Boston Housing prediction

Jie Li

December 25, 2018

Abstract

The goal of this analysis is to fit a regression model that best explains the variation in Boston house price. Various statistical techniques were used to eliminate predictors and extraneous.

The first thing that can be interpreted is that the average number of rooms is positively correlated with house price. Second, there is also a positive correlation if the house is next to the Charles River. It is reasonable that more people like to live closer to the river for the great view on offer and that this should raise the house prices. Similarly, negative correlations with lower status of the population and pupil-teacher ratios are also to be expected. People would prefer to live in areas that have a lower pupil-teacher ratio and high status of population.

Most importantly, higher levels of pollution decrease house prices. People would prefer to live further away from high nitric oxides area. In the years since, there is no doubt that pollution levels have risen and it would be interesting to examine the ways in which that affects house pricing in Boston today.

Boston Housing Dataset

The Boston housing data to be analyzed were collected by Harrison and Rubinfeld in 1978, and contains 506 census tracts of Boston from the 1970 census. The purpose of discovering whether or not clean air influenced the value of houses in Boston.

The data `BostonHousing2` is the corrected version with additional spatial information. Therefore, I am using `BostonHousing2` dataset from `mlbench` library. The following is a brief description of each feature and the outcome in our dataset:

- cmedv - target variable, correctly version median value of owner-occupied homes in USD 1000’s
- 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 centres
- rad - index of accessibility to radial highways
- tax - full-value property-tax rate per USD 10,000
- ptratio- pupil-teacher ratio by town
- b - 1000(Black - 0.63)^2, where Black is the proportion of blacks by town
- lstat - percentage of lower status of the population
- town - name of town
- tract - census tract
- lon - longitude of census tract
- lat - latitude of census tract

```
## 'data.frame':    506 obs. of  19 variables:
##  $ town   : Factor w/ 92 levels "Arlington","Ashland",..: 54 77 77 46 46 46 69 69 69 69 ...
##  $ tract  : int  2011 2021 2022 2031 2032 2033 2041 2042 2043 2044 ...
##  $ lon    : num  -71 -71 -70.9 -70.9 -70.9 ...
##  $ lat    : num  42.3 42.3 42.3 42.3 42.3 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
##  $ cmedv  : num  24 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 ...
##  $ 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   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
```

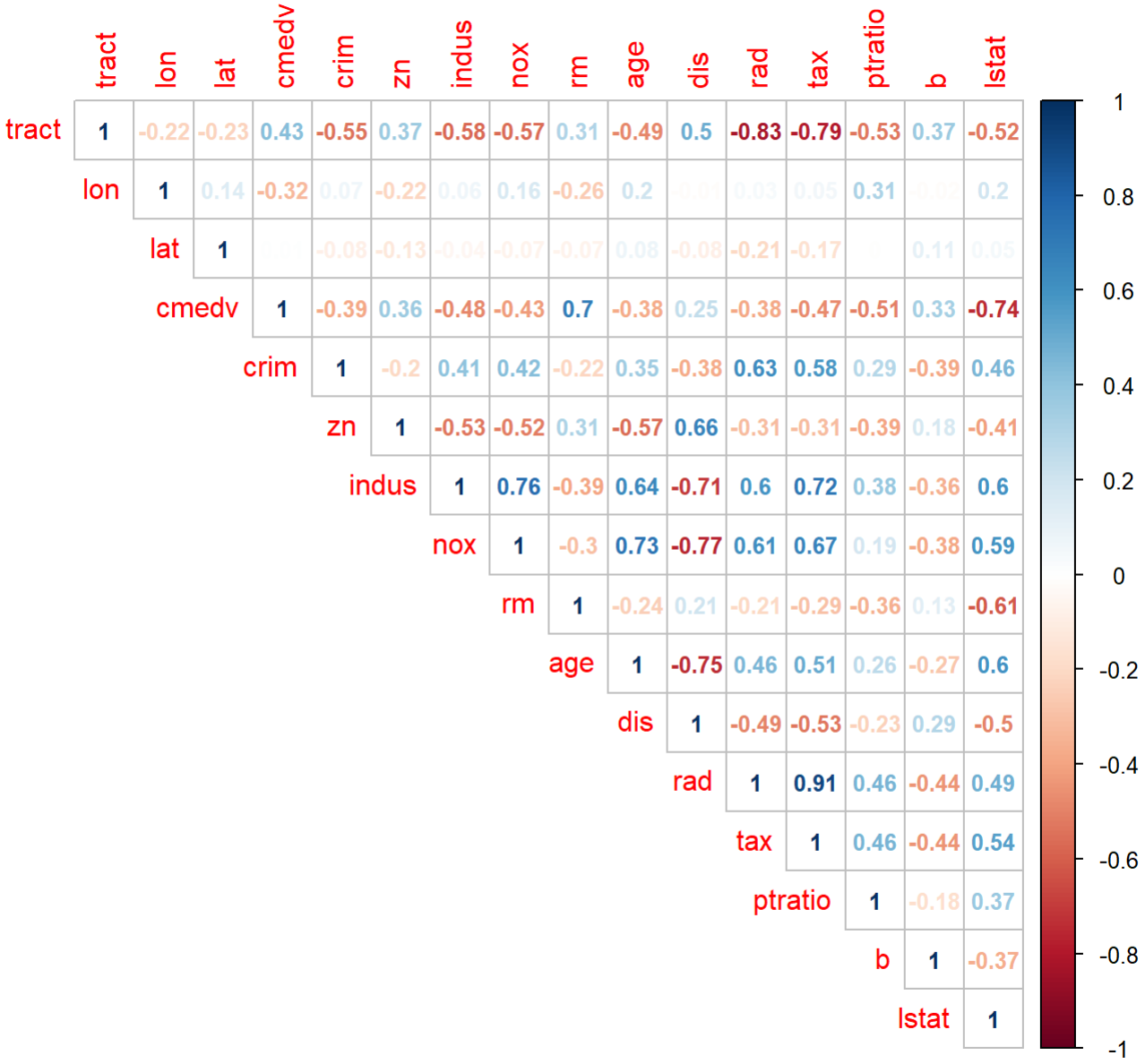
Data Pre-processing

First of all, there is no missing data found in the Boston dataset

```
##    town   tract     lon     lat    medv   cmedv    crim      zn   indus
##       0       0       0       0       0       0       0       0       0
##    chas     nox      rm     age     dis     rad     tax ptratio       b
##       0       0       0       0       0       0       0       0       0
##   lstat
##       0
```

Exploratory Data Analysis

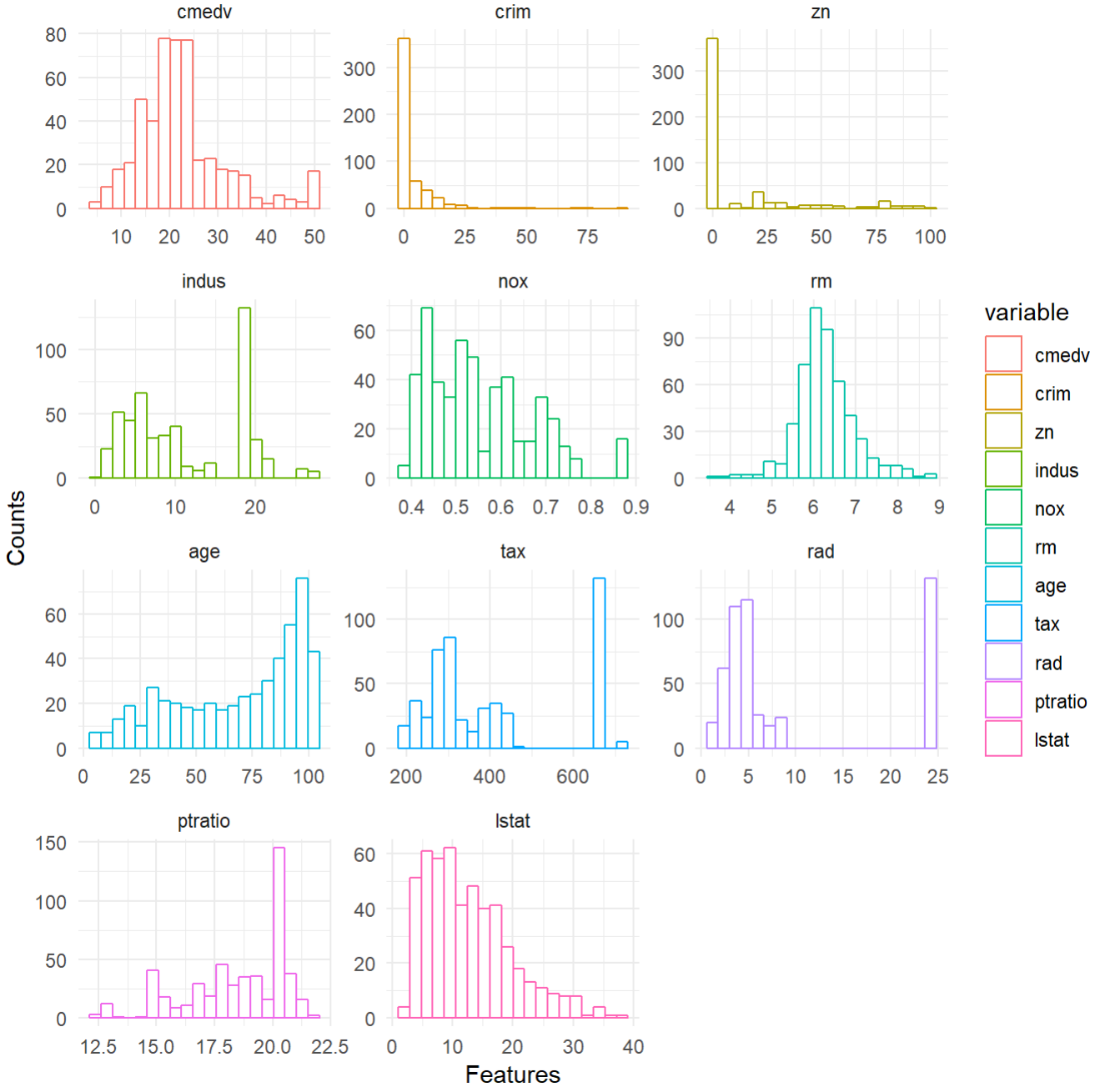
Correlation plots are a great way of exploring data and seeing any features are highly correlated with `cmedv` and any interaction terms.



