Universidad Nacional Autónoma de México

Proyecto Final

Inferencia Bayesiana

Ángel Fernando Escalante López
Shadanna Ortega Hernández

Dr. Eduardo Gutiérrez Peña

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Índice

Introducción

En la actualidad, una de las principales causas de muerte es por la enfermedad conocida como cáncer. El cáncer comienza en una célula normal que cambia a una célula neoplásica[1] a través de varias mutaciones en varios genes a lo largo de mucho tiempo, podrían ser años, de estar expuesto a un agente carcinogenético [2]. No obstante, las mutaciones inducidas por los carcinógenos no son la única vía que afecta a la célula, sino que a lo largo de cada división celular se producen errores espontáneos en cada duplicación y los mismos se van acumulando constituyendo un factor intrínseco de riesgo (Martín de Civetta y Civetta, 2011). Por lo cual, es de suma importancia estudiar la cura para esta enfermedad.

En este contexto, el presente trabajo analizará desde una perspectiva bayesiana de un experimento de un tipo de tumor en un grupo de ratas, dadas diferentes dosis de una droga. En otras palabras, estudiar la tasa a la que el riesgo de tumor crece o decrece como función de la dosis.

Para ello, se examinarán tres perspectivas de acuerdo al tipo de información inicial, después se hará la una comparación entre modelos y por último unos comentarios finales.

Ejercicio 1:

Con el propósito de estudiar la relación entre la dosis y la respuesta, se tienen los datos del experimento en el cuadro 1, donde x representa el nivel de la dosis, mientras que n_x y y_x denotan respectivamente, el número de ratas tratadas y el número de ratas que presentan tumor en cada nivel $(x = 0, 1, 2)$.

x	n_x	y_x
0	14	4
1	34	4
2	34	2

Tabla 1: Datos

Sea π_x la probabilidad de que una rata en el grupo x desarrolle un tumor. Entonces, se considera el modelo

$$Y_x \sim Bin(\pi_x, n_x) \qquad (x = 0, 1, 2).$$

Dado que las investigadoras están interesadas en la forma como varía π_x en función de la dosis x, propusieron el modelo

$$logit(\pi_x) = \alpha + \beta x \qquad (x = 0, 1, 2).$$

El parámetro de interés para las investigadoras es la pendiente β, pero no cuentan con información inicial sobre su valor.

Entonces, se realizará un resumen de la distribución final de β suponiendo una distribución inicial no informativa en la que α y β se asumen independientes, con $\alpha \sim N(0, 1000)$ y $\beta \sim N(0, 1000)$; esto es, con media 0 y varianza 1000.

[1] Célula con una multiplicación o crecimiento anormal en un tejido del organismo.
[2] Agente capaz de causar cáncer.

De forma analítica el modelado de la relación dosis-respuesta sería mediante un modelo lineal generalizado, en específico, un modelo de regresión logística. Sin embargo, en este trabajo se utilizarán métodos de simulación[3], en particular, Métodos Monte Carlo de Cadenas de Markov (MCMC)[4]. Por lo cual, se realizan un número considerable de simulaciones, así como su diagnóstico de la convergencia de los valores obtenidos. Lo anterior, para asegurar que la inferencia se realice sobre simulaciones que son representativas de la distribución de interés. Por último, el modelado se realizará con *Stan*.

Definiendo modelo con `Stan`

Considerando el contexto del problema, se selecciona un modelo binomial. Entonces, la declaración del modelo binomial se hace a través de la función `binomial_logit()` de `Stan`, que recibe como segundo parámetro la inversa de la función logit[5], además se usa $\sigma = \sqrt{\sigma^2} = \sqrt{1000} = 31.62278$ ya que `Stan` recibe desviación estándar como parámetro y no la varianza.

```
model_string <-
"
data {
  int<lower=0> N;
  int<lower=0> n[N];
  int<lower=0> y[N];
  vector[N] x;
}

parameters {
  real alpha;
  real beta;
}

model {
  alpha ~ normal(0, 31.62278);
  beta ~ normal(0, 31.62278);

  for (i in 1:N) {
    y[i] ~ binomial_logit(n[i], alpha + beta * x[i]);
  }
}
"
```

Muestreando el modelo con datos del estudio

Para este modelo, se utilizaron los siguientes parámetros en `Stan`:

[3]Los métodos de simulación se refieren a la obtención de pseudo-muestras que se originan de una distribución de probabilidad en una computadora, también conocidos como Métodos de Monte Carlo, pues e introduce un nivel de aleatoredad en el análisis (Bravo *et. al.*, 2008).

[4]Los Métodos Monte Carlo de Cadenas de Markov tienen como objetivo encontrar una cadena de Markov en el espacio de parámetros, de manera tal que la distribución de equilibrio o estacionaria e la cadena coincida con la distribución posterior (Bravo *et. al.*, 2008).

