Bayesian Methods Final Project

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

Aráiztegui, Aránzazu
Ferrara, Lorenzo
Lucchini, Marco

03 January, 2023

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Índice

1 Introducción

El sesgo de género está presente en numerosos ámbitos de nuestra vida y se aprecia de forma notable en el ámbito educativo, más concretamente en las materias del ámbito STEAM.
Parece ser que aunque no hay una evidente brecha de género en el comienzo de los estudios primarios, el sesgo de género comienza a aparecer de forma clara a lo largo del proceso educativo, agudizándose en las últimas etapas de la educación obligatoria y evidenciándose de forma clara en la educación universitaria.

...

1.1 Descripción del problema

Con el objetivo de evaluar la capacidad y el nivel de competencia en las diferentes áreas del conocimiento que tiene el alumnado de Cataluña, el Departament d'Educació de la Generalitat de Cataluña realiza una prueba de competencias y conocimientos básicos en las áreas lingüísticas, matemáticas y científico-tecnológicas en los últimos cursos de la educación primaria y secundaria.
Según el Departament se trata de una evaluación de carácter formativo y orientador que pueda servir tanto a los centros como al profesorado y al propio Departament para impulsar las mejoras en el sistema educativo catalán.

1.2 Objectivos del modelo

Este trabajo tiene como objetivo principal ver si existe un sesgo de género en los resultados obtenidos en la competencia matemática con respecto al sexo y a las competencias humanísticas para ello intentaremos crear un modelo que relacione la puntuación obtenida en la competencia matemática con respecto al sexo, a las competencias lingüísticas e incluso con respecto al tipo de centro educativo o al tamaño de la población. De esta manera podríamos ver si en el sexo femenino no se da una diferencia entre el rendimiento en humanidades y matemáticas, personas con bajo rendimiento en humanidades también lo tendrían en matemáticas.
Y en cambio en el sexo masculino personas con bajo rendimiento en humanidades tendrían buenos resultados en matemáticas.
Si ampliamos los datos y vemos la tabla de resultados para un mismo individuo en sexto de primaria y cuarto de la podríamos establecer un segundo objetivo que sería ver si se mantienen los resultados en ambos sexos o si hay diferencias significativas en cuanto al rendimiento en el área de matemáticas al aumentar la edad en relación al sexo.

2 Descripción de la base de datos

La base de datos que hemos utilizado en este trabajo procede del catálogo de datos en abierto que proporciona la Generalitat de Catalunya en su página web. **IS a merge between QUARTZ and SISE on common id of the student.** Se puede acceder a la base de datos completa en el siguiente enlace:
Avaluació de quart d'Educació Secundària Obligatòria | Dades obertes de Catalunya
El dataset contiene los resultados obtenidos por el alumnado de cuarto curso de ESO en la evaluación de competencias básicas al final de la educación secundaria desde el año 2012.
Se incluye un código de alumno para poder hacer comparativas con los resultados obtenidos en sextos de primaria. Dado que el código solo está disponible a partir del año 2016 se utilizarán únicamente los datos del alunado a partir de este año.
La base de datos ha sido actualizada el 20 de octubre de 2022
We have 46384 students

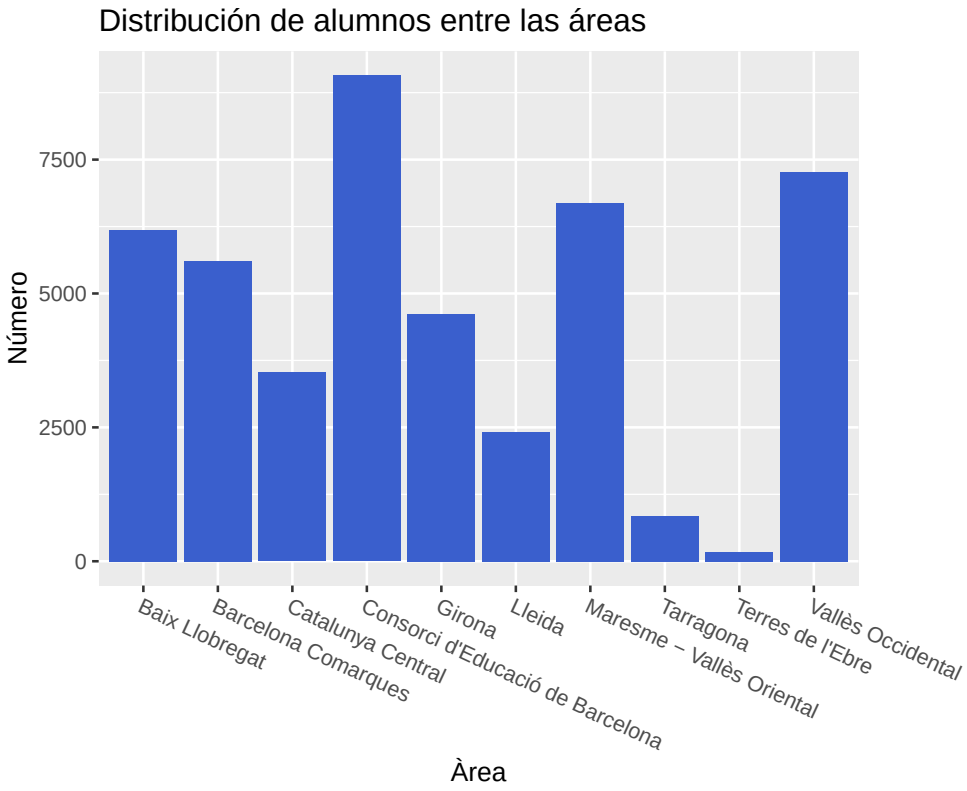
2.1 Definición de las variables utilizadas

Base de dades

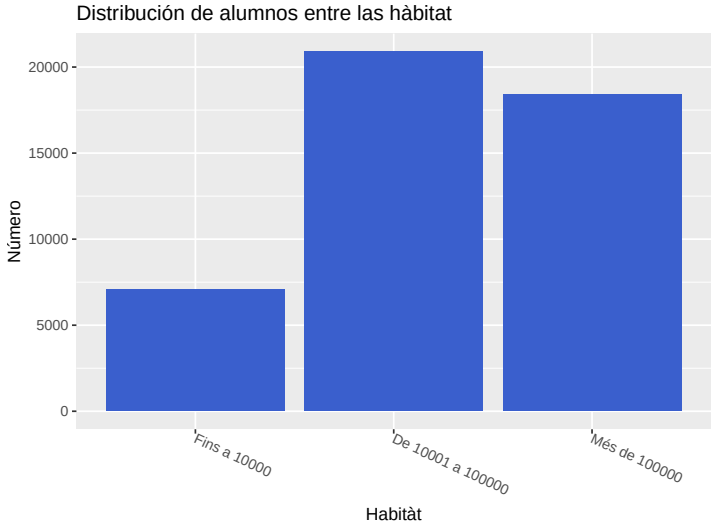
Nom de columna	Descripció	Tipus
PMAT_4	Puntuació global ponderada de competència matemàtica en el examen de Quart	Nombre
PMAT_6	Puntuació global ponderada de competència matemàtica en el examen de Sisè	Nombre
PLENG_4	Puntuació global ponderada de la competència lingüística en llengua catalana y castellana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

