

Bayesian Methods Final Project

SESGO DE GÉNERO EN EL RENDIMIENTO ACADÉMICO EN LA MATERIA DE
MATEMÁTICAS

Aráiztegui, Aránzazu
Ferrara, Lorenzo
Lucchini, Marco

06 January, 2023



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística



UNIVERSITAT DE
BARCELONA

Índice

1 Introducción	1
1.1 Descripción del problema	1
1.2 Objectivos del modelo	1
2 Descripción de la base de datos	2
2.1 Definición de las variables utilizadas	2
2.2 Análisis exploratorio de los datos	2
3 Análisis bayesiano	5
3.1 Modelo 1	5
3.1.1 Presentacion del modelo	5
3.1.2 Simulación de modelo	6
3.1.3 Interpretación y validación de los modelos	6
3.2 Modelo 2 sin areas territoriales	8
3.2.1 Presentacion de los modelos 2	8
3.2.2 Simulación de modelo 2	8
3.2.3 Interpretación y validación de los modelos 2	8
3.3 Modelo 3: quito “ciudad”	9
3.3.1 Presentacion de los modelos 3	10
3.3.2 Simulación de modelo 3	10
3.3.3 Interpretación y validación de los modelos 3	11
4 Implementación del modelo	13
5 Conclusions	16
6 Apéndice	17
6.1 Código R	17
6.2 Tablas parámetros estimados	17
7 Referencias	18

1 Introducción

El sesgo de género está presente en numerosos ámbitos de nuestra vida y se aprecia de forma notable en el ámbito educativo, más concretamente en las materias del ámbito STEAM.

Parece ser que aunque no hay una evidente brecha de género en el comienzo de los estudios primarios, el sesgo de género comienza a aparecer de forma clara a lo largo del proceso educativo, agudizándose en las últimas etapas de la educación obligatoria y evidenciándose de forma clara en la educación universitaria.

1.1 Descripción del problema

Con el objetivo de evaluar la capacidad y el nivel de competencia en las diferentes áreas del conocimiento que tiene el alumnado de Cataluña, el Departament d'Educació de la Generalitat de Cataluña realiza una prueba de competencias y conocimientos básicos en las áreas lingüísticas, matemáticas y científico-tecnológicas en los últimos cursos de la educación primaria y secundaria.

Según el Departament se trata de una evaluación de carácter formativo y orientador que pueda servir tanto a los centros como al profesorado y al propio Departament para impulsar las mejoras en el sistema educativo catalán.

1.2 Objectivos del modelo

Este trabajo tiene como objetivo principal ver si existe un sesgo de género en los resultados obtenidos en la competencia matemática con respecto al sexo y a las competencias humanísticas para ello intentaremos crear un modelo que relacione la puntuación obtenida en la competencia matemática con respecto al sexo, a las competencias lingüísticas e incluso con respecto al tipo de centro educativo o al tamaño de la población. De esta manera podríamos ver si en el sexo femenino no se da una diferencia entre el rendimiento en humanidades y matemáticas, personas con bajo rendimiento en humanidades también lo tendrían en matemáticas.

Y en cambio en el sexo masculino personas con bajo rendimiento en humanidades tendrían buenos resultados en matemáticas.

Si ampliamos los datos y vemos la tabla de resultados para un mismo individuo en sexto de primaria y cuarto de la podríamos establecer un segundo objetivo que sería ver si se mantienen los resultados en ambos sexos o si hay diferencias significativas en cuanto al rendimiento en el área de matemáticas al aumentar la edad.

2 Descripción de la base de datos

La base de datos utilizada es una mezcla de los datos ofrecidos para cuarto curso de ESO y los datos que se ofrecen para sexto curso de primaria. La razón de realizar la fusión de las dos tablas es poder evaluar la evolución de una misma persona desde el final de la educación primaria hasta el final de la educación secundaria. Se puede acceder a la base de datos completa en el siguiente enlace:

Avaluació de quart d'Educació Secundària Obligatòria | Dades obertes de Catalunya

Avaluació de sisè d'educació primària | Dades obertes de Catalunya

El dataset contiene los resultados obtenidos por el alumnado de cuarto curso de ESO y los datos que se ofrecen para sexto curso de primaria en la evaluación de competencias básicas desde el año 2012.

El código de alumno se utilizó para hacer comparativas con los resultados obtenidos.

La base de datos ha sido actualizada el 20 de octubre de 2022 y contiene los datos de 46384 estudiantes.

2.1 Definición de las variables utilizadas

Base de dades

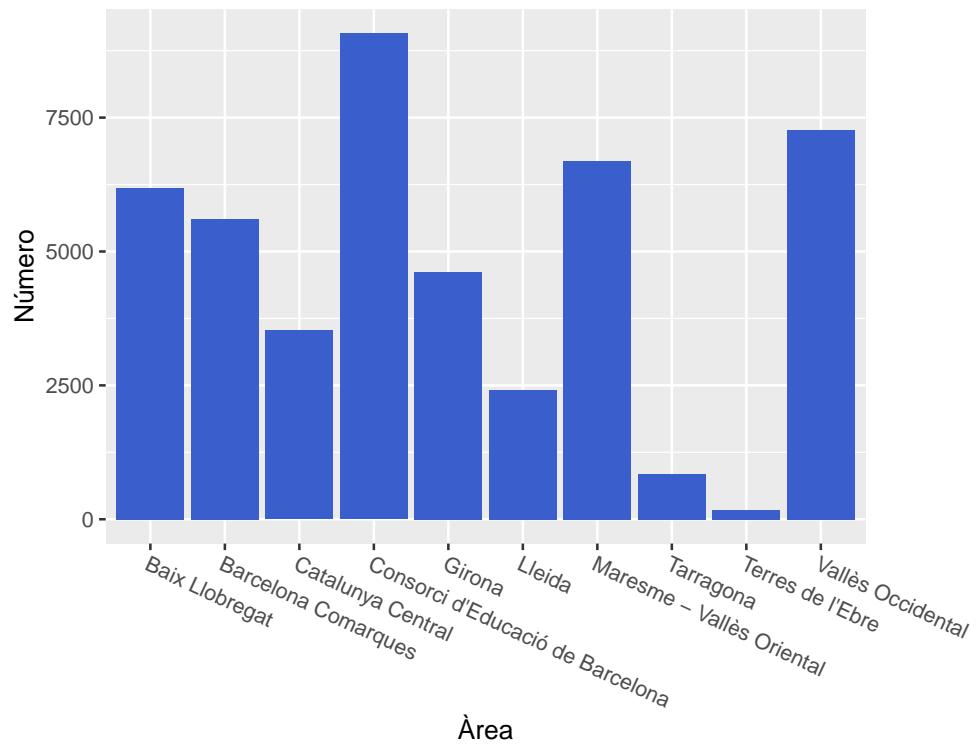
Nom de columna	Descripció	Tipus
PMAT_4	Puntuació global ponderada de competència matemàtica en el examen de Quart	Nombre
PMAT_6	Puntuació global ponderada de competència matemàtica en el examen de Sisè	Nombre
PLENG_4	Puntuació global ponderada de la competència lingüística en llengua catalana y castellana en el examen de Quart	Nombre
PLENG_6	Puntuació global ponderada de la competència lingüística en llengua catalana y castellana en el examen de Sisè	Nombre
PANG_4	Puntuació global ponderada de la competència lingüística en llengua anglesa en el examen de Quart	Nombre
PANG_6	Puntuació global ponderada de la competència lingüística en llengua anglesa en el examen de Sisè	Nombre
GENERE	Gènere de l'alumne/a que es presenta a l'avaluació	Text Pla
AREA_TERRITORIAL	Regió on es troba el centre de l'alumne/a que es presenta a l'avaluació	Text Pla
NATURALESA	Determina si el centre de l'alumne/a és públic, privat o concertat	Text Pla
HÀBITAT	Municipis per trams de població	Text Pla