[5]Stan User Guide

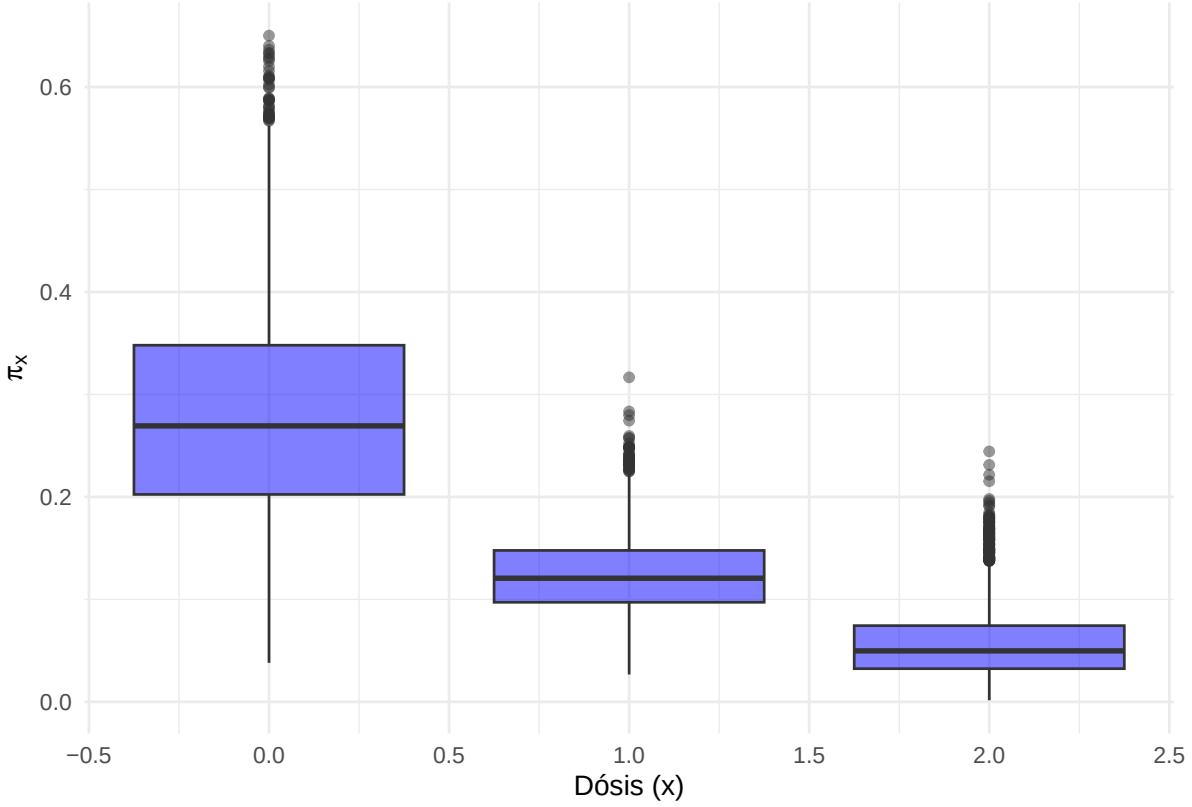
- Número de iteraciones = 5000
- Warmup (calentamiento) = 2000
- Thin = 3
- Número de cadenas = 4

	mean	*se_mean*	*sd*	*2.5 %*	*25 %*	*50 %*	*75 %*	*97.5 %*	*n_eff*	*Rhat*
alpha	-1.01	0.01	0.56	-2.16	-1.37	-1	-0.63	0.03	2604	1
beta	-0.99	0.01	0.5	-1.98	-1.31	-0.98	-0.64	-0.06	2582	1

Tabla 2: Ajuste del estudio 1

Diagnóstico del modelo

Con base en la tabla anterior, se puede ver que el **Rhat**[6] es exactamente 1 para las cadenas de α y β, lo cuál representaría convergencia para ambos parámetros. De igual manera, esto se aprecia a través de los gráficos de las trazas:



[6] $\hat{R}$ estima la reducción de la escala potencial. Es decir, para monitorear la convergencia en los algoritmos de simulación MCMC se realiza mediante la estimación de un factor por el cual la escala de la distribución actual del parámetro ψ puede ser reducida suponiendo que se continuan las simulaciones en el límite $n \to \infty$, lo cual va a 1 si $n \to \infty$ (Bravo *et. al.*, 2008).

