Audrey Kimball

Project 1 Report

The goal of the project was to explore and apply different linear regression models to predict the median value of owner-occupied homes (medv) in Boston, Massachusetts, based on a set of 13 predictors. The dataset comes from the MASS library in R, the dataset is loaded as Boston, the full name of the set is called Boston Housing Dataset. This data set because it is available publicly it used a lot to study the relationship between socio-economic factors and housing prices. The dataset contains 506 observations across 14 variables where the response variable is medv which represents median value of homes in thousands of dollars, and predictor variables include factors like crime rate (crim), proportion of residential land zoned for large lots (zn), average number of rooms per dwelling (rm) and other indicators. Understanding the relationship between the predictors is important for urban planning and real estate development.

The project will focus on three different linear modeling approaches to predict medv. The first method will be best subset selection. This method is used to identify the predictive subset of available predictors by evaluating all combinations. The other method is LASSO (least absolute shrinkage and selection operator) this is a regularization technique that not only performs variable but also shrinks the coefficients of less important predicators to zero. The final method is ridge regression, another regularization method, like LASSO but it constrains the size of the coefficients rather than setting them to zero, so all predictors are retained but with reduced impact.

The data set will be divided into training and validation sets, and each of the methos will be applied to the training set. The performance of each model is assessed using the Mean Squared Error (MSE) on the validation set. The hypothesis test will be conducted on the coefficients to evaluate the significance of each predictor. The analysis aims to compare the predication accuracy of the models and identify the most effective approach for predicting housing prices in Boston, considering the predictive power and model simplicity.

To split the data set I started by setting a seed, this helps ensure that the results can be reproduced by random number generator to a fixed state. I then had to determine the size of the training and validation sets. The training set used 80% of the data and the validation set used 20% of the data. We also needed random sampling, this needed to be done without replacement to make sure the training and validation sets do not overlap.

The best subset selection is used to identify the subset of the predictors that provides the best linear model for predicting the response variable medv. I used regsubsets() from the leaps package which helps search for all the possible subset of predictors. The function was applied to training data (train\_set) with the response variable medv and all other variables as predictors. The adjusted was used to identify the best model. It then said that 6 of the included predictors, chas, nox, rm, dis, ptratio, and lstat. The linear model was fitted with using the training dataset with the selected predictors. Model performance and significance of predictors were assessed using the output of the summary function which reported the coefficients, standard errors, t-values, and p-values. There is multiple 0.7155, adjusted 0.7112 and a residual standard error of 5.105. The hypothesis test p-values examined to test the null hypothesis of each coefficient is zero. Predictors chas, nox, rm, dis, ptratio, and lstat are statistically significant at the 0.05 level because the p-values were all < 0.05. The predictions were made using the validation dataset using the fitted model. The result was a mean square error of 18.42.

For the LASSO method we first needed to take the training data and extract the predictor variables excluding the medv and convert to a matrix and have a vector for medv. The goal is to select the optimal regularization parameter lambada that minimizes the prediction error. I used cv.glmnet to select the optimal lambda. The alpha variable specifies that we are fitting a LASSO model as opposed to ridge regression. The best lambda was decided through cross-validation and model coefficients using lambda.min. The optimal lambda value is 0.03180808. The most important non-zero coefficients are crim, nox, rm, dis, ptratio and lstat. For predicting I did the same method as the training set, the predictions I used the fitted LASSO model with the optimal lambda. The mean square error is 17.31. I then wanted to assess the variability of the LASSO coefficients and perform hypothesis testing. I started with a bootstrap procedure and applied 100 times using the training data set. I then took the coefficients and computed the standard error. I computed the p values using the t-statistics and the results were that coefficients crim, nox, rm, ptratio and lstat had significant p-values they were less than 0.05 and some of the other coefficients like dis has high p-values. The model’s optimal lambda value was 0.03180808, and the mean square error was 17.31 which is a reasonable prediction accuracy for the lasso model. The significant predictors were crim, nox, ptratio, and lstat all were found to be significant based on hypothesis testing the p-values were < 0.05. Bootstrapping revealed that some of the coefficients have high variability with significant coefficients being dependable for interpretation in the model.

For ridge regression the method is remarkably like the LASSO method in fact the first few steps are the same for turning the training data into a matrix and a vector. After repeating the steps of the LASSO method I knew the goal of ridge regression was to identify the optimal value of the regularization parameter lambda using cross-validation. I used cv.glmmet() to fit the ridge regression model which is why the argument alpha is 0 so it would not do LASSO. The function performs cross validation to determine the optimal value of lambda. The optimal value of lambda was identified by cross validation as 0.7014097. All the coefficients corresponding to the optimal lambda.min were extracted. The coefficients represent the relationship between each predictor and the target variable, with ridge regression shrinking the coefficients toward zero without them hitting zero. The validation set was used in the same way as the training set and the predictions were made using the fitted ridge model with optimal lambda. The mean square error was 17.14339 which means that the average squared difference between the predicted and actual values on the validation set. I then performed bootstrapping, the goal for this is to assess the variability of the ridge coefficients and perform hypothesis testing. The starting process was the same was the same but then used confidence intervals. I used 2.5th and the 97.5th percentiles bootstrapped coefficients. I looked at only coefficients that did not include zero and I thought those were important predictors these included, crim, nox, rm, dis, ptratio, lstat and tax. The other coefficients included zero so these might have less influence on medv.

In conclusion after completing the model’s best subset selection, LASSO, and ridge regression we can now decide on the model that will work best. The first model done was best subset selection which had 6 predictors chas, nox, rm, dis, ptratio and lstat, with a mean square error on the validation set as 18.42, the model explained 71.12% variance of the training data. The second model done was LASSO this had an optimal lambda of 0.03180808, there were a 11 predictors used crim, zn, indus, chas, nox, rm, dis, tax, ptratio, black, and lstat the reason that there were 11 was because LASSO shrunk some of the coefficients to 0, the mean square error was 17.31. The final model was ridge regression. The optimal lambda was 0.7014097, the model included all 14 predictors, but their coefficients were regularized to reduce their magnitude, the mean square error was 17.14. Overall the trade off between number of predictors and prediction can have an effect on the accuracy, for example best subset selection had the fewest predictors 6 but had the highest error, lasso used 11 had only a slightly worse error compared to best error rate, ridge regression had the best error rate and used all the predictors. If you want the most predication accurate ridge regression would be used because it had the lowest error. If you want a balance between predictors and accuracy it would be LASSO only used 11 predictors and only a slight worse accuracy than ridge regression.