PROJECT ASSIGNMENT
Bayesian Data Analysis

Christel Faes & Oswaldo Gressani

2024-2025

INSTRUCTIONS

- Establish groups of the size of about 3 students
- Make a report in question-answer style
- Describe the flow of your procedures and the reasoning behind them.
- Limit your report to 20 pages (excluding title page, but including tables and figures)
- **Deadline of report is 07/01/2025**
- Send also your programs so that we can check that the programs really work!
- Submit report and R program via Blackboard

PART 1

In this part, you have to use Bayesian software such as OpenBUGS, JAGS, R2OpenBUGS, R2jags or Nimble and any additional R packages such as CODA.

Research Question

We want to investigate the participation rate in the screening of colorectal cancer in Flanders. Colorectal cancer screening is a preventive medical test used to detect early signs of colon or rectal cancer. The goal of this screening is to identify cancerous or precancerous changes in the colon (large intestine) or rectum before symptoms develop, making treatment more effective. Colorectal cancer is one of the most common forms of cancer, and early detection significantly improves survival rates. In Flanders, the program targets adults aged 50 to 74. The screening test is free of charge for the target population, as it is funded by the government.

Data description

The dataset contains the following variables:

- The number of invited individuals per gender, age group (5 age-years per group) and municipality.
- The number of individuals participating per gender, age group (5 age-years per group) and municipality.
- Naam = municipality

Note that values in the above dataset in red are imputed values.

Specific Questions

1. **Model 1**. Model the number of participants in the screening program as:

$$Y_{iag} \sim Binom(\pi_i, N_{iag}),$$

where Y_{iag} is the number of participants amongst the N_{iag} number of invited individuals, in municipality i of age a and gender g. π_i is the participation rate per municipality. Assume conjugate non-informative priors for the participation rate π_i.

   - What do you conclude from this model?
   - Present a caterpillar plot of the posterior participation rates.
   - For which regions is $P(\pi_i < 0.30|Y) > 0.9$?

2. **Model 2.** Model the number of participants in the screening program as:

$$Y_{iag} \sim Binom(\pi_{iag}, N_{iag}),$$

and assume that the participation rates are impacted by age and gender, but are similar in some way amongst municipalities:

$$logit(\pi_{iag}) = \alpha + \beta a + \gamma g + b_i$$

with $b_i \sim N(0, \sigma^2)$. This is equivalent to to specifying a logistic random effects model. The effect βa represents the effect of age (modelled linearly or using dummies) and γg the effect of gender. Assume vague priors for all the parameters.

- Compare convergence for the hierarchically centered versus uncentered model.
- Investigate the amount of shrinkage in this analysis.
- What do you conclude from this model?
- For which regions and age-gender-groups is $P(\pi_{iag} < 0.30|Y) > 0.9$?

3. Which of the above models is better? Do a model comparison using information criteria. Investigate whether there are any outlying observations (based on the selected model).

Important notes:

- For questions 1 and 2, take a look at the following example code: `https://webbugs.psychstat.org/wiki/Manuals/Examples/Surgical.html`
- Give details about the assumed model, including the used prior information.
- Give and explain the BUGS code (= model specification) in each question.
- Make sure that you interpret all results in a Bayesian fashion.

PART 2

The Gibbs algorithm

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$ and consider the following (unscaled) posterior density:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\left(-8\theta_1^2\theta_2^2 - 0.5\theta_1^2 - 0.5\theta_2^2 + \cos(2\pi + 0.3)\theta_1\theta_2 + 0.3\theta_1 + 0.2\theta_2\right).$$

1. Derive the conditional posteriors $p(\theta_1|\theta_2, \mathcal{D})$ and $p(\theta_2|\theta_1, \mathcal{D})$. To what well-known parametric family do these conditional posterior distributions belong? Explain mathematically how you computed the conditional posterior mean $\mathbb{E}(\theta_1|\theta_2, \mathcal{D})$ and the conditional posterior variance $\mathbb{V}(\theta_1|\theta_2, \mathcal{D})$.

2. Based on the full conditionals obtained in the previous step, write a pseudo-code to obtain a random sample of total size 50 000 (including a burn-in of size 20 000) from $p(\boldsymbol{\theta}|\mathcal{D})$ using the Gibbs sampler.

3. Using R and without relying on external software/packages (except the *coda* package), write the Gibbs sampler based on your pseudo-code in step 2. Please specify *set.seed(2025)* at the beginning of your code for reproducibility of your results. What is your acceptance rate?

4. Use the Geweke convergence diagnostic test on the generated chains and interpret your results.

5. Compute a point estimate for θ_1 and a 95% quantile-based credible interval.

6. Using your generated chains, estimate the probability $\mathbb{P}(\theta_1 > 0.5)$.

The Metropolis algorithm

1. Write the log posterior distribution $\log p(\boldsymbol{\theta}|\mathcal{D})$ and obtain analytically the gradient $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}) = (\partial \log p(\boldsymbol{\theta}|\mathcal{D})/\partial\theta_1, \partial \log p(\boldsymbol{\theta}|\mathcal{D})/\partial\theta_2)^\top$ and Hessian matrix:

$$\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}) = \begin{pmatrix} \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_1^2} & \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_2^2} \end{pmatrix}.$$

2. Using the gradient and Hessian matrix obtained in the previous step, implement a Newton-Raphson algorithm in R to find the posterior mode of $p(\boldsymbol{\theta}|\mathcal{D})$ and denote by $\boldsymbol{\theta}^*_{NR} = (\theta_1^*, \theta_2^*)^\top$ the mode after convergence of the algorithm (Note: round $\boldsymbol{\theta}^*_{NR}$ to five digits after the decimal point).

3. Compute a Laplace approximation to $p(\boldsymbol{\theta}|\mathcal{D})$ around $\boldsymbol{\theta}^*_{NR}$. Report the covariance matrix Σ^* of the Laplace approximation (Note: round your results to five digits after the decimal point).

4. In R write a random-walk Metropolis algorithm to explore the joint posterior $p(\boldsymbol{\theta}|\mathcal{D})$ using a Gaussian proposal with covariance matrix $\widetilde{\Sigma} = c\Sigma^*$. Use a chain of length $M = 50\ 000$ and tune c to (approximately) reach the optimal acceptance rate of 23%. Please specify *set.seed(1993)* for the sake of reproducibility of your results.

5. Using your generated chains, estimate of the probability $\mathbb{P}\left((\theta_1/\theta_2) > 0.45\right)$.