

Bayesian Methods Final Project

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

Aráiztegui, Aránzazu
Ferrara, Lorenzo
Lucchini, Marco

12 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	2
2 Descripción de la base de datos	3
2.1 Definición de las variables utilizadas	3
2.2 Análisis exploratorio de los datos	3
3 Análisis bayesiano	6
3.1 Modelo 1	6
3.1.1 Presentacion del modelo	6
3.1.2 Simulación de modelo	7
3.1.3 Interpretación y validación de los modelos	7
3.2 Modelo 2 sin áreas territoriales	9
3.2.1 Presentacion de los modelos 2	9
3.2.2 Simulación de modelo 2	9
3.2.3 Interpretación y validación de los modelos 2	10
3.3 Modelo 3 se eliminan Servicio Territorial y Ciudad	11
3.3.1 Presentacion de los modelos 3	11
3.3.2 Simulación de modelo 3	12
3.3.3 Interpretación y validación de los modelos 3	12
4 Implementación del modelo	14
5 Conclusions	17
6 Apéndice	18
6.1 Código R	18
6.2 Tablas parámetros estimados	32
7 Referencias	33

diferentes pruebas depende de variables como el sexo, la zona en la que viven o el tipo de escuela a la que asisten.

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. Nuestro primer objetivo va a ser ver si realmente es cierto que “las chicas son más de letras y no tanto de ciencias”, de ser así en el sexo femenino no se dará una diferencia entre el rendimiento en humanidades y matemáticas, personas con bajo rendimiento en humanidades también lo tendrían en matemáticas, pero 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 los siguientes enlaces:

[Avaluació de quart d'Educació Secundària Obligatòria | Dades obertes de Catalunya](#)

[Avaluació de sisè d'educació primària | Dades obertes de Catalunya](#)

Los ficheros de datos contienen 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. Las bases de datos han sido actualizadas el 20 de octubre de 2022 y contienen los datos de 46384 estudiantes. Dada la extensión de la base de datos y el deseo de estudiar la evolución de un mismo alumno en el tiempo se ha realizado una fusión de las dos tablas manteniéndose sólo algunas variables que hemos considerado significativas para evaluar los objetivos propuestos. El código de alumno se ha utilizado para unir las dos tablas y poder hacer estas comparativas.

2.1 Definición de las variables utilizadas

Base de dades

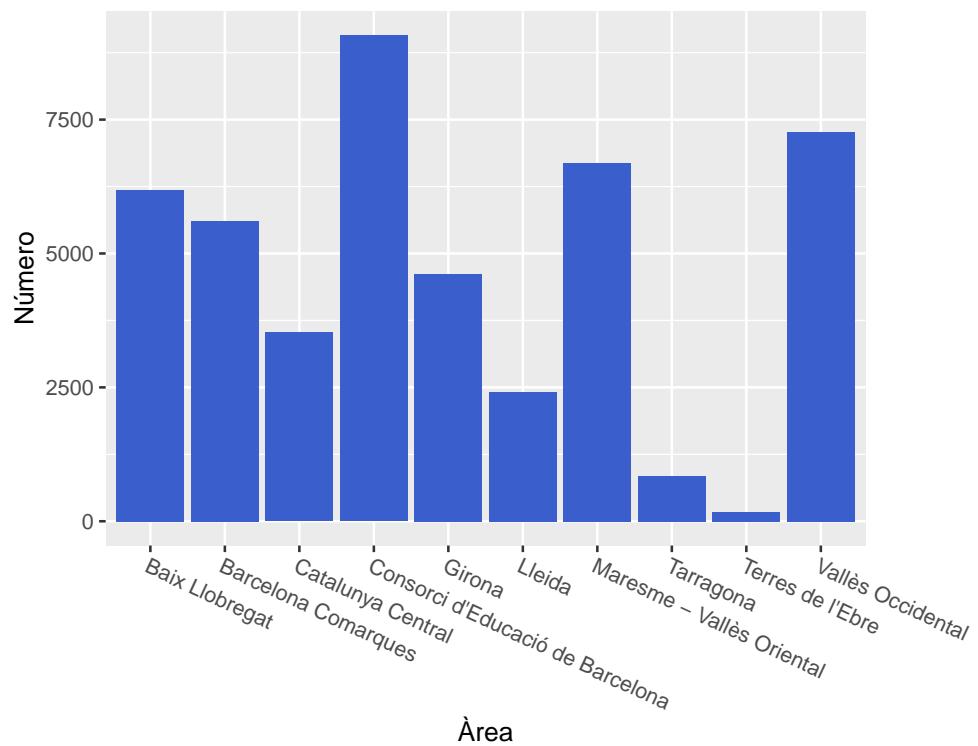
Nom de columna	Descripció	Tipus
PMAT_4	Puntuació global ponderada de competència matemàtica en el examen de Quart	Nombre
PMAT_6	Puntuació global ponderada de competència matemàtica en el examen de Sisè	Nombre
PLENG_4	Puntuació global ponderada de la competència lingüística en llengua catalana y castellana 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.

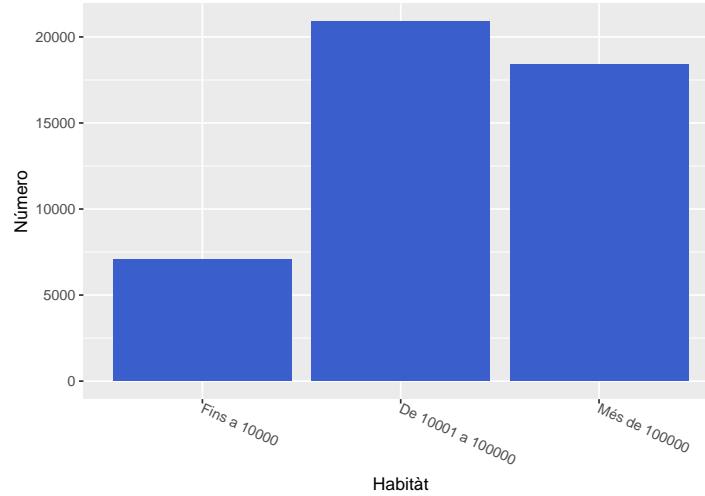
La distribución de alumnos por áreas es la siguiente.

Distribución de alumnos entre las áreas



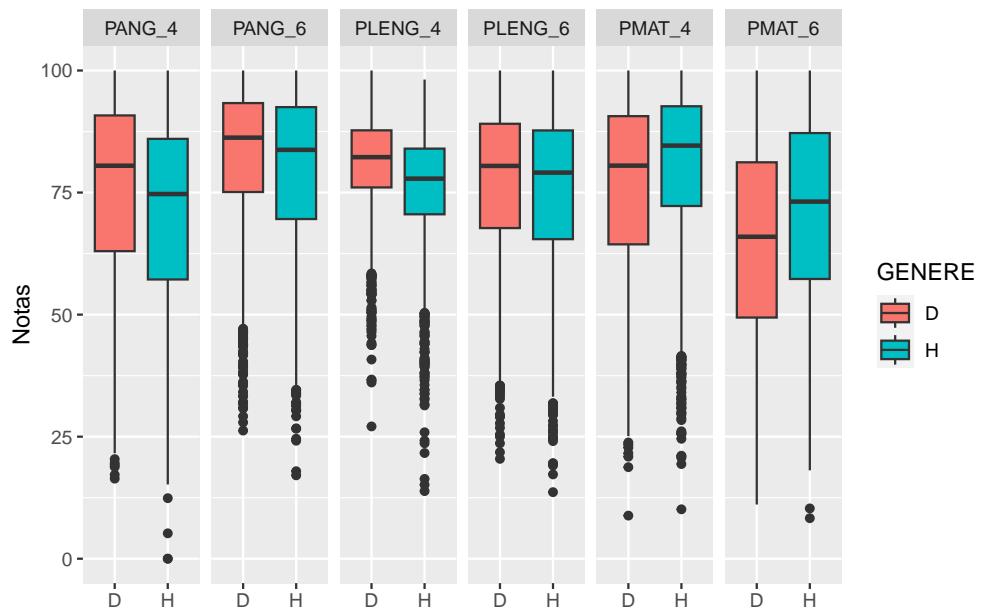
Según el tamaño de la población los datos de los que disponemos se concentran en individuos de ciudades de tamaño medio o grande.

Distribución de alumnos entre las hàbitat



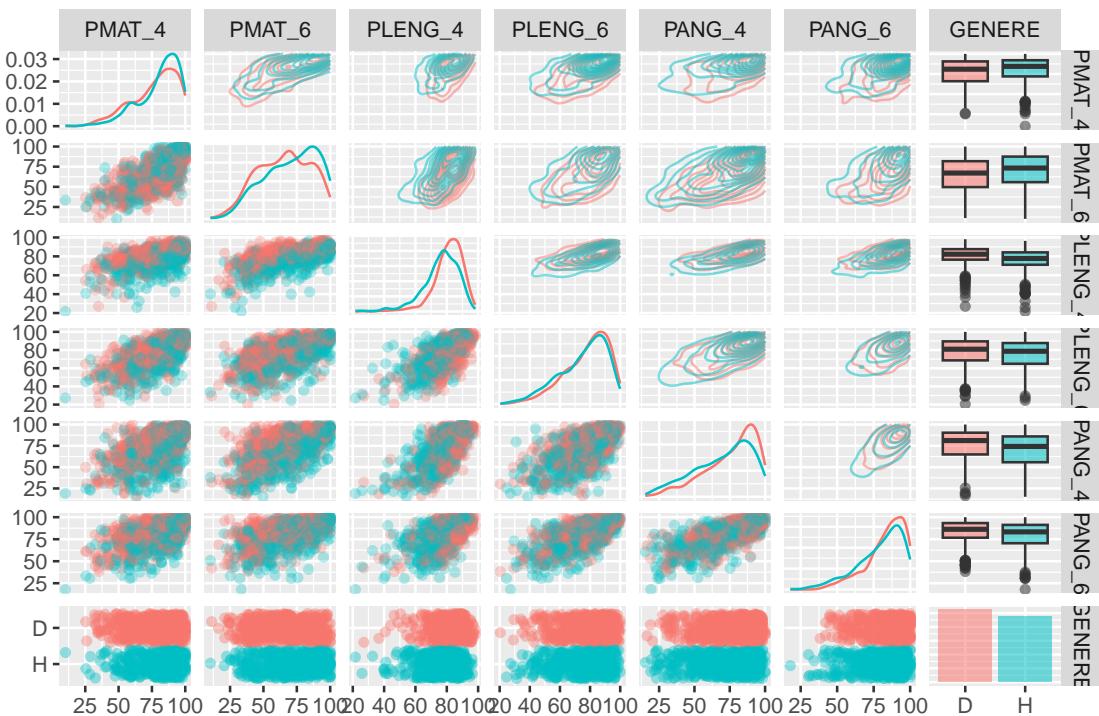
Pasamos ahora a la nota obtenida en el examen de inglés, lengua y matemáticas en el examen de Sisè y Quart. Los grados se dividen entre niños y niñas. Del gráfico suponemos que los chicos tienden a tener peores notas en inglés y lengua mientras que obtienen mejores resultados en matemáticas con respecto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

