Hosuing Sale Price Regression Analysis

Joey Bringley, Mike Salinas, Franics Dadson, and Joseph Olenraweju

(Dated: 22 November 2016)

At some point or another, most people end up in the market for a house. It might be your first house, it could be your third, who knows? When looking for a home, the biggest factor is almost always the price. We want all of these characteristics of a house and essentially wonder, how much would my dream home cost me? Well, you could look around on the market and hope you find it, and then boom, there is your price. Or we could gather data on home prices, their characteristics, and use regression techniques to estimate what our dream house should cost us.

Using data gathered and sorted by Kaggle, we are able to apply regression analysis to help predict what a home with a given set of characteristics should cost. The data consists of 2,919 homes in Ames, Iowa with 80 variables. The variables range from square footage to the quality of the pool with the price of the home as the dependent variable. Using this data, our goal is to create a regression model that accurately predicts the price of a home in Ames, Iowa given a specific set of features.

I. DATA

The data set we are using is comprised of housing features and prices from Ames, Iowa. There are 1,460 observations in the training data set and 1459 in the test set. There are 80 explanatory variables and our goal is to find out what features of a home best lead us to predict the price of a home in Ames, Iowa.

II. ANALYSIS

To begin the analysis, a linear model was fit on the training data containing all 80 predictors. While initial summary statistics showed high R^2 values ($R^2 = 0.8503, AdjR^2 = 0.842$), looking further into the fit we see major issues. Two of the predictors returned with coefficients 'NA', meaning they are linearly related to other variables in the data. These predictors, 'TotalBsmtSF' and 'GrLivArea' were dropped from the dataset. While this won't necessarily throw off the model's prediction accuracy, the coefficient estimates will be wrong. A quick glance at the coefficients can confirm this suspicion. For example, 'Lot Frontage', actually has a negative coefficient (-90.546). We would expect the price of the house to increase as lot frontage increased.

Furthermore if we examine the residual plot (figure 1), we see while it is not too bad, there is room for improvement. Computing the test MSE on the entire training set yields a value of $943,330,445$. While this value may seem similar to other models we will discuss soon, this is only the MSE of the training set, which is commonly known to underestimate the MSE of the test set.

Due to our problem with collinearity in the data, and

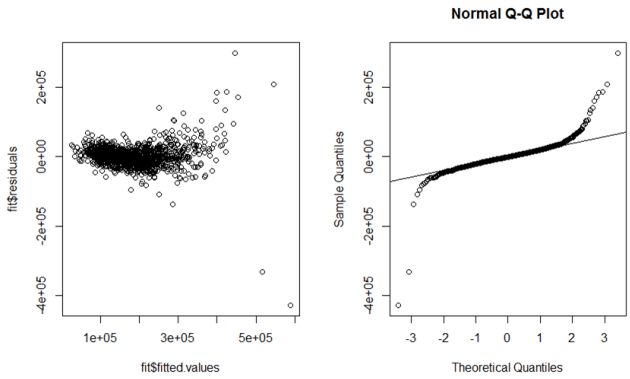

FIG. 1. Model Diagnostics for Ordinary Least Squares

a large number of predictor variables, subset selection methods were next explored, using the package *regsubsets* in R. While the best subset selection approach is the most thorough, looking at every possible model, the data is much too large. With 80 predictors, there are 2^{80} possible models. Even if we could fit $100,000$ models per second, it would take 380 *billion years* to complete.

Next considered was forward and backwards selection. The forward model was fit with a limit of 78 predictors, which would be the full model. The following graph (figure 2) shows the following model measures: RSS, $AdjR^2$, C_p, BIC for the fits selected.

Here we see some conflicting results. Depending on the measure we choose, we will select a different model with a different number of predictors for each. $AdjR^2$ finds the best model is that with 54 predictors, C_p best model has 46 predictors, and the best model according to BIC only contains 22 predictors. For the sake of this paper, we will use $Mallow's\ C_p$ as the final criterion for model selection.

If we look at the plots for backwards stepwise selection we see similar results. The minimum value of $Mallow's C_p$ occurs at the 44 variable model, while the highest $AdjustedR^2$ occurs at the 54 variable model.