2.2 Análisis exploratorio de los datos

En el análisis inicial de los datos podemos ver que la distribución de hombres y mujeres es uniforme con una proporción de 50,1% niños y 49,9% niñasen.

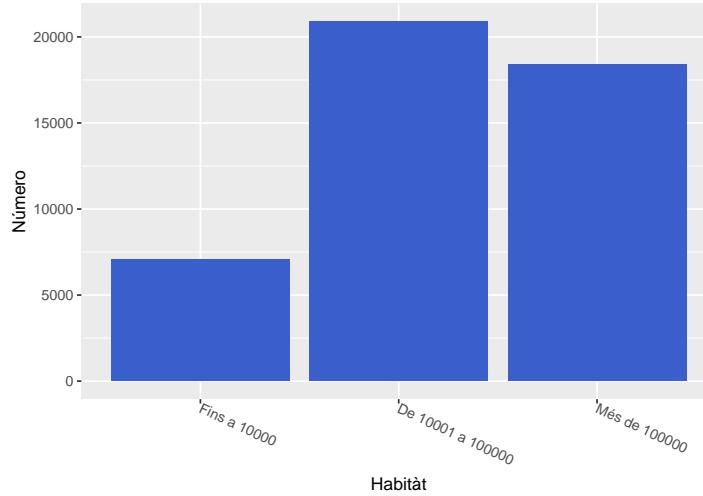
La distribución de alumnos por áreas es la siguiente.

Distribución de alumnos entre las áreas



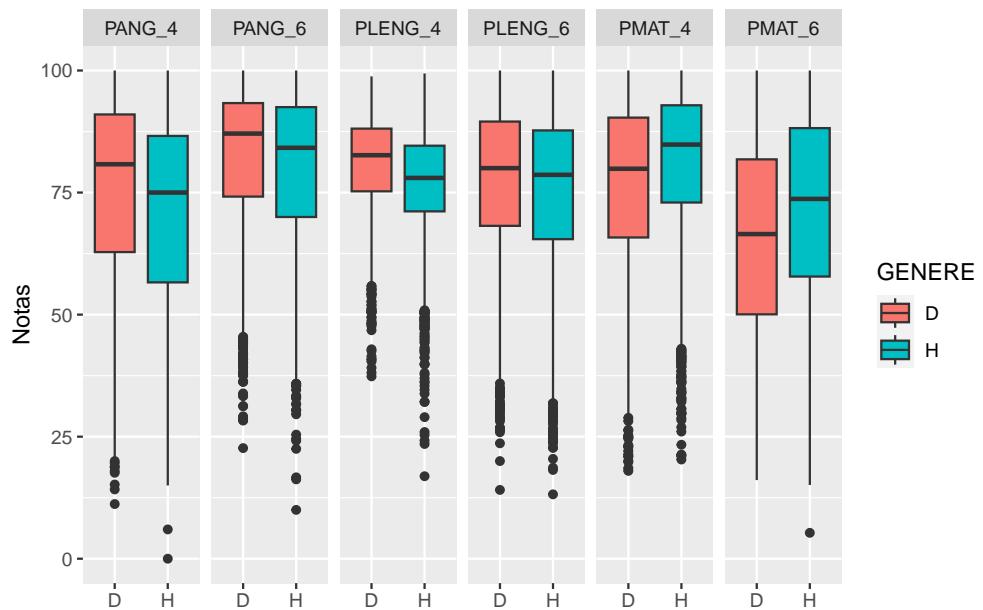
Según el tamaño de la población los datos de los que disponemos se concentran en individuos de ciudades de tamaño medio o grande.

Distribución de alumnos entre las hàbitat



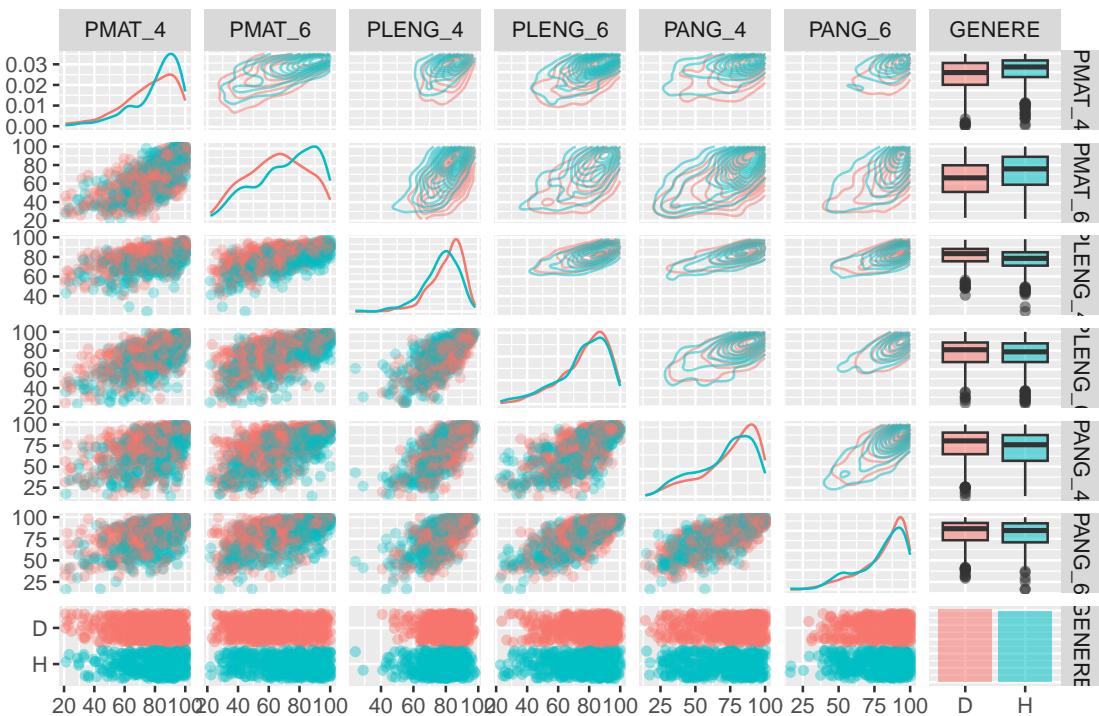
Pasamos ahora a la nota obtenida en el examen de inglés, lengua y matemáticas en el examen de Sisè y Quart. Los grados se dividen entre niños y niñas. Del gráfico suponemos que los chicos tienden a tener peores notas en inglés y lengua mientras que obtienen mejores resultados en matemáticas con respecto a las chicas.

