# STA 478 Fall 2025 Assignment #7

*Dr. Robert Buscaglia*

*October 24, 2025*

## Due Date: Tuesday, Nobember 9th, 2021 before 8:00 AM.

**Exam 2 Tentative Date: Weekend of November 12th, 2021 - Due Monday, 11/15/2021, before 5:00 PM**

## Exercises

### Exercise 1

Construct a function that will calculate the Variance Inflation Factors of all predictors. You may choose how the data is read in, but I recommend that the function take a matrix or data.frame of which each predictor is a unique column. The VIF function should return the VIF score for each predictor. Recall that VIF is calculated as:

$$VIF = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the Coefficient of Determination for regression of predictor $x_j$ against all other predictors.

### Exercise 2

**a.** Return to the `Boston` data analyzed in a previous homework. Using your VIF function, calculate the VIF for each predictor in the `Boston` set (all variables but `crim`). Display a table of the VIF for each predictor.

**b.** Iteratively remove the predictor with the highest VIF until no predictors have a VIF greater than 5. Call this predictor set `VIF.5`.

**c.** Prepare a predictor set for VIF less than 3 (`VIF.3`) and 2 (`VIF.2`)

**d.** Run 10-repeats of 10-fold cross-validation on the three different VIF predictor sets. Estimate logistic regression models for above/below median crime rate. Report the mean test set accuracy/error within a table. How well did your model perform compared to your results from Assignment 5?

### Exercise 3

Review the `glmnet` vignette found at https://glmnet.stanford.edu/articles/glmnet.html. This was recently updated with the current form of the `glmnet` package and is full of useful information! Review through `Guassian` linear regression at a minimum, most should review through `Logistic Regression`. Reviewing the `glmnet` vignette is required to help you run the analysis below.

## Exercise 4

We will predict the number of applications received using all other variables in the `College` data set.

**a.** Load the data and give a brief evaluation of the predictors (we should never blindly apply models without an understanding of the data). Remove `Accept`, `Enroll`, `Top10perc`, and `Top25perc` from the data frame.

**b.** Produce a two-thirds/one-third split of the `College` data for training and testing.

**c.** Estimate a backward hybrid stewpise linear model using the training set. Report the test set root MSE (RMSE).

**d.** The `glmnet` functions will require you to produce design matrices. Produce the design matrix for the training and test set. You may use `model.matrix()` from base R, or the built in `makeX()` function from the glmnet package. It does not use the `y ~ x` form we commonly use.

**e.** Estimate a regularized model using ridge regression. Use the design matrix you produce in **d.** and let `y` be the response vector `Apps`. Validate the penalty coefficient, $\lambda$, using `cv.glmnet()`. Plot the `cv.glmnet` object to visualize the cross-validation result.

Report the $\lambda$ and test set RMSE for the one-third left out set using the penalty coefficient $\lambda$ taken from `lambda.min`. Export the coefficients for the model at `lambda.min`.

**f.** Repeat **e.** using LASSO regression.

**g.** Repeat **e.** using the Elastic-Net. Set $\alpha = 0.5$.

**h.** Make a table comparing the coefficients from each model type. What are some differences between the estimated models?

**i.** Compare the single iteration test set MSE values? Is there much of a difference between the methods?

**j.** Discuss what is happening to the regularized models when we let $\lambda = 0$ and what changes when we increase the size of the penalty coefficient $\lambda$. What is the limiting behavior as $\lambda \to \infty$?

*Note: this is just a single split, so do not give the RMSE values too much credit. We would want to repeat this many times to get a better estimate. Focus more on the models being output.*