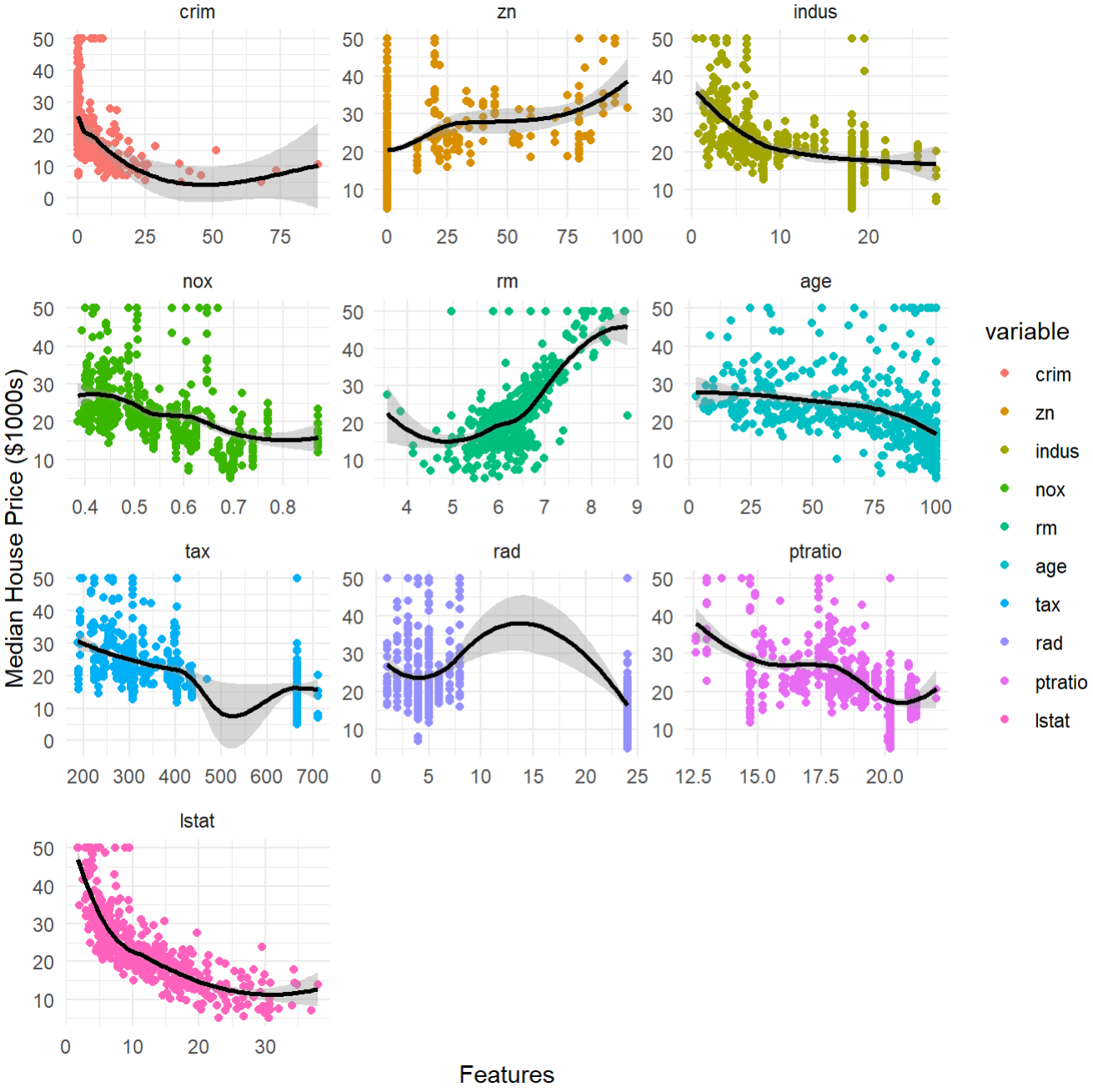
`cmedv` has positive relationship with `tract` (*medium*) `zn` (*medium*), `rm` (*high*), `dis` (*low*) and `b` (*low*)

`cmedv` has negative relationship with `crim` (*medium*), `indus` (*medium*), `nox` (*medium*), `age` (*medium*), `rad` (*medium*), `tax` (*medium*), `ptratio` (*high*), `lstat` , `lon` (*low*) and `lat` (*low*)

Assume that the correlation is lower than 0.33, it means the features did not contain much information to explain `cmedv` . Therefore, I focus on the features have medium and high correlation. The following histograms are showing the distribution of selected features.

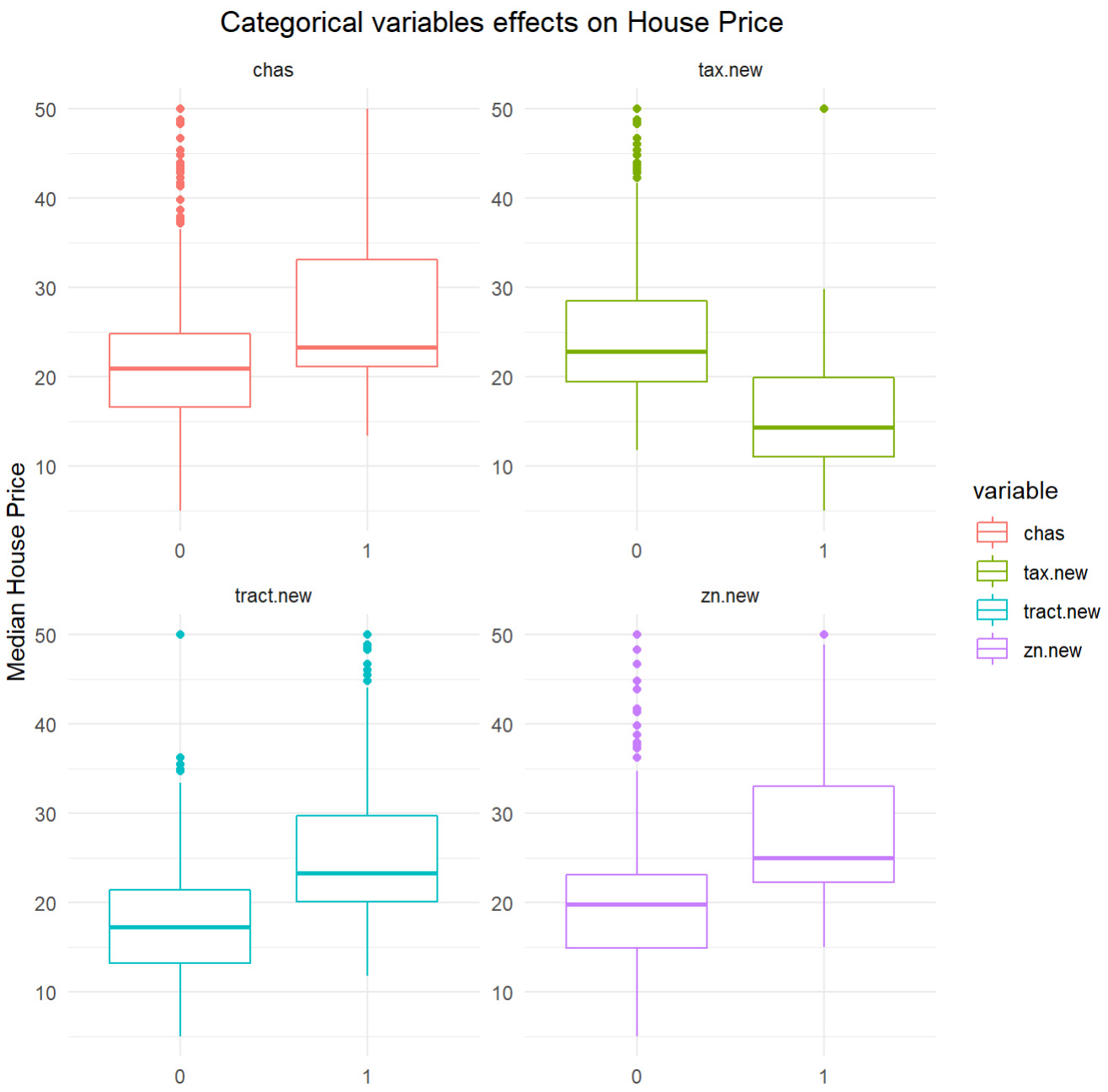


Let's see the scatter plots of `cmedv` versus each feature



The features (`cmedv` , `crim` , `indus` , `nox` , `age` and `lstat`) are not following normal distribution, and exist curvatures (non-linearly association). Also, the features (`zn` , `tax` , `tract`) are showing that the `cmedv` might be depended on two different groups. In addition, `rad` should be treated as multivariate categorical variable.

Create categorical variables to separate two groups, and use box plots and `T-test` to show the significant housing price difference between two groups.




```
##
##  Welch Two Sample t-test
##
## data:  cmedv by chas
## t = -3.1156, df = 36.865, p-value = 0.003546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.480805  -2.220002
## sample estimates:
## mean in group 0 mean in group 1
##         22.0896         28.4400
```

```
##
##  Welch Two Sample t-test
##
## data:  cmedv by tax.new
## t = 10.238, df = 239.45, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.96593 10.28515
## sample estimates:
## mean in group 0 mean in group 1
##        24.86423        16.23869
```

```
##
##  Welch Two Sample t-test
##
## data:  cmedv by tract.new
## t = -10.638, df = 482.06, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.309274 -6.406586
## sample estimates:
## mean in group 0 mean in group 1
##        18.00977        25.86770
```

```
##
##  Welch Two Sample t-test
##
## data:  cmedv by zn.new
## t = -9.1098, df = 240.22, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.415366 -6.067405
## sample estimates:
## mean in group 0 mean in group 1
##        20.47876        28.22015
```

From the box-plots above, two different groups people have significantly difference on the average housing price in Boston. Also, the p-value from `T-test` supports my hypothesis. Therefore, add those categorical features into model.

