Regresión logística desde la perspectiva bayesiana.

Diego Hernández

Esta entrega consiste en que realicéis un contraste de hipótesis bayesiano utilizando una base de datos que conozcáis. El objetivo de esta entrega es que demostréis que habéis comprendido la estructura básica de la inferencia bayesiana y que sepáis aplicarla a una pregunta de investigación concreta que tengáis.

Los datos con los que se realiza la tarea provienen de la investigación realizada por Sap et. al (2020). Se puede acceder al archivo a través del enlace que proporcionan los autores, aunque también se puede descargar desde la web de kaggle. Contiene 6854 observaciones, aunque son 5535 participantes. Aproximadamente la mitad de ellos debían redactar brevemente un evento real memorable o saliente que hubieran experimentado en los últimos 6 meses. La otra mitad debía redactar brevemente una historia a partir de alguno de los resúmenes de las historias reales que se les proporcionaba. Parte de los participantes que escribieron una historia real tuvieron que reescribirla pasados 2-3 meses, de ahí que haya más observaciones que participantes en el conjunto de datos. Con independencia de la condición a las que se les sometió, todos tuvieron que responder algunos autoinformes (información demográfica y sobre la historia que habían escrito).

(1) Pregunta de investigación e hipótesis

En el trabajo original de Sap et. al (2020), el interés estaba en el contenido de las historias y se analizaba con técnicas propias del área de Procesamiento del Lenguaje Natural. Pero es posible estudiar la diferencia entre el tipo de evento (real/recordado e imaginado) desde un enfoque analítico más convencional. Se propone estudiar la relación entre tipo de evento y dos de las variables recogidas, la distracción durante la escritura de la historia y la importancia atribuida al evento descrito. Se asume que ambas variables son continuas, aunque podría ser una decisión cuestionable para el caso de importancia, pues se ha medido con un ítem con respuesta tipo Likert con solo 5 alternativas de respuesta.

Se hipotetiza una relación positiva entre distracción y la probabilidad de que el evento narrado sea recordado. En principio, la creación de una historia debería suponer un mayor esfuerzo cognitivo frente a la recuperación de recuerdos. De igual modo, entre los que escriben el evento estando más distraídos, se espera que haya una proporción mayor de eventos recordados.

Por otro lado, se plantea la existencia de una relación positiva entre importancia y la probabilidad de que el evento sea real. Esta hipótesis se basa más en la concepción intuitiva de

que es más razonable atribuir relevancia a sucesos o recuerdos que son reales, en vez de simplemente imaginados.

No se va a evaluar el ajuste de distintos modelos, sino que el objetivo es simplemente examinar el modelo propuesto.

(2) Técnica de contraste de hipótesis y especificaciones

El modelo apropiado para estudiar la relación entre una serie de predictores y una variable dependiente dicotómica es la regresión logística. Cada caso i de la variable dicotómica tipo de evento, que aquí se llama $\mathbf{y}$, sigue una distribución Bernoulli con probabilidad π_i de que el caso sea un éxito (que pertenezca a la categoría 1 cuando se ha hecho codificación *dummy*). Aquí indica la probabilidad de que el evento sea recordado. Formalmente:

$$\mathbf{y}_{\mathrm{i}} \sim Bernoulli(\pi_i)$$

Por otro lado, esa probabilidad es la que se modela por medio de la ecuación logística:

$$\pi_i \coloneqq \mathrm{p}\left(\mathbf{x}_1^{(\mathrm{i})}, \mathbf{x}_2^{(\mathrm{i})}\right) = \frac{1}{1+\exp\left(-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i})\right)}$$

Donde:

x_{1i} : valor de la variable estandarizada distracción (z_distracted) para el caso i.

x_{2i} : valor de la variable estandarizada importancia del evento (z_importance) para el caso i.

β_0 : intersección.

β_1 : coeficiente de regresión asociado a la variable distracción estandarizada.

β_2 : coeficiente de regresión asociado a la variable importancia del evento.

