ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

BUSINESS
ANALYTICS
Master of Science

Statistics for Business Analytics I

Lab Assignment #2

Due Date: December 23th, 2020

Anastasios Theodorou

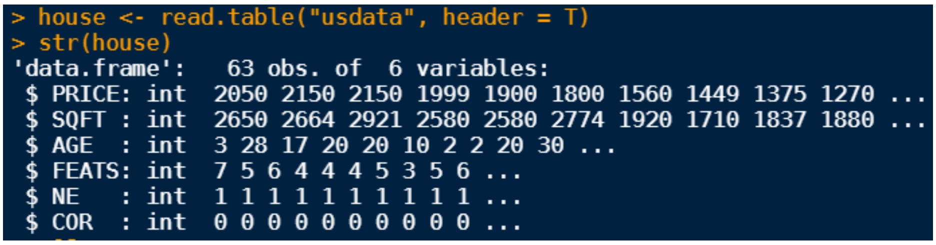
(p2822007)

Contents

1st Question

Read the “usdata” dataset and use str() to understand its structure.

1.1 Output in R



```
> house <- read.table("usdata", header = T)
> str(house)
'data.frame':	63 obs. of  6 variables:
 $ PRICE: int  2050 2150 2150 1999 1900 1800 1560 1449 1375 1270 ...
 $ SQFT : int  2650 2664 2921 2580 2580 2774 1920 1710 1837 1880 ...
 $ AGE  : int  3 28 17 20 20 10 2 2 20 30 ...
 $ FEATS: int  7 5 6 4 4 4 5 3 5 6 ...
 $ NE   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ COR  : int  0 0 0 0 0 0 0 0 0 0 ...
```

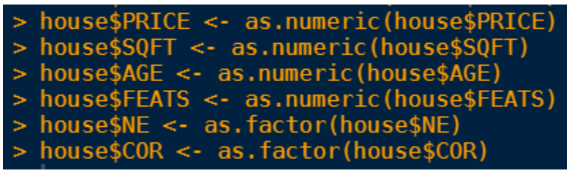
1.2 Comment

This data frame, which is called "*house*" contains 63 observations (house sales) and 6 variables. These 63 observations form the data frame’s rows, while the 6 variables, which are all defined as integers are the columns of it. It can be clarified, that it would be useful if some of the variables have been assigned as factors, because they have visible levels, for, example, the NE (if one house is located in the northeast part of the city or not) and the COR (corner location or not) have two levels: 0 and 1 or even FEATS (features of the houses).

2nd Question

Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and NE, COR to be factors.

2.1 Output in R

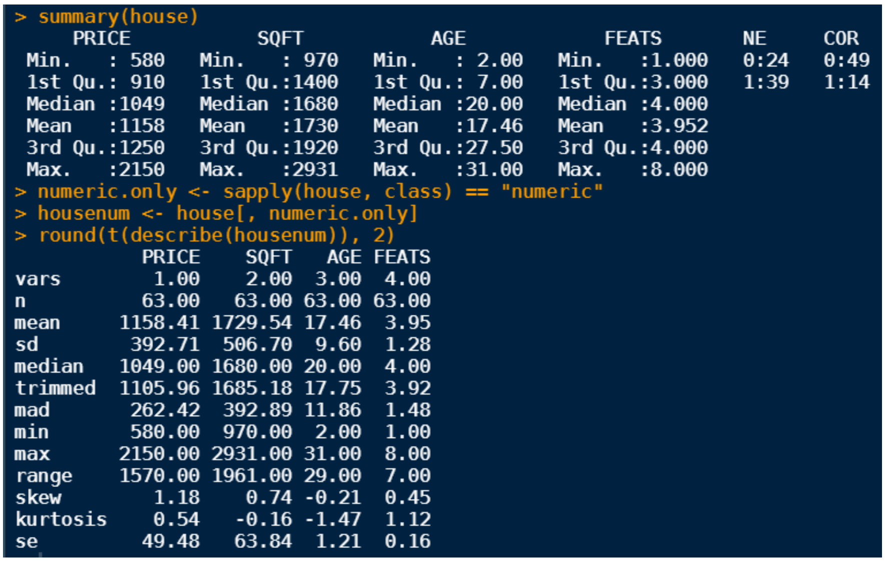


```
> house$PRICE <- as.numeric(house$PRICE)
> house$SQFT <- as.numeric(house$SQFT)
> house$AGE <- as.numeric(house$AGE)
> house$FEATS <- as.numeric(house$FEATS)
> house$NE <- as.factor(house$NE)
> house$COR <- as.factor(house$COR)
```

