DATA 605 Final Project

Ilya Kats

December 20, 2017

Contents


```
# Required libraries
library(MASS)
library(psych)
```

This project is based on data from the *House Prices* competition on Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). Full data description is available here.

Data Import and Overview

```
# Import training data
train <- read.csv('https://raw.githubusercontent.com/ilyakats/CUNY-DATA605/master/Project/train.csv')

# Get general size of the data set
dim(train)
```

```
## [1] 1460   81
```

Some variables, such as `LotArea`, should be correlated with the sale price. One variable that caught my eye as less obvious one is `LotFrontage`, linear feet of street connected to property.

```
summary(train$LotFrontage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.00   59.00   69.00   70.05   80.00  313.00     259
```

There are 259 NAs out of 1,460 observations - still a decent amount of data to consider. The numbers seem like this may work for analysis. Let us assign our X variable, `LotFrontage`, and our Y variable, `SalePrice`.

```
X <- train$LotFrontage
Y <- train$SalePrice
```

Probability

My chosen variable, `LotFrontage`, has some NA values. For this section, I have decided to remove all observations with NA.

```
probdata <- train[, c("LotFrontage", "SalePrice")]
probdata <- probdata[!is.na(probdata$LotFrontage),]

summary(probdata$LotFrontage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   59.00   69.00   70.05   80.00  313.00
```

```
summary(probdata$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  127500  159500  180800  213500  755000
```

```
# First quartile of X variable
x <- quantile(probdata$LotFrontage)[2]
# Second quartile / median of Y variable
y <- median(probdata$SalePrice)
```

```
t <- c(nrow(probdata[probdata$LotFrontage<x & probdata$SalePrice<y,]),
       nrow(probdata[probdata$LotFrontage<x & probdata$SalePrice==y,]),
       nrow(probdata[probdata$LotFrontage<x & probdata$SalePrice>y,]))
t <- rbind(t,c(nrow(probdata[probdata$LotFrontage==x & probdata$SalePrice<y,]),
               nrow(probdata[probdata$LotFrontage==x & probdata$SalePrice==y,]),
               nrow(probdata[probdata$LotFrontage==x & probdata$SalePrice>y,])))
t <- rbind(t,c(nrow(probdata[probdata$LotFrontage>x & probdata$SalePrice<y,]),
               nrow(probdata[probdata$LotFrontage>x & probdata$SalePrice==y,]),
               nrow(probdata[probdata$LotFrontage>x & probdata$SalePrice>y,])))
t <- cbind(t, t[,1] + t[,2] + t[,3])
t <- rbind(t, t[1,] + t[2,] + t[3,])
colnames(t) <- c("Y<y", "Y=y", "Y>y", "Total")
rownames(t) <- c("X<x", "X=x", "X>x", "Total")
knitr::kable(t)
```

	Y<y	Y=y	Y>y	Total
X<x	190	0	101	291
X=x	5	0	8	13
X>x	405	3	489	897
Total	600	3	598	1201

a. $P(X > x|Y > y) = \frac{489}{598} \approx 0.8177$
b. $P(X > x \text{ and } Y > y) = \frac{489}{1201} \approx 0.4072$
c. $P(X < x|Y > y) = \frac{101}{598} \approx 0.1689$

Additional probabilities can be calculated using the following probabilities table.

```
knitr::kable(round(t/1201,4))
```

	Y<y	Y=y	Y>y	Total
X<x	0.1582	0.0000	0.0841	0.2423
X=x	0.0042	0.0000	0.0067	0.0108
X>x	0.3372	0.0025	0.4072	0.7469

	Y<y	Y=y	Y>y	Total
Total	0.4996	0.0025	0.4979	1.0000

Consider probability (a) above: $P(X > x|Y > y) = 0.8177$. $P(X > x) = 0.7469$.

Since $P(X > x|Y > y) \neq P(X > x)$, these events are **not independent**.

Chi-squared Test

Let us the chi-squared test to evaluate the null hypothesis that $X > x$ (lot frontage is greater than first quartile or 59) and $Y > y$ (sale price is greater than median or 159,500) are independent events. Because we have very few events such that $X = x$ or $Y = y$, we will combine those values and only use two categories - $\leq$ and $>$.

```
chisq.test(table(probdata$LotFrontage>x, probdata$SalePrice>y))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(probdata$LotFrontage > x, probdata$SalePrice > y)
## X-squared = 30.881, df = 1, p-value = 0.00000002743
```

The p-value is nearly zero. Therefore, we reject the null hypothesis. Two events are not independent.

Descriptive and Inferential Statistics

Let us get some basic statistics about the `LotFrontage` variable.

```
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.00   59.00   69.00   70.05   80.00  313.00     259
```

```
describe(X)
```

```
##    vars    n  mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 1201 70.05 24.28     69   68.94 16.31  21 313   292 2.16    17.34
##     se
## X1 0.7
```

There are 1,201 valid observations between a very small/narrow lot of 21 feet and large lot of 313 feet. Average frontage is 70.05 feet.

Let us get some basic statistics about the `SalePrice` variable.

```
summary(Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  130000  163000  180900  214000  755000
```

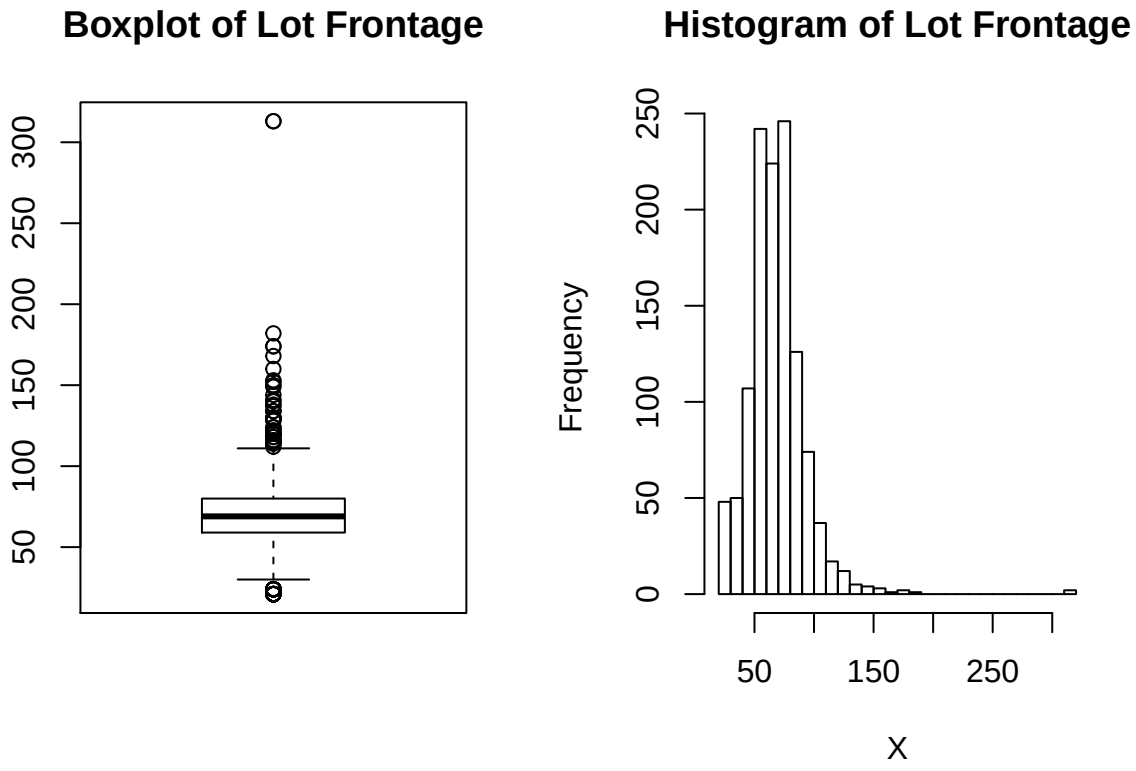
```
describe(Y)
```

```
##    vars    n     mean      sd median  trimmed     mad   min    max  range
## X1    1 1460 180921.2 79442.5 163000 170783.3 56338.8 34900 755000 720100
##    skew kurtosis      se
## X1 1.88      6.5 2079.11
```

Sale price is availavble for all 1,460 observations. It ranges from just shy of $35,000 to just over $750,000. Average sale price is $180,900.

