Report

Artem Solomko

Data description

Each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are defined as follows (taken from the UCI Machine Learning Repository)

Parameters:
X: the sequence number of the line
crim: per capita crime rate by town
zn: proportion of residential land zoned for lots over 25,000 sq.ft.
indus: proportion of non-retail business acres per town
chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox: nitric oxides concentration (parts per 10 million)
rm: average number of rooms per dwelling
age: proportion of owner-occupied units built prior to 1940
dis: weighted distances to five Boston employment centers
rad: index of accessibility to radial highways
tax: full-value property-tax rate per $10,000
ptratio: pupil-teacher ratio by town
black: 1000(Bk−0.63) where Bk is the proportion of blacks by town
lstat: % lower status of the population
medv: Median value of owner-occupied homes in $1000s

Data loading

```
data_init <- read.csv("Boston.csv", sep=",", dec=".", header=1)
data <- data_init
```

Data analysis

Data structure

```
str(data)
```

```
## 'data.frame':    506 obs. of  15 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
```

```
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : int  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

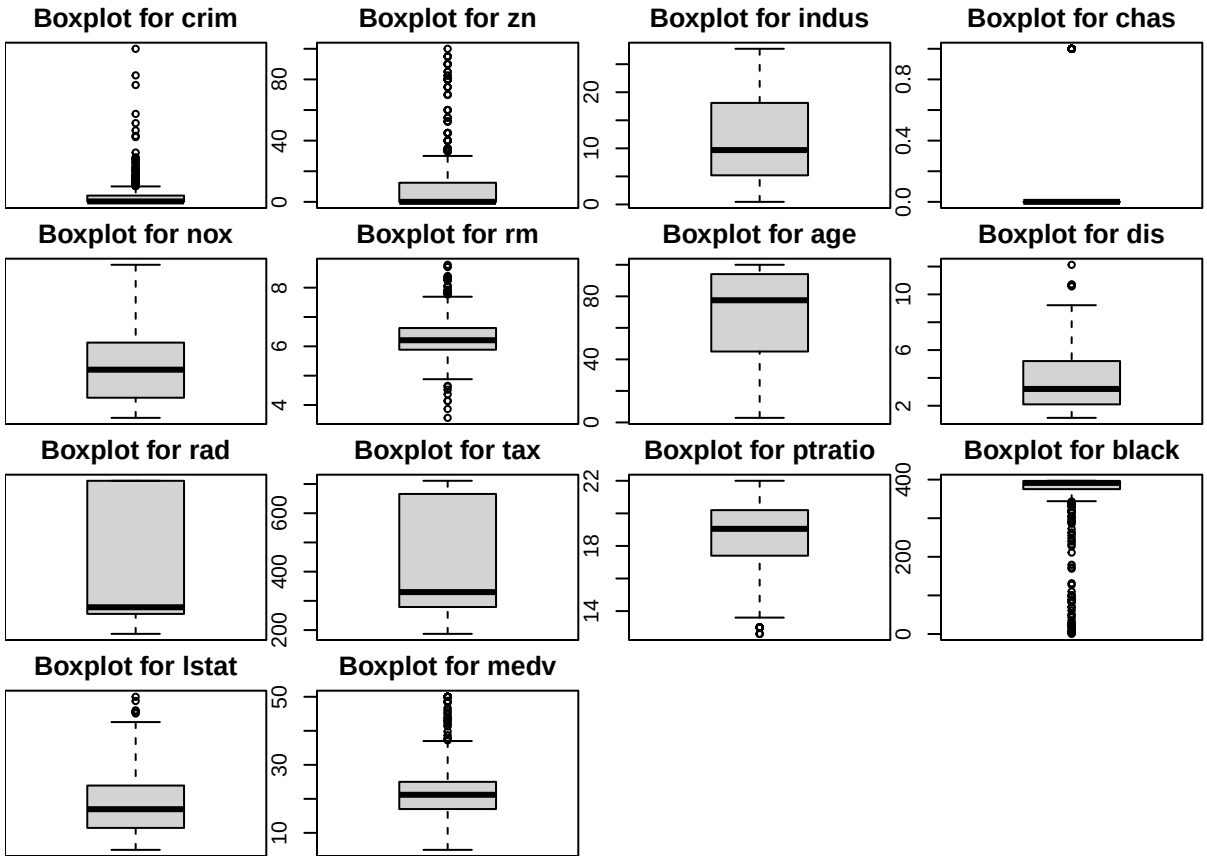
Data summary

```
summary(data)
```

```
##        X              crim                zn             indus            chas              nox
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.385
##  1st Qu.:127.2   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.449
##  Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.538
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.554
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.624
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.871
##       age              dis              rad              tax            ptratio          black
##  Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

Boxplot for each column to detect outlires

```
par(mfrow=c(4,4), mar=c(0,0,2,2))
for (i in 2:ncol(data)) {
  boxplot(data[,i],
          main = paste("Boxplot for", names(data)[i]),
          ylab = names(data)[i],
          outline = TRUE)
}
```

