ADIM: Bayesian Computation and Mixed models

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

Joaquín Martínez-Minaya, 2024-12-09

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

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Motivation example

UNIVERSITA
POLITÈCNIC
DE VALÈNC

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:

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

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

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

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

Table of contents

1. Bayesian computation. MCMC methods

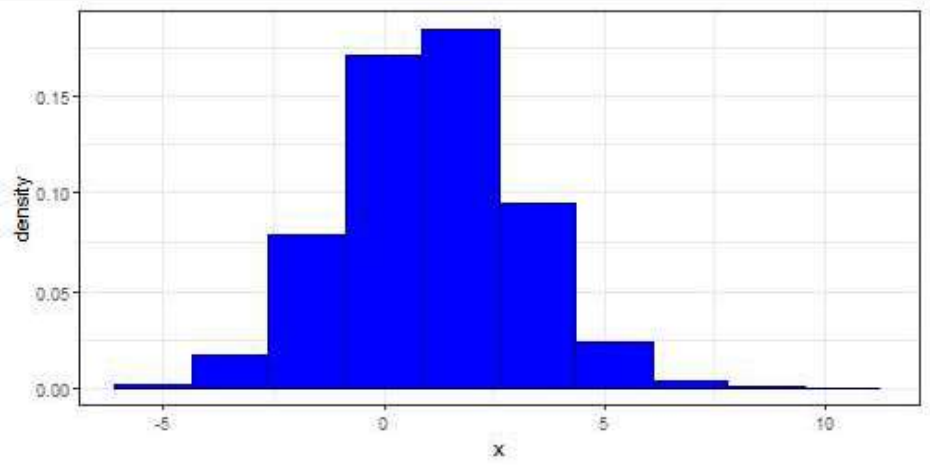
UNIVERSITA
POLITÈCNIC
DE VALÈNC

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.
-5.824	-0.296	1.04	1.031	2.342	9.789

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 $t = 1, \ldots, T$

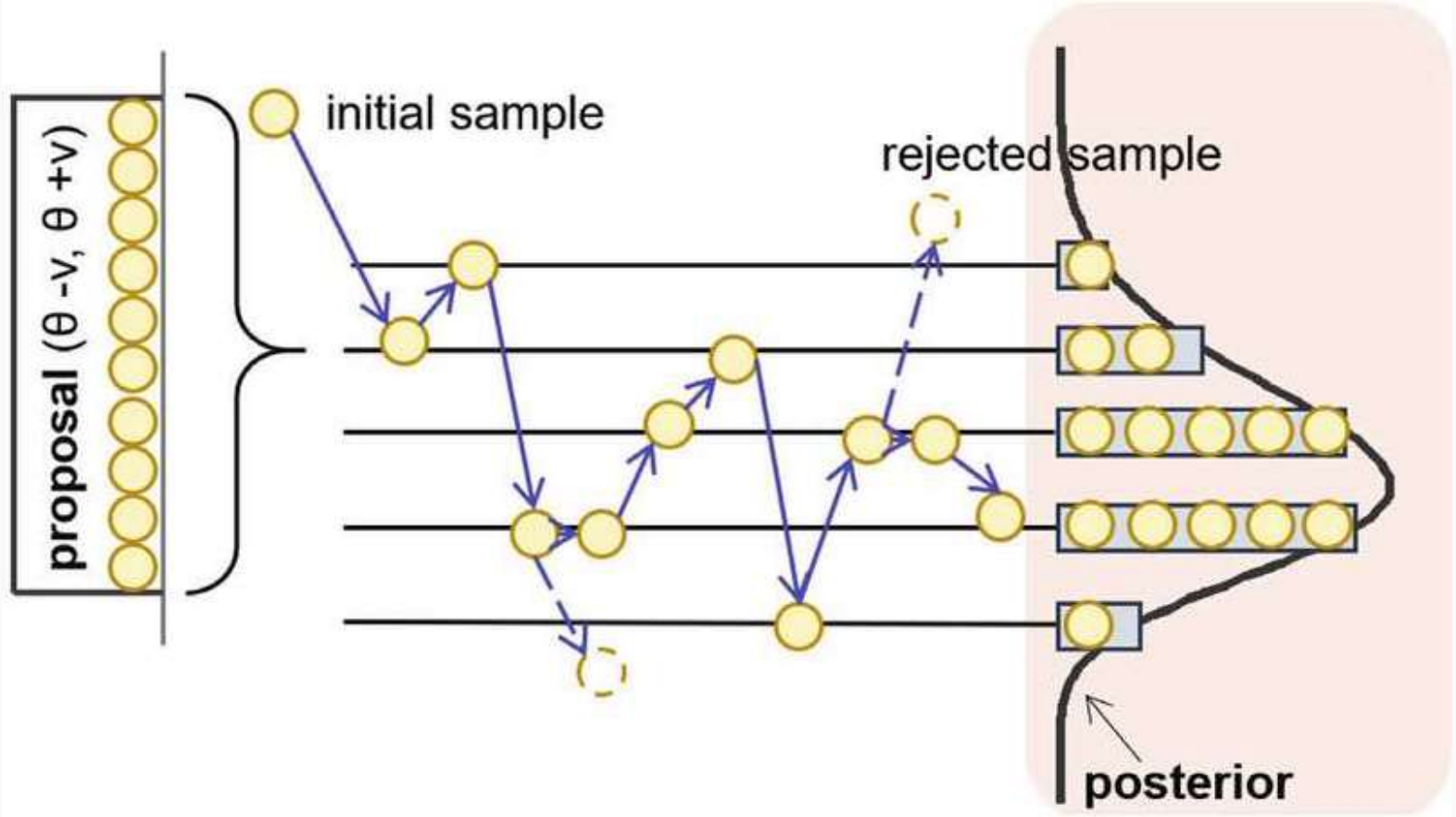
   - **We define a proposal distribution** (Usually similar to the objective distribution). In this case, $q(\pi \mid \pi^{(t-1)}) \sim logit - N(\pi^{(t-1)}, \sigma = 0.5)$. **Simulate** $\pi^{(prop)}$ from it.
   - Compute **probability of acceptance**:

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

   - 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 $u < \alpha$

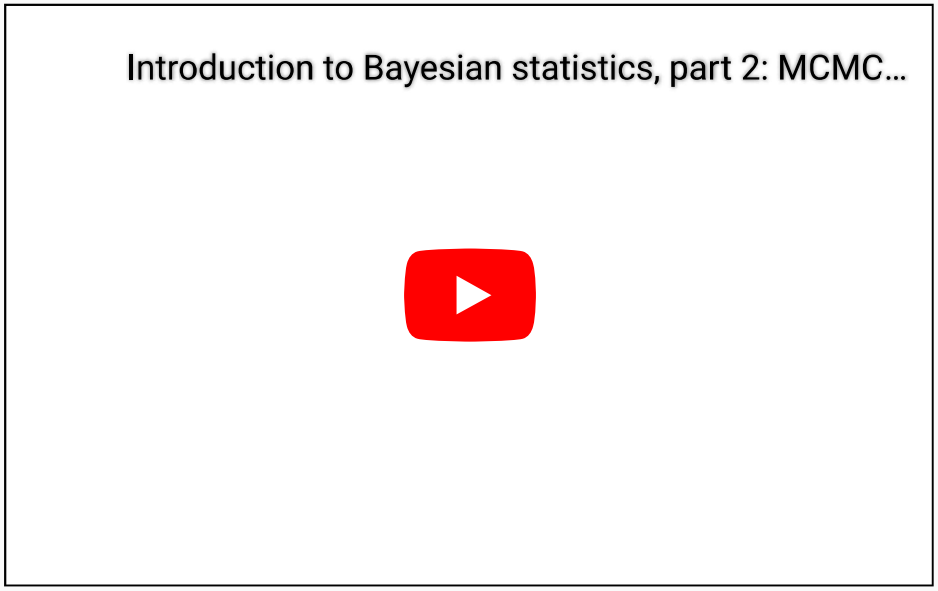
3. Finally, we **obtain** $\pi^0, \pi^1, \ldots, \pi^T$ which is **a simulation of the posterior distribution**.

