Step1-4 Summary

Liuqian Bao

2023-12-12

Introduction

Throughout the quarter, we have investigated the housing Data from residential homes in Ames, Iowa, in an attempt to learn what influences the sale price of each property. Our response variable was therefore: SalePrice - the property's sale price in dollars. Originally, there were many predictor variables, both categorical and quantitative. However, as we worked, we used various methods to use only the most influencial variables, and to minimize the noise created by irrelevant information.

The citation for our original data source is: Anna Montoya, DataCanary. (2016).House Prices - Advanced Regression Techniques. Kaggle.Obtained from https://kaggle.com/competitions/house-prices-advanced-regression-techniques (https://kaggle.com/competitions/house-prices^advanced-regression-techniques)

In step 1, we explored each variable, attempting to understand clearly what we were investigating, and to have visuals of any potential correlation.

The following are two graphs we found the most informative:

The Scatterplot we Based Step 2 on: SalePrice vs BsmtFinSF1

The scatterplot of SalePrice versus BsmtFinSF1 indicates a strong association between BsmtFinSF1 and SalePrice. When BsmtFinSF1 increases, there is a clear corresponding increase in SalePrice, as indicated by the tightly packed data points. We note that there is also a clustering of data points at the 0 value for BsmtFinSF1, suggesting many properties without finished basement space.

A Histogram Showing the Distribution of SalePrice

The distribution of SalePrice for properties in Ames, Iowa shows a very right-skewed shape, which implies that a significant number of houses have lower sale prices, and there are relatively fewer houses with very high sale prices. This may reflect income disparities or variations in property values within the area, with a concentration of less expensive homes and only a few high-end properties.

Next, Explore the Effect of One Variable

Since we noticed that the finished basement square footage seemed to have a positive correlation to the sale price, we decided to fit a linear model with it as the predictor. After exploring the model assumptions, we found that in order to obtain a more constant variance, we needed to transform the response with a log function.

The following graphics display the scatter plot of our predictor vs response, and a residual panel to check the model assumptions. Additionally, we checked a qqplot for normality, and saw that with the log transformation, the points closely follow the line.

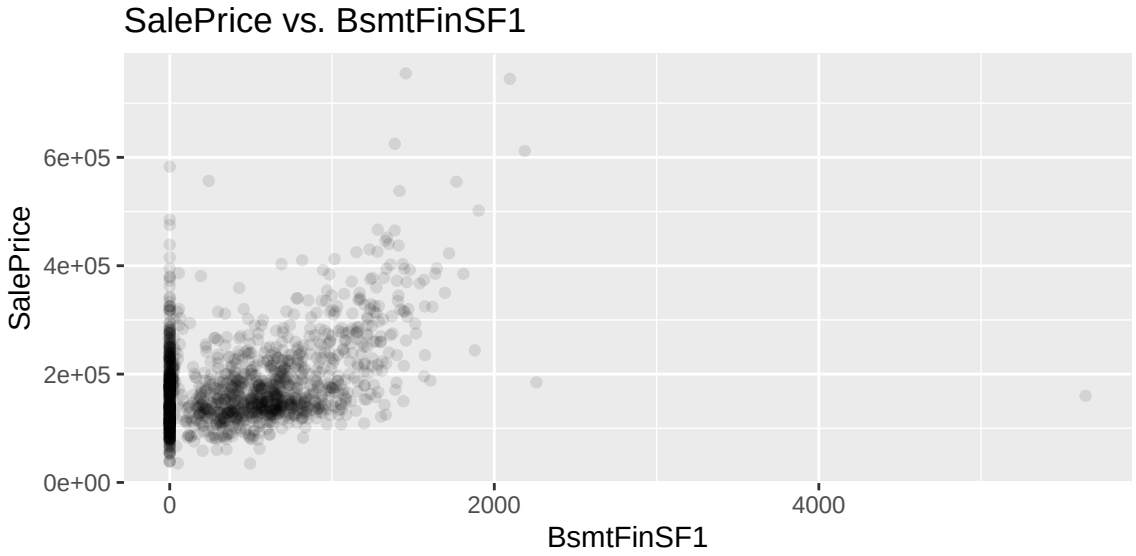


Figure 1: Scatter plot of SalePrice($) against Basement Area(squared ft)

