ADIM: Bayesian GLM. MCMC methods

Master’s Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-12-02

VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante
jmarmin@eio.upv.es

Motivation example

Heart Disease

- The study examines the relationship between:
  - the **myocardial infarction** (MI): $y = 1$ if MI occurrence, or $y = 0$ if No MI occurrence; and
  - **Age60**: Patients aged ≥ 60 (1) versus < 60 (0).
  - **Systolic blood pressure (SBP140)**: SBP ≥ 140 mmHg (1) versus < 140 mmHg (0).
- **Objective**:
  - Evaluate the association of age60 and sbp140 with MI probability.
  - Interpret the odds ratio (OR) for both predictors.

Table: Summary of Data from Study

y	age60	sbp140
0	<60	>=140
0	>=60	<140
0	<60	>=140
0	>=60	>=140
0	>=60	<140
1	<60	<140

Bayesian Logistic Regression Model

- **Logistic regression** is used to model MI's probability based on age60 and sbp140.
- **Likelihood**

$$y_i \sim \text{Bernoulli}(\pi_i)\,, i = 1, \ldots, 400\,,$$

using logit link:

$$\boldsymbol{logit}(\pi_i) = \beta_0 + \beta_1 \boldsymbol{age60}_i + \beta_2 \boldsymbol{sbp140}_i$$

- **Prior distributions** (weakly-informative):

$$\beta_0 \sim \mathcal{N}(0, 10^3),\ \beta_1\ \sim \mathcal{N}(0, 10^3),\ \beta_2\ \sim \mathcal{N}(0, 10^3),$$

Note: There are no conjugate priors available for the logistic regression model.

Table of contents

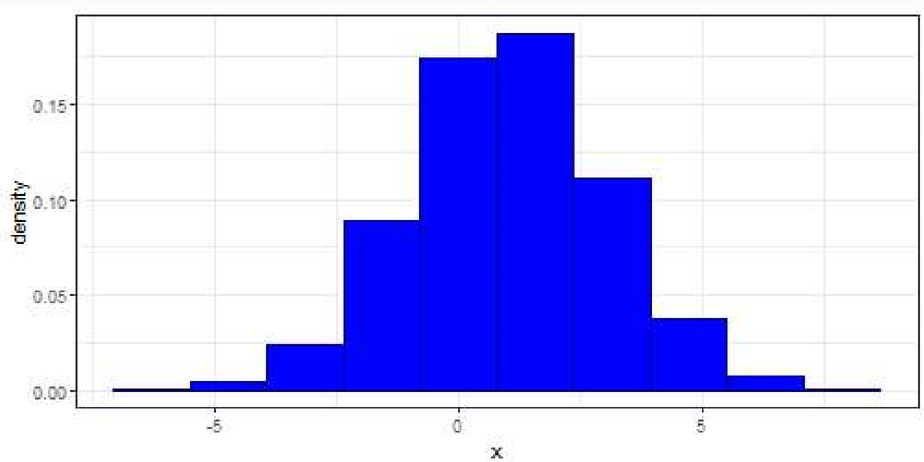
1. Bayesian computation. MCMC methods

Monte Carlo Methods

Monte Carlo Simulation

- Draw **realizations of a random variable** for which only its density function is (fully or partially) known.

```
x ← rnorm(1000, mean = 1, sd = 2)
```



Monte Carlo Integration

- Computing the mean of a N(1, 2),
  - $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$
- Using **Monte Carlo integration**:
  - Simulate from N(1, 2^2): $\phi^1, \ldots, \phi^N$.
  - Compute the mean of the simulated values: $E(X) \approx \frac{1}{N}\sum_{i=1}^{N} \phi^i$
- Doing `summary` of the simulation, we compute more measures:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-6.065	-0.334	1	1	2.332	8.086

