ECON 4750: Project 1

AUTHOR
Spencer Katzman

Model Descriptions

In this section, I introduce five different models. They are:

```
Model 1: SalePrice ~ GrLivArea
```

Model 1 is a uni-variate model of SalePrice on GrLivArea. GrLivArea represents the square footage of the house that is above ground. I included this as a measure of the size of the house based on the idea that larger houses, on average, sell for more than smaller houses.

```
Model 2: SalePrice ~ BedroomAbvGr + FullBath
```

Model 2 is a bi-variate model of Sales Price on BedroomAbvGr and FullBath. BedroomAbvGr represents the number of bedrooms above ground (not including those in basements) and FullBath represents the number of full bathrooms above ground. I included these two variables because I often hear houses advertised based on the number of bedrooms and bathrooms they have. The idea is that houses with more bedrooms/bathrooms sell for more, on average, than houses with fewer bedrooms/bathrooms.

```
Model 3: SalePrice ~ GrLivArea + BedroomAbvGr + FullBath
```

Model 3 combines GrLivArea, BedroomAbvGr, and FullBath into a single model as I would speculate that these are probably the three most important characteristics of a house that determine its selling price.

```
Model 4: SalePrice ~ GrLivArea + BedroomAbvGr + FullBath + as.factor(Neighborhood)
                                + GrLivArea * as.factor(Neighborhood)
```

Model 4 adds the Neighborhood variable and an interaction of Neighborhoods with GrLivArea to Model 3. The intuition is that the neighborhood in which a house is located will affect its value. The neighborhood information was operationalized using the as.factor function to create 24 dichotomous variables representing the 25 different neighborhoods with the Bloomington Heights (Blmngtn) neighborhood as the baseline (thereby avoiding the dummy variable trap).

```
Model 5: SalePrice ~ (GrLivArea + BedroomAbvGr + FullBath + as.factor(Neighborhood))^3
```

Model 5 is the complicated model requested in the assignment. It takes the variables GrLivArea, BedroomAbvGr, Fullbath, and Neighborhood used in model 4 and cubes them, thereby creating many interactions and higher-order terms. Although inclusion of these variables makes sense in terms of determining the value of a house, I had no definitive reason for cubing them other than to create a very complicated model with 176 coefficients to be estimated.

Model Selection based on Training Data

I began my analysis by estimating the five models using linear regressions. I followed this by estimating my complicated model (5) using Lasso and Ridge regressions. Table 2 lists the various model selection criteria results for these seven regressions. All regressions used the training data and 10 folds for cross-validation.

Table 1: Model Assessment Using Training Data

Model	R-Squared	Adj R-Squared	aic	bic	cv*
LM 1	0.5402211	0.5397604	28720.58	28730.39	54562.54
LM 2	0.3149282	0.3135539	29121.35	29136.08	66581.56
LM 3	0.6157400	0.6145826	28545.15	28564.78	50086.40
LM 4	0.8230613	0.8135425	27913.63	28286.62	37025.29
LM 5	0.8711099	0.8472584	27796.79	28660.56	85792.34
Lasso 5	NA	NA	NA	NA	38863.43
Ridge 5	NA	NA	NA	NA	39028.53

*Note: Values in this column are the square roots of cvs (see Comment 1 below).

Table 1 shows that R-squared, Adjusted R-squared, and aic all rank the five linear regression models in the following order (from best to worst):

```
LM 5, LM 4, LM 3, LM 1, LM 2
```

LM 5 performs the best and LM 2 performs the worst. It is not surprising that LM 1, LM 2, and LM 3 are ranked lowest because they are relatively simple models. I was a bit surprised that LM 5 ranked highest in terms of adjusted R-squared since it uses so many variables but that seems not to have hurt LM 5 very much because $n = 1000$ making the penalty term weight, $\frac{(n-1)}{(n-k)}$, relatively close to 1 despite the k = 176 regressors. In terms of adjusted R-squared, it was encouraging that LM 4 and LM 5 generated adjusted R-square values great than 0.80, indicating that their regressors explained more than 80% of the variance in selling prices. I thought this was very good for models I built based on very limited knowledge of real estate markets.

Things change with bic where LM 4 overtakes LM 5 as the highest ranked model with the three simple linear models ranking in the same order as before. I speculate that LM 4 performs so well because it includes the most relevant information and interaction terms without over complicating the analysis.

Comment 1: Before discussing the cross-validation results, I note that the project instructions asked us to calculate cv $= \sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2$. I calculated cv in this way but the values are very large, so I present their square roots in the Table 1. Because the square root function is monotonic increasing, the ranking based on cv is preserved with this transformation. These cv rankings change dramatically from the previous rankings, yielding the following (from best to worst):

```
LM 4, Lasso 5, Ridge 5 LM 3, LM 1, LM 2 LM 5
```

Interestingly, LM 4 is at the top of the cv rankings with Lasso 5 and Ridge 5 coming in second and third. Still, Ridge 5 and Lasso 5 outperform all the other linear regression models. Of particular interest is that LM 5 was the best model based on R-squared, adjusted R-squared, and aic, but is now ranked last based on cv. This is undoubtedly because of overfitting. At this point, LM 4, Lasso 5, and Ridge 5 are the front runners, the next section will try to hone in on one of these models as the model of choice by considering out-of-sample cross-validation.

