STAT462 Data Analysis Project: Ames Housing Price Prediction

Yiqi Xiong

Introduction

Housing price is one of the indexes for the economy. Sharply decreased housing prices and over-issued sub-prime mortgage unbalanced the relationship between the global real estate market and the baking system, then it ultimately triggered the global financial crisis in 2008. Housing evaluation is also crucial for different groups with a multitude of purposes: homeowners, investors, tax assessors, and other real estate market participants. (Frew. & Jud., 2003) Housing prices can be influenced by various factors, such as location, neighborhood, and total living area. Therefore, it is important to predict the housing prices to provide a practical suggestion for both buyer and seller. Moreover, the development of a housing price prediction model would greatly assist in the prediction of future housing prices and the establishment of real estate policies. (Park. & Bae., 2015) This project uses regression analysis as a study methodology to develop housing price prediction models. The *"AmesHousing"* data set was collected from the Ames Assessorís Office, and it contains information on computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The data set has 2930 observations and 80 variables(exclude 2 observation identifiers): 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, and they are the direct description of the quality and quantity of many physical attributes of the property. (De Cock, 2011) This project uses 20 continuous variables for the construct regression models. The goal of this project is to select important features for predicting housing prices and to find which model can achieve better performance.

Data Preprocessing

We extract 20 continuous variables (Table 1)from the original *"AmesHousing"* data set, and We drop observations with missing value. (Output 1) Since the variable *"Total Bsmt SF"* is the sum of the *"BsmtFin SF 1"* and *"BsmtFin SF 2"*, we drop *"Total Bsmt SF"*, and similar situation also applies to the variable *"Gr Liv Area"* to considering keep more information. The final data set dimension 2421 observations and 18 variables. The response is the *"SalePrice"*, and the other variables are predictors.

Exploratory Data Analysis

After we fit histogram and boxplot to every18 variables, we observe that all variables are not normally distributed, and they are all skewed to the right. From boxplots and the summary statistics (Output 2), we can see there exists large variation and some outliers among the dataset, especially for *"SalePrice"*. Therefore, checking for influential points and normality for regression model is necessary. From Figure 2, we observe negative relationship between the combination of *"SalePrice"* with "*Low.Qual.Fin.SF*", "*Misc.Val*", and "*Low.Qual.Fin.SF*" respectively. No relationship between *"BsmtFin.SF.2"* and *"SalePrice"* because of many "*0*" value in *"BsmtFin.SF.2"*. Other predictors have positive relationship with *"SalePrice"*. We do not find multicollinearity issue from the plot, but further validation is needed. We also split the dataset into a 25% testing set and a 75% training set. The following models will perform predictions based on training set and compare accuracy based on the testing set.

