Universidad Nacional Autónoma de México

Facultad de Ciencias

Estadistica Bayesiana

2021-2

Proyecto:
Enfermedad del corazón.

Integrantes:
22. Hernández Joaquín Leopoldo
29. Luna Gutiérrez Yanely
Oyente. Ortiz Morán Raquel

15 de agosto del 2021

1. Introducción

La enfermedad de las arterias coronarias (CHD: *Coronary Heart Disease*) se desarrolla cuando dichas arterias se vuelven muy estrechas o quedan bloqueadas por el colesterol que se acumula en sus paredes. Las arterias suministran sangre, oxígeno y nutrientes al corazón, por lo que al estrecharse dismunuye el flujo sanguíneo y esto puede causar con el tiempo dolor de pecho. Una obstrucción completa de las arterias coronarias puede originar un ataque cardíaco. Algunas de los factores asociados como causas de esta enfermedad son el tabaquismo, la presión arterial alta y el colesterol alto.

La base de datos `SAheart`: *South Africa Heart Disease Data* recuperada de `http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data` contiene información sobre 462 hombres de la región Western Cape en Sudáfrica relacionada con los posibles factores de riesgo de la CHD y el diagnóstico de si tienen o no la enfermedad.

En este trabajo ajustamos una regresión logística sobre la variable que indica si la persona fue diagnosticada o no con CHD; primero usando un modelo de estadística clásica y posteriormente implementando el modelo seleccionado de estadística bayesiana.

2. Análisis descriptivo

La base de datos `SAheart` contiene 462 observaciones sobre las 10 variables siguientes:

- `sbp`: (*systolic blood pressure*) presión sanguínea medida en mmHg.
- `tobacco`: (*cumulative tobacco*) medido en kg.
- `ldl`: (*low density lipoprotein*) lipoproteína de baja densidad o colesterol malo.
- `adiposity`: Mide la severidad del sobrepeso de la persona.
- `famhist`: Historial familiar de la presencia de enfermedades del corazón. Toma los valores `Absent` y `Present`.
- `typea`: (*type-A behavior*)
- `obesity`: Indica la severidad de la obesidad en la persona.
- `alcohol`: Consumo actual de alcohol.
- `age`: Edad. Entre 15 y 64 años.
- `chd`: (*coronary heart disease*) ausencia (0) o presencia (1) de la enfermedad de las arterias coronarias. Hay 160 personas con la enfermedad y 302 de control.

Notamos que la variable respuesta `chd` y la variable `famhist` son categóricas por lo que las convertimos en factor. En la Figura 1 se muestran las gráficas de dispersión por pares de el resto de las variables coloreadas según el valor de `chd` 0 (verde) o 1 (rojo). Notamos que parece haber una correlación alta entre `adiposity` y `obesity`, por lo que podríamos tener problemas de colinealidad.

