Fall 2021 Jingchen (Monika) Hu

MATH 301 Homework 5

Due: Sunday 10/10, 11:59pm; submission on Moodle under Discussion Forum

1. The utility result of our synthesized Race in class seem to suggest that there might exist more suitable predictor variables for Race compared to our chosen LogIncome. Experiment with LogExpenditure in the CE sample and report your findings.
2. Create $m = 20$ synthetic datasets with synthesized KidsCount. First, write an R script to create $m = 20$ synthetic datasets with synthesized KidsCount and the unsynthesized LogIncome and synthesized LogExpenditure (from a linear regression synthesis model given LogIncome from last lecture) as the two predictors, following the synthesis R script covered in class and the example R script of generating $m > 1$ synthetic datasets for Expenditure covered in last class. Next, apply the written script to generate $m = 20$ synthetic datasets with synthesized KidsCount, save it as a list, and create histograms of confidential and synthetic KidsCount for the first 3 synthetic datasets (all on one plot). Discuss whether the generated synthetic datasets vary from each other and how these results inform us about the necessity of simulating multiple $m > 1$ synthetic datasets. [Hint: You need to create the synthetic design matrix for the second step modeling in the loop over $m = 20$ since that design matrix depends on the synthetic predictor simulated from the first step.]
3. Use the DPMPM model with the `NPBayesImputeCat` R package, create $m = 5$ partially synthetic datasets of the ACS sample where `MIG` and `HISP` are synthesized. Check the MCMC diagnostics of `kstar` and the utility results of synthesized `MIG` vs confidential `MIG` and those for `HISP` using appropriate functions from the `NPBayesImputeCat` R package.

Be prepared to discuss these questions in class on Monday 10/11.