El objetivo de nuestro modelo será predecir la nota de un estudiante en el examen de competencia matemática de ultimo curso de secundaria basandose en los resultados obtenidos en todas las competencias en sexton de primaria y en las competencias de las materias de lenguas castellana, catalana e inglesa de cuarto de secundaria.

Dado que nuestro objetivo no es solo hacer una predicción de la nota sino si saber si el género va a marcar una diferencia en la nota obtenida estudiaremos no la nota absoluta sino la desviación de cada individuo con respecto a la media:

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

En el modelo se tienen también en cuenta las variables género del estudiante, servei territorial, tamaño de la ciudad al que pertenece y la interacción entre el género y las puntuaciones en las pruebas de lenguas. Pensamos que esta última covariable nos ayudará a explicar si el alumnado tiene rendimientos opuestos en las dos materias con respecto al rendimiento medio.

Este primer modelo es un modelo jerárquico respecto al servicio territorial del que se depende para agregar un efecto aleatorio debido al área. Suponemos que todos los servicios territoriales tienen una distribución normal de media 0 mientras que la varianza cambia para cada una. Consideramos, sin embargo, que los datos se obtienen a partir de una única distribución común a todos los servicios.

3.1.1 Presentacion del modelo

El modelo utilizado es el siguiente.

$$\begin{aligned} y_i &\sim N(\beta_0 + mat_4 * x_i^1 + leng_4 * x_i^2 + leng_6 * x_i^3 + ing_4 * x_i^4 + ing_6 * x_i^5 + publica * x_i^6 \\ &+ ciudPeq * x_i^7 + ciudGrande * x_i^8 + sexo_h * x_i^9 + interaccion * x_i^2 * x_i^9 + G[\text{area}[i]], tau_y) \end{aligned}$$

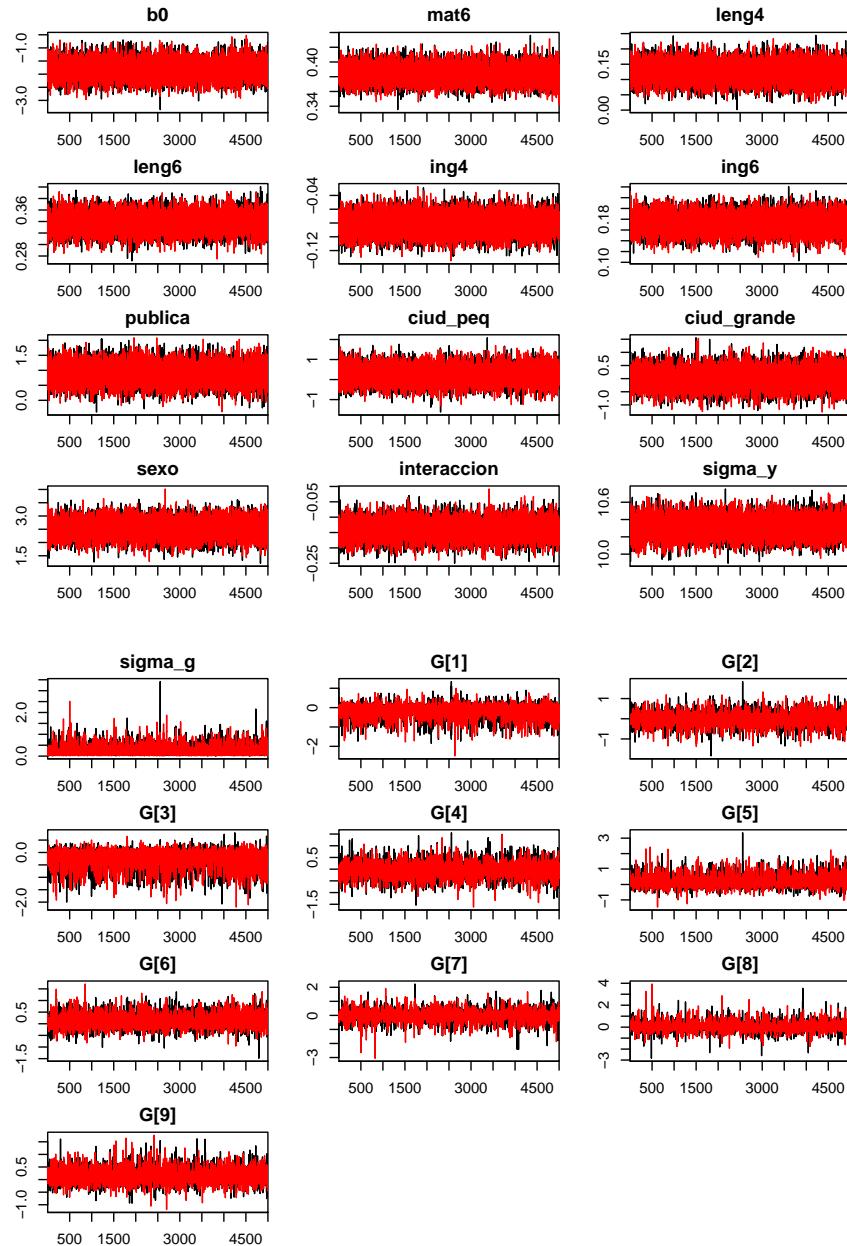
Suponemos que los parámetros para cada variable tienen las siguientes distribuciones. En todos los casos elegimos la distribución normal estándar ya que las variables son un sesgo de la media de cada examen.

$$\begin{aligned} \beta_0 &\sim N(0, 1) \\ mat_4 &\sim N(0, 1) \\ leng_4 &\sim N(0, 1) \\ leng_6 &\sim N(0, 1) \\ ing_4 &\sim N(0, 1) \\ ing_6 &\sim N(0, 1) \\ publica &\sim N(0, 1) \\ ciudPeq &\sim N(0, 1) \\ ciudGrande &\sim N(0, 1) \\ sexo_h &\sim N(0, 1) \quad x^9 \text{ es } 1 \text{ si es hombre} \\ interaccion &\sim N(0, 1) \\ tau_y &\sim \Gamma(0.001, 0.001) \\ sigma_y &= \frac{1}{\sqrt{tau_y}} \\ G_i &\sim N(0, tau_g) \\ tau_g &\sim \Gamma(0.001, 0.001) \\ sigma_g &= \frac{1}{\sqrt{tau_g}} \end{aligned}$$

3.1.2 Simulación de modelo

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

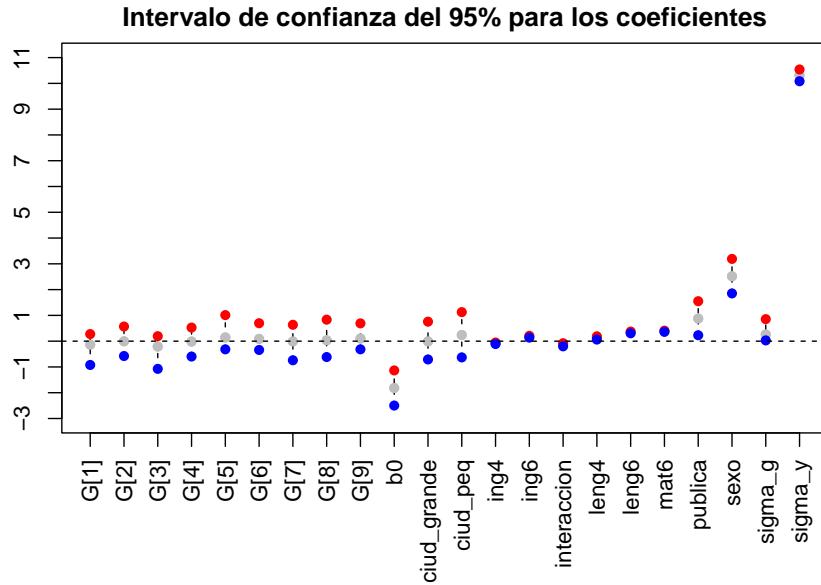
Usando dos cadenas diferentes de 5000 iteraciones de largo y eliminando las primeras 400 iteraciones, obtenemos los siguientes resultados. Como puede verse en los siguientes gráficos las dos cadenas convergen y, por lo tanto, consideramos que el modelo es estable y que se pueden tener en cuenta los resultados como válidos.



3.1.3 Interpretación y validación de los modelos

En el siguiente cuadro se pueden observar las distribuciones de los parámetros obtenidos en nuestro modelo. Si bien las distribuciones relativas a las puntuaciones parecen muy cercanas a cero, todas son significativas.