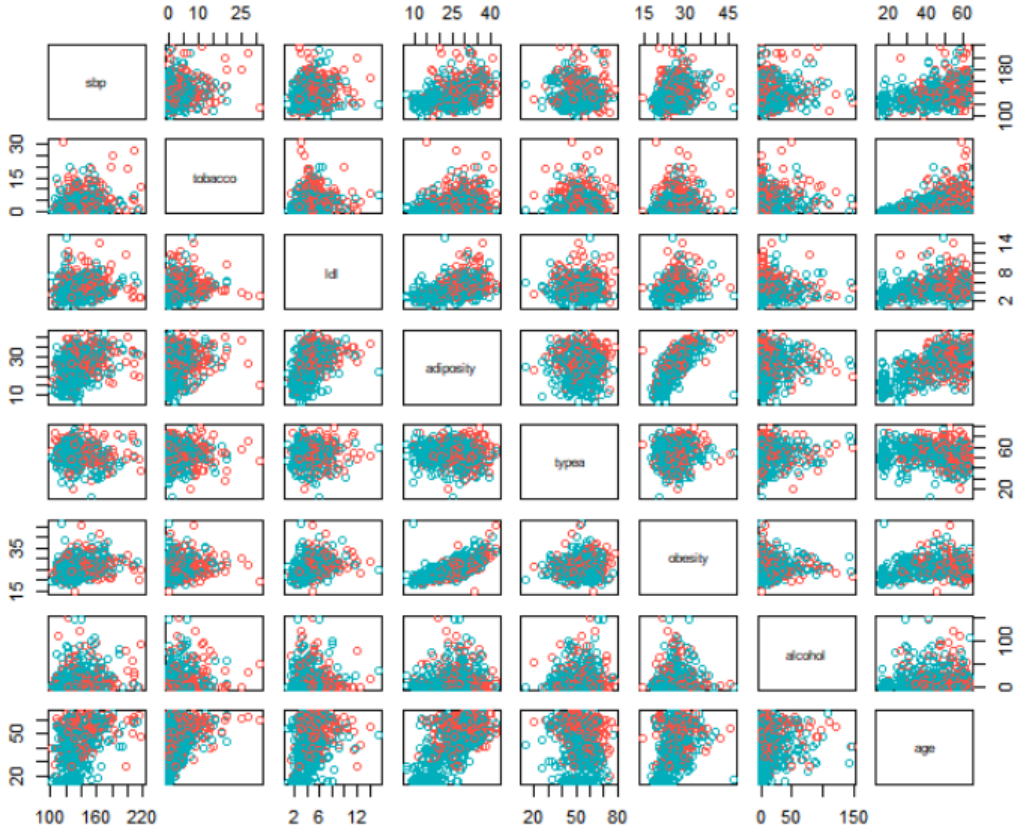

Figura 1: Gráficas de dispersión por pares de variables continuas.

En la Tabla 1 comprobamos que la correlación más alta es entre `adiposity` y `obesity`, con un valor de 0.72.

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
sbp	1.00	0.21	0.16	0.36	-0.06	0.24	0.14	0.39
tobacco	0.21	1.00	0.16	0.29	-0.01	0.12	0.20	0.45
ldl	0.16	0.16	1.00	0.44	0.04	0.33	-0.03	0.31
adiposity	0.36	0.29	0.44	1.00	-0.04	0.72	0.10	0.63
typea	-0.06	-0.01	0.04	-0.04	1.00	0.07	0.04	-0.10
obesity	0.24	0.12	0.33	0.72	0.07	1.00	0.05	0.29
alcohol	0.14	0.20	-0.03	0.10	0.04	0.05	1.00	0.10
age	0.39	0.45	0.31	0.63	-0.10	0.29	0.10	1.00

Cuadro 1: Correlaciones entre las variables continuas.

En la Figura 2 notamos que los valores de `sbp` están más dispersos cuando hay presencia de CHD que cuando no la hay, además de que la mediana es un poco más alta en presencia de CHD. Algo similar ocurre para las variables `tobacco`, `ldl` y `typea`.

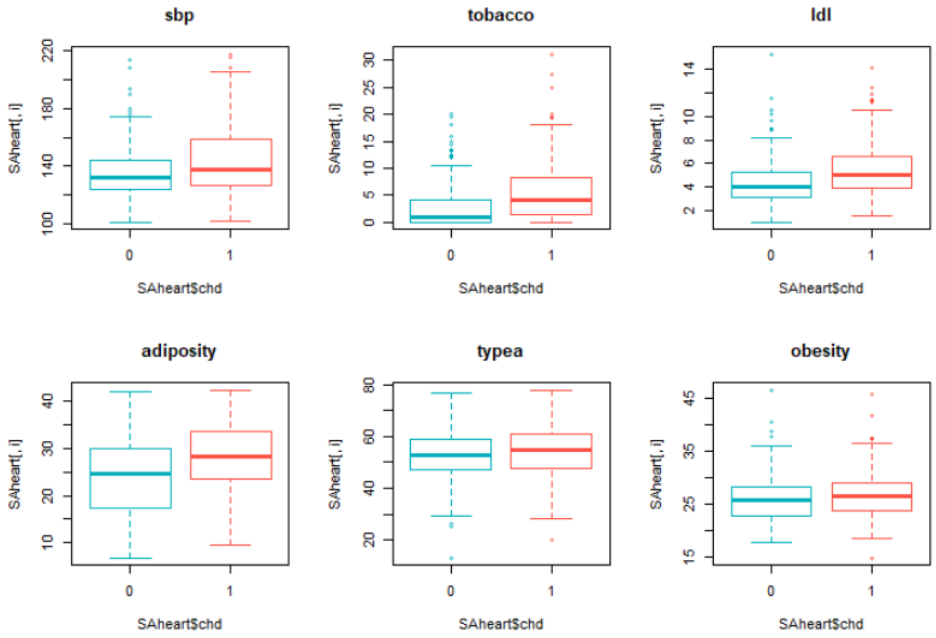

Figura 2: Distribución de algunas variables explicativas según el valor de `chd`.

