

Project Proposal - Predict House Price in Ames, Iowa

Yue Zhou

Monday, December 19, 2016

1. Introduction

In the housing bubble, house price rose beyond all reason, which has been a problem that affecting over half of U.S. states. Housing prices peaked in early 2006 and reached new lows in 2012 in U.S. Housing bubble has caused all sort of problems. One of them is causing a large number of home owners not able to pay their mortgages, also known as mortgage and credit crisis.

The goal of this project is to analyze the information of the house and to predict the final price of that house. It is important to have a sense of what variables would influence the housing price in order for residents to choose their real estate wisely. Moreover, predicting housing price accurately would help preventing the inability of home owners to pay their mortgages later on. Furthermore, since U.S. is not the only country that has been experiencing housing bubble, it is possible to apply analysis of house pricing dataset to other countries, such as China and Canada, to help residents in other countries as well.

2. Data Sets

The house price dataset is available on Kaggle website. This house price train dataset has 1460 data points consists of 81 explanatory variables describing almost every aspect of residential homes in Ames, Iowa. Variables include type of road access to property, type of utilities available, slope of property, roof material, etc.

3. Exploratory Data Analysis

Before modeling the Titanic data, it is necessary to first prepare the data, which includes data cleaning, missing data imputation, checking correlation between variables. I used package “mice” in R to summarize the missing data in Titanic dataset. I also use package “Amelia” to visualize the missing data. Roughly 99 percent of the PoolQC data is missing, 96 percent of the MiscFeature data is missing, 80 percent of the Fence data is missing, 94 percent of the Alley data is missing, etc. As shown in the chart below.

Variable Name	Percent of missing data
PoolQC	0.9952055
MiscFeature	0.9630137
Alley	0.9376712
Fence	0.8075342
FireplaceQu	0.4726027
LotFrontage	0.1773973
MasVnrType	0.005479452
MasVnrArea	0.005479452
BsmtQual	0.02534247
BsmtCond	0.02534247
BsmtExposure	0.0260274
BsmtFinType1	0.02534247
BsmtFinType2	0.0260274
Electrical	0.0006849315
GarageType	0.05547945

Variable Name	Percent of missing data
GarageYrBlt	0.05547945
GarageFinish	0.05547945
GarageQual	0.05547945
GarageCond	0.05547945

First of all, let's remove the variable "Id", since it is essentially recording the row number and would not add much value to our model.

In common sense, if the missing data is more than 5 percent of the whole dataset, we should leave the feature out. I will follow the thumb rule here and to discard the following variables: "Id", "PoolQC", "MiscFeature", "Alley", "Fence", "FireplaceQu", "LotFrontage", "MasVnrType", "MasVnrArea", "BsmtQual", "BsmtCond", "BsmtExposure", "GarageType", "GarageYrBlt", "GarageFinish", "GarageQual", "GarageCond"

For those variable that has missing data less than 5 percent of the whole dataset("Electrical", "BsmtFinType1", "BsmtFinType2", "BsmtExposure", "BsmtCond", "BsmtQual", "LotFrontage"), I would use the `complete.cases()` function in r to remove the rows that have missing datas. After removing rows with missing data and removing variables with missing data more than 5 percent of the whole dataset, we obtain a dataset with 1167 data points and 67 variables, which is not bad.

Since the housing dataset contains mostly the factor variables, and we want to have variables that is useful that provide information, we need to identify the variables that has only one level, which also have zero variance. After analyzing, we found that the variable Utilities has zero variance, therefore we remove this variable from data.



Figure 1: Caption for the picture.

Now we want to look at the relationship between each variables and check if there is some interactions that may be worth adding to the model. The visualization for correlation between all these variables is massive and hard to capture the important information from, thus we extract the variables with the highest ten correlations. The correlation plot are shown above.

Variable Exterior2nd (Exterior covering on house (if more than one material)) and Exterior1st (Exterior covering on house) tends to have really high correlation with each other, which makes sense because only we have certain Exterior covering on house, we could have more than one material for Exterior covering on house. The response variable SalePrice tends to have high correlations with variable OverallQual (Overall

material and finish quality), GarageCars (Size of garage in car capacity) and GarageArea(Size of garage in square feet), and 1stFlrSF(First Floor square feet). These findings follow common sense.

Variable1 Name	Variable2 Name	Correlation
Exterior2nd	Exterior1st	0.8615358
SalePrice	OverallQual	0.7993647
BldgType	MSSubClass	0.7577731
Foundation	YearBuilt	0.7191483
TotRmsAbvGrd	BedroomAbvGr	0.6547721
KitchenQual	ExterQual	0.6533928
SalePrice	GarageCars	0.6500270
BsmtFullBath	BsmtFinSF1	0.6477560
SalePrice	GarageArea	0.6309583
SalePrice	X1stFlrSF	0.6290195

We conduct a PCA analysis. Results are shown below We can conclude that a fair amount of variance is explained by the first five principal components and there is an elbow after the sixth component. After all, the sixth, seventh, eighth and ninth component explains less than 10 percent of the variance in the variance therefore not significant. Therefore, five components seem to be significant.

