Mini Project: 2

Loading Boston Housing data from MASS library:

Library(MASS)
data(Boston)

All the variables in the dataset are shown below and are as follows: crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv. Where medv is the response variable and all the others are predictor variables.

`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`
: nitrogen 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 mean of distances to five Boston employment centres.

`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)^2* where *Bk* is the proportion of blacks by town.

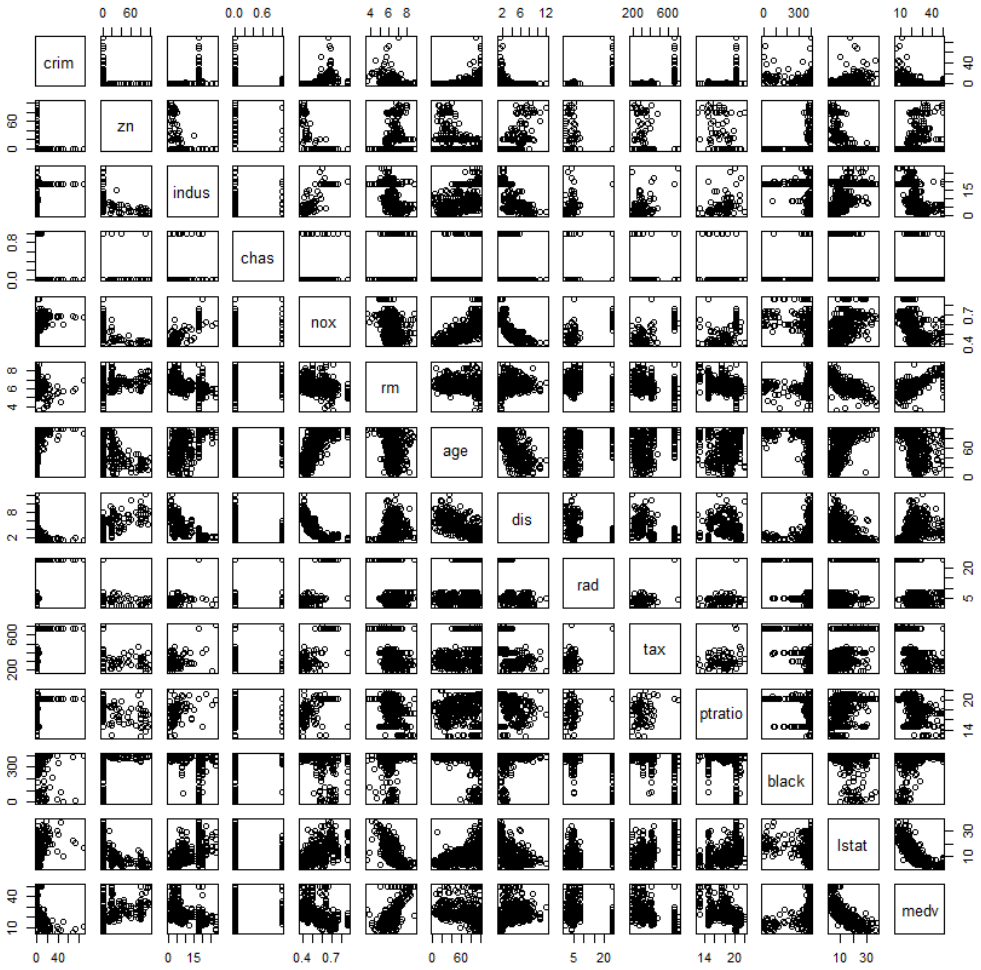
`lstat`
: lower status of the population (percent).

`medv`
: median value of owner-occupied homes in \$1000s.

Now, lets find a strong model for predicting the house prices. For that we'll first check the scatter plot for all the predictors to see if we can find the variables that have a good co-relation with medv. The scatterplot for them is shown below.

From the image below, we can infer the following:

1. Lstat has a strongest negative correlation with medv and rm has a strongest positive correlation with medv.
2. Nox, indus, ptratio are the other variable that seem to be negatively correlated with medv.
3. Dis, chas seems to be positively correlated with medv.
4. All the other predictors seem to have lower or no correlation with medv.