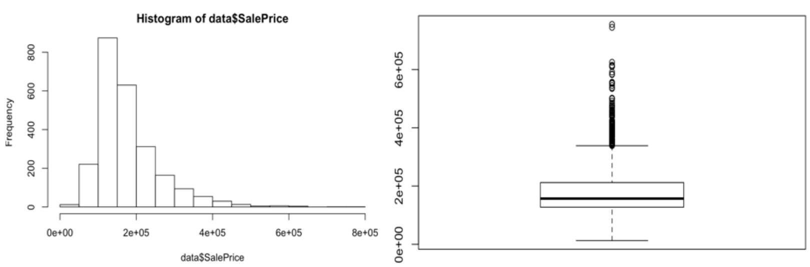


Figure 1: Distribution of *"SalePrice"*

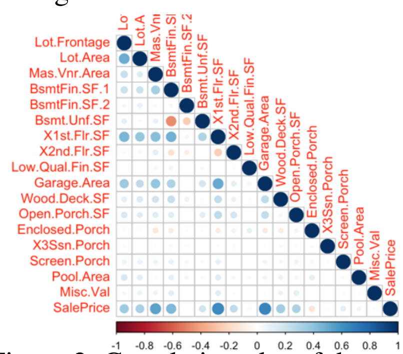


Figure 2: Correlation plot of data set

Analysis

First, we fit the full linear model. The variable *"Lot.Frontage", " Low.Qual.Fin.SF",* and *"3Ssn.Porch"* are not statistically significant at 0.05 level from model summary (Output 3) using T test, but *"Lot.Frontage"* is significant and *"BsmtFin.SF.2"* is insignificant from the ANOVA (Analysis of Variance) using F test.(Output 4)The full linear model is significant, but the normality assumption is not satisfied from Shapiro-Wilk test. The linearity and equal variance assumption are satisfied if we ignore the labeled outliers.(Output 5) Also, the VIF result (Output 6) confirms there is no serious multicollinearity problem among predictors. The following fitted models have same diagnostic conclusion. After we remove outliers, we apply full model again. *"BsmtFin.SF.2"* becomes insignificant. Since *"Lot.Frontage"* is very close to significant. Then, we drop *"BsmtFin.SF.2", " Low.Qual.Fin.SF",* and *"3Ssn.Porch"* to perform a sub linear model. Since the response has an extreme wide range and huge variation, we decide to apply cubic root transformation on *"SalePrice"* to reduce homoscedasticity. The same procedures are applied to cubic root transformation models. Second, We apply general lest squared regression with an autoregressive process of order 1 on cubic_root_sub model. Third, we perform AIC and BIC selection based on full cubic root model. The AIC for AIC selection is 10195.16 and for BIC selection is 10195.26. Lastly, we fit models based on selected variables using the ridge, lasso, and elastic net.

Result

We can overserve coefficient estimates of each model in Table 1. The values of coefficients from models with cubic root transformation are very close. The sub linear model, cubic root sub model, and generalized least squares regression based on the cubic root sub model do not contain *"BsmtFin.SF.2", " Low.Qual.Fin.SF",* and *"3Ssn.Porch"*. The AIC selection, Lasso regression and Elastic Net do not contain *"Lot.Frontage", " Low.Qual.Fin.SF",* and *"3Ssn.Porch"*. The BIC selection does not choose *"Lot.Frontage", "Lor.Area" " Low.Qual.Fin.SF",* and *"3Ssn.Porch"*. Since all model selection drop the predictor *" Low.Qual.Fin.SF",* and *"3Ssn.Porch"*, it implies these two variables may not influence sale price. Also, the increasing total area of the house, including basement, living area, and garage result higher sale prices. The smaller pool area, less expensive miscellaneous feature, shorter distance between the property and the street, and smaller enclosed porch area will also increase the sale price.

	Full_Linea	Sub_Linear	Full_Cubic_rt	Sub_Cubic_rt	gls	AIC	BIC	Ridge	Lasso	Elastic Net
(Intercept)	-15971.5	-14407.47	36.30244	36.49613	36.49747356	36.11337	36.1165525	37.16751	36.79766	36.86352
Lot Frontage	-97.4461	-104.2353	-0.00660804	-0.0073275	-0.00732889	.	.	-0.001785847	.	.
Lot Area	0.504395	0.573146	2.9385E-05	3.7405E-05	3.74907E-05	2.20635E-05	.	3.74565E-05	1.2967E-05	1.4966E-05
Mas Vnr Area	58.26353	59.02929	0.003148997	0.00325371	0.003252351	0.003146292	0.00307769	0.004016454	0.00327806	0.00335835
BsmtFin SF 1	54.13022	45.32046	0.005327887	0.00428518	0.004284577	0.005364535	0.0053726	0.004309428	0.00410146	0.00408961
BsmtFin SF 2	32.63058	.	0.003854303	.	.	0.003877597	0.00395946	0.002720816	0.00199865	0.00200858
Bsmt Unf SF	38.07184	28.64135	0.00407384	0.00294743	0.002948476	0.004095717	0.00407978	0.00322113	0.00283962	0.00284368
1st Flr SF	65.22939	72.19611	0.005994319	0.00680967	0.006808081	0.005848842	0.00601404	0.005765341	0.00648787	0.00641697
2nd Flr SF	64.50233	62.82666	0.006388789	0.00617878	0.006179138	0.006354505	0.00641462	0.005608478	0.00590152	0.00584953
Low Qual Fin	-0.32961	.	-0.0014195	.	.	.	.	-0.001286164	.	.
Garage Area	90.28449	91.9141	0.009443645	0.00965156	0.00965102	0.009376428	0.00939557	0.009501216	0.00977071	0.00975925
Wood Deck S	62.62332	67.23239	0.006516175	0.00707654	0.00707719	0.006525632	0.0065283	0.006815707	0.00608722	0.0061398
Open Porch S	47.95113	52.36256	0.005158114	0.00563465	0.005631578	0.005087379	0.0050007	0.006601976	0.00460899	0.00478358
Enclosed Por	-58.8158	-59.01903	-0.00778937	-0.0078509	-0.00786324	-0.0080504	-0.008003	-0.007585307	-0.0067516	-0.0067872
3-Ssn Porch	24.20309	.	0.004372154	.	.	.	.	0.004206297	.	.
Screen Porch	60.22199	64.39311	0.006256544	0.0066953	0.006698939	0.006138448	0.00620639	0.006258218	0.00457075	0.00466178
Pool Area	-94.6223	-86.58328	-0.01301816	-0.1206262	-0.01207044	-0.01327337	-0.0132775	-0.01060477	-0.0083985	-0.0084201
Misc Val	-19.1513	-19.24978	-0.00167625	-0.0016868	-0.00168654	-0.00166788	-0.0016592	-0.001520685	-0.0013866	-0.0013878

