Applied Empirical Economics
Task 5

Gizem Kilicgedik

October 4, 2023

1. Exercise 1

In this task, we are instructed to utilize the cross-fit partial method to assess the impact of an unemployment insurance program on the duration of unemployment. You can see the outcomes in the lable located at the end of this section. The dependent variable in this analysis is the natural logarithm of the duration of unemployment. In the first column, we apply the cross-fit partial method with the plug-in estimator. In the second column, we compel the model to include all the original variables and again estimate it using the plug-in estimator. In the third column, we estimate lambda through cross-validation. Each time, we divide the dataset into 5 subsets (folds) and resample it 15 times. Overall, we observe that the results remain robust.

Furthermore, in exercise 1a, we are informed that the variables included in the model will vary across different folds and samples. We are asked to report the variables selected for a specific fold, which can be achieved using the "lassoknots" command. As an example, the variables selected in the 4th fold of the 10th sample are: q6, q2*q6, black*recall, recall*recall, agelt35*husd, agelt35*agelt35, and recall*nondurable.

Moving on to the exercise 1b, we need to determine how many variables with non-zero coefficients are included in the model. The post-estimation results reveal that the selected control variables are: agegt54, agelt35, black, dep, durable, female, hispanic, hispanic*lusd, hispanic*q6, husd, lusd, nondurable, q5, othrace, q1, q2, q3, q4, q6, q6*durable, recall, and v14.

Lastly, in 1c, we are asked to explain why the cross-validation (CV) method results in a higher number of selected control variables compared to the plug-in method. This difference arises because in the CV approach, we can minimize the objective function by either reducing the number of selected control variables for a given lambda value or by reducing the value of lambda itself. The CV method, which selects a lower lambda value than the plug-in method, therefore includes a larger number of control variables in the model.

	(1)	(2)	(3)
Treated	-0.0923***	-0.0864***	-0.0961***
	(0.0302)	(0.0300)	(0.0301)
Plug in or CV	Plug in	Plug in	CV
Force raw vars	No	Yes	No

2. Exercise 2

In this task, we are instructed to revisit Task 4, but this time utilizing random forest. Initially, we apply random forest for making predictions within the time period from 1992 to 2002. To this end, we implement 1000 repetitions and set the number of variables to be chosen as 8, which is consistent with the number of variables selected by the other three methods, namely subset selection, lasso, and ridge. The 8 most significant variables identified are: presidential, inflation, female employment, corruption, parliamentary, effectiveness, lexp, and regulation. When it comes to making predictions within the sample, using 1000 iterations, we find that the random forest method does not perform as well as the other three methods. However, it still outperforms naive predictions.

Nevertheless, when we apply the model recommended by the random forest algorithm to the out-of-sample dataset, we discover that this suggested model surpasses both the lasso and ridge models. This suggests that there may have been some overfitting of the data with those models.

Finally, we repeat the same analysis using the dataset covering the period from 2002 to 2011. In this iteration, we identify the 8 most important variables as: competitiveness legislation, voice, parliamentary, corruption, regulation, legal aspects, effectiveness, and age dependence among the youth. It's noteworthy that only half of the variables selected for the 1992-2002 dataset are retained in the 2002-2011 dataset.