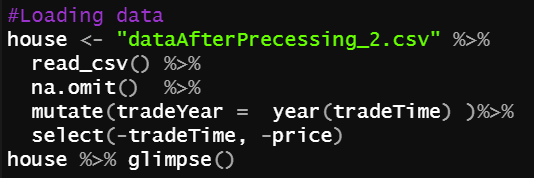
# Data Loading & Wrangling

Change all data type into numeric type then load CSV



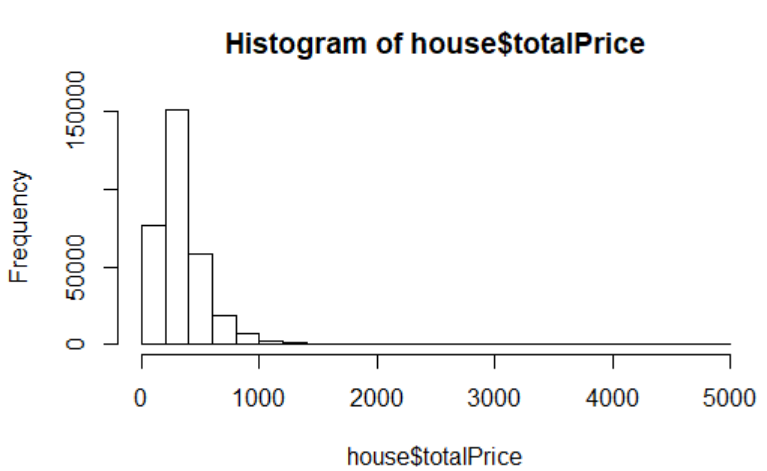
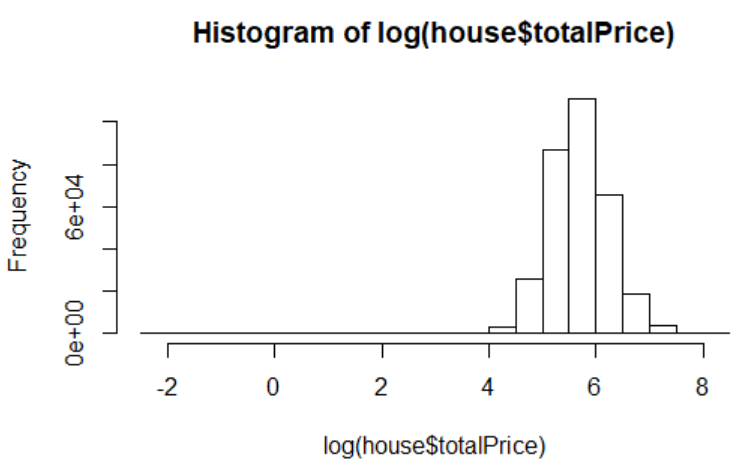
# Assumptions of Multiple Regression

Multiple linear regression analysis makes several key assumptions:

1. There must be a linear relationship between the outcome variable and the independent variables.
2. Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
3. No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
4. Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

