36-350: Data Mining

Homework 10
Date: November 9, 2001 **Due: start of class November 16, 2001**

---

1. In this problem and the next, you will use regression techniques to model a time series. The predictor is time. The file `uspop.dat` gives the population of the United States (in millions) as recorded by the census every ten years from 1790–1970.

   (a) Plot the population versus time. What anomalous years do you see?

   (b) Break the series into pre-1860 and post-1860. If the data frame is x, this can be accomplished by

   ```
   x1 <- x[(x$time <= 1860),]
   x2 <- x[(x$time > 1860),]
   ```

   Make a transformation and use linear regression to fit a model for the pre-1860 population as a function of time. Give a plot or two to argue that a linear model fits well under the transformation. State the model as a formula in the original variables.

   (c) Plot the residuals for your pre-1860 model. Which two years pre-1860 are most unlike the others (have the largest residuals)?

2. (a) Make a transformation and use linear regression to fit a model for the post-1860 population as a function of time. Give a plot or two to argue that a linear model fits well under the transformation. State the model as a formula in the original variables.

   (b) Plot the residuals for your post-1860 model. Two years are outliers. Which are they?

   (c) Remove the outlier years. You can remove a year as follows:

   ```
   x2 <- x2[(x2$time != year),]
   ```

   Refit the model and plot residuals. One year is unusually high. Which one? (The page `http://www.missouri.edu/~socbrent/immigr.htm` may be of interest.)

   (d) What does your refitted model predict for the population in year 2000?

3. The file `States.dat` contains descriptive statistics about the 50 states. We would like to know what influences the life expectancy (`Life.Exp`) of people in a state. The other variables are total population, per capita income, illiteracy rate, homocide rate, percent high-school graduates, number of cold days per year, and land area.

   (a) Make a predict.plot. Which variables seem relevant to life expectancy?

   (b) Fit a linear model and plot the partial residuals. Which variables seem relevant now? How do you explain the difference with part (a)?

(c) Use `step` to reduce the model. Do the remaining variables match those you identified in part (a) or those in part (b)?

(d) Many of the remaining predictors have smaller p-values in the reduced model than they had in the full model. Explain how this can happen.

(e) Give a scenario in which a predictor's p-value could increase substantially when the model is reduced, i.e. after running `step`.

(f) Use `step.up` to test for interactions. Make a profile plot and describe the interactions found. For example: "predictor A has a large effect on the response only when predictor B is large."