

## Assignment 4

**Due: in class, on or before Wednesday, May 6**

(Data and description from: Kutner, M. H., C. J. Nachtsheim, and J. Neter. 2004. *Applied linear regression models, 4th edition*. McGraw-Hill/Irwin)

The data set accompanying this assignment (link, class website) provides selected county demographic information for 440 of the most populous counties in the United States (some counties were deleted because of missing data). Each case (county) has an identification number, a county name, and state abbreviation, along with values for 14 other variables. The information pertains to the years 1990-1992. The 17 variables are:

- 1 Identification number
- 2 County name
- 3 State
- 4 Land area (square miles)
- 5 Total population in 1990
- 6 Percent of population aged 18-34
- 7 Percent of population aged 65 or older
- 8 Number of active physicians in 1990
- 9 Number of hospital beds in 1990
- 10 Total serious crimes in 1990
- 11 Percent high school graduates
- 12 Percent bachelor's degrees
- 13 Percent below poverty level
- 14 Percent unemployment
- 15 Per capita income
- 16 Total personal income
- 17 Geographic region (1=northeast, 2=northcentral, 3=south, 4=west)

(continue to page 2 ↓)

Your task is to develop a model, using SAS, to predict the *per capita* number of serious crimes in these counties. This is a variable you will have to create in the SAS DATA step as a ratio of variables 10 and 5. You should also create and study a new predictor variable, population *density*, by dividing total population by the county land area. Minimal SAS code to read the data (once you copy the data onto portable memory) is:

```
data;  
infile 'F:\county_demographics.txt';  
input id county $ state $ area pop pctyng pctold docs beds  
      crimes hsgrads badegs pctpov pctunem income totinc  
      region @@;
```

The above assumes that the data reside on the "F" drive in a file named county\_demographics.txt ; substitute the appropriate drive letter for the computer you are using.

*You* are a policy advisor for a candidate running for the office of President of the United States! What variables best predict county crime rates, and what can counties do, if anything, to reduce their crime rates? Boil your results down into a simple executive summary for your candidate (who, like your instructor, has a limited attention span). This assignment is open-ended and exploratory: your instructor has little idea what the results will be!

Hints: start with the numerical (quantitative) predictor variables, using a model selection routine. Subsequently, look at adding categorical predictor variables into the model using an analysis of covariance routine.

Hand in, stapled:

One-page cover sheet with your name and typed executive summary of the analysis and results, interpreted in your own words

Printed **program(s)** and **output** of the central SAS analyses supporting your results