Comparación de resultados educativos por género



En el siguiente gráfico podemos ver las diferencias entre las calificaciones en matemáticas, inglés y lengua de cada sexo dependiendo del nivel educativo. Si comparamos la misma asignatura en los dos cursos los datos muestran una distribución casi lineal y con poca variabilidad, lo cual nos hace pensar que el alumnado que obtiene buenos resultados en sexto de primaria también obtienen buenos resultados al acabar la secundaria, como cabría esperar. También cabe destacar la distribución de las frecuencias por género, en especial en el caso de matemáticas donde parece observarse una diferencia más evidente entre chicos y chicas.

Comparación de resultados entre sujetos



3 Análisis bayesiano

3.1 Modelo 1

Para describir qué tan bien obtendrá un estudiante en el examen Quart Maths, usamos las calificaciones que obtuvo en los exámenes Sise y las que obtuvo en los exámenes Quart English and Language. Dado que nuestro objetivo es explicar si el género influye en el desempeño de un estudiante, llevándolo a obtener mejores resultados que sus compañeros, utilizamos las calificaciones desde una perspectiva diferencial con respecto al promedio, es decir,

$$MAT_4 = MAT_4 - \text{media}(MAT_4)$$

y lo mismo para todos los demás grados utilizados dentro del modelo. También usamos el género del estudiante, la región de origen, el tamaño de la ciudad y la interacción entre el género y los puntajes de los exámenes de idioma de la escuela secundaria en el modelo. Esta última covariable nos ayudará a explicar si los chicos tienen rendimientos opuestos en las dos materias con respecto al rendimiento medio.

Se crea un modelo jerárquico con la región de origen para agregar un efecto aleatorio debido al área. Se supone que todas las regiones tienen una distribución normal con media 0 mientras que la varianza cambia para cada una. Sin embargo, se obtiene a partir de una única distribución común a todas las regiones.

3.1.1 Presentación del modelo

El modelo utilizado es el siguiente.

$$\begin{aligned} y_i &\sim N(b_0 + mat_4 * x_i^1 + leng_4 * x_i^2 + leng_6 * x_i^3 + ing_4 * x_i^4 + ing_6 * x_i^5 + publica * x_i^6 \\ &+ ciudPeq * x_i^7 + ciudGrande * x_i^8 + sexo_h * x_i^9 + interaccion * x_i^2 * x_i^9 + G[\text{area}[i]], \tau_y) \end{aligned}$$

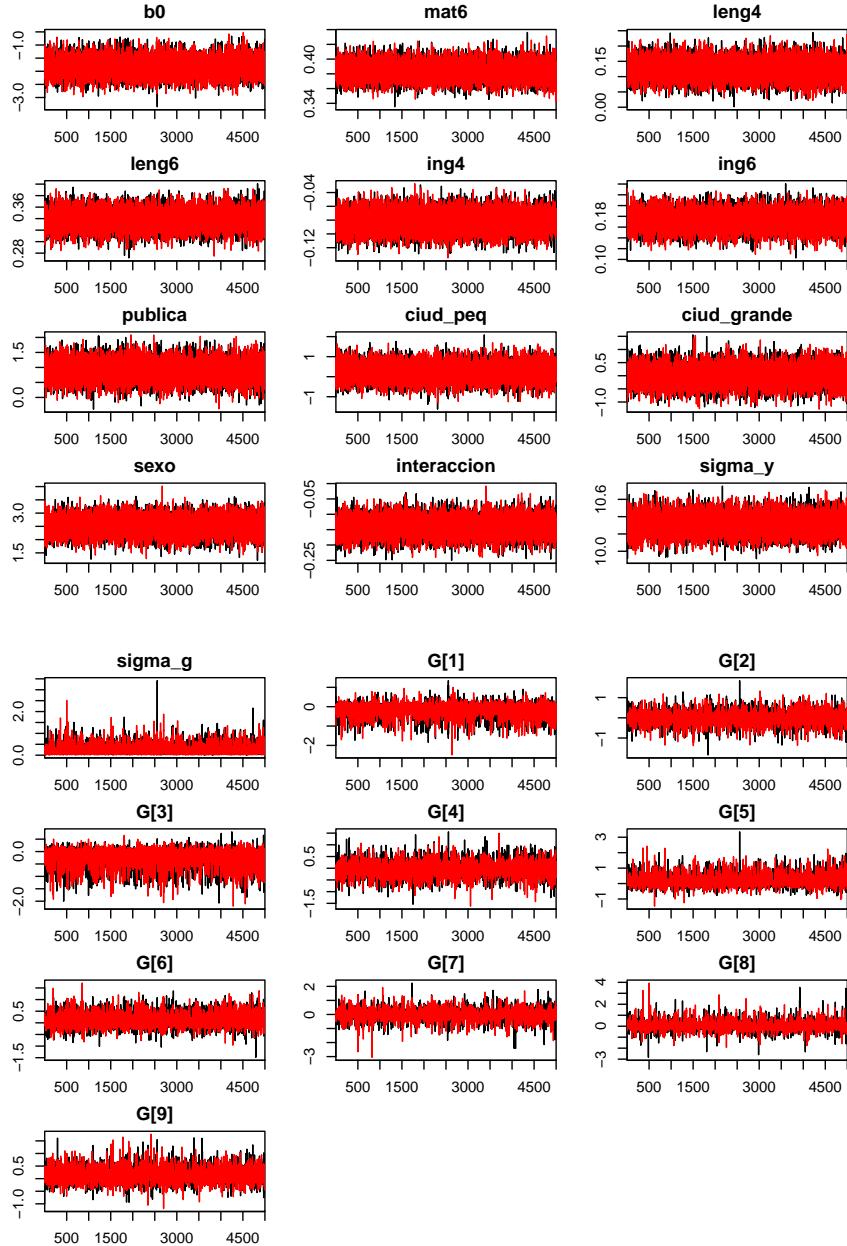
Suponemos que el coeficiente tiene la siguiente distribución. Elegimos la distribución normal estándar ya que las variables son un sesgo de la media de cada examen.