Using 10-fold C.V. we measure the MSE on the training set. The result is rather large(1.4 billion), so since we still see a large test MSE using forwards stepwise regres-

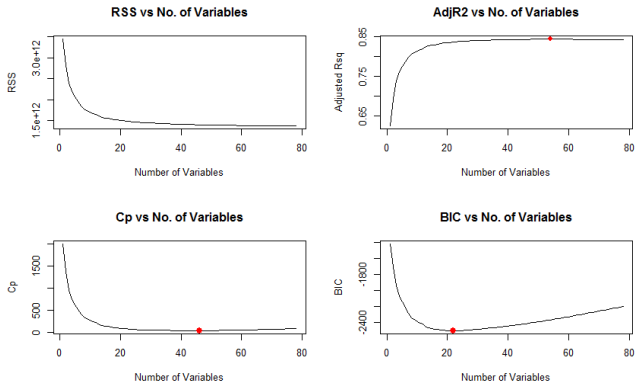

FIG. 2. Model Measures for Forward Selection

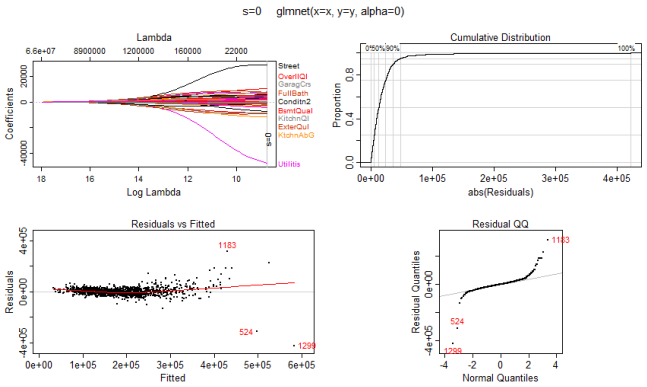

FIG. 4. Ridge Regression Diagnostics

sion, we decied to turn to shrinkage methods for improved performance, namely Ridge and Lasso Regression.

With 10-fold cross validation, our ridge regression model was fit and the best lambda value was determined to be $55,844$. The model showed a significant improvement in terms of model accuracy, producing a Root MSE of 30,000 less than the ordinary least squares estimate.

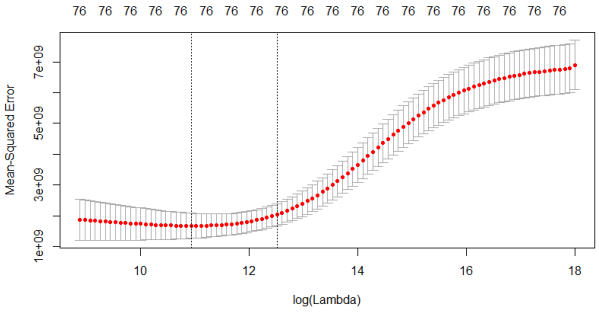

FIG. 3. Ridge Regression MSE for corresponding values of λ

Further examining the diagnostic of the ridge fit, observing the plot below (figure 4), we see a decent residual plot. While there could be more spread, it seems to indicate constant error variance. The normal probabilty plot could be better as well, but it doesn't look too bad.

Ridge Regression also improved our score on Kaggle! Measured by RMSLE (Root Mean Squared Logarithmic Error), we improved by almost 3 percentage points, bringing us up 130 positions, to where we sit at 1165 out of 1600. Comparing our results with OLS, we see intuitively our coefficients have improved. For example, 'Lot Frontage' which should probably have a possible correlation with housing price, now has a positive coefficient.

Lastly, LASSO regression was explored to see if there was any improvement over ridge regression. Using 10-fold cross validation, the best λ for our model was determined to be 3,348.

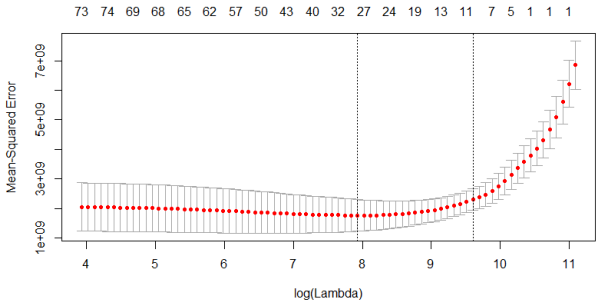

FIG. 5. Lambda vs MSE for Lasso Model

Test RMSE with this value for λ was 31,902.75. While this is an improvement over ordinary least squares, it is not in comparison to ridge. Our test RMSE increased by 10,696 with LASSO.

