Real estate selling price prediction

28/04/2020

Background

The task for this research project is to develop a model to predict the selling price of a given home in Ames, Iowa. Real estate investors may use this information to assess whether the asking price of a house is higher or lower than the true value of the house. If the home is undervalued, it may be a good investment for the firm.

Training Data and relevant packages

In order to better assess the quality of the model, the data have been randomly divided into three separate pieces: a training data set, a testing data set, and a validation data set. For now we will load the training data set, the others will be loaded and used later.

```
load("ames_train.Rdata")
```

Loading packages

```
library(statsr)
library(dplyr)
library(ggplot2)
library(BAS)
library(MASS)
library(e1071)
library(corrplot)
library(forcats)
```

Part 1 - Exploratory Data Analysis (EDA)

In the EDA section, we will try to better understand the data structure as well as detect any patterns and relationships.

Before creating graphs we will need to filter by normal sales conditions as the test data only include these observations.

```
# filtering by "Normal Sales Condition"
ames_train <- ames_train %>%
  filter(Sale.Condition == "Normal")
```

After examining the variables provided in the data set we have chosen the following plots.

Figure 1. `Area` variable distribution

```
# library(cowplot)
# theme_set(theme_cowplot())

# checking skewness
skewness(log(ames_train$area))
```

```
## [1] -0.05410451
```

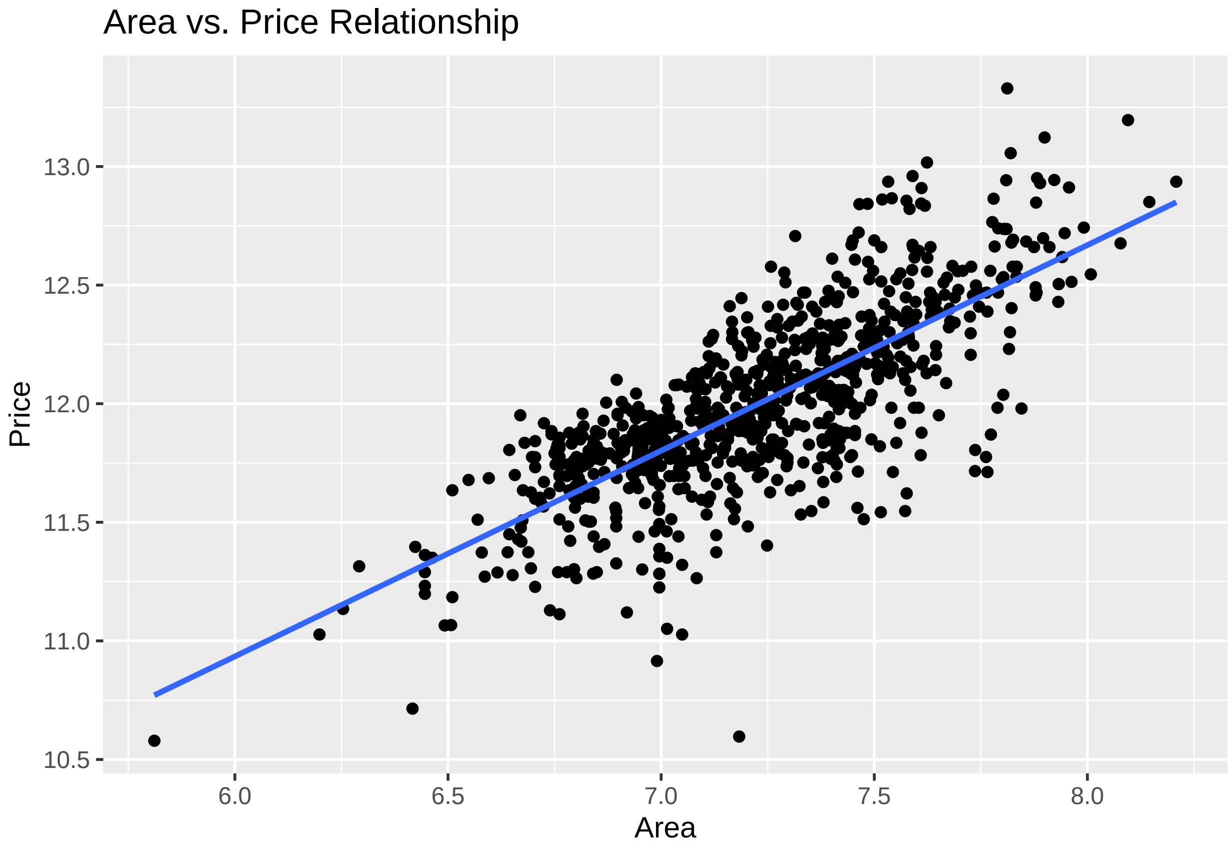
```
# checking collinearity
cor(log(ames_train$price), log(ames_train$area))
```

```
## [1] 0.7579522
```

```
par(mfrow=c(1,2))

# log-transformation
hist(log(ames_train$area),
     main="log(‘area‘) distribution",
     xlab="log(‘area‘)",
     ylab="Frequency",
     col="#F8766D",
     breaks = 30)

# checking linear relationship
ggplot(data = ames_train, aes(x = log(area), y = log(price))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Area vs. Price Relationship", x = "Area", y = "Price")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

```
#plot_grid(x, y)
```

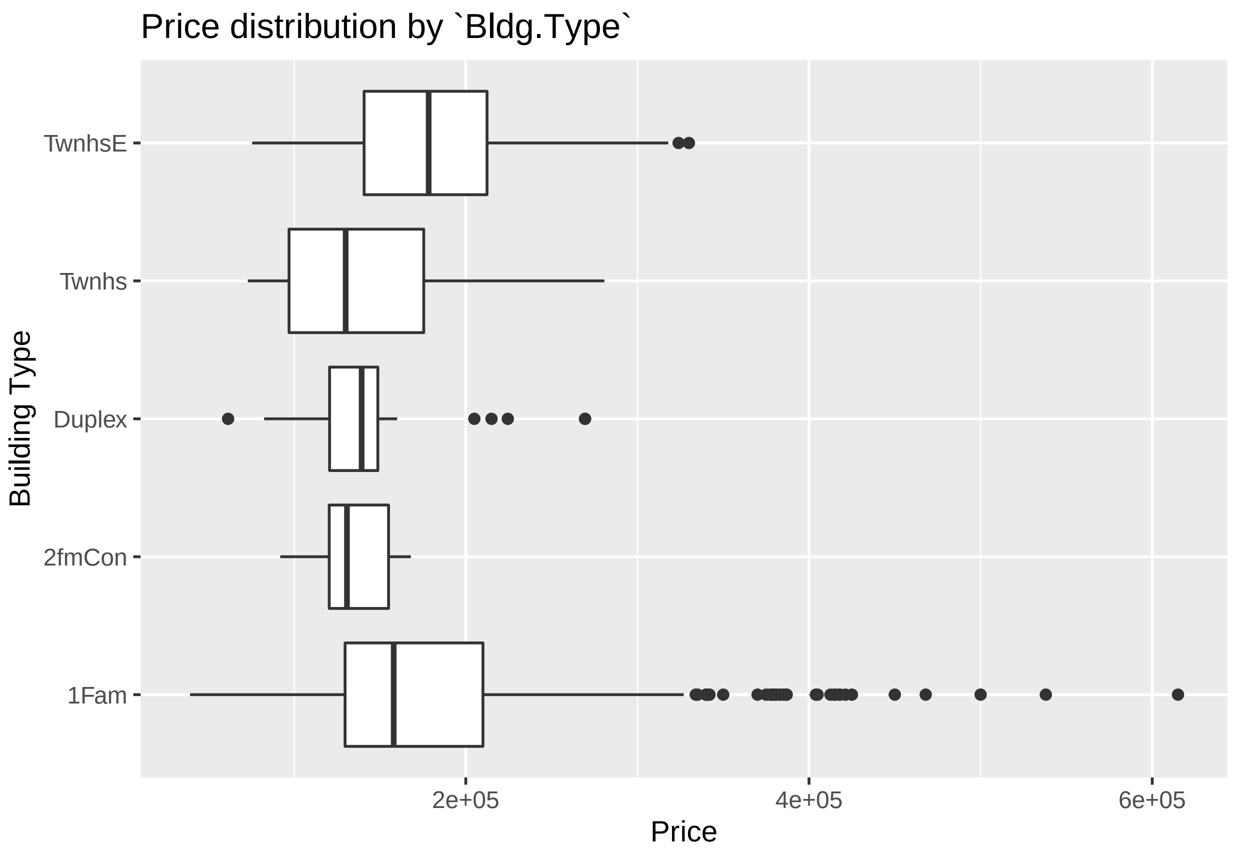


From the plots and summary statistics above we can see that after log-transforming the variable `area` we achieve a generally normal distribution and note quite a high correlation between the explanatory and the response variable.

This variable will be included in the model.

Figure 2. `Bldg.Type` variable

```
# flipped side-by-side boxplot
ggplot(ames_train, aes(x = Bldg.Type, y = price)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Price distribution by ‘Bldg.Type‘", x = "Building Type", y = "Price")
```



```r
# number of properties in each category
table(ames_train$Bldg.Type)
```

```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##    687     18     26     33     70
```

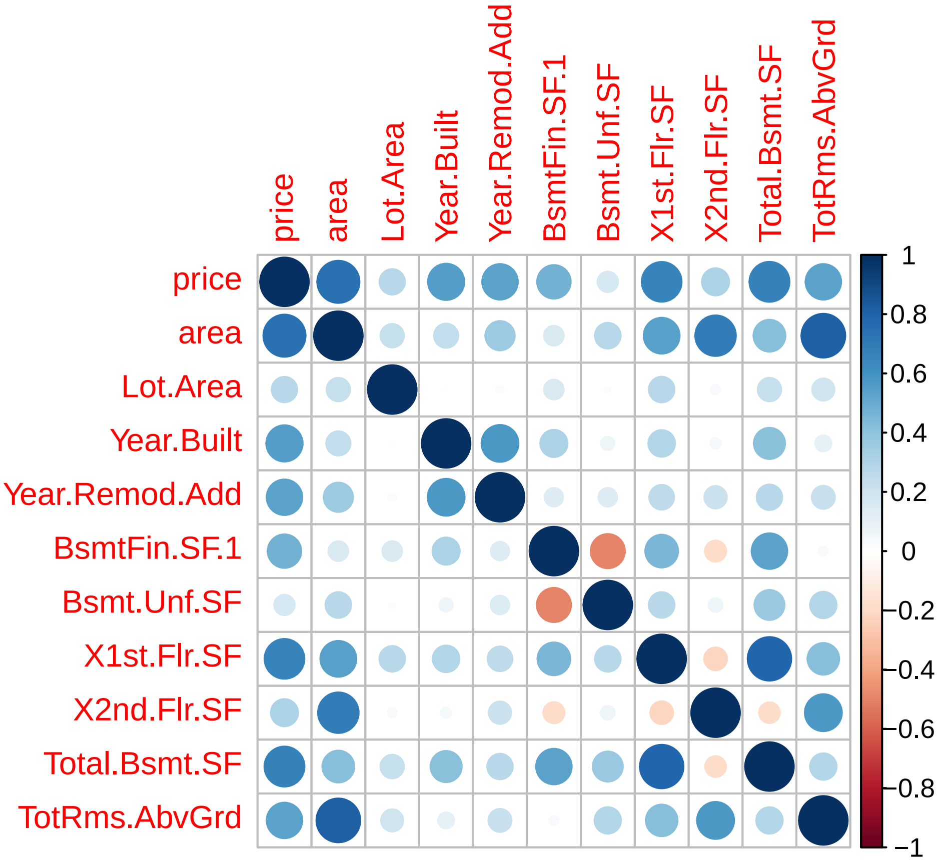
Unlike the price distribution by `Neighborhood`, grouping by `Bldg.Type` does not show that the medians are significantly different.

Secondly, there are many more Single Family homes which may lead to bias in the variable interpretation.

This variable will not be included in the model.

Figure 3. Correlation plot.

```r
# preparing the data
cor_var <- ames_train %>%
  dplyr::select(price, area, Lot.Area, Year.Built, Year.Remod.Add, BsmtFin.SF.1, Bsmt.Unf.SF, X1st.Flr.S

# correlation plot between some numeric variables
corrplot(cor(cor_var))
```



From the plot above we see there is a relatively high correlation between `price` response variable and `area`, `Year.Built`, `Year.Remod.Add`, `X1st.Flr.SF`, `Total.Bsmt.SF` explanatory variables.

At the same time, some of the explanatory variables are collinear. Such is the case of `area` and `X2nd.Flr.SF` and `TotRms.AbvGrd`, `X1st.Flr.SF` and `Total.Bsmt.SF`. One of the two collinear variables adds nothing new to the model and should not be considered.

Part 2 - Development and assessment of an initial model

Section 2.1 An Initial Model

Creating a simple, intuitive initial model based on the results of the exploratory data analysis.

Choice of the variables. One of the criteria for choosing the variables is based on the idea of breaking down the explanatory variables into groups. This allows to

(a) create a linear model that explains a large amount of variability using different factors (the initial model includes area, neighborhood, quality, year of remodeling, exterior and lot factors) and

(b) avoid collinearity as, for example, `area` and `TotRms.AbvGrd` are both very good predictors of `price` being at the same time highly correlated.

For the purpose of future model selection we will use two methods for creating the model.