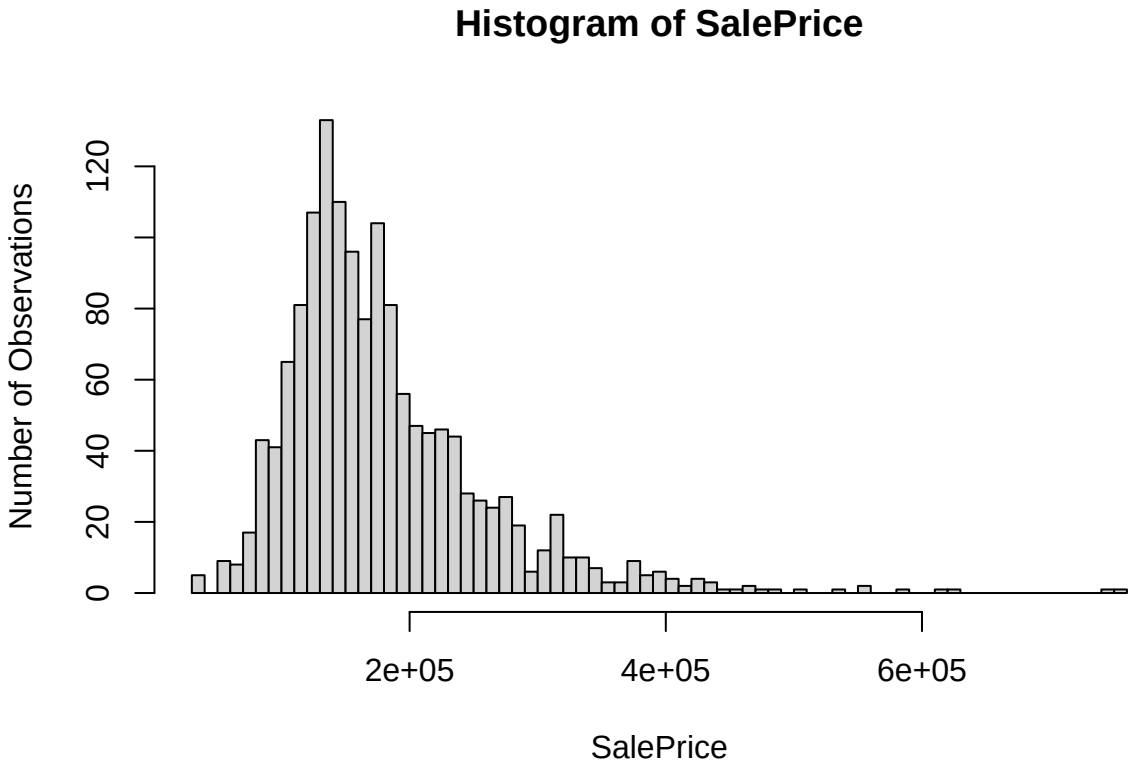


Figure 2: Histogram of SalePrice($)