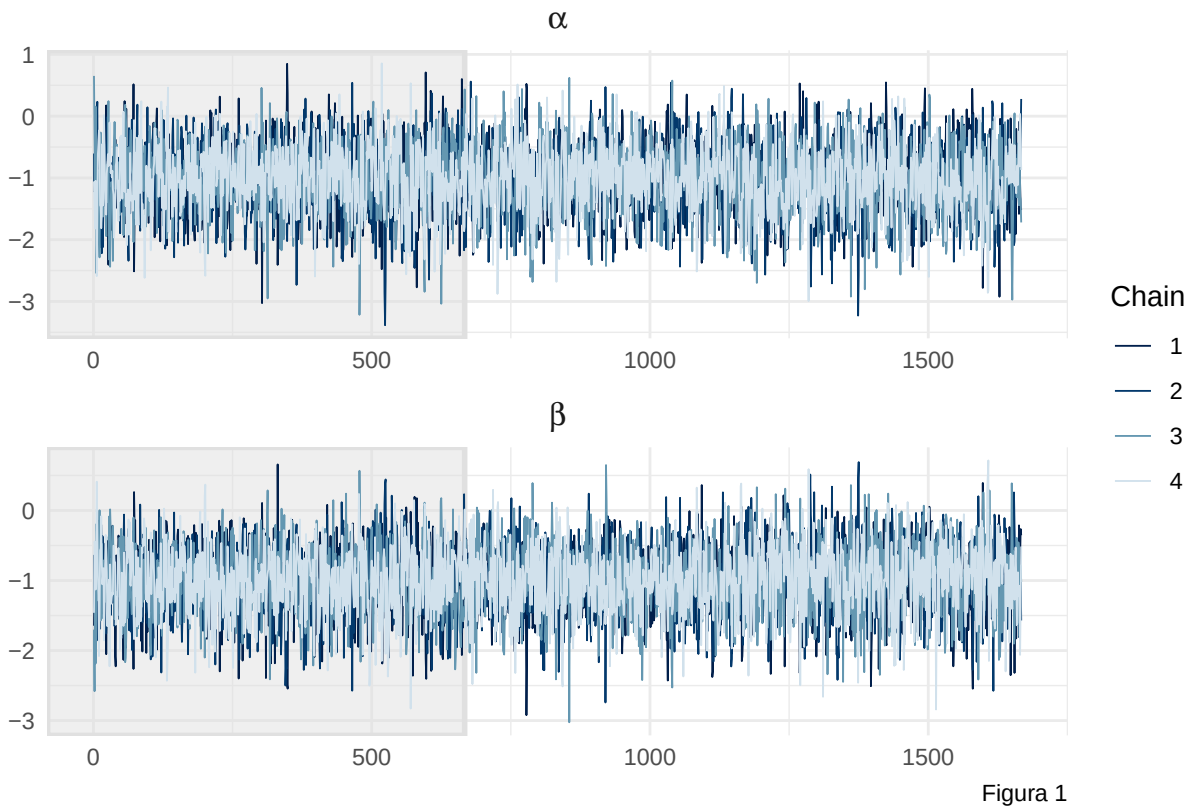


Figura 1

Figura 2

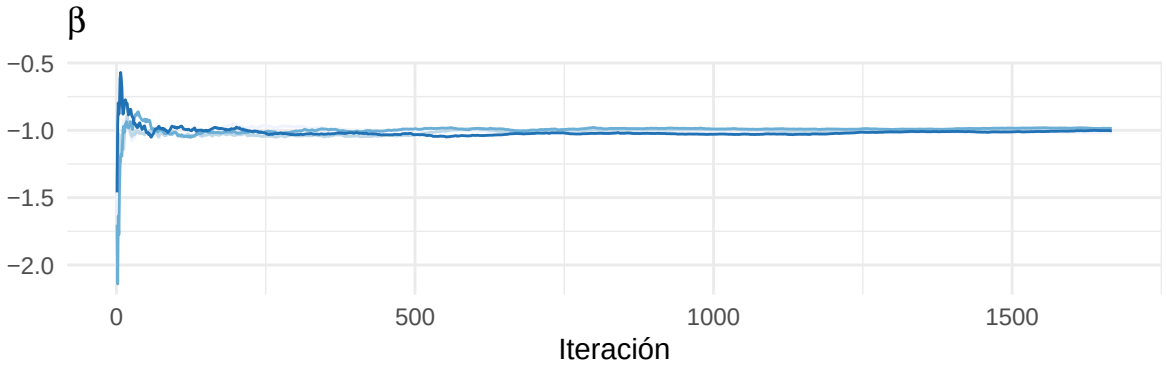


Figura 3

En la figura 1, el área sombreada representa las muestras del calentamiento ($n = 667$), pues de esta manera se pueden omitir los primeros valores de las cadenas para que salgan de una

primera fase de inestabilidad. Después de ese periodo, e incluso un poco antes, se puede ver el comportamiento estacionario. Este comportamiento se esclarece con los promedios ergódigos (veáse figuras 2 y 3).

Posterior de α y β

En este contexto, en la figura 4, visualizamos las muestras de las posteriores de α y β del modelo.

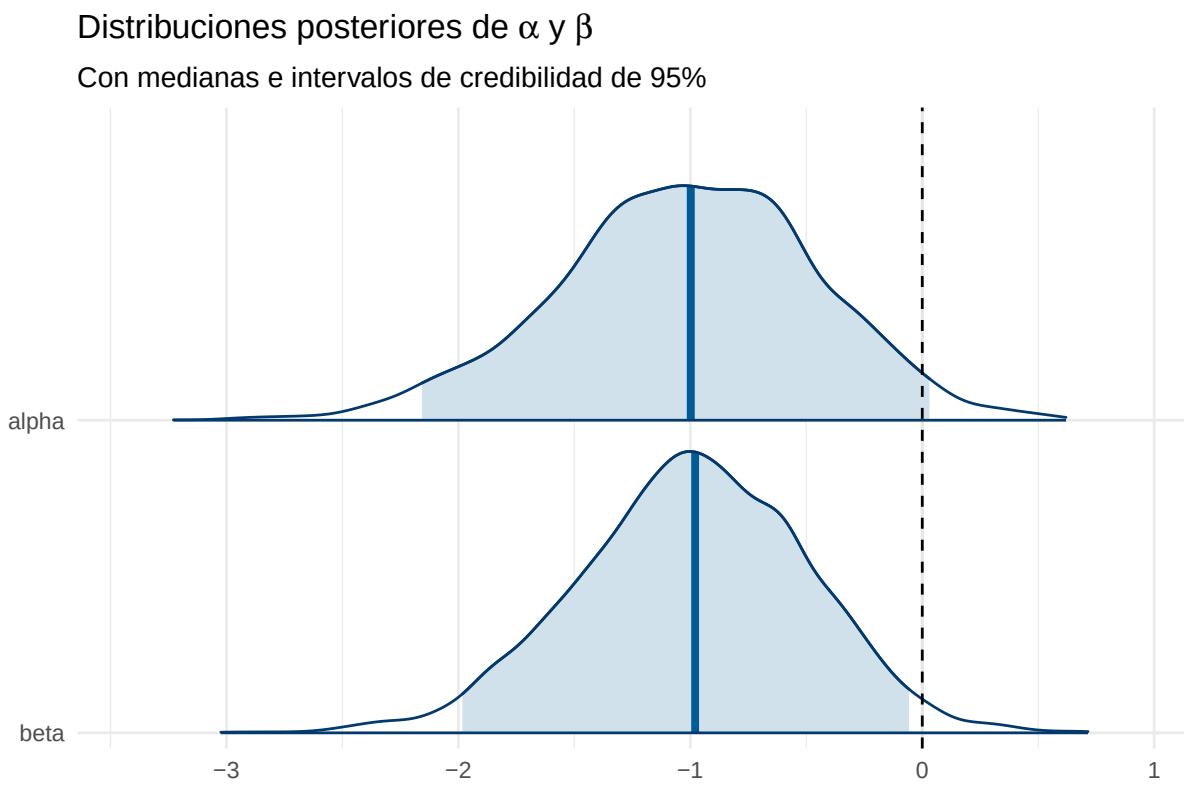


Figura 4

Intervalo de credibilidad

De la gráfica anterior, es fácil ver que la media de la distribución final de β es muy cercana a -1. En específico, la estimación puntual de la esperanza es:

$$E[\beta|y] \approx -0.99$$

El intervalo de credibilidad del 95 % de β sería de:

$$p(1.98 \leq \beta \leq -0.057|y) \approx 0.95$$

Planteamiento de hipótesis

En particular, para la β notamos que el intervalo de credibilidad del 95 % se encuentra del lado izquierdo del cero, lo que significa que $p(\beta < 0|y) > 0.95$. Si plantearamos una hipótesis sobre la β diríamos que tenemos evidencia para afirmar que hay un efecto negativo entre los niveles de dósis y la probabilidad de que la rata desarrolle un tumor, de hecho, la probabilidad de que la pendiente β sea negativa sería $p(\beta < 0|y) \approx 0.98$.

En un planteamiento de hipótesis como el siguiente

$$H_0 : \beta < 0 \quad \text{vs} \quad H_a : \beta \geq 0$$

la hipótesis H_0 tiene mayor plausibilidad.

