RM 294 – Optimization I

Project 3 – Non-Linear Programming

Group 22

Abhinav Sharma, Emily McCullough, Serena Wu, Luna(Yue) Cui

Objective

Variable selection is a crucial step for regression models as it can help find the most relevant variables, prevent overfitting, and also increase the accuracy of the overall models. Although the most commonly used method for variable selection is through the “shrinkage component” of Lasso regression, with the improvement on optimization field, now we can also derive variable selection through optimization method with less computational difficulties. In this report, our objective is to figure out which method works better on variable selection by direct comparison on the same dataset. We are also going to summarize their advantages and disadvantages so people can make decision choices based on the situation and what they need.

Direct Variable Selection – MIQP Problem

1. Transform variable selection to optimization question:

The objective function of variable selection is to minimize the squared difference between the model predicted value and the true value. Suppose we have m variables in the model, we will have 2m+1 decision variables, including m+1 values of beta and m value of z, which is a binary variable that can determine if the beta is equal to zero, to achieve the purpose of variable selection. The whole problem and constraints is formulated as below:

$$
\begin{gathered}
\min_{\beta,z} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} - y_i)^2 \\
s.t. -Mz_j \leq \beta_j \leq Mz_j \quad for\ j = 1, 2, 3, \ldots, m \\
\sum_{j=1}^{m} z_j \leq k \\
z_j \ \ are\ binary.
\end{gathered}
$$

By using the knowledge of linear algebra, we transformed our objective function into below format so it is easier for programming:

$$\min_{\beta,z} \beta^T (X^T X)\, \beta + (-2\, y^T X)\, \beta$$

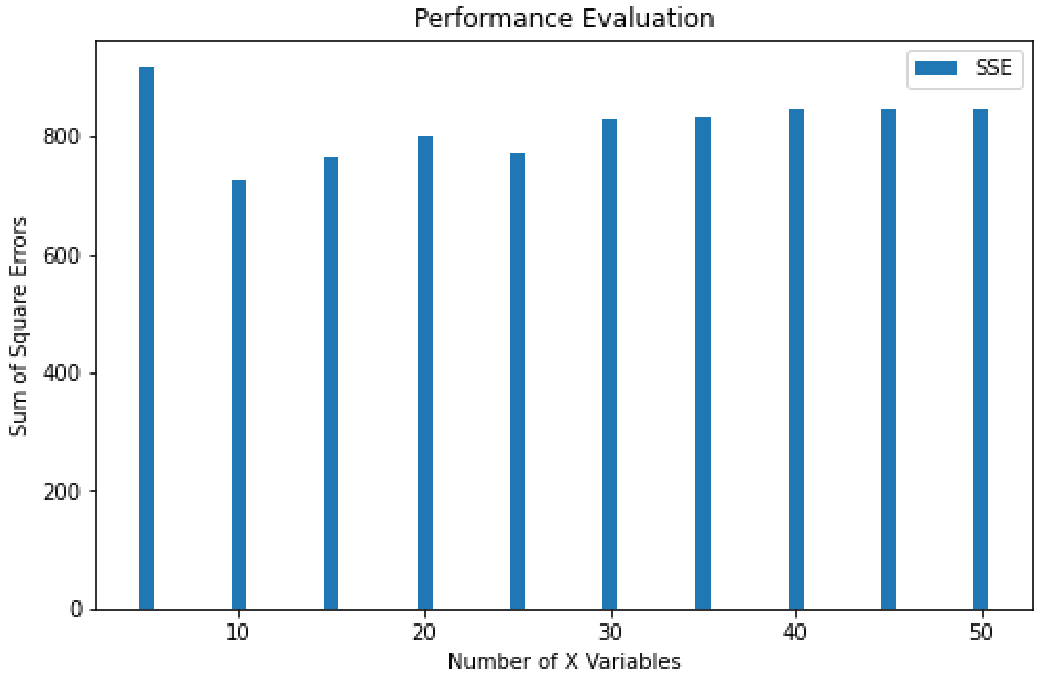
2. Method:

We first did 10-fold cross validation on a training data set with all possible values of k (5, 10, 15, 20, 25, 30, 35, 40, 45, 50) to determine the optimal value of k by comparing the cross validation error.

