The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Sukjin Han

Assignment 4

Due: Nov. 15th 11:00am

- Report all the codes and the outputs in answering the questions.

1. Consider the datset "credit.csv" which contains information about individual's credit scores and other characteristics. Using this dataset, we want to understand which characteristics are important in predicting average credit card debt (*balance*). Specifically, we want to consider the following (nonlinear) regression model:

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{J} \gamma_{jk} x_{ij} x_{ik} + e_i$$

where y_i is the credit score and $\{x_{ij}\}_{j=1}^{J}$ are the continuous characteristics (*standardized*) and characteristic dummies.

   (a) What are the number of observations (n) and the number of predictors (p) in this regression? Make an argument why lasso may be the better procedure in this context, compared to OLS.

   (b) Conduct the estimation of the model using lasso. For the first regression, set $\lambda = 0.5$. Report the result.

   (c) Calculate the training MSE given the estimation results.

   (d) Calculate CV_n defined in our note, using a 5-fold CV.

   (e) Compare the answer in (c) and (d) and explain the discrepancy.

   (f) We want to choose an optimal λ using a 5-fold CV as in (d).

      i. Creat a grid for λ with 100 grid points. The range should include zero as a starting value. The range should be determined at your discretion. For example, try to run several lassos with, say, $\lambda = 10, 100, 1000$, and see how sentive the result changes. You want to pick a good range that yields different lasso estimates, i.e., different number of variables selected.

      ii. Given the grid, calculate CV_n for each λ and draw a picture with λ on the x-axis and CV_n on the y-axis. You may want to use a loop in your code to implement this. Determine the optimal choice of λ.

   (g) Conclude by reporting the final estimation results with the optimal λ.