

Bayesian Methods Final Project

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

Aráiztegui, Aránzazu
Ferrara, Lorenzo
Lucchini, Marco

15 January, 2023



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística



UNIVERSITAT DE
BARCELONA

Índice

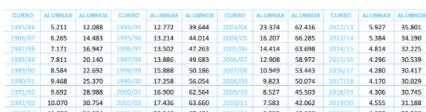
1 Introducción	1
1.1 Descripción del estudio	1
1.2 Objectivos del modelo	2
2 Descripción de la base de datos	3
2.1 Definición de las variables utilizadas	3
2.2 Análisis exploratorio de los datos	3
3 Análisis bayesiano	6
3.1 Modelo 1	6
3.1.1 Presentación del modelo	6
3.1.2 Simulación del modelo	7
3.1.3 Interpretación y validación de los modelos	8
3.2 Modelo 2: sin Servicios Territoriales	9
3.2.1 Presentación del modelo 2	9
3.2.2 Simulación del modelo 2	9
3.2.3 Interpretación y validación del modelo 2	10
3.3 Modelo 3: sin Servicios Territoriales ni Localidades	11
3.3.1 Presentación del modelo 3	11
3.3.2 Simulación de modelo 3	12
3.3.3 Interpretación y validación del modelo 3	12
4 Implementación del modelo definitivo	14
5 Conclusiones	17
6 Apéndice	18
6.1 Código R	18
6.2 Tablas parámetros estimados	30
7 Referencias	31

1 Introducción

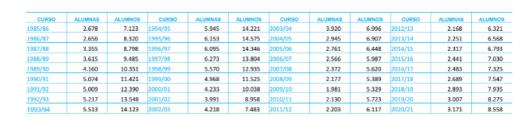
Según los últimos datos del Ministerio de Educación y Formación Profesional del Gobierno de España, la evolución del número de mujeres que se matriculan en estudios de grado de Informática, Matemáticas, Estadística o Física sigue una tendencia a la baja con respecto al número de hombres matriculados a lo largo de los años.

Además, en algunas titulaciones como Matemáticas o Estadística se puede apreciar que la diferencia entre hombres y mujeres matriculados ha aumentado en los últimos años.

Evolución y distribución porcentual del alumnado matriculado en Ciclo Corto, Ciclo Largo y Grado Universitario en Universidades públicas por sexo, modalidad (presencial y no presencial) y campo de estudio: Informática. Cursos 1985/86 a 2020/21



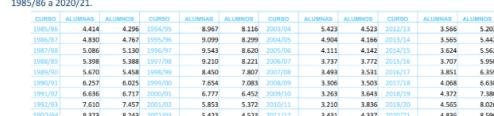
Evolución y distribución porcentual del alumnado matriculado en Ciclo Largo y Grado Universitario en Universidades públicas por sexo, modalidad (presencial y no presencial) y campo de estudio: Física. Cursos 1985/86 a 2020/21



INFORME ALIANZA STEAM / 1
RADIOGRAFÍA DE LA BRECHA DE GÉNERO
EN LA FORMACIÓN STEM

FUENTE: ELABORACIÓN UNIDAD DE IGUALDAD DEL MISP A PARTIR DE LAS ESTADÍSTICAS DE LAS ENSEÑANZAS UNIVERSITARIAS

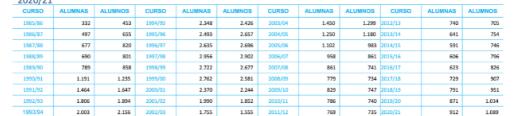
Gráfico 124. Evolución y distribución porcentual del alumnado matriculado en Ciclo Largo y Grado Universitario en Universidades públicas por sexo, modalidad (presencial y no presencial) y campo de estudio: Matemáticas. Cursos 1985/86 a 2020/21.



INFORME ALIANZA STEAM / 1
RADIOGRAFÍA DE LA BRECHA DE GÉNERO
EN LA FORMACIÓN STEM

FUENTE: ELABORACIÓN UNIDAD DE IGUALDAD DEL MISP A PARTIR DE LAS ESTADÍSTICAS DE LAS ENSEÑANZAS UNIVERSITARIAS

Evolución y distribución porcentual del alumnado matriculado en Ciclo Corto, solo Segundo Ciclo y Grado Universitario en Universidades públicas por sexo, modalidad (presencial y no presencial) y campo de estudio: Estadística. Cursos 1985/86 a 2020/21



INFORME ALIANZA STEAM / 1
RADIOGRAFÍA DE LA BRECHA DE GÉNERO
EN LA FORMACIÓN STEM

FUENTE: ELABORACIÓN UNIDAD DE IGUALDAD DEL MISP A PARTIR DE LAS ESTADÍSTICAS DE LAS ENSEÑANZAS UNIVERSITARIAS

Diversos estudios sugieren que este sesgo de género presente en la educación universitaria conlleva asimismo un sesgo en el ámbito laboral. Esta escasa presencia de mujeres en los estudios relacionados con las matemáticas repercute directamente en unas diferencias salariales importantes, ya que son las profesiones relacionadas con las materias del ámbito STEAM las que presentan mejores salidas profesionales y, por tanto, mejores salarios.

Pensamos que una de las posibles causas que determinan la no elección de estudios superiores relacionados con las matemáticas por parte de las mujeres podría ser la diferencia en el rendimiento académico en esta materia en las etapas escolares previas y, más concretamente, en las de escolarización obligatoria: primaria y secundaria obligatoria.

1.1 Descripción del estudio

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 realiza una prueba de competencias y conocimientos básicos en las áreas lingüísticas, matemática y científico-tecnológica en sexto de primaria y cuarto de ESO, últimos cursos de la educación primaria y secundaria obligatoria.

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.

Nuestro estudio se va a centrar en determinar si las puntuaciones que obtiene el alumnado de Cataluña en las diferentes pruebas depende de variables como el género, la zona en la que viven o el tipo de escuela a la que asisten.

1.2 Objectivos del modelo

Este trabajo presenta como objetivo principal analizar si existe un sesgo de género en los resultados obtenidos en la competencia matemática con respecto al sexo y a las competencias lingüísticas. 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 relación al tipo de centro educativo o al tamaño de la población. Nuestro primer objetivo consistirá en determinar si realmente es cierto el estereotipo de que “las chicas son de letras y los chicos son de ciencias”. De ser así, en el sexo femenino no debería apreciarse una diferencia entre el rendimiento en lenguas y en matemáticas: las chicas con bajo o alto rendimiento en lenguas también presentarán el mismo resultado en matemáticas; en cambio, en el sexo masculino, los chicos con bajo rendimiento en lenguas deberían obtener mejores resultados en matemáticas.

Si ampliamos los datos y vemos la tabla de resultados para un mismo individuo en sexto de primaria y cuarto de ESO, podríamos establecer un segundo objetivo: comprobar si se mantienen los resultados en ambos sexos o si, por el contrario, aparecen diferencias significativas con respecto al rendimiento en el área de matemáticas.

2 Descripción de la base de datos

La base de datos utilizada es una mezcla de los datos ofrecidos para cuarto de ESO y para sexto de primaria. La razón de realizar esta fusión 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 de ESO y sexto 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 muestran registros de 46384 estudiantes.