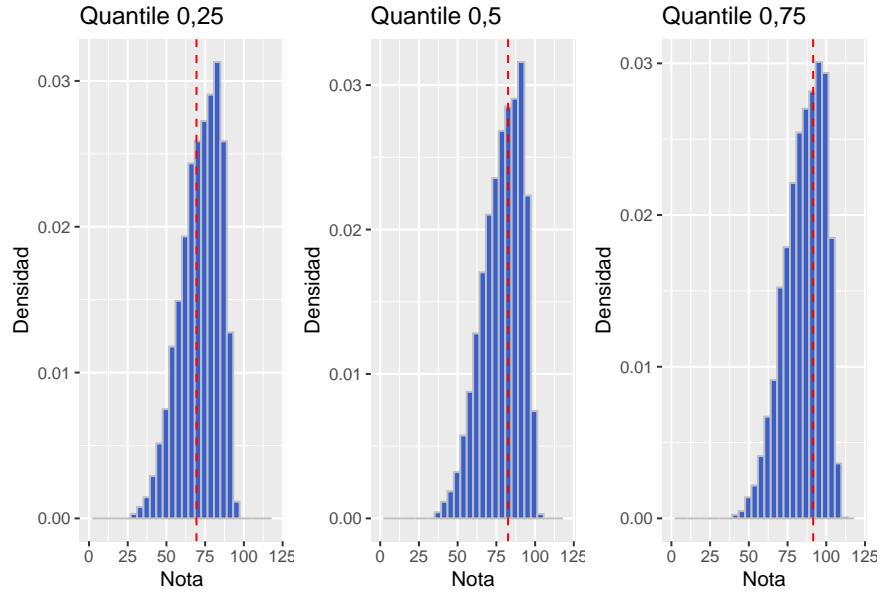
El rango tan reducido se debe al valor de las puntuaciones que varía en un rango de alrededor de ± 30 . Sin embargo, es interesante observar que el efecto aleatorio debido al servicio territorial tiene una distribución centrada alrededor de 0 y que ocurre lo mismo con la variable tamaño de ciudad. Por lo tanto, no podemos considerar su contribución relevante dentro del modelo y, por ello, en los siguientes modelos decidimos eliminar dichas variables ya que no son decisorias para la consecución de nuestro objetivo inicial.



Ahora validamos nuestro modelo usando los cuantiles 25%, 50% y 75% como estadísticas.

De los datos disponibles para nosotros sabemos que el valor de estas estadísticas es $q_{0.25} = 69.39 q_{0.5} = 82.54 q_{0.75} = 91.51$

Simulemos ahora la distribución de estas estadísticas. Creamos \$ 4 000 \$ distribuciones normales teniendo como media el valor obtenido al combinar una n-ésima simulación con un estudiante aleatorio del testset, y como varianza la varianza relativa a los coeficientes utilizados. De cada uno de estos generamos aleatoriamente \$ 1 000 \$ puntos y en estos vectores calculamos los cuantiles requeridos. Luego mostramos los histogramas que representan las distribuciones obtenidas de esta manera.



Como puede verse, los tres estadísticos parecen estar bien dispuestos en la distribución de referencia, aunque el quantile 0,25 se mantienen ligeramente por debajo de los valores reales.

Intentemos eliminar los efectos del servicio territorial para que nuestro modelo sea más interpretable.

3.2 Modelo 2 sin áreas territoriales

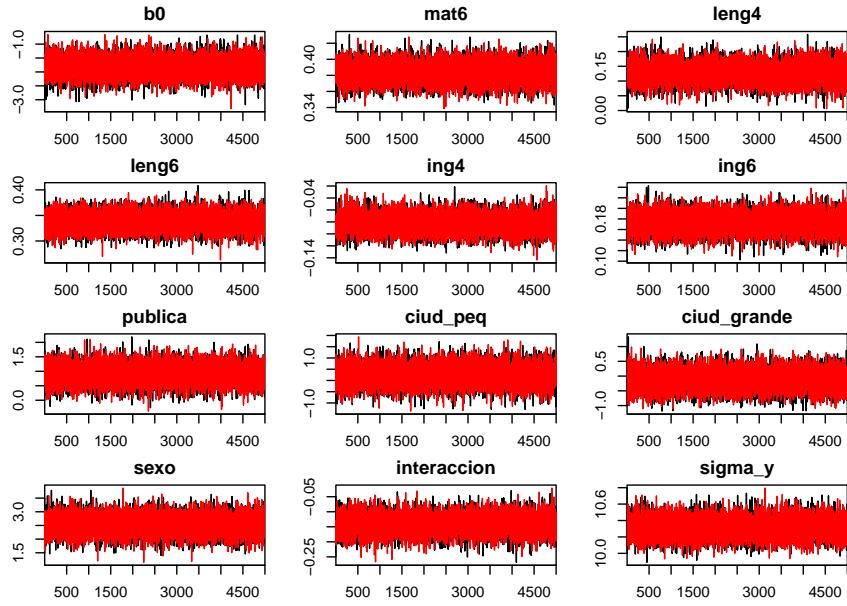
Dado que el parámetro relativo a la variable servicio territorial resultó con una distribución muy centrada alrededor de cero, vamos a eliminar esa covariable del modelo buscando de esta manera mejores resultados y mayor interpretabilidad con respecto al objetivo marcado.

3.2.1 Presentación de los modelos 2

Este segundo modelo utiliza el mismo formato que el modelo anterior pero eliminando 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

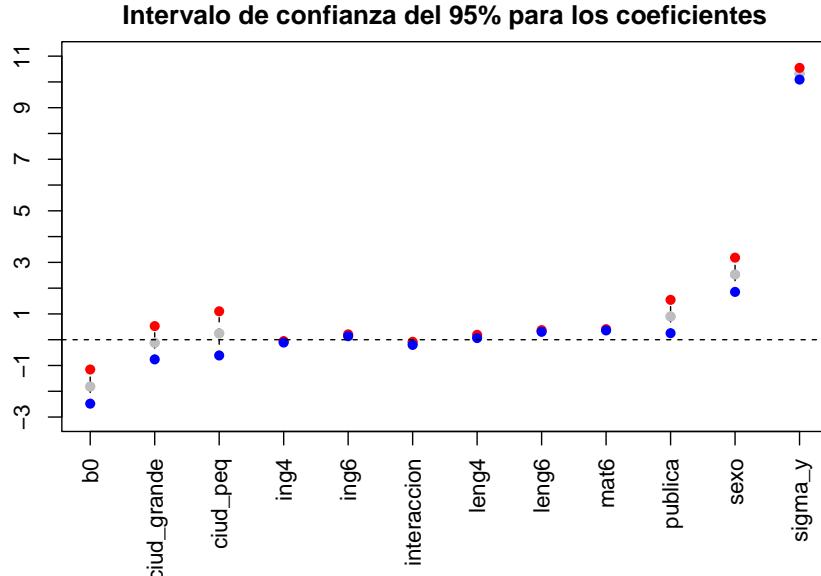
Al igual que hicimos en el modelo anterior, aquí también usamos el paquete JAGS para simular el modelo a través de MCMC. El resultado para los distintos parámetros es el siguiente.



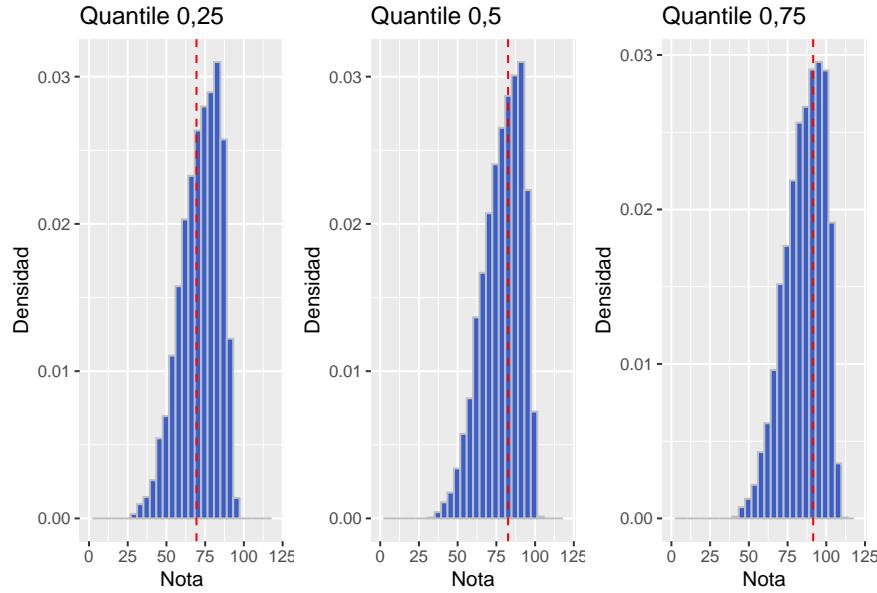
También en este caso las cadenas convergen lo que nos permite validar el modelo y proceder a su análisis.

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 el mismo que el anterior y además en este caso las estadísticas obtenidas de la simulación con el testset 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 se eliminan Servicio Territorial y Ciudad

Reducimos aún más nuestro modelo manteniendo únicamente las calificaciones obtenidas en los exámenes de matemáticas de cuarto curso de secundaria, lengua en los dos cursos (sexto de primaria y cuarto de ESO), inglés en también en sexto y cuarto, el género, el tipo de escuela (pública o privada) y la interacción entre género y nota de idioma en cuarto de secundaria.