```
# fitting the Ordinary Least Square (OLS) initial model
initial.model <- lm(log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add + Exterior.1s

summary(initial.model)
```

```
##
## Call:
## lm(formula = log(price) ~ log(area) + Neighborhood + Overall.Qual +
##     Year.Remod.Add + Exterior.1st + X1st.Flr.SF + Paved.Drive,
##     data = ames_train, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82238 -0.07167  0.00259  0.08355  0.52060
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.759e+00  6.420e-01   5.856 6.96e-09 ***
## log(area)            3.481e-01  2.133e-02  16.317  < 2e-16 ***
## NeighborhoodBlueste -1.466e-02  9.442e-02  -0.155 0.876630
## NeighborhoodBrDale  -1.732e-01  7.466e-02  -2.320 0.020601 *
## NeighborhoodBrkSide  2.909e-03  5.856e-02   0.050 0.960394
## NeighborhoodClearCr  2.209e-01  6.686e-02   3.305 0.000994 ***
## NeighborhoodCollgCr  9.285e-02  5.301e-02   1.752 0.080243 .
## NeighborhoodCrawfor  1.590e-01  5.998e-02   2.651 0.008197 **
## NeighborhoodEdwards -2.676e-02  5.642e-02  -0.474 0.635368
## NeighborhoodGilbert  1.031e-01  5.623e-02   1.834 0.067047 .
## NeighborhoodGreens   1.451e-01  8.580e-02   1.691 0.091165 .
## NeighborhoodGrnHill  4.171e-01  1.083e-01   3.853 0.000126 ***
## NeighborhoodIDOTRR  -7.511e-02  6.111e-02  -1.229 0.219379
## NeighborhoodMeadowV -2.763e-01  7.472e-02  -3.698 0.000233 ***
## NeighborhoodMitchel  1.280e-01  5.598e-02   2.286 0.022535 *
## NeighborhoodNAmes    6.187e-02  5.411e-02   1.143 0.253261
## NeighborhoodNoRidge  2.373e-01  5.764e-02   4.117 4.25e-05 ***
## NeighborhoodNPkVill -1.055e-02  8.688e-02  -0.121 0.903420
## NeighborhoodNridgHt  2.183e-01  5.524e-02   3.951 8.47e-05 ***
## NeighborhoodNWAmes   1.022e-01  5.751e-02   1.777 0.075886 .
## NeighborhoodOldTown -6.450e-02  5.699e-02  -1.132 0.258019
## NeighborhoodSawyer   6.611e-02  5.594e-02   1.182 0.237667
## NeighborhoodSawyerW  3.534e-02  5.597e-02   0.631 0.527903
## NeighborhoodSomerst  1.277e-01  5.483e-02   2.329 0.020128 *
## NeighborhoodStoneBr  1.411e-01  6.522e-02   2.164 0.030775 *
## NeighborhoodSWISU   -6.611e-02  6.922e-02  -0.955 0.339832
## NeighborhoodTimber   1.511e-01  6.007e-02   2.515 0.012108 *
## NeighborhoodVeenker  2.228e-01  6.914e-02   3.223 0.001322 **
## Overall.Qual         9.001e-02  6.130e-03  14.684  < 2e-16 ***
## Year.Remod.Add       2.382e-03  3.246e-04   7.339 5.33e-13 ***
## Exterior.1stBrkComm  2.621e-01  1.433e-01   1.829 0.067741 .
## Exterior.1stBrkFace  1.131e-01  5.450e-02   2.075 0.038302 *
## Exterior.1stCemntBd  1.953e-01  6.200e-02   3.149 0.001699 **
## Exterior.1stHdBoard  9.546e-02  5.047e-02   1.891 0.058940 .
## Exterior.1stImStucc  3.978e-02  1.445e-01   0.275 0.783136
## Exterior.1stMetalSd  9.343e-02  4.947e-02   1.889 0.059290 .
## Exterior.1stPlywood  6.912e-02  5.212e-02   1.326 0.185156
## Exterior.1stStucco   1.093e-01  5.922e-02   1.845 0.065386 .
## Exterior.1stVinylSd  1.082e-01  5.035e-02   2.148 0.032002 *
## Exterior.1stWd Sdng  8.858e-02  4.965e-02   1.784 0.074780 .
## Exterior.1stWdShing  7.152e-02  5.768e-02   1.240 0.215401
## X1st.Flr.SF          1.794e-04  1.851e-05   9.694  < 2e-16 ***
## Paved.DriveP         1.688e-02  3.158e-02   0.535 0.592991
## Paved.DriveY         1.029e-01  2.253e-02   4.568 5.70e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1329 on 790 degrees of freedom
## Multiple R-squared:  0.8852, Adjusted R-squared:  0.8789
## F-statistic: 141.6 on 43 and 790 DF,  p-value: < 2.2e-16
```

```
BIC(initial.model)
```

```
## [1] -742.1167
```

Model results. From the summary table we see that many of the variables and the model itself are highly statistically significant with a very low p-value and the Adjusted R-squared, *i.e.* the explained variability, at the level of 88%.

All the predictors except `Neighborhood` increase the price all other variables held constant. In the case of the `Neighborhood` categorical variable it depends on the specific neighborhood we are using as input.

Section 2.2 Model Selection

Choosing the “best” model using the initial model as a starting point.

As mentioned in Section 2.1, we will use two approaches. Under the first approach, will will use the BIC as the criterion for selecting the best model. Under the second approach, we will use the Bayesian model averaging (BMA) to create posteriors from the data we have in the data set.

```
# BIC backwards elimination setting k = log(n)
BIC.model <- stepAIC(initial.model,
                     scale = 0,
                     direction = c("backward"),
                     trace = 1,
                     keep = NULL,
                     steps = 1000,
                     use.start = FALSE,
                     k = log(nrow(ames_train)))
```

```
## Start:  AIC=-3115.63
## log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add +
##     Exterior.1st + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Exterior.1st   11    0.2966 14.248 -3172.1
## <none>                        13.952 -3115.6
## - Paved.Drive     2    0.4527 14.404 -3102.5
## - Neighborhood   26    4.2062 18.158 -3070.8
## - Year.Remod.Add  1    0.9513 14.903 -3067.3
## - X1st.Flr.SF     1    1.6595 15.611 -3028.6
## - Overall.Qual    1    3.8080 17.760 -2921.1
## - log(area)       1    4.7020 18.654 -2880.1
##
## Step:  AIC=-3172.07
## log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add +
##     X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        14.248 -3172.1
## - Paved.Drive     2    0.4743 14.723 -3158.2
## - Neighborhood   26    4.3124 18.561 -3126.4
## - Year.Remod.Add  1    1.1598 15.408 -3113.5
## - X1st.Flr.SF     1    1.6694 15.918 -3086.4
## - Overall.Qual    1    4.0592 18.308 -2969.7
## - log(area)       1    4.9720 19.220 -2929.2
```

```
summary(BIC.model)
```

```
##
## Call:
## lm(formula = log(price) ~ log(area) + Neighborhood + Overall.Qual +
##     Year.Remod.Add + X1st.Flr.SF + Paved.Drive, data = ames_train,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80352 -0.07180  0.00249  0.08265  0.53470
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.533e+00  6.198e-01   5.700 1.68e-08 ***
## log(area)            3.545e-01  2.120e-02  16.719  < 2e-16 ***
## NeighborhoodBlueste -2.627e-02  9.339e-02  -0.281 0.778553
## NeighborhoodBrDale  -1.804e-01  7.337e-02  -2.459 0.014156 *
## NeighborhoodBrkSide -2.373e-03  5.750e-02  -0.041 0.967094
## NeighborhoodClearCr  2.002e-01  6.564e-02   3.050 0.002361 **
## NeighborhoodCollgCr  9.315e-02  5.312e-02   1.754 0.079872 .
## NeighborhoodCrawfor  1.484e-01  5.873e-02   2.527 0.011682 *
## NeighborhoodEdwards -3.655e-02  5.577e-02  -0.655 0.512437
## NeighborhoodGilbert  9.785e-02  5.613e-02   1.743 0.081684 .
## NeighborhoodGreens   1.277e-01  8.431e-02   1.515 0.130211
## NeighborhoodGrnHill  4.119e-01  1.070e-01   3.850 0.000128 ***
## NeighborhoodIDOTRR  -7.881e-02  6.050e-02  -1.303 0.193049
## NeighborhoodMeadowV -1.820e-01  6.311e-02  -2.884 0.004036 **
## NeighborhoodMitchel  1.215e-01  5.549e-02   2.189 0.028860 *
## NeighborhoodNAmes    5.551e-02  5.334e-02   1.041 0.298328
## NeighborhoodNoRidge  2.267e-01  5.730e-02   3.956 8.29e-05 ***
## NeighborhoodNPkVill -4.485e-02  8.442e-02  -0.531 0.595338
## NeighborhoodNridgHt  2.133e-01  5.533e-02   3.856 0.000125 ***
## NeighborhoodNWAmes   8.746e-02  5.630e-02   1.554 0.120693
## NeighborhoodOldTown -7.619e-02  5.609e-02  -1.358 0.174695
## NeighborhoodSawyer   5.791e-02  5.501e-02   1.053 0.292720
## NeighborhoodSawyerW  2.465e-02  5.560e-02   0.443 0.657615
## NeighborhoodSomerst  1.249e-01  5.462e-02   2.286 0.022495 *
## NeighborhoodStoneBr  1.635e-01  6.369e-02   2.567 0.010428 *
## NeighborhoodSWISU   -7.411e-02  6.831e-02  -1.085 0.278292
## NeighborhoodTimber   1.537e-01  6.002e-02   2.562 0.010601 *
## NeighborhoodVeenker  2.073e-01  6.753e-02   3.069 0.002217 **
## Overall.Qual         9.151e-02  6.058e-03  15.106  < 2e-16 ***
## Year.Remod.Add       2.523e-03  3.124e-04   8.075 2.48e-15 ***
## X1st.Flr.SF          1.756e-04  1.813e-05   9.687  < 2e-16 ***
## Paved.DriveP         1.455e-02  3.155e-02   0.461 0.644699
## Paved.DriveY         1.030e-01  2.231e-02   4.617 4.54e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1334 on 801 degrees of freedom
## Multiple R-squared:  0.8827, Adjusted R-squared:  0.878
## F-statistic: 188.4 on 32 and 801 DF,  p-value: < 2.2e-16
```

```
BIC(BIC.model)
```

```
## [1] -798.5587
```

```
# implementing Bayesian model averaging (BMA)
model.bas <- bas.lm(log(price) ~ log(area) + Neighborhood +
                      Overall.Qual + Year.Remod.Add + Exterior.1st +
                      X1st.Flr.SF + Paved.Drive,
                    data = ames_train,
                    na.action = na.omit,
                    prior = "BIC",
                    modelprior=uniform())

model.bas
```

```
##
## Call:
## bas.lm(formula = log(price) ~ log(area) + Neighborhood + Overall.Qual +
##     Year.Remod.Add + Exterior.1st + X1st.Flr.SF + Paved.Drive,
##     data = ames_train, na.action = na.omit, prior = "BIC", modelprior = uniform())
##
##
##  Marginal Posterior Inclusion Probabilities:
##           Intercept            log(area)  NeighborhoodBlueste
##             1.00000              1.00000              0.08879
##  NeighborhoodBrDale  NeighborhoodBrkSide  NeighborhoodClearCr
##             0.99994              0.94766              0.56626
## NeighborhoodCollgCr  NeighborhoodCrawfor  NeighborhoodEdwards
##             0.03710              0.30042              0.99999
## NeighborhoodGilbert   NeighborhoodGreens  NeighborhoodGrnHill
##             0.02706              0.02660              0.89380
##  NeighborhoodIDOTRR  NeighborhoodMeadowV  NeighborhoodMitchel
##             0.99997              1.00000              0.11172
##   NeighborhoodNAmes  NeighborhoodNoRidge  NeighborhoodNPkVill
##             0.54165              0.99791              0.21653
## NeighborhoodNridgHt   NeighborhoodNWAmes  NeighborhoodOldTown
##             0.99501              0.03918              1.00000
##  NeighborhoodSawyer  NeighborhoodSawyerW  NeighborhoodSomerst
##             0.26237              0.79198              0.05381
## NeighborhoodStoneBr    NeighborhoodSWISU   NeighborhoodTimber
##             0.04207              0.96464              0.08280
## NeighborhoodVeenker         Overall.Qual       Year.Remod.Add
##             0.41616              1.00000              1.00000
## Exterior.1stBrkComm  Exterior.1stBrkFace  Exterior.1stCemntBd
##             0.05151              0.05140              0.68775
## Exterior.1stHdBoard  Exterior.1stImStucc  Exterior.1stMetalSd
##             0.02614              0.02722              0.02504
## Exterior.1stPlywood   Exterior.1stStucco  Exterior.1stVinylSd
##             0.13078              0.03178              0.04883
## Exterior.1stWd Sdng  Exterior.1stWdShing          X1st.Flr.SF
##             0.02956              0.02760              1.00000
##        Paved.DriveP         Paved.DriveY
##             0.02880              0.99983
```

