Machine Learning Methods andApplications HW - Week 2

Uğur DAR

08 03 2021

Contents

Linear Regression Model - Boston Housing Data

Abstract

This week's homework is linear regression, one of the simplest models in machine learning and statistical learning. In my last week's paper, I showed that there is a linear relationship between the medv target variable and some of the other variables in the Boston data set. In this document, fitting the linear regression model to the Boston data set, interpretation of the model outputs can be found.

Packages

```
library(dplyr)
library(mlbench)
library(car)
library(caret)
library(lmtest)
```

```
data(BostonHousing) # Calling the data from mlbench
```

Train-Test Split

```
glimpse(BostonHousing)
```

```
## Rows: 506
## Columns: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.088...
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5...
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87,...
## $ chas    <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.5...
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.6...
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9...
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9...
## $ rad     <dbl> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4,...
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311,...
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2,...
## $ b       <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396...
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17...
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9,...
```

Different variables can be selected as target variables in this data set. Crime rates in different neighborhoods in the city of Boston can be modeled. For this, the variable crime can be selected as the target variable. Determining the target variable is the subject of the research. In this data set, the properties of the houses and neighborhoods are given and the main purpose is to estimate the prices of the houses. So, I chose **medv**(median value of owner-occupied homes in USD 1000's) variable as target variable.

Train-Test Split

```
set.seed(26) # reproducbility
index <- sample(nrow(BostonHousing),nrow(BostonHousing)*0.8)
train <- BostonHousing[index,]
test <- BostonHousing[-index,]
```

```
dim(train)
```

```
## [1] 404  14
```

```
dim(test)
```

```
## [1] 102  14
```

BostonHousing data has 506 instances(rows). 0.8 of the data is train, 0.2 of the data is test set. So, I choose randomly 404 instances from the data set as train data, 102 instances as test data.

Modelling

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

```
##
## Call:
## lm(formula = medv ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.091  -2.890  -0.565   1.956  25.166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.903426   5.625781   7.448 6.13e-13 ***
## crim         -0.102668   0.036096  -2.844 0.004686 **
## zn            0.044163   0.015533   2.843 0.004701 **
## indus         0.008889   0.070357   0.126 0.899528
## chas1         2.191928   1.019087   2.151 0.032100 *
## nox         -17.159257   4.557631  -3.765 0.000192 ***
## rm            3.299570   0.470133   7.018 1.00e-11 ***
## age           0.008360   0.015215   0.549 0.583010
## dis          -1.391811   0.228993  -6.078 2.90e-09 ***
## rad           0.347897   0.076666   4.538 7.58e-06 ***
## tax          -0.014009   0.004280  -3.273 0.001159 **
## ptratio      -1.070237   0.147733  -7.244 2.34e-12 ***
## b             0.008403   0.003016   2.786 0.005596 **
## lstat        -0.563333   0.056366  -9.994  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.828 on 390 degrees of freedom
## Multiple R-squared:  0.734,  Adjusted R-squared:  0.7252
## F-statistic: 82.79 on 13 and 390 DF,  p-value: < 2.2e-16
```

$\widehat{medv} = 41.903426 - 0.102668 * crim + 0.044163 * zn + 0.008889 * indus + 2.191928 * chas1 - 17.159257 * nox + 3.299570 * rm + 0.008360 * age - 1.391811 * dis + 0.347897 * rad - 0.014009 * tax - 1.070237 * ptratio + 0.008403 * b - 0.563333 * lstat$

R gives a very good regression model output. Firstly, it gives some statistics about residuals. Secondly, we can see coefficients part. In this output, we see the coefficients estimated in the regression model, the standard deviation of these coefficients, the t statistisc and the test result of the coefficient significance, the p-value. We see stars sign for each coefficient next to p-value. This points to the Signif.code section below the output, for example, zn is significant feature at 0.05 significance level or crim is significant at 0.01 significance level. So, all features except indus and age significant at 0.05 level. Lastly, this section is about the significance of the model in general. As we can see, $R^2 = 0.734$, $R^2_{Adj} = 0.7252$. Theoretically, no matter how many explanatory variables we add to the model, the value of R^2 in the model increases or remains constant. Therefore R^2_{Adj} gives us more reliable results. In summary, in this model, the target variable is explained by the features at a rate of 0.7348. In general, the F test is used for the significance of the model. The last part shows the F statistics and the p-value in the F-test, $2.210^{-16} < 0.001$ it is too close to 0, therefore we can say that the model is significance at 0.001 level.