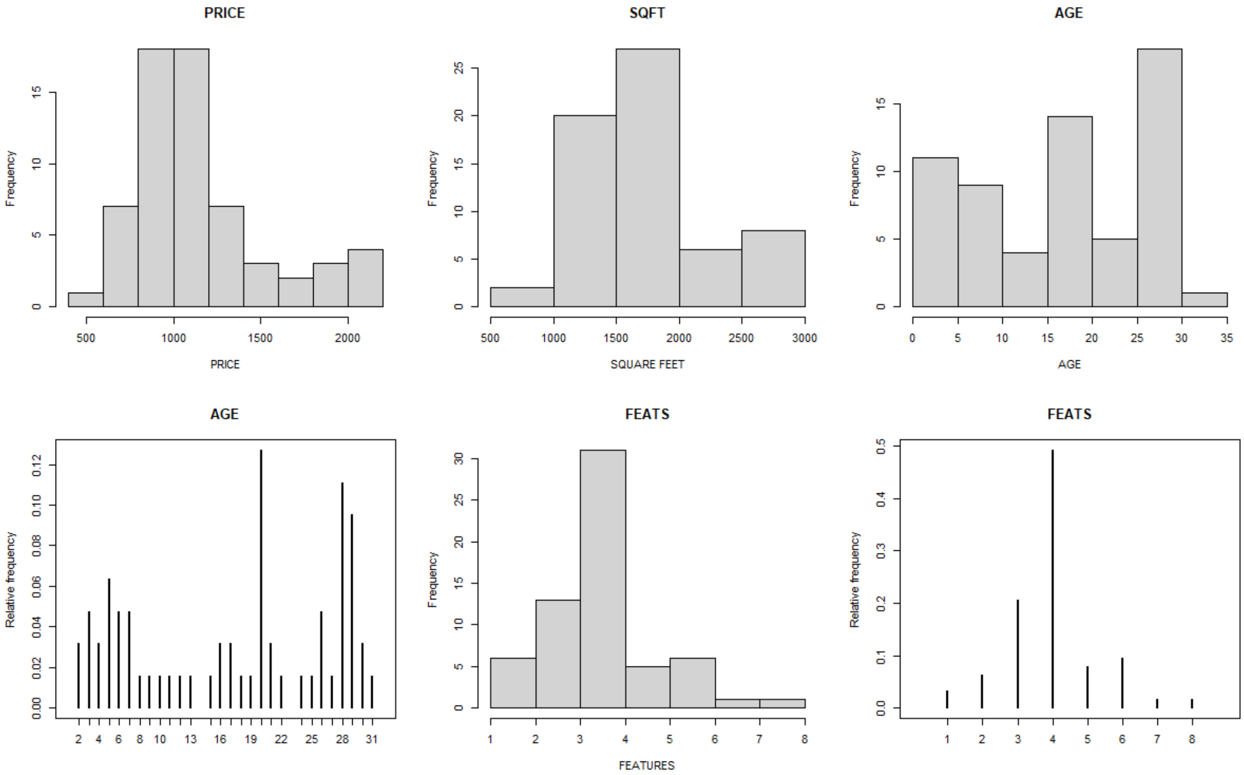
3rd Question

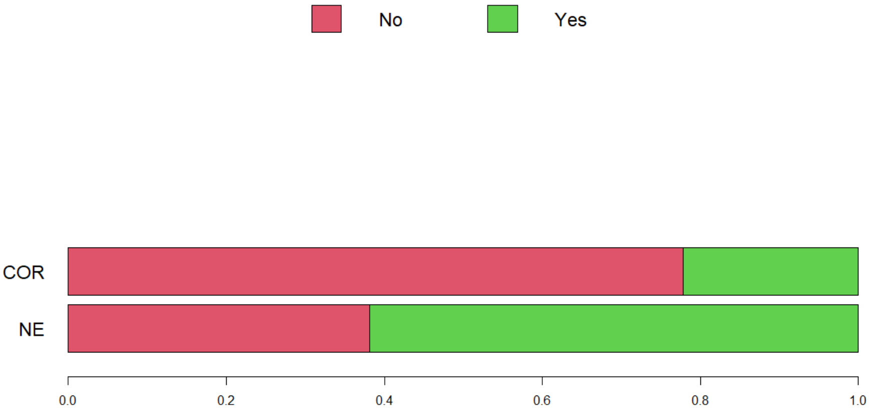
Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings.

3.1 Output in R



```
> summary(house)
     PRICE           SQFT             AGE             FEATS          NE      COR
 Min.   : 580   Min.   : 970   Min.   : 2.00   Min.   :1.000   0:24   0:49
 1st Qu.: 910   1st Qu.:1400   1st Qu.: 7.00   1st Qu.:3.000   1:39   1:14
 Median :1049   Median :1680   Median :20.00   Median :4.000
 Mean   :1158   Mean   :1730   Mean   :17.46   Mean   :3.952
 3rd Qu.:1250   3rd Qu.:1920   3rd Qu.:27.50   3rd Qu.:4.000
 Max.   :2150   Max.   :2931   Max.   :31.00   Max.   :8.000
> numeric.only <- sapply(house, class) == "numeric"
> housenum <- house[, numeric.only]
> round(t(describe(housenum)), 2)
           PRICE    SQFT   AGE FEATS
vars        1.00    2.00  3.00  4.00
n          63.00   63.00 63.00 63.00
mean     1158.41 1729.54 17.46  3.95
sd        392.71  506.70  9.60  1.28
median   1049.00 1680.00 20.00  4.00
trimmed  1105.96 1685.18 17.75  3.92
mad       262.42  392.89 11.86  1.48
min       580.00  970.00  2.00  1.00
max      2150.00 2931.00 31.00  8.00
range    1570.00 1961.00 29.00  7.00
skew        1.18    0.74 -0.21  0.45
kurtosis    0.54   -0.16 -1.47  1.12
se         49.48   63.84  1.21  0.16
```





3.2 Comment

First of all, we used summary in order to get a more generalized insight of our data and then we separated the numeric variables from the factors, to get a more profound view of the first variables that we are more interested in. In addition with these visualizations, we can conclude that:

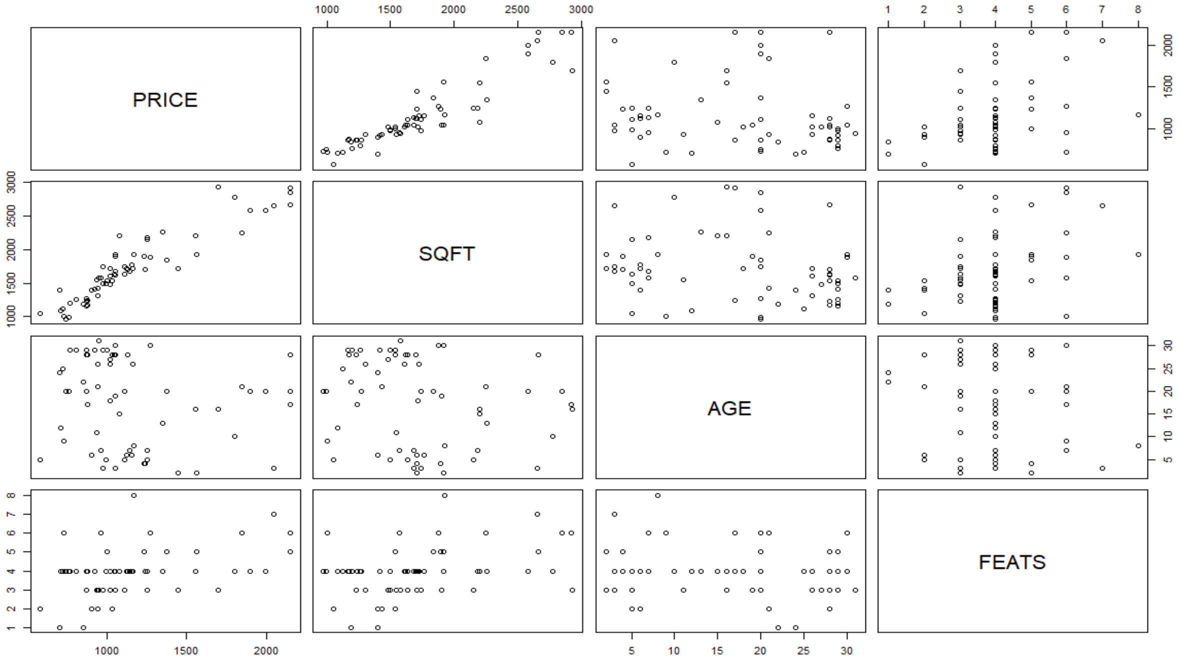
- Price:
    - Median ≠ Mean: asymmetry and no normally distributed variable, the average house is sold about 1158 thousand dollars
    - Trimmed mean: it has not many outliers, because it is close to the real mean
    - Skewness: asymmetry and specifically positive skewness (right tale)
    - Histogram: the most houses are sold more or less, but near to 1000 thousand dollars
- Square Feet:
    - Median ≠ Mean: asymmetry variable, the average house has 1730 sq ft of living space
    - Trimmed mean: it has not many outliers, because it is close to the real mean
    - Skewness: close to symmetric, but it is slightly right tale distribution
    - Kurtosis: the negative value shows that is more flattened than the normal distribution
    - Histogram: the most houses have approximately between 1500 and 2000 sq feet of living space
- Age:
    - Median ≠ Mean: asymmetry variable, the average house is about 17 years old
    - Trimmed mean is not very close to real mean and as we can observe from the graphs this variable has some outliers
    - Skewness: close to symmetric, but has a slightly negative skewness (left tale)
    - Kurtosis: the negative value shows that is more flattened than the normal distribution and this specifically variable can be described as platykurtic (like cube)
    - Histogram & Relative Frequencies plot: the majority of houses are 20 years old, but there is a significant number of houses that are 28 and 29 years old
- Features:
    - Median ≠ Mean: there is some kind of asymmetry here
    - Trimmed mean is very close to real mean, so there aren’t many outliers
    - Skewness: asymmetry and especially positive skewness (left tale)
    - Kurtosis: this value shows that is mesokyrtic, which means that is more pointed than the normal distribution
    - Histogram & Relative Frequencies plot: the majority of houses has 4 features, but there is an important amount of houses that has 3 features