3.3.1 Presentacion de los modelos 3

El modelo implementado es el siguiente.

$$y_i \sim N(\beta_0 + mat_4 * x_i^1 + leng_4 * x_i^2 + leng_6 * x_i^3 + ing_4 * x_i^4 + ing_6 * x_i^5 + publica * x_i^6 + sexo_h * x_i^9 + interaccion * x_i^2 * x_i^9, tau_y)$$

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.

$$\begin{aligned} \beta_0 &\sim N(0, 1) \\ mat_4 &\sim N(0, 1) \\ leng_4 &\sim N(0, 1) \\ leng_6 &\sim N(0, 1) \\ ing_4 &\sim N(0, 1) \\ ing_6 &\sim N(0, 1) \\ publica &\sim N(0, 1) \end{aligned}$$

$$sexoh \sim N(0, 1) \quad x^9 \text{ es } 1 \text{ si es hombre}$$

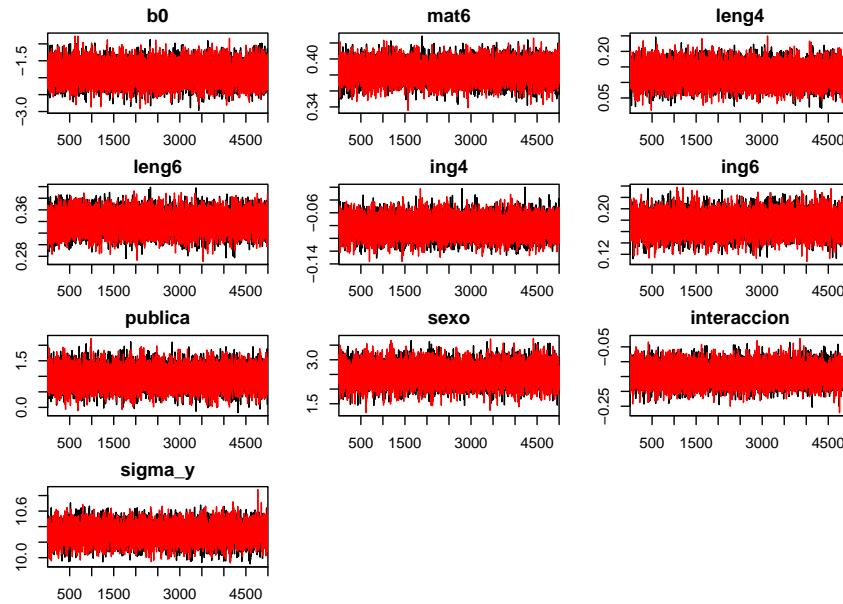
$$interaccion \sim N(0, 1)$$

$$\tau_{\alpha} \sim \Gamma(0.001, 0.001)$$

$$\sigma_y = \frac{1}{\sqrt{\tau_y}}$$

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.

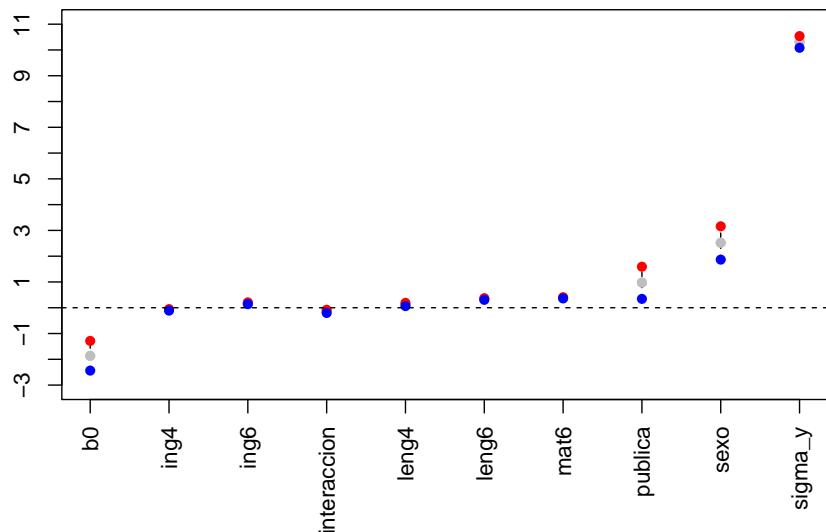


Como puede verse también en este caso el modelo converge.

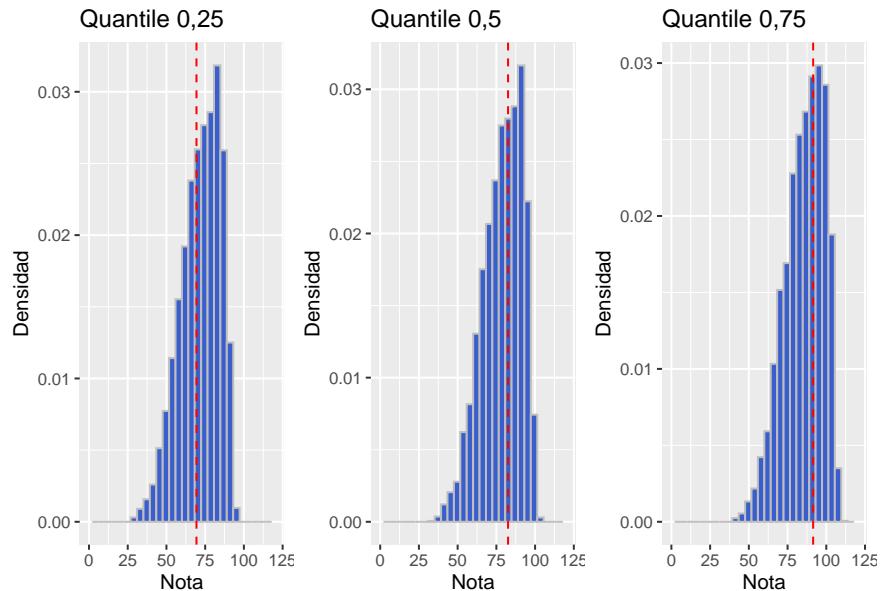
3.3.3 Interpretación y validación de los modelos 3

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.

Intervalo de confianza del 95% para los coeficientes



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 $\text{mod_3} = 3.002972 \times 10^4$.

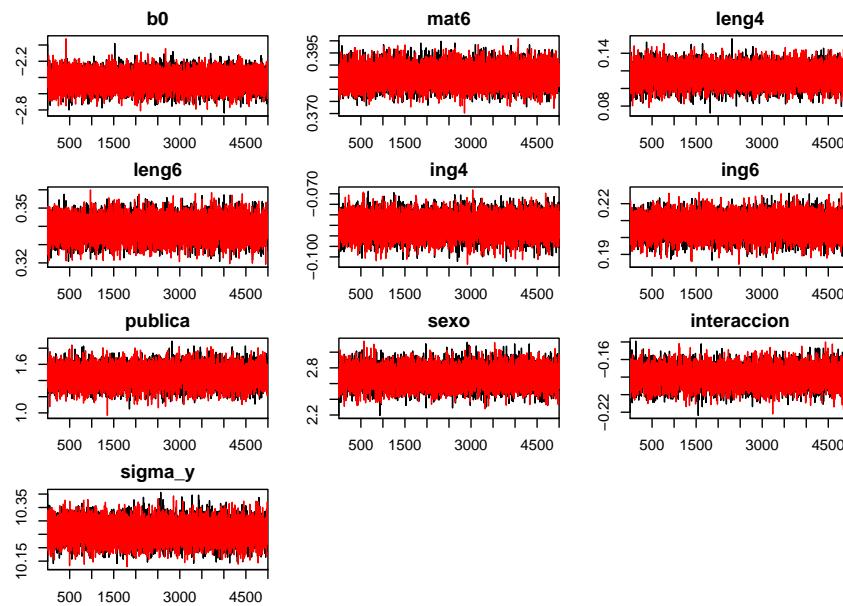
Recordamos que un índice DIC más bajo es mejor por lo que confirmamos la elección del modelo 3 como modelo para evaluar nuestro objetivo inicial. En el siguiente apartado procederemos al análisis en detalle de este modelo.

4 Implementación del modelo

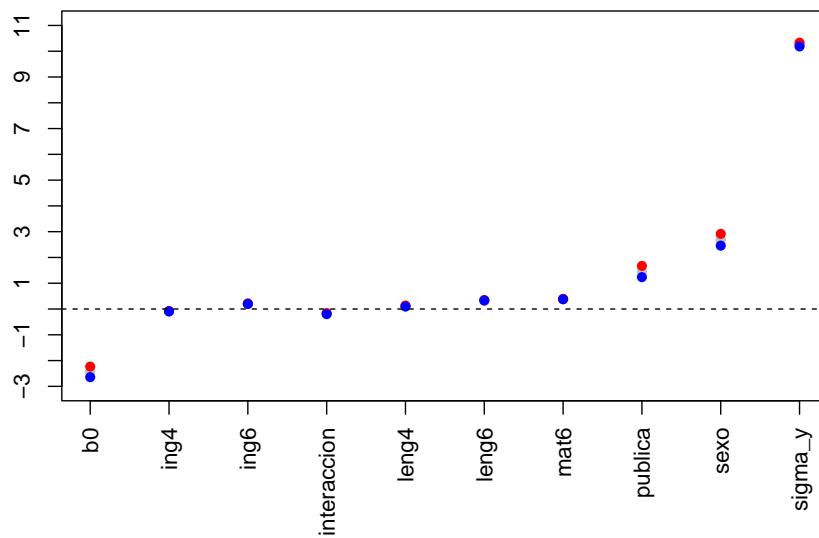
Analicemos ahora el modelo 3 en detalle. Dado que es un modelo más ligero, podemos entrenarlo en una mayor parte de los estudiantes. Usamos un conjunto de entrenamiento esta vez compuesto por el 80% de los estudiantes (alrededor de 37 000).

