Universidad Nacional Autónoma de México

Proyecto Final

Inferencia Bayesiana

Ángel Fernando Escalante López
Shadanna Ortega Hernández

Dr. Eduardo Gutiérrez Peña

Instituto de Investigaciones en Matem?ticas Aplicadas y en Sistemas

Índice

Introducción **2**

Pregunta 1: **2**
Respuesta . . . 3
Definiendo modelo con `Stan` . . . 3
Sampleando el modelo con datos del estudio . . . 3
Diagnóstico del modelo . . . 4
Posterior de α y β . . . 4

Ejercicio 2 **5**
Respuesta . . . 7
Definiendo nuevos modelos en `Stan` . . . 7
Ajustando el modelo con los datos . . . 8

Ejercicio 3: Modelo Jerárquico **14**

Ejercicio 4: Modelos **15**
Comparación de modelos . . . 15
Discusión de resultados . . . 15

Comentarios finales **16**

Referencias **17**

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.

Pregunta 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.

Respuesta

Definiendo modelo con `Stan`

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[3], 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]);
  }
}
"
```

Sampleando el modelo con datos del estudio

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

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

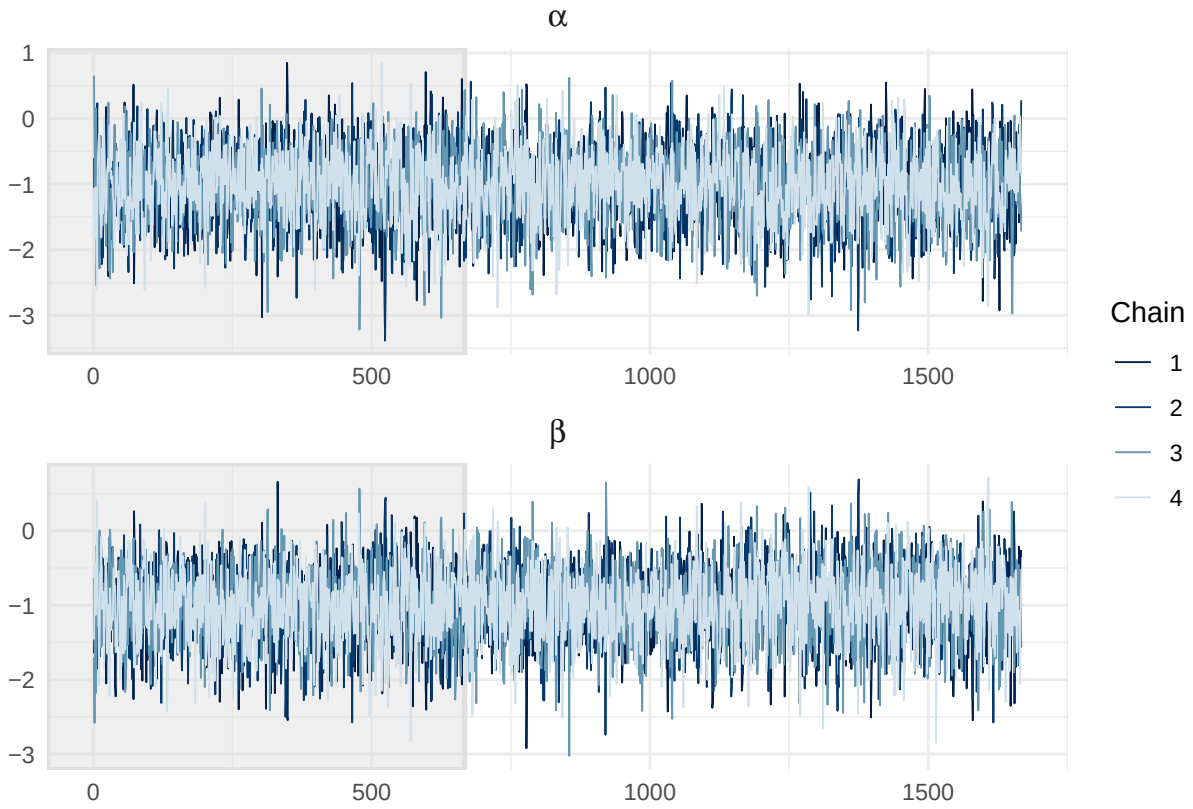
```
## Inference for Stan model: anon_model.
## 4 chains, each with iter=5000; warmup=2000; thin=3;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##        mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## alpha -1.01    0.01 0.56 -2.16 -1.37 -1.00 -0.63  0.03  2604    1
## beta  -0.99    0.01 0.50 -1.98 -1.31 -0.98 -0.64 -0.06  2582    1
```

[3] Stan User Guide