Los predictores han sido estandarizados para facilitar la interpretación. Aunque ambos estaban medidos en una escala de 5 puntos, eso no significa que compartan unidades de medida, pues un incremento de un punto en distracción no necesariamente es equivalente a un aumento de un punto en importancia.

Ahora bien, desde el punto de vista bayesiano, los coeficientes de regresión son variables aleatorias, y por tanto tienen asociada una función de densidad. Para la especificación de las distribuciones previas de los parámetros se ha adoptado el enfoque propuesto por Gelman et. al (2008) de utilizar *priors* débilmente informativos, lo que se considera una posición intermedia entre distribuciones planas, nada informativas, y distribuciones muy informativas, fundamentadas en conocimiento previo:

$$\beta_0 \sim Cauchy(0,10)$$
$$\beta_1 \sim Cauchy(0,2.5)$$
$$\beta_2 \sim Cauchy(0,2.5)$$

Según Gelman et. al (2008), la distribución Cauchy presenta algunas ventajas con respecto a la distribución normal, que se suele utilizar por defecto. De todas formas, aquí se probaron ambas opciones, distribuciones a priori normales y distribuciones a priori Cauchy. La diferencia en

términos de DIC era mínima, pero se acabó eligiendo Cauchy por tener una ligera superioridad y ser la recomendación de Gelman et. al (2008).

Para el ajuste del modelo desde el enfoque bayesiano se emplean cadenas de Markov Montecarlo (MCMC), específicamente el algoritmo de muestreo de Gibbs. La implementación se hace mediante el software JAGS, a través de R mediante el paquete R2jags (ver anexo). En JAGS no se dispone de una función para la distribución Cauchy, pero es equivalente a una distribución t con 1 grado de libertad y mismos parámetros de posición y escala.

En código:

```
bayes_logit <- function(){
   # priors normales
   # beta0 ~ dnorm(0,0.01)
   # beta[1] ~ dnorm(0,0.16)
   # beta[2] ~ dnorm(0,0.16)
   # recomendaciones Gelman et. al (2008)
   # Cauchy(loc=0,scale=2.5) (t student con 1 gl == Cauchy)
   beta0 ~ dt(0,0.01,1) # prior menos informativo para intersección
   beta[1] ~ dt(0,0.16,1) # (1/2.5^2) = 0.16
   beta[2] ~ dt(0,0.16,1)
   # likelihood
   for (i in 1:N){
      prob[i] <- ilogit(beta0 + beta[1]*z_distracted[i] + beta[2]*z_importance[i])
      y[i] ~ dbern(prob[i])
   }
}
```

Se estableció el número de muestras (tamaño de las cadenas) en 10000 y un *burn-in* de 500. Para cada parámetro se construyeron 3 cadenas distintas de forma paralela (la función jags.parallel lo permite). Se fijó la semilla de aleatorización en 13 para permitir la reproducibilidad de los resultados.

(3) Resultados del análisis

La tabla de resultados que proporciona R2jags es la siguiente:

```
Inference for Bugs model at "bayes_logit", fit using jags,
 3 chains, each with 10500 iterations (first 500 discarded)
 n.sims = 30000 iterations saved
         mu.vect sd.vect     2.5%      25%      50%      75%    97.5%  Rhat n.eff
beta[1]   0.259   0.034    0.195    0.237    0.259    0.282    0.326 1.001 12000
beta[2]   1.382   0.044    1.298    1.352    1.382    1.412    1.469 1.001  8600
beta0    -0.057   0.033   -0.123   -0.079   -0.057   -0.035    0.007 1.001  4600
```

```
deviance 5830.271   2.424 5827.490 5828.492 5829.657 5831.398 5836.547 1.001 30000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 2.9 and DIC = 5833.2
DIC is an estimate of expected predictive error (lower deviance is better).
```

A partir de la tabla podemos evaluar una primera medida de convergencia. El estadístico de Gelman-Rubin (Rhat) para los tres parámetros estimados es prácticamente 1, lo cual es indicativo de convergencia.