Reference：<https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/>

# Normal Distribution Check

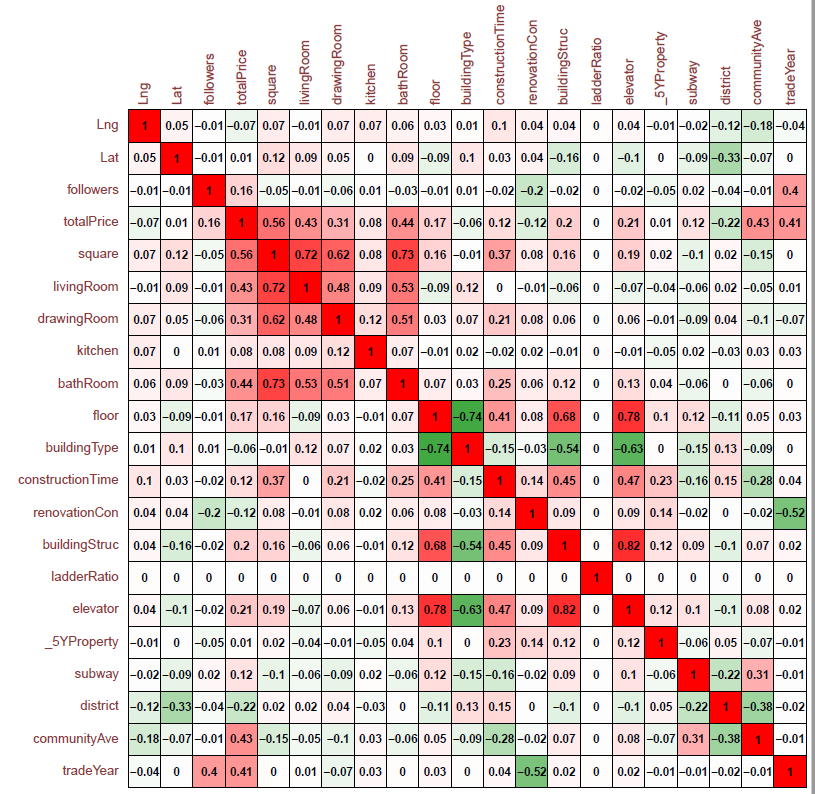
 

The ‘totoalPrice’ is not normal distribution, and log(totalPrice).is close to the bell shape,

So we assume that y=log(totalPrice)

# Check cor, R^2, and choose variables as X.

4.1 Cor Matric

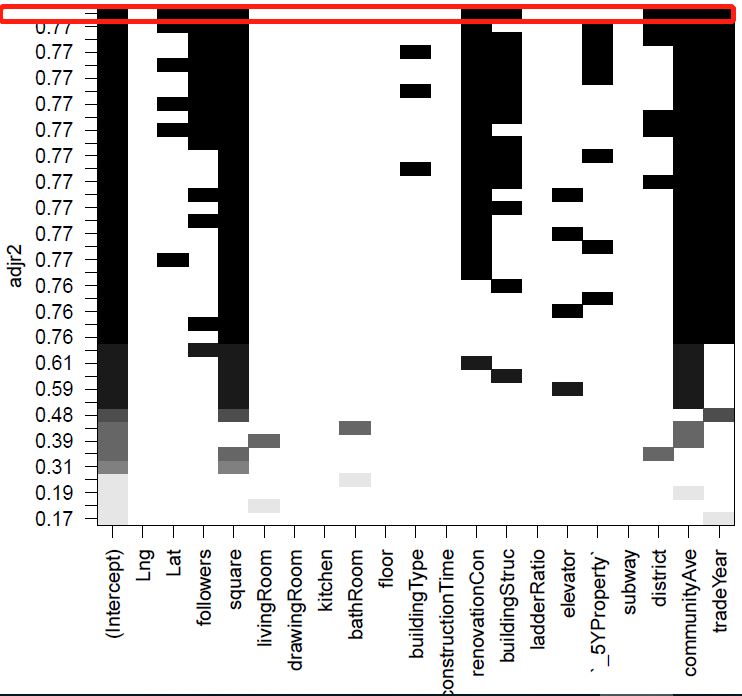


1. the 'square' has strong relationship with 'livingRoom', 'drawingRoom' and 'bathRoom'.

2. 'floor' has strong relationship with‘buildingType’， 'elevator' and 'buildingStructure'.

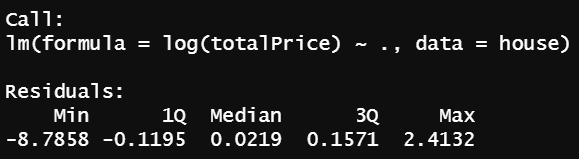
So we can keep 'square', 'floor', while remove the other variables mentioned above in the next step.

