P8130: Biostatistical Methods I

Final Project (Fall 2020)

Due, December 18th @ 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 (not the R file), otherwise it will not be graded. In a separate attachment, you also must submit the 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 five summary pages. Be selective in your output and visual displays!

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

- Abstract

- Introduction (context, 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.

A hate crime is a traditional offense that is motivated by bias. A person commits a hate crime when one of a specified set of crimes (e.g., murder, arson, vandalism, etc.) is committed targeting a victim because of a perception or belief about their race, color, national origin, ancestry, gender, religion, religious practice, age, disability or sexual orientation, or when such an act is committed as a result of that type of perception or belief. These crimes can have devasting effects on the individual and/or community and they are currently the highest priority of the FBI's civil rights program.

For this project you will work with a dataset containing details on hate crimes occurring in the United States (by state) and **the goal is to identify variables that are associated (or mostly associated) with these unfortunate outcomes.** This dataset has been adapted based on original records used in a FiveThirtyEight article entitled: "Higher Rates Of Hate Crimes Are Tied To Income Inequality". You can access the original article and look for more context.

Dataset "HateCrimes.csv" contains the following variables recorded at the state level:

- **Outcome of interest**:
  - Hate_crimes_per_100k_splc: hate crime rate per 100,000 population; data was recorded during the first weeks of November, 2016 by the Southern Poverty Law Center (SPLC). The SPLC uses media accounts and self-report entries including both hate crimes and non-prosecutable hate incidents.
- **Covariates**:
  - Unemployment (2016): level of state unemployment - high/low
  - Urbanization (2015): level of state urbanization – high/low (based on the percentage of population living in metro areas)
  - Median_household_income (2016): median household income per state
  - Perc_population_with_high_school_degree (2009): percentage of adults (>25 yrs.) with a high school degree
  - Perc_non_citizen (2015): percentage of population that are not US citizens
  - Perc_non_white (2015): percentage of population that are non-white
  - Gini_index (2015): index measuring income inequality (range 0-100)
    https://www.investopedia.com/terms/g/gini-index.asp

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.
- The initial analysis published by FiveThirtyEight showed that income inequality was the main predictor of hate crimes rates.
  - Verify if this association holds true in this dataset, but also explore associations of all the other covariates mentioned above and draw your own conclusions.
  - 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 interactions between variables (e.g., urbanization), and if the case, fit stratified models.

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