Table 1: Model coefficients

To choose a model that has better performance for predicting house sale price, we use R squared, RMSE(root mean squared error), and MAE(mean absolute error) to evaluate the model prediction performance. From Table 2, we choose lasso regression for our final prediction model, since lasso regression has relatively high R squared, lowest RMSE, and relatively small MEA value by comparing with other models. We also take mean, median, and maximum from lasso regression to predict sale price with 95% confidence interval and prediction interval. (Figure 3) For example, we have 95% confidence that a house with a true value \$169278 (=55.31803^3) will fall in a range between \$166290 and \$172301.

model	r.squared	RMSE	MAE
full_linear	0.7573448	43242.770000	2.282734e+05
sub_linear	0.7521192	43705.910000	2.853415e+04
AIC	0.7652462	4.048899	2.759633e+00
BIC	0.7612804	4.082956	2.784051e+00
cubic_root _full	0.7659834	4.042537	2.768526e+00
cubic_root_sub1	0.7594337	4.098719	2.787835e+00
gls	0.7601454	4.141466	2.866811e+00
ridge	0.7672392	3.699669	2.831422e+00
lasso	0.7666096	3.680018	2.847055e+00
elastic	0.7669758	3.681160	2.848590e+00

Table 2: Model Performance

```
predict(lasso, new = data.frame(t(u)), interval = 'confidence', level = 0.95)
predict(lasso, new = data.frame(t(med)), interval = 'confidence', level = 0.95)
predict(lasso, new = data.frame(t(max)),interval = 'confidence', level = 0.95)
predict(lasso, new = data.frame(t(u)), interval = 'prediction', level = 0.95)
predict(lasso, new = data.frame(t(med)),interval = 'prediction', level = 0.95)
predict(lasso, new = data.frame(t(max)),interval = 'prediction', level = 0.95)

       fit      lwr      upr
1 55.31803 54.99061 55.64545
       fit      lwr      upr
1 50.73429 50.13795 51.33062
       fit     lwr      upr
1 109.8442 97.0728 122.6155
       fit      lwr      upr
1 55.31803 47.25121 63.38486
       fit      lwr      upr
1 50.73429 42.65208 58.81649
       fit      lwr      upr
1 109.8442 94.74204 124.9463
```

Figure 3: Prediction on Lasso Regression

Conclusion

We perform various regression models in this project. The lasso regression model has a better performance in predicting housing prices since it can explain more observations and have high prediction accuracy when we only use the continuous variables in the original data set. To further improve our model, we can use all predictors to build models. In this way, we can consider more factors that can influence sale price, then we may achieve a better prediction performance. Moreover, since the real estate market closely relates to the economy. We can also connect the economic environment to housing prices, which will generate a spatial-temporal relationship for prediction models.

Reference:

Frew, J., & Jud, G. (2003). Estimating the value of apartment buildings. *Journal of Real Estate Research*, 25(1), 77-86.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934.

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, *19*(3).

Code Soucre:

kassambara. (2018) Penalized Regression Essentials: Ridge, Lasso & Elastic Net. *Statistical tools for high-throughput data analysis*. Retrieved from: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/#elastic-net

Luis. (2011) Linear regression with correlated data. Retrieved from: https://www.r-bloggers.com/linear-regression-with-correlated-data/

Appendix

Fields	Description
Lot Frontage	Linear feet of street connected to property
Lot Area	Lot size in square feet
Mas Vnr Area	Masonry veneer area in square feet
BsmtFin SF 1	Type 1 finished square feet
BsmtFin SF 2	Type 2 finished square feet
Bsmt Unf SF	Unfinished square feet of basement area
Total Bsmt SF	Total square feet of basement area
1st Flr SF	First Floor square feet
2nd Flr SF	Second floor square feet
Low Qual Fin SF	Low quality finished square feet (all floors)
Gr Liv Area	Above grade (ground) living area square feet
Garage Area	Size of garage in square feet
Wood Deck SF	Wood deck area in square feet
Open Porch SF	Open porch area in square feet
Enclosed Porch	Enclosed porch area in square feet
3-Ssn Porch	Three season porch area in square feet
Screen Porch	Screen porch area in square feet
Pool Area	Pool area in square feet
Misc Val	$Value of miscellaneous feature
SalePrice	Sale price $$