Dada la extensión de la base de datos, hemos decidido mantener 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 realizar diferentes comparativas.

2.1 Definición de las variables utilizadas

Base de datos

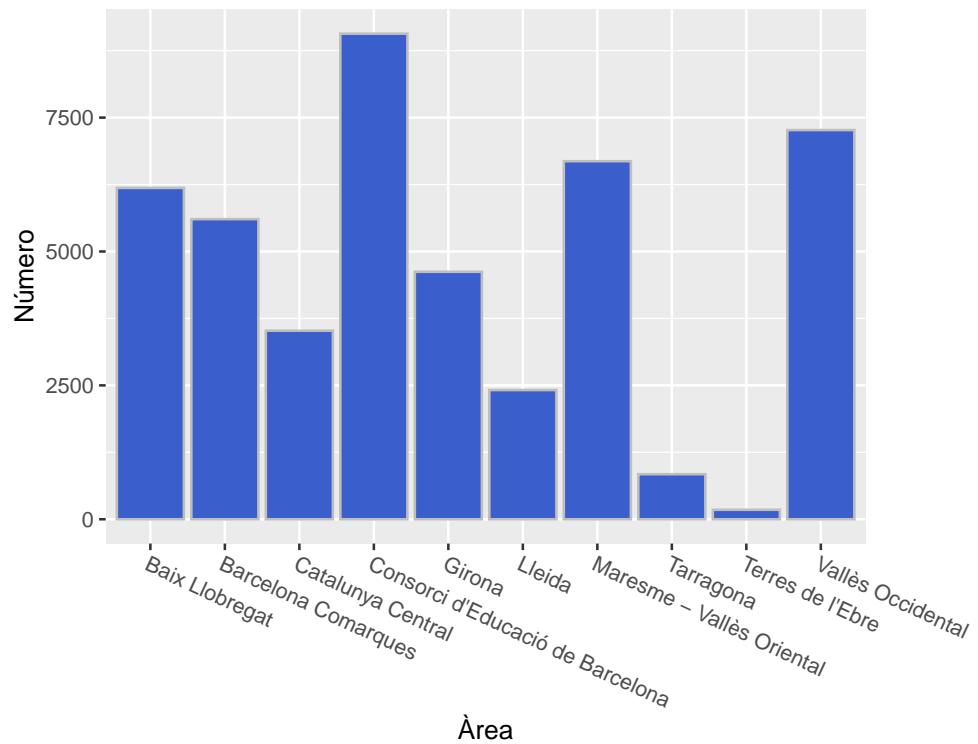
Nombre de columna	Descripción	Tipo
PMAT_4	Puntuación global ponderada de competencia matemática en el examen de Cuarto	Número
PMAT_6	Puntuación global ponderada de competencia matemática en el examen de Sexto	Número
PLENG_4	Puntuación global ponderada de la competencia lingüística en lenguas catalana y castellana en el examen de Cuarto	Número
PLENG_6	Puntuación global ponderada de la competencia lingüística en lenguas catalana y castellana en el examen de Sexto	Número
PANG_4	Puntuación global ponderada de la competencia lingüística en lengua inglesa en el examen de Cuarto	Número
PANG_6	Puntuación global ponderada de la competencia lingüística en lengua inglesa en el examen de Sexto	Número
GENERE	Género del alumno/a que se presenta a la evaluación	Texto
AREA_TERRITORIAL	Servicio territorial donde se encuentra el centro del alumno que se presenta a la evaluación	Texto
NATURALESA	Determina si el centro del alumno es público, privado o concertado	Texto
HÀBITAT	Municipios por tramos de población	Texto

2.2 Análisis exploratorio de los datos

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

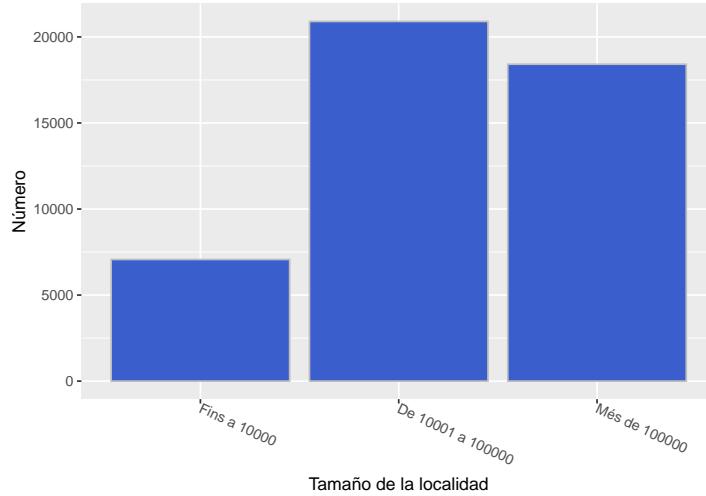
La distribución de los alumnos por áreas es la siguiente:

Distribución de los alumnos entre las áreas



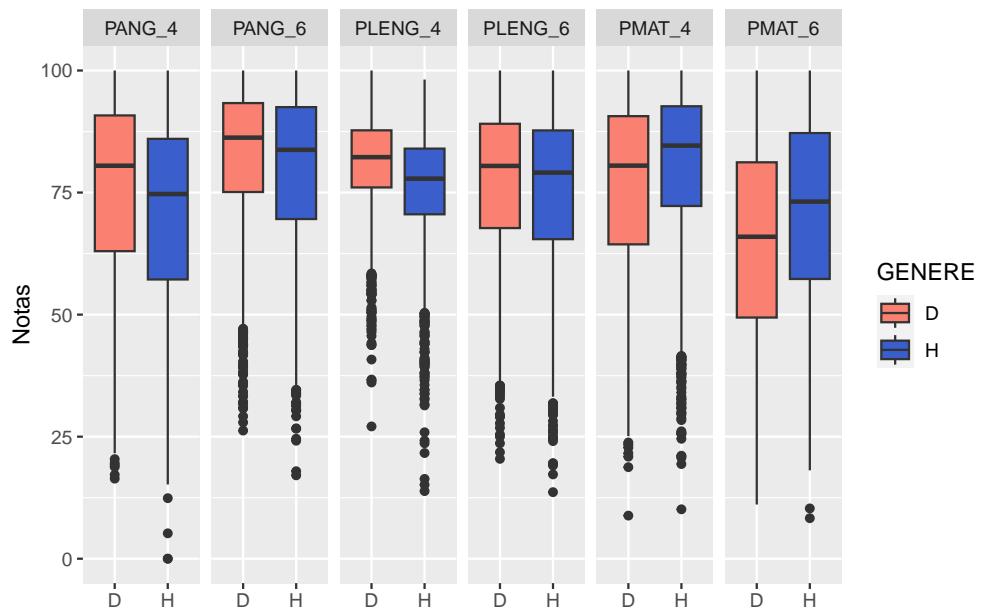
Según el tamaño de la población, los datos se concentran en los individuos de ciudades de tamaño medio o grande.