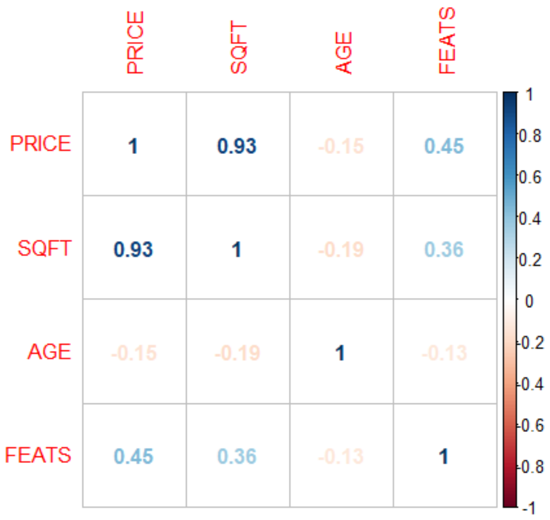
- Cornered houses: the 80% of the houses are not placed in a corner location (especially 49)
- Northeast Location: more than 60% of the houses that are sold are located in this section of the city (exactly 39)

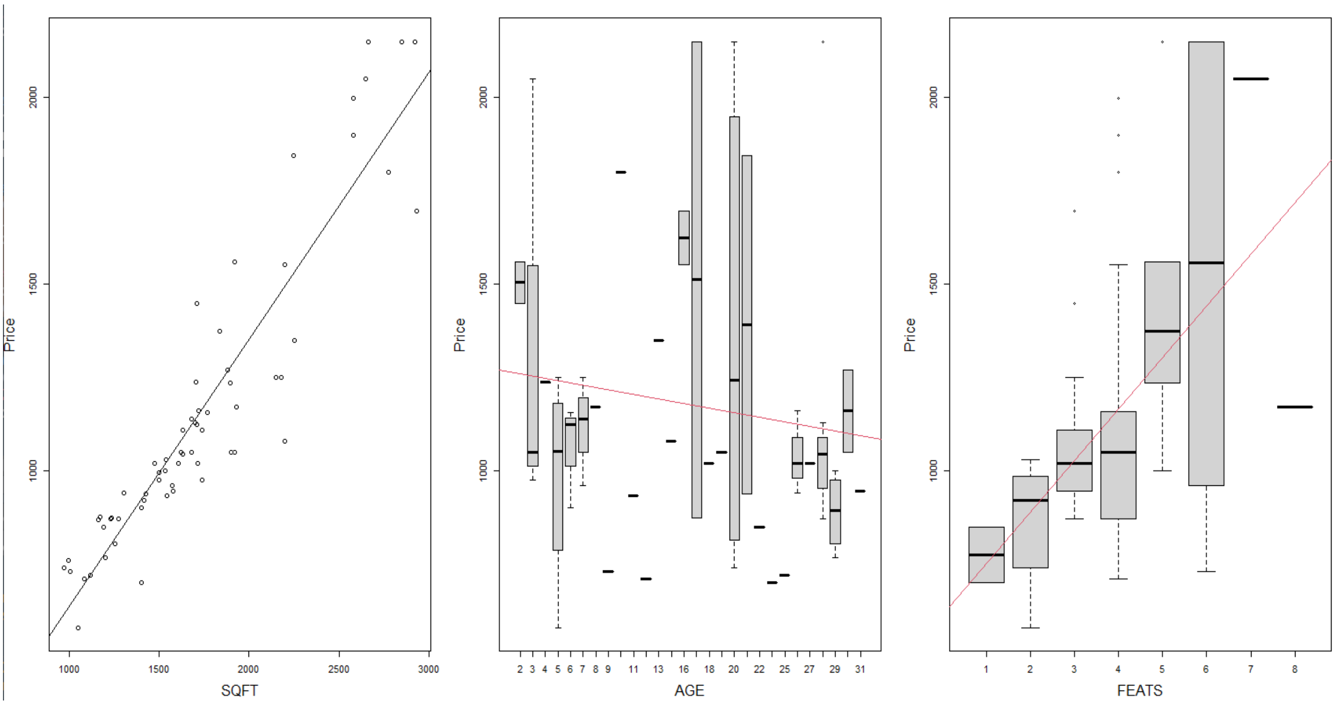
4th Question

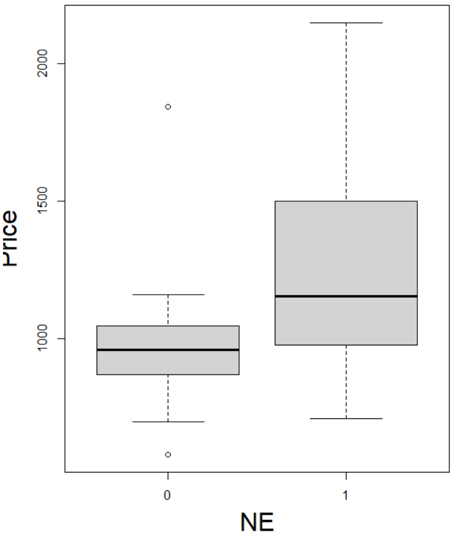
Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset. (Hint: Plot variables against one another and use correlation plots and measures for the numerical variables.). Comment on your findings. Is there a linear relationship between PRICE and any of the variables in the dataset?

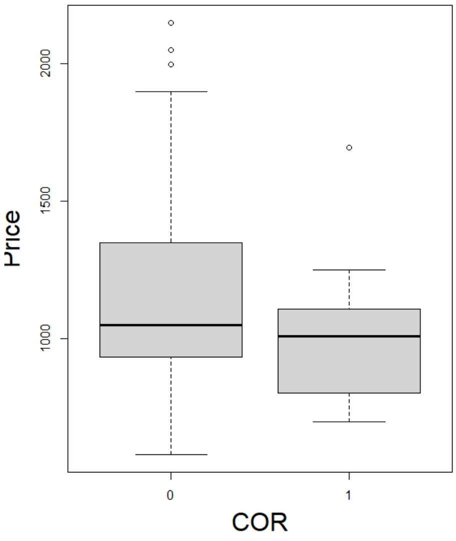
4.1 Output in R











4.2 Comment

To begin with, first it was plotted a general graph to see all the associations between the numeric variables, which is combined with the coloured table of the values of each correlation and then it had been plotted 3 more graphs that focuses on comparing price with squared feet, age and features. Finally, we created boxplots of the 2 factors in order to investigate if there is any association between them and the price, which in our case is the response variable.

