SIM Project 2. Model fitting

Adrià Casanova Víctor Garcia Zhengyong Ji

November, 19th 2023

Table of Contents

# Clean workspace  
if(!is.null(dev.list())) dev.off()

## null device   
## 1

rm(list = ls())

# Load libraries  
library(car)

## 载入需要的程辑包：carData

library(mice)

##   
## 载入程辑包：'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(dplyr)

##   
## 载入程辑包：'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(missMDA)  
library(FactoMineR)  
library(chemometrics)

## 载入需要的程辑包：rpart

library(DataExplorer)

## Warning: 程辑包'DataExplorer'是用R版本4.3.2 来建造的

library(corrplot)

## corrplot 0.92 loaded

library(MASS)

##   
## 载入程辑包：'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(effects)

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

# Load data  
df = read.csv("train\_impute.csv")  
df\_test = read.csv("test\_impute.csv")  
  
# Declare factors  
df$OverallQual <- as.factor(df$OverallQual)  
df$MSSubClass <- as.factor(df$MSSubClass)  
char\_var <- which(sapply(df, is.character))  
df[,char\_var] <- lapply(df[, char\_var], as.factor)  
  
# Declare factors  
df\_test$OverallQual <- as.factor(df\_test$OverallQual)  
df\_test$MSSubClass <- as.factor(df\_test$MSSubClass)  
char\_var <- which(sapply(df\_test, is.character))  
df\_test[,char\_var] <- lapply(df\_test[, char\_var], as.factor)

# 8. First model building

We create a first model with all the numerical variables that we selected previously.

df\_num <- df[, which(sapply(df, is.numeric))]  
m0 = lm(SalePrice ~ ., data=df\_num)  
  
summary(m0)

##   
## Call:  
## lm(formula = SalePrice ~ ., data = df\_num)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -160830 -15890 -1092 14377 164012   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.783e+06 1.122e+05 -15.887 < 2e-16 \*\*\*  
## LotFrontage 8.821e+01 2.535e+01 3.480 0.000516 \*\*\*  
## LotArea 1.508e+00 2.702e-01 5.581 2.86e-08 \*\*\*  
## YearBuilt 4.597e+02 5.315e+01 8.650 < 2e-16 \*\*\*  
## YearRemodAdd 5.492e+02 5.246e+01 10.469 < 2e-16 \*\*\*  
## MasVnrArea 2.637e+01 6.338e+00 4.160 3.37e-05 \*\*\*  
## BsmtFinSF1 1.975e+01 4.851e+00 4.072 4.91e-05 \*\*\*  
## BsmtUnfSF 2.595e+00 4.872e+00 0.533 0.594302   
## TotalBsmtSF 3.133e+01 5.792e+00 5.408 7.45e-08 \*\*\*  
## X1stFlrSF -3.742e+01 1.263e+01 -2.964 0.003092 \*\*   
## X2ndFlrSF -2.545e+01 1.219e+01 -2.087 0.037071 \*   
## GrLivArea 9.523e+01 1.189e+01 8.008 2.40e-15 \*\*\*  
## BsmtFullBath 5.513e+02 2.138e+03 0.258 0.796537   
## FullBath -3.287e+03 2.399e+03 -1.370 0.170788   
## HalfBath -3.521e+03 2.304e+03 -1.528 0.126693   
## BedroomAbvGr -1.013e+04 1.447e+03 -6.998 3.99e-12 \*\*\*  
## TotRmsAbvGrd 4.475e+02 1.040e+03 0.430 0.667190   
## Fireplaces 8.700e+03 1.491e+03 5.836 6.59e-09 \*\*\*  
## GarageYrBlt -9.735e+01 6.592e+01 -1.477 0.139962   
## GarageCars 6.429e+03 2.464e+03 2.609 0.009169 \*\*   
## GarageArea 2.753e+01 8.866e+00 3.105 0.001943 \*\*   
## WoodDeckSF 2.326e+01 7.061e+00 3.294 0.001013 \*\*   
## OpenPorchSF 4.699e+01 1.561e+01 3.009 0.002664 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29940 on 1425 degrees of freedom  
## Multiple R-squared: 0.8233, Adjusted R-squared: 0.8206   
## F-statistic: 301.9 on 22 and 1425 DF, p-value: < 2.2e-16

vif(m0)

## LotFrontage LotArea YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1   
## 1.141502 1.463022 4.139242 1.895906 1.313313 6.894967   
## BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF GrLivArea BsmtFullBath   
## 7.381445 8.459562 33.083818 44.080849 53.497983 1.987127   
## FullBath HalfBath BedroomAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt   
## 2.743664 2.161873 2.172081 4.340452 1.467970 4.270628   
## GarageCars GarageArea WoodDeckSF OpenPorchSF   
## 5.396088 5.481694 1.173485 1.223935

There is a lot of features with a vif correlation bigger than 5. So in order to reduce the amount of workload, we decided to keep those that are less than 5 and has a high correlation to our target.

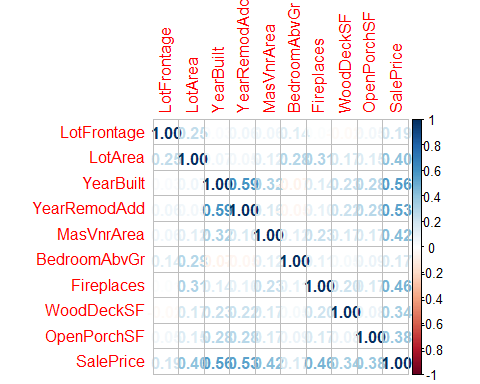
# Let's store the indices of the variables with at least one star in the lm and vif<5  
id\_num\_star1 = c(1:5,15,17,21:23)  
df\_num1 <- df\_num[, id\_num\_star1]  
# And build a new model only with significance features  
m1 = lm(SalePrice ~., data=df\_num1)  
summary(m1)

##   
## Call:  
## lm(formula = SalePrice ~ ., data = df\_num1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -170973 -25349 -4048 18791 207331   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.876e+06 1.159e+05 -24.820 < 2e-16 \*\*\*  
## LotFrontage 1.951e+02 3.456e+01 5.645 1.99e-08 \*\*\*  
## LotArea 3.960e+00 3.532e-01 11.212 < 2e-16 \*\*\*  
## YearBuilt 5.558e+02 4.789e+01 11.607 < 2e-16 \*\*\*  
## YearRemodAdd 9.359e+02 6.734e+01 13.897 < 2e-16 \*\*\*  
## MasVnrArea 8.678e+01 8.411e+00 10.317 < 2e-16 \*\*\*  
## BedroomAbvGr 5.492e+03 1.453e+03 3.781 0.000163 \*\*\*  
## Fireplaces 2.779e+04 1.875e+03 14.826 < 2e-16 \*\*\*  
## WoodDeckSF 5.749e+01 9.669e+00 5.946 3.44e-09 \*\*\*  
## OpenPorchSF 1.478e+02 2.112e+01 6.997 4.00e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 41800 on 1438 degrees of freedom  
## Multiple R-squared: 0.6525, Adjusted R-squared: 0.6503   
## F-statistic: 300 on 9 and 1438 DF, p-value: < 2.2e-16

vif(m1)

## LotFrontage LotArea YearBuilt YearRemodAdd MasVnrArea BedroomAbvGr   
## 1.088404 1.282087 1.723739 1.602800 1.186320 1.122948   
## Fireplaces WoodDeckSF OpenPorchSF   
## 1.190834 1.128839 1.149369

# As we can observe, the VIF correlation are much better, all the values are less than 2. So the next step is to chech the correlation beatween the features.  
corr\_mat <- cor(df\_num1)  
corrplot(corr\_mat, method = "number")



Feature “YearBuilt” and “YearRemodAdd” are highly correlated, and “YearBuilt” is more correlated to our target SalePrice. Hence, we remove YearRemodAdd in the next model.

# Building the model without "YearRemodAdd"  
id\_num\_star2 = c(1:3,5,15,17,21:23)  
df\_num2 <- df\_num[, id\_num\_star2]  
m2 = lm(SalePrice ~., data=df\_num2)  
summary(m2)

##   
## Call:  
## lm(formula = SalePrice ~ ., data = df\_num2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -165690 -27970 -5057 19803 205977   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.714e+06 8.539e+04 -20.069 < 2e-16 \*\*\*  
## LotFrontage 2.282e+02 3.670e+01 6.218 6.59e-10 \*\*\*  
## LotArea 3.810e+00 3.759e-01 10.136 < 2e-16 \*\*\*  
## YearBuilt 9.075e+02 4.329e+01 20.966 < 2e-16 \*\*\*  
## MasVnrArea 8.050e+01 8.942e+00 9.002 < 2e-16 \*\*\*  
## BedroomAbvGr 5.014e+03 1.546e+03 3.243 0.00121 \*\*   
## Fireplaces 2.795e+04 1.996e+03 14.006 < 2e-16 \*\*\*  
## WoodDeckSF 7.267e+01 1.023e+01 7.105 1.88e-12 \*\*\*  
## OpenPorchSF 1.917e+02 2.224e+01 8.619 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44500 on 1439 degrees of freedom  
## Multiple R-squared: 0.6058, Adjusted R-squared: 0.6036   
## F-statistic: 276.4 on 8 and 1439 DF, p-value: < 2.2e-16

Now, the most correlated variables in our model have at most a coefficient of correlation of 0.315, which in the context of real estate it is weak. We have obtained this information from <https://37parallel.com/real-estate-correlation/>.

Anova(m2)

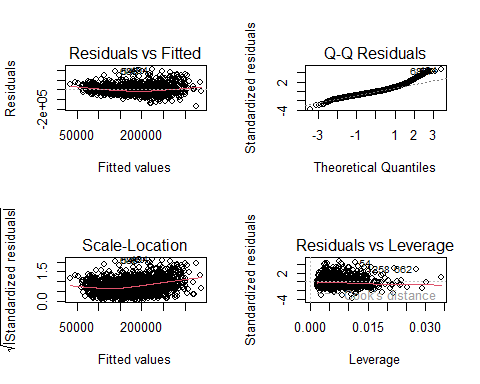
## Anova Table (Type II tests)  
##   
## Response: SalePrice  
## Sum Sq Df F value Pr(>F)   
## LotFrontage 7.6561e+10 1 38.662 6.588e-10 \*\*\*  
## LotArea 2.0346e+11 1 102.744 < 2.2e-16 \*\*\*  
## YearBuilt 8.7042e+11 1 439.553 < 2.2e-16 \*\*\*  
## MasVnrArea 1.6047e+11 1 81.034 < 2.2e-16 \*\*\*  
## BedroomAbvGr 2.0821e+10 1 10.515 0.001212 \*\*   
## Fireplaces 3.8845e+11 1 196.163 < 2.2e-16 \*\*\*  
## WoodDeckSF 9.9972e+10 1 50.485 1.883e-12 \*\*\*  
## OpenPorchSF 1.4712e+11 1 74.292 < 2.2e-16 \*\*\*  
## Residuals 2.8496e+12 1439   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova shows that all the variables we have kept are relevant.

