Introducción a la estadística Bayesiana con aplicaciones de estimación en áreas pequeñas usando software STAN

Ignacio Alvarez-Castro   Juan José Goyeneche

Instituto de Estadística, Facultad de Ciencias Económicas y Administración, UdelaR.

XV Congreso Latinoamericano de Sociedades de Estadística
9 al 13 de Octubre 2023
Santiago de Cali, Colombia

Hasta ahora los modelos que vimos consisten de dos niveles:

- Modelo para los datos o verosimilitud $p(y|\theta)$
- Previa para los parámetros *que aparecen en el modelo para los datos*, $p(\theta)$

Llamamos **modelo jerárquico** a modelos en los que agregamos previas para los parámetros de las distribuciones que **NO** aparecen en el modelo para los datos.

Modelos jerárquicos

Tres niveles:

$$\begin{aligned} y_j = (y_{j,1}, \ldots, y_{j,n_j}) &\overset{ind}{\sim} p(y|\theta_j) \\ \theta_j &\overset{ind}{\sim} p(\theta|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

- $y_j = (y_{j,1}, \ldots, y_{j,n_j})$ observaciones para el grupo j
- n_j es la cantidad de observaciones en el grupo j

Tres niveles:

$$
\begin{aligned}
y_j = (y_{j,1}, \ldots, y_{j,n_j}) &\overset{ind}{\sim} p(y|\theta_j) \\
\theta_j &\overset{ind}{\sim} p(\theta|\phi) \\
\phi &\sim p(\phi)
\end{aligned}
$$

- $y_j = (y_{j,1}, \ldots, y_{j,n_j})$ observaciones para el grupo j
- n_j es la cantidad de observaciones en el grupo j

- $p(y|\theta_j)$ controla la variabilidad *al interior* de cada grupo
- $p(\theta|\phi)$ controla la variabilidad *entre* grupos
- $p(\phi)$ representa información previa sobre ϕ

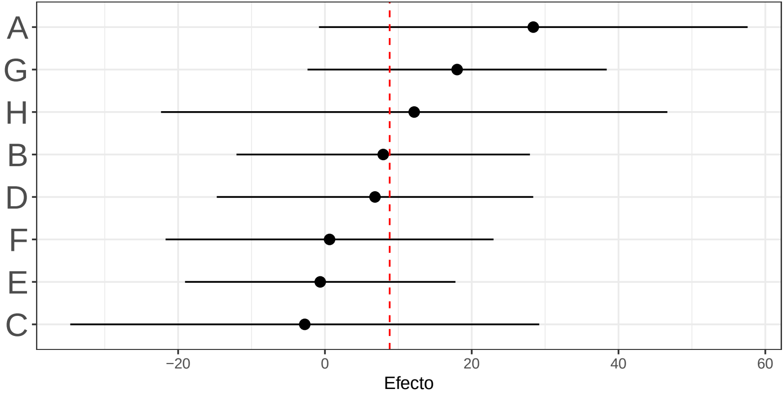
¿Cuál es la previa en estos modelos ?

Datos:

- Efectos de 'Coaching' para SAT en 8 escuelas
- Los experimentos son independientes por escuela
- Ejemplo clásico: Rubin 81, y BDA

	A	B	C	D	E	F	G	H
effect	28.39	7.94	-2.75	6.82	-0.64	0.63	18.01	12.16
see	14.90	10.20	16.30	11.00	9.40	11.40	10.40	17.60

8 escuelas



Estimación de modelos jerárquico

La estimación (clásica) de modelos jerárquicos puede ser difícil

- teoría asintótica requiere que n_j y J sean grandes
- la cantidad de parámetros a estimar crece con los datos
- no hay p-valor en `lmer`, Douglas Bates explica porque

Inferencia Bayesiana tiene ventajas aqui

- es válido para muestras finitas
- aprovecha estructura de dependencia entre parámetros

Usando MCMC el método de inferencia (Bayesiana) no cambia, pero puede ser computacionalmente costoso

Posterior conjunta

$$y_{j,i} \overset{ind}{\sim} p(y|\theta_j) \quad \theta_j \overset{ind}{\sim} p(\theta|\phi) \quad \phi \sim p(\phi)$$

La distribución posterior conjunta se refiere a la posterior de todos los parámetros en el modelo, dados los datos observados (lo mismo de siempre !!).

$$\begin{aligned} p(\theta, \phi|y) &\propto p(y|\theta, \phi)p(\theta, \phi) \\ &= p(y|\theta)p(\theta|\phi)p(\phi) \end{aligned}$$

Posterior conjunta

$$y_{j,i} \overset{ind}{\sim} p(y|\theta_j) \quad \theta_j \overset{ind}{\sim} p(\theta|\phi) \quad \phi \sim p(\phi)$$

La distribución posterior conjunta se refiere a la posterior de todos los parámetros en el modelo, dados los datos observados (lo mismo de siempre !!).

$$\begin{aligned} p(\theta, \phi|y) &\propto p(y|\theta, \phi)p(\theta, \phi) \\ &= p(y|\theta)p(\theta|\phi)p(\phi) \\ &= \left[\prod_{j=1}^{J} p(y_j|\theta_j)p(\theta_j|\phi)\right]p(\phi). \end{aligned}$$

Posterior conjunta

Se puede descomponer como

$$p(\theta, \phi|y) = p(\theta|\phi, y)p(\phi|y)$$

lo cual puede ser útil para obtener la expresión analítica de $p(\theta, \phi|y)$.