Ejercicio 2

Dado que el tamaño de las muestras en el problema anterior es muy pequeño, y en vista de la falta de información inicial, las investigadoras se dieron a la tarea de buscar información relevante en la literatura. Como producto de esta labor, encontraron datos de 10 estudios similares con ratas de la misma cepa. Desafortunadamente, todos estos datos correspondían a *controles*; es decir, ratas a las que no se les aplicó la droga. Los datos se presentan en la Tabla 2a. Aquí $n_{0,i}$ y $y_{0,i}$ denotan, respectivamente, el número total de ratas y el número de ratas que presentaron un tumor en el i-ésimo estudio ($i = 1, 2, ..., 10$).

$Estudio \quad i$	$n_{0,i}$	$y_{0,i}$
1	10	1
2	13	2
3	48	10
4	19	5
5	20	0
6	18	0
7	25	2
8	49	5
9	48	9
10	19	4

Tabla 3: Tabla 2a

No satisfechas con estos datos, las investigadoras siguieron buscando trabajos recientes (no publicados). Finalmente encontraron dos reportes muy relevantes, de donde extrajeron los siguientes datos:

x	$n_{x,11}$	$y_{x,11}$
0	7	3
1	16	5
2	18	2

Tabla 4: Tabla 2b

En vista de que para los datos de la Tabla 2a sólo se recabó información de controles, el nivel de la dosis es $x = 0$ en todos esos casos. Por lo tanto el modelo que propusieron para esos datos es

$$Y_x \sim Bin(\pi_{0,1}, n_{0,i}), \qquad i = 1, 2, ..., 10,$$

donde

x	$n_{x,12}$	$y_{x,12}$
0	5	2
1	11	1
2	9	0

Tabla 5: Tabla 2c

$$logit(\pi_{0,i}) = \alpha_i, \qquad i = 1, 2, ..., 10.$$

Por otra parte, para los datos de las Tablas 2b y 2c (Estudios 11 y 12), las investigadoras supusieron un modelo de la misma forma que el del problema 1, es decir:

$$Y_x \sim Bin(\pi_{x,i}, n_{x,i}) \qquad (x = 0, 1, 2); \qquad i = 11, 12,$$

con

$$logit(\pi_x) = \alpha_i + \beta_i x \qquad (x = 0, 1, 2); \qquad i = 11, 12.$$

Para simplificar el análisis en esta etapa, las investigadoras decidieron considerar todos estos estudios suficientemenete similares como para suponer que los datos de las Tablas 1,2a,2b y 2c *provienen de un solo experimento*, de manera que $\alpha_1 = \alpha_2 = ... = \alpha_{12} = \alpha$ y $\beta_{11} = \beta_{12} = \beta$.

Utilizando la misma distribución inicial que en el Ejercicio 1, se proporciona un resumen de la distribución final de β.

Para abordar este problema, vamos a combinar los datos de todos los estudios (Tablas 1, 2a, 2b y 2c) en un solo modelo y realizar la inferencia bayesiana utilizando Stan. Mantendremos la misma distribución inicial no informativa para α y β, y modelaremos los datos de manera que α y β sean comumes a todos los estudios.

Definiendo nuevos modelos en `Stan`

Primero, definimos todos los datos del experimento tal como se presentan en las tablas.

```
model_string_2 <- "
data {
  int<lower=0> N1;
  int<lower=0> n1[N1];
  int<lower=0> y1[N1];
  vector[N1] x1;

  int<lower=0> N0;
  int<lower=0> n0[N0];
  int<lower=0> y0[N0];

  int<lower=0> N11;
  int<lower=0> n11[N11];
```

```
  int<lower=0> y11[N11];
  vector[N11] x11;

  int<lower=0> N12;
  int<lower=0> n12[N12];
  int<lower=0> y12[N12];
  vector[N12] x12;
}

parameters {
  real alpha;
  real beta;
}

model {
  alpha ~ normal(0, 31.62278);
  beta ~ normal(0, 31.62278);

  for (i in 1:N1) {
    y1[i] ~ binomial_logit(n1[i], alpha + beta * x1[i]);
  }

  for (i in 1:N0) {
    y0[i] ~ binomial_logit(n0[i], alpha);
  }

  for (i in 1:N11) {
    y11[i] ~ binomial_logit(n11[i], alpha + beta * x11[i]);
  }

  for (i in 1:N12) {
    y12[i] ~ binomial_logit(n12[i], alpha + beta * x12[i]);
  }
}
"
```

Aquí, al igual que en el modelo anterior, definimos las distribuciones a priori de los parámetros a estimar: $\alpha \sim N(0, 31.62278)$ y $\beta \sim N(0, 31.62278)$, donde $\sigma = \sqrt{\sigma^2} = \sqrt{1000} = 31.62278$ ya que `Stan` recibe desviación estándar como parámetro y no la varianza. Además, definimos 4 verosimilitudes distintas para:

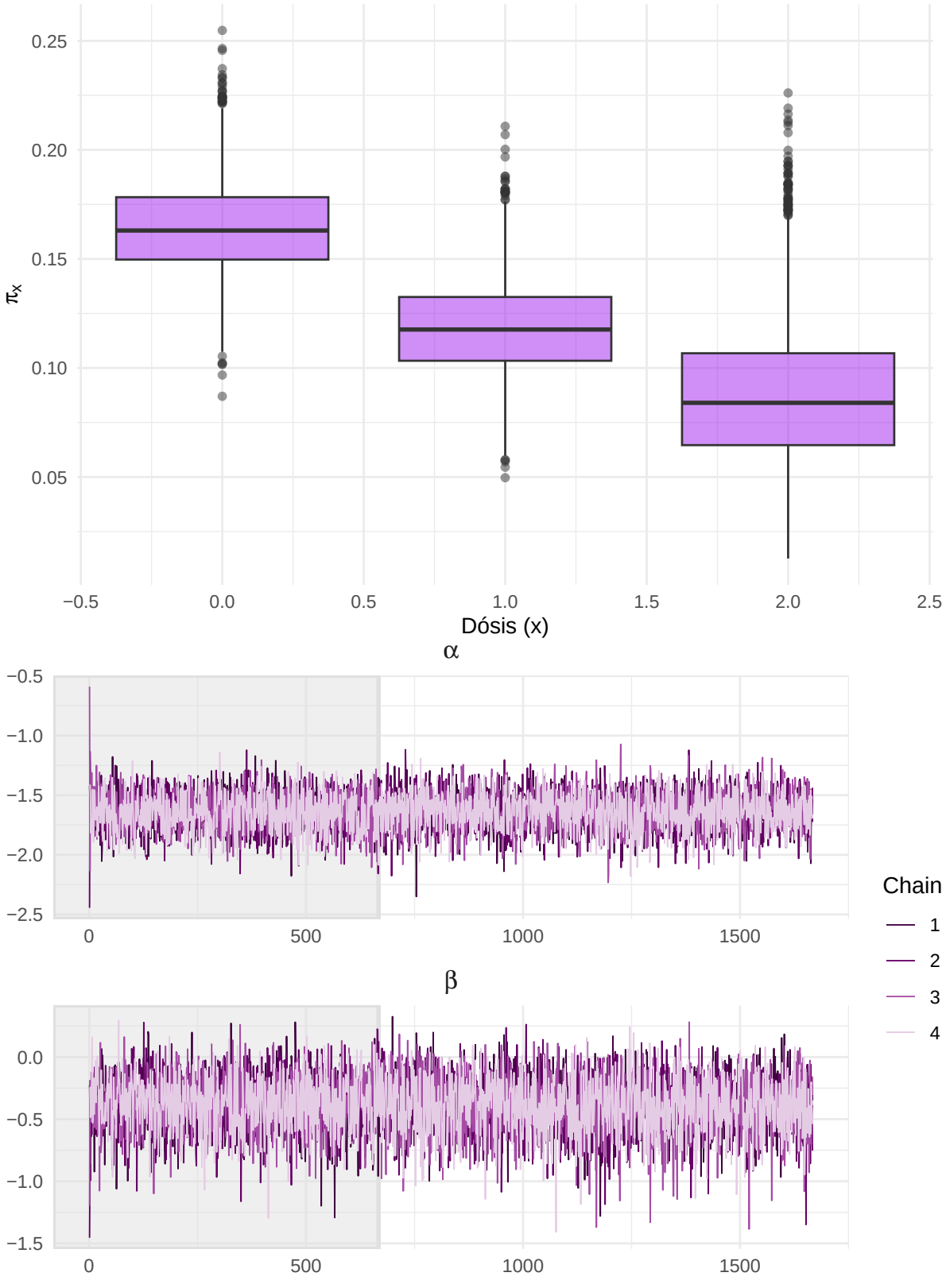
- Datos originales: `y1[i] ~ binomial_logit(n1[i], alpha + beta * x1[i]);`
- Estudios de controles: `y0[i] ~ binomial_logit(n0[i], alpha);`
- Estudio 11: `y11[i] ~ binomial_logit(n11[i], alpha + beta * x11[i]);`
- Estuio 12: `y12[i] ~ binomial_logit(n12[i], alpha + beta * x12[i]);`

Estos son usados en el modelo para estimar las distribuciones de α y β.

