36-350: Data Mining

Lab 14
Date: December 5, 2003 **Due: end of lab**

This is a lab exam. Do your work individually and email your code, plots, and answers to `fangc@stat.cmu.edu` by the end of the lab hour. Some relevant labs to this one are labs 5, 7, and 9.

```
source("lab14.r")
```

This defines a data frame `x` which contains the population of the United States (in millions) as recorded by the census every ten years from 1790–1970.

1. Plot the population versus time, with trend curve overlaid. What anomalous years do you see?
2. (a) Define a matrix `x1` which contains only the data up to and including 1860.
   (b) Transform the series appropriately so that the pre-1860 population can be predicted by a linear function of time. Fit a linear model and give a plot or two to argue that the model fits well. State the model as a formula for (untransformed) `uspop`.
   (c) Plot the residuals for your pre-1860 model. Which two years pre-1860 are most unlike the others (have the largest residuals)?
3. (a) Define a matrix `x2` which contains only the data after 1860.
   (b) Transform the series appropriately so that the post-1860 population can be predicted by a linear function of time. Fit a linear model and give a plot or two to argue that the model fits well. State the model as a formula for (untransformed) `uspop`.
   (c) Plot the residuals for your post-1860 model. Two years are outliers. Which are they?
   (d) Remove the outlier years. You can remove a year as follows:

   ```
   x2 = x2[(x2[,"time"] != year),]
   ```

   Refit the model and plot residuals. One year should be unusually high. Which one?
   (e) What does your refitted model predict for the population in year 2000?