House Price Prediction Using MLR

Group Members

Aditya Jain (160047)

Shyam Sundhar R (150709)

Venkatramana Manikanta Patnana (171173)

April 12, 2019

Acknowledgement

We would like to thank Dr. Sharmishtha Mitra who gave us an opportunity to apply the theory and concepts we learnt in regression analysis course to a practical problem. The project exposed us to practical issues and challenges in applying regression analysis techniques and tackling them in a systematic way.

About the Data Set

The data set for the project has been sourced from a Kaggle Competition[1]. The description of the data fields (regressors) is as follows:

LotArea : Area of the Lot

OverallQual : Overall Quality of the House

YearBuilt : Built in year

GrLivArea : Ground Floor Living Area

BedroomAbvGr : Number of bedrooms above ground

KitchenAbvGr : Kitchen Above Ground

TotRmsAbvGrd : Total number of rooms above ground

GarageArea : Area of the Garage

WoodDeckSF : Surface Area of the Wood Deck

TotalBsmtSF : Surface Area of the Basement

1stFlrSF : Surface Area of the 1st Floor

HouseStyle : Style of the House

RoofStyle : Style of the Roof

Abstract

An MLR (Multiple Linear Regression) model is fitted to predict the price of a house based on various features as mentioned in "About the Data Set".

Various regression analysis techniques involving exploratory data analysis, calculation of various parameters of the fit, statistics and hypothesis tests are performed.

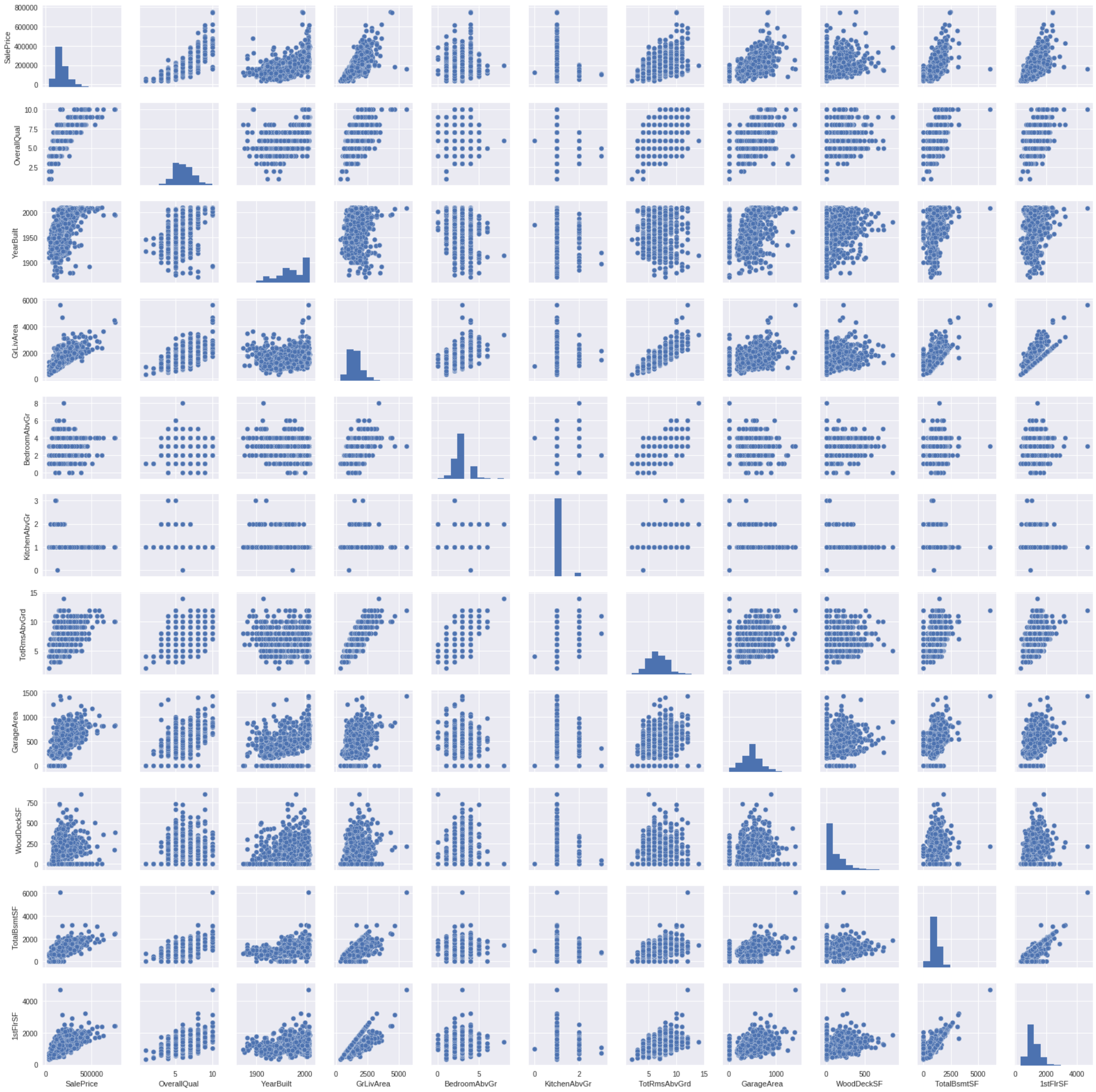
Then, model adequacy checks of the fitted MLR model like normality assumption check, residual analysis are performed followed by suitable transformation of variable and outlier treatment.

Then, the multi-collinearity check is performed, the categorical variables are handled using dummy variables and variable selection using backward elimination is performed.

In the end, graphical analysis of residuals and Q-Q plot is done based on which the adequacy of the fit of the regression model is concluded.

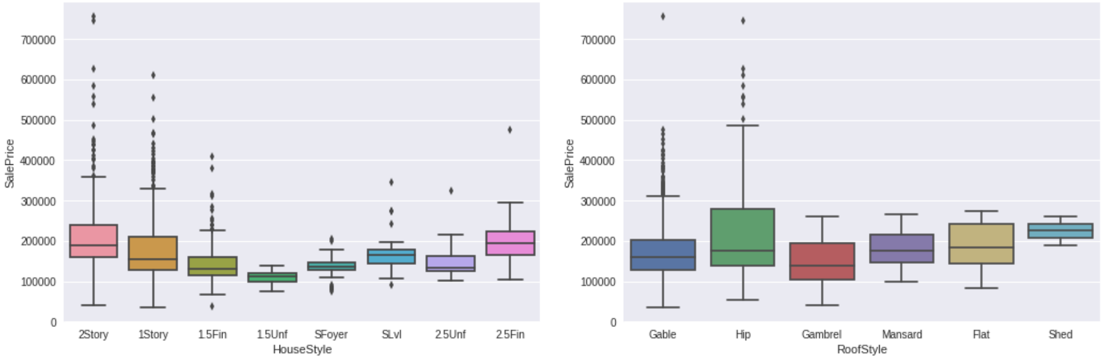
Exploratory Data Analysis :

The plot below shows the pairwise-relationships between the **continuous regressors** in the model.



Some interesting Obervations :
'SalePrice' vs 'GrLivArea' : Linear relationship
'SalePrice' vs 'TotalBsmtSF' : Linear relationship
'SalePrice' vs '1stFlrSF' : Linear relationship
'GrLivArea', 'TotRmsAbvGrd', 'TotalBsmtSF', '1stFlrSF' : Linear relationship (Possible Multicollinearity)

The plot below shows the box-plots corresponding to the **categorical variables**.



If we go by the traditional dummy variable encoding for both the categorical variables, we will end up with 48 (8*6) categories. Since, the number of observations in our dataset is small, such an encoding faces the curse of dimensionality. Therefore, we come up with a different kind of encoding (mean encoding) for one of the categorical variable. Mean encoding replaces each category value by the mean of the target values corresponding to the same category value.