$$\begin{aligned} b_0 &\sim N(0, 1) \\ mat_4 &\sim N(0, 1) \\ leng_4 &\sim N(0, 1) \\ leng_6 &\sim N(0, 1) \\ ing_4 &\sim N(0, 1) \\ ing_6 &\sim N(0, 1) \\ publica &\sim N(0, 1) \\ ciudPeq &\sim N(0, 1) \\ ciudGrande &\sim N(0, 1) \\ sexo_h &\sim N(0, 1) \quad x^9 \text{ es } 1 \text{ si es hombre} \\ interaccion &\sim N(0, 1) \\ \tau_y &\sim \Gamma(0.001, 0.001) \\ sigma_y &= \frac{1}{\sqrt{\tau_y}} \\ G_i &\sim N(0, \tau_g) \\ \tau_g &\sim \Gamma(0.001, 0.001) \\ sigma_g &= \frac{1}{\sqrt{\tau_g}} \end{aligned}$$

3.1.2 Simulación de modelo

Ejecutamos nuestro modelo en un conjunto de datos de prueba de 4 000 estudiantes. Usamos el paquete JAGS para realizar simulaciones utilizando algoritmos de MCMC ya que realizar el problema analíticamente sería casi imposible.

Usando dos cadenas diferentes de 5000 iteraciones de largo y eliminando las primeras 400 iteraciones, obtenemos los siguientes resultados. Como puede ver, las dos cadenas convergen y, por lo tanto, el resultado de nuestro modelo es estable.

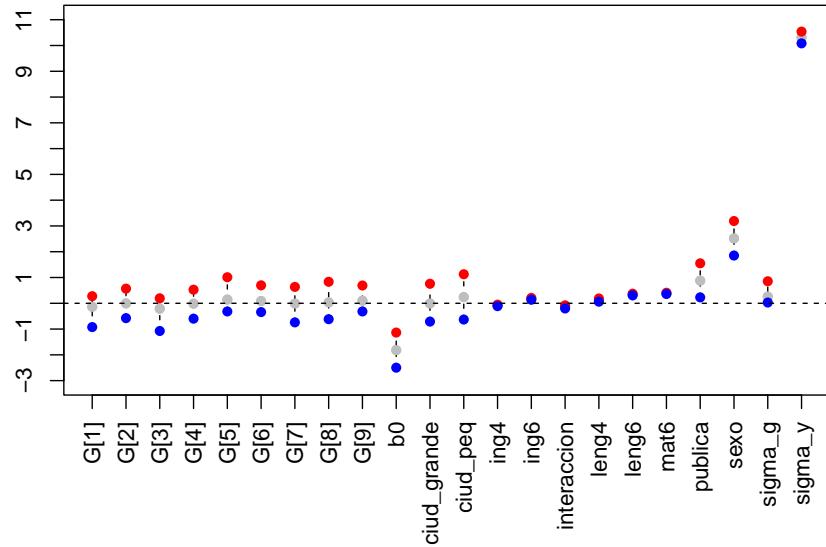


3.1.3 Interpretación y validación de los modelos

Los coeficientes obtenidos de nuestro modelo tienen las siguientes distribuciones. Si bien las distribuciones relativas a los votos parecen muy cercanas a cero, todas son significativas, el rango tan reducido se debe al

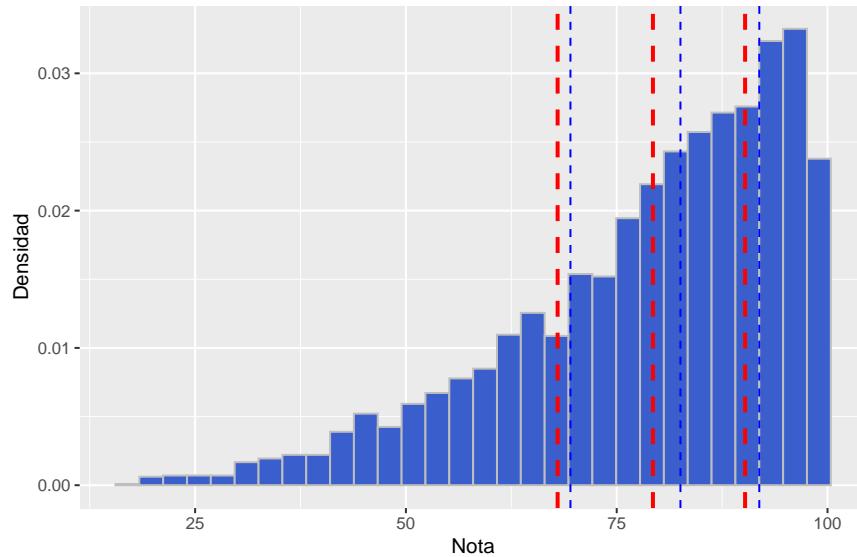
valor de los votos que varía en un rango de alrededor de ± 30 votos. Sin embargo, es interesante observar que el efecto aleatorio debido a todas las regiones tiene una distribución centrada alrededor de 0. Lo mismo ocurre con el tamaño de la ciudad. Por lo tanto, no podemos considerar su contribución relevante dentro del modelo y, posteriormente, intentaremos eliminar su contribución.

Intervalo de confianza del 95% para los coeficientes



Ahora validamos nuestro modelo usando los cuantiles 25%, 50% y 75% como estadísticas. Estos valores los obtenemos simulando nuestro modelo sobre un subconjunto del testset obtenido como complemento del conjunto de entrenamiento inicial. Nuevamente usamos 4 000 estudiantes.