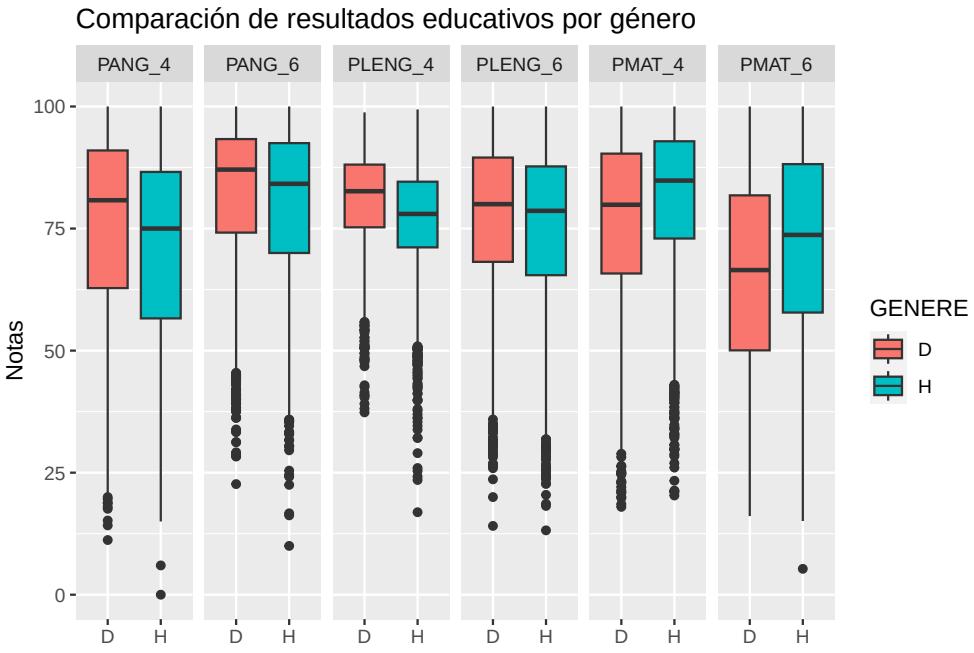
We start by looking at our dataset, the students are evenly spread among boys and girls. Their distribution in the areas is the following.



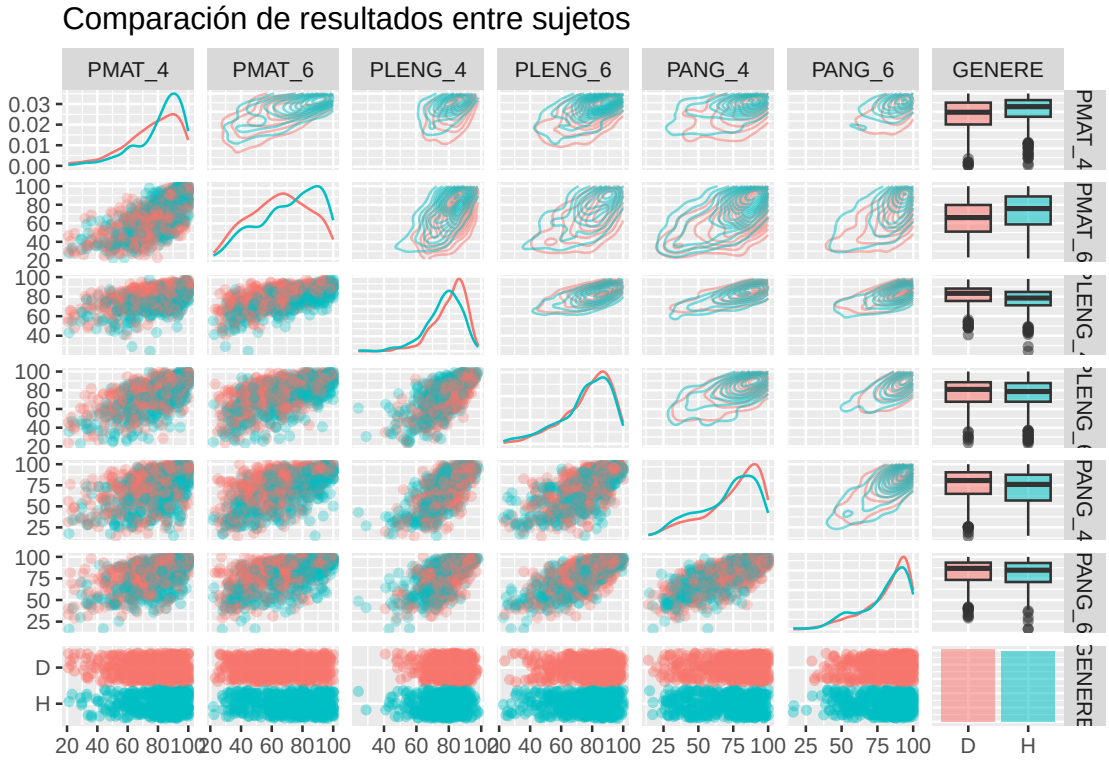
Analyzing then the division in city we find that most of them lives in medium or big cities.



We now look at the grade obtained in the exam of english, lenguage and math in the Sisè and Quart exam. The grades are divided between boys and girls. From the plot we suppose that boys tend to have lower grades in english and lenguage while obtain better result in math with respect to girls.



We then look at the data in pairs. As espected the data are almost linearly distributed with low variability when comparing the same subject among the years, this means that students who where good in a subject in the Sisè exam were good also in the Quart exam and the opposite. In the other comparison the variability is greater. We also point out the distribution of frequency in the groups by gender, particularly in math where the difference between male and female is more evident.