Markov Chain Monte Carlo

- A Markov chain is a **stochastic sequence of numbers** where each value in the sequence depends only upon the last.
- If $\phi^1, \phi^2, \ldots, \phi^N$ is a sequence of numbers, then ϕ^2 is only a function of ϕ^1, ϕ^3 of ϕ^2, etc.
- Under certain conditions, the distribution over the states of the **Markov chain** will **converge to a stationary distribution**.
- The **stationary distribution is independent of the initial starting values** specified for the chains.
- AIM: construct a Markov chain such that **the stationary distribution is equal to the posterior distribution** $p(\theta \mid x)$.
- We combine Markov Chain with Monte Carlo simulation --> **Markov chain Monte Carlo (MCMC)**.
- They were proposed by first time in the Statistics area by Gelfand and Smith (1990) .

Posterior distribution

Estimating the probability to score a penalty

- **Likelihood**

$$p(\boldsymbol{y} \mid \pi) = \pi^{k}(1-\pi)^{N-k}$$

- **Prior distribution**

$$p(\pi) = \pi^{a-1}(1-\pi)^{b-1}$$

- **Posterior distribution**

$$p(\pi \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \pi) \times p(\pi) \propto \pi^{k+a-1}(1-\pi)^{N-k+b-1} = p^{*}(\pi)$$

MCMC: Metropolis-Hastings (MH)

1. Starting value $\pi^{(0)}$

2. For $t = 1, \dots, T$

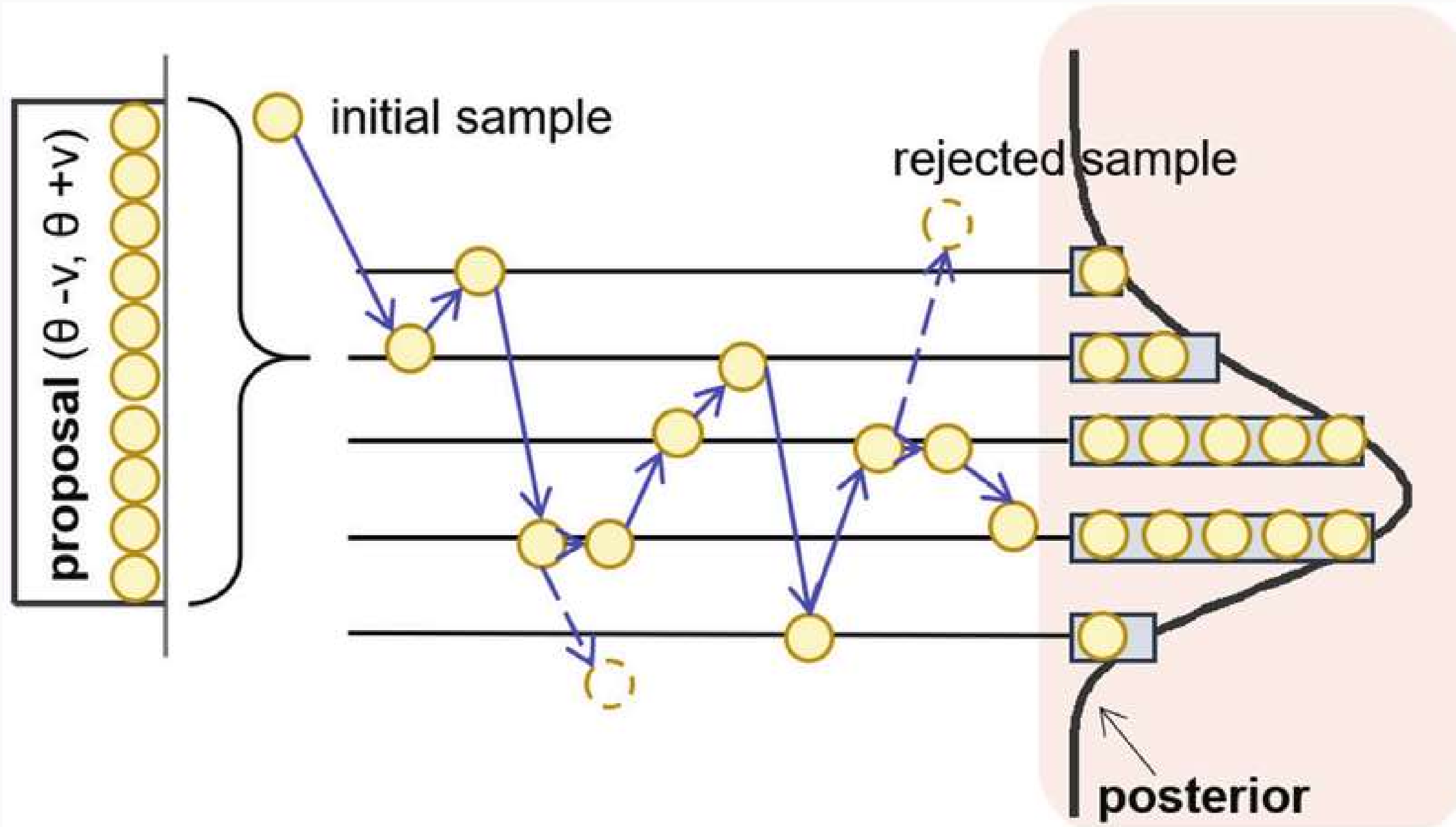
   - **We define a proposal distribution** (Usually similar to the objective distribution). In this case, $q(\pi \mid \pi^{(t-1)}) \sim logit - N(\pi^{(t-1)}, \sigma = 0.5)$. **Simulate** $\pi^{(prop)}$ from it.
   - Compute **probability of acceptance**:

$$\alpha = \min\left(1, \frac{p^*(\pi^{(prop)})q(\pi^{(t-1)} \mid \pi^{(prop)})}{p^*(\pi^{(t-1)})q(\pi^{(prop)} \mid \pi^{(t-1)})}\right)$$

   - Generate a **random number** u from the Uniform(0, 1).
     - $\pi^{(t+1)} = \pi^{(prop)}$, if $u \geq \alpha$,
     - $\pi^{(t+1)} = \pi^{(t)}$, if $u < \alpha$

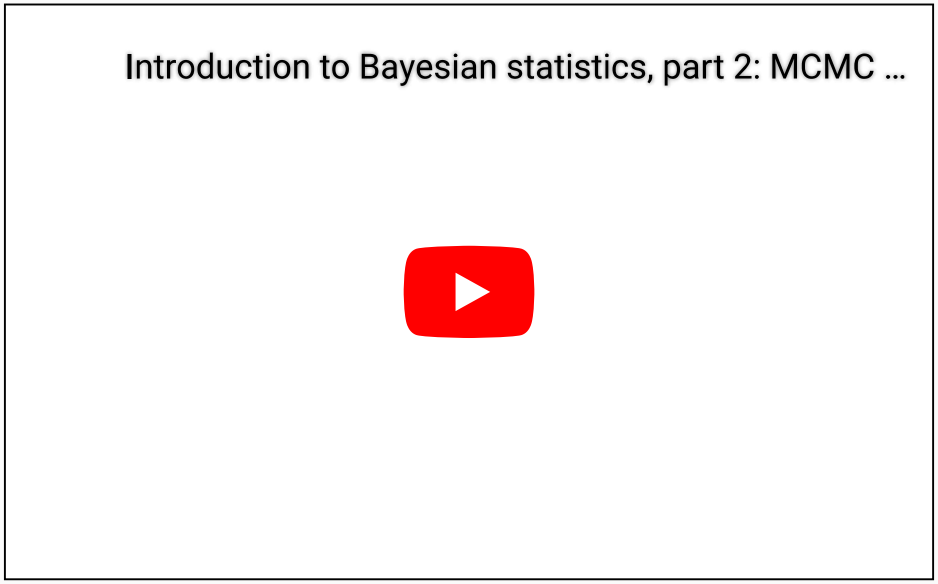
3. Finally, we **obtain** $\pi^0, \pi^1, \dots, \pi^T$ which is **a simulation of the posterior distribution**.