# 9. Model analysis and iteration

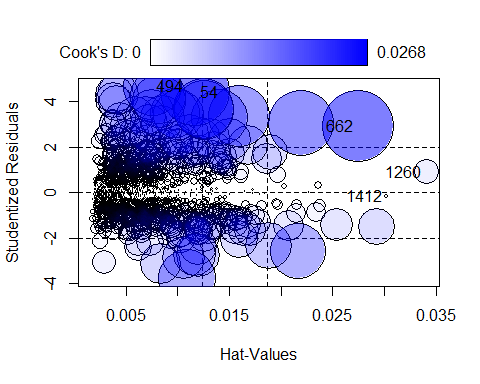
Let’s test the model with the residual plot and QQ plot. Later, we will study the possible transformations needed for each variable.

par(mfrow = c(2,2))  
plot(m2)



And we analyse if there are influent data and found that there are 3 observations that with a bigger Cook’s distance than the threshold (considered as 2 / sqrt(n)). Consequently, we decided to remove those observations.

# Check the influential plot before removing the influential observation.  
influencePlot(m2)

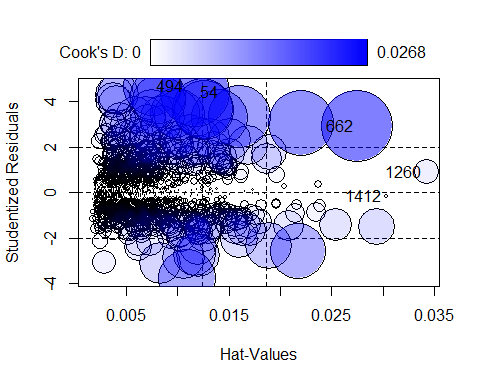


## StudRes Hat CookD  
## 54 4.4351943 0.011705638 2.555600e-02  
## 494 4.6798339 0.007521628 1.817802e-02  
## 662 2.9293234 0.027500038 2.681970e-02  
## 1260 0.9231784 0.034051251 3.338507e-03  
## 1412 -0.1455309 0.030239831 7.343087e-05

# Calculate D's threshold  
D\_thresh <- 2/sqrt(dim(df\_num2)[1]); D\_thresh

## [1] 0.05255883

#Remove the points and fit the model again  
influent <- c(1183, 692, 186)  
  
df <- df[-influent,]  
df\_num <- df[, which(sapply(df, is.numeric))]  
df\_num2 <- df\_num[, id\_num\_star2]  
m2 = lm(SalePrice ~., data=df\_num2)  
  
# Make influential plot after removing the influential observation.  
influencePlot(m2)



## StudRes Hat CookD  
## 54 4.4294986 0.011723124 0.0255291338  
## 494 4.6763230 0.007523824 0.0181559616  
## 662 2.9243598 0.027523168 0.0267522494  
## 1260 0.9172023 0.034157772 0.0033061236  
## 1412 -0.1449652 0.030259525 0.0000729101

###### Aixo jo no faria per aquest model

#step(m2)  
#Anova(m2)

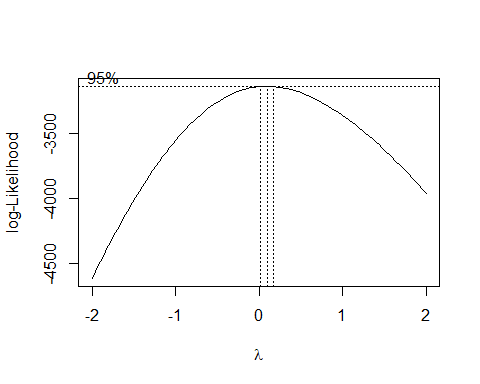
Errors aren’t normally distributed, so we will use boxTidwell() to test whether some transformation of the variables should be carried.

#residualPlots(m2)  
#avPlots(m2)  
#crPlot(m2)

###### Aixo jo no faria per aquest model

Firstly, we check with boxcox function to check if there is any needed transformation

boxcox(m2)

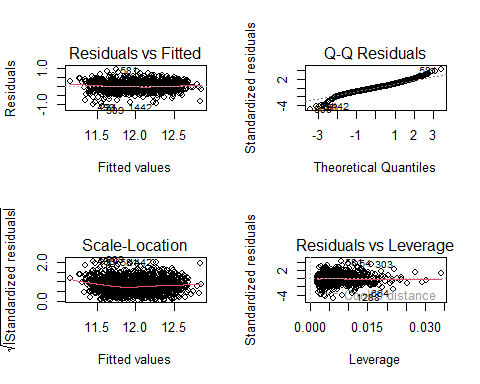


# As the lambda is over the 0, We should apply logaritmic transformation to SalePrice  
m3 = lm(log(SalePrice)~., data=df\_num2)  
summary(m3)

##   
## Call:  
## lm(formula = log(SalePrice) ~ ., data = df\_num2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.09727 -0.13827 -0.00372 0.12799 0.94710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.822e-02 4.444e-01 0.221 0.825   
## LotFrontage 8.611e-04 1.910e-04 4.507 7.09e-06 \*\*\*  
## LotArea 2.011e-05 1.956e-06 10.281 < 2e-16 \*\*\*  
## YearBuilt 5.743e-03 2.253e-04 25.489 < 2e-16 \*\*\*  
## MasVnrArea 2.988e-04 4.662e-05 6.410 1.97e-10 \*\*\*  
## BedroomAbvGr 5.453e-02 8.047e-03 6.777 1.78e-11 \*\*\*  
## Fireplaces 1.625e-01 1.038e-02 15.648 < 2e-16 \*\*\*  
## WoodDeckSF 3.624e-04 5.321e-05 6.811 1.42e-11 \*\*\*  
## OpenPorchSF 9.790e-04 1.157e-04 8.462 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2314 on 1436 degrees of freedom  
## Multiple R-squared: 0.6412, Adjusted R-squared: 0.6392   
## F-statistic: 320.8 on 8 and 1436 DF, p-value: < 2.2e-16

Compared with the m2, the Ajusted R-squared has increased about 4%.

par(mfrow=c(2,2))  
plot(m3, id.n=5)



The residual’s distribution is closer to normal now. However, it isn’t normal yet.

#boxTidwell(SalePrice ~ ., data=df\_num2) THIS GIVES ERROR because most variables have null values  
# We'll assign 10^(-6) to all cells equal to 0 to be able to use boxTidwell without altering too much the model  
  
df\_num2 = replace(df\_num2, df\_num2 == 0, 1e-6)  
summary(df\_num2)

## LotFrontage LotArea YearBuilt MasVnrArea   
## Min. : 0.00 Min. : 1300 Min. :1872 Min. : 0.00   
## 1st Qu.: 42.00 1st Qu.: 7500 1st Qu.:1954 1st Qu.: 0.00   
## Median : 63.00 Median : 9375 Median :1972 Median : 0.00   
## Mean : 57.05 Mean : 9493 Mean :1971 Mean : 90.18   
## 3rd Qu.: 78.00 3rd Qu.:11316 3rd Qu.:2000 3rd Qu.:158.99   
## Max. :182.00 Max. :23595 Max. :2010 Max. :664.00   
## BedroomAbvGr Fireplaces WoodDeckSF OpenPorchSF   
## Min. :0.000001 Min. :0.000001 Min. : 0.00 Min. : 0.00   
## 1st Qu.:2.000000 1st Qu.:0.000001 1st Qu.: 0.00 1st Qu.: 0.00   
## Median :3.000000 Median :1.000000 Median : 0.00 Median : 24.00   
## Mean :2.861519 Mean :0.605537 Mean : 92.28 Mean : 42.55   
## 3rd Qu.:3.000000 3rd Qu.:1.000000 3rd Qu.:168.00 3rd Qu.: 65.00   
## Max. :6.000000 Max. :3.000000 Max. :670.00 Max. :267.00   
## SalePrice   
## Min. : 34900   
## 1st Qu.:129900   
## Median :162000   
## Mean :177697   
## 3rd Qu.:213000   
## Max. :465000

# boxTidwell(log(SalePrice)~., data=df\_num2) THIS GIVES ERROR AS WELL because the model has too many variables  
  
boxTidwell(log(SalePrice) ~ LotArea+YearBuilt+MasVnrArea, data = df\_num2)

## MLE of lambda Score Statistic (t) Pr(>|t|)   
## LotArea 0.46268 -4.3123 1.725e-05 \*\*\*  
## YearBuilt 66.57971 14.5973 < 2.2e-16 \*\*\*  
## MasVnrArea 1.01690 0.0152 0.9879   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 5   
##   
## Score test for null hypothesis that all lambdas = 1:  
## F = 77.374, df = 3 and 1438, Pr(>F) = < 2.2e-16

# We should apply sqrt(LotArea). The lambda for YearBuilt is too large, so it would difficult to interpet the model using it. MasVnrArea has a too large p-value, so we cannot reject the null hypothesis that lambda = 1.  
boxTidwell(log(SalePrice)~LotFrontage, data = df\_num2)

## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :  
## maximum iterations exceeded

## MLE of lambda Score Statistic (t) Pr(>|t|)   
## -3.1109 11.028 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 26

# Too small lambda  
boxTidwell(log(SalePrice)~BedroomAbvGr, data = df\_num2)

## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :  
## maximum iterations exceeded

## MLE of lambda Score Statistic (t) Pr(>|t|)  
## 0.98657 0.3194 0.7494  
##   
## iterations = 26

# Too large p-value  
boxTidwell(log(SalePrice)~Fireplaces, data =df\_num2)

## MLE of lambda Score Statistic (t) Pr(>|t|)   
## 0.17624 -8.0252 2.083e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 3

# We apply log() to Fireplaces  
boxTidwell(log(SalePrice)~WoodDeckSF, data = df\_num2)

## MLE of lambda Score Statistic (t) Pr(>|t|)   
## 0.50697 -5.2996 1.341e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 7

# We apply sqrt() to WoodDeckSF  
boxTidwell(log(SalePrice)~OpenPorchSF, data = df\_num2)

## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :  
## maximum iterations exceeded

## MLE of lambda Score Statistic (t) Pr(>|t|)   
## -7.8358 -11.723 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 26

# Too small lambda

Using the boxTidwell method, below transformation can be applied in the model m4.

m4 = lm(log(SalePrice) ~ LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+  
 BedroomAbvGr+log(Fireplaces)+sqrt(WoodDeckSF)+OpenPorchSF,  
 data=df\_num2)  
summary(m4)

##   
## Call:  
## lm(formula = log(SalePrice) ~ LotFrontage + sqrt(LotArea) + YearBuilt +   
## MasVnrArea + BedroomAbvGr + log(Fireplaces) + sqrt(WoodDeckSF) +   
## OpenPorchSF, data = df\_num2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.10276 -0.14161 -0.00581 0.13022 0.87128   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.403e-01 4.544e-01 1.629 0.103443   
## LotFrontage 6.910e-04 1.919e-04 3.601 0.000327 \*\*\*  
## sqrt(LotArea) 4.130e-03 3.631e-04 11.373 < 2e-16 \*\*\*  
## YearBuilt 5.418e-03 2.293e-04 23.623 < 2e-16 \*\*\*  
## MasVnrArea 3.151e-04 4.654e-05 6.770 1.87e-11 \*\*\*  
## BedroomAbvGr 5.103e-02 8.086e-03 6.311 3.70e-10 \*\*\*  
## log(Fireplaces) 1.467e-02 9.639e-04 15.218 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 6.185e-03 9.073e-04 6.817 1.37e-11 \*\*\*  
## OpenPorchSF 9.839e-04 1.157e-04 8.505 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2313 on 1436 degrees of freedom  
## Multiple R-squared: 0.6416, Adjusted R-squared: 0.6396   
## F-statistic: 321.3 on 8 and 1436 DF, p-value: < 2.2e-16

Adjusted R-squared has increased slightly. Since we cannot find a significant improvement, we will compare m3 and m4 with more advanced tools such as AIC and BIC.

