DSC 241 - Homework 6

Problem 1. Implement k-fold cross-validation and sequential model selection for linear regression models.

a. Write a function **cv.lm(x, y, k)** which estimates the prediction error of the linear regression model with **y** as response using k-fold cross-validation

b. Write a function **SequentialSelection(x, y, method)** which computes the forward selection path for linear regression from 'intercept only' to 'full model' and chooses the model on that path using different criteria specified by **method**. The function should support these methods:

- **method** = "AdjR2": Sequentially include the columns of **x** and choose the model that gives the largest adjusted R^2.
- **method** = "AIC": Sequentially include the columns of **x** and choose the model that gives the smallest AIC.
- **method** = "CV5": Sequentially include the columns of **x** and choose the model that gives the smallest 5-fold cross-validation prediction error.

Problem 2. Consider a regression setting where the predictor variable is real valued and the goal is to fit a polynomial model. Specifically, we assume that $x_1, ..., x_n$ are iid uniform in $[0, 2\pi]$ and conditional on these, $y_1, ..., y_n$ are independent, with y_i normal with mean $\sin(3x_i) + x_i$ and variance 1. Take $n = 200$ and set the maximum degree at 20. Perform simulations (at least 100 data instances) to compare the choice of degree by the sequential model selection methods in Problem 1. Produce plots of 3 example data instances and their best model fits according to different methods. Produce plots of the distribution of the polynomial degrees chosen by the different methods over all simulated instances. Offer comments on what you observe.