3 Análisis bayesiano

3.1 Modelo 1

3.1.1 Presentacion del modelo

We buid the following model

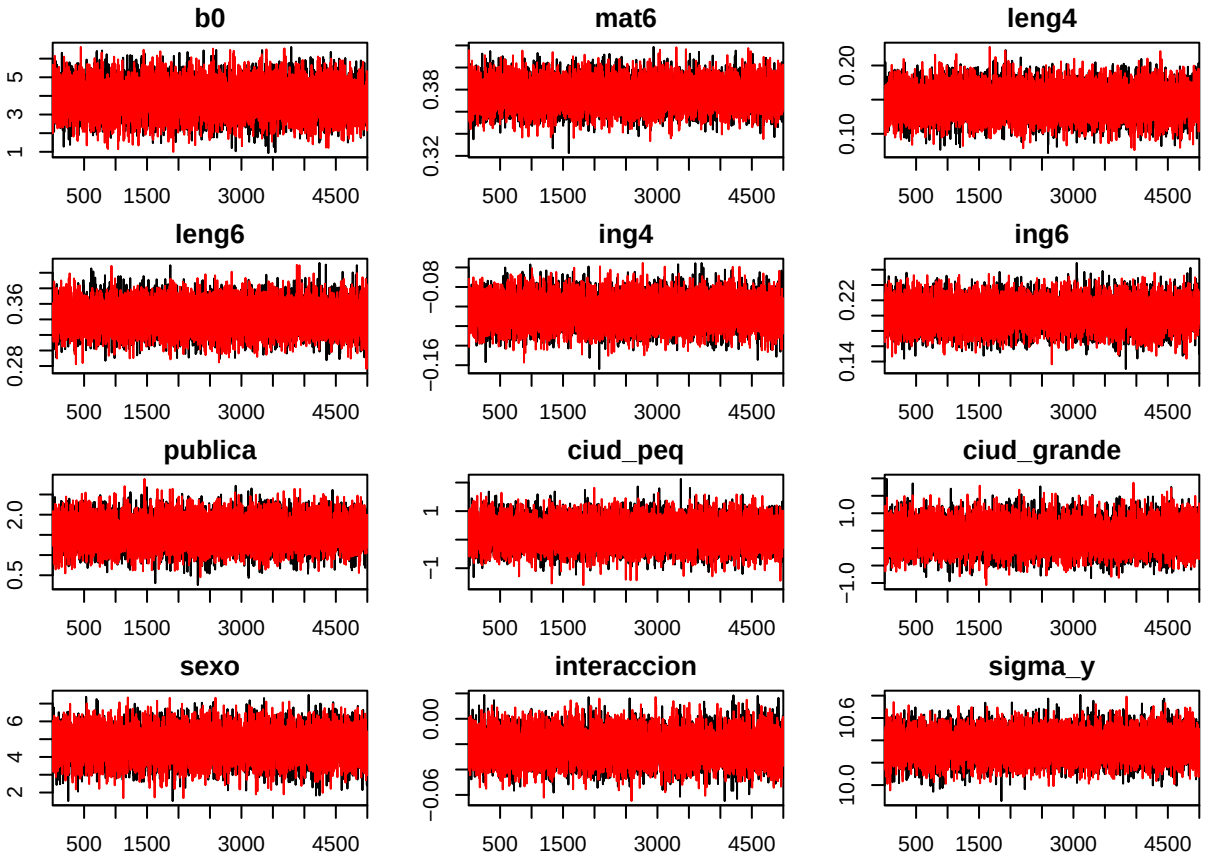
$$y_i \sim N(b_0 + mat_4 * x_i^1 + leng_4 * x_i^2 + leng_6 * x_i^3 + ing_4 * x_i^4 + ing^6 * x_i^5 + \ publica * x_i^6$$
$$+ ciudPeq * x_i^7 + ciudGrande * x_i^8 + sexo * x_i^9 + interaccion * x_i^3 * x_i^9 + G[area[i]], tau_y)$$

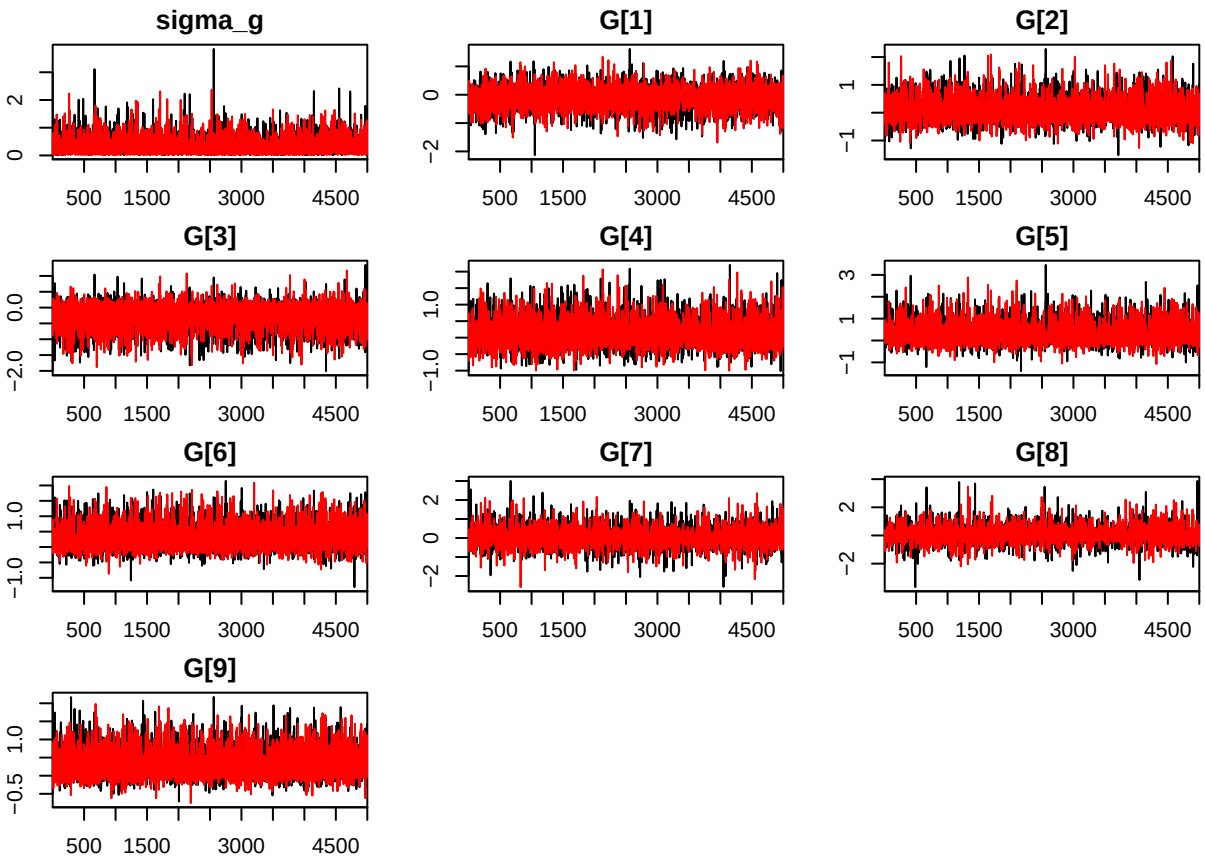
We suppose the coefficient having the following distribution.