Now let us also check the standard error for the variables. The summary of the whole model is shown below that also includes the standard error:

```
Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

From the summary above we also see that the standard error is lowest for the variables chas, rm, dis, ptratio, lstat. While comparatively other predictors seem to have more standard error.

From the scatter plot and standard error. We see that the predictors lstat, rm, nox, indus, ptratio, dis, chas are the most correlated to predict the housing prices than the other 6 predictors. We use the combination of these predictors further to make our candidate models and compare them to see which candidate model fits the best.

Candidate Models:

1 candidate model with all the predictors and 4 candidate models with the combinations of lstat, rm, nox, indus, ptratio, dis, chas predictors are shown below:

Candidate Model 1: With all the predictors

```
> model1 <- lm(medv ~ . , data =Boston)
> summary(model1)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16

> |
```

From the summary of the model 1 above, we see that the R-squared value is 0.7338 which was expected to be this high since it includes all the predictors. But out goal it have a candidate model which is parsimonious. So, let’s go ahead with making the model with combinations of the best set of predictors that we chose before.

Candidate Model 2: With the predictors rm and lstat that we found to be strongly correlated positively and negatively with medv respectively

```
> model2 <- lm(medv ~ lstat:rm, data =Boston)
> summary(model2)

Call:
lm(formula = medv ~ lstat:rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.673  -4.068  -1.206   2.554  24.904

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.850307   0.622626   57.58   <2e-16 ***
lstat:rm    -0.174207   0.007275  -23.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.297 on 504 degrees of freedom
Multiple R-squared:  0.5322,    Adjusted R-squared:  0.5313
F-statistic: 573.4 on 1 and 504 DF,  p-value: < 2.2e-16
```

Looking at the summary for this model we see that the R-squared value is 0.5313 which is less than the previous model. Looking at just this criteria we can already say that this model doesn't seem to a good choice. But again, we will further confirm this by using AIC and BIC selection criteria later. Lets move on to the next model with a few more predictors than this.

Candidate Model 3: With predictors indus, rm, nox, ptratio, lstat

```
> model3 <- lm(medv~indus+rm+nox+ptratio+lstat, data = Boston)
> summary(model3)

Call:
lm(formula = medv ~ indus + rm + nox + ptratio + lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-14.111  -3.078  -0.831   1.748  29.924

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.60281    4.16490   4.947 1.03e-06 ***
indus        0.06188    0.05796   1.068    0.286
rm           4.58538    0.42871  10.696  < 2e-16 ***
nox         -4.72784    3.32874  -1.420    0.156
ptratio     -0.97272    0.12322  -7.894 1.86e-14 ***
lstat       -0.55338    0.05035 -10.990  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.229 on 500 degrees of freedom
Multiple R-squared:  0.6799,    Adjusted R-squared:  0.6767
F-statistic: 212.4 on 5 and 500 DF,  p-value: < 2.2e-16
```

From the summary, we see that the R-squared value is 0.6767 which is more than model 2 but still less than model one. From this we see that we are getting closer to the right combination of predictors whose R-squared value will be more than model 1 with much lesser predictors than model 1. R-squared values being one of our selection criteria.

Candidate Model 4: With predictors lstat, chas, ptratio, dis and second order terms of lstat and rm

```
> model4 <- lm(medv~lstat+I(rm^2)+I(lstat^2)+chas+ptratio+dis, data = Boston)
> summary(model4)

Call:
lm(formula = medv ~ lstat + I(rm^2) + I(lstat^2) + chas + ptratio +
    dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-21.059  -2.749  -0.393   2.360  26.065

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.532365   2.810826  14.776  < 2e-16 ***
lstat       -1.767237   0.120985 -14.607  < 2e-16 ***
I(rm^2)      0.290678   0.029638   9.808  < 2e-16 ***
I(lstat^2)   0.032571   0.003265   9.976  < 2e-16 ***
chas         2.853775   0.808559   3.529 0.000455 ***
ptratio     -0.704258   0.104661  -6.729 4.70e-11 ***
dis         -0.615218   0.114773  -5.360 1.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.524 on 499 degrees of freedom
Multiple R-squared:  0.7609,    Adjusted R-squared:  0.7581
F-statistic: 264.7 on 6 and 499 DF,  p-value: < 2.2e-16

>
```

We see from the summary above that the R-squared value for this model is 0.7581 which is the highest so far and only has 6 predictor variables. So far just considering the R-squared values, model 4 seem to be the best. Let's now move on to the last model.

Candidate Model 5: With predictors lstat, rm and second order terms of lstat and rm

```
> model5 <- lm(medv ~  lstat + I(rm^2)+ I(lstat^2) + rm, data =Boston)
> summary(model5)

Call:
lm(formula = medv ~ lstat + I(rm^2) + I(lstat^2) + rm, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5670  -2.8232  -0.4123   2.2523  27.2530