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



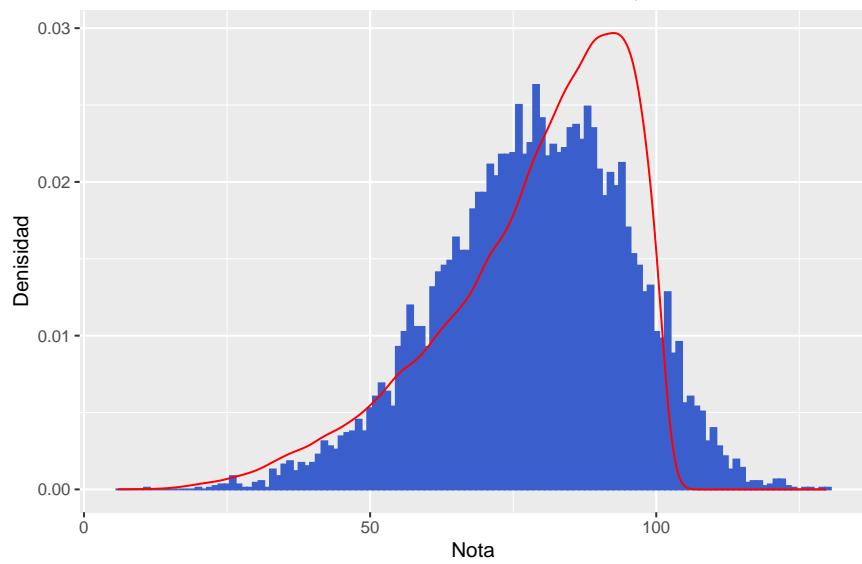
Los valores numéricos de los coeficientes y todos los rangos se pueden ver en la tabla de coeficientes del anexo. Aquí presentamos un análisis de los valores medios obtenidos de la distribución de los coeficientes:

- $\beta_0 = -2.44$: El valor β_0 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 de cuarto el coeficiente también es negativo por lo que podemos suponer que una persona que no 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 de sexto 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 resultado más bajo en lenguas 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 chica que en promedio obtiene buenos resultados en lenguas en el examen de cuarto 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, lo que significa que la puntuación de matemáticas en sexto curso de primaria con respecto a la media es la que contribuye de manera más importante a la hora de predecir el resultado de la puntuación que se obtendrá en la prueba de cuarto de la ESO, lo cual no deja de tener bastante sentido.
- $publica = 1.46$: 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 de sexo masculino obtendrá un resultado medio aproximadamente 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. Usaremos los coeficientes promedio para la simulación y extraeremos un valor aleatorio de cada distribución generada para un estudiante.

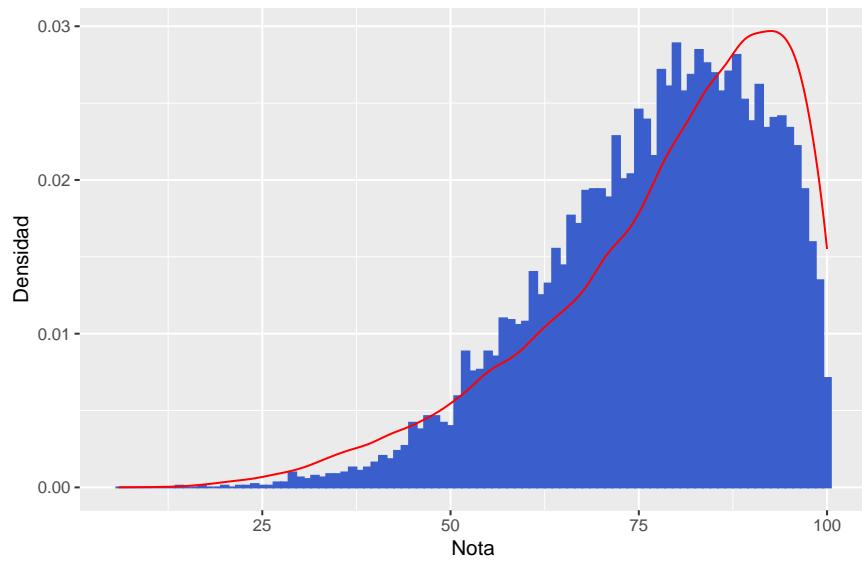
La distribución resultante es la siguiente.

Comparación entre distribución predictiva posterior y PMAT_4



Como podemos ver, la distribución es bastante descriptiva de los datos, pero muchas de las observaciones pronosticadas superan el límite de 100 puntos. Por lo tanto, decidimos usar una distribución normal truncada entre 0 y 100 con la misma media y varianza que se usaron para la predicción anterior para predecir los resultados de los estudiantes.

Comparación entre la distribución posterior predictiva truncada y PMAT_4



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

5 Conclusions

El objetivo principal de nuestro trabajo era poder evaluar cuál es el resultado obtenido por un estudiante de Cataluña en la prueba de evaluación de competencia matemática de cuarto curso de ESO en función de diferentes variables con ello se pretendía saber si existía una diferencia evidente en la puntuación esperada por el simple hecho de pertenecer a un determinado género, de ser así, podríamos afirmar la existencia de un sesgo de género en la materia de matemáticas para la población estudiada.

Después de evaluar diferentes modelos y de descartar variables que inicialmente pensábamos que podían influir en los resultados finales, hemos validado el último modelo como el que mejor explicaría nuestros datos y mejores predicciones podría ofrecer para estimar la puntuación de una persona en la prueba de competencia matemática de Cataluña.

Las puntuaciones se han evaluado como diferencias con respecto a la media y no como puntuaciones absolutas con lo cual la interpretación de los coeficientes se deberá realizar teniendo en cuenta esta circunstancia. Teniendo eso en cuenta procedemos dar respuesta a las preguntas con las que iniciábamos este trabajo:

- ¿Podemos considerar el género un factor relevante en el resultado de las pruebas de competencia matemática al acabar la educación secundaria?

Pensamos que dados los resultados podríamos considerar que sí que sería un factor relevante. Sería interesante realizar alguna otro análisis para confirmar esta creencia pero el coeficiente de sexo muestra que los niños obtendrían un promedio de 2,52 puntos más que las niñas en esta prueba.

- ¿Es cierto que las chicas no demuestran una habilidad más especial en determinados campos mientras que los chicos demuestran mejor competencia en matemáticas que en lenguas?

Pensamos que la respuesta a esta pregunta es que sí. El coeficiente de leng4 muestra que para las niñas la puntuación obtenida en matemáticas sería similar a la obtenida en lenguas mientras que para los niños la interacción reduce este coeficiente y se vuelve negativo ($0.11 - 0.18 = -0.07$), así que podríamos decir que para los niños el resultado de las matemáticas está inversamente correlacionado con el resultado de la lengua y que niños que no tienen buena competencia en materia lingüística sí que la tienen en materia matemática.

Entonces en nuestra opinión para aumentar la tendencia de las chicas a estudiar asignaturas científicas para luego matricularse en carreras STEM, en primer lugar es importante fomentar una cultura de igualdad de género en las escuelas desde temprana edad. Esto puede lograrse a través de campañas educativas, programas de mentoría y liderazgo femenino, y la inclusión de más mujeres en los cuerpos docentes y de investigación.

Además, es esencial romper los estereotipos de género asociados a estos campos. Esto incluye mostrar a las jóvenes ejemplos de mujeres en posiciones de liderazgo en campos científicos, asegurando que los materiales educativos y las actividades extracurriculares promuevan una representación equilibrada de hombres y mujeres en carreras científicas.

6 Apéndice

6.1 Código R

```
load("alumni.Rdata")

##### Histogram by area #####
area = alumni$AREA_TERRITORIAL
ggplot(data.frame(area), aes(x = area)) + geom_bar(fill = "royalblue3") +
  theme(axis.text.x = element_text(angle = 335, vjust = 1,
    hjust = 0)) + theme(plot.margin = margin(0,
  2, 0, 2, "cm")) + ggtitle("Distribución de alumnos entre las áreas",
) + ylab("Número") + xlab("Área")

##### Histogram by city #####
ciudad = factor(alumni$HÀBITAT, levels = c("Fins a 10000",
  "De 10001 a 100000", "Més de 100000"))
ggplot(data.frame(ciudad), aes(x = ciudad)) + geom_bar(fill = "royalblue3") +
  theme(axis.text.x = element_text(angle = 335, vjust = 0.5,
    hjust = 0)) + theme(plot.margin = margin(0,
  1, 0, 0, "cm")) + ggtitle("Distribución de alumnos entre las hàbitat",
) + ylab("Número") + xlab("Habitat")

##### Boxplot #####
set.seed(1234)
df = alumni[sample(1:length(alumni$PMAT_6), replace = FALSE,
  size = 4000), ]

library(tidyverse)

df_tmp = df[, c("PMAT_4", "PMAT_6", "PLENG_4", "PLENG_6",
  "PANG_4", "PANG_6", "GENERE")]
keep = c("PMAT_4", "PMAT_6", "PLENG_4", "PLENG_6",
  "PANG_4", "PANG_6", "GENERE")
df2 = df_tmp[, keep] %>%
  pivot_longer(c(1, 2, 3, 4, 5, 6), names_to = "Test") # select relevant columns

library(ggplot2)
library(GGally)

ggplot(data = df2, aes(x = GENESE, y = value, fill = GENESE)) +
  geom_boxplot() + labs(title = "Comparación de resultados educativos por género",
  y = "Notas", x = "") + facet_wrap(~Test, nrow = 1)

##### GGpairs #####
keep = c("PMAT_4", "PMAT_6", "PLENG_4", "PLENG_6",
  "PANG_4", "PANG_6", "GENERE")
set.seed(1999)
df_sampl <- df[sample(1:dim(df)[1], 1000), keep]
```

```

my_dens <- function(data, mapping, ...) {
  ggplot(data = data, mapping = mapping) + geom_density(...,
    mapping = ggplot2::aes(color = GENERE, alpha = 0.7),
    fill = NA)
}

ggpairs(df_sampl[, ], mapping = ggplot2::aes(color = GENERE,
  alpha = 0.8), diag = list(continuous = my_dens),
  upper = list(continuous = wrap("density", alpha = 0.5),
    combo = "box_no_facet"), lower = list(continuous = wrap("points",
    alpha = 0.3), combo = wrap("dot_no_facet",
    alpha = 0.4)), title = "Comparación de resultados entre sujetos")

##### Preprocess before models #####
names(alumni)[4] = "HABITAT"
alumni = dummy_cols(alumni, select_columns = c("NATURALES", "GENERE", "HABITAT"), remove_first_dummy = T, remove_selected_columns = T)
areas = dummy_cols(data.frame(Area = alumni$AREA_TERRITORIAL), remove_first_dummy = T, remove_selected_columns = T)
names(areas) = levels(alumni$AREA_TERRITORIAL)[-1]
m = dim(areas)[2]

m4 = mean(alumni$PMAT_4)

alumni$PLENG_6 = alumni$PLENG_6 - mean(alumni$PLENG_6)
alumni$PLENG_4 = alumni$PLENG_4 - mean(alumni$PLENG_4)
alumni$PANG_6 = alumni$PANG_6 - mean(alumni$PANG_6)
alumni$PANG_4 = alumni$PANG_4 - mean(alumni$PANG_4)
alumni$PMAT_6 = alumni$PMAT_6 - mean(alumni$PMAT_6)
alumni$PMAT_4 = alumni$PMAT_4 - mean(alumni$PMAT_4)

n = 4000
set.seed(1234)
id_train = sample(1:length(alumni$PMAT_6), replace = FALSE, size = n)
train = alumni[id_train, ]
test = alumni[-id_train, ]

areas_train = areas[id_train, ]
areas_test = areas[-id_train, ]

#####
Model 1 #####
mod.alumni.1 <- "
model {

  for (i in 1:n) {
    y[i] ~ dnorm(b0 + mat6*x1[i] + leng4*x2[i] + leng6*x3[i] + ing4*x4[i] +
      ing6*x5[i] + publica*x6[i] + ciud_peq*x7[i] + ciud_grande*x8[i] + sexo*x9[i] +
      interaccion*x2[i]*x9[i] + sum(G*geo[i,]), tau_y)
  }
}