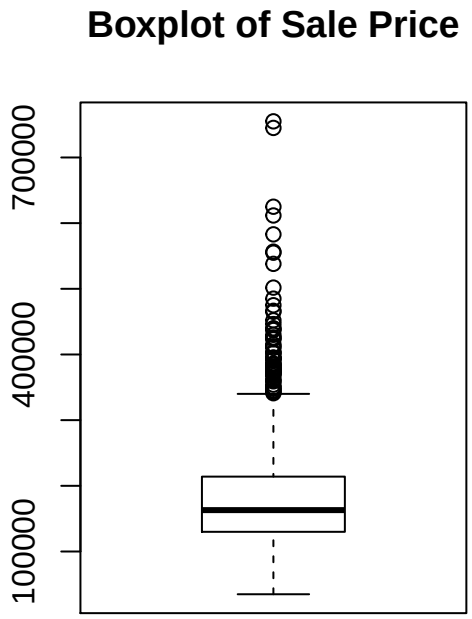
Let us evaluate a few plots.

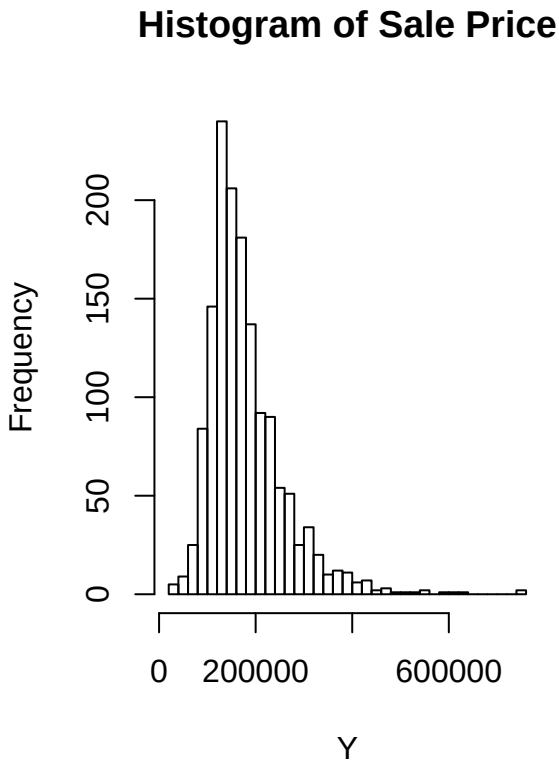
```
par(mfrow=c(1,2))
boxplot(X, main="Boxplot of Lot Frontage")
hist(X, breaks=40, main="Histogram of Lot Frontage")
```



Looking at the `LotFrontage` boxplot and histogram, we can see that there are definitely outliers and distibution is right-skewed.

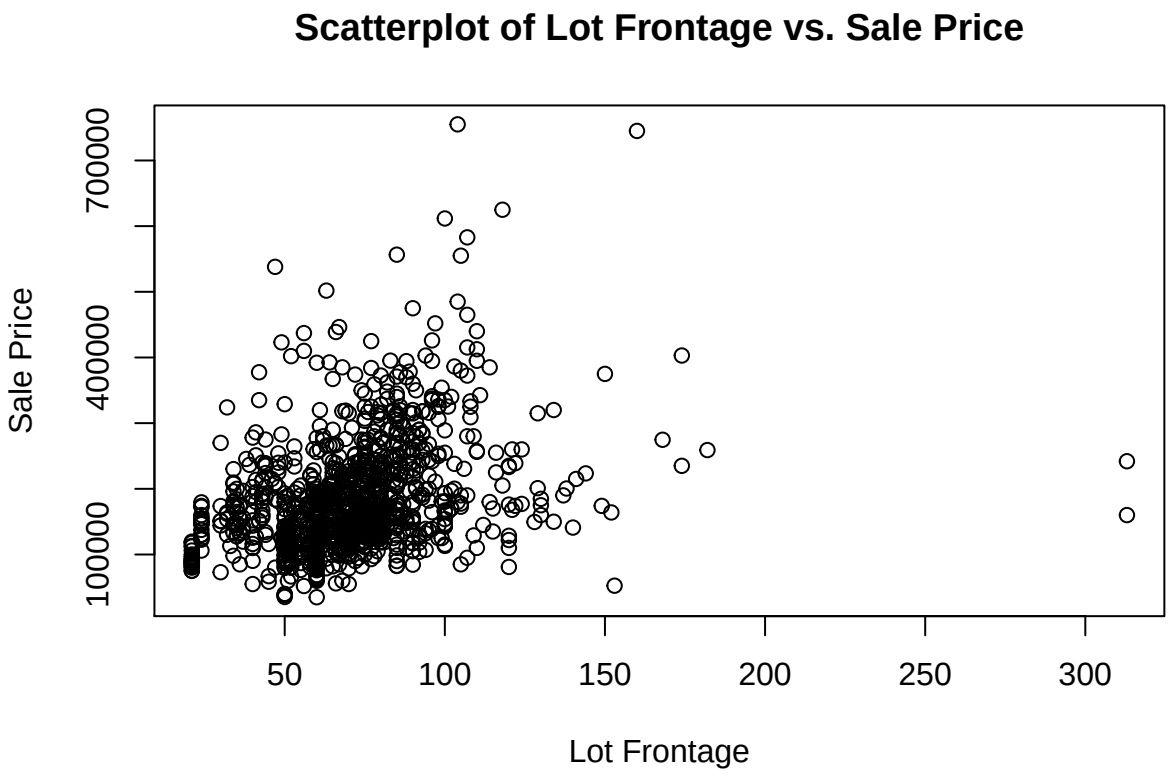
```
par(mfrow=c(1,2))
boxplot(Y, main="Boxplot of Sale Price")
hist(Y, breaks=40, main="Histogram of Sale Price")
```





Distribution of `SalePrice` is close to the distribution of `LotFrontage`. It is also right-skewed with some number of outliers.

```
plot(X, Y, xlab="Lot Frontage", ylab="Sale Price",
     main="Scatterplot of Lot Frontage vs. Sale Price")
```



Looking at the scatter plot, there is no obvious correlation, but it is a bit hard to say definitely. Let us build and evaluate a linear regression model.

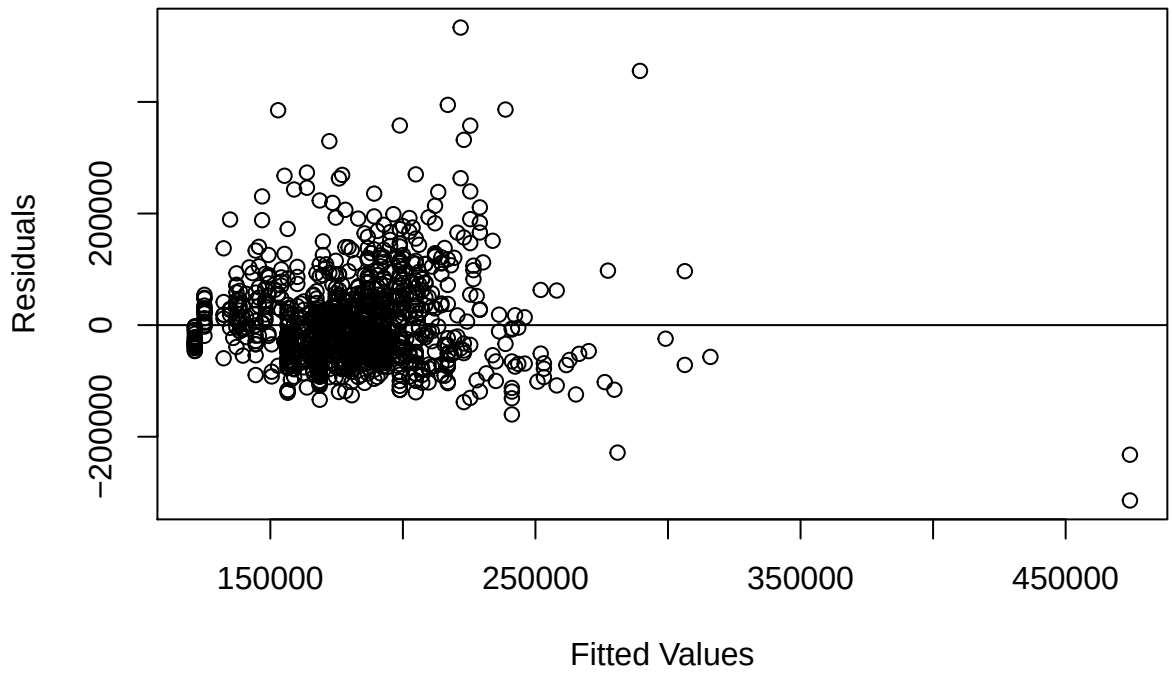
```
lm1 <- lm(Y ~ X)
summary(lm1)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -314258  -48878  -19402   33290  533217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96149.04    6881.97   13.97   <2e-16 ***
## X            1208.02      92.83   13.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78090 on 1199 degrees of freedom
##   (259 observations deleted due to missingness)
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.123
## F-statistic: 169.4 on 1 and 1199 DF,  p-value: < 2.2e-16
```

