MSAN 601: Linear Regression Analysis
Quiz 2

by James D. Wilson (University of San Francisco)

There are 5 questions. Each question is weighted equally. Grades will be out of 100 points. You have until 9:45. Good luck!

1. Suppose that a linear regression model perfectly fits the training data so that the MSE is 0. Which of the following statements is true? (thought I would give a nice multiple choice question to start things off)
   (a) You will always have a test error of zero
   (b) You can not have test error zero
   (c) None of the above

2. Let X be an $n \times p$ data matrix as described in class, and y an n-dimensional continuous vector. You decide to fit a linear regression model of y on X:

$$y = X\beta + \epsilon$$

   According to what we've talked about in class, what two possible assumptions can you make about ϵ?

3. Suppose that you observe the data $(x_1, y_1), \ldots, (x_n, y_n)$ and you fit the model

$$y_i = f(x_i) + \epsilon_i.$$

   Through estimation, you find an estimator $\widehat{f}$. Write the equation for the mean squared error of $\widehat{f}$.

4. Suppose you take the data observed in Question (2) and randomly split the data into training data $(\mathbf{x}_{train}, \mathbf{y}_{train})$ and test set data $(\mathbf{x}_{test}, \mathbf{y}_{test})$. From the training set, you fit the model $\widehat{g}$ according to

$$y_i = g(x_i) + \epsilon_i.$$

   (a) Write the equation for the MSPE of $\widehat{g}$.
   (b) What can you say about the relationship between the MSE and MSPE of $\widehat{g}$?
   (c) Suppose that you observe a new data point (x_0, y_0). Write the equation for the $\mathbb{E}[MSPE(\widehat{g}(x_0))]$ in terms of the bias and variance of $\widehat{g}$ and the error terms ϵ_i.

5. Suppose that we fit polynomial models of data $\mathbf{y}$ against $\mathbf{x}$ of order $p = 1, 2, \ldots, 99$ to data where we observe 150 data points. In particular, we fit functions of the following type on a randomly selected training set of size 100:

$$f(\mathbf{x}) = \sum_{j=0}^{p} a_j x^j$$

   In terms of the bias-variance tradeoff, explain what happens to the prediction accuracy of the fitted model on the test set as a function of p.