PROJECT 2017

BST 232

ORGANIZATIONAL ASPECTS

We have structured this project as a data challenge using the Kaggle platform. You are welcome to work in groups of 3 or 4 people.

Your challenge is to analyze a sample dataset with two goals: provide predictions and estimate the average causal effect of interest. Furthermore, you will address the questions below. As in every data challenge, the groups are going to be ranked according to a certain metric. We consider two metrics. The first metric that we consider here is the average prediction error directly calculated in Kaggle. Note that you will have the possibility to submit your predictions twice every day. Moreover, we will consider as additional metric the discrepancy of your proposed estimate relative to the population exposure effect as defined below. More information is found the Kaggle website. *Note that the final model that minimizes the prediction error might differ from the final model that minimizes the discrepancy of your proposed estimate relative to the population exposure effect.* Please, follow the instructions below and the instructions given in the Kaggle website so that we have all the information needed to compute these metrics for you.

You will be evaluated based on your answers to the questions listed below and you will obtain an extra credit if you will rank among the top 3 groups according to one or both of the two metrics. Please, be brief in your answers.

The project is due on **November 27th at 11:59pm**. Online submissions only are accepted. The answers to the questions and code should be uploaded on Canvas in the designated project tab and on Kaggle, as described in the website.

DESCRIPTION OF THE SCIENTIFIC QUESTION and of THE DATASET

A group of researchers is interested in investigating the effect of maternal exposure to manganese at high doses on child neuro-development in a sample of mother-infant pairs in Bangladesh. Bangladesh is experiencing an epidemic of metals and heavy metals poisoning. The soil of the country is contaminated with arsenic and manganese. Occupational and environmental exposure to lead is also a concern. The levels of arsenic, manganese, and lead found in Bangladesh can be extremely high and the support of these environmental toxicants poorly overlaps with what observed in Europe and in the USA.

In a sample of 400 mother infant pairs, the investigators measured child cognitive development (Y) at 24 months using the Bayley scales for infant neurodevelopment. Furthermore, they obtained a measure of concentrations of manganese (X1, μ/dl), the primary exposure of interest, as well as of arsenic (X2, μ/dl) and lead (X3, μ/dl) in cord blood. This measure reflects maternal exposure to these metals during pregnancy. An extensive questionnaire was filled at baseline by the mothers to collect information on potential confounders of the manganese-cognitive development relationship. These include child age at testing (Z1, months), mother education (Z2, years), mother age (Z3, centered at the mean), mother IQ (Z4, z-scored), home quality score (Z5), daily nutritional intake of egg (Z6), meat (Z7), fish (Z8), and maternal smoking status (Z9). The data is generated so that observations are independent of each other and identically distributed (conditional on covariates). Under a certain model selection NUCA will hold. Note that not all covariates will be true confounders.

QUESTIONS

1. Find your best model to estimate the marginal causal effect of a change in manganese from 4 (roughly corresponding to the 25th percentile) to 17 (roughly corresponding to the 75th percentile) on cognitive development.

E[YX1=17]-EYX1=4]

1. Explain your strategy of model building in concise steps. Be sure to clarify the criteria you employ. Check whether the assumptions that these criteria rely upon are met in the sample.
2. Describe in two sentences the results of each step in your data analysis.
3. Interpret all the coefficients in your final model.
4. Provide the r code of your final model (see example).
5. Give an estimate of the distribution of the estimator of the marginal causal effect.
6. Provide the point estimate and confidence interval of the marginal causal effect. What do you conclude? Justify the appropriateness of the causal interpretation of your analysis.
7. What is the predictive ability of your model?
8. Find your best model to maximize prediction accuracy.
9. Explain your strategy of model building in concise steps. Be sure to clarify the criteria you employ. Check whether the assumptions that these criteria rely upon are met in the sample.
10. Describe in two sentences the results of each step in your data analysis.
11. Interpret all the coefficients in your final model.
12. Provide the r code of your final model (see example below).

Example code:

Using the variable names given above, say that the final model takes the form:

Y= β0+β1X1+β2X2+β3sqrt(Z1)+e.

Then, the code will look like:

sqrt\_Z1<-Z1^(-1/2)

lm(Y~X1+X2+sqrt\_Z1)