Distribución de los alumnos según tamaño de la localidad



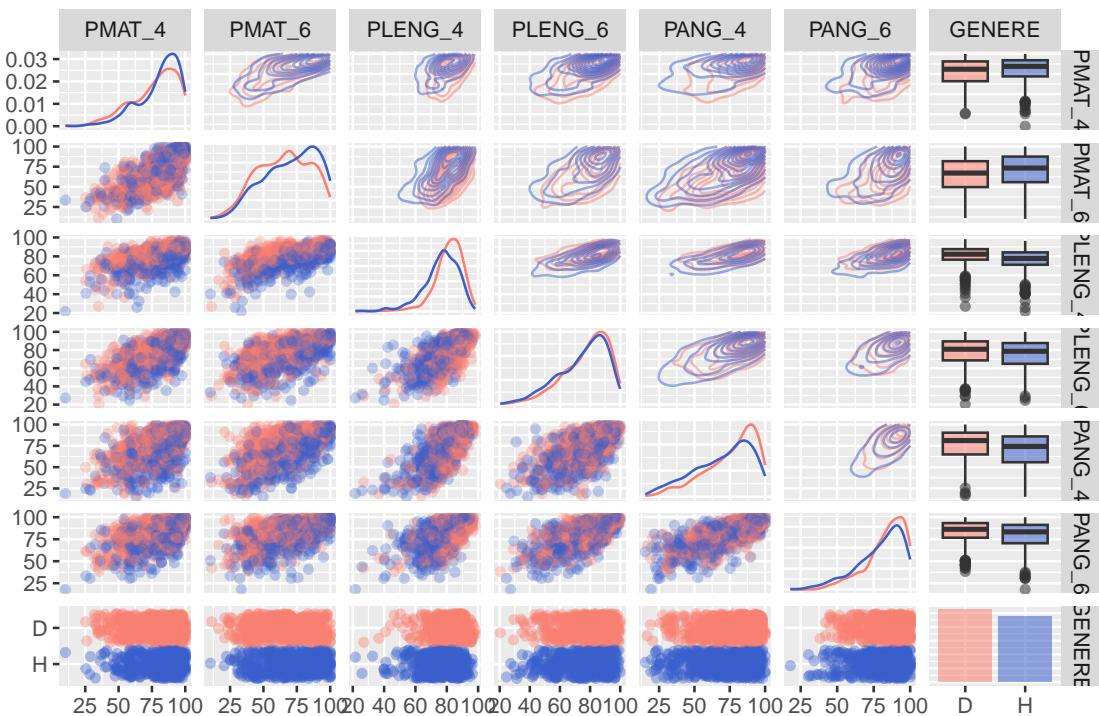
A continuación analizamos las notas obtenidas en las pruebas de inglés, lenguas y matemáticas. Al evaluar las puntuaciones, se ha tenido en cuenta la variable sexo y los resultados se muestran para niños y niñas por separado. Del gráfico que aparece a continuación se puede deducir que los chicos parecen tener peores resultados medios en inglés y lenguas mientras que estarían obteniendo mejores puntuaciones en matemáticas con respecto a las chicas.

Comparación de resultados por género



En el siguiente gráfico se aprecian las diferencias entre las calificaciones en matemáticas, inglés y lenguas de cada sexo en función 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 con buenos resultados en sexto de primaria también obtiene buenos resultados al acabar la secundaria. Asimismo, cabe destacar la distribución de las frecuencias por género, en especial en el caso de matemáticas, donde se observa una diferencia más evidente entre chicos y chicas.

Comparación de los resultados entre individuos



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 del último curso de secundaria basándonos en los resultados obtenidos en todas las competencias en sexto de primaria y en las competencias de las materias de lenguas castellana, catalana e inglesa de cuarto de ESO. Dado que nuestro objetivo no es solo hacer una predicción de la nota, sino saber si el género va a determinar el resultado obtenido, no estudiaremos las puntuaciones absolutas, sino la desviación de cada individuo con respecto a la media:

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

En el modelo se tienen en cuenta igualmente las variables siguientes: género del estudiante, servicio territorial, tamaño de la localidad 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 lenguas y matemáticas 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 uno. Consideramos, sin embargo, que los datos se obtienen a partir de una única distribución común a todos los servicios.

3.1.1 Presentación del modelo

El modelo utilizado es el siguiente:

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

Suponemos que los parámetros para cada variable presentan las siguientes distribuciones. En todos los casos elegimos la distribución normal estándar, ya que las variables están construidas como diferencias respecto de la media para cada prueba.