From these graphs it can be easily concluded that the only variables that are highly associated is price and sq feet. This can be resulted both in the table of the correlations, which shows that it is 0.93 from 1 (nearly to perfect correlation) and in the other plots where can be derived that most of the values are near the line of an assumed perfect linear association. This association in other words can be said that is linear relationship.

Also, from the boxplot and from the correlation coefficient (r = 0.45) it can be derived that there is an assumption of linear relationship between price and features and to be exact they seem to be medium associated. This cannot be concluded in the plot because the features as variable is discreet.

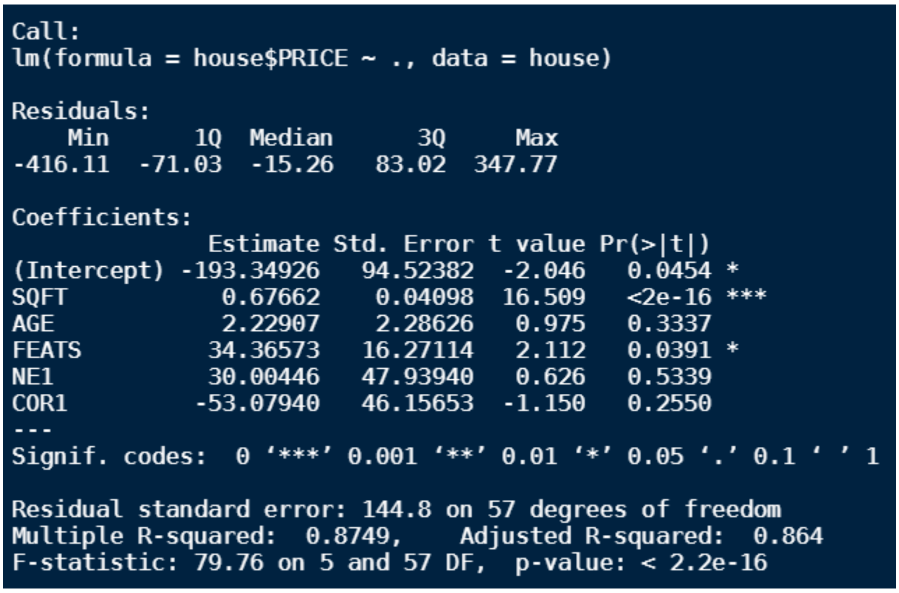
The other variables are not associated because not only their values in the plots are too dispersed, but their correlation coefficient is too small.

As far as the factors and price are concerned, we see that if one house is in the northeast side of the city, then the price is increased and so it seems to happen and if the house that is sold is not in the corner.

5th Question

Construct a model for the expected selling prices (PRICE) according to the remaining features.(hint: Conduct multiple regression having PRICE as a response and all the other variables as predictors). Does this linear model fit well to the data? (Hint: Comment on R^2 adj).

5.1 Output in R



```
Call:
lm(formula = house$PRICE ~ ., data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-416.11  -71.03  -15.26   83.02  347.77 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -193.34926   94.52382  -2.046   0.0454 *  
SQFT           0.67662    0.04098  16.509   <2e-16 ***
AGE            2.22907    2.28626   0.975   0.3337    
FEATS         34.36573   16.27114   2.112   0.0391 *  
NE1           30.00446   47.93940   0.626   0.5339    
COR1         -53.07940   46.15653  -1.150   0.2550    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 144.8 on 57 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.864 
F-statistic: 79.76 on 5 and 57 DF,  p-value: < 2.2e-16
```

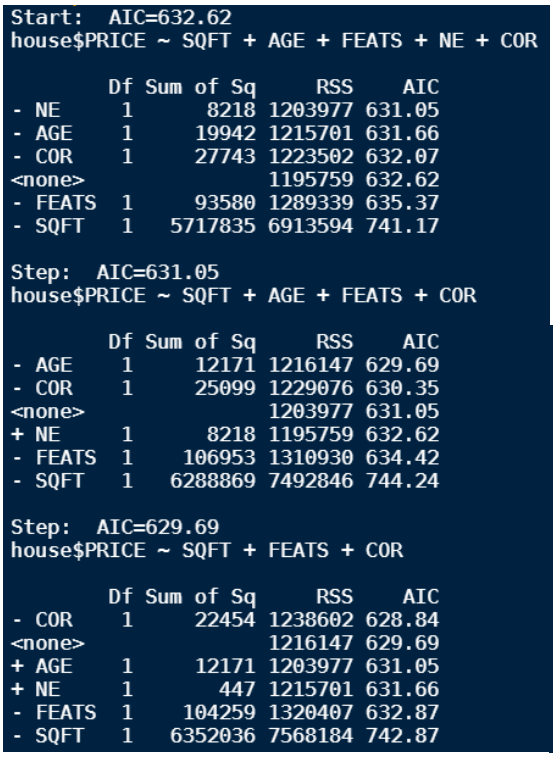
5.2 Comment

This linear model seems to fit well to the data, because the Adjusted R^2 is high (86.4%), which means good fit. Practically this means that nearly 86% of the variance is explained only using the price as covariate.

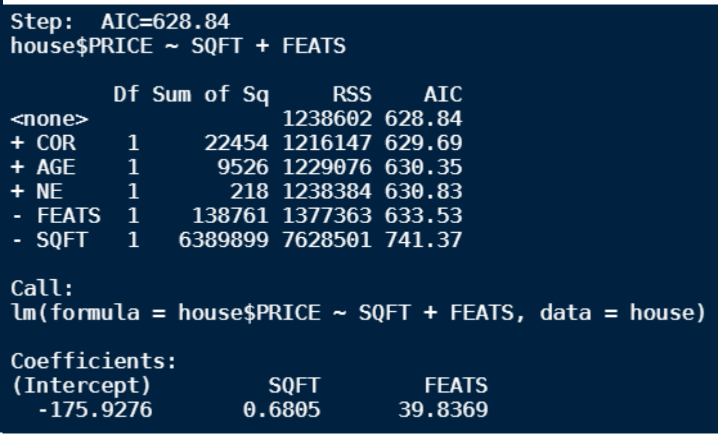
6th Question

Find the best model for predicting the selling prices (PRICE). Select the appropriate features using stepwise methods. (Hint: Use Forward, Backward or Stepwise procedure according to AIC or BIC to choose. Which variables appear to be more significant for predicting selling PRICES).