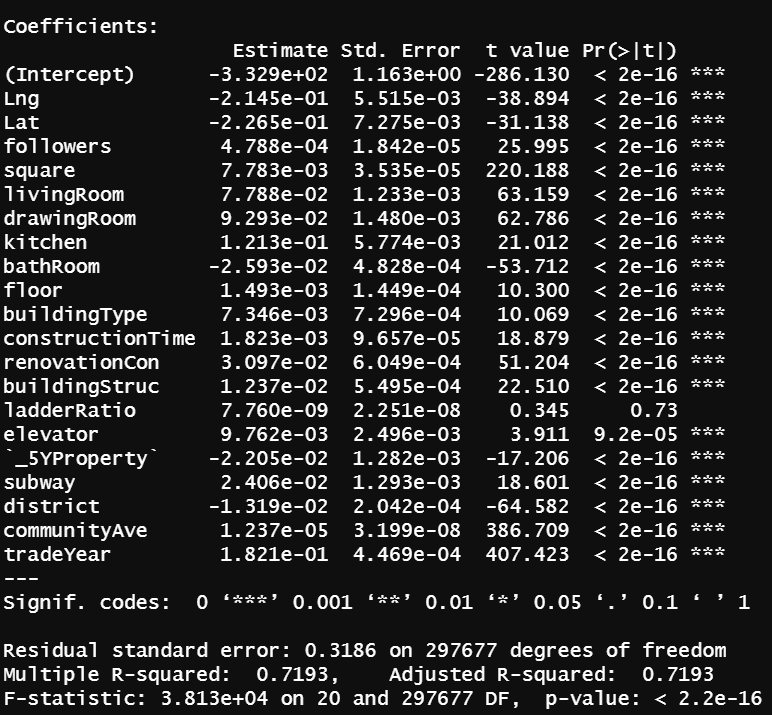
4.2 Adjust R^2

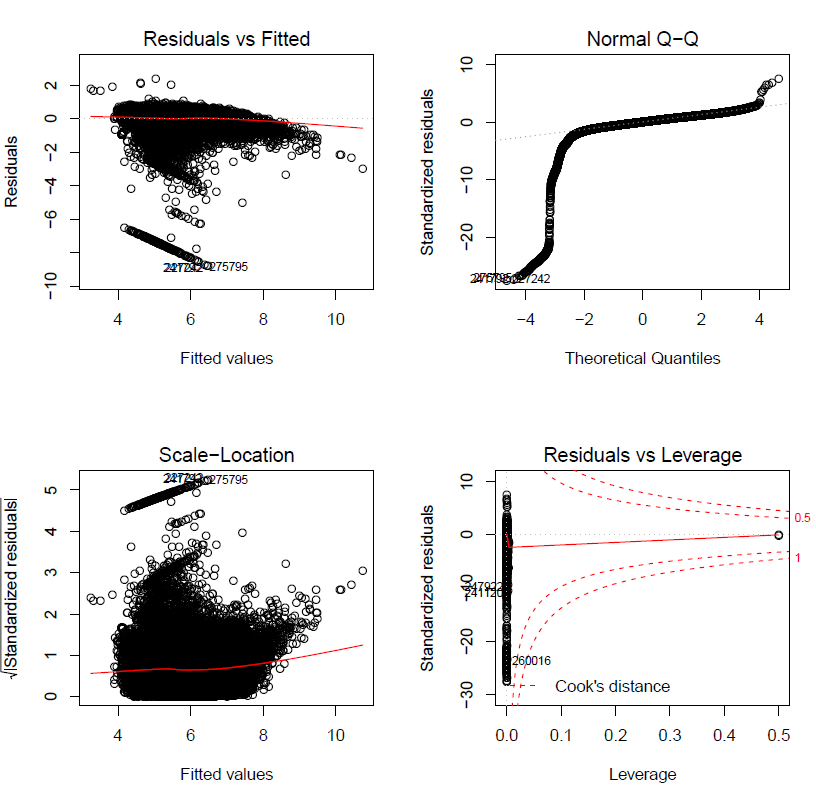


The best model for adjust R^2 is：log(totalPrice) ~ Lat + followers + square + renovationCondition + buildingStruc + district + communityAverage + tradeYear

