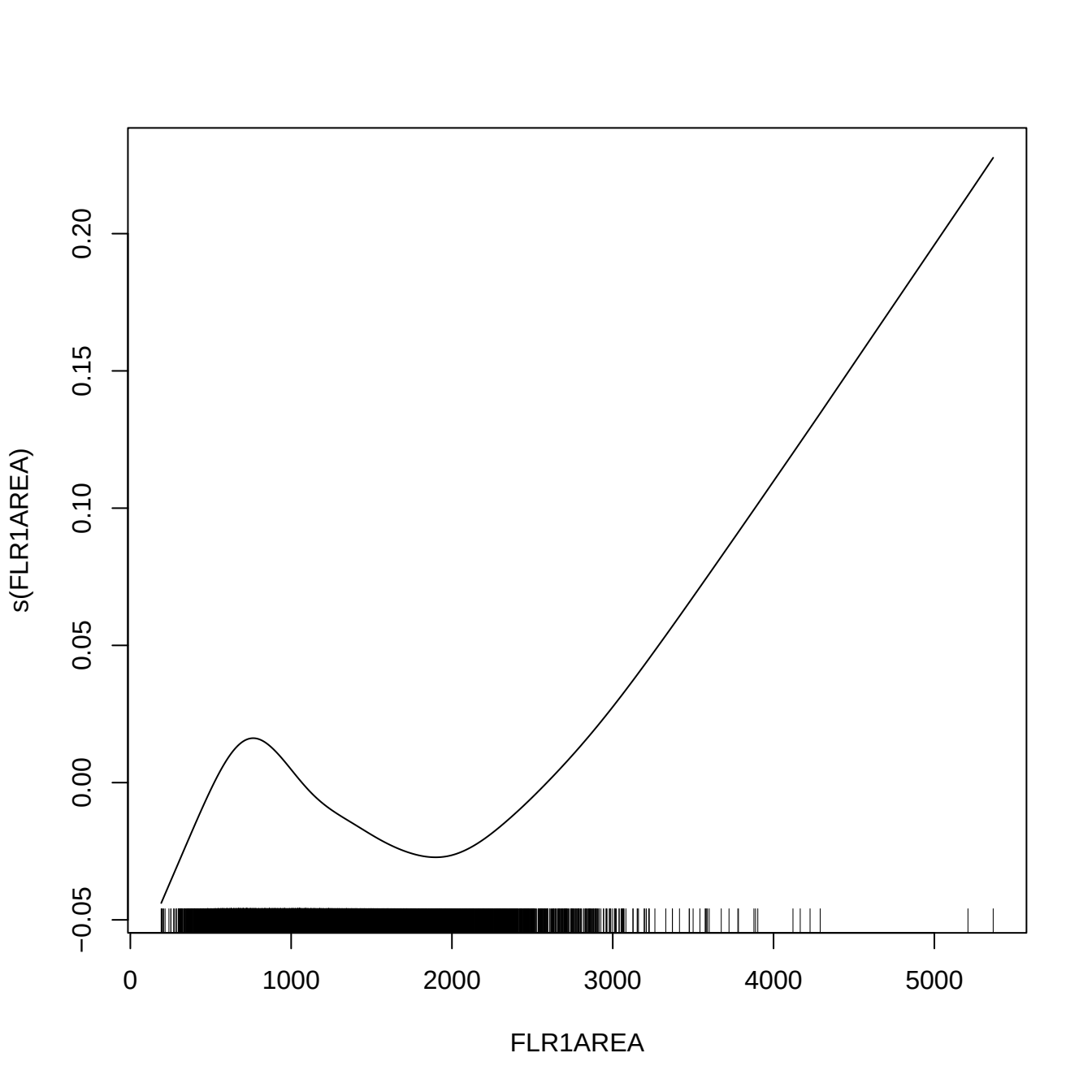

0.20
0.15
0.10
0.05
0.00
s(RMBED)
2
4
6
8
10
12
RMBED

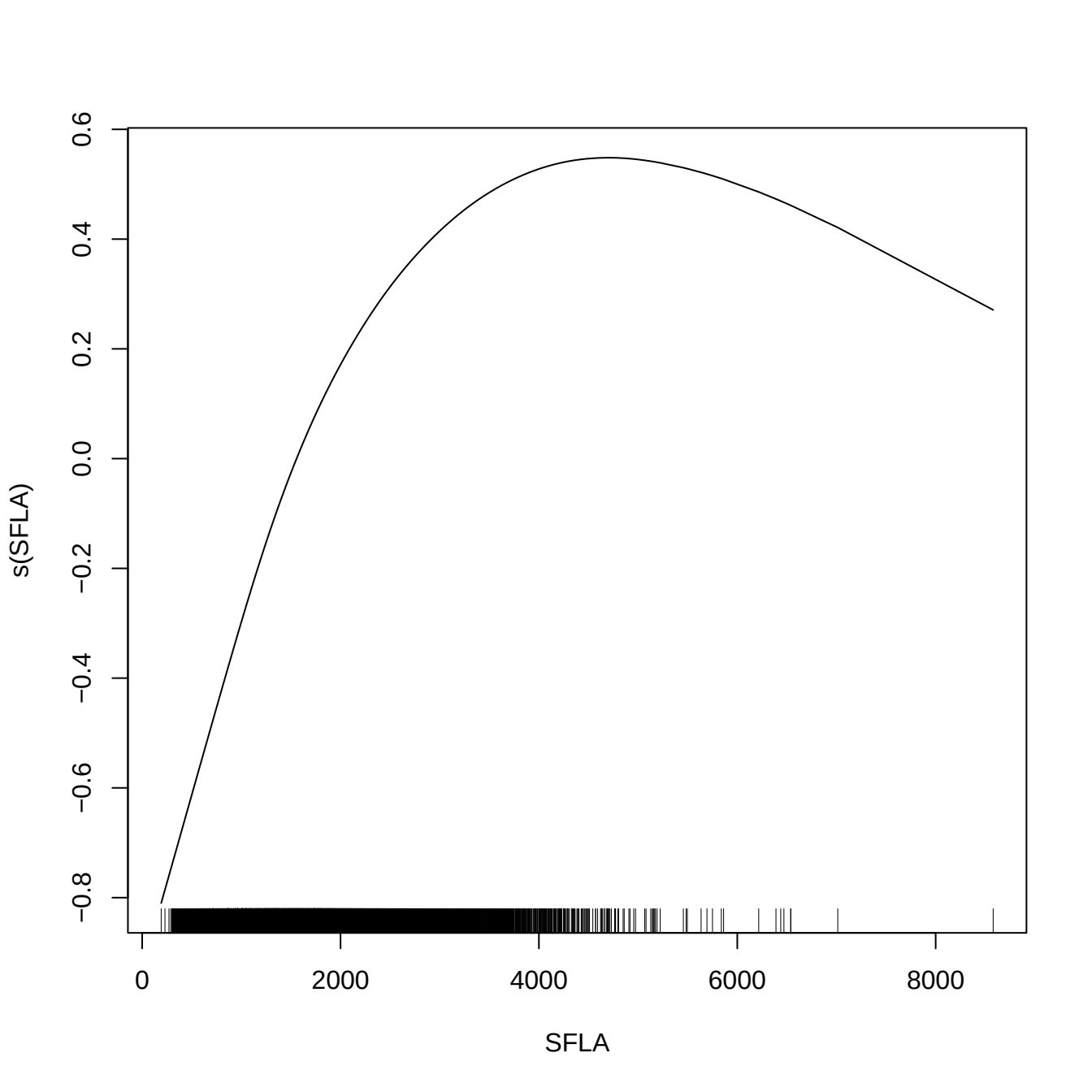

s(BATH)
0.15
0.10
0.05
0.00
−0.05
1
2
3
4
5
6
7
BATH

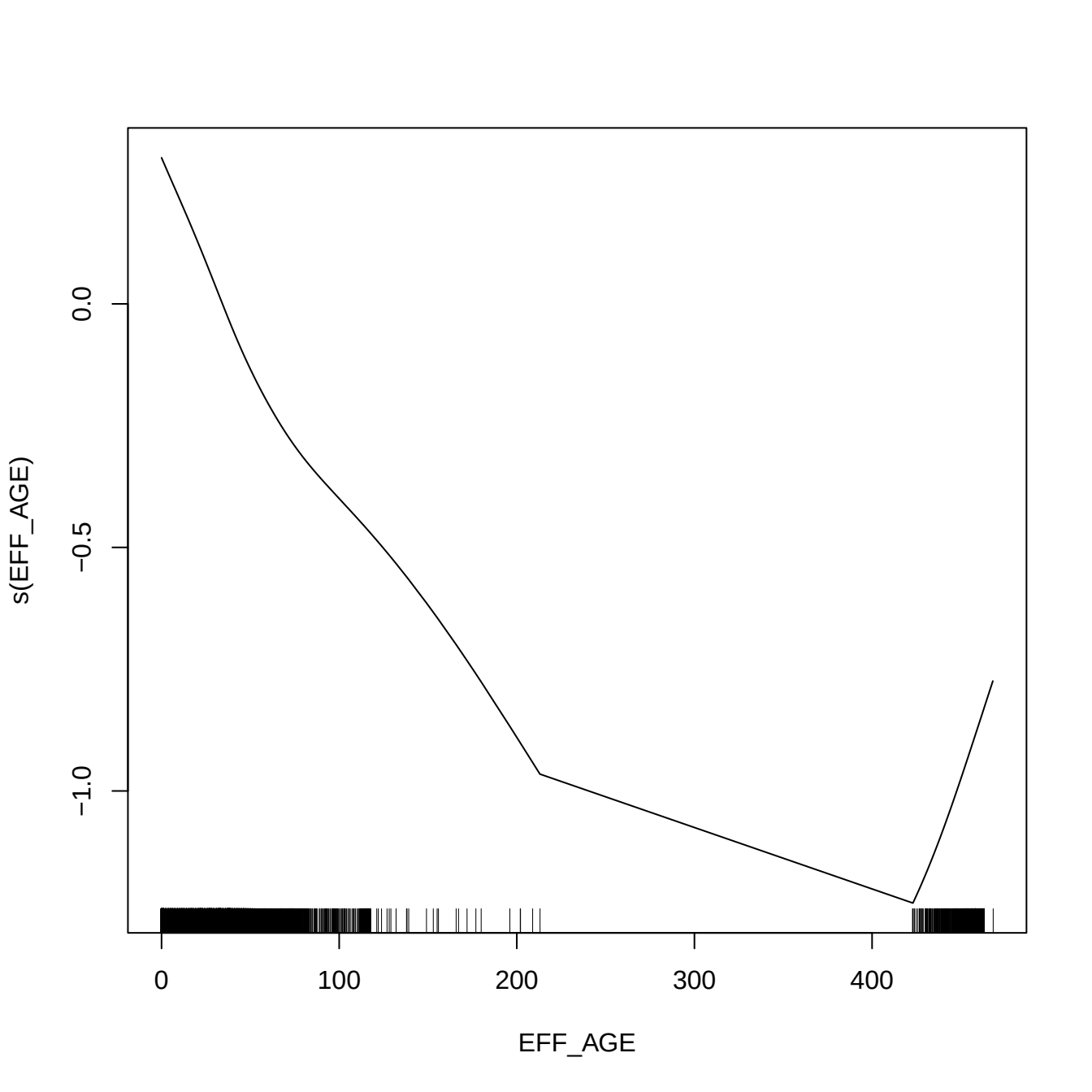

partial for BSMT
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
1
2
3
4
5
6
7
BSMT