```

```

b0 ~ dnorm(0, 1)
mat6 ~ dnorm(0, 1)
leng4 ~ dnorm(0, 1)
leng6 ~ dnorm(0, 1)
ing4 ~ dnorm(0, 1)
ing6 ~ dnorm(0, 1)
publica ~ dnorm(0, 1)
ciud_peq ~ dnorm(0, 1)
ciud_grande ~ dnorm(0, 1)
sexo ~ dnorm(0, 1)
interaccion ~ dnorm(0, 1)

tau_y ~ dgamma(0.001, 0.001)
sigma_y <- 1/sqrt(tau_y)

for(i in 1:m) {
  G[i] ~ dnorm(0, tau_g)
}

tau_g ~ dgamma(0.001, 0.001)
sigma_g <- 1/sqrt(tau_g)
}
"

Iter <- 5000
Burn <- 400
Chain <- 2
Thin <- 1

data.1 <- list(y = train$PMAT_4, x1 = train$PMAT_6,
  x2 = train$PLENG_4, x3 = train$PLENG_6, x4 = train$PANG_4,
  x5 = train$PANG_6, x6 = train$NATURALES_A_Pública,
  x7 = train$`HABITAT_Fins_a_10000`, x8 = train$`HABITAT_Més_de_100000`,
  x9 = train$GENERE_H, geo = areas_train, n = dim(train)[1],
  m = m)

parameters.1 <- c("b0", "mat6", "leng4", "leng6", "ing4",
  "ing6", "publica", "ciud_peq", "ciud_grande", "sexo",
  "interaccion", "sigma_y", "sigma_g", "G")

initials.1 = list(list(b0 = 3, mat6 = 1, leng4 = 0.1,
  leng6 = 0, ing4 = 0, ing6 = 0, publica = 0, ciud_peq = 0,
  ciud_grande = 0, sexo = 0, interaccion = 0, tau_y = 4,
  tau_g = 4, G = rep(0, m)), list(b0 = 2, mat6 = 1.5,
  leng4 = 0.2, leng6 = 0, ing4 = 1, ing6 = 1, publica = 1,
  ciud_peq = 1, ciud_grande = 1, sexo = 1, interaccion = 1,
  tau_y = 2, tau_g = 2, G = rep(0, m)))
alumni.sim.1 <- jags(data.1, inits = initials.1, parameters.to.save = parameters.1,
  n.iter = (Iter * Thin + Burn), n.burnin = Burn,
  n.thin = Thin, n.chains = Chain, model = textConnection(mod.alumni.1))

##### Traceplot model 1 #####

```

```

par(mar = c(2, 2, 2, 2))
traceplot(alumni.sim.1, mfrrow = c(4, 3), varname = parameters.1,
  col = c("black", "red"), ask = F)

##### Confidence interval model 1 #####
output = alumni.sim.1
significance = !(output$BUGSoutput$summary[, 3] < 0 &
  output$BUGSoutput$summary[, 7] > 0)

linea = which(rownames(output$BUGSoutput$summary) ==
  "deviance")
confid.inter = output$BUGSoutput$summary[-linea, c(3,
  7, 1)] # small, big, mean

par(mfrrow = c(1, 1), mar = c(6, 3, 2, 1))
plot(1:length(confid.inter[, 1]), confid.inter[, 3],
  col = "grey", pch = 16, axes = F, xlab = "", ylab = "",
  ylim = c(-3, 11), main = "Intervalo de confianza del 95% para los coeficientes")
segments(x0 = 1:length(confid.inter[, 1]), y0 = confid.inter[, 1], x1 = 1:length(confid.inter[, 1]), y1 = confid.inter[, 2], lty = 2)
points(1:length(confid.inter[, 1]), confid.inter[, 3], col = "grey", pch = 16)
box()
axis(1, 1:length(rownames(confid.inter)), rownames(confid.inter),
  las = 2)
axis(2, -3:11, -3:11)
points(1:length(confid.inter[, 1]), confid.inter[, 2], col = "red", pch = 16)
points(1:length(confid.inter[, 1]), confid.inter[, 1], col = "blue", pch = 16)
abline(h = 0, lty = 2)

print(alumni.sim.1, digits = 2)
quant = quantile(alumni$PMAT_4 + m4, c(0.25, 0.5, 0.75))
attach.jags(alumni.sim.1)
coeff = data.frame(b0)
coeff$mat6 = mat6
coeff$leng4 = leng4
coeff$leng6 = leng6
coeff$ing4 = ing4
coeff$ing6 = ing6
coeff$publica = publica
coeff$sexo = sexo
coeff$ciud_peq = ciud_peq
coeff$ciud_grande = ciud_grande
coeff$interaccion = interaccion
coeff$G = G
coeff$sigma_y = sigma_y
coeff$sigma_g = sigma_g
detach.jags()

```

```

##### Prediction model 1 #####
M = 4000
set.seed(2023)
id_test_pred = sample(1:dim(test)[1], M)
test_pred = test[id_test_pred, -c(2, 3)]
areas_test_pred = areas_test[id_test_pred, ]
X = cbind(rep(1, n), as.matrix(test_pred), as.matrix(areas_test_pred),
          as.matrix(test_pred$PLENG_4 * (test_pred$GENERE_H)))

# C = confid.inter[,3] C = C[c('b0', 'mat6',
# 'leng4', 'leng6', 'ing4', 'ing6', 'publica',
# 'sexo', 'ciud_peq', 'ciud_grande', 'G[1]',
# 'G[2]', 'G[3]', 'G[4]', 'G[5]', 'G[6]', 'G[7]',
# 'G[8]', 'G[9]', 'interaccion')]

C = as.matrix(coeff[2000 + 1:M, c(1:10, 12, 11)])

Mu = rowSums(X * C)
Sigma = coeff[1:M, 13]
q_sim = NULL
for (i in 1:4000) {
  y_t = (rnorm(1000, Mu[i], Sigma[i]) + m4)
  q_sim = rbind(q_sim, quantile(y_t, c(0.25, 0.5,
    0.75)))
}
q_sim = data.frame(q_sim)
colnames(q_sim) = c("Q25", "Q50", "Q75")

# q_5 = data.frame(qnorm(0.5, Mu, Sigma) + m4)
# colnames(q_5) = 'Val' q_75 =
# data.frame(qnorm(0.75, Mu, Sigma) + m4)
# colnames(q_75) = 'Val'
per <- quantile(y_t, c(0.25, 0.5, 0.75))

#####
Validation model 1 #####
plot1 = ggplot() + geom_histogram(data = q_sim, aes(x = Q25,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,25", ) + geom_vline(xintercept = quant[1],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot2 = ggplot() + geom_histogram(data = q_sim, aes(x = Q50,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,5", ) + geom_vline(xintercept = quant[2],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot3 = ggplot() + geom_histogram(data = q_sim, aes(x = Q75,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,75", ) + geom_vline(xintercept = quant[3],

