Final project

Statistical methods for data science

Cappiello, Costanzo, De Cleva, Tarchi

Data Science and Scientific Computing

February 27, 2024

Table of Contents

Brief Introduction

The State of Wisconsin Medicaid program funds nursing home care for individuals qualifying on the basis of need and financial status. As part of the conditions for participation, Medicaid-certified nursing homes must file an annual cost report, summarizing the volume and cost of care provided to all of its residents, Medicaid funded and otherwise. These cost reports are audited and form the basis for facility-specific Medicaid daily payment rates for subsequent periods.

Data

The data here is in the cost report years 2000 and 2001. There are 362 facilities in 2000 and 355 facilities in 2001. Typically, utilization of nursing home care is measured in patient days ("patient days" is the number of days each patient was in the facility, summed over all patients).

Objective

The aim of this project is to develop a predictive model that provides a reliable utilization forecast to update the Medicaid funding rate schedule of nursing facilities.

Table of Contents

Information about the variables

File Name:WiscNursingHome	Number of obs: 717	Number of variables: 12
Variable (type of variable)	**Number of Obs Missing**	**Description**
hospID (categorical)		Hospital identification number
CRYEAR (categorical binary)		Cost report year
TPY (quantitative continuous)		Total patient years
NUMBED (quantitative discrete)		Number of beds
SQRFOOT (quantitative continuous)	10	Square footage of the nursing home
MSA (categorical)		Metropol Statist Area code, 1-13, 0 for rural
URBAN (categorical Binary)		1 if urban, 0 if rural
PRO (categorical Binary)		1 if for profit, 0 for non-profit
TAXEXEMPT (categorical Binary)		1 if tax-exempt
SELFFUNDINS (categorical Binary)		1 if self-funded for insurance
MCERT (categorical Binary)		1 if Medicare certified
ORGSTR (categorical)		1 for profit, 2 for tax-exempt,3 for gov unit

Since the dataset only have 10 missing values in the variable SQRFOOT (1,4 perc. of all obs.), we decided to delete this rows.

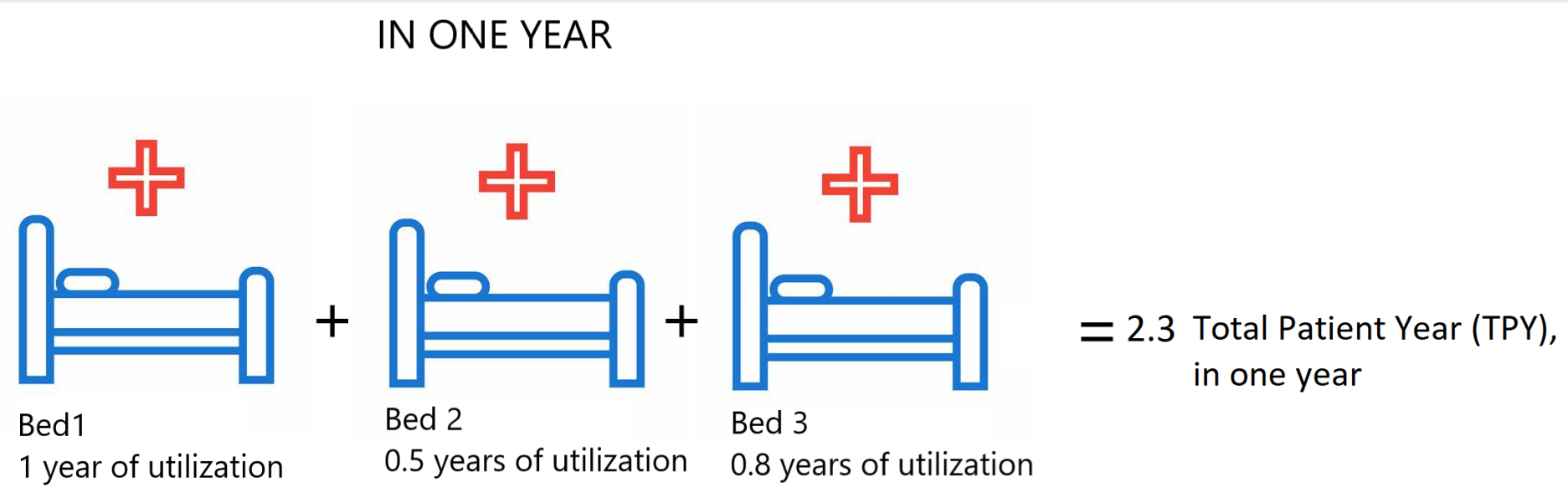
Response variable

Formula for TPY by nursing house

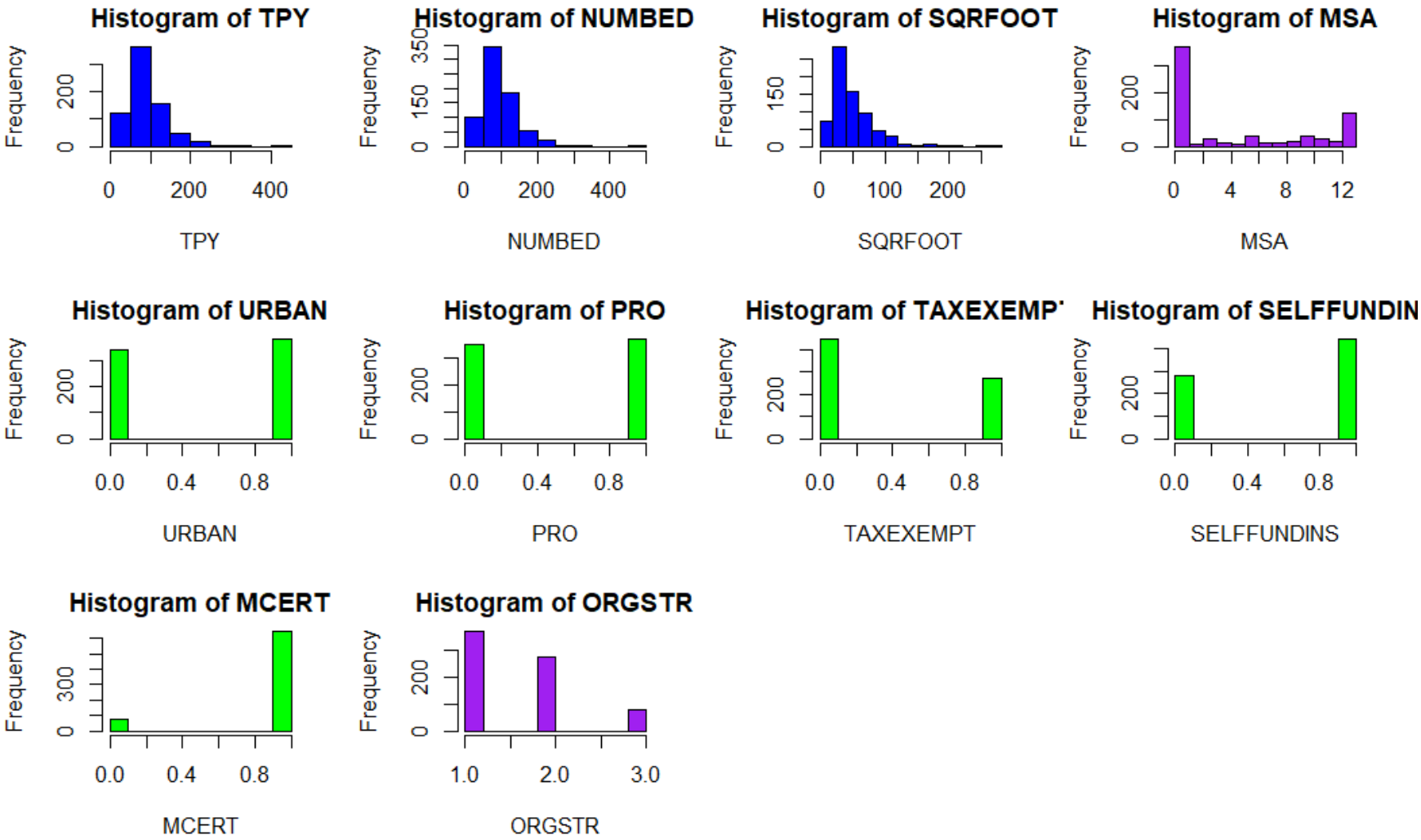
calculated in this equation

$$\text{Total patients years in one year (TPY)} = \sum_{i}^{n} (\text{Bed time utilization in one year}) \tag{1}$$

Example



Analyzing variables' distributions

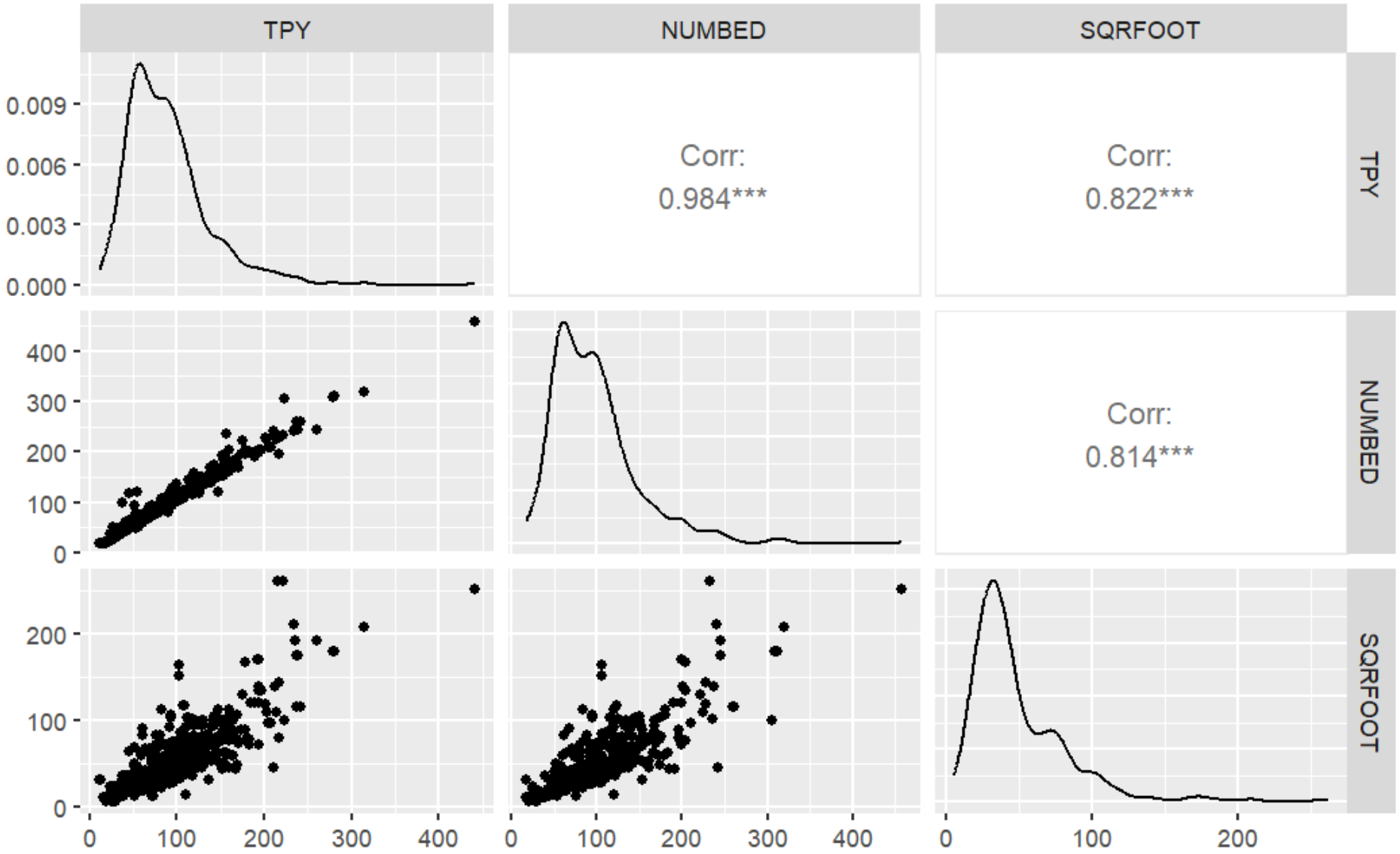


For categorical variables, we can observe that they are unbalanced. On the other hand, quantitative variables seem to be right-skewed.

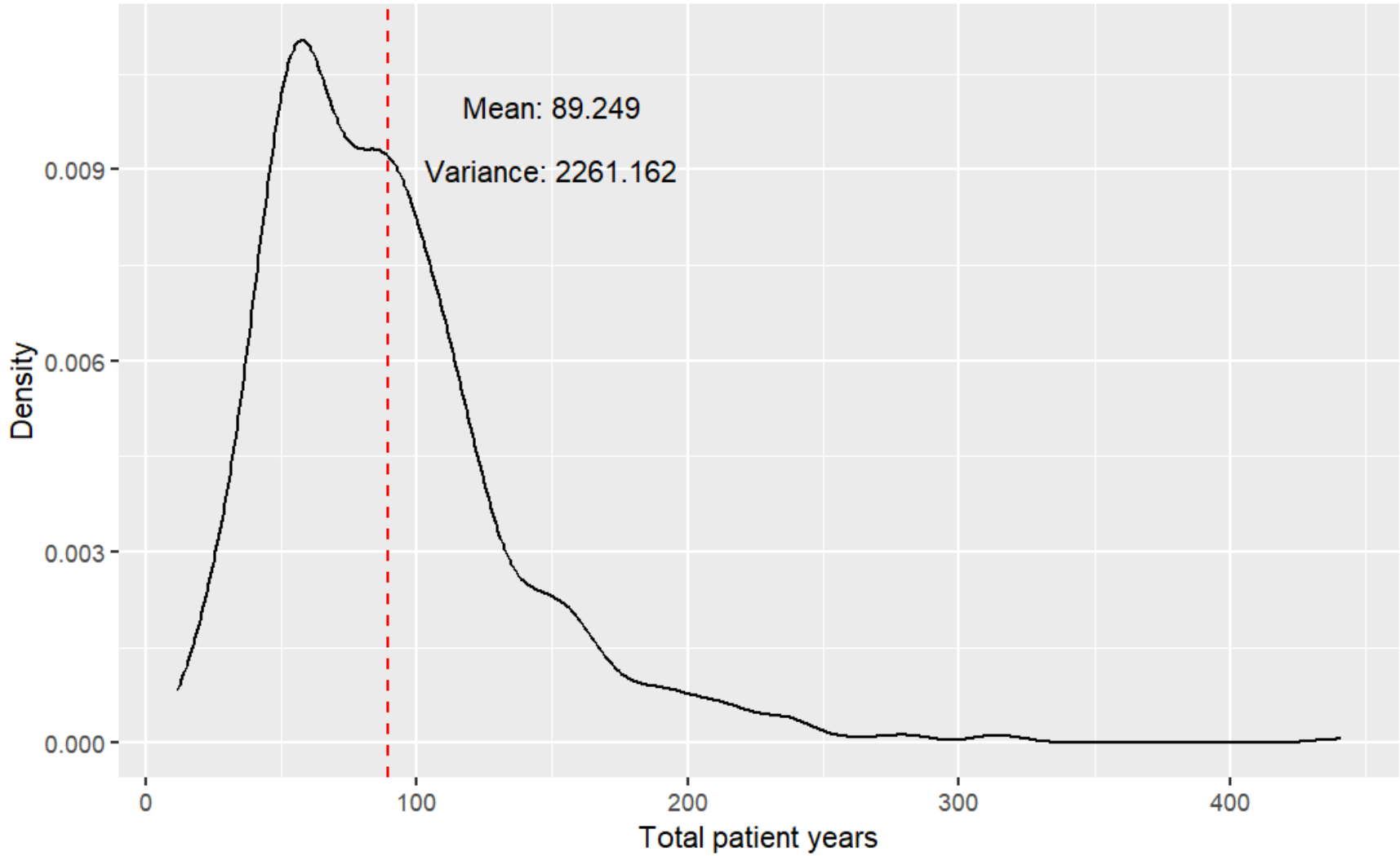
Statistics summary for quantitative variables

```
      TPY              SQRFOOT           NUMBED
Min.   : 11.57   Min.   :  5.644   Min.   : 18.0
1st Qu.: 56.71   1st Qu.: 28.638   1st Qu.: 60.0
Median : 80.93   Median : 39.883   Median : 90.0
Mean   : 89.25   Mean   : 50.257   Mean   : 97.2
3rd Qu.:109.12   3rd Qu.: 64.281   3rd Qu.:119.0
Max.   :440.67   Max.   :262.000   Max.   :457.0
                 NA's   :10
```