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.084032   9.938929  10.573  < 2e-16 ***
lstat        -1.416229   0.120312 -11.771  < 2e-16 ***
I(rm^2)       2.356069   0.239998   9.817  < 2e-16 ***
I(lstat^2)    0.021850   0.003515   6.217 1.07e-09 ***
rm          -26.009362   3.103058  -8.382 5.30e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.608 on 501 degrees of freedom
Multiple R-squared:  0.751,     Adjusted R-squared:  0.749
F-statistic: 377.7 on 4 and 501 DF,  p-value: < 2.2e-16

>
```

We see from the summary above that R-squared value for this model is 0.749 which is still less than model 4. Hence, considering just the R-squared values so far. Candidate model 4 is our best choice.

Let's perform some more selection criteria on our models to see if the model 4 is indeed the best one out these.

We're now calculating RSS for feeding it to AIC and BIC selection criteria as shown below:

```
> # calculating RSS
> RSSmodel1 <- sum((Boston$medv - fitted(model1))^2)
> RSSmodel2 <- sum((Boston$medv - fitted(model2))^2)
> RSSmodel3 <- sum((Boston$medv - fitted(model3))^2)
> RSSmodel4 <- sum((Boston$medv - fitted(model4))^2)
> RSSmodel5 <- sum((Boston$medv - fitted(model5))^2)
>
```

Selection criteria: AIC

AIC values for all the models are shown below:

```
> # caculating AIC and BIC
> model_1 <- 13
> model_2 <- 2
> model_3 <- 5
> model_4 <- 6
> model_5 <- 4
> n <- nrow(Boston)
>
> AICmodel1 <- n*log(RSSmodel1/n)+2*model_1
> AICmodel2 <- n*log(RSSmodel2/n)+2*model_2
> AICmodel3 <- n*log(RSSmodel3/n)+2*model_3
> AICmodel4 <- n*log(RSSmodel4/n)+2*model_4
> AICmodel5 <- n*log(RSSmodel5/n)+2*model_5
>
> c(AICmodel1,AICmodel2,AICmodel3,AICmodel4,AICmodel5)
[1] 1587.643 1864.093 1678.064 1532.422 1549.075
>
```

From this we can see that the AIC value for model 4 is 1532.422 which is less than all the other models. This also gives us an additional proof that model 4 is our best choice.

Selection criteria: BIC

BIC values for all the models are shown below:

```
> BICmodel1 <- n*log(RSSmodel1/n)+log(n)*model_1
> BICmodel2 <- n*log(RSSmodel2/n)+log(n)*model_2
> BICmodel3 <- n*log(RSSmodel3/n)+log(n)*model_3
> BICmodel4 <- n*log(RSSmodel4/n)+log(n)*model_4
> BICmodel5 <- n*log(RSSmodel5/n)+log(n)*model_5
>
> c(BICmodel1, BICmodel2, BICmodel3, BICmodel4,BICmodel5)
[1] 1642.588 1872.546 1699.196 1557.781 1565.981
>
```

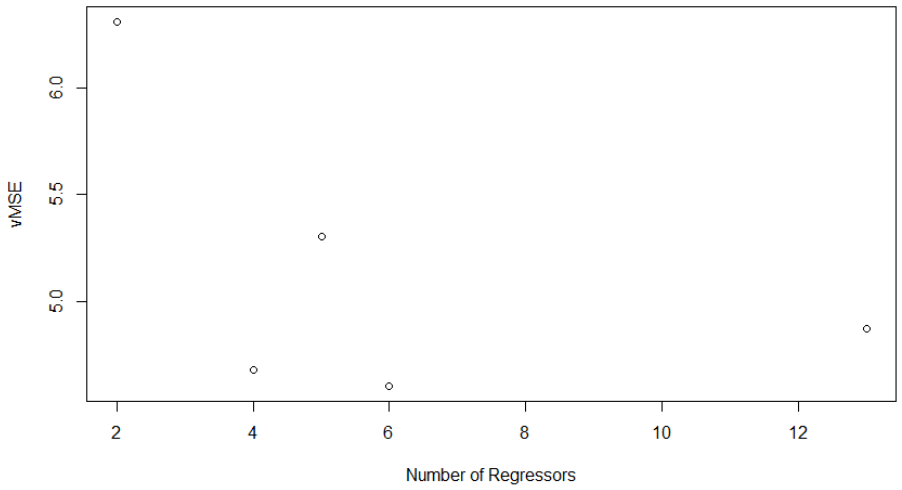
From this we can see that the BIC value for model 4 is 1557.781 which is also less than all the other models. This further confirms why model 4 is our best choice.

Cross-validation: Leave one out cross validation error

Mean square error values for all the models is shown below:

```
> c(mse1, mse2, mse3, mse4,mse5)
[1] 23.72575 39.84042 28.11180 21.18164 21.90228
>
```

We see above that the MSE value for the model 4 is the lowest again out of all the models we see. We now confirm that the model 4 is the best choice even according to the cross-validation criteria. Plot for square root of MSE vs number of regressors for all the models is shown below:



Selecting one candidate model out of all the candidate models:

From the above selection criteria used and shown. We have compared the resulting respective values of the models with each other with reasoning. From every selection criteria used: R-squared, AIC, BIC, cross-validation. We can only infer how model 4 is our best choice for predicting the housing prices.

Coefficients of model 4 are: lstat, chas, ptratio, dis, second order term of lstat, second order term of rm. The same is shown below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.532365   2.810826  14.776  < 2e-16 ***
lstat       -1.767237   0.120985 -14.607  < 2e-16 ***
I(rm^2)      0.290678   0.029638   9.808  < 2e-16 ***
I(lstat^2)   0.032571   0.003265   9.976  < 2e-16 ***
chas         2.853775   0.808559   3.529 0.000455 ***
ptratio     -0.704258   0.104661  -6.729 4.70e-11 ***
dis         -0.615218   0.114773  -5.360 1.27e-07 ***
---
```

Inferential analysis:

Let's consider β1 to be the coefficient of lstat.

Hypothesis testing:

Null hypothesis: Ho: $\beta 1 = 0$

Alternate hypothesis: Ha: $\beta 1 \neq 0$

Test statistics:

T value of lstat = -14.607

P value:

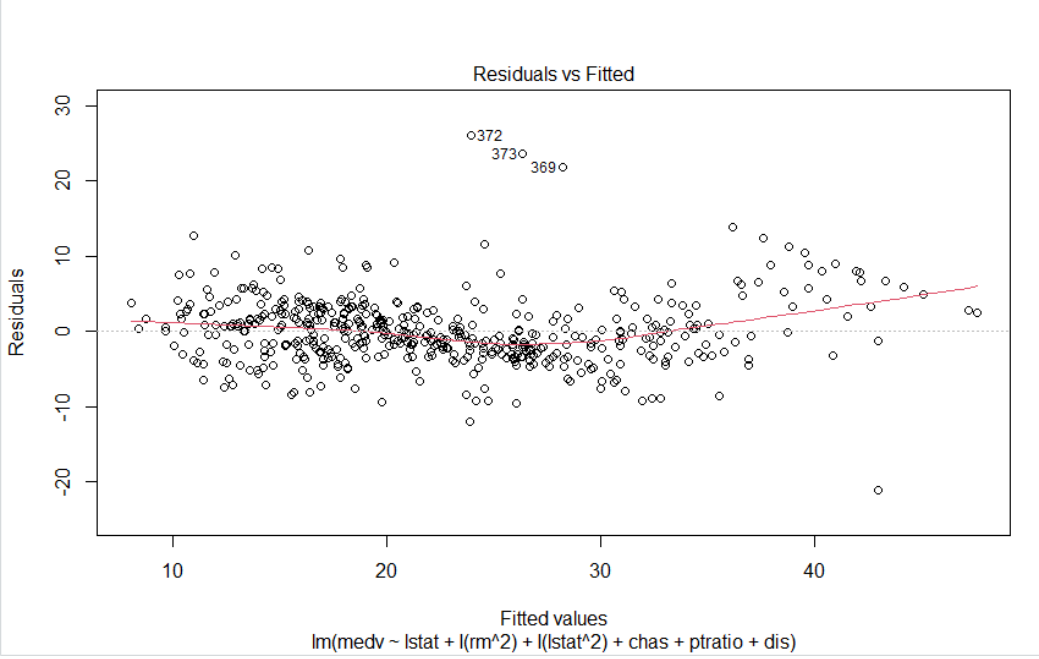
P value of lstat = <2e-16

Now, assuming that the alpha value is 0.05, we see that the p value is less than the alpha value. Hence, we reject the null hypothesis.

We can see the evidence that after all the adjustments, lstat has non-zero contribution to mean function. We can also say that all the variables in our model 4 also has non-zero contributions to mean function since all the p values of them is less than the alpha.

