**Assignment 3: Predicting Claim Sizes**

**Part 1: Importing and Fine Tuning Data**

To begin our process, we must first import our necessary packages and libraries. I imported both “corrplot” and “glmnet” packages and libraries. Next, we import our ‘claim size’ data from All-State into R and put it into a data frame. We also want to remove the ID column, as it has no influence on the claim size outcome data. The code to do so follow below.

Text

Description automatically generated with medium confidence

Because we will be using our models to make business decisions, we care far less about small claim sizes then we do about large claim sizes. To all-state, a claim size under 100$ is far less influential than a claim size of say 20,000 dollars. Therefore, we want to eliminate all claims under 100$ from our data set so they don’t influence our regressions. Our QQ plot illustrates these small claims outliers, highlighted in yellow.

Chart, line chart

Description automatically generated

We will now drop our claims under a hundred dollars, and re-evaluate the new QQ plot minus those small value claims. The code for that is as follows:

A picture containing chart

Description automatically generated

Chart, line chart

Description automatically generated

We see the small claims outliers are removed, and we are now ready to begin building our regressions.

**Part 1I: Running Our OLS Regression**

We will start our comparison by building an OLS regression model. To start, we must split our data into a training and test set. In this instance, we will only put 20% of our data into our training set. This is not ideal, but due to time constraints we must sacrifice more data in our training set for faster computing power.

Graphical user interface, text

Description automatically generated with medium confidence

We will also move our data from data frames into matrixes. This isn’t necessary for our OLS regression but will be when we begin building our LASSO regression.

Text

Description automatically generated

We will now run our basic linear (OLS) regression. The code is as follows:

Text

Description automatically generated with medium confidence A picture containing text

Description automatically generated

We will use these metrics to compare to our LASSO regression, to see which is more accuratley prdicting losses. We will also want our Root Mean Squared Error, as this will also be a helpful comapriosn metric when working to compare our two models. The code for RMSE is as follows:

Text

Description automatically generated

Our RMSE for the training and test sets of our OLS Regression are as follows:



We will use these figures to compare the efficacy of our OLS regression model to our upcoming LASSO model.

**Part 1II: Running Our LASSO Regression**

We will now build a LASSO regression model to see if we can improve on the efficacy of our OLS regression. We will start with a ten-fold Lasso Regression, with the code is as follows;

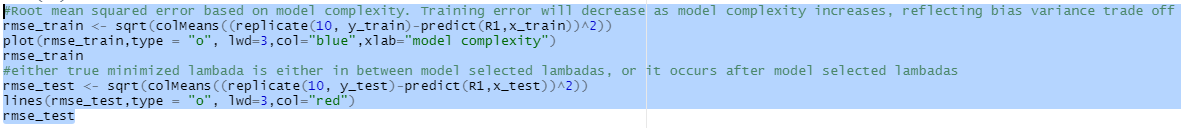


We now examine our LASSO regression model, and how the lambdas effect the selection and shrinkage of different variables:

 Chart

Description automatically generated

We will now try and discover our optimal lambda. The code for this is as follows:



Our outputs from the code above show that we have a continually decreasing RMSE. Ideally, we would want our output to show a low point of training and test RMSE (the bottom of the U) and then see the RMSE increase again. We would select our lambda from the model showing the lowest test RMSE. There could be two explanations for this not occurring; either the true best lambda is between our model selected lambdas, or it after model selected lambdas.

Chart, line chart

Description automatically generated

We will attempt to find our optimal lambda by cross validating our model.

A picture containing graphical user interface

Description automatically generated

Our optimal lambda for our cross validated model is .8068. We will then use this to run a third “optimal lasso” model to see if we can improve on our R-squared and RMSE from our OLS regression

Graphical user interface, text

Description automatically generated

Our Lasso Model shows a higher R-squared (.69) and drastically lower test set RSME(2121.9) then did our simple OLS regression model with all variables included (.54 multiple R-squared and test RMSE of 3584.2) Therefore, we should utilize our LASSO regression model when predicting claim sizes.