Nazarbayev University

Math 440: Regression Analysis

Group Project

**Boston Housing Data Analysis**

*(Using R)*

*Done by:*

Alibi Jangeldin

Anel Mukasheva

Aizhan Akhmetbek

Azat Gabdolla

November 23, 2016

1. **Executive Overview**

The topic that our group has decided to explore is very focused since it regards the Boston Housing in suburb regions. In particular, it refers to the prices of housing.

Our response variable (Y) is median value of owner-occupied homes in $1000's (medv). The goal is to predict which attributes (out of 13 response variables that are given with the initial model) affect the pricing of Boston suburb housing the most. We have explored that the following predictors explains the housing values the best: per capita crime rate by town (crim), nitric oxides concentration (nox), weighted distances to five Boston employment centres (dis), index of accessibility to radial highways (rad), pupil-teacher ratio by town (ptratio), and percentage of lower status of the population (lstat).

After conducting the study to measure the effect of our predictor variables on the housing values, we obtained the following model:

*Yi = 3.056 - 0.036X1 (lstat.c) - 0.049X2 (ptratio.c) - 0.317X3 (log(dis.c)) -*

* *0.884X4 (log(nox.c)) + 0.067X5 (log(rad.c)) - 0.009X6 (crim.c)*

Taking the logarithm makes the variable's distribution more symmetric. Since most of the variables exhibit an asymmetry with a higher density on the left side, we proposed *log(dis.c), log(nox.c),* and *log(rad.c)* transformations. This is due to the fact that lower values move further away from each other, whereas the distance between greater values is reduced by these transformations.

It seems that our model adequately fits the data because we have relevant (1) *p-*value which equals 0.1713 and (2) coefficient of multiple determination (or R-squared) which equals 0.7759.

(1) The *p*-value assesses normality assumption, by estimating the probability that residuals are normally distributed. If this assumption do not hold we will not be able to use confidence intervals and inferences based on normal distribution. The *p*-value of 0.1713 (> 0.05) indicates that residuals seem to be normally distributed.

(2) R-squared is defined as a statistical measure of how close the data is to the fitted regression line. It can take values between 0 and 1 meaning that closer the value to 1 implies higher degree of linear relationship between predictors and the response. In our case the fitted model explains 77,59% of linear association that seems to be high enough to explain patterns in price formulation.

According to this model, we can conclude that the index of accessibility to radial highways (log(rad.c)) positively influences on the price of Boston housing. However, percentage of lower status of the population (lstat.c), pupil-teacher ratio by town (ptratio.c), weighted distances to five Boston employment centres (log(dis.c)), nitric oxides concentration (log(nox.c)), and per capita crime rate by town (crim.c) have negative impact on our response variable. Consequently, accessibility to radial highways implies higher price of houses. On the other hand, other factors will lead to lower marginal price of the housing values.

The following model was chosen based on Forward Feature Selection algorithm. This is automated processes of building model by successively adding the most informative variables. Then, as it is mentioned above, the model has been checked based on *p*-value and R-squared.

**2. Overview**

The data used to develop the model was taken from [*https://archive.ics.uci.edu/ml/datasets/Housing*](https://archive.ics.uci.edu/ml/datasets/Housing) website which originally obtained data from StatLib library managed by the Carnegie Mellon University. In addition, the article on  'Hedonic prices and the demand for ‘clean air', by Harrison, D. and Rubinfeld, D.L. J. Environ. Economics & Management, vol.5, 81-102, 1978 gives a general overview of the trends.

The choice has been made due to the availability of the data, which  provides all information that we are interested in. Moreover, the data set contains the variables that we are familiar with. Besides scientific interest we have personal interest concerning house prices in Boston as the potential renters.

Initially, we worked with all 13 numerous potential predictor variables of the data that were provided by the website. However, Forward Feature selection and Shapiro-Wilk test we have chosen to work with six predictors that fit the model the best.