Assessing the fit for the selected model, Including an assessment of collinearity:

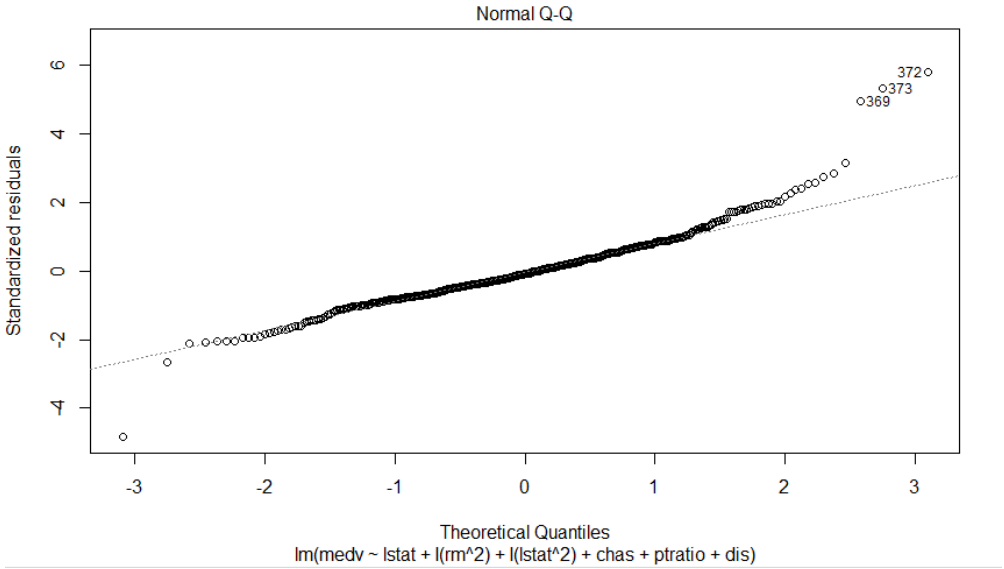
Residual vs Fitted plot for the selected model 4 is shown below:



Here, from the above figure we notice that the deviation of the residual points from the referenced line is very less. The only deviations we observe is in the higher fitted values. Which when compared to all the

other fitted values, is much lower. This shows us that the linear mean function supposition is a better fit for this model.

Now, lets see more clearly if the data is normally distributed. For that, Normal Q-Q plot is shown below:



From the plot above we can see and say that the data is more normally distributed.

Collinearity:

VIF values for all the coefficients in the model 4 is shown below:

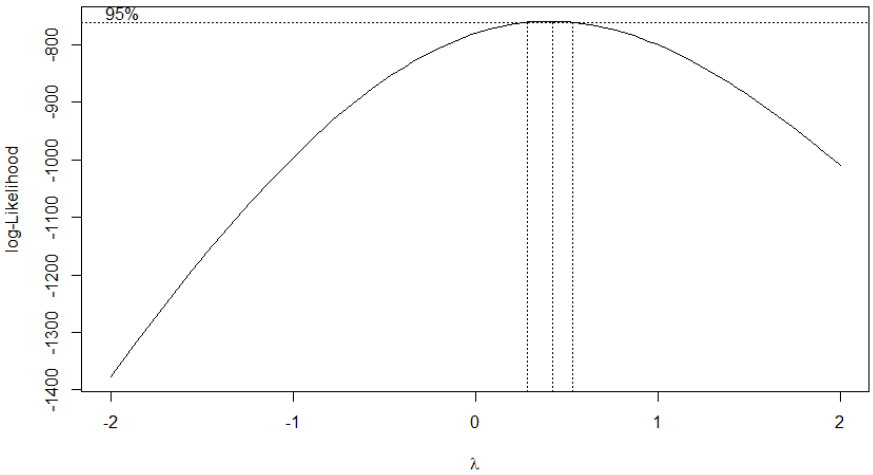
```
> vif(model4)
     lstat    I(rm^2) I(lstat^2)       chas    ptratio        dis
 18.418970   1.786939  14.656834   1.040757   1.266898   1.441296
> 
```

We know that if the VIF value is closer to 1 then those coefficients are uncorrelated with the other predictors of the model. And for those whose VIF value is closer to 10, have a very strong correlation with the other predictors of the model.

From the above picture we see that both lstat and lstat^2 is closer to 10 which makes them strongly correlated. All the other coefficients are closer to 1 making them not strongly correlated. Thus, we can say that there is evidence of having no problem due to collinearity.