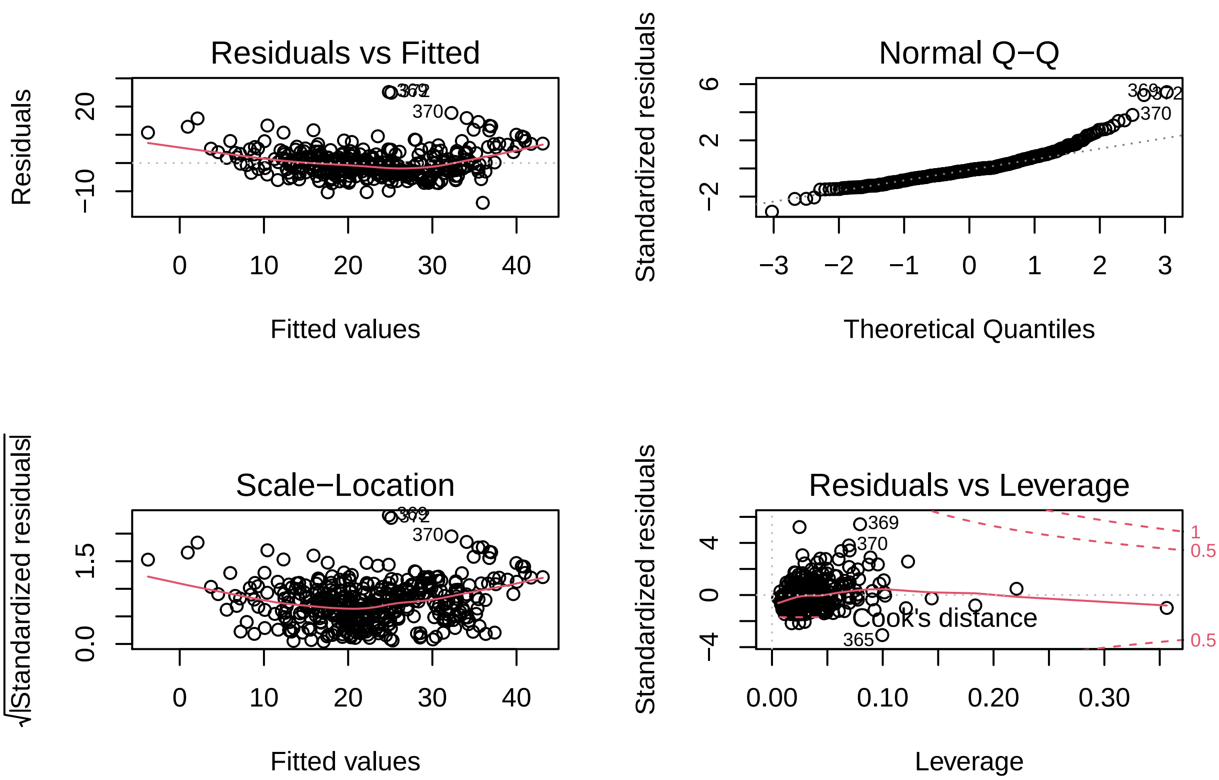
Regression Diagnostics

Potential Problems in RMs

1. Non-linearity of the target-feature relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points

Last week, I examined relationships between target and features. In that homework, plots shows that some features and target have linear relationship. Also, linear regression model is not seems to bad. It is significant and it explains 72.52% of the relationship. See the HW1

```
par(mfrow=c(2,2))
plot(model1)
```



