Regression III: Homework 3

The goal of this homework is to get you to think about how some of the flexible models we have been talking about might help us overcome non-linearity. As the assignments are mainly meant as a tool to allow you to better assimilate the information presented in the class, you may present them in whatever method is most useful to you. I would ask, however if you're writing math to please use math characters (preferably in an equation environment). From my point of view, it is always easier to read proper tables than statistical output, so that is certainly appreciated, but not an absolute requirement.

There are two sections to this homework assignment. The data for the first section are exactly the same as the previous homework. For the second section, you will get to play around with some generated data.

PART I.

1. Using the data from the second homework, estimate the GAM of logged aid using population, GDP per capita, and polity. Try simple smoother splines for each of the variables (remember from last time there was nonlinearity). Does the model fit better as a GAM than a linear model?

2. Estimate another GAM of logged aid using population, GDP per capita, and polity. This time, use tensor splines for each of the variables. If you choose to specify a value for "k" be sure to explain *why* you choose that value. Relative to the first GAM, what has changed, and which type of spline is better?

PART 2.

3. In this problem you'll be playing with some generated data with non-linearities, interactions, and a number of irrelevant variables. The data is deterministic (i.e. generated without an error term), and meant to allow you to explore the advantages and limitations of the techniques introduced over the past few lectures. Do the following:

   (a) Load in the generated dataset (y is the dependent variable). Run a simple linear model, and assess both model fit and non-linearity. Do not report these results; just briefly describe what you find.

   (b) Fit MARS to the data. Comment on which variables were selected/important, how much model fit has improved, and the extent to which the data generation process is found to be non-linear and interactive. Plot the surface of any interaction you find. Repeat this process using the method of your choice (CART, RF, BART, GBM, XGBM). Discuss how the results differ and why.

   (c) Finally, fit an appropriate GAM to the data using what you have learned from the previous step (which variables seem to matter and how). Plot the interaction surface. *Optional*: Find the parametric functional form used to generate the data.