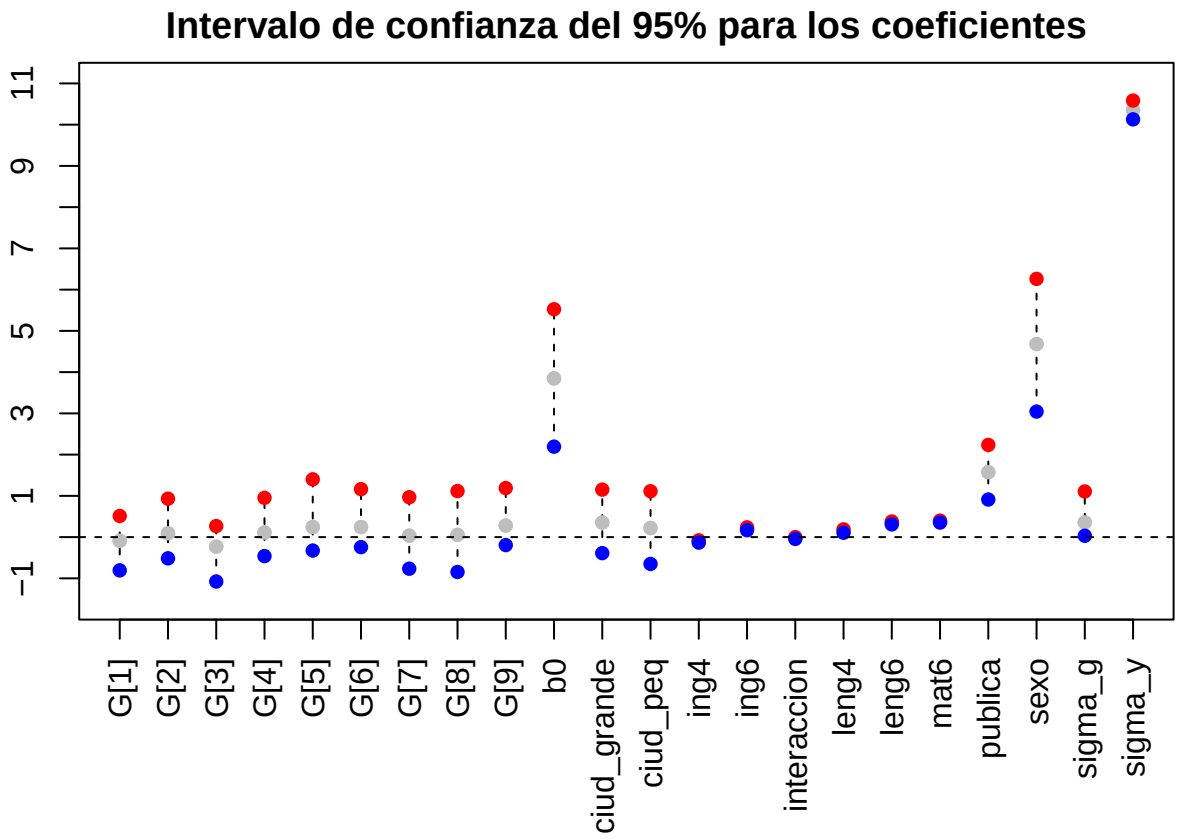
$$b_0 \sim N(0,1)$$
$$mat_4 \sim N(1,1)$$
$$leng_4 \sim N(1,1)$$
$$leng_6 \sim N(1,1)$$
$$ing_4 \sim N(1,1)$$
$$ing_6 \sim N(0,1)$$
$$publica \sim N(0,1)$$
$$ciudPeq \sim N(0,1)$$
$$ciudGrande \sim N(0,1)$$
$$sexo \sim N(0,1)$$
$$interaccion \sim N(0,1)$$
$$tau_y \sim \Gamma(0.001, 0.001)$$
$$sigma_y = \frac{1}{\sqrt{tau_y}}$$
$$G_i \sim N(0, tau_g)$$
$$tau_g \sim \Gamma(0.001, 0.001)$$
$$sigma_g = \frac{1}{\sqrt{tau_g}}$$