MCMC: Metropolis-Hastings (MH)



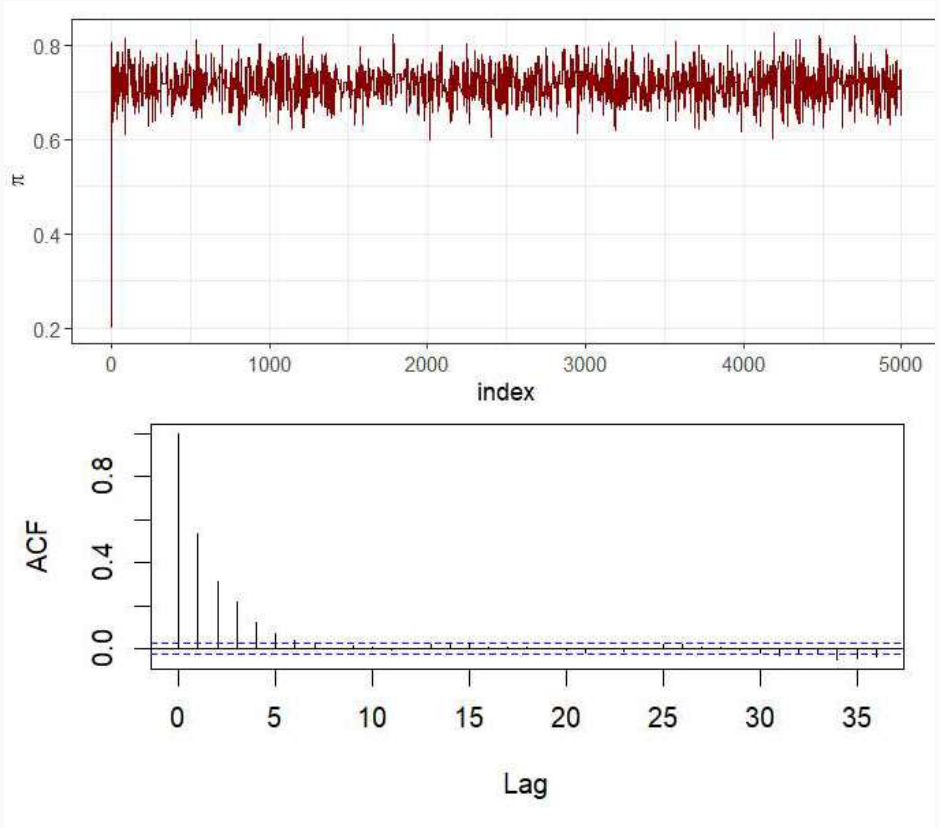
Approaching probability of score using MH