Then, we used the optimal k value to fit the MIQP model on the entire training data, and evaluated its performance by making predictions on y values in test data as well as calculating the MSE of the test data set.

3. Result:

The result of different k choice and corresponding cross validation error (SSE) is shown below:



According to this graph we determined that our optimal k value is:

k =10 (SSE~=724.78, the smallest cross validation error)

With optimal of total 10 variables (k=10), the MIQP model gave us the following variable selection result and value of beta, as shown below:

	x9	x15	x16	x23	x24	x26	x34	x45	x47	x48
Beta	-2.30820726	-0.51832612	-0.20416201	-1.55914318	0.86697336	-1.31191942	0.4081653	1.78147489	0.88738292	-0.28229213

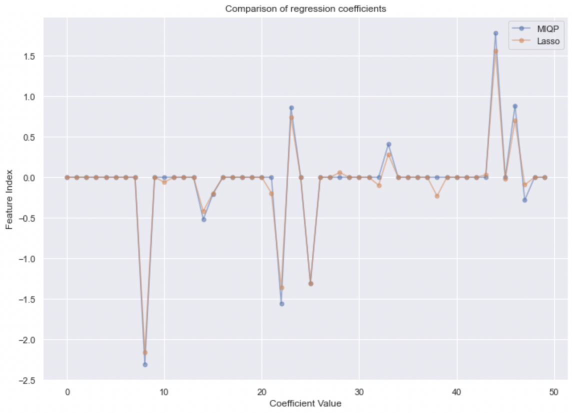
(intercept = beta 0 = 0.97252408)

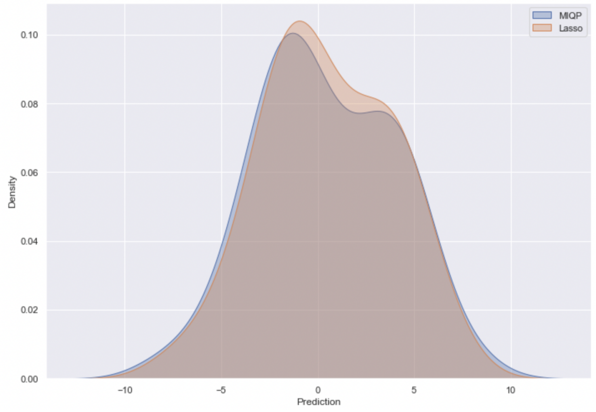
The MSE based on this model's prediction on y value of test data is: **2.33654**

Indirect Variable Selection – LASSO :

To compare our MIQP results with other indirect feature selection methods, we introduced the lasso regularization on a linear regression model trained on our training set. Lasso helps reduce the regression coefficients by introducing the absolute sum of coefficients penalty in the regression loss function. As this penalty term increases, it results in more and more variables with zero coefficients. The regularization parameter is used to account for the penalty term. This is how we can use Lasso for indirectly selecting our features. Leveraging cross-validation across 10 folds, the ideal regularization parameter is obtained as 0.076. The model selects 17 features compared to the ideal subset of 10 features recommended by the MIQP implementation earlier.

We compare the regression betas from both the techniques and find a huge similarity as shown in the graph below. The predictions for the dependent variable on the holdout sample are very similar from both the techniques as shown in the density plots below.





From above plots, it’s evident that the MIQP method is able to create a simpler model with a lesser number of features that essentially is able to predict the holdout data with similar predictive power. In comparison, the MSE for MIQP is marginally lower than the lasso regression model.

Recommendation

As computation resources become more accessible, we are in a better position to apply custom optimization solutions instead of simply relying upon readily available ML libraries. The existing modules and packages might have certain limitations due to the assumptions they employ for faster convergence in their internal optimizations. In our use-case our custom optimization is able to return a simpler model with a better holdout sample MSE. We also observe that our MIQP solution does not consider features from lasso which only had coefficient values closer to zero in the lasso regression, thereby formulating a much more robust model.