MCMC: Metropolis-Hastings (MH)



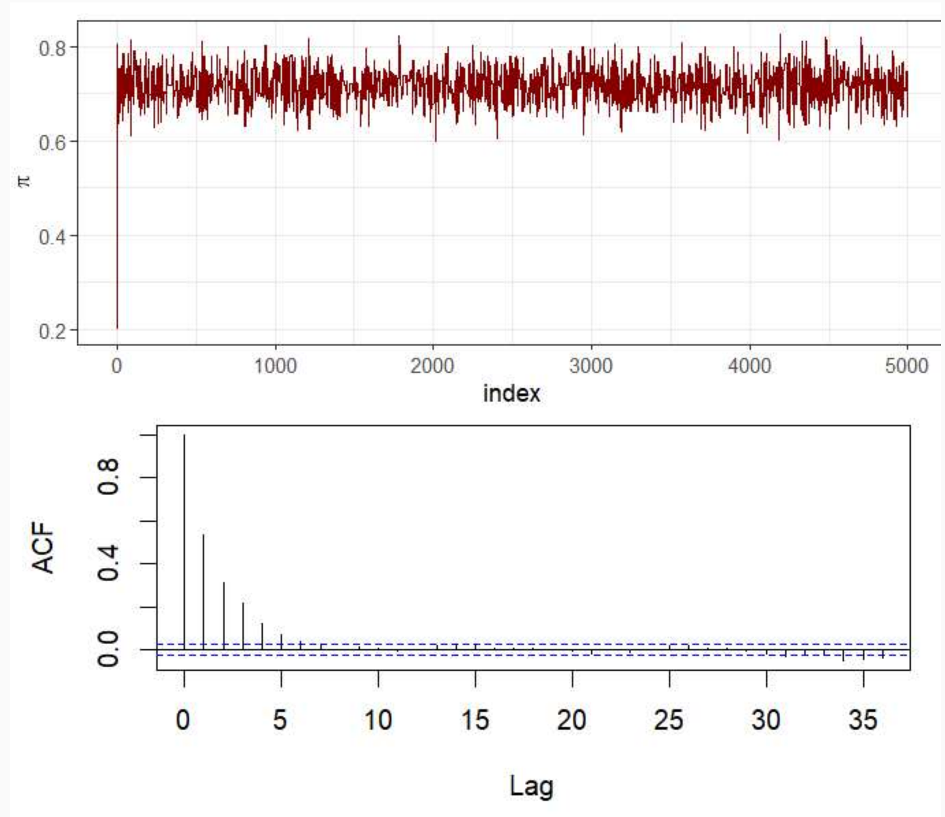
Approaching probability of score using MH