```

```

linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

grid.arrange(plot1, plot2, plot3, ncol = 3)

##### Model 2 #####
mod.alumni.2 <- "
model {

  for (i in 1:n) {
    y[i] ~ dnorm(b0 + mat6*x1[i] + leng4*x2[i] + leng6*x3[i] + ing4*x4[i] +
      ing6*x5[i] + publica*x6[i] + ciud_peq*x7[i] + ciud_grande*x8[i] + sexo*x9[i] +
      interaccion*x2[i]*x9[i], tau_y)
  }

  b0 ~ dnorm(0, 1)
  mat6 ~ dnorm(0, 1)
  leng4 ~ dnorm(0, 1)
  leng6 ~ dnorm(0, 1)
  ing4 ~ dnorm(0, 1)
  ing6 ~ dnorm(0, 1)
  publica ~ dnorm(0, 1)
  ciud_peq ~ dnorm(0, 1)
  ciud_grande ~ dnorm(0, 1)
  sexo ~ dnorm(0, 1)
  interaccion ~ dnorm(0, 1)

  tau_y ~ dgamma(0.001, 0.001)
  sigma_y <- 1/sqrt(tau_y)

}

"
Iter <- 5000
Burn <- 400
Chain <- 2
Thin <- 1

data.2 <- with(train, list(y = PMAT_4, x1 = PMAT_6,
  x2 = PLENG_4, x3 = PLENG_6, x4 = PANG_4, x5 = PANG_6,
  x6 = NATURALES_Pública, x7 = `HABITAT_Fins a 10000`,
  x8 = `HABITAT_Més de 100000`, x9 = GENERE_H, n = dim(train)[1]))

parameters.2 <- c("b0", "mat6", "leng4", "leng6", "ing4",
  "ing6", "publica", "ciud_peq", "ciud_grande", "sexo",
  "interaccion", "sigma_y")

initials.2 = list(list(b0 = 3, mat6 = 1, leng4 = 0.1,
  leng6 = 0, ing4 = 0, ing6 = 0, publica = 0, ciud_peq = 0,
  ciud_grande = 0, sexo = 0, interaccion = 0, tau_y = 4),
  list(b0 = 2, mat6 = 1.5, leng4 = 0.2, leng6 = 0,
    ing4 = 1, ing6 = 1, publica = 1, ciud_peq = 1,

```

```

    ciud_grande = 1, sexo = 1, interaccion = 1,
    tau_y = 2))
alumni.sim.2 <- jags(data.2, inits = initials.2, parameters.to.save = parameters.2,
  n.iter = (Iter * Thin + Burn), n.burnin = Burn,
  n.thin = Thin, n.chains = Chain, model = textConnection(mod.alumni.2))

##### Traceplot model 2 #####
par(mar = c(2, 2, 2, 2))
traceplot(alumni.sim.2, mfrw = c(4, 3), varname = parameters.2,
  col = c("black", "red"), ask = F)
print(alumni.sim.2, digits = 2)

##### Confidence interval model 2 #####
output = alumni.sim.2

significance = !(output$BUGSoutput$summary[, 3] < 0 &
  output$BUGSoutput$summary[, 7] > 0)
# print(significance)

linea = which(rownames(output$BUGSoutput$summary) ==
  "deviance")
confid.inter = output$BUGSoutput$summary[-linea, c(3,
  7, 1)]

par(mfrw = c(1, 1), mar = c(6, 3, 2, 1))
plot(1:length(confid.inter[, 1]), confid.inter[, 3],
  col = "grey", pch = 16, axes = F, xlab = "", ylab = "",
  ylim = c(-3, 11), main = "Intervalo de confianza del 95% para los coeficientes")
segments(x0 = 1:length(confid.inter[, 1]), y0 = confid.inter[, 1], x1 = 1:length(confid.inter[, 1]), y1 = confid.inter[, 2], lty = 2)
points(1:length(confid.inter[, 1]), confid.inter[, 3], col = "grey", pch = 16)
box()
axis(1, 1:length(rownames(confid.inter)), rownames(confid.inter),
  las = 2)
axis(2, -3:11, -3:11)
points(1:length(confid.inter[, 1]), confid.inter[, 2], col = "red", pch = 16)
points(1:length(confid.inter[, 1]), confid.inter[, 1], col = "blue", pch = 16)
abline(h = 0, lty = 2)
attach.jags(alumni.sim.2)
coeff2 = data.frame(b0)
coeff2$mat6 = mat6
coeff2$leng4 = leng4
coeff2$leng6 = leng6
coeff2$ing4 = ing4
coeff2$ing6 = ing6
coeff2$publica = publica
coeff2$sexo = sexo

```

```

coeff2$ciud_peq = ciud_peq
coeff2$ciud_grande = ciud_grande
coeff2$interaccion = interaccion
coeff2$sigma_y = sigma_y
detach.jags()

##### Prediction model 2 #####
M = 4000
set.seed(2023)
id_test_pred = sample(1:dim(test)[1], M)
test_pred = test[id_test_pred, -c(2, 3)]
areas_test_pred = areas_test[id_test_pred, ]
X = cbind(rep(1, n), as.matrix(test_pred), as.matrix(test_pred$PLENG_4 *
  (test_pred$GENERE_H)))
C = as.matrix(coeff2[2000 + 1:M, 1:11])

Mu = rowSums(X * C)
Sigma = coeff2[1:M, 12]
q_sim = NULL
for (i in 1:4000) {
  y_t = rnorm(1000, Mu[i], Sigma[i]) + m4
  q_sim = rbind(q_sim, quantile(y_t, c(0.25, 0.5,
    0.75)))
}
q_sim = data.frame(q_sim)
colnames(q_sim) = c("Q25", "Q50", "Q75")

##### Validation model 2 #####
plot1 = ggplot() + geom_histogram(data = q_sim, aes(x = Q25,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,25", ) + geom_vline(xintercept = quant[1],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot2 = ggplot() + geom_histogram(data = q_sim, aes(x = Q50,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,5", ) + geom_vline(xintercept = quant[2],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot3 = ggplot() + geom_histogram(data = q_sim, aes(x = Q75,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,75", ) + geom_vline(xintercept = quant[3],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

grid.arrange(plot1, plot2, plot3, ncol = 3)

##### Model 3 #####

```

```

mod.alumni.3 <- "
model {

  for (i in 1:n) {
    y[i] ~ dnorm(b0 + mat6*x1[i] + leng4*x2[i] + leng6*x3[i] + ing4*x4[i] +
      ing6*x5[i] + publica*x6[i] + sexo*x9[i] + interaccion*x2[i]*x9[i], tau_y)
  }

  b0 ~ dnorm(0, 1)
  mat6 ~ dnorm(0, 1)
  leng4 ~ dnorm(0, 1)
  leng6 ~ dnorm(0, 1)
  ing4 ~ dnorm(0, 1)
  ing6 ~ dnorm(0, 1)
  publica ~ dnorm(0, 1)
  sexo ~ dnorm(0, 1)
  interaccion ~ dnorm(0, 1)

  tau_y ~ dgamma(0.001, 0.001)
  sigma_y <- 1/sqrt(tau_y)

}

"

Iter <- 5000
Burn <- 400
Chain <- 2
Thin <- 1 #per eliminare l'effetto dell'autocorrelazione delle simulazioni

data.3 <- with(train, list(y = PMAT_4, x1 = PMAT_6,
  x2 = PLENG_4, x3 = PLENG_6, x4 = PANG_4, x5 = PANG_6,
  x6 = NATURALES_Pública, x9 = GENERE_H, n = dim(train)[1]))

parameters.3 <- c("b0", "mat6", "leng4", "leng6", "ing4",
  "ing6", "publica", "sexo", "interaccion", "sigma_y")

initials.3 = list(list(b0 = 3, mat6 = 1, leng4 = 0.1,
  leng6 = 0, ing4 = 0, ing6 = 0, publica = 0, sexo = 0,
  interaccion = 0, tau_y = 4), list(b0 = 2, mat6 = 1.5,
  leng4 = 0.3, leng6 = 0, ing4 = 1, ing6 = 1, publica = 1,
  sexo = 1, interaccion = 1, tau_y = 2))

alumni.sim.3 <- jags(data.3, inits = initials.3, parameters.to.save = parameters.3,
  n.iter = (Iter * Thin + Burn), n.burnin = Burn,
  n.thin = Thin, n.chains = Chain, model = textConnection(mod.alumni.3))

##### Traceplot model 3 #####
par(mar = c(2, 2, 2, 2))
traceplot(alumni.sim.3, mfrrow = c(4, 3), varname = parameters.3,
  col = c("black", "red"), ask = F)
print(alumni.sim.3, digits = 2)

##### Confidence intervals model 3 #####

```

```

output = alumni.sim.3
significance = !(output$BUGSoutput$summary[, 3] < 0 &
                 output$BUGSoutput$summary[, 7] > 0)
# print(significance)

linea = which(rownames(output$BUGSoutput$summary) ==
              "deviance")
confid.inter = output$BUGSoutput$summary[-linea, c(3,
                                                7, 1)]

par(mfrow = c(1, 1), mar = c(6, 3, 2, 1))
plot(1:length(confid.inter[, 1]), confid.inter[, 3],
      col = "grey", pch = 16, axes = F, xlab = "", ylab = "",
      ylim = c(-3, 11), main = "Intervalo de confianza del 95% para los coeficientes")
segments(x0 = 1:length(confid.inter[, 1]), y0 = confid.inter[, 1], x1 = 1:length(confid.inter[, 1]), y1 = confid.inter[, 2], lty = 2)
points(1:length(confid.inter[, 1]), confid.inter[, 3], col = "grey", pch = 16)
box()
axis(1, 1:length(rownames(confid.inter)), rownames(confid.inter),
      las = 2)
axis(2, -3:11, -3:11)
points(1:length(confid.inter[, 1]), confid.inter[, 2], col = "red", pch = 16)
points(1:length(confid.inter[, 1]), confid.inter[, 1], col = "blue", pch = 16)
abline(h = 0, lty = 2)
attach.jags(alumni.sim.3)
coeff3 = data.frame(b0)
coeff3$mat6 = mat6
coeff3$leng4 = leng4
coeff3$leng6 = leng6
coeff3$ing4 = ing4
coeff3$ing6 = ing6
coeff3$publica = publica
coeff3$sexo = sexo
coeff3$interaccion = interaccion
coeff3$sigma_y = sigma_y
detach.jags()

##### Prediction model 3 #####
M = 4000
set.seed(2023)
id_test_pred = sample(1:dim(test)[1], M)
test_pred = test[id_test_pred, -c(2, 3, 10, 11)]
X = cbind(rep(1, n), as.matrix(test_pred), as.matrix(test_pred$PLENG_4 *
    (test_pred$GENERE_H)))

C = as.matrix(coeff3[2000 + 1:M, 1:9])

Mu = rowSums(X * C)