```
summary(model.bas)
```

```
##                      P(B != 0 | Y)    model 1       model 2       model 3
## Intercept              1.00000000     1.0000     1.0000000     1.0000000
## log(area)              1.00000000     1.0000     1.0000000     1.0000000
## NeighborhoodBlueste    0.08878815     0.0000     0.0000000     0.0000000
## NeighborhoodBrDale     0.99994014     1.0000     1.0000000     1.0000000
## NeighborhoodBrkSide    0.94765963     1.0000     1.0000000     1.0000000
## NeighborhoodClearCr    0.56625864     0.0000     0.0000000     1.0000000
## NeighborhoodCollgCr    0.03709951     0.0000     0.0000000     0.0000000
## NeighborhoodCrawfor    0.30042284     0.0000     0.0000000     0.0000000
## NeighborhoodEdwards    0.99998888     1.0000     1.0000000     1.0000000
## NeighborhoodGilbert    0.02705664     0.0000     0.0000000     0.0000000
## NeighborhoodGreens     0.02660389     0.0000     0.0000000     0.0000000
## NeighborhoodGrnHill    0.89380171     1.0000     1.0000000     1.0000000
## NeighborhoodIDOTRR     0.99997121     1.0000     1.0000000     1.0000000
## NeighborhoodMeadowV    1.00000000     1.0000     1.0000000     1.0000000
## NeighborhoodMitchel    0.11172379     0.0000     0.0000000     0.0000000
## NeighborhoodNAmes      0.54164910     1.0000     1.0000000     1.0000000
## NeighborhoodNoRidge    0.99791126     1.0000     1.0000000     1.0000000
## NeighborhoodNPkVill    0.21653332     0.0000     0.0000000     0.0000000
## NeighborhoodNridgHt    0.99500941     1.0000     1.0000000     1.0000000
## NeighborhoodNWAmes     0.03918267     0.0000     0.0000000     0.0000000
## NeighborhoodOldTown    1.00000000     1.0000     1.0000000     1.0000000
## NeighborhoodSawyer     0.26236976     1.0000     0.0000000     0.0000000
## NeighborhoodSawyerW    0.79198309     1.0000     1.0000000     1.0000000
## NeighborhoodSomerst    0.05381294     0.0000     0.0000000     0.0000000
## NeighborhoodStoneBr    0.04206538     0.0000     0.0000000     0.0000000
## NeighborhoodSWISU      0.96464059     1.0000     1.0000000     1.0000000
## NeighborhoodTimber     0.08280101     0.0000     0.0000000     0.0000000
## NeighborhoodVeenker    0.41615794     0.0000     0.0000000     0.0000000
## Overall.Qual           1.00000000     1.0000     1.0000000     1.0000000
## Year.Remod.Add         1.00000000     1.0000     1.0000000     1.0000000
## Exterior.1stBrkComm    0.05151076     0.0000     0.0000000     0.0000000
## Exterior.1stBrkFace    0.05139808     0.0000     0.0000000     0.0000000
## Exterior.1stCemntBd    0.68775435     1.0000     1.0000000     1.0000000
## Exterior.1stHdBoard    0.02613763     0.0000     0.0000000     0.0000000
## Exterior.1stImStucc    0.02721757     0.0000     0.0000000     0.0000000
## Exterior.1stMetalSd    0.02503754     0.0000     0.0000000     0.0000000
## Exterior.1stPlywood    0.13078454     0.0000     0.0000000     0.0000000
## Exterior.1stStucco     0.03178205     0.0000     0.0000000     0.0000000
## Exterior.1stVinylSd    0.04883134     0.0000     0.0000000     0.0000000
## Exterior.1stWd Sdng    0.02956183     0.0000     0.0000000     0.0000000
## Exterior.1stWdShing    0.02759648     0.0000     0.0000000     0.0000000
## X1st.Flr.SF            1.00000000     1.0000     1.0000000     1.0000000
## Paved.DriveP           0.02879790     0.0000     0.0000000     0.0000000
## Paved.DriveY           0.99982833     1.0000     1.0000000     1.0000000
## BF                             NA     1.0000     0.9309327     0.7691037
## PostProbs                      NA     0.0105     0.0098000     0.0081000
## R2                             NA     0.8792     0.8782000     0.8791000
## dim                            NA    20.0000    19.0000000    20.0000000
## logmarg                        NA -1187.5415 -1187.6130832 -1187.8040444
##                           model 4     model 5
## Intercept               1.0000000     1.00000
## log(area)               1.0000000     1.00000
## NeighborhoodBlueste     0.0000000     0.00000
## NeighborhoodBrDale      1.0000000     1.00000
## NeighborhoodBrkSide     1.0000000     1.00000
## NeighborhoodClearCr     0.0000000     1.00000
## NeighborhoodCollgCr     0.0000000     0.00000
## NeighborhoodCrawfor     0.0000000     0.00000
## NeighborhoodEdwards     1.0000000     1.00000
## NeighborhoodGilbert     0.0000000     0.00000
## NeighborhoodGreens      0.0000000     0.00000
## NeighborhoodGrnHill     1.0000000     1.00000
## NeighborhoodIDOTRR      1.0000000     1.00000
## NeighborhoodMeadowV     1.0000000     1.00000
## NeighborhoodMitchel     0.0000000     0.00000
## NeighborhoodNAmes       1.0000000     0.00000
## NeighborhoodNoRidge     1.0000000     1.00000
## NeighborhoodNPkVill     0.0000000     0.00000
## NeighborhoodNridgHt     1.0000000     1.00000
## NeighborhoodNWAmes      0.0000000     0.00000
## NeighborhoodOldTown     1.0000000     1.00000
## NeighborhoodSawyer      1.0000000     0.00000
## NeighborhoodSawyerW     1.0000000     1.00000
## NeighborhoodSomerst     0.0000000     0.00000
## NeighborhoodStoneBr     0.0000000     0.00000
## NeighborhoodSWISU       1.0000000     1.00000
## NeighborhoodTimber      0.0000000     0.00000
## NeighborhoodVeenker     0.0000000     0.00000
## Overall.Qual            1.0000000     1.00000
## Year.Remod.Add          1.0000000     1.00000
## Exterior.1stBrkComm     0.0000000     0.00000
## Exterior.1stBrkFace     0.0000000     0.00000
## Exterior.1stCemntBd     0.0000000     1.00000
## Exterior.1stHdBoard     0.0000000     0.00000
## Exterior.1stImStucc     0.0000000     0.00000
## Exterior.1stMetalSd     0.0000000     0.00000
## Exterior.1stPlywood     0.0000000     0.00000
## Exterior.1stStucco      0.0000000     0.00000
## Exterior.1stVinylSd     0.0000000     0.00000
## Exterior.1stWd Sdng     0.0000000     0.00000
## Exterior.1stWdShing     0.0000000     0.00000
## X1st.Flr.SF             1.0000000     1.00000
## Paved.DriveP            0.0000000     0.00000
## Paved.DriveY            1.0000000     1.00000
## BF                      0.7635298     0.69434
## PostProbs               0.0080000     0.00730
## R2                      0.8781000     0.87810
## dim                    19.0000000    19.00000
## logmarg             -1187.8113179 -1187.90631
```

Conclusion. Using the BIC selection method, we see that we are left with a model with 6 predictors (`Exterior.1st` was dropped), the BIC improved and the Adjusted R-squared remained the same.

Most likely, it means that the dropped variable made no contribution to the model.

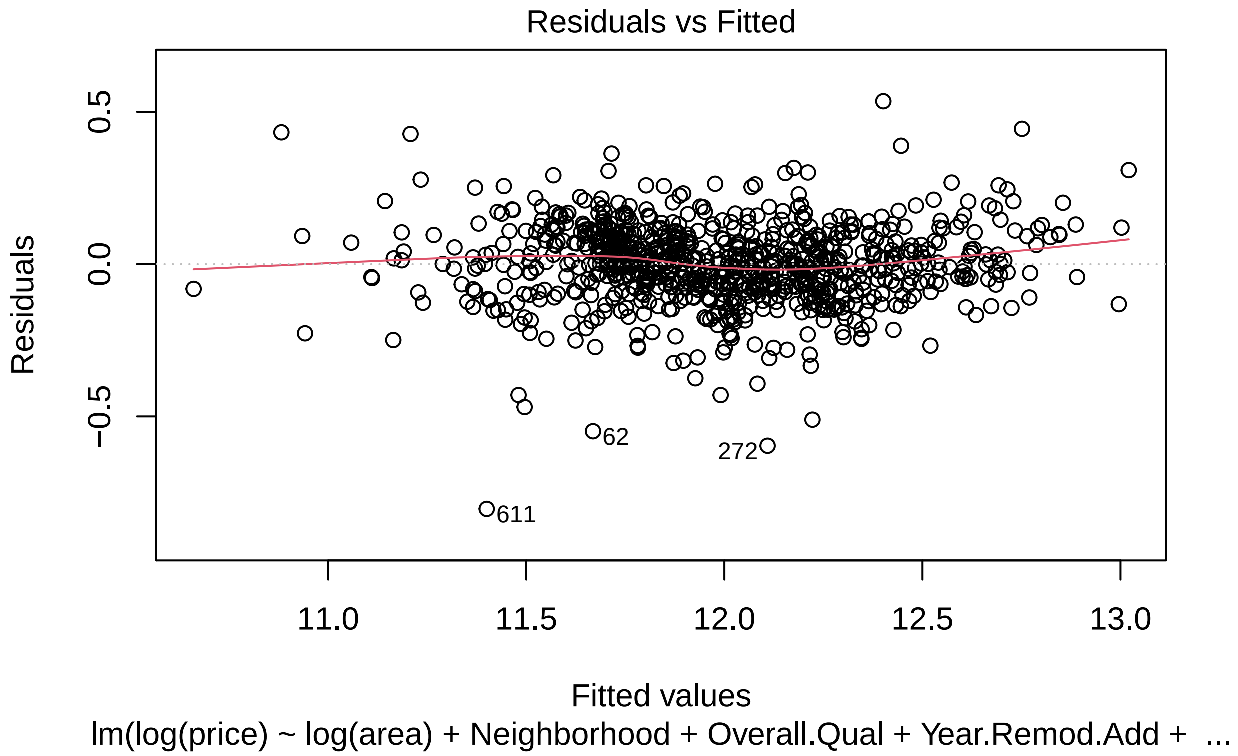
Under the BAS method, we have several models we will use for prediction. Model 1 has the posterior probability of inclusion of 0.01 and uses the following variables: intercept, area, neighborhood, exterior covering on house, overall quality, remodel date, type of drive and first floor area.

Section 2.3 Initial Model Residuals

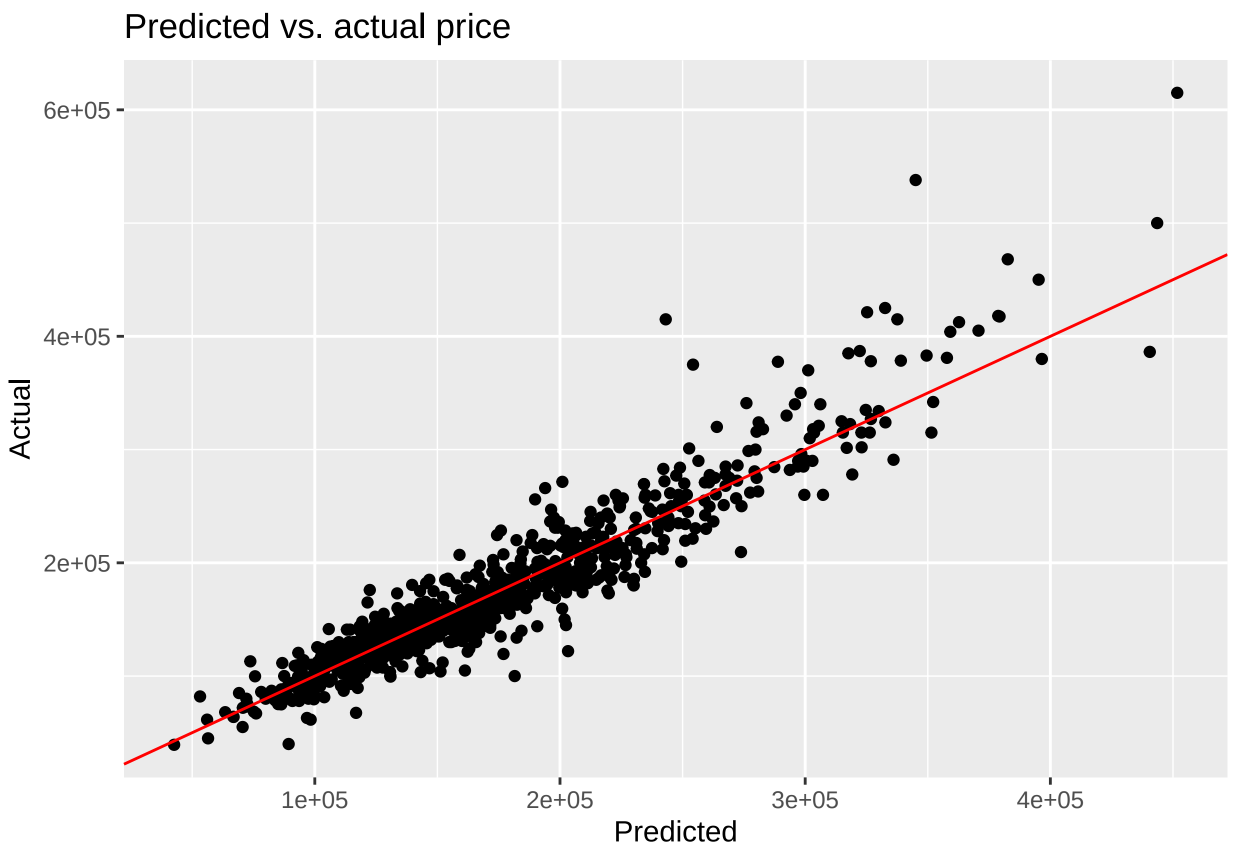
Model performance assessment.

For the purpose of examining the residuals, we will construct several plots. We will be using the `BIC.model`.

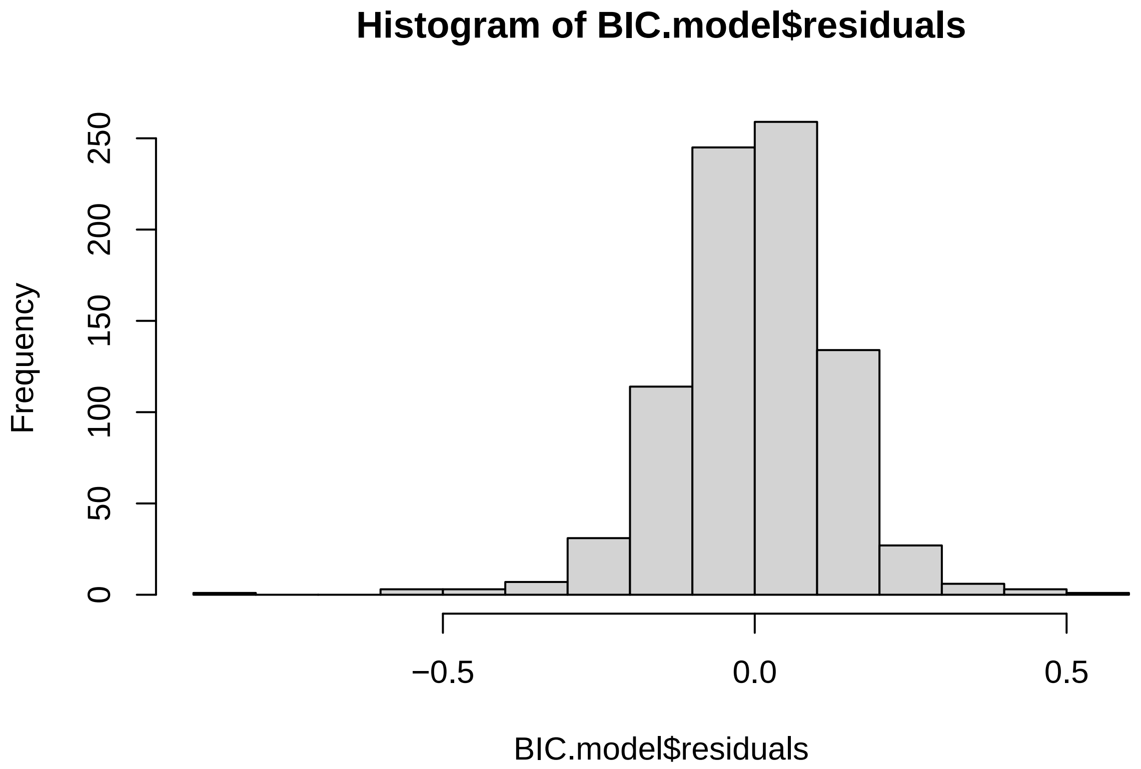
```
# residuals vs. fitted
plot(BIC.model, which = 1)
```



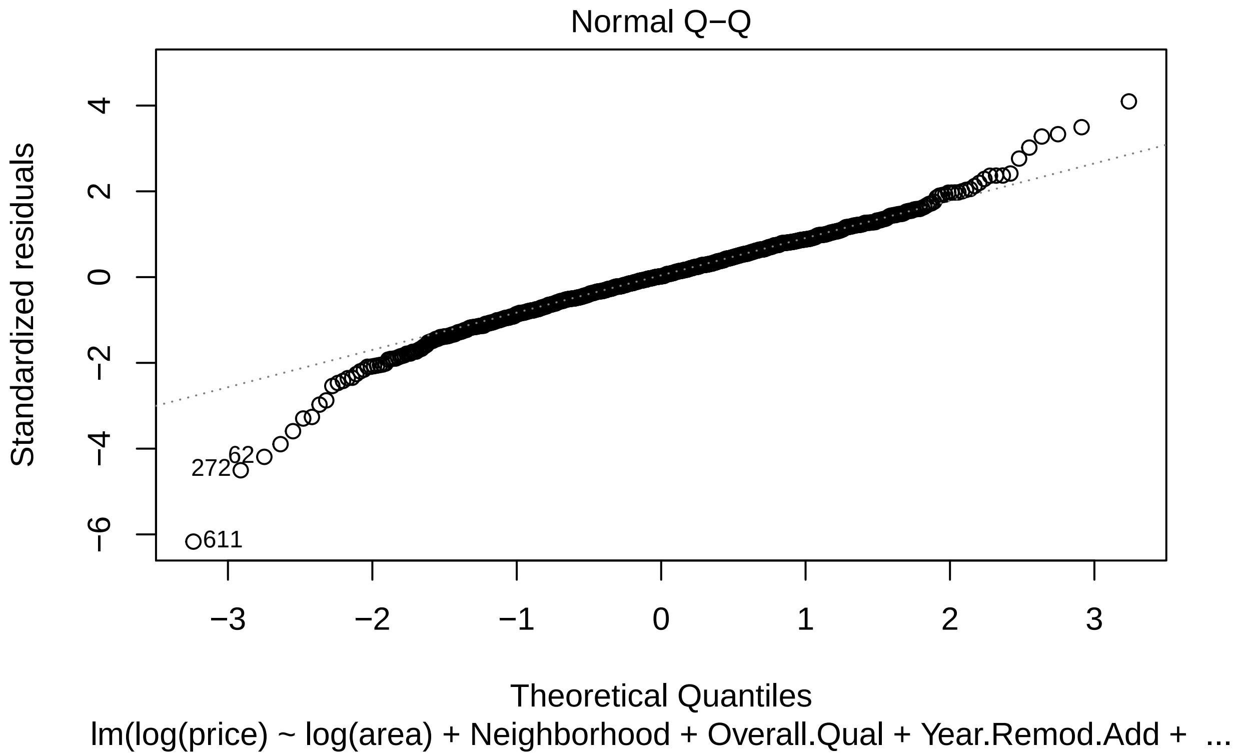
```
# predicted vs. actual
ames_train$prediction <- predict(BIC.model)
ames_train$prediction = exp(ames_train$prediction)