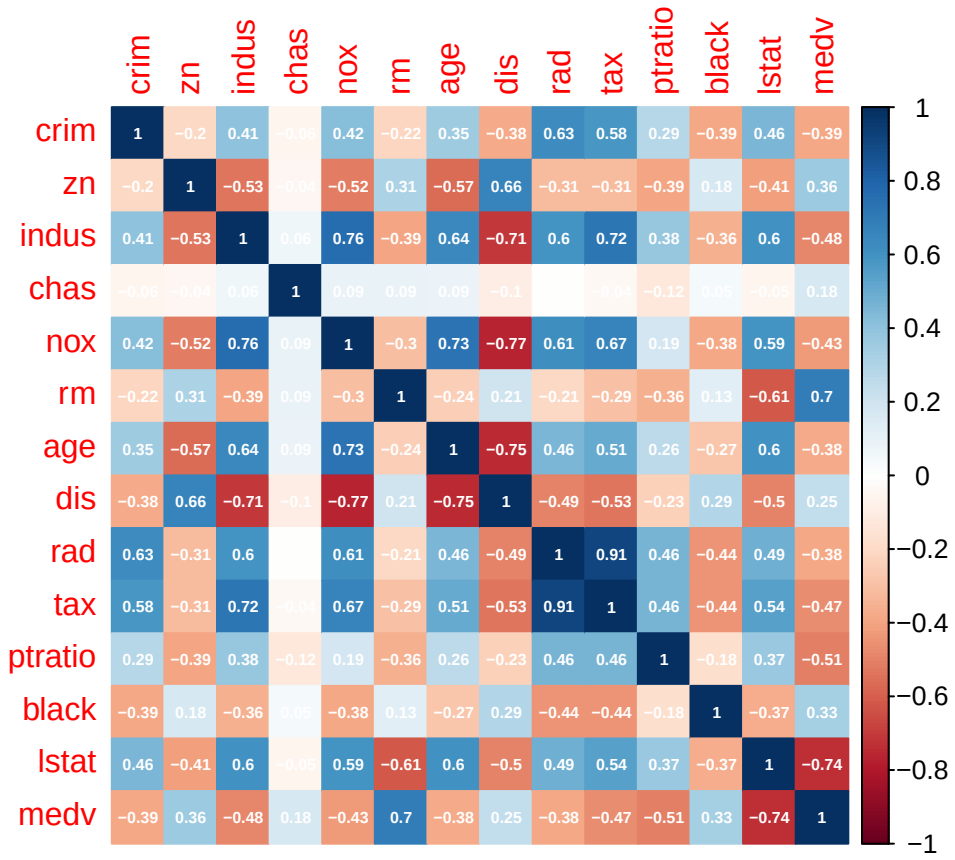


Correlation matrix

```
if (!requireNamespace("corrplot", quietly = TRUE)) {
  install.packages("corrplot")
}
library(corrplot)

correlation_matrix <- cor(data[,-1])

corrplot(correlation_matrix, method = "color", addCoef.col = "white", number.cex = 0.5)
```



By correlation matrix we can see huge correlation between tax and rad. We can delete from data one of them (see Data preparing)

Data preparing

Deleting tax (by huge correlation between it and rad)

```
data <- subset(data, select = -tax)
```

Outlires deleting

```
remove_outliers <- function(x) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
  iqr <- IQR(x, na.rm = TRUE)
  lower <- qnt[1] - 1.5 * iqr
  upper <- qnt[2] + 1.5 * iqr
  return(ifelse(x < lower | x > upper, NA, x))
}

for (i in 2:ncol(data)) {
  data[,i] <- remove_outliers(data[,i])
}

data <- na.omit(data)
```

Regression analysis

We will try to predict medv by other parameters

```
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

library(ggplot2)
library(dplyr)

model <- lm(medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + ptratio + black + lstat, data

summary(model)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
##     dis + rad + ptratio + black + lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7048 -1.4314 -0.2203  1.3159 11.8841
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.554634   7.613977   1.518 0.130360
## crim        -0.900177   0.219890  -4.094 5.69e-05 ***
## zn          -0.023093   0.024449  -0.945 0.345766
## indus       -0.075966   0.039643  -1.916 0.056445 .
## chas               NA         NA      NA       NA
## nox         -2.322106   3.312010  -0.701 0.483867
## rm           5.190437   0.535450   9.694  < 2e-16 ***
## age         -0.047060   0.009587  -4.909 1.64e-06 ***
## dis         -0.639091   0.168534  -3.792 0.000186 ***
## rad          0.206938   0.059379   3.485 0.000579 ***
## ptratio     -0.706143   0.106111  -6.655 1.71e-10 ***
## black        0.000643   0.015336   0.042 0.966591
## lstat       -0.155125   0.052386  -2.961 0.003352 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.682 on 256 degrees of freedom
## Multiple R-squared:  0.6898, Adjusted R-squared:  0.6765
## F-statistic: 51.75 on 11 and 256 DF,  p-value: < 2.2e-16
```

Adjusting a set of parameters based on their significance

```r
model <- lm(medv ~ rm + age + dis + rad + ptratio + lstat, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm + age + dis + rad + ptratio + lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1252 -1.5955 -0.2781  1.2160 12.7612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.554672   4.224268   2.262  0.02453 *
## rm           5.196553   0.552637   9.403  < 2e-16 ***
## age         -0.051670   0.009591  -5.387 1.60e-07 ***
## dis         -0.409254   0.124169  -3.296  0.00112 **
## rad         -0.045725   0.027538  -1.660  0.09803 .
## ptratio     -0.650198   0.103215  -6.299 1.26e-09 ***
## lstat       -0.217535   0.052270  -4.162 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.791 on 261 degrees of freedom
## Multiple R-squared:  0.6574, Adjusted R-squared:  0.6495
## F-statistic: 83.48 on 6 and 261 DF,  p-value: < 2.2e-16
```

The model shows poor results, let's try to use the logarithm

```r
data$medv_log <- log(data$medv)

model <- lm(medv_log ~ crim + zn + indus + chas + nox + rm + age + dis + rad + ptratio + black + lstat,

summary(model)
```

```
##
## Call:
## lm(formula = medv_log ~ crim + zn + indus + chas + nox + rm +
##     age + dis + rad + ptratio + black + lstat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48848 -0.06739 -0.00111  0.06077  0.42210
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.744e+00  3.519e-01   7.796 1.61e-13 ***
## crim        -5.813e-02  1.016e-02  -5.720 2.97e-08 ***
## zn          -1.140e-03  1.130e-03  -1.009 0.313950
## indus       -2.008e-03  1.832e-03  -1.096 0.274156
## chas                NA         NA      NA       NA
```

```
## nox         -8.534e-02  1.531e-01  -0.557 0.577677
## rm           2.137e-01  2.475e-02   8.635 6.39e-16 ***
## age         -2.233e-03  4.431e-04  -5.038 8.88e-07 ***
## dis         -2.760e-02  7.790e-03  -3.543 0.000469 ***
## rad          1.292e-02  2.744e-03   4.707 4.12e-06 ***
## ptratio     -3.243e-02  4.904e-03  -6.612 2.20e-10 ***
## black        2.388e-06  7.088e-04   0.003 0.997314
## lstat       -9.631e-03  2.421e-03  -3.978 9.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.124 on 256 degrees of freedom
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.6918
## F-statistic: 55.49 on 11 and 256 DF,  p-value: < 2.2e-16
```

