Summary and Problem Statement

In this exercise, we are evaluating different machine learning modeling techniques in order to recommend one to a real estate firm. We are using data from Boston which were collected as part of a study examining whether air pollution impacted home values. We are fitting multiple regression models to the data and evaluating the results based on cross-validation testing; the root mean-squared error value (RMSE) is used as a comparative index of fit. The regression techniques are fitted trying both the response variable and then the log of the response variable to compare the results of transforming the response variable.

Methodology

The data are described as a "market response study of sorts", where the "market" is composed of 506 census tracts. Per the problem brief, the non-numeric field, neighborhood, was dropped from the data set. The data are complete, so no handling of missing data is required. The distribution of the response variable (mv) is skewed[1] (skewness = 1.1); log(mv) has a more normal distribution[2]. Both were tried when fitting models to see the effect of the transformation. Models for all 4 regression techniques: linear, ridge, lasso, and elastic net were fit to provide between-model comparisons. Comparisons were done based on RMSE; the smaller the RSME the better the fit.

Code Overview

The program has 3 major sections: exploration and visualization of the data, iteration over a set of methods to build and evaluate the models, and finally outputting the results of running cross-validation on the models. Models are built and evaluated as follows. A list variable is initialized, containing the regression methods to use and all required parameters for each

method. Next, the data are iteratively split into training and test "folds"; a total of 10 folds were specified. The program then iterates over the list of regression method once for each fold. Per-fold summary data are output for each model[3]. The per-fold summaries include: intercepts and coefficients for the fitted line, R^2 value, and RSME. Finally, a standardized average of the per-fold results for each method is output[4]. The modelling is run once with the unmodified target variable, then re-run using log(mv) as the target variable. Finally, the average mv and log(mv) results are repeated to make comparisons easier.

Results and Recommendations

The summary of average results[5] clearly shows the benefit of normalizing the response variable by taking its logarithm. The RSME values for the log(mv) models are less than half of those for the original (mv) variable. The results support using log(mv) as a good choice for transformation of the target variable, and I recommend using it in model building.

Linear, ridge and elastic net regression techniques had approximately equivalent results; the maximum difference in the averages for these 3 models is .001 when using the log(mv) target variable. The lasso technique performed the least well, with an RMSE nearly .02 higher than the others. I recommend using the elastic net regression method, with an alpha of 1.0, and an l1 ratio of .5. Its results are on par with linear and ridge regression models and it has the method also brings some regularization into the model. Because elastic net was able to have competitive results after driving some variables to 0, we know that there are superfluous variables in the model. Elastic net will give us the simplest and most easily interpretable final model.

[1] See mv-dist.pdf
[2] See mv-log-dist.png
[3] See console_output.html for examples of the per-fold output
[4] See console_output.html
[5] See summary_results.png