Visual Metropolis-Hastings

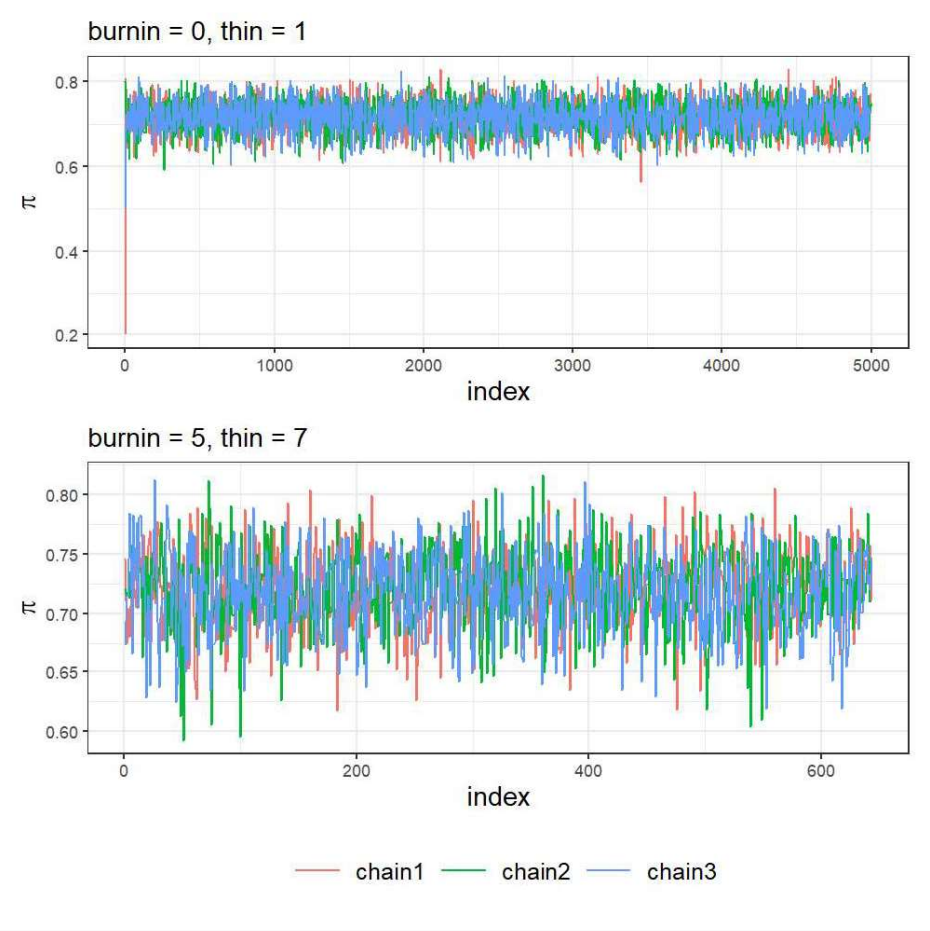


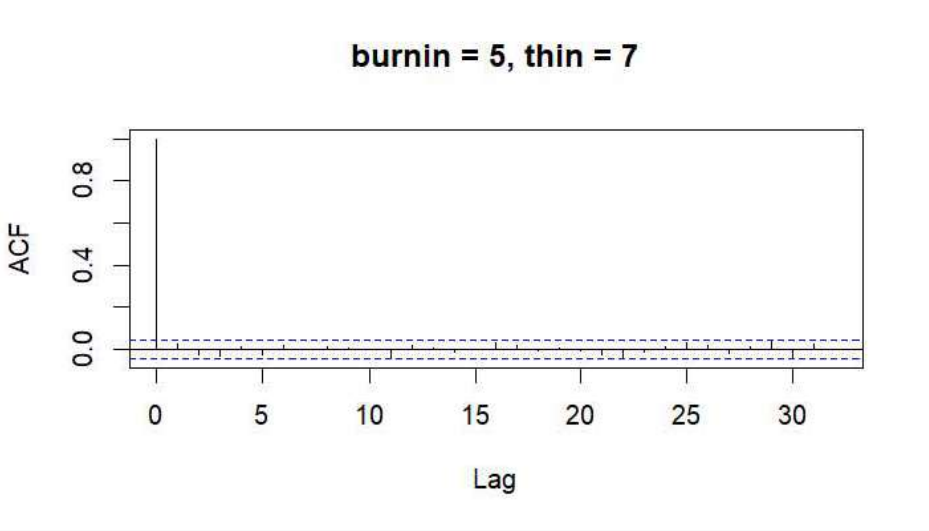
- Play the video from **minute 4:44**.

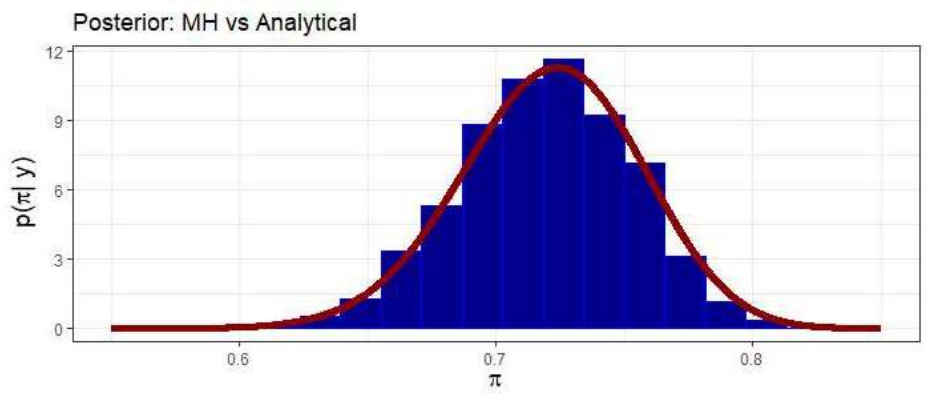
Tracing the chain. Is the chain autocorrelated?