Adjusting a set of parameters based on their significance for log regression

```
model <- lm(medv_log ~ nox + rm + age + dis + rad + ptratio + lstat, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = medv_log ~ nox + rm + age + dis + rad + ptratio +
##     lstat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59706 -0.06777 -0.00262  0.06571  0.42801
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.9322633  0.2298078  12.760  < 2e-16 ***
## nox         -0.3432963  0.1491583  -2.302 0.022151 *
## rm           0.2075437  0.0260762   7.959 5.38e-14 ***
## age         -0.0022306  0.0004639  -4.808 2.58e-06 ***
## dis         -0.0270202  0.0069866  -3.867 0.000139 ***
## rad         -0.0008172  0.0014700  -0.556 0.578734
## ptratio     -0.0312500  0.0049618  -6.298 1.28e-09 ***
## lstat       -0.0117662  0.0024743  -4.755 3.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1309 on 260 degrees of freedom
## Multiple R-squared:  0.6651, Adjusted R-squared:  0.6561
## F-statistic: 73.76 on 7 and 260 DF,  p-value: < 2.2e-16
```

Logarithmic regression also shown bad results

Trying to build square model

```
data$rm_2 = data$rm^2
data$age_2 = data$age^2
```

```r
data$dis_2 = data$dis^2
data$rad_2 = data$rad^2
data$ptratio_2 = data$ptratio^2
data$lstat_2 = data$lstat^2


model <- lm(medv ~ rm + age + dis + rad + ptratio + lstat + rm_2 + age_2 + dis_2 + rad_2 + ptratio_2 + l

summary(model)
```

```
##
## Call:
## lm(formula = medv ~ rm + age + dis + rad + ptratio + lstat +
##     rm_2 + age_2 + dis_2 + rad_2 + ptratio_2 + lstat_2, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.078 -1.399 -0.185  1.111 12.406
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.674e+01  3.211e+01   1.144 0.253627
## rm          -1.223e+01  8.401e+00  -1.455 0.146818
## age          6.323e-02  3.151e-02   2.007 0.045801 *
## dis         -6.580e-01  5.480e-01  -1.201 0.230941
## rad          2.547e-01  1.788e-01   1.425 0.155499
## ptratio      2.118e+00  2.004e+00   1.057 0.291494
## lstat       -5.321e-01  1.641e-01  -3.243 0.001343 **
## rm_2         1.380e+00  6.697e-01   2.061 0.040355 *
## age_2       -9.702e-04  2.801e-04  -3.464 0.000625 ***
## dis_2        1.660e-02  5.489e-02   0.302 0.762545
## rad_2       -1.090e-02  6.515e-03  -1.673 0.095526 .
## ptratio_2   -7.193e-02  5.509e-02  -1.306 0.192849
## lstat_2      1.058e-02  4.976e-03   2.127 0.034400 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.655 on 255 degrees of freedom
## Multiple R-squared:  0.6972, Adjusted R-squared:  0.6829
## F-statistic: 48.92 on 12 and 255 DF,  p-value: < 2.2e-16
```

Adjusting a set of parameters based on their significance for square regression

```r
model <- lm(medv_log ~ rm_2 + age_2 + rm, data = data)

summary(model)
```

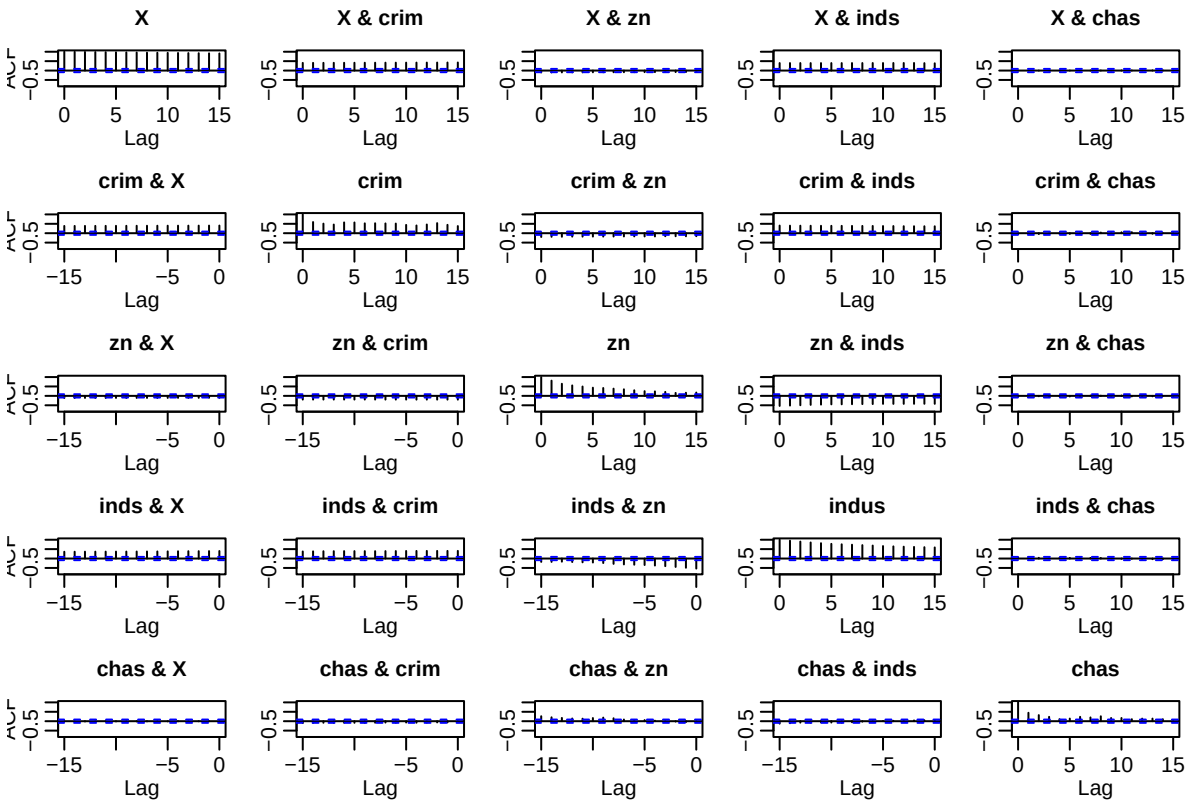
```
##
## Call:
## lm(formula = medv_log ~ rm_2 + age_2 + rm, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.63048 -0.06587  0.00304  0.06065  0.52767
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.541e+00  1.187e+00   2.982  0.00313 **
## rm_2         5.426e-02  3.045e-02   1.782  0.07589 .
## age_2       -3.214e-05  2.812e-06 -11.428  < 2e-16 ***
## rm          -3.911e-01  3.802e-01  -1.029  0.30457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1443 on 264 degrees of freedom
## Multiple R-squared:  0.5873, Adjusted R-squared:  0.5826
## F-statistic: 125.2 on 3 and 264 DF,  p-value: < 2.2e-16
```