Por otro lado, los valores de `adiposity` están más dispersos en ausencia de CHD, mientras que en presencia de CHD se concentran más en valores altos. En la Figura 3 notamos que los valores de `alcohol` se encuentran muy concentrados, por lo que tal vez nos convendría aplicar alguna transformación a esta variable. Para `age` notamos que hay mayor presencia de CHD en edades mayores.

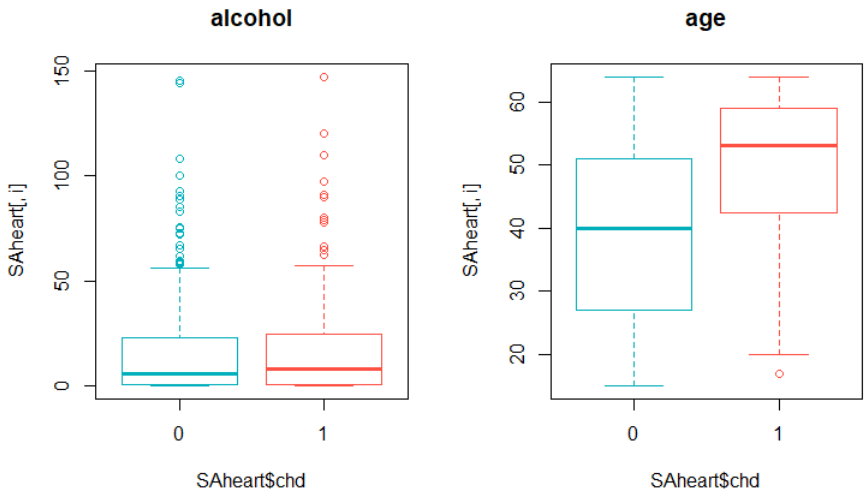

Figura 3: Distribución de las variables `alcohol` y `age` según los valores de `chd`.

En la Figura 4 notamos que cuando ya hay un historial familiar de enfermedades del corazón, existe una proporción cercana al 50 % de personas que presentaron CHD, mientras que en ausencia de historial familiar de enfermedades del corazón es mayor la proporción que no tiene CHD.

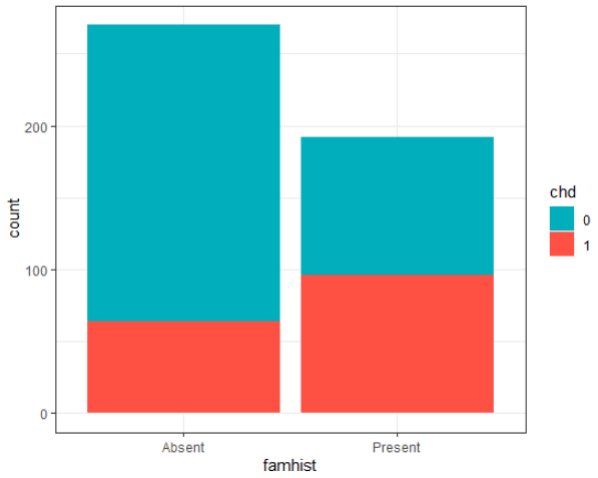

Figura 4: Valores de `famhist` y su relación con `chd`.

Como se ha visto en las Figuras 1, 2 y 3, los valores de `alcohol`, `tobacco` y `obesity` están muy concentrados, por lo cual aplicar una transformación puede ayudar a resolver estos problemas. La transformación que usaremos (es de las más usuales) es aplicar logaritmo natural a los valores de `alcohol`, `tobacco` y `obesity`, aunque antes se le sumara 0.1 a todos los valores para evitar errores cuando se evalua en el cero.

