ECON 626: Empirical Microeconomics

Introduction to LASSO

Department of Economics
University of Maryland
Fall 2019

The do file `econ626-2019-L6-A2-lasso.do` generates a data set containing 200 observations of an outcome Y and a set of covariates $A1$–$A9$, $B1$–$B9$, C–$C9$, and $D1$–$D20$. The A^*, B^*, and C^* all predict the outcome, Y (to varying degrees).

1. Run the do file. The last line of code estimates an OLS regression of Y on A^*, B^*, and C^* in half the sample. Extend the program so that it also estimates OLS in the other half of the sample. Record the R^2 and root mean squared error (RMSE) from these regressions in the table on the next page.

2. Next use a stepwise selection approach: estimate the full model in half the sample and then drop any covariates that are not statistically significant at the 95 percent confidence level. Re-estimate the model (in the same half of the data) and repeat. Iterate until all covariates have p-values below 0.05. After using stepwise selection in half the sample, run the restricted model (including only the covariates chosen through stepwise selection) in the other half of the sample. Record the R^2 and RMSE in the table.

3. Stata's lasso2 command to choose a set of covariates that best predict Y in half the sample, and then use this set of covariates to predict Y in the other half of the sample. First, run the lasso procedure by typing:

   ```
   lasso2 Y A* B* C*
   ```

   Which variables enter the model first, and why? What is the λ associated with the most predictive covariate? What is the λ associated with the last A^*, B^*, or C^* variable to enter the model?

4. Type

   ```
   lasso2, lic(ebic)
   ```

   to see the covariates selected using the extended Bayesian information criterion. Use (only) these variables to predict Y in the two halves of the data (in-sample and out-of-sample) and record the R^2 and RMSE in the table.

5. Now estimate the default LASSO model, but use Aikike's information criterion to choose the λ for covariate selection. Again, use (only) these covariates to predict Y in the two halves of the data (in-sample and out-of-sample) and record the R^2 and RMSE in the table.

6. Repeat the exercise using `lasso2`'s `sqrt` option and the EBIC criterion for choosing λ.
7. Repeat the exercise using Stata's `cvlasso` command (setting the number of folds to 10, which is the default).
8. Repeat the exercise using Stata's `rlasso` command to identify the theoretically-grounded λ.

	In-Sample		**Out-of-Sample**	
Method	R^2	**RMSE**	R^2	**RMSE**
OLS (kitchen sink)				
Stepwise selection				
LASSO (EBIC)				
LASSO (AIC)				
Square-root LASSO (EBIC)				
Cross-validated LASSO				
`rlasso`				