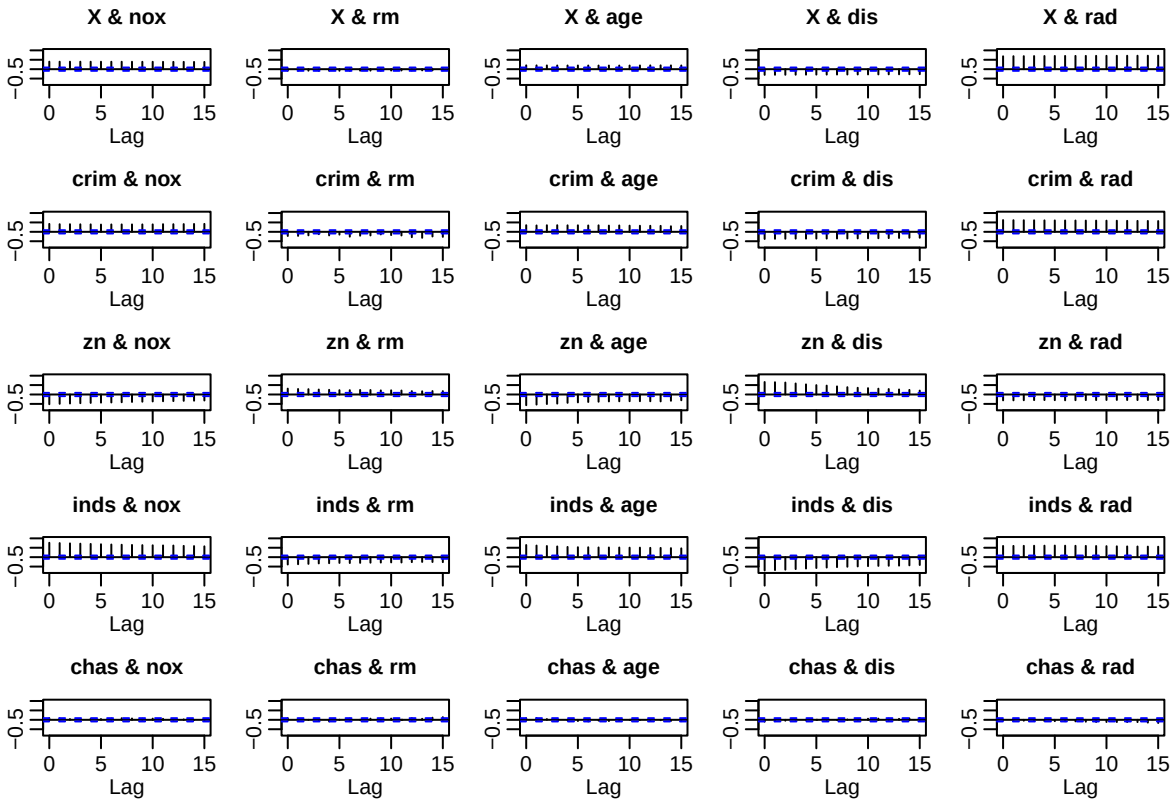
Model became less quality then it was before

