P8130: Biostatistical Methods I
Final Project (Fall 2021)
Due, December 17th @ 5:00pm

Guidelines for Project Submission

This group project must be submitted through CourseWorks before the deadline. Email submissions WILL NOT be accepted and will receive a score of 'Zero' for all group members!!

All graphs, output and interpretations must be included in ONE PDF file, otherwise it will not be graded. In a separate attachment, you also must submit your R/Rmd code used in your project.

General Writing Instructions

Your project should not exceed 5 double-spaced pages using 11 or 12-point font, EXCLUDING figures and tables, references, appendix, that can be placed at the end of the main text. Be selective in your output and visual displays!

Your brief report should be structured as a publishable research article containing the following sections:

- Abstract
- Introduction (brief context and background of the problem)
- Methods (data description and statistical methods)
- Results
- Conclusions/Discussion

Your findings should be written as for an informed (but non-statistical) audience (no formulae!). Each figure and table should be of publishable quality and well notated, i.e., labeled and/or captioned.

Grading Instructions

The rubric attached will be used to evaluate the project. This is a group project and collaborations within your group are essential and great practice for your career.

Academic dishonesty or lack of contribution to the team effort will be penalized and reflected in individual grades.

You will be analyzing data from the “County Demographic Information” (CDI) data set, which contains characteristics of 440 counties in the United States collected from 1990-1992. **The primary objective of this investigation is to develop insight relevant to predicting the crime rate in counties**, which you might want to summarize as the crime rate per 1,000 population (CRM_1000). You may use any of the other variables as predictors, and you may want to consider transformations of variables, derived variables (in an attempt to extract interpretable information from correlated predictors), as well as interaction effects and/or polynomial terms. The variables in the CDI data set are as follows:

Variable name	Description
ID number	1-440
County name	Text string containing name of county
State name	Two-letter text string containing abbreviation of state
Land area	Land area measured in square miles
Total population	Estimated 1990 population
Percent of population aged 18-34	Percent of total population in age range from 18-34
Percent of population aged 65+	Percent of total population in aged 65 or older
Number of active physicians	Number of professionally active nonfederal physicians,1990
Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
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
Percent high school graduates	Percent of persons 25 years old or older who completed 12 or more years of school
Percent bachelor’s degrees	Percent of persons 25 years old or older with bachelor’s degrees
Percent below poverty level	Percent of 1990 total population with income below poverty level
Percent unemployment	Percent of labor force that is unemployed
Per capita income	Income (in dollars) per person among those in 1990 population
Total personal income	Total personal income (in millions of dollars) among those in 1990 total population
Geographic region	Classification (according to U.S. Census Bureau) of region of the U.S. (1=Northeast, 2=North Central, 3=South, 4=West)

You will turn a report of your findings (with a five-page limit) including an abstract that condenses the results into a one-paragraph summary. In your report, you should describe a final model and interpret its parameters in an accurate and useful manner. It is expected that you would first examine the marginal distributions and pairwise relationships between variables (e.g., to check to see whether any nonlinearities are immediately obvious), that you would explore several candidate models for predicting the crime rate, and that you would check for violations of regression model assumptions, influential observations, and multicollinearity.

Your report will be evaluated based on your ability to communicate insights about how population characteristics relate to the crime rate, so it would be helpful both to be clear about your motivation for carrying out certain analyses as well as to be clear about the interpretation of fitted model parameters. It is expected that your report would include a table summarizing parameter estimates associated with your final fitted model, characterizing predictor variables in a way that a reader can clearly understand.

Below you'll find some aspects to be addressed in your report. These are just a few suggestions, but feel free to add your own input/creativity to the analysis:

Data exploration: descriptive statistics and visualization.

- Explore the distribution of the outcome and consider potential transformations (if the case).
- Identify states with unusual rates and consider them as potential outliers/influential points.

In your multiple regression model, be careful of variables that are highly correlated and be selective of the ones that you choose to include in the analysis.

Consider selective interactions between variables.

Lastly, do not ignore the model diagnostics (check model assumptions and goodness of fit).