Visual Metropolis-Hastings

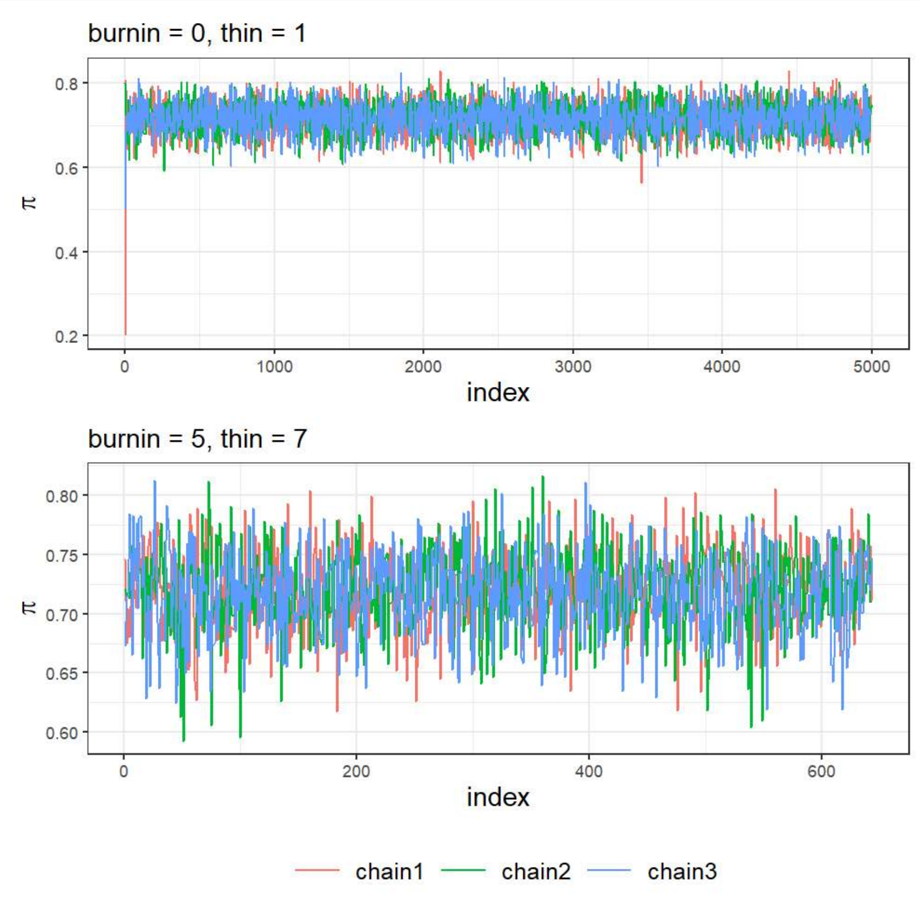


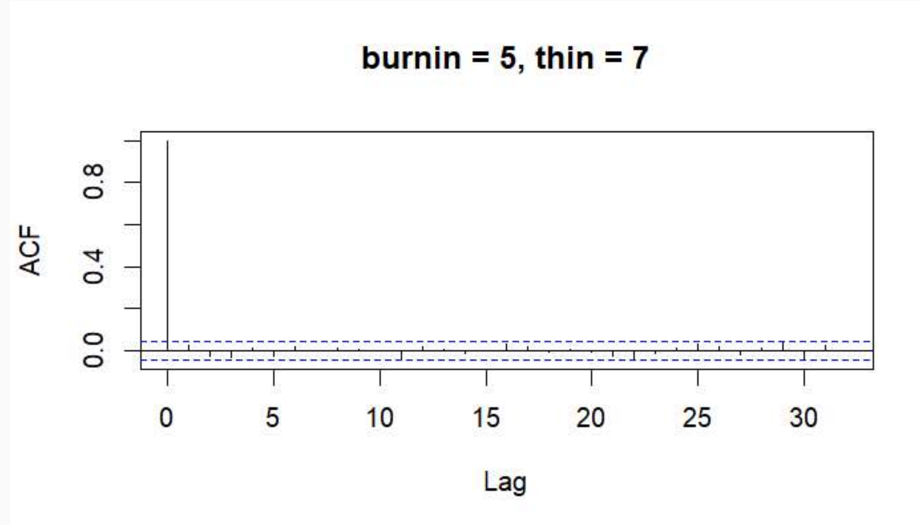
- Play the video from **minute 4:44**.

