Stat 6021: Homework Set 12

1. You will use the `College` data set from the `ISLR` package for this question. The data set comes from the 1995 issue of the US News and World Report, and contains information on 777 US Colleges on a number of variables. Please use the documentation in R to read the description of the variables. You will use ridge regression and lasso regression to improve upon a model that is predicting the number of applications a college receives, using the other 17 variables in this data set.

    (a) Before fitting any model(s), explain the circumstances that result in ridge regression and lasso regression to improve the accuracy of the model (compared to ordinary least squares, OLS).

    (b) Before fitting any model(s), discuss whether you think ridge regression or lasso regression will perform better in predicting the number of applications a college receives (in terms of model accuracy). Briefly explain.

    (c) Split your data into a training set and a test set with (roughly) equal numbers. Use `set.seed(2019)`.

    (d) Fit a ridge regression model on the training set, with λ chosen by cross-validation using the `cv.glmnet()` function. Before using this function, use `set.seed(4630)`. Report the test MSE based on this value of λ.

    (e) Fit a lasso model on the training set, with λ chosen by cross-validation using the `cv.glmnet()` function. Before using this function, use `set.seed(4630)`. Report the test MSE based on this value of λ.

    (f) Find the test MSE with OLS.

    (g) Comment on the test MSE with ridge regression, the lasso regression, and OLS. Which model has the best accuracy? Is this result surprising?

    (h) Create ridge plots to see how the values of the estimated coefficients vary with λ, for both ridge and lasso regression. Comment on how these plots explain why these methods are called "shrinkage methods".

2. For this second question, we will use the `swiss` data set that you have worked on before. You will perform principal component analysis (PCA) on the quantitative variables for this data set and answer the following questions:

(a) What are the loading vectors for the principal components (PCs)?

(b) How would you interpret the first and second PCs contextually?

(c) Use the `biplot()` function to create a plot of the first two PCs. Locate the province of La Vallee. How would you characterize this province, based on this plot?

(d) Produce a scree plot. How many PCs would you consider using? Briefly explain.