Residuals vs Fitted plot shows that, features and target have linear relationships but it hasn't exactly straight red line. So, non-linear models can also be tried. As we can see at Q-Q plot, the dots are supposed to follow a more or less straight line, which they clearly don't here, residuals are not exactly normally distributed. Scale-Location plot, we check for homoskedasticity we would want the red line on the plot to be more or less straight and horizontal, homoskedasticity(constant variance) assumptions can be considered to be fulfilled. Residuals vs Levarage plot shows that there is no leverage point, every instances past dotted red lines. In a nutshell, looking at the plot, we can't say that there is a problem with our model, but we need to do the necessary tests for assumptions.

Multicollinearity

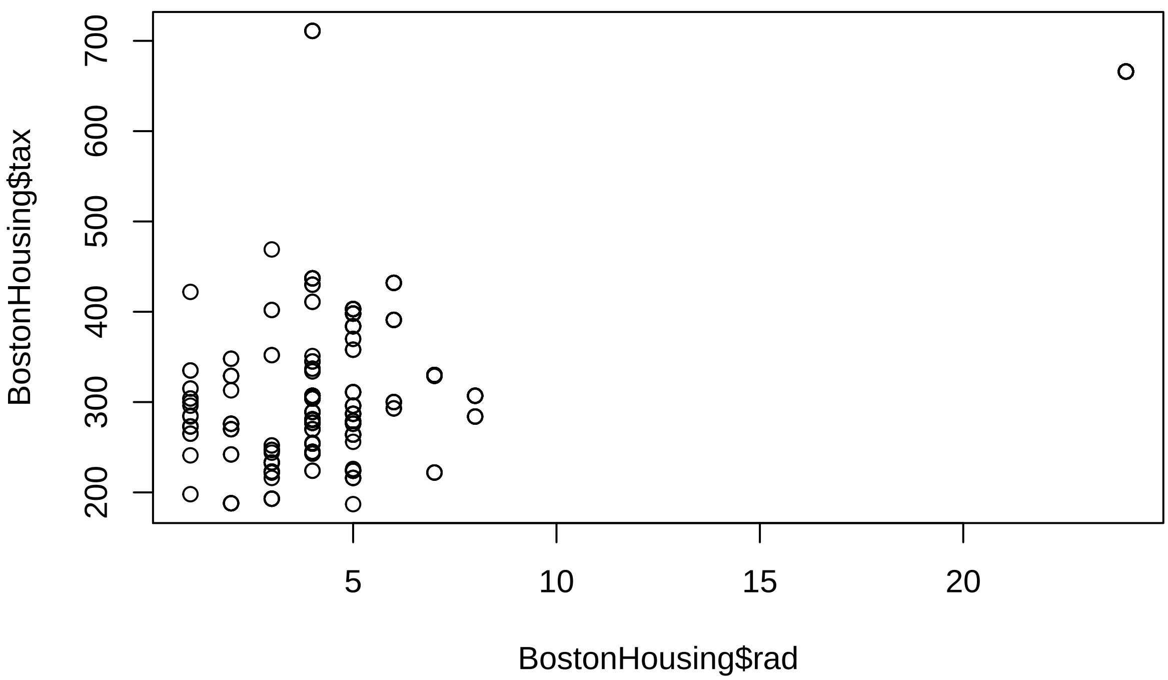
Multicollinearity can examine with VIF(Variance Inflation Factors). If VIF is 1, then there is no correlation, if it's between 1 and 10, there is moderate correlation, and if it's greater than 10, there is high correlation and there is serious multicollinearity problem.

```
vif(model1)
```

```
##     crim       zn    indus     chas      nox       rm      age      dis 
## 1.753824 2.355805 3.956009 1.084063 4.440576 1.870232 3.126411 3.966104 
##      rad      tax  ptratio        b    lstat 
## 7.516475 8.800637 1.763016 1.371206 2.799379
```

tax and *rad* features can cause multicollinearity problem. Let's look at on plot.

```
plot(BostonHousing$rad,BostonHousing$tax)
```