Since ”HouseStyle” variable has 8 categories and ”RoofStyle” has 6 categories, so we choose to mean encode the former and dummy variable encoding for the latter.

Log transformation of target variable

We first fit the OLS Model without any transformation and check all the statistics of the model.

R^2	0.78603
$Adj R^2$	0.784678
MS_{res}	1358913886.616555
F_{obs}	484.357672

Since the F-statistic is large we conclude that the regression is significant. We now check the significance of each variable.

Regressor	t-statistic
LotArea	1.06880866e-04
OverallQual	5.13651391e+01
YearBuilt	1.43017970e-02
GrLivArea	1.86855809e-01
BedroomAbvGr	2.76066778e+00
KitchenAbvGr	3.56476437e-03
TotRmsAbvGrd	1.14209078e-03
GarageArea	3.55732715e-03
WoodDeckSF	1.02782002e+00
TotalBsmtSF	4.85946903e-01
1stFlrSF	9.08977020e-01

Note: for alpha = 0.05, p = 11, n-p-1 = 1448, t-value is 1.96

This implies that only two of the regressors are significant which is very poor because we chose many important regressors based on the domain knowledge. So, it implies that the normal error assumptions are not even close.

Therefore, we try to transform the response variable by taking log(”SalePrice”+1) to match the scaling between the response variable and the regressors. Repeating the same procedure as before, we get the following statistics of the transformed model :

R^2	0.823297
$Adj R^2$	0.821955
MS_{res}	0.028409
F_{obs}	613.322910

Since the F-statistic is large we conclude that the regression is significant. We now check the significance of each variable.

Regressor	t-statistic
LotArea	4.18681536e+01
OverallQual	1.26021194e+07
YearBuilt	3.71751206e+03
GrLivArea	6.86173978e+04
BedroomAbvGr	5.26415845e+05
KitchenAbvGr	6.81221871e+01
TotRmsAbvGrd	2.70555447e+02
GarageArea	5.90952525e+02
WoodDeckSF	3.06531675e+05
TotalBsmtSF	1.26443667e+05
1stFlrSF	1.56407177e+05

We now notice that all the regressors have become significant after log transformation of the response variable which is a good sign. Hence, we now proceed to work on this transformed model.

Residual Analysis (before Variable Selection)

The aim of doing residual analysis before doing any further analysis is to clearly see if there is some graphical pattern between the predicted responses and the residuals and to check for outliers which causes disruption to the regression line. We further check the graph of residuals and regressors for heteroscedasticity so that the equation can be scaled w.r.t to that particular regressor. As already mentioned we check whether the log transformation of the regressand is justified based on the residual analysis.

We first fit the data to the log(house price+1) and find the residuals. And then we find the standardized, student residuals and R-student residuals to find the outliers.

Outliers

Outliers identified by :

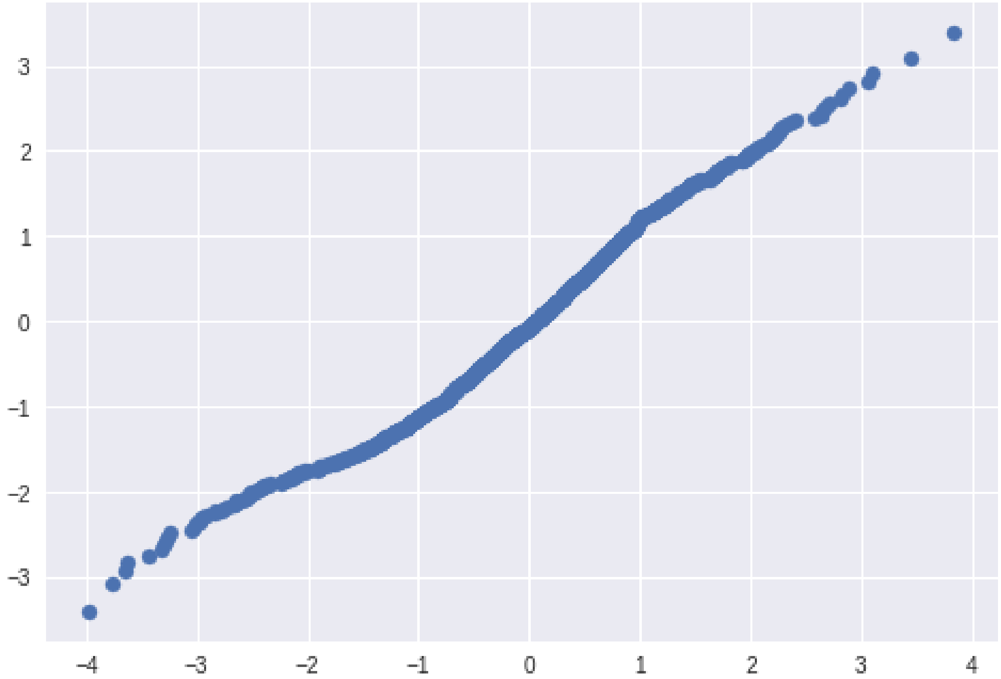
Standardized residuals	student residuals	R-student residuals
30	30	30
218	218	218
398	398	398
410	410	410
462	462	462
495	495	495
523	523	523
632	632	632
812	812	812
916	916	916
968	968	968
1298	1298	1298
1324	1324	1324

Since all the methods give the same indices of outlier observations, we delete the corresponding observations from our dataset and fit OLS again.

R^2	0.881746
$Adj R^2$	0.880839
MS_{res}	0.018022
F_{obs}	972.717496

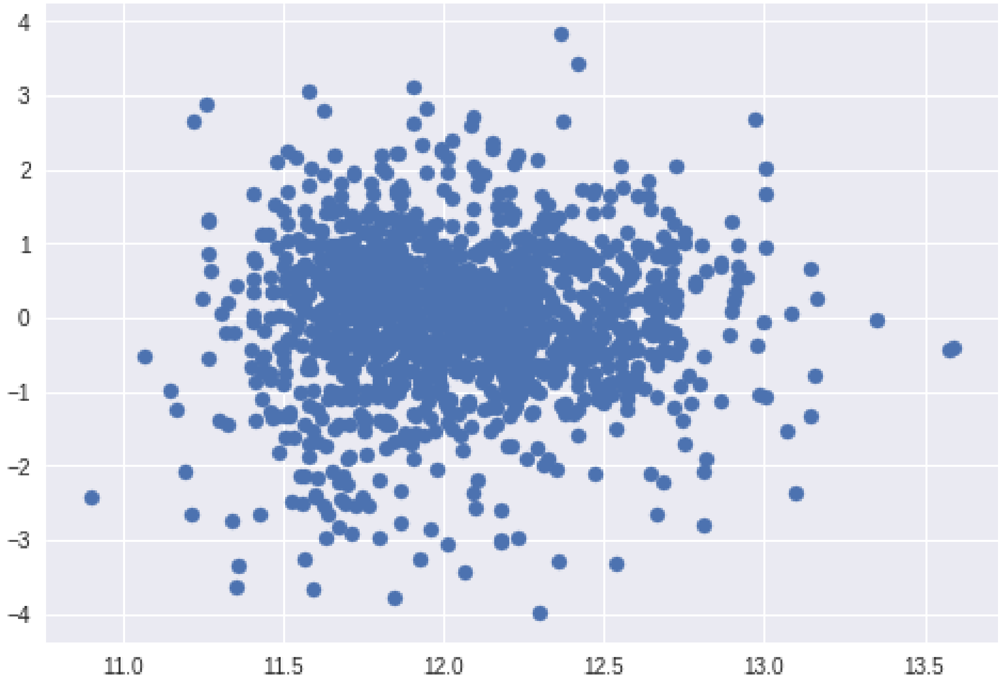
we find that all the parameters have improved.

We now check the **Q-Q plot** to check normality assumption of residuals.



