Homework 9

36-401, Fall 2015 , Section B

Due at the start of class on **Tuesday, 8 December 2015**

In this assignment, we will revisit the bicycle-rental data from DAP 2. You mayy ignore the "feeling temperature" variable, along with the number of rentals by registered users. The total number of rentals is still the response variable.

1. *Stepwise regression*

   (a) (10) Fit a model predicting daily rentals, using linear terms for all available variables, plus product interactions between year and temperature and between weather and humidity, plus second-order polynomial terms for temperature, humidity and windspeed, plus the interaction of year with squared temperature and of weather with squared humidity. Report the summary output.

   (b) (10) Fit an intercept-only model, then use forward stepwise regression to add variables. (Make sure that all the variables and interactions from part (a) are included in the scope.) Report the summary output. Can you trust the p-values?

   (c) Use backward selection on your model from part (a) to choose two models:

      i. (10) one model with no constraints forcing any terms to stay in the model; and

      ii. (10) another model constrained to include temperature, windspeed, and humidity.

      Report the summary outputs. Can you trust the p-values?

   (d) (10) Use forward-backward selection on your model from part (a) to choose a model. Report its summary output.

   (e) (10) Compare and contrast the five models. Are any of them the same? Are there any terms which are always included? Which model would you choose? Why?

2. *Regression trees* We will now fit a regression tree model to the same data. There is no need to explicitly include nonlinearities or interactions, since the tree will find those automatically.

(a) (10) Build and plot a regression tree predicting daily bike rentals from all available variables. How many leaves does the tree have? Into how many different groups of days does the tree divide the data? Which variables appear in the tree? Which variables are important? What are the predicted number of rentals for the Thanksgiving of 2011 and the Independence Day of 2012? What is the in-sample MSE for predicting all days in the two years?

(b) (10) Now re-code the months so that January and February share one code, May through October shares another, and March, April, November and December share a third. Re-estimate the regression tree and plot it again. How does your tree change (if at all)? What is the MSE? Did we improve the fit?

3. (10) Looking at your results from problems 1 and 2, compare the various sets of "important" variables you selected. Which variables were most useful in prediction for linear models? Which were most useful in prediction for trees? Does anything surprise you?

Rubric (10 pts): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, please only submit the knitted file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.