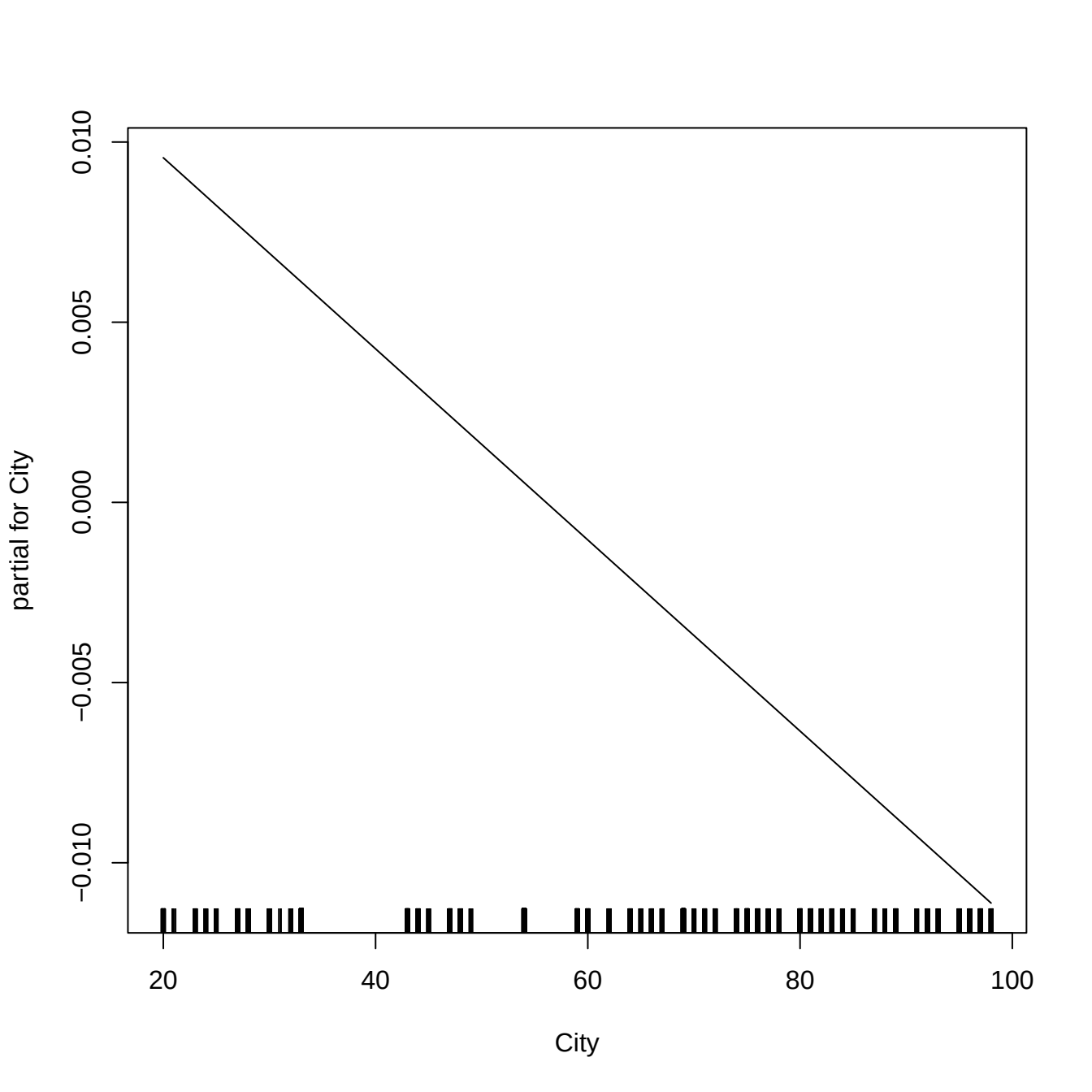

partial for ATTIC
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ATTIC

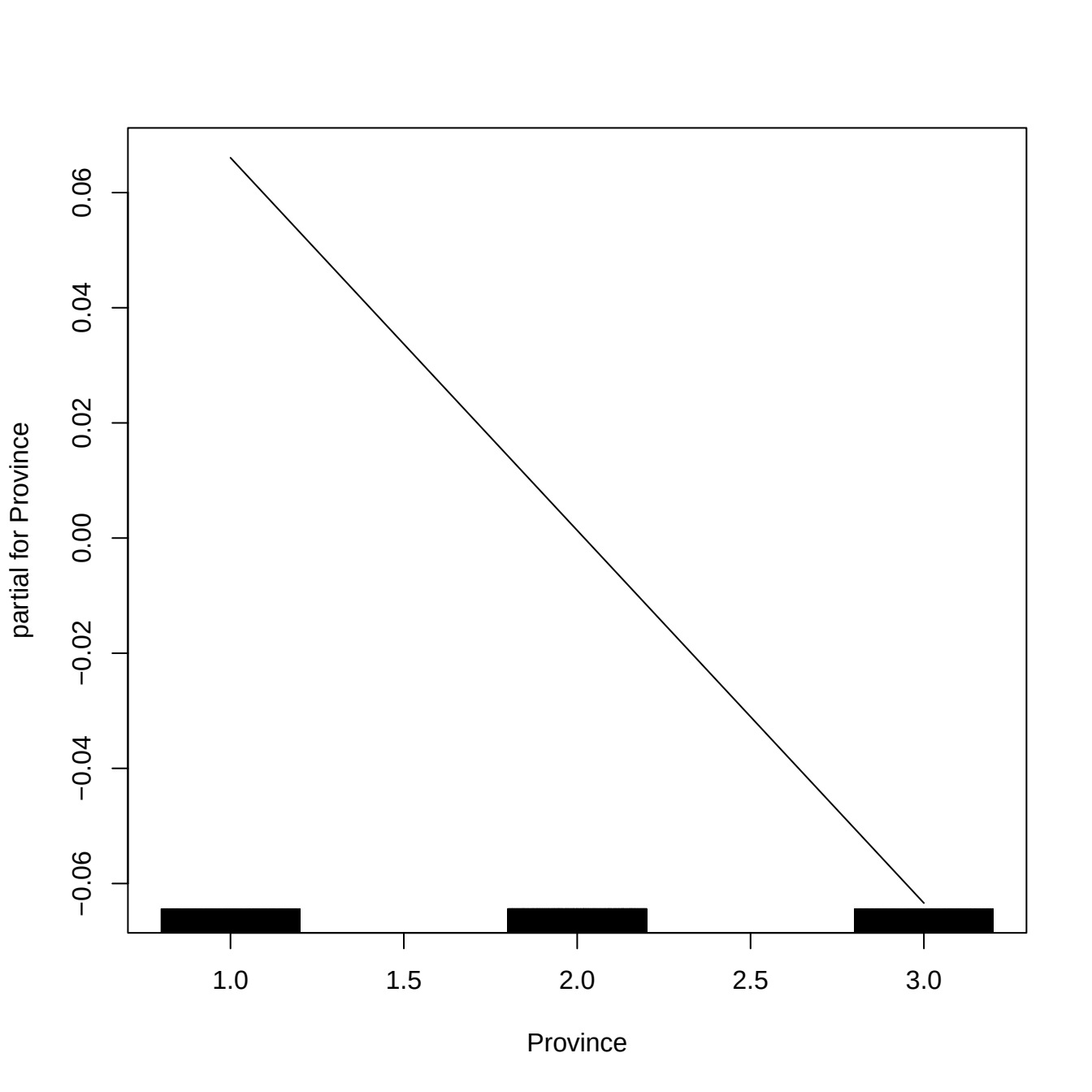

s(SQFT)
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
SQFT