Lets try to check autocorrelation

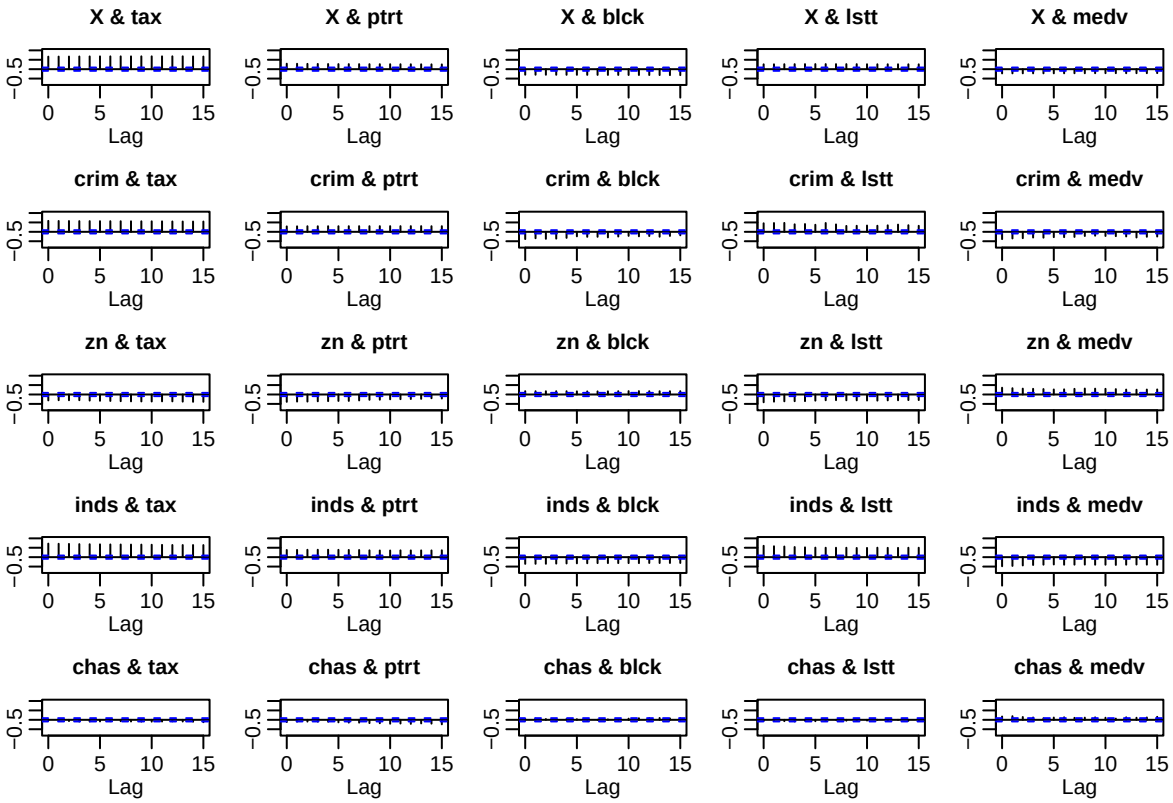
```
data <- data_init
acf(data)
```



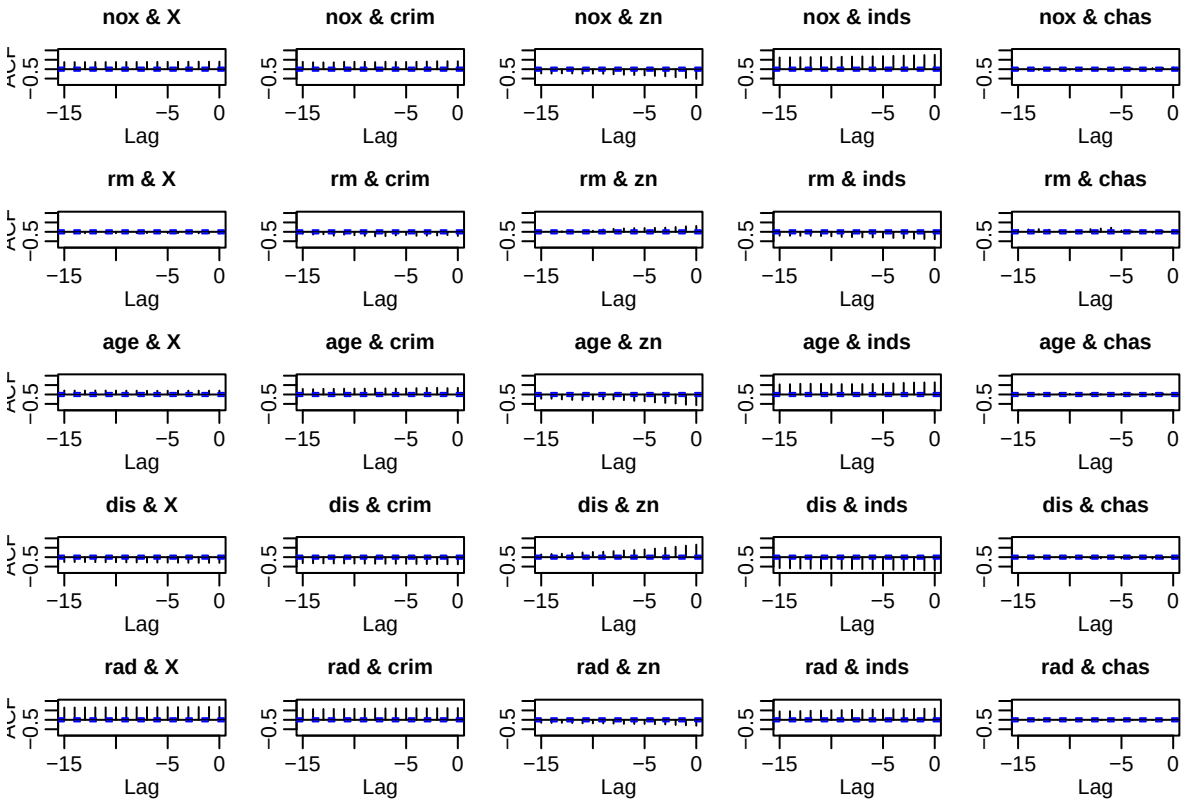
[1 , 1]


X & nox
X & rm
X & age
X & dis
X & rad
crim & nox
crim & rm
crim & age
crim & dis
crim & rad
zn & nox
zn & rm
zn & age
zn & dis
zn & rad
inds & nox
inds & rm
inds & age
inds & dis
inds & rad
chas & nox
chas & rm
chas & age
chas & dis
chas & rad
−0.5
0
5
10
15
Lag

