Biostatistics 201A, Fall 2021
Data Analysis Assignment 2
Due Friday, December 10, 2021

As part of a group (ideally of 3-4 people, although 2 or 5 people will be considered), your task is to analyze the "County Demographic Information" (CDI) data set, which contains information from 440 counties in the United States collected from 1990-1992. The primary objective of the investigation is to develop insights into predicting the county crime rate, which you might summarize as the number of crimes per 1,000 population (CRM_1000). You may use any other variables as predictors, and you may want to consider transformations of variables, derived variables (in an attempt to obtain interpretable findings from correlated predictors), and interaction effects and/or polynomial terms. Working as part of a group is meant to encourage sharing ideas.

The variables in the CDI data set are as follows:

Var	Variable name	Description
1	ID number	1-440
2	County name	Text string containing name of county
3	State	Two-letter text string containing abbreviation of state name
4	Land area	Land area measured in square miles
5	Total population	Estimated 1990 population
6	Percent of population aged 18-34	Percent of total population in age range from 18-34
7	Percent of population aged 65+	Percent of total population in aged 65 or older
8	Number of active physicians	Number of professionally active nonfederal physicians,1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of persons 25 years old or older who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of persons 25 years old or older with bachelor's degrees
13	Percent below poverty level	Percent of 1990 population with income below poverty level
14	Percent unemployment	Percent of labor force that is unemployed
15	Per capita income	Income (in dollars) per person among those in 1990 population
16	Total personal income	Total personal income (in millions of dollars) among those in 1990 total population
17	Geographic region	U.S. Census Bureau classification of region of the U.S. (1=Northeast, 2=North Central, 3=South, 4=West)

The assignment is for each group to produce a report providing insight into predicting crime rates. The report can have up to six pages of text; additional tables, figures, or other summaries may be appended. The report will be evaluated based on your ability to communicate how population characteristics relate to crime, so it would be helpful to motivate analysis choices and to interpret fitted model parameters clearly and carefully. It is expected that the analysis would examine marginal distributions as well as pairwise relationships between variables (e.g., to check for apparent nonlinearities). It is also expected that several candidate models for predicting crime rates would be considered and that the analysis would check for influential observations, multicollinearity, and possible violations of regression model assumptions. In line with the format of many other scientific reports, the report might have an introduction, a "Methods" section describing analysis techniques, a "Results" section summarizing findings, and a "Discussion" section highlighting key conclusions or interpretations. (Without carrying out a formal evaluation, the Discussion section might also comment on how model predictions could be evaluated based on more recent data.) The report should also state that everyone in the group contributed to and approved the final report.