6.1 Output in R



```
Start:  AIC=632.62
house$PRICE ~ SQFT + AGE + FEATS + NE + COR

        Df Sum of Sq     RSS    AIC
- NE     1      8218 1203977 631.05
- AGE    1     19942 1215701 631.66
- COR    1     27743 1223502 632.07
<none>               1195759 632.62
- FEATS  1     93580 1289339 635.37
- SQFT   1   5717835 6913594 741.17

Step:  AIC=631.05
house$PRICE ~ SQFT + AGE + FEATS + COR

        Df Sum of Sq     RSS    AIC
- AGE    1     12171 1216147 629.69
- COR    1     25099 1229076 630.35
<none>               1203977 631.05
+ NE     1      8218 1195759 632.62
- FEATS  1    106953 1310930 634.42
- SQFT   1   6288869 7492846 744.24

Step:  AIC=629.69
house$PRICE ~ SQFT + FEATS + COR

        Df Sum of Sq     RSS    AIC
- COR    1     22454 1238602 628.84
<none>               1216147 629.69
+ AGE    1     12171 1203977 631.05
+ NE     1       447 1215701 631.66
- FEATS  1    104259 1320407 632.87
- SQFT   1   6352036 7568184 742.87
```



```
Step:  AIC=628.84
house$PRICE ~ SQFT + FEATS

        Df Sum of Sq     RSS    AIC
<none>               1238602 628.84
+ COR    1     22454 1216147 629.69
+ AGE    1      9526 1229076 630.35
+ NE     1       218 1238384 630.83
- FEATS  1    138761 1377363 633.53
- SQFT   1   6389899 7628501 741.37

Call:
lm(formula = house$PRICE ~ SQFT + FEATS, data = house)

Coefficients:
(Intercept)         SQFT        FEATS
  -175.9276       0.6805      39.8369
```

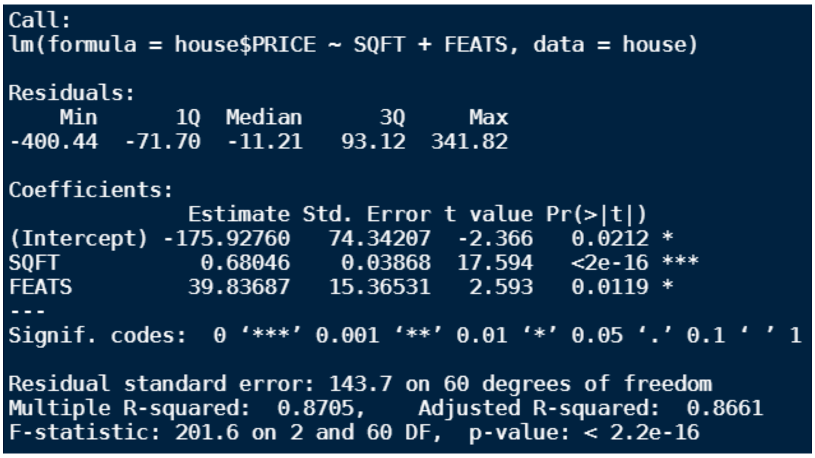
6.2 Comment

For predicting the selling prices we use Akaike's Information Criterion (AIC) with stepwise procedure in order to see which variable should be included or excluded in our model. Finally, we end up to exclude the age variable and the factors from the initial model, because they were too insignificant. So, the model consists of the price (as response) and the sq feet and the features as independent variables with constant -175.93 and coefficients 0.68 and 39.84 accordingly.

7th Question

Get the summary of your final model, (the model that you ended up having after conducting the stepwise procedure) and comment on the output. **Interpret the coefficients**. Comment on the significance of each coefficient and write down the mathematical formulation of the model (e.g PRICES = Intercept + coef1*Variable1 + coef2*Variable2 +…. + ε where ε ~ N(0, ...)). Should the intercept be excluded from our model?

7.1 Output in R



```
Call:
lm(formula = house$PRICE ~ SQFT + FEATS, data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-400.44  -71.70  -11.21   93.12  341.82 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -175.92760   74.34207  -2.366   0.0212 *  
SQFT           0.68046    0.03868  17.594   <2e-16 ***
FEATS         39.83687   15.36531   2.593   0.0119 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 143.7 on 60 degrees of freedom
Multiple R-squared:  0.8705,    Adjusted R-squared:  0.8661 
F-statistic: 201.6 on 2 and 60 DF,  p-value: < 2.2e-16
```

7.2 Comment

The full model which was created using stepwise procedure is:

PRICE = -175.93 + 0.68SQFT + 39.84FEATS + ε with ε ~ N(0, 143.7^2)

As we can see this model is pretty good for two reasons. First, the Adjusted R^2 is high enough and equal to 86.6% which means good fit and second the p-value of the intercept and the remaining variables is under the significant α level (5%), which means that these covariates are significant (Ho: βo = β1 = β2 = 0, where β is the coefficient of the covariates). Also, we have to mention that ε is the error and the sigma hat (= 143,7) measure the precision of model predictions.

The interpretation of the model is:

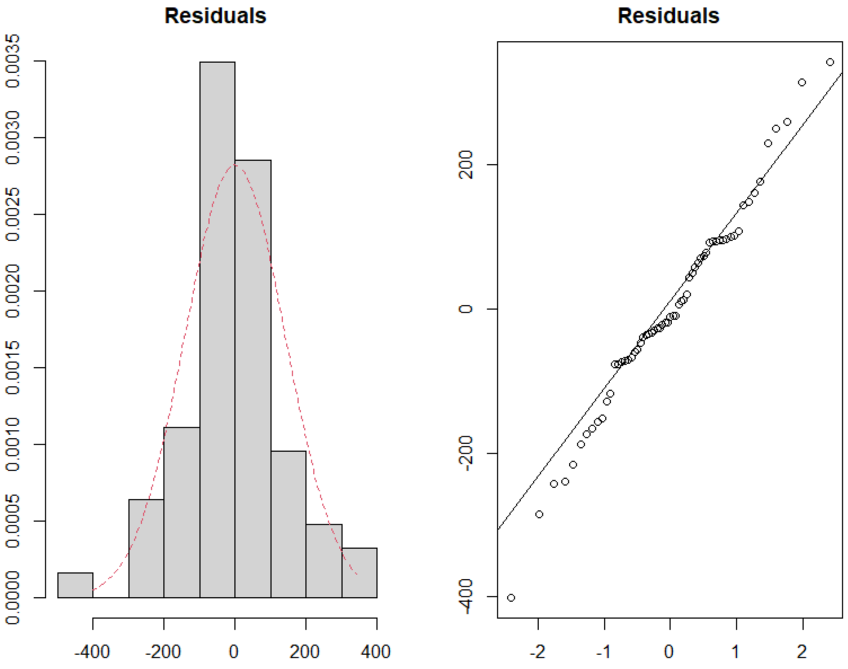
- if we have no sq feet and no features the price would be negative and equal to –175.93 $
- if we compare two houses with the same characteristics which differ only by 1 sq.ft, then the expected difference in the price will be 0.68$ (in thousands) in favour of the larger house.
- if we compare two houses with the same characteristics which differ only by 1 feature, then the expected difference in the price will be 39.84$ (in thousands) in favour of the larger house.