Table 1: Data filed description.

```
colSums(sapply(housing, is.na))
```

```
##     Lot Frontage        Lot Area    Mas Vnr Area    BsmtFin SF 1    BsmtFin SF 2
##              490               0              23               1               1
##      Bsmt Unf SF   Total Bsmt SF      1st Flr SF      2nd Flr SF Low Qual Fin SF
##                1               1               0               0               0
##      Gr Liv Area     Garage Area    Wood Deck SF   Open Porch SF  Enclosed Porch
##                0               1               0               0               0
##       3Ssn Porch    Screen Porch       Pool Area        Misc Val       SalePrice
##                0               0               0               0               0
```

Output 1: Report for missing value

```
 Lot.Frontage      Lot.Area        Mas.Vnr.Area      BsmtFin.SF.1     BsmtFin.SF.2
Min.   : 21.00   Min.   :  1300   Min.   :   0.00   Min.   :   0.0   Min.   :   0.00
1st Qu.: 58.00   1st Qu.:  7207   1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:   0.00
Median : 68.00   Median :  9247   Median :   0.00   Median : 338.0   Median :   0.00
Mean   : 69.18   Mean   :  9708   Mean   :  99.92   Mean   : 426.4   Mean   :  46.93
3rd Qu.: 80.00   3rd Qu.: 11202   3rd Qu.: 158.00   3rd Qu.: 716.0   3rd Qu.:   0.00
Max.   :313.00   Max.   :215245   Max.   :1600.00   Max.   :5644.0   Max.   :1474.00
 Bsmt.Unf.SF      X1st.Flr.SF    X2nd.Flr.SF     Low.Qual.Fin.SF    Garage.Area
Min.   :   0.0   Min.   : 334   Min.   :   0.0   Min.   :   0.000   Min.   :   0.0
1st Qu.: 228.0   1st Qu.: 866   1st Qu.:   0.0   1st Qu.:   0.000   1st Qu.: 308.0
Median : 486.0   Median :1073   Median :   0.0   Median :   0.000   Median : 477.0
Mean   : 576.1   Mean   :1153   Mean   : 330.6   Mean   :   5.151   Mean   : 468.7
3rd Qu.: 817.0   3rd Qu.:1378   3rd Qu.: 688.0   3rd Qu.:   0.000   3rd Qu.: 576.0
Max.   :2336.0   Max.   :5095   Max.   :2065.0   Max.   :1064.000   Max.   :1488.0
 Wood.Deck.SF     Open.Porch.SF    Enclosed.Porch     X3Ssn.Porch       Screen.Porch
Min.   :  0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000   Min.   :  0.00
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:  0.00
Median :  0.00   Median : 25.00   Median :   0.00   Median :  0.000   Median :  0.00
Mean   : 89.21   Mean   : 46.78   Mean   :  23.67   Mean   :  2.434   Mean   : 16.17
3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:   0.00   3rd Qu.:  0.000   3rd Qu.:  0.00
Max.   :870.00   Max.   :742.00   Max.   :1012.00   Max.   :508.000   Max.   :576.00
  Pool.Area         Misc.Val           SalePrice
Min.   :  0.00   Min.   :    0.00   Min.   : 12789
1st Qu.:  0.00   1st Qu.:    0.00   1st Qu.:127500
Median :  0.00   Median :    0.00   Median :157000
Mean   :  2.41   Mean   :   44.48   Mean   :179706
3rd Qu.:  0.00   3rd Qu.:    0.00   3rd Qu.:212000
Max.   :800.00   Max.   :17000.00   Max.   :755000
```

sd(data$Lot.Frontage)	23.36302
sd(data$Lot.Area)	6443.529
sd(data$Mas.Vnr.Area)	180.0843
sd(data$BsmtFin.SF.1)	462.835
sd(data$BsmtFin.SF.2)	162.3563
sd(data$Bsmt.Unf.SF)	444.0202
sd(data$X1st.Flr.SF)	397.63
sd(data$X2nd.Flr.SF)	421.4685
sd(data$Low.Qual.Fin.SF)	48.60975
sd(data$Garage.Area)	222.0557
sd(data$Wood.Deck.SF)	120.8667
sd(data$Open.Porch.SF)	67.95858
sd(data$Enclosed.Porch)	64.38114
sd(data$X3Ssn.Porch)	24.69179
sd(data$Screen.Porch)	56.38898
sd(data$Open.Porch.SF)	67.95858
sd(data$Pool.Area)	36.17878
sd(data$Misc.Val)	503.2732
sd(data$SalePrice	83348.92