Ajustando el modelo con los datos

```
## Inference for Stan model: anon_model.
## 4 chains, each with iter=5000; warmup=2000; thin=3;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##        mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## alpha -1.64       0 0.16 -1.95 -1.74 -1.64 -1.53 -1.34  3369    1
## beta  -0.39       0 0.23 -0.86 -0.53 -0.38 -0.23  0.02  3308    1
##
## Samples were drawn using NUTS(diag_e) at Thu May 30 23:21:01 2024.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Diagnóstico del modelo



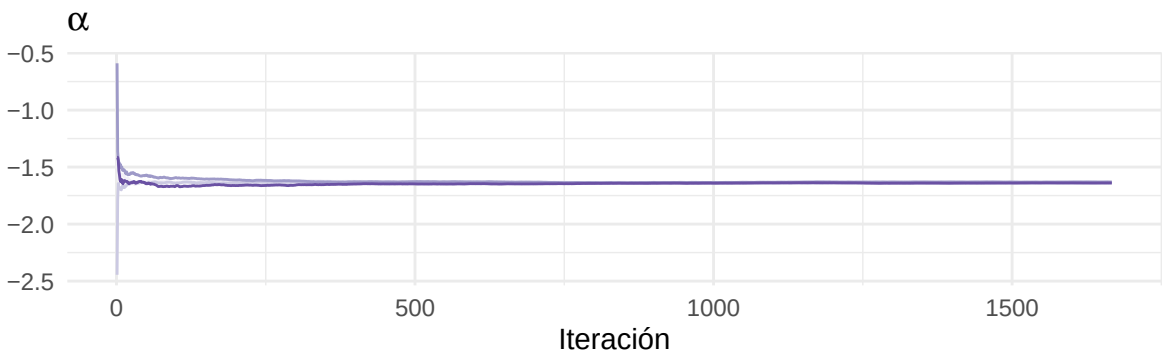


Figura 2

Figura 3

Posterior de α y β

A continuación, visualizamos los muestras de las posteriores de α y β con el modelo bayesiano incluyendo los 12 estudios adicionales. Se aprecia que α se diferencia de β volviéndose más negativo que en el resultado de `Stan` anterior, siendo aquí $E[\alpha|x] \approx -1.635$

Intervalo de credibilidad

De la gráfica anterior, es fácil ver que la media de la distribución final de β es muy cercana a -1. En específico, la estimación puntual de la esperanza es:

$$E[\beta|y] \approx -0.39$$

De la gráfica anterior es fácil calcular el intervalo de credibilidad o high density interval (HDI por sus siglas en inglés).

$$P[-0.859 \leq \beta \leq 0.0226|y] \approx 0.95$$

Planteamiento de hipótesis

Podemos plantear una hipótesis sobre la probabilidad de β como lo hicimos en la pregunta anterior. Si las investigadoras se plantearan si existe una relación negativa entre el nivel de dosis y la probabilidad de presentar el tumor para una rata, tendrían que calcular:

$$H_0 : \beta < 0 \quad \text{vs} \quad H_a : \beta \geq 0$$

desde el enfoque bayesiano, esto equivaldría a calcular la plausibilidad de que la β sea negativa.

Con este modelo propuesto, observamos que: $p(\beta < 0) \approx 0.9665$

Ejercicio 3: Modelo Jerárquico

Poco tiempo después, una de las investigadoras tuvo la oportunidad de asistir a un curso de Análisis Bayesiano de Modelos Jerárquicos y convenció al resto del equipo de que ésa es la manera más apropiada de analizar los datos con los que contaban. Específicamente, dado que todos los estudios eran similares, consideraron que podían utilizar los 12 estudios que encontraron en la literatura para complementar la información de su experimento original (ver Tabla 1).

Las investigadoras supusieron entonces que los parámetros $\alpha_1, \alpha_2, ..., \alpha_{12}$ eran intercambiables, con distribución común $N(\alpha^*, \sigma_\alpha^2)$, y también que los parámetros $\beta, \beta_{11}, \beta_{12}$ eran intercambiables con distribución común $N(\beta^*, \sigma_\beta^2)$. Finalmente, tanto para α^* como para β^* supusieron una distribución $N(0, 100)$, mientras que para $\tau_\alpha = 1/\sigma_\alpha^2$ y $\tau_\beta = 1/\sigma_\beta^2$ consideraron una distribución $Gamma(0.01, 0.01)$.

Se propone un resumen de la distribución final de β (la correspondiente al Ejercicio 1) bajo estas condiciones.

Definiendo el modelo en `Stan`

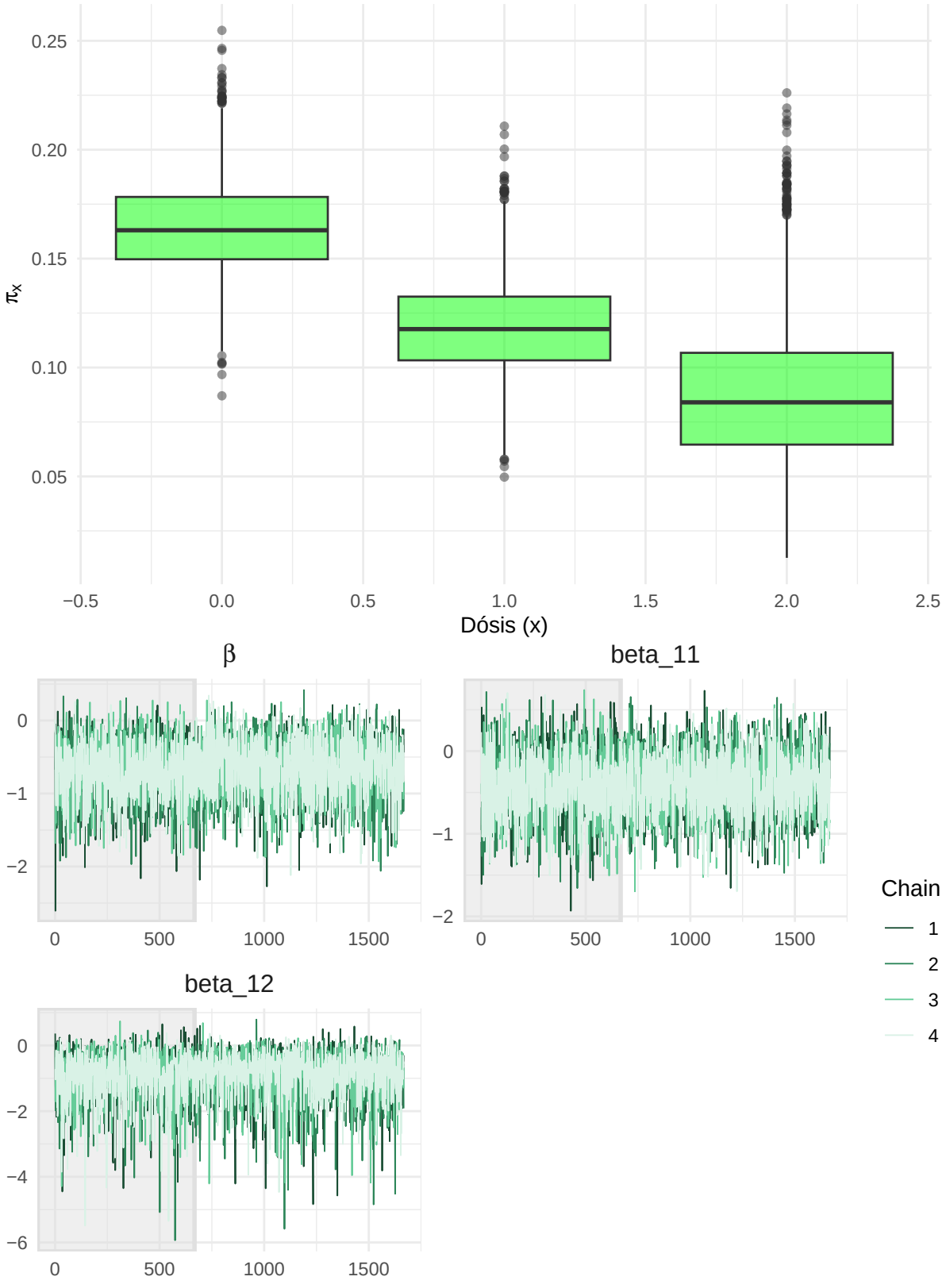
```
model_string_3 <- "
data {
  int<lower=0> N1;
  int<lower=0> n1[N1];
  int<lower=0> y1[N1];
  vector[N1] x1;

  int<lower=0> N0;
  int<lower=0> n0[N0];
  int<lower=0> y0[N0];

  int<lower=0> N11;
  int<lower=0> n11[N11];
  int<lower=0> y11[N11];
  vector[N11] x11;

  int<lower=0> N12;
  int<lower=0> n12[N12];
  int<lower=0> y12[N12];
  vector[N12] x12;
}

parameters {
  real alpha_estrella;
  real<lower=0> tau_alpha;
  vector[12] alpha;

  real beta_estrella;
  real<lower=0> tau_beta;
  real beta;
  real beta_11;
  real beta_12;
}

transformed parameters {
  real sigma_alpha = 1 / sqrt(tau_alpha);
  real sigma_beta = 1 / sqrt(tau_beta);
}

model {
  alpha_estrella ~ normal(0, 10);
  beta_estrella ~ normal(0, 10);
  tau_alpha ~ gamma(0.01, 0.01);
  tau_beta ~ gamma(0.01, 0.01);