There does not appear to be a linear relationship between these variables in the plot. Let's look correlation matrix.

```
cor(BostonHousing %>% select_if(is.numeric)) > 0.90  # if correlation between features greater than 0.9
```

```
##          crim    zn indus   nox    rm   age   dis   rad   tax ptratio     b
## crim     TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## zn      FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## indus   FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## nox     FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## rm      FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## age     FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE   FALSE FALSE
```

```
## dis     FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE   FALSE FALSE
## rad     FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE   FALSE FALSE
## tax     FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE   FALSE FALSE
## ptratio FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE    TRUE FALSE
## b       FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE  TRUE
## lstat   FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## medv    FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
##         lstat  medv
## crim    FALSE FALSE
## zn      FALSE FALSE
## indus   FALSE FALSE
## nox     FALSE FALSE
## rm      FALSE FALSE
## age     FALSE FALSE
## dis     FALSE FALSE
## rad     FALSE FALSE
## tax     FALSE FALSE
## ptratio FALSE FALSE
## b       FALSE FALSE
## lstat    TRUE FALSE
## medv    FALSE  TRUE
```

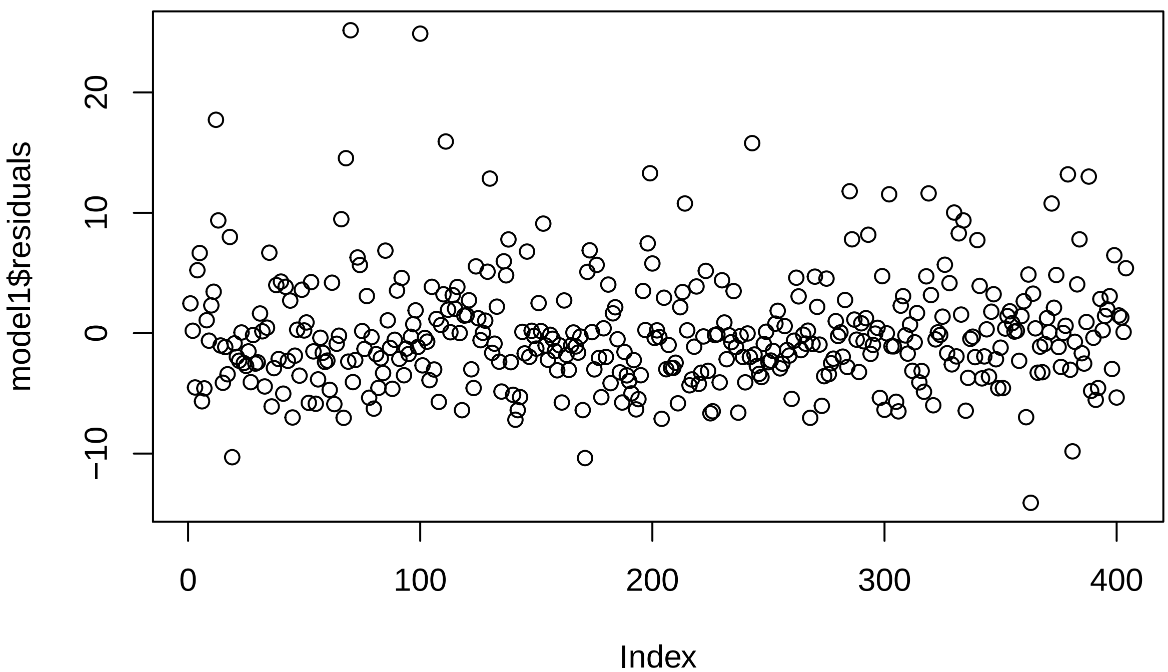
```
cor(BostonHousing$rad,BostonHousing$tax)
```

```
## [1] 0.9102282
```

As I said above, visual comments are subjective. There seems to be a very high correlation between *rad* and *tax*. One of these features can be omitted from the model.

