Regression Modeling

Ali Harb

May 25, 2017

Libraries

Load the required libraries to perform statistical analysis on the dataset.

```
library(stats)
library(MASS)
library(fifer)
library(moments)
library(ggplot2)
library(ggpubr)
library(psych)
library(leaps)
```

Load Dataset

Upload the train dataset from the GitHub to R environment.

```
train_data = read.csv("https://raw.githubusercontent.com/aliharb/IS-605-Computational-Mathematics/maste
colnames(train_data)
```

```
##  [1] "Id"            "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"       "Street"        "Alley"         "LotShape"
##  [9] "LandContour"   "Utilities"     "LotConfig"     "LandSlope"
## [13] "Neighborhood"  "Condition1"    "Condition2"    "BldgType"
## [17] "HouseStyle"    "OverallQual"   "OverallCond"   "YearBuilt"
## [21] "YearRemodAdd"  "RoofStyle"     "RoofMatl"      "Exterior1st"
## [25] "Exterior2nd"   "MasVnrType"    "MasVnrArea"    "ExterQual"
## [29] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
## [33] "BsmtExposure"  "BsmtFinType1"  "BsmtFinSF1"    "BsmtFinType2"
## [37] "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"   "Heating"
## [41] "HeatingQC"     "CentralAir"    "Electrical"    "X1stFlrSF"
## [45] "X2ndFlrSF"     "LowQualFinSF"  "GrLivArea"     "BsmtFullBath"
## [49] "BsmtHalfBath"  "FullBath"      "HalfBath"      "BedroomAbvGr"
## [53] "KitchenAbvGr"  "KitchenQual"   "TotRmsAbvGrd"  "Functional"
## [57] "Fireplaces"    "FireplaceQu"   "GarageType"    "GarageYrBlt"
## [61] "GarageFinish"  "GarageCars"    "GarageArea"    "GarageQual"
## [65] "GarageCond"    "PavedDrive"    "WoodDeckSF"    "OpenPorchSF"
## [69] "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"   "PoolArea"
## [73] "PoolQC"        "Fence"         "MiscFeature"   "MiscVal"
## [77] "MoSold"        "YrSold"        "SaleType"      "SaleCondition"
## [81] "SalePrice"
```

Check for best variable to do regression modeling

```
m <- lm(SalePrice ~ GrLivArea + LotArea + TotalBsmtSF, data=train_data)
step <- stepAIC(m, direction="both")
```

```
## Start:  AIC=31578.65
```

```
## SalePrice ~ GrLivArea + LotArea + TotalBsmtSF
##
##               Df  Sum of Sq        RSS   AIC
## <none>                      3.5927e+12 31579
## - LotArea      1 6.6413e+09 3.5993e+12 31579
## - TotalBsmtSF  1 9.3223e+11 4.5249e+12 31913
## - GrLivArea    1 2.0422e+12 5.6349e+12 32234
```