Transformation regressors & response

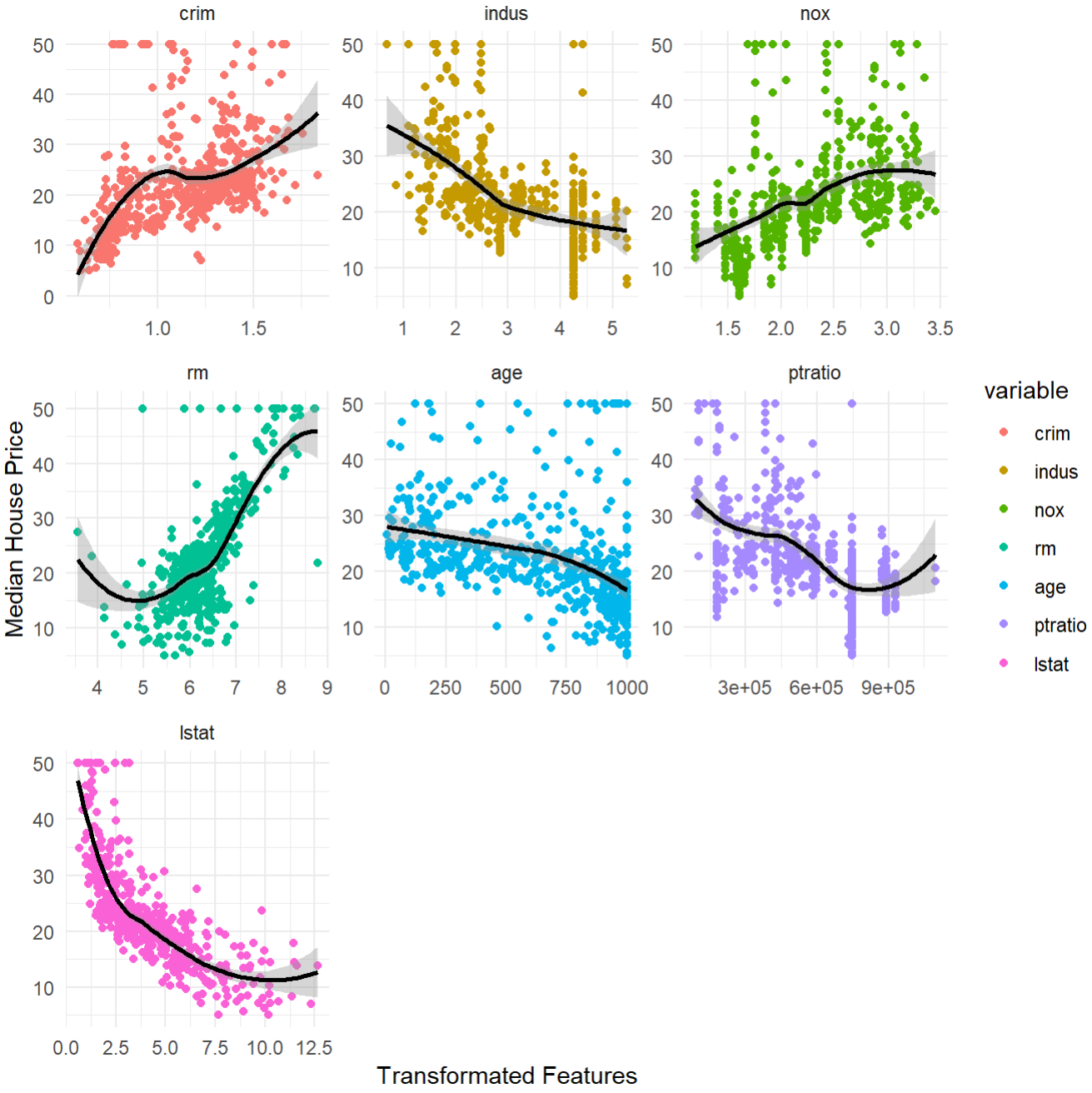
The purpose of transformations is to achieve a mean function that is linear in the transformed scale. The `Box-Cox` method provided a general method for selecting a transformation from a family indexed by a parameter λ. It is not transforming for linearity, but rather it is transforming for normality since X is multivariate normal is much stronger than linearly related regressors.

Use `Box-Cox` power transformation on the features to make for a better fit.

```
## bcPower Transformations to Multinormality
##         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## crim      -0.1224       -0.12      -0.1563      -0.0885
## indus      0.4032        0.50       0.3062       0.5003
## nox       -1.3391       -1.34      -1.6531      -1.0252
## rm         1.0258        1.00       0.6228       1.4288
## age        1.4720        1.47       1.3021       1.6419
## ptratio    4.5219        4.52       3.7435       5.3004
## lstat      0.2313        0.33       0.1257       0.3369
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0) 848.1541  7 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1) 3318.489  7 < 2.22e-16
```

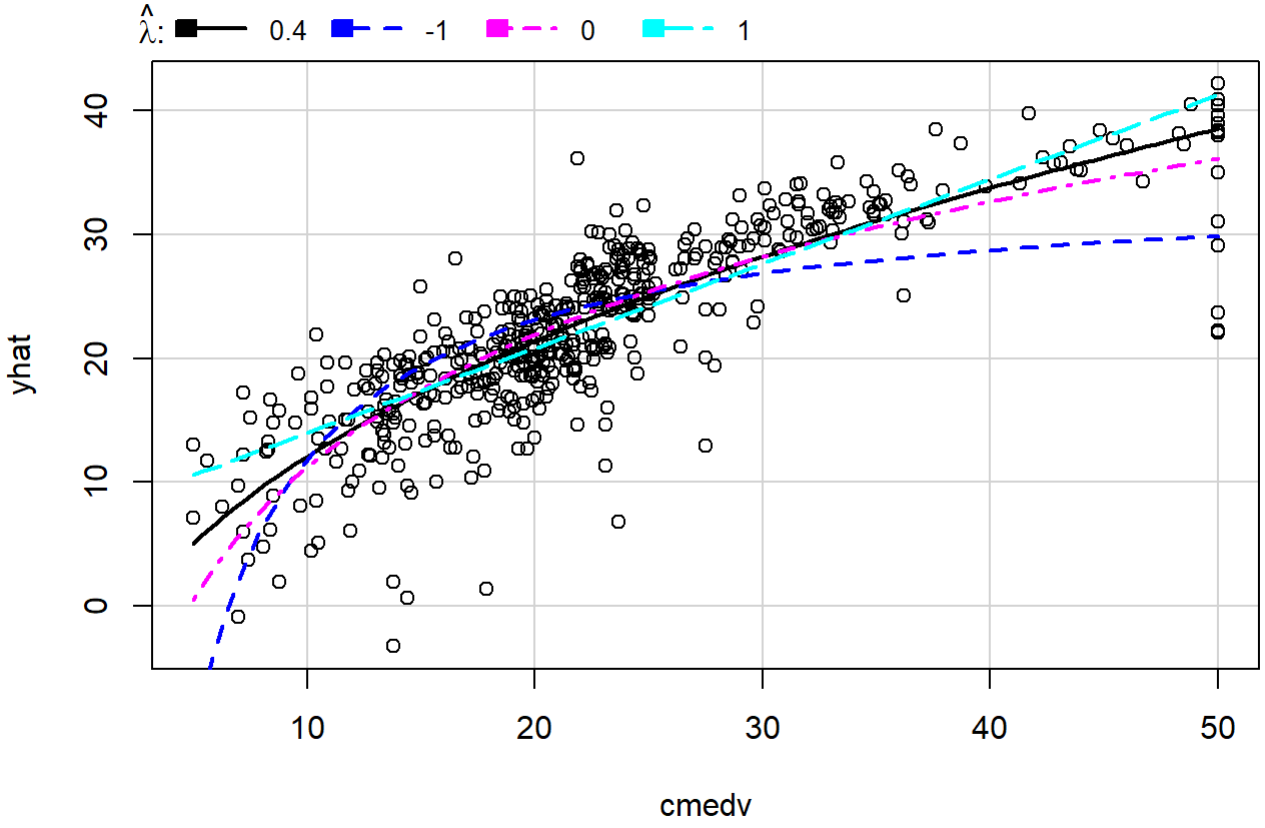
```
##    crim   indus     nox      rm     age ptratio   lstat
##   -0.12    0.50   -1.30    1.00    1.50    4.50    0.33
```

```
##                                                       LRT df    pval
## LR test, lambda = (-0.12 0.5 -1.3 1 1.5 4.5 0.33) 7.237873  7 0.40454
```