- Dado ϕ, $p(\theta_j|\phi, y)$ es sencilla de obtener en modelos conjugados
- Cuando ϕ es de baja dimensión, se pueden obtener aproximaciones numéricas para $p(\phi|y)$

$$p(\phi|y) \propto p(y|\phi)p(\phi)$$

donde

$$p(y|\phi) = \int p(y|\theta)p(\theta|\phi)d\theta$$

Más posteriores de interés

Posteriores Marginales:

$$
\begin{aligned}
p(\theta_j|y) &= \int \dots \int p(\theta, \phi|y) d\theta_{-j} d\phi \\
p(\phi_k|y) &= \int \dots \int p(\theta, \phi|y) d\theta d\phi_{-k}
\end{aligned}
$$

- θ_{-j} son todos menos θ_j
- $p(\theta_j, |y)$ considera la incertidumbre en todos los parámetros del modelo

Posterior condicional completa (Full conditional):

$$
\begin{aligned}
p(\theta_j|\theta_{-j}, \phi, y) &\propto p(\theta, \phi|y) \\
p(\phi_k|\theta, \phi_{-k}, y) &\propto p(\theta, \phi|y)
\end{aligned}
$$

- Posterior de θ_j dado los datos **y el resto de los parámetros**
- La base del algorítmo de GIBBS

Predictiva posterior

Queremos realizar predicciones de nuevas observaciones $\widetilde{y}$.

Se abren dos posibilidades:

- Hacer predicción de observaciones para un **grupo en la muestra**.
- Hacer predicción de observaciones para un **nuevo grupo**.

Predecir el total de renacuajos que sobreviven *en uno de los 48 estanques de la muestra*

Tenemos $(\theta^k, \phi^k)_{k=1}^S$ de la posterior conjunta, entonces para cada k

1. tenemos $\theta_j^k \sim p(\theta_j|y)$ para todos los j
2. Dado θ^k, simular $\widetilde{y}_j^k \sim p(y|\theta_j^k)$

obtenemos:

1. $(\widetilde{y}_j^k)_{k=1}^S$

Predictiva posterior: grupo nuevo

Predecir el resultado del programa *en la escuela* $J+1$

Para cada k

1. tenemos $\phi^k \sim p(\phi|y)$
2. dado ϕ^k, simulamos $\widetilde{\theta}^k_{J+1} \sim p(\theta|\phi^k)$
3. dado $\widetilde{\theta}^k_{J+1}$, simulamos $\widetilde{y}^k_{J+1} \sim p(y|\widetilde{\theta}^k)$

Obtenemos:

1. $(\widetilde{\theta}^k_{J+1})^S_{k=1}$
2. $(\widetilde{y}^k_{J+1})^S_{k=1}$

Lo que vuelve jerárquico al modelo es $p(\phi)$.

Aprender sobre la distribución de los hiperparámetros, ϕ, nos permite aprender sobre la distribución poblacional de $\theta|\phi$

Nos sirve para la inferencia posterior de $(\theta_1, \ldots, \theta_J)$, pero más importante, nos permite hacer inferencia de nuevos $\widetilde{\theta}$

Compartir información entre escuelas

$$y_j \sim N(\cdot, \sigma_j^2)$$

- En cada escuela hay suficientes datos para trabajar bajo *normalidad*
- Los parámetros σ_j son *conocidos*
- Cómo estimamos los efectos por escuela?

Consideremos 3 modelos:

- No se comparte información: las escuelas son independientes
- Se comparte toda la información: las escuelas son iguales
- Se comparte parte de la información: escuelas intercambiales, son distintas pero comparte la distributión de probabilidad.

No compartir información / *shrinkage* nulo

- Cada escuela representa un experimento independiente

$$y_j \sim N(\mu_j, \sigma_j^2) \quad \mu_j \sim N(m, t^2) \text{ } m, t \text{ conocidos}$$

$$\mu_j | D \sim N(m_1, t_1^2) \quad \frac{1}{t_1^2} = \frac{1}{t^2} + \frac{1}{\sigma_j^2} \quad m_1 = \frac{\frac{1}{t^2}}{\frac{1}{t^2} + \frac{1}{\sigma_j^2}} m + \frac{\frac{1}{\sigma_j^2}}{\frac{1}{t^2} + \frac{1}{\sigma_j^2}} y_j$$

- Sólo los datos de la escuela j tienen información sobre μ_j

$$E(\mu_j | D) = m_1 = \gamma_j m + (1 - \gamma_j) y_j$$

Igual efecto en todas las escuelas $\mu_j = \mu \; \forall \; j$.

$$y_j \sim N(\mu, \sigma_j^2) \quad \mu \sim N(m, t^2) \; m, t \text{ conocidos}$$

$$\mu | D \sim N(m_1, t_1^2) \quad \frac{1}{t_1^2} = \frac{1}{t^2} + \frac{1}{\sum \sigma_j^2} \quad m_1 = \frac{\frac{1}{t^2}}{\frac{1}{t^2} + \frac{1}{\sum \sigma_j^2}} m + \frac{\frac{1}{\sum \sigma_j^2} y_j}{\frac{1}{t^2} + \frac{1}{\sum \sigma_j^2}}$$

Todos los datos observados tienen información sobre θ, el único parámetro del modelo.