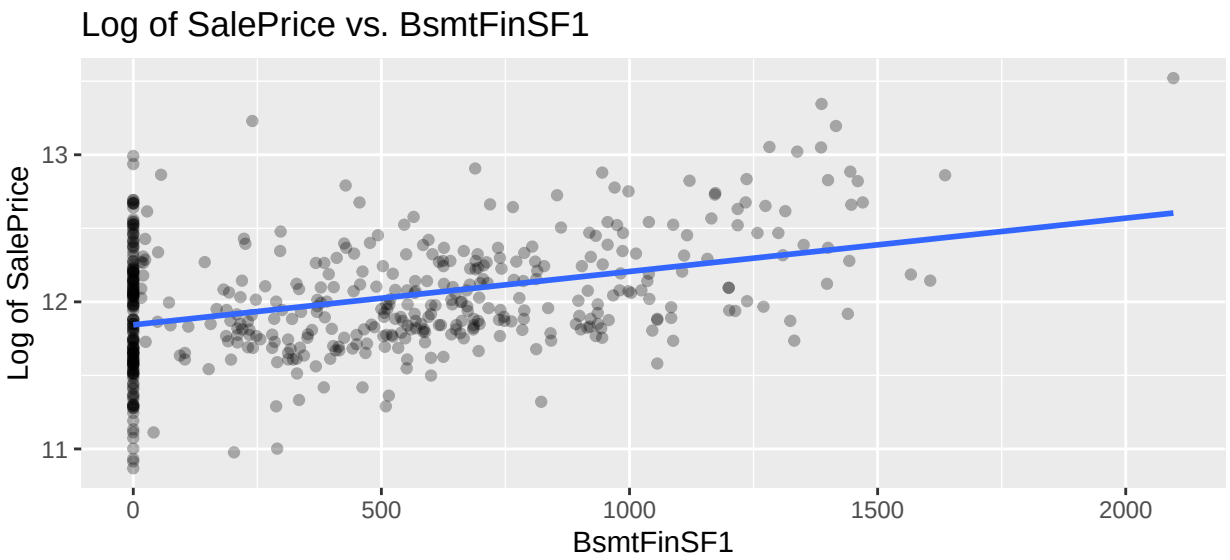


Figure 3: Model with the Log Transformation