  alpha ~ normal(alpha_estrella, sigma_alpha);
```

```
  beta ~ normal(beta_estrella, sigma_beta);
  beta_11 ~ normal(beta_estrella, sigma_beta);
  beta_12 ~ normal(beta_estrella, sigma_beta);

  for (i in 1:N1) {
    y1[i] ~ binomial_logit(n1[i], alpha[1] + beta * x1[i]);
  }

  for (i in 1:N0) {
    y0[i] ~ binomial_logit(n0[i], alpha[i+1]);
  }

  for (i in 1:N11) {
    y11[i] ~ binomial_logit(n11[i], alpha[11] + beta_11 * x11[i]);
  }

  for (i in 1:N12) {
    y12[i] ~ binomial_logit(n12[i], alpha[12] + beta_12 * x12[i]);
  }
}
"
```

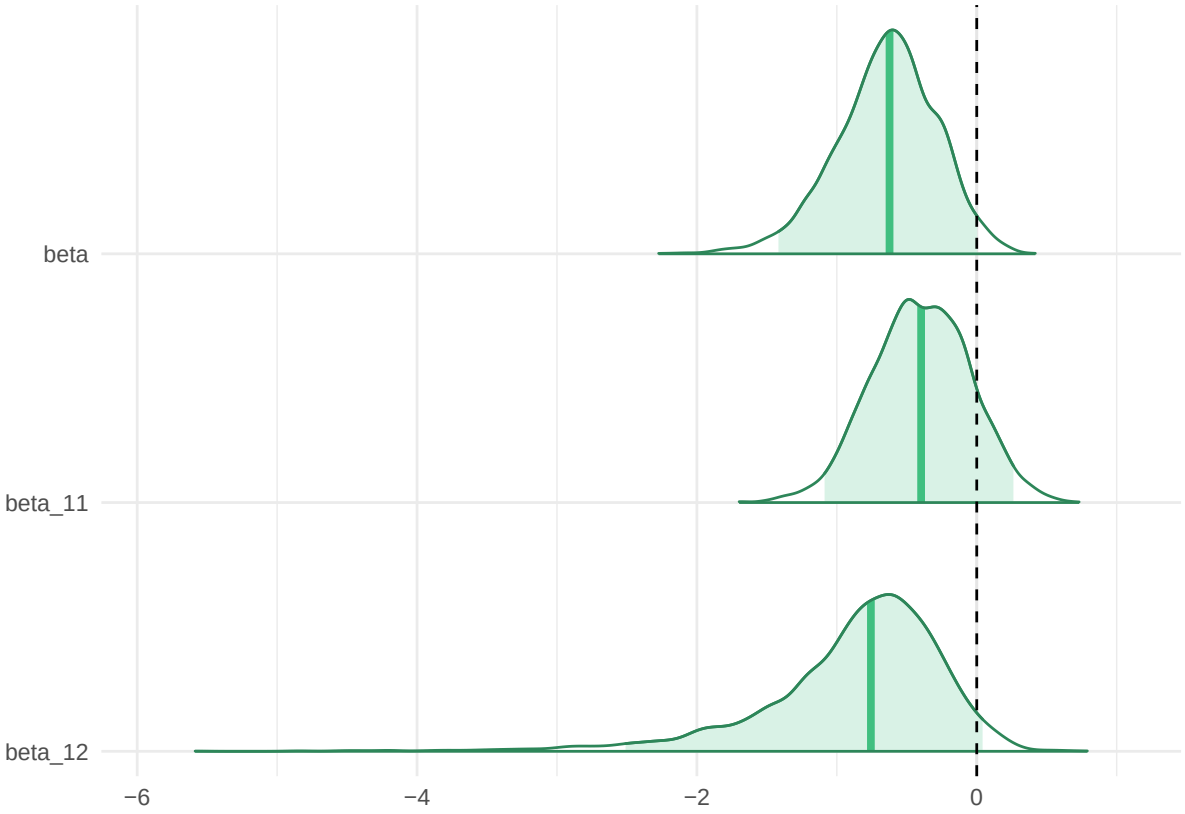
```
## Warning: There were 89 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