AIC(m3, m4)

## df AIC  
## m3 10 -118.1671  
## m4 10 -119.6517

BIC(m3, m4)

## df BIC  
## m3 10 -65.40847  
## m4 10 -66.89307

Checking the AIC and BIC, the overall improvement of applying all the changes simultaneously is small, so we decided to check different combination to find a better result.

We’ll try now to only apply a few of the relevant transformations.

m5 = lm(log(SalePrice) ~ LotFrontage+LotArea+YearBuilt+MasVnrArea+BedroomAbvGr+log(Fireplaces)+sqrt(WoodDeckSF)+OpenPorchSF,data=df\_num2)  
m6 = lm(log(SalePrice) ~ LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF,data=df\_num2)  
m7 = lm(log(SalePrice) ~ LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+log(Fireplaces)+WoodDeckSF+OpenPorchSF,data=df\_num2)  
m8 = lm(log(SalePrice)~LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+WoodDeckSF+OpenPorchSF, data=df\_num2)  
m9 = lm(log(SalePrice)~LotFrontage+LotArea+YearBuilt+MasVnrArea+BedroomAbvGr+log(Fireplaces)+WoodDeckSF+OpenPorchSF, data=df\_num2)  
m10 = lm(log(SalePrice)~LotFrontage+LotArea+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF, data=df\_num2)  
AIC(m4,m5,m6,m7,m8,m9,m10)

## df AIC  
## m4 10 -119.6517  
## m5 10 -109.4902  
## m6 10 -130.8424  
## m7 10 -117.0384  
## m8 10 -127.5111  
## m9 10 -107.0498  
## m10 10 -121.3265

BIC(m4,m5,m6,m7,m8,m9,m10)

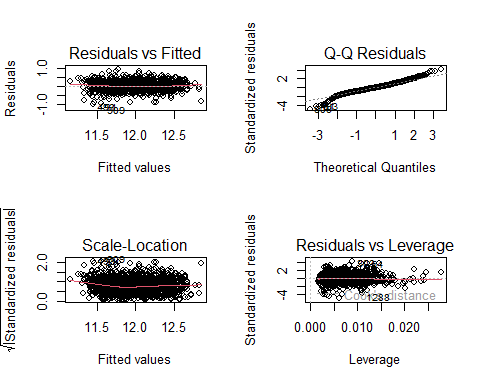
## df BIC  
## m4 10 -66.89307  
## m5 10 -56.73155  
## m6 10 -78.08378  
## m7 10 -64.27974  
## m8 10 -74.75244  
## m9 10 -54.29115  
## m10 10 -68.56782

The best model is m6, that is only applying the sqrt transformations in LotArea and WoodDeckSF. For this model we have compared the distribution of residuals and saw that it was very similar to the original model

par(mfrow=c(2,2))  
m11 = lm(log(SalePrice) ~ LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF,data=df\_num)  
BIC(m3,m11)

## df BIC  
## m3 10 -65.40847  
## m11 10 -78.08351

plot(m11)



# 10. Adding Factors to the numerical model

As there were an important number of numeric variables we tried to add factor variables one by one. We started with the most correlated variable with the target and continued in decreasing order. To test if they increase the model forecasting capability we analysed the BIC and the distribution of residuals.

For example, we added “ExterQual” and saw that it increased R^2 while reducing the BIC. Moreover, Anova and step methods suggested that adding the variable was relevant. Hence, we are following a heuristic approach.

m12 = lm(log(SalePrice)~LotFrontage+sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual, data=df)  
BIC(m11,m12)

## df BIC  
## m11 10 -78.08351  
## m12 14 -641.83724

Anova(m12)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## LotFrontage 0.015 1 0.4279 0.5131   
## sqrt(LotArea) 6.022 1 170.5308 < 2.2e-16 \*\*\*  
## YearBuilt 7.605 1 215.3492 < 2.2e-16 \*\*\*  
## MasVnrArea 1.006 1 28.5009 1.088e-07 \*\*\*  
## BedroomAbvGr 1.036 1 29.3444 7.098e-08 \*\*\*  
## Fireplaces 5.526 1 156.4860 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 1.302 1 36.8707 1.614e-09 \*\*\*  
## OpenPorchSF 1.374 1 38.9232 5.792e-10 \*\*\*  
## OverallQual 25.650 4 181.5928 < 2.2e-16 \*\*\*  
## Residuals 50.568 1432   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m12, k = log(nrow(df)))

## Start: AIC=-4749.85  
## log(SalePrice) ~ LotFrontage + sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual  
##   
## Df Sum of Sq RSS AIC  
## - LotFrontage 1 0.0151 50.583 -4756.7  
## <none> 50.568 -4749.8  
## - MasVnrArea 1 1.0064 51.574 -4728.6  
## - BedroomAbvGr 1 1.0362 51.604 -4727.8  
## - sqrt(WoodDeckSF) 1 1.3020 51.870 -4720.4  
## - OpenPorchSF 1 1.3745 51.942 -4718.4  
## - Fireplaces 1 5.5260 56.094 -4607.3  
## - sqrt(LotArea) 1 6.0219 56.590 -4594.5  
## - YearBuilt 1 7.6046 58.172 -4554.7  
## - OverallQual 4 25.6502 76.218 -4186.1  
##   
## Step: AIC=-4756.69  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 50.583 -4756.7  
## - MasVnrArea 1 1.0094 51.592 -4735.4  
## - BedroomAbvGr 1 1.0552 51.638 -4734.1  
## - sqrt(WoodDeckSF) 1 1.2880 51.871 -4727.6  
## - OpenPorchSF 1 1.3702 51.953 -4725.3  
## - Fireplaces 1 5.5291 56.112 -4614.1  
## - sqrt(LotArea) 1 6.5588 57.142 -4587.8  
## - YearBuilt 1 7.5898 58.173 -4562.0  
## - OverallQual 4 26.4801 77.063 -4177.4

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 5.0182743 0.0039333 0.0030914   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0002047 0.0367297 0.1078671   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0044644 0.0005940 0.4616675   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1992740 -0.5850392 0.7094287

Comparing the m11 and m12, there was a huge improvement in terms of BIC and Adjusted R-squared as we expected.

The Anova test indicates that LotFrontage has lose significance once we add the categorical feature, and the step method suggest to remove it.

m12.1 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual, data=df); summary(m12.1)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0228 -0.1026 0.0022 0.1080 0.6162   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.018e+00 4.139e-01 12.126 < 2e-16 \*\*\*  
## sqrt(LotArea) 3.933e-03 2.885e-04 13.631 < 2e-16 \*\*\*  
## YearBuilt 3.091e-03 2.108e-04 14.663 < 2e-16 \*\*\*  
## MasVnrArea 2.047e-04 3.829e-05 5.347 1.04e-07 \*\*\*  
## BedroomAbvGr 3.673e-02 6.718e-03 5.468 5.38e-08 \*\*\*  
## Fireplaces 1.079e-01 8.619e-03 12.515 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 4.464e-03 7.390e-04 6.041 1.95e-09 \*\*\*  
## OpenPorchSF 5.940e-04 9.534e-05 6.230 6.10e-10 \*\*\*  
## OverallQualGood 4.617e-01 2.157e-02 21.403 < 2e-16 \*\*\*  
## OverallQualModerate 1.993e-01 1.820e-02 10.951 < 2e-16 \*\*\*  
## OverallQualVBad -5.850e-01 8.625e-02 -6.783 1.71e-11 \*\*\*  
## OverallQualVGood 7.094e-01 3.526e-02 20.121 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1879 on 1433 degrees of freedom  
## Multiple R-squared: 0.764, Adjusted R-squared: 0.7621   
## F-statistic: 421.6 on 11 and 1433 DF, p-value: < 2.2e-16

BIC(m10,m12,m12.1)

## df BIC  
## m10 10 -68.56782  
## m12 14 -641.83724  
## m12.1 13 -648.68143

After removing the LotFrontage, although the R\_sqrt didn’t change, the BIC of m12.1 increased because we are using less variable (avoid ovefitting), which is good.