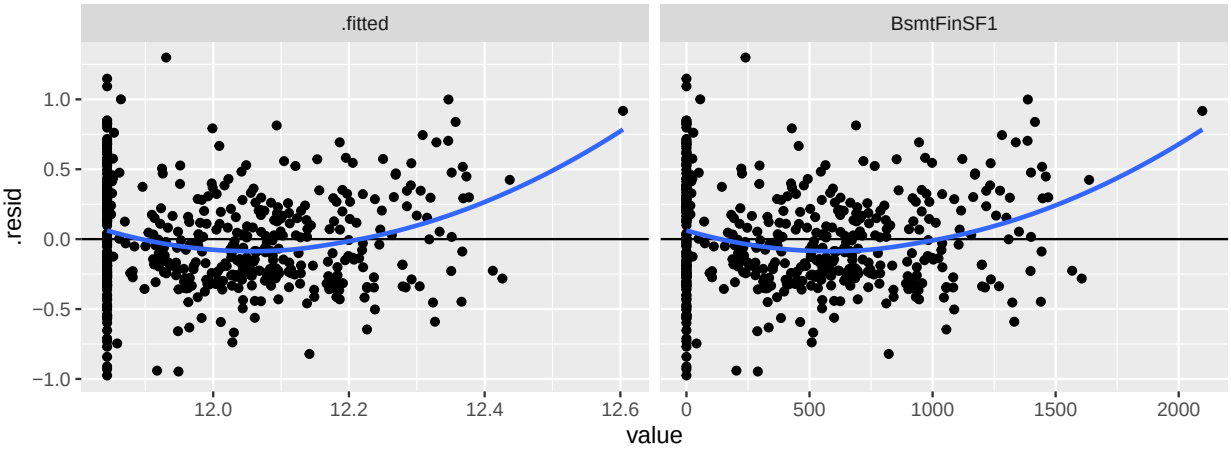
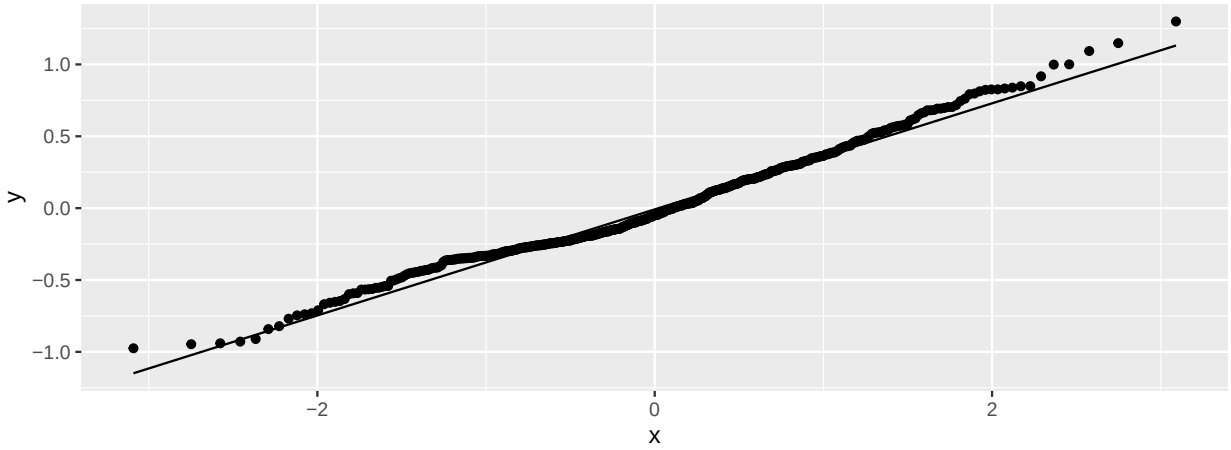