Comparación entre la distribución PMAT_4 y el valor obtenido por las estat



Como puede verse, los tres estadísticos parecen estar bien dispuestos en la distribución de referencia, aunque se mantienen ligeramente por debajo de los valores reales.

Intentemos eliminar los efectos de área para que nuestro modelo sea más interpretable.

3.2 Modelo 2 sin áreas territoriales

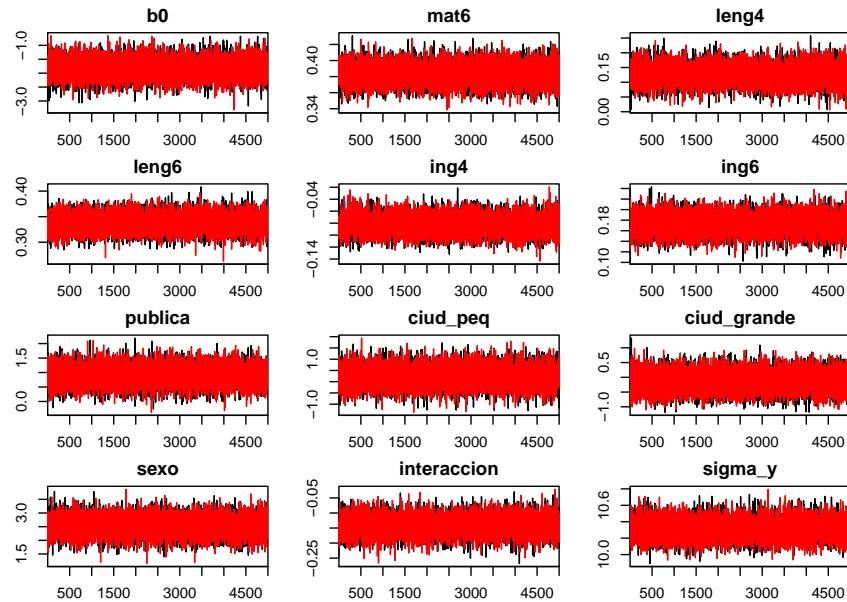
Dado que los coeficientes de las áreas territoriales resultaron con una distribución centrada alrededor de cero, tratamos de eliminar esa covariable del modelo para que sea más fácil de interpretar.

3.2.1 Presentación de los modelos 2

Usamos el mismo modelo ilustrado arriba en el que eliminamos las covariables relacionadas con el área geográfica. Para los coeficientes asumimos siempre una distribución normal estándar para partir de la hipótesis de que los alumnos se comportan con respecto a la media de los alumnos de la misma forma que se comportaron en los demás exámenes.

3.2.2 Simulación de modelo 2

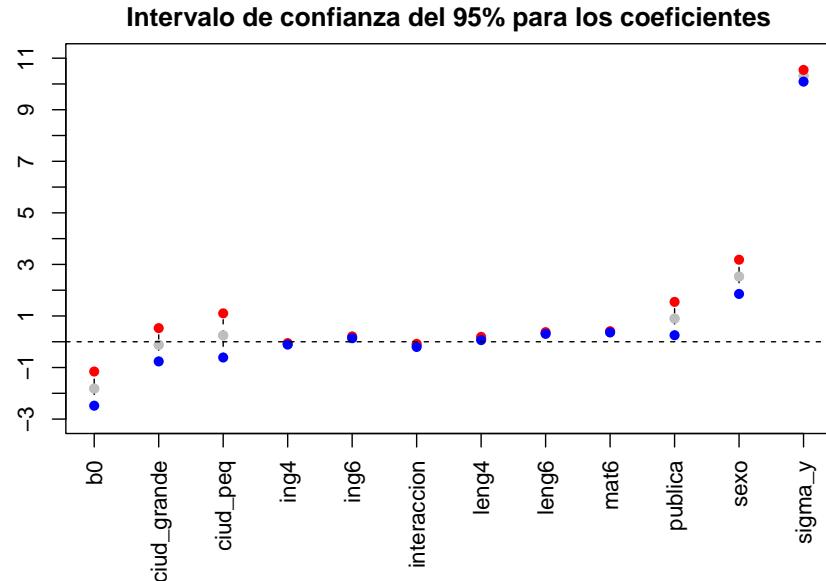
Al igual que hicimos en el modelo anterior, aquí también usamos el paquete JAGS para simular el modelo a través de MCMC. El resultado es el siguiente.



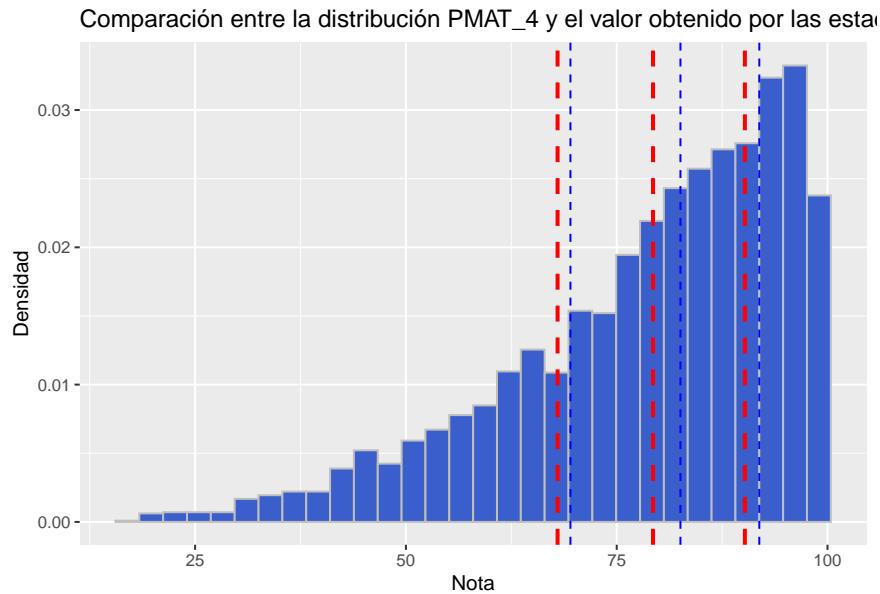
También en este caso las cadenas convergen y luego procedemos a analizar el modelo.

3.2.3 Interpretación y validación de los modelos 2

En cuanto al modelo 1, recordamos que las covariables relativas a las notas se expresan en términos de la diferencia entre el resultado y la media. Podemos ver que incluso después de eliminar las covariables relacionadas con el área geográfica, el coeficiente relacionado con la ciudad siempre se distribuye normalmente alrededor de 0.