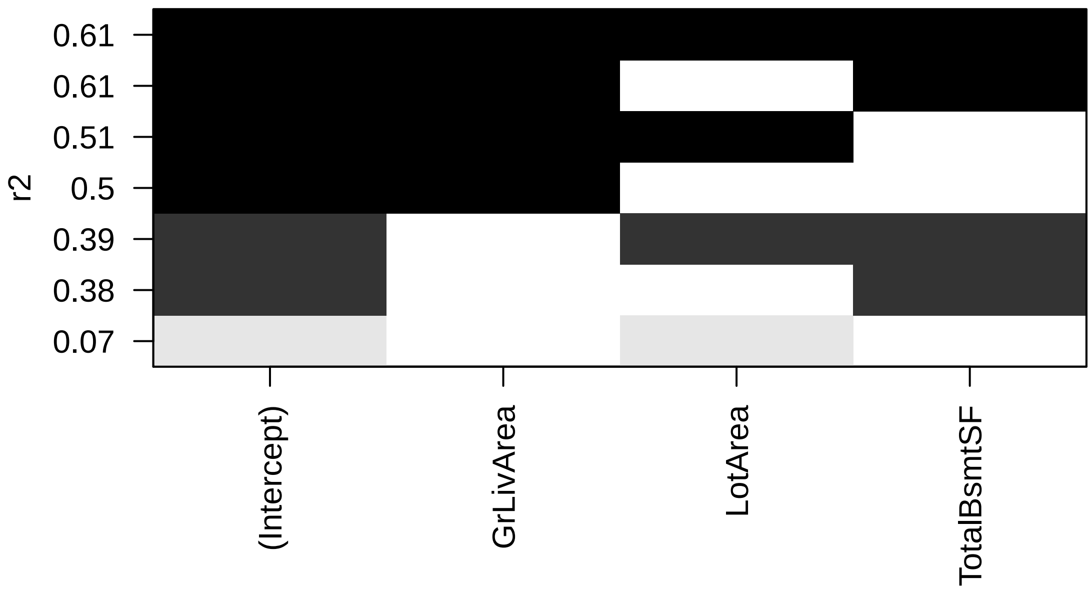
```r
step$anova # display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## SalePrice ~ GrLivArea + LotArea + TotalBsmtSF
##
## Final Model:
## SalePrice ~ GrLivArea + LotArea + TotalBsmtSF
##
##
##   Step Df Deviance Resid. Df   Resid. Dev      AIC
## 1                       1456 3.592686e+12 31578.65
```

```r
attach(train_data)
leaps<-regsubsets(SalePrice ~ GrLivArea + LotArea + TotalBsmtSF,data=train_data,nbest=3)
summary(leaps)
```

```
## Subset selection object
## Call: regsubsets.formula(SalePrice ~ GrLivArea + LotArea + TotalBsmtSF,
##     data = train_data, nbest = 3)
## 3 Variables  (and intercept)
##             Forced in Forced out
## GrLivArea       FALSE      FALSE
## LotArea         FALSE      FALSE
## TotalBsmtSF     FALSE      FALSE
## 3 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          GrLivArea LotArea TotalBsmtSF
## 1  ( 1 ) "*"       " "     " "
## 1  ( 2 ) " "       " "     "*"
## 1  ( 3 ) " "       "*"     " "
## 2  ( 1 ) "*"       " "     "*"
## 2  ( 2 ) "*"       "*"     " "
## 2  ( 3 ) " "       "*"     "*"
## 3  ( 1 ) "*"       "*"     "*"
```

```r
plot(leaps,scale="r2")
```



Based on the leaps results of the r-squared, i will choose the Ground living Area for my analysis.

Let's subset the variables and get the summary statitics

```
data <- subset(train_data, select = c("GrLivArea","SalePrice"))
summary(data)
```

```
##    GrLivArea      SalePrice
##  Min.   : 334   Min.   : 34900
##  1st Qu.:1130   1st Qu.:129975
##  Median :1464   Median :163000
##  Mean   :1515   Mean   :180921
##  3rd Qu.:1777   3rd Qu.:214000
##  Max.   :5642   Max.   :755000
```