MCMC. Burnin and thin





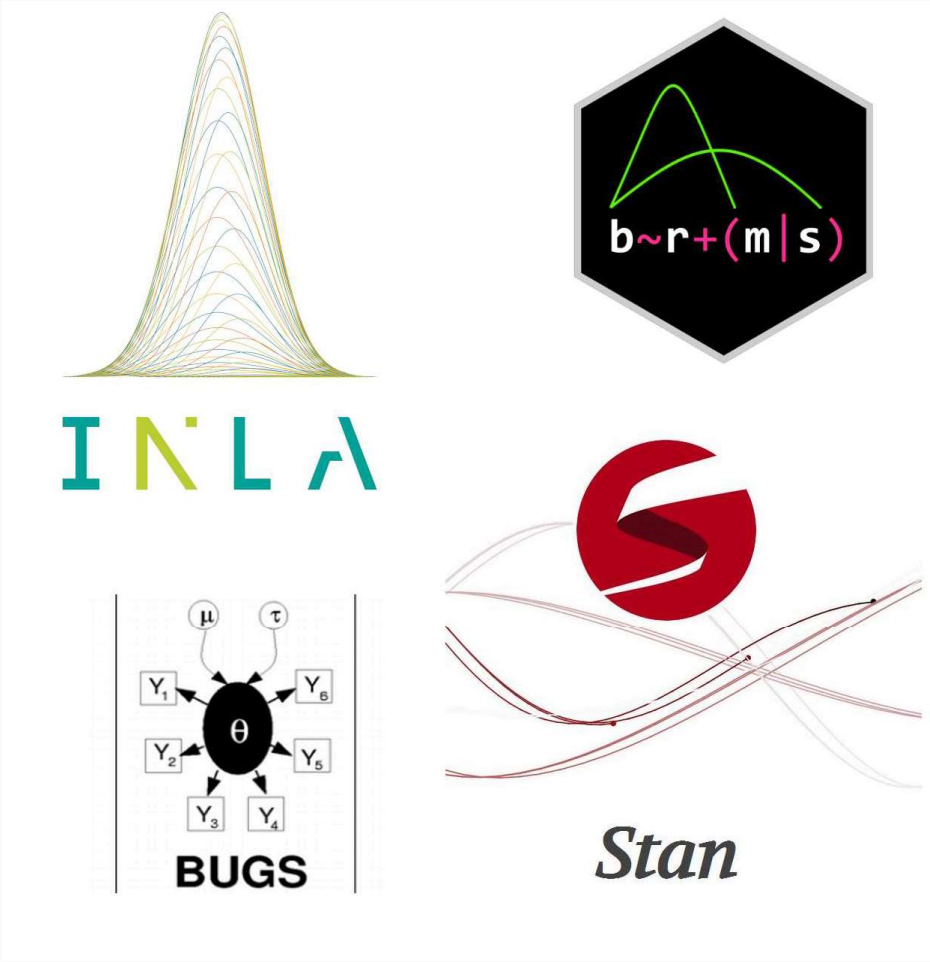


2. Bayesian Software

UNIVERSITA
POLITÈCNIC
DE VALÈNC

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:

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

- **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:

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

- **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 reproc
  )
```

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

Exercise: Predicting diabetes

Diabetes in **Pima Indian women aged 21 and older**. The dataset, contains diagnostic medical features and an additional zone variable (`zona1` to `zona30`), which categorizes patients geographically (Areas has been simulated).

Aim:

Examine the relationships between medical predictors and the likelihood of diabetes.

The dataset includes features such as:

- **Pregnancies**: Number of pregnancies.
- **Glucose**: Plasma glucose levels.
- **BloodPressure**: Diastolic blood pressure (mm Hg).
- **SkinThickness**: Triceps skin fold thickness (mm).
- **Insulin**: 2-hour serum insulin $\mu U/ml$.
- **BMI**: Body mass index kg/m^2.
- **DiabetesPedigreeFunction**: Family history risk score.
- **Age**: Age in years.
- **Zone**: Zone of residence, from `zona1` to `zona30`.
- **Outcome**: Indicator of diabetes diagnosis (0 = No, 1 = Yes).

Exercise: Predicting diabetes

```
# Load dataset
data_diab ← readxl::read_excel("../data/diabetes.xlsx")
kable(head(data_diab))
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
6.000000	148	72	35.00000	155.5482	33.6	0.627	50
1.000000	85	66	29.00000	155.5482	26.6	0.351	31
8.000000	183	64	29.15342	155.5482	23.3	0.672	32
1.000000	89	66	23.00000	94.0000	28.1	0.167	21
4.494673	137	40	35.00000	168.0000	43.1	2.288	33
5.000000	116	74	29.15342	155.5482	25.6	0.201	30

Predicting diabetes: Bayesian GLM

Logistic regression models the probability of diabetes based on selected predictors.