```
##
## Samples were drawn using NUTS(diag_e) at Sun May 26 14:07:01 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

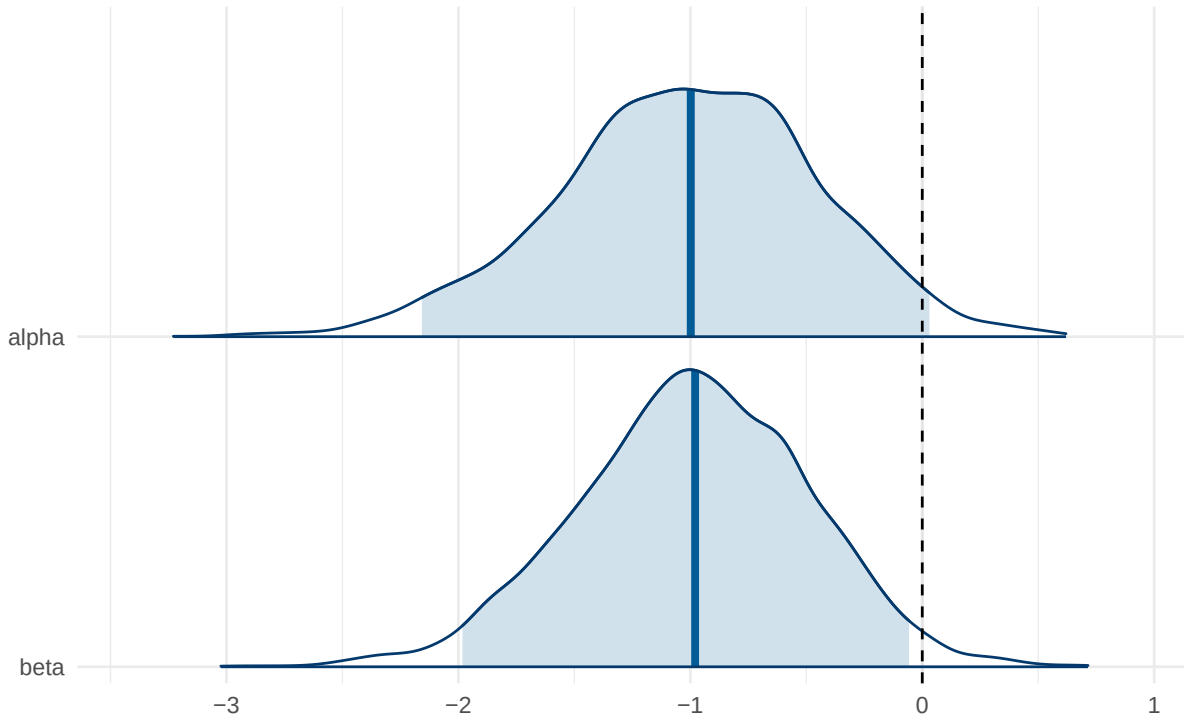
Notamos que el **Rhat** es exactamente 1 para las cadenas de α y β, lo cuál representaría convergencia para ambos parámetros. Esto se aprecia de igual manera en forma visual a través de los gráficos de las trazas:



aquí, el área sombreada representa los samplings del calentamiento ($n = 667$). Después de ese periodo, e incluso un poco antes, se aprecia el comportamiento estacionario.

Posterior de α y β

A continuación, visualizamos los muestras de las posteriores de α y β con el modelo bayesiano. En particular, para la β notamos que el intervalo de credibilidad del 95 % se encuentra del lado izquierdo del cero, lo que significa que $0.95 \leq p(\beta < 0)$. 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 $\beta \approx 0.98$



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 2: 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 3: Tabla 2b

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

Tabla 4: Tabla 2c

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

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

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.

Respuesta

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 β, y modelaremos los datos de manera que α sea común a todos los estudios, mientras que β es específico para los estudios que incluyen diferentes niveles de dosis (estudios 11 y 12).