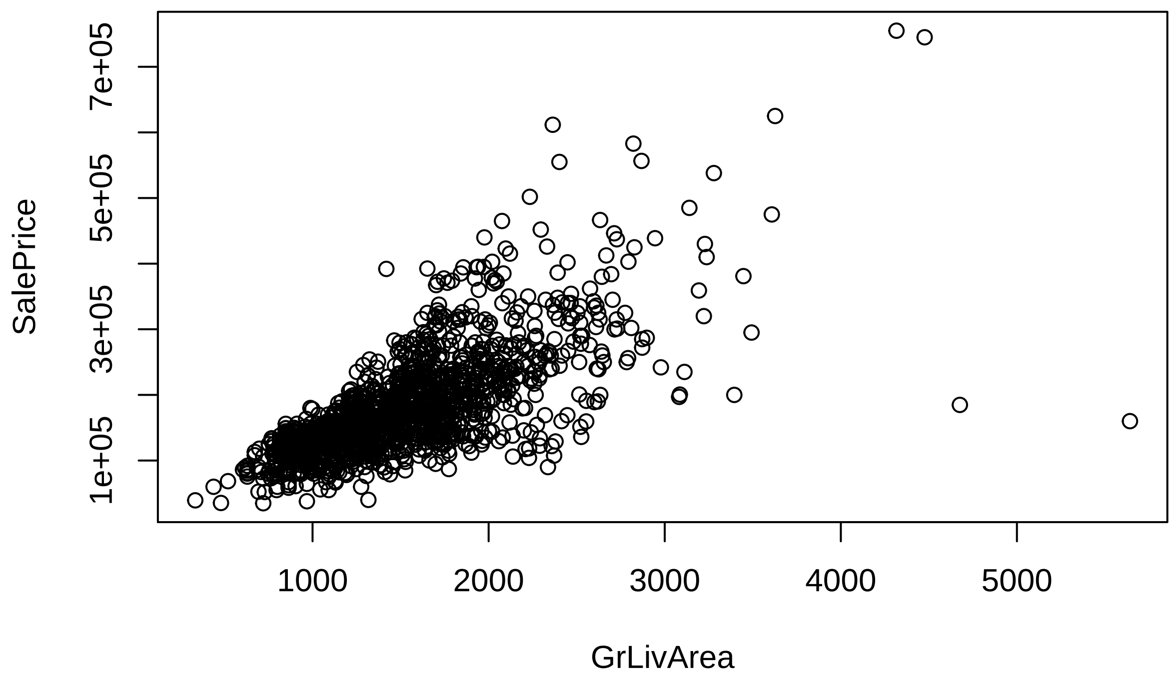
```
str(data)
```

```
## 'data.frame':    1460 obs. of  2 variables:
##  $ GrLivArea: int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ SalePrice: int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
describe(data)
```

```
##           vars    n      mean       sd median   trimmed      mad   min
## GrLivArea    1 1460   1515.46   525.48   1464   1467.67   483.33   334
## SalePrice    2 1460 180921.20 79442.50 163000 170783.29 56338.80 34900
##              max  range skew kurtosis      se
## GrLivArea   5642   5308 1.36     4.86   13.75
## SalePrice 755000 720100 1.88     6.50 2079.11
```

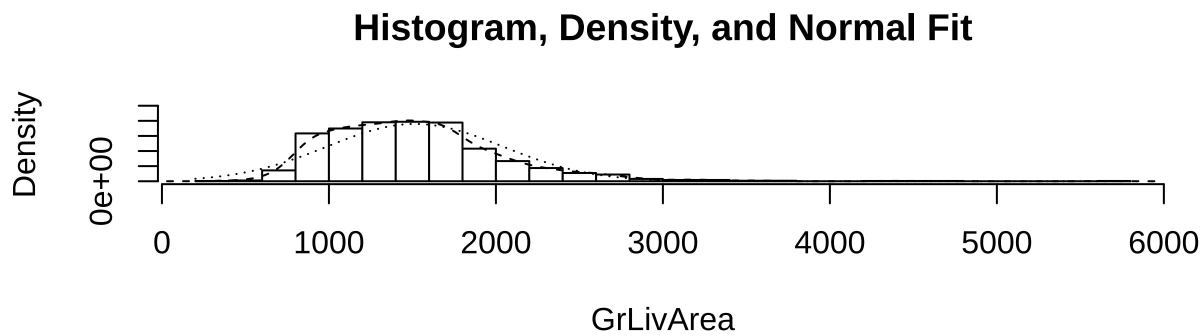
```r
plot(data)
```

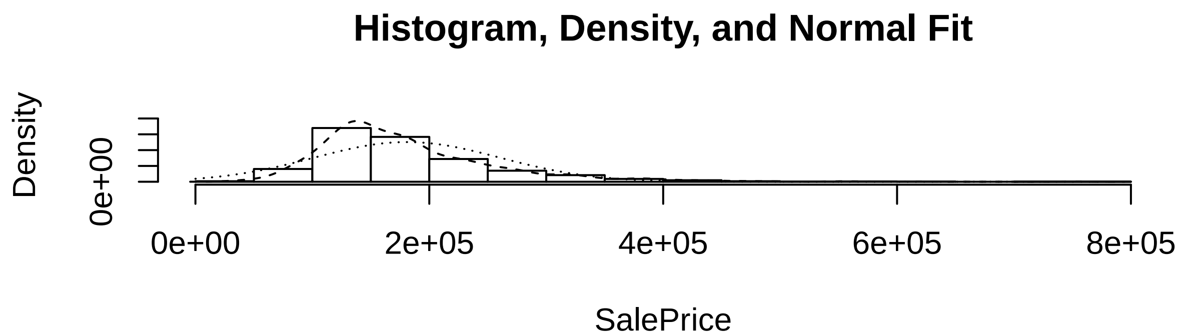


The scatter plot illustrates a possible positive linear relationship between ground living areas and sale prices. The scatter plot exhibit outliers specially at the high prices.