3.1.2 Justificación e interpretación de los modelos

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 4000
##    Unobserved stochastic nodes: 22
##    Total graph size: 87242
##
## Initializing model
```


b0
mat6
leng4
leng6
ing4
ing6
publica
ciud_peq
ciud_grande
sexo
interaccion
sigma_y
500 1500 3000 4500

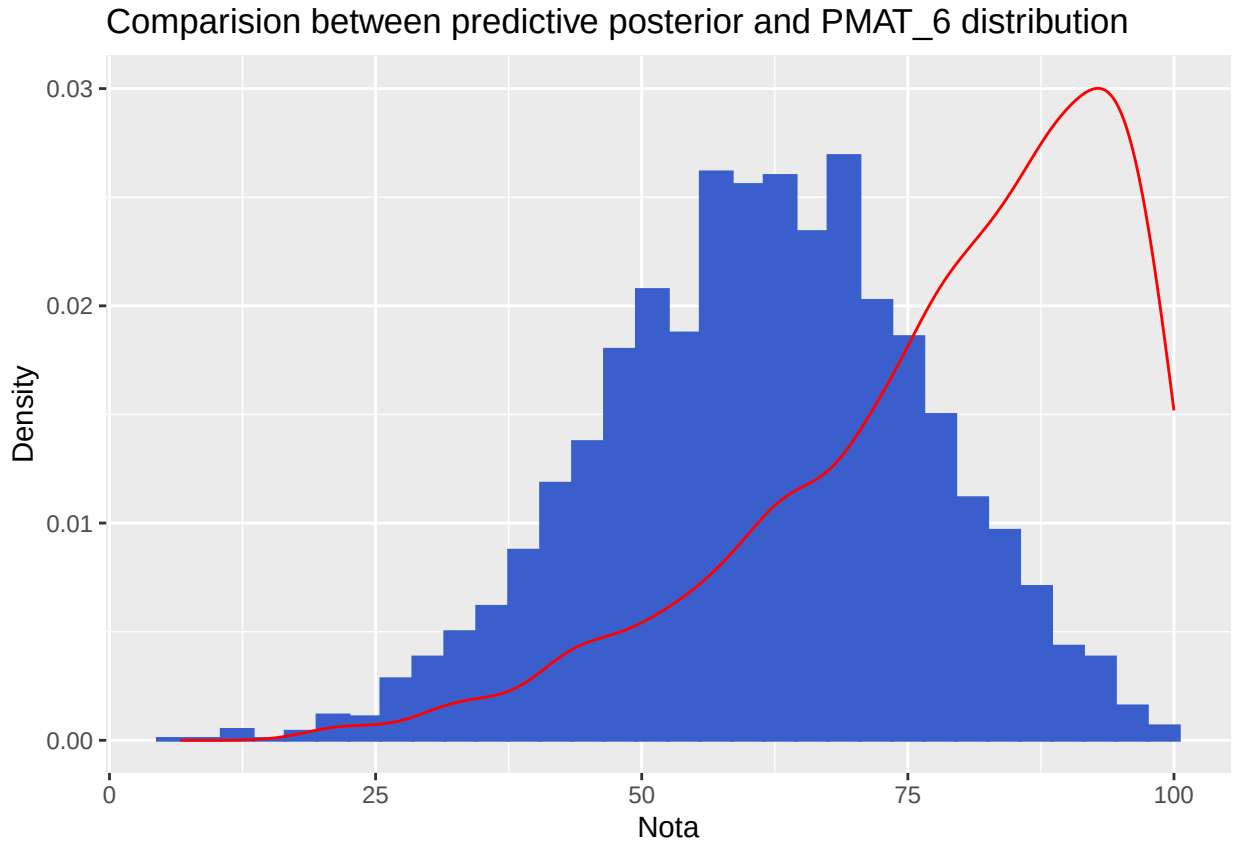

sigma_g
G[1]
G[2]
G[3]
G[4]
G[5]
G[6]
G[7]
G[8]
G[9]
500
1500
3000
4500



3.1.3 Validación de los modelos

We now predict the values of PMAT_4 for the students in the test set and compare the results.

Using Truncated Noemal 0-100 Blue predicted, Red Dataset



3.2 Modelo 2

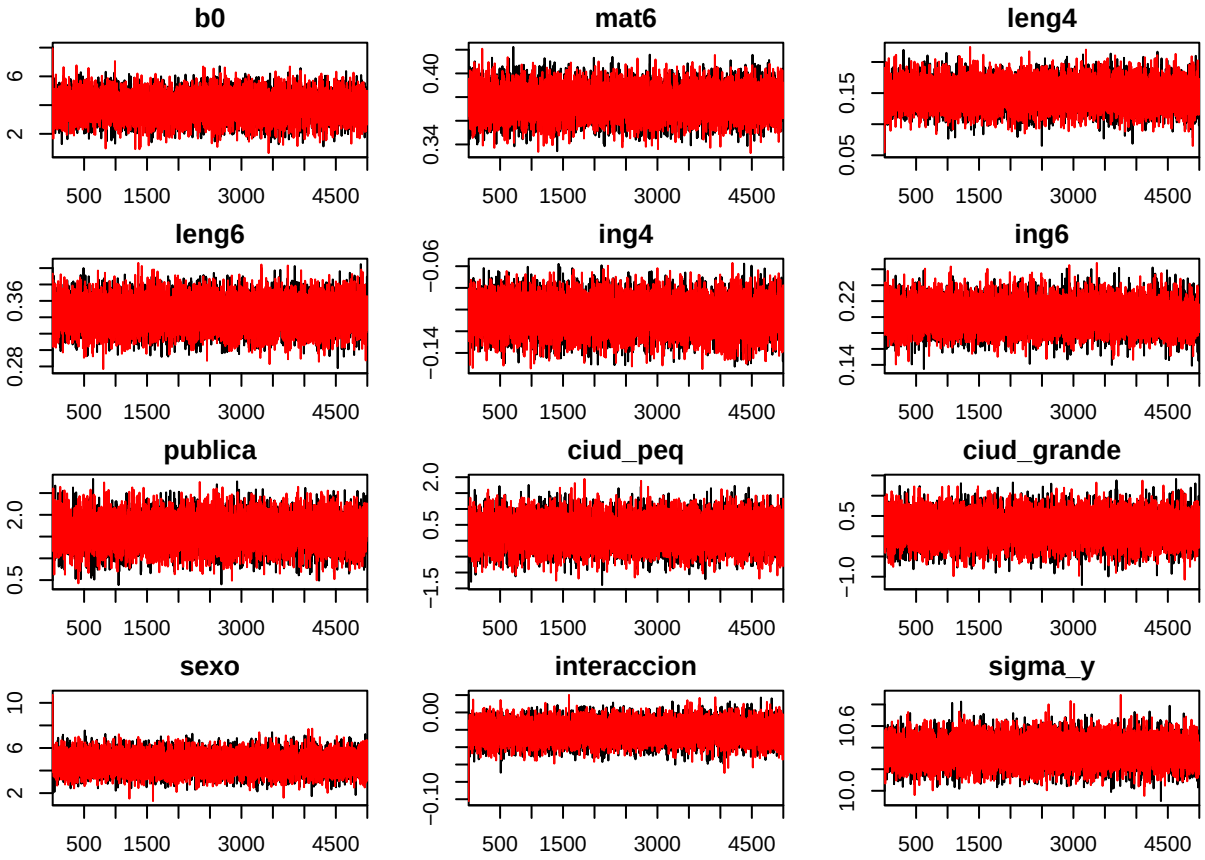
3.2.1 Presentacion de los modelos 2

3.2.2 Justificación e interpretación de los modelos 2

3.2.3 Validación de los modelos 2

3.3 2) Modelo sin areas territoriales, con interaccion sexo:lenguas

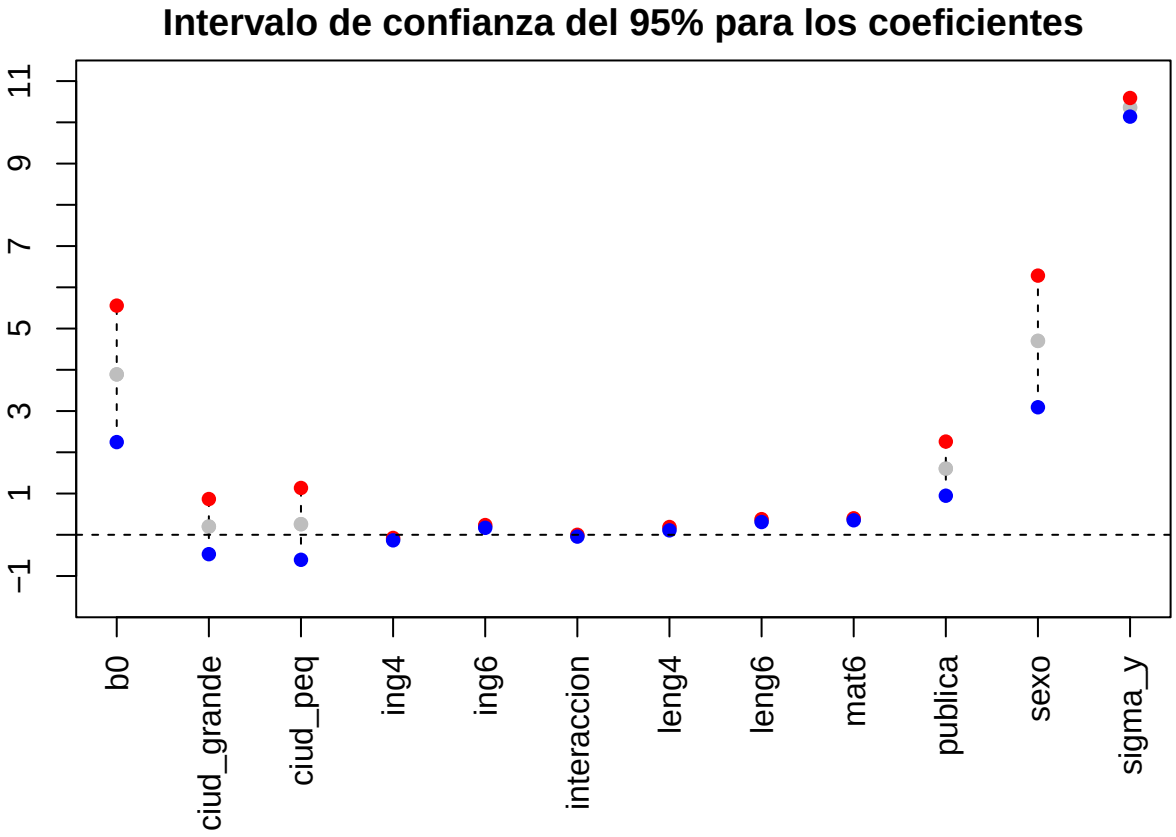
```
## module glm loaded