- **Likelihood**

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

using a logit link:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \, \text{Glucose}_i + \beta_2 \, \text{BMI}_i + \beta_3 \, \text{Age}_i.$$

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

$$\beta_0 \sim \mathcal{N}(0, 10^3), \; \beta_1 \sim \mathcal{N}(0, 10^3), \; \beta_2 \sim \mathcal{N}(0, 10^3), \; \beta_3 \sim \mathcal{N}(0, 10^3).$$

3. Hierarchical Bayesian Models

UNIVERSITA
POLITÈCNIC
DE VALÈNC

Example: Scoring penalties Valencia C. F.

- Liga Santander is one of the famous league around the world. In this example, we use data of the last 10 seasons in order to know the chance of success (π) to score a penalty for **Valencia Club de Fútbol**.

Again we talk about football

- We consider same experiment in **10 different teams**

- How can we model this situation? and what can we conclude?

- More generally, how can we incorporate **random effects**?

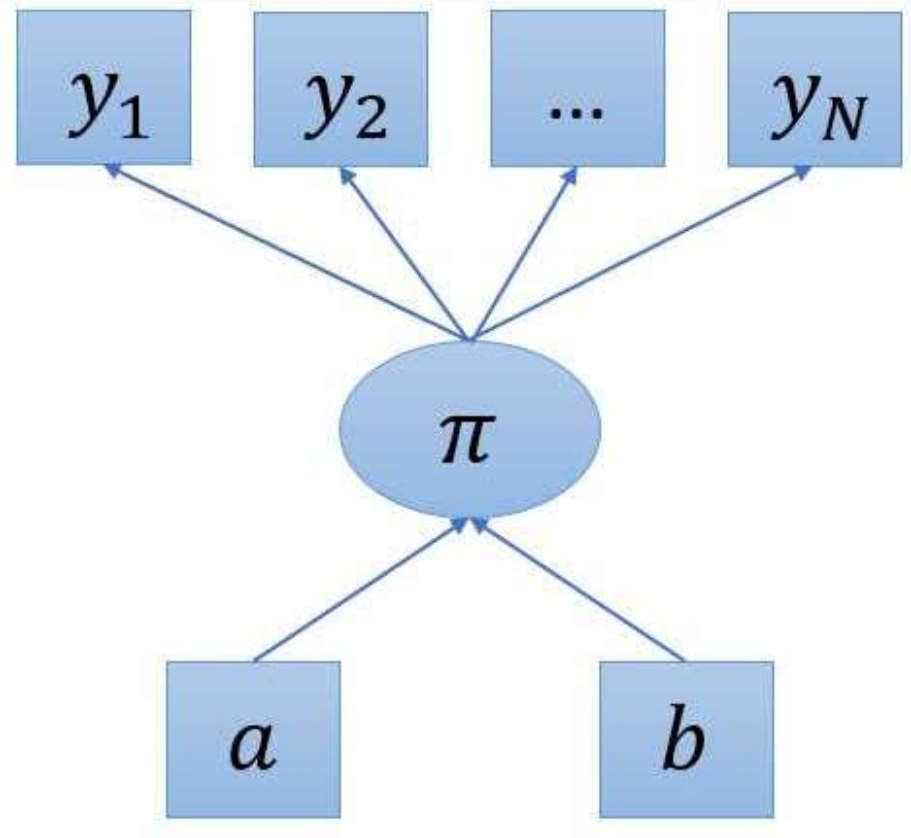
Three ways to do so

1. **All teams have the same characteristics**.

- Apply a **joint analysis** to all the teams.
- The probability of score a penalty (π) is the **same in all teams**.
- Observations are independent and identically distributed.