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

$$G_i \sim N(0, \tau_g)$$

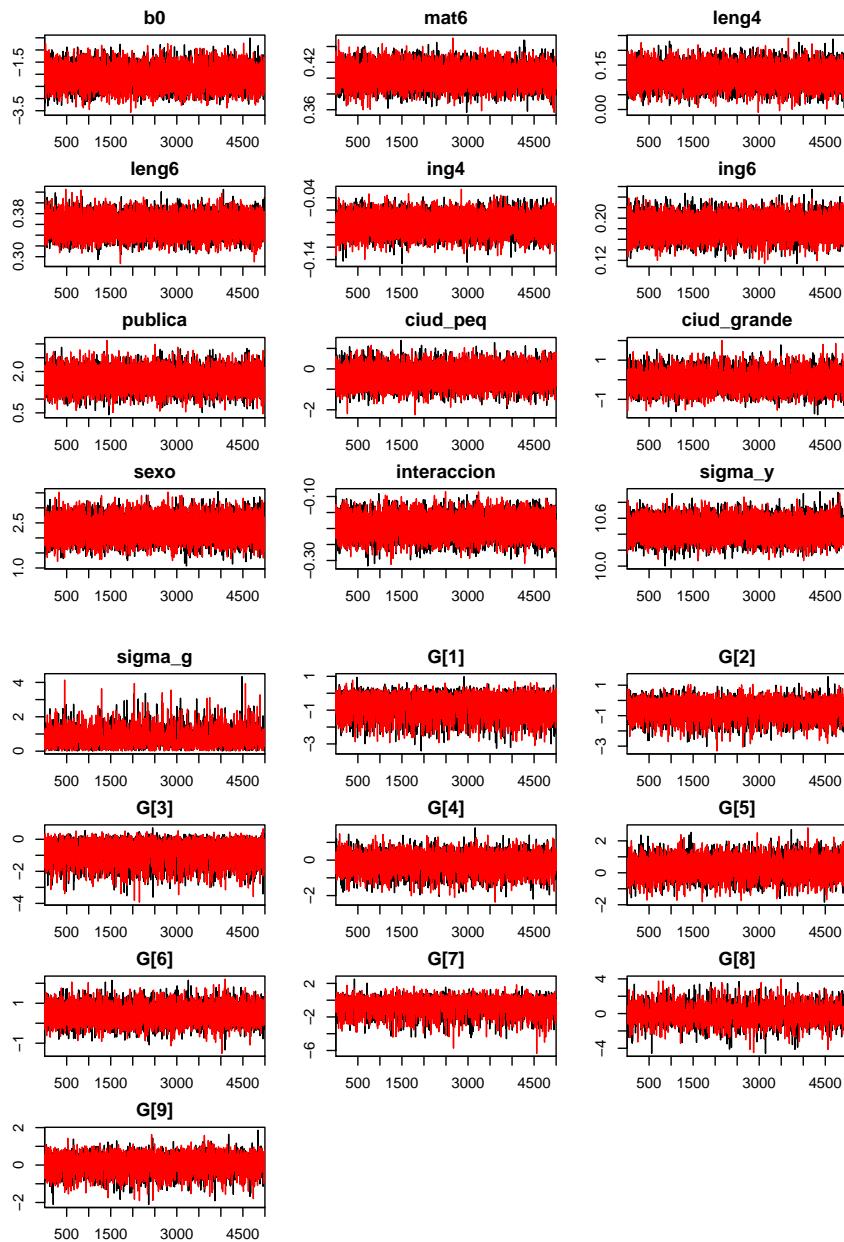
$$\tau_g \sim \Gamma(0.001, 0.001)$$

$$\sigma_g = \frac{1}{\sqrt{\tau_g}}$$

3.1.2 Simulación del modelo

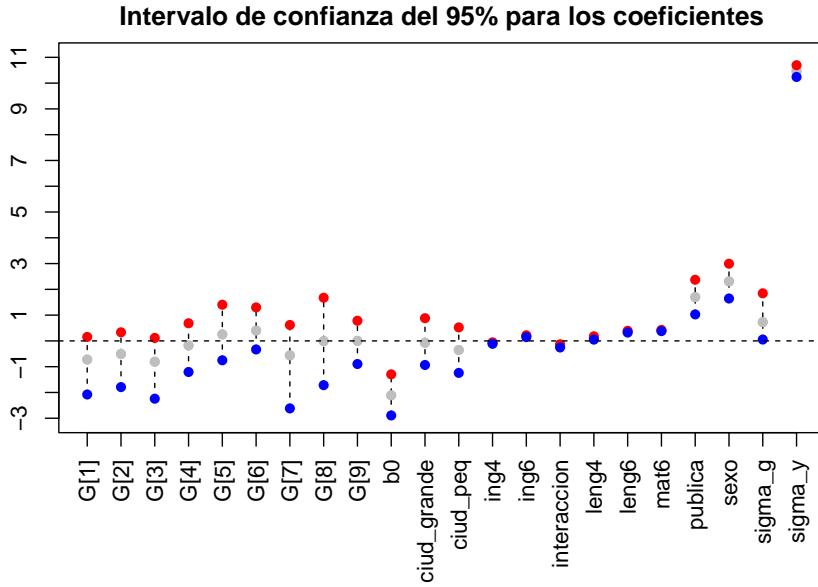
Ejecutamos nuestro modelo en un conjunto de datos de prueba de 4000 estudiantes usando el paquete JAGS de R para realizar las simulaciones. Se utilizan algoritmos de MCMC, ya que la realización del problema analíticamente resultaría casi imposible.

Utilizando dos cadenas diferentes de 5000 iteraciones y eliminando las primeras 400 iteraciones, obtenemos los siguientes resultados. Como puede verse en los 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

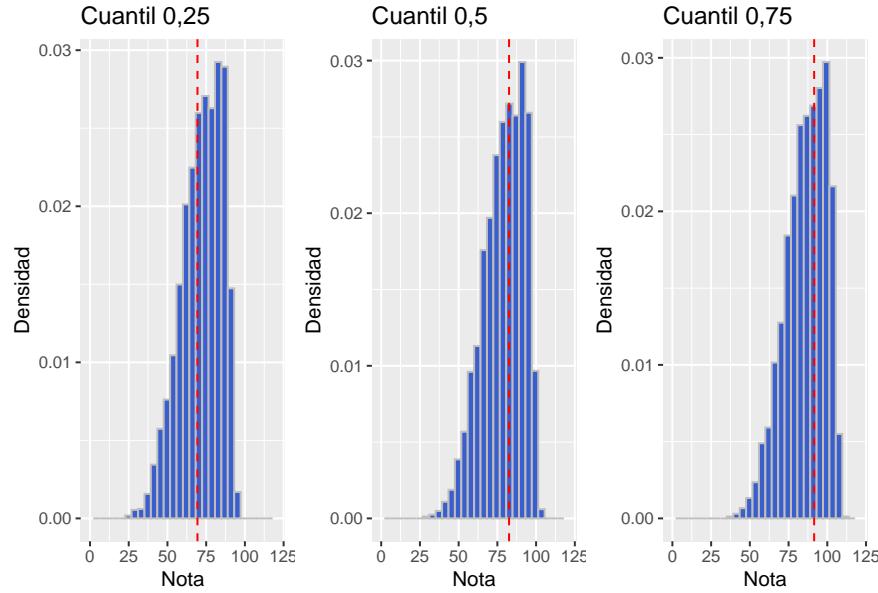
En el siguiente cuadro se pueden observar las distribuciones de los parámetros obtenidos en nuestro modelo. Si bien las distribuciones relativas a las puntuaciones parecen muy cercanas a cero, todas resultan significativas. El rango tan reducido se debe al valor de las puntuaciones, que varía en torno a ± 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 la localidad. 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 resultan decisivas para la consecución de nuestro objetivo inicial.



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

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

Simulemos ahora la distribución de estos estadísticos. Creamos 4000 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 1000 valores y con ellos calculamos los cuantiles requeridos. A continuación mostramos los histogramas que representan las distribuciones obtenidas de esta manera.



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

3.2 Modelo 2: sin Servicios Territoriales

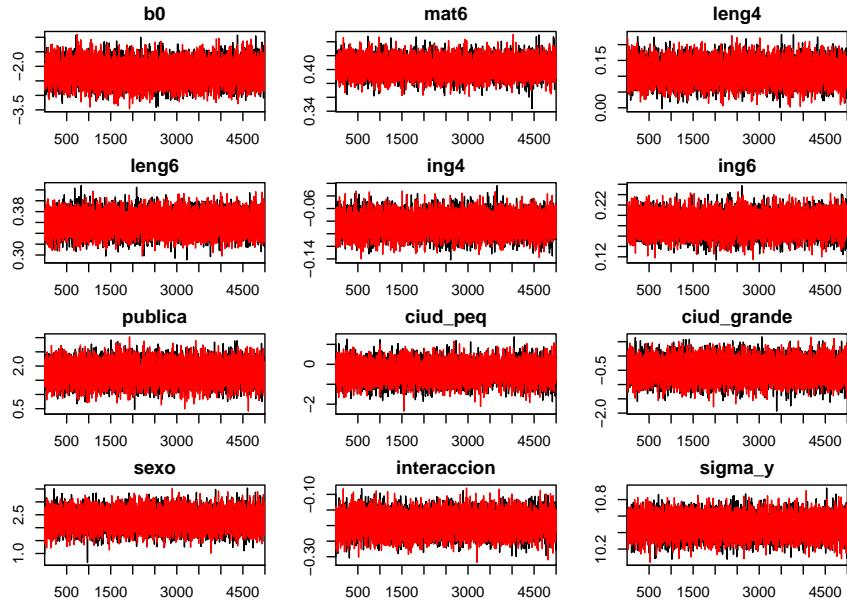
Dado que el parámetro relativo a la variable servicio territorial resultó con una distribución muy centrada alrededor de cero, vamos a eliminar esa covariable del modelo buscando de esta manera mejores resultados y una interpretación más efectiva con respecto al objetivo marcado.

3.2.1 Presentación del modelo 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 a 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 el resto de las pruebas.