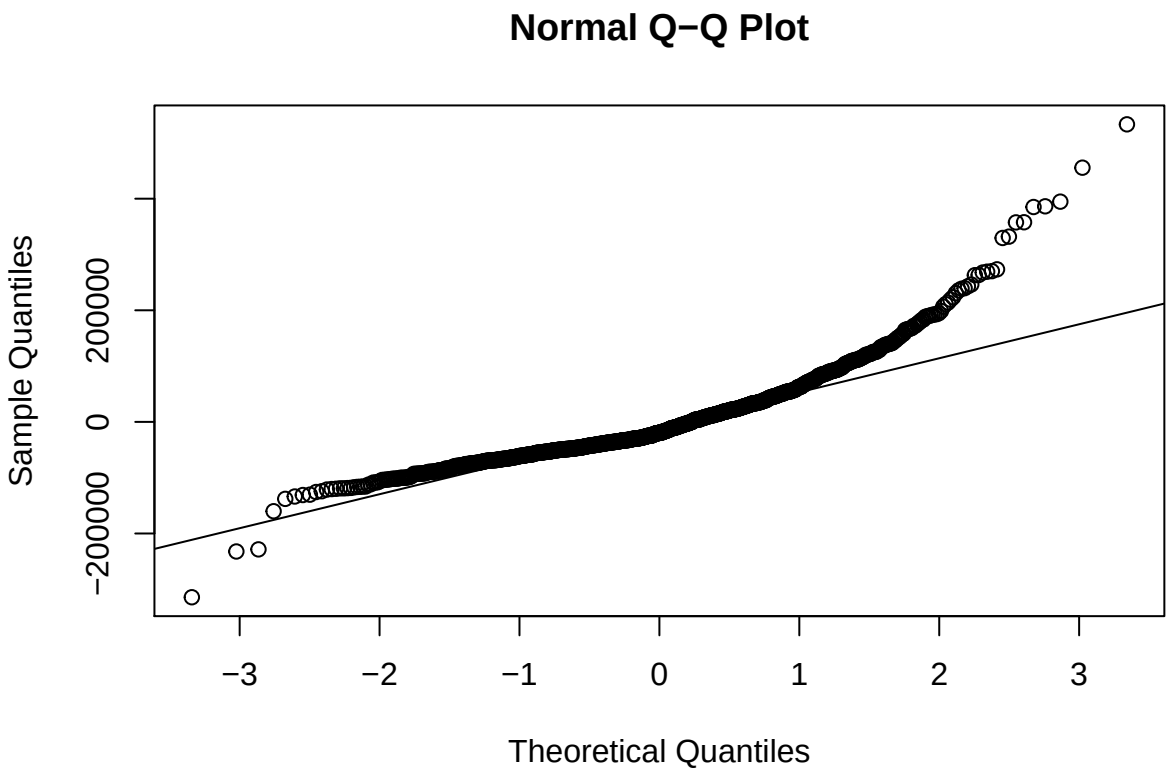
Looking at the summary, we see that `LotFrontage` may carry some significance, but at the same time R^2 is

very small and the model covers only about 12% of variability. More importantly looking at the analysis of residuals below it is clear that variability of residuals is not constant and distribution deviates from normal distribution. Because of these issues it would not be appropriate to use this model for analysis.

```
plot(lm1$fitted.values, lm1$residuals,
     xlab="Fitted Values", ylab="Residuals")
abline(h=0)
```



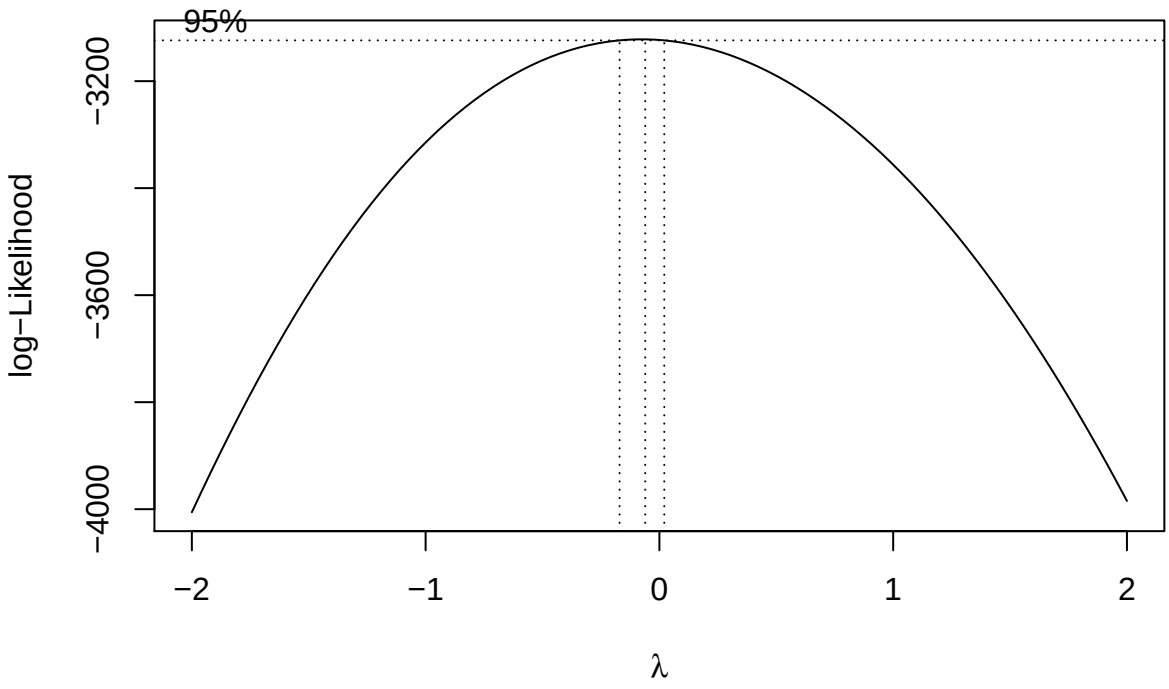

```
qqnorm(lm1$residuals); qqline(lm1$residuals)
```



Box Cox Transformation

Let us perform the Box Cox transformation to see if this model can be improved.

```
bc <- boxcox(lm1)
```




```
(lambda <- bc$x[which.max(bc$y)])
```

```
## [1] -0.06060606
```

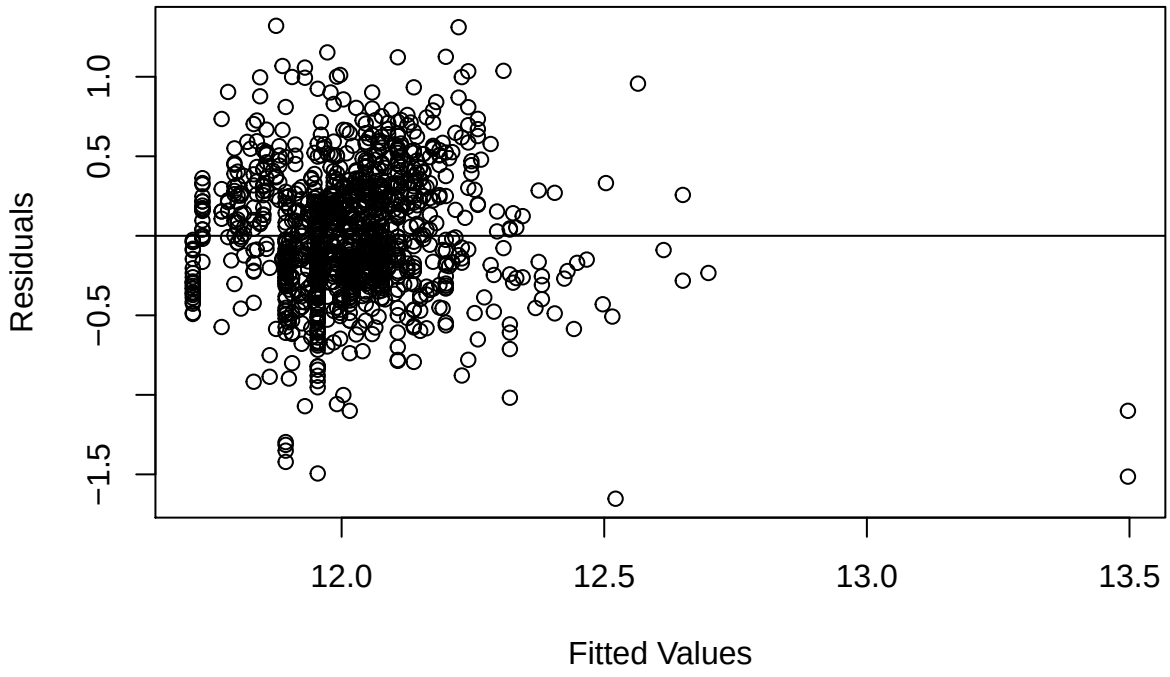
It looks like the optimal λ value is close to 0. In fact 0 is included in the 95% confidence interval. Let us try the *log* transformation.

```
lm2 <- lm(log(Y) ~ X)
summary(lm2)
```