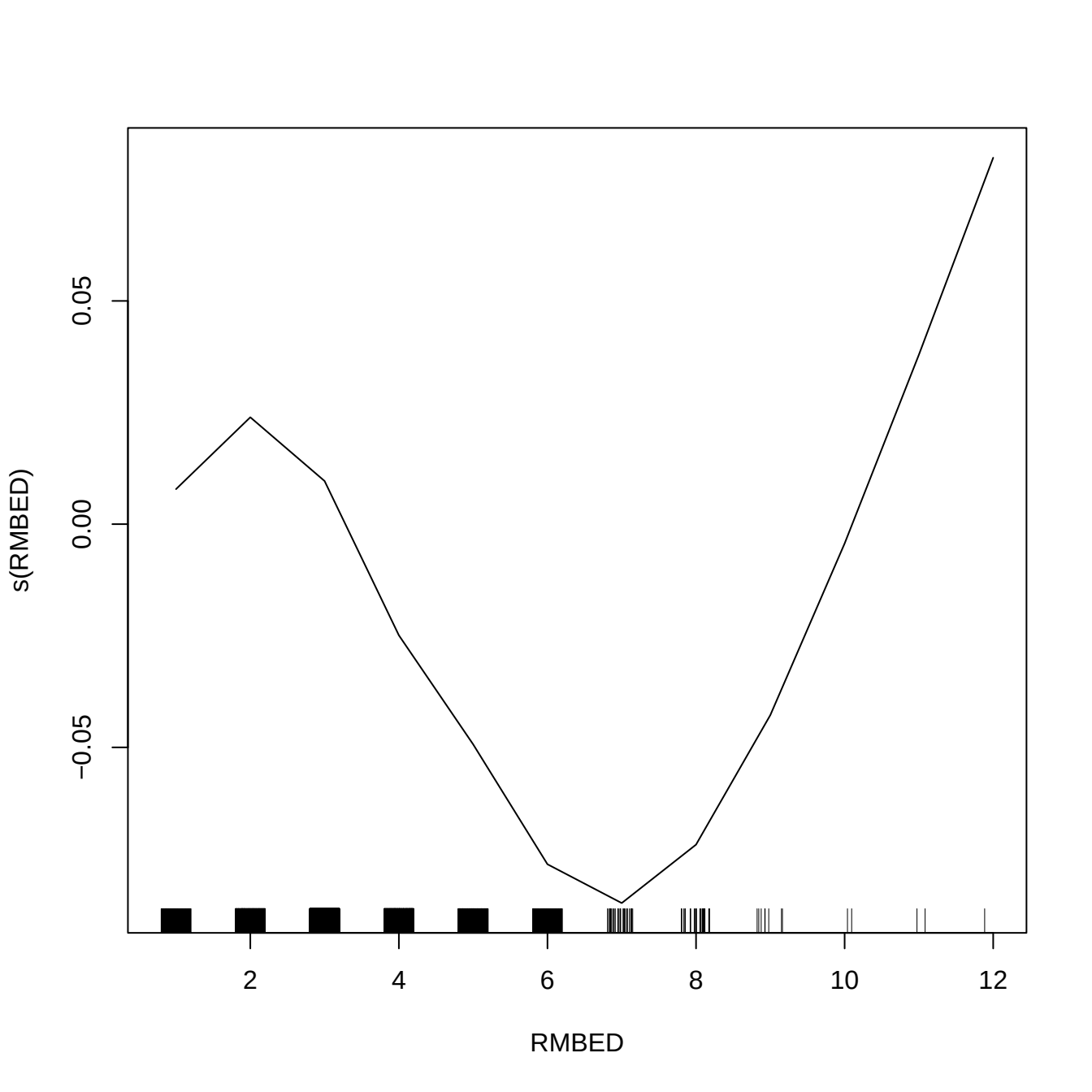

s(FLR1AREA)
0.20
0.15
0.10
0.05
0.00
−0.05
0
1000
2000
3000
4000
5000
FLR1AREA

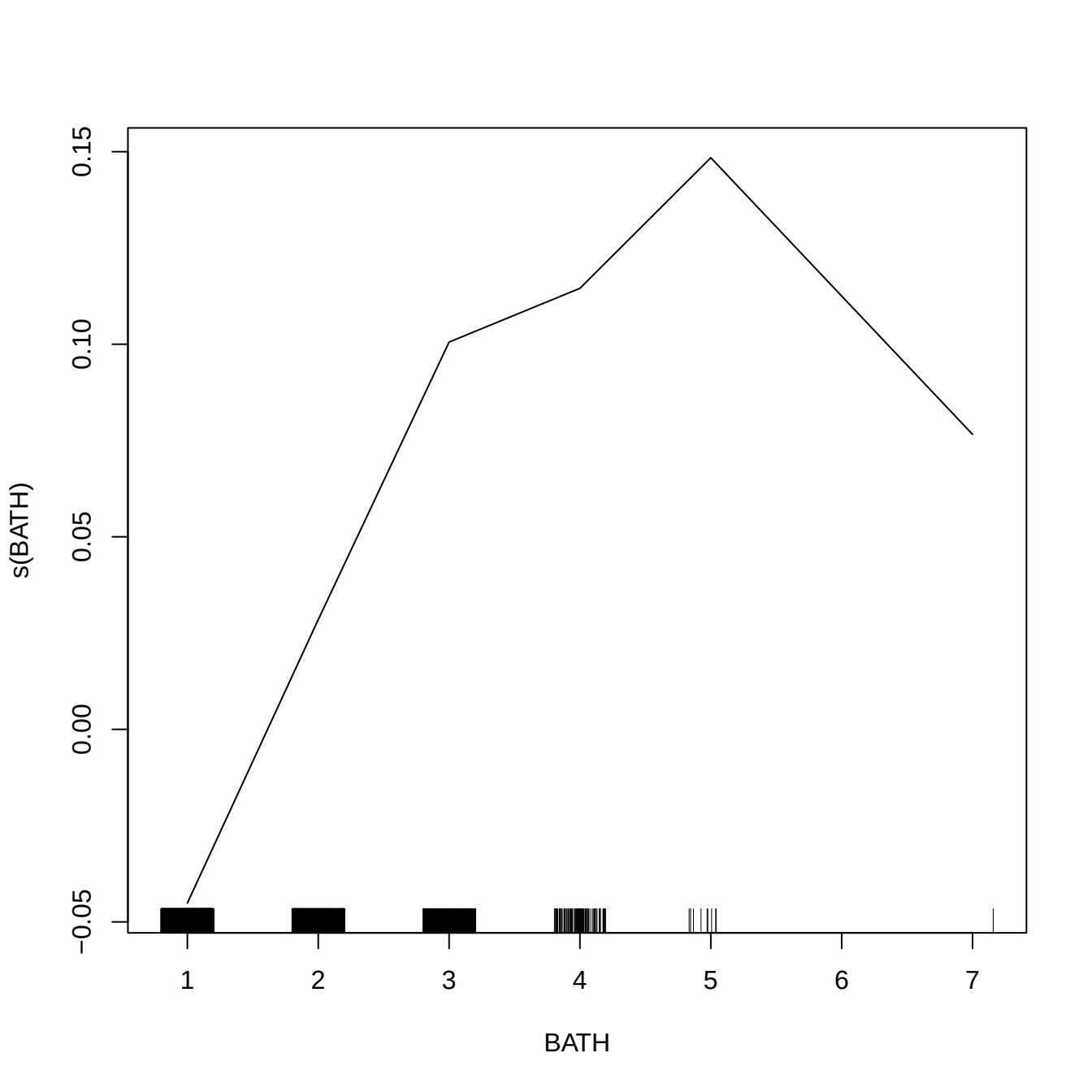

s(SFLA)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
2000
4000
6000
8000
SFLA