ggplot(ames_train, aes(x = prediction, y = price)) +
geom_point() +
geom_abline(color = "red") +
labs(title="Predicted vs. actual price", x="Predicted", y="Actual")
```



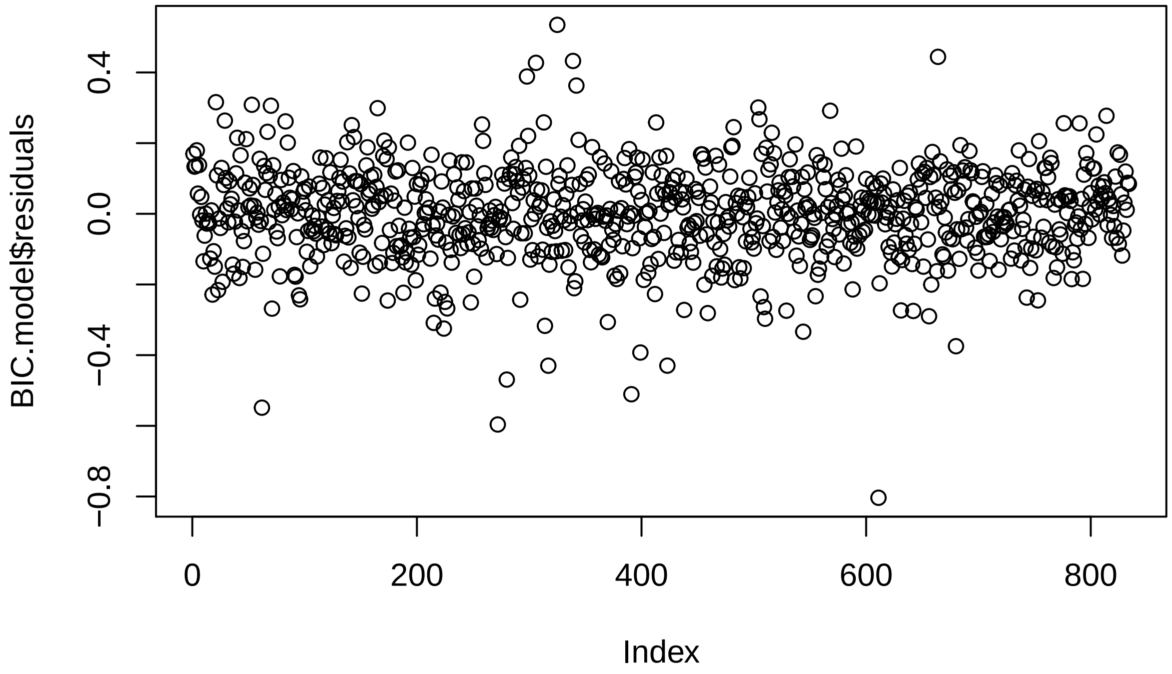
```r
# residuals distribution
hist(BIC.model$residuals)
```



```
plot(BIC.model, which = 2)
```



```
# independence of residuals
plot(BIC.model$residuals)
```



As we can see from the plots, the model appears to fit the data well, there is a clear trend which is however weaker at the higher level of prices.

At the same time, there are several outliers that affect the normality of the residuals distribution. These are the observations 272 and 611.

```
# looking at the outliers
outliers.df <- ames_train %>%
  dplyr::select(price, prediction, area, Neighborhood, Overall.Qual, Year.Remod.Add, X1st.Flr.SF, Paved

outliers.df[c(272,611), ]
```

```
## # A tibble: 2 x 8
##    price prediction  area Neighborhood Overall.Qual Year.Remod.Add X1st.Flr.SF
##    <int>      <dbl> <int> <fct>               <int>          <int>       <int>
## 1 100000    181527.  1764 NAmes                   5           1982        1764
## 2  40000     89336.  1317 IDOTRR                  4           1950         649
## # ... with 1 more variable: Paved.Drive <fct>
```

Given the area, the overall quality, and the remodel date the actual price is very low. These outliers might affect the model predictive capacity.

```
# deleting outliers
ames_train <- ames_train[-c(272,611),]

# new initial model
```

```
initial.model.2 <- lm(log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add + Exterior.

# new BIC model
BIC.model.2 <- stepAIC(initial.model.2,
                       scale = 0,
                       direction = c("backward"),
                       trace = 1,
                       keep = NULL,
                       steps = 1000,
                       use.start = FALSE,
                       k = log(nrow(ames_train)))
```

```
## Start:  AIC=-3173.25
## log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add +
##     Exterior.1st + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Exterior.1st   11    0.3490 13.210 -3224.9
## <none>                        12.861 -3173.3
## - Paved.Drive     2    0.3979 13.259 -3161.3
## - Year.Remod.Add  1    0.9418 13.803 -3121.2
## - Neighborhood   26    4.1703 17.032 -3114.4
## - X1st.Flr.SF     1    1.6307 14.492 -3080.7
## - Overall.Qual    1    3.5999 16.461 -2974.7
## - log(area)       1    4.9796 17.841 -2907.7
##
## Step:  AIC=-3224.94
## log(price) ~ log(area) + Neighborhood + Overall.Qual + Year.Remod.Add +
##     X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        13.210 -3224.9
## - Paved.Drive     2    0.4357 13.646 -3211.4
## - Neighborhood   26    4.3229 17.533 -3164.2
## - Year.Remod.Add  1    1.1170 14.328 -3164.1
## - X1st.Flr.SF     1    1.6756 14.886 -3132.3
## - Overall.Qual    1    3.8337 17.044 -3019.7
## - log(area)       1    5.2589 18.469 -2952.9
```

```
summary(BIC.model.2)
```

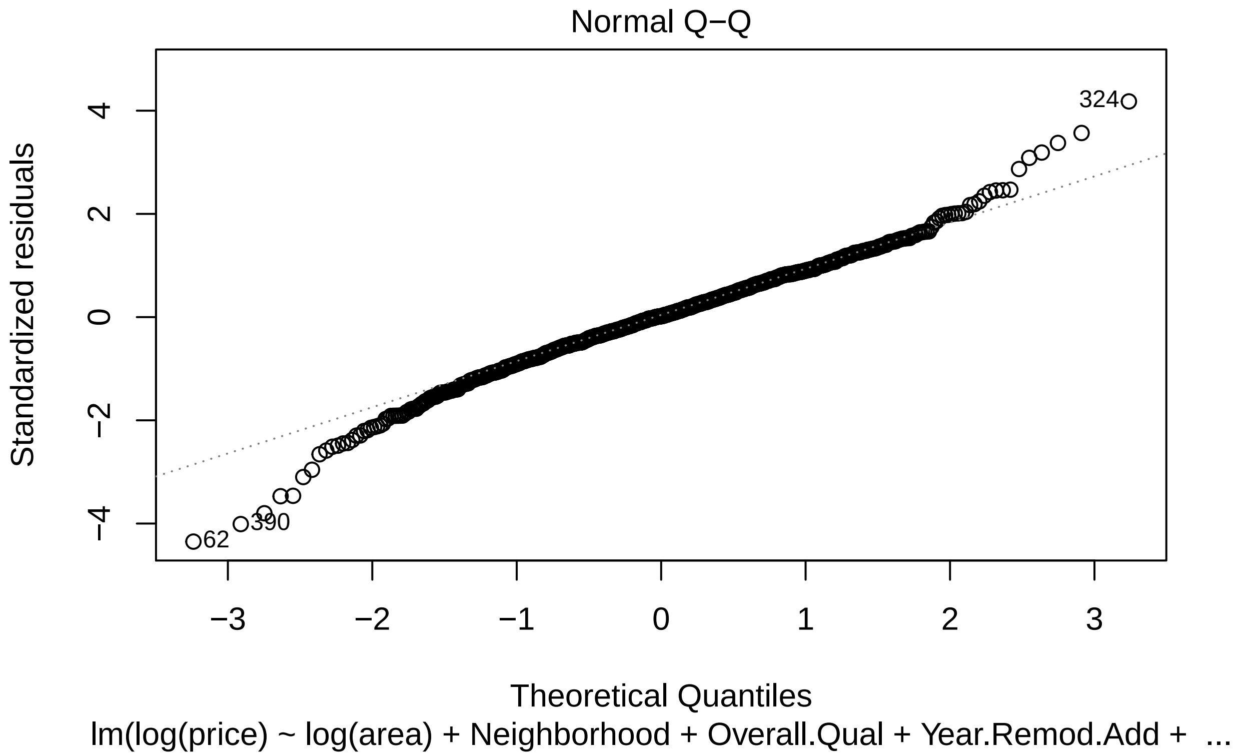
```
##
## Call:
## lm(formula = log(price) ~ log(area) + Neighborhood + Overall.Qual +
##     Year.Remod.Add + X1st.Flr.SF + Paved.Drive, data = ames_train,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54885 -0.07145  0.00230  0.08232  0.52591
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          3.567e+00  5.980e-01   5.965 3.67e-09 ***
## log(area)            3.654e-01  2.049e-02  17.834  < 2e-16 ***
## NeighborhoodBlueste -2.660e-02  9.004e-02  -0.295  0.76777    
## NeighborhoodBrDale  -1.814e-01  7.075e-02  -2.564  0.01054 *  
## NeighborhoodBrkSide -6.533e-03  5.545e-02  -0.118  0.90624    
## NeighborhoodClearCr  1.942e-01  6.329e-02   3.068  0.00222 ** 
## NeighborhoodCollgCr  9.154e-02  5.121e-02   1.787  0.07424 .  
## NeighborhoodCrawfor  1.432e-01  5.663e-02   2.529  0.01162 *  
## NeighborhoodEdwards -4.199e-02  5.378e-02  -0.781  0.43521    
## NeighborhoodGilbert  9.489e-02  5.412e-02   1.753  0.07994 .  
## NeighborhoodGreens   1.317e-01  8.128e-02   1.620  0.10562    
## NeighborhoodGrnHill  4.111e-01  1.032e-01   3.985 7.35e-05 ***
## NeighborhoodIDOTRR  -5.597e-02  5.842e-02  -0.958  0.33829    
## NeighborhoodMeadowV -1.855e-01  6.086e-02  -3.049  0.00237 ** 
## NeighborhoodMitchel  1.185e-01  5.350e-02   2.215  0.02705 *  
## NeighborhoodNAmes    5.587e-02  5.144e-02   1.086  0.27775    
## NeighborhoodNoRidge  2.228e-01  5.525e-02   4.032 6.06e-05 ***
## NeighborhoodNPkVill -4.662e-02  8.139e-02  -0.573  0.56691    
## NeighborhoodNridgHt  2.125e-01  5.335e-02   3.983 7.43e-05 ***
## NeighborhoodNWAmes   8.205e-02  5.428e-02   1.512  0.13102    
## NeighborhoodOldTown -8.400e-02  5.409e-02  -1.553  0.12081    
## NeighborhoodSawyer   5.405e-02  5.304e-02   1.019  0.30842    
## NeighborhoodSawyerW  2.114e-02  5.361e-02   0.394  0.69347    
## NeighborhoodSomerst  1.243e-01  5.266e-02   2.361  0.01848 *  
## NeighborhoodStoneBr  1.632e-01  6.141e-02   2.658  0.00801 ** 
## NeighborhoodSWISU   -7.857e-02  6.587e-02  -1.193  0.23329    
## NeighborhoodTimber   1.527e-01  5.786e-02   2.639  0.00847 ** 
## NeighborhoodVeenker  2.047e-01  6.511e-02   3.144  0.00173 ** 
## Overall.Qual         8.907e-02  5.849e-03  15.227  < 2e-16 ***
## Year.Remod.Add       2.478e-03  3.015e-04   8.220 8.23e-16 ***
## X1st.Flr.SF          1.764e-04  1.752e-05  10.067  < 2e-16 ***
## Paved.DriveP         5.218e-03  3.045e-02   0.171  0.86400    
## Paved.DriveY         9.585e-02  2.154e-02   4.450 9.83e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.1286 on 799 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8848 
## F-statistic: 200.5 on 32 and 799 DF,  p-value: < 2.2e-16
```

```r
BIC(BIC.model.2)
```

```
## [1] -857.1033
```

```r
# checking the residuals
plot(BIC.model.2, which = 2)
```



Removing the two outliers we improved the Adjusted R-squared and the BIC.

At the same time, it would be wise to investigate further why these two properties have such a low price. These particular properties may be a good investment.

Section 2.4 Initial Model RMSE

Calculating the initial model RMSE.

The RMSE for the second BIC model is $24323.9. It measures the error of a model in predicting quantitative data. The smaller the error the better.

```
# extracting predictions
predict.BIC <- exp(predict(BIC.model.2, ames_train))

# extracting residuals
resid.BIC <- ames_train$price - predict.BIC

# calculating RMSE
rmse.BIC <- sqrt(mean(resid.BIC^2))
rmse.BIC
```

```
## [1] 24301.57
```

Section 2.5 Overfitting and out-of-sample data

Comparing the performance of the model on both in-sample and out-of-sample data sets.

```
# loading out-of-sample data set
load("ames_test.Rdata")
```

First, we will have to remove one observation in the Landmark neighborhood as otherwise R makes it impossible to compare the two data sets.

```
# removing ‘Landmrk‘ observation
ames_test <- subset(ames_test, Neighborhood != "Landmrk")
```

Then, we will calculate the RMSE using the test data.

```
# extracting predictions
predict.BIC.test <- exp(predict(BIC.model.2, ames_test))

# extracting residuals
resid.BIC.test <- ames_test$price - predict.BIC.test

# calculating RMSE
rmse.BIC.test <- sqrt(mean(resid.BIC.test^2))
rmse.BIC.test
```

```
## [1] 25486.81
```

It would also be useful to calculate the coverage probability.

```
# predicting prices
predict.BIC.test <- exp(predict(BIC.model.2, ames_test, interval = "prediction"))

# calculating proportion of observations that fall within prediction intervals
coverage.prob.BIC.test <- mean(ames_test$price > predict.BIC.test[,"lwr"] &
                              ames_test$price < predict.BIC.test[,"upr"])
coverage.prob.BIC.test
```

```
## [1] 0.939951
```

```
# proportion of observations (rows) in ‘ames_test‘ have sales prices that fall outside the prediction i
1-coverage.prob.BIC.test
```

```
## [1] 0.06004902
```

Although both the RMSE and the coverage probability are not perfect, they do not significantly diverge from the training data calculations and thus the model fits the test data reasonably well.

Part 3 Development of a Final Model

Creating a final model to predict housing prices in Ames, IA.

Section 3.1 Final Model

In this section we will fit the final model using the same criteria of breaking down the variables by groups thus achieving representativeness and avoiding collinearity.