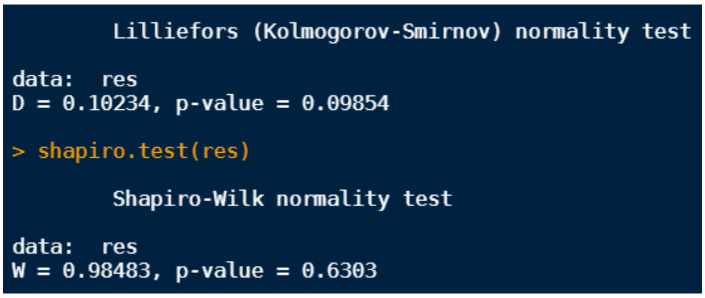
The intercept has not big p-value, so it should not be excluded from our model, but in our case where the constant variable has negative value it would be a good idea to be investigated more in order to examine if this value is reliable. The removal of the intercept causes many problems in the model and sometimes end up to wrong conclusions. This value that the constant covariate has is meaningless and is probably caused due to extrapolation.

8th Question

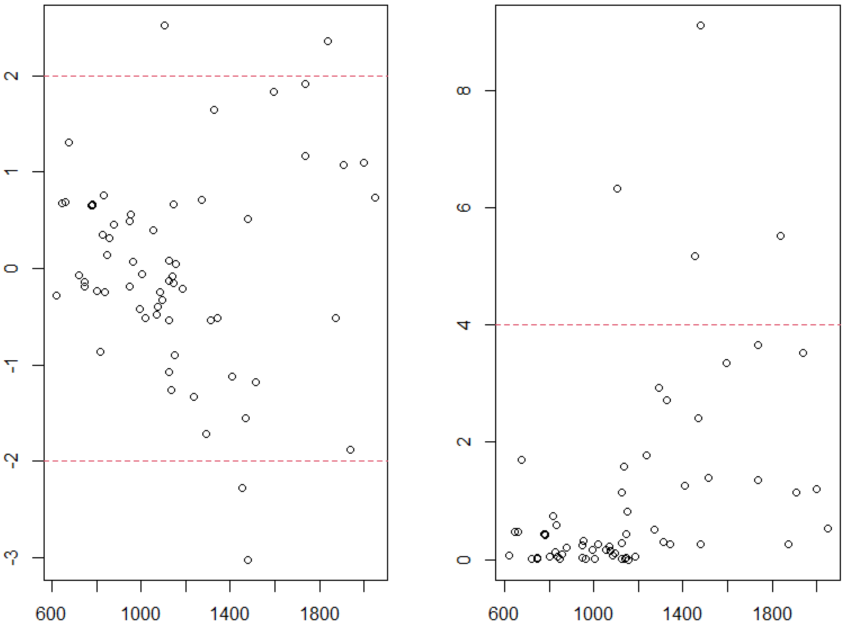
Check the assumptions of your final model. Are the assumptions satisfied? If not, what is the impact of the violation of the assumption not satisfied in terms of inference? What could someone do about it?

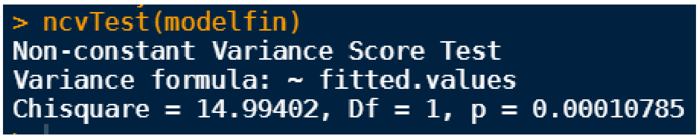
8.1 Output in R



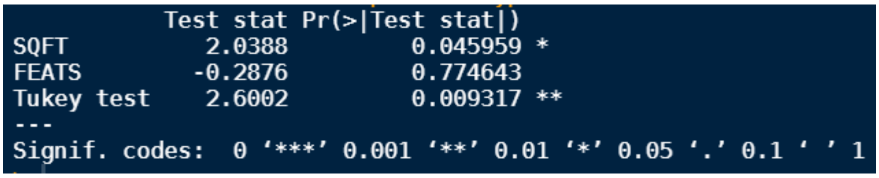


```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.10234, p-value = 0.09854

> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.98483, p-value = 0.6303
```

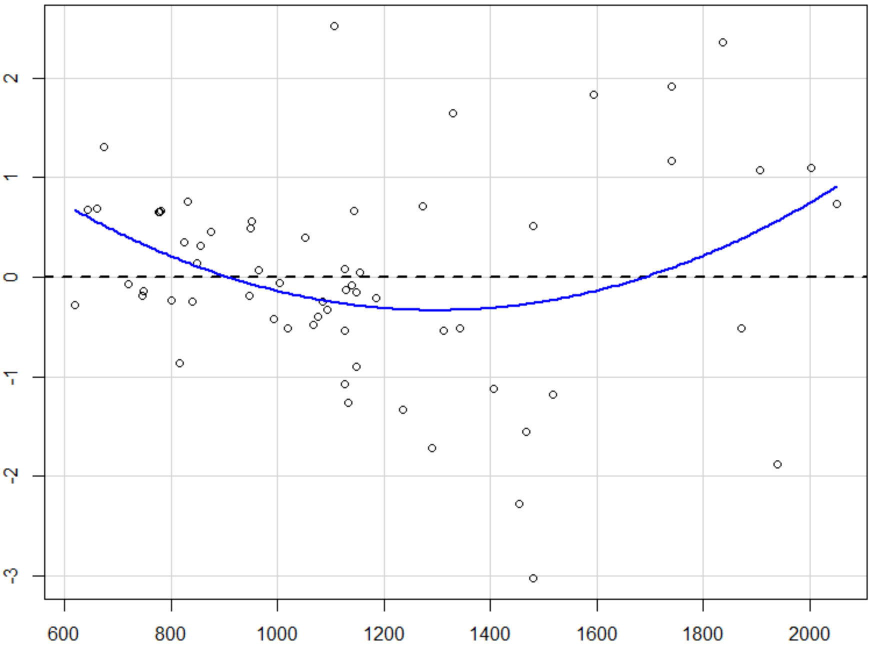


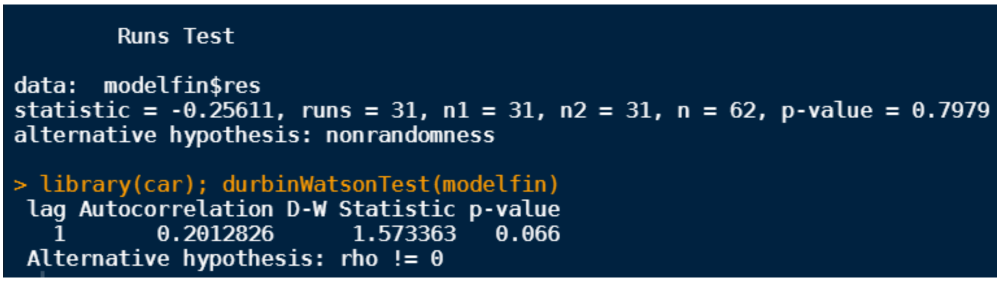


```
> ncvTest(modelfin)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 14.99402, Df = 1, p = 0.00010785
```



```
            Test stat Pr(>|Test stat|)
SQFT           2.0388         0.045959 *
FEATS         -0.2876         0.774643
Tukey test     2.6002         0.009317 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





```
        Runs Test

data:  modelfin$res
statistic = -0.25611, runs = 31, n1 = 31, n2 = 31, n = 62, p-value = 0.7979
alternative hypothesis: nonrandomness