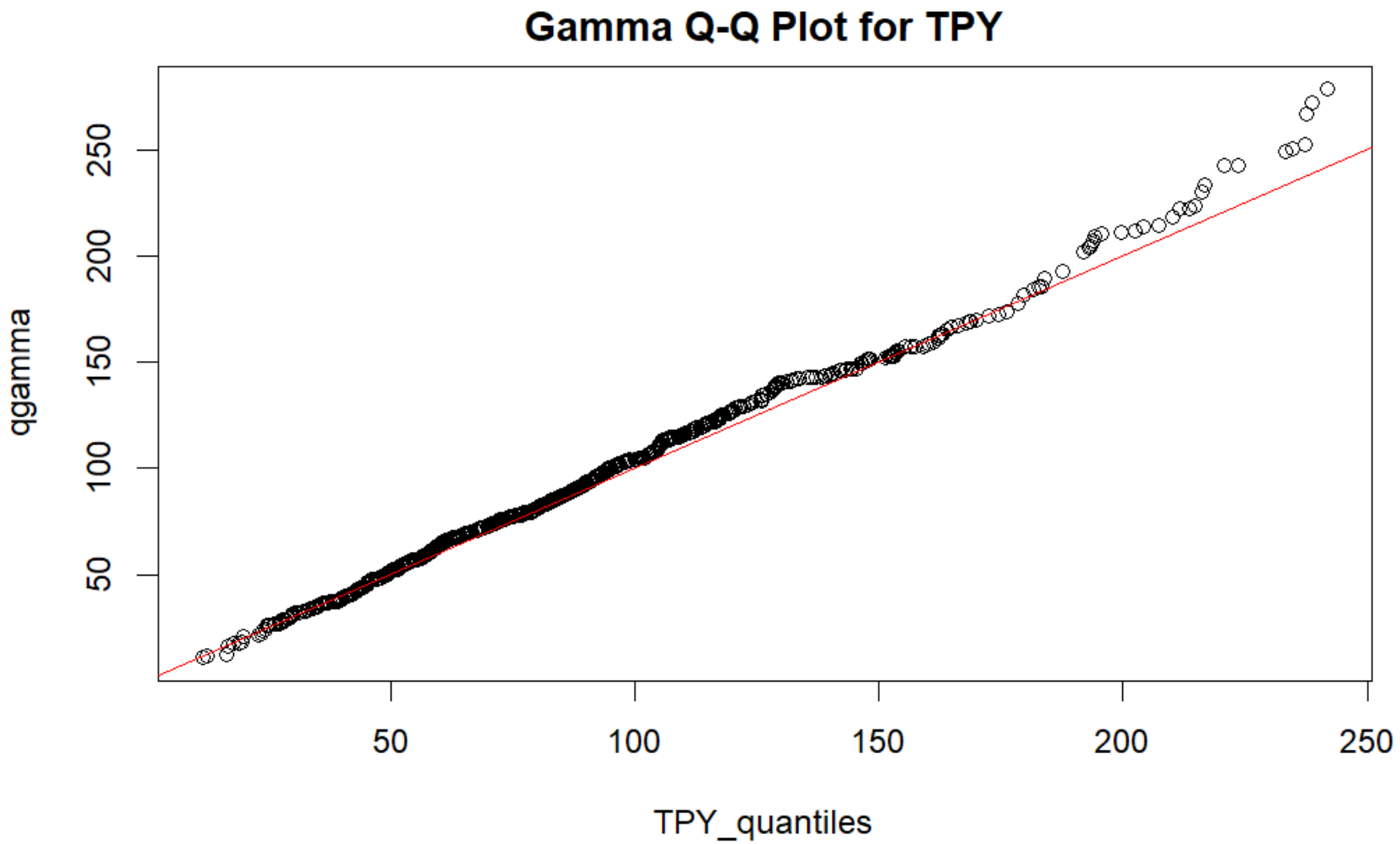
Quantitative variables' pair plots



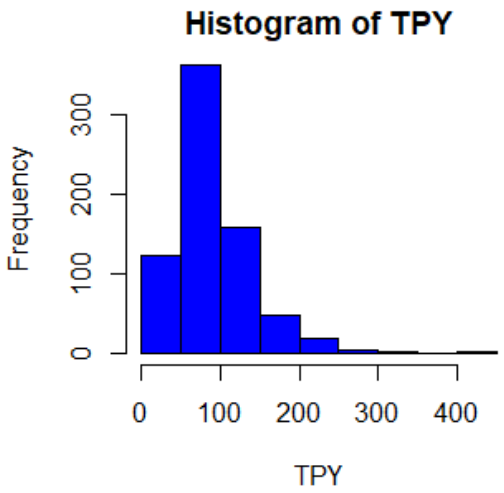
Hypothesis about response variable distribution

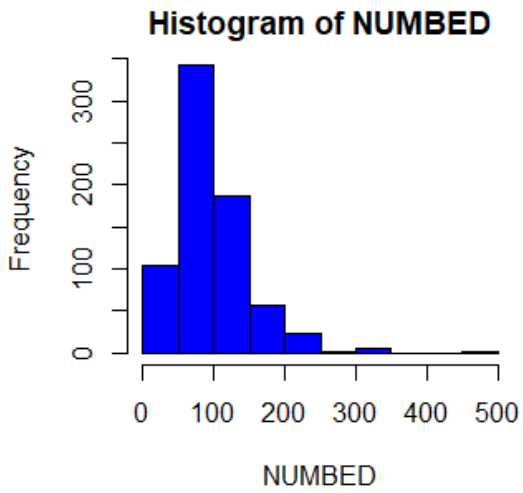


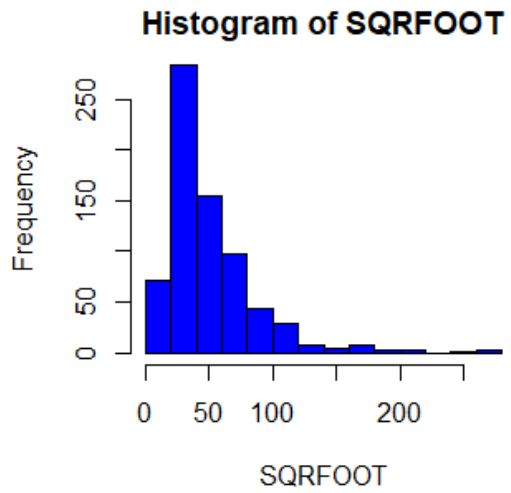
Since this is a continuous variable with a right-skewed distribution and a large difference between its mean and variance, we hypothesized that this variable follows a Gamma distribution

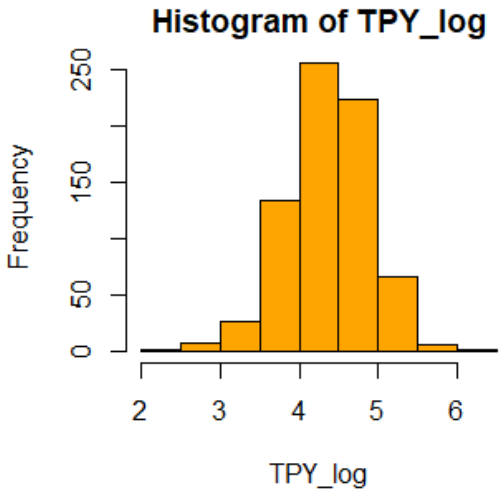

Gamma Q-Q Plot for TPY
qgamma
50
100
150
200
250
TPY_quantiles
50
100
150
200
250

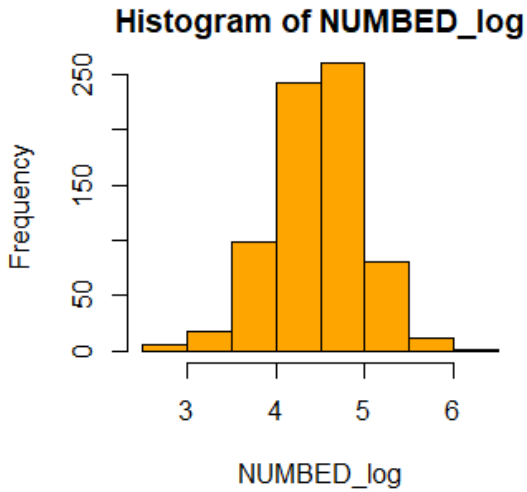
Log transformation of the quantitative variables.

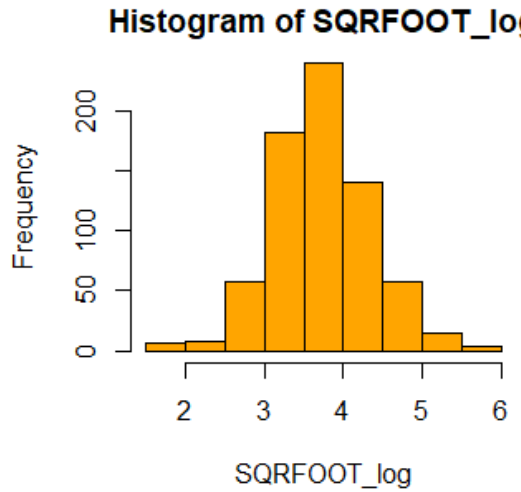




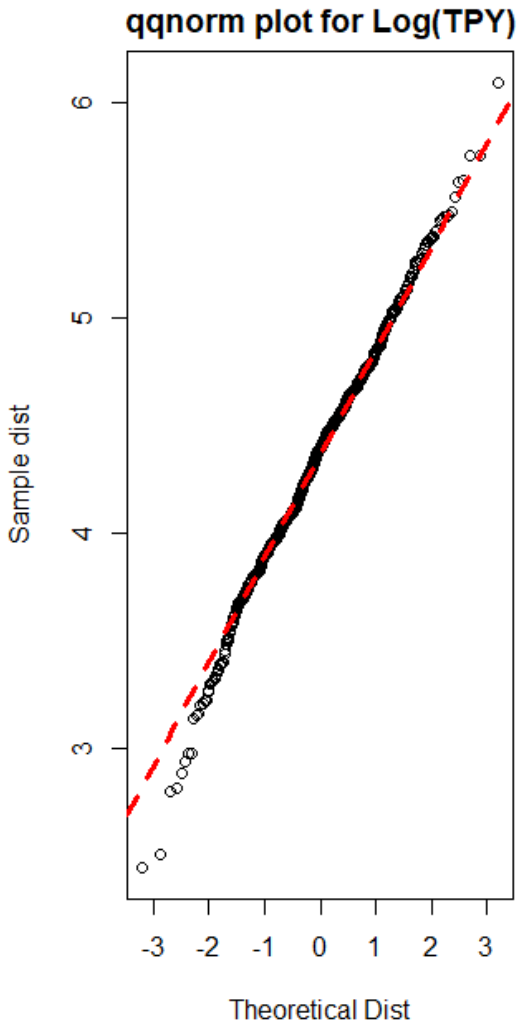


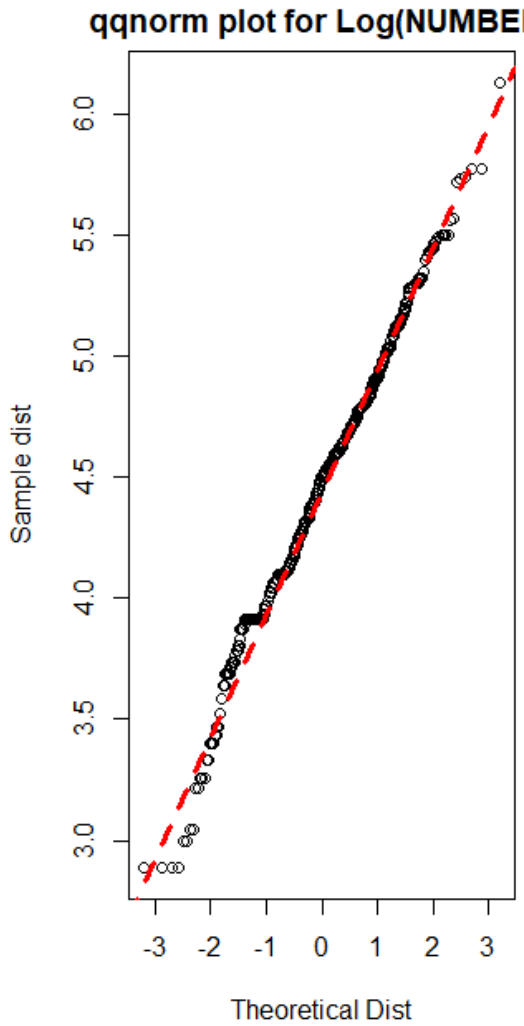


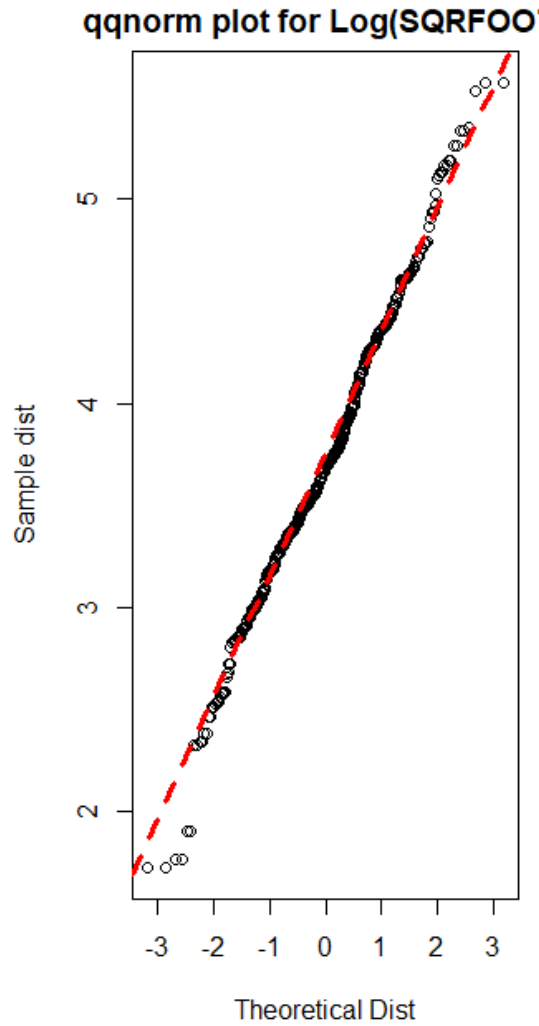


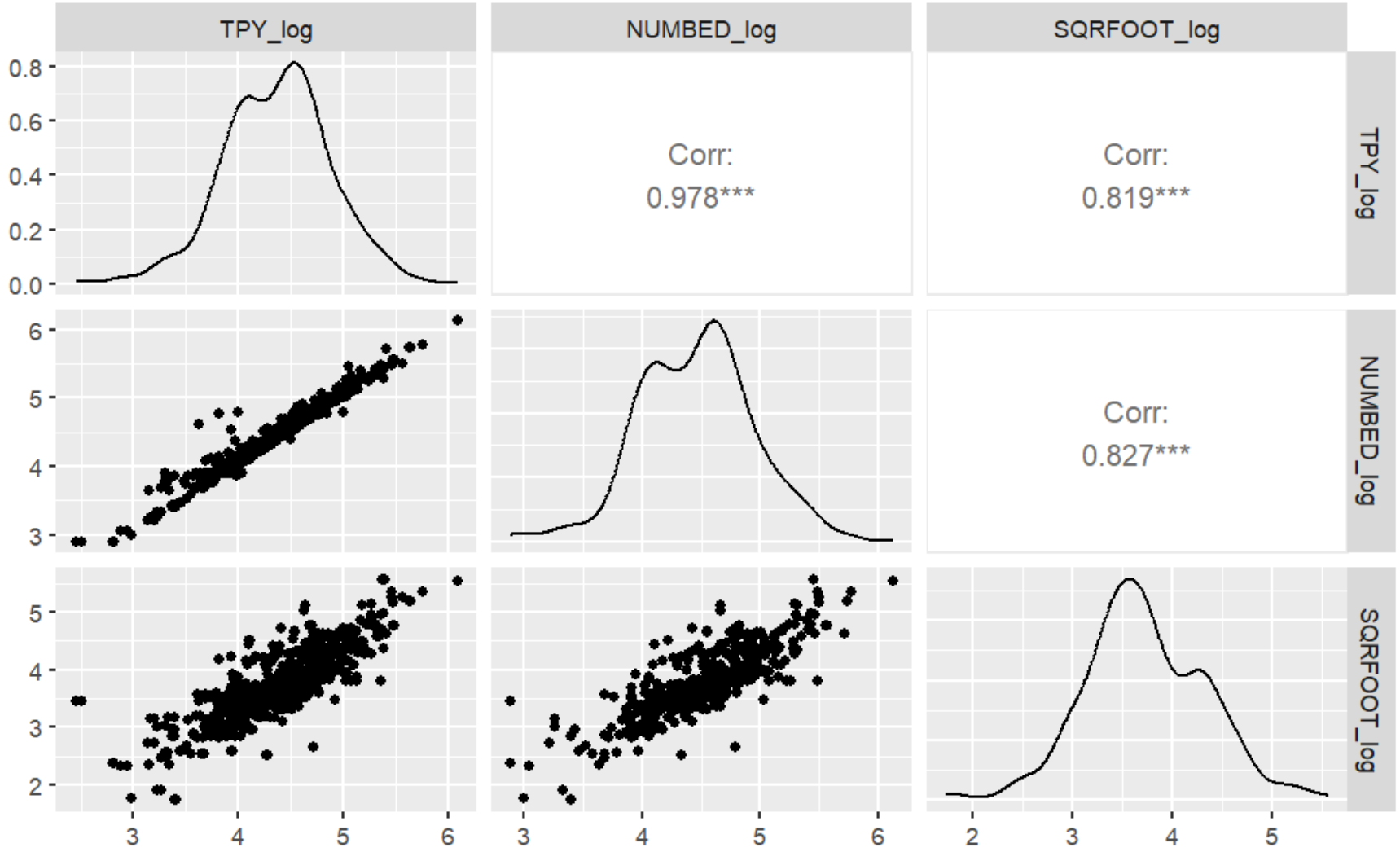


Normality evaluation of the log trasf. distrib.