Los gráficos *traceplot* y las curvas de densidad también aportan evidencia en esta dirección. En la figura 1 puede verse que las cadenas oscilan en torno al mismo rango de valores, no hay desviaciones. En la figura 2 puede comprobarse el claro solapamiento entre las curvas de densidad, lo que indica que las 3 cadenas (de cada paramétro) generan la misma distribución a posteriori.

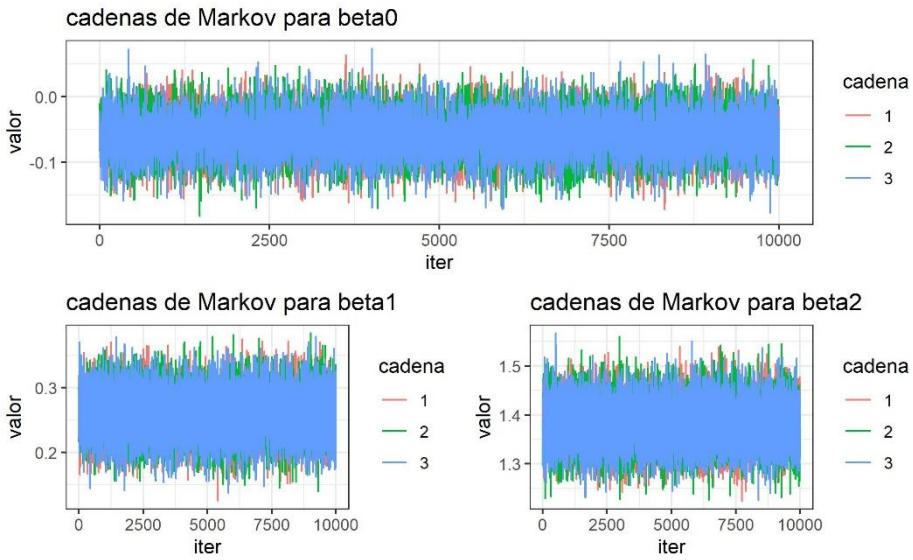

Figura 1. Gráficos traceplot generados manualmente (ver anexo)

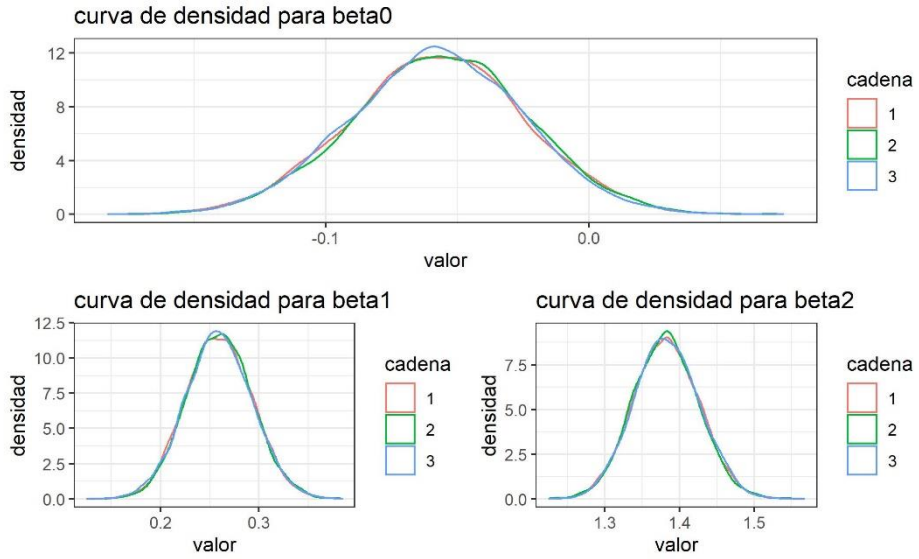

Figura 2. Curvas de densidad generadas manualmente (ver anexo)

La dependencia en las cadenas generadas parece ser baja, a juzgar por los elevados tamaños muestrales efectivos. Otra prueba de ello son los gráficos de autocorrelación, que muestran una mínima relación con una lag menor a 10.