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

Diagnósticos del modelo



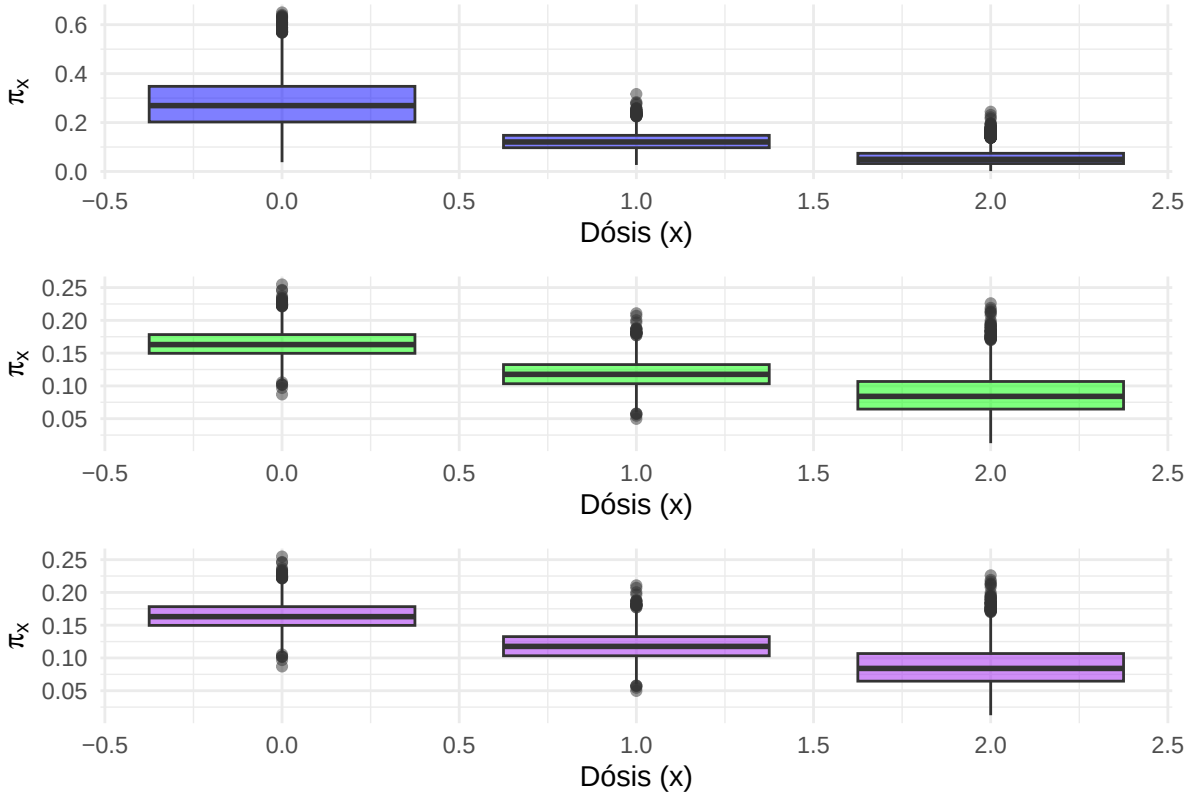
Posteriores de las β's



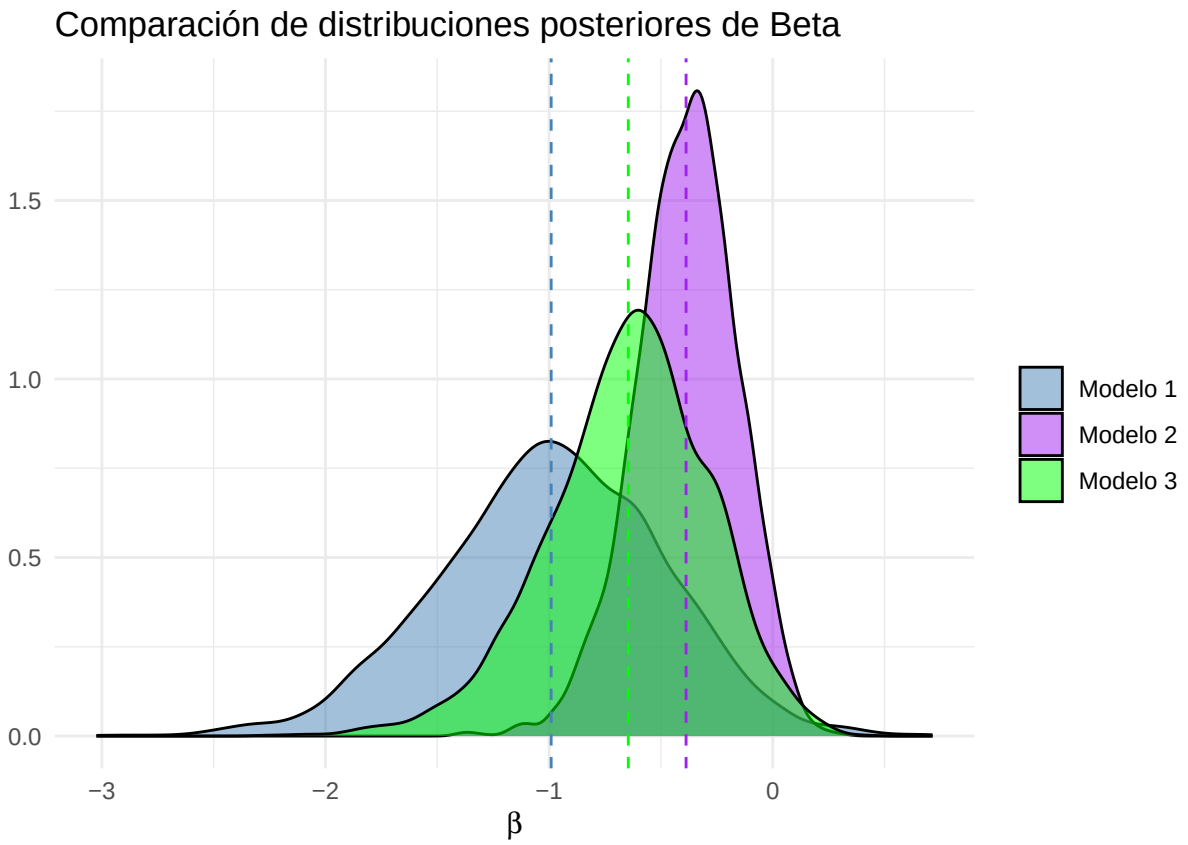
Ejercicio 4: Modelos

Comparación de modelos

Boxplots del ajuste del modelo (curva de predicción).



En comparativa, notamos que la distribución posterior de la β del modelo 2 tiene menor incertidumbre dado que estamos considerando la información de los otros estudios de manera conjunta, es decir, asumimos intercambiabilidad en toda la información. La distribución de la β del modelo 1 tiene mayor dispersión dado que hay menor información para el estudio orginal, a comparación del modelo 2. Además, esta beta tiene la media más negativa de los tres modelos. Para el modelo 3, notamos que, si bien hay mayor dispersión que el modelo 2, su media es más grande que la media del modelo 1 y menor que la del modelo 2. Siendo así, una distribución "intermedia" entre el modelo original y el modelo combinado.



Discusión de resultados

Debemos notar que partimos de distibuciones iniciales no informativas, lo cuál se ve en los modelos con mayor información puesto que sus estimaciones tienen menor variabilidad, es decir, se nota el peso de la verosimilitud de los datos. Por otra parte, nos hace mayor sentido una estructura jerárquica que permita ponderar los resultados obtenidos por las investigadoras con los otros estudios (meta-análisis); la metodología para la obtención de los datos en los otros estudios no es del completo conocimiento de las investigadoras y por tanto, podría ser riesgoso asumir que los estudios son suficientemente similares como para condensar su información y sea proveniente de un sólo experimento.

Dicho lo anterior, sugeriríamos a las investigadoras adoptar el modelo jerarquico y buscar mayor información para la definición de las distribuciones a priori.

Reflexiones adicionales: Los métodos de Montecarlo de Cadenas de Markov facilitaron en gran medida la definición y obtención de las distribuciones posteriores de los parámetros de interés de las investigadoreas.

Comentarios finales

Referencias

- Martín de Civetta MT y Civetta JD.(2011). Carcinogénesis. Salúd Pública Mex;53:405-414.