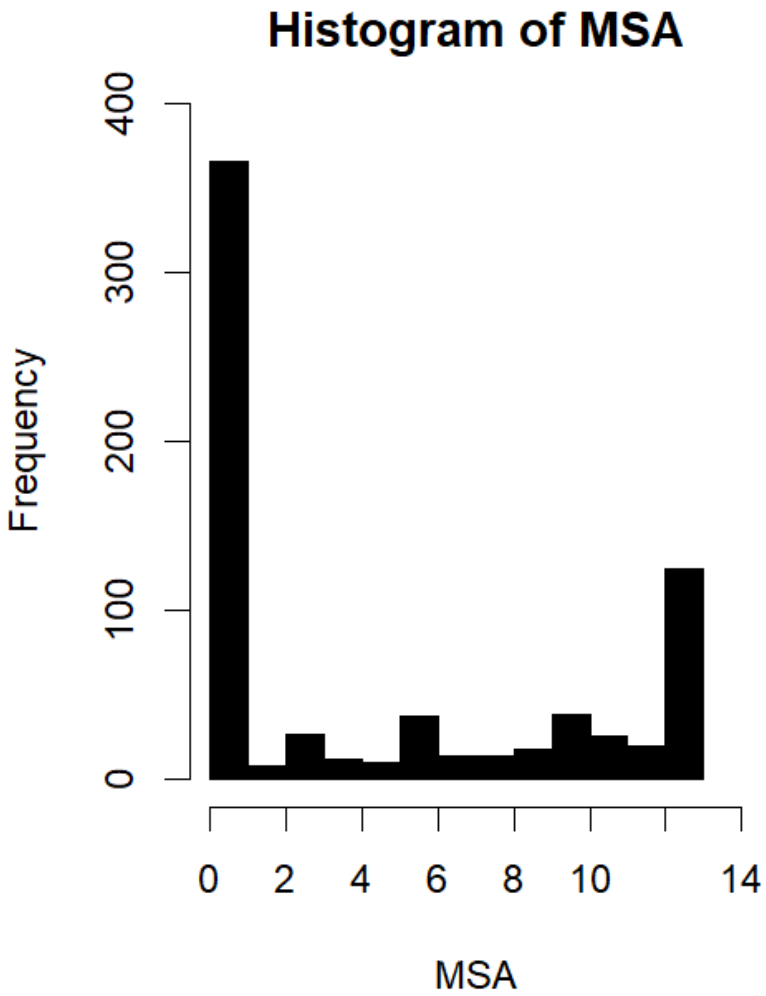


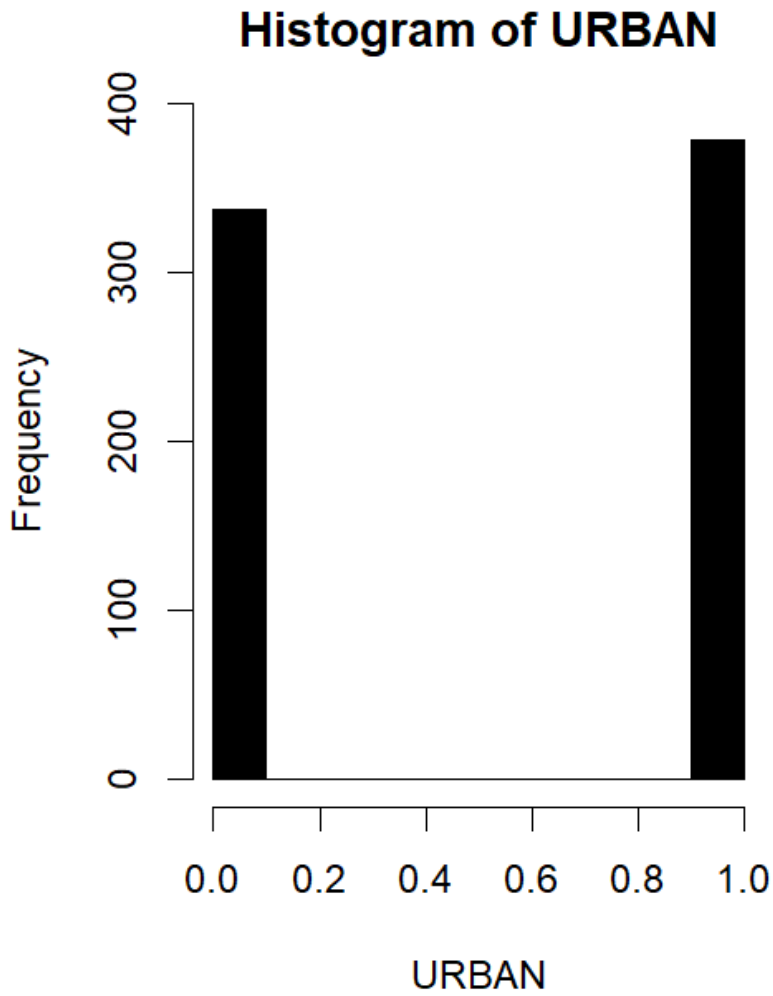





TPY_log
NUMBED_log
SQRFOOT_log
Corr:
0.978***
Corr:
0.819***
Corr:
0.827***
TPY_log
NUMBED_log
SQRFOOT_log
0.8
0.6
0.4
0.2
0.0
6
5
4
3
5
4
3
2
3
4
5
6
3
4
5
6
2
3
4
5

Analysis of categorical variables

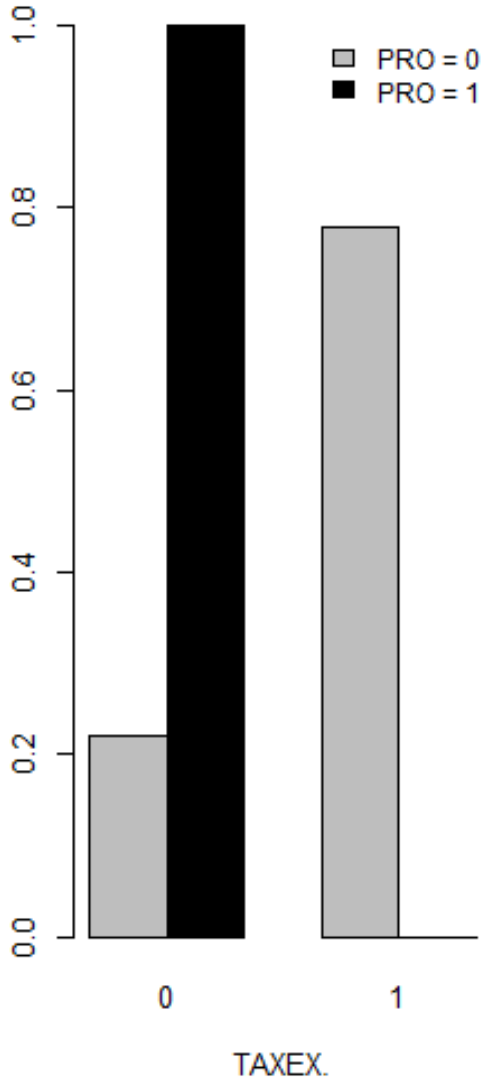


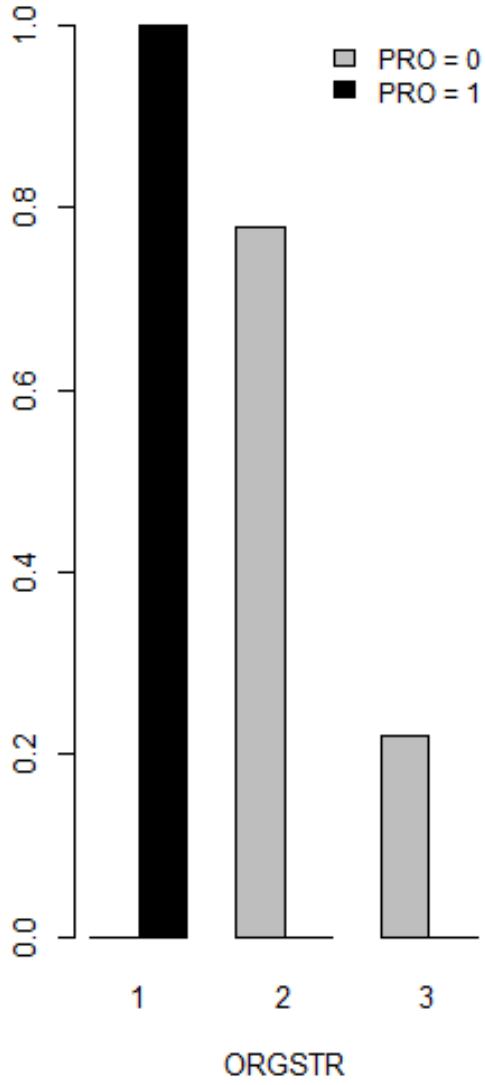


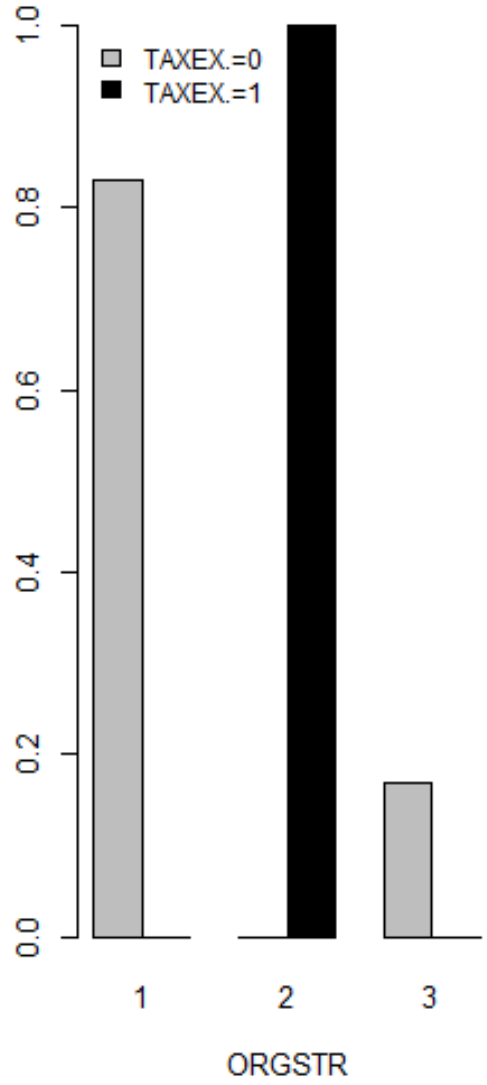
DATA REDUNDANCY: MSA the observations are concentrated in levels 0 (rural) and 13 (Metropol. Statist. Area code). As we can see the variable URBAN shows the same information.

We observe that variables TAXEXEMPT, ORGSTR and PRO carry similar information. The ORGSTR, in particular, is not informative if we consider TAXEXEMPT and PRO combined.

Relation between ORGSTR, PRO, and TAXEXEMPT







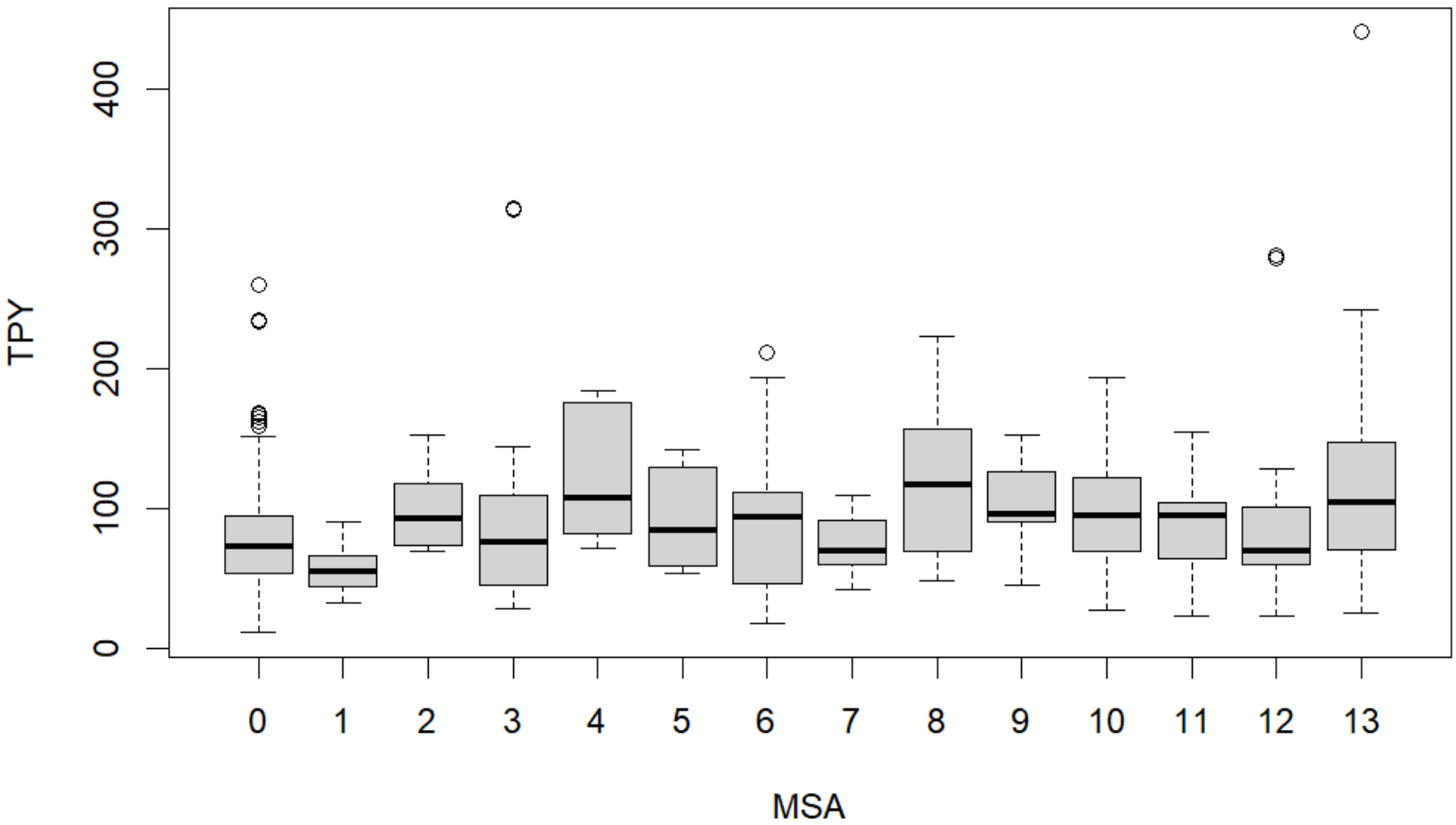
Relation between ORGSTR, PRO, and TAXEXEMPT summary