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

```
# Preparar los datos para Stan
data_combinada <- list(
  N1 = length(x_1), n1 = n_x_1, y1 = y_x_1, x1 = x_1,
  N0 = length(n_0), n0 = n_0, y0 = y_0,
  N11 = length(x_11), n11 = n_x_11, y11 = y_x_11, x11 = x_11,
  N12 = length(x_12), n12 = n_x_12, y12 = y_x_12, x12 = x_12
)

# Compilar el modelo
modelo_combinado <- stan_model(model_code = model_string_2)
```

```
## Trying to compile a simple C file
```

```
## Running /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/bin/R \
##   CMD SHLIB foo.c
## clang -arch arm64 -I"/Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/include
## In file included from <built-in>:1:
## In file included from /Users/angelescalante/Library/R/arm64/4.2/library/StanHeaders/inclu
## In file included from /Users/angelescalante/Library/R/arm64/4.2/library/RcppEigen/include
## In file included from /Users/angelescalante/Library/R/arm64/4.2/library/RcppEigen/include
## /Users/angelescalante/Library/R/arm64/4.2/library/RcppEigen/include/Eigen/src/Core/util/M
## #include <cmath>
##          ^~~~~~~
```

```
## 1 error generated.
## make: *** [foo.o] Error 1
```

```
# Ajustar el modelo a los datos
fit_combinado <- sampling(
  object = modelo_combinado,
  data = data_combinada,
  iter = 5000,
  warmup = 2000,
  chains = 4,
  thin = 3,
  seed = 123
)
```

```
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 1.7e-05 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.17 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 5000 [  0%]  (Warmup)
## Chain 1: Iteration:  500 / 5000 [ 10%]  (Warmup)
## Chain 1: Iteration: 1000 / 5000 [ 20%]  (Warmup)
## Chain 1: Iteration: 1500 / 5000 [ 30%]  (Warmup)
## Chain 1: Iteration: 2000 / 5000 [ 40%]  (Warmup)
## Chain 1: Iteration: 2001 / 5000 [ 40%]  (Sampling)
## Chain 1: Iteration: 2500 / 5000 [ 50%]  (Sampling)
## Chain 1: Iteration: 3000 / 5000 [ 60%]  (Sampling)
## Chain 1: Iteration: 3500 / 5000 [ 70%]  (Sampling)
## Chain 1: Iteration: 4000 / 5000 [ 80%]  (Sampling)
## Chain 1: Iteration: 4500 / 5000 [ 90%]  (Sampling)
## Chain 1: Iteration: 5000 / 5000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 0.012 seconds (Warm-up)
## Chain 1:                0.018 seconds (Sampling)
## Chain 1:                0.03 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 1e-06 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.01 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 5000 [  0%]  (Warmup)
## Chain 2: Iteration:  500 / 5000 [ 10%]  (Warmup)
```

```
## Chain 2: Iteration: 1000 / 5000 [ 20%]  (Warmup)
## Chain 2: Iteration: 1500 / 5000 [ 30%]  (Warmup)
## Chain 2: Iteration: 2000 / 5000 [ 40%]  (Warmup)
## Chain 2: Iteration: 2001 / 5000 [ 40%]  (Sampling)
## Chain 2: Iteration: 2500 / 5000 [ 50%]  (Sampling)
## Chain 2: Iteration: 3000 / 5000 [ 60%]  (Sampling)
## Chain 2: Iteration: 3500 / 5000 [ 70%]  (Sampling)
## Chain 2: Iteration: 4000 / 5000 [ 80%]  (Sampling)
## Chain 2: Iteration: 4500 / 5000 [ 90%]  (Sampling)
## Chain 2: Iteration: 5000 / 5000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 0.013 seconds (Warm-up)
## Chain 2:                0.021 seconds (Sampling)
## Chain 2:                0.034 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 1e-06 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.01 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 5000 [  0%]  (Warmup)
## Chain 3: Iteration:  500 / 5000 [ 10%]  (Warmup)
## Chain 3: Iteration: 1000 / 5000 [ 20%]  (Warmup)
## Chain 3: Iteration: 1500 / 5000 [ 30%]  (Warmup)
## Chain 3: Iteration: 2000 / 5000 [ 40%]  (Warmup)
## Chain 3: Iteration: 2001 / 5000 [ 40%]  (Sampling)
## Chain 3: Iteration: 2500 / 5000 [ 50%]  (Sampling)
## Chain 3: Iteration: 3000 / 5000 [ 60%]  (Sampling)
## Chain 3: Iteration: 3500 / 5000 [ 70%]  (Sampling)
## Chain 3: Iteration: 4000 / 5000 [ 80%]  (Sampling)
## Chain 3: Iteration: 4500 / 5000 [ 90%]  (Sampling)
## Chain 3: Iteration: 5000 / 5000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 0.013 seconds (Warm-up)
## Chain 3:                0.02 seconds (Sampling)
## Chain 3:                0.033 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 1e-06 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.01 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
```

```
## Chain 4: Iteration:    1 / 5000 [  0%]  (Warmup)
## Chain 4: Iteration:  500 / 5000 [ 10%]  (Warmup)
## Chain 4: Iteration: 1000 / 5000 [ 20%]  (Warmup)
## Chain 4: Iteration: 1500 / 5000 [ 30%]  (Warmup)
## Chain 4: Iteration: 2000 / 5000 [ 40%]  (Warmup)
## Chain 4: Iteration: 2001 / 5000 [ 40%]  (Sampling)
## Chain 4: Iteration: 2500 / 5000 [ 50%]  (Sampling)
## Chain 4: Iteration: 3000 / 5000 [ 60%]  (Sampling)
## Chain 4: Iteration: 3500 / 5000 [ 70%]  (Sampling)
## Chain 4: Iteration: 4000 / 5000 [ 80%]  (Sampling)
## Chain 4: Iteration: 4500 / 5000 [ 90%]  (Sampling)
## Chain 4: Iteration: 5000 / 5000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 0.013 seconds (Warm-up)
## Chain 4:                0.021 seconds (Sampling)
## Chain 4:                0.034 seconds (Total)
## Chain 4:
```

