BASKIN SCHOOL OF ENGINEERING

Department of Applied Mathematics and Statistics

2015 First Year Exam, Take Home Question (Statistics)

Due by 5:30PM, Sunday June 7, 2015

Instructions:

Please work individually on this problem. You are allowed to consult any material you wish, but do not share with any other individual any information or comments about your findings or the models and methods you use. You are required to write a report using a word processing software (i.e., LaTeX or Microsoft Word). You are required to email your report as **one pdf file** to the graduate director at `thanos@soe.ucsc.edu`

by 5:30PM, Sunday June 7, 2015

Please organize and present the material in the best possible way. Be informative but concise. You should include a summary of your work at the beginning of the report, include and annotate all relevant figures and tables in the body of the report, write your conclusions in a separate section, and list your references (if any). Your report should consist of no more than 8 letter-size pages (typeset with 11pt or larger font and margins on all four sides of at least 1 inch), including all figures, tables, and appendices (but excluding the numerical codes); answers longer than 8 pages will lose credit for excess length. For those of you that will type your report in LaTeX, it is suggested (but not required) to use the template from `https://courses.soe.ucsc.edu/courses/ams207/Spring15/01`

For the implementation of the models included in this problem, you can use any language you feel comfortable with, but you are **not** allowed to use a pre-programmed sampler, such as the ones implemented in BUGS or STAN. You must include your MCMC codes (in R or other programming languages) at the end of your report; the codes do not count toward the 8-page limit.

Exam Problem:

For this project you will use the data from: `http://www.ams.ucsc.edu/~bruno/Take_Home.R` to explore the relationship between life expectancy (in years), gross domestic product (GDP) per capita (a measure of how rich a country is, recorded in US dollars) and population size in the Americas.

1. Perform a descriptive analysis of the data. What are the main insights from this analysis?

2. Fit separate (independent) Bayesian linear models for each country to explain life expectancy as a function of GDP per capita and population size. The model should have the form:

$$y_{it} = \beta_{0,i} + \beta_{1,i} v_{it} + \beta_{2,i} w_{it} + \epsilon_{it} \qquad \epsilon_{it} \mid \sigma_i^2 \overset{ind.}{\sim} \mathsf{N}(0, \sigma_i^2)$$

where i denotes the country and t denotes the year; y_{it} is the life expectancy; v_{it} is the base-10 log transformation of the GDP per capita, and w_{it} is the base-10 log transformation of the population size.

   (a) Explore the posterior distribution of all model parameters, $(\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \sigma_i^2)$, by running your own set of routines that implement a posterior simulation approach.

   (b) Justify your choice of priors and provide some analysis of the goodness of fit.

   (c) Provide graphs with point and interval estimates of the regression coefficients across all countries. Think carefully about how to present these graphs!

   (d) Remove the data point corresponding to the last year, for each country. Fit the models with the rest of the data, and perform a posterior predictive analysis of the life expectancy for the last year.

3. Let $\boldsymbol{x}_{it}' = (1, v_{it}, w_{it})$ and $\boldsymbol{\beta}_i = (\beta_{0,i}, \beta_{1,i}, \beta_{2,i})'$. Consider the hierarchical model

$$y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta}_i + \epsilon_{it} \qquad \epsilon_{it} \mid \sigma_i^2 \overset{ind.}{\sim} \mathsf{N}(0, \sigma_i^2), \quad \sigma_i^2 \overset{ind.}{\sim} \mathsf{IGam}(1, b)$$

$$\boldsymbol{\beta}_i = \boldsymbol{\theta} + \boldsymbol{v}_i \qquad \boldsymbol{v}_i \mid \boldsymbol{\Sigma} \overset{ind.}{\sim} \mathsf{N}_3(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \sim \mathsf{IW}(\nu, \boldsymbol{S})$$

where $\boldsymbol{\theta} \sim \mathsf{N}_3(\boldsymbol{A}, \boldsymbol{D})$, and $b, \nu, \boldsymbol{S}, \boldsymbol{A}$ and $\boldsymbol{D}$ are fixed parameters for which you have to specify appropriate values.

   (a) Provide the posterior full conditional distributions for all model parameters. Implement a set of routines that use those distributions to obtain samples from the joint posterior distribution. Again, make sure to justify your choice of priors and provide some analysis of the goodness of fit.

   (b) Provide graphs comparing the point and interval estimates of the regression coefficients you obtained under the independent model from part 2 to those you obtained with the hierarchical model. Explain the reasons for the differences.

   (c) Repeat the prediction exercise described above (part 2(d)). Compare the results and discuss.