**3. Model Description**

Final Regression Model: log(medv) = 4.171362 - 0.036041\* lstat - 0.049920\* ptratio - 0.316984\* log(dis) - 0.883956\* log(nox) + 0.066430\* log(rad) - 0.009244 crim

*Coefficients interpretation:*

* β0: *the average median value for housing prices when all predictor variables are 0*
* β1 - β7: *the average change in median value for housing prices for each unit increase in* Xi *when all remaining predictor variables are kept fixed*

When the data was obtained, it contained large data set, so, it was divided into train and test sets. All test and investigations were applied on train set.

During the process of investigation of the sufficient regression model, all models were evaluated on the basis of normality assumption by using Shapiro-Wilk test. The aim was to achieve model’s p-value to reach the point 0.05 or above.

Initially, starting model contained 13 predictor variables, which were listed before (see the section 2):

*Model0<-lm(medv ~crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat)*

Shapiro test p-value = 2.523e-09

Conclusion: too low p-value, model should be modified

However, since the data related to several response variables, namely, nox for pollution and medv for house pricing, and it was decided to investigate only the latter one (medv), it is obvious that some of predictor variables could become useless for prediction of Boston housing prices. That is why, the model was evaluated via automatic feature elimination methods (i.e. forward and backward selection); using these procedures, 5 variables were eliminated: zn, indus, age, tax and black. So, the model became:

*Model1 <- lm(medv ~ lstat + rm + ptratio + dis + nox + rad + crim + chas)*

p-value = 6.208e-09

Conclusion: too low p-value, model should be modified.

Thus, as it can be deduced from the p-value, normality assumption was violated, however, feature scaling tests (R2, adjR2, BIC, Cp) considered the remaining features as significant. Because of that, it was decided  to evaluate the model using boxcox plot:

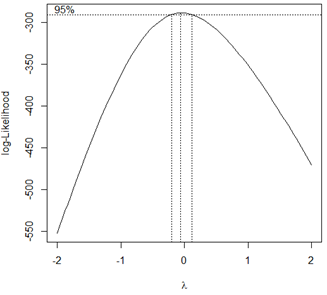


Figure 1. Boxcox plot of the Model1

Plot, in turn, revealed that applying logarithmic transformation on Y could improve the model (since the value of 𝞴 is close to 0):

*Model2<-lm(log(medv) ~ lstat + rm + ptratio + dis + nox + rad + crim + chas)*

p-value = 5.665e-06

Conclusion: too low p-value, model should be modified

All insignificant variables were eliminated, that is why, and for further improvements of the model, the need for deeper examination of remained variables individually arose. Hence, all features were plotted versus residuals:

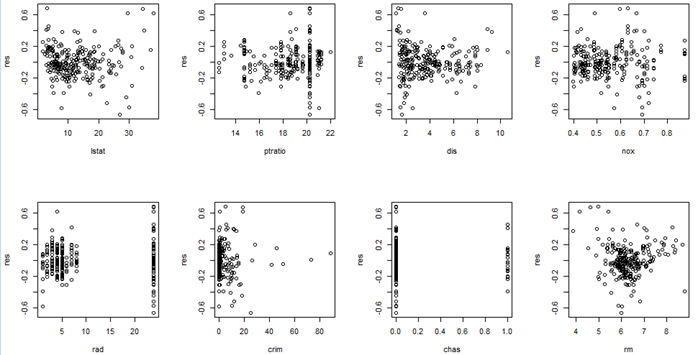


Figure 2. Predictors variables versus residuals for Model2

From the plot, it can be observed that lstat, nox, crim and rm can cause some problems, so, several models were investigated whether removal of these variables one by one could improve the situation. Here are results (all p-values presented are fpr Shapiro-Wilk test):

* Lstat removed: p-value = 3.334e-09
* Nox removed: p-value = 2.29e-08
* Crim removed: p-value = 8.319e-06
* Rm removed: p-value = 0.01026