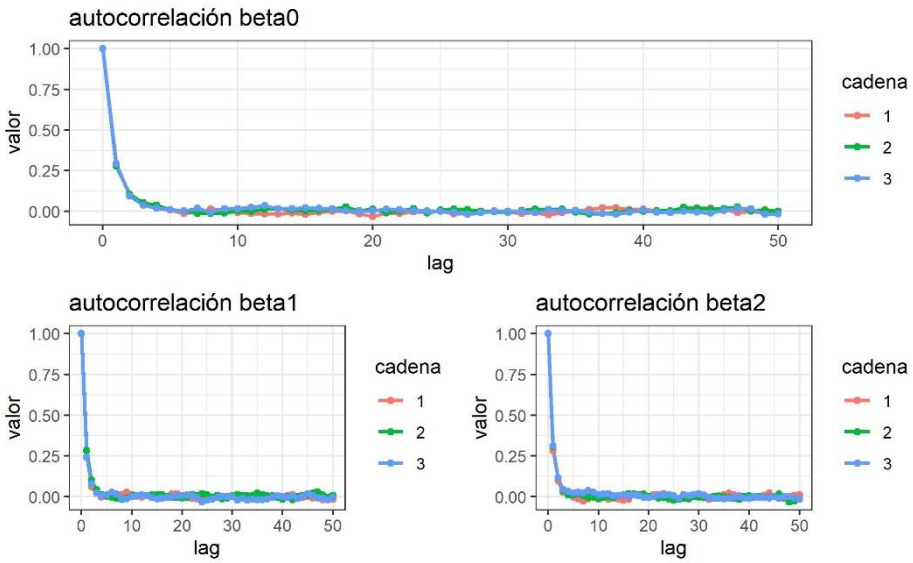

Figura 3. Gráficos de autocorrelación generados manualmente (ver anexo)

Para una representación gráfica del proceso de "actualización de creencias" se han impuesto en un mismo gráfico las curvas de densidad a priori y a posteriori (éstas con los datos de las 3 cadenas). Se reproduces aquí con el eje de abscisas "truncado" para que se muestren solo los valores más probables.