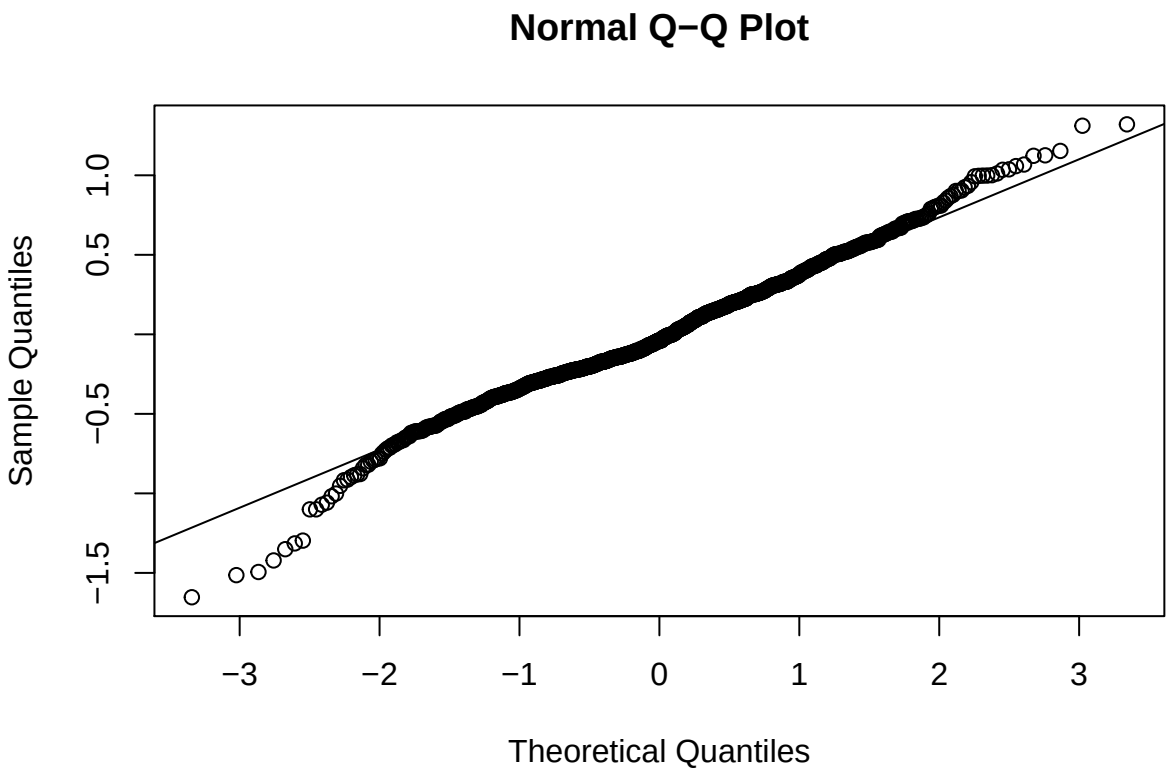
```
##
## Call:
## lm(formula = log(Y) ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65299 -0.24099 -0.03926  0.25161  1.32033
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.5887337  0.0342787  338.07   <2e-16 ***
## X            0.0060969  0.0004624   13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.389 on 1199 degrees of freedom
##   (259 observations deleted due to missingness)
## Multiple R-squared:  0.1266, Adjusted R-squared:  0.1259
```

```
## F-statistic: 173.9 on 1 and 1199 DF,  p-value: < 2.2e-16
```

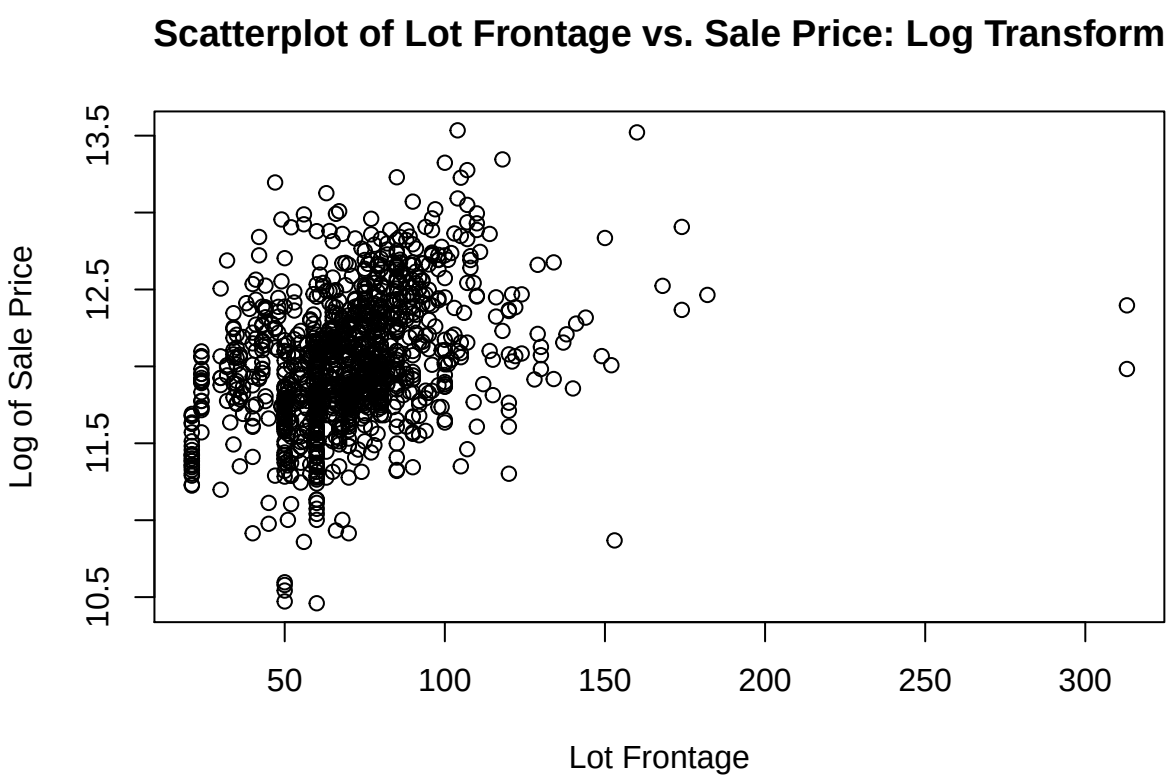
```
plot(lm2$fitted.values, lm2$residuals,
     xlab="Fitted Values", ylab="Residuals")
abline(h=0)
```




```
qqnorm(lm2$residuals); qqline(lm2$residuals)
```




```
plot(X, log(Y), xlab="Lot Frontage", ylab="Log of Sale Price",
     main="Scatterplot of Lot Frontage vs. Sale Price: Log Transform")
```



After *log* transformaion of `SalePrice`, distribution of residuals is significantly closer to normal and there is noticeable improvement in variability. R^2 of the new model has not increased, but with Box Cox transformation the new model adheres closer to necessary assumptions. It is stil probably not enough, but there is improvement.

