# MSAN 601: Linear Regression Analysis
# Homework 1

## Due: Thursday, September 7th by 9:00 AM

## Computational Problems

For each of these problems, use `R`

1. Write a function *kfold.cv.lm()* which performs the following. This function is to be made from scratch, i.e. you cannot simply use a function for cross validation in any package.

   **Input Arguments**:

   - *k*: integer number of disjoint sets
   - *seed*: numeric value to set random number generator seed for reproducability
   - *X*: $n \times p$ design matrix
   - *y*: $n \times 1$ numeric response
   - *which.betas* : $p \times 1$ logical specifying which predictors to be included in a regression

   **Output**: *Avg.MSPE*, *Avg.MSE*

   **Description**: Function performs k-fold cross-validation on the linear regression model of $y$ on $X$ for predictors *which.betas*. Returns both the average MSE of the training data and the average MSPE of the test data.

2. Download the *College* data set from the following link:

   `http://www-bcf.usc.edu/~gareth/ISL/College.csv`

   This data describes several interesting summary characteristics of American colleges and universities in 2013, including the University of San Francisco!

   Suppose that we are curious about what factors at a university play an important role in the room and board each semester (column *Room.Board*). Answer the following questions.

   (a) Based on some research into the area, you believe that the five most important predictors for the room and board amount are

   - the number of students who accepted admission *Accept*
   - the number of students who are currently enrolled *Enroll*
   - the out of state tuition for a semester *Outstate*
   - the average cost of books per year *Books*
   - the graduation rate of the students *Grad.Rate*

Plot a pairwise scatterplot of these variables along with the room and board cost. Also, summarize each of these variables in terms of correlation, expectation, and variance and comment on any trends.

(b) Use your *kfold.cv.lm()* function from the first question to run 10 - fold cross-validation on each of the $2^5 = 32$ possible regression models of *Room.Board* on the every subset of the above 5 predictors. For each model, run cross validation 100 times to get a distribution of the average MSPE. Which model would you choose? What are the estimates and standard errors of your parameter estimates? Plot a histogram of the average MSE and MSPE from the 100 runs of 10-fold cross validation for your chosen model.

# Conceptual Problems

1. Recall that the variance of a random variable $X$ is defined by $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$. Establish the following.

   (a) $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$

   (b) If $a, b$ are constants, then $\text{Var}(aX + b) = a^2 \text{Var}(X)$

   (c) $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

2. Consider the traditional multiple linear regression model of response $Y = (y_1, \ldots, y_n)^T$ on data $X \in \mathbb{R}^{n \times p}$:

$$Y = X\beta + \epsilon$$

   a) What possible assumptions can you make on the error terms $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$?

   b) The least squares estimates of $\hat{\beta}$ are given by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

      i) Do the least squares estimates above depend on any of the assumptions we can make about $\epsilon$?

      ii) Using the least squares estimates and an appropriate assumption on $\epsilon$, show that

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$$

      It follows that making an appropriate assumption on $\epsilon$ allows us to make inference about our model. We will revisit this with other regression models in the future.

3. In this problem we will prove the relationship between bias, variance, and MSPE in a regression problem. The proof just relies on properties of expectation. Let $Y$ and $Z$ be two independent random variables with means $\mu_Y$ and $\mu_Z$, respectively. Answer the following questions.

(a) By expansion, show that

$$\mathbb{E}[(Z-\mu_Y)^2] = \mathrm{Var}(Z) + (\mathbb{E}[Y-Z])^2$$

*Hint*: think carefully about what is needed to obtain $\mathrm{Var}(Z)$ on the right hand side of the equality above.

(b) Use part (a) and expansion to show that

$$\mathbb{E}[(Y-Z)^2] = \mathrm{Var}(Y) + \mathrm{Var}(Z) + (\mathbb{E}[Y-Z])^2$$

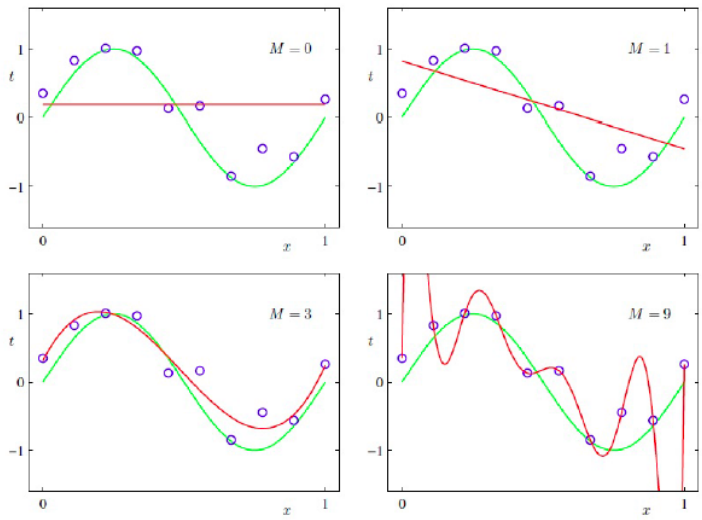
(c) Consider setting $Y = f(X) + \epsilon$, and $Z = \hat{f}(X)$. Now use parts (a) and (b) to conclude that

$$\mathbb{E}[\mathrm{MSPE}(\hat{f}(X))] = \mathrm{Var}(\hat{f}(X)) + \mathrm{Var}(\epsilon) + \mathrm{Bias}(\hat{f}(X))^2$$

4. Suppose that we observe data $(x, y)$ and that we fit polynomials of increasing degree $M$ to the data. Namely, we fit models like

$$f(X) = \sum_{j=0}^{M} a_j x^j$$

In the plots below, we show the data (blue points), the fitted model (red line) and the true model (green line). Let $\hat{f}_M$ be the model fitted in each




plot.

(a) Suppose that all of the data observed in the plots are used as training data. Suppose that we observed a new data point $(X_o, y_o)$ *outside* the original data set. Rank the models in terms of $\mathrm{MSPE}(\hat{f}_M)$. Explain your answer.

(b) Suppose that all data observed in the plots are used as training data. Rank the models in terms of $\mathrm{MSE}(\hat{f}_M)$. Explain your answer.

(c) Rank the models in terms of $\mathrm{Var}(\hat{f}_M)$. Explain your answer.