$$y_i \mid \pi \sim \text{Ber}(\pi)$$

$$\pi \sim \text{Beta}(a, b)\,, \text{ with a and b fixed}$$



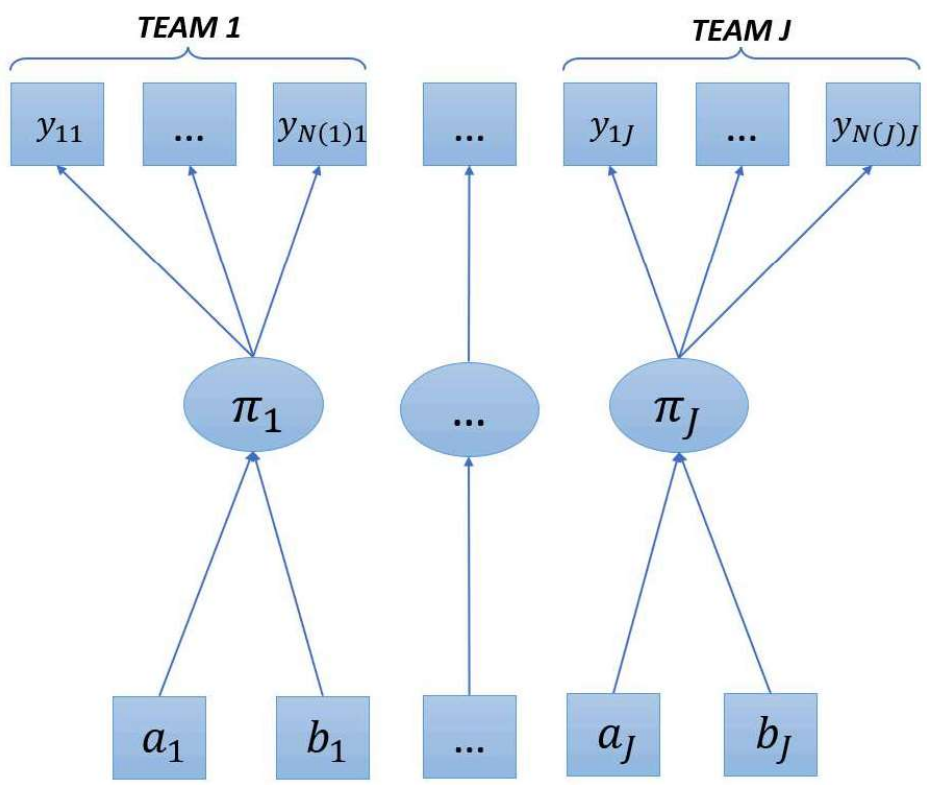
Three ways to do so

2. **Each team is different and has nothing in common with the others**.

- Apply an analysis to **each team separately**.
- Assume a **different proportion of presence** in each one: π_j, $j = 1, \ldots, J$. In this case, $J = 10$.
- Observations are independent but are **distributed differently in each team**.
- **Likelihood** is different for each team. For each j

$$y_{ij} \mid \pi_j \sim \text{Ber}(\pi_j)$$

$$\pi_j \sim \text{Beta}(a_j, b_j)\,, \text{ with } a_j \text{ and } b_j \text{ fixed}$$



In view of the two possible modelings

- Is it reasonable to assume **the same proportion of presence** in all teams?

- There are reasons to suggest that **there is variability in those proportions**:

  - The teams do not behave the same way.
  - The observations of the same team are more similar among themselves than when they are from different teams.

- Is it reasonable to think that **there is no relationship between the proportions of presence** of the different teams?

Although not identical, **teams** are at least **similar**.

Three ways to do so

3. Consider a **hierarchical model**.

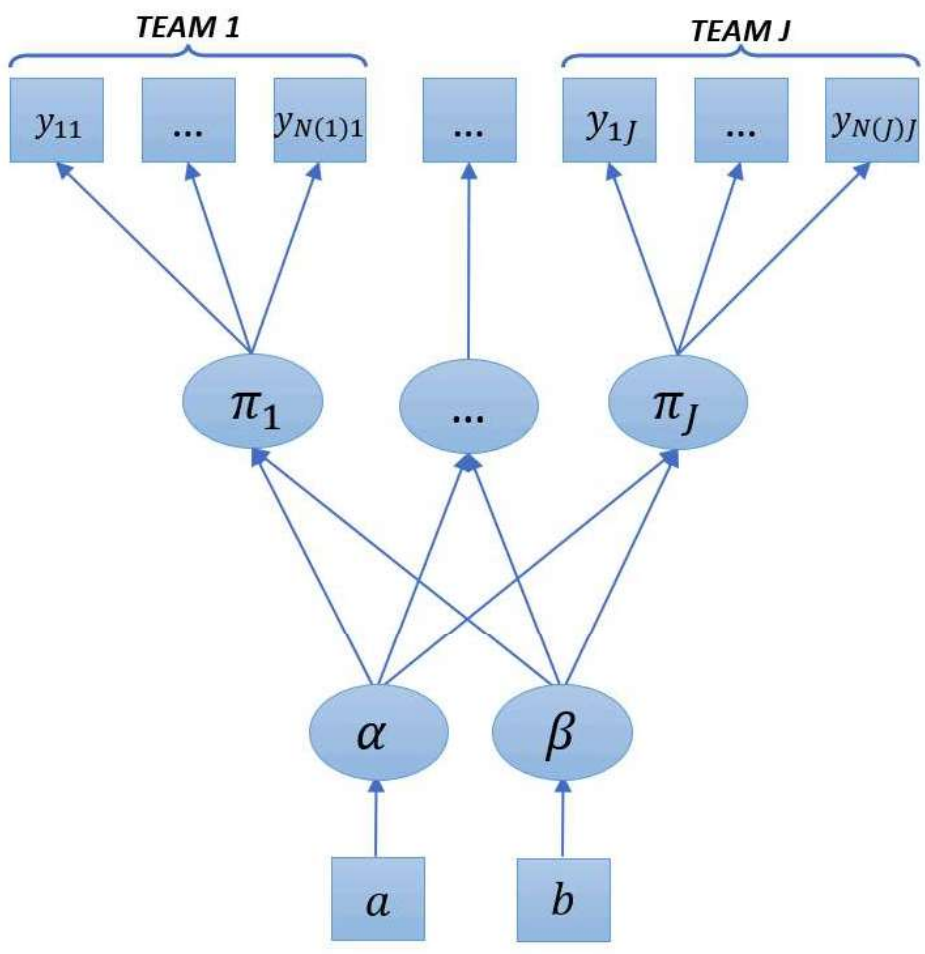
- The parametric vector $\boldsymbol{\pi} = \pi_1, \ldots, \pi_J$ is a **random sampling from a common distribution** that depends on a vector of **hyperparameters**, α and β, partial or totally unknown.