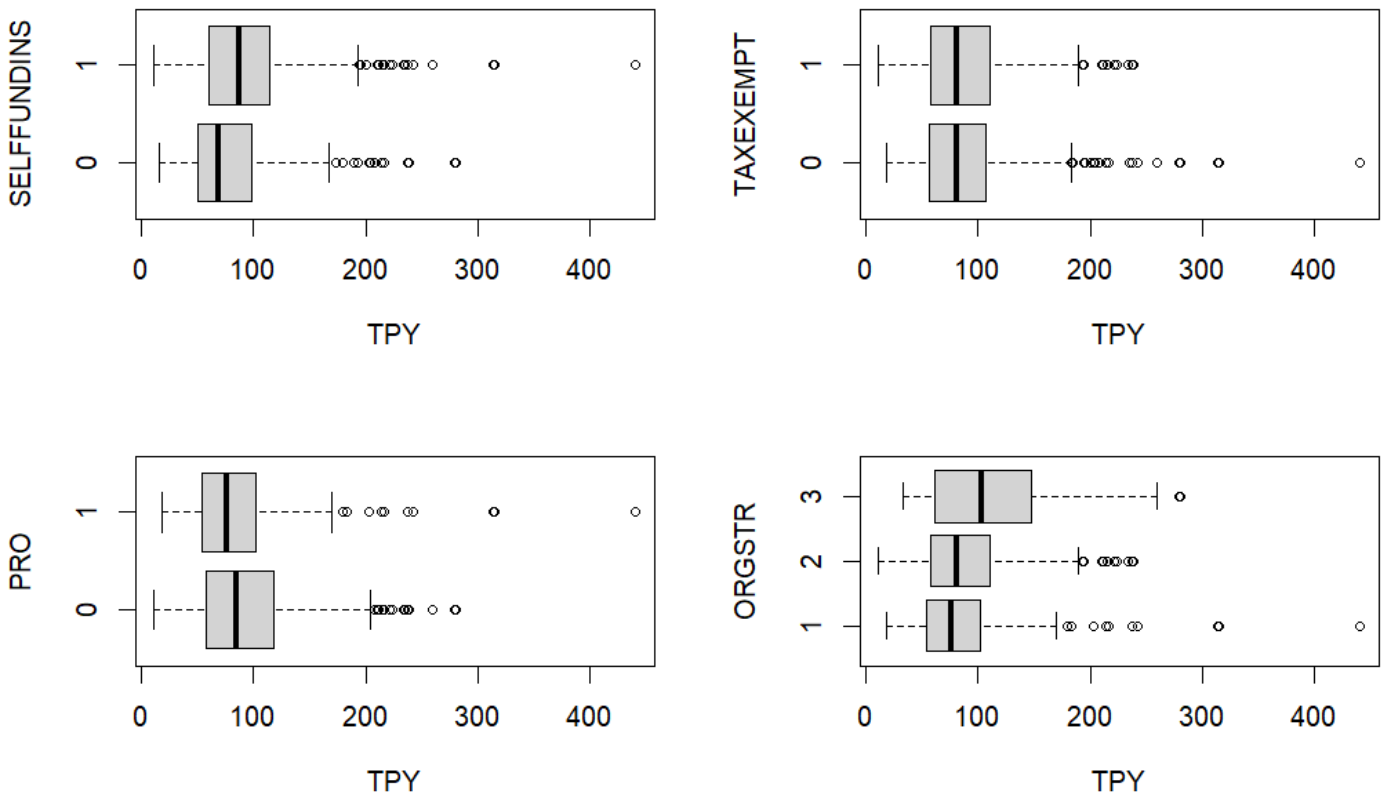
$$ORGSTR = 1 \approx (PRO = 1 \ \ and \ \ TAXEXEMP = 0)$$

$$ORGSTR = 2 \approx (PRO = 0 \ \ and \ \ TAXEXEMP = 1)$$

$$ORGSTR = 3 \approx (PRO = 0 \ \ and \ \ TAXEXEMP = 0)$$

Distribution of the response var. through categorical var.





From the boxplots, we can see that there are no significant differences in the TPY distribution over the categorical variables.

Table of Contents

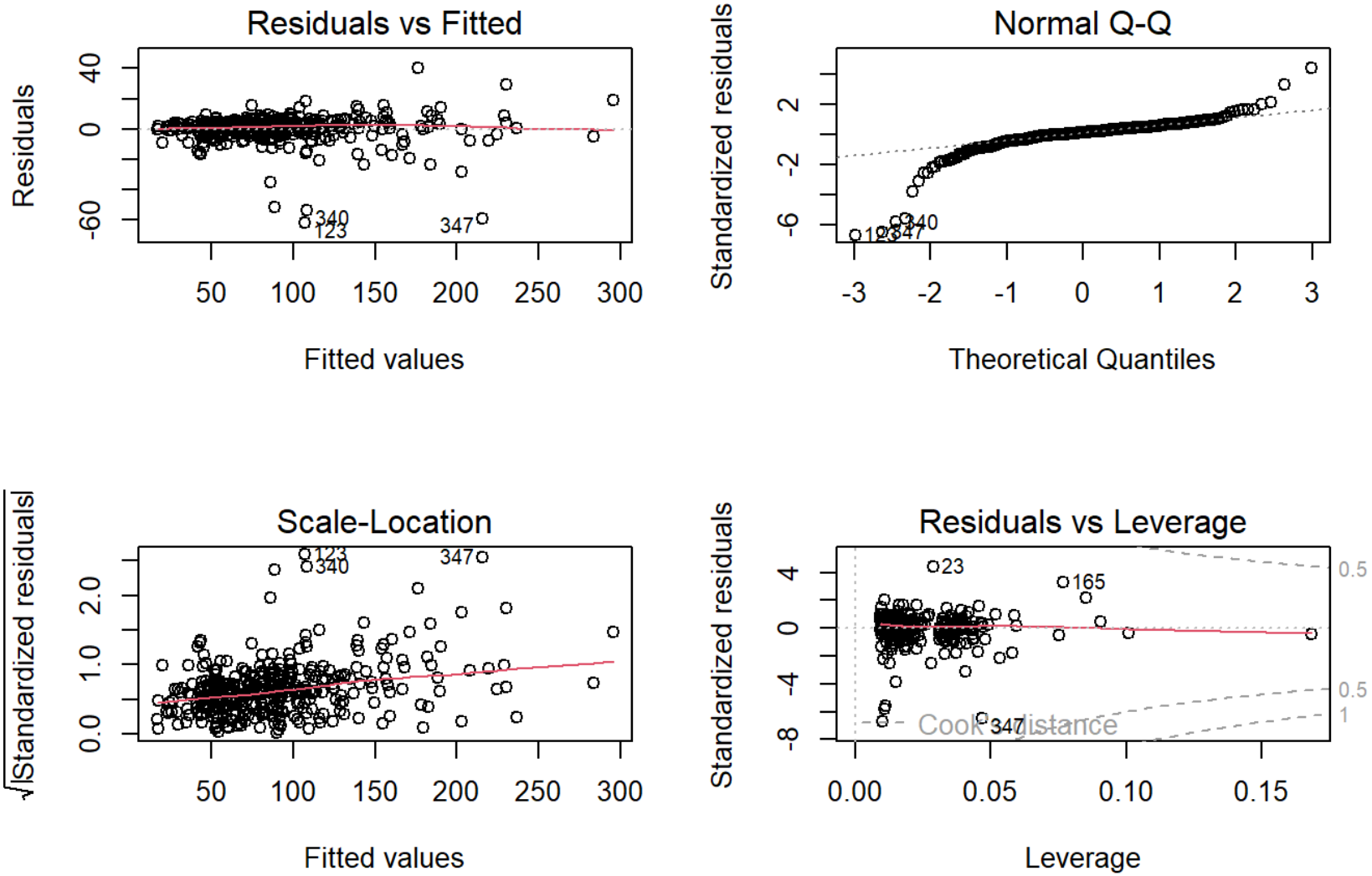
Model with all the variables

To ensure independence between observations we only used data from the year 2000. As a starting point, let's fit a linear model with all the variables:

```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT + PRO + URBAN + SELFFUNDINS +
##     MCERT + TAXEXEMPT, data = data_2000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.873  -2.179   1.245   3.998  40.282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.64970    2.38553  -0.692  0.48968
## NUMBED       0.87792    0.01827  48.041  < 2e-16 ***
## SQRFOOT      0.07677    0.02682   2.863  0.00445 **
## PRO         -0.26603    1.75470  -0.152  0.87958
## URBAN       -1.42231    1.03244  -1.378  0.16920
## SELFFUNDINS  0.54617    1.03711   0.527  0.59879
## MCERT        1.37646    1.73630   0.793  0.42846
## TAXEXEMPT    1.91295    1.74329   1.097  0.27326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.263 on 349 degrees of freedom
## Multiple R-squared:  0.9605, Adjusted R-squared:  0.9597
## F-statistic:  1212 on 7 and 349 DF,  p-value: < 2.2e-16
```

- Only NUMBED and (barely) SQRFOOT are statistically significant for predicting TPY.
- Even though quantitative variables NUMBED and SQRFOOT are strongly correlated, we get low standard errors on coefficients' estimates (computing VIFs confirms the negligible effect of collinearity: $VIF_{NUMBED} = 3.3$, $VIF_{SQRFOOT} = 3.55$)

Let's check the diagnostic plots:



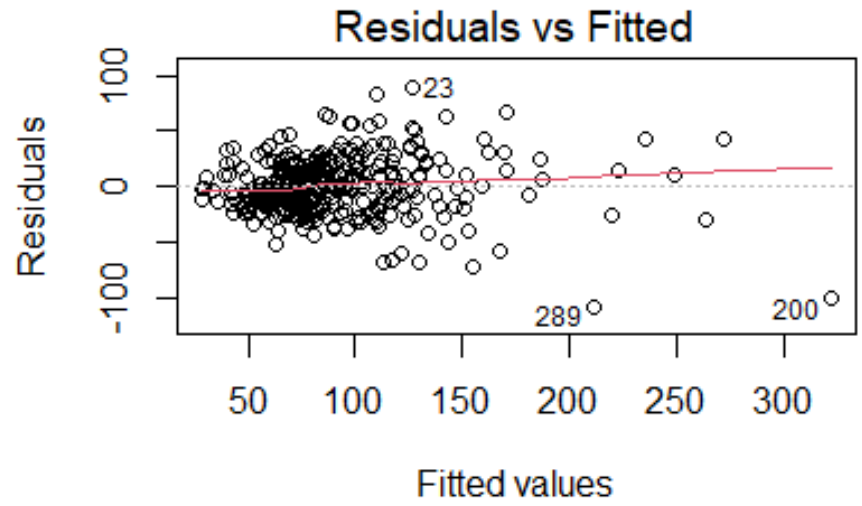
The assumptions on homoscedasticity and normality of error are not fully and clearly met. Let's try models using one variable at a time.

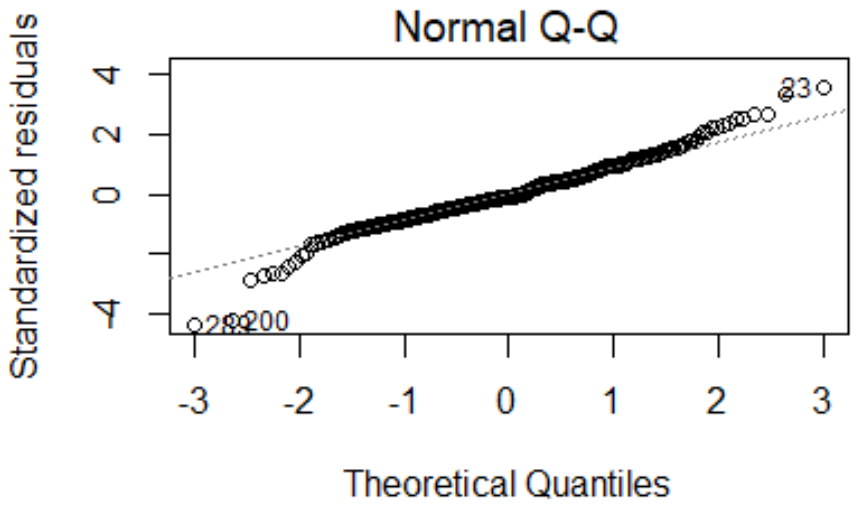
Simpler model

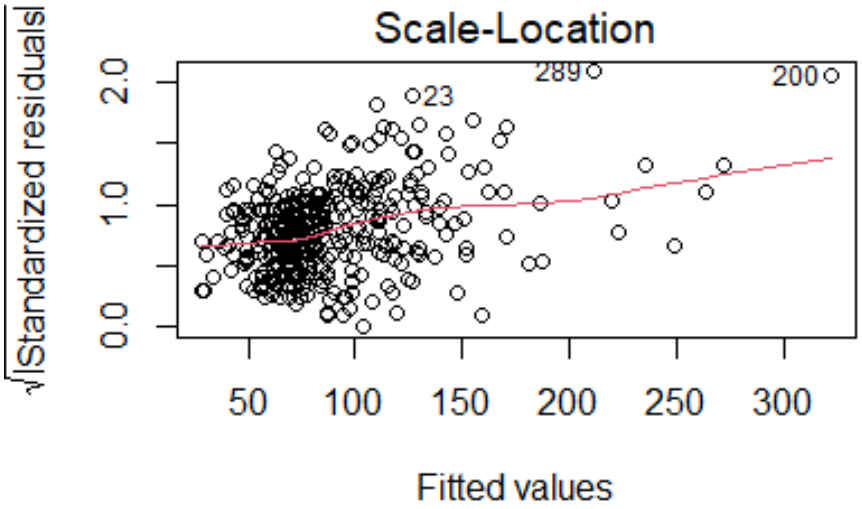
Here's a model with just SQRFOOT + categorical variables as predictors:

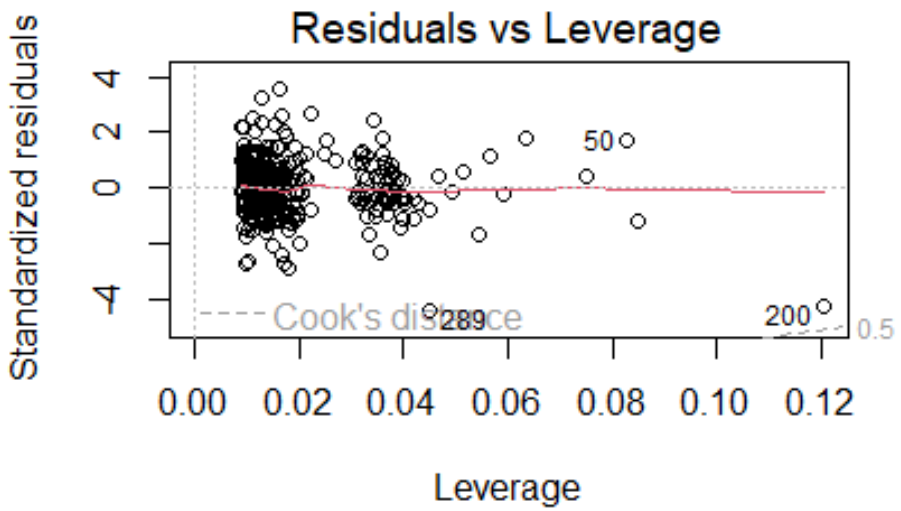
```
Call:
lm(formula = TPY ~ SQRFOOT + TAXEXEMPT + MCERT + SELFFUNDINS +
    PRO + URBAN, data = data_2000)

Residuals:
     Min       1Q   Median       3Q      Max
-108.294  -14.661   -2.574   14.608   89.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.2624     6.4155   3.626 0.000331 ***
SQRFOOT       1.1143     0.0438  25.443  < 2e-16 ***
TAXEXEMPT    -6.5432     4.7786  -1.369 0.171793
MCERT         8.8839     4.7645   1.865 0.063073 .
SELFFUNDINS   2.0734     2.8561   0.726 0.468359
PRO           3.9722     4.8285   0.823 0.411263
URBAN         2.3222     2.8365   0.819 0.413530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.52 on 350 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6992,    Adjusted R-squared:  0.6941
F-statistic: 135.6 on 6 and 350 DF,  p-value: < 2.2e-16
```

We inspect the diagnostic plots:



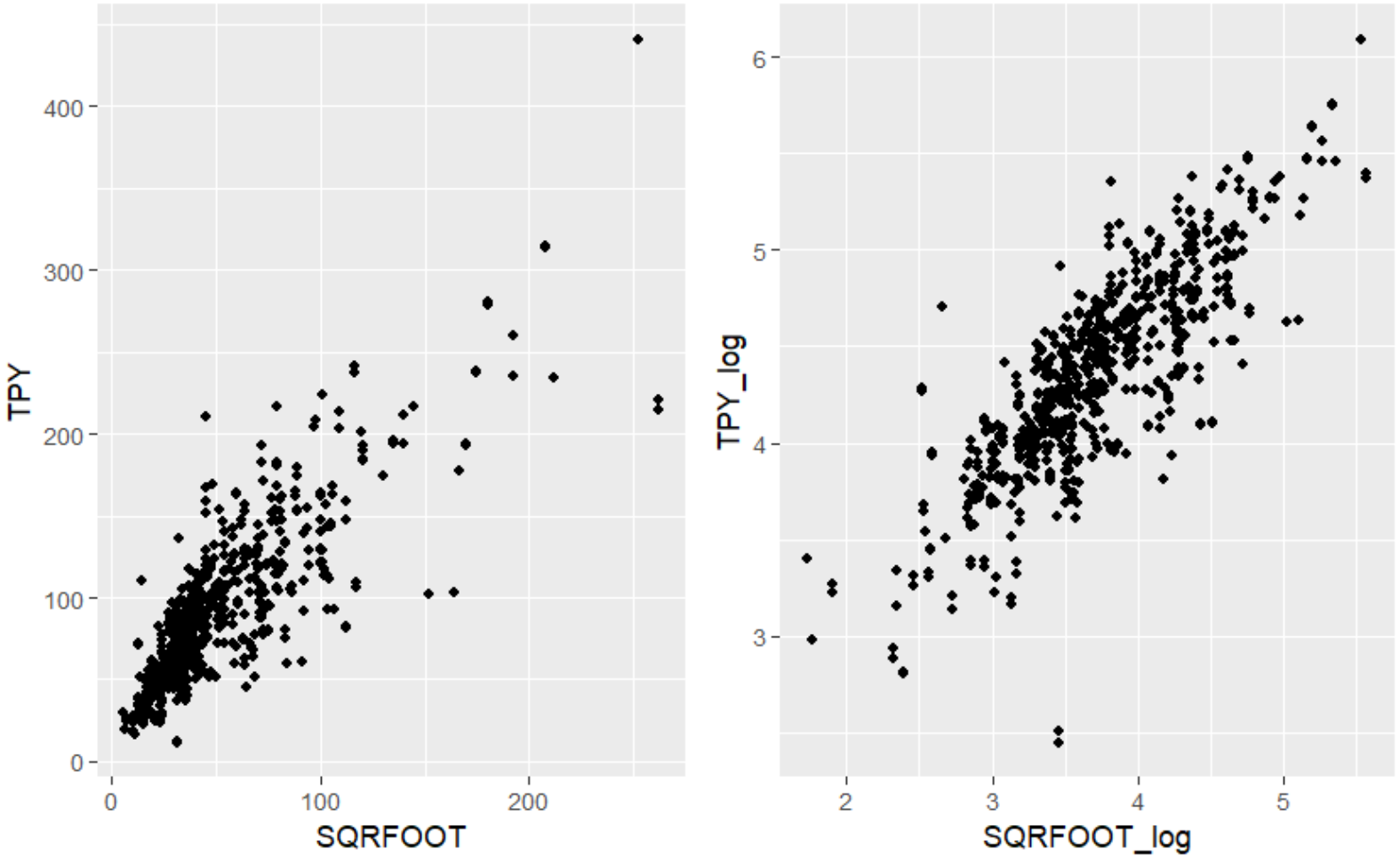






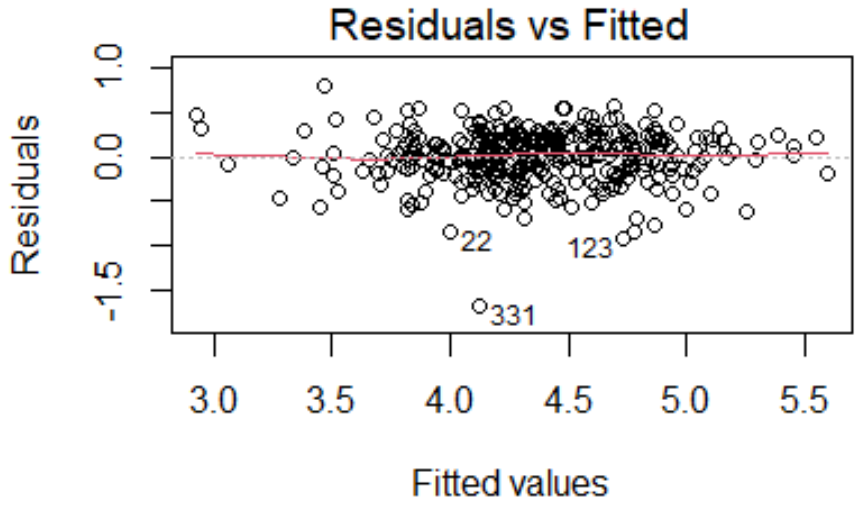
Homoscedasticity may still be lacking, distribution of error becomes more normal.

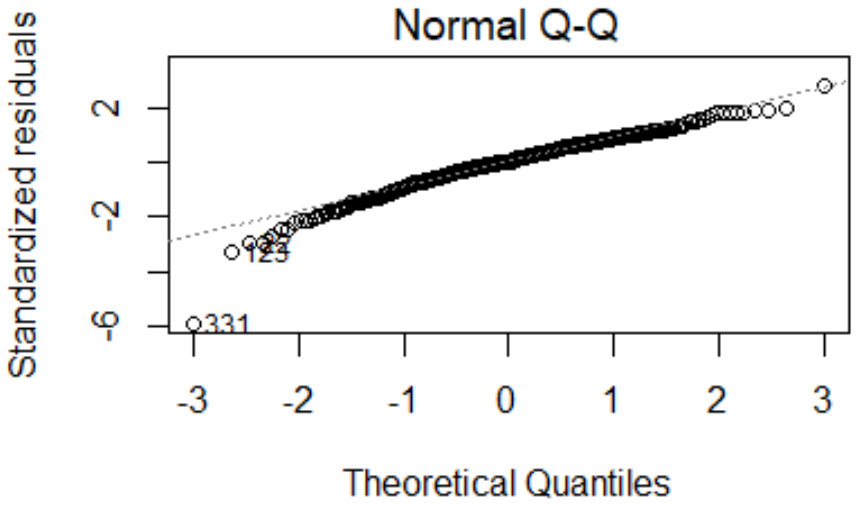
Logarithmic transformation

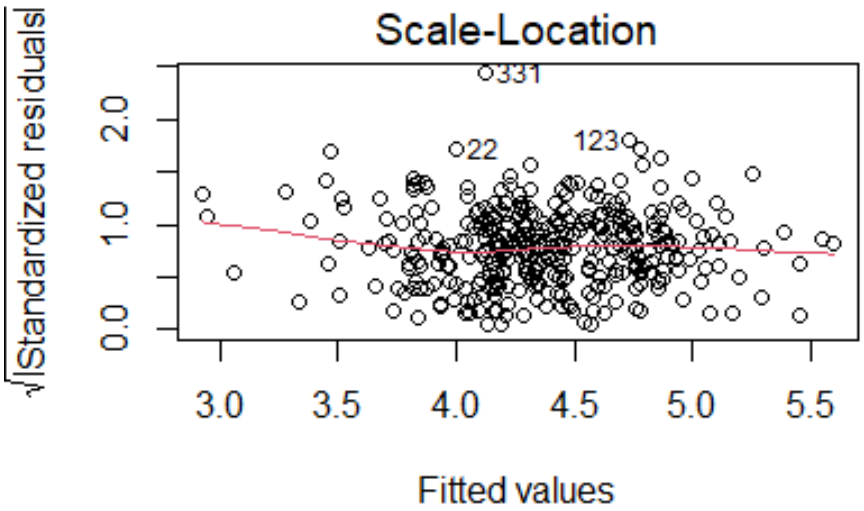


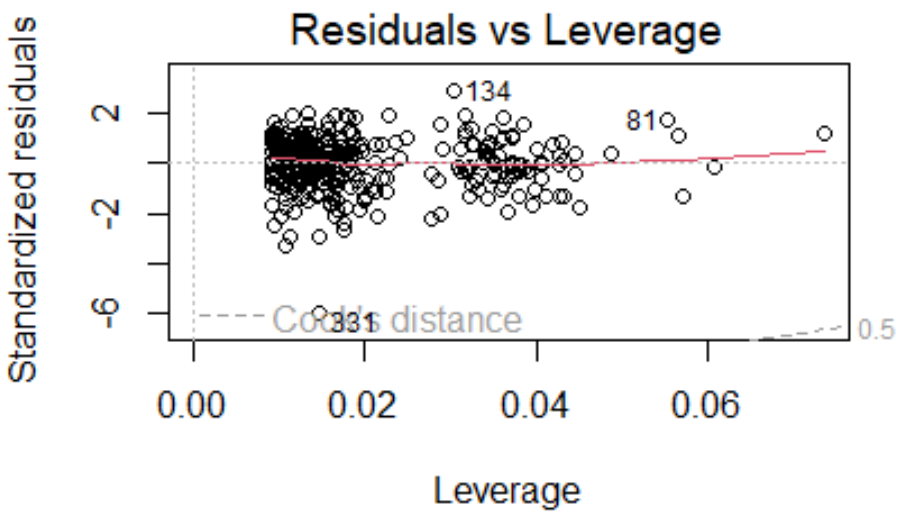
We expect an improvement in both homoscedasticity and normality of the error.

```
Call:
lm(formula = log(TPY) ~ log(SQRFOOT) + TAXEXEMPT + PRO + URBAN +
    MCERT + SELFFUNDINS, data = data_2000)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67954 -0.15593  0.01403  0.19054  0.80652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.685388   0.118982  14.165   <2e-16 ***
log(SQRFOOT)  0.695540   0.027796  25.023   <2e-16 ***
TAXEXEMPT    -0.066653   0.053226  -1.252   0.2113
PRO           0.047028   0.053418   0.880   0.3793
URBAN        -0.007322   0.031701  -0.231   0.8175
MCERT         0.097763   0.053751   1.819   0.0698 .
SELFFUNDINS   0.009171   0.031992   0.287   0.7745
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.285 on 350 degrees of freedom
Multiple R-squared:  0.693,  Adjusted R-squared:  0.6877
F-statistic: 131.7 on 6 and 350 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
22
123
331


Normal Q-Q
Standardized residuals
Theoretical Quantiles
123
331


Scale-Location
√|Standardized residuals|
Fitted values
331
22
123


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
134
81
331
0.5

Consequences of adding NUMBED as predictor

Trying to add NUMBED we get:

```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED) + log(SQRFOOT) + TAXEXEMPT,
##     data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87174 -0.01698  0.01996  0.05573  0.21374
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.15359    0.05990  -2.564   0.0108 *
## log(NUMBED)   1.00057    0.02370  42.219   <2e-16 ***
## log(SQRFOOT)  0.01149    0.01926   0.597   0.5510
## TAXEXEMPT     0.02964    0.01332   2.225   0.0267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1161 on 353 degrees of freedom
## Multiple R-squared:  0.9486, Adjusted R-squared:  0.9481
## F-statistic:  2170 on 3 and 353 DF,  p-value: < 2.2e-16
```

These results are confirmed by ANOVA and AIC:

```
## Analysis of Variance Table
##
## Response: log(TPY)
##               Df Sum Sq Mean Sq   F value  Pr(>F)
## log(NUMBED)    1 87.718  87.718 6503.0259 < 2e-16 ***
## log(SQRFOOT)   1  0.023   0.023    1.7402 0.18797
## TAXEXEMPT      1  0.067   0.067    4.9528 0.02668 *
## Residuals    353  4.762   0.013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                                                                                AIC
## lm(log(TPY) ~ log(NUMBED) + log(SQRFOOT) + TAXEXEMPT, data = data_2000) -518.1060
## lm(log(TPY) ~ log(NUMBED) + TAXEXEMPT, data = data_2000)                -519.7459
```

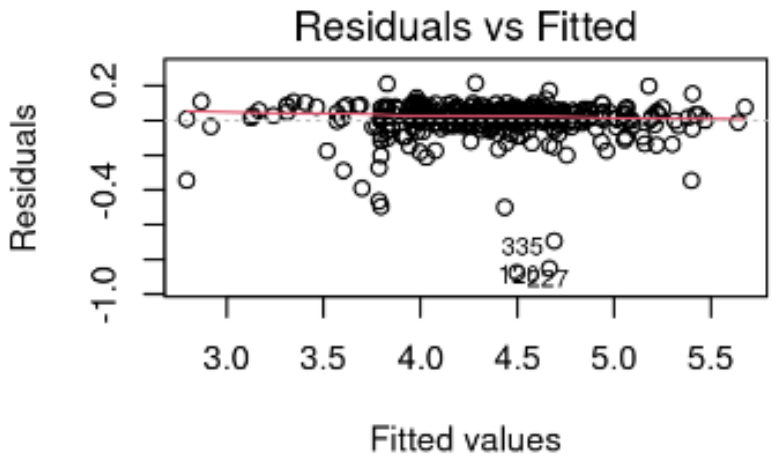
NUMBED seems to be a much stronger predictor than SQRFOOT, and the evidence is enough to drop the latter.

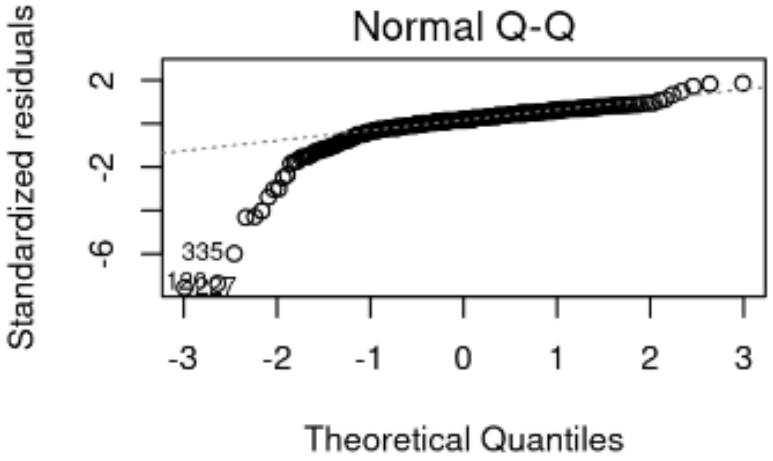
This outcome is not surprising, since (as we saw during the exploration of the dataset) NUMBED has a much stronger correlatation to TPY than SQRFOOT.

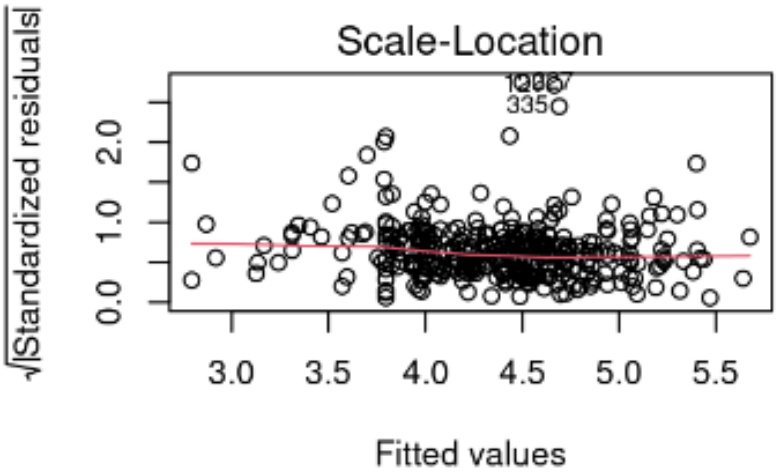
R^2 with just NUMBED is very high...

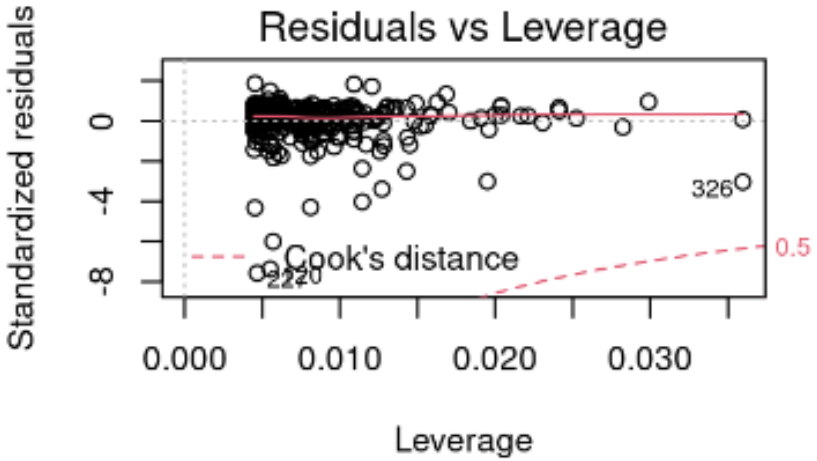
```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED) + TAXEXEMPT, data = data_2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87585 -0.01786  0.02008  0.05573  0.21599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16520    0.05660  -2.919  0.00374 **
## log(NUMBED)  1.01257    0.01254  80.749  < 2e-16 ***
## TAXEXEMPT    0.03201    0.01270   2.520  0.01219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.116 on 354 degrees of freedom
## Multiple R-squared:  0.9485, Adjusted R-squared:  0.9482
## F-statistic:  3261 on 2 and 354 DF,  p-value: < 2.2e-16
```

...though the error shows not normally distributed tails.









Summary of fitted models

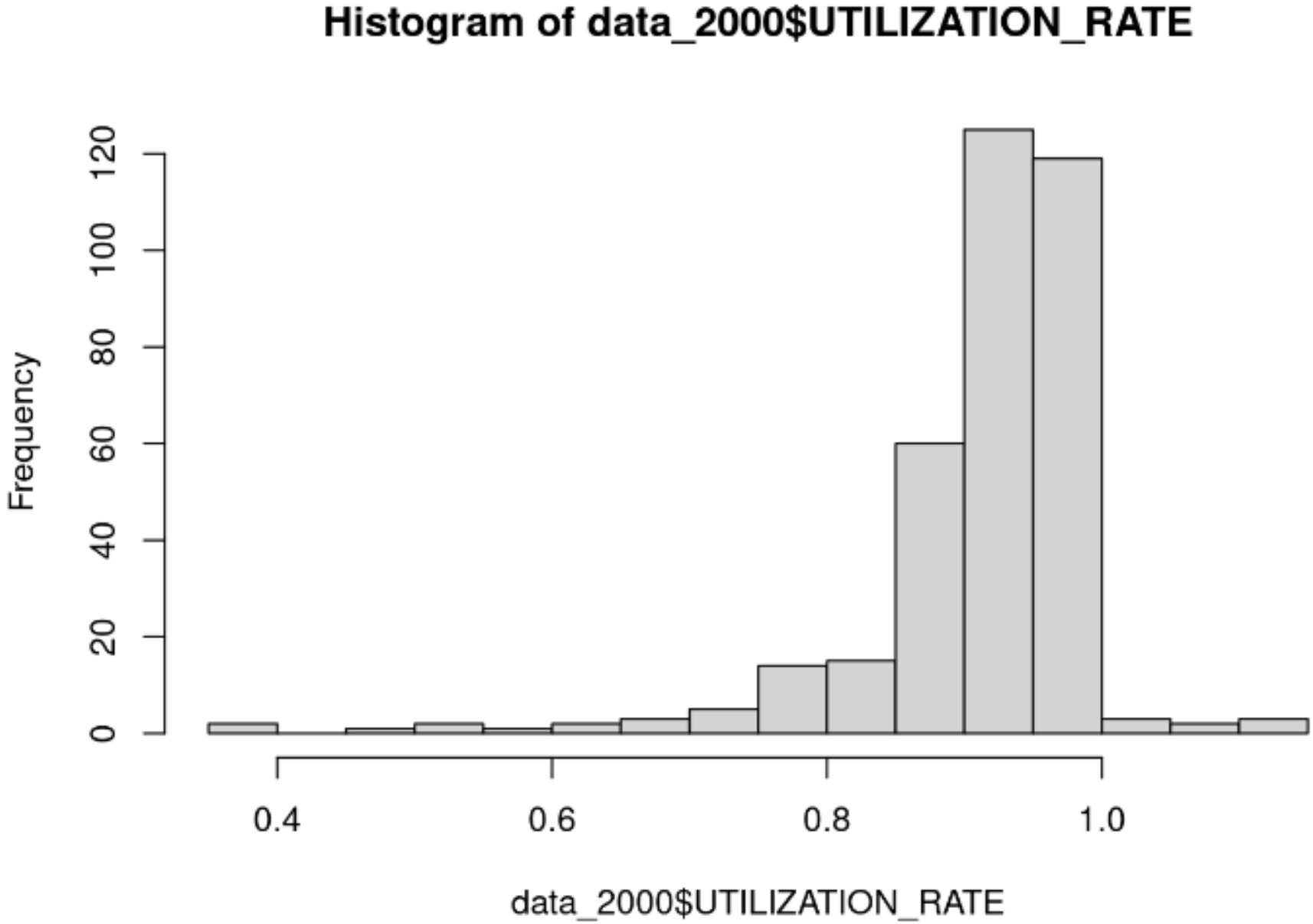
Resp. Variable	Independent Variables	R^2	AIC	MSE	Residuals
TPY	NUMBED + SQRFOOT	0.96	2609	85.47	Some heteroscedasticity and partially normal error
TPY	NUMBED	0.96	2620	88.75	Some heteroscedasticity and partially normal error
TPY	SQRFOOT	0.68	3347	680	Some heteroscedasticity and normal error
log(TPY)	log(NUMBED) + log(SQRFOOT)	0.95	-515	0.013	Homoscedasticity and partially normal error