```
# converting NAs to a category
ames_train$Garage.Qual <- fct_explicit_na(ames_train$Garage.Qual, "No garage")

# formula
fmla <- log(price) ~ log(area) + log(Lot.Area) + Lot.Config + Neighborhood + House.Style + Overall.Qual

# fitting the final model
final.model <- lm(fmla, ames_train, na.action = na.omit)

summary(final.model)
```

```
##
## Call:
## lm(formula = fmla, data = ames_train, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50493 -0.06059 -0.00016  0.06874  0.39527
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.725e+00  6.004e-01   4.539 6.55e-06 ***
## log(area)              3.451e-01  4.600e-02   7.503 1.73e-13 ***
## log(Lot.Area)          9.142e-02  1.172e-02   7.802 1.98e-14 ***
## Lot.ConfigCulDSac      2.061e-02  1.716e-02   1.201 0.230218
## Lot.ConfigFR2         -3.301e-02  2.204e-02  -1.498 0.134561
## Lot.ConfigFR3         -8.197e-02  5.539e-02  -1.480 0.139317
## Lot.ConfigInside       1.277e-02  1.028e-02   1.242 0.214636
## NeighborhoodBlueste   -2.329e-02  7.650e-02  -0.304 0.760904
## NeighborhoodBrDale    -1.211e-01  6.084e-02  -1.991 0.046815 *
## NeighborhoodBrkSide   -5.737e-02  4.880e-02  -1.176 0.240090
## NeighborhoodClearCr   -1.386e-03  5.724e-02  -0.024 0.980692
## NeighborhoodCollgCr   -2.234e-02  4.495e-02  -0.497 0.619274
## NeighborhoodCrawfor    8.034e-02  4.950e-02   1.623 0.104983
## NeighborhoodEdwards   -7.615e-02  4.684e-02  -1.626 0.104402
## NeighborhoodGilbert   -3.071e-02  4.748e-02  -0.647 0.517932
## NeighborhoodGreens     8.573e-02  6.978e-02   1.229 0.219598
## NeighborhoodGrnHill    4.439e-01  8.746e-02   5.076 4.84e-07 ***
## NeighborhoodIDOTRR    -1.029e-01  5.169e-02  -1.990 0.046921 *
## NeighborhoodMeadowV   -1.357e-01  5.266e-02  -2.577 0.010139 *
## NeighborhoodMitchel   -1.202e-02  4.727e-02  -0.254 0.799406
## NeighborhoodNAmes     -1.785e-02  4.566e-02  -0.391 0.695867
## NeighborhoodNoRidge    7.527e-02  4.812e-02   1.564 0.118224
## NeighborhoodNPkVill   -1.720e-02  6.887e-02  -0.250 0.802851
## NeighborhoodNridgHt   9.177e-02  4.583e-02   2.002 0.045583 *  
## NeighborhoodNWAmes   -7.167e-03  4.765e-02  -0.150 0.880490    
## NeighborhoodOldTown  -1.393e-01  4.738e-02  -2.941 0.003372 ** 
## NeighborhoodSawyer   -1.970e-02  4.734e-02  -0.416 0.677382    
## NeighborhoodSawyerW  -6.405e-02  4.661e-02  -1.374 0.169787    
## NeighborhoodSomerst   5.304e-02  4.472e-02   1.186 0.235970    
## NeighborhoodStoneBr   1.030e-01  5.241e-02   1.965 0.049831 *  
## NeighborhoodSWISU    -8.037e-02  5.660e-02  -1.420 0.156026    
## NeighborhoodTimber    4.252e-02  5.026e-02   0.846 0.397813    
## NeighborhoodVeenker   5.226e-02  5.773e-02   0.905 0.365573    
## House.Style1.5Unf    -7.159e-03  4.800e-02  -0.149 0.881480    
## House.Style1Story    -9.025e-04  2.363e-02  -0.038 0.969544    
## House.Style2.5Unf    -4.915e-02  3.822e-02  -1.286 0.198774    
## House.Style2Story     2.588e-02  1.775e-02   1.458 0.145215    
## House.StyleSFoyer     9.718e-02  3.061e-02   3.175 0.001559 ** 
## House.StyleSLvl       3.540e-02  2.685e-02   1.318 0.187853    
## Overall.Qual          6.977e-02  5.257e-03  13.272  < 2e-16 ***
## Year.Remod.Add        2.616e-03  2.679e-04   9.766  < 2e-16 ***
## FunctionalMaj2        1.474e-01  1.235e-01   1.193 0.233053    
## FunctionalMin1        1.565e-01  6.225e-02   2.513 0.012159 *  
## FunctionalMin2        1.387e-01  6.041e-02   2.296 0.021936 *  
## FunctionalMod         9.848e-02  6.294e-02   1.565 0.118064    
## FunctionalTyp         2.079e-01  5.605e-02   3.709 0.000223 ***
## Heating.QCFa         -1.021e-01  2.802e-02  -3.644 0.000287 ***
## Heating.QCGd         -1.300e-02  1.179e-02  -1.102 0.270838    
## Heating.QCPo         -9.076e-02  1.174e-01  -0.773 0.439697    
## Heating.QCTA         -3.821e-02  1.110e-02  -3.442 0.000608 ***
## Total.Bsmt.SF         1.609e-04  1.827e-05   8.807  < 2e-16 ***
## Full.Bath            -2.184e-03  1.136e-02  -0.192 0.847622    
## Bedroom.AbvGr        -1.261e-02  7.312e-03  -1.725 0.084928 .  
## Fireplaces            3.869e-02  7.561e-03   5.117 3.93e-07 ***
## Garage.Area           1.343e-04  2.854e-05   4.708 2.97e-06 ***
## Garage.QualFa        -9.410e-02  1.139e-01  -0.826 0.409131    
## Garage.QualGd        -8.307e-02  1.226e-01  -0.677 0.498336    
## Garage.QualPo        -3.332e-01  1.304e-01  -2.556 0.010785 *  
## Garage.QualTA        -7.947e-02  1.119e-01  -0.710 0.477663    
## Garage.QualNo garage -1.188e-01  1.151e-01  -1.032 0.302490    
## X1st.Flr.SF           1.063e-06  3.626e-05   0.029 0.976616    
## Paved.DriveP         -3.973e-02  2.607e-02  -1.524 0.128037    
## Paved.DriveY          3.306e-02  1.924e-02   1.718 0.086215 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1063 on 769 degrees of freedom
## Multiple R-squared:  0.9272, Adjusted R-squared:  0.9213 
## F-statistic: 157.9 on 62 and 769 DF,  p-value: < 2.2e-16
```

```
BIC(final.model)
```

```
## [1] -1004.033
```

Section 3.2 Transformation

Before fitting the model several variables were transformed (see the code above).

Log transformation was applied to numerical variables `area` and `Lot.Area` to achieve more normal distribution.

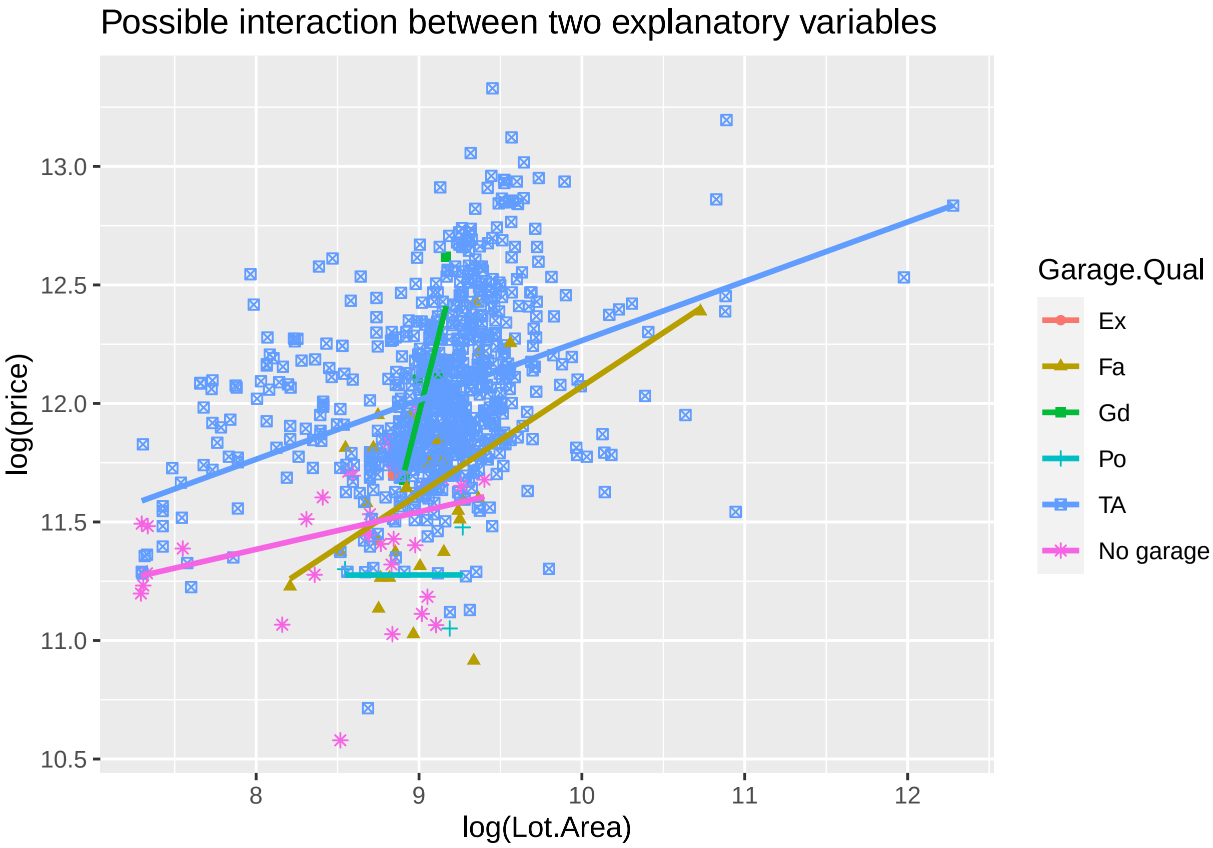
At the same time the NAs in the `Garage.Qual` variable were converted to a category since they represent the absence of a garage in a property.

Section 3.3 Variable Interaction

After visually examining the interaction between several explanatory variables (mostly between a categorical and a continuous variable) we have come to the conclusion that there are no significant interaction effects to consider in this model.

```
# check interaction between ‘Lot.Area‘ and ‘Garage.Qual‘
ggplot(data = ames_train, aes(x = log(Lot.Area), y = log(price), color=Garage.Qual, shape=Garage.Qual))
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Possible interaction between two explanatory variables")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



As an example, the above figure shows that when breaking down by `Garage.Qual` variable there are no significant interaction effects between this variable and `Lot.Area` as the trends are clearly not parallel.