3.2.2 Simulación del modelo 2

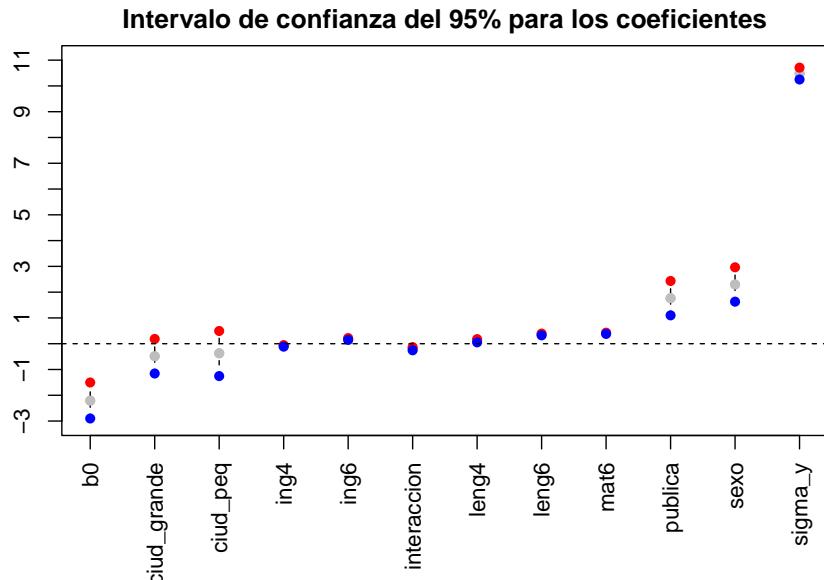
Al igual que hicimos en el modelo anterior, aquí también usamos el paquete JAGS para simular el modelo a través de MCMC. El resultado para los distintos parámetros es el siguiente:



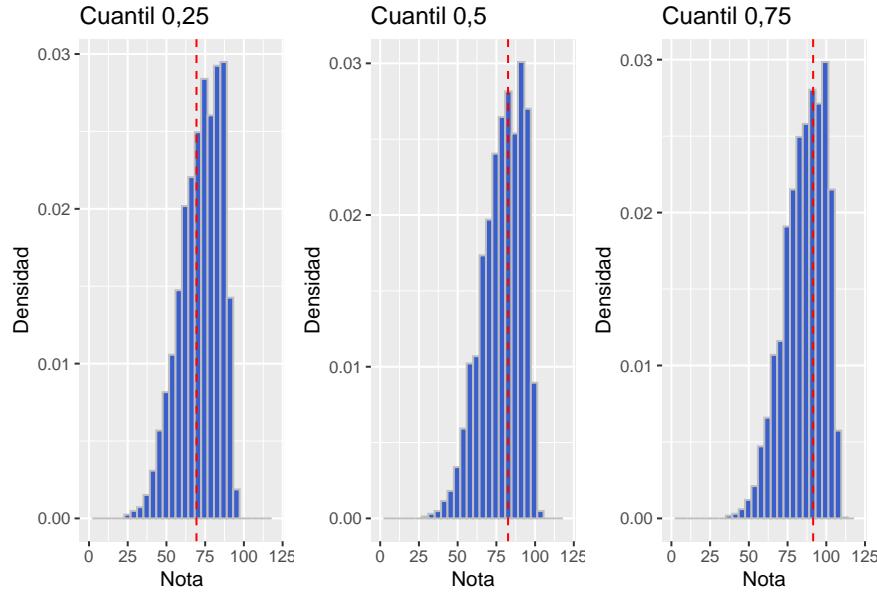
También en este caso las cadenas convergen, lo que nos permite validar el modelo y proceder a su análisis.

3.2.3 Interpretación y validación del modelo 2

En cuanto al modelo 1, recordemos que las covariables relativas a las puntuaciones se expresan en términos de 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 localidad siempre se distribuye normalmente alrededor de 0.



Como en el caso anterior, analizamos los estadísticos relativos a los percentiles 25%, 50% y 75%.



El modelo resultante parece ser casi el mismo que el anterior y, además en este caso, los estadísticos obtenidos en la simulación con el *testset* presentan valores ligeramente inferiores a los estadísticos reales. También en este caso se procede a eliminar la covariable de localidad para obtener un modelo más fácilmente interpretable.

3.3 Modelo 3: sin Servicios Territoriales ni Localidades

Reducimos aún más nuestro modelo manteniendo únicamente las calificaciones obtenidas en las pruebas de matemáticas de cuarto, lengua en los dos cursos (sexto de primaria y cuarto de ESO), inglés también en ambos cursos, género, tipo de escuela (pública o privada) e interacción entre género y lenguas en cuarto de ESO.

3.3.1 Presentación del modelo 3

El modelo implementado es el siguiente:

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

Suponemos que los coeficiente tienen las siguientes distribuciones. Elegimos una distribución normal estándar, ya que las variables son diferencias en las puntuaciones de cada prueba con respecto a la media.

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

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

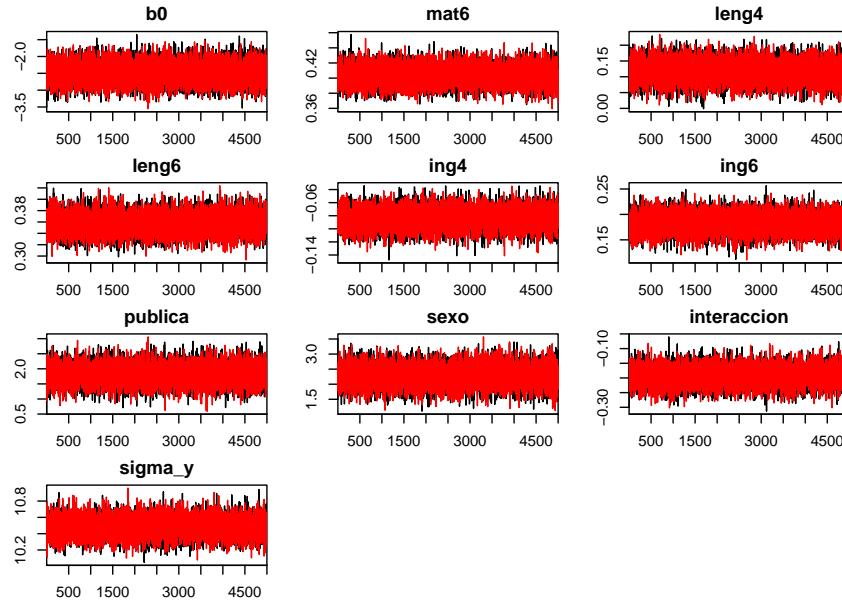
$$interaccion \sim N(0, 1)$$

$$\tau_y \sim \Gamma(0.001, 0.001)$$

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

3.3.2 Simulación de modelo 3

Como en los casos anteriores, simularemos el modelo, ya que una resolución analítica requeriría un esfuerzo computacional considerable.