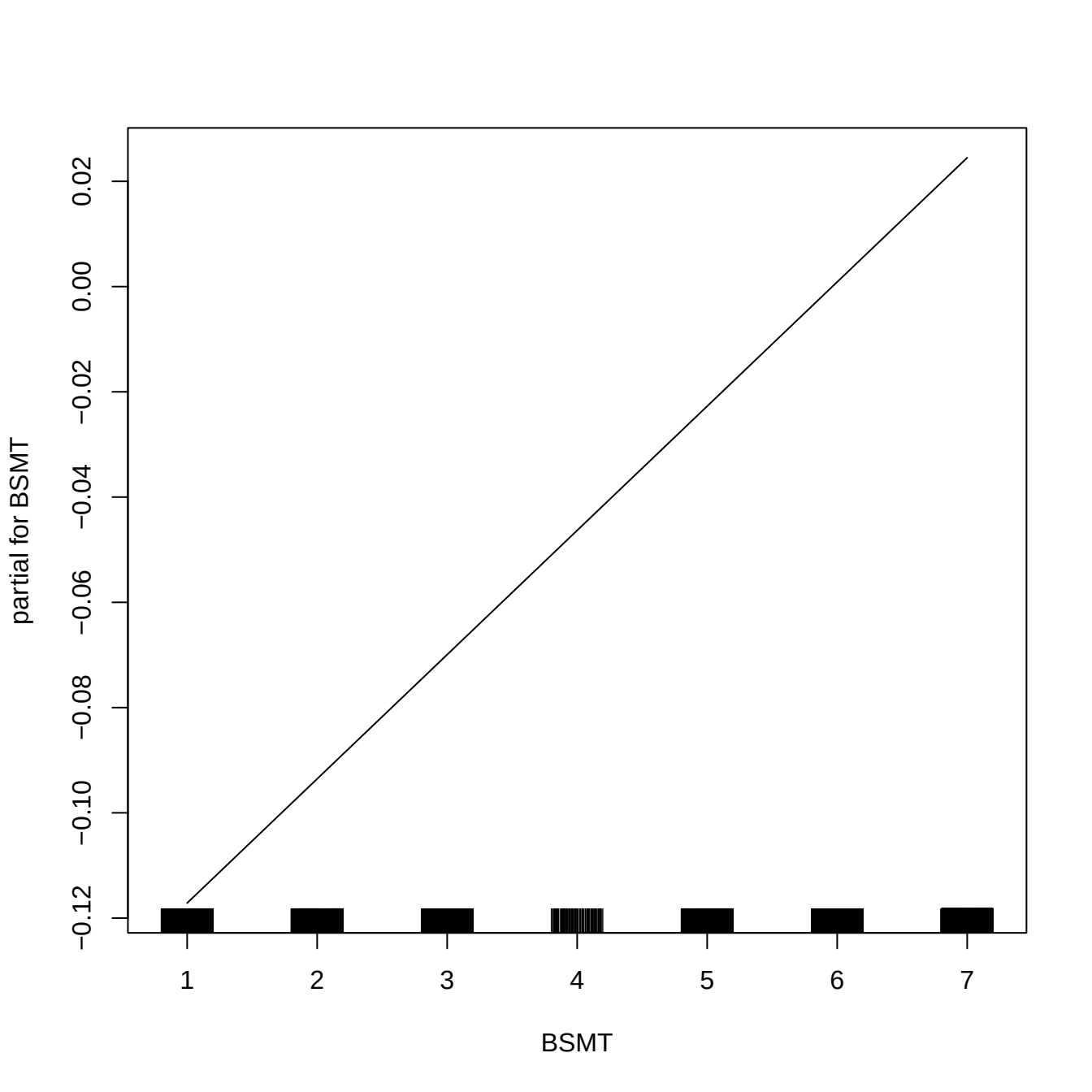

s(EFF_AGE)
0.0
−0.5
−1.0
0
100
200
300
400
EFF_AGE

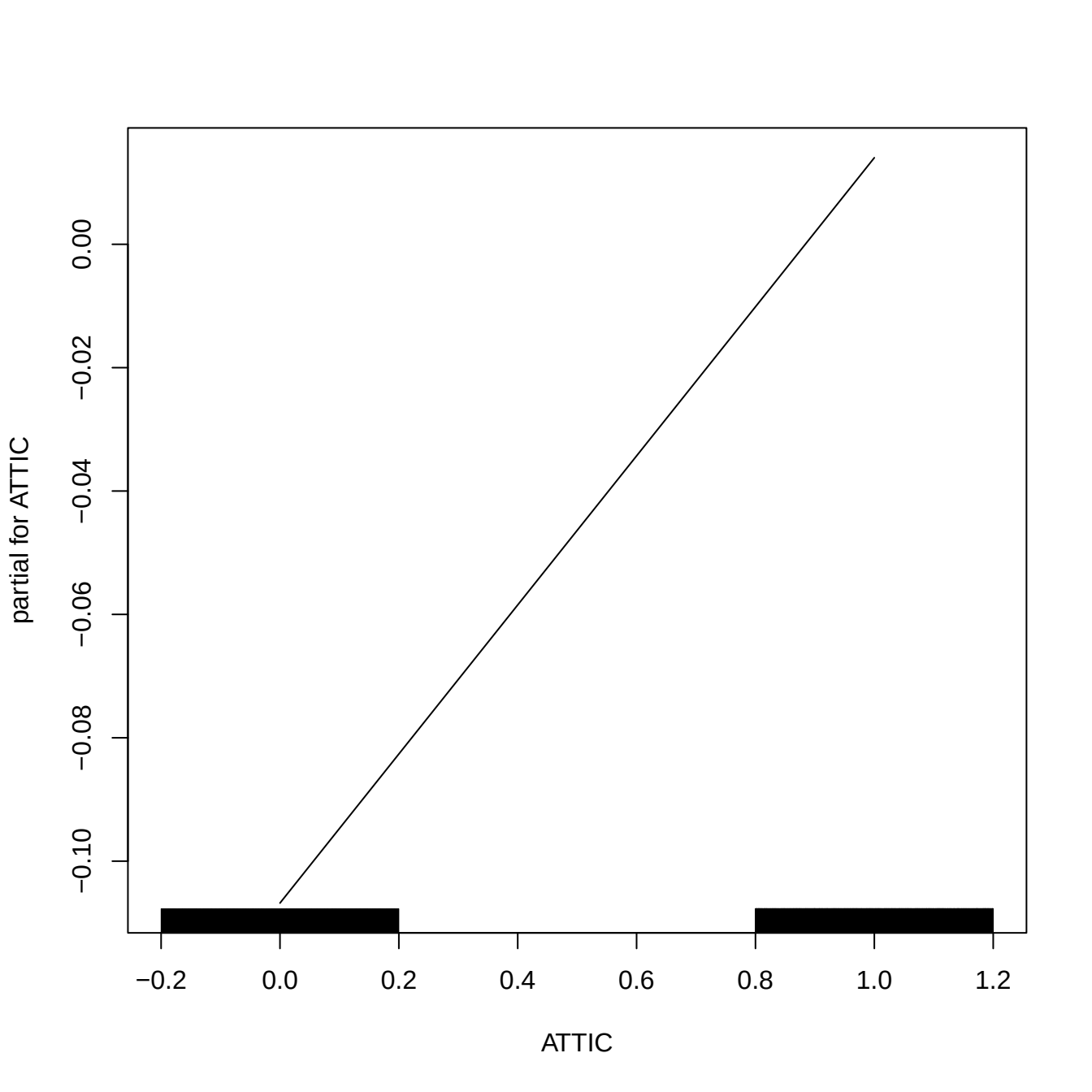

partial for City
0.010
0.005
0.000
−0.005
−0.010
20
40
60
80
100
City

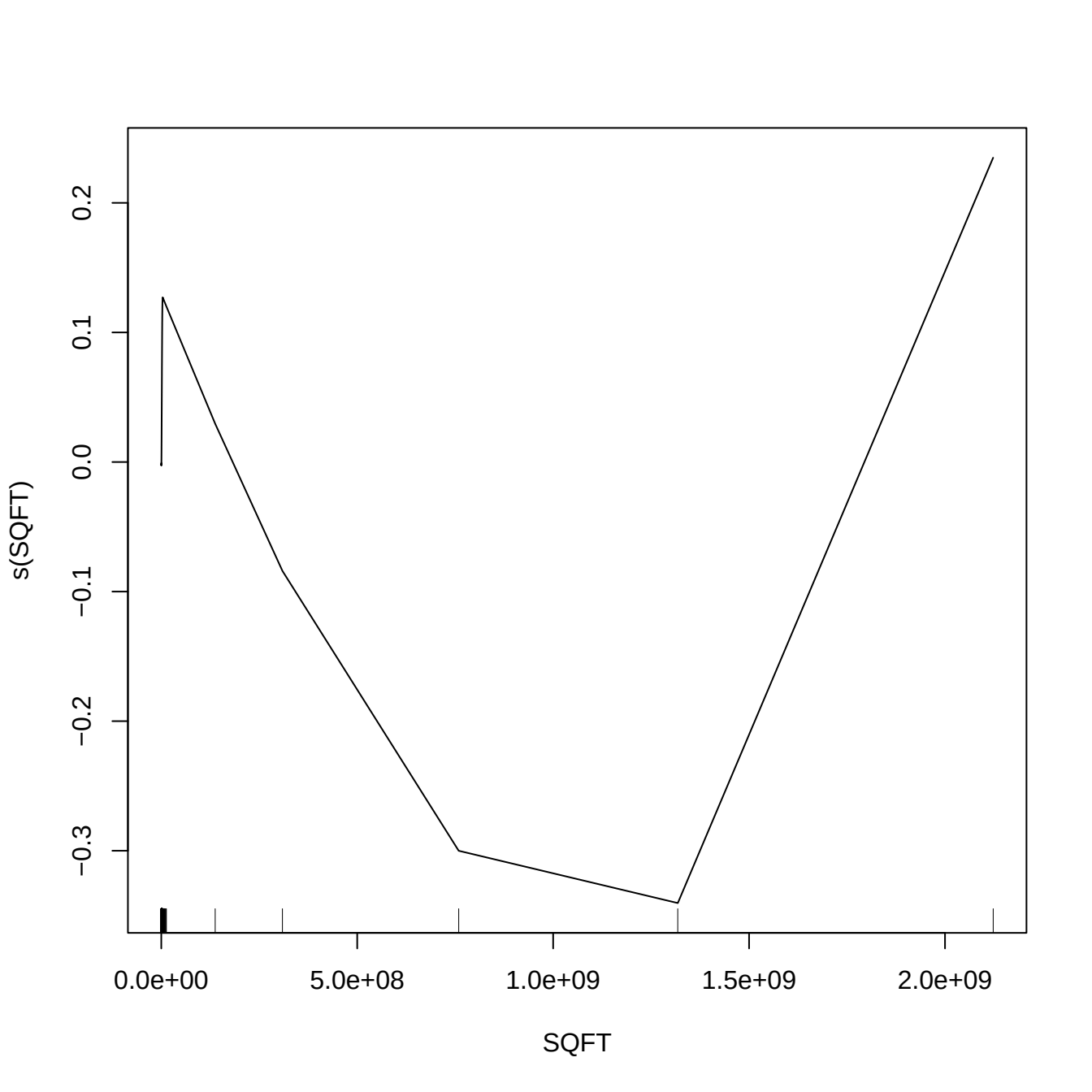

partial for Province
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
1.0
1.5
2.0
2.5
3.0
Province