Linear Algebra and Correlation

Let us select the following 4 variables from the data and build a correlation matrix:

- `TotalBsmtSF`: Total square feet of basement area
- `MoSold`: Month sold
- `OverallCond`: Rates the overall condition of the house (from *1-Very Poor* to *10-Very Excellent*)
- `SalePrice`: Sale price

```
cordata <- train[, c("TotalBsmtSF", "MoSold", "OverallCond", "SalePrice")]
cormatrix <- cor(cordata)
round(cormatrix,2)
```

```
##             TotalBsmtSF MoSold OverallCond SalePrice
## TotalBsmtSF        1.00   0.01       -0.17      0.61
## MoSold             0.01   1.00        0.00      0.05
## OverallCond       -0.17   0.00        1.00     -0.08
## SalePrice          0.61   0.05       -0.08      1.00
```

Before analysis I suspected that month the house was sold in correlates to the sale price since sale decisions may be driven by school schedule and weather. This turned out to be not the case. Less surprisingly basement size does seem to correlate with the sale price (larger basement suggests larger house and higher sale price).

Let us invert the correlation matrix to get the precision matrix.

```
precmatrix <- solve(cormatrix)
round(precmatrix,2)
```

```
##             TotalBsmtSF MoSold OverallCond SalePrice
## TotalBsmtSF        1.64   0.03        0.20     -0.99
## MoSold             0.03   1.00        0.00     -0.06
## OverallCond        0.20   0.00        1.03     -0.05
## SalePrice         -0.99  -0.06       -0.05      1.61
```

```
round(diag(precmatrix),2)
```

```
## TotalBsmtSF      MoSold OverallCond   SalePrice
##        1.64        1.00        1.03        1.61
```

Variance inflation factors from the diagonal of the precision matrix indicate that `MoSold` and `OverallCond` are not correlated among 4 variables chosen for this analysis while `TotalBsmtSF` and `SalePrice` may have moderate correlation.

Since $[Precision] = [Correlation]^{-1}$, then $[Precision] \times [Correlation]$ should be equal to **identity** matrix. Let us confirm.

```
round(cormatrix %*% precmatrix,4)
```

```
##             TotalBsmtSF MoSold OverallCond SalePrice
## TotalBsmtSF           1      0           0         0
## MoSold                0      1           0         0
## OverallCond           0      0           1         0
## SalePrice             0      0           0         1
```

```
round(precmatrix %*% cormatrix,4) == round(cormatrix %*% precmatrix,4)
```

```
##             TotalBsmtSF MoSold OverallCond SalePrice
## TotalBsmtSF        TRUE   TRUE        TRUE      TRUE
## MoSold             TRUE   TRUE        TRUE      TRUE
## OverallCond        TRUE   TRUE        TRUE      TRUE
## SalePrice          TRUE   TRUE        TRUE      TRUE
```

Calculus-Based Probability and Statistics

Let us compare the actual distribution of `LotFrontage` against gamma distribution using the `fitdistr` method of the `MASS` library.

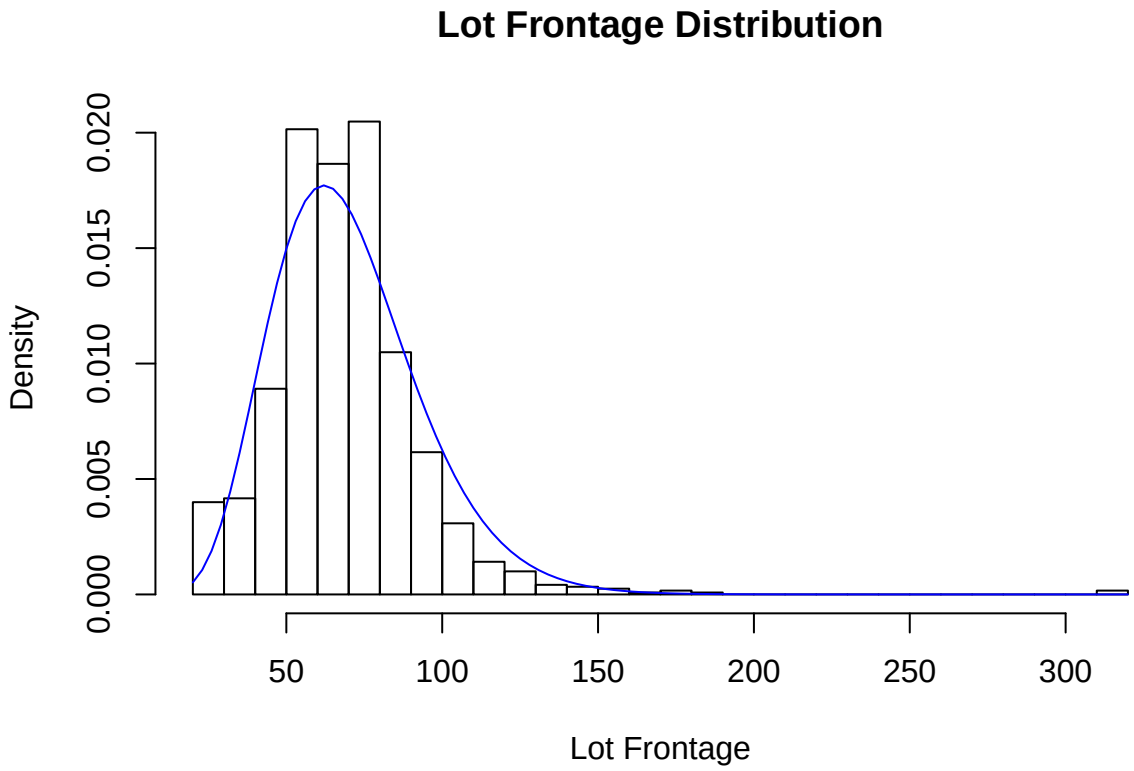
```
# Remove NAs
X <- X[!is.na(X)]

# Fitting of univariate distribution
(fd <- fitdistr(X, "gamma"))
```

```
##       shape         rate
##   8.760347516   0.125058578
##  (0.350766218) (0.005153394)
```

```
# Actual vs simulated distribution
hist(X, breaks=40, prob=TRUE, xlab="Lot Frontage",
     main="Lot Frontage Distribution")
```

```
curve(dgamma(x, shape = fd$estimate['shape'], rate = fd$estimate['rate']),
      col="blue", add=TRUE)
```



Looks like a very good fit. I picked the gamma distribution because the histogram of `LotFrontage` above seemed to resemble it. Now I am curious if perhaps another distribution would be a better fit based on `fitdistr`.

```
distributions <- c("cauchy", "exponential", "gamma", "geometric", "log-normal", "lognormal",
                   "logistic", "negative binomial", "normal", "Poisson", "t", "weibull")
logliks <- c()
for (d in distributions) {
  logliks <- c(logliks, fitdistr(X, d)$loglik)
}

logtable <- as.data.frame(cbind(distributions, logliks))
knitr::kable(logtable[order(logtable$logliks),],
             row.names=FALSE, col.names=c("Distribution", "Log-Likelihood"))
```

Distribution	Log-Likelihood
t	-5396.744870698
logistic	-5419.29549094738
negative binomial	-5453.90319970878
gamma	-5457.20013064461
log-normal	-5485.89040512713
lognormal	-5485.89040512713
cauchy	-5519.19936068162

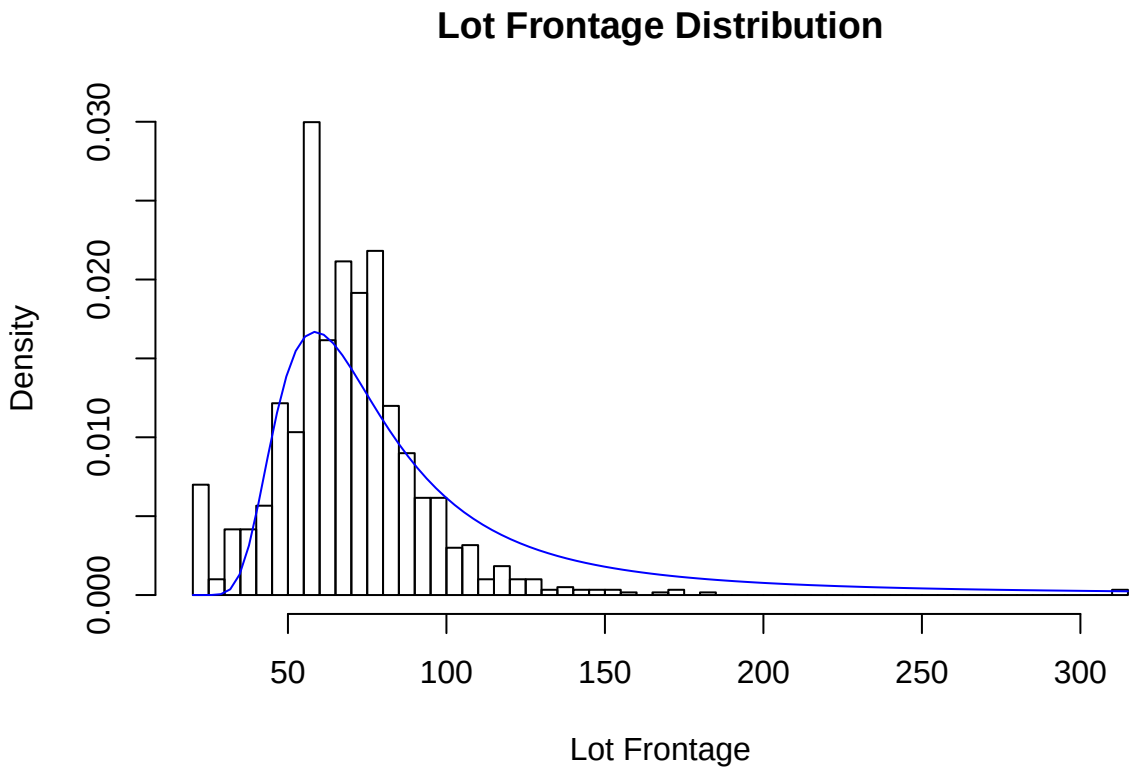
Distribution	Log-Likelihood
normal	-5534.6532040307
weibull	-5559.41415360689
exponential	-6304.29962283632
geometric	-6312.83157278274
Poisson	-8309.11486947954