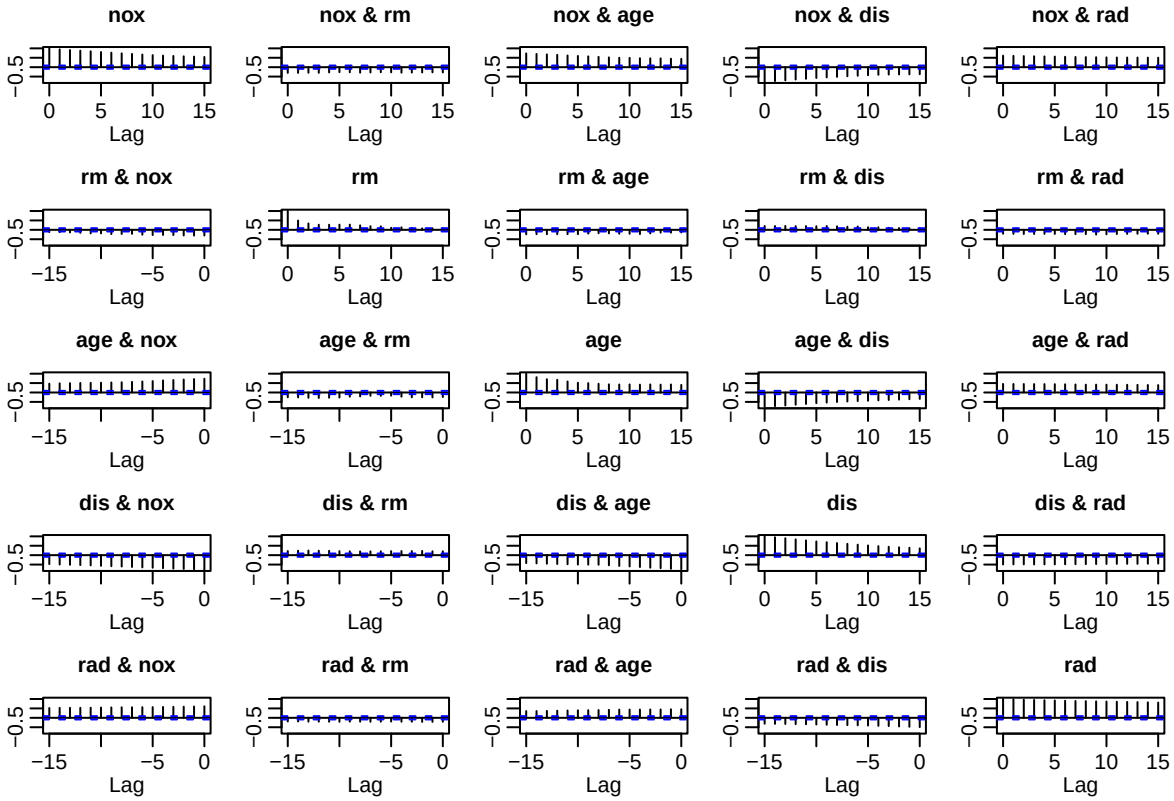
[1 , 2]


X & tax
X & ptrt
X & blck
X & lstt
X & medv
crim & tax
crim & ptrt
crim & blck
crim & lstt
crim & medv
zn & tax
zn & ptrt
zn & blck
zn & lstt
zn & medv
inds & tax
inds & ptrt
inds & blck
inds & lstt
inds & medv
chas & tax
chas & ptrt
chas & blck
chas & lstt
chas & medv
−0.5
0
5
10
15
Lag

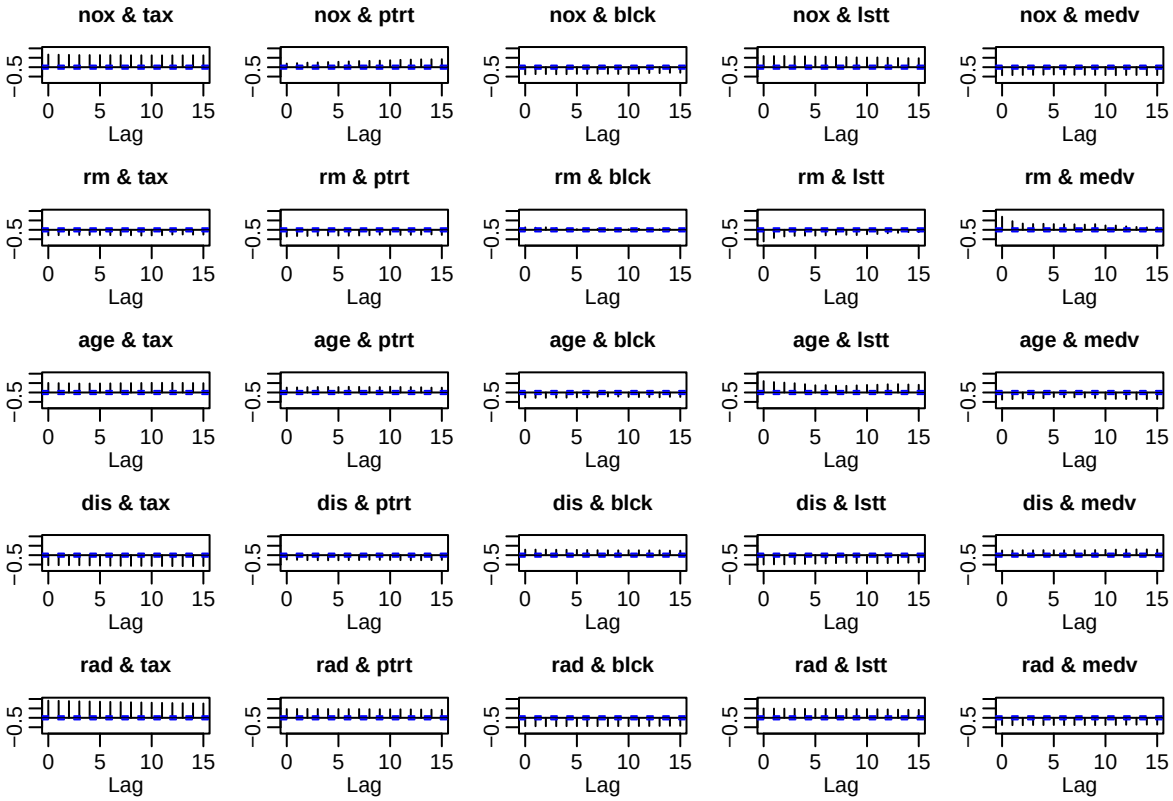
[1 , 3]


nox & X
nox & crim
nox & zn
nox & inds
nox & chas
rm & X
rm & crim
rm & zn
rm & inds
rm & chas
age & X
age & crim
age & zn
age & inds
age & chas
dis & X
dis & crim
dis & zn
dis & inds
dis & chas
rad & X
rad & crim
rad & zn
rad & inds
rad & chas
ACF
−0.5
−15
−5
0
Lag