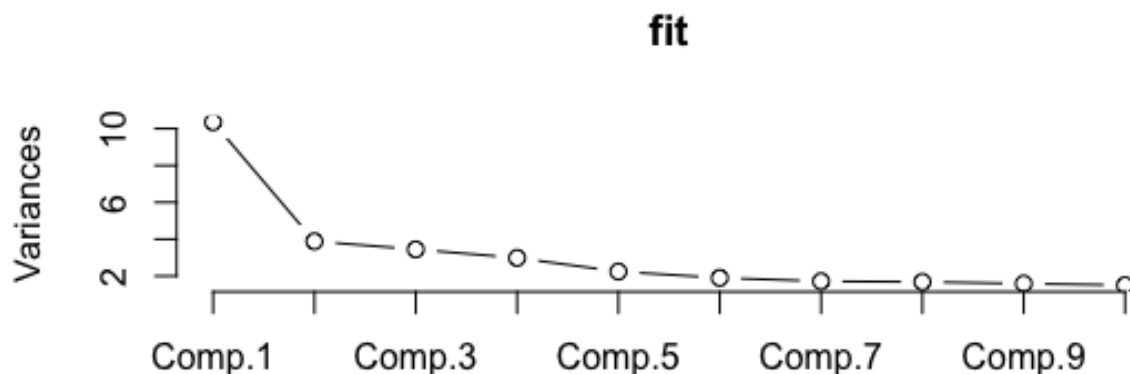


Figure 2: Caption for the picture.

4. Model Selection

We usually use Best Subset Selection to do model selection, however, since in this housing dataset we have too many variables to have every possible combination of models tested. We use Lasso to do model selection.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large.

The lowest MSE is achieved with large values of the penalty parameter lambda and the lasso regression gives us the model with all the variable, because lasso regression does not set any of the coefficient estimates to be exactly equal to zero.

5. Modeling

Let us first start by fitting the linear model for response variable SalePrice with all the other variables. By checking the t-value and p-value in the summary chart of linear model, we conclude that the variable MSSubClass, LotArea, LandContour, Neighborhood, Condition2, HouseStyle, OverallQual, OverallCond, RoofStyle, RoofMatl, ExterQual, BsmtQual, BsmtExposure, 1stFlrSF, 2ndFlrSF, BsmtFullBath, BedroomAbvGr, KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, GarageCars, SaleCondition, ScreenPorch, PoolArea are statistically significant at significance level of 0.05, for their p-value are less than 0.05.

Non-linear Effect

After checking the plot of the residuals with respect to each covariate including the variables that are not in the model, we see there are obvious non-linearity with respect to LotArea, LotFrontage, GarageArea and 1stFlrSF. as shown in the graph below.

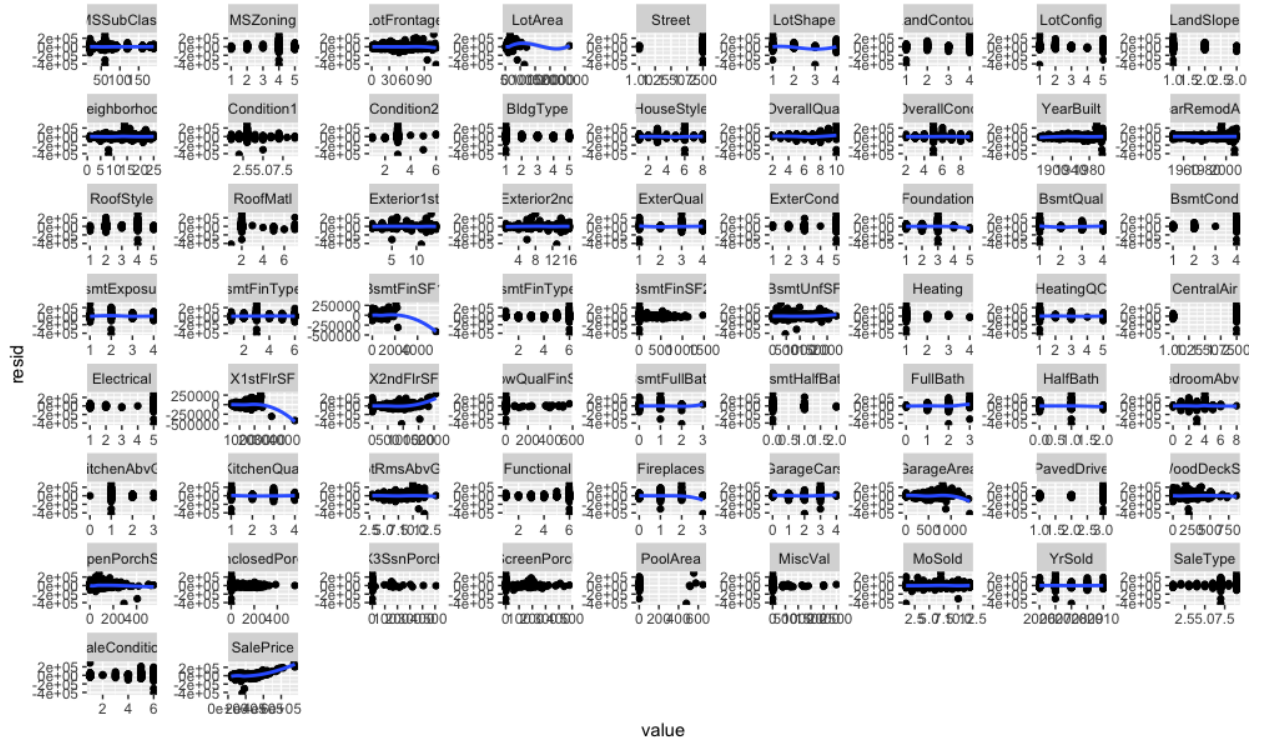


Figure 3: Caption for the picture.