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 4000
##    Unobserved stochastic nodes: 12
##    Total graph size: 47208
##
## Initializing model
```




```
## 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% Rhat
## b0               3.89    0.84     2.25     3.33     3.88     4.45     5.56    1
## ciud_grande      0.20    0.34    -0.47    -0.03     0.20     0.43     0.87    1
## ciud_peq         0.26    0.45    -0.61    -0.05     0.25     0.56     1.14    1
## ing4            -0.11    0.01    -0.13    -0.11    -0.11    -0.10    -0.08    1
## ing6             0.20    0.02     0.17     0.19     0.20     0.21     0.24    1
## interaccion     -0.02    0.01    -0.04    -0.03    -0.02    -0.01     0.00    1
## leng4            0.15    0.02     0.11     0.13     0.15     0.16     0.19    1
## leng6            0.34    0.02     0.31     0.33     0.34     0.36     0.38    1
## mat6             0.38    0.01     0.35     0.37     0.38     0.38     0.40    1
## publica          1.60    0.34     0.95     1.38     1.60     1.83     2.26    1
## sexo             4.70    0.82     3.09     4.16     4.71     5.25     6.28    1
## sigma_y         10.36    0.12    10.14    10.28    10.36    10.44    10.59    1
## deviance    30055.31    9.92 30037.19 30048.50 30054.77 30061.68 30076.28    1
##               n.eff
## b0            10000
## ciud_grande    6700
## ciud_peq      10000
## ing4           3000
## ing6           5600
## interaccion    1600
## leng4          7700
## leng6          2700
```

```
## mat6         3100
## publica      2600
## sexo         1000
## sigma_y     10000
## deviance     1600
##
## 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 = 49.2 and DIC = 30104.5
## DIC is an estimate of expected predictive error (lower deviance is better).
```

```
##          b0 ciud_grande    ciud_peq    deviance        ing4        ing6
##        TRUE       FALSE       FALSE        TRUE        TRUE        TRUE
## interaccion       leng4       leng6        mat6     publica        sexo
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##     sigma_y
##        TRUE
```

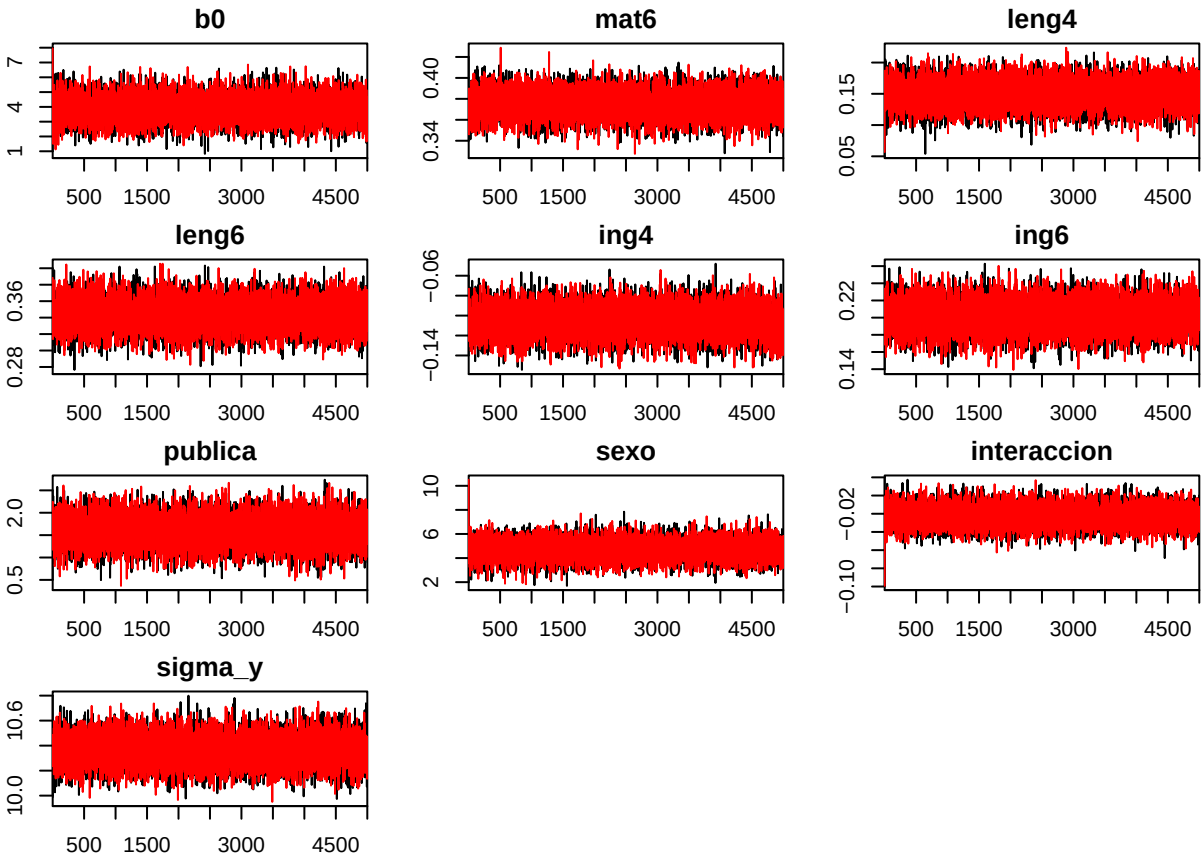


3.3.1 3) Modelo: quito "ciudad"