Figure 4: Residual vs. fitted and BsmtFinSF1, Logged Model


1.0
0.5
0.0
−0.5
−1.0
y
−2
0
2
x

Thus, we chose to proceed by logging the response in each model used. We performed a double t-test on the coefficient of BsmtFinSF1. As you can see, the resulting p value allows us to reject our null hypothesis($\beta_1 = 0$) and accept our alternative hypothesis that $\beta_1 > 0$, confirming that BsmtFinSF1 and SalePrice are positively correlated.

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 1.184366e+01 2.395519e-02 494.408910 0.000000e+00
## BsmtFinSF1  3.626989e-04 3.868634e-05   9.375375 2.413948e-19
```

Because the goal is to view how our predictors change the sale price of a property, we constructed a confidence interval for the mean.

The following graph offers a visualization of the mean line.

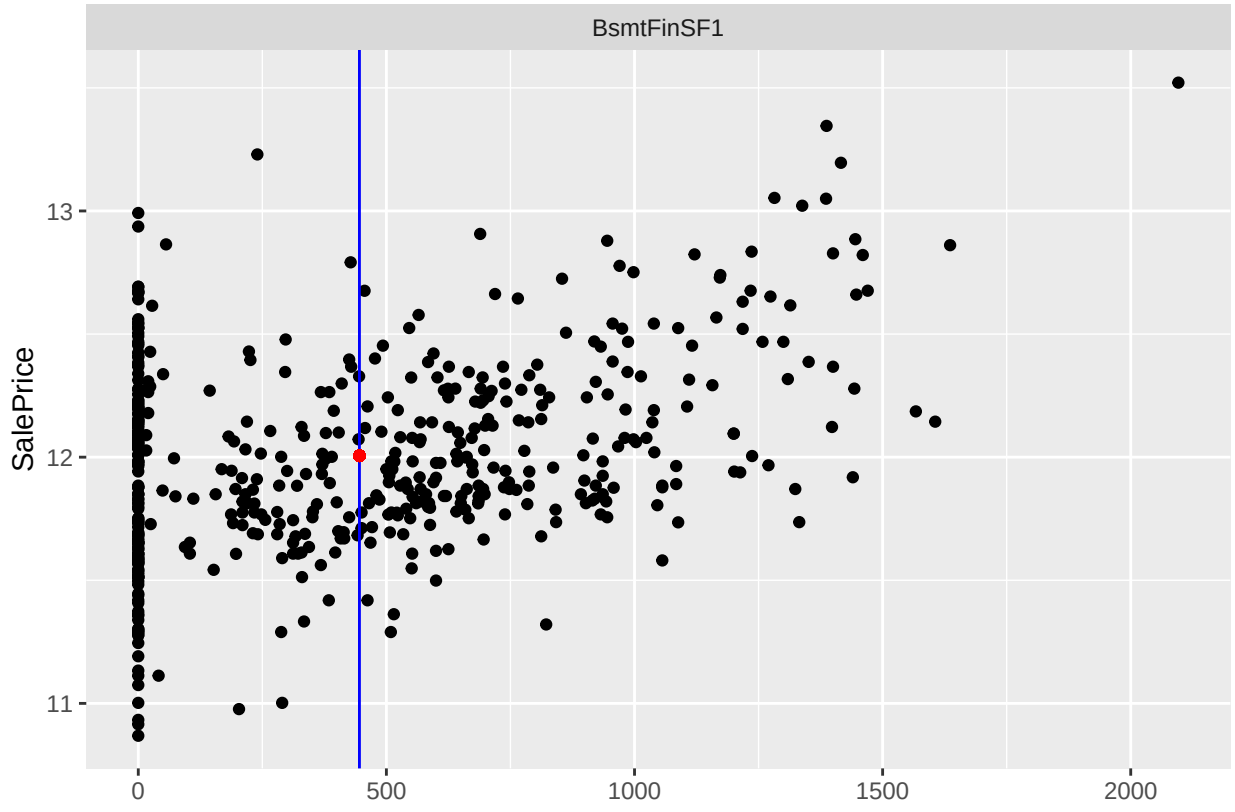


Figure 5: Visualization of Mean Line

According to our test, we determined that with 95% confidence, the mean of the log SalePrice for a house with basement area equal to the average in the data is estimated to be between 11.97263 and 12.03798.

However, the r^2 value of this model was only 14.83%, and thus not much of the variance in the log of SalePrice was explained.

Now that we'd explored how BsmtFinSF1 may affect the response when it is the sole predictor, we decided to construct a model that conveyed more information.

Constructing a More Accurate Model

In order to investigate which predictors held the most relevance, we performed both backward and forward selection on our model, keeping the response as the log of SalePrice.

Each of these methods suggested that the variables with the most significance are external quality(ExterQual), LotArea, GarageCars,finished basement square footage (BsmtFinSF1), HalfBath, FullBath, kitchen quality (KitchenQual), andopen porch square footage (OpenPorchSF).

Fitting this model to our data, we obtained the following summary:

```
##
## Call:
## lm(formula = log(SalePrice) ~ ExterQual + LotArea + GarageCars +
##     BsmtFinSF1 + HalfBath + FullBath + KitchenQual + OpenPorchSF,
##     data = h2_partition$test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62300 -0.09981  0.00000  0.09839  0.59175
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.178e+01  1.009e-01 116.735  < 2e-16 ***
## ExterQualFa   -8.778e-01  2.016e-01  -4.353 2.59e-05 ***
## ExterQualGd   -1.905e-01  8.960e-02  -2.126 0.035242 *
## ExterQualTA   -3.509e-01  9.376e-02  -3.743 0.000266 ***
## LotArea        4.491e-06  8.398e-07   5.348 3.60e-07 ***
## GarageCars     1.080e-01  2.677e-02   4.034 9.04e-05 ***
## BsmtFinSF1     1.950e-04  3.560e-05   5.476 1.99e-07 ***
## HalfBath       1.002e-01  3.104e-02   3.229 0.001550 **
## FullBath       1.725e-01  3.130e-02   5.512 1.69e-07 ***
## KitchenQualFa -2.845e-01  1.116e-01  -2.548 0.011924 *
## KitchenQualGd -7.999e-02  7.493e-02  -1.068 0.287600
## KitchenQualTA -1.889e-01  7.870e-02  -2.400 0.017734 *
## OpenPorchSF    4.452e-04  2.449e-04   1.818 0.071169 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1723 on 138 degrees of freedom
## Multiple R-squared:  0.8135, Adjusted R-squared:  0.7973
## F-statistic: 50.16 on 12 and 138 DF,  p-value: < 2.2e-16
```

As you can see, 81.35% of the variation in SalePrice is explained with this model, and it is therefore a much better fit.

Ensuring We Have Used Variables of Influence

We performed both LASSO and Ridge Regression, attempting to ensure that none of these eight variables was irrelevant, or too closely correlated with another that it incorrectly estimated the response.

After using the optimal lambda values to find the best models, we constructed this graph to compare the three.

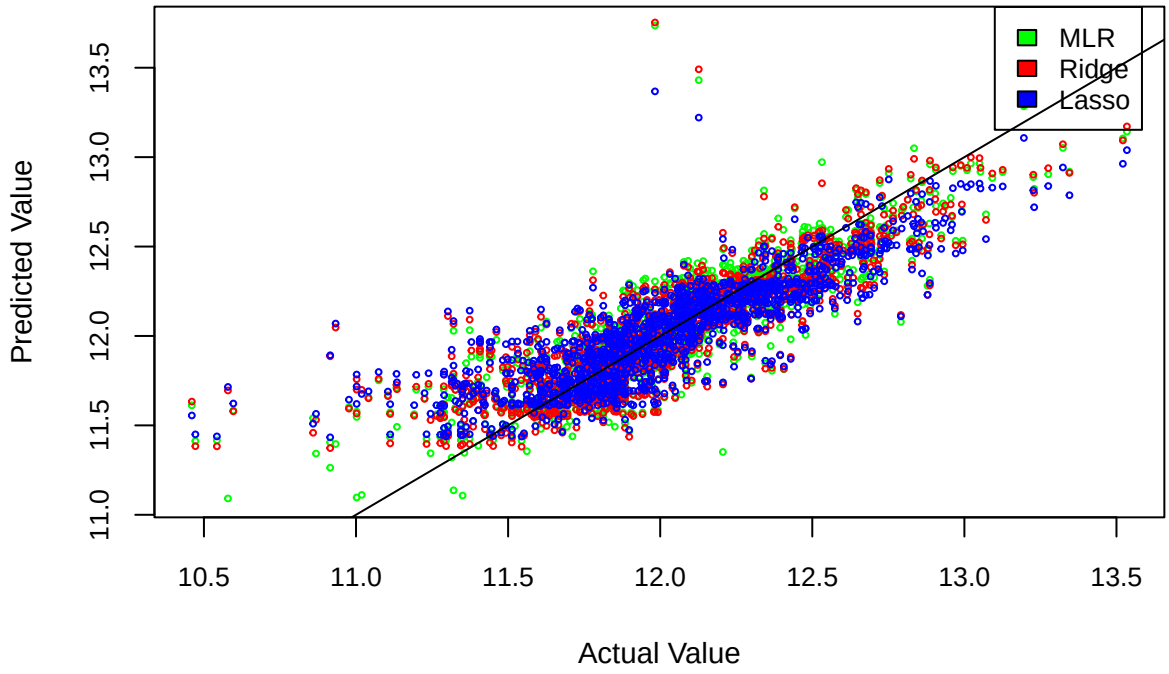


Figure 6: scatter plot of the predicted vs. actual values for MLR, Rridge, and Lasso Models

In summary, the MLR, Ridge, and Lasso models have R-squared values of 0.828, 0.7132, and 0.6972, respectively, suggesting that MLR achieves the best fit. However, by analyzing the predicted vs. response graph, Ridge and Lasso have lower variances in predictions and might generalize better to unseen data due to their reduced complexity.

Conclusion

In conclusion, we found that the most statistically significant variables from our dataset are external quality(ExterQual), LotArea, GarageCars,finished basement square footage (BsmtFinSF1), HalfBath, FullBath, kitchen quality (KitchenQual), andopen porch square footage (OpenPorchSF). Due to this conclusion, we suggest clients use these as factors to be considered when deciding which properties to invest in, or how to price a property for the market.