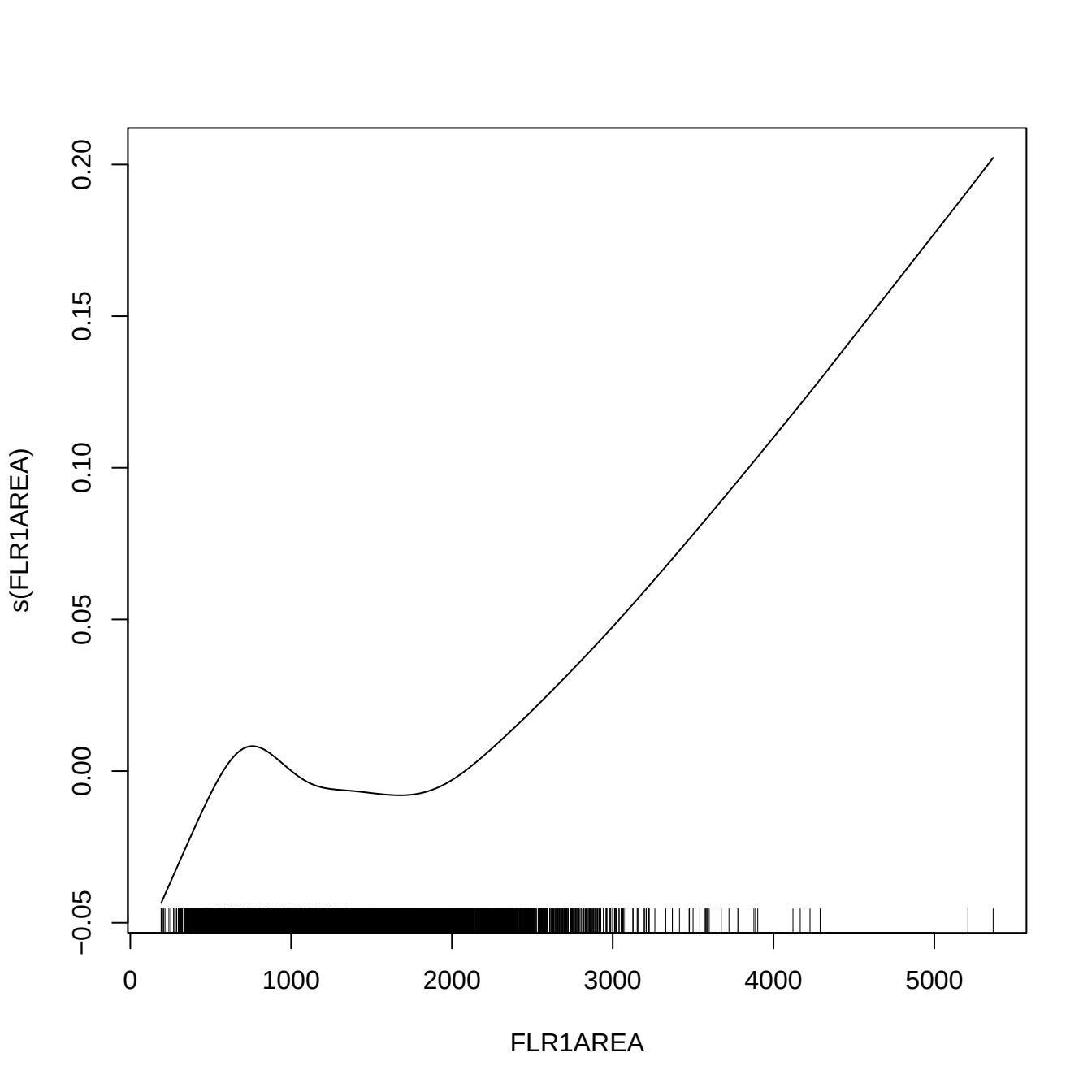

s(RMBED)
0.05
0.00
−0.05
2
4
6
8
10
12
RMBED

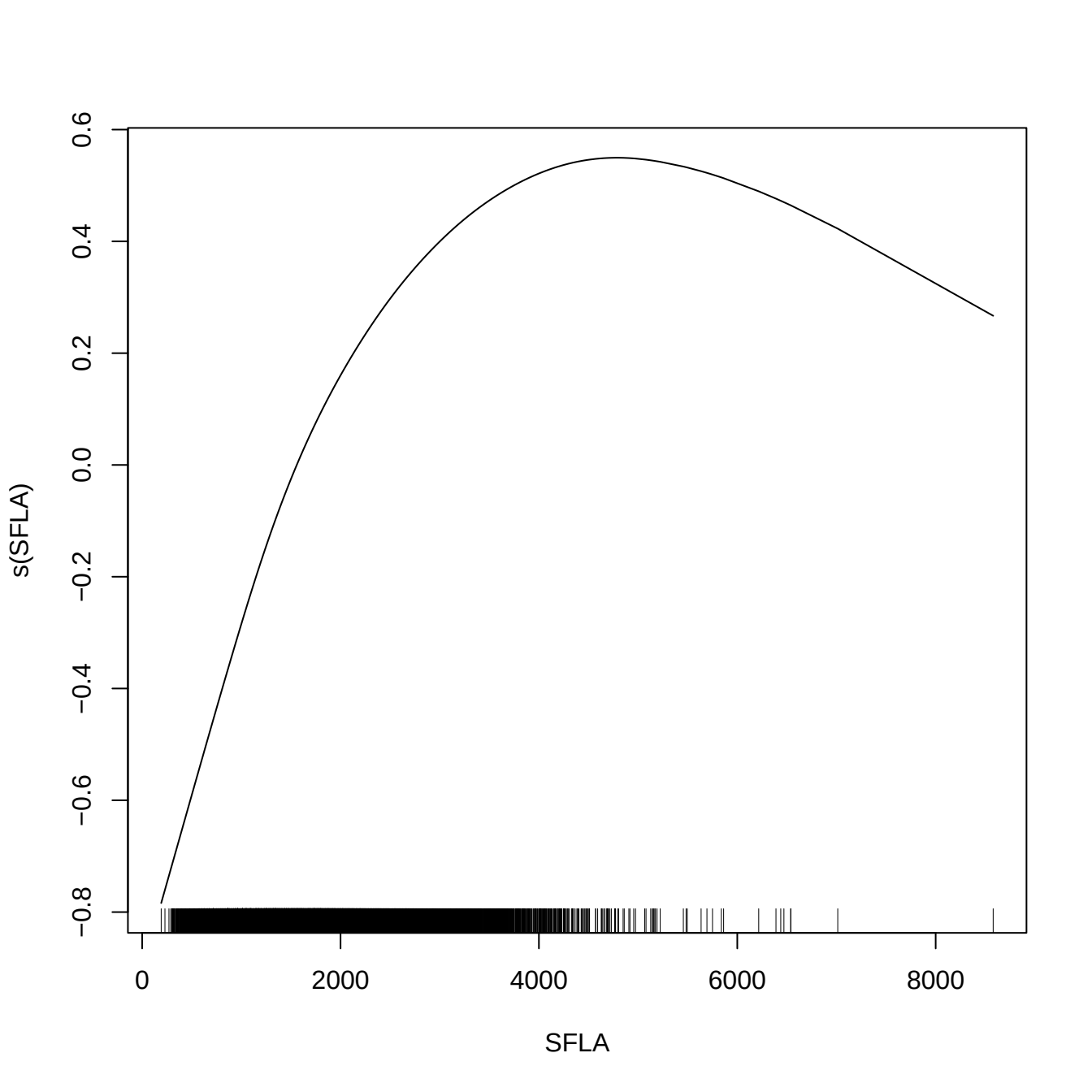

s(BATH)
0.15
0.10
0.05
0.00
−0.05
1
2
3
4
5
6
7
BATH

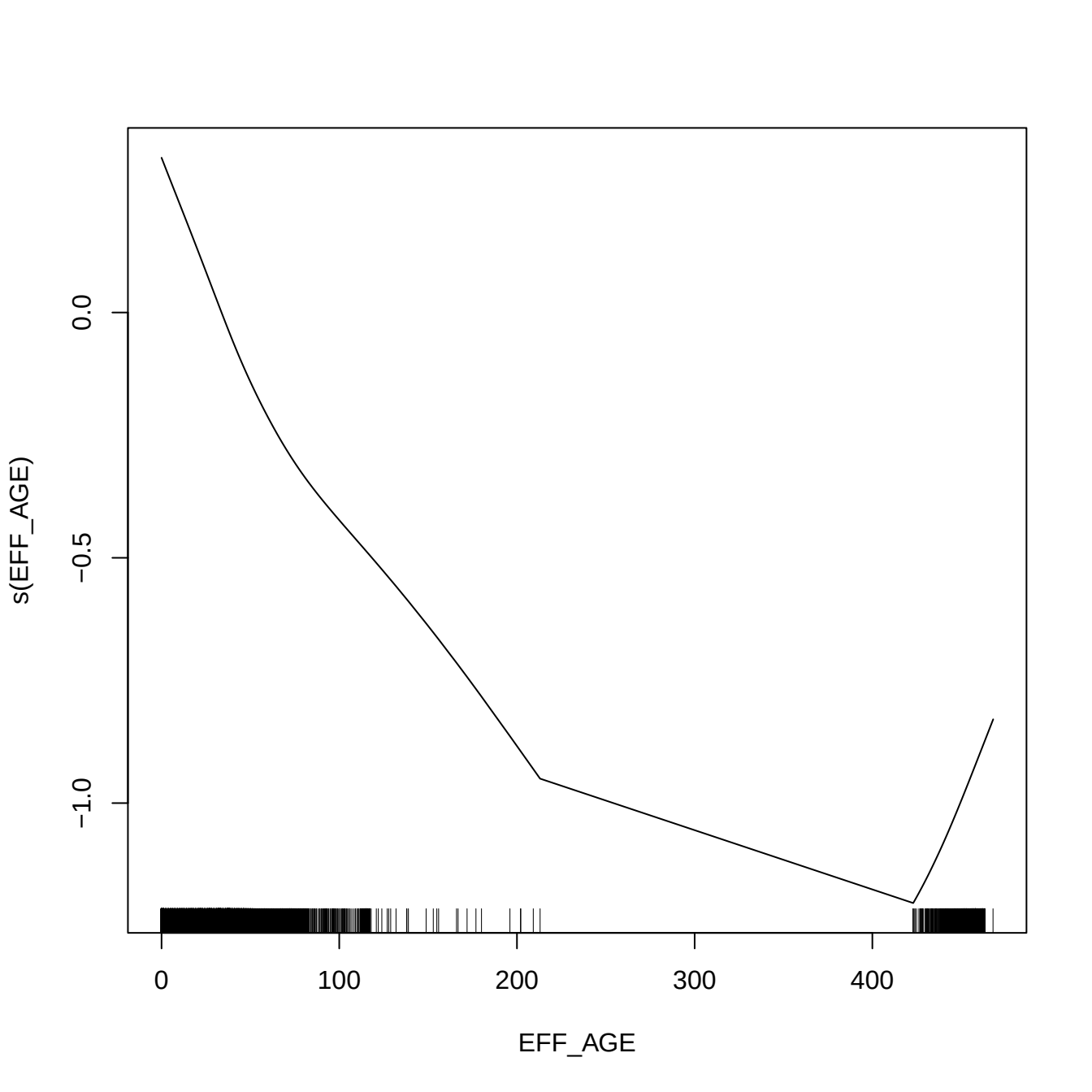

partial for BSMT
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
1
2
3
4
5
6
7
BSMT