- The model

$$y_{ij} \mid \pi_j \sim \text{Ber}(\pi_j)\ , j = 1, \ldots, J$$

$$\pi_j \sim \text{Beta}(\alpha, \beta)$$

$$\alpha \sim p(a), \beta \sim p(b)$$



Predicting diabetes: Bayesian GLM

Logistic regression models the probability of diabetes based on selected predictors.

- **Likelihood**

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

using a logit link:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \, \text{Glucose}_i + \beta_2 \, \text{BMI}_i + \beta_3 \, \text{Age}_i.$$

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

$$\beta_0 \sim \mathcal{N}(0, 10^3), \ \beta_1 \ \sim \mathcal{N}(0, 10^3), \ \beta_2 \ \sim \mathcal{N}(0, 10^3), \ \beta_3 \ \sim \mathcal{N}(0, 10^3).$$

- **Geographical zone: Additional variability is captured by including the patient's zone (`zona1` to `zona30`) as a random effect.**

Diabetes. Bayesian Mixed GLMs

Logistic regression models the probability of diabetes, including a random effect for geographical zones to account for variability.

- **Likelihood**

$$
\begin{aligned}
y_i &\sim \text{Bernoulli}(\pi_i), \quad i = 1, \ldots, n, \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \, \text{Glucose}_i + \beta_2 \, \text{BMI}_i + \beta_3 \, \text{Age}_i + u_{\text{zone}[i]}.
\end{aligned}
$$

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

$$
\begin{aligned}
\beta_j &\sim \mathcal{N}(0, 10^3), \; j = 0, \ldots, 3 \\
u_{\text{zone}[i]} &\sim \mathcal{N}(0, \sigma_u^2),
\end{aligned}
$$

- **Hyperparameters**: The precision σ_u is modeled to capture the degree of heterogeneity across geographical zones (`zona1` to `zona30`).

$$
\sigma_u \sim \text{Cauchy}(0, 2).
$$

Diabetes using `brms`

```
# Define formula with a random effect for zone
formula ← bf(Outcome ~ Glucose + BMI + Age + (1 | zone), family = bernoulli())

# Set priors
priors ← c(
  prior(normal(0, 1000), class = "b"),          # Priors for fixed effects
  prior(cauchy(0, 2), class = "sd", group = "zone"),  # Prior for random effect
  prior(normal(0, 1000), class = "Intercept")   # Prior for intercept
)

# Fit the model using brms
fit_brms ← brm(
  formula = formula,
  data = data_diab,
  prior = priors,
  chains = 4,
  iter = 4000,
  warmup = 1000)
```

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 posterior 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: ADIM: Bayesian Computation and Mixed models

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

Joaquín Martínez-Minaya, 2024-12-09
VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante

jmarmin@eio.upv.es

UNIVERSITA
POLITÈCNIC
DE VALÈNC