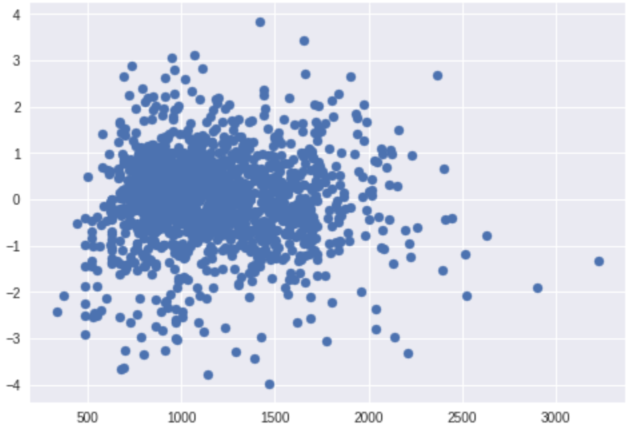
Since it is almost a straight line with slope = 1, we can conclude that all our assumptions hold and log transformation is a suitable one.

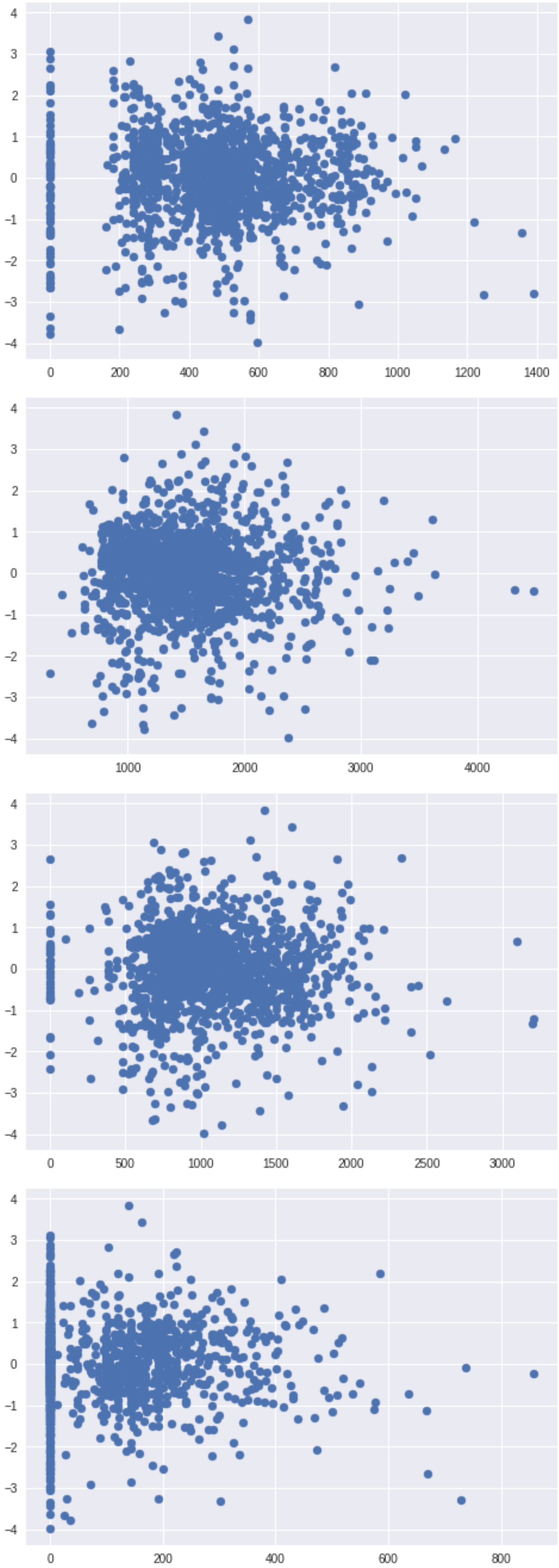
We further plot the **residuals versus predicted-responses**.

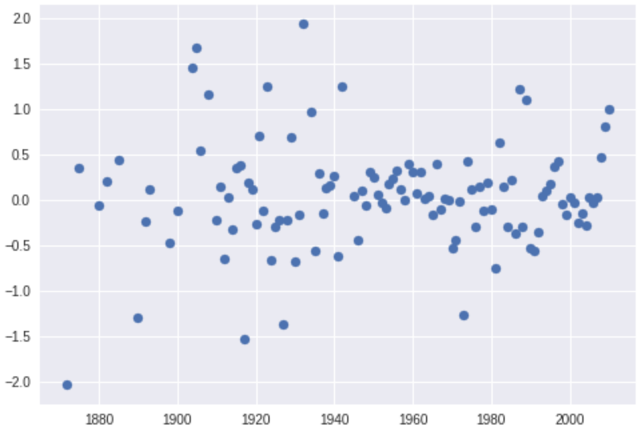


Since, they are randomly distributed within a horizontal band, so there is no model deficiency.

We now plot the **residuals and regressors** to check whether they show some pattern.



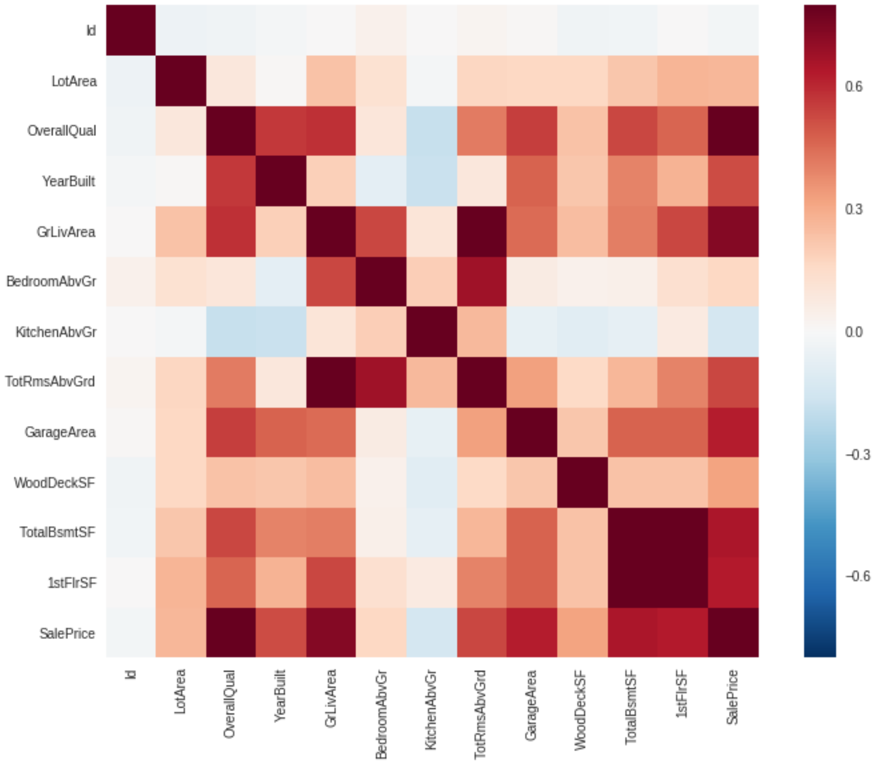

4
3
2
1
0
−1
−2
−3
−4
0
200
400
600
800
1000
1200
1400
4
3
2
1
0
−1
−2
−3
−4
1000
2000
3000
4000
4
3
2
1
0
−1
−2
−3
−4
0
500
1000
1500
2000
2500
3000
4
3
2
1
0
−1
−2
−3
−4
0
200
400
600
800