Model Selection based on Testing Data

This section compares the seven regressions using out-of-sample data from the house_price_test.csv file. Table 3 presents the results, ordered from best to worst in terms of cv.

Table 3: Out of Sample cv Comparison

Model	cv
Ridge 5	37628.23
LM 4	37660.24
Lasso 5	40093.09
LM 3	55138.63
LM 1	59977.54
LM 2	64109.34
LM 5	78144.33

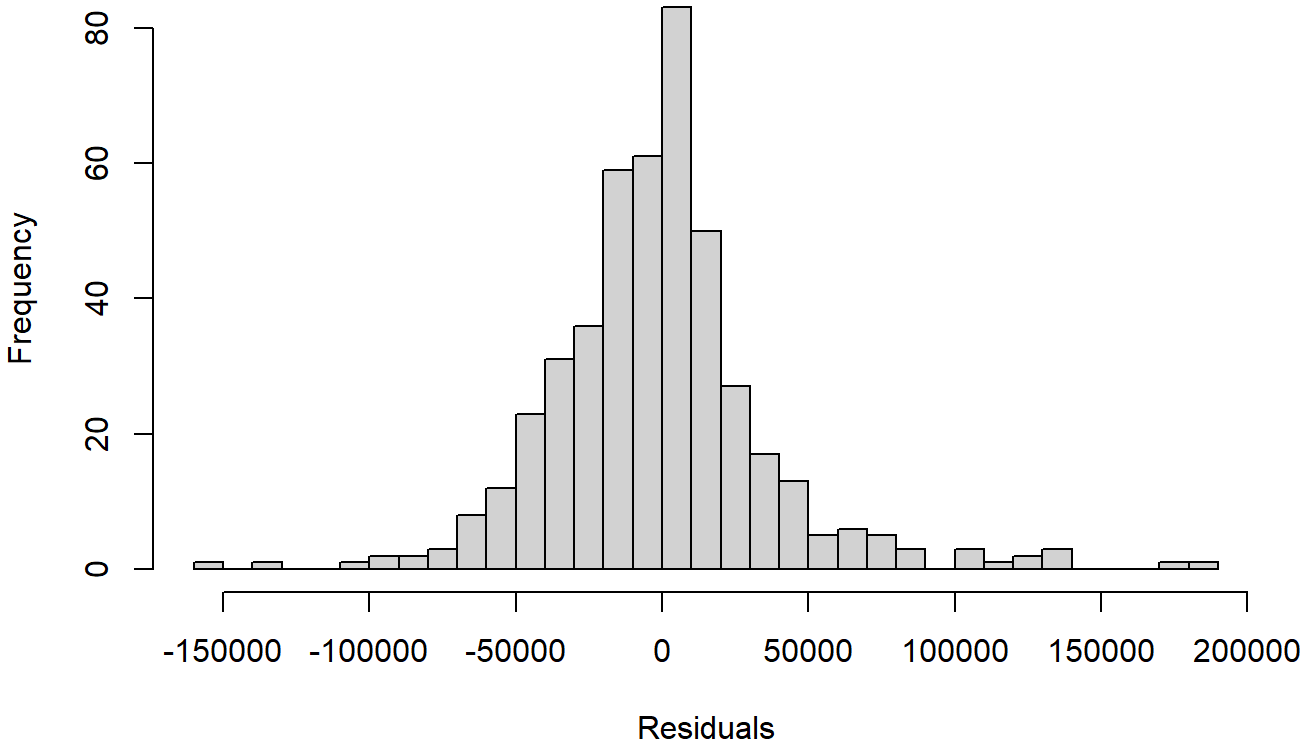
Ridge 5 is ranked highest, overtaking LM 4 which is ranked a very close second. LM 5 is still ranked last, once again likely attributable to the complexity of the model and overfitting. Lasso 5 is now ranked third in terms of cross-validation. The high ranking of Lasso 5 makes sense since it is designed to detect and use the more meaningful interaction and non-linear terms and ignore those that are less meaningful. Regressions LM 3, LM 1, and LM 2 are ranked in the same order as before and are ranked low by cv which is, once again, not surprising because they are relatively simple models. Still, it is worth pointing out that all three linear regressions outperform the bloated regression LM 5.

In the end, Ridge 5 appears to be the best model based on out-of-sample cross-validation and would be my choice to predict house selling prices in Ames, IA. I conclude by investigating just how good Ridge 5 is in terms of out-of-sample prediction quality.

Conclusions

Above, I identified Ridge 5 as the best of my models. I conclude this project with more detailed examination of how well Ridge 5 predicts selling prices by examining how "off the mark" its predictions were using the testing data.

Figure 1: Test Data Residual Histogram (Ridge 5)



Here, I plotted the Ridge 5 residuals in the histogram above for a more granular view than the aggregated cv measure. The histogram in Figure 1 is divided into bins of $10,000. We see that most residuals are between -$50,000 and $50,000. The mean and median absolute prediction error (using the test data) are $25,818.99 and $17,827.64 respectively. Because of a few fairly large residuals, the median provides a better measure of central tendency than does the mean. Given a median selling price of $163,995, $17,827.64 seems reasonably good. Still, I would be hesitant to use any of these models to actually predict house selling prices in Ames, IA if money was on the line. Still, my results suggest that someone who knows more about real estate and the areas in Ames could likely help build a useful model using the most important factors that influence selling price.