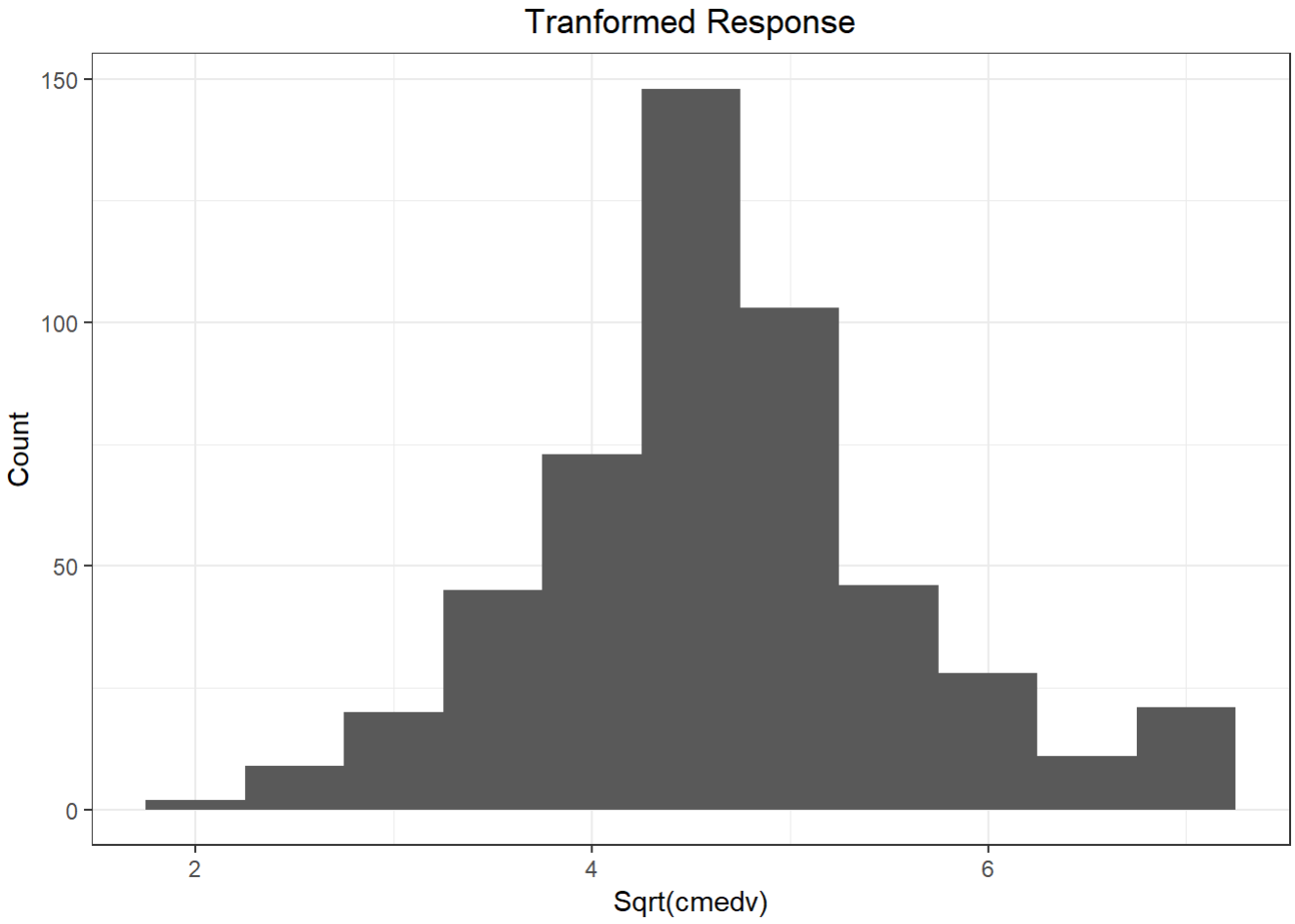


Based on `Box-Cox` method and `likelihood ratio test (LRT)` above, there is no evidence to reject our estimated power transformation (p-value > 0.05). Then, the scatter plots of transformated features are showing that it fixs some curvatures (`crim` , `indus` , `nox` and `age`), but `rm` and `lstat` still exist curvatures, so add second-order regressors such as polynomial or interaction terms in the model.

Once the predictors are transformed, let's transform the response by inverse fitted value plot




```
##        lambda       RSS
## 1   0.4049464  8454.053
## 2  -1.0000000 13284.567
## 3   0.0000000  8873.525
## 4   1.0000000  9218.243
```



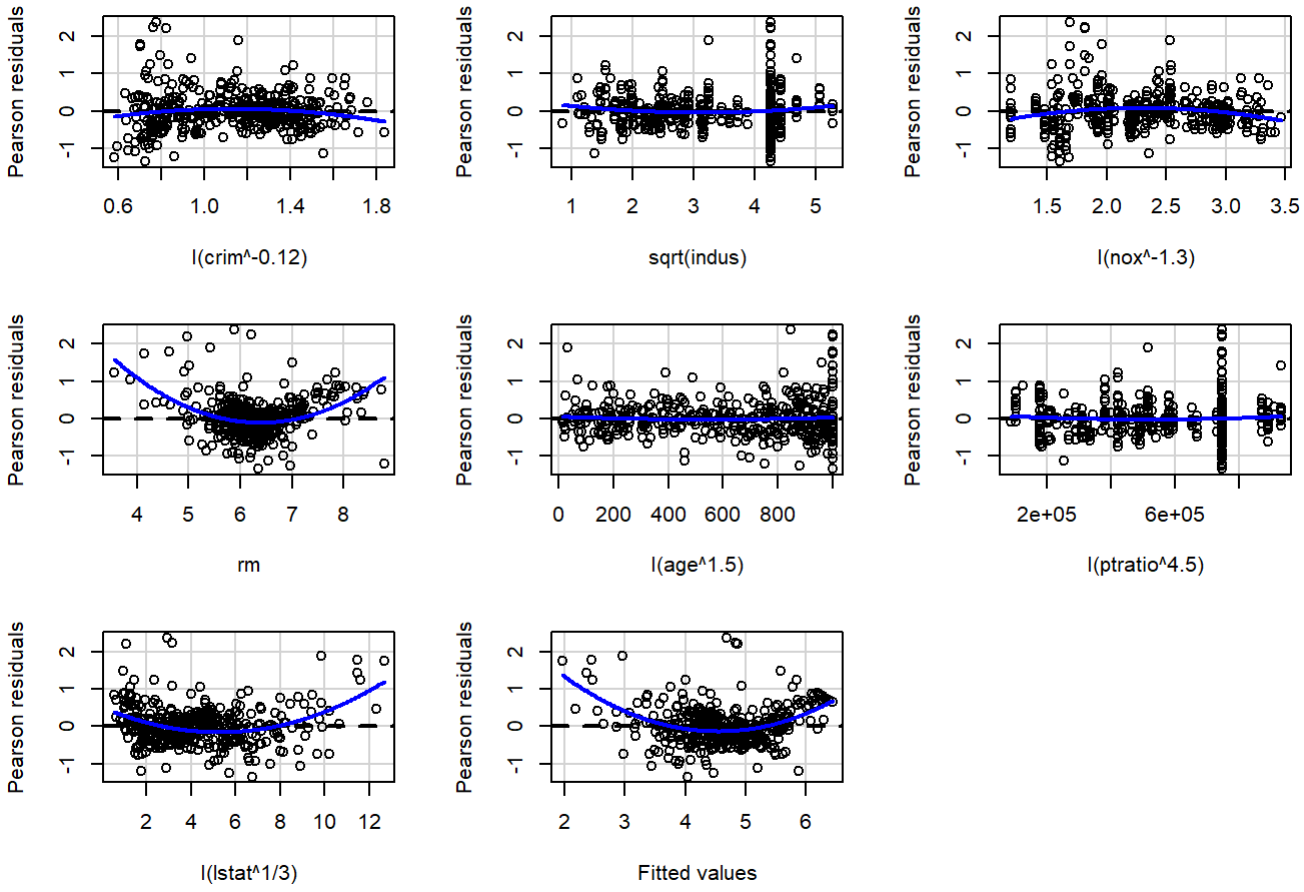
Based on the table above, when $\lambda = 0.4$, the residuals sum of squared (`RSS`) is the lowest. To be simple, take square root of `cmedv` is approximately following the normal distribution.

Modeling

Split our Boston data into train set (80%) and test set (20%). Let's build a initial linear regression model M_1

$$M_1 : E[\sqrt{cmedv}|x] = crim^{-0.12} + \sqrt{indus} + nox^{-1.3} + rm + age^{1.5} + ptratio^{4.5} + lstat^{1/3}$$

The residual plots and fitted value plot help us to detect non-constant variance and non-linearity. If the model is correct, residual plots should expect null plots.

