PSYC 560
Due: February 23, 2023

Assignment #3: Shrinkage and Data Reduction Methods

[Q1] The data file **mobilephone2_training.csv** is part of large European survey data for mobile phone customers. It includes the following variables:

1. dura24: the duration of remaining a customer in the past 24 months
2. age
3. Ssound: satisfaction with sound quality (1 - 10)
4. Sglob: satisfaction with overall network functioning (1 - 10)
5. Svmail: satisfaction with voicemail (1 - 10)
6. Sinfor: satisfaction with information (1 - 10)
7. Sprice: satisfaction on price (1 - 10)
8. Spromo: satisfaction on promotion (1 - 10)
9. Sadver: satisfaction with the advertisement (1 - 10)
10. Sperson: satisfaction with the salesperson (1 - 10)
11. Simage: satisfaction on company image (1 - 10)
12. Sglobal: global satisfaction (1 – 10)
13. Lintent: intention of loyalty (1 - 10)

Suppose that you are interested in predicting the duration of remaining a customer (dura24) using 12 predictors (variables 2 – 12).

(1) Apply a series of ridge regression to the data, considering 100 candidate values of the tuning parameter λ within the interval [10^{-5}, 1]. Based on 10-fold cross validation, plot all MSE values against the candidate values of λ and choose the optimal λ value. Given the chosen λ value, re-run a ridge regression for the data and report its coefficient estimates. (2 points)

(2) Apply a series of the lasso to the data, considering 100 candidate values of the tunning parameter λ within the interval [10^{-5}, 1]. Based on 10-fold cross validation, plot all MSE values against the candidate values of λ and choose the optimal λ value. Given the chosen λ value, re-run the lasso for the data and report its coefficient estimates. (2 points)

(3) Apply a principal component regression to the data and choose the number of components that minimizes the cross-validated RMSE. Plot the RMSE values against the number of components. Also, describe how much variance of the predictors and the response the chosen number of components explain (2 points)

(4) Apply a partial least squares regression to the data and choose the number of components that minimizes the cross-validated RMSE. Plot the RMSE values against the number of components. Also, describe how much variance of the predictors and the response the chosen number of components explain (2 points)

[Q2] Using the estimated solutions obtained from the above four methods, report their MSE values in the test sample (**mobilephone2_test.csv)** and conclude which method appears to perform best in terms of the test MSE values (2 points).