**Homework Week 5**

**PSC 204B**

For this homework, we will be using a simulated dataset to illustrate issues related to under- and over-fitting. You can use the following code to simulate the true model, which is a cubic model relating x and y.

set.seed(02082018)

x = rnorm(200)

y = x + x^2 + x^3 + rnorm(200, 0, 4)

dat = data.frame(y, x)

You will use this generated data to answer the following questions.

1. Create three models with polynomial parameters of different degrees: an under-fit model, a true model and an over-fit model.
2. Plot the predicted curves over the sample data points in three separate plots (one for each model).

1. Calculate mean squared error for each model. Your over-fit model should produce the lowest MSE.
2. Now, using ridge regression in R, determine the best value for lambda (hint: you will need to install the glmnet package in R). Report best lambda and corresponding regression coefficients.
3. Repeat the above steps, using lasso regression in R.
4. Compare your results found in steps 4 and 5. Which method do you think provides a better predictive model? Explain your reasoning.

BONUS 1: Prepare the data to be used for *k* = 5 folds cross-validation **without** the assistance of an R package specifically designed to do so.

BONUS 2: Using k-fold cross validation, calculate both Train MSE and Test MSE for 10 separate models with the degree of polynomial term increasing with each subsequent model. In a single figure, plot MSE (y-axis) vs model degree (x-axis) for both the training and testing sets.