Based on log-loglikelihood values, it seems that the t distribution would be a better fit.

```
(fd <- fitdistr(X, "t"))
```

```
##         m             s            df
##    68.6302728    16.4861002     3.8687709
##   ( 0.5662556) ( 0.5999284) ( 0.4532258)
```

```
# Actual vs simulated distribution
hist(X, breaks=50, prob=TRUE, xlab="Lot Frontage",
     main="Lot Frontage Distribution")
curve(dt(x, df = fd$estimate['df'], ncp=fd$estimate['m']),
      col="blue", add=TRUE)
```



The graph is a bit inconclusive. I feel that I am missing something here in properly interpreting the t distribution, but I have decided to include the work anyway.

Modeling

Modeling Summary

The following are key steps taken in building the model:

1. Data was reviewed and number of variables were eliminated because they contained a lot of missing data or there was not enough variability in the data.
2. Several variables were eliminated because they contained either repetetive data (represented by other variables) or data that is more likely to confuse the model rather than improve it.
3. Categorical variables were converted to numerica values.
4. Target variable, `SalePrice`, was log-transformed to bring it to the scale of other variables.
5. Training and testing data sets were combined for data clean-up, but then separated for modeling.

Final model formula is as follows:

> formula = SalePrice ~ MSZoning + LotFrontage + LotArea + BldgType + OverallQual + OverallCond + YearBuilt + RoofMatl + Exterior1st + Exterior2nd + ExterCond + BsmtCond + BsmtFinType1 + BsmtFinSF1 + HeatingQC + CentralAir + X1stFlrSF + GrLivArea + FullBath + KitchenQual + Functional + GarageFinish + GarageArea + GarageCond + PavedDrive + WoodDeckSF + SaleCondition

Kaggle username is **IlyaKats**. Final score is **0.13513**. Submission data is available at https://github.com/ilyakats/CUNY-DATA605/tree/master/Project.

Modeling Work

```
# Read test data and add SalePrice column
test <- read.csv('https://raw.githubusercontent.com/ilyakats/CUNY-DATA605/master/Project/test.csv')
test <- cbind(test, SalePrice=rep(0,nrow(test)))

# Get training data and review summary statistics
md <- train
summary(md)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##                                                  NA's   :259
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48
##                FR3    :   4              Edwards:100   RRAn   :  26
##                Inside :1052              Somerst: 86   PosN   :  19
```

```
##                                          Gilbert: 79   RRAe   :  11
##                                          (Other):707   (Other):  15
##    Condition2     BldgType      HouseStyle   OverallQual
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000
##  (Other):   2                 (Other): 19
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2
##
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7
##  ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
##  (Other):   2   (Other):128   (Other):136                 NA's   :8
##  ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
##  Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
##  Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
##  TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
##            TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00
##  BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00
##  GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00
##  LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55
##  Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00
##  Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00
##  NA's: 37                      NA's:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
```

```
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
##  Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
##                   Typ :1360                               NA's   : 81
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3
##  1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48
##  Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd  :  14
##  Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
##  NA's   :81
##  GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  Ex  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Fa  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Gd  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  Po  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  TA  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NA's:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch        PoolArea        PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex  :   2
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Gd  :   3
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  GdPrv:  59   Gar2:   2   Min.   :    0.00   Min.   : 1.000
```

```
##  GdWo :  54   Othr:   2   1st Qu.:    0.00   1st Qu.: 5.000
##  MnPrv: 157   Shed:  49   Median :    0.00   Median : 6.000
##  MnWw :  11   TenC:   1   Mean   :   43.49   Mean   : 6.322
##  NA's :1179   NA's:1406   3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :15500.00   Max.   :12.000
##
##      YrSold        SaleType    SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
```

```
# Combine with testing data to do global replacements
md <- rbind(md, test)

# Eliminate features with limited or missing data
md <- subset(md, select=-c(Street, Alley, LandContour, Utilities,
                           LandSlope, Condition2, MasVnrArea, Heating,
                           BsmtFinSF2, X2ndFlrSF, LowQualFinSF, BsmtFullBath,
                           BsmtHalfBath, HalfBath, PoolQC, PoolArea, MiscVal,
                           MiscFeature, Fence, ScreenPorch, Fireplaces,
                           EnclosedPorch, MoSold, YrSold))
```

Based on summary statistics above the following fields were eliminated from modeling because of a lot of missing data - `Street`, `Alley`, `LandContour`, `Utilities`, `LandSlope`, `Condition2`, `MasVnrArea`, `Heating`, `BsmtFinSF2`, `X2ndFlrSF`, `LowQualFinSF`, `BsmtFullBath`, `BsmtHalfBath`, `HalfBath`, `PoolQC`, `PoolArea`, `MiscVal`, `MiscFeature`, `Fence`, `ScreenPorch`, `Fireplaces`, `EnclosedPorch`. Additionally, `Id` was eliminated because it carries no relevant information. We have established above that `MoSold` does not correlate with sale price, so it was eliminated. Finally, `YrSold` was eliminated. Although, year may play a factor, data covers 2006 through 2010 - a relatively short period that is unlikely to contain significant patterns.

We are left with the following columns.

```
colnames(md)
```

```
##  [1] "Id"            "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"       "LotShape"      "LotConfig"     "Neighborhood"
##  [9] "Condition1"    "BldgType"      "HouseStyle"    "OverallQual"
## [13] "OverallCond"   "YearBuilt"     "YearRemodAdd"  "RoofStyle"
## [17] "RoofMatl"      "Exterior1st"   "Exterior2nd"   "MasVnrType"
## [21] "ExterQual"     "ExterCond"     "Foundation"    "BsmtQual"
## [25] "BsmtCond"      "BsmtExposure"  "BsmtFinType1"  "BsmtFinSF1"
## [29] "BsmtFinType2"  "BsmtUnfSF"     "TotalBsmtSF"   "HeatingQC"
## [33] "CentralAir"    "Electrical"    "X1stFlrSF"     "GrLivArea"
## [37] "FullBath"      "BedroomAbvGr"  "KitchenAbvGr"  "KitchenQual"
## [41] "TotRmsAbvGrd"  "Functional"    "FireplaceQu"   "GarageType"
## [45] "GarageYrBlt"   "GarageFinish"  "GarageCars"    "GarageArea"
## [49] "GarageQual"    "GarageCond"    "PavedDrive"    "WoodDeckSF"
## [53] "OpenPorchSF"   "X3SsnPorch"    "SaleType"      "SaleCondition"
## [57] "SalePrice"
```

```
md <- subset(md, select=-c(LotShape, YearRemodAdd, BsmtExposure,
                           BsmtFinType2, TotalBsmtSF, TotRmsAbvGrd,
                           FireplaceQu, GarageYrBlt, GarageCars))
```

After reviewing data dictionary, the following fields were eliminated - `LotShape`, `YearRemodAdd` (contains construction year if no remodeling was done which may negatively interfere with the model), `BsmtExposure`, `BsmtFinType2`, `TotalBsmtSF` (included in other variables), `TotRmsAbvGrd`, `FireplaceQu`, `GarageYrBlt`, `GarageCars`.