Homoskedasticity

```
plot(model1$residuals)
```

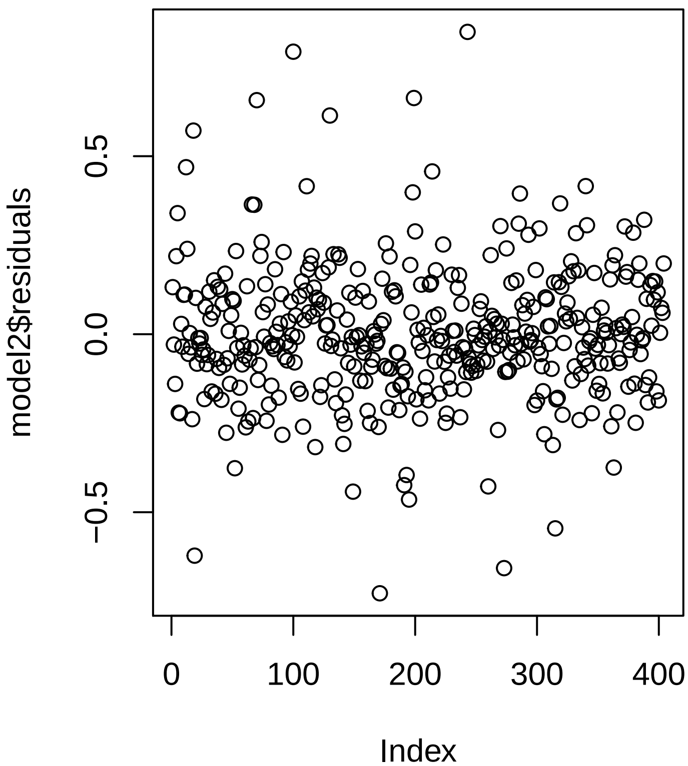



```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 54.053, df = 13, p-value = 5.92e-07
```

If the test statistic has a p-value below an appropriate threshold (e.g. $p < 0.05$) then the null hypothesis of homoskedasticity is rejected and heteroskedasticity assumed.

```
model2 <- lm(log(medv)~.,data=train)
plot(model2$residuals)
```




```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 57.731, df = 13, p-value = 1.332e-07
```

It doesn't work. Log transformation might use on other features. Other models can also be tested, even if one of the assumptions is violated, the linear regression model may not give bad results compared to other models.

Model Evaluation

Some Evaluation Metrics

$e_t = y_t - \hat{y}_t$
Mean squared error(MAE) $\quad \text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Mean absolute error(MSE) $\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t|$
Root mean squared error(RMSE) $\quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$

```r
mse <- function(y_actual,y_pred){
  mean((y_actual-y_pred)^2)
}
rmse <- function(y_actual,y_pred){
  sqrt(mean((y_actual-y_pred)^2))
}
mae <- function(y_actual,y_pred){
  mean(abs(y_actual-y_pred))
}
```

Prediction

```r
train_pred <- predict(model1,train)
test_pred <- predict(model1,test)
```

Evaluation

```r
train_metrics <- data.frame(MSE = mse(train$medv,train_pred),
                            RMSE = rmse(train$medv,train_pred),
                            MAE = mae(train$medv,train_pred))

test_metrics <- data.frame(MSE = mse(test$medv,test_pred),
                           RMSE = rmse(test$medv,test_pred),
                           MAE = mae(test$medv,test_pred))

results <- data.frame(rbind(train_metrics,test_metrics))
rownames(results) <- c("Train","Test")
results
```

```
##            MSE     RMSE      MAE
## Train 22.49754 4.743157 3.369471
## Test  20.20724 4.495246 3.029191
```

The results came out close to each other. Maybe underfitting has occurred, because train error metrics greater than test's metrics.

The linear regression model, some assumptions have been violated like normality of error terms, multicollineratiy and homoskedasticity but when I try other regression models, it can be seen that linear regression works well.

Click to see the other models on my Kaggle Notebook