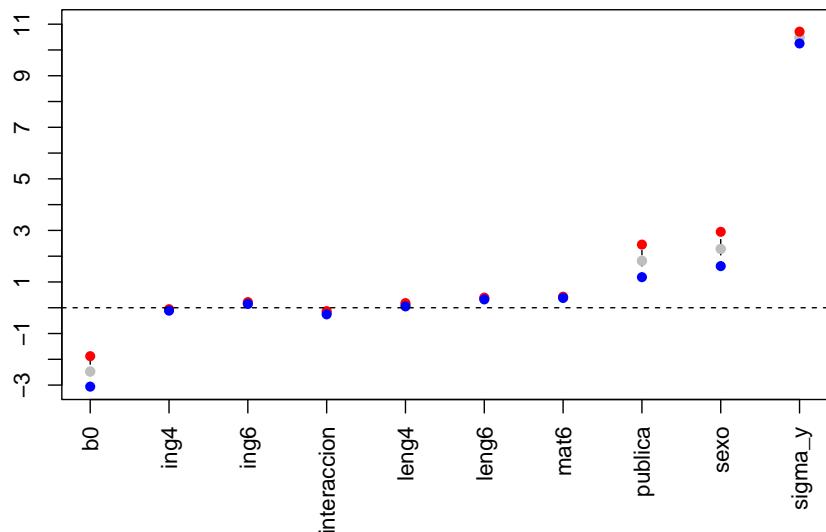
Como se aprecia en los gráficos, también en este caso el modelo converge y, por tanto, se puede validar.

3.3.3 Interpretación y validación del modelo 3

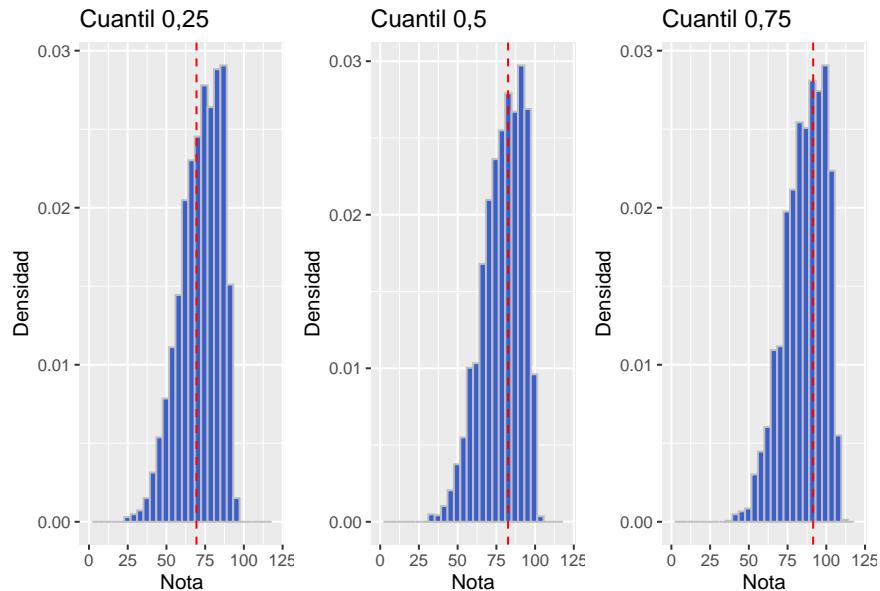
En este modelo observamos que los coeficientes obtenidos para las diferentes covariables presentan valores muy cercanos a 0 pero sin contenerlo en el intervalo de confianza del 95%, por lo que consideramos significativas todas las covariables escogidas.

De nuevo recordamos que las puntuaciones han sido evaluadas como diferencias entre la puntuación absoluta y la media de todo el alumnado en cada una de las pruebas.

Intervalo de confianza del 95% para los coeficientes



También en este caso analizamos los estadísticos siguientes: percentiles 25%, 50% y 75%. Nos referimos siempre a la distribución real de las puntuaciones obtenidas por los alumnos.



Los estadísticos siempre están a la izquierda de los estadísticos reales, pero aún así consideramos que el modelo es satisfactorio y que describe bien nuestros datos en comparación con el modelo inicial, que era mucho más pesado.

Comparando también los índices DIC vemos que los modelos presentan los siguientes valores $mod_1 = 30165.7$, $mod_2 = 30161.74$ and $mod_3 = 30160.36$.

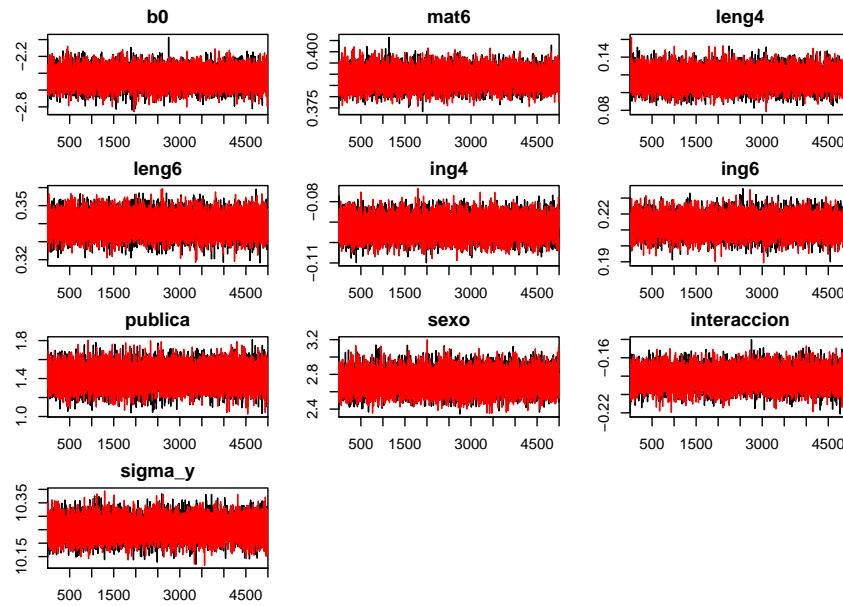
Recordamos que un índice DIC más bajo es mejor, por lo que confirmamos la elección del modelo 3 para evaluar nuestro objetivo inicial. En el siguiente apartado procederemos al análisis detallado de este modelo.

4 Implementación del modelo definitivo

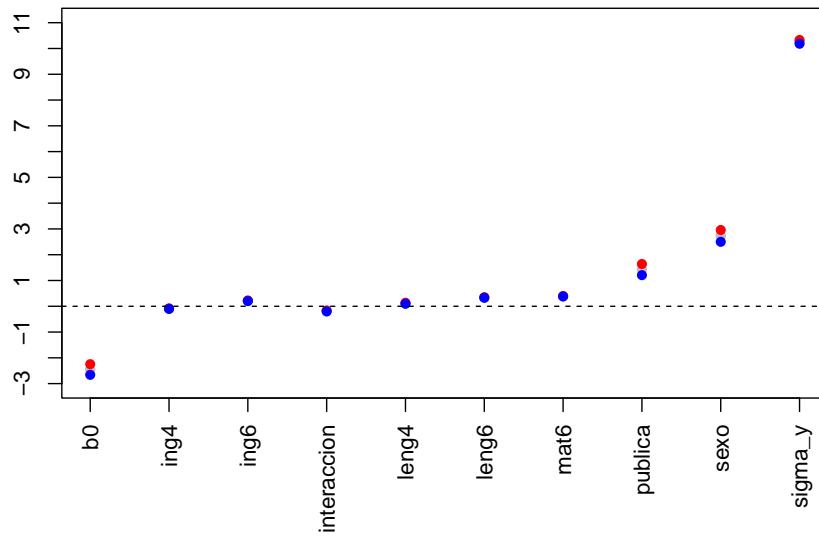
Dado que el modelo escogido es un modelo más ligero que el modelo con el que iniciábamos este trabajo, podemos entrenarlo en una mayor parte de los estudiantes. En esta ocasión usamos un conjunto de entrenamiento compuesto por el 80% de los estudiantes (alrededor de 37000).