Categorical variables were converted to numerical values. Remaining NAs were replaced with zeros.

```
md$Neighborhood <- as.integer(factor(md$Neighborhood))
md$MSZoning <- as.integer(factor(md$MSZoning))
md$LotConfig <- as.integer(factor(md$LotConfig))
md$Condition1 <- as.integer(factor(md$Condition1))
md$BldgType <- as.integer(factor(md$BldgType))
md$HouseStyle <- as.integer(factor(md$HouseStyle))
md$RoofStyle <- as.integer(factor(md$RoofStyle))
md$RoofMatl <- as.integer(factor(md$RoofMatl))
md$Exterior1st <- as.integer(factor(md$Exterior1st))
md$Exterior2nd <- as.integer(factor(md$Exterior2nd))
md$MasVnrType <- as.integer(factor(md$MasVnrType))
md$ExterQual <- as.integer(factor(md$ExterQual))
md$ExterCond <- as.integer(factor(md$ExterCond))
md$BsmtQual <- as.integer(factor(md$BsmtQual))
md$BsmtCond <- as.integer(factor(md$BsmtCond))
md$Electrical <- as.integer(factor(md$Electrical))
md$KitchenQual <- as.integer(factor(md$KitchenQual))
md$Functional <- as.integer(factor(md$Functional))
md$GarageType <- as.integer(factor(md$GarageType))
md$GarageFinish <- as.integer(factor(md$GarageFinish))
md$GarageCond <- as.integer(factor(md$GarageCond))
md$BsmtFinType1 <- as.integer(factor(md$BsmtFinType1))
md$PavedDrive <- as.integer(factor(md$PavedDrive))
md$SaleType <- as.integer(factor(md$SaleType))
md$SaleCondition <- as.integer(factor(md$SaleCondition))
md$Foundation <- as.integer(factor(md$Foundation))
md$HeatingQC <- as.integer(factor(md$HeatingQC))
md$GarageQual <- as.integer(factor(md$GarageQual))

md[is.na(md)] <- 0
```

Separate data into training and testing sets. Log-transform sales price in the training set.

```
test <- md[md$SalePrice==0,]
md <- md[md$SalePrice>0,]

md$SalePrice <- log(md$SalePrice)

# Remove ID column from training data
md <- subset(md, select=-c(Id))

# Build initial model with all fields
sale_lm <- lm(SalePrice ~ . , data=md)
summary(sale_lm)
```

```
##
## Call:
```

```
## lm(formula = SalePrice ~ ., data = md)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22119 -0.06729  0.00481  0.07152  0.57745
##
## Coefficients:
##                   Estimate    Std. Error t value       Pr(>|t|)
## (Intercept)    6.8245491893  0.5866086932  11.634        < 2e-16 ***
## MSSubClass    -0.0001499385  0.0002062562  -0.727       0.467375
## MSZoning      -0.0113649010  0.0069254126  -1.641       0.101011
## LotFrontage   -0.0003569532  0.0001241554  -2.875       0.004100 **
## LotArea        0.0000021439  0.0000004345   4.934 0.000000900481 ***
## LotConfig     -0.0024205201  0.0024645119  -0.982       0.326194
## Neighborhood   0.0006136411  0.0007167650   0.856       0.392073
## Condition1     0.0023177047  0.0046531586   0.498       0.618496
## BldgType      -0.0141640724  0.0067806927  -2.089       0.036897 *
## HouseStyle    -0.0006281607  0.0029067379  -0.216       0.828937
## OverallQual    0.0841413908  0.0053662685  15.680        < 2e-16 ***
## OverallCond    0.0377076166  0.0043970790   8.576        < 2e-16 ***
## YearBuilt      0.0018044387  0.0003013258   5.988 0.000000002685 ***
## RoofStyle      0.0054973694  0.0051203865   1.074       0.283175
## RoofMatl       0.0142457900  0.0068518449   2.079       0.037787 *
## Exterior1st   -0.0044228762  0.0024047674  -1.839       0.066094 .
## Exterior2nd    0.0051075939  0.0021846799   2.338       0.019531 *
## MasVnrType     0.0066061971  0.0063761285   1.036       0.300341
## ExterQual     -0.0071693436  0.0089708972  -0.799       0.424322
## ExterCond      0.0126996544  0.0057910534   2.193       0.028471 *
## Foundation     0.0031759507  0.0076838702   0.413       0.679430
## BsmtQual      -0.0082611036  0.0057893473  -1.427       0.153816
## BsmtCond       0.0190941148  0.0056795398   3.362       0.000795 ***
## BsmtFinType1  -0.0070904709  0.0028044244  -2.528       0.011569 *
## BsmtFinSF1     0.0000453185  0.0000175868   2.577       0.010071 *
## BsmtUnfSF     -0.0000022554  0.0000185693  -0.121       0.903347
## HeatingQC     -0.0111870208  0.0027751580  -4.031 0.000058470318 ***
## CentralAirY    0.0890470168  0.0196187285   4.539 0.000006135533 ***
## Electrical     0.0022272879  0.0041701417   0.534       0.593354
## X1stFlrSF      0.0000761228  0.0000222085   3.428       0.000626 ***
## GrLivArea      0.0002137006  0.0000161953  13.195        < 2e-16 ***
## FullBath       0.0277160050  0.0112363373   2.467       0.013757 *
## BedroomAbvGr   0.0095702265  0.0068565312   1.396       0.162999
## KitchenAbvGr  -0.0269103105  0.0219370079  -1.227       0.220137
## KitchenQual   -0.0299140191  0.0066383046  -4.506 0.000007141896 ***
## Functional     0.0164999406  0.0043389159   3.803       0.000149 ***
## GarageType    -0.0054905209  0.0028665986  -1.915       0.055650 .
## GarageFinish  -0.0125810692  0.0064817604  -1.941       0.052457 .
## GarageArea     0.0001610149  0.0000294003   5.477 0.000000051244 ***
## GarageQual     0.0045933999  0.0076827806   0.598       0.550014
## GarageCond     0.0162230798  0.0079750856   2.034       0.042116 *
## PavedDrive     0.0201491987  0.0095357241   2.113       0.034774 *
## WoodDeckSF     0.0001147680  0.0000340577   3.370       0.000772 ***
## OpenPorchSF   -0.0000148112  0.0000654207  -0.226       0.820923
## X3SsnPorch     0.0001175080  0.0001348675   0.871       0.383747
## SaleType      -0.0005982979  0.0026325612  -0.227       0.820248
```

```
## SaleCondition  0.0238176500  0.0038100777   6.251 0.000000000538 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1486 on 1413 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.8617
## F-statistic: 198.6 on 46 and 1413 DF,  p-value: < 2.2e-16
```

Optimize the model using `stepAIC` method.