Next, in the model m13, we have add the ExterQual to out model.

m13 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual+ExterQual, data=df); summary(m13)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + ExterQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02914 -0.10041 -0.00212 0.10935 0.62350   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.429e+00 4.361e-01 14.742 < 2e-16 \*\*\*  
## sqrt(LotArea) 3.979e-03 2.816e-04 14.128 < 2e-16 \*\*\*  
## YearBuilt 2.468e-03 2.195e-04 11.240 < 2e-16 \*\*\*  
## MasVnrArea 1.916e-04 3.751e-05 5.109 3.67e-07 \*\*\*  
## BedroomAbvGr 3.943e-02 6.569e-03 6.001 2.48e-09 \*\*\*  
## Fireplaces 1.070e-01 8.414e-03 12.715 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 4.280e-03 7.222e-04 5.927 3.86e-09 \*\*\*  
## OpenPorchSF 5.112e-04 9.360e-05 5.461 5.58e-08 \*\*\*  
## OverallQualGood 3.931e-01 2.261e-02 17.390 < 2e-16 \*\*\*  
## OverallQualModerate 1.911e-01 1.789e-02 10.682 < 2e-16 \*\*\*  
## OverallQualVBad -5.164e-01 8.897e-02 -5.803 7.99e-09 \*\*\*  
## OverallQualVGood 5.814e-01 4.176e-02 13.922 < 2e-16 \*\*\*  
## ExterQualFa -3.314e-01 6.645e-02 -4.987 6.87e-07 \*\*\*  
## ExterQualGd -7.916e-02 3.738e-02 -2.118 0.0344 \*   
## ExterQualTA -2.013e-01 3.948e-02 -5.097 3.90e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1833 on 1430 degrees of freedom  
## Multiple R-squared: 0.7757, Adjusted R-squared: 0.7735   
## F-statistic: 353.3 on 14 and 1430 DF, p-value: < 2.2e-16

BIC(m13,m12.1)

## df BIC  
## m13 16 -700.8136  
## m12.1 13 -648.6814

Anova(m13)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 6.708 1 199.604 < 2.2e-16 \*\*\*  
## YearBuilt 4.246 1 126.341 < 2.2e-16 \*\*\*  
## MasVnrArea 0.877 1 26.105 3.670e-07 \*\*\*  
## BedroomAbvGr 1.210 1 36.016 2.477e-09 \*\*\*  
## Fireplaces 5.433 1 161.659 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 1.181 1 35.128 3.863e-09 \*\*\*  
## OpenPorchSF 1.002 1 29.823 5.576e-08 \*\*\*  
## OverallQual 13.417 4 99.809 < 2.2e-16 \*\*\*  
## ExterQual 2.524 3 25.032 8.731e-16 \*\*\*  
## Residuals 48.059 1430   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m13, k = log(nrow(df)))

## Start: AIC=-4808.82  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## ExterQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 48.059 -4808.8  
## - MasVnrArea 1 0.8773 48.936 -4790.0  
## - OpenPorchSF 1 1.0023 49.061 -4786.3  
## - sqrt(WoodDeckSF) 1 1.1806 49.240 -4781.0  
## - BedroomAbvGr 1 1.2104 49.270 -4780.2  
## - ExterQual 3 2.5239 50.583 -4756.7  
## - YearBuilt 1 4.2460 52.305 -4693.8  
## - Fireplaces 1 5.4330 53.492 -4661.3  
## - sqrt(LotArea) 1 6.7082 54.767 -4627.3  
## - OverallQual 4 13.4175 61.477 -4482.1

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + ExterQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 6.4292204 0.0039786 0.0024676   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0001916 0.0394251 0.1069818   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0042804 0.0005112 0.3931219   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1911459 -0.5163512 0.5814474   
## ExterQualFa ExterQualGd ExterQualTA   
## -0.3314113 -0.0791602 -0.2012559

All the parameters shows that is correct to add the ExterQual, so we keep adding the BsmtQual feature.

m14 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual+ExterQual+BsmtQual, data=df); summary(m14)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + ExterQual + BsmtQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02441 -0.09243 0.00530 0.10366 0.63010   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.789e+00 5.125e-01 15.200 < 2e-16 \*\*\*  
## sqrt(LotArea) 3.960e-03 2.756e-04 14.368 < 2e-16 \*\*\*  
## YearBuilt 1.842e-03 2.558e-04 7.202 9.60e-13 \*\*\*  
## MasVnrArea 1.878e-04 3.702e-05 5.071 4.47e-07 \*\*\*  
## BedroomAbvGr 4.463e-02 6.461e-03 6.908 7.40e-12 \*\*\*  
## Fireplaces 1.048e-01 8.232e-03 12.737 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 3.639e-03 7.110e-04 5.118 3.51e-07 \*\*\*  
## OpenPorchSF 4.253e-04 9.205e-05 4.621 4.17e-06 \*\*\*  
## OverallQualGood 3.542e-01 2.268e-02 15.614 < 2e-16 \*\*\*  
## OverallQualModerate 1.716e-01 1.789e-02 9.592 < 2e-16 \*\*\*  
## OverallQualVBad -5.028e-01 8.695e-02 -5.783 9.02e-09 \*\*\*  
## OverallQualVGood 4.833e-01 4.306e-02 11.225 < 2e-16 \*\*\*  
## ExterQualFa -2.827e-01 6.536e-02 -4.326 1.63e-05 \*\*\*  
## ExterQualGd -4.773e-02 3.724e-02 -1.282 0.2   
## ExterQualTA -1.544e-01 3.935e-02 -3.923 9.17e-05 \*\*\*  
## BsmtQualFa -1.954e-01 4.293e-02 -4.551 5.81e-06 \*\*\*  
## BsmtQualGd -1.158e-01 2.342e-02 -4.943 8.61e-07 \*\*\*  
## BsmtQualNBsmt -3.282e-01 4.034e-02 -8.136 8.83e-16 \*\*\*  
## BsmtQualTA -1.842e-01 2.746e-02 -6.707 2.85e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1791 on 1426 degrees of freedom  
## Multiple R-squared: 0.7867, Adjusted R-squared: 0.784   
## F-statistic: 292.1 on 18 and 1426 DF, p-value: < 2.2e-16

BIC(m14,m13)

## df BIC  
## m14 20 -743.8592  
## m13 16 -700.8136

Anova(m14)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 6.619 1 206.452 < 2.2e-16 \*\*\*  
## YearBuilt 1.663 1 51.863 9.600e-13 \*\*\*  
## MasVnrArea 0.824 1 25.716 4.473e-07 \*\*\*  
## BedroomAbvGr 1.530 1 47.715 7.404e-12 \*\*\*  
## Fireplaces 5.201 1 162.221 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.840 1 26.193 3.510e-07 \*\*\*  
## OpenPorchSF 0.684 1 21.350 4.172e-06 \*\*\*  
## OverallQual 10.095 4 78.716 < 2.2e-16 \*\*\*  
## ExterQual 1.842 3 19.155 3.525e-12 \*\*\*  
## BsmtQual 2.341 4 18.252 1.236e-14 \*\*\*  
## Residuals 45.718 1426   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m14, k = log(nrow(df)))

## Start: AIC=-4851.87  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## ExterQual + BsmtQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 45.718 -4851.9  
## - OpenPorchSF 1 0.6845 46.403 -4837.7  
## - MasVnrArea 1 0.8245 46.543 -4833.3  
## - sqrt(WoodDeckSF) 1 0.8398 46.558 -4832.8  
## - ExterQual 3 1.8424 47.561 -4816.6  
## - BedroomAbvGr 1 1.5298 47.248 -4811.6  
## - BsmtQual 4 2.3407 48.059 -4808.8  
## - YearBuilt 1 1.6628 47.381 -4807.5  
## - Fireplaces 1 5.2009 50.919 -4703.5  
## - sqrt(LotArea) 1 6.6190 52.337 -4663.8  
## - OverallQual 4 10.0947 55.813 -4592.7

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + ExterQual + BsmtQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 7.7893913 0.0039597 0.0018424   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0001878 0.0446301 0.1048475   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0036391 0.0004253 0.3542044   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1715928 -0.5028148 0.4833112   
## ExterQualFa ExterQualGd ExterQualTA   
## -0.2827456 -0.0477254 -0.1543754   
## BsmtQualFa BsmtQualGd BsmtQualNBsmt   
## -0.1953599 -0.1157561 -0.3282078   
## BsmtQualTA   
## -0.1841878

After this, we add the KitcheQual.

m15 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual+ExterQual+BsmtQual+KitchenQual, data=df); summary(m15)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + ExterQual + BsmtQual + KitchenQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02630 -0.09110 0.00329 0.09918 0.63236   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.137e+00 5.067e-01 16.059 < 2e-16 \*\*\*  
## sqrt(LotArea) 3.964e-03 2.697e-04 14.698 < 2e-16 \*\*\*  
## YearBuilt 1.710e-03 2.526e-04 6.770 1.88e-11 \*\*\*  
## MasVnrArea 1.917e-04 3.620e-05 5.297 1.37e-07 \*\*\*  
## BedroomAbvGr 4.665e-02 6.328e-03 7.372 2.84e-13 \*\*\*  
## Fireplaces 1.030e-01 8.062e-03 12.778 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 3.198e-03 6.990e-04 4.575 5.18e-06 \*\*\*  
## OpenPorchSF 3.712e-04 9.031e-05 4.111 4.17e-05 \*\*\*  
## OverallQualGood 3.200e-01 2.260e-02 14.161 < 2e-16 \*\*\*  
## OverallQualModerate 1.596e-01 1.761e-02 9.063 < 2e-16 \*\*\*  
## OverallQualVBad -4.814e-01 8.630e-02 -5.578 2.91e-08 \*\*\*  
## OverallQualVGood 4.121e-01 4.387e-02 9.395 < 2e-16 \*\*\*  
## ExterQualFa -1.880e-01 6.580e-02 -2.858 0.00433 \*\*   
## ExterQualGd -3.182e-02 3.711e-02 -0.857 0.39135   
## ExterQualTA -8.772e-02 3.959e-02 -2.216 0.02688 \*   
## BsmtQualFa -1.858e-01 4.203e-02 -4.419 1.07e-05 \*\*\*  
## BsmtQualGd -1.062e-01 2.319e-02 -4.580 5.05e-06 \*\*\*  
## BsmtQualNBsmt -3.137e-01 3.958e-02 -7.925 4.56e-15 \*\*\*  
## BsmtQualTA -1.723e-01 2.706e-02 -6.368 2.58e-10 \*\*\*  
## KitchenQualFa -2.202e-01 4.122e-02 -5.343 1.07e-07 \*\*\*  
## KitchenQualGd -7.567e-02 2.519e-02 -3.004 0.00271 \*\*   
## KitchenQualTA -1.752e-01 2.738e-02 -6.396 2.16e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1751 on 1423 degrees of freedom  
## Multiple R-squared: 0.7965, Adjusted R-squared: 0.7935   
## F-statistic: 265.2 on 21 and 1423 DF, p-value: < 2.2e-16

BIC(m15,m14)

## df BIC  
## m15 23 -790.2038  
## m14 20 -743.8592

Anova(m15)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 6.621 1 216.0306 < 2.2e-16 \*\*\*  
## YearBuilt 1.405 1 45.8290 1.881e-11 \*\*\*  
## MasVnrArea 0.860 1 28.0542 1.365e-07 \*\*\*  
## BedroomAbvGr 1.666 1 54.3468 2.840e-13 \*\*\*  
## Fireplaces 5.004 1 163.2860 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.641 1 20.9298 5.179e-06 \*\*\*  
## OpenPorchSF 0.518 1 16.8974 4.171e-05 \*\*\*  
## OverallQual 7.733 4 63.0801 < 2.2e-16 \*\*\*  
## ExterQual 0.501 3 5.4497 0.001002 \*\*   
## BsmtQual 2.120 4 17.2925 7.325e-14 \*\*\*  
## KitchenQual 2.107 3 22.9143 1.733e-14 \*\*\*  
## Residuals 43.612 1423   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m15, k = log(nrow(df)))