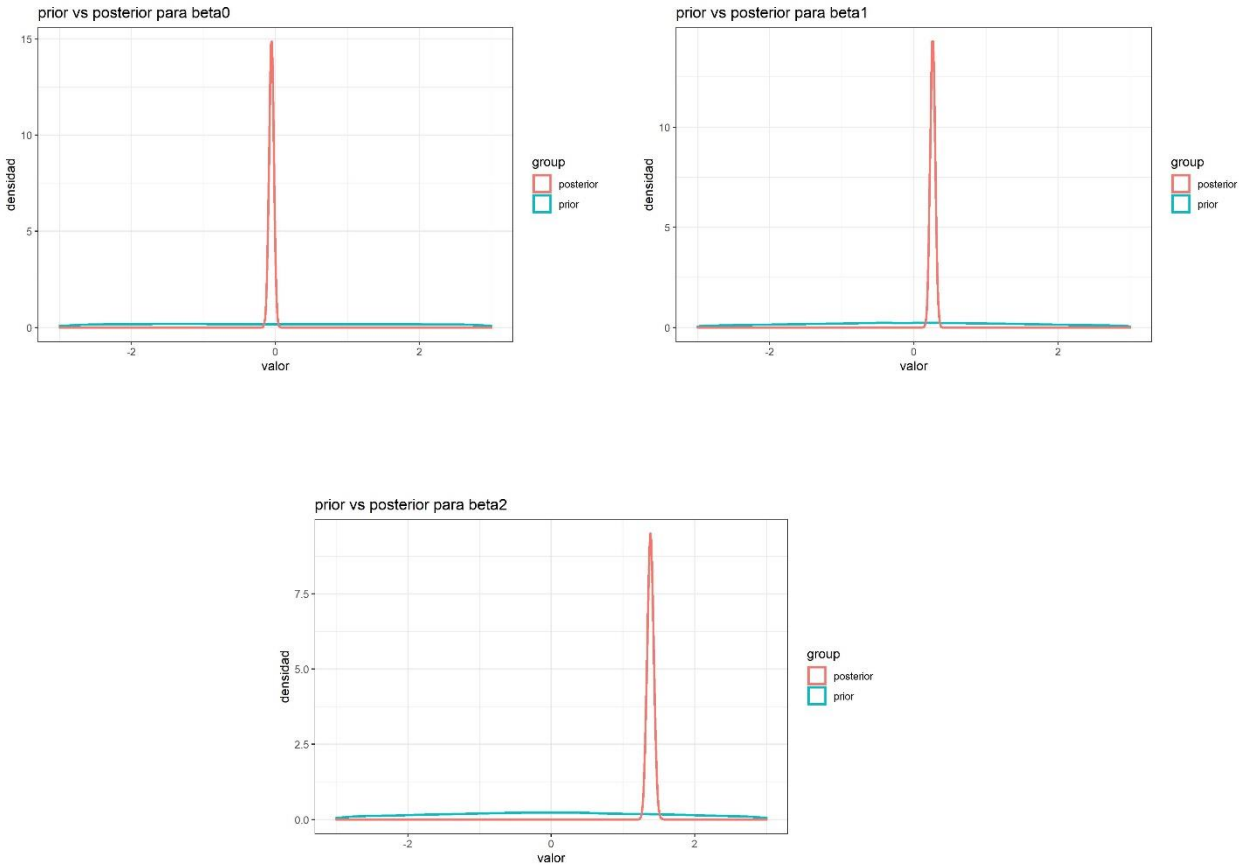

Figura 4. Gráficos prior vs posterior generados manualmente (ver anexo)

Resulta evidente que se ha producido una reducción de la incertidumbre con respecto a los valores de los parámetros, aunque no es sorprendente teniendo en cuenta que los priors no eran muy informativos.

(4) Interpretación: ¿Qué significan los resultados?

La interpretación del modelo suele hacerse reformulándolo en términos de las *odds*:

$$odds_i := \frac{\pi_i}{1-\pi_i} = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$

Teniendo en cuenta que la categoría de referencia es el evento imaginado, podemos concluir lo siguiente sobre los coeficientes de regresión:

- Con valores promedio tanto en distracción como en importancia atribuida al evento, la odds media de evento recordado es aproximadamente 0.944. Esto implica que la probabilidad de que el evento sea real es prácticamente la misma de que sea imaginado, tan solo es algo inferior. La interpretación va en la misma dirección si atendemos al intervalo de credibilidad. El valor de β_0 en la distribución a posteriori se encuentra entre -0.123 y -0.035 con una probabilidad de 0.95. Esto supone que la odds se encuentra entre 0.884 y 0.965

- Con respecto β_1, indica que la odds de evento real se incrementa 1.295 puntos por cada puntuación típica extra en la variable distracción, manteniendo constante la variable importancia. Por ejemplo, conociendo que un sujeto puntúa muy alto en ese rasgo (supongamos que está una desviación típica por encima de la media), se pronostica un aumento promedio del 30% en la odds de que el suceso descrito sea real con respecto a la odds "base" en la que asumimos un valor medio en distracción.
- El coeficiente β_2 tiene una magnitud promedio mucho mayor, de 1.382, con un intervalo de credibilidad al 95% de (1.298, 1.469). Manteniendo la distracción constante, la odds de evento real aumenta en un factor de 3.982 por cada desviación típica de incremento en la variable importancia. Conociendo que un sujeto considera muy importante para él/ella la historia narrada (puntuación una desviación típica por encima de la media, por ejemplo), la probabilidad pasa de ser aproximadamente 0.5 (habiendo asumido puntuación media en importancia) a 0.79. Todo ello considerando el valor promedio de β_2.

La bondad de ajuste global no se comenta, ya que no se compara el modelo con ninguna alternativa.