Section 3.4 Variable Selection

For the purpose of variable selection we will use the BIC backwards selection method that increases the model fit, at the same time, introducing penalty for adding new parameters.

```r
# BIC final model
final.model.BIC <- stepAIC(final.model,
                     scale = 0,
                     direction = c("backward"),
                     trace = 1,
                     keep = NULL,
                     steps = 1000,
                     use.start = FALSE,
                     k = log(nrow(ames_train)))
```

```
## Start:  AIC=-3371.87
## log(price) ~ log(area) + log(Lot.Area) + Lot.Config + Neighborhood + 
##     House.Style + Overall.Qual + Year.Remod.Add + Functional + 
##     Heating.QC + Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr + 
##     Fireplaces + Garage.Area + Garage.Qual + X1st.Flr.SF + Paved.Drive
## 
##                   Df Sum of Sq     RSS     AIC
## - Lot.Config       4   0.10872  8.7969 -3388.4
## - Garage.Qual      5   0.19006  8.8783 -3387.5
## - House.Style      6   0.30188  8.9901 -3383.8
## - X1st.Flr.SF      1   0.00001  8.6882 -3378.6
## - Full.Bath        1   0.00042  8.6886 -3378.6
## - Heating.QC       4   0.23555  8.9238 -3376.5
## - Bedroom.AbvGr    1   0.03362  8.7218 -3375.4
## - Neighborhood    26   2.01343 10.7016 -3373.3
## - Paved.Drive      2   0.13087  8.8191 -3372.9
## <none>                          8.6882 -3371.9
## - Functional       5   0.36931  9.0575 -3370.9
## - Garage.Area      1   0.25039  8.9386 -3355.0
## - Fireplaces       1   0.29577  8.9840 -3350.7
## - log(area)        1   0.63595  9.3242 -3319.8
## - log(Lot.Area)    1   0.68777  9.3760 -3315.2
## - Total.Bsmt.SF    1   0.87635  9.5646 -3298.6
## - Year.Remod.Add   1   1.07754  9.7658 -3281.3
## - Overall.Qual     1   1.99020 10.6784 -3207.0
## 
## Step:  AIC=-3388.42
## log(price) ~ log(area) + log(Lot.Area) + Neighborhood + House.Style + 
##     Overall.Qual + Year.Remod.Add + Functional + Heating.QC + 
##     Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr + Fireplaces + 
##     Garage.Area + Garage.Qual + X1st.Flr.SF + Paved.Drive
## 
##                   Df Sum of Sq     RSS     AIC
## - Garage.Qual     5   0.18817  8.9851 -3404.4
## - House.Style     6   0.31644  9.1134 -3399.4
## - X1st.Flr.SF     1   0.00010  8.7970 -3395.1
## - Full.Bath       1   0.00149  8.7984 -3395.0
## - Heating.QC      4   0.23178  9.0287 -3393.7
## - Neighborhood   26   2.00521 10.8021 -3392.4
## - Bedroom.AbvGr   1   0.02951  8.8264 -3392.4
## - Paved.Drive     2   0.12666  8.9236 -3390.0
## <none>                         8.7969 -3388.4
## - Functional      5   0.36904  9.1660 -3387.8
## - Garage.Area     1   0.24907  9.0460 -3371.9
## - Fireplaces      1   0.32825  9.1252 -3364.7
## - log(area)       1   0.65871  9.4556 -3335.1
## - log(Lot.Area)   1   0.68331  9.4802 -3332.9
## - Total.Bsmt.SF   1   0.91975  9.7167 -3312.4
## - Year.Remod.Add  1   1.07483  9.8718 -3299.2
## - Overall.Qual    1   1.99718 10.7941 -3224.9
## 
## Step:  AIC=-3404.43
## log(price) ~ log(area) + log(Lot.Area) + Neighborhood + House.Style + 
##     Overall.Qual + Year.Remod.Add + Functional + Heating.QC + 
##     Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr + Fireplaces + 
##     Garage.Area + X1st.Flr.SF + Paved.Drive
## 
##                  Df Sum of Sq     RSS     AIC
## - Neighborhood   26   1.93898 10.9241 -3416.7
## - House.Style     6   0.31673  9.3018 -3415.9
## - X1st.Flr.SF     1   0.00127  8.9864 -3411.0
## - Full.Bath       1   0.00392  8.9890 -3410.8
## - Bedroom.AbvGr   1   0.03433  9.0194 -3408.0
## - Functional      5   0.33993  9.3250 -3407.2
## <none>                         8.9851 -3404.4
## - Heating.QC      4   0.30410  9.2892 -3403.6
## - Paved.Drive     2   0.16582  9.1509 -3402.7
## - Fireplaces      1   0.36730  9.3524 -3377.8
## - Garage.Area     1   0.42562  9.4107 -3372.6
## - log(Lot.Area)   1   0.67772  9.6628 -3350.7
## - log(area)       1   0.71549  9.7006 -3347.4
## - Total.Bsmt.SF   1   0.89998  9.8851 -3331.7
## - Year.Remod.Add  1   1.03875 10.0238 -3320.1
## - Overall.Qual    1   2.06089 11.0460 -3239.3
## 
## Step:  AIC=-3416.68
## log(price) ~ log(area) + log(Lot.Area) + House.Style + Overall.Qual + 
##     Year.Remod.Add + Functional + Heating.QC + Total.Bsmt.SF + 
##     Full.Bath + Bedroom.AbvGr + Fireplaces + Garage.Area + X1st.Flr.SF + 
##     Paved.Drive
## 
##                  Df Sum of Sq    RSS     AIC
## - Functional      5    0.3350 11.259 -3425.2
## - House.Style     6    0.4309 11.355 -3424.8
## - Full.Bath       1    0.0015 10.926 -3423.3
## - X1st.Flr.SF     1    0.0017 10.926 -3423.3
## <none>                        10.924 -3416.7
## - Bedroom.AbvGr   1    0.1257 11.050 -3413.9
## - Heating.QC      4    0.4318 11.356 -3411.3
## - Paved.Drive     2    0.5699 11.494 -3387.8
## - Garage.Area     1    0.5577 11.482 -3382.0
## - Fireplaces      1    0.5667 11.491 -3381.3
## - log(area)       1    0.8723 11.796 -3359.5
## - Total.Bsmt.SF   1    0.9183 11.842 -3356.2
## - Year.Remod.Add  1    1.1682 12.092 -3338.9
## - log(Lot.Area)   1    1.6159 12.540 -3308.6
## - Overall.Qual    1    3.9884 14.912 -3164.5
##
## Step:  AIC=-3425.17
## log(price) ~ log(area) + log(Lot.Area) + House.Style + Overall.Qual +
##     Year.Remod.Add + Heating.QC + Total.Bsmt.SF + Full.Bath +
##     Bedroom.AbvGr + Fireplaces + Garage.Area + X1st.Flr.SF +
##     Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - House.Style     6    0.4542 11.713 -3432.6
## - Full.Bath       1    0.0005 11.259 -3431.9
## - X1st.Flr.SF     1    0.0110 11.270 -3431.1
## - Bedroom.AbvGr   1    0.0755 11.335 -3426.3
## <none>                        11.259 -3425.2
## - Heating.QC      4    0.4415 11.700 -3420.1
## - Garage.Area     1    0.5520 11.811 -3392.1
## - Paved.Drive     2    0.6643 11.923 -3390.9
## - Fireplaces      1    0.5958 11.855 -3389.0
## - log(area)       1    0.7838 12.043 -3375.9
## - Total.Bsmt.SF   1    1.1609 12.420 -3350.2
## - Year.Remod.Add  1    1.1774 12.436 -3349.1
## - log(Lot.Area)   1    1.5693 12.828 -3323.3
## - Overall.Qual    1    4.6622 15.921 -3143.6
##
## Step:  AIC=-3432.61
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Heating.QC + Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Full.Bath       1    0.0001 11.713 -3439.3
## - X1st.Flr.SF     1    0.0026 11.716 -3439.1
## <none>                        11.713 -3432.6
## - Heating.QC      4    0.3865 12.100 -3432.5
## - Bedroom.AbvGr   1    0.1114 11.825 -3431.5
## - Fireplaces      1    0.6378 12.351 -3395.2
## - Garage.Area     1    0.7756 12.489 -3386.0
## - Paved.Drive     2    0.9527 12.666 -3381.0
## - Total.Bsmt.SF   1    1.1261 12.839 -3363.0
## - Year.Remod.Add  1    1.5031 13.216 -3338.9
## - log(Lot.Area)   1    1.5784 13.292 -3334.2
## - log(area)       1    2.0821 13.795 -3303.2
## - Overall.Qual    1    4.6361 16.349 -3161.9
##
## Step:  AIC=-3439.33
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr + Fireplaces +
##     Garage.Area + X1st.Flr.SF + Paved.Drive
##
##                   Df Sum of Sq    RSS     AIC
## - X1st.Flr.SF      1    0.0026 11.716 -3445.9
## <none>                         11.713 -3439.3
## - Heating.QC       4    0.3873 12.101 -3439.2
## - Bedroom.AbvGr    1    0.1129 11.826 -3438.1
## - Fireplaces       1    0.6379 12.351 -3401.9
## - Garage.Area      1    0.7794 12.493 -3392.5
## - Paved.Drive      2    0.9542 12.668 -3387.6
## - Total.Bsmt.SF    1    1.1281 12.841 -3369.5
## - Year.Remod.Add  1    1.5653 13.279 -3341.7
## - log(Lot.Area)    1    1.5914 13.305 -3340.1
## - log(area)        1    2.3040 14.017 -3296.6
## - Overall.Qual     1    4.6764 16.390 -3166.6
##
## Step:  AIC=-3445.87
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr + Fireplaces +
##     Garage.Area + Paved.Drive
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         11.716 -3445.9
## - Bedroom.AbvGr    1    0.1103 11.826 -3444.8
## - Heating.QC       4    0.4039 12.120 -3444.6
## - Fireplaces       1    0.6360 12.352 -3408.6
## - Garage.Area      1    0.7773 12.493 -3399.1
## - Paved.Drive      2    0.9569 12.673 -3394.0
## - Year.Remod.Add  1    1.5627 13.279 -3348.4
## - log(Lot.Area)    1    1.6392 13.355 -3343.6
## - Total.Bsmt.SF    1    2.0947 13.810 -3315.7
## - log(area)        1    2.4628 14.179 -3293.9
## - Overall.Qual     1    4.7291 16.445 -3170.5
```

```
summary(final.model.BIC)
```

```
##
## Call:
## lm(formula = log(price) ~ log(area) + log(Lot.Area) + Overall.Qual +
##     Year.Remod.Add + Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + Paved.Drive, data = ames_train,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53899 -0.06877  0.00260  0.07619  0.45768
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.423e+00  5.361e-01   4.521 7.08e-06 ***
## log(area)       3.102e-01  2.367e-02  13.105  < 2e-16 ***
## log(Lot.Area)   9.895e-02  9.255e-03  10.692  < 2e-16 ***
## Overall.Qual    9.207e-02  5.070e-03  18.160  < 2e-16 ***
## Year.Remod.Add  2.806e-03  2.688e-04  10.439  < 2e-16 ***
## Heating.QCFa   -1.073e-01  3.020e-02  -3.554 0.000401 ***
## Heating.QCGd   -7.432e-03  1.242e-02  -0.598 0.549695
## Heating.QCPo   -1.658e-01  1.206e-01  -1.375 0.169630
## Heating.QCTA   -4.784e-02  1.133e-02  -4.222 2.69e-05 ***
## Total.Bsmt.SF   1.632e-04  1.350e-05  12.086  < 2e-16 ***
## Bedroom.AbvGr  -1.975e-02  7.122e-03  -2.773 0.005673 **
## Fireplaces      5.170e-02  7.763e-03   6.660 5.04e-11 ***
## Garage.Area     1.920e-04  2.608e-05   7.362 4.42e-13 ***
## Paved.DriveP   -5.773e-03  2.777e-02  -0.208 0.835351
## Paved.DriveY    1.182e-01  1.747e-02   6.767 2.50e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1198 on 817 degrees of freedom
## Multiple R-squared:  0.9018, Adjusted R-squared:  0.9001
## F-statistic: 535.8 on 14 and 817 DF,  p-value: < 2.2e-16
```

```r
BIC(final.model.BIC)
```

```
## [1] -1078.03
```

The BIC model selection left 10 variables `log(area)`, `log(Lot.Area)`, `Overall.Qual`, `Year.Remod.Add`, `Heating.QC`, `Total.Bsmt.SF`, `Bedroom.AbvGr`, `Fireplaces`, `Garage.Area`, `Paved.Drive`.

The Adjusted R-squared now explains less variability while the BIC improved.

Section 3.5 Model Testing

Based on the out-of-sample data, we have not changed the model since although the RMSE increased, the coverage probability shows a relatively high performance of the model.

Possible changes will be made when creating and examining the residuals plot.

```r
# dropping 2 observations from House.Style to test the model
ames_test <- subset(ames_test, House.Style != "2.5Fin")

# calculating RMSE using training data
predict.final.BIC.train <- exp(predict(final.model.BIC, ames_train))
resid.final.BIC.train <- ames_train$price - predict.final.BIC.train
rmse.final.BIC.train <- sqrt(mean(resid.final.BIC.train^2))
rmse.final.BIC.train
```

```
## [1] 21976.79
```

```r
# calculating RMSE using test data
predict.final.BIC.test <- exp(predict(final.model.BIC, ames_test))
resid.final.BIC.test <- ames_test$price - predict.final.BIC.test
rmse.final.BIC.test <- sqrt(mean(resid.final.BIC.test^2))
rmse.final.BIC.test
```

```
## [1] 23542.26
```

```
# predicting prices
predict.final.BIC.test <- exp(predict(final.model.BIC, ames_test, interval = "prediction"))

# calculating proportion of observations that fall within prediction intervals
coverage.prob.final.BIC.test <- mean(ames_test$price > predict.final.BIC.test[,"lwr"] &
                            ames_test$price < predict.final.BIC.test[,"upr"])
coverage.prob.final.BIC.test
```

```
## [1] 0.9398034
```

```
# proportion of observations (rows) in ‘ames_test‘ have sales prices that fall outside the prediction i
1-coverage.prob.final.BIC.test
```

```
## [1] 0.06019656
```

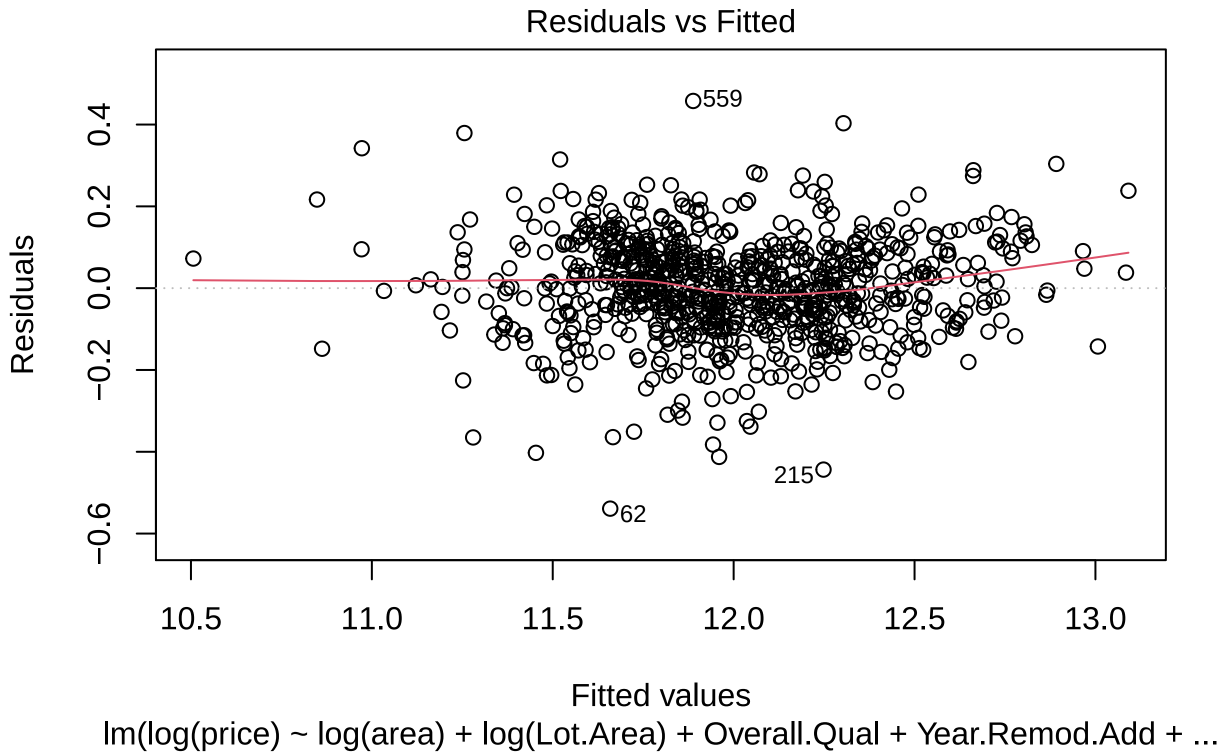
Part 4 Final Model Assessment

Section 4.1 Final Model Residuals

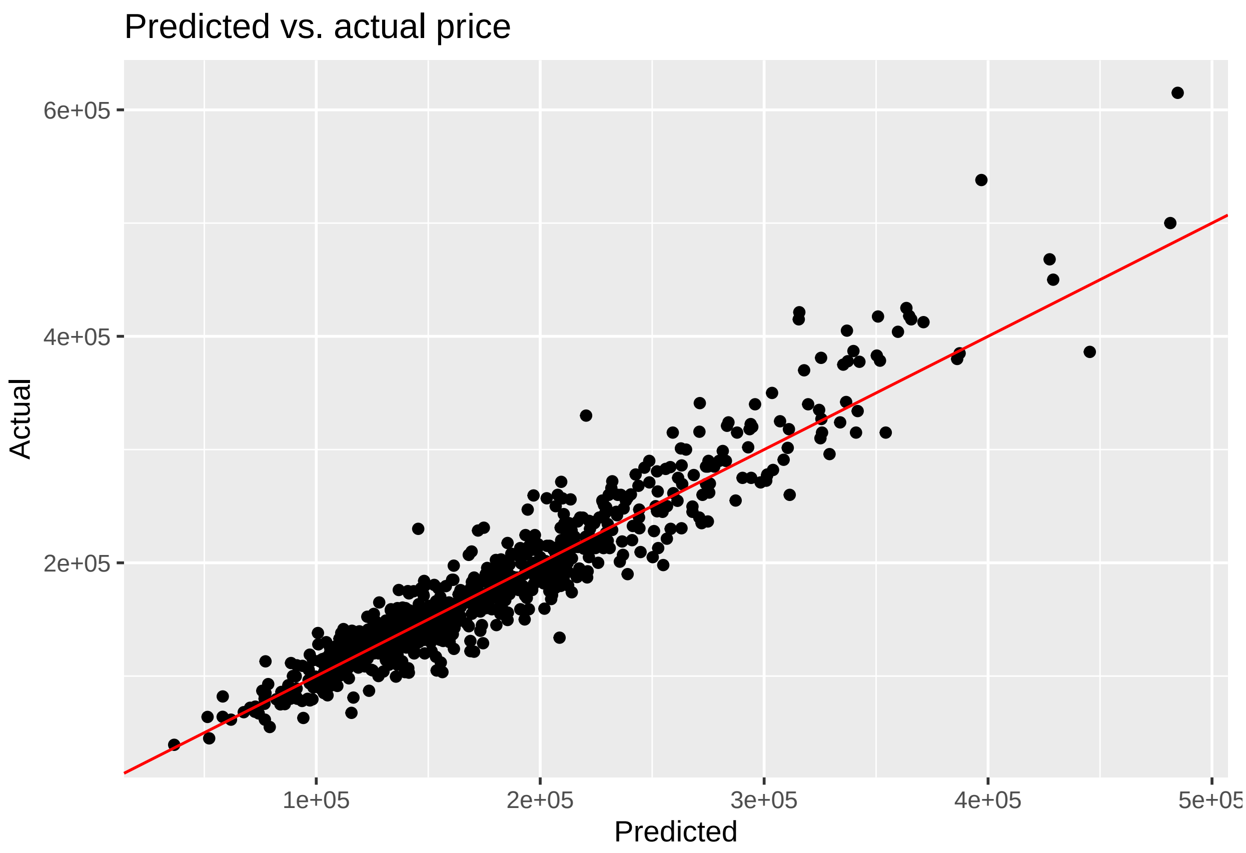
On average, residuals are normally distributed with several outliers that affect the fit of the model.

As the next step, we will try to fit the data without the outliers.