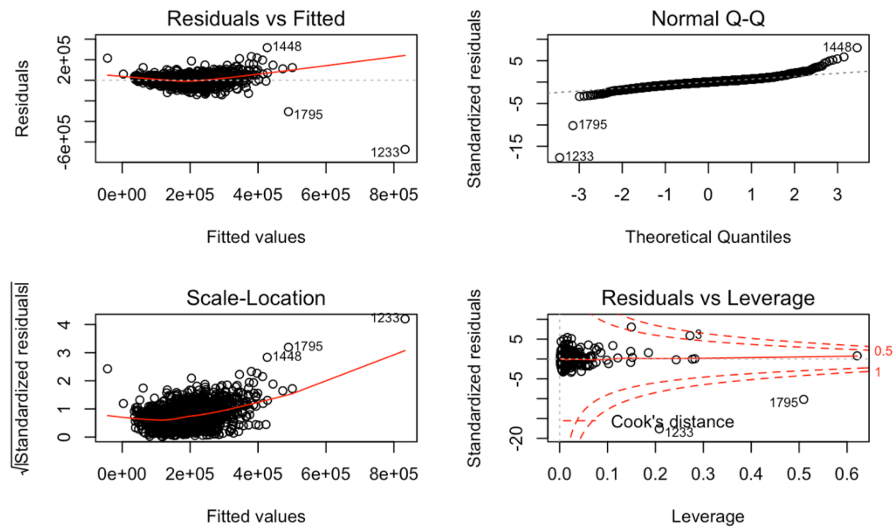
Output 2: Summary statistics

```
Call:
lm(formula = SalePrice ~ ., data = housing_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-672724  -20059    -159   19555  318004 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.597e+04  4.195e+03  -3.808 0.000145 ***
Lot.Frontage    -9.745e+01  5.350e+01  -1.821 0.068709 .  
Lot.Area         5.044e-01  1.741e-01   2.897 0.003819 ** 
Mas.Vnr.Area     5.826e+01  6.942e+00   8.393  < 2e-16 ***
BsmtFin.SF.1     5.413e+01  4.317e+00  12.540  < 2e-16 ***
BsmtFin.SF.2     3.263e+01  7.325e+00   4.455 8.92e-06 ***
Bsmt.Unf.SF      3.807e+01  4.200e+00   9.064  < 2e-16 ***
X1st.Flr.SF      6.523e+01  4.932e+00  13.226  < 2e-16 ***
X2nd.Flr.SF      6.450e+01  2.905e+00  22.205  < 2e-16 ***
Low.Qual.Fin.SF -3.296e-01  2.102e+01  -0.016 0.987488    
Garage.Area      9.028e+01  5.910e+00  15.275  < 2e-16 ***
Wood.Deck.SF     6.262e+01  8.941e+00   7.004 3.50e-12 ***
Open.Porch.SF    4.795e+01  1.652e+01   2.903 0.003746 ** 
Enclosed.Porch  -5.882e+01  1.736e+01  -3.388 0.000720 ***
X3Ssn.Porch      2.420e+01  4.166e+01   0.581 0.561311    
Screen.Porch     6.022e+01  1.791e+01   3.362 0.000791 ***
Pool.Area       -9.462e+01  2.865e+01  -3.302 0.000977 ***
Misc.Val        -1.915e+01  1.807e+00 -10.598  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42900 on 1797 degrees of freedom
Multiple R-squared:  0.7276, Adjusted R-squared:  0.725 
F-statistic: 282.3 on 17 and 1797 DF,  p-value: < 2.2e-16
```