Obviously, rm leads to major problems related to normality assumption; exclusion of this variable gives:

*Model3<-lm(log(medv) ~ lstat + ptratio + dis + nox + rad + crim + chas)*

p-value = 0.01026

Conclusion: p-value significantly improved, but there is still place for improvement

After this step, various combinations of predictors’ elimination in addition with transformations on them were examined and the best result (based on p-value) was obtained by this model (all changes were made intuitively by observing the plots of transformed features versus log(medv)):

*Model4 <- lm(log(medv) ~ lstat+ ptratio + log(dis) + log(nox) + log(rad) + crim )*

p-value = 0.1713

Conclusion: p-value > 0.05, normality assumption seems to be satisfied

Tests on interaction and quadratic terms

Moreover, it was examined whether interaction and quadratic terms could improve the model even further. 6 interaction terms and 6 quadratic terms were evaluated using their own p-values and formal F-tests. Interactions were selected on the basis of correlation matrix, thus, terms which had the correlation values higher than 50% were considered.

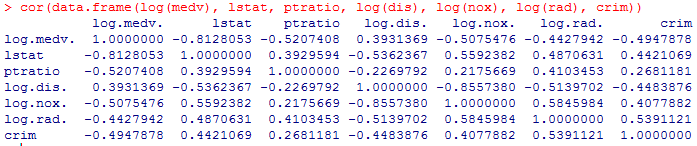


Figure 3. Correlation matrix for Model4

As it turned out after all tests, the only additional term needed was lstat.c.sq (squared centered value for lstat), but the model containing this extra term proved to have lower p-value for Shapiro-Wilk test.

Therefore, after implementation of different procedures, transformations and tests, it was decided to use the model 4 as the final one.

Evaluation of the model (verifying assumptions)

* Normality assumption

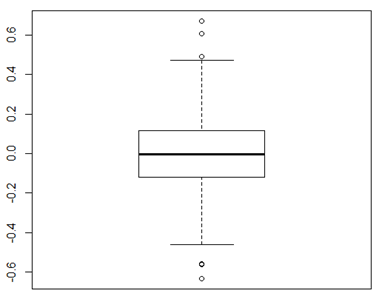


Figure 4. Boxplot of residuals for Model4

From the boxplot of residuals it can be observed that, Distribution appears to be symmetrical, so, it does not contradict the normality assumption.

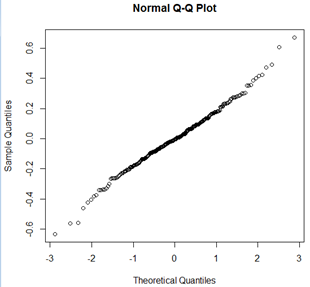


Figure 5. Normal probability plot of residuals for Model 4

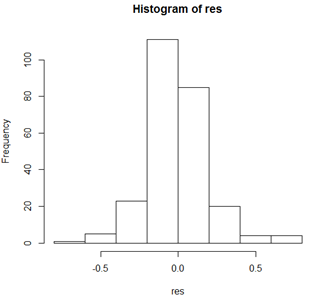


Figure 6. Histogram of residuals for Model 4

Normal probability plot and the histogram of residuals do not appear to reveal major issues with normality assumption (no extremely heavy tail and no significant skewness).

* Constant variance assumption

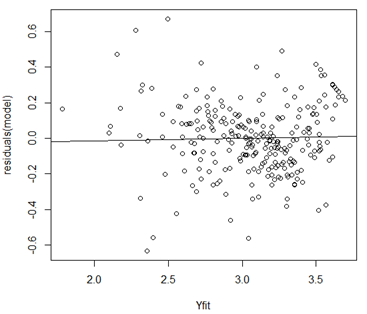


Figure 7. Plot of Fitted valued versus residuals for Model 4

From the plot of residuals versus fitted values, it can be seen that there is no obvious pattern present, thus the multiple regression function seems to be appropriate and the change in variance of residuals does not seem to occur.