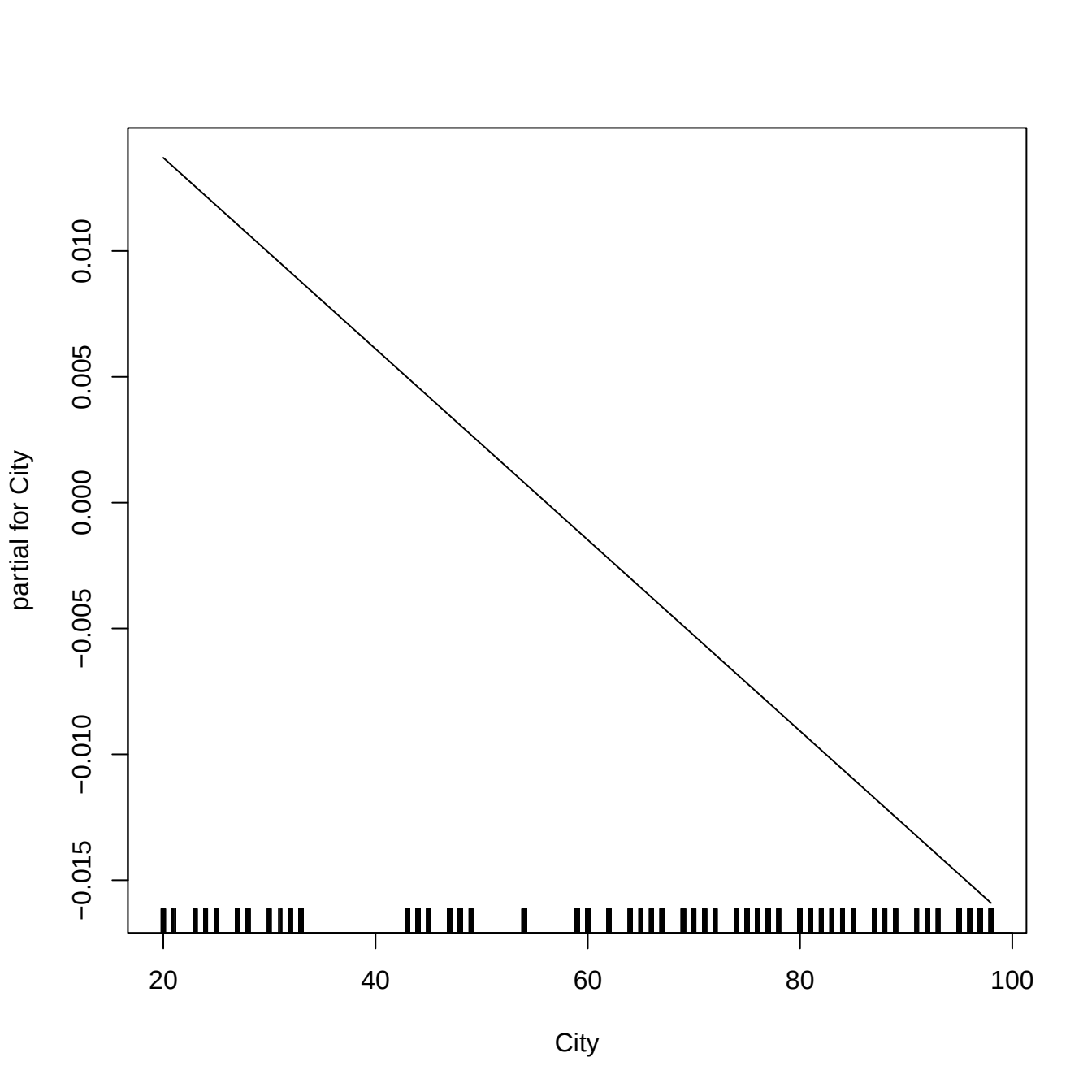

partial for ATTIC
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ATTIC

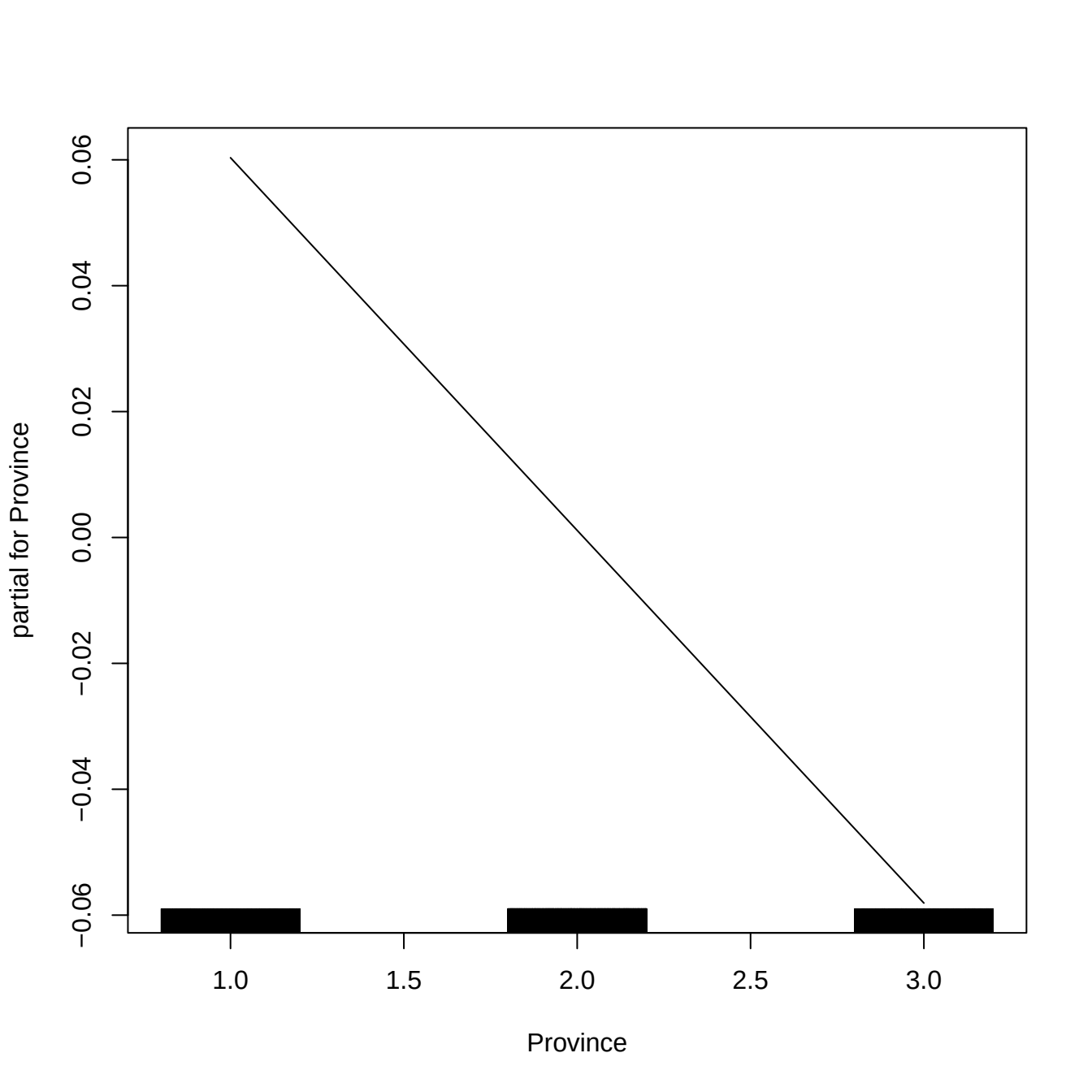

s(SQFT)
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
SQFT