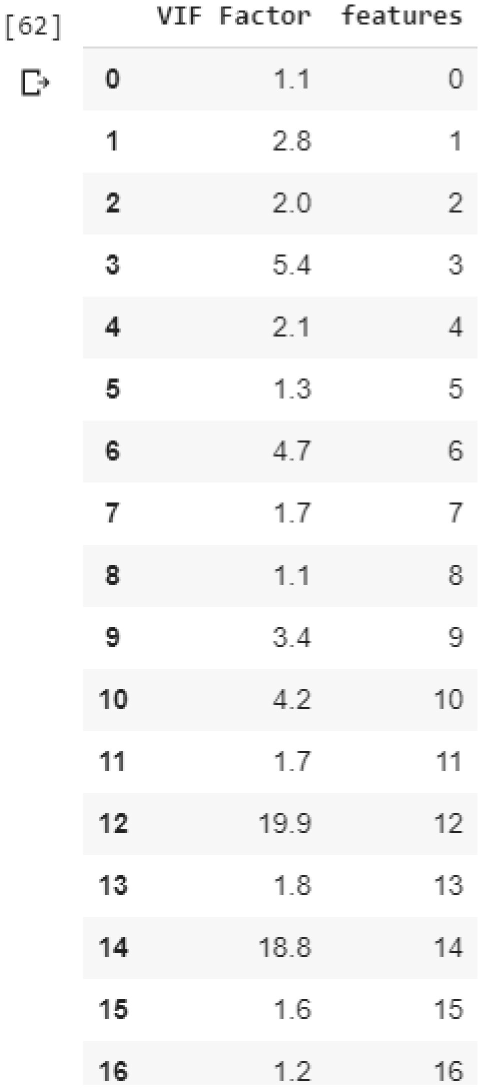
We clearly see that none of the regressors show any kind of pattern w.r.t the residuals. *Mostly, the residuals are fluctuating randomly within a horizontal band without any order. So, no corrective measures are required.*
Hence we stand by our log transformation of target variable and proceed.

Multicollinearity

We first plot the **correlation matrix** between all the regressors to check which of the regressors show dominant correlation.



We now check the **VIFs** corresponding to each regressor.



	VIF Factor	features
0	1.1	0
1	2.8	1
2	2.0	2
3	5.4	3
4	2.1	4
5	1.3	5
6	4.7	6
7	1.7	7
8	1.1	8
9	3.4	9
10	4.2	10
11	1.7	11
12	19.9	12
13	1.8	13
14	18.8	14
15	1.6	15
16	1.2	16

We observe that 3 regressors have VIFs greater than 5.

We now apply **Variance Decomposition Method** to remove the regressors causing multicollinearity.

eigenvalues	Condition Number
4.06893229	1
1.94519842	1.446
1.07665231	1.944
0.97351195	2.044
0.812676	2.237
0.67771429	2.450
0.49628356	2.863
0.43077655	3.073
0.23598477	4.152
0.15693549	5.092
0.12533436	5.700

We notice that none of the condition numbers are more than 30. So, we can't remove any of the regressors according to variance decomposition method.

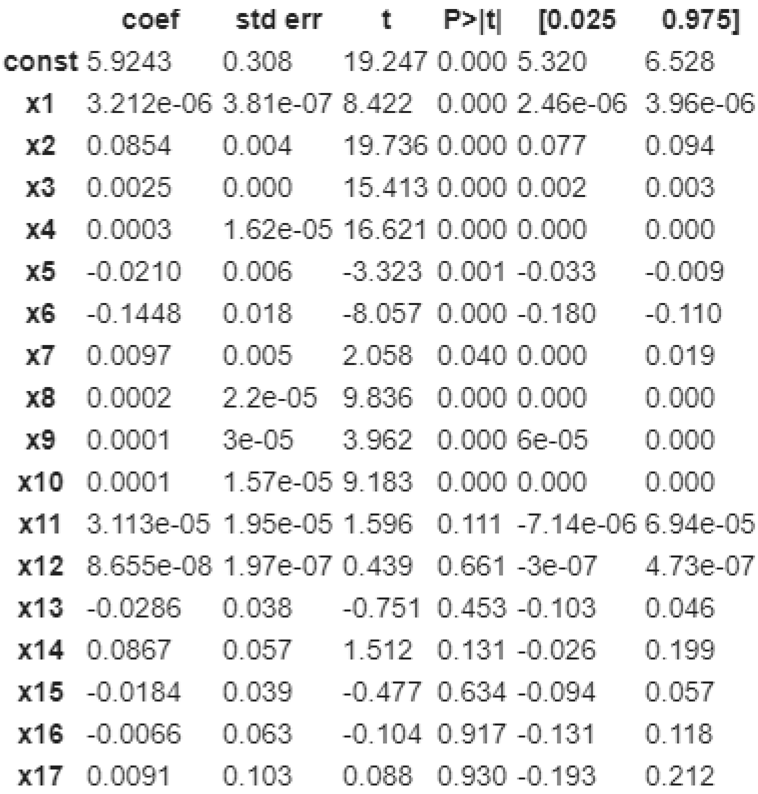
Hence we proceed to variable selection.

Variable Selection

We now need to perform variable selection as the next step for model building. The aim of this part is to identify and successfully remove insignificant regressors which do not contribute much in explaining the response variable. We use the backward elimination procedure to find the required set of regressors. Here we include the 5 dummy variables and one mean encoding variable for roofstyle and housestyle respectively.

Step 1

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

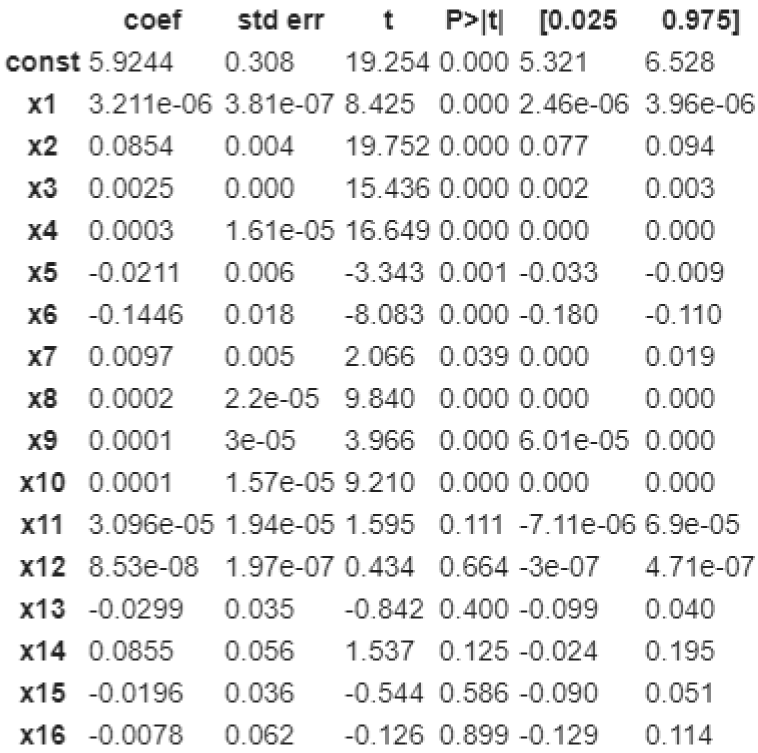


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9243	0.308	19.247	0.000	5.320	6.528
x1	3.212e-06	3.81e-07	8.422	0.000	2.46e-06	3.96e-06
x2	0.0854	0.004	19.736	0.000	0.077	0.094
x3	0.0025	0.000	15.413	0.000	0.002	0.003
x4	0.0003	1.62e-05	16.621	0.000	0.000	0.000
x5	-0.0210	0.006	-3.323	0.001	-0.033	-0.009
x6	-0.1448	0.018	-8.057	0.000	-0.180	-0.110
x7	0.0097	0.005	2.058	0.040	0.000	0.019
x8	0.0002	2.2e-05	9.836	0.000	0.000	0.000
x9	0.0001	3e-05	3.962	0.000	6e-05	0.000
x10	0.0001	1.57e-05	9.183	0.000	0.000	0.000
x11	3.113e-05	1.95e-05	1.596	0.111	-7.14e-06	6.94e-05
x12	8.655e-08	1.97e-07	0.439	0.661	-3e-07	4.73e-07
x13	-0.0286	0.038	-0.751	0.453	-0.103	0.046
x14	0.0867	0.057	1.512	0.131	-0.026	0.199
x15	-0.0184	0.039	-0.477	0.634	-0.094	0.057
x16	-0.0066	0.063	-0.104	0.917	-0.131	0.118
x17	0.0091	0.103	0.088	0.930	-0.193	0.212

We notice that 7 of the regressors have p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the dummy variable corresponding to roofstyle-shed. We now repeat the procedure after removing this regressor.