log(TPY)	log(NUMBED)	0.95	-515	0.013	Homoscedasticity and partially normal error
log(TPY)	log(SQRFOOT)	0.68	134	0.084	Homoscedasticity and normal error

UTILIZATION_RATE

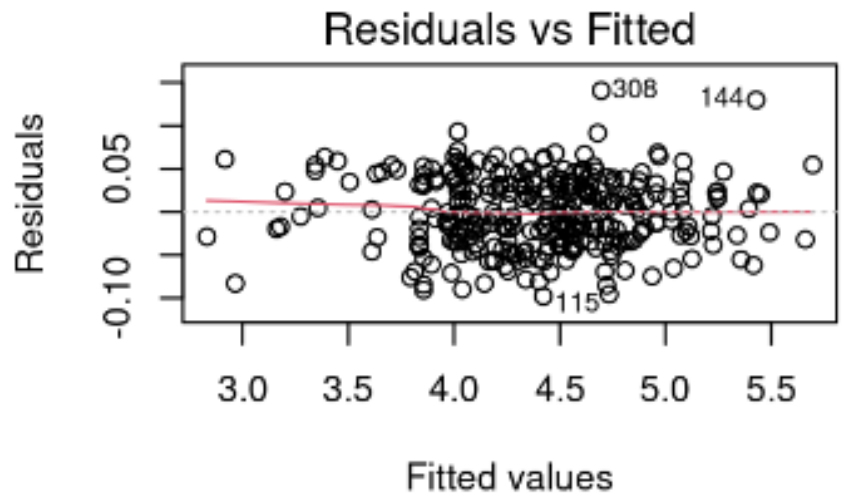
As we saw in the previous table, the model with log(NUMBED) would be the best, if it wasn't for several outliers compromising normality and to some extent homoscedasticity.

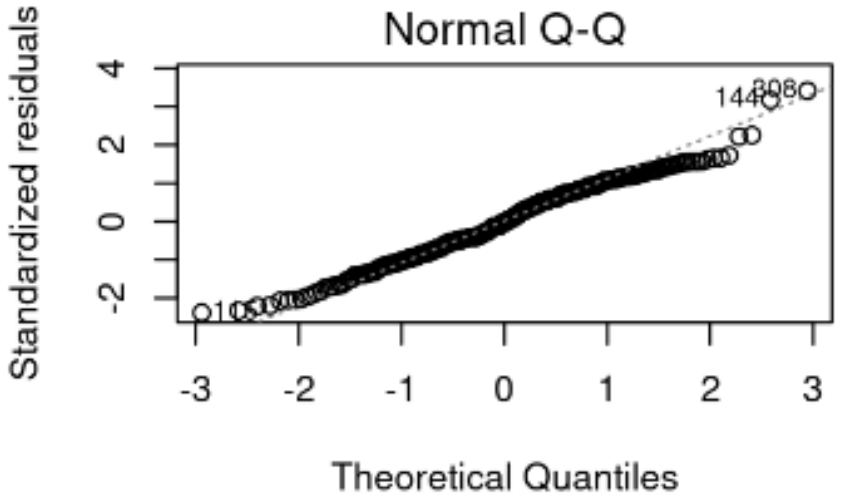
These outliers are observations for which TPY is a lot lower (and sometimes higher) than expected, meaning that these nursing houses have been underusing (or overusing) their facilities. We can measure this by defining a new variable:

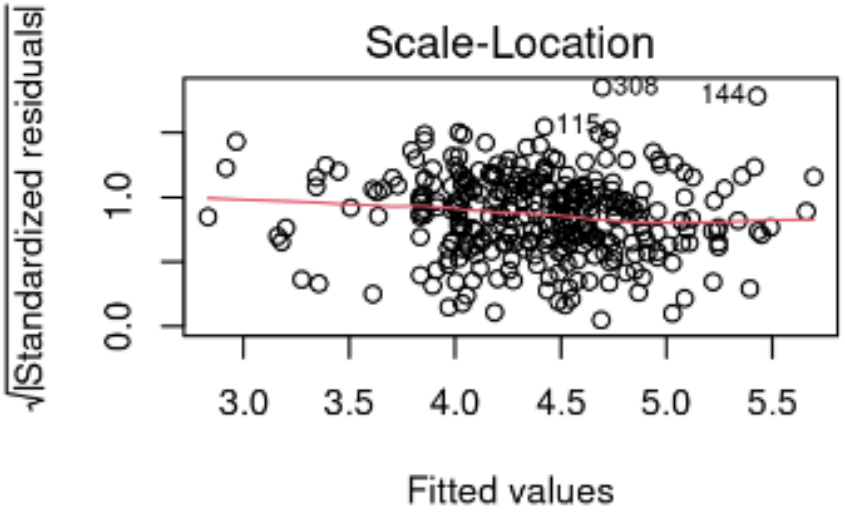
$$UTILIZATION\_RATE = \frac{TPY}{NUMBED}.$$

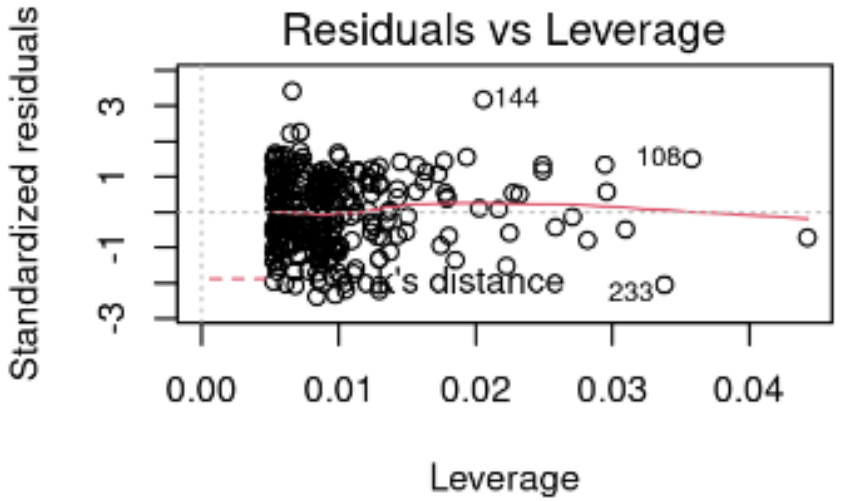

Histogram of data_2000$UTILIZATION_RATE
Frequency
0
20
40
60
80
100
120
0.4
0.6
0.8
1.0
data_2000$UTILIZATION_RATE

In fact, if we choose a suitable interval for *UTILIZATION_RATE* (in this case [0.85, 1.07]) and remove the observations with the most extreme values, then we get the following residuals:





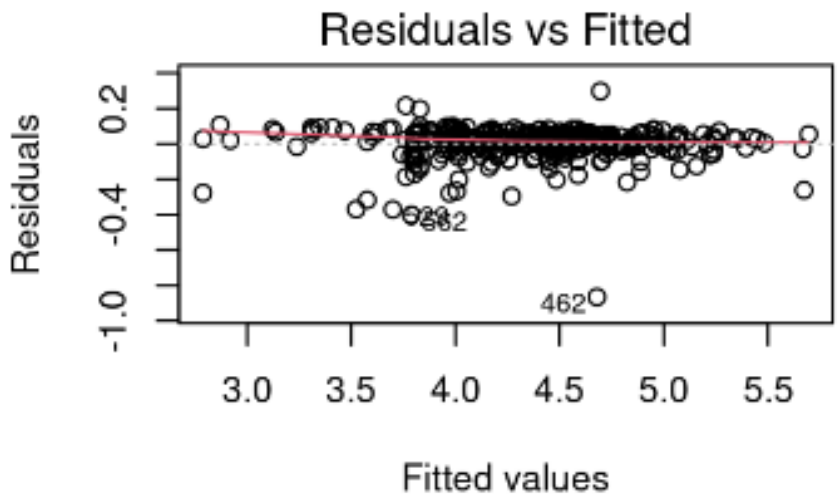


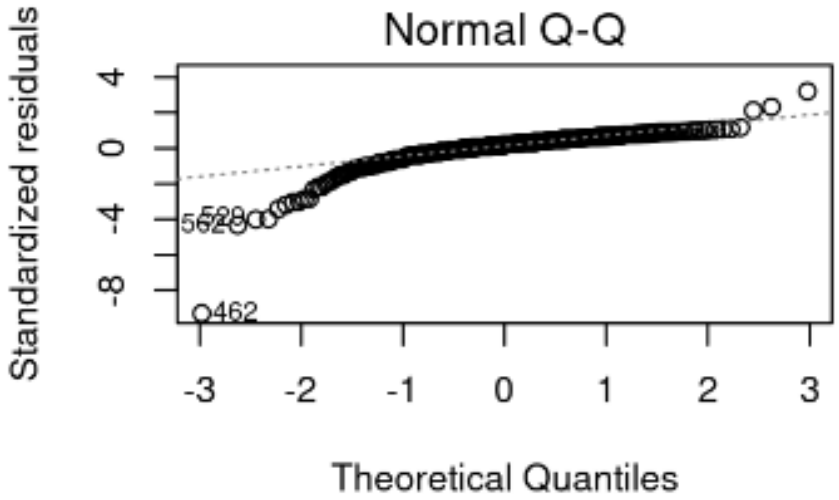


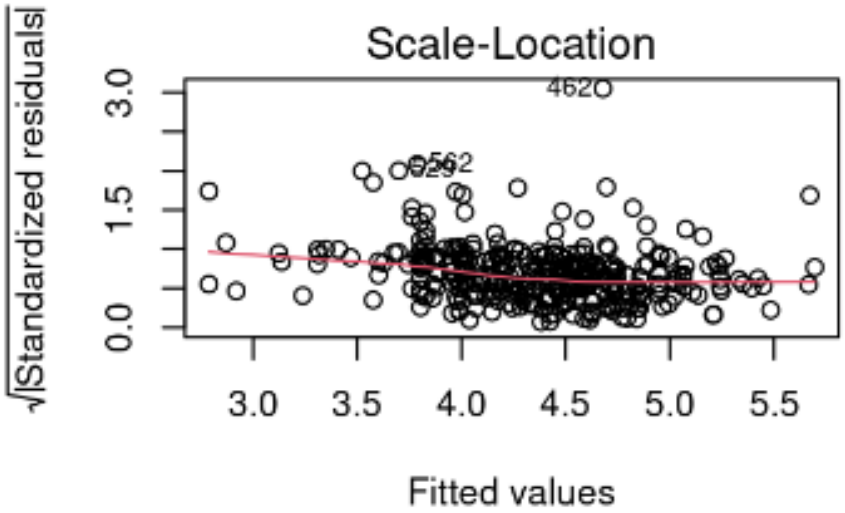
Of course this is not a usable model, since the selection of data requires to know TPY for the current year in advance.

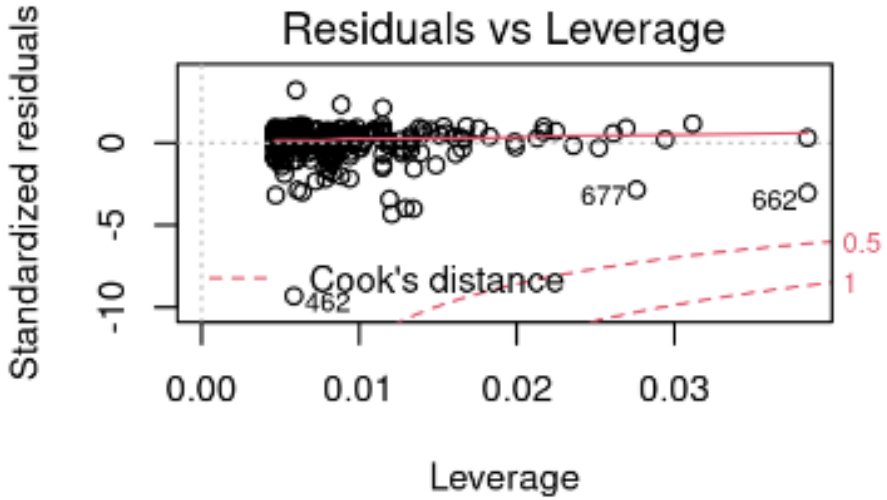
However, our dataset contains data from both 2000 and 2001. Therefore, for the following model we will assume to have, at the moment of estimation, a recording of TPY from the previous year.

Without filtering

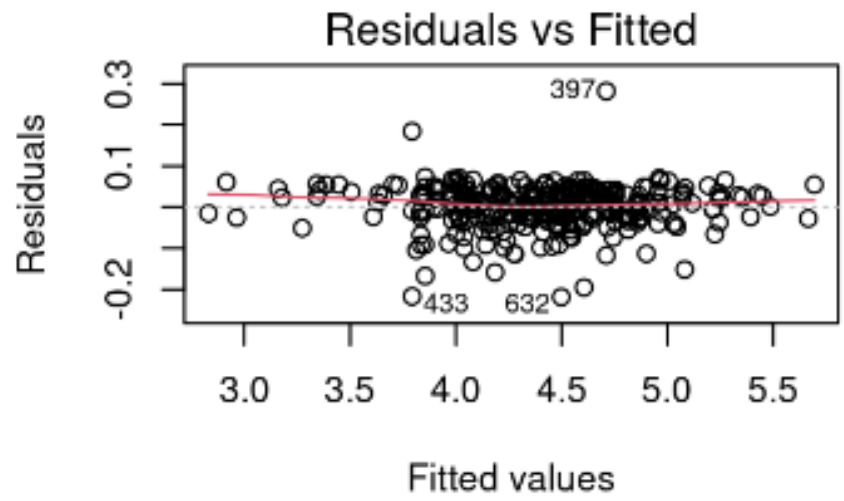


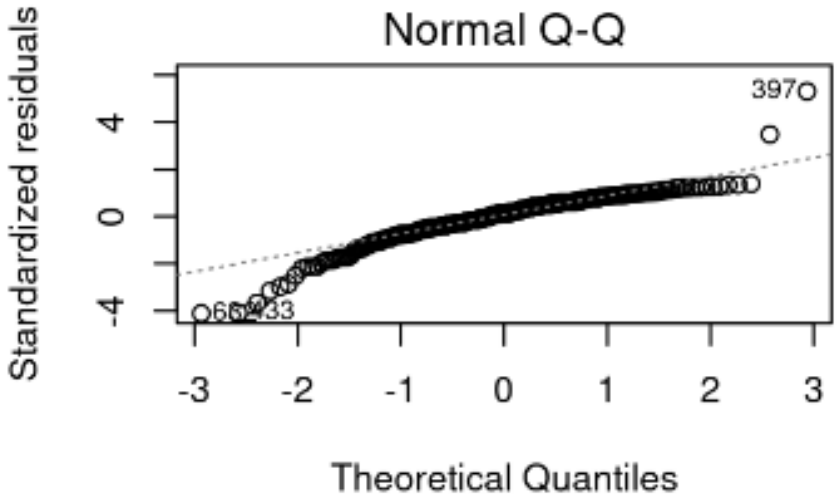


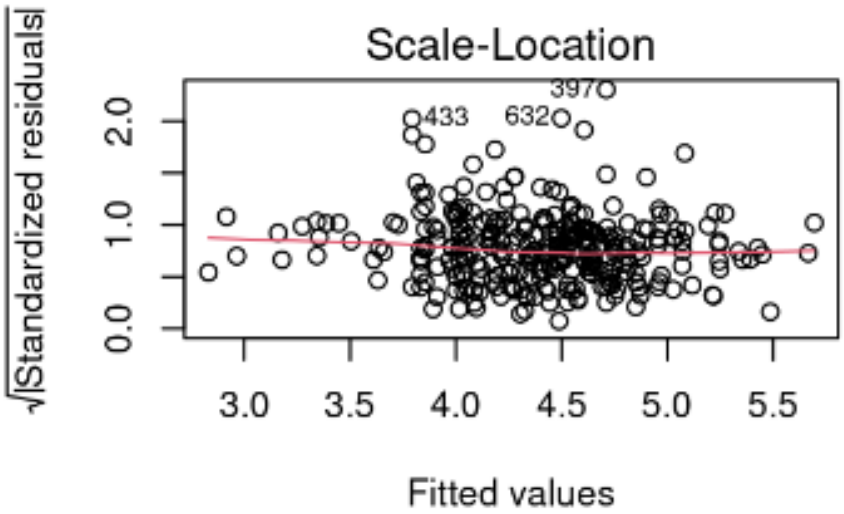


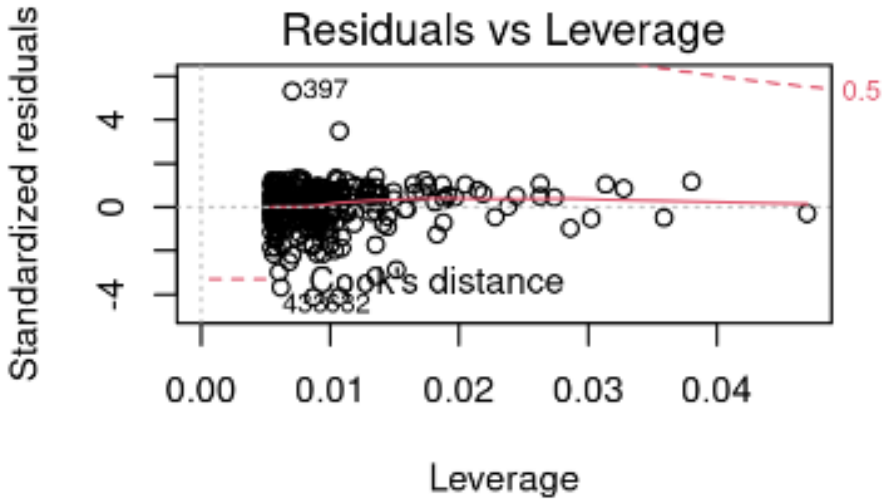


With filtering

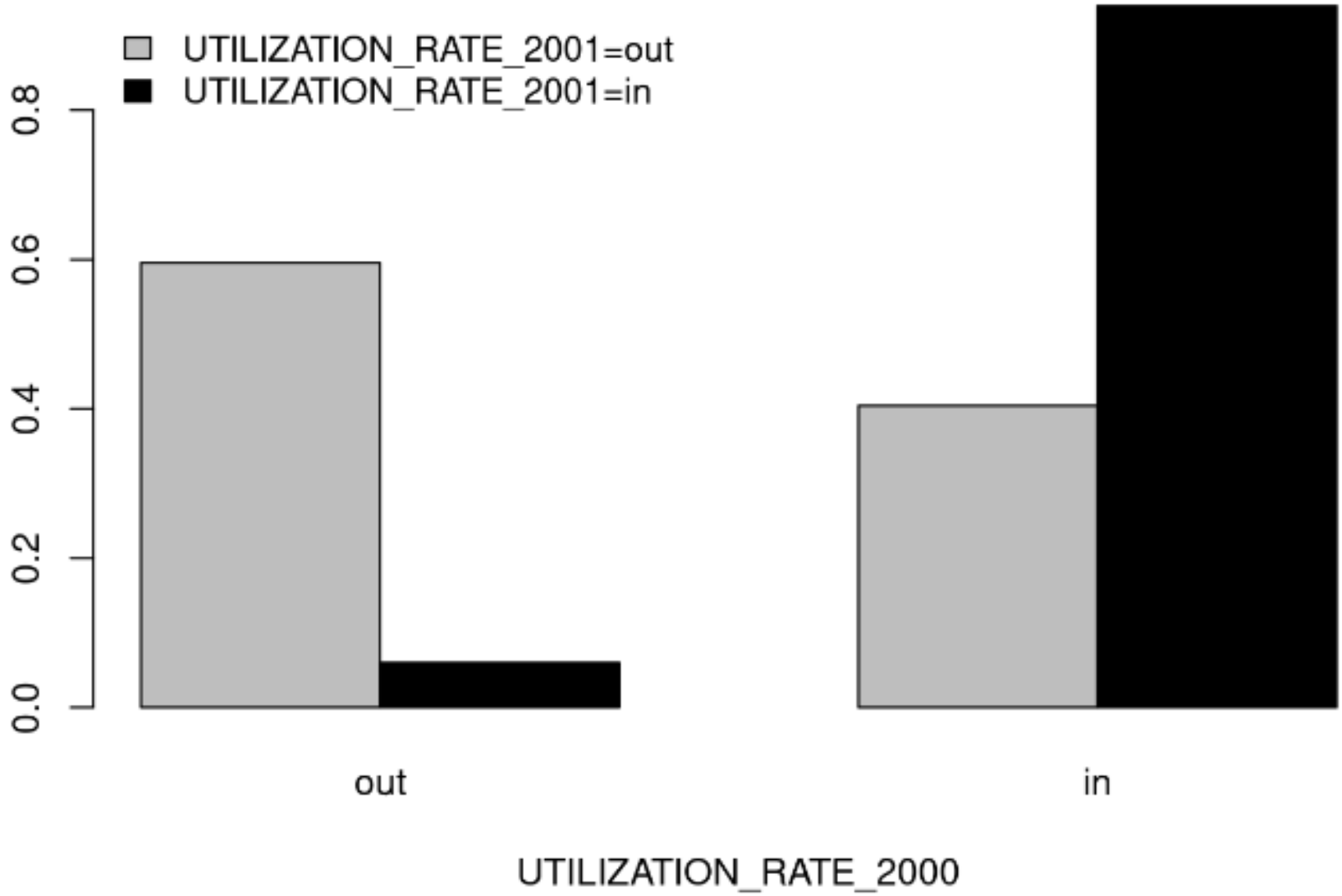








We will try to fit a model on 2001 data, filtered using the optimal interval found for the previous year.



Gamma GLM

Let's go back to modeling only data from 2000.

The nature and the distribution of the variable TPY suggests that another appropriate model could be the gamma generalized linear model. The best link function turns out to be the logarithm.

Here the AIC of some models:

```
##                                                                                                  AIC
## glm(TPY ~ NUMBED + SQRFOOT + TAXEXEMPT, data = data_2000, family = Gamma(link = "log"))          2961.018
## glm(TPY ~ log(NUMBED) + log(SQRFOOT) + TAXEXEMPT, data = data_2000, family = Gamma(link = "log")) 2550.667
## glm(TPY ~ log(NUMBED) + TAXEXEMPT, data = data_2000, family = Gamma(link = "log"))               2548.986
## glm(TPY ~ log(SQRFOOT) + TAXEXEMPT, data = data_2000, family = Gamma(link = "log"))              3217.750
```

As we can see, the outcome is very similar to the one we got using linear models: the log-transformed variables are more appropriate and NUMBED is a much stronger predictor than SQRFOOT.

Actually, a gamma GLM with a logarithmic link function and log-transformed predictors is very similar to a linear model with all log-transformed variables.

In fact, for instance, the GLM with log(SQRFOOT) and TAXEXEXMPT gives parameters' estimations that are very close to the ones we got using the linear model with the same predictors.

Linear model:

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.79793    0.09357  19.215  < 2e-16 ***
## log(SQRFOOT)  0.70110    0.02505  27.992  < 2e-16 ***
## TAXEXEMPT    -0.11807    0.03157  -3.741 0.000214 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gamma model:

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.85251    0.08644  21.432  < 2e-16 ***
## log(SQRFOOT)  0.69426    0.02314  30.007  < 2e-16 ***
## TAXEXEMPT    -0.09582    0.02916  -3.286  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table of Contents

Tree-based techniques

The techniques

- Regression Tree
- Tree Bagging
- Random Forest

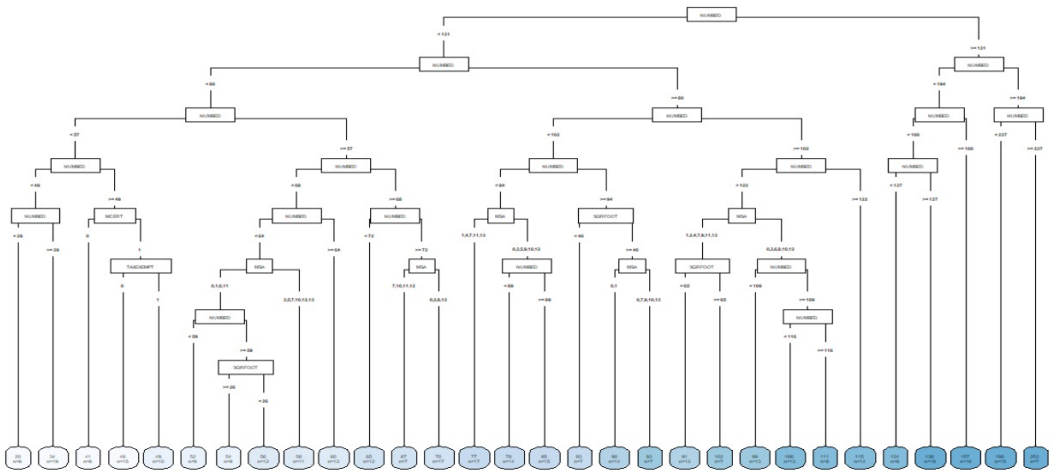
The data

We used year 2000 for the model training, and year 2001 for testing.

Model evaluation

The evaluation of how well the model can predict a value was done with the calculation of the RMSE (Root Mean Squared Error), i.e the average difference between the observed known values of the outcome and the predicted value by the model.

Regression tree



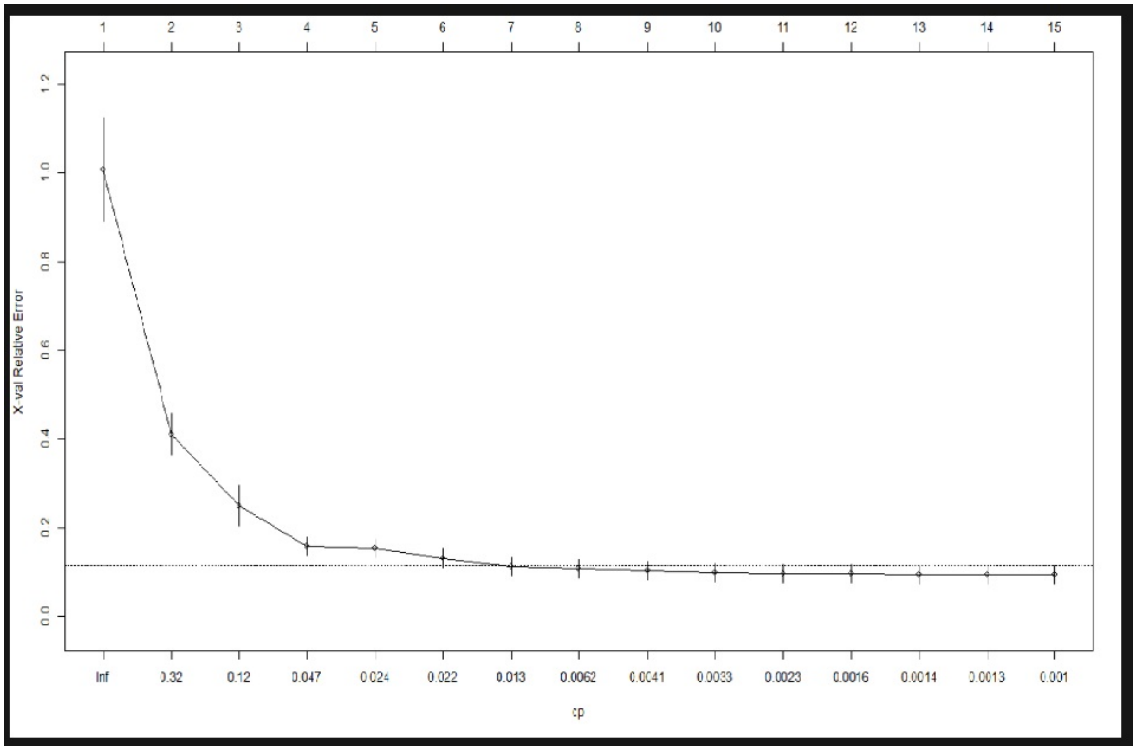
The most complex tree (cp = 0):

- 29 terminal nodes
- 5 variables used: NUMBED, SQRFOOT, MSA, MCERT, TAXEXEMPT
- RMSE = 13.41

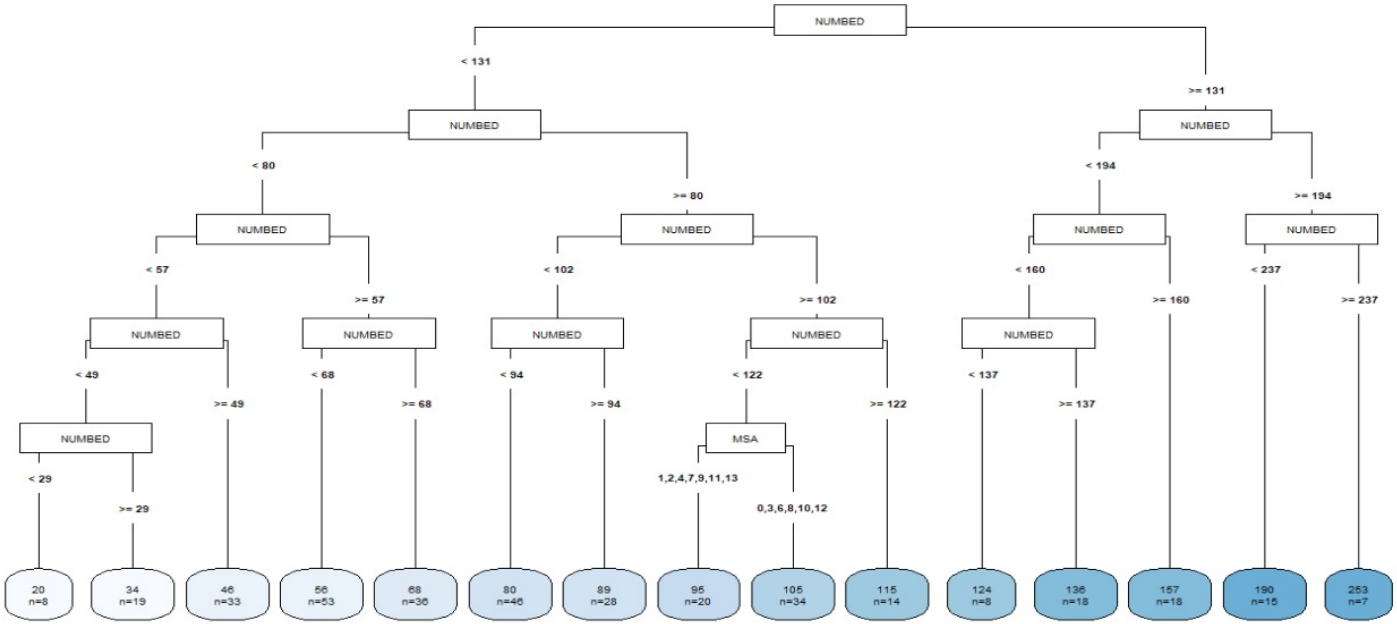
This model, despite the RMSE value is quite good, is not easily interpretable. We pruned the tree whit a greater value of cp in order to have a more comprehensible model.

Pruned tree

Looking at the cp-table and cp-plot, a cp value lower than 0.00092 does not appear to make a significant improvement in the model. So, let's improve the tree by applying this parameter.



Pruned tree



- 15 terminal nodes
- 2 variables used: NUMBED, MSA
- RMSE = 13.41

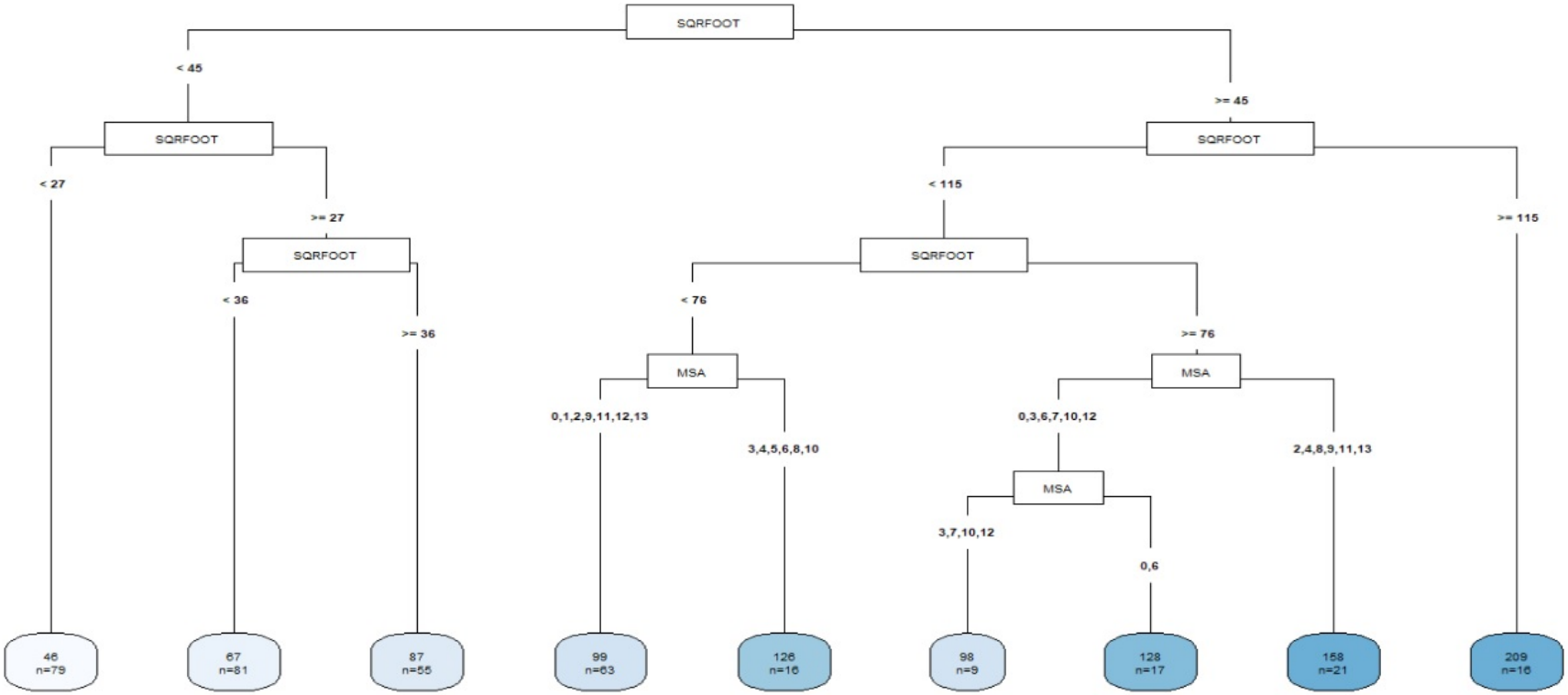
Regression tree without NUMBED

From the linear model analysis and looking at the previous tree, we already know that the variable NUMBED is highly correlated with TPY. Trying to build a tree without including it, we have further confirmation.
After pruning the most complex tree (with a cp=0.0055), this is a summary of the model:

- 9 terminal nodes
- 2 variables used: SQRFOOT, MSA
- RMSE = 25.85

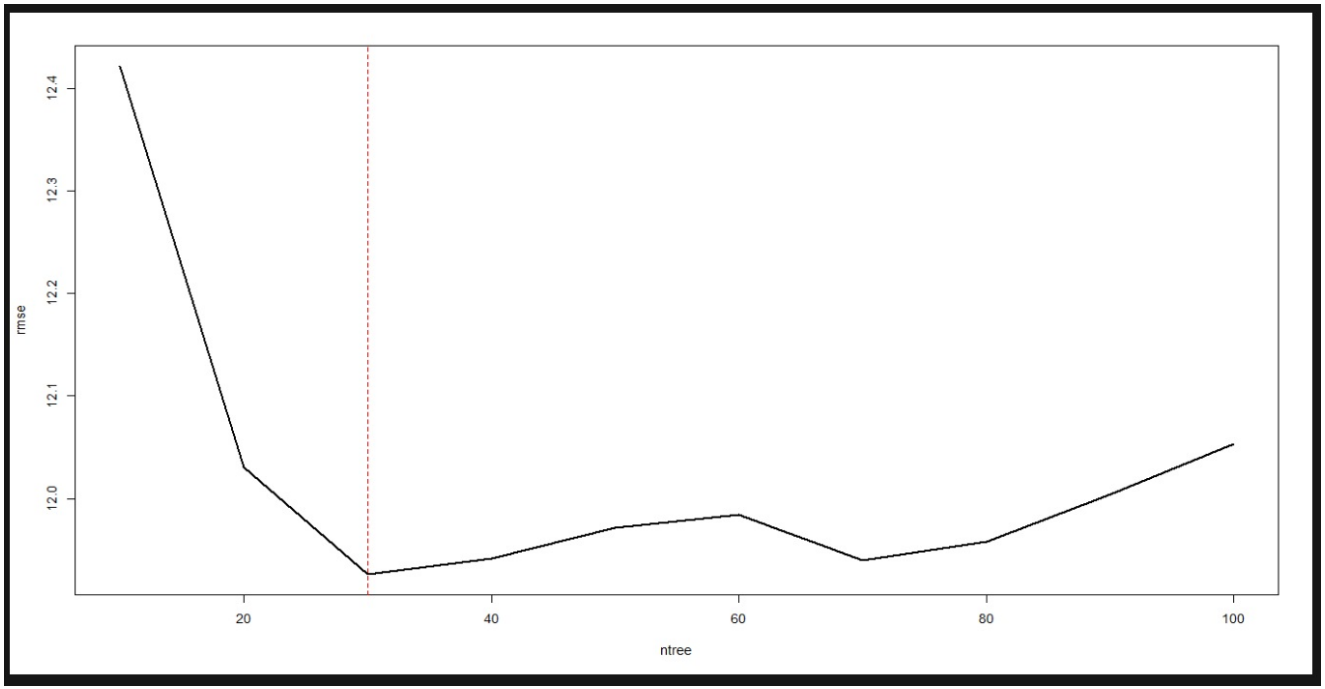
Also from this model it emerges that, if we do not consider NUMBED, the second variable with the most impact on the prediction is SQRFOOT.

Pruned tree without NUMBED



Bagging

First, let's calculate the best number of trees to use for the Bagging

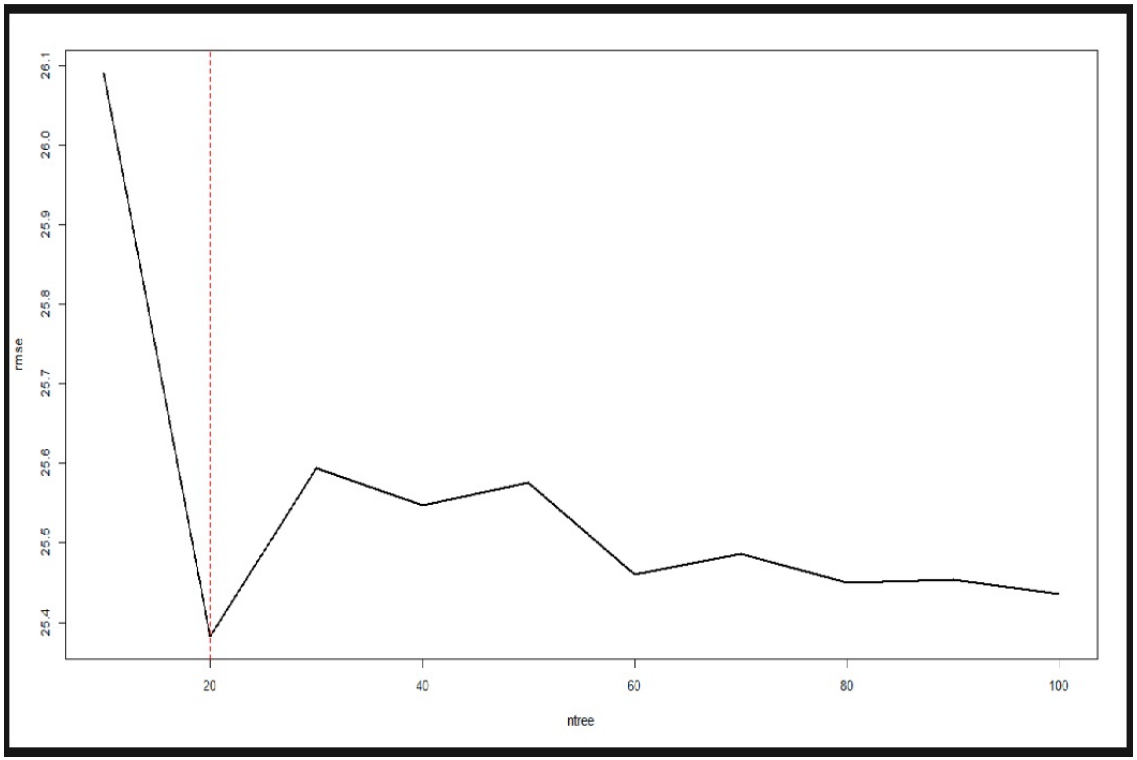


Bagging with 30 trees:

- RMSE = 14.51
- OOB = 11.93

Bagging without NUMBED

The same for the model without the variable NUMBED:

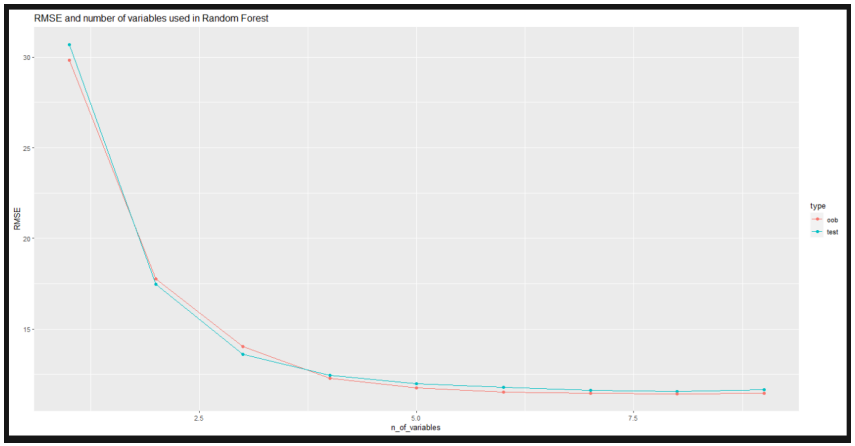


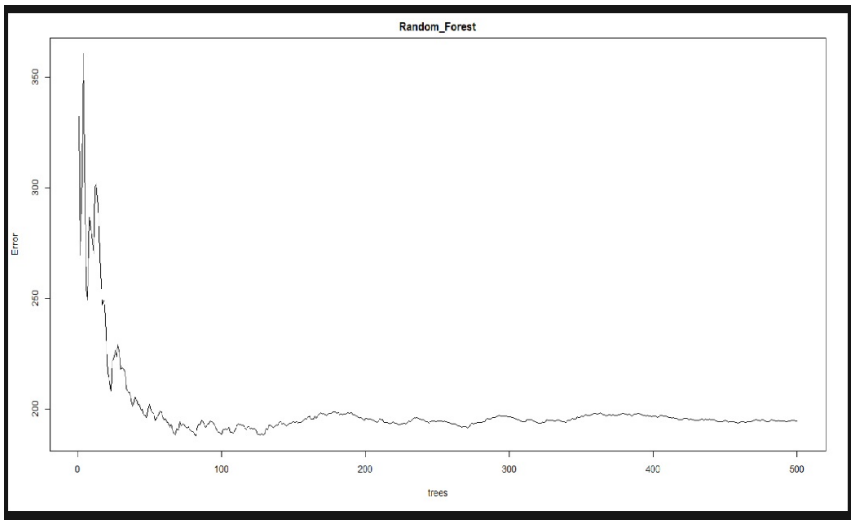
Bagging with 20 trees and without NUMBED:

- RMSE = 25.85
- OOB = 25.38

Random Forest

The last model we have fitted is the Random Forest.





As we can see from the plot on the left, the best number of variables to consider is 6. A Random Forest with 500 trees has the following performance:

- RMSE $= 11.74$

Random Forest without NUMBED

Considering only SQRFOOT and the categorical variables, as observed in the previous models, the error is quite larger

- RMSE = 19.81

Considerations about Random Forests
As expected, the model based on the Random Forest is the one with the best performance among the tree based techniques.

Table of Contents

Performance

We can compare the RMSE on predicting 2001 data of linear models trained with 2000 data with the results obtained from the trees:

Technique	Resp. variable	Independent variables	RMSE
Linear Regression	TPY	NUMBED + SQRFOOT	7.46
Linear Regression	TPY	SQRFOOT + Categorical variables	27.24
Regression Tree	TPY	NUMBED + MSA	13.41
Regression Tree	TPY	SQRFOOT + MSA	25.85
Tree Bagging	TPY	With NUMBED	14.51
Tree Bagging	TPY	Without NUMBED	25.85
Random Forest	TPY	With NUMBED	11.74
Random Forest	TPY	Without NUMBED	19.81

Conclusions

The clear linear relation between TPY and NUMBED suggested liner regression to be the most suitable technique for our problem.

This idea was confirmed by the comparison between several models, obtained using linear regression, generalized linear models and tree-based techniques.

However, from the analysis of residuals we noticed that the assumptions for linear models were not fully met. Assuming the knowledge of the response variable from the previous year, we showed that the linear model could be slightly improved.