```
## module glm loaded
```

```
## Compiling model graph
```

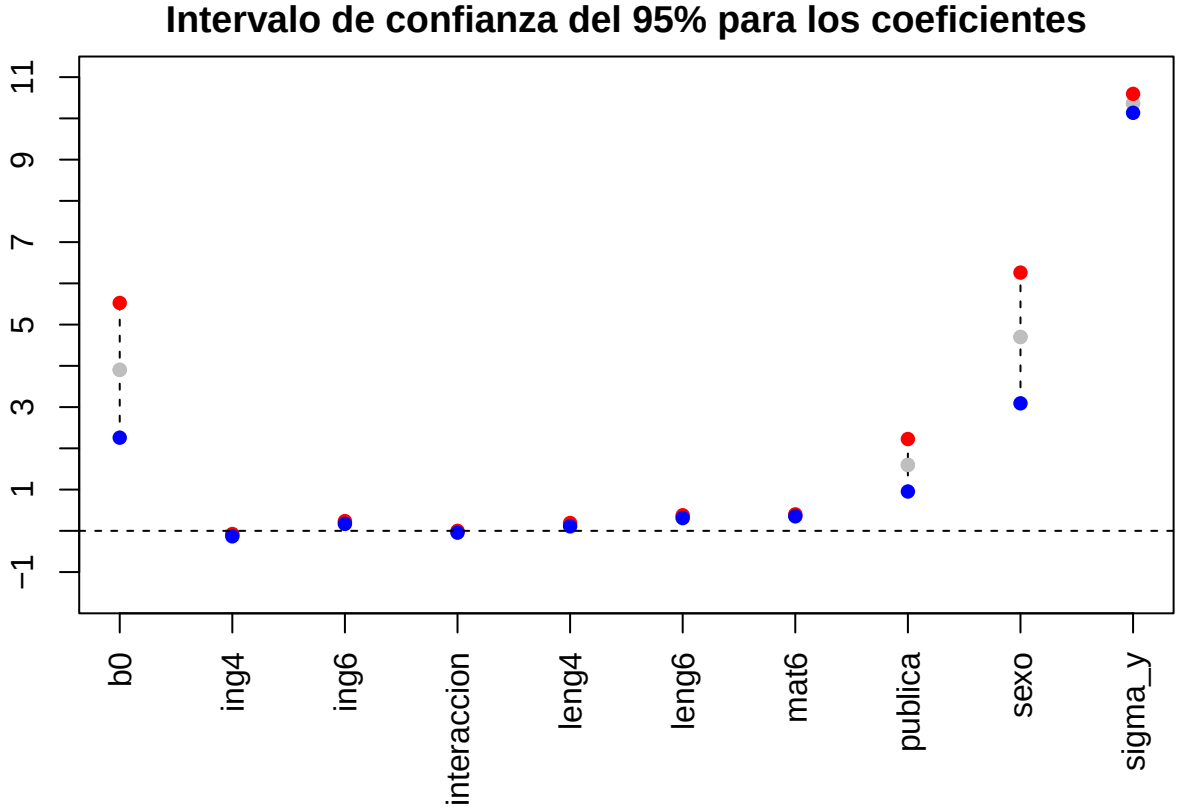
```
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 4000
##    Unobserved stochastic nodes: 10
##    Total graph size: 39202
##
## Initializing model
```




```
## 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% Rhat
## b0              3.90    0.83     2.26     3.34     3.91     4.45     5.53    1
## ing4           -0.11    0.01    -0.13    -0.12    -0.11    -0.10    -0.08    1
## ing6            0.20    0.02     0.17     0.19     0.20     0.21     0.24    1
## interaccion    -0.02    0.01    -0.04    -0.03    -0.02    -0.01     0.00    1
## leng4           0.15    0.02     0.11     0.14     0.15     0.16     0.19    1
## leng6           0.34    0.02     0.31     0.33     0.34     0.36     0.38    1
## mat6            0.38    0.01     0.35     0.37     0.38     0.38     0.40    1
## publica         1.60    0.32     0.95     1.38     1.60     1.81     2.23    1
## sexo            4.70    0.81     3.09     4.14     4.70     5.26     6.26    1
## sigma_y        10.36    0.12    10.13    10.28    10.36    10.44    10.59    1
## deviance    30054.19    9.81 30036.41 30047.34 30053.71 30060.67 30074.48    1
##             n.eff
```

