Math 0218 Final Project

By Kyra Gray and Will Ernst

For our final project, we explored regression approaches to predict the revenue of restaurant in Turkey, using a data set found on Kaggle. Our data comes from TFI, a company that invests in developing new restaurant sites. The data is split into a train and test set consisting of 137 and 100,000 observations (restaurants), respectively. Each observation includes data on the open date, location, city type, and three categories of obfuscated data: Demographic data, Real estate data, and Commercial data that result in P1-P37 numerical predictor variables.

In our analysis, we compare the results of five classes of regression: Linear Regression, Ridge Regression, Elastic Net Regularization, LASSO Regularization, and Principal Components Regression (PCR). We begin our analysis by selecting the variables we would like to use in our models. PCR has issues using categorical data (as it is difficult to determine 'distances' between categorical variables), so we drop any categorical variables in our PCR model formula. Next we find the optimal knob values for our methods. For Ridge Regression, Elastic Net Regularization, and LASSO Regularization, we use cross validation to compute the lambda value that minimizes the MSE. For PCR, we use cross validation to compute the optimal number of components used in the model. Although we could try different selections of variables (i.e. use a dimension reduction approach) for Linear Regression as a 'knob', this is captured in our regularization methods so that analysis would be redundant.

According to our results, the three regression regularization methods give us the lowest pseudo-scores (see *Simulation RSME Results without Lm*). After running our code multiple times, it appears that Ridge generally has the highest frequency of ranking first (see *Method Ranking Comparison*) and has the lowest average RMSE across the simulations (see *Average RMSE Scores across Simulations without Lm*). Although our cross validation efforts have determined Ridge to be the best predictor, the PCR method gave us the lowest Kaggle score. While this seems odd, we believe that having a small training dataset (and thus an even smaller pseudo-test and pseudo-train set) generates these odd results. We also hypothesize that the small training data set may not be a good representation of the overall population. Perhaps the training data is just the comprisal of all the restaurants in Turkey that publicly report their revenue. Given how close the average scores end up being (see *Average RMSE Scores across Simulations without Lm*) and the scoring on Kaggle, we are confident concluding that PCR may be a better method on the larger set of test data.

In our "Extra" section, we also attempted to manually conduct PCA and created fewer principle component variables based off the initial P1-P37 numerical predictors. Then we cross-validated across Lm, Ridge, and LASSO to determine which regression method worked best on this new data set comprised of the principal components. However, the Kaggle score generated by this approach was not better than the final method outlined above.