Step 2

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

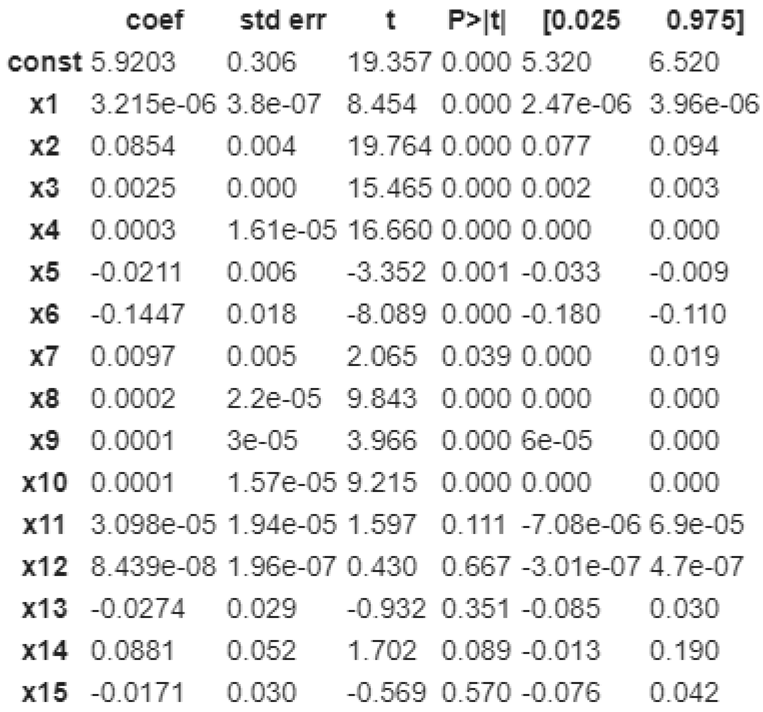


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9244	0.308	19.254	0.000	5.321	6.528
x1	3.211e-06	3.81e-07	8.425	0.000	2.46e-06	3.96e-06
x2	0.0854	0.004	19.752	0.000	0.077	0.094
x3	0.0025	0.000	15.436	0.000	0.002	0.003
x4	0.0003	1.61e-05	16.649	0.000	0.000	0.000
x5	-0.0211	0.006	-3.343	0.001	-0.033	-0.009
x6	-0.1446	0.018	-8.083	0.000	-0.180	-0.110
x7	0.0097	0.005	2.066	0.039	0.000	0.019
x8	0.0002	2.2e-05	9.840	0.000	0.000	0.000
x9	0.0001	3e-05	3.966	0.000	6.01e-05	0.000
x10	0.0001	1.57e-05	9.210	0.000	0.000	0.000
x11	3.096e-05	1.94e-05	1.595	0.111	-7.11e-06	6.9e-05
x12	8.53e-08	1.97e-07	0.434	0.664	-3e-07	4.71e-07
x13	-0.0299	0.035	-0.842	0.400	-0.099	0.040
x14	0.0855	0.056	1.537	0.125	-0.024	0.195
x15	-0.0196	0.036	-0.544	0.586	-0.090	0.051
x16	-0.0078	0.062	-0.126	0.899	-0.129	0.114

We notice that 6 of the regressors have p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the dummy variable corresponding to roofstyle-Mansard. We now repeat the procedure after removing this regressor.

Step 3

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

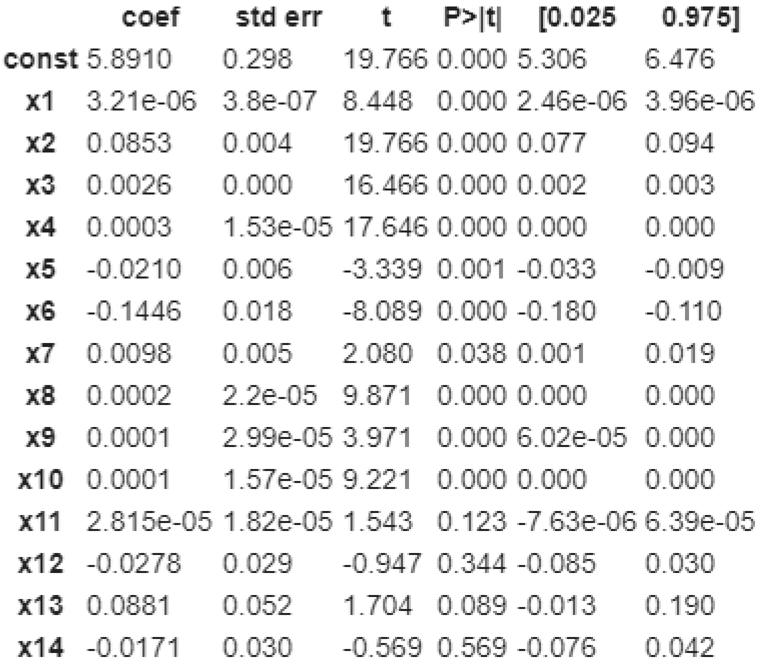


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9203	0.306	19.357	0.000	5.320	6.520
x1	3.215e-06	3.8e-07	8.454	0.000	2.47e-06	3.96e-06
x2	0.0854	0.004	19.764	0.000	0.077	0.094
x3	0.0025	0.000	15.465	0.000	0.002	0.003
x4	0.0003	1.61e-05	16.660	0.000	0.000	0.000
x5	-0.0211	0.006	-3.352	0.001	-0.033	-0.009
x6	-0.1447	0.018	-8.089	0.000	-0.180	-0.110
x7	0.0097	0.005	2.065	0.039	0.000	0.019
x8	0.0002	2.2e-05	9.843	0.000	0.000	0.000
x9	0.0001	3e-05	3.966	0.000	6e-05	0.000
x10	0.0001	1.57e-05	9.215	0.000	0.000	0.000
x11	3.098e-05	1.94e-05	1.597	0.111	-7.08e-06	6.9e-05
x12	8.439e-08	1.96e-07	0.430	0.667	-3.01e-07	4.7e-07
x13	-0.0274	0.029	-0.932	0.351	-0.085	0.030
x14	0.0881	0.052	1.702	0.089	-0.013	0.190
x15	-0.0171	0.030	-0.569	0.570	-0.076	0.042

We notice that 5 of the regressors have p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the mean encoding variable corresponding to housestyle.
We now repeat the procedure after removing this regressor.