En este caso, como en los anteriores, 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. Asimismo observamos 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



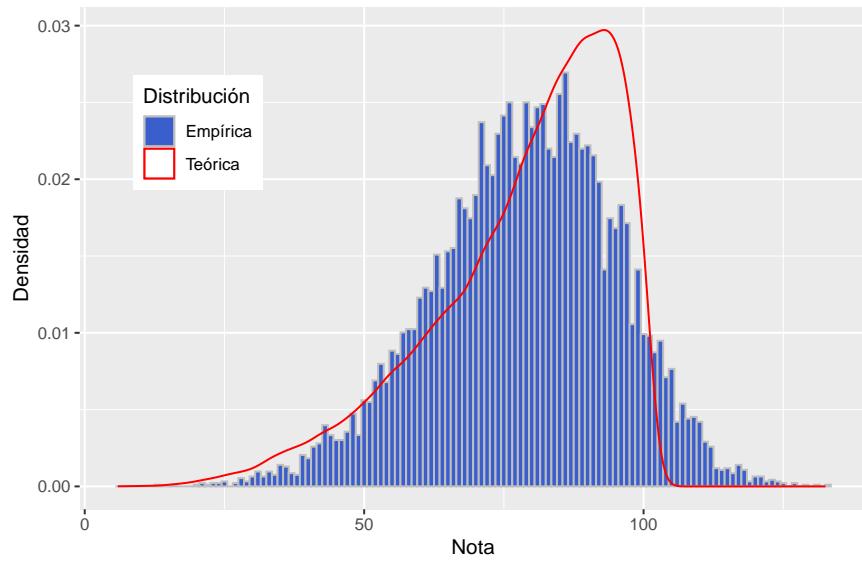
Los valores numéricos de los coeficientes y todos los rangos aparecen en la tabla de coeficientes del anexo. Las puntuaciones se han evaluado como diferencias con respecto a la media y no como puntuaciones absolutas, con lo cual la interpretación de los coeficientes se deberá realizar teniendo en cuenta esta circunstancia. Aquí presentamos el análisis de los valores medios obtenidos de la distribución de los coeficientes:

- $\beta_0 = -2.45$: El valor β_0 es negativo, así que supondremos que un estudiante promedio (covariables numéricas = 0), mujer, en una escuela privada obtendrá una calificación más baja que el promedio de los alumnos en la prueba de matemáticas de cuarto.
- $ing4 = -0.09$: Para la calificación de inglés en el examen de cuarto, el coeficiente también es negativo, por lo que podemos suponer que una persona que está por encima del promedio en inglés estará un poco por debajo del promedio en la prueba de 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 obtener un buen resultado también en matemáticas de cuarto.
- $leng4 = 0.12$: El valor del coeficiente es positivo. Teniendo en cuenta la interacción esto significa que una chica que en promedio obtiene buenos resultados en lenguas en la prueba de cuarto también tendrá un resultado por encima del promedio en la 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.
- $interacción = -0.18$: El valor del coeficiente es negativo. Este resultado es muy interesante cuando se compara con el de leng4 ya que el coeficiente se vuelve negativo en promedio para leng4. Por lo tanto, según el modelo, un estudiante varón mostrará un rendimiento opuesto al promedio de los alumnos en ambas pruebas.
- $mate6 = 0.39$: También en este caso el coeficiente es positivo, lo que significa que el alumnado que destaca en matemáticas al final de la primaria sigue estando por encima de la media en esta materia al final de la ESO. Asimismo, observamos que éste es el mayor coeficiente de todo el modelo, lo que implica que la puntuación de matemáticas en sexto es la que más contribuye en la predicción del resultado de la prueba de matemáticas de cuarto de ESO.
- $pública = 1.42$: 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 un alumno del mismo perfil de un colegio privado.
- $sexo = 2.73$: También en este caso la variable es binaria, por lo que podemos afirmar 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 desviación 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. Utilizaremos los coeficientes promedio para la simulación y extraeremos un valor aleatorio de cada distribución generada para un estudiante.

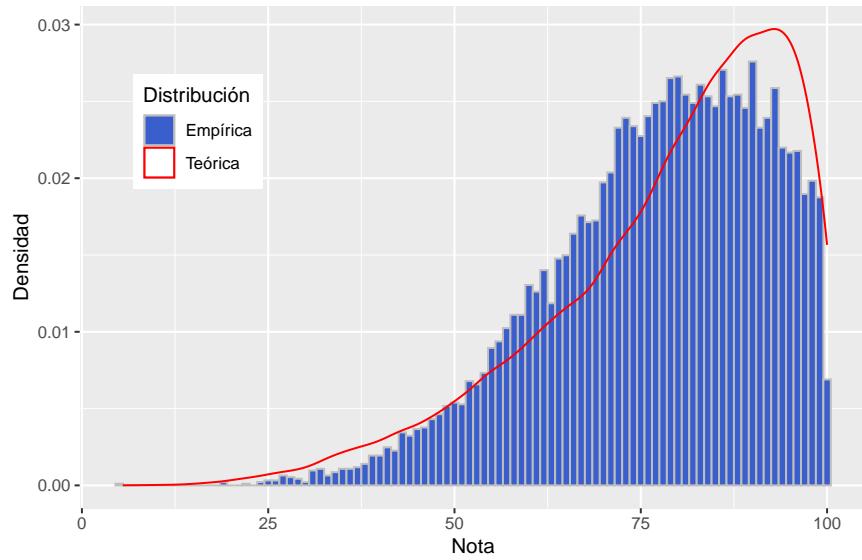
La distribución resultante es la siguiente:

Comparación entre distribución predictiva posterior y PMAT_4



Como podemos observar, la distribución describe bastante bien los datos aunque muchas de las observaciones pronosticadas superan el límite de 100. Por lo tanto, decidimos usar una distribución normal truncada entre 0 y 100 con la misma media y varianza que se usaron anteriormente para predecir los resultados de los estudiantes.

Comparación entre la distribución posterior predictiva truncada y PMAT_4



Se aprecia que nuestro modelo sigue bastante de cerca la distribución teórica, por lo que estamos satisfechos de su efectividad.

5 Conclusiones

El objetivo principal de nuestro trabajo consistía en 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 averiguar si existe una diferencia evidente en la puntuación esperada por el simple hecho de ser chica o chico. 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 explica nuestros datos y el que mejores predicciones puede ofrecer para estimar la puntuación de una persona en la prueba de competencia matemática de Cataluña.

Teniendo eso en cuenta procedemos a 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?

Dados los resultados, podemos considerar que, efectivamente, se trata de un factor relevante. Sería interesante realizar algún 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?

La respuesta a esta pregunta sería de nuevo afirmativa. 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 ($0.12 - 0.18 = -0.06$). Por tanto, cabría afirmar que para los niños el resultado en matemáticas está inversamente correlacionado con el resultado en lenguas y que niños que no tienen buena competencia en materia lingüística sí que la tendrían en matemáticas.