Lets take a visual look at the normal distribution

```r
multi.hist(data)
```





Even though the distribution not symmetric we will apply a simple regression and look at the result and correlation

```
m <- lm(SalePrice ~ GrLivArea, data=train_data)
coeffs <- coefficients(m)

print(paste0("SalesPrice = ", round(coeffs[2],3),"x + ",round(coeffs[1],3)))
```

```
## [1] "SalesPrice = 107.13x + 18569.026"
```

```
summary(m)
```

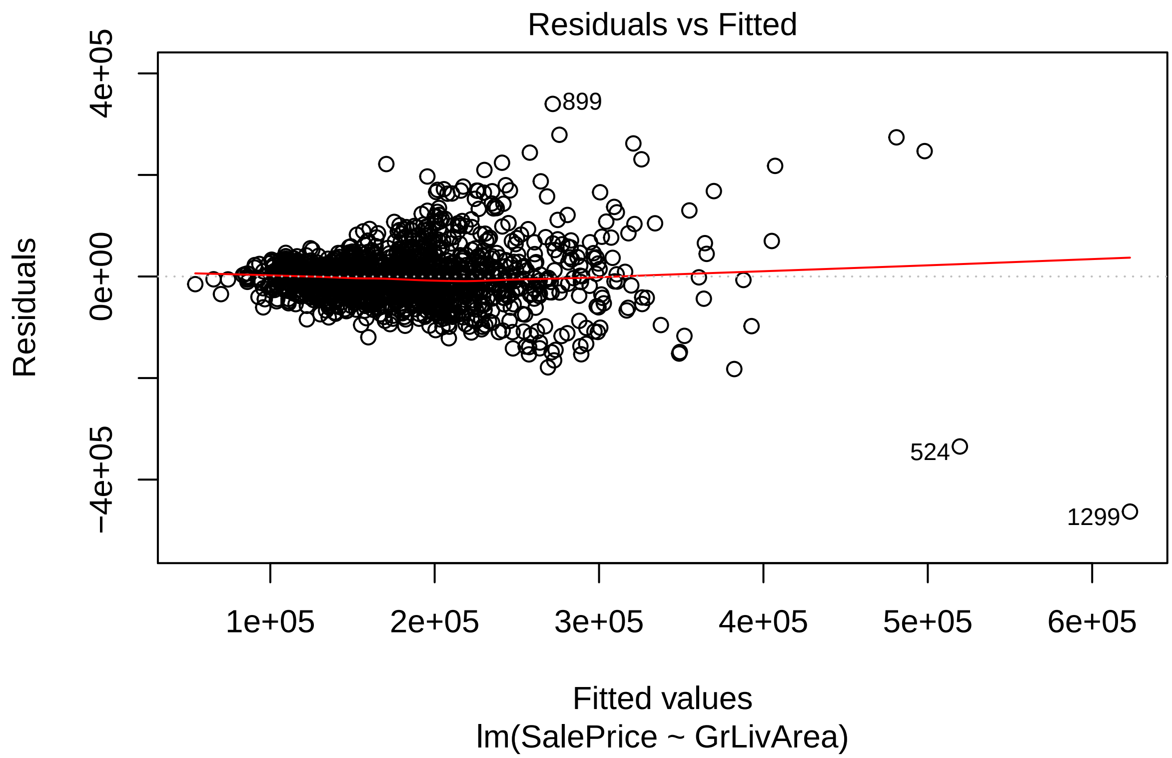
```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462999  -29800   -1124   21957  339832
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18569.026   4480.755   4.144 3.61e-05 ***
## GrLivArea     107.130      2.794  38.348  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56070 on 1458 degrees of freedom
```