Step 4

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

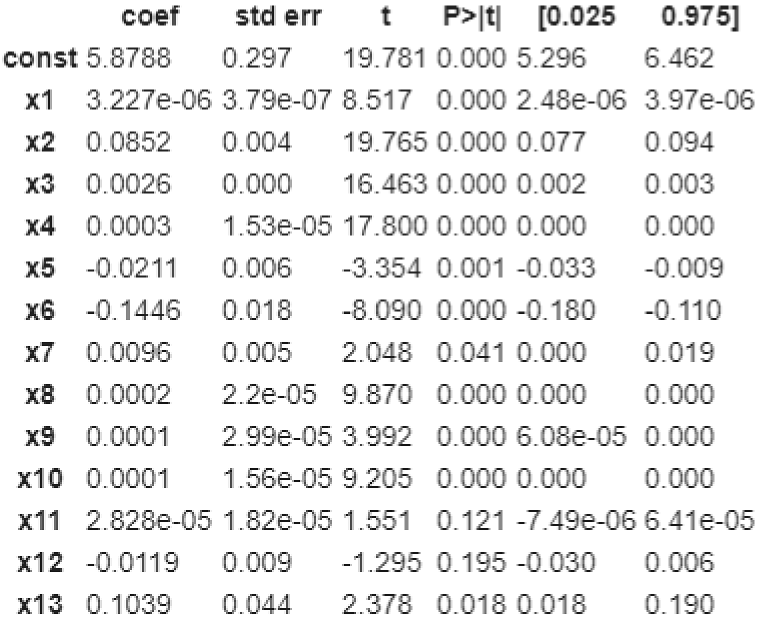


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8910	0.298	19.766	0.000	5.306	6.476
x1	3.21e-06	3.8e-07	8.448	0.000	2.46e-06	3.96e-06
x2	0.0853	0.004	19.766	0.000	0.077	0.094
x3	0.0026	0.000	16.466	0.000	0.002	0.003
x4	0.0003	1.53e-05	17.646	0.000	0.000	0.000
x5	-0.0210	0.006	-3.339	0.001	-0.033	-0.009
x6	-0.1446	0.018	-8.089	0.000	-0.180	-0.110
x7	0.0098	0.005	2.080	0.038	0.001	0.019
x8	0.0002	2.2e-05	9.871	0.000	0.000	0.000
x9	0.0001	2.99e-05	3.971	0.000	6.02e-05	0.000
x10	0.0001	1.57e-05	9.221	0.000	0.000	0.000
x11	2.815e-05	1.82e-05	1.543	0.123	-7.63e-06	6.39e-05
x12	-0.0278	0.029	-0.947	0.344	-0.085	0.030
x13	0.0881	0.052	1.704	0.089	-0.013	0.190
x14	-0.0171	0.030	-0.569	0.569	-0.076	0.042

We notice that 4 of the regressors have p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the dummy variable corresponding to roofstyle-Hip
We now repeat the procedure after removing this regressor.

Step 5

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

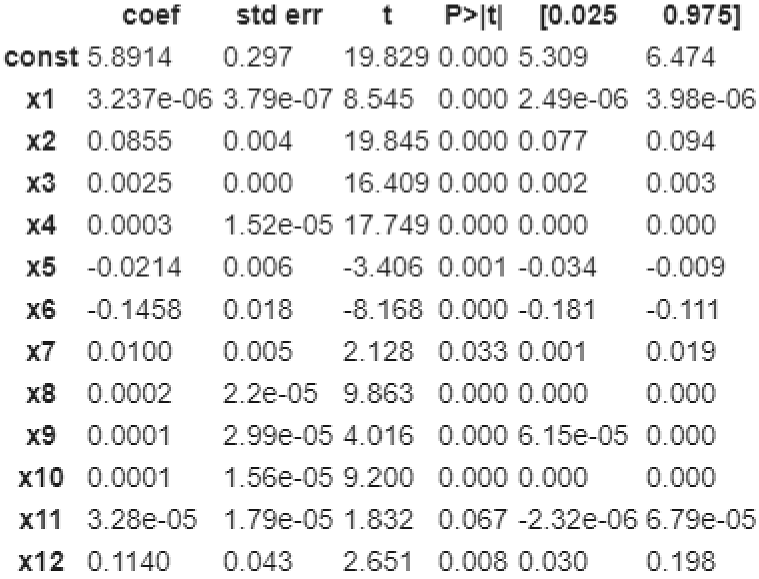


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8788	0.297	19.781	0.000	5.296	6.462
x1	3.227e-06	3.79e-07	8.517	0.000	2.48e-06	3.97e-06
x2	0.0852	0.004	19.765	0.000	0.077	0.094
x3	0.0026	0.000	16.463	0.000	0.002	0.003
x4	0.0003	1.53e-05	17.800	0.000	0.000	0.000
x5	-0.0211	0.006	-3.354	0.001	-0.033	-0.009
x6	-0.1446	0.018	-8.090	0.000	-0.180	-0.110
x7	0.0096	0.005	2.048	0.041	0.000	0.019
x8	0.0002	2.2e-05	9.870	0.000	0.000	0.000
x9	0.0001	2.99e-05	3.992	0.000	6.08e-05	0.000
x10	0.0001	1.56e-05	9.205	0.000	0.000	0.000
x11	2.828e-05	1.82e-05	1.551	0.121	-7.49e-06	6.41e-05
x12	-0.0119	0.009	-1.295	0.195	-0.030	0.006
x13	0.1039	0.044	2.378	0.018	0.018	0.190

We notice that 2 of the regressors have p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the dummy variable corresponding to roofstyle-Gable
We now repeat the procedure after removing this regressor.

Step 6

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.

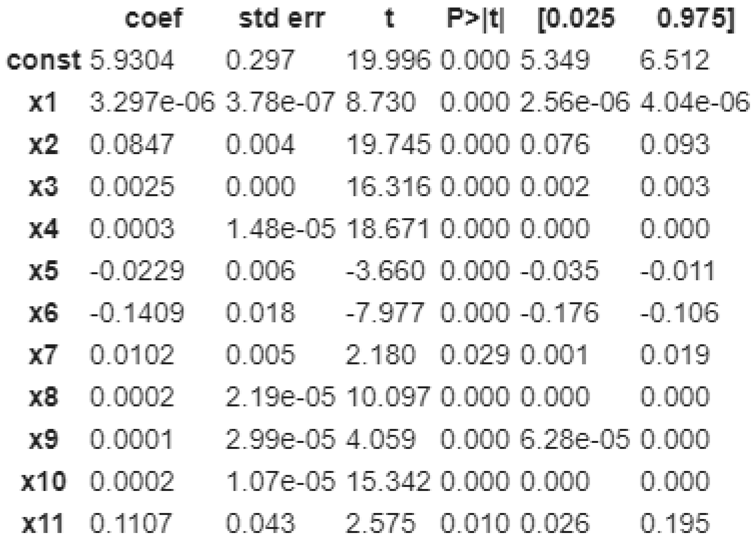


	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8914	0.297	19.829	0.000	5.309	6.474
x1	3.237e-06	3.79e-07	8.545	0.000	2.49e-06	3.98e-06
x2	0.0855	0.004	19.845	0.000	0.077	0.094
x3	0.0025	0.000	16.409	0.000	0.002	0.003
x4	0.0003	1.52e-05	17.749	0.000	0.000	0.000
x5	-0.0214	0.006	-3.406	0.001	-0.034	-0.009
x6	-0.1458	0.018	-8.168	0.000	-0.181	-0.111
x7	0.0100	0.005	2.128	0.033	0.001	0.019
x8	0.0002	2.2e-05	9.863	0.000	0.000	0.000
x9	0.0001	2.99e-05	4.016	0.000	6.15e-05	0.000
x10	0.0001	1.56e-05	9.200	0.000	0.000	0.000
x11	3.28e-05	1.79e-05	1.832	0.067	-2.32e-06	6.79e-05
x12	0.1140	0.043	2.651	0.008	0.030	0.198

We notice that only one regressor has p-value greater than 0.05 and are insignificant. We now remove the regressor having the largest p-value which is the regressor corresponding to 1st floor surface area
We now repeat the procedure after removing this regressor.

Step 7

We include all the regressors and fit OLS. We then look at t-statistic for each regressor and their significance.



	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9304	0.297	19.996	0.000	5.349	6.512
x1	3.297e-06	3.78e-07	8.730	0.000	2.56e-06	4.04e-06
x2	0.0847	0.004	19.745	0.000	0.076	0.093
x3	0.0025	0.000	16.316	0.000	0.002	0.003
x4	0.0003	1.48e-05	18.671	0.000	0.000	0.000
x5	-0.0229	0.006	-3.660	0.000	-0.035	-0.011
x6	-0.1409	0.018	-7.977	0.000	-0.176	-0.106
x7	0.0102	0.005	2.180	0.029	0.001	0.019
x8	0.0002	2.19e-05	10.097	0.000	0.000	0.000
x9	0.0001	2.99e-05	4.059	0.000	6.28e-05	0.000
x10	0.0002	1.07e-05	15.342	0.000	0.000	0.000
x11	0.1107	0.043	2.575	0.010	0.026	0.195

We notice that all the regressors are now significant. We now stop the Backward elimination.

