# COM S 573: Home work 4 Spring 2014

**Write your name on each page**. Maximum score is 25 points, due date is **Friday, April 4, 2014** . Please hand in the solutions (CLEAN version) on the due date in class (**hard copy**). Also paste the results of your R code and the code itself into your homework. Make sure your homework is stapled!

1. [8 points] A cubic regression spline with one knot at $\xi$ can be obtained using a basis of the form $1, x, x^2, x^3, (x-\xi)^3_+$ where $(x-\xi)^3_+ = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise. Show that a function of the form
$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+$$
is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

2. It was mentioned that GAMs are generally fit using a backfitting approach. The idea behind backfitting is actually quite simple. We will now explore backfitting in the context of multiple linear regression. Suppose that we would like to perform multiple linear regression, but we do not have software to do so. Instead, we only have software to perform simple linear regression. Therefore, we take the following iterative approach: we repeatedly hold all but one coefficient estimate fixed at its current value, and update only that coefficient estimate using a simple linear regression. The process is continued until convergence–that is, until the coefficient estimates stop changing. The process flow is sketched next.

   1. Download the *adv.dat* data set ($n = 200$) with response $Y$ and 2 predictors $X_1$ and $X_2$ on BlackBoard
   2. Initialize $\hat{\beta}_1$ (estimated coefficient of $X_1$) to take on a value of your choice, say 0.
   3. Keeping $\hat{\beta}_1$ **fixed**, fit the model
   $$Y - \hat{\beta}_1 X_1 = \beta_0 + \beta_2 X_2 + e$$
   4. Keeping $\hat{\beta}_2$ **fixed**, fit the model
   $$Y - \hat{\beta}_2 X_2 = \beta_0 + \beta_1 X_1 + e.$$

   (a) [6 points] Write a for loop to repeat (3) and (4) 1,000 times. Report the estimates of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ at each iteration of the for loop. Create a plot in which each of these values is displayed, with $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ each shown in a different color.

   (b) [2 points] Compare your answer in (a) to the results of simply performing multiple linear regression to predict $Y$ using $X_1$ and $X_2$. Use the `abline()` function to overlay those multiple linear regression coefficient estimates on the plot obtained in (a).

(c) [1 point] On this data set, how many backfitting iterations were required in order to obtain a "good" approximation to the multiple regression coefficient estimates? What would be a good stopping criterion?

3. [5 points] Show that the Nadaraya-Watson estimator is equal to **local constant** fitting. *Hint*: Use the local polynomial cost function to start and adapt where necessary.

4. [3 points] Show that the kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

with kernel $K$ and bandwidth $h > 0$, is a bonafide density. Did you need any condition(s) on $K$? If so, which one(s).