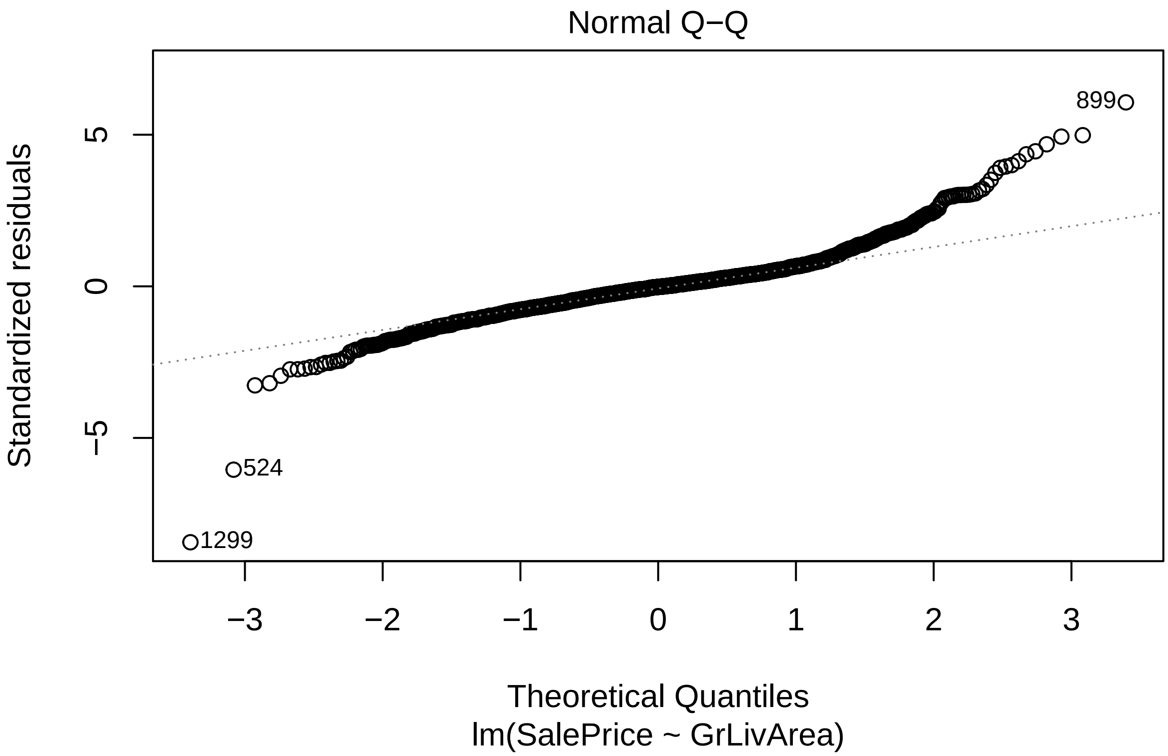
```
## Multiple R-squared:  0.5021, Adjusted R-squared:  0.5018
## F-statistic:  1471 on 1 and 1458 DF,  p-value: < 2.2e-16
```