Now we notice that exceot for the dummy variable corresponding to Gambrel all other dummy variables have been removed in the variable selection. We now check the mean of the response variable in each category.

Category	Mean of response variable
Flat	194690.0
Gable	171483.95617879054
Gambrel	148909.0909090909
Hip	218876.93356643355
Mansard	180568.42857142858
Shed	225000.0

We notice that mean in Gambrel category is small compared to other category means. This justifies retaining dummy variable corresponding to Gambrel.

Hence the final set of regressors are ’LotArea’, ’OverallQual’, ’YearBuilt’, ’RoofStyle’(dummy variable of Gambrel category), ’GrLivArea’, ’BedroomAbvGr’, ’KitchenAbvGr’, ’TotRmsAbvGrd’, ’GarageArea’, ’WoodDeckSF’, ’TotalBsmtSF’, ’SalePrice’

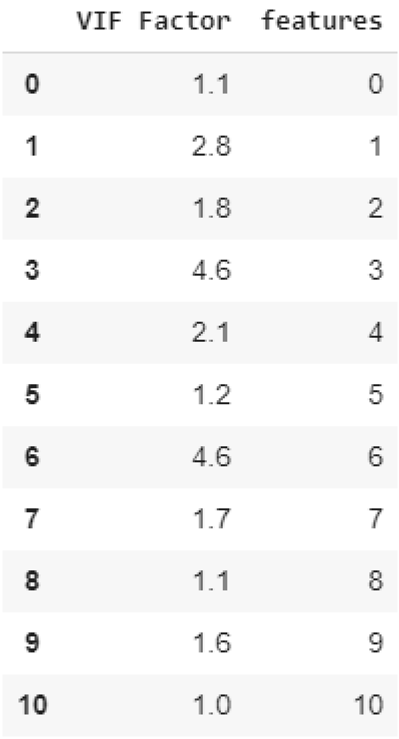
Multi-Collinearity Check

We again do Variance Decoposition to check for regressors causing multicollinearity.

eigenvalues	Condition index
3.55294943	1
1.95565187	1.34787227
1.03990972	1.84840312
0.99454011	1.89009381
0.83725616	2.05999107
0.80757353	2.09750728
0.54476385	2.55382052
0.49477659	2.67972322
0.42308536	2.89788076
0.21925903	4.02546232
0.13023434	5.22314113

We notice that none of the condition indices are greater than 30 which implies we can't remove any of the variables.

We now check the VIFs for each regressor.

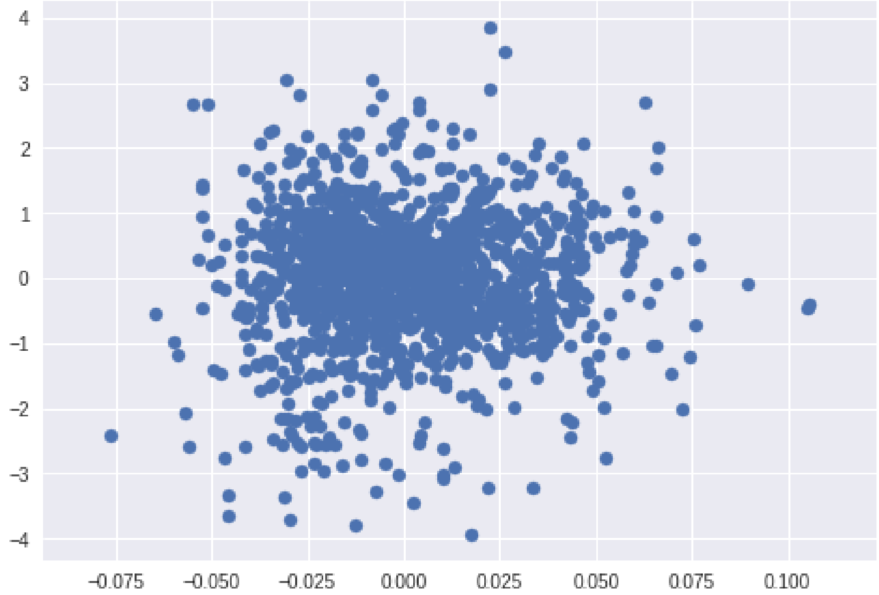


	VIF Factor	features
0	1.1	0
1	2.8	1
2	1.8	2
3	4.6	3
4	2.1	4
5	1.2	5
6	4.6	6
7	1.7	7
8	1.1	8
9	1.6	9
10	1.0	10

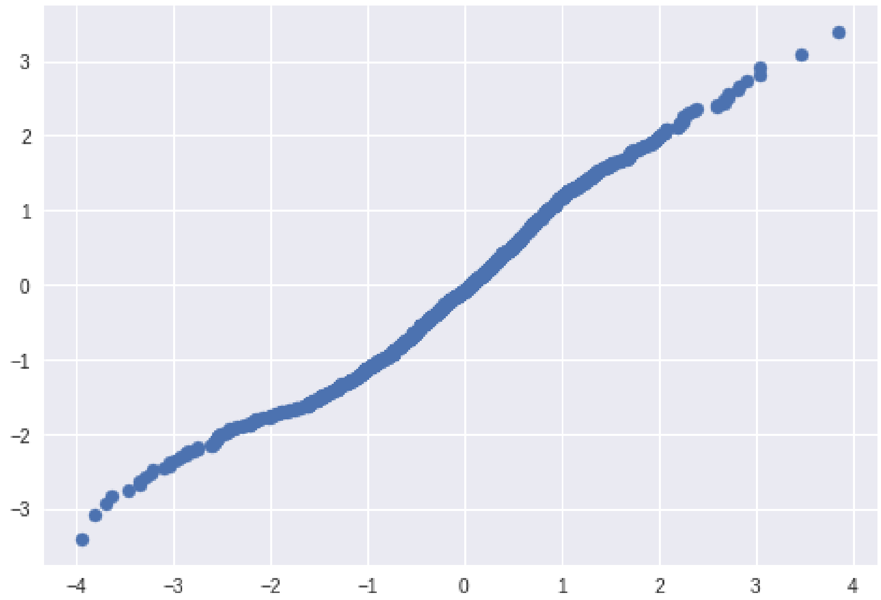
We observe that none of the VIFs are greater than 5 which implies there is no multi-collinearity problem after variable selection.

We conclude that from Variable Decomposition and VIFs that there is no multi-collinearity after dropping relevant regressors in Variable selection.

Final Residual Analysis



The reason for doing residual analysis after we fit the model is to see if there is some graphical pattern between the predicted responses and the residuals. In case there is a pattern in the graph, it indicates that there is some inherent relationship between the response variable and the regressors. In that case we would have to perform suitable transformations on the regressors and the response variable depending on the nature of the pattern obtained in the graph. We should watch out for patterns such as outward or inward opening funnel, double bow pattern in our plot for residual analysis. However we know that under the assumption of homoscedasticity, independence and normality of errors, the covariance between predicted responses and obtained residuals is a null matrix indicating absolutely no correlation between them. Thus we should expect to get a graph in which the residuals are randomly distributed with the predicted responses. This is exactly what happens in the graph we have obtained. The residuals are randomly and symmetrically distributed against the predicted responses in a thin band extending on either side of the X-axis. Therefore there is no need for us to apply any transformation on the regressors or the observations. Perhaps taking the log of the continous regressors as the regressors itself in the initial model ensured that we get a residual plot devoid of any patterned structure.



Q-Q plot

To ensure that the residuals and hence the predicted responses follow normal distribution, we plot the Q-Q plot. From the figure given above, we see that the graph of standardized residuals vs the quantiles of standard normal distribution N(0,1) resembles a 45 degree line as expected. Thus we conclude that the normality assumptions hold and thus it would not be wrong to use the standard test statistics as a measure of the goodness of the fitted model.