We will go in that order and try polynomial regression with different terms for variable Neighborhood, LotFrontage, GarageArea as well as WoodDeckSF, and choose the order of polynomials using cross validation. By running cross validation, we believe that the 6th order of polynomials for variable GarageArea would give us the smallest MSE from cross validation. Similarly, 9th order of polynomials for variable LotFrontage, 9th order of polynomials for variable Neighborhood and 8th order of polynomials for variable WoodDeckSF would give us the smallest MSE from cross validation.

Instead of fitting a high-degree polynomial over the entire range of variables, splines regression involves fitting separate low-degree polynomials over different regions. Therefore, I also try spline for variable Neighborhood, LotFrontage, GarageArea as well as WoodDeckSF, and the spline model gives MSE of 0.02808415 from cross validation, which is much smaller than the MSE given by the polynomial regression, which is 0.156416

Check Assumption

In order to check if these p-values are meaningful, we will do some diagnostics. There are three major assumptions we need to check:

1. Residuals have a Gaussian distribution.
2. Residuals are homoskedastic (have the same variance).
3. Residuals are independent.

The diagnostics plots are at the end of the report. First four diagnostics plots are for linear model with response variable log-transformed. Then next four diagnostics plots are for polynomial regression. The last two diagnostics plots are for spline model.

For linear model with response variable log-transformed

The first graph check for the homogeneity of the variance and the linear relation, we see no pattern in this graph and then the assumption is met. The second graphs check for the normal distribution of the residuals, the points falls on the diagonol line. The third graph, the y-axis is changed, the residuals are square-root standardized making it easier to see heterogeneity of the variance. The fourth one allow detecting points that have a too big impact on the regression coefficients and that should be removed, i.e. outliers. From the fourth plot, we can see clear outliers, data point with row number 524, 1183 and 1299.

For polynomial regression

The first graph check for the homogeneity of the variance and the linear relation, we see no pattern in this graph and then the assumption is met. The second graphs check for the normal distribution of the residuals, the points falls on the diagonol line. The third graph, the y-axis is changed, the residuals are square-root standardized making it easier to see heterogeneity of the variance. The fourth one allow detecting points that have a too big impact on the regression coefficients and that should be removed, i.e. outliers. From the fourth plot, we can see clear outliers, data point with row number 633, 1325 and 524

For spline model

As the qqplot shown, there are heavy tails in ends. Therefore, we do not have strong evidence shown that residual is normal. As the third plot shown, we see a clear pattern which indicates heteroscedasticity.

We try removing these outliers and fitting the model again.

Below are the results from linear regression, polynomial regression, splines, and model with response variable transformed, so far linear regression with response variable log-transformed, outlier removed and spline with variable GarageArea, LotFrontage, Neighborhood, WoodDeckSF obtain the smallest cross validation MSE, 0.02779583. Therefore splines with outlier removed give us the best model.

Method	CV MSE
Linear Regression	1654684782
Polynomial Regression with log transformation	0.156416
Splines	0.02779583
model with response variable log transformed	0.02808415
Spline with outlier removed	0.02606843

Graphical Model Building

BsmtUnfSF

Variable BsmtUnfSF indicates the unfinished square feet of basement area and it is a continuous variable. For variable BsmtUnfSF, the best model selected by lasso regression includes all other variables, except for variable Exterior1st, SaleType, ScreenPorch, X2ndFlrSF, SalePrice, MiscVal, LowQualFinSF and GarageCars. After fitting a linear model to it, we find among all variables, variable PoolArea, OpenPorchSF, KitchenAbvGr, X1stFlrSF, BsmtFinSF2, BsmtFinSF1, BsmtFinType1, ExterQual, YearBuilt, HouseStyle, LotArea and Functional are with most significant small p-value, indicates most statistically significant at significance level of 0.05. The model has an adjusted R-square of 0.9019.

LotArea

Variable LotArea indicates the lot size in square feet and it is a continuous variable. For variable LotArea, the best model selected by lasso regression includes all other variables. After fitting a linear model to it, we find among all these variables, variable BsmtFinSF1, LandSlope, LotFrontage, BsmtFinSF2, Heating and BsmtUnfSF are with most significant small p-value, indicates most statistically significant at significance level of 0.05. The model has an adjusted R-square of 0.3602.

KitchenQual

Variable KitchenQual indicates the kitchen quality and it is a categorical variable. Therefore we fit a logistic regression to it. For variable KitchenQual, the best model selected by lasso regression includes all other variables. After fitting a linear model to it, we find among all these variables, variable HouseStyle, YearRemodAdd, BsmtQual, WoodDeckSF, TotRmsAbvGrd, SaleCondition, Electrical and ExterQual are with significant small p-value, indicates statistically significant at significance level of 0.05. The model has an AIC of 396.8

Logistic Regression

BsmtHalfBath

We choose the variable to be BsmtHalfBath, the results only take three values, we need to fit logistic regression. I pass the features provided in the training dataset through the R function to fit general linearized models. By setting the argument family to binomial with a logit link. I am asking glm() to produce a logistic regression. After checking the z-values and associated p values, we see that variable Condition1, OverallCond, Exterior1st, Exterior2nd, BsmtQual, BsmtFinSF1, BsmtFullBath, FullBath, X3SsnPorch, and BedroomAbvGr are statistically significant. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. For example, for every one unit change in BsmtFullBath, the log odds of admission (versus non-admission) decreased by 4.473.

