INTA 4803 Homework 1

Instructions

Please submit all analyses using `R Markdown` .html format unless otherwise noted. Follow the instructions specified below in each section. All answers are to be submitted to Canvas by the due date and time specified on the syllabus and Canvas. All analysis is to be completed individually, but you may utilize the class Slack channel to ask questions and collaborate on finding answers (i.e. directing students to Stack Exchange and the like).

Conceptual

Question 1
Atlanta's National Center for Human and Civil Rights has partnered with the U.S. Holocaust Museum and Yad Vashem to institute a prediction competition for scholars of political violence. These institutions wish to determine whether scholars should utilize traditional, parametric regression models or nonparamteric statistical learning models to determine answers to the following questions:

- Whether ethnic minorities which exist in a society ruled by a non-coethnic majority group are more likely to suffer higher risks of ethnic violence.
- What is the likelihood that Myanmar will suffer ethnic violence next year which will leave 1,000 or more dead.
- Whether poorer countries on average suffer higher levels of ethnic violence.
- What region of the world is at highest risk for outbreaks of ethnic violence?

Answer for each question whether a regression or statistical learning model is the most appropriate method to determine an answer. Justify your response theoretically.

Question 2
A first year student shows you his logistic regression model's output with 3 variables at or lower than the 5% threshold of statistical significance. This student

then proceeds to explain to you that these three variables must also predict his dependent variable with a high degree of accuracy because these three variables are statistically significant. Explain why this student is mistaken.

Question 3
The Curse of Dimensionality states than the possible number of combinations of a statistical model is given by 2^p where p is the number of parameters in the model. Why is the CoD fundamentally at odds with mainstream methodological practice in most quantitative social science? What are some common issues that arise because of the CoD, and what problems do these issues pose for estimating parametric regression models like logit or OLS? Why might a statistical model estimated on one set of data generalize poorly to new data, given what we know about the CoD? What can analysts do to overcome these issues?

Applied

Question 4
Use the dataset posted to Canvas to answer the following question. Provide the results to sections a & c of the question in a properly formatted regression table using `stargazer` or provide this in a separate Word document.

- Using the Fearon and Laitin (2003) dataset, replicated Model 1 from Fearon and Laitin's (2003) paper "Ethnicity, Insurgency, and Civil War".
- Did Fearon and Laitin account for the Time Series Cross-Sectional relationships in their data by accounting for autocorrelation among the x_i's?
- Re-estimate Model 1 using clustered standard errors or fixed effects in the `lme4` and `miceadds` Libraries respectively. What did Fearon and Laitin get wrong about their model, if anything?