```
## b0          10000
## ing4        10000
## ing6        10000
## interaccion 10000
## leng4        2200
## leng6       10000
## mat6         2800
## publica     10000
## sexo        10000
## sigma_y     10000
## deviance    10000
##
## 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 = 48.1 and DIC = 30102.3
## DIC is an estimate of expected predictive error (lower deviance is better).
```

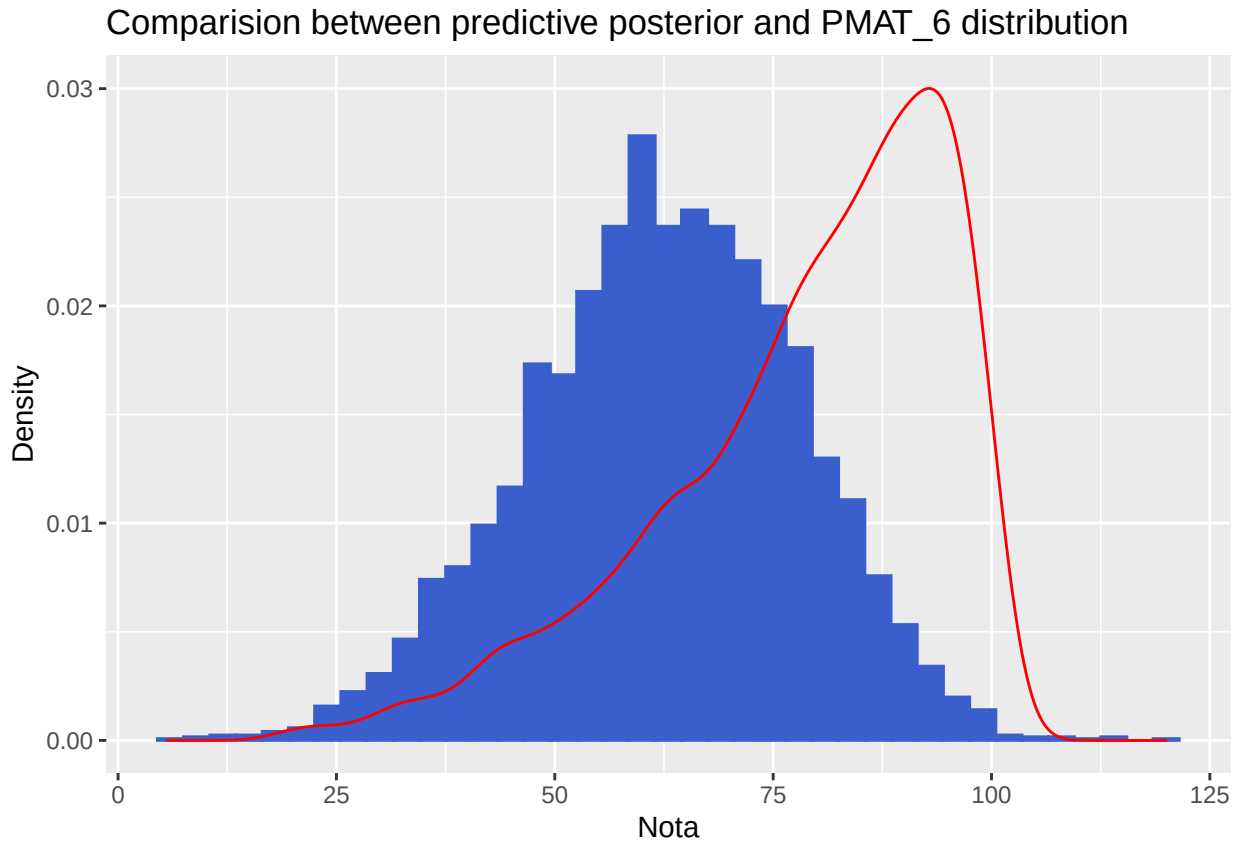
```
##          b0    deviance        ing4        ing6 interaccion       leng4
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##       leng6        mat6     publica        sexo     sigma_y
##        TRUE        TRUE        TRUE        TRUE        TRUE
```



3.3.2 Validación de los modelos

We now predict the values of PMAT_4 for the students in the test set and compare the results.

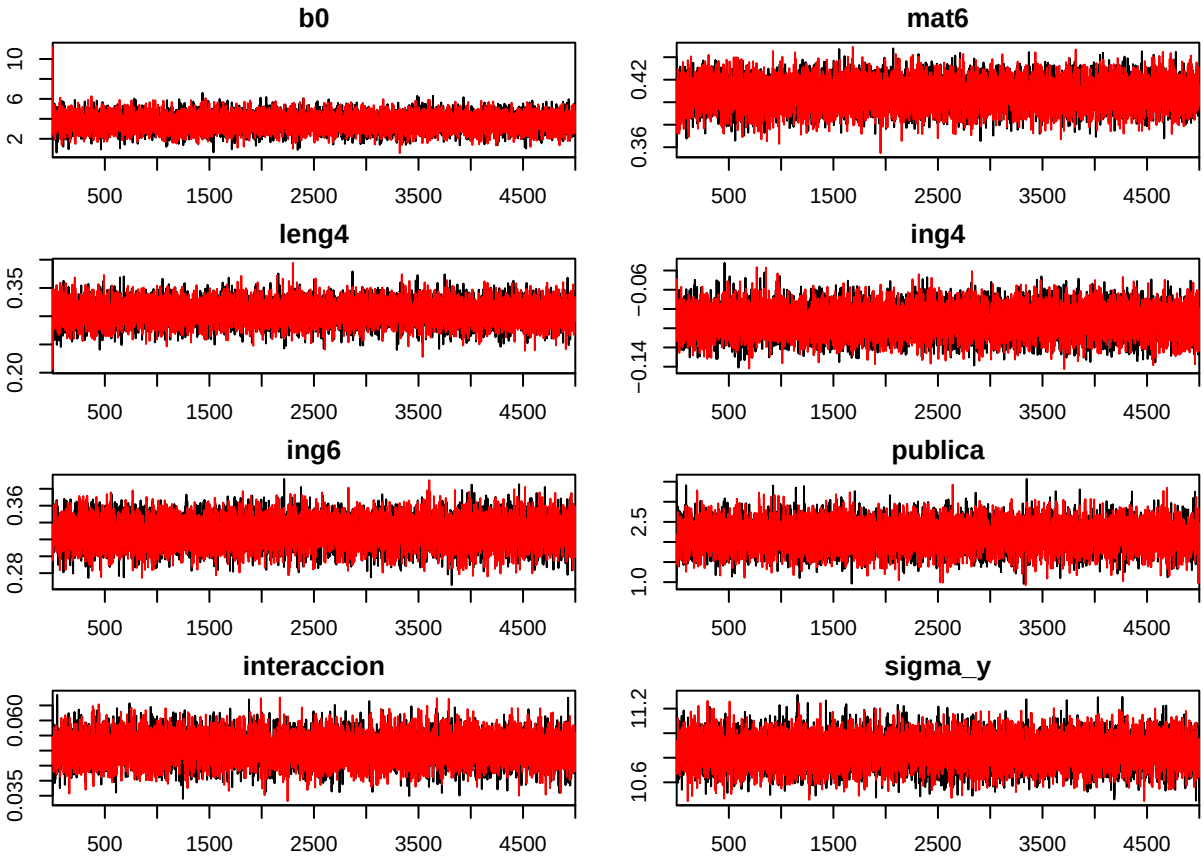
Using Truncated Noemal 0-100 Blue predicted, Red Dataset

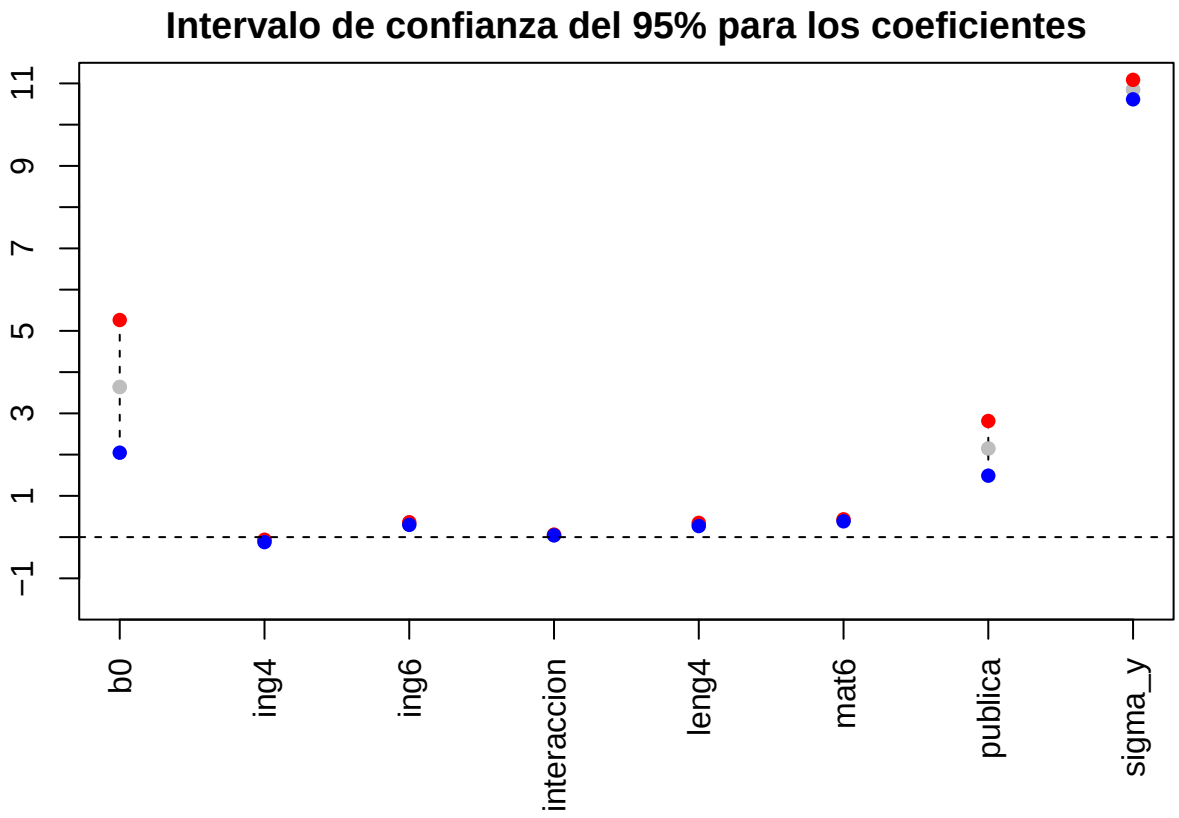


3.4 5) Modelo: quito "leng6"

We dont need this model since in model 3 all variable are significative

```
## module glm loaded

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 4000
##    Unobserved stochastic nodes: 8
##    Total graph size: 38862
##
## Initializing model
```


b0
mat6
leng4
ing4
ing6
publica
interaccion
sigma_y
500
1500
2500
3500
4500


Intervalo de confianza del 95% para los coeficientes
11
9
7
5
3
1
−1
b0
ing4
ing6
interaccion
leng4
mat6
publica
sigma_y

4 Implementación del modelo

5 Conclusions

6 Apéndice

6.1 Código R

6.2 Tablas parámetros estimados

7 Referencias