To assess the first model and the logistic regressions, we will use the chi-square statistics, which is a measure of the goodness of fit of observed value to expected values. The smaller the difference (or deviance) of the observed value from the expected values, the better the fit of the model. The null deviance shows how well BsmtHalfBath is predicted by a null model using only grand mean. As we adjust the model's formula by adding or removing variables, we'll look for those changes which decrease in the residual deviance, indicating an improvement in fit. The deviance was reduced by 199.39 points from 490.34 to 290.95 on $1166 - 1102 = 64$ degrees of freedom (DF), which is a significant reduction.

The summary of the fit can be found at the last page of this report.

We need to compare the result we obtained from the above five methods. Instead of calculating MSE for these models, we usually get the confusion matrix for Logistic regression, which is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. But it is not necessary for these questions, and we will go ahead and use cv.

The chart below shows 5 models with different variables and I evaluate all of them using CV. As the chart shown, the model with variables Condition1, BsmtFullBath ,FullBath, X3SsnPorch, and BedroomAbvGr gives the smallest CV mse, thus consider to be the best model in fitting response variable BsmtHalfBath.

Model	CV mse
With All Variables	0.05047529
With Variables Condition1, BsmtFullBath ,FullBath, X3SsnPorch, and BedroomAbvGr	0.05045228
With Variables SaleType, SalePrice, FullBath, X3SsnPorch, and BedroomAbvGr	0.05181958
With Variables YearRemodAdd, CentralAir, RoofStyle, Fireplaces, and Neighborhood	0.05213484
With Variables Street, Heating, BsmtCond, and OpenPorchSF	0.05130155

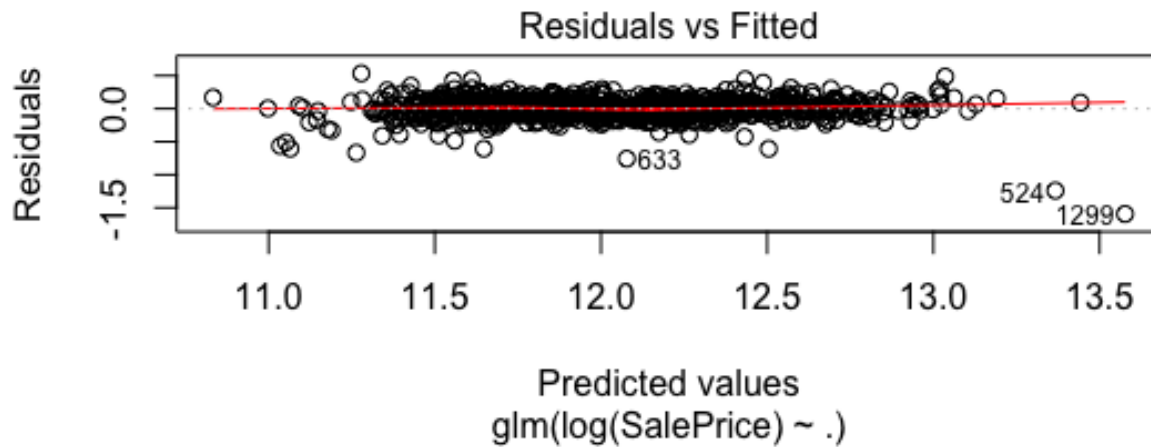


Figure 4: Caption for the picture.

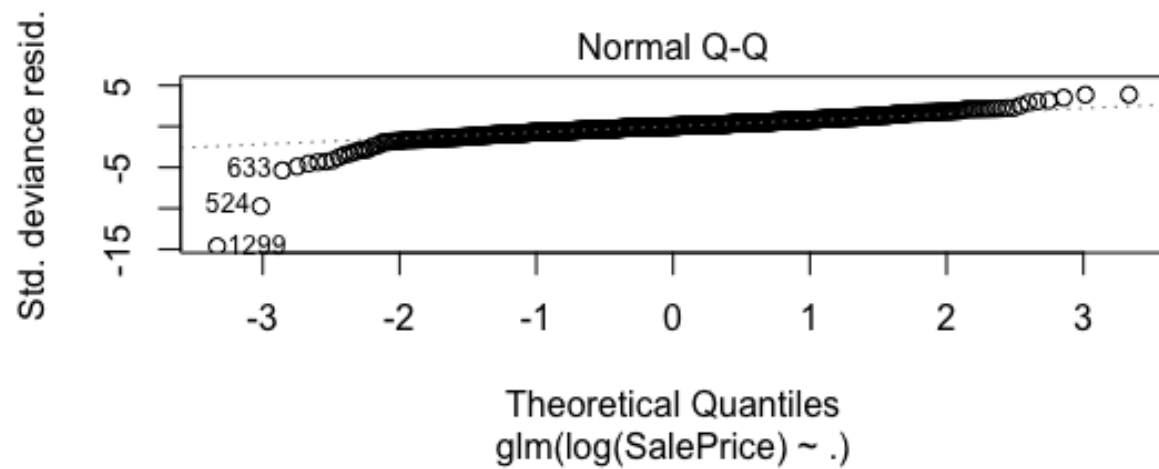


Figure 5: Caption for the picture.