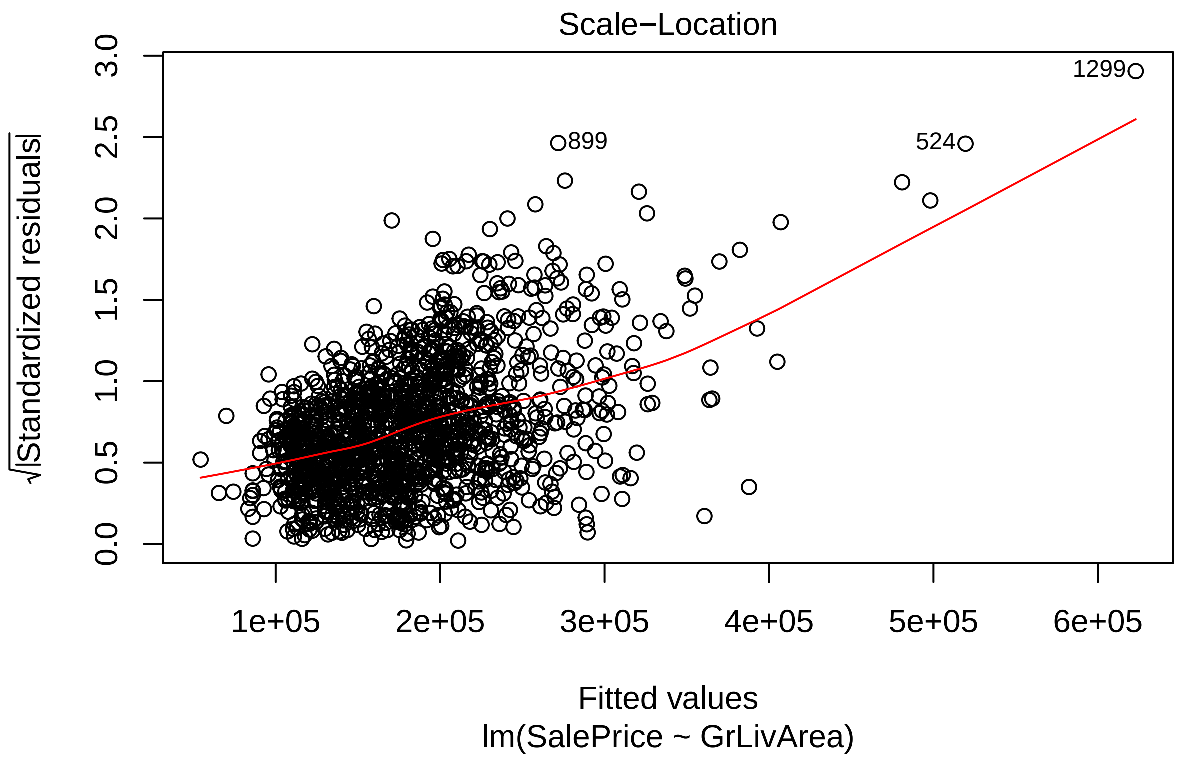
```r
cor(data$GrLivArea,data$SalePrice)
```

```
## [1] 0.7086245
```