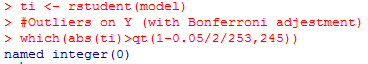
Conclusion: In addition with Shapiro-Wilk test (which tests the normality assumption), the regression model seems to satisfy normality and constant variance assumptions.

Tests on outliers and influential points:

* Formal test for outliers with respect to Y

H0: Yi is not an outlier

Ha : Yi is an outlier with respect to Y



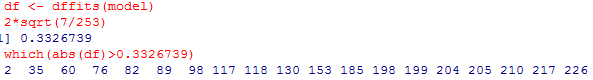
There is no any points for which |ti| > qt(1-0.05/2/253, 253-7-1=245). Therefore, on the basis of Bonferroni outlier test with α = 0.05, it can be concluded that the model does not seem to have any outliers with respect to Y.

* Formal test for outliers with respect to X

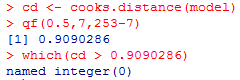
https://lh6.googleusercontent.com/CDqrdn-aYamPKfqrxoTkcmJtObGElDYWNVr8-A1VhDTjPCObgPGyjMRwrzDY9Qw98nt5tSQf5gRP_ScregvihcXb8-2w5QuCL-_3OdmfKqBYQN7UkIw2fmt3xFXX3fwxO9ZrABoL

The leverage of these points is exceeding 2p/n value, that is why they should be considered as outliers with respect to X.

* Influential points



Since the dffits values for these points are exceeding 2(p/n)^1/2 they are considered as influential on their own fitted values.



Since there is no any point for which the cook’s distance value exceeds 50th percentile, it can be concluded that there is no influential points on model coefficients.

Finally, applying the model of decision to the test set, we get MSE=23.41823 and MSPR = 24.53965 and that result shows good generalization with only 5 percent difference.

**4. Summary**

In real life this model can be used in two ways:

1.   The person who planned to buy apartments in Boston in 1980s could predict the amount of money that will be enough to buy home by the preferences he has.  The model is outdated, so currently people can evaluate the prices of real estate market of that period by this model. To validate the model in current period new data should be gathered and new model selection should be performed. It is so, because the real estate market changes and consumers’ preferences influence to the price in new way.

2.   Social research institution can use this model as rough estimator of life conditions of that period and can make social portraits of citizens of Boston suburb.  For example, by knowing that someone was the member of lower class researches can roughly identify the crime rate and ratio of teachers to pupils, this knowledge can give picture of social climate of his neighborhood.

For example, let’s assume that young family planned to buy the apartments in Boston. They are interested in three qualities: short distance to work center (4 miles) to reach their workplace fast, number of teachers (1 teacher to 16 pupils) within the area to provide children with proper education, low crime rate ( 0.14455 crimes per capita) to make own self secured. They could use the model that we got in section 3 to predict the amount of money that they have to have, so they can be 95% sure that the price of apartments will be within the range of (3 373.372,3 774,342) dollars, which is precise enough.

**5. Conclusion**

Prior to building the model, we expected that the following variables would affect the price of homes:

CRIM   per capita crime rate by town

ZN                  proportion of residential land zoned for lots over 25,000 sq.ft.

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 $10,000

PTRATIO         pupil-teacher ratio by town

LSTAT              % lower status of the population

But, out of these variables crim, ptratio, dis, rad and lstat were left only. Surprisingly variables that describe the size of the apartments were not influential. Moreover the tax rates did not met on final model. But nox that describes the nitric acid concentration was left on final model. It let us think that for us environmental conditions are less important than for members of US real estate market.

As it was expected the coefficients of lstat, ptratio, logdis.c, lognox.c and crim.c are negative. Because the higher values of each variables make area less comfortable for living as a result it becomes less attractive for buyers. For example, the higher index of crim make area more dangerous and higher ratio of pupils to teachers describe the deficiency of teachers within the area.