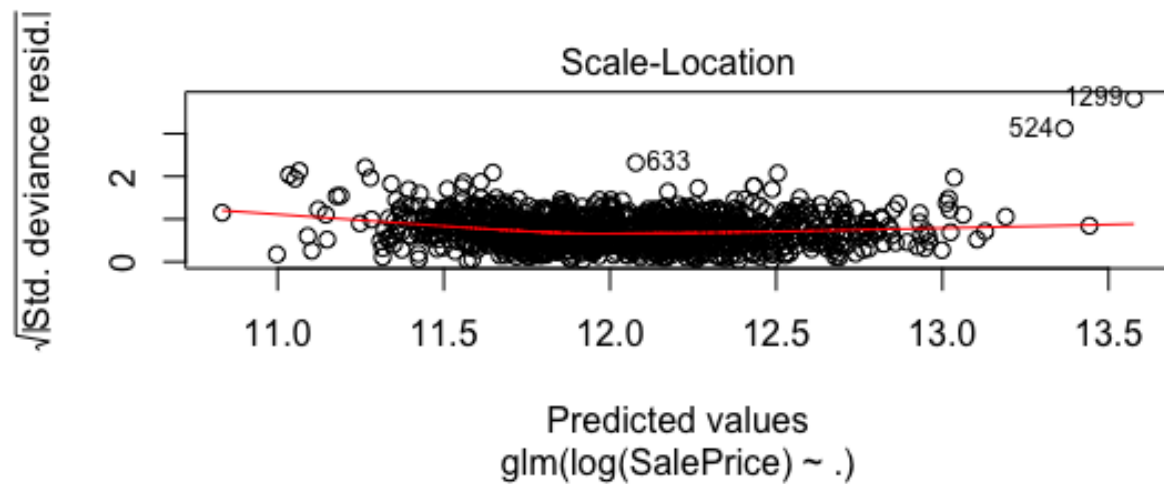


Figure 6: Caption for the picture.

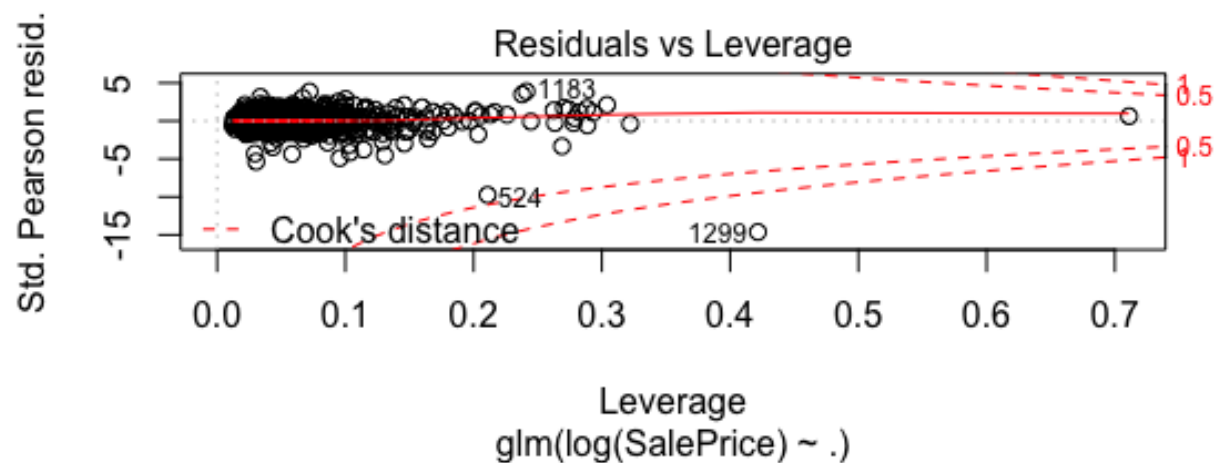


Figure 7: Caption for the picture.

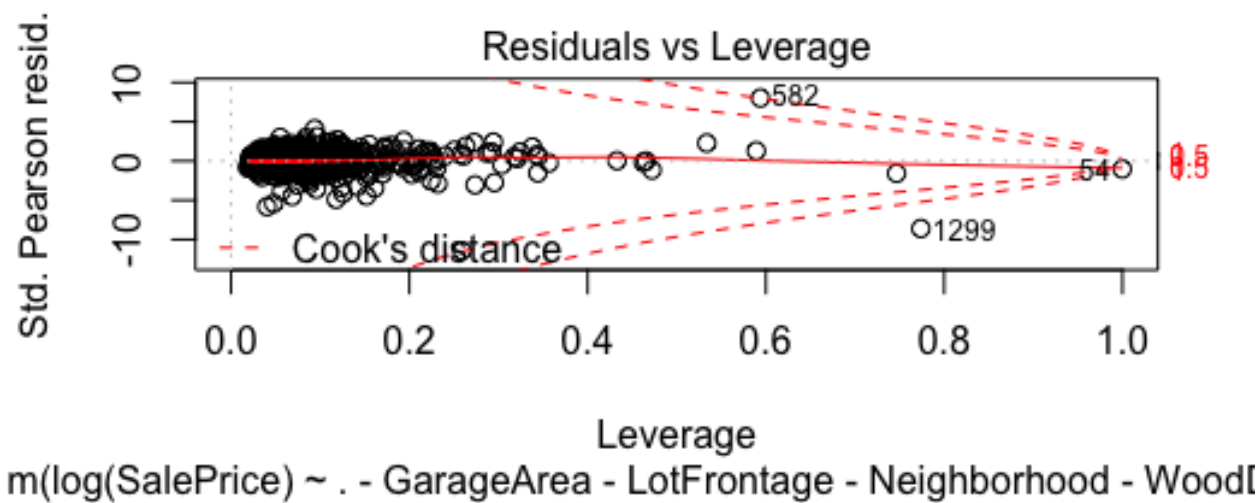


Figure 8: Caption for the picture.