## Start: AIC=-4898.21  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## ExterQual + BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## - ExterQual 3 0.5011 44.113 -4903.5  
## <none> 43.612 -4898.2  
## - OpenPorchSF 1 0.5179 44.130 -4888.4  
## - sqrt(WoodDeckSF) 1 0.6414 44.253 -4884.4  
## - MasVnrArea 1 0.8598 44.471 -4877.3  
## - YearBuilt 1 1.4046 45.016 -4859.7  
## - BsmtQual 4 2.1199 45.732 -4858.7  
## - KitchenQual 3 2.1068 45.718 -4851.9  
## - BedroomAbvGr 1 1.6656 45.277 -4851.3  
## - Fireplaces 1 5.0043 48.616 -4748.5  
## - sqrt(LotArea) 1 6.6208 50.232 -4701.3  
## - OverallQual 4 7.7330 51.345 -4691.4  
##   
## Step: AIC=-4903.53  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 44.113 -4903.5  
## - OpenPorchSF 1 0.5650 44.678 -4892.4  
## - sqrt(WoodDeckSF) 1 0.6140 44.727 -4890.8  
## - MasVnrArea 1 0.9196 45.032 -4881.0  
## - BedroomAbvGr 1 1.6555 45.768 -4857.6  
## - BsmtQual 4 2.3882 46.501 -4856.5  
## - YearBuilt 1 1.7309 45.844 -4855.2  
## - KitchenQual 3 3.4481 47.561 -4816.6  
## - Fireplaces 1 5.0300 49.143 -4754.8  
## - sqrt(LotArea) 1 6.5636 50.676 -4710.4  
## - OverallQual 4 9.6902 53.803 -4645.7

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + BsmtQual + KitchenQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 7.7754232 0.0039460 0.0018685   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0001978 0.0464577 0.1032420   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0031271 0.0003871 0.3383180   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1615787 -0.5262979 0.4460879   
## BsmtQualFa BsmtQualGd BsmtQualNBsmt   
## -0.1956825 -0.1122630 -0.3274001   
## BsmtQualTA KitchenQualFa KitchenQualGd   
## -0.1842038 -0.2561414 -0.0810783   
## KitchenQualTA   
## -0.2000847

The test shows that ExterQual, after adding the KitchenQual, has lose significance and is suggest to be removed. And comparing the model m15 and m15.1, indeed the AIC get better.

m15.1 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual+BsmtQual+KitchenQual, data=df); summary(m15.1)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + BsmtQual + KitchenQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.01612 -0.09105 0.00407 0.10208 0.62044   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.775e+00 4.998e-01 15.557 < 2e-16 \*\*\*  
## sqrt(LotArea) 3.946e-03 2.709e-04 14.566 < 2e-16 \*\*\*  
## YearBuilt 1.869e-03 2.498e-04 7.480 1.29e-13 \*\*\*  
## MasVnrArea 1.978e-04 3.628e-05 5.452 5.85e-08 \*\*\*  
## BedroomAbvGr 4.646e-02 6.351e-03 7.315 4.27e-13 \*\*\*  
## Fireplaces 1.032e-01 8.096e-03 12.751 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 3.127e-03 7.019e-04 4.455 9.03e-06 \*\*\*  
## OpenPorchSF 3.871e-04 9.058e-05 4.274 2.05e-05 \*\*\*  
## OverallQualGood 3.383e-01 2.216e-02 15.268 < 2e-16 \*\*\*  
## OverallQualModerate 1.616e-01 1.762e-02 9.171 < 2e-16 \*\*\*  
## OverallQualVBad -5.263e-01 8.347e-02 -6.305 3.83e-10 \*\*\*  
## OverallQualVGood 4.461e-01 4.092e-02 10.902 < 2e-16 \*\*\*  
## BsmtQualFa -1.957e-01 4.202e-02 -4.656 3.52e-06 \*\*\*  
## BsmtQualGd -1.123e-01 2.299e-02 -4.884 1.16e-06 \*\*\*  
## BsmtQualNBsmt -3.274e-01 3.952e-02 -8.285 2.69e-16 \*\*\*  
## BsmtQualTA -1.842e-01 2.685e-02 -6.860 1.02e-11 \*\*\*  
## KitchenQualFa -2.561e-01 4.015e-02 -6.380 2.39e-10 \*\*\*  
## KitchenQualGd -8.108e-02 2.484e-02 -3.264 0.00112 \*\*   
## KitchenQualTA -2.001e-01 2.650e-02 -7.551 7.67e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1759 on 1426 degrees of freedom  
## Multiple R-squared: 0.7942, Adjusted R-squared: 0.7916   
## F-statistic: 305.6 on 18 and 1426 DF, p-value: < 2.2e-16

BIC(m15.1,m15)

## df BIC  
## m15.1 20 -795.5242  
## m15 23 -790.2038

Anova(m15.1)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 6.564 1 212.176 < 2.2e-16 \*\*\*  
## YearBuilt 1.731 1 55.953 1.292e-13 \*\*\*  
## MasVnrArea 0.920 1 29.727 5.853e-08 \*\*\*  
## BedroomAbvGr 1.655 1 53.515 4.265e-13 \*\*\*  
## Fireplaces 5.030 1 162.600 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.614 1 19.850 9.031e-06 \*\*\*  
## OpenPorchSF 0.565 1 18.264 2.051e-05 \*\*\*  
## OverallQual 9.690 4 78.312 < 2.2e-16 \*\*\*  
## BsmtQual 2.388 4 19.300 1.776e-15 \*\*\*  
## KitchenQual 3.448 3 37.155 < 2.2e-16 \*\*\*  
## Residuals 44.113 1426   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m15.1, k = log(nrow(df)))

## Start: AIC=-4903.53  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 44.113 -4903.5  
## - OpenPorchSF 1 0.5650 44.678 -4892.4  
## - sqrt(WoodDeckSF) 1 0.6140 44.727 -4890.8  
## - MasVnrArea 1 0.9196 45.032 -4881.0  
## - BedroomAbvGr 1 1.6555 45.768 -4857.6  
## - BsmtQual 4 2.3882 46.501 -4856.5  
## - YearBuilt 1 1.7309 45.844 -4855.2  
## - KitchenQual 3 3.4481 47.561 -4816.6  
## - Fireplaces 1 5.0300 49.143 -4754.8  
## - sqrt(LotArea) 1 6.5636 50.676 -4710.4  
## - OverallQual 4 9.6902 53.803 -4645.7

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + BsmtQual + KitchenQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 7.7754232 0.0039460 0.0018685   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0001978 0.0464577 0.1032420   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0031271 0.0003871 0.3383180   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1615787 -0.5262979 0.4460879   
## BsmtQualFa BsmtQualGd BsmtQualNBsmt   
## -0.1956825 -0.1122630 -0.3274001   
## BsmtQualTA KitchenQualFa KitchenQualGd   
## -0.1842038 -0.2561414 -0.0810783   
## KitchenQualTA   
## -0.2000847

Adding the Neighbourhood and GarageFinish to the model.

m16 = lm(log(SalePrice)~sqrt(LotArea)+YearBuilt+MasVnrArea+BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+OverallQual+BsmtQual+KitchenQual+Neighborhood+GarageFinish, data=df); summary(m16)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + BsmtQual + KitchenQual + Neighborhood + GarageFinish,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.94507 -0.08482 0.00502 0.09269 0.55190   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.0726845 0.5108987 17.758 < 2e-16 \*\*\*  
## sqrt(LotArea) 0.0035680 0.0002559 13.941 < 2e-16 \*\*\*  
## YearBuilt 0.0012324 0.0002529 4.873 1.22e-06 \*\*\*  
## MasVnrArea 0.0001410 0.0000342 4.122 3.98e-05 \*\*\*  
## BedroomAbvGr 0.0538917 0.0059628 9.038 < 2e-16 \*\*\*  
## Fireplaces 0.0805645 0.0077364 10.414 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.0031694 0.0006553 4.837 1.46e-06 \*\*\*  
## OpenPorchSF 0.0003211 0.0000847 3.791 0.000156 \*\*\*  
## OverallQualGood 0.2763502 0.0211637 13.058 < 2e-16 \*\*\*  
## OverallQualModerate 0.1367402 0.0166894 8.193 5.61e-16 \*\*\*  
## OverallQualVBad -0.4854769 0.0779016 -6.232 6.06e-10 \*\*\*  
## OverallQualVGood 0.3797439 0.0384279 9.882 < 2e-16 \*\*\*  
## BsmtQualFa -0.1596736 0.0395162 -4.041 5.61e-05 \*\*\*  
## BsmtQualGd -0.0830346 0.0215893 -3.846 0.000125 \*\*\*  
## BsmtQualNBsmt -0.2611066 0.0372542 -7.009 3.70e-12 \*\*\*  
## BsmtQualTA -0.1305402 0.0255610 -5.107 3.72e-07 \*\*\*  
## KitchenQualFa -0.2111850 0.0376490 -5.609 2.44e-08 \*\*\*  
## KitchenQualGd -0.0727347 0.0232570 -3.127 0.001799 \*\*   
## KitchenQualTA -0.1692437 0.0248729 -6.804 1.49e-11 \*\*\*  
## NeighborhoodPoor -0.0400814 0.0127404 -3.146 0.001689 \*\*   
## NeighborhoodRich 0.1339629 0.0131859 10.160 < 2e-16 \*\*\*  
## GarageFinishNGar -0.1900531 0.0241084 -7.883 6.30e-15 \*\*\*  
## GarageFinishRFn -0.0174421 0.0125007 -1.395 0.163145   
## GarageFinishUnf -0.0653116 0.0142690 -4.577 5.12e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1637 on 1421 degrees of freedom  
## Multiple R-squared: 0.8224, Adjusted R-squared: 0.8195   
## F-statistic: 286 on 23 and 1421 DF, p-value: < 2.2e-16

BIC(m16,m15.1)

## df BIC  
## m16 25 -972.2267  
## m15.1 20 -795.5242

Anova(m16)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 5.206 1 194.347 < 2.2e-16 \*\*\*  
## YearBuilt 0.636 1 23.747 1.222e-06 \*\*\*  
## MasVnrArea 0.455 1 16.990 3.975e-05 \*\*\*  
## BedroomAbvGr 2.188 1 81.685 < 2.2e-16 \*\*\*  
## Fireplaces 2.905 1 108.445 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.627 1 23.393 1.464e-06 \*\*\*  
## OpenPorchSF 0.385 1 14.371 0.0001564 \*\*\*  
## OverallQual 6.406 4 59.783 < 2.2e-16 \*\*\*  
## BsmtQual 1.378 4 12.859 2.757e-10 \*\*\*  
## KitchenQual 2.270 3 28.247 < 2.2e-16 \*\*\*  
## Neighborhood 4.251 2 79.345 < 2.2e-16 \*\*\*  
## GarageFinish 1.751 3 21.791 8.469e-14 \*\*\*  
## Residuals 38.065 1421   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m16, k = log(nrow(df)))