```

```

Sigma = coeff3[1:M, 10]
q_sim = NULL
for (i in 1:4000) {
  y_t = (rnorm(1000, Mu[i], Sigma[i]) + m4)
  q_sim = rbind(q_sim, quantile(y_t, c(0.25, 0.5,
    0.75)))
}
q_sim = data.frame(q_sim)
colnames(q_sim) = c("Q25", "Q50", "Q75")

##### Validation model 3 #####
plot1 = ggplot() + geom_histogram(data = q_sim, aes(x = Q25,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,25", ) + geom_vline(xintercept = quant[1],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot2 = ggplot() + geom_histogram(data = q_sim, aes(x = Q50,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,5", ) + geom_vline(xintercept = quant[2],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

plot3 = ggplot() + geom_histogram(data = q_sim, aes(x = Q75,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Quantile 0,75", ) + geom_vline(xintercept = quant[3],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")

grid.arrange(plot1, plot2, plot3, ncol = 3)

##### Preprocessing model 3 big #####
load("alumni.Rdata")
alumni$AREA_TERRITORIAL[alumni$AREA_TERRITORIAL ==
  "Maresme Vallès Oriental" | alumni$AREA_TERRITORIAL ==
  "Maresme-Vallès Oriental"] = "Maresme - Vallès Oriental"

alumni$AREA_TERRITORIAL = factor(alumni$AREA_TERRITORIAL)

names(alumni)[4] = "HABITAT"
alumni = dummy_cols(alumni, select_columns = c("NATURALES", "GENERE", "HABITAT"), remove_first_dummy = T, remove_selected_columns = T)
areas = dummy_cols(data.frame(Area = alumni$AREA_TERRITORIAL),
  remove_first_dummy = T, remove_selected_columns = T)
names(areas) = levels(alumni$AREA_TERRITORIAL)[-1]
m = dim(areas)[2]

m4 = mean(alumni$PMAT_4)

alumni$PLENG_6 = alumni$PLENG_6 - mean(alumni$PLENG_6)

```

```

alumni$PLENG_4 = alumni$PLENG_4 - mean(alumni$PLENG_4)
alumni$PANG_6 = alumni$PANG_6 - mean(alumni$PANG_6)
alumni$PANG_4 = alumni$PANG_4 - mean(alumni$PANG_4)
alumni$PMAT_6 = alumni$PMAT_6 - mean(alumni$PMAT_6)
alumni$PMAT_4 = alumni$PMAT_4 - mean(alumni$PMAT_4)

n = round(0.8 * dim(alumni)[1])
set.seed(1234)
id_train = sample(1:length(alumni$PMAT_6), replace = F,
                  size = n)
train = alumni[id_train, ]
test = alumni[-id_train, ]

areas_train = areas[id_train, ]
areas_test = areas[-id_train, ]

##### Model 3 big #####
mod.alumni.3_big <- "
model {

  for (i in 1:n) {
    y[i] ~ dnorm(b0 + mat6*x1[i] + leng4*x2[i] + leng6*x3[i] + ing4*x4[i] +
      ing6*x5[i] + publica*x6[i] + sexo*x9[i] + interaccion*x2[i]*x9[i], tau_y)
  }

  b0 ~ dnorm(0, 1)
  mat6 ~ dnorm(0, 1)
  leng4 ~ dnorm(0, 1)
  leng6 ~ dnorm(0, 1)
  ing4 ~ dnorm(0, 1)
  ing6 ~ dnorm(0, 1)
  publica ~ dnorm(0, 1)
  sexo ~ dnorm(0, 1)
  interaccion ~ dnorm(0, 1)

  tau_y ~ dgamma(0.001, 0.001)
  sigma_y <- 1/sqrt(tau_y)

}

"
Iter <- 5000
Burn <- 400
Chain <- 2
Thin <- 1 #per eliminare l'effetto dell'autocorrelazione delle simulazioni

data.3 <- with(train, list(y = PMAT_4, x1 = PMAT_6,
                           x2 = PLENG_4, x3 = PLENG_6, x4 = PANG_4, x5 = PANG_6,
                           x6 = NATURALES_Pública, x9 = GENERE_H, n = dim(train)[1]))

parameters.3 <- c("b0", "mat6", "leng4", "leng6", "ing4",

```

```

"ing6", "publica", "sexo", "interaccion", "sigma_y")

initials.3 = list(list(b0 = 3, mat6 = 1, leng4 = 0.1,
  leng6 = 0, ing4 = 0, ing6 = 0, publica = 0, sexo = 0,
  interaccion = 0, tau_y = 4), list(b0 = 2, mat6 = 1.5,
  leng4 = 0.3, leng6 = 0, ing4 = 1, ing6 = 1, publica = 1,
  sexo = 1, interaccion = 1, tau_y = 2))
alumni.sim.3_big <- jags(data.3, inits = initials.3,
  parameters.to.save = parameters.3, n.iter = (Iter *
  Thin + Burn), n.burnin = Burn, n.thin = Thin,
  n.chains = Chain, model = textConnection(mod.alumni.3_big))

##### Traceplot model 3 big #####
par(mar = c(2, 2, 2, 2))
traceplot(alumni.sim.3_big, mfrow = c(4, 3), varname = parameters.3,
  col = c("black", "red"), ask = F)

##### Conf. intervals mod. 3 big #####
output = alumni.sim.3_big
significance = !(output$BUGSoutput$summary[, 3] < 0 &
  output$BUGSoutput$summary[, 7] > 0)
# print(significance)

linea = which(rownames(output$BUGSoutput$summary) ==
  "deviance")
confid.inter = output$BUGSoutput$summary[-linea, c(3,
  7, 1)]

par(mfrow = c(1, 1), mar = c(6, 3, 2, 1))
plot(1:length(confid.inter[, 1]), confid.inter[, 3],
  col = "grey", pch = 16, axes = F, xlab = "", ylab = "",
  ylim = c(-3, 11), main = "Intervalo de confianza del 95% para los coeficientes")
segments(x0 = 1:length(confid.inter[, 1]), y0 = confid.inter[, 1], x1 = 1:length(confid.inter[, 1]), y1 = confid.inter[, 2], lty = 2)
points(1:length(confid.inter[, 1]), confid.inter[, 3], col = "grey", pch = 16)
box()
axis(1, 1:length(rownames(confid.inter)), rownames(confid.inter),
  las = 2)
axis(2, -3:11, -3:11)
points(1:length(confid.inter[, 1]), confid.inter[, 2], col = "red", pch = 16)
points(1:length(confid.inter[, 1]), confid.inter[, 1], col = "blue", pch = 16)
abline(h = 0, lty = 2)

##### Prediciton model 3 big #####
M = dim(test)[1]
set.seed(2023)

```

```

id_test_pred = sample(1:dim(test)[1], M)
test_pred = test[id_test_pred, -c(2, 3, 10, 11)]
X = cbind(rep(1, n), as.matrix(test_pred), as.matrix(test_pred$PLENG_4 *
  (test_pred$GENERE_H)))

C = confid.inter[, 3]
C = C[c("b0", "mat6", "leng4", "leng6", "ing4", "ing6",
  "publica", "sexo", "interaccion")]

y_t = rnorm(M, X %*% C, confid.inter["sigma_y", 3]) +
  m4
# y_t <- rtruncnorm(M, a = -m4, b = 100-m4, X %*%
# C, confid.inter['sigma_y',3]) + m4

##### Prevision model 3 big #####
ggplot() + geom_histogram(data = data.frame(y_t), aes(x = y_t,
  y = ..density..), colour = "royalblue3", fill = "royalblue3",
  binwidth = 1) + geom_density(data = train, aes(x = PMAT_4 +
  m4), color = "red") + ggtitle("Comparación entre distribución predictiva posterior y PMAT_4",
) + ylab("Denisidad") + xlab("Nota")

##### Prevision 3 big truncated #####
y_t <- rtruncnorm(M, a = -m4, b = 100 - m4, X %*% C,
  confid.inter["sigma_y", 3]) + m4
ggplot() + geom_histogram(data = data.frame(y_t), aes(x = y_t,
  y = ..density..), colour = "royalblue3", fill = "royalblue3",
  binwidth = 1) + geom_density(data = train, aes(x = PMAT_4 +
  m4), color = "red") + ggtitle("Comparación entre la distribución posterior predictiva truncada y PMAT_4",
) + ylab("Denisidad") + xlab("Nota")
print(alumni.sim.3_big, digits = 2)

```

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

- Comunicaciones MIPP, *Miradas MIPP*, El sesgo de género en la educación afecta la elección de carrera y el futuro laboral de las mujeres, Julio 5 2021, consultado 11 de Enero, <http://www.mipp.cl/miradas/2021/07/05/sesgo-genero-en-la-educacion/>
- Maria Caprile Elola-Olas, ‘El Sesgo De Género En El Sistema Educativo. Su Repercusion En Las áreas De Matematicas Y Tecnologia En Secundaria (Theano)’, Ministerio de Igualdad, 2008.
- Redacción, *Mastermania*, Las 15 carreras mejor pagadas y más demandadas en 2021 en España, Marzo 2 2021, consultado 10 de Enero, https://www.mastermania.com/noticias_masters/las-15-carreras-mejor-pagadas-y-mas-demandadas-en-2021-en-espana-org-6484.html
- Méndez, I. (2020) Sobre los orígenes del sesgo de género en matemáticas. Papeles de Economía Española nº166
- Marina Velasco, *Huffingtonpost*, Por qué es una buena noticia que las Matemáticas incluyan perspectiva de géne, Agosto 14 2021, consultado 9 de Enero, https://www.huffingtonpost.es/entry/por-que-es-una-buena-noticia-que-las-matematicas-incluyan-perspectiva-de-genero_es_61153844e4b07c14031252de