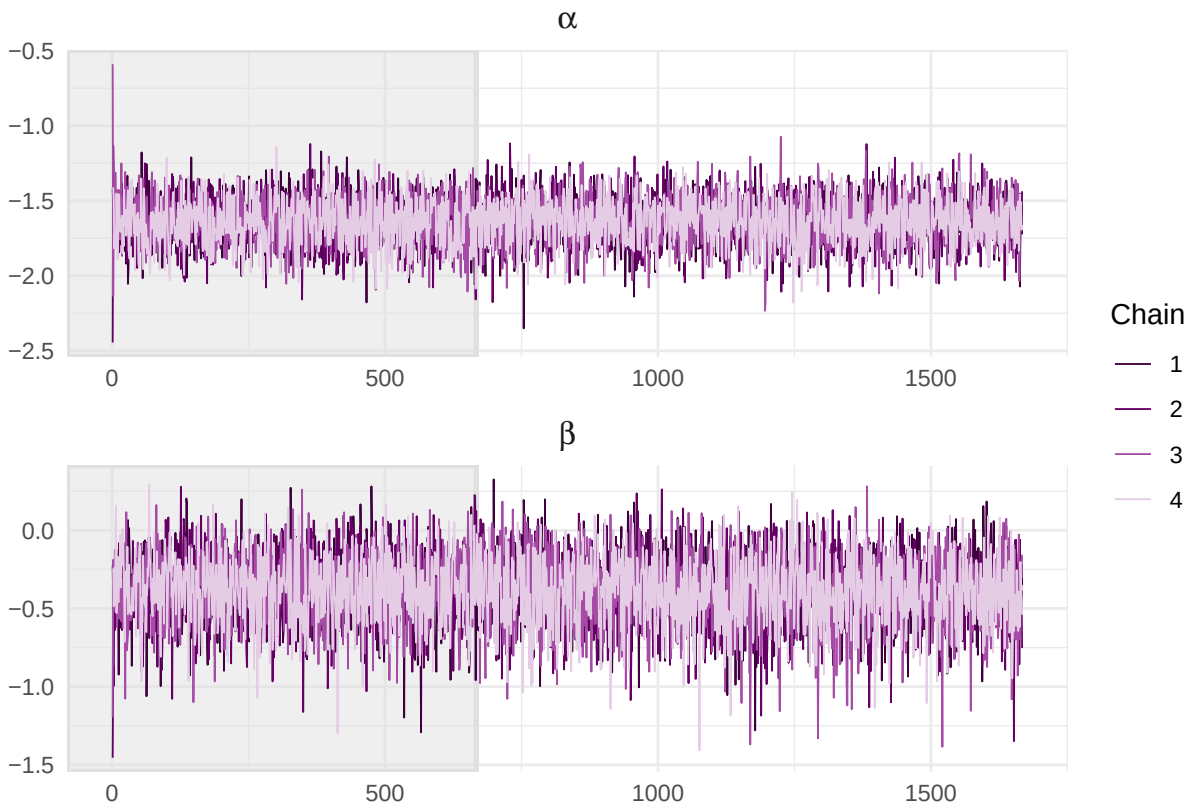
```r
# Resumen de los resultados
print(fit_combinado, pars = c("alpha", "beta"))
```

```
## 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 Sun May 26 14:07:22 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).
```