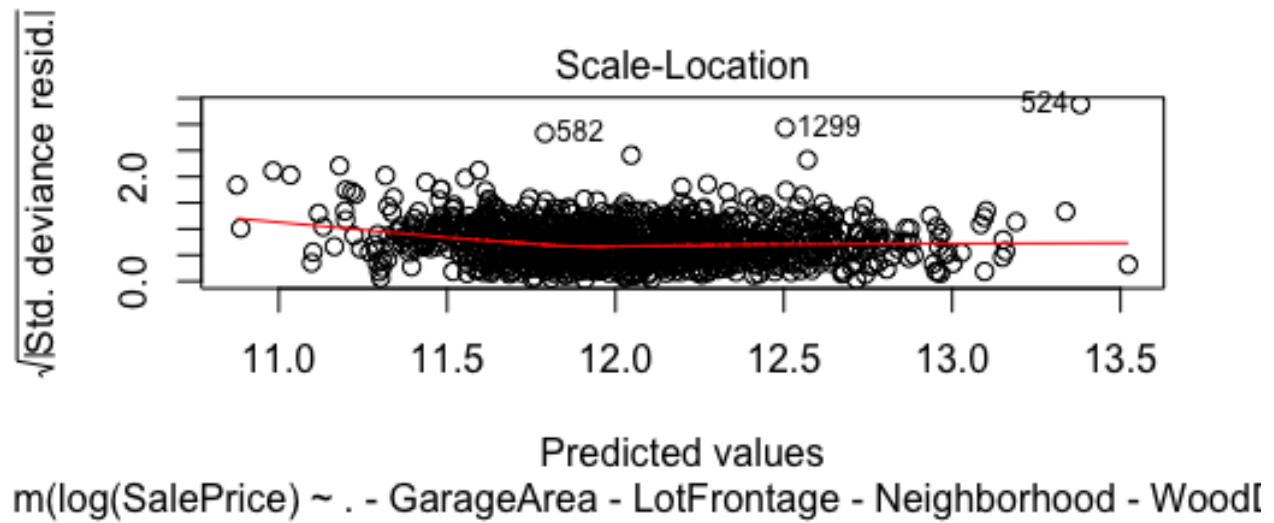


Figure 9: Caption for the picture.

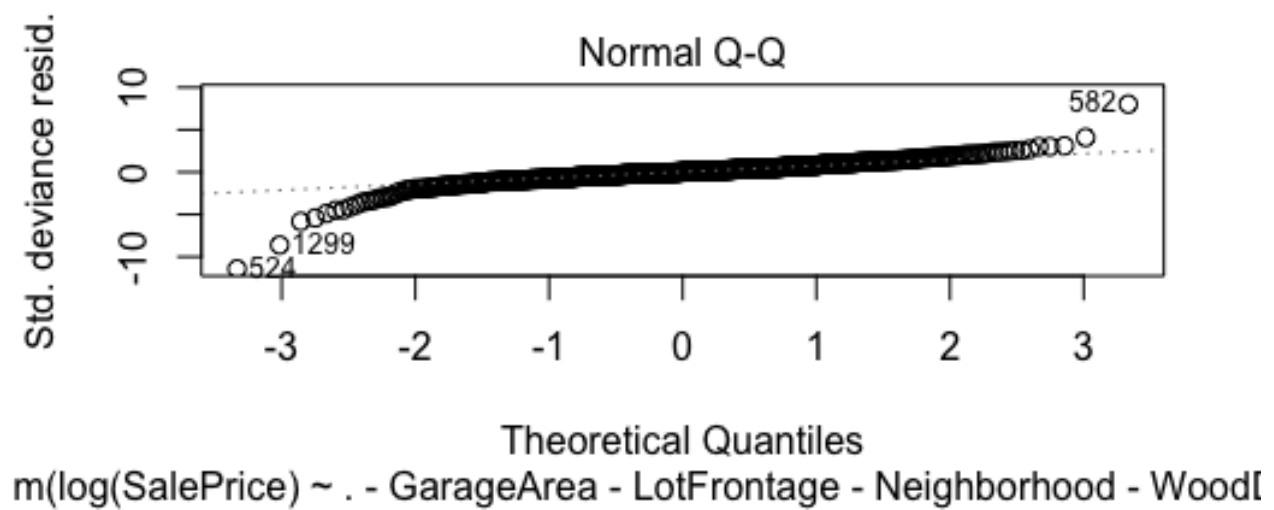


Figure 10: Caption for the picture.