Conclusion

We tried to model house price based on the available data. We used Multiple Linear Regression to build our model. Considering differences in scale of price, the regressors and using our intuition based on the field knowledge, we tried log transformation on the price (target variable) and it worked based on the analysis we did .

Then we tried to solve the problem of multicollinearity to ensure stable estimates of the parameters of our model. This made us to remove 2 continuous regressors and some regressors corresponding to dummy variables introduced for the categorical variable.

The next step included variable selection to include only significant regressors. We carried out variable selection using backward selection method.
Then we performed the multi-collinearity check again and concluded that there is no multicollinearity problem based on the results, so ridge regression was not needed.

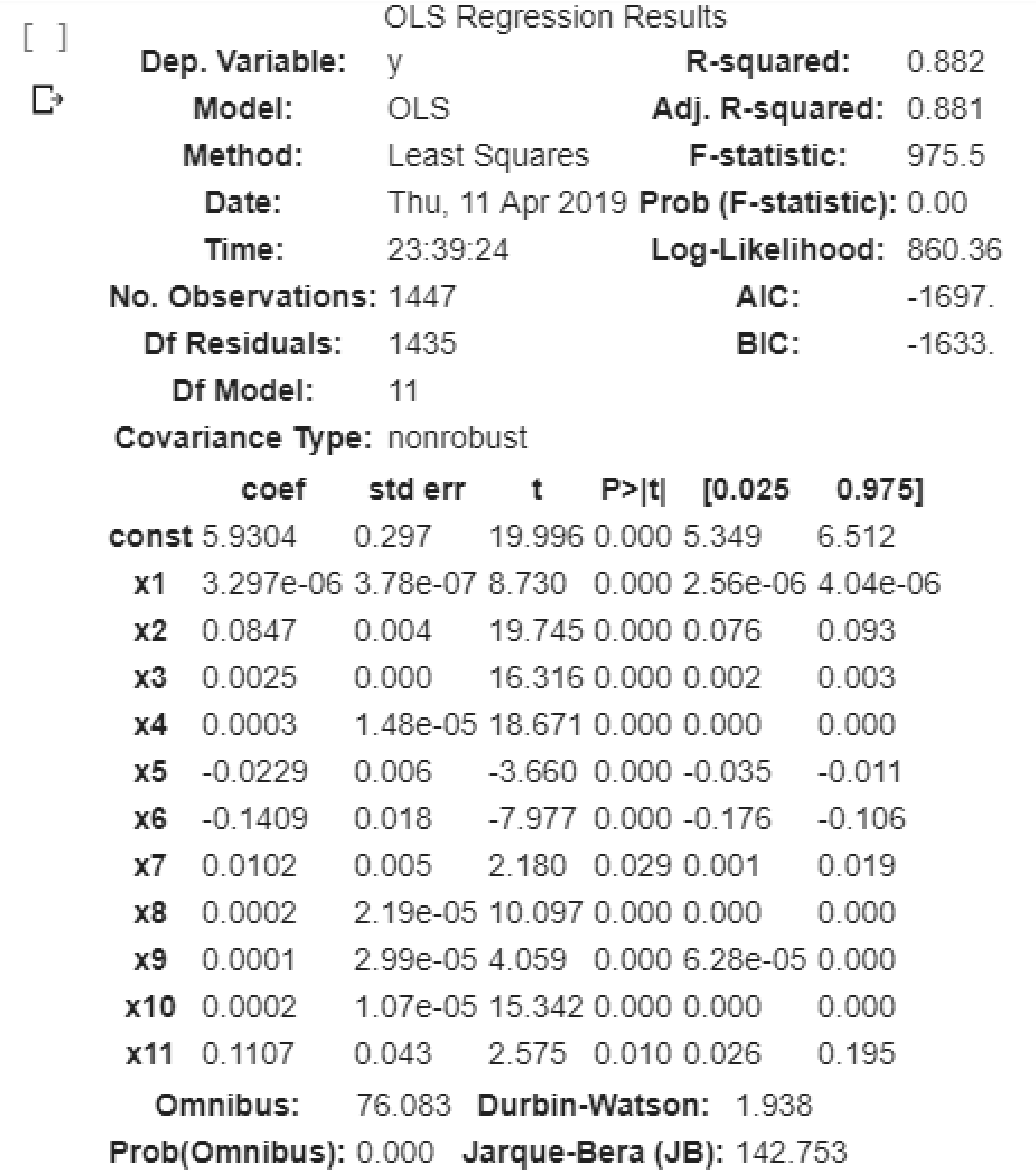
We finally carried out Residual Analysis to check if the regressors and reponse variable had any inherent relation and if any suitable transformation was required. The plot of residual analysis indicated that no transformation was required. The QQ plot shows that the normality conditions of the predicted reponses and residuals is satisfied and hence it is justified to use the standard tests and statistics of linear regression such as F-statistic.

After being certain that our model is within safe limits of assumptions of multiple linear regression model,. we therefore conclude our linear model to be

$$\begin{aligned} log(Houseprice + 1) = {} & LotArea + OverallQual + YearBuilt + GrLivArea + \\ & BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd + Garage \\ & WoodDeckSF + TotalBsmtSF + SalePrice + \\ & RoofStyle(\text{dummy variable of Gambrel category}) \end{aligned}$$

This is obviously a representative model and each regressor is multiplied by a unique coefficient.

The set of coefficients , their corresponding standard errors and all other relevant information is provided in the below mentioned table.



```
                         OLS Regression Results
    Dep. Variable:      y                        R-squared:  0.882
           Model:       OLS                 Adj. R-squared:  0.881
          Method:       Least Squares          F-statistic:  975.5
            Date:       Thu, 11 Apr 2019 Prob (F-statistic): 0.00
            Time:       23:39:24            Log-Likelihood:  860.36
No. Observations:       1447                           AIC:  -1697.
    Df Residuals:       1435                           BIC:  -1633.
        Df Model:       11
 Covariance Type:       nonrobust
             coef       std err    t      P>|t|  [0.025    0.975]
     const 5.9304       0.297      19.996 0.000  5.349     6.512
     x1    3.297e-06    3.78e-07   8.730  0.000  2.56e-06  4.04e-06
     x2    0.0847       0.004      19.745 0.000  0.076     0.093
     x3    0.0025       0.000      16.316 0.000  0.002     0.003
     x4    0.0003       1.48e-05   18.671 0.000  0.000     0.000
     x5    -0.0229      0.006      -3.660 0.000  -0.035    -0.011
     x6    -0.1409      0.018      -7.977 0.000  -0.176    -0.106
     x7    0.0102       0.005      2.180  0.029  0.001     0.019
     x8    0.0002       2.19e-05   10.097 0.000  0.000     0.000
     x9    0.0001       2.99e-05   4.059  0.000  6.28e-05  0.000
     x10   0.0002       1.07e-05   15.342 0.000  0.000     0.000
     x11   0.1107       0.043      2.575  0.010  0.026     0.195
      Omnibus:      76.083  Durbin-Watson:     1.938
Prob(Omnibus):      0.000   Jarque-Bera (JB):  142.753
```

0.1 Future Scope

We used mean encoding for one categorical variable , to avoid a large number of parameters (55+) in our model. Considering the fact that we have only 1400 Observations, this may lead to compromise on the goodness or reliability of the fit. Incase we manage to obtain more data , it is possible to work with 55+ parameters without compromise on the goodness of fit. And we will definitely get better resultsif we do so.

We removed 13 observations after the outlier detection. So to improve further in this aspect, we may try to get M estimate of the parameters which is more efficient and has a larger breakdown point.

References

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- https://urlzs.com/RHC6
- Introduction to Linear Regression Analysis, 5th ed.
  Douglas C. Montgomery, Elizabeth A. Peck, and G.