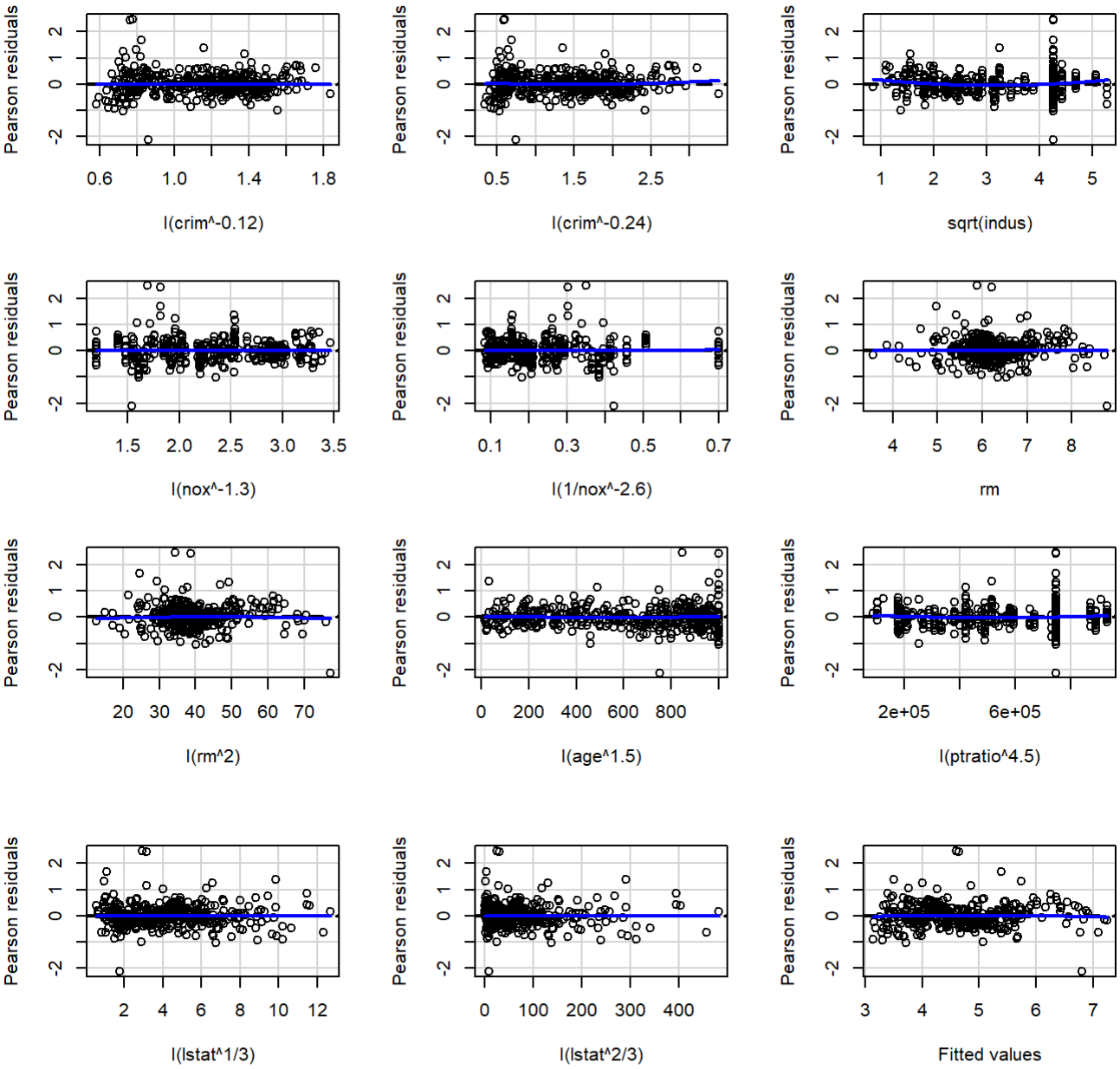



```
##                Test stat Pr(>|Test stat|)
## I(crim^-0.12)    -2.1370        0.0332108 *
## sqrt(indus)       1.6588        0.0979520 .
## I(nox^-1.3)      -3.4319        0.0006627 ***
## rm                9.6871        < 2.2e-16 ***
## I(age^1.5)        0.8299        0.4070714
## I(ptratio^4.5)    1.0195        0.3086088
## I(lstat^1/3)      8.3441        1.216e-15 ***
## Tukey test       10.5386        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I assume that the constant variance is hold, but there are four regressors (`crim` , `nox` , `rm` , `lstat`) have significantly non-linear, so adding polynomial terms in the models.

Re-fit the linear regression model with second-order terms called M_2

$$M_2 : E[\sqrt{cmedv}|x] = crim^{-0.12} + crim^{-0.24} + \sqrt{indus} + nox^{-1.3} + nox^{-2.6} + rm + rm^2 + age^{1.5} + ptratio^{4.5} + lstat^{1/3} + lstat^{2/3}$$


Pearson residuals
-2 0 1 2
0.6 1.0 1.4 1.8
I(crim^-0.12)
0.5 1.5 2.5
I(crim^-0.24)
1 2 3 4 5
sqrt(indus)
1.5 2.0 2.5 3.0 3.5
I(nox^-1.3)
0.1 0.3 0.5 0.7
I(1/nox^-2.6)
4 5 6 7 8
rm
20 30 40 50 60 70
I(rm^2)
0 200 400 600 800
I(age^1.5)
2e+05 6e+05
I(ptratio^4.5)
2 4 6 8 10 12
I(lstat^1/3)
0 100 200 300 400
I(lstat^2/3)
3 4 5 6 7
Fitted values

```
##                Test stat Pr(>|Test stat|)
## I(crim^-0.12)     0.8700        0.3848100
## I(crim^-0.24)     3.4703        0.0005778 ***
## sqrt(indus)       2.4806        0.0135343 *
## I(nox^-1.3)      -0.3642        0.7159004
## I(1/nox^-2.6)     1.0442        0.2970574
## rm                0.1163        0.9074734
## I(rm^2)          -2.2413        0.0255641 *
## I(age^1.5)        0.2977        0.7660920
## I(ptratio^4.5)    1.1369        0.2562727
## I(lstat^1/3)     -2.4100        0.0164140 *
## I(lstat^2/3)     -0.0763        0.9392544
## Tukey test       -1.0632        0.2876765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of `Tukey test` is given 0.48 > 0.05, and most regressors are significantly linear.

Add categorical variables `Tax` and `ANOVA` tests shows p-value

Add categorical variables `Tract` and `ANOVA` tests shows p-value

Add categorical variables `Chas` and `ANOVA` tests shows p-value

Add categorical variables `zn` and `ANOVA` tests shows p-value

Add categorical variables `rad` and `ANOVA` tests shows p-value

```
## Analysis of Variance Table
##
## Model 1: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(1/nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3)
## Model 2: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3) + factor(tax.new)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    392 73.406
## 2    391 72.909  1   0.49725 2.6667 0.1033
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(1/nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3)
## Model 2: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3) + factor(tract.new)
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    392 73.406
## 2    391 73.488  1 -0.081701
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(1/nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3)
## Model 2: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3) + factor(chas)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    392 73.406
## 2    391 71.988  1    1.4175 7.6991 0.00579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(1/nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3)
## Model 2: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3) + factor(zn.new)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    392 73.406
## 2    391 73.404  1 0.0014326 0.0076 0.9304
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(1/nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3)
## Model 2: sqrt(cmedv) ~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) +
##     I(nox^-2.6) + rm + I(rm^2) + I(age^1.5) + I(ptratio^4.5) +
##     I(lstat^1/3) + I(lstat^2/3) + factor(rad)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    392 73.406
## 2    384 70.404  8    3.0023 2.0469 0.04014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the `ANOVA` table, the p-value of `chas` is close to zero, so it provides strong evidence to reject the null hypothesis that `chas` is contributed on housing price. Therefore, adding these two categorical regressor into model M_2

However, the residuals analysis claims that `crim`, `rm` and `indus` have significantly curvatures, even added polynomial terms. Adding interaction terms with features above, and using `Type II ANOVA` to verify the results.

$$E[\sqrt{cmedv}|x] = crim^{-0.12} + crim^{-0.24} + \sqrt{indus} + indus + nox^{-1.3} + nox^{-2.6} + rm + rm^2 + age^{1.5} + ptratio^{4.5} + lstat^{1/3} + lstat^{2/3} + rm * chas + \sqrt{indus} * chas + crim^{-0.12} * chas + chas$$

```
## Analysis of Variance Table
##
## Response: sqrt(cmedv)
##                             Df Sum Sq Mean Sq  F value    Pr(>F)
## I(crim^-0.12)                1 81.888  81.888 460.0149 < 2.2e-16 ***
## I(crim^-0.24)                1  7.199   7.199  40.4440 5.696e-10 ***
## sqrt(indus)                  1 20.071  20.071 112.7505 < 2.2e-16 ***
## I(nox^-1.3)                  1  0.027   0.027   0.1531  0.695818
## I(nox^-2.6)                  1  0.306   0.306   1.7182  0.190704
## rm                           1 82.377  82.377 462.7641 < 2.2e-16 ***
## I(rm^2)                      1 25.994  25.994 146.0274 < 2.2e-16 ***
## I(age^1.5)                   1  3.539   3.539  19.8826 1.080e-05 ***
## I(ptratio^4.5)               1 11.179  11.179  62.8015 2.448e-14 ***
## I(lstat^1/3)                 1 32.591  32.591 183.0867 < 2.2e-16 ***
## I(lstat^2/3)                 1  4.921   4.921  27.6463 2.411e-07 ***
## factor(chas)                 1  1.499   1.499   8.4226  0.003917 **
## rm:factor(chas)              1  1.702   1.702   9.5626  0.002129 **
## sqrt(indus):factor(chas)     1  0.555   0.555   3.1169  0.078272 .
## I(crim^-0.12):factor(chas)   1  0.663   0.663   3.7243  0.054355 .
## Residuals                  388 69.068   0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Selection

Now, use `stepwise` and `best subset selection` with AIC, BIC and mallow's cp criteria to select our models

Choosen the lowest AIC scores, `Model.AIC` shows $R^2 = 0.786$, and all regressors are significantly contributed.

$$Model.\ AIC = E[\sqrt{cmedv}|x] = rm + rm^2 + age^{1.5} + ptratio^{4.5} + lstat^{1/3} + lstat^{2/3} + nox^{-1.3} + nox^{-2.6} + rm * chas + chas$$

```
summary(model.AIC)
```

```
##
## Call:
## lm(formula = sqrt(cmedv) ~ rm + I(lstat^1/3) + I(rm^2) + I(ptratio^4.5) +
##     I(lstat^2/3) + factor(chas) + I(age^1.5) + I(nox^-1.3) +
##     I(nox^-2.6) + rm:factor(chas), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75134 -0.23572 -0.01776  0.21600  2.38682
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.845e+00  1.069e+00   9.208  < 2e-16 ***
## rm                -2.028e+00  3.291e-01  -6.162 1.78e-09 ***
## I(lstat^1/3)      -4.173e-01  4.508e-02  -9.257  < 2e-16 ***
## I(rm^2)            1.837e-01  2.607e-02   7.048 8.21e-12 ***
## I(ptratio^4.5)    -6.209e-07  1.092e-07  -5.684 2.57e-08 ***
## I(lstat^2/3)       5.806e-03  1.199e-03   4.841 1.86e-06 ***
## factor(chas)1      2.290e+00  6.603e-01   3.468 0.000582 ***
## I(age^1.5)         1.851e-04  1.238e-04   1.496 0.135451
## I(nox^-1.3)        1.570e+00  3.225e-01   4.866 1.65e-06 ***
## I(nox^-2.6)       -3.260e-01  7.084e-02  -4.602 5.65e-06 ***
## rm:factor(chas)1 -3.071e-01  1.000e-01  -3.071 0.002284 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4268 on 393 degrees of freedom
## Multiple R-squared:  0.7917, Adjusted R-squared:  0.7864
## F-statistic: 149.4 on 10 and 393 DF,  p-value: < 2.2e-16
```

Choosen the lowest BIC -567.735 and mallow's cp scores 28.203, `Model.BIC` shows $R^2 = 0.781$

$$Model.BIC = E[\sqrt{cmedv}|x] = rm + rm^2 + age^{1.5} + ptratio^{4.5} + lstat^{1/3} + lstat^{2/3} + chas$$