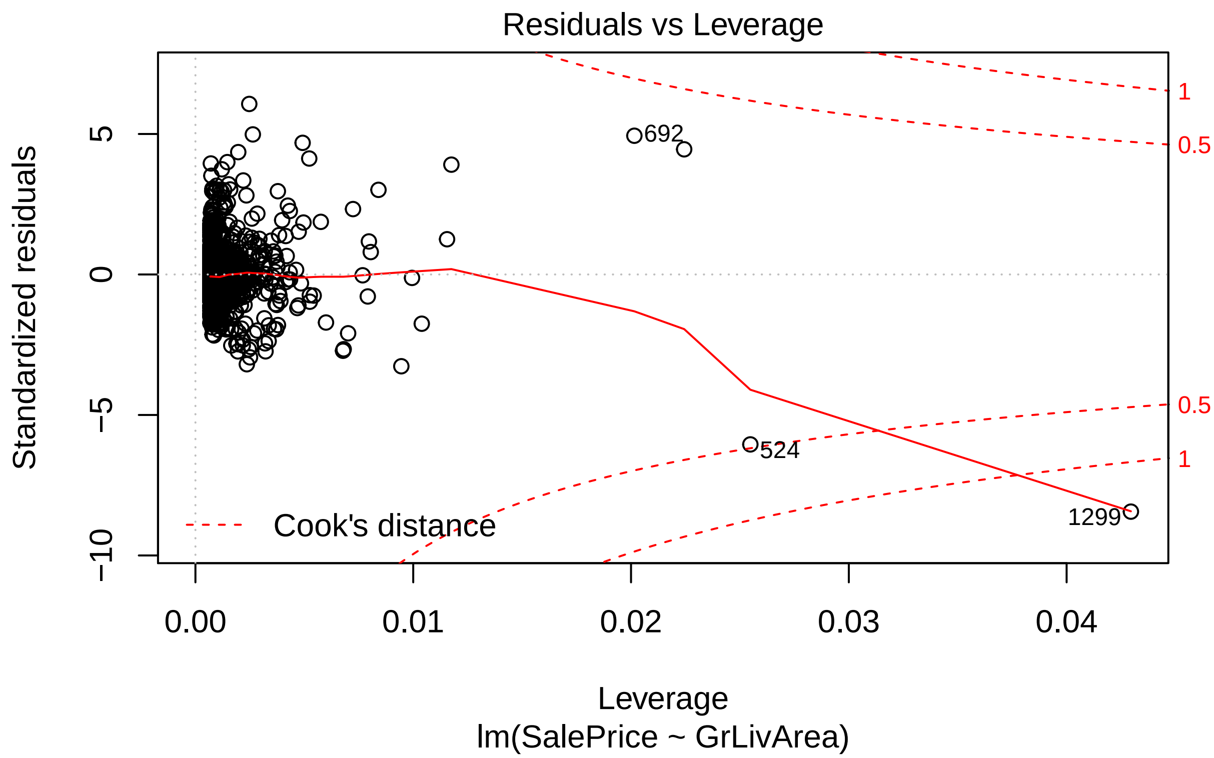
```r
plot(m)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SalePrice ~ GrLivArea)
899
524
1299
−5
0
5
−3
−2
−1
0
1
2
3



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
899
524
1299
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Fitted values
lm(SalePrice ~ GrLivArea)

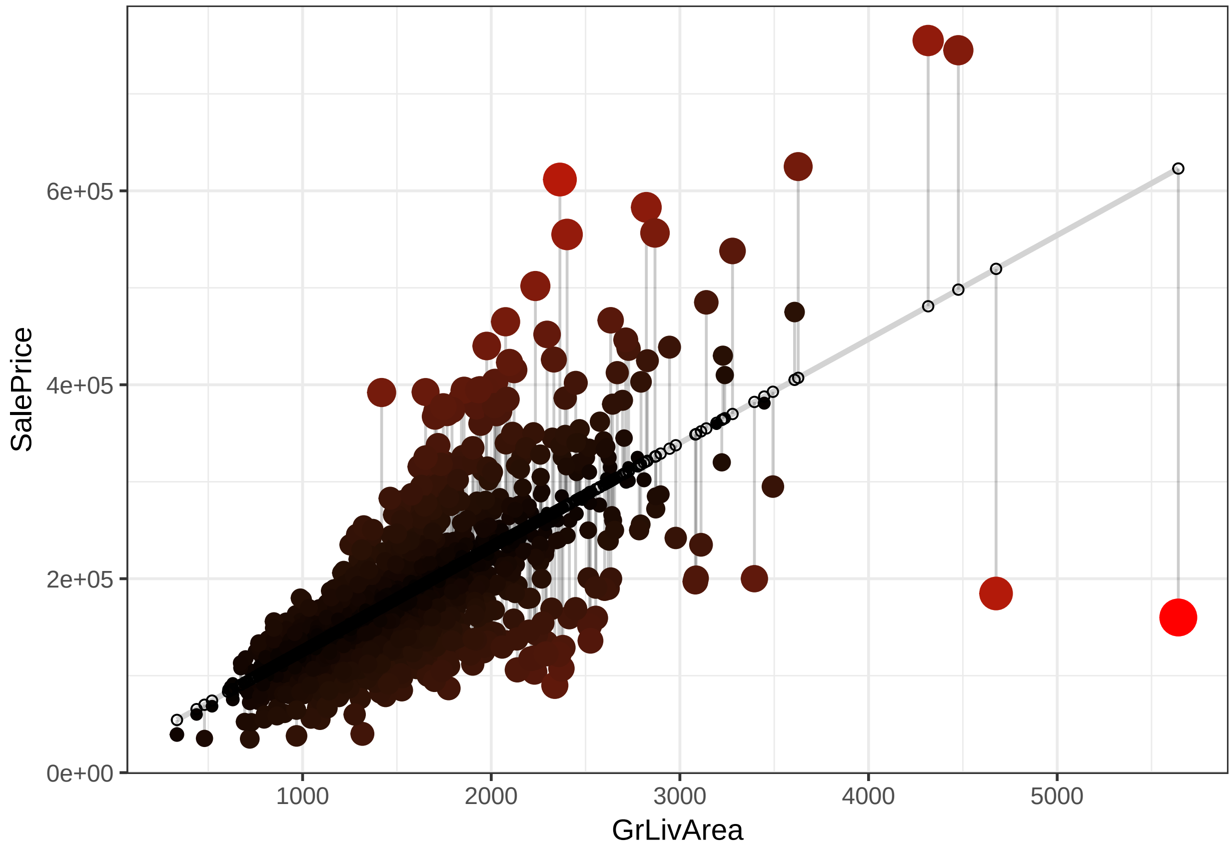