Como en el caso anterior, analicemos las estadísticas relativas a los percentiles 25%, 50% y 75%.



El modelo resultante parece ser casi tan fiel como el anterior y además en este caso las estadísticas obtenidas de la simulación con el test set tienen valores ligeramente inferiores a las estadísticas reales.

También en este caso se procede a eliminar la covariable de ciudad para obtener un modelo más fácilmente interpretable.

3.3 Modelo 3: quito “ciudad”

Reducimos aún más nuestro modelo manteniendo únicamente las calificaciones obtenidas en los exámenes de Mate, idioma e inglés en los exámenes de Sisè, idioma e inglés en los exámenes de Quart, el género, el tipo de escuela (pública o privada) y la interacción entre género y nota de idioma en Quart.

3.3.1 Presentacion de los modelos 3

El modelo implementado es el siguiente.

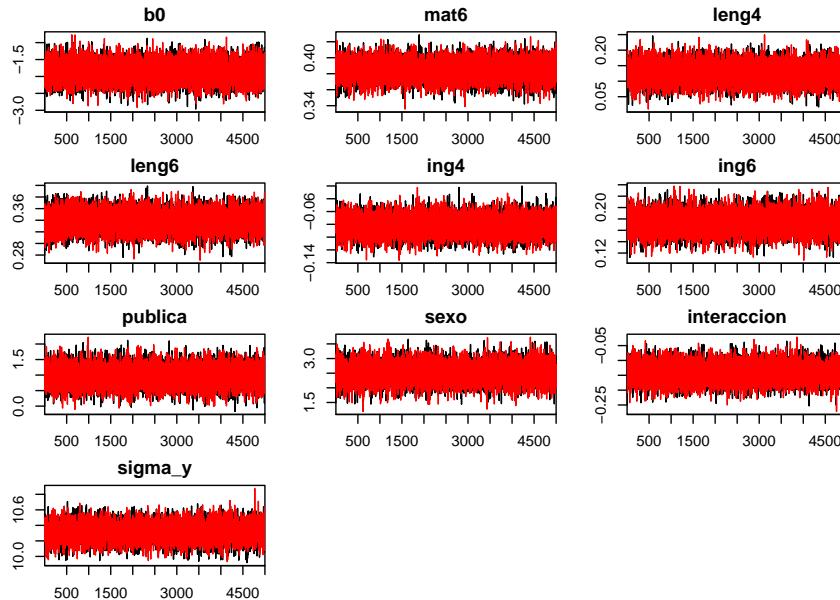
$$y_i \sim N(b_0 + mat_4 * x_i^1 + leng_4 * x_i^2 + leng_6 * x_i^3 + ing_4 * x_i^4 + ing_6 * x_i^5 + publica * x_i^6 + sexo_h * x_i^9 + interaccion * x_i^3 * x_i^9, tau_y)$$

Suponemos que el coeficiente tiene la siguiente distribución. Elegimos la distribución normal estándar ya que las variables son un sesgo de la media de cada examen.

$$\begin{aligned} b_0 &\sim N(0, 1) \\ mat_4 &\sim N(0, 1) \\ leng_4 &\sim N(0, 1) \\ leng_6 &\sim N(0, 1) \\ ing_4 &\sim N(0, 1) \\ ing_6 &\sim N(0, 1) \\ publica &\sim N(0, 1) \\ sexo_h &\sim N(0, 1) \quad x^9 \text{ es } 1 \text{ si es hombre} \\ interaccion &\sim N(0, 1) \\ tau_y &\sim \Gamma(0.001, 0.001) \\ sigma_y &= \frac{1}{\sqrt{tau_y}} \end{aligned}$$

3.3.2 Simulación de modelo 3

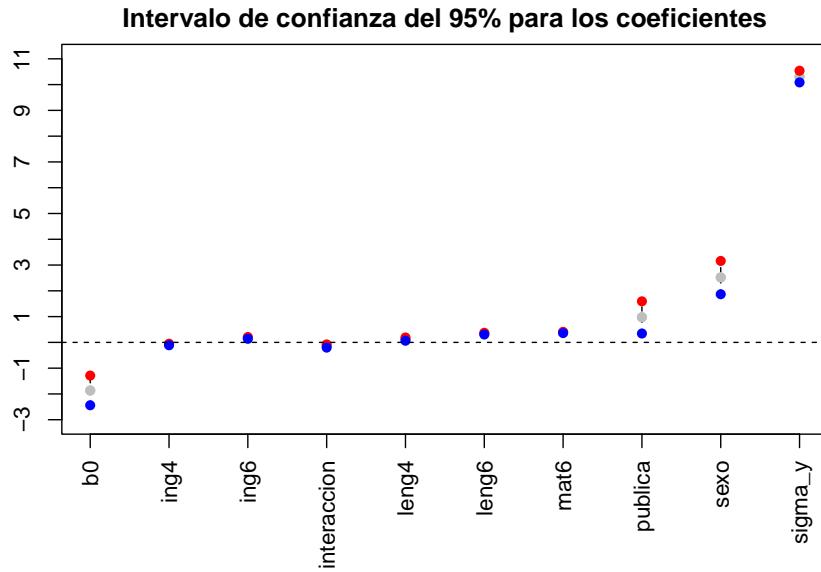
También en este caso se simula el modelo ya que una resolución analítica requeriría un esfuerzo computacional más que considerable.



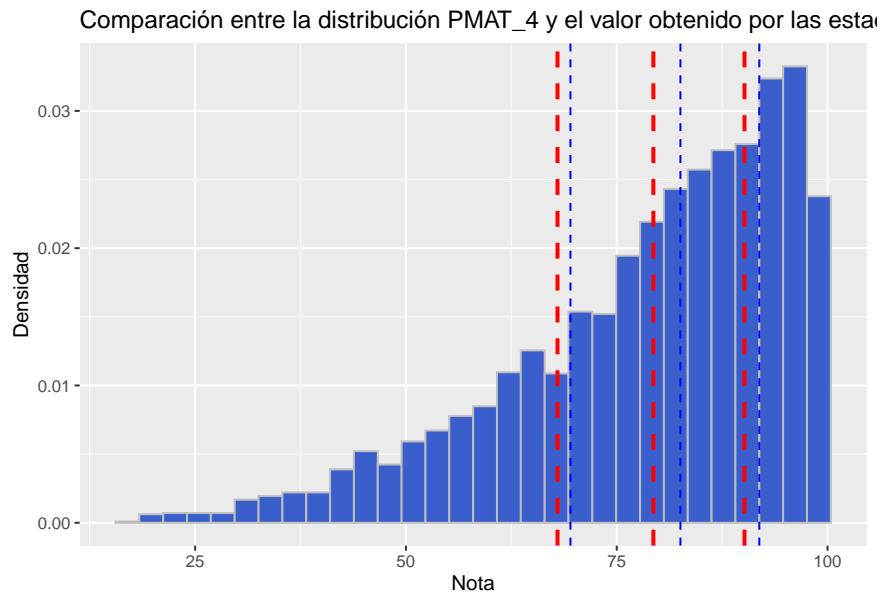
Como puede verse también en este caso el modelo converge.