Output 3: Full Linear Model Summary

```
Analysis of Variance Table

Response: SalePrice
                  Df     Sum Sq    Mean Sq   F value    Pr(>F)    
Lot.Frontage       1 1.5631e+12 1.5631e+12  849.3121 < 2.2e-16 ***
Lot.Area           1 3.0669e+11 3.0669e+11  166.6389 < 2.2e-16 ***
Mas.Vnr.Area       1 2.4919e+12 2.4919e+12 1353.9554 < 2.2e-16 ***
BsmtFin.SF.1       1 5.4662e+11 5.4662e+11  297.0017 < 2.2e-16 ***
BsmtFin.SF.2       1 5.3178e+07 5.3178e+07    0.0289  0.865043    
Bsmt.Unf.SF        1 1.2642e+12 1.2642e+12  686.9036 < 2.2e-16 ***
X1st.Flr.SF        1 1.6326e+11 1.6326e+11   88.7057 < 2.2e-16 ***
X2nd.Flr.SF        1 1.5782e+12 1.5782e+12  857.5166 < 2.2e-16 ***
Low.Qual.Fin.SF    1 2.7285e+09 2.7285e+09    1.4825  0.223538    
Garage.Area        1 5.6426e+11 5.6426e+11  306.5852 < 2.2e-16 ***
Wood.Deck.SF       1 6.7888e+10 6.7888e+10   36.8866 1.524e-09 ***
Open.Porch.SF      1 1.0428e+10 1.0428e+10    5.6661  0.017400 *  
Enclosed.Porch     1 2.8440e+10 2.8440e+10   15.4527 8.780e-05 ***
X3Ssn.Porch        1 3.6403e+08 3.6403e+08    0.1978  0.656563    
Screen.Porch       1 1.9468e+10 1.9468e+10   10.5779  0.001166 ** 
Pool.Area          1 1.8485e+10 1.8485e+10   10.0439  0.001554 ** 
Misc.Val           1 2.0672e+11 2.0672e+11  112.3199 < 2.2e-16 ***
Residuals       1797 3.3073e+12 1.8405e+09                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output 4: ANOVA for full linear model



Output 5: Diagnostic plots for full linear model

```
vif(full_linear)
```

```
##    Lot.Frontage        Lot.Area    Mas.Vnr.Area    BsmtFin.SF.1    BsmtFin.SF.2 
##        1.535619        1.386764        1.434204        4.053617        1.439785 
##     Bsmt.Unf.SF     X1st.Flr.SF     X2nd.Flr.SF Low.Qual.Fin.SF     Garage.Area 
##        3.399242        3.841746        1.437851        1.023035        1.619682 
##    Wood.Deck.SF   Open.Porch.SF  Enclosed.Porch     X3Ssn.Porch    Screen.Porch 
##        1.174250        1.185472        1.069250        1.007203        1.037652 
##       Pool.Area        Misc.Val 
##        1.069327        1.055056
```

Output 6: VIF result for full linear model

```
Call:
lm(formula = SalePrice ~ . - BsmtFin.SF.2 - Low.Qual.Fin.SF - 
    X3Ssn.Porch - Lot.Frontage, data = housing_train_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-672015  -20733    -189   19952  314215 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.753e+04  3.888e+03  -4.508 6.96e-06 ***
Lot.Area        4.567e-01  1.639e-01   2.787 0.005379 ** 
Mas.Vnr.Area    5.901e+01  6.977e+00   8.458  < 2e-16 ***
BsmtFin.SF.1    4.546e+01  3.842e+00  11.831  < 2e-16 ***
Bsmt.Unf.SF     2.883e+01  3.641e+00   7.918 4.19e-15 ***
X1st.Flr.SF     7.017e+01  4.582e+00  15.315  < 2e-16 ***
X2nd.Flr.SF     6.243e+01  2.893e+00  21.576  < 2e-16 ***
Garage.Area     9.053e+01  5.896e+00  15.353  < 2e-16 ***
Wood.Deck.SF    6.787e+01  8.925e+00   7.605 4.55e-14 ***
Open.Porch.SF   5.246e+01  1.658e+01   3.164 0.001583 ** 
Enclosed.Porch -6.049e+01  1.740e+01  -3.477 0.000519 ***
Screen.Porch    6.404e+01  1.798e+01   3.562 0.000378 ***
Pool.Area      -9.035e+01  2.870e+01  -3.148 0.001672 ** 
Misc.Val       -1.915e+01  1.817e+00 -10.535  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43170 on 1798 degrees of freedom
Multiple R-squared:  0.7238,	Adjusted R-squared:  0.7218 
F-statistic: 362.4 on 13 and 1798 DF,  p-value: < 2.2e-16
```

Output 7: Model Summary for Sub Linear Model