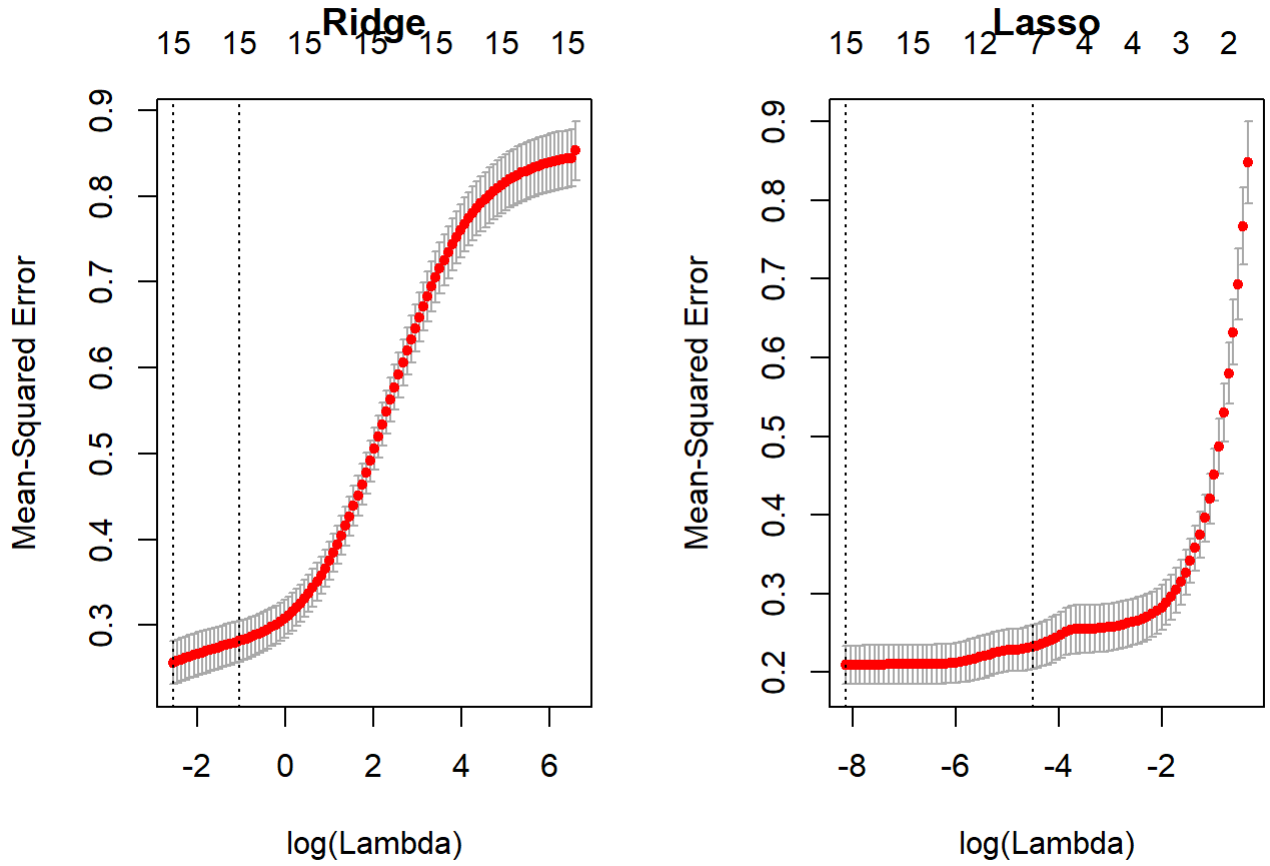
```
##
## Call:
## lm(formula = sqrt(cmedv) ~ I(nox^-1.3) + I(nox^-2.6) + rm + I(rm^2) +
##     I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + factor(chas),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30358 -0.22941 -0.01759  0.22016  2.43764
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.609e+00  1.057e+00   9.087  < 2e-16 ***
## I(nox^-1.3)     1.586e+00  3.258e-01   4.868 1.63e-06 ***
## I(nox^-2.6)    -3.440e-01  7.032e-02  -4.892 1.46e-06 ***
## rm             -1.891e+00  3.241e-01  -5.835 1.12e-08 ***
## I(rm^2)         1.713e-01  2.537e-02   6.752 5.24e-11 ***
## I(ptratio^4.5) -6.625e-07  1.097e-07  -6.039 3.59e-09 ***
## I(lstat^1/3)   -3.951e-01  4.385e-02  -9.011  < 2e-16 ***
## I(lstat^2/3)    5.414e-03  1.195e-03   4.530 7.83e-06 ***
## factor(chas)1   2.885e-01  9.019e-02   3.199  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.432 on 395 degrees of freedom
## Multiple R-squared:  0.7854, Adjusted R-squared:  0.7811
## F-statistic: 180.7 on 8 and 395 DF,  p-value: < 2.2e-16
```

Add regularization to perform `Ridge` and `Lasso` models

```
# Lasso
x.train = model.matrix(~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) + I(nox^-2.6) + rm + I
(rm^2) + I(age^1.5) + I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + rm:factor(chas) + sqrt(indus):factor
(chas) + I(crim^-0.12):factor(chas) + factor(chas), train)

x.test = model.matrix(~ I(crim^-0.12) + I(crim^-0.24) + sqrt(indus) + I(nox^-1.3) + I(nox^-2.6) + rm + I(r
m^2) + I(age^1.5) + I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + rm:factor(chas) + sqrt(indus):factor(ch
as) + I(crim^-0.12):factor(chas) + factor(chas), test)

y.train = sqrt(train[, "cmedv"])
y.test = sqrt(test[, "cmedv"])

model.ridge.cv = cv.glmnet(x=x.train, y=y.train, type.measure = "mse", alpha = 0, nfolds = 10)
model.lasso.cv = cv.glmnet(x=x.train, y=y.train, type.measure = "mse", alpha = 1, nfolds = 10)
```

Use `cross validation` to choose optimal λ 0.351 and 0.011

```
par(mfrow = c(1,2))
plot(model.ridge.cv, main = "Ridge")
plot(model.lasso.cv, main = "Lasso")
```



Let's evaluate our four models by square root of mean square error RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y - \hat{y}_{pred})^2}$$

Prediction & Model selection

```
## Linear Regression
##
## 404 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 364, 364, 364, 362, 363, 365, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.4404066  0.7841329  0.3142545
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 404 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 363, 365, 363, 364, 362, 364, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.4302916  0.7902331  0.3094658
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## RMSE of Ridge: 0.5247687
```

```
## RMSE of Lasso: 0.4752173
```

Based on the `RMSE`, I claim that the model `Model.BIC` is the best fit.

$$Model.BIC = E[\sqrt{cmedv}|x] = rm + rm^2 + ptratio^{4.5} + lstat^{1/3} + lstat^{2/3} + nox^{-1.3} + nox^{-2.6} + chas$$

Show `Model.BIC` performance on test set

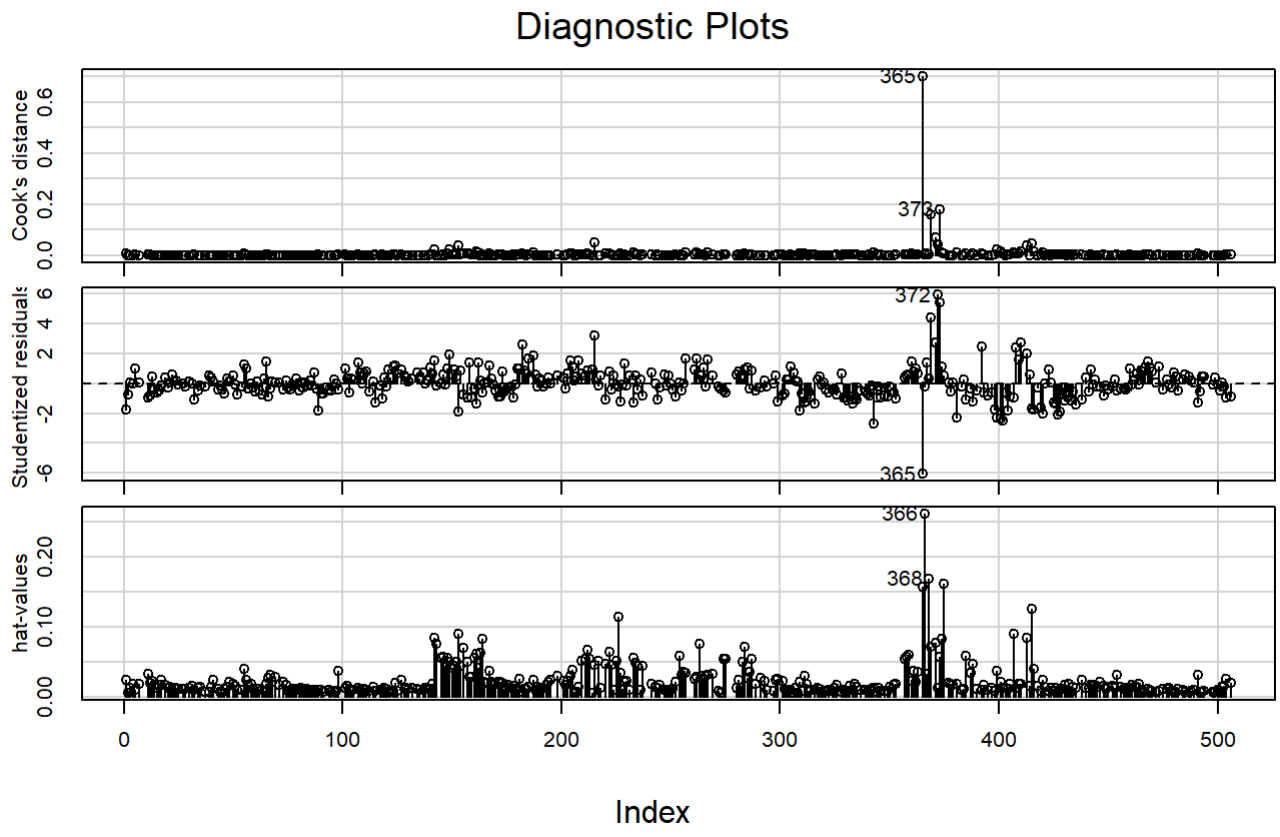
```
model.test = lm(sqrt(cmedv) ~ I(nox^-1.3) + I(nox^-2.6) + rm + I(rm^2) + I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + factor(chas), test)

summary(model.test)
```

```
##
## Call:
## lm(formula = sqrt(cmedv) ~ I(nox^-1.3) + I(nox^-2.6) + rm + I(rm^2) +
##     I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + factor(chas),
##     data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10380 -0.26377 -0.03414  0.24177  1.53865
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.215e+01  3.149e+00   3.858 0.000211 ***
## I(nox^-1.3)     2.335e+00  6.501e-01   3.591 0.000528 ***
## I(nox^-2.6)    -4.724e-01  1.392e-01  -3.394 0.001015 **
## rm             -3.034e+00  8.922e-01  -3.400 0.000993 ***
## I(rm^2)         2.592e-01  6.678e-02   3.881 0.000194 ***
## I(ptratio^4.5) -8.171e-07  2.032e-07  -4.021 0.000118 ***
## I(lstat^1/3)   -3.200e-01  8.493e-02  -3.768 0.000288 ***
## I(lstat^2/3)    3.343e-03  2.228e-03   1.501 0.136838
## factor(chas)1   3.916e-01  1.472e-01   2.661 0.009181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4096 on 93 degrees of freedom
## Multiple R-squared:   0.84,  Adjusted R-squared:  0.8262
## F-statistic: 61.03 on 8 and 93 DF,  p-value: < 2.2e-16
```