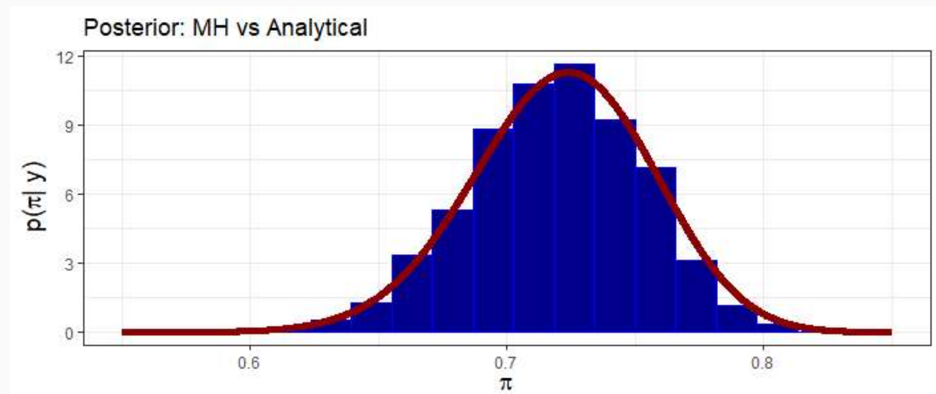
Tracing the chain. Is the chain autocorrelated?



MCMC. Burnin and thin





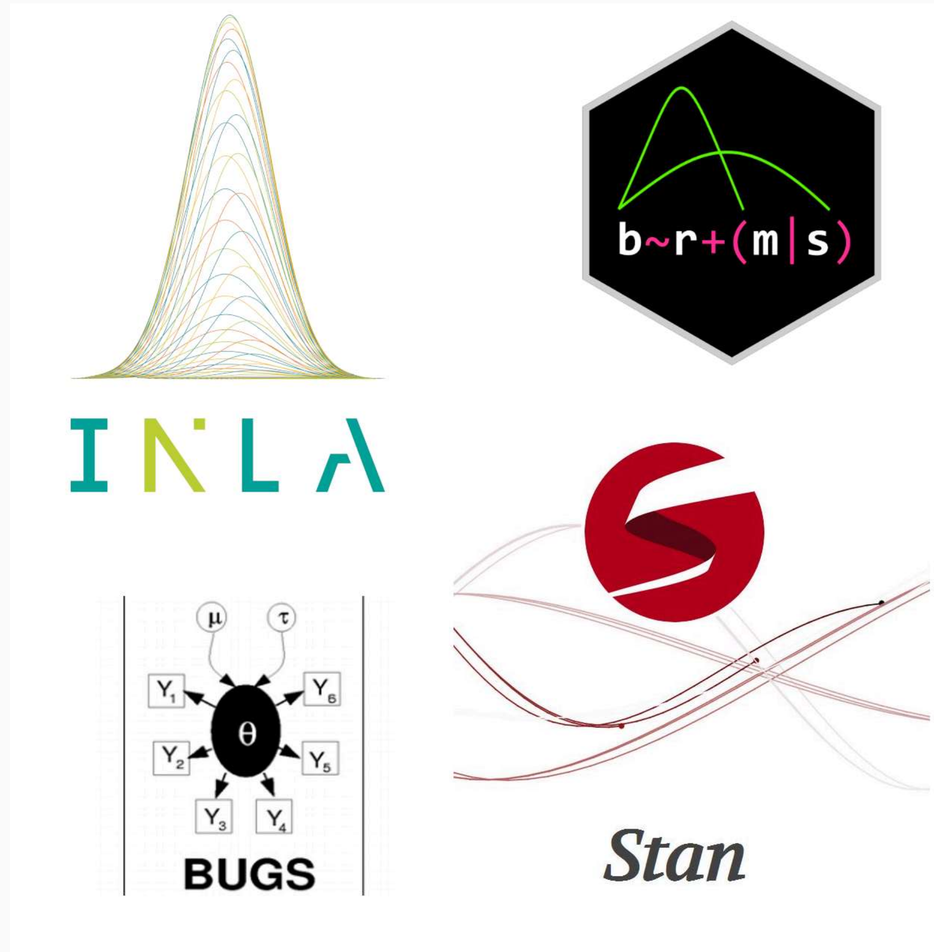


2. Bayesian Software for MCMC

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Software

- JAGS
- Stan
    - brms : allow for easy Bayesian Inference using Hamiltonian Monte Carlo
- INLA (it does not use MCMC methods, but it is a very powerful tool)
    - inlabru : facilitate Bayesian Inference in Spatio-temporal models
- MCMC pack (With this R-package, you can use MCMC methods with similar notation as usually use in R)
- Nimble



