Fall 2021 Jingchen (Monika) Hu

MATH 301 Homework 4

Due: Sunday 10/3, 11:59pm; submission on Moodle under Discussion Forum

1. In class, we demonstrated using default priors for β_0, β_1, σ in the MCMC estimation. Now we consider independent and weakly informative priors: $\beta_0 \sim \text{Normal}(0, 10)$, $\beta_1 \sim \text{Normal}(0, 10)$, and $\sigma \sim \text{Cauchy}^{+}(0, 1)$. This last distribution is called a half Cauchy distribution. It consists of the re-scaled density of the Cauchy distribution over the positive real line, and is a popular prior for standard deviation parameters.

   In the `brm()` function of the `brms` package, these priors can be specified in the prior section.

   ```
   SLR_wi_fit <- brms::brm(data = CEdata,
                           family = gaussian,
                           LogExpenditure ~ 1 + LogIncome,
                           prior = c(prior(normal(0, 10), class = Intercept),
                                     prior(normal(0, 10), class = b),
                                     prior(cauchy(0, 1), class = sigma)),
                           iter = 5000,
                           warmup = 3000,
                           thin = 1,
                           chains = 1,
                           seed = 539)
   ```

   Now using these priors to perform a Bayesian data synthesis analysis for continuous LogExpenditure using LogIncome as the single predictor. Make sure to check MCMC diagnostics. Report your findings and discuss your preference of the prior choice in terms of utility preservation of the generated synthetic data.

2. Revisit the Bayesian simple linear regression synthesis model for continuous LogExpenditure in class. Consider now in addition to LogExpenditure, using another predictor, Race. Below a sample script for MCMC estimation is included.

   ```
   ff <- stats::as.formula(LogExpenditure ~ 1 + LogIncome + as.factor(Race))
   model <- stats::model.frame(ff, CEdata)
   X <- data.frame(stats::model.matrix(ff, model))
   ```

   ```
   MLR_2vars_fit <- brms::brm(data = CEdata,
                              family = gaussian,
                              LogExpenditure ~ 1 + LogIncome + as.factor(Race),
                              iter = 5000,
                              warmup = 3000,
                              thin = 1,
                              chains = 1,
                              seed = 127)
   ```

   Note that we save the output in `MLR_2vars_fit` where `MLR` stands for multiple linear regression (compared to `SLR` referring to simple linear regression) as we have more than one predictors. Also, in the sample script we exclude the `prior` statement in `brm()`, which indicates that the default priors will be used. A new `MLR_2vars_synthesize()` function is required, and the `SLR_synthesize()` function covered in class can be helpful.

Make sure to check MCMC diagnostics. By comparing the density plots of the confidential LogExpenditure and synthetic LogExpenditure, discuss whether the utility improves from using a simple linear regression model to a multiple linear regression model.

3. Create $m = 20$ Synthetic Datasets with Synthesized UrbanRural. First, write an R script to create $m = 20$ synthetic datasets with synthesized UrbanRural and unsynthesized LogExpenditure as the predictor, following the synthesis R script covered in class and the example R script of generating $m > 1$ synthetic datasets for Expenditure covered in class. Next, apply the written script to generate $m = 20$ synthetic datasets with synthesized UrbanRural, save it as a list, and create histograms of confidential and synthetic UrbanRural 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.

Be prepared to discuss these questions in class on Monday 10/4.