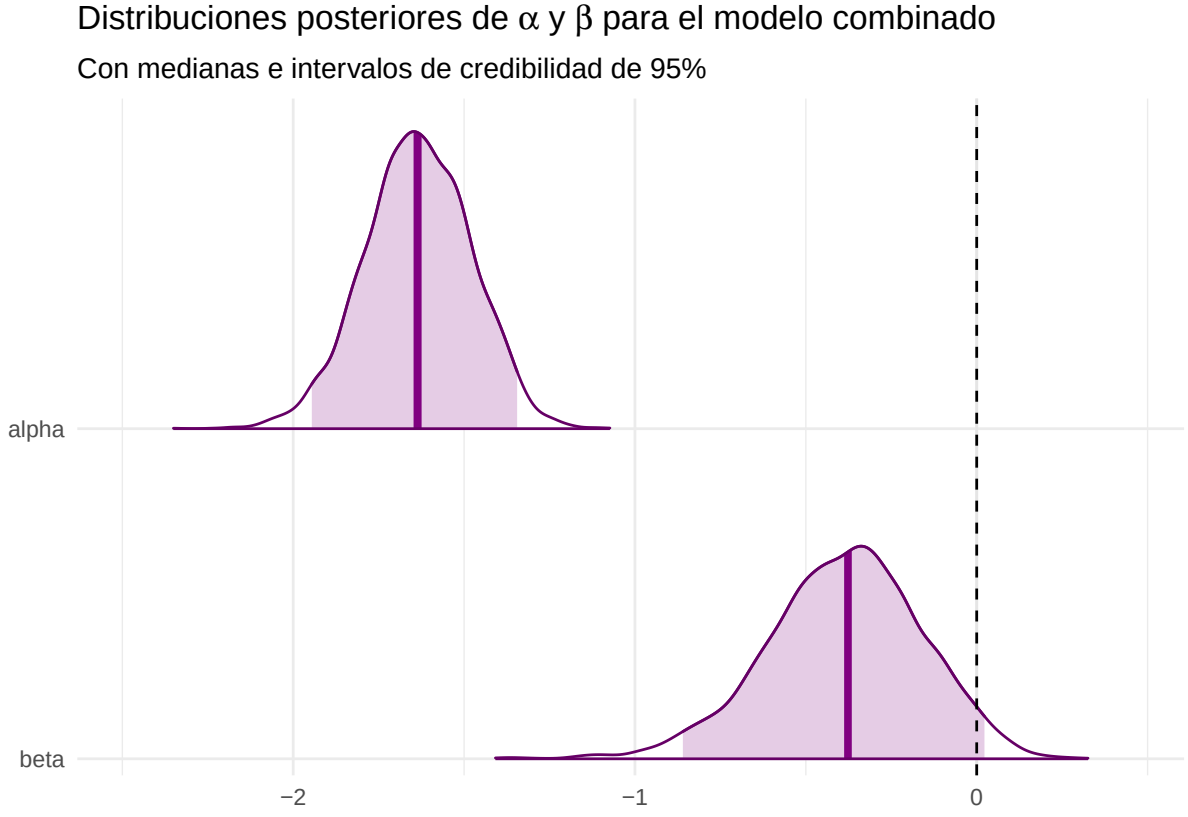
```r
# Extraer las muestras de la distribución posterior
posterior_samples_combinado <- extract(fit_combinado)
posterior_samples_combinado_2 <- extract(fit_combinado, inc_warmup = TRUE, permuted = FALSE)

# Cambiando el color para diferenciar del modelo anterior
color_scheme_set("purple")

mcmc_trace(
  posterior_samples_combinado_2,
  pars = c("alpha", "beta"),
  n_warmup = 667,
  facet_args = list(nrow = 2, labeller = label_parsed)
) + theme_minimal() + facet_text(size = 12)
```




```r
mcmc_areas(fit_combinado, pars = c("alpha", "beta"), prob = 0.95) +
  theme_minimal() +
  labs(
    title = expression(paste("Distribuciones posteriores de ", alpha, " y ", beta, " para el
    subtitle = "Con medianas e intervalos de credibilidad de 95%"
  ) +
  geom_vline(xintercept = 0, linetype = "dashed")
```


Distribuciones posteriores de α y β para el modelo combinado
Con medianas e intervalos de credibilidad de 95%
alpha
beta
−2
−1
0

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.

Ejercicio 4: Modelos

Comparación de modelos

Discusión de resultados

Comentarios finales

Referencias

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