De manera complementaria, resulta interesante ver en qué se traduce el modelo en una situación práctica en la que debemos decidir (clasificar) si un evento narrado es real o imaginado. El siguiente gráfico muestra la frontera de decisión que genera el modelo cuando el criterio para clasificar un evento como real es que se le pronostique una probabilidad superior a 0.5.

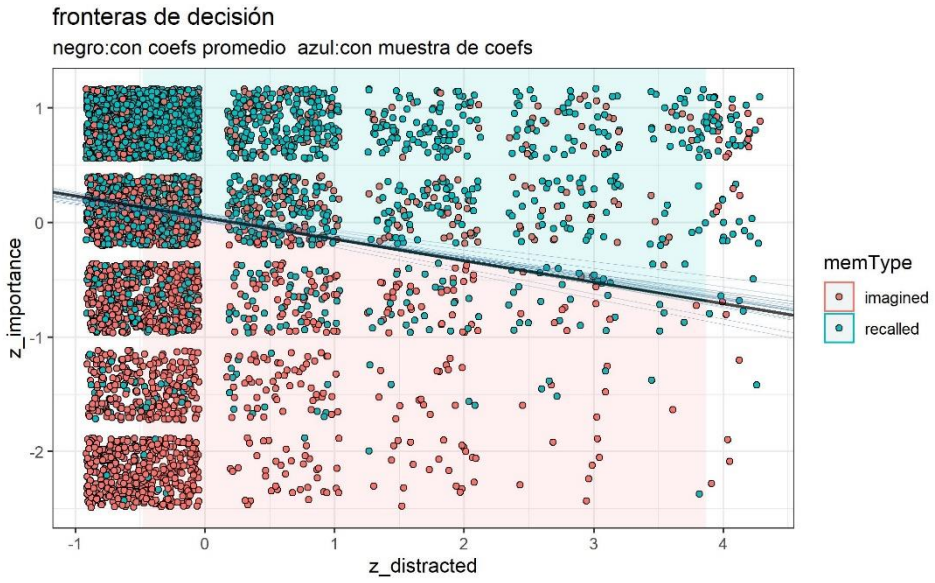

Figura 5. Gráfico con fronteras de decisión generado manualmente (ver anexo)

En azul se ha representado una muestra 20 fronteras de decisión construidas a partir de 20 conjuntos de coeficientes generados en el muestreo de Gibbs. Pero no representan necesariamente coeficientes que pertenezcan a los intervalos de credibilidad al 95%. Cabe añadir además que se ha añadido ruido artificialmente (*jitter*) a cada punto para mejorar la visualización.

En línea con lo comentado anteriormente, se observa como la mayor puntuación en importancia es lo que más ayuda a discriminar entre eventos reales e imaginados. Por último, en este gráfico puede verse que el modelo hace una separación razonable de las categorías, de hecho, la exactitud o proporción de aciertos en la clasificación es de aproximadamente 0.73. Esta medida de exactitud de ningún modo puede utilizarse como referencia para juzgar la capacidad generalización, pues se ha obtenido haciendo predicciones con los mismos datos que se han utilizado para ajustar el modelo. Es muy probable que se haya producido sobreajuste y que la exactitud con observaciones nuevas sea menor. De todas formas, indica que el modelo parece adecuado para la clasificación.

(5) Conclusiones: ¿Qué respuesta habéis encontrado a vuestra investigación?

Parece que existe una evidencia clara de la relación entre el tipo de evento y las variables de distracción e importancia atribuida. De los dos predictores, la importancia es que parece explicar mejor la variable dependiente, si atendemos al tamaño del coeficiente de regresión. Este hecho es consistente con la hipótesis que se barajaba al inicio. A mayor importancia atribuida a un suceso que se narra, más probable es que ese suceso sea real, y que se haya recordado. Resultaría contraintuitivo que se diese más importancia a eventos que tan solo han sido imaginados.

Referencias

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, *2*(4), 1360-1383.

Kruschke, J. (2015). Dichotomous predicted variable. In J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (pp. 621-649). Elsevier.

Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. W. (2020, July). Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (pp. 1970–1978)*. Association for Computational Linguistics.