> library(car); durbinWatsonTest(modelfin)
 lag Autocorrelation D-W Statistic p-value
   1       0.2012826      1.573363   0.066
 Alternative hypothesis: rho != 0
```

8.2 Comment

First, we check the normality of the residuals. We drew a histogram and a QQ plot and from them we can assume that the residuals are normally distributed. This assumption is confirmed and by the execution of Kolmogorov – Smirnov and Shapiro – Wilk normality tests, where the null Hypothesis, that the residuals follow the Normal distribution, is not rejected.

After that, we check if the variance is constant. From both the plots, the stundentized residuals and the squares of them, we end up that the variance is increasing. The exact same conclusion has been drawn after executing the Non-constant Variance score test, where we reject the null Hypothesis that the variance is non-constant. That conclusion violate our first assumption that the variance of the errors is constant and that may have many problems to our model, such as the estimators of the coefficients are still unbiased, the error variance estimator is not estimated correctly, standard errors are not estimated appropriately and may also affects the performance of the hypothesis tests and confidence intervals. So, in order to solve this problem someone should use weighted least squares regression models or use transformed response or use GLMs with more complicated distributions and if the above do not work then use GAMLSS to use covariates in the variance components.

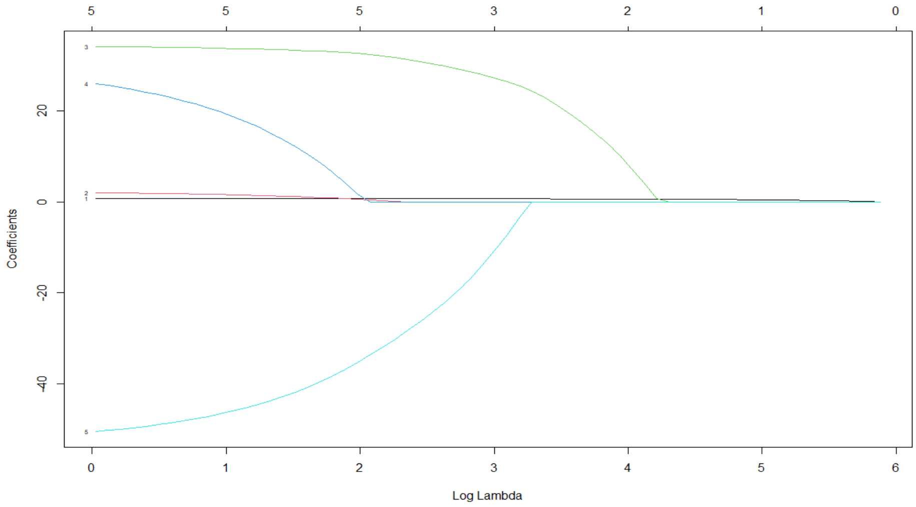
Next, we check if the residuals' linearity is violated. After the interpretation of the Tukey test we cannot be sure at this point that they have a linear relationship. This assumption is confirmed and in the plot where we can easily conclude that they are not linear related, because the blue line of the residuals do not follow the straight black line of linearity. This violation in simple words mean that the model is inadequate, especially for prediction and maybe the error variance will appear as non-constant, even if it is constant due to the model misspecification. So, someone in order to fix the problem should first try to transform the response, then if this does not work should try to transform the covariates, after that it has to be used polynomial regression or non-parametric regression models and finally if none of the above works then he/she has to use non-linear models.

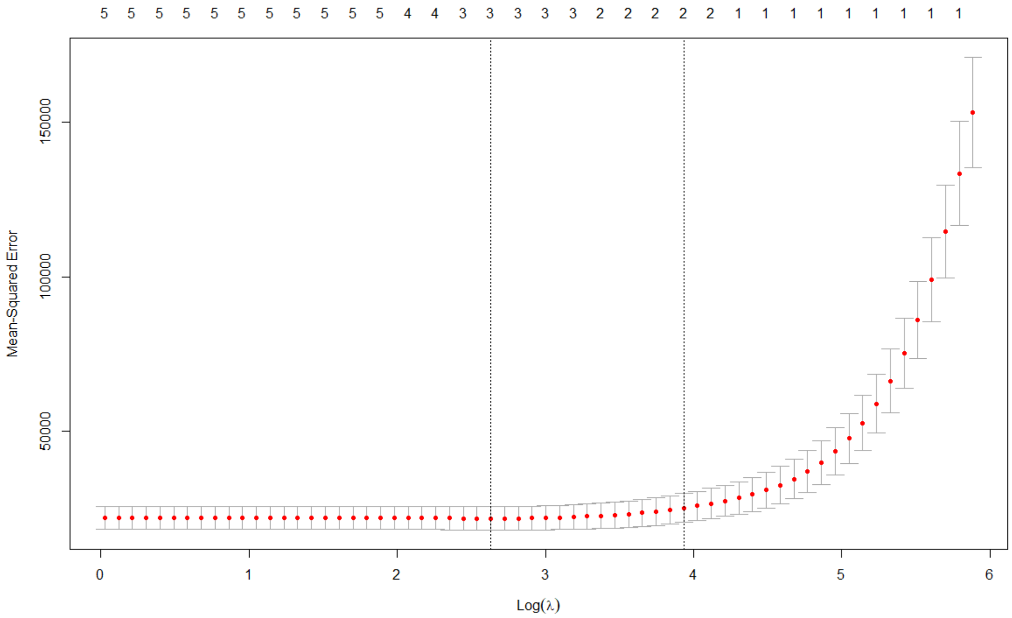
Finally, we check the independency of the errors. After executing the Runs Test and the Auto-correlation Durbin Watson test we take as a result that we do not reject the null Hypothesis that the errors are produced in a random manner or as the latter test says we do not reject the Ho that the autocorrelation of the disturbances is 0.

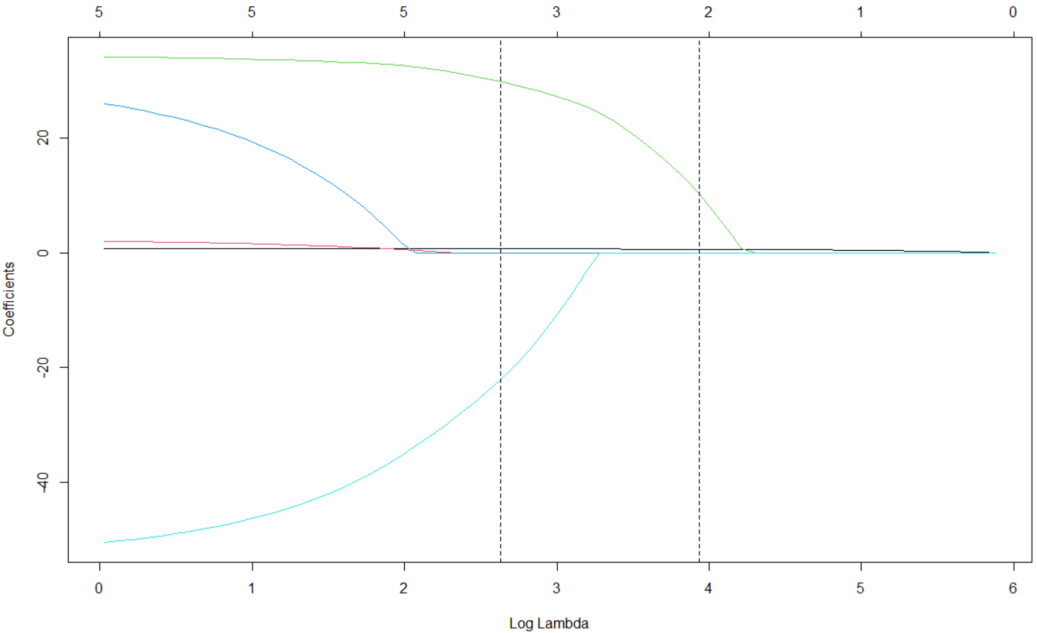
9th Question

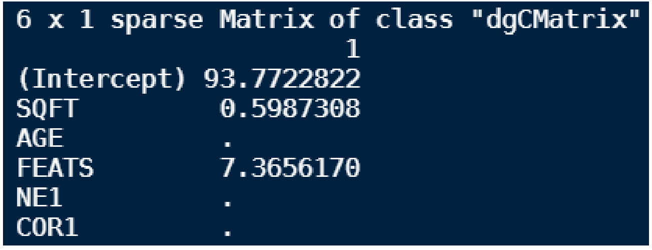
Conduct LASSO as a variable selection technique and compare the variables that you end up having using LASSO to the variables that you ended up having using stepwise methods in (VI). Are you getting the same results? Comment.

9.1 Output in R