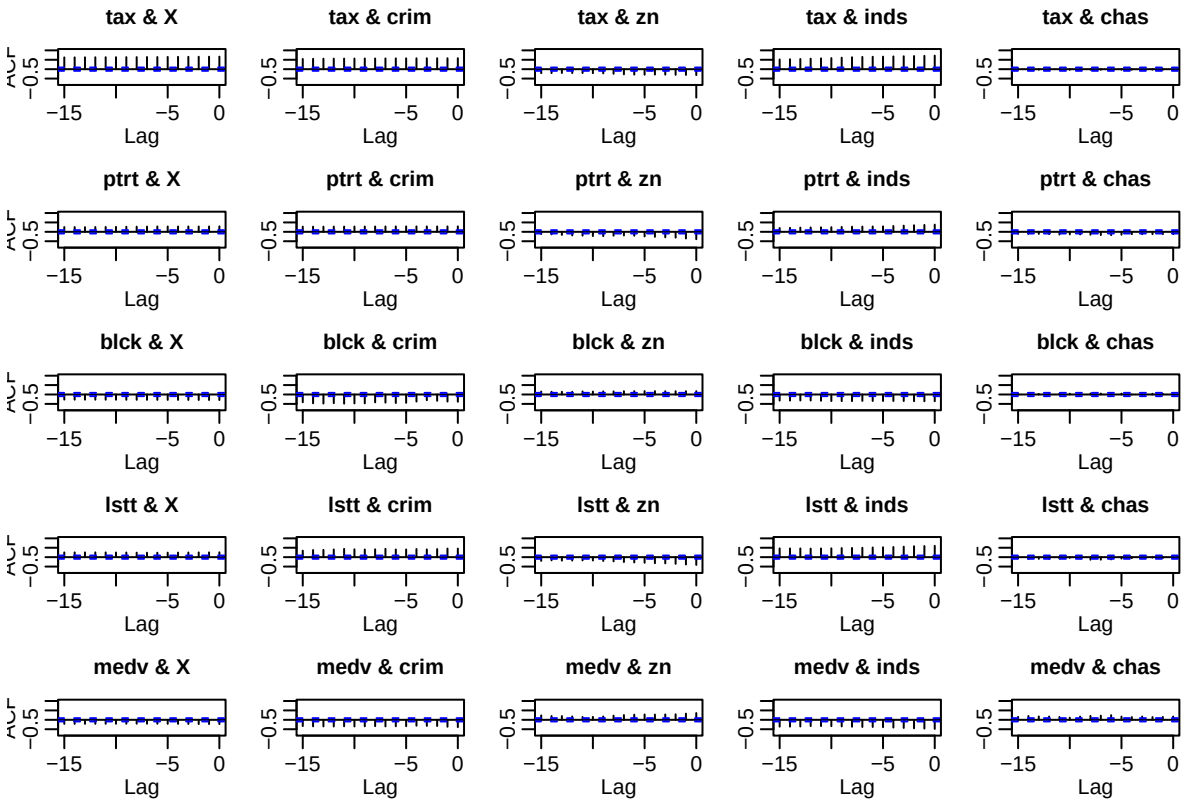
[2 , 1]


nox
nox & rm
nox & age
nox & dis
nox & rad
rm & nox
rm
rm & age
rm & dis
rm & rad
age & nox
age & rm
age
age & dis
age & rad
dis & nox
dis & rm
dis & age
dis
dis & rad
rad & nox
rad & rm
rad & age
rad & dis
rad
−0.5
Lag
0
5
10
15
−15
−5

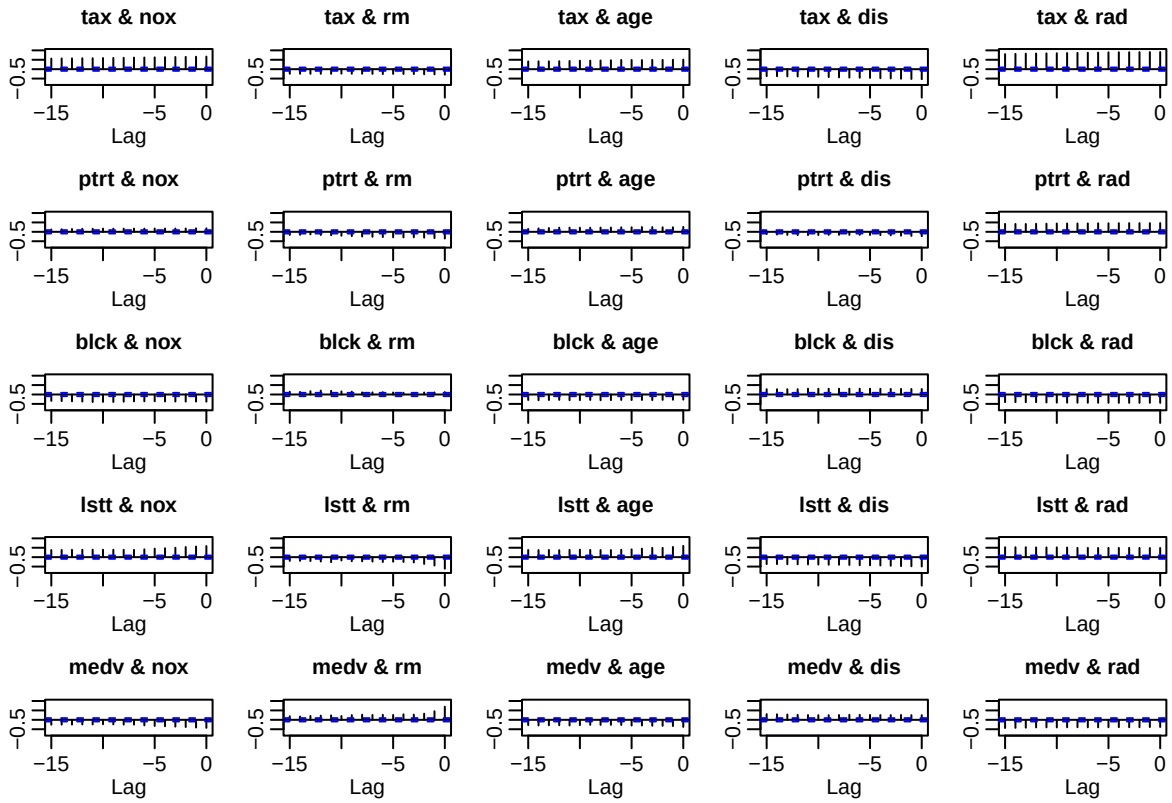
[2 , 2]