4.3 All variables model:



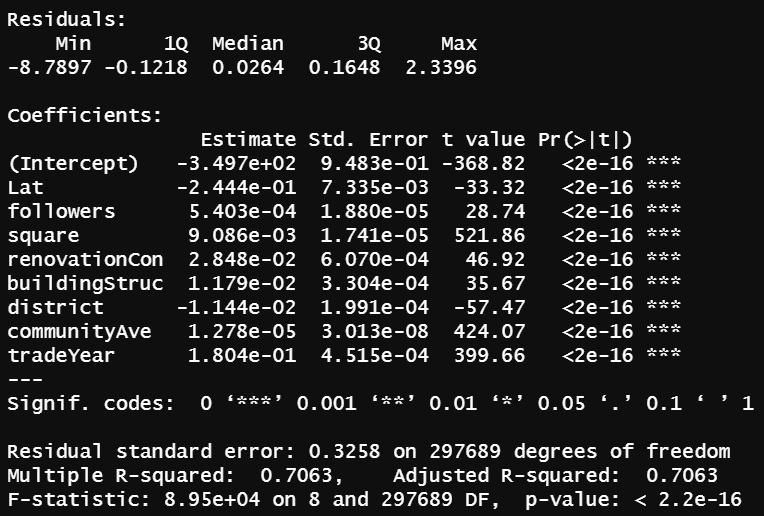


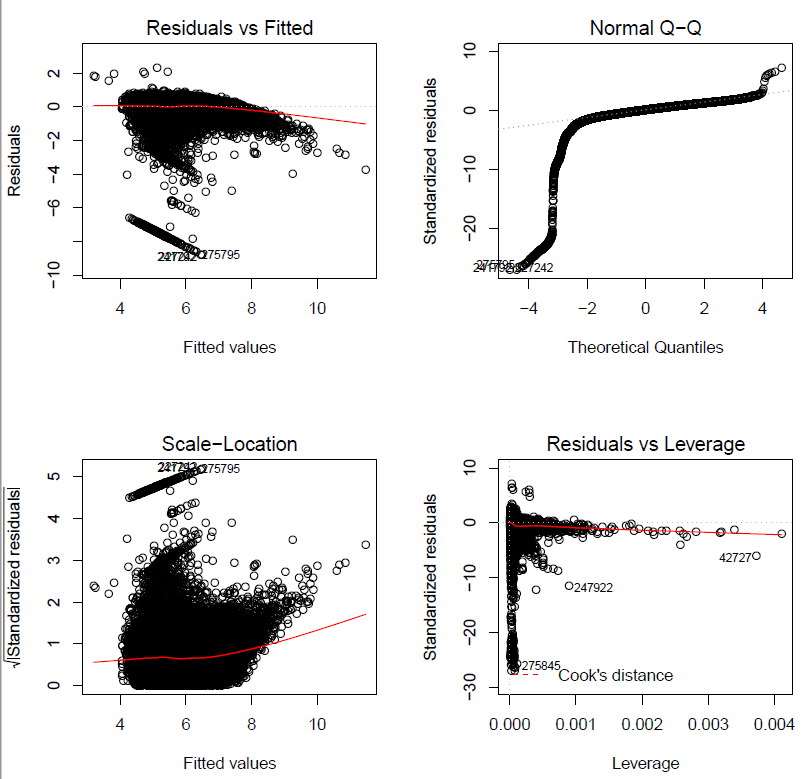


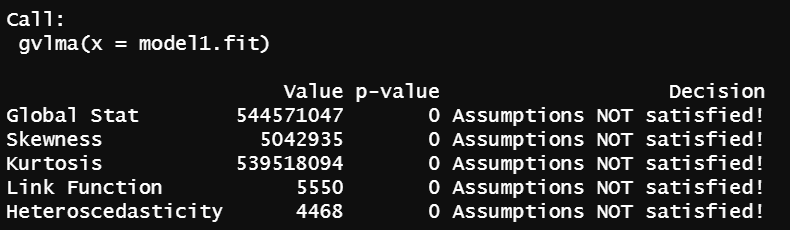
The all variables model does not match the assumptions of multiple regression. Let’s try the best model for adjust R^2.