## Start: AIC=-5080.23  
## log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + OverallQual +   
## BsmtQual + KitchenQual + Neighborhood + GarageFinish  
##   
## Df Sum of Sq RSS AIC  
## <none> 38.065 -5080.2  
## - OpenPorchSF 1 0.3850 38.450 -5073.0  
## - MasVnrArea 1 0.4551 38.520 -5070.3  
## - sqrt(WoodDeckSF) 1 0.6266 38.691 -5063.9  
## - YearBuilt 1 0.6361 38.701 -5063.6  
## - BsmtQual 4 1.3778 39.442 -5058.0  
## - GarageFinish 3 1.7512 39.816 -5037.1  
## - KitchenQual 3 2.2700 40.335 -5018.4  
## - BedroomAbvGr 1 2.1881 40.253 -5006.7  
## - Fireplaces 1 2.9049 40.970 -4981.2  
## - Neighborhood 2 4.2509 42.316 -4941.8  
## - sqrt(LotArea) 1 5.2060 43.271 -4902.3  
## - OverallQual 4 6.4057 44.470 -4884.6

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF +   
## OverallQual + BsmtQual + KitchenQual + Neighborhood + GarageFinish,   
## data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea) YearBuilt   
## 9.0726845 0.0035680 0.0012324   
## MasVnrArea BedroomAbvGr Fireplaces   
## 0.0001410 0.0538917 0.0805645   
## sqrt(WoodDeckSF) OpenPorchSF OverallQualGood   
## 0.0031694 0.0003211 0.2763502   
## OverallQualModerate OverallQualVBad OverallQualVGood   
## 0.1367402 -0.4854769 0.3797439   
## BsmtQualFa BsmtQualGd BsmtQualNBsmt   
## -0.1596736 -0.0830346 -0.2611066   
## BsmtQualTA KitchenQualFa KitchenQualGd   
## -0.1305402 -0.2111849 -0.0727347   
## KitchenQualTA NeighborhoodPoor NeighborhoodRich   
## -0.1692437 -0.0400814 0.1339629   
## GarageFinishNGar GarageFinishRFn GarageFinishUnf   
## -0.1900531 -0.0174421 -0.0653116

df BIC

m16 22 -929.0607 m15.1 20 -795.5242

# 11. Checking the possible Interactions

1. YearBuilt and OverallQual intuitively should interact because of inflation. Indeed, all variables could interact with YearBuilt, but OverallQual summarizes them all.

m17 = lm(log(SalePrice)~sqrt(LotArea)+MasVnrArea+  
 BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+YearBuilt\*OverallQual+BsmtQual+KitchenQual, data=df); summary(m17)

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + YearBuilt \*   
## OverallQual + BsmtQual + KitchenQual, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98975 -0.09217 0.00322 0.09815 0.63437   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.465e+00 1.293e+00 2.680 0.007441 \*\*   
## sqrt(LotArea) 3.949e-03 2.686e-04 14.705 < 2e-16 \*\*\*  
## MasVnrArea 2.081e-04 3.601e-05 5.778 9.27e-09 \*\*\*  
## BedroomAbvGr 4.654e-02 6.295e-03 7.392 2.45e-13 \*\*\*  
## Fireplaces 1.003e-01 8.116e-03 12.354 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 3.069e-03 6.957e-04 4.412 1.10e-05 \*\*\*  
## OpenPorchSF 3.995e-04 8.988e-05 4.445 9.46e-06 \*\*\*  
## YearBuilt 4.096e-03 6.623e-04 6.185 8.12e-10 \*\*\*  
## OverallQualGood 5.871e+00 1.445e+00 4.062 5.12e-05 \*\*\*  
## OverallQualModerate 4.337e+00 1.360e+00 3.189 0.001460 \*\*   
## OverallQualVBad 3.668e+01 1.281e+01 2.863 0.004263 \*\*   
## OverallQualVGood 1.443e+01 3.337e+00 4.324 1.64e-05 \*\*\*  
## BsmtQualFa -1.996e-01 4.172e-02 -4.784 1.90e-06 \*\*\*  
## BsmtQualGd -1.235e-01 2.299e-02 -5.371 9.11e-08 \*\*\*  
## BsmtQualNBsmt -3.381e-01 3.937e-02 -8.588 < 2e-16 \*\*\*  
## BsmtQualTA -1.993e-01 2.682e-02 -7.434 1.81e-13 \*\*\*  
## KitchenQualFa -2.663e-01 3.997e-02 -6.661 3.87e-11 \*\*\*  
## KitchenQualGd -9.059e-02 2.485e-02 -3.645 0.000277 \*\*\*  
## KitchenQualTA -2.118e-01 2.643e-02 -8.012 2.33e-15 \*\*\*  
## YearBuilt:OverallQualGood -2.832e-03 7.395e-04 -3.830 0.000134 \*\*\*  
## YearBuilt:OverallQualModerate -2.146e-03 6.983e-04 -3.074 0.002155 \*\*   
## YearBuilt:OverallQualVBad -1.922e-02 6.623e-03 -2.902 0.003759 \*\*   
## YearBuilt:OverallQualVGood -7.057e-03 1.675e-03 -4.212 2.69e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1742 on 1422 degrees of freedom  
## Multiple R-squared: 0.7986, Adjusted R-squared: 0.7955   
## F-statistic: 256.3 on 22 and 1422 DF, p-value: < 2.2e-16

BIC(m17,m16,m15.1,m15,m14,m13,m12.1,m12,m11)

## df BIC  
## m17 24 -797.99229  
## m16 25 -972.22665  
## m15.1 20 -795.52420  
## m15 23 -790.20379  
## m14 20 -743.85924  
## m13 16 -700.81362  
## m12.1 13 -648.68143  
## m12 14 -641.83724  
## m11 10 -78.08351

Anova(m17)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## sqrt(LotArea) 6.563 1 216.2291 < 2.2e-16 \*\*\*  
## MasVnrArea 1.013 1 33.3855 9.273e-09 \*\*\*  
## BedroomAbvGr 1.659 1 54.6477 2.451e-13 \*\*\*  
## Fireplaces 4.632 1 152.6255 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.591 1 19.4680 1.100e-05 \*\*\*  
## OpenPorchSF 0.600 1 19.7598 9.463e-06 \*\*\*  
## YearBuilt 1.731 1 57.0285 7.644e-14 \*\*\*  
## OverallQual 9.690 4 79.8171 < 2.2e-16 \*\*\*  
## BsmtQual 2.566 4 21.1347 < 2.2e-16 \*\*\*  
## KitchenQual 3.636 3 39.9369 < 2.2e-16 \*\*\*  
## YearBuilt:OverallQual 0.953 4 7.8527 2.934e-06 \*\*\*  
## Residuals 43.159 1422   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

step(m17, k = log(nrow(df)))

## Start: AIC=-4906  
## log(SalePrice) ~ sqrt(LotArea) + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + YearBuilt \*   
## OverallQual + BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 43.159 -4906.0  
## - YearBuilt:OverallQual 4 0.9534 44.113 -4903.5  
## - sqrt(WoodDeckSF) 1 0.5909 43.750 -4893.6  
## - OpenPorchSF 1 0.5997 43.759 -4893.3  
## - MasVnrArea 1 1.0133 44.173 -4879.7  
## - BedroomAbvGr 1 1.6586 44.818 -4858.8  
## - BsmtQual 4 2.5658 45.725 -4851.7  
## - KitchenQual 3 3.6364 46.796 -4810.9  
## - Fireplaces 1 4.6324 47.792 -4766.0  
## - sqrt(LotArea) 1 6.5628 49.722 -4708.7

##   
## Call:  
## lm(formula = log(SalePrice) ~ sqrt(LotArea) + MasVnrArea + BedroomAbvGr +   
## Fireplaces + sqrt(WoodDeckSF) + OpenPorchSF + YearBuilt \*   
## OverallQual + BsmtQual + KitchenQual, data = df)  
##   
## Coefficients:  
## (Intercept) sqrt(LotArea)   
## 3.4648765 0.0039493   
## MasVnrArea BedroomAbvGr   
## 0.0002081 0.0465388   
## Fireplaces sqrt(WoodDeckSF)   
## 0.1002693 0.0030694   
## OpenPorchSF YearBuilt   
## 0.0003995 0.0040961   
## OverallQualGood OverallQualModerate   
## 5.8714822 4.3372588   
## OverallQualVBad OverallQualVGood   
## 36.6841275 14.4323011   
## BsmtQualFa BsmtQualGd   
## -0.1995871 -0.1234950   
## BsmtQualNBsmt BsmtQualTA   
## -0.3381408 -0.1993441   
## KitchenQualFa KitchenQualGd   
## -0.2662522 -0.0905916   
## KitchenQualTA YearBuilt:OverallQualGood   
## -0.2117534 -0.0028318   
## YearBuilt:OverallQualModerate YearBuilt:OverallQualVBad   
## -0.0021463 -0.0192219   
## YearBuilt:OverallQualVGood   
## -0.0070565

1. LotArea and YearBuilt should interact as well because of inflation.

m18 = lm(log(SalePrice)~MasVnrArea+  
 BedroomAbvGr+Fireplaces+sqrt(WoodDeckSF)+OpenPorchSF+YearBuilt\*OverallQual+sqrt(LotArea)\*YearBuilt+OverallQual+BsmtQual+KitchenQual, data=df); summary(m18)