```
lm(formula = ((SalePrice)^(1/3)) ~ ., data = housing_train_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-60.536  -1.856   0.287   2.232  19.809 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.630e+01  3.927e-01  92.448  < 2e-16 ***
Lot.Frontage    -6.608e-03  5.006e-03  -1.320 0.187018    
Lot.Area         2.939e-05  1.629e-05   1.804 0.071411 .  
Mas.Vnr.Area     3.149e-03  6.495e-04   4.849 1.35e-06 ***
BsmtFin.SF.1     5.328e-03  4.047e-04  13.166  < 2e-16 ***
BsmtFin.SF.2     3.854e-03  6.860e-04   5.619 2.23e-08 ***
Bsmt.Unf.SF      4.074e-03  3.938e-04  10.345  < 2e-16 ***
X1st.Flr.SF      5.994e-03  4.617e-04  12.984  < 2e-16 ***
X2nd.Flr.SF      6.389e-03  2.722e-04  23.472  < 2e-16 ***
Low.Qual.Fin.SF -1.419e-03  1.966e-03  -0.722 0.470347    
Garage.Area      9.444e-03  5.538e-04  17.053  < 2e-16 ***
Wood.Deck.SF     6.516e-03  8.371e-04   7.784 1.18e-14 ***
Open.Porch.SF    5.158e-03  1.545e-03   3.338 0.000862 ***
Enclosed.Porch  -7.789e-03  1.627e-03  -4.789 1.81e-06 ***
X3Ssn.Porch      4.372e-03  3.897e-03   1.122 0.261994    
Screen.Porch     6.257e-03  1.676e-03   3.734 0.000195 ***
Pool.Area       -1.302e-02  2.680e-03  -4.857 1.29e-06 ***
Misc.Val        -1.676e-03  1.690e-04  -9.917  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.013 on 1794 degrees of freedom
Multiple R-squared:  0.7337,	Adjusted R-squared:  0.7312 
F-statistic: 290.8 on 17 and 1794 DF,  p-value: < 2.2e-16
```

Output 8: Model Summary for Full Model with Cubic Root Transformation

```r
modAIC <- MASS::stepAIC(cubic_linear, k = 2,trace = FALSE)
modBIC<- MASS::stepAIC(cubic_linear, k = log(nrow(housing_train_1)),trace = FALSE)
summary(modAIC)
```

```
##
## Call:
## lm(formula = ((SalePrice)^(1/3)) ~ Lot.Area + Mas.Vnr.Area +
##     BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF + X1st.Flr.SF +
##     X2nd.Flr.SF + Garage.Area + Wood.Deck.SF + Open.Porch.SF +
##     Enclosed.Porch + Screen.Porch + Pool.Area + Misc.Val, data = housing_train_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.211  -1.875   0.280   2.223  19.633
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.611e+01  3.626e-01  99.590  < 2e-16 ***
## Lot.Area        2.206e-05  1.529e-05   1.443 0.149099
## Mas.Vnr.Area    3.146e-03  6.489e-04   4.849 1.35e-06 ***
## BsmtFin.SF.1    5.365e-03  4.042e-04  13.272  < 2e-16 ***
## BsmtFin.SF.2    3.878e-03  6.856e-04   5.656 1.80e-08 ***
## Bsmt.Unf.SF     4.096e-03  3.935e-04  10.408  < 2e-16 ***
## X1st.Flr.SF     5.849e-03  4.499e-04  13.001  < 2e-16 ***
## X2nd.Flr.SF     6.355e-03  2.714e-04  23.414  < 2e-16 ***
## Garage.Area     9.376e-03  5.491e-04  17.077  < 2e-16 ***
## Wood.Deck.SF    6.526e-03  8.364e-04   7.802 1.02e-14 ***
## Open.Porch.SF   5.087e-03  1.545e-03   3.294 0.001009 **
## Enclosed.Porch -8.050e-03  1.617e-03  -4.977 7.07e-07 ***
## Screen.Porch    6.138e-03  1.674e-03   3.666 0.000253 ***
## Pool.Area      -1.327e-02  2.674e-03  -4.964 7.55e-07 ***
## Misc.Val       -1.668e-03  1.690e-04  -9.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.013 on 1797 degrees of freedom
## Multiple R-squared:  0.7332, Adjusted R-squared:  0.7311
## F-statistic: 352.8 on 14 and 1797 DF,  p-value: < 2.2e-16
```