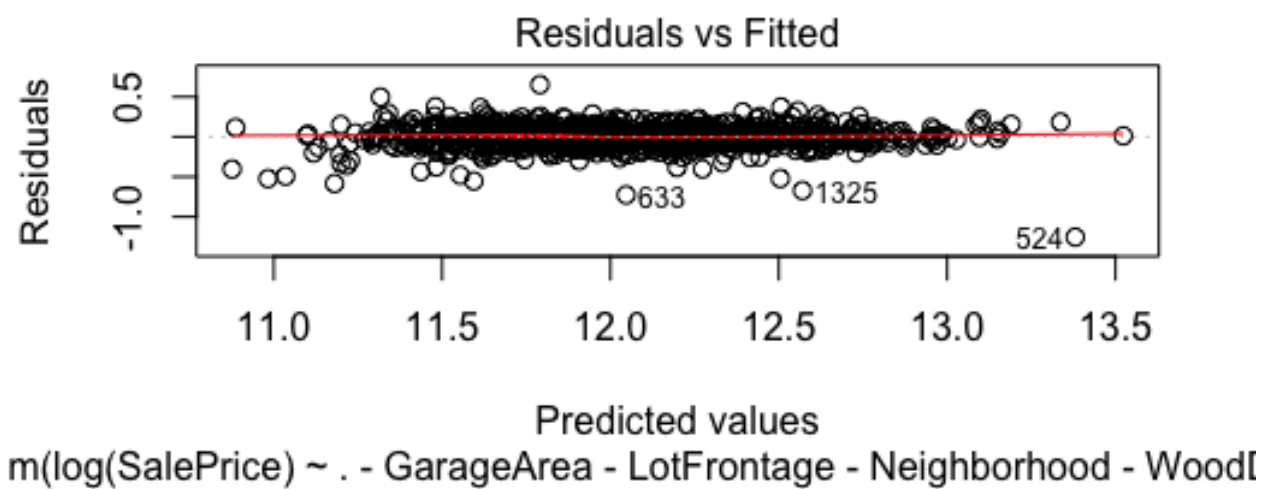


Figure 11: Caption for the picture.

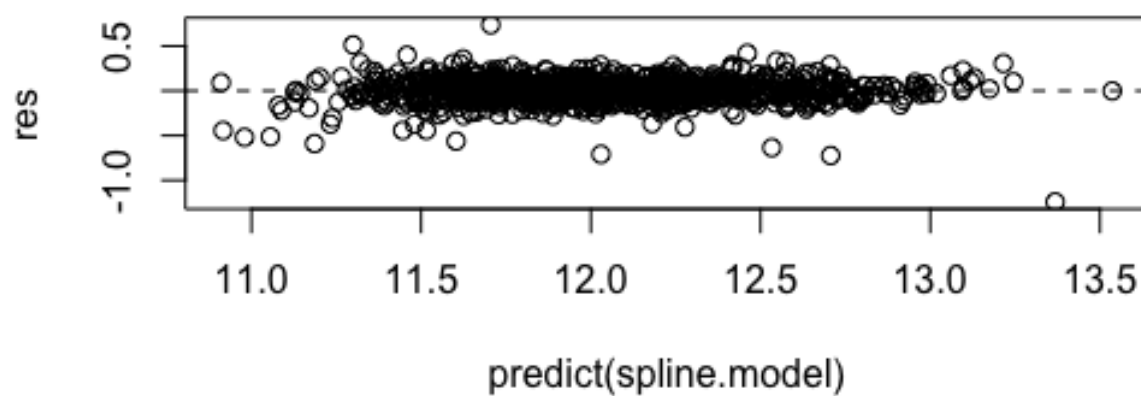


Figure 12: Caption for the picture.

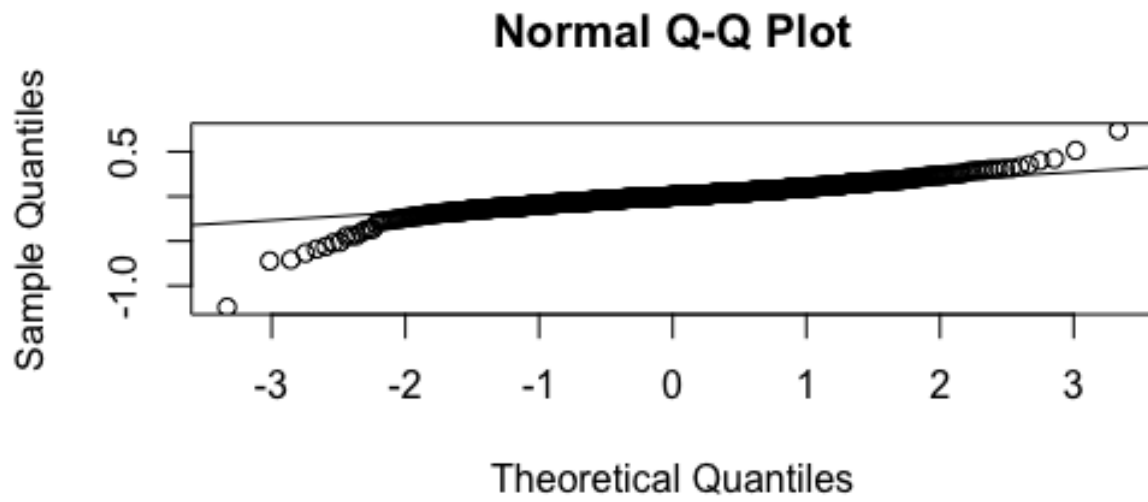


Figure 13: Caption for the picture.

