CS&SS/STAT 563 — Statistical Demography — Spring 2022 - Homework no. 4

Due Monday May 9 at 2:15pm on the course Canvas website.

1. The purpose of this question is to give you some practice in fitting the simplest form of Bayesian hierarchical model, namely the random intercept model, or Bayesian random effects one-way analysis of variance model.

   This question uses data from the 1975 U.S. Sustaining Effects Study of elementary education, available as the `egsingle` dataset in the `mlmRev R` package. This gives data on 1,721 students in 60 schools. We will take `math` (Mathematics achievement score) as the outcome variable, and `schoolid` (the code for the school the student attends) as the grouping variable. We will use only the data for year 0.5 (for which there are data on 1672 students). Your task is to estimate the Bayesian random effects one-way analysis of variance model for these data and interpret the results.

   (a) Write out the Bayesian random effects one-way analysis of variance model for these data as a set of equations. What are the unknown parameters to be estimated?

   (b) Specify and write down a reasonable prior distribution for the parameters. Explain your reasoning.

   (c) Estimate the model in a Bayesian way via Markov chain Monte Carlo using an appropriate software package.

   (d) Assess the convergence of your algorithm and whether it has run for enough iterations. If not, run it for longer.

   (e) Summarize the posterior distribution you obtain in graphical and tabular form.

2. Obtain the values of TFR for Peru for 1950-2020 from the 2019 World Population Prospects.

   (a) Fit a version of the double logistic decline model by nonlinear least squares to these data, assuming that the error variance remains constant over time (this will just give one set of double logistic parameter values and an estimated error variance). Note: Peru is still in Phase II up to 2020.

   (b) Plot the observed five-year declines against their fitted values, and comment on the fit.

   (c) Find the predictive distribution of Peru TFR for 2020-2025 *conditionally* on this model and on the estimated parameters, analytically or by simulation. Plot the distribution and give its median and a 95% prediction interval.

3. Obtain the values of TFR for the Netherlands for 1950-2020 from the 2019 World Population Prospects.

   (a) Identify the period in which the Netherlands entered Phase III of the fertility model.

   (b) Fit a (non-Bayesian) AR(1) model to the Phase III data, estimating the long-term mean, autoregressive parameter, and error variance.

   (c) Find the predictive distribution of Netherlands TFR for 2020-2025 conditional on this model, analytically or by simulation. Plot the distribution and give its median and a 95% prediction interval.

4. A fully converged simulation containing three MCMC chains from phase II, each of length 62,000, and three MCMC chains of Phase III, each of length 50,000, both thinned by 30, and 1,000 projection trajectories for all countries, is available as TFRsim.tgz on the Homework page. This is based on WPP 2017 data, and treats 2015 as the present year. After unpacking you'll find a README file that contains the code used to generate the simulation.

   (a) Use the get.tfr.mcmc, get.tfr3.mcmc, and get.tfr.prediction functions to obtain the MCMC (II and III) and the prediction objects, respectively.

   (b) Using the converged simulation, assess the fit of the double logistic model for Peru and Ecuador. In which country has fertility been declining faster according to the double logistic fit, and according to the raw data?

   (c) For each five-year period from 2020-2025 to 2095-2100, find the posterior predictive probability that the TFR of Peru will be higher than that of Ecuador. Find the probability that the TFR of Peru will be higher than that of Ecuador in all 16 five-year periods.