Escuelas intercambiables: μ_j son intercambiables, entonces existe una variable aleatoria, ϕ, tal que los μ_j son i.i.d condicional en ϕ

$$\begin{aligned} \mu_j &\overset{ind}{\sim} a(\mu, \tau^2) \\ \phi = (\mu, \tau) &\sim p(\mu, \tau) \end{aligned}$$

- El *shrinkage* parcial implica un *modelo jerárquico*
- En la posterior los μ_j NO son independientes, la inferencia de un μ_j depende de los demás μs, a través de su dependencia de μ y τ.

$$E(\mu_j|D) = E[E(\mu_j|D, \mu, \tau)] = \int \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma_j^2}} \mu + \frac{\frac{1}{\sigma_j^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma_j^2}} y_j p(\mu, \tau|D) d\mu d\tau$$

Modelo completo

Modelo:

$$
\begin{aligned}
y_j &\overset{ind}{\sim} Normal(\mu_j, \sigma_j^2) \\
\mu_j &\overset{ind}{\sim} Normal(\mu, \tau^2) \\
p(\mu) &\propto 1 \\
\tau & \quad Normal(0, 15^2)
\end{aligned}
$$

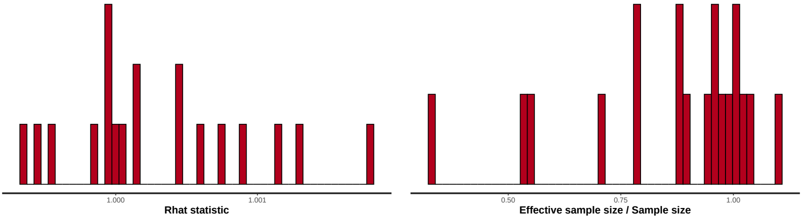
```
data {
  int<lower=0> J; // cantidad de escuelas
  real y[J]; // efectos individuales
  real<lower=0> sigma[J]; // se de los efectos
}
parameters {
  real mu; // media poblacional del programa
  real<lower=0> tau; // variabilidad entre escuelas
  vector[J] eta; // error a nivel de escuela
}
transformed parameters {
  vector[J] muj; // efectos estimados
  muj <- mu + tau*eta;
}
model {
  tau ~ normal(0, 15);
  eta ~ normal(0, 1); // muj ~ normal(mu, tau)
  y ~ normal(muj, sigma);
}
```

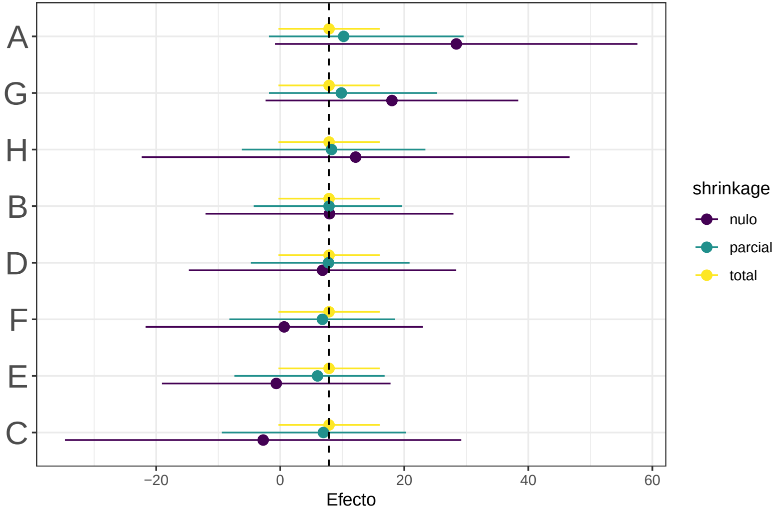
```
res <- stan(file = "../rcode/escuela_taunormal.stan",
            data = list(J=nrow(dt), y = dt$effect, sigma=dt$see))

## Warning: There were 1 divergent transitions after warmup. See
##
https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warm
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling
problems
```

```r
plot(res, plotfun='rhat', bins=50)
plot(res, plotfun='ess', bins=50)
```




A
G
H
B
D
F
E
C
−20
0
20
40
60
Efecto
shrinkage
nulo
parcial
total

- En el modelo normal, la previa IG es conjugada para varianzas
- $\tau \sim IG(\epsilon, \epsilon)$ puede ser muy informativa
- *Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper).*
- *Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. arXiv preprint arXiv:1408.4050.*

```
## `stat_bin()` using `bins = 30`. Pick better value with
`binwidth`.
```