##   
## Call:  
## lm(formula = log(SalePrice) ~ MasVnrArea + BedroomAbvGr + Fireplaces +   
## sqrt(WoodDeckSF) + OpenPorchSF + YearBuilt \* OverallQual +   
## sqrt(LotArea) \* YearBuilt + OverallQual + BsmtQual + KitchenQual,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98632 -0.09213 0.00553 0.09971 0.63145   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.134e+00 2.081e+00 -1.506 0.132318   
## MasVnrArea 2.081e-04 3.582e-05 5.810 7.68e-09 \*\*\*  
## BedroomAbvGr 4.916e-02 6.296e-03 7.808 1.12e-14 \*\*\*  
## Fireplaces 9.869e-02 8.083e-03 12.210 < 2e-16 \*\*\*  
## sqrt(WoodDeckSF) 3.083e-03 6.920e-04 4.456 9.01e-06 \*\*\*  
## OpenPorchSF 4.049e-04 8.941e-05 4.529 6.43e-06 \*\*\*  
## YearBuilt 7.441e-03 1.059e-03 7.024 3.32e-12 \*\*\*  
## OverallQualGood 5.507e+00 1.440e+00 3.823 0.000138 \*\*\*  
## OverallQualModerate 4.338e+00 1.353e+00 3.207 0.001372 \*\*   
## OverallQualVBad 3.520e+01 1.275e+01 2.760 0.005854 \*\*   
## OverallQualVGood 1.302e+01 3.338e+00 3.901 0.000100 \*\*\*  
## sqrt(LotArea) 7.508e-02 1.764e-02 4.255 2.22e-05 \*\*\*  
## BsmtQualFa -1.985e-01 4.150e-02 -4.783 1.91e-06 \*\*\*  
## BsmtQualGd -1.294e-01 2.292e-02 -5.649 1.95e-08 \*\*\*  
## BsmtQualNBsmt -3.462e-01 3.921e-02 -8.828 < 2e-16 \*\*\*  
## BsmtQualTA -2.033e-01 2.669e-02 -7.615 4.78e-14 \*\*\*  
## KitchenQualFa -2.670e-01 3.976e-02 -6.717 2.68e-11 \*\*\*  
## KitchenQualGd -9.585e-02 2.476e-02 -3.872 0.000113 \*\*\*  
## KitchenQualTA -2.153e-01 2.630e-02 -8.186 5.94e-16 \*\*\*  
## YearBuilt:OverallQualGood -2.648e-03 7.369e-04 -3.594 0.000337 \*\*\*  
## YearBuilt:OverallQualModerate -2.149e-03 6.946e-04 -3.094 0.002017 \*\*   
## YearBuilt:OverallQualVBad -1.845e-02 6.590e-03 -2.800 0.005185 \*\*   
## YearBuilt:OverallQualVGood -6.347e-03 1.676e-03 -3.788 0.000158 \*\*\*  
## YearBuilt:sqrt(LotArea) -3.603e-05 8.936e-06 -4.032 5.82e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1733 on 1421 degrees of freedom  
## Multiple R-squared: 0.8009, Adjusted R-squared: 0.7977   
## F-statistic: 248.5 on 23 and 1421 DF, p-value: < 2.2e-16

BIC(m18,m17,m16,m15.1,m15,m14,m13,m12.1,m12,m11)

## df BIC  
## m18 25 -807.15432  
## m17 24 -797.99229  
## m16 25 -972.22665  
## m15.1 20 -795.52420  
## m15 23 -790.20379  
## m14 20 -743.85924  
## m13 16 -700.81362  
## m12.1 13 -648.68143  
## m12 14 -641.83724  
## m11 10 -78.08351

Anova(m18)

## Anova Table (Type II tests)  
##   
## Response: log(SalePrice)  
## Sum Sq Df F value Pr(>F)   
## MasVnrArea 1.014 1 33.7613 7.679e-09 \*\*\*  
## BedroomAbvGr 1.831 1 60.9645 1.121e-14 \*\*\*  
## Fireplaces 4.477 1 149.0919 < 2.2e-16 \*\*\*  
## sqrt(WoodDeckSF) 0.596 1 19.8549 9.010e-06 \*\*\*  
## OpenPorchSF 0.616 1 20.5084 6.434e-06 \*\*\*  
## YearBuilt 1.731 1 57.6404 5.671e-14 \*\*\*  
## OverallQual 9.626 4 80.1403 < 2.2e-16 \*\*\*  
## sqrt(LotArea) 6.563 1 218.5491 < 2.2e-16 \*\*\*  
## BsmtQual 2.649 4 22.0501 < 2.2e-16 \*\*\*  
## KitchenQual 3.607 3 40.0345 < 2.2e-16 \*\*\*  
## YearBuilt:OverallQual 0.797 4 6.6333 2.744e-05 \*\*\*  
## YearBuilt:sqrt(LotArea) 0.488 1 16.2572 5.824e-05 \*\*\*  
## Residuals 42.671 1421   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

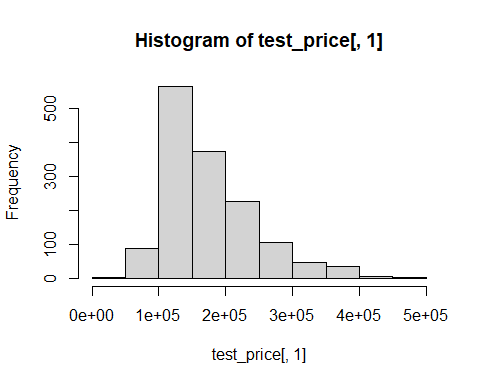
step(m18, k = log(nrow(df)))

## Start: AIC=-4915.16  
## log(SalePrice) ~ MasVnrArea + BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) +   
## OpenPorchSF + YearBuilt \* OverallQual + sqrt(LotArea) \* YearBuilt +   
## OverallQual + BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## - YearBuilt:OverallQual 4 0.7968 43.468 -4917.5  
## <none> 42.671 -4915.2  
## - YearBuilt:sqrt(LotArea) 1 0.4882 43.159 -4906.0  
## - sqrt(WoodDeckSF) 1 0.5962 43.267 -4902.4  
## - OpenPorchSF 1 0.6158 43.287 -4901.7  
## - MasVnrArea 1 1.0138 43.685 -4888.5  
## - BedroomAbvGr 1 1.8307 44.502 -4861.7  
## - BsmtQual 4 2.6486 45.320 -4857.2  
## - KitchenQual 3 3.6066 46.278 -4819.7  
## - Fireplaces 1 4.4771 47.148 -4778.3  
##   
## Step: AIC=-4917.53  
## log(SalePrice) ~ MasVnrArea + BedroomAbvGr + Fireplaces + sqrt(WoodDeckSF) +   
## OpenPorchSF + YearBuilt + OverallQual + sqrt(LotArea) + BsmtQual +   
## KitchenQual + YearBuilt:sqrt(LotArea)  
##   
## Df Sum of Sq RSS AIC  
## <none> 43.468 -4917.5  
## - OpenPorchSF 1 0.5848 44.053 -4905.5  
## - sqrt(WoodDeckSF) 1 0.6212 44.089 -4904.3  
## - YearBuilt:sqrt(LotArea) 1 0.6448 44.113 -4903.5  
## - MasVnrArea 1 0.9287 44.397 -4894.3  
## - BsmtQual 4 2.5049 45.973 -4865.7  
## - BedroomAbvGr 1 1.8563 45.324 -4864.4  
## - KitchenQual 3 3.4391 46.907 -4829.3  
## - Fireplaces 1 4.7754 48.243 -4774.2  
## - OverallQual 4 9.6261 53.094 -4657.6

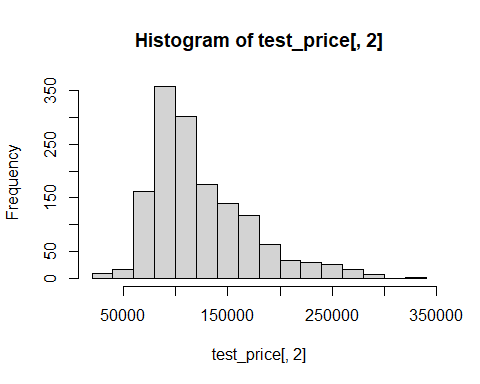
##   
## Call:  
## lm(formula = log(SalePrice) ~ MasVnrArea + BedroomAbvGr + Fireplaces +   
## sqrt(WoodDeckSF) + OpenPorchSF + YearBuilt + OverallQual +   
## sqrt(LotArea) + BsmtQual + KitchenQual + YearBuilt:sqrt(LotArea),   
## data = df)  
##   
## Coefficients:  
## (Intercept) MasVnrArea BedroomAbvGr   
## 1.331e-01 1.988e-04 4.946e-02   
## Fireplaces sqrt(WoodDeckSF) OpenPorchSF   
## 1.008e-01 3.145e-03 3.939e-04   
## YearBuilt OverallQualGood OverallQualModerate   
## 5.745e-03 3.350e-01 1.567e-01   
## OverallQualVBad OverallQualVGood sqrt(LotArea)   
## -5.209e-01 4.507e-01 8.472e-02   
## BsmtQualFa BsmtQualGd BsmtQualNBsmt   
## -1.951e-01 -1.212e-01 -3.385e-01   
## BsmtQualTA KitchenQualFa KitchenQualGd   
## -1.908e-01 -2.576e-01 -8.870e-02   
## KitchenQualTA YearBuilt:sqrt(LotArea)   
## -2.061e-01 -4.091e-05

Any of these interactions have improved much the model, so we won’t keep any. No other interaction would make sense, so we won’t try anymore.

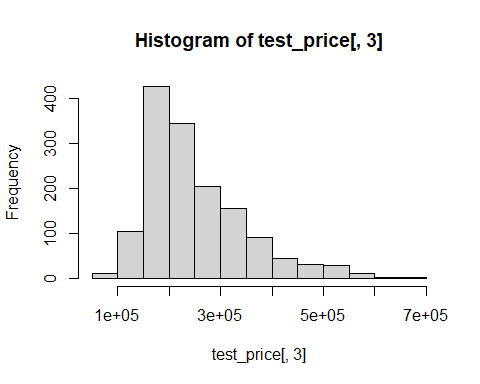
predicted\_values = predict.lm(m15.1, df\_test, se.fit=TRUE, interval="prediction", level=0.95)  
  
test\_price = exp(predicted\_values$fit)  
  
hist(test\_price[,1])



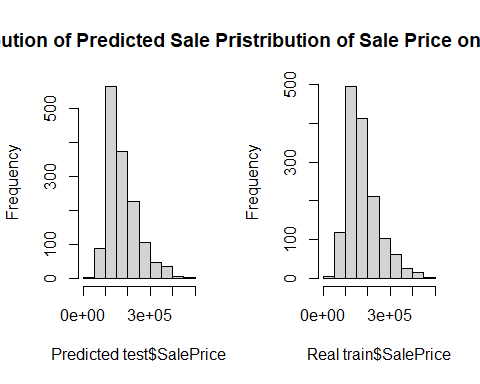
hist(test\_price[,2])



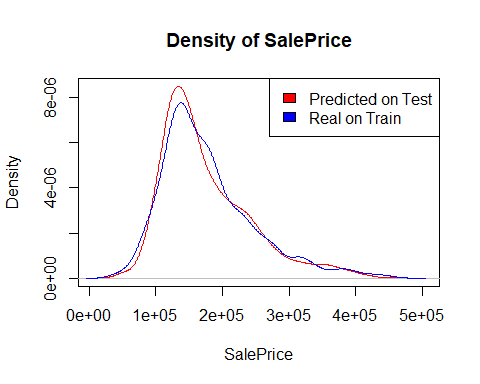
hist(test\_price[,3])



par(mfrow=c(1,2))  
hist(test\_price[,1], main = "Distribution of Predicted Sale Price on Test", xlab = "Predicted test$SalePrice")  
hist(df$SalePrice, main = "Distribution of Sale Price on Train", xlab = "Real train$SalePrice")