3.3.3 Interpretación y validación de los modelos 3

También en este modelo recordamos que los coeficientes relativos a los resultados escolares se obtienen como diferencia de la media relativa al examen. Observamos que son valores muy cercanos a 0 pero a pesar de esto las distribuciones no contienen 0 en el intervalo de confianza del 95%. Por lo tanto, todas las covariables son significativas.



También en este caso analizamos las estadísticas de los percentiles 25%, 50%, 75%. Nos referimos siempre a la distribución real de las notas obtenidas por los alumnos.



Las estadísticas siempre están a la izquierda de las estadísticas reales, pero aún consideramos que el modelo es satisfactorio y sin una pérdida significativa de descriptividad en comparación con el modelo inicial, que era mucho más pesado. Comparando también los índices DIC vemos que los modelos tienen los siguientes

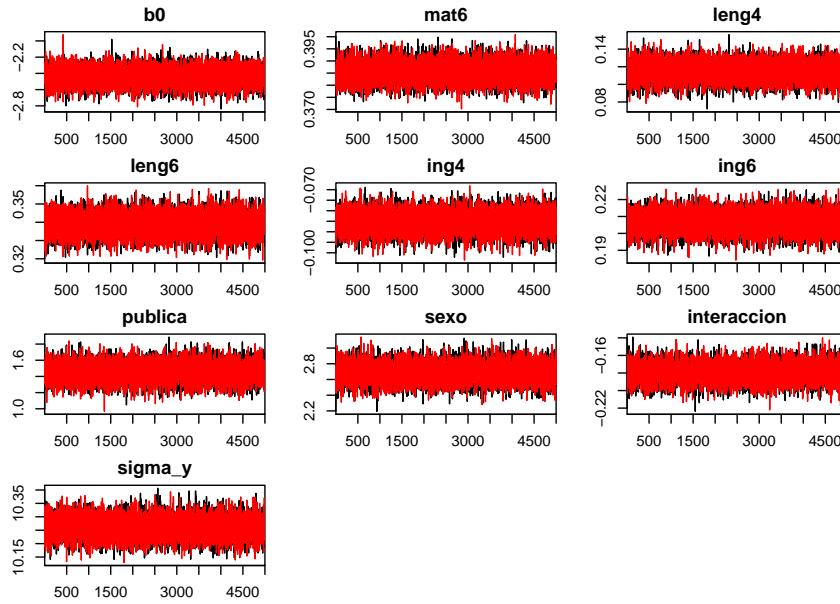
valores $\text{mod_1} = 3.003341 \times 10^4$, $\text{mod_2} = 3.003222 \times 10^4$ and $\text{mod_3} = 3.002972 \times 10^4$. Les recordamos que para el índice DIC más bajo es mejor por lo que confirmamos la elección del modelo 3. Así que vamos a analizarlo en detalle.

4 Implementación del modelo

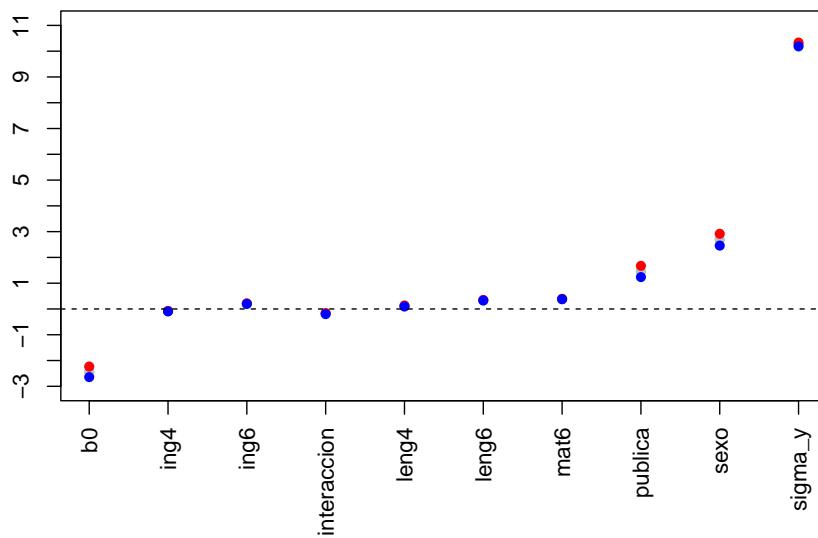
Analicemos ahora el modelo 3 en detalle. Dado que es un modelo más ligero, podemos entrenarlo en una mayor parte de los estudiantes. Usamos un conjunto de entrenamiento esta vez compuesto por el 80% de los estudiantes (alrededor de 37 000).

También en este caso simulamos el modelo a través del paquete JAGS generando MCMCs. Al igual que en los modelos anteriores, aquí también convergen las dos cadenas de prueba.

Los coeficientes resultantes de este modelo son siempre significativos. También notamos que la variabilidad de las distribuciones es menor gracias al mayor tamaño del conjunto de datos de entrenamiento.



Intervalo de confianza del 95% para los coeficientes



Los valores numéricos de los coeficientes y todos los rangos se pueden ver en la tabla de coeficientes del anexo.