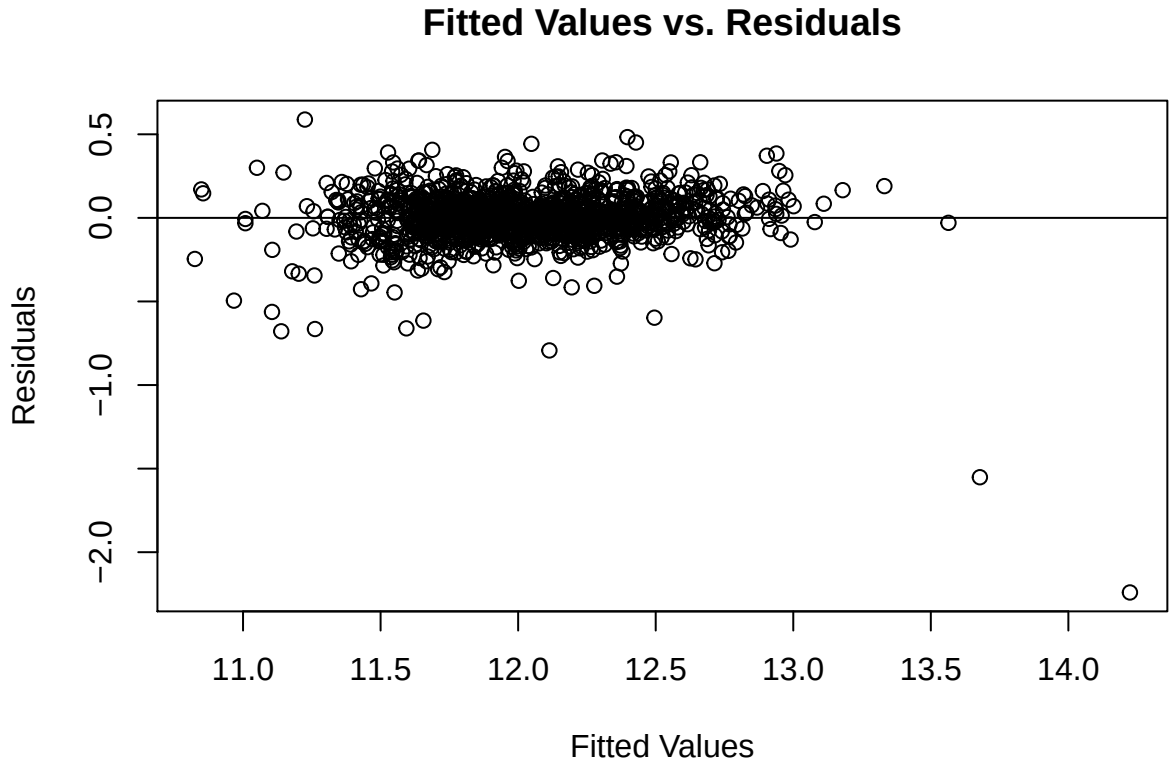
```
step_lm <- stepAIC(sale_lm, trace=FALSE)
summary(step_lm)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSZoning + LotFrontage + LotArea + BldgType +
##     OverallQual + OverallCond + YearBuilt + RoofMatl + Exterior1st +
##     Exterior2nd + ExterCond + BsmtQual + BsmtCond + BsmtFinType1 +
##     BsmtFinSF1 + HeatingQC + CentralAir + X1stFlrSF + GrLivArea +
##     FullBath + KitchenQual + Functional + GarageType + GarageFinish +
##     GarageArea + GarageCond + PavedDrive + WoodDeckSF + SaleCondition,
##     data = md)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24086 -0.06623  0.00565  0.07351  0.58820
##
## Coefficients:
##                    Estimate    Std. Error t value          Pr(>|t|)
## (Intercept)    6.6009060484  0.5122756066  12.885           < 2e-16 ***
## MSZoning      -0.0135208963  0.0066378388  -2.037          0.041840 *
## LotFrontage   -0.0003484392  0.0001218122  -2.860          0.004292 **
## LotArea        0.0000021492  0.0000004291   5.009 0.000000616289858 ***
## BldgType      -0.0212957859  0.0036613594  -5.816 0.000000007408787 ***
## OverallQual    0.0863505864  0.0050137850  17.223           < 2e-16 ***
## OverallCond    0.0382499520  0.0042948768   8.906           < 2e-16 ***
## YearBuilt      0.0019266454  0.0002610902   7.379 0.000000000000269 ***
## RoofMatl       0.0128618101  0.0067259235   1.912          0.056041 .
## Exterior1st   -0.0042612207  0.0023702750  -1.798          0.072424 .
## Exterior2nd    0.0047060558  0.0021372511   2.202          0.027830 *
## ExterCond      0.0122796320  0.0057027593   2.153          0.031464 *
## BsmtQual      -0.0085045662  0.0054468491  -1.561          0.118657
## BsmtCond       0.0185608151  0.0054753119   3.390          0.000718 ***
## BsmtFinType1  -0.0072791564  0.0026341330  -2.763          0.005794 **
## BsmtFinSF1     0.0000450241  0.0000118126   3.812          0.000144 ***
## HeatingQC     -0.0113964832  0.0027047614  -4.213 0.000026719179231 ***
## CentralAirY    0.0937487472  0.0190062531   4.933 0.000000906709975 ***
## X1stFlrSF      0.0000842612  0.0000148898   5.659 0.000000018369101 ***
## GrLivArea      0.0002151555  0.0000126530  17.004           < 2e-16 ***
## FullBath       0.0268033002  0.0107570156   2.492          0.012826 *
## KitchenQual   -0.0312929251  0.0061043433  -5.126 0.000000335857780 ***
## Functional     0.0159773907  0.0042064430   3.798          0.000152 ***
## GarageType    -0.0051096355  0.0028169164  -1.814          0.069901 .
## GarageFinish  -0.0125104973  0.0063597975  -1.967          0.049362 *
## GarageArea     0.0001617009  0.0000284126   5.691 0.000000015284918 ***
```

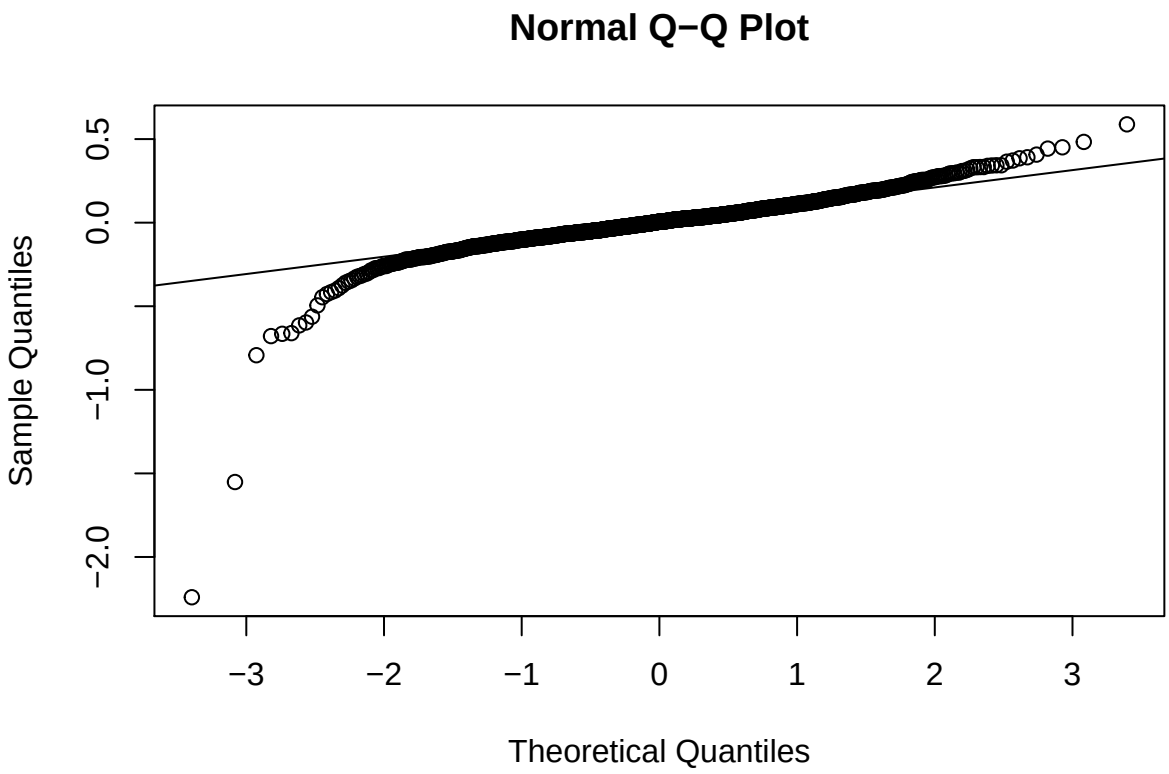
```
## GarageCond     0.0200256677  0.0049076281   4.081 0.000047417532552 ***
## PavedDrive     0.0205367649  0.0094337381   2.177           0.029647 *
## WoodDeckSF     0.0001163306  0.0000336756   3.454           0.000568 ***
## SaleCondition  0.0242121985  0.0036888630   6.564 0.000000000073266 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1483 on 1430 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8622
## F-statistic: 315.8 on 29 and 1430 DF,  p-value: < 2.2e-16
```

R^2 is over 0.86 which seems like a good fit. Check residuals.

```
plot(step_lm$fitted.values, step_lm$residuals,
     xlab="Fitted Values", ylab="Residuals", main="Fitted Values vs. Residuals")
abline(h=0)
```




```
qqnorm(step_lm$residuals); qqline(step_lm$residuals)
```



Residuals definitely point to a few outliers. The model can probably be improved if outliers are removed from the data. Residuals are approximately normally distributed.

```
# Predict prices for test data
pred_saleprice <- predict(step_lm, test)
# Convert from log back to real world number
pred_saleprice <- sapply(pred_saleprice, exp)
# Prepare data frame for submission
kaggle <- data.frame(Id=test$Id, SalePrice=pred_saleprice)
write.csv(kaggle, file = "submission.csv", row.names=FALSE)
```

Kaggle Submission

Kaggle username is **IlyaKats**. Final score is **0.13513**.

1333	new	**IlyaKats**		0.13513

Figure 1: