Crime Data: Introductory and Overview Comments

1. Data and data source

The crime data set has more than 144 variables with candidate dependent variables in the last columns. The dependent variable for analysis in this assignment is ViolentCrimesPerPop.

The data come from the UC Irvine Machine learning Repository http://archive.ics.uci.edu/ml/ . The web site cites two papers that have analyzed the data. The web sites suggests that published papers cite the repository.

2. Transformations

Many exploratory analyses use transformations to make the data for different cases comparable. Some examples shown in class include:

> The study of population change: In the study of Louisiana parishes populations before and after hurricanes Katrina and Rita the population data was expressed as yearly **percent population change.**
>
> The study of mortality rates: In the study of lung cancer mortality, more deaths are expected in regions that have more people and more people in age intervals associated with higher risk. For a small age range intervals the death counts are re-expressed as age specific rates (deaths per 10000 persons in the age interval population). The age intervals are often (0, 1-4, 5-9, 10-14, …). When a unified view is of interest over large age range, deaths are often re-expressed as age-adjusted rates based on a specified standard US age distribution for the large age interval. This is typical based on a decennial census. A common convention expresses rates as death per 100000.

The different community sizes in the crime study makes the direct comparison crime counts of little value, unless the focus is restrict to set of communities with about the same population size. To analyze all of communities of the together, the number of crimes is re-expressed as a rate, the number of crimes per person in the 100000.

In a linear regression context, the boxcox transformation of the dependent variable is helpful in terms of model fitting. This can be helpful in a random forest model as well. A drawback is that the units be analyzed seem strange to many audiences.

Explanatory variables can also be transformed.

3. Units and communication

The use of deaths per 100000 in mortality studies seem reasonable because deaths from many causes are infrequent and rates per 100000 are in the familiar interval from from 0 to 100. Violent crimes are more common. I think the number violent crimes per 1000 would lead to smaller numbers and would help people related to the rates more directly. The rate for one large community in the study was about 4800 per 100000 or 4.8 per thousand. If the rate were 1 per thousand and the GMU student population is around 32000, the thought of expecting in 32 violent crimes in our community during a year is uncomfortable. I don't tend to identify will populations as large as 100000. 4800 per 1000 is just a number. (This is not to say that the rate is anywhere near this 1 per 1000 at GMU.)

4. Case and variable selection.

The web site says: "The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault."

"There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in missing values for per capita violent crime. Many of these omitted communities were from the Midwestern USA (Minnesota, Illinois, and Michigan have many of these). "

The choice here is to omit cases with missing values for dependent, variable crime rate. The rape counts and hence the violent crimes counts could be imputed, but that is not done in this assignment.

The web site briefly mentions omitting variables for the analysis of the crime rate. It says "It would not be interesting or appropriate to predict total crime (e.g. violent crime) while including subtotals (e.g. murders) as independent variables. There are also identifying variables (community name, county code, community code) that are not predictive." We also omit these variables. We will also omit cases that have missing values for explanatory variables that are chosen for the initial model.

Both linear models and random forests can be used for data imputation. This will not be done. In general different cases may have missing values for quite different regions. This complicates the choice of the modeling methods and interpretation of results. Of course just working with cases that have complete data can lead to biased results. So can working with cases that include those with imputed values.

As far as I know, the R version of random forests does not directly support analysis when cases have missing values for variables being used in the model. There tree models that will use surrogate values or other means to deal with missing data.

I seek parsimonious models that I have some hope of understanding. In the assignment variables are removed until the model begins to degrade more than a couple of percent points in terms of mean square error. The linear regression example illustrates the use of backward step wise regression. There are other good options. Random forests provide a measure of variable importance. I select subsets of variables to keep based on higher variable importance and, to a lesser extent, my knowledge about variables in terms of quality and availability and my notions of how they related to crime.

5. Graphics

Graphics help by calling attention to outliers, skewed distributions, shapes of bivariate distributions, variable correlations, and possible functional relationships. Some of the graphics can be improved. More graphics can be added. For example the rates can be shown for communities on a map.