Also, the positive coefficient of rad was expected, because higher index means that it is easier to get the highways and move more freely.

We learnt from the project to work with data with many variables and make the best model. Also we learnt to make necessary transformations to get rid of the obstacles that prevent from getting necessary conclusions and read the data. Moreover, we learnt to make checking and found the balance between Shapiro-Wilk test and multiple R-squared. Finally, we learnt to work in group efficiently, so everyone was working in the field he or she is the best.

**Appendix (R code)**

library(MASS)

library(leaps)

# read data

data(Boston,package="MASS")

data<-Boston

# 50% of the sample size

smp\_size <- floor(0.5 \* nrow(data))

# set the seed to make your partition reproductible

set.seed(12)

train\_ind <- sample(seq\_len(nrow(data)), size = smp\_size)

train <- data[train\_ind, ]

test <- data[-train\_ind, ]

attach(train)

summary(train)

head(train)

#Forward selection

Null\_Model<-lm(medv~1, data=data.frame(train[,1:13]))

Full\_Model<-lm(medv~., data=data.frame(train[,1:13]))

step(Null\_Model, scope=list(lower=Null\_Model, upper=Full\_Model), direction="forward", k=log(54))

#lm(formula = medv ~ lstat + rm + ptratio + dis + nox + rad + crim + chas, data = data.frame(train[, 1:13]))

#Backward selection

Full\_Model<-lm(medv~., data=data.frame(train[,1:13]))

step(Full\_Model, direction="backward", k=log(54))

lm(formula = medv ~ lstat + rm + ptratio + chas + dis + nox + rad + crim + black + tax + zn, data = data.frame(train[,1:13]))

# p-value = 1.698e-08

model<-lm(medv ~ lstat + rm + ptratio + dis + nox + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

# Feature selection using different criteria

leaps<-regsubsets(medv ~ lstat + rm + ptratio + dis + nox + rad + crim + chas, data=data.frame(train), nbest=4)

plot(leaps, scale="adjr2")

# p-value = 1.698e-08

model<-lm(medv ~ lstat + rm + ptratio + dis + nox + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

plot(rad, residuals(model))

boxcox(model)

# log(medv) p-value = 7.937e-06

model<-lm(log(medv) ~ lstat + rm + ptratio + dis + nox + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

plot(rm, residuals(model))

# exclude rm p-value = 0.01863

model<-lm(log(medv) ~ lstat + ptratio + dis + nox + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

# Feature selection using different criteria

leaps<-regsubsets(log(medv) ~ lstat + ptratio + dis + nox + rad + crim + chas, data=data.frame(train), nbest=4)

plot(leaps, scale="adjr2")

plot(ptratio, log(medv))

plot(log(ptratio), log(medv))

plot(dis, log(medv))

plot(log(dis), log(medv))

# log(dis) p-value = 0.03766

model<-lm(log(medv) ~ lstat+ ptratio + log(dis) + nox + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

plot(nox, log(medv))

plot(log(nox), log(medv))

# log(nox) p-value = 0.04375

model<-lm(log(medv) ~ lstat+ ptratio + log(dis) + log(nox) + rad + crim + chas)

summary(model)

shapiro.test(model$residuals)

plot(rad, log(medv))

plot(log(rad), log(medv))

# log(rad) p-value = 0.05499

model<-lm(log(medv) ~ lstat+ ptratio + log(dis) + log(nox) + log(rad) + crim + chas)

summary(model)

shapiro.test(model$residuals)

# Feature selection using different criteria

leaps<-regsubsets(log(medv) ~ lstat+ ptratio + log(dis) + log(nox) + log(rad) + crim + chas, data=data.frame(train), nbest=4)

plot(leaps, scale="bic")

# exclude chas p-value = 0.0631

model<-lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim)

summary(model)

shapiro.test(model$residuals)