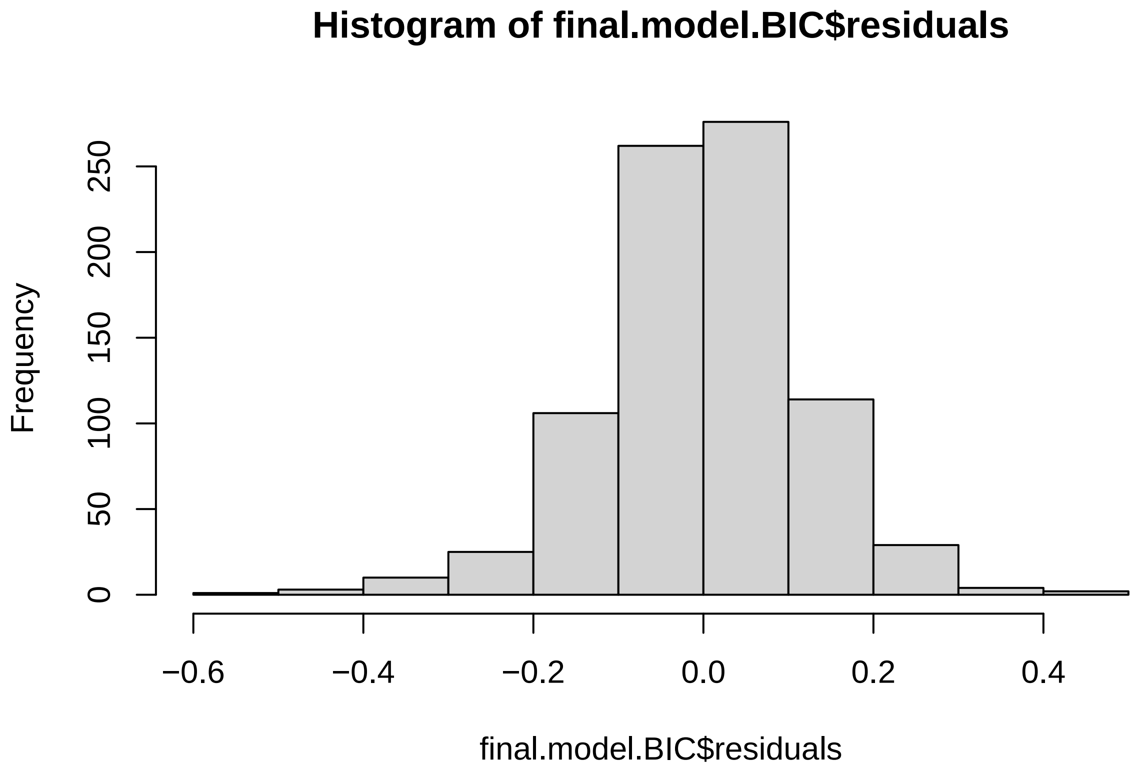
```
# residuals vs. fitted
plot(final.model.BIC, which = 1)
```



```r
# predicted vs. actual
ames_train$prediction <- predict(final.model.BIC)
ames_train$prediction = exp(ames_train$prediction)

ggplot(ames_train, aes(x = prediction, y = price)) +
geom_point() +
geom_abline(color = "red") +
labs(title="Predicted vs. actual price", x="Predicted", y="Actual")
```

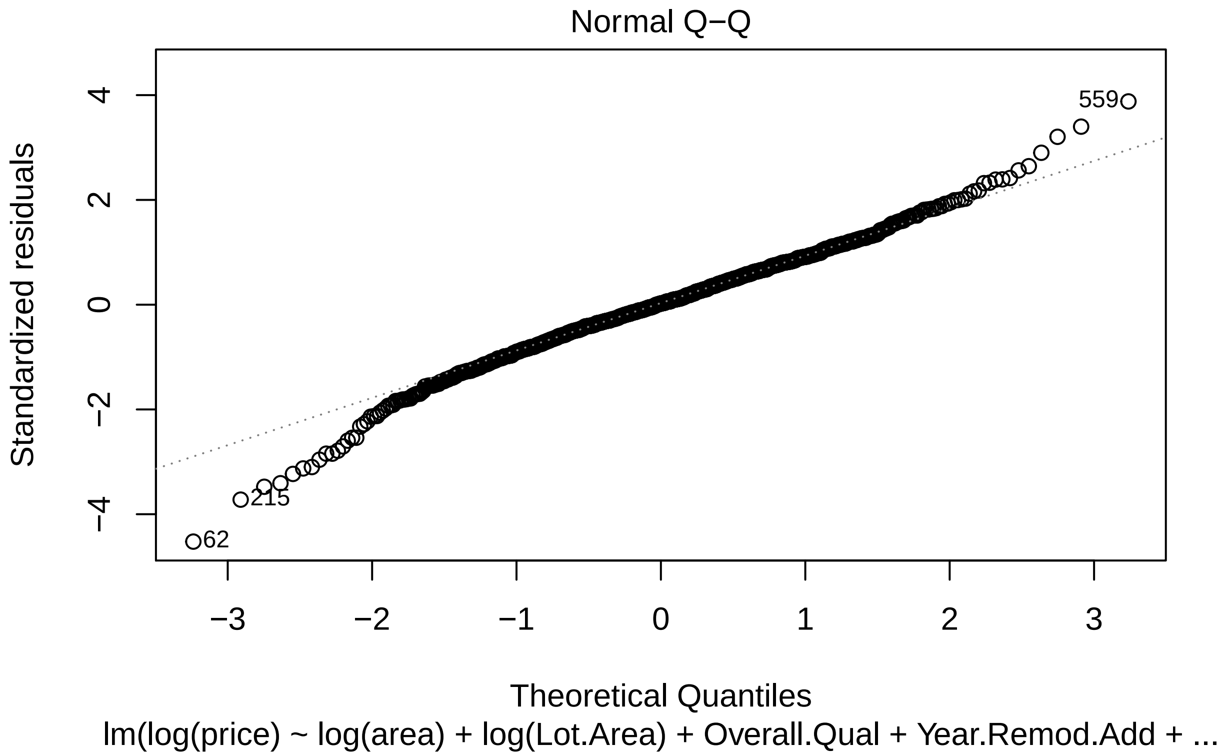


```r
# residuals distribution
hist(final.model.BIC$residuals)
```

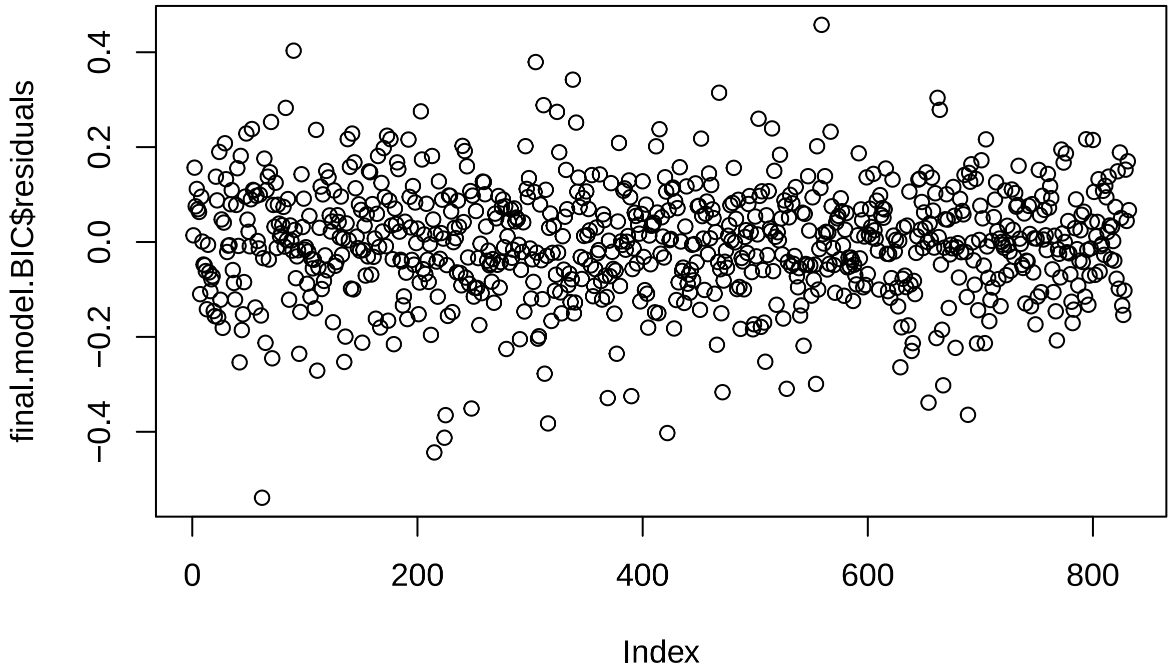


```r
plot(final.model.BIC, which = 2)
```

```
## Warning: not plotting observations with leverage one:
##   653
```



```
# independence of residuals
plot(final.model.BIC$residuals)
```



Data transformation post residual analysis.

```
# deleting outliers
ames_train <- ames_train[-c(62,215,559),]

# new initial model
final.model.2 <- lm(fmla, ames_train, na.action = na.omit)

# new BIC model
final.model.BIC.2 <- stepAIC(final.model.2,
                         scale = 0,
                         direction = c("backward"),
                         trace = 1,
                         keep = NULL,
                         steps = 1000,
                         use.start = FALSE,
                         k = log(nrow(ames_train)))
```

```
## Start:  AIC=-3389.71
## log(price) ~ log(area) + log(Lot.Area) + Lot.Config + Neighborhood +
##     House.Style + Overall.Qual + Year.Remod.Add + Functional +
##     Heating.QC + Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + Garage.Qual + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq     RSS     AIC
## - Neighborhood   26   1.73696 10.0731 -3407.5
## - Lot.Config      4   0.09864  8.4348 -3406.8
## - Garage.Qual     5   0.19087  8.5270 -3404.5
## - House.Style     6   0.26225  8.5984 -3404.4
## - X1st.Flr.SF     1   0.00002  8.3362 -3396.4
## - Full.Bath       1   0.00216  8.3383 -3396.2
## - Heating.QC      4   0.22913  8.5653 -3394.1
## - Bedroom.AbvGr   1   0.04147  8.3776 -3392.3
## <none>                         8.3361 -3389.7
## - Paved.Drive     2   0.14151  8.4777 -3389.2
## - Functional      5   0.38139  8.7175 -3386.2
## - Garage.Area     1   0.24739  8.5835 -3372.2
## - Fireplaces      1   0.28601  8.6222 -3368.5
## - log(area)       1   0.67092  9.0071 -3332.3
## - log(Lot.Area)   1   0.67185  9.0080 -3332.2
## - Total.Bsmt.SF   1   0.85330  9.1895 -3315.6
## - Year.Remod.Add  1   1.02918  9.3653 -3299.9
## - Overall.Qual    1   1.90064 10.2368 -3226.2
##
## Step:  AIC=-3407.53
## log(price) ~ log(area) + log(Lot.Area) + Lot.Config + House.Style +
##     Overall.Qual + Year.Remod.Add + Functional + Heating.QC +
##     Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr + Fireplaces +
##     Garage.Area + Garage.Qual + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Garage.Qual     5    0.1270 10.200 -3430.7
## - Lot.Config      4    0.0753 10.148 -3428.2
## - House.Style     6    0.3281 10.401 -3421.3
## - Full.Bath       1    0.0006 10.074 -3414.2
## - X1st.Flr.SF     1    0.0023 10.075 -3414.1
## - Functional      5    0.3764 10.450 -3410.7
## <none>                        10.073 -3407.5
## - Heating.QC      4    0.3557 10.429 -3405.6
## - Bedroom.AbvGr   1    0.1417 10.215 -3402.7
## - Garage.Area     1    0.3572 10.430 -3385.4
## - Paved.Drive     2    0.4467 10.520 -3385.0
## - Fireplaces      1    0.5020 10.575 -3373.9
## - log(area)       1    0.8486 10.922 -3347.2
## - Total.Bsmt.SF   1    1.0323 11.105 -3333.4
## - Year.Remod.Add  1    1.1656 11.239 -3323.5
## - log(Lot.Area)   1    1.5134 11.586 -3298.2
## - Overall.Qual    1    3.5872 13.660 -3161.7
##
## Step:  AIC=-3430.74
## log(price) ~ log(area) + log(Lot.Area) + Lot.Config + House.Style +
##     Overall.Qual + Year.Remod.Add + Functional + Heating.QC +
##     Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr + Fireplaces +
##     Garage.Area + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Lot.Config      4    0.0745 10.275 -3451.6
## - House.Style     6    0.3485 10.549 -3443.2
## - Full.Bath       1    0.0033 10.204 -3437.2
## - X1st.Flr.SF     1    0.0059 10.206 -3437.0
## - Functional      5    0.3521 10.552 -3436.2
## <none>                        10.200 -3430.7
## - Bedroom.AbvGr   1    0.1439 10.344 -3425.9
## - Heating.QC      4    0.4011 10.601 -3425.6
## - Paved.Drive     2    0.5814 10.781 -3398.2
## - Fireplaces      1    0.5349 10.735 -3395.1
## - Garage.Area     1    0.6070 10.807 -3389.5
## - log(area)       1    0.9075 11.108 -3366.8
## - Total.Bsmt.SF   1    1.0108 11.211 -3359.1
## - Year.Remod.Add  1    1.1501 11.350 -3348.9
## - log(Lot.Area)   1    1.5322 11.732 -3321.4
## - Overall.Qual    1    3.6772 13.877 -3182.3
##
## Step:  AIC=-3451.59
## log(price) ~ log(area) + log(Lot.Area) + House.Style + Overall.Qual +
##     Year.Remod.Add + Functional + Heating.QC + Total.Bsmt.SF +
##     Full.Bath + Bedroom.AbvGr + Fireplaces + Garage.Area + X1st.Flr.SF +
##     Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - House.Style     6    0.3609 10.636 -3463.3
## - Functional      5    0.3425 10.617 -3458.0
## - Full.Bath       1    0.0045 10.279 -3457.9
## - X1st.Flr.SF     1    0.0073 10.282 -3457.7
## <none>                        10.275 -3451.6
## - Bedroom.AbvGr   1    0.1340 10.409 -3447.6
## - Heating.QC      4    0.3997 10.674 -3446.8
## - Paved.Drive     2    0.5757 10.850 -3419.8
## - Fireplaces      1    0.5743 10.849 -3413.2
## - Garage.Area     1    0.6011 10.876 -3411.2
## - log(area)       1    0.9121 11.187 -3387.8
## - Total.Bsmt.SF   1    1.0452 11.320 -3378.0
## - Year.Remod.Add  1    1.1766 11.451 -3368.4
## - log(Lot.Area)   1    1.5746 11.849 -3340.1
## - Overall.Qual    1    3.6975 13.972 -3203.5
##
## Step:  AIC=-3463.29
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Functional + Heating.QC + Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + X1st.Flr.SF + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - X1st.Flr.SF     1    0.0005 10.636 -3470.0
## - Full.Bath       1    0.0016 10.637 -3469.9
## - Functional      5    0.3580 10.994 -3469.4
## <none>                        10.636 -3463.3
## - Heating.QC      4    0.3687 11.004 -3461.9
## - Bedroom.AbvGr   1    0.1667 10.802 -3457.1
## - Fireplaces      1    0.6279 11.263 -3422.5
## - Paved.Drive     2    0.8180 11.454 -3415.3
## - Garage.Area     1    0.8216 11.457 -3408.3
## - Total.Bsmt.SF   1    0.9905 11.626 -3396.2
## - Year.Remod.Add  1    1.4461 12.082 -3364.3
## - log(Lot.Area)   1    1.5810 12.217 -3355.1
## - log(area)       1    2.4562 13.092 -3297.8
## - Overall.Qual    1    3.6885 14.324 -3223.2
##
## Step:  AIC=-3469.97
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Functional + Heating.QC + Total.Bsmt.SF + Full.Bath + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Full.Bath       1    0.0016 10.638 -3476.6
## - Functional      5    0.3729 11.009 -3475.0
## <none>                        10.636 -3470.0
## - Heating.QC      4    0.3811 11.017 -3467.7
## - Bedroom.AbvGr   1    0.1665 10.803 -3463.8
## - Fireplaces      1    0.6311 11.267 -3428.9
## - Paved.Drive     2    0.8188 11.455 -3421.9
## - Garage.Area     1    0.8234 11.460 -3414.9
## - Year.Remod.Add  1    1.4468 12.083 -3371.0
## - log(Lot.Area)   1    1.6309 12.267 -3358.4
## - Total.Bsmt.SF   1    2.0232 12.659 -3332.3
## - log(area)       1    2.5575 13.194 -3298.1
## - Overall.Qual    1    3.7023 14.338 -3229.1
##
## Step:  AIC=-3476.57
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Functional + Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## - Functional      5    0.3718 11.009 -3481.7
## <none>                        10.638 -3476.6
## - Heating.QC      4    0.3799 11.018 -3474.4
## - Bedroom.AbvGr   1    0.1740 10.812 -3469.8
## - Fireplaces      1    0.6331 11.271 -3435.4
## - Paved.Drive     2    0.8241 11.462 -3428.2
## - Garage.Area     1    0.8221 11.460 -3421.6
## - Year.Remod.Add  1    1.4831 12.121 -3375.1
## - log(Lot.Area)   1    1.6584 12.296 -3363.2
## - Total.Bsmt.SF   1    2.0230 12.661 -3339.0
## - log(area)       1    2.7961 13.434 -3289.8
## - Overall.Qual    1    3.7105 14.348 -3235.2
##
## Step:  AIC=-3481.69
## log(price) ~ log(area) + log(Lot.Area) + Overall.Qual + Year.Remod.Add +
##     Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr + Fireplaces +
##     Garage.Area + Paved.Drive
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        11.009 -3481.7
## - Bedroom.AbvGr   1    0.1021 11.112 -3480.8
## - Heating.QC      4    0.3964 11.406 -3479.2
## - Fireplaces      1    0.6439 11.653 -3441.3
## - Garage.Area     1    0.8024 11.812 -3430.1
## - Paved.Drive     2    0.9553 11.965 -3426.1
## - Year.Remod.Add  1    1.5219 12.531 -3381.1
## - log(Lot.Area)   1    1.5521 12.562 -3379.1
## - Total.Bsmt.SF   1    2.2252 13.235 -3335.8
## - log(area)       1    2.4981 13.508 -3318.9
## - Overall.Qual    1    4.4400 15.450 -3207.5
```

```
summary(final.model.BIC.2)
```