```
data$predicted <- predict(m)   # Save the predicted values
data$residuals <- residuals(m) # Save the residual values

ggplot(data, aes(x = GrLivArea, y = SalePrice)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  geom_segment(aes(xend = GrLivArea, yend = predicted), alpha = .2) +

  # > Color AND size adjustments made here...
  geom_point(aes(color = abs(residuals), size = abs(residuals))) + # size also mapped
  scale_color_continuous(low = "black", high = "red") +
  guides(color = FALSE, size = FALSE) +  # Size legend also removed
  # <

  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```



Looking the summary values and plots, we can determine the following:

The R-square value of ~0.5 is indicate nonsymmetrical residual even though the correlation is acceptable ~ 0.71 The P-value is very small and under 0.05 significance.

The cook's distance illustrates the points that influence our simple regression model result that is located farther away from the other points on the graph.

The residuals are not symmetric and localized which explained the low value of r-square. The difference between the fitting value and the predict values vary along the regression line.

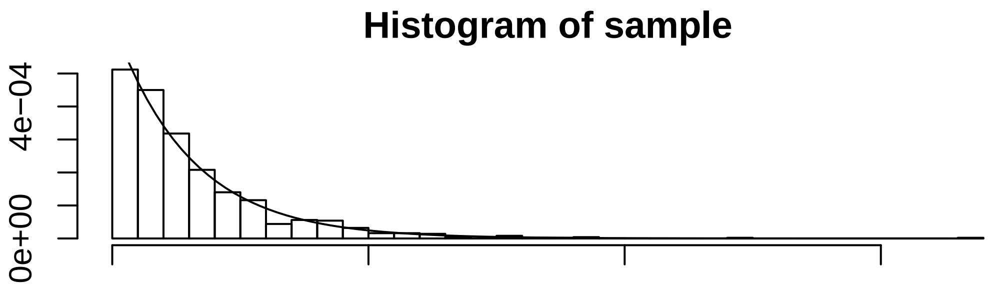
The qqplot indicate a right skewed distribution which would be an indication of using the square or exponential model will produce a better result.

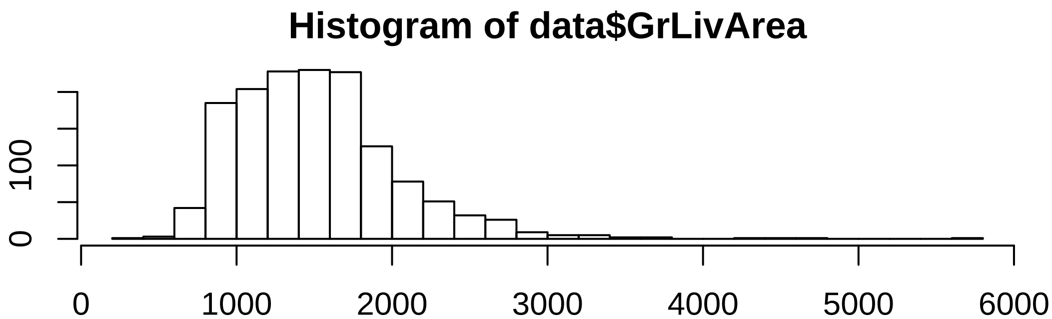
Based on the above results, we will do an exponential transformation model to see if have a better result.

```
X<-data$GrLivArea
fit <- fitdistr(X,"exponential")
lambda <- fit$estimate
sample <- rexp(1000,lambda)
```

```
# plot historgram of exponential function
par(oma=c(3,3,0,0),mar=c(3,3,2,2),mfrow=c(2,1))
hist(sample,prob=TRUE,breaks=25)
curve(dexp(x,lambda),add=T)

# plot histogram of original x
hist(data$GrLivArea, breaks=25)
```






```r
coef(fit)
```

```
##        rate
## 0.000659864
```

```r
print(fit)
```

```
##        rate
##   6.598640e-04
##  (1.726943e-05)
```

```r
vcov(fit)
```

```
##              rate
## rate 2.982333e-10
```

As shown above the density distribution exponential regression produce a very good fit to the datawith a small change of rate and very small covariance.