```r
summary(modBIC)
```

```
##
## Call:
## lm(formula = ((SalePrice)^(1/3)) ~ Mas.Vnr.Area + BsmtFin.SF.1 +
##     BsmtFin.SF.2 + Bsmt.Unf.SF + X1st.Flr.SF + X2nd.Flr.SF +
##     Garage.Area + Wood.Deck.SF + Open.Porch.SF + Enclosed.Porch +
##     Screen.Porch + Pool.Area + Misc.Val, data = housing_train_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.621  -1.889   0.281   2.237  19.619
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.1165525  0.3627244  99.570  < 2e-16 ***
## Mas.Vnr.Area    0.0030777  0.0006473   4.754 2.15e-06 ***
## BsmtFin.SF.1    0.0053726  0.0004043  13.289  < 2e-16 ***
## BsmtFin.SF.2    0.0039595  0.0006834   5.793 8.12e-09 ***
## Bsmt.Unf.SF     0.0040798  0.0003935  10.368  < 2e-16 ***
## X1st.Flr.SF     0.0060140  0.0004352  13.819  < 2e-16 ***
## X2nd.Flr.SF     0.0064146  0.0002683  23.912  < 2e-16 ***
## Garage.Area     0.0093956  0.0005491  17.111  < 2e-16 ***
## Wood.Deck.SF    0.0065283  0.0008366   7.803 1.02e-14 ***
## Open.Porch.SF   0.0050007  0.0015440   3.239 0.001222 **
## Enclosed.Porch -0.0080030  0.0016176  -4.947 8.23e-07 ***
## Screen.Porch    0.0062064  0.0016742   3.707 0.000216 ***
## Pool.Area      -0.0132775  0.0026746  -4.964 7.55e-07 ***
## Misc.Val       -0.0016592  0.0001689  -9.823  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.015 on 1798 degrees of freedom
## Multiple R-squared:  0.7329, Adjusted R-squared:  0.731
## F-statistic: 379.5 on 13 and 1798 DF,  p-value: < 2.2e-16
```

Output 9: Model Summary for AIC and BIC selection

```r
library(nlme)
model_gls = gls((SalePrice)^(1/3)~.-BsmtFin.SF.2
                -Low.Qual.Fin.SF-X3Ssn.Porch,correlation = corAR1(form=~1)
                ,data=housing_train_1)
summary(model_gls)
```

```
## Generalized least squares fit by REML
##   Model: (SalePrice)^(1/3) ~ . - BsmtFin.SF.2 - Low.Qual.Fin.SF - X3Ssn.Porch
##   Data: housing_train_1
##        AIC      BIC    logLik
##   10414.77 10508.17 -5190.385
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##         Phi
## 0.003381842
##
## Coefficients:
##                   Value Std.Error  t-value p-value
## (Intercept)    36.49747 0.3946071 92.49066  0.0000
## Lot.Frontage   -0.00733 0.0050446 -1.45282  0.1464
## Lot.Area        0.00004 0.0000164  2.29145  0.0221
## Mas.Vnr.Area    0.00325 0.0006539  4.97363  0.0000
## BsmtFin.SF.1    0.00428 0.0003603 11.89323  0.0000
## Bsmt.Unf.SF     0.00295 0.0003415  8.63508  0.0000
## X1st.Flr.SF     0.00681 0.0004406 15.45352  0.0000
## X2nd.Flr.SF     0.00618 0.0002719 22.72309  0.0000
## Garage.Area     0.00965 0.0005568 17.33435  0.0000
## Wood.Deck.SF    0.00708 0.0008371  8.45395  0.0000
## Open.Porch.SF   0.00563 0.0015543  3.62326  0.0003
## Enclosed.Porch -0.00786 0.0016323 -4.81719  0.0000
## Screen.Porch    0.00670 0.0016856  3.97423  0.0001
## Pool.Area      -0.01207 0.0026967 -4.47605  0.0000
## Misc.Val       -0.00169 0.0001704 -9.89579  0.0000
```

Output 10: Model Summary for General Least Squared Regression

The following code is for ridge, lasso, and elastic net:

```{r}
lambda <- 10^seq(-3, 3, length = 100)
# Ridge
set.seed(123)
ridge <- train(
  (SalePrice)^(1/3) ~., data = housing_train_1, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 0, lambda = lambda)
  )
coef(ridge$finalModel, ridge$bestTune$lambda)
predictions <- ridge %>% predict(housing_test)
data.frame(
 Rsquare = R2(predictions, housing_test$SalePrice),
  RMSE = RMSE(predictions, housing_test$SalePrice),
  MAE = RMSE(predictions, housing_test$SalePrice)
)
# Lasso
set.seed(123)
lasso <- train(
  (SalePrice)^(1/3) ~., data = housing_train_1, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 1, lambda = lambda)
  )
coef(lasso$finalModel, lasso$bestTune$lambda)
predictions <- lasso %>% predict(housing_test)
data.frame(
  Rsquare = R2(predictions, housing_test$SalePrice),
  RMSE = RMSE(predictions, housing_test$SalePrice),
  MAE = RMSE(predictions, housing_test$SalePrice)
)
# Elastic Net
set.seed(123)
elastic <- train(
  (SalePrice)^(1/3) ~., data = housing_train_1, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneLength = 10
  )
coef(elastic$finalModel, elastic$bestTune$lambda)
predictions <- elastic %>% predict(housing_test)
data.frame(
  Rsquare = R2(predictions, housing_test$SalePrice),
  RMSE = RMSE(predictions, housing_test$SalePrice),
  MAE = RMSE(predictions, housing_test$SalePrice)
)
```