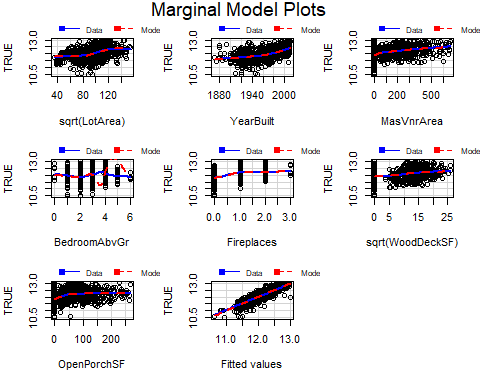


par(mfrow=c(1,1))  
plot(density(test\_price[,1]), col="red", main = "Density of SalePrice", xlab = "SalePrice")  
lines(density(df$SalePrice), col="blue")  
legend("topright",fill = c("red", "blue"), c("Predicted on Test","Real on Train"))

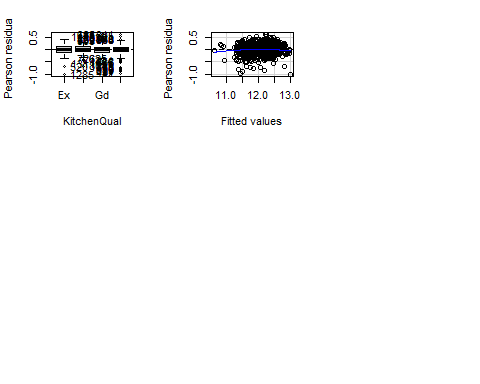
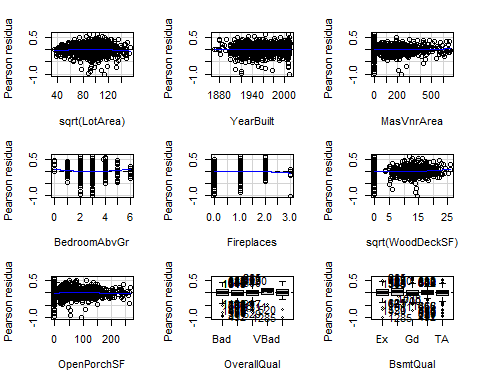


marginalModelPlots(m15.1, id=list(n=0))

## Warning in mmps(...): Interactions and/or factors skipped



residualPlots( m15.1, id=list(n=0))



## Test stat Pr(>|Test stat|)   
## sqrt(LotArea) -2.2578 0.0241120 \*   
## YearBuilt 0.9240 0.3556360   
## MasVnrArea 1.2181 0.2233933   
## BedroomAbvGr 2.7063 0.0068852 \*\*   
## Fireplaces -1.8111 0.0703410 .   
## sqrt(WoodDeckSF) 2.5817 0.0099297 \*\*   
## OpenPorchSF -0.8800 0.3790024   
## OverallQual   
## BsmtQual   
## KitchenQual   
## Tukey test -3.7167 0.0002018 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

In general, using the marginal model plots, we can see that the residuals distribution for most variables are close to 0. However, sqrt(LotArea) seems to have bad residuals in marginalModelPlots(), but not in residualPlots. This could simply mean the first plot doesn’t properly represent the residuals of this variable. As for categorical variables, all errors are close to 0, except for the level “VBad” of OverallQual, which is due to the fact that it contains few individuals.

ks\_test\_result <- ks.test(test\_price[,1], df$SalePrice)

## Warning in ks.test.default(test\_price[, 1], df$SalePrice): 并列的时候P-值将近似

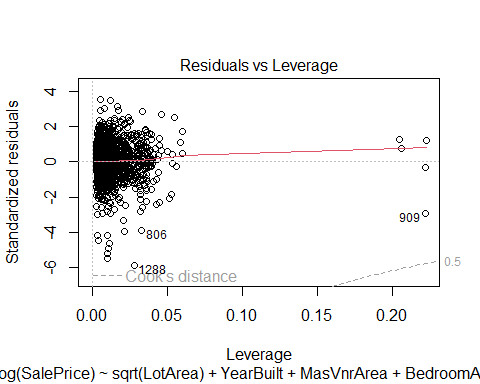
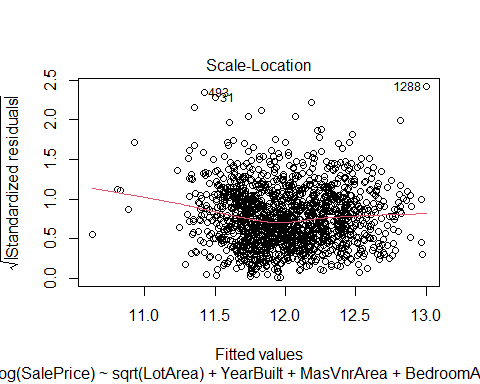
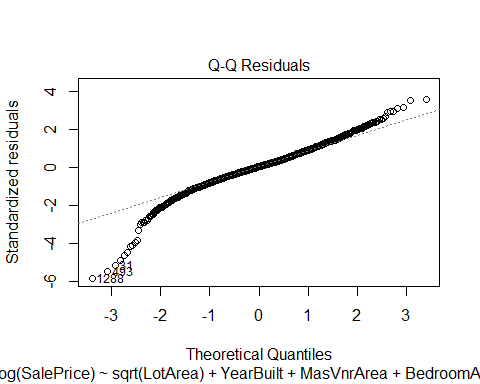
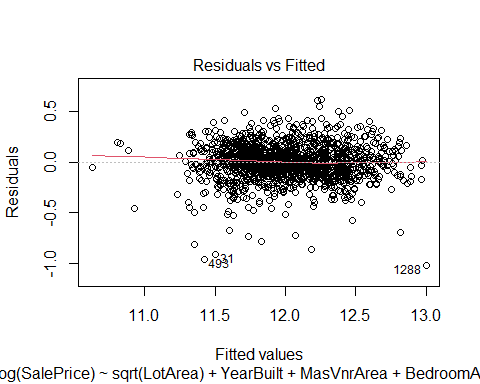
ks\_test\_result

##   
## Asymptotic two-sample Kolmogorov-Smirnov test  
##   
## data: test\_price[, 1] and df$SalePrice  
## D = 0.04363, p-value = 0.1261  
## alternative hypothesis: two-sided

The Kolmogorov-Smirnov test shows that predicted and real distributions of SalePrice should be accepted to be different.

Finally, we will check the normality of the residuals.

plot(m15.1)



shapiro.test(m15.1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: m15.1$residuals  
## W = 0.96191, p-value < 2.2e-16

Residuals don’t follow a normal distribution, so the model won’t give very accurate results. Nevertheless, we are happy with our resulting model, so we will not do any more changes.

# 12. Model selection

From the resulting model, we tried to reduce the number of variables using “step” method with BIC (Bayesian information criteria) to consider a greater effect of the complexity on the model considering that there are a lot of variables. Additionally we performed an “Anova” method, that compares a base model with a nested model of one less variable.

Both methods suggested that the best possible model is obtained without removing any variable. For example, in AIC, removing “BedroomAbvGr” makes a simpler model but it is less predictive. As a consequence BIC is increased from 31053 to 31056. In “Anova” the p-value for this value is not significant and, as a consequence, we cannot remove this variable as the resulting model is not equivalent.

# we use k = log(n) to use BIC instead of AIC   
step(m2, k = log(nrow(df)))

## Start: AIC=30990.46  
## SalePrice ~ LotFrontage + LotArea + YearBuilt + MasVnrArea +   
## BedroomAbvGr + Fireplaces + WoodDeckSF + OpenPorchSF  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.8468e+12 30991  
## - BedroomAbvGr 1 2.0454e+10 2.8673e+12 30994  
## - LotFrontage 1 7.6210e+10 2.9231e+12 31021  
## - WoodDeckSF 1 1.0001e+11 2.9469e+12 31033  
## - OpenPorchSF 1 1.4655e+11 2.9934e+12 31056  
## - MasVnrArea 1 1.5969e+11 3.0065e+12 31062  
## - LotArea 1 2.0346e+11 3.0503e+12 31083  
## - Fireplaces 1 3.8922e+11 3.2361e+12 31168  
## - YearBuilt 1 8.7282e+11 3.7197e+12 31370

##   
## Call:  
## lm(formula = SalePrice ~ LotFrontage + LotArea + YearBuilt +   
## MasVnrArea + BedroomAbvGr + Fireplaces + WoodDeckSF + OpenPorchSF,   
## data = df\_num2)  
##   
## Coefficients:  
## (Intercept) LotFrontage LotArea YearBuilt MasVnrArea   
## -1.718e+06 2.279e+02 3.814e+00 9.096e+02 8.051e+01   
## BedroomAbvGr Fireplaces WoodDeckSF OpenPorchSF   
## 4.973e+03 2.800e+04 7.272e+01 1.914e+02

Anova(m0)

## Anova Table (Type II tests)  
##   
## Response: SalePrice  
## Sum Sq Df F value Pr(>F)   
## LotFrontage 1.0854e+10 1 12.1118 0.0005161 \*\*\*  
## LotArea 2.7910e+10 1 31.1439 2.864e-08 \*\*\*  
## YearBuilt 6.7049e+10 1 74.8180 < 2.2e-16 \*\*\*  
## YearRemodAdd 9.8216e+10 1 109.5967 < 2.2e-16 \*\*\*  
## MasVnrArea 1.5511e+10 1 17.3081 3.368e-05 \*\*\*  
## BsmtFinSF1 1.4861e+10 1 16.5829 4.913e-05 \*\*\*  
## BsmtUnfSF 2.5433e+08 1 0.2838 0.5943018   
## TotalBsmtSF 2.6212e+10 1 29.2499 7.452e-08 \*\*\*  
## X1stFlrSF 7.8706e+09 1 8.7826 0.0030916 \*\*   
## X2ndFlrSF 3.9030e+09 1 4.3553 0.0370711 \*   
## GrLivArea 5.7471e+10 1 64.1307 2.398e-15 \*\*\*  
## BsmtFullBath 5.9596e+07 1 0.0665 0.7965373   
## FullBath 1.6829e+09 1 1.8779 0.1707884   
## HalfBath 2.0928e+09 1 2.3353 0.1266930   
## BedroomAbvGr 4.3882e+10 1 48.9674 3.993e-12 \*\*\*  
## TotRmsAbvGrd 1.6577e+08 1 0.1850 0.6671900   
## Fireplaces 3.0526e+10 1 34.0634 6.595e-09 \*\*\*  
## GarageYrBlt 1.9543e+09 1 2.1808 0.1399622   
## GarageCars 6.1012e+09 1 6.8082 0.0091690 \*\*   
## GarageArea 8.6377e+09 1 9.6387 0.0019428 \*\*   
## WoodDeckSF 9.7220e+09 1 10.8486 0.0010130 \*\*   
## OpenPorchSF 8.1156e+09 1 9.0560 0.0026643 \*\*   
## Residuals 1.2770e+12 1425   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1