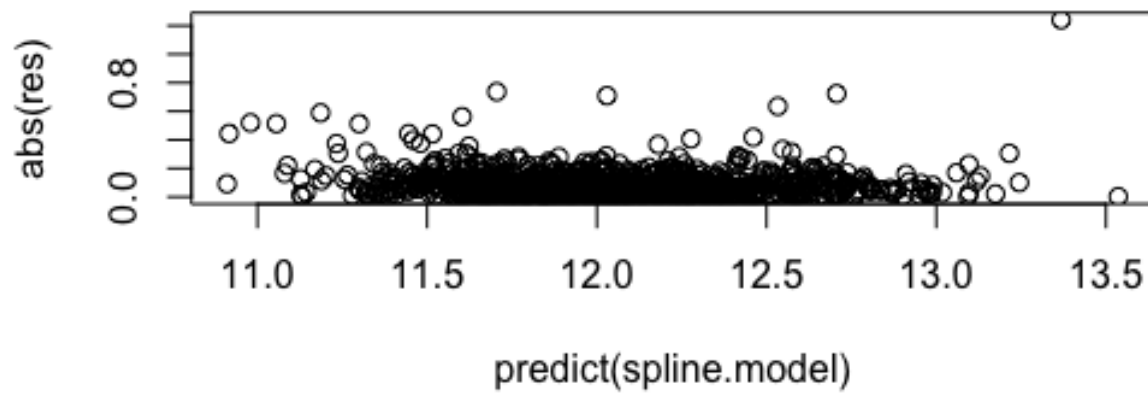


Figure 14: Caption for the picture.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	191.0901	1804.6448	0.11	0.9157
MSSubClass	0.0008	0.0132	0.06	0.9531
MSZoning	-0.3695	0.3378	-1.09	0.2740
LotFrontage	0.0049	0.0123	0.40	0.6894
LotArea	-0.0000	0.0000	-0.33	0.7423
Street	11.6357	891.4370	0.01	0.9896
LotShape	-0.0472	0.1461	-0.32	0.7465
LandContour	0.2426	0.3055	0.79	0.4270
LotConfig	0.1159	0.1208	0.96	0.3373
LandSlope	0.7255	0.7052	1.03	0.3036
Neighborhood	-0.0518	0.0361	-1.43	0.1515
Condition1	-1.0582	0.3388	-3.12	0.0018
Condition2	0.0863	1.7298	0.05	0.9602
BldgType	0.2750	0.3914	0.70	0.4823
HouseStyle	0.0111	0.1468	0.08	0.9399
OverallQual	0.0802	0.2640	0.30	0.7613
OverallCond	0.3952	0.2001	1.98	0.0482
YearBuilt	0.0140	0.0156	0.90	0.3673
YearRemodAdd	0.0152	0.0128	1.19	0.2359
RoofStyle	0.2025	0.2117	0.96	0.3388
RoofMatl	0.6072	0.3203	1.90	0.0580
Exterior1st	0.2309	0.1134	2.04	0.0417
Exterior2nd	-0.2060	0.0961	-2.14	0.0321
ExterQual	0.6942	0.4517	1.54	0.1244
ExterCond	-0.3989	0.2223	-1.79	0.0728
Foundation	-0.1708	0.4144	-0.41	0.6803
BsmtQual	-0.5639	0.2836	-1.99	0.0468
BsmtCond	-0.3793	0.2322	-1.63	0.1024
BsmtExposure	-0.0760	0.1844	-0.41	0.6802
BsmtFinType1	-0.1520	0.1164	-1.31	0.1915
BsmtFinSF1	0.0021	0.0010	2.09	0.0369
BsmtFinType2	-0.0944	0.2431	-0.39	0.6979
BsmtFinSF2	0.0020	0.0018	1.15	0.2502
BsmtUnfSF	-0.0013	0.0010	-1.26	0.2080
Heating	1.3313	1.0027	1.33	0.1843
HeatingQC	-0.0715	0.1136	-0.63	0.5293
CentralAir	1.3236	1.4967	0.88	0.3765
Electrical	0.1812	0.2052	0.88	0.3773
X1stFlrSF	-0.0001	0.0013	-0.10	0.9233
X2ndFlrSF	-0.0005	0.0011	-0.47	0.6414
LowQualFinSF	0.0002	0.0038	0.06	0.9539
BsmtFullBath	-4.4733	0.6430	-6.96	0.0000
FullBath	-1.6514	0.5475	-3.02	0.0026
HalfBath	-0.3587	0.5337	-0.67	0.5015
BedroomAbvGr	1.0097	0.3675	2.75	0.0060
KitchenAbvGr	0.8368	1.3728	0.61	0.5421
KitchenQual	0.0734	0.3198	0.23	0.8184
TotRmsAbvGrd	-0.2784	0.2724	-1.02	0.3069
Functional	-0.1452	0.2247	-0.65	0.5180
Fireplaces	0.4374	0.3452	1.27	0.2051
GarageCars	0.3320	0.5509	0.60	0.5467
GarageArea	0.0004	0.0021	0.19	0.8496
PavedDrive	0.2707	0.5964	0.45	0.6499
WoodDeckSF	0.0028	0.0015	1.89	0.0591
OpenPorchSF	0.0022	0.0033	0.68	0.4996
EnclosedPorch	0.0006	13 0.0030	0.22	0.8289
X3SsnPorch	0.0094	0.0034	2.79	0.0053
ScreenPorch	-0.0013	0.0034	-0.40	0.6904
PoolArea	0.0020	0.0030	0.64	0.5211