Model Diagnostics

Next, run model diagnostics to check the assumptions of linear regression that errors have a constant variance, are normally distributed and not independent.

Apply Cook’s distance to find outliers and influent points.

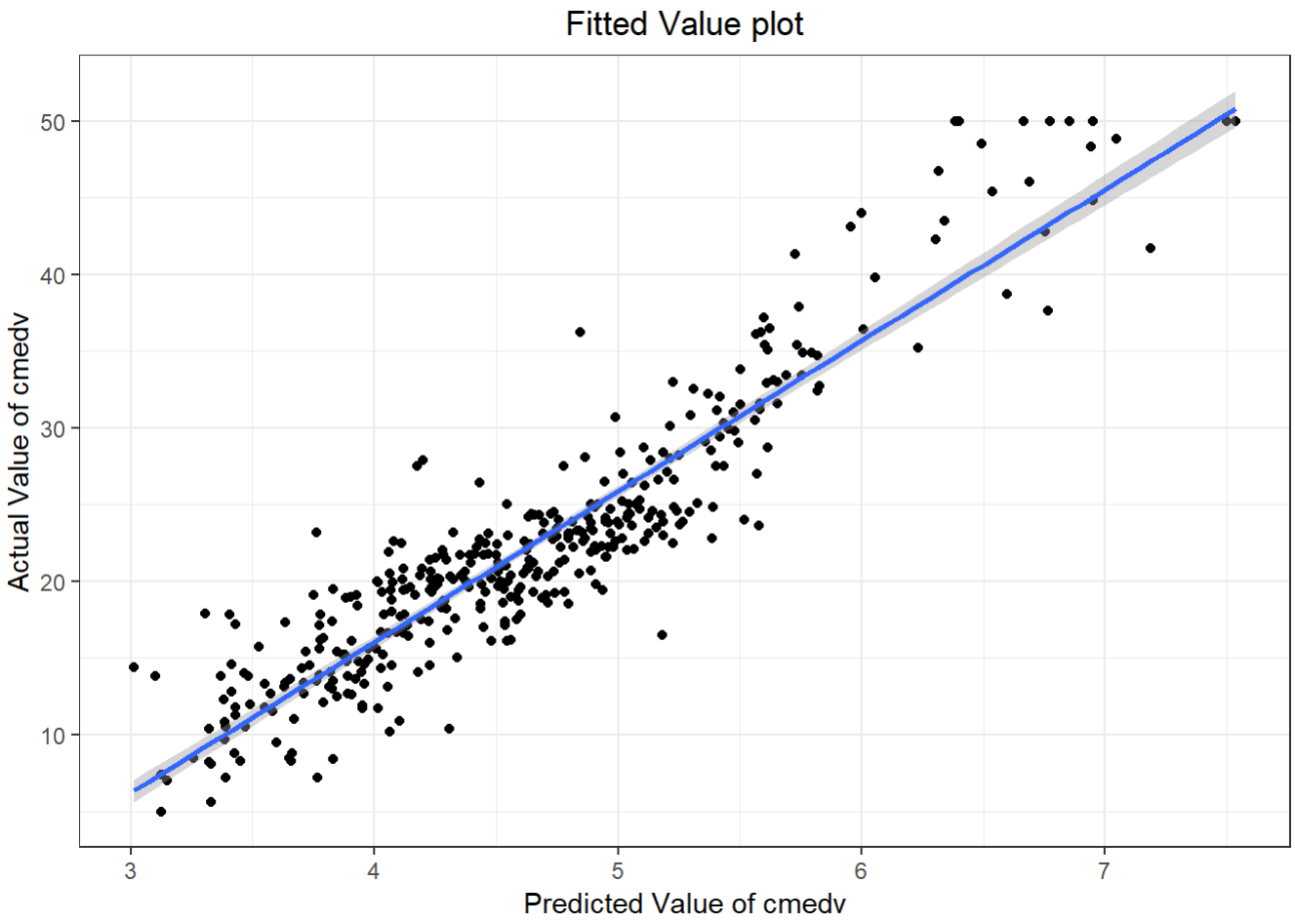



```
##      rstudent unadjusted p-value Bonferonni p
## 365 -6.066679         3.0662e-09   1.2387e-06
## 372  5.914713         7.2197e-09   2.9167e-06
## 373  5.415081         1.0675e-07   4.3127e-05
## 369  4.392461         1.4423e-05   5.8267e-03
## 215  3.171129         1.6373e-03   6.6148e-01
```

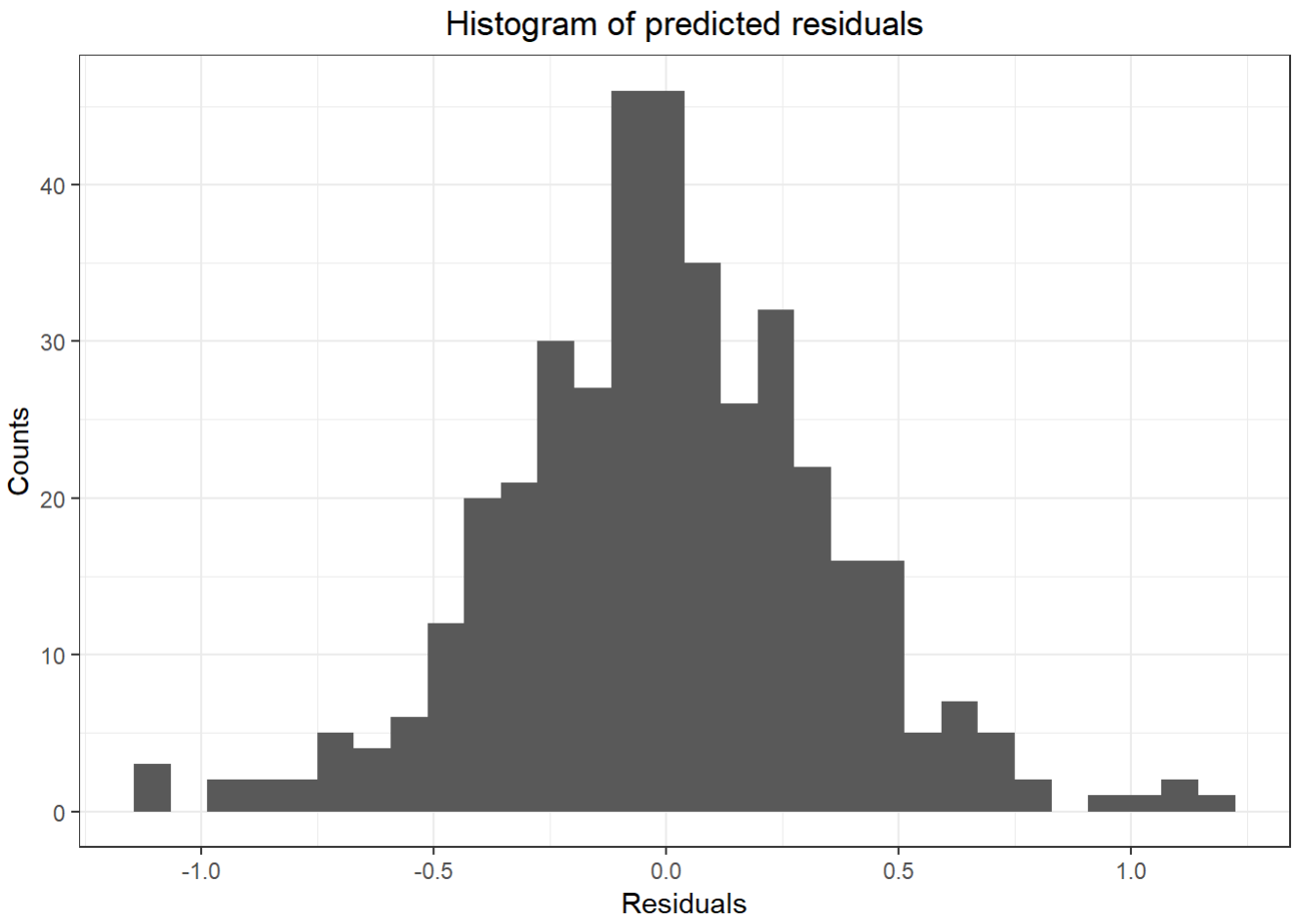
Based on the Cook's distance plots and outlier tests, it gives five outliers which have signifianctly evidences. Removed those five outliers and re-fit the `Model.AIC` model.`