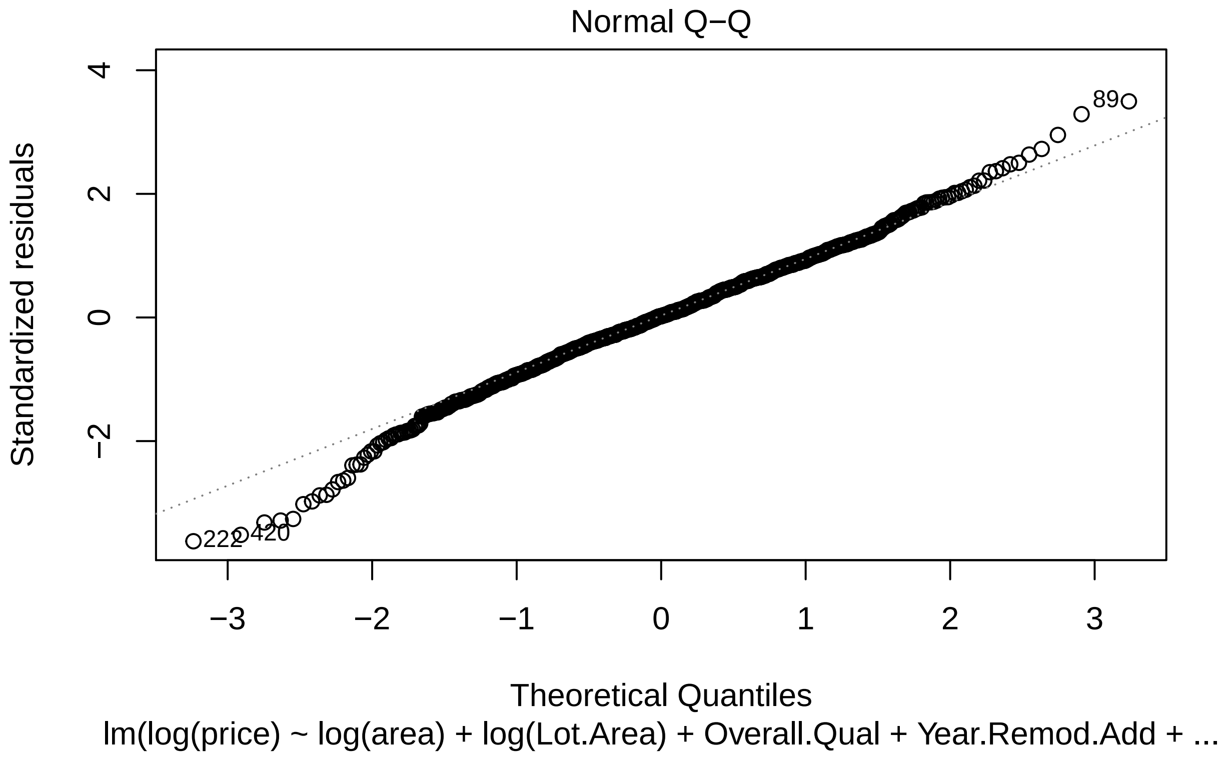
```
##
## Call:
## lm(formula = log(price) ~ log(area) + log(Lot.Area) + Overall.Qual +
##     Year.Remod.Add + Heating.QC + Total.Bsmt.SF + Bedroom.AbvGr +
##     Fireplaces + Garage.Area + Paved.Drive, data = ames_train,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41724 -0.06783  0.00222  0.07504  0.40292
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.497e+00  5.212e-01   4.792 1.96e-06 ***
## log(area)       3.129e-01  2.302e-02  13.591  < 2e-16 ***
## log(Lot.Area)   9.655e-02  9.013e-03  10.712  < 2e-16 ***
## Overall.Qual    8.964e-02  4.947e-03  18.118  < 2e-16 ***
## Year.Remod.Add  2.773e-03  2.615e-04  10.608  < 2e-16 ***
## Heating.QCFa   -1.093e-01  2.933e-02  -3.728 0.000207 ***
## Heating.QCGd   -1.293e-02  1.210e-02  -1.069 0.285530
## Heating.QCPo   -1.680e-01  1.171e-01  -1.434 0.151957
## Heating.QCTA   -4.847e-02  1.101e-02  -4.402 1.22e-05 ***
## Total.Bsmt.SF   1.696e-04  1.322e-05  12.827  < 2e-16 ***
## Bedroom.AbvGr  -1.906e-02  6.936e-03  -2.748 0.006125 **
## Fireplaces      5.211e-02  7.552e-03   6.900 1.05e-11 ***
## Garage.Area     1.952e-04  2.534e-05   7.702 3.88e-14 ***
## Paved.DriveP   -5.710e-03  2.697e-02  -0.212 0.832348
## Paved.DriveY    1.183e-01  1.698e-02   6.966 6.73e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1163 on 814 degrees of freedom
## Multiple R-squared:  0.907,  Adjusted R-squared:  0.9054
## F-statistic: 566.9 on 14 and 814 DF,  p-value: < 2.2e-16
```

```
BIC(final.model.BIC.2)
```

```
## [1] -1122.368
```

```
# checking the residuals
plot(final.model.BIC.2, which = 2)
```

```
## Warning: not plotting observations with leverage one:
##   650
```



```r
# calculating RMSE using test data with ‘final.model.BIC.2‘
predict.final.BIC.2.test <- exp(predict(final.model.BIC.2, ames_test))
resid.final.BIC.2.test <- ames_test$price - predict.final.BIC.2.test
rmse.final.BIC.2.test <- sqrt(mean(resid.final.BIC.2.test^2))
rmse.final.BIC.2.test
```

```
## [1] 23450.57
```

As we can see, these transformations did not improve the BIC and left the Adjusted R-Squared almost the same. RMSE remained almost the same. We will continue with the `final.model.BIC`.

Section 4.2 Final Model RMSE

The RMSE, that is the standard deviation of the residuals, is 23542.26 which is slightly worse than the RMSE calculated using the training data.

```r
# calculating RMSE using test data
predict.final.BIC.test <- exp(predict(final.model.BIC, ames_test))
resid.final.BIC.test <- ames_test$price - predict.final.BIC.test
rmse.final.BIC.test <- sqrt(mean(resid.final.BIC.test^2))
rmse.final.BIC.test
```

```
## [1] 23542.26
```

Section 4.3 Final Model Evaluation

The strength of the model is that it is able to explain over 90% of the variability in the data and performs well with the test data. In addition, only 6 per cent of sales prices in the test data set fall outside the prediction intervals (94 per cent coverage probability).

Weaknesses include possible overfitting (we will examine that during model validation). In addition, not all the important predictors are included in the model.

Section 4.4 Final Model Validation

Testing the final model on a separate, validation data set.

```
# loading the validation data set
load("ames_validation.Rdata")
```

Validation results are as follows.

- The RMSE of the final model when applied to the validation data is 22012.34.
- Validation data RMSE is above the training data RMSE (21976.79) and below the test data RMSE (23542.26).
- 95% of the 95% predictive confidence intervals contain the true price of the house in the validation data set.
- Given better validation results (compared to test results) in terms of RMSE and coverage probability, we can conclude that the final model reflects the uncertainty properly.

```
# calculating RMSE using validation data
predict.final.BIC.val <- exp(predict(final.model.BIC, ames_validation))
resid.final.BIC.val <- ames_validation$price - predict.final.BIC.val
rmse.final.BIC.val <- sqrt(mean(resid.final.BIC.val^2))
rmse.final.BIC.val
```

```
## [1] 22012.34
```

```
# predicting prices
predict.final.BIC.val <- exp(predict(final.model.BIC, ames_validation, interval = "prediction"))

# calculating proportion of observations that fall within prediction intervals
coverage.prob.final.BIC.val <- mean(ames_validation$price > predict.final.BIC.val[,"lwr"] &
                               ames_validation$price < predict.final.BIC.val[,"upr"])
coverage.prob.final.BIC.val
```

```
## [1] 0.9475754
```

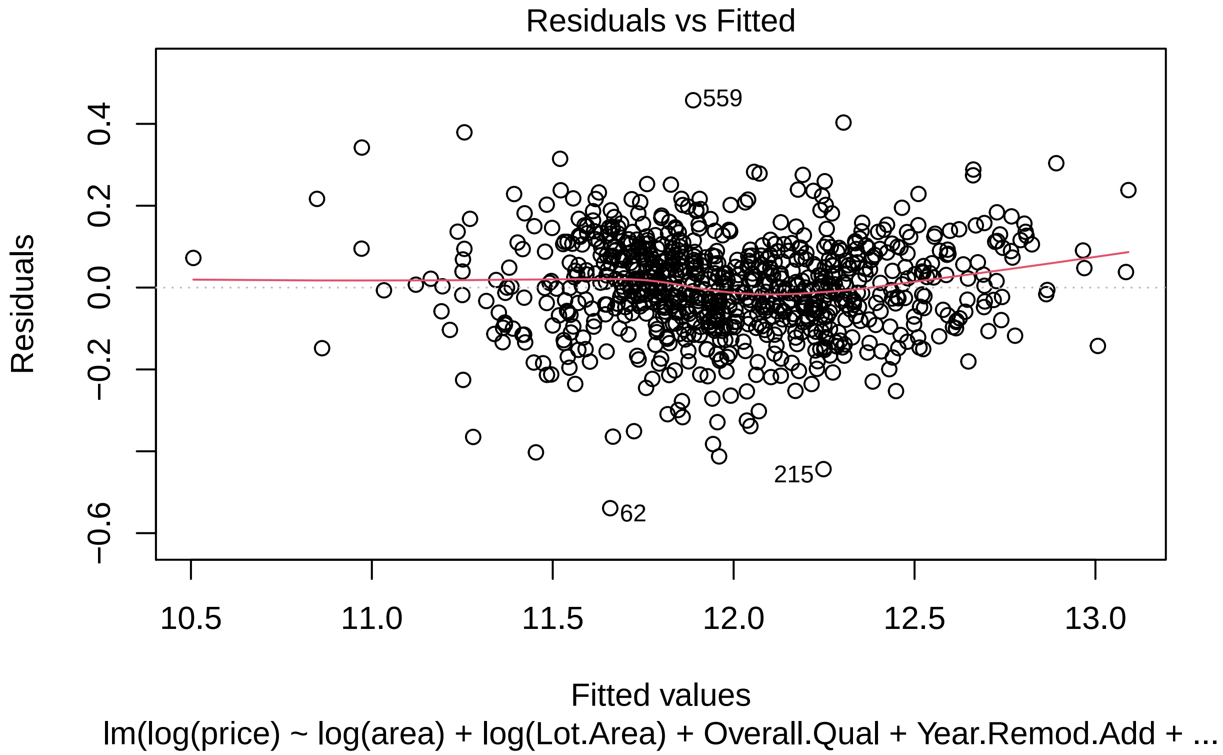
```
# proportion of observations (rows) in ‘ames_test‘ have sales prices that fall outside the prediction i
1-coverage.prob.final.BIC.val
```

```
## [1] 0.05242464
```

Section 4.5 Final Model Result

To see which properties in the validation data set are undervalued and overvalued we will need to once again examine the residuals this time with the validation data.

```
# residuals vs. fitted
plot(final.model.BIC, which = 1)
```



```
# predicted prices
ames_validation$prediction <- exp(predict(final.model.BIC, ames_validation))

# looking at the outliers
valuation.df <- ames_validation %>%
  dplyr::select(price, prediction, area, Lot.Area, Overall.Qual, Year.Remod.Add, Heating.QC,  Total.Bsm

outliers.df[c(62,215,559), ]
```

```
## # A tibble: 3 x 8
##    price prediction  area Neighborhood Overall.Qual Year.Remod.Add X1st.Flr.SF
##    <int>      <dbl> <int> <fct>               <int>          <int>       <int>
## 1  67500    116833.  1012 SawyerW                 5           1950        1012
## 2 133900    182311.  2291 OldTown                 6           1998        1260
## 3 378000    326784.  1780 NridgHt                 9           2005        1780
## # ... with 1 more variable: Paved.Drive <fct>
```

Observations 62 and 215 are significantly undervalued and may be a good investment. At the same time, observation 559 is overvalued.

Part 5 Conclusion

Despite several complications which introduce additional uncertainty brought in by the outliers and possibly insufficient number of predictors, the model performs well on training, test and validation data.

Additional research is required on properties that are undervalued and may represent a good investment.

Moreover, although the data allows to create a representative model that is relatively easy to fit observing all the criteria, there is still uncertainty around the predictions which requires us to cautiously apply the modeling results.
