STA 290 — Final Project — due Dec. 12 at 5PM

How well can you predict groundlevel ozone in Los Angeles? To address this your goal is to develop a model using the ozone data of Breiman and Friedman (1985). The data are available on the course web page, at
`http://www.stat.duke.edu/courses/Fall02/sta290/datasets/ozone.dat` and contain:

1. Response: ozone ground level ozone
2. Meteorological Variables:
   (a) VH, the altitude at which the pressure is 500 millibars
   (b) WIND, the wind speed (mph)
   (c) HUM, percent humidity
   (d) TEMP, the temperature (degrees F)
   (e) IBH, the temperature inversion base height (feet)
   (f) DPG, the pressure gradient (mm Hg)
   (g) IBT, the inversion base temperature (degrees F)
   (h) VIS, the visibility (miles)
3. DAY, the day of the year.

Read in the data. Construct a variable and two dataframes: ozone.test and ozone.val

```
leaveout <- round(seq(1,300, length=30))
ozone.train <- ozone[-leaveout,]
ozone.val <- ozone[leaveout,]
```

to use for model training and model validation, respectively.

Conduct exploratory data analyses, and then model the relationship between ozone and the other variables and/or transformations of them using ONLY the TRAINING DATA (ozone.train). Is there a DAY of the week effect? (Construct a new variable day of the week from DAY and conduct an appropriate test. Find an appropriate model using the training data only. Once you are satisfied with your model proceed to the next step.

Use the data in ozone.val to validate your model. Construct predictions using your model and compute the MSE for prediction, $\sum_{i=1}^{30}(Y_i - \hat{Y}_i)^2/30$ where y_i are the 30 ozone values in the validation set, and $\hat{Y}_i$ are the predictions from your model developed with the training data, ozone.train. Compute the same using observed and fitted values for the training data. How do they compare? Construct 95% prediction intervals for the for the training and validation data. For both compute the percent of intervals that contain the observed data. How do they compare? DO NOT REVISE YOUR MODEL AFTER CONDUCTING THIS STEP - THAT IS CHEATING!

Ozone is known to cause respiratory problems, and vulnerable individuals are advised to limit there exposure during high ozone episodes. Describe how your model can be used for forecasting ozone on a given day (assume you have forecasts of the explanatory variables or values from the previous day) for individuals that are interested in deciding whether to venture outside. What are the most important factors for predicting high ozone levels? Use your validation results to describe how accurate your predictions are. Limit your write up to 5 pages with text and figures.