```
6 x 1 sparse Matrix of class "dgCMatrix"
                     1
(Intercept) 93.7722822
SQFT         0.5987308
AGE          .        
FEATS        7.3656170
NE1          .        
COR1         .        
```

9.2 Comment

To begin with, we drew some LASSO plots to have a first look on how many variables the model will remove and what confidence intervals the mean-squared errors have. After that, we used lambda 1se that is the standard deviation away from the minimum lambda, but with higher penalty, because the minimum lambda usually leads to overfitted models. So, we fit the LASSO model and at the end we have the below equation:

$$\text{PRICE} = 93.77 + 0.6\text{SQFT} + 7.37\text{FEATS} + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2)$$

From the above, we can conclude that both the two models (AIC in question (VI) and LASSO) kept the same covariates, but with different coefficients. This is expected because by definition the LASSO model shrinks the coefficients in order to exclude some of the covariates. That is why and in our case, it seems that the constant in absolute value is much more smaller than the previous one in the AIC model and the same is applied and in the SQFT, where the difference is not very significant and in the FEATS variable.

R Code

The code that has been produced for this Lab Assignment is the following:

```
#Q1
require(foreign)
house <- read.table("usdata", header = T)
str(house)
#Q2
house$PRICE <- as.numeric(house$PRICE)
house$SQFT <- as.numeric(house$SQFT)
house$AGE <- as.numeric(house$AGE)
house$FEATS <- as.numeric(house$FEATS)
house$NE <- as.factor(house$NE)
house$COR <- as.factor(house$COR)
#Q3
summary(house)
library(psych)
#Numeric
numeric.only <- sapply(house, class) == "numeric"
housenum <- house[, numeric.only]
round(t(describe(housenum)), 2)
par(mfrow=c(2,3))
hist(housenum[,1], main=names(housenum)[1], xlab = names(housenum[1]))
hist(housenum[,2], main=names(housenum)[2], xlab = "SQUARE FEET")
hist(housenum[,3], main=names(housenum)[3], xlab = names(housenum[3]))
plot(table(housenum[,3])/nrow(housenum), type='h', xlim=range(housenum[,3])+c(-
1,1), main=names(housenum)[3], ylab='Relative frequency')
hist(housenum[,4], main=names(housenum)[4], xlab = "FEATURES")
plot(table(housenum[,4])/nrow(housenum), type='h', xlim=range(housenum[,4])+c(-
1,1), main=names(housenum)[4], ylab='Relative frequency')
#Factors
housefac <- house[ , !numeric.only]
par(mfrow=c(1,1))
par(mai=c(2.0,1.5,0.5,0.5))
```

```
barplot(sapply(housefac,table)/nrow(housefac), horiz=T, las=1, col=2:3, ylim=c(0,8),
cex.names=1.5)
legend('top', fil=2:3, legend=c("No","Yes"), ncol=2, bty='n',cex=1.5)
#Q4
#Pairs of numerical variables
pairs(housenum)
require(corrplot)
corrplot(cor(housenum), method = "number")
#More focused pairwise correlation
par(mfrow=c(1,3))
plot(housenum[,2], housenum[,1], xlab=names(housenum)[2],
ylab='Price',cex.lab=1.5)
abline(lm(housenum[,1]~housenum[,2]))
for(j in 3:4){
  boxplot(housenum[,1]~housenum[,j], xlab=names(housenum)[j],
ylab='Price',cex.lab=1.5)
  abline(lm(housenum[,1]~housenum[,j]),col=2)
}
#Pairs of price and factor variables
par(mfrow=c(1,2))
for(j in 1:2){
  boxplot(housenum[,1]~housefac[,j], xlab=names(housefac)[j],
ylab='Price',cex.lab=2.0)
}
#Q5
model <- lm(house$PRICE ~., data = house)
summary(model)
#Q6
modelfin <- step(model, direction = "both")
#Q7
summary(modelfin)
#Q8
#Normality of the residuals
res <- modelfin$residuals
par(mfrow=c(1,2), mai = c(0.5,0.5,0.5,0.5) )
mt <- 'Residuals'
hist(res, probability=T, main=mt)
x0<-seq(min(res), max(res),length.out=100)
y0<-dnorm(x0, mean(res),sd(res))
lines(x0,y0, col=2, lty=2)
qqnorm(res, main=mt)
qqline(res)

par(mfrow=c(1,2))
plot(yhat, Stud.residuals)
abline(h=c(-2,2), col=2, lty=2)
plot(yhat, Stud.residuals^2)
abline(h=4, col=2, lty=2)
```

```
library(nortest)
lillie.test(res)
shapiro.test(res)
library(car)
ncvTest(modelfin)
#Non linearity
library(car)
par(mfrow=c(1,1))
residualPlot(modelfin, type='rstudent')
residualPlots(modelfin, plot=F, type = "rstudent")
#Independence
library(randtests)
runs.test(modelfin$res)
library(car)
durbinWatsonTest(modelfin)
#Q9 - LASSO
require(glmnet)
X <- model.matrix(model)[,-1]
lasso <- glmnet(X, house$PRICE)
plot(lasso, xvar = "lambda", label = T)
lasso1 <- cv.glmnet(X, house$PRICE, alpha = 1)
plot(lasso1)
coef(lasso1, s = "lambda.1se")
plot(lasso1$glmnet.fit, xvar = "lambda")
abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)
```