Examining the Lasso diagnostics (figure 6), we see a

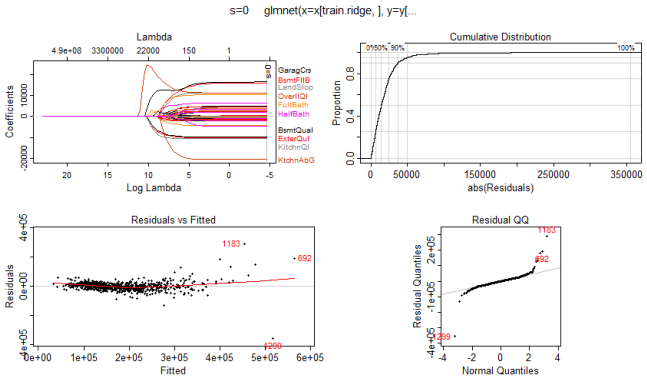

FIG. 6. Lasso Regression Diagnostics

similar results in line with ridge regression. The residual

plot, while it could be better, seems pretty good. The normal probability plot shows that a box plot transform could possibly be used here, but because this model under performed compared to ridge regression, this measures was not taken.

With Lasso Regression only, 35 predictors are kept. The rest are shrunk to zero. Submitting our predictions to Kaggle, our test MSE was a good indicator. While it performed better than OLS, the RMSLE was signifcantly worse than our score for ridge regression.

III. CONCLUSION

While the output from Lasso regression is often easier to interpret due to the subset selection process, here we see ridge regression has better prediction accuracy. Lasso's problem can be that it implicitly assumes some of the coefficients to be equal to zero. If all predictors are meaningful and related to the response, we will often see Lasso perform worse than ridge. This is the case here.

Generally speaking, we may state that we believe Lasso to perform better when there are only a few significant predictors in the data set. Both however are an improvement over least squares in terms of prediction accuracy. As we see λ grow in both cases, the coefficients shrink towards zero. This causes a reduction in the variance of the predictions, only costing us a slight increase in bias.

Our final model fit, that being the one using Ridge Regression, is shown in the appendix.

Appendix A: Final Model

(Intercept) 5.337906e+05
Id -1.353103e+00
MSSubClass -4.102425e+01
MSZoning -2.663063e+02
LotFrontage 4.063544e+01
LotArea 2.668741e-01
Street 1.948114e+04
Alley 4.271676e+03
LotShape -1.118865e+03
LandContour 7.707143e+02
Utilities -1.974456e+04
LotConfig -1.764281e+02
LandSlope 3.771087e+03
Neighborhood 2.422258e+02
Condition1 -2.152272e+02
Condition2 -3.458509e+03
BldgType -1.616997e+03
HouseStyle -7.357745e+01
OverallQual 6.041064e+03
OverallCond 1.969990e+03
YearBuilt 7.088864e+01
YearRemodAdd 1.377973e+02
RoofStyle 2.654762e+03
RoofMatl 3.895436e+03
Exterior1st -1.955577e+02
Exterior2nd 3.287614e+01
MasVnrType 2.484954e+03
MasVnrArea 2.485154e+01
ExterQual -8.009146e+03
ExterCond 4.152374e+02
Foundation 2.119204e+03
BsmtQual -6.422088e+03
BsmtCond 8.822989e+02
BsmtExposure -2.379551e+03
BsmtFinType1 -5.474917e+02
BsmtFinSF1 7.808070e+00
BsmtFinType2 5.512379e+01
BsmtFinSF2 2.830381e+00
BsmtUnfSF 1.490878e+00
TotalBsmtSF 1.034584e+01
Heating -2.923268e+02
HeatingQC -1.342692e+03
CentralAir 4.326266e+03
Electrical 2.543874e+02
X1stFlrSF 1.307266e+01
X2ndFlrSF 9.588265e+00
LowQualFinSF -2.446641e+00
GrLivArea 1.367772e+01
BsmtFullBath 4.488749e+03
BsmtHalfBath -9.988887e+01
FullBath 6.786929e+03
HalfBath 4.161897e+03
BedroomAbvGr 3.499633e+02
KitchenAbvGr -9.343653e+03
KitchenQual -7.195114e+03
TotRmsAbvGrd 3.013372e+03

Functional 2.136047e+03
Fireplaces 6.507699e+03
FireplaceQu -1.140577e+03
GarageType -5.428254e+02
GarageYrBlt 5.072145e+01
GarageFinish -3.024353e+03
GarageCars 6.028786e+03
GarageArea 1.743020e+01
GarageQual -2.753471e+02
GarageCond 1.127075e+03
PavedDrive 1.886437e+03
WoodDeckSF 1.899849e+01
OpenPorchSF 1.629161e+01
EnclosedPorch 3.466658e+00
X3SsnPorch 1.580019e+01
ScreenPorch 3.393583e+01
PoolArea 1.759399e-01
PoolQC -2.454014e+02
Fence -4.007113e+02
MiscFeature 3.010772e+02
MiscVal -4.066390e-01
MoSold 8.140928e+01
YrSold -5.030500e+02
SaleType -2.722595e+02
SaleCondition 1.693910e+03