# Model：log(totalPrice) ~ Lat + followers + square + renovationCondition + buildingStruc + district + communityAverage + tradeYear

5.1 Cor Matric

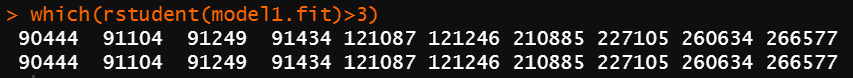


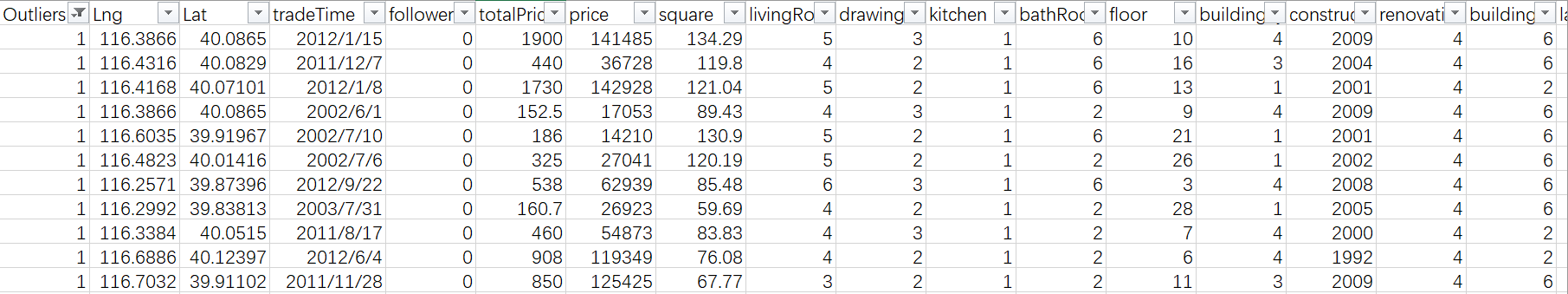




Global Stat is a comprehensive verification of the four assumptions. In this case, the model1 not pass. We can see that from the Normal Q-Q plot in the four plots as well, in which the points are not in a line. It means the residuals are not normally distributed, which doesn’t match the assumption 2.

Let’s find out the residual outliers:



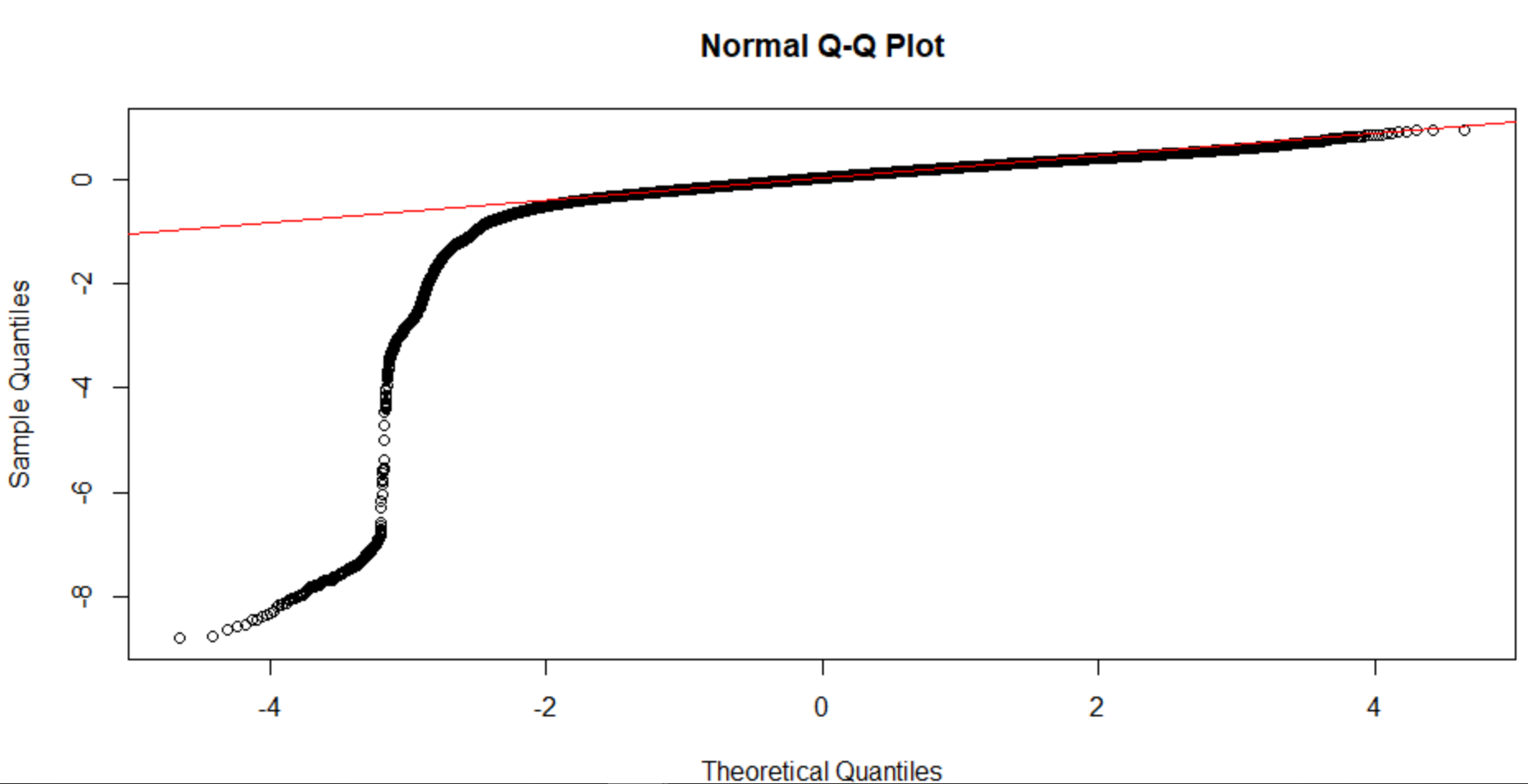


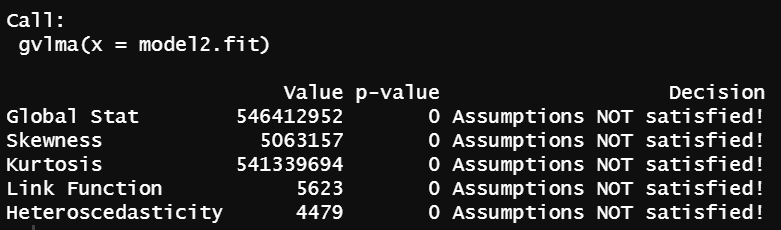
The outliers’ prices are ridiculous high.

5.2 Optimization Model (deleted residual outliers)

The model1 without outlier observations:







It still can’t match the four assumptions.

In conclusion, this dataset is not suitable for linear regression.

Reference:

* <http://www.xiaofandajie.top/2017/12/22/%E5%A4%9A%E5%85%83%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92%E5%88%86%E6%9E%90%E5%AE%9E%E4%BE%8B%E2%80%94%E2%80%94%E5%9F%BA%E4%BA%8ER/#%E6%95%B0%E6%8D%AE%E6%8E%A2%E7%B4%A2>
* <https://www.cnblogs.com/lafengdatascientist/p/5554167.html>