Box-Cox transformation:

Box-Cox transformation is shown below:



Optimal lambda value after transformation of our model 4 is 0.4242424 as shown below:

```
> t <- r$x[which.max(r$y)]
> t
[1] 0.4242424
>
```

We can now say that the model will fit at lambda value of 0.26 after Box-Cox transformation.

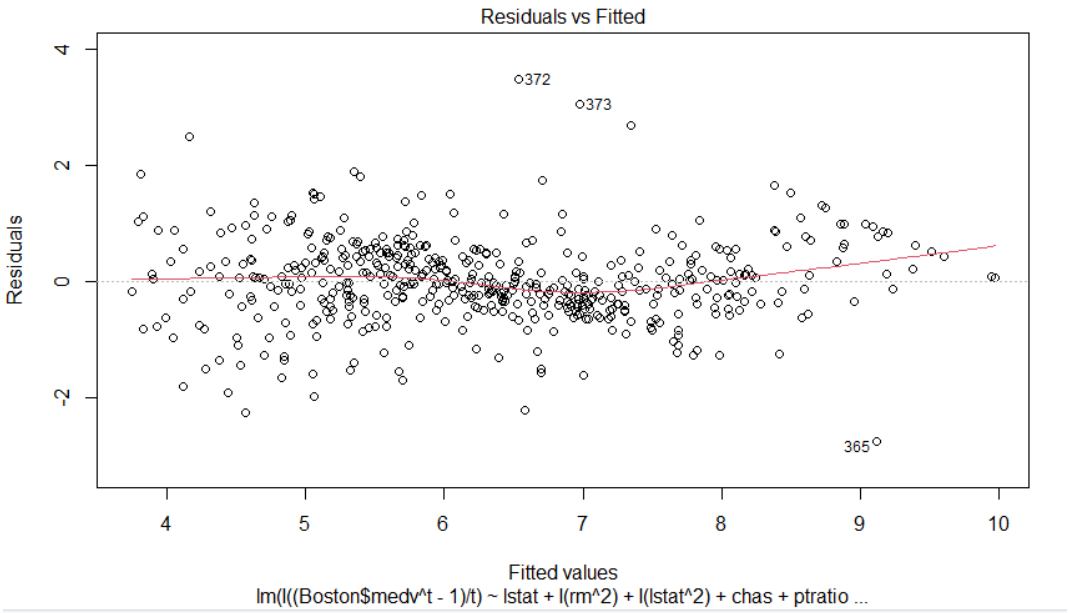
Now, let's fit the Box-Cox transformed model. The results are shown below:

```
>
> #Fitting Box-Cox model
> F <- lm(I((Boston$medv^t-1)/t)~lstat+I(rm^2)+I(lstat^2)+chas+ptratio+dis,data=Boston )
> summary(F)

Call:
lm(formula = I((Boston$medv^t - 1)/t) ~ lstat + I(rm^2) + I(lstat^2) +
    chas + ptratio + dis, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7450 -0.3897 -0.0301  0.4068  3.4951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.7718594  0.4516274  21.637  < 2e-16 ***
lstat       -0.2569949  0.0194391 -13.221  < 2e-16 ***
I(rm^2)      0.0346540  0.0047621   7.277 1.33e-12 ***
I(lstat^2)   0.0039221  0.0005246   7.477 3.45e-13 ***
chas         0.4543284  0.1299145   3.497 0.000512 ***
ptratio     -0.1185982  0.0168163  -7.053 5.89e-12 ***
dis         -0.0700393  0.0184410  -3.798 0.000164 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7269 on 499 degrees of freedom
Multiple R-squared:  0.7616,    Adjusted R-squared:  0.7587
F-statistic: 265.7 on 6 and 499 DF,  p-value: < 2.2e-16
```

We see that the R-squared value is 0.7587 which was the same before the transformation. Thus, it does not improve the model fit.

Conclusion:

To further conclude that the transformation does not improve the model fit. We will see the Residual vs Fitted Plot for the transformed model below:



From the above figure we can say that there is not much significant change in the model after Box-Cox transformation. Hence, the transformation did not improve out model fitting.

Now, in a whole, from the above shown candidate model comparisons and the selection criteria used, we can say that our model 4 consisting of the variables lstat, chas, ptratio, dis is the right fit because of the highest R-squared value and lowest AIC, BIC values and lowest mean square error value. The variables lstat, chas, ptratio, dis in model 4 are providing us helpful prediction of housing prices.