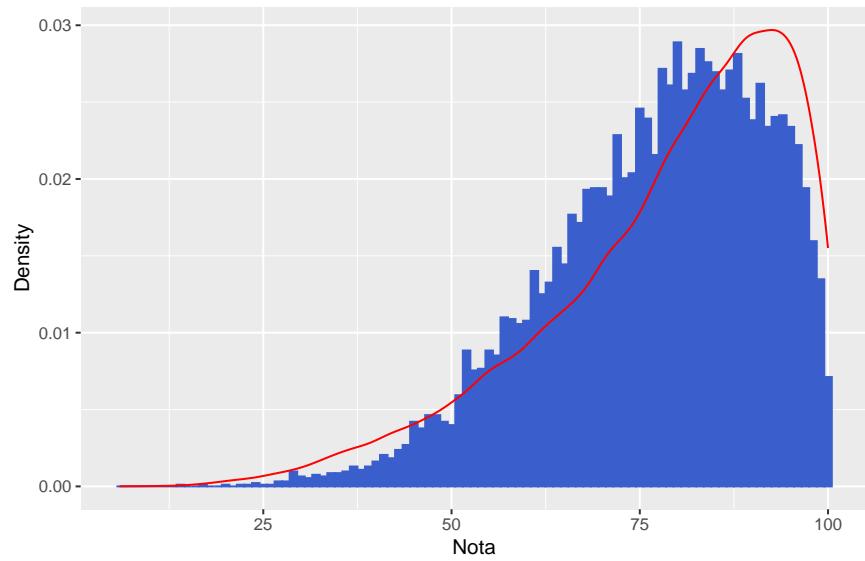
Aquí presentamos un análisis de los valores medios obtenidos de la distribución de los coeficientes:

- $b_0 = -2.44$: El intercepto es negativo, así que supongamos que un estudiante promedio (covariables numéricas = 0), mujer, en una escuela privada obtendrá una calificación más baja que el promedio en el examen mat4.
- $ing4 = -0.09$: Para la calificación de inglés en el examen Quart el coeficiente es negativo por lo que podemos suponer que una persona que está por encima del promedio en inglés tampoco estará por encima del promedio en matemáticas.
- $ing6 = 0.21$: El coeficiente en este caso es positivo, por lo que un alumno que haya obtenido un buen resultado en inglés en el examen Sisè tenderá a tener un buen resultado también en matemáticas.
- $interaccion = -0.18$: El valor del coeficiente es negativo. Esto significa que un estudiante varón tendrá a tener un desempeño más bajo en lenguaje que en matemáticas que el estudiante promedio. Este resultado es aún más interesante cuando se compara con el de leng4 ya que el coeficiente se vuelve negativo en promedio para leng4. De acuerdo con el modelo, por lo tanto, un estudiante varón tiene un desempeño opuesto al promedio en los dos exámenes.
- $leng4 = 0.11$: El coeficiente es positivo, teniendo en cuenta la interacción significa que una niña que en promedio obtiene buenos resultados en lenguaje en el examen de Quart también tendrá un resultado por encima del promedio en el examen de matemáticas del mismo año.
- $leng6 = 0.34$: Siendo positivo el resultado de mate4 es consistente con el de leng 6 con respecto a la media.
- $mate6 = 0.39$: También en este caso el coeficiente es positivo por lo que el resultado de las dos pruebas está de acuerdo. También notamos que el coeficiente es el más grande entre los presentes. La marca de mat6 con respecto a la media es, por tanto, la mayor contribución a la hora de predecir el resultado de mat4.
- $publica = 1.46$: El coeficiente también es positivo en este caso. Al ser una variable binaria podemos considerar que un alumno de un colegio público obtendrá una nota media superior a este valor en comparación con el mismo alumno de un colegio privado.
- $sexo = 2.69$: También en este caso la variable es binaria por lo que observamos que un alumno varón obtendrá un resultado medio 2,5 puntos superior al de una chica.
- $sigma_y = 10.26$: El valor estimado para la varianza de la distribución es bajo y parece ser un valor probable.

Ahora usamos el modelo obtenido para simular los resultados obtenidos en el conjunto de prueba. Usamos los coeficientes promedio para la simulación y extraemos un valor aleatorio de cada distribución generada para un estudiante.

La distribución resultante es la siguiente.

Comparision between predictive posterior and PMAT_4 distribution



Vemos que sigue bastante de cerca la distribución teórica, por lo que estamos satisfechos con este modelo.

5 Conclusions

Explicación rápida para nosotros:

El último modelo es el que se elige. A partir de este modelo podemos responder a nuestro objetivo del proyecto:

Toda esta conclusión debe ser explicada en términos de diferencia con el resultado medio ya que nuestros covariados se refieren a la diferencia con la media

- 1 ¿Es el género un factor relevante en el resultado de matemáticas?

Sí lo es, tenemos que hacer alguna prueba para confirmar esto, pero el coeficiente de sexo muestra que los niños deberían obtener un promedio de 2,52 puntos más que las niñas en el examen.

- 2 ¿Es cierto que las chicas son coherentes en las notas de longitud y matemáticas mientras que los chicos tienden a tener un comportamiento opuesto?

Sí, el coeficiente de leng4 muestra que para las niñas el resultado es el mismo tanto para leng como para matemáticas.

Mientras que si para los niños la interacción reduce este coeficiente y se vuelve negativo ($0.11 - 0.18 = -0.07$), para los niños el resultado de las matemáticas está inversamente correlacionado con el resultado de la lengua.

6 Apéndice

6.1 Código R

6.2 Tablas parámetros estimados

```
## Inference for Bugs model at "4", fit using jags,
## 2 chains, each with 5400 iterations (first 400 discarded)
## n.sims = 10000 iterations saved
##          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%
## b0        -2.44   0.11  -2.64  -2.51  -2.43  -2.36  -2.23
## ing4      -0.09   0.00  -0.09  -0.09  -0.09  -0.08  -0.08
## ing6       0.21   0.01   0.19   0.20   0.21   0.21   0.22
## interaccion -0.18   0.01  -0.20  -0.19  -0.18  -0.17  -0.16
## leng4      0.11   0.01   0.09   0.11   0.11   0.12   0.14
## leng6      0.34   0.01   0.33   0.33   0.34   0.34   0.35
## mat6       0.39   0.00   0.38   0.38   0.39   0.39   0.39
## publica    1.46   0.11   1.24   1.38   1.46   1.53   1.67
## sexo       2.69   0.12   2.46   2.61   2.69   2.77   2.92
## sigma_y    10.26  0.04  10.19  10.23  10.26  10.29  10.33
## deviance   278092.89 4.60 278085.90 278089.49 278092.22 278095.62 278103.62
##          Rhat n.eff
## b0        1 10000
## ing4      1 10000
## ing6       1  3700
## interaccion 1  5200
## leng4      1 10000
## leng6      1 10000
## mat6       1 10000
## publica    1  7200
## sexo       1  6900
## sigma_y    1 10000
## deviance   1     1
##
## 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 = 10.6 and DIC = 278103.5
## DIC is an estimate of expected predictive error (lower deviance is better).
```

7 Referencias