En definitiva, el sesgo de género es evidente en la materia de matemáticas y el rendimiento de las alumnas al finalizar la secundaria obligatoria es menor que el de los alumnos. Existen distintas hipótesis, que incluyen factores psicosociales, y que podrían explicar la razón de esta diferencia. Nosotros pensamos que para aumentar la tendencia de las chicas a estudiar asignaturas científicas, y más concretamente matemáticas, que las conduciría a matricularse en carreras STEAM, es importante fomentar la cultura de la igualdad de género en las escuelas desde edades tempranas.

Además, es esencial romper los estereotipos de género asociados a estos campos. Esto incluye mostrar a las niñas y a las jóvenes ejemplos de mujeres relevantes de antes y de ahora en los campos científico-tecnológicos asegurando que los materiales educativos y las actividades extraescolares promueven una representación equilibrada de hombres y mujeres.

Asimismo necesitamos mostrar a más mujeres en posiciones de liderazgo en estas disciplinas. Esto puede lograrse a través de las campañas de comunicación de las empresas, universidades y otros medios para que sirvan como referente a las nuevas generaciones.

6 Apéndice

6.1 Código R

```
load("alumni.Rdata")
##### Histogram by area #####
area = alumni$AREA_TERRITORIAL
ggplot(data.frame(area), aes(x = area)) + geom_bar(fill = "royalblue3",
  color = "grey") + theme(axis.text.x = element_text(angle = 335,
  vjust = 1, hjust = 0)) + theme(plot.margin = margin(0,
  2, 0, 2, "cm")) + ggtitle("Distribución de los alumnos entre las áreas",
) + ylab("Número") + xlab("Área")
#####
Histogram by city #####
ciudad = factor(alumni$HABITAT, levels = c("Fins a 10000",
  "De 10001 a 100000", "Més de 100000"))
ggplot(data.frame(ciudad), aes(x = ciudad)) + geom_bar(fill = "royalblue3",
  color = "grey") + 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 los alumnos según tamaño de la localidad",
) + ylab("Número") + xlab("Tamaño de la localidad")
#####
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 = GENERE, y = value, fill = GENERE)) +
  geom_boxplot() + labs(title = "Comparación de resultados por género",
y = "Notas", x = "") + facet_wrap(~Test, nrow = 1) +
  scale_fill_manual(values = c("salmon", "royalblue3"))

#####
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 los resultados entre individuos") +
  scale_fill_manual(values = c("salmon", "royalblue3")) +
  scale_colour_manual(values = c("salmon", "royalblue3"))
##### 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, mfrow = 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(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[, 3])

```

```

  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("Cuantil 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("Cuantil 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("Cuantil 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 = NATURALESA_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, mfrow = 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)
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.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("Cuantil 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("Cuantil 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("Cuantil 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)
#####
Modelo 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
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, mfrow = c(4, 3), varname = parameters.3,
  col = c("black", "red"), ask = F)
print(alumni.sim.3, digits = 2)
##### Intervalos de confianza para el modelo
#####
output = alumni.sim.3
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)]
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$GENERE_H)))
C = as.matrix(coeff3[2000 + 1:M, 1:9])
Mu = rowSums(X * C)
Sigma = coeff3[1:M, 10]
q_sim = NULL
for (i in 1:4000) {
  y_t = (rnorm(1000, Mu[i], Sigma[i]) + m4)
  q_sim = rbind(q_sim, quantile(y_t, c(0.25, 0.5,
    0.75)))
}
q_sim = data.frame(q_sim)
colnames(q_sim) = c("Q25", "Q50", "Q75")
##### Validation model 3 #####
plot1 = ggplot() + geom_histogram(data = q_sim, aes(x = Q25,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Cuantil 0,25", ) + geom_vline(xintercept = quant[1],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")
plot2 = ggplot() + geom_histogram(data = q_sim, aes(x = Q50,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Cuantil 0,5", ) + geom_vline(xintercept = quant[2],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")
plot3 = ggplot() + geom_histogram(data = q_sim, aes(x = Q75,
  y = ..density..), color = "grey", fill = "royalblue3") +
  ggtitle("Cuantil 0,75", ) + geom_vline(xintercept = quant[3],
  linetype = "dashed", color = "red", size = 0.5) +
  xlim(c(0, 120)) + ylab("Densidad") + xlab("Nota")
grid.arrange(plot1, plot2, plot3, ncol = 3)
##### Preprocessing model 3 big #####
load("alumni.Rdata")
alumni$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
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)
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
##### Prevision model 3 big #####
ggplot() + geom_histogram(data = data.frame(y_t), aes(x = y_t,

```

```

y = ..density.., fill = "royalblue3", colour = "grey"),
binwidth = 1) + geom_density(data = train, aes(x = PMAT_4 +
m4, fill = "#00000000", color = "red")) + ggtitle("Comparación entre distribución predictiva posterior")
) + ylab("Densidad") + xlab("Nota") + scale_colour_identity(name = "Distribución",
guide = "legend", labels = c("Empírica", "Teórica")) +
theme(legend.position = c(0.15, 0.75)) + guides(color = guide_legend	override.aes = list(fill = c("white"))))
+ scale_fill_identity(guide = "none")
#####
##### Previsión 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 = "grey", fill = "royalblue3"),
binwidth = 1) + geom_density(data = train, aes(x = PMAT_4 +
m4, fill = "00000000", color = "red")) + ggtitle("Comparación entre la distribución posterior predictiva")
) + ylab("Densidad") + xlab("Nota") + scale_colour_identity(name = "Distribución",
guide = "legend", labels = c("Empírica", "Teórica")) +
theme(legend.position = c(0.15, 0.75)) + guides(color = guide_legend	override.aes = list(fill = c("white"))))
+ scale_fill_identity(guide = "none")
print(alumni.sim.3_big, digits = 2)

```

6.2 Tablas parámetros estimados

```
## Inference for Bugs model at "7", 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        -2.45   0.11  -2.66  -2.52  -2.45  -2.38  -2.25
## ing4      -0.09   0.00  -0.10  -0.10  -0.09  -0.09  -0.08
## ing6       0.21   0.01   0.20   0.21   0.21   0.22   0.22
## interaccion -0.18   0.01  -0.20  -0.19  -0.18  -0.17  -0.16
## leng4      0.12   0.01   0.10   0.11   0.12   0.12   0.14
## leng6      0.34   0.01   0.33   0.34   0.34   0.34   0.35
## mat6       0.39   0.00   0.38   0.39   0.39   0.39   0.40
## publica    1.42   0.11   1.21   1.35   1.43   1.50   1.64
## sexo       2.73   0.12   2.50   2.65   2.73   2.81   2.96
## sigma_y    10.26  0.04  10.18  10.23  10.26  10.28  10.33
## deviance   278065.92 4.64 278058.87 278062.56 278065.25 278068.56 278076.73
##          Rhat n.eff
## b0        1 3600
## ing4      1 10000
## ing6      1 10000
## interaccion 1 10000
## leng4      1 10000
## leng6      1 10000
## mat6       1 10000
## publica    1 10000
## sexo       1 2700
## 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 = 10.8 and DIC = 278076.7
## 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