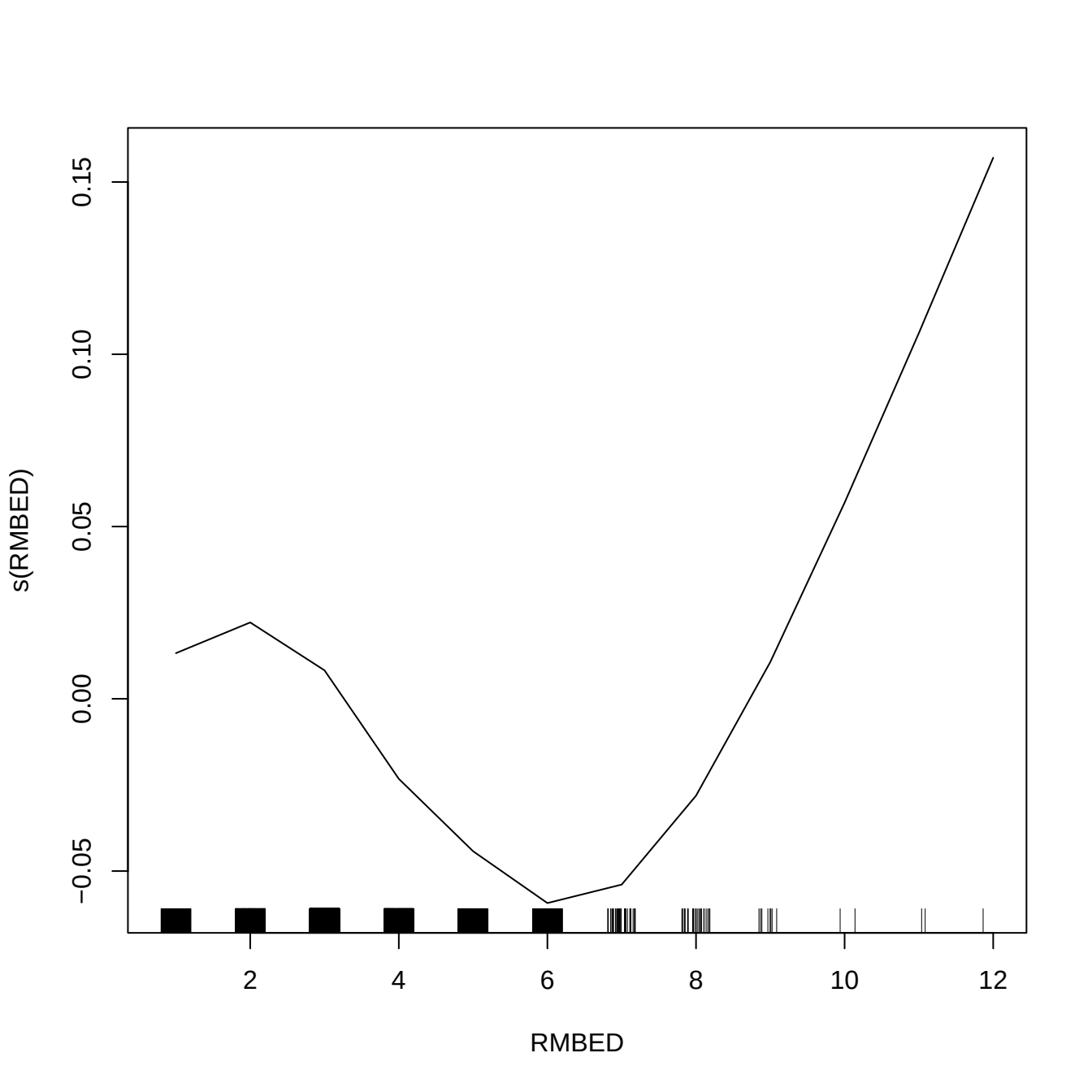

s(FLR1AREA)
0.20
0.15
0.10
0.05
0.00
−0.05
0
1000
2000
3000
4000
5000
FLR1AREA

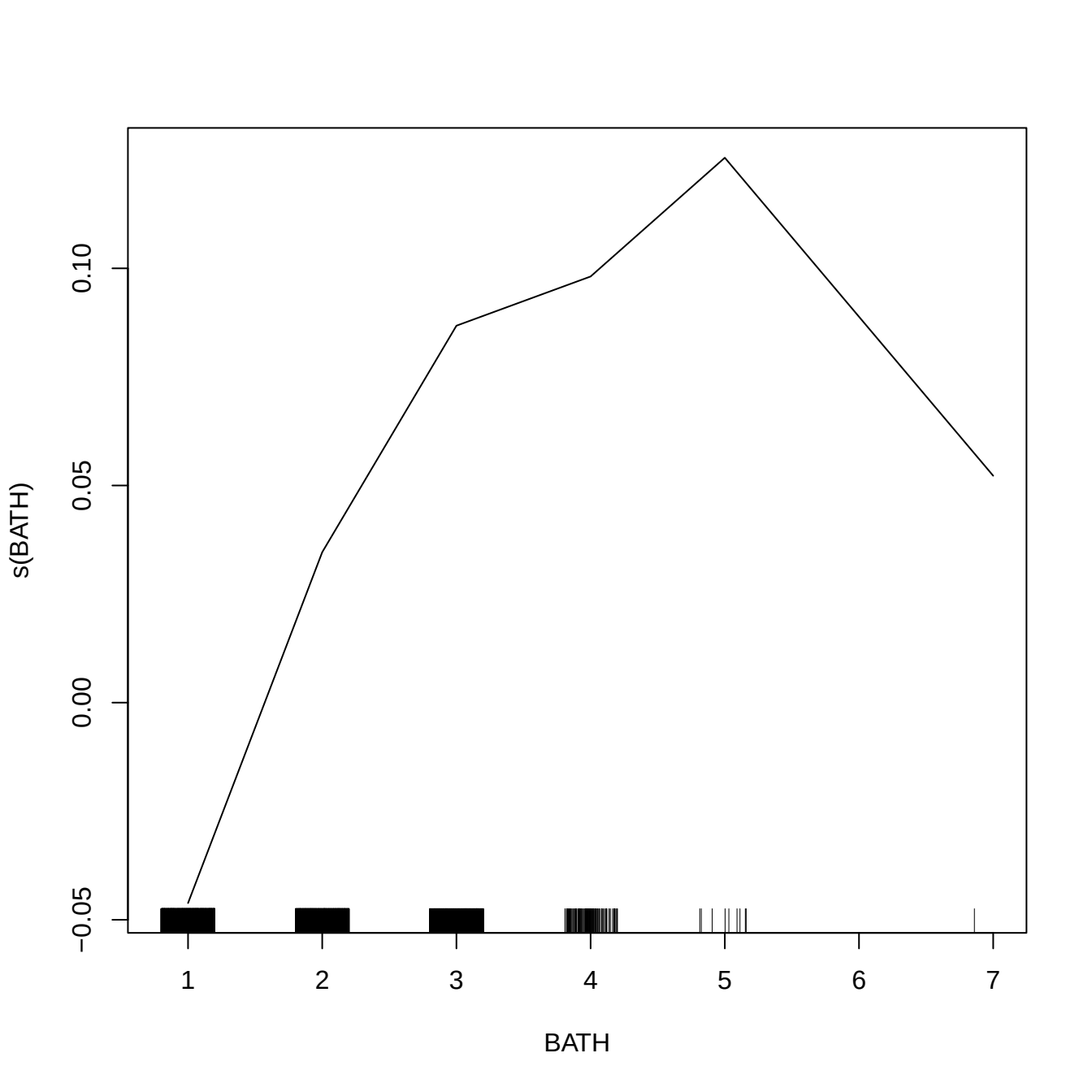

s(SFLA)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
2000
4000
6000
8000
SFLA