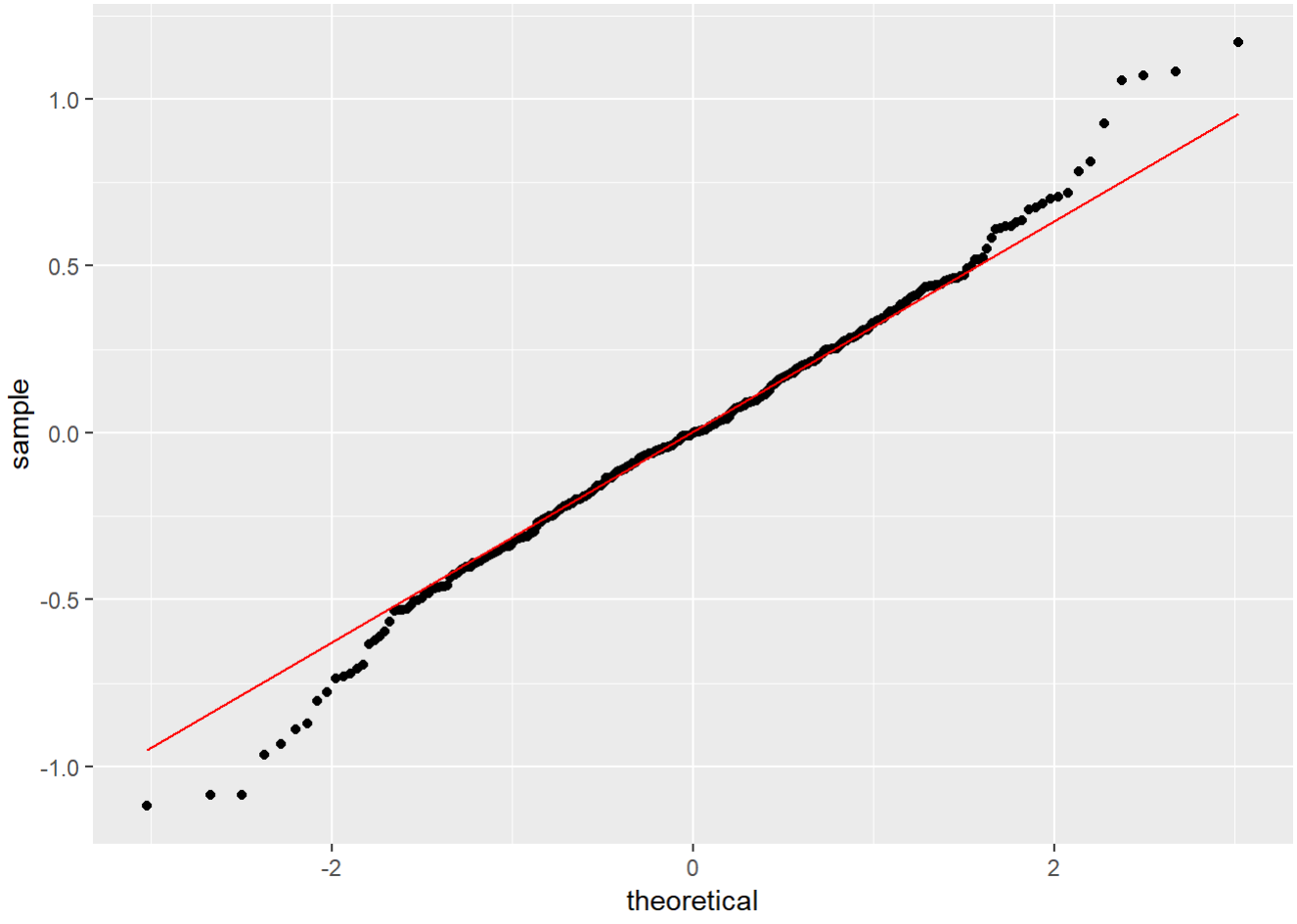
```
##
## Call:
## lm(formula = sqrt(cmedv) ~ I(nox^-1.3) + I(nox^-2.6) + rm + I(rm^2) +
##     I(ptratio^4.5) + I(lstat^1/3) + I(lstat^2/3) + factor(chas),
##     data = train.rm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11883 -0.21117 -0.00053  0.21466  1.16943
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.127e+01  9.081e-01  12.408  < 2e-16 ***
## I(nox^-1.3)     1.403e+00  2.725e-01   5.148 4.19e-07 ***
## I(nox^-2.6)    -2.913e-01  5.855e-02  -4.976 9.78e-07 ***
## rm             -2.555e+00  2.765e-01  -9.241  < 2e-16 ***
## I(rm^2)         2.307e-01  2.175e-02  10.607  < 2e-16 ***
## I(ptratio^4.5) -7.117e-07  9.249e-08  -7.695 1.18e-13 ***
## I(lstat^1/3)   -2.831e-01  3.744e-02  -7.561 2.93e-13 ***
## I(lstat^2/3)    2.714e-03  1.008e-03   2.692  0.00740 **
## factor(chas)1   2.607e-01  7.957e-02   3.277  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3544 on 388 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8444
## F-statistic: 269.5 on 8 and 388 DF,  p-value: < 2.2e-16
```

Let's create a fitted value plot to evaluate the result

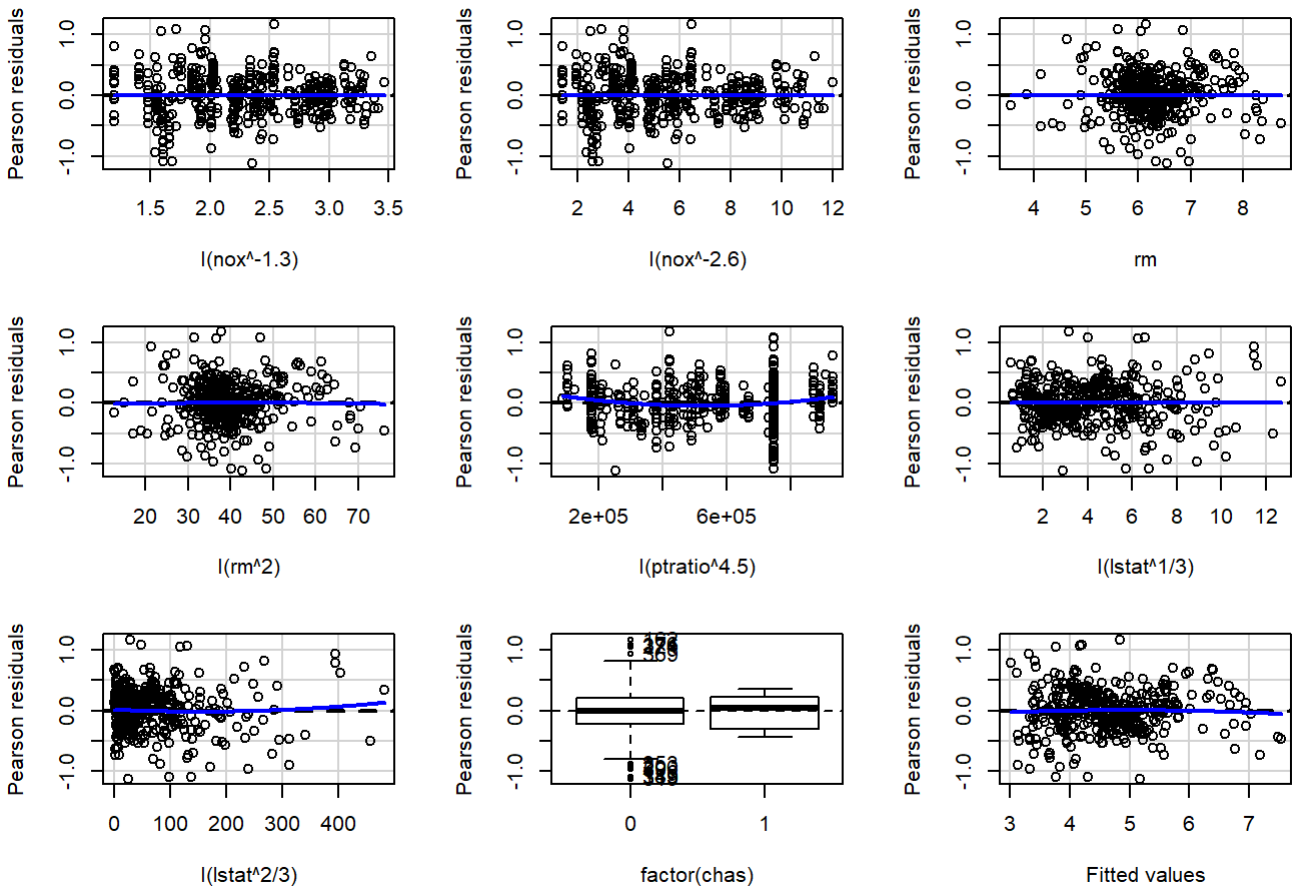


Make a histgorm and Q-Q plot of residuals to check normality assumption


Histogram of predicted residuals
Counts
40
30
20
10
0
-1.0
-0.5
0.0
0.5
1.0
Residuals


sample
1.0
0.5
0.0
-0.5
-1.0
-2
0
2
theoretical

Make residuals plots to check non-constant variance and curvatures




```
##                Test stat Pr(>|Test stat|)
## I(nox^-1.3)      -1.2394         0.215950
## I(nox^-2.6)      -0.4003         0.689153
## rm                1.2694         0.205061
## I(rm^2)          -0.7987         0.424948
## I(ptratio^4.5)    2.6379         0.008678 **
## I(lstat^1/3)      0.5447         0.586253
## I(lstat^2/3)      1.8123         0.070718 .
## factor(chas)
## Tukey test       -1.0751         0.282311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```