Mid-proposal: analysis of suicide rates

Pratheepa Jeganathan

5/17/2019

This provides comments on the mid-term project proposal.

Grade: 9/10.

Your project has an interesting research question. The proposed workflow could be improved to complete your project by June 5, 2019.

Data: yealry-wise sucide rate in XX countries by sex and age, HDI, GDP measures.

You proposed to combine four different datasets. I could not find in the proposal what are the datasets related to variables you have listed in the project proposal. If you know the variables and the corresponding dataset, then okay.

A suggested workflow (no need to be exactly as below)

- To make your project to be done in a short time, I suggest using the data from WHO suicide statistics from Kaggle. This gives population-based statistics on suicide rate, so we need to use the crude rate of suicide per 100k population. This analysis provides information on age-standardized rates, so I recommend to use the same measure.
- Filter and save countries with missing suicide rate.
- After filtering countries with missing suicide rate, take a random sample of 100 countries and make sure each continent has approximately equal countries. Use this sample for the following analyses.
- It seems you're interested in comparing nonparametric and parametric methods.
- Do not forget to write acknowledgment for the data sources.
- Let's say your goal is to predict the suicide rate based on age-category, sex, year-cutpoints (1996-2005, 2006 - 2015), HDI, and GDP.
    1) Test whether crude suicide rate (rate per 100k population) associated with sex, age-category: 15-24 years, 35-54 years, 75+ years, HDI, and GDP. Within each sex, age-category, year-cutpoints (1996-2005, 2006 - 2015) test the correlation between suicide rate and HDI and GDP (use rank-based correlation and Pearson correlation). For each correlation measure, construct a confidence interval using nonparametric bootstrap.
    2) Use one of the nonparametric regression methods to model the relationship between suicide rate and sex, age-category, year-cutpoints, HDI, and GDP.
        - Use cross-validation to find the smoothing parameter.
        - Using the estimated model, predict the suicide rate in 10 countries that were not selected in this analysis (by sex, age-category, year-cutpoints). Compare the actual suicide rate and predicted value (by sex, age-category, year-cutpoints).
    3) Use parametric multiple linear regression to model the relationship between suicide rate and sex, age-category, year-cutpoints, HDI, and GDP.
        - Using the estimated model, predict the suicide rate in (the same countries that were considered in nonparametric regression) 10 countries that were not selected in this analysis (by sex, age-category, year-cutpoints). Compare the actual suicide rate and predicted value (by sex, age-category, year-cutpoints).

4) Compare the predicted suicide rate (by sex, age-category, year-cutpoints) using nonparametric and parametric regression methods.

5) Use our textbook/lecture notes to describe the parametric and nonparametric bootstrap, rank-based correlation, cross-validation, nonparametric regression, and write down the assumptions.

Comments on the proposed project:

5. A proposed solution to the problem using the nonparametric method.

1. For each column of data, produce confidence intervals using both statistical methods (parametric and nonparametric bootstrap), training on some of the data (cross-validation).
2. Find significant correlation for each column / category of data vs. suicide rate.
3. Cross validate on test data.
4. Repeat many times using different subsets for testing and training (cross-validation).
5. Compare results.

This workflow has not been clearly written.

I suggest using the above workflow.

6. A flowchart of the various tasks to complete the project

You do not need to collect new data. I recommend following the workflow provided above (select a random sample of 100 countries and predict the suicide rate in 10 of the countries that were not included in the model fitting).

Programming

Possibly use pandas library in Python for cross-validation of bootstrap intervals.

If you use cross-validation, you can simply implement in R (most of the nonparametric regression methods in R has a built-in function for doing cross-validation).