Bayesian Logistic Regression using JAGS

- **Likelihood**

  $$y_i \sim \text{Bernoulli}(\pi_i)\,, i = 1, \ldots, 400\,,$$

  using logit link:

  $$logit(\pi_i) = \beta_0 + \beta_1 age60_i + \beta_2 sbp140_i$$

- **Prior distributions** (weakly-informative):

  $$\beta_0 \sim \mathcal{N}(0, 10^3),$$

  $$\beta_1 \sim \mathcal{N}(0, 10^3),$$

  $$\beta_2 \sim \mathcal{N}(0, 10^3).$$

```
model_string <- "
model {
  for (i in 1:N) {
    y[i] ~ dbern(pi[i])
    logit(pi[i]) <- beta0 +
      beta1 * age60[i] +
      beta2 * sbp140[i]
  }
  # Priors for regression coefficients
  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
}"
```

Check `S1-JAGS-heart_attack.Rmd` for the complete solution

Bayesian Logistic Regression using brms

- **Likelihood**

  $$y_i \sim \text{Bernoulli}(\pi_i)\,, i = 1, \ldots, 400\,,$$

  using logit link:

  $$\text{logit}(\pi_i) = \beta_0 + \beta_1\,\mathbf{age60}_i + \beta_2\,\mathbf{sbp140}_i$$

- **Prior distributions** (weakly-informative):

  $$\beta_0 \sim \mathcal{N}(0, 10^3),$$

  $$\beta_1 \sim \mathcal{N}(0, 10^3),$$

  $$\beta_2 \sim \mathcal{N}(0, 10^3).$$

```
formula <- bf(y ~ age60 + sbp140, fami

# Fit the model using brms
fit_brms <- brm(formula,
    data = data_hattack,
    prior = priors,
    chains = 3,        # Number of MCMC
    iter = 5000,       # Total number of
    warmup = 1000,     # Number of itera
    thin = 1,          # Thinning interv
    seed = 123,        # Seed for reprod
  )

# Summary of the fitted model
summary(fit)
```

Check `S1-brms-heart_attack.Rmd` for the complete solution

What we have learned so far

- The biggest challenge for Bayesian inference has always been the **computational power** that more complex models require.
- **MCMC methods allow computationally estimating posterior distributions** that are analytically intractable.
- Today, there are many, many teams of researchers developing computational techniques for computing a posteriori distributions.

References

Books

- **Stan & R**: Lambert, B. (2018). A Student's Guide to Bayesian Statistics. SAGE Publications.
- **JAGS**: Plummer, M. (2019). JAGS User Manual Version 4.3.0.
- **OpenBUGS**: Cowles, M. K. (2013). Applied Bayesian statistics: with R and OpenBUGS examples (Vol. 98). Springer Science & Business Media.
- **WinBUGS**: Ntzoufras, I. (2011). Bayesian modeling using WinBUGS. John Wiley & Sons.
- **Stan**: Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin (2013). Bayesian Data Analysis. Chapman and Hall/CRC
- **INLA**: Gómez-Rubio, V. (2020). Bayesian inference with INLA. CRC Press.

References

Blogs

- http://wlm.userweb.mwn.de/R/wlmRcoda.htm
- https://rpubs.com/FJRubio/IntroMCMC
- https://darrenjw.wordpress.com/tag/mcmc/
- https://www.tweag.io/blog/2019-10-25-mcmc-intro1/

ADIM: Bayesian GLM. MCMC methods

Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-10-21
VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante

jmarmin@eio.upv.es