#interaction terms

cor(data.frame(log(medv), lstat, ptratio, log(dis), log(nox), log(rad), crim))

#take interactions with correlation larger than 50%

#lstat\*log(dis), lstat\*log(nox), log(dis)\*log(nox), log(dis)\*log(rad), log(nox)\*log(rad), log(rad)\*crim

int1 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + lstat\*log(dis))

summary(int1)

anova(int1)

0.0264/(9.5443/245)

qf(0.95, 1, 245)

int2 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + lstat\*log(nox))

summary(int2)

anova(int2)

0.1223/(9.4485/245)

int3 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + log(dis)\*log(nox))

summary(int3)

anova(int3)

0.0269/(9.5438/245)

int4 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + log(dis)\*log(rad))

summary(int4)

anova(int4)

int5 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + log(nox)\*log(rad))

summary(int5)

anova(int5)

int6 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + log(rad)\*crim)

summary(int6)

anova(int6)

#Conclusion: no interaction terms are needed

#quadratic terms

#model<-lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim)

lstat.c <- lstat - mean(lstat)

ptratio.c <- ptratio - mean(ptratio)

logdis.c <- log(dis) - mean(log(dis))

lognox.c <- log(nox) - mean(log(nox))

lograd.c <- log(rad) - mean(log(rad))

crim.c <- crim - mean(crim)

lstat.c.sq <- lstat.c^2

ptratio.c.sq <- ptratio.c^2

logdis.c.sq <- logdis.c^2

lognox.c.sq <- lognox.c^2

lograd.c.sq <- lograd.c^2

crim.c.sq <- crim.c^2

q1 <- lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c + lstat.c.sq)

q2 <- lm(log(medv) ~ lstat.c + ptratio + log(dis) + log(nox) + log(rad) + crim + ptratio.c.sq)

q3 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + logdis.c.sq)

q4 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + lognox.c.sq)

q5 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + lograd.c.sq)

q6 <- lm(log(medv) ~ lstat + ptratio + log(dis) + log(nox) + log(rad) + crim + crim.c.sq)

#Conclusion: lstat.c.sq is the only quadratic term needed

#new model

model.new<-lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c + lstat.c.sq)

# Stepwise Regression

step <- stepAIC(model.new, direction="both")

collection <- regsubsets(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c + lstat.c.sq, data=data.frame(train\_ind), nbest=4)

plot(collection, scale="adjr2")

#with adjR2 = 0.79 lograd.c can be dropped

plot(collection, scale="bic")

#with BIC = -360 lograd.c can be dropped

model.new<-lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c + lstat.c.sq)

model.new2<-lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + crim.c + lstat.c.sq)

model.old<-lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c)

shapiro.test(model.old$residuals)

#Since old model has the best p-val for Shapiro test, it will be counted as the final one

model<-lm(log(medv) ~ lstat.c + ptratio.c + logdis.c + lognox.c + lograd.c + crim.c)

#Outliers Tests

res <- residuals(model)

H <- hatvalues(model)

ti <- rstudent(model)

#Outliers on Y (with Bonferroni adjestment)

which(abs(ti)>qt(1-0.05/2/253,245))

#result: named integer(0)

#Outliers on Y (without Bonferroni adjestment)

which(abs(ti)>qt(1-0.05/2,245))

#result: 2  32  35  60  92 107 118 130 153 158 198 205 210 217

#Outliers with respect to X

2\*7/253

which(H>0.05533597)

#result: 2  33  54  60  69  76  94  98 112 117 122 134 147 159 165 178 198 199 204 210 217 221 226 236

#Influential Observations

df <- dffits(model)

2\*sqrt(7/253)

which(abs(df)>0.3326739)

#result: 2  35  60  76  82  89  98 117 118 130 153 185 198 199 204 205 210 217 226

cd <- cooks.distance(model)

qf(0.5,7,253-7)

which(cd > 0.9090286)

#result: named integer(0)