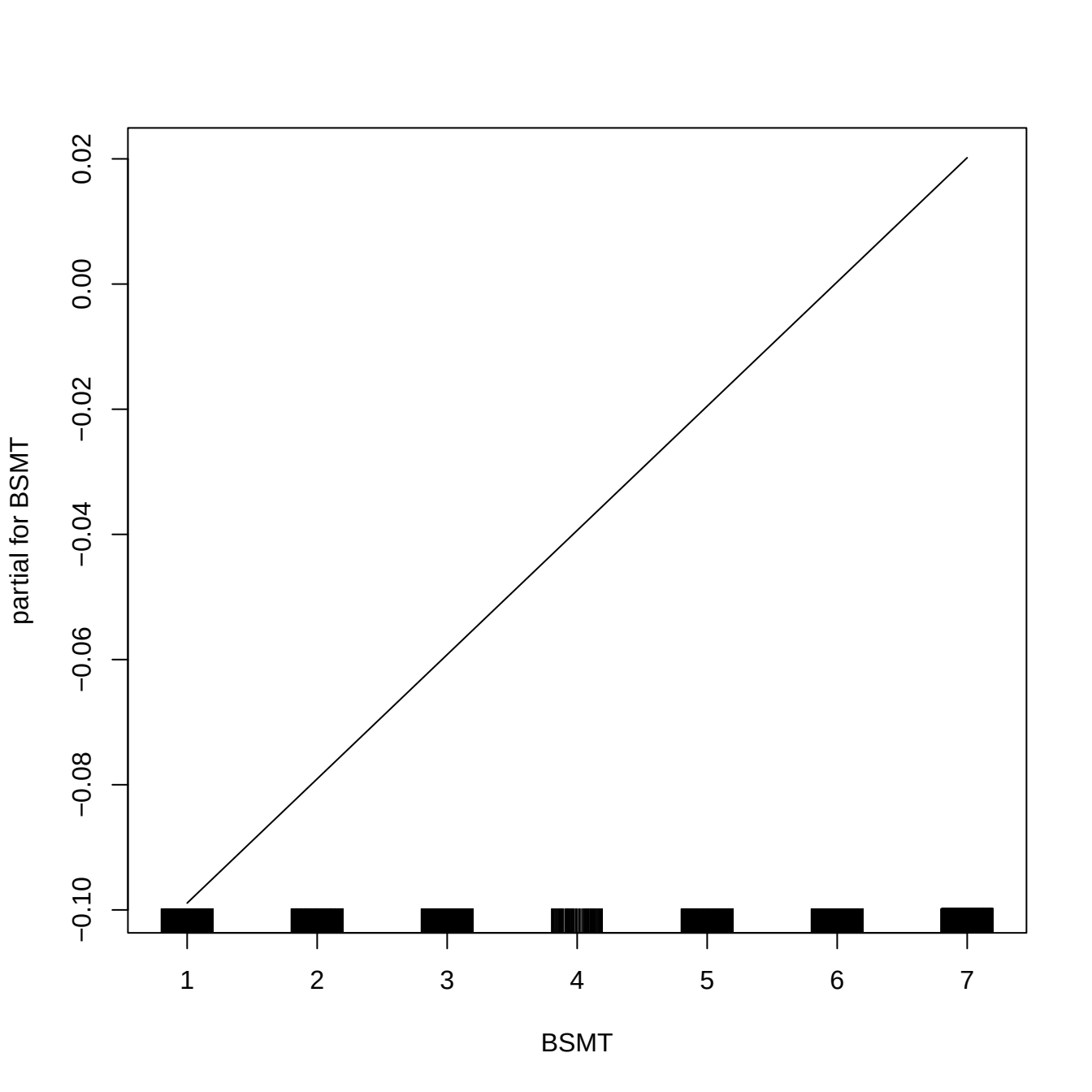

s(EFF_AGE)
0.0
−0.5
−1.0
0
100
200
300
400
EFF_AGE

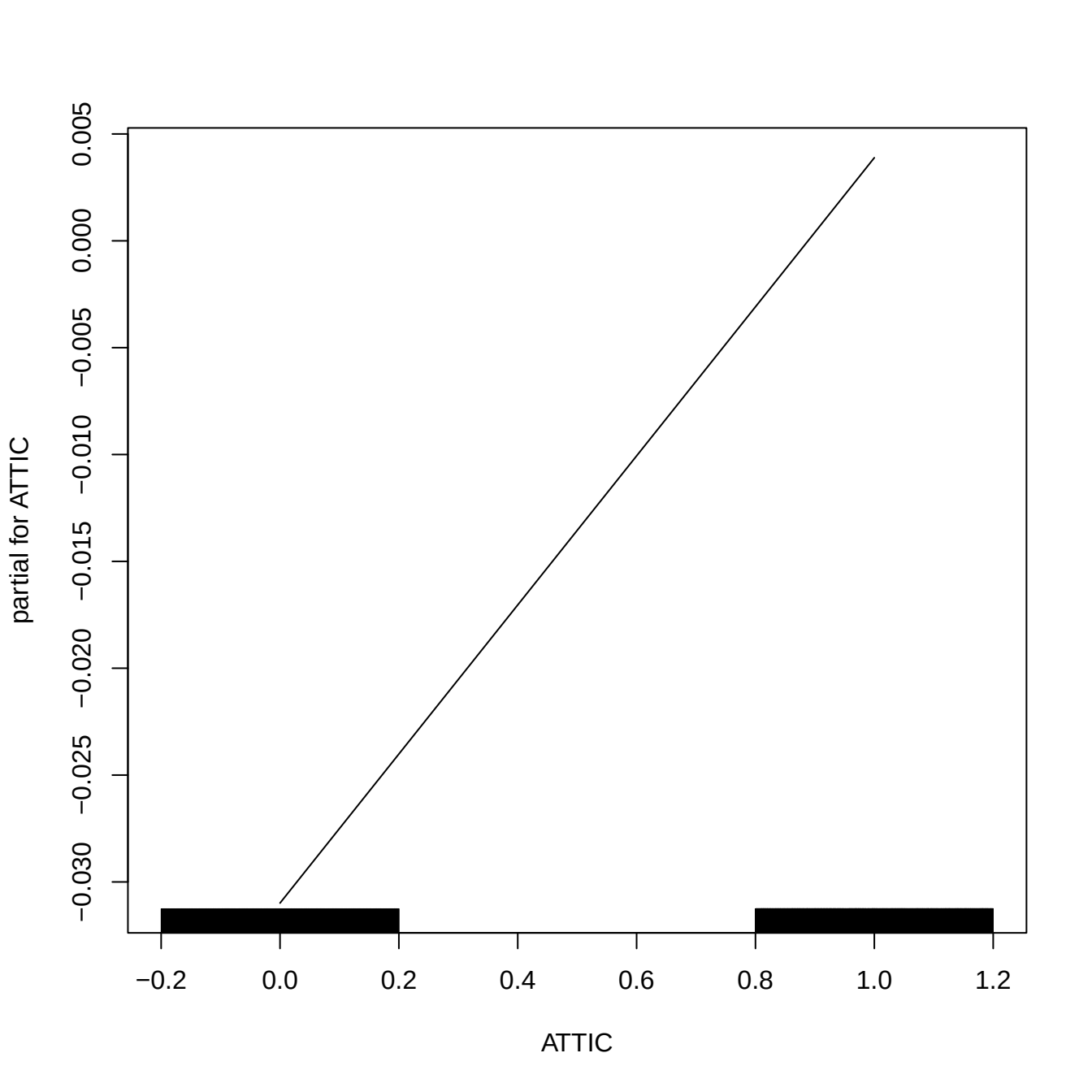

partial for City
0.010
0.005
0.000
−0.005
−0.010
−0.015
20
40
60
80
100
City


partial for Province
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
1.0
1.5
2.0
2.5
3.0
Province


s(RMBED)
0.15
0.10
0.05
0.00
−0.05
2
4
6
8
10
12
RMBED


s(BATH)
0.10
0.05
0.00
−0.05
1
2
3
4
5
6
7
BATH


partial for BSMT
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
1
2
3
4
5
6
7
BSMT


partial for ATTIC
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−0.025
−0.030
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ATTIC