nox & tax
nox & ptrt
nox & blck
nox & lstt
nox & medv
rm & tax
rm & ptrt
rm & blck
rm & lstt
rm & medv
age & tax
age & ptrt
age & blck
age & lstt
age & medv
dis & tax
dis & ptrt
dis & blck
dis & lstt
dis & medv
rad & tax
rad & ptrt
rad & blck
rad & lstt
rad & medv
−0.5
0
5
10
15
Lag

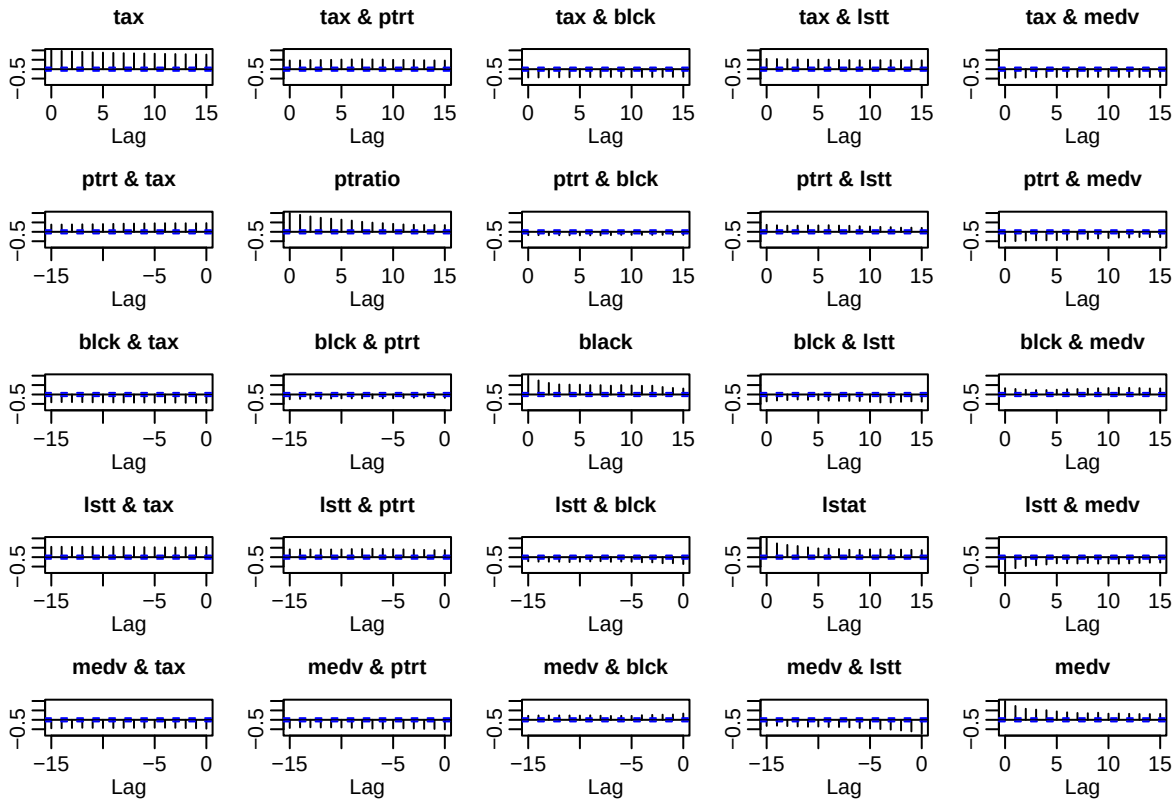
[2 , 3]


tax & X
tax & crim
tax & zn
tax & inds
tax & chas
ptrt & X
ptrt & crim
ptrt & zn
ptrt & inds
ptrt & chas
blck & X
blck & crim
blck & zn
blck & inds
blck & chas
lstt & X
lstt & crim
lstt & zn
lstt & inds
lstt & chas
medv & X
medv & crim
medv & zn
medv & inds
medv & chas
ACF
−0.5
−15
−5
0
Lag

[3 , 1]


tax & nox
tax & rm
tax & age
tax & dis
tax & rad
ptrt & nox
ptrt & rm
ptrt & age
ptrt & dis
ptrt & rad
blck & nox
blck & rm
blck & age
blck & dis
blck & rad
lstt & nox
lstt & rm
lstt & age
lstt & dis
lstt & rad
medv & nox
medv & rm
medv & age
medv & dis
medv & rad
−0.5
−15
−5
0
Lag

[3 , 2]



Autocorrelation graphs show its presence in many places. Perhaps this was the reason for the poor results for the constructed models.

Excluding its influence will improve the quality of the model.

Conclusion

The author analyzed and processed the input data. There have been attempts to build a regression model for the processed data. None of the regressions obtained gave satisfactory results. One of the reasons for this is the presence of autocorrelation. Perhaps, for the presented data, the best solution would be to use a different model specification (for example, machine learning models, neural networks, or others).