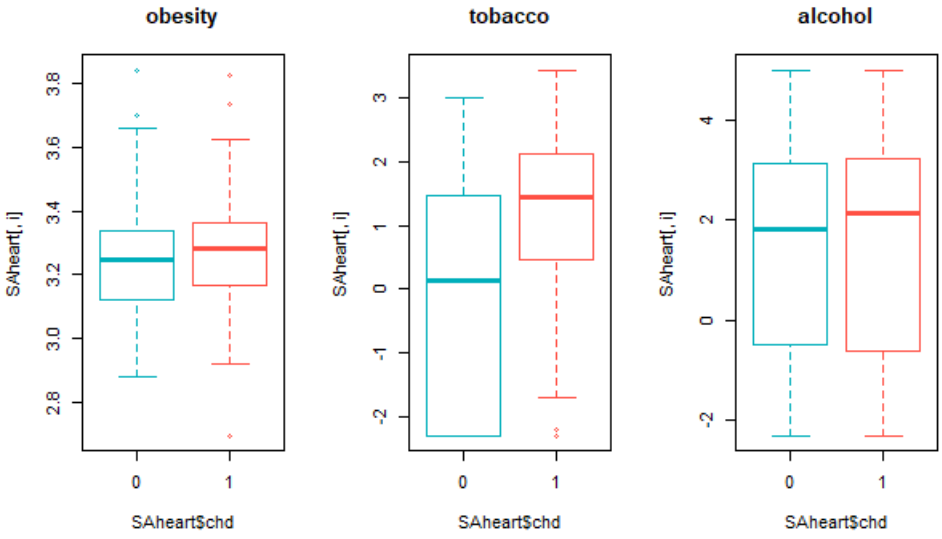

Figura 5: Distribución de `alcohol`, `tobacco` y `obesity` para cada valor de `chd` considerando la transformación.

En la Figura 5, se puede observar que los valores de `alcohol`, `tobacco` y `obesity` aplicando la transformación ahora se encuentran más dispersos.

3. Ajuste del modelo de Regresión logística

Para el ajuste del modelo consideramos tres posibles modelos:

- Modelo 1: Todas las variables.
- Modelo 2: Solo las variables `tobacco,ldl,famhist, typea` y `age`.
- Modelo 3: Las variables `tobacco,ldl,famhist, typea` y `age`, y las interacciones `tobacco*typea` y `ldl*famhist`.
- Modelo 4: Las variables `tobacco, typea` y `age`, y las interacciones `tobacco*typea` y `ldl*famhist`.

En el Modelo 1 notamos que solo las variables `tobacco,ldl,famhist, typea` y `age` resultaban ser significativas, por lo que usamos estas para ajustar el Modelo 2 y en éste modelo todas las variables explicativas son significativas al nivel 0.05. Además, realizamos la prueba de devianza en la que obtuvimos un *p-value* de 0.47, lo que nos indica que es mejor el Modelo 2, es decir, no es necesario incluir todas las variables del Modelo 1.

En el Modelo 3 agregamos las interacciones entre `tobacco` y `typea` y entre `ldl` y `famhist`. Notamos que en este modelo las variables `ldl` y `famhistPresent` dejan de ser significativas, sin embargo, al hacer la prueba de devianza contra el Modelo 2, resulta un *p-value* de 0.0013, por lo que elegimos el Modelo 3 sobre el Modelo 2. Dado que las variables `ldl` y `famhistPresent` no son significativas preferimos quitarlas, pues aunque tanto la devianza como las tasas de error aumentaban al quitarlas, este cambio no era significativo. Que estas variables dejaran de ser significativas una vez se agrego una interacción entre ellas indica que posiblemente que una persona tenga antecedente familiar de enfermendades coronarias o un alto colesterol malo no influiran por si solos en que esta desarrolle la enfermedad, pero una combiación de ambas sí podría indicar la presencia futura de la enfermedad.

De este modo, ajustamos el cuarto modelo. Notamos que existian problemas de colinearidad con las variables `tobacco` y `tobacco*typea`, pero al quitar una de las variables notamos un aumento en los indicadores y en las tasas de error por lo que preferimos dejarlas. No fue posible utilizar la función `anova` en este caso, pues tenemos modelos no anidados.

Para seleccionar el modelo consideramos los criterios de Devianza, AIC y BIC, así como las tasas de error aparentes y de prueba, utilizando el método de *train/test*, con un conjunto de entrenamiento del 85 % del total de los datos. La Tabla 2 resume los resultados obtenidos.

	Devianza	AIC	BIC	ApGlob	Ap0	Ap1	TestGlob	Test0	Test1
Modelo 1	470	490	532	26.0	15.6	45.6	29.0	15.6	54.2
Modelo 2	475	487	512	25.2	15.6	43.4	31.9	20.0	54.2
Modelo 3	460	476	509	22.9	12.5	42.6	27.5	13.3	54.2
Modelo 4	462	476	505	24.2	14.4	42.6	27.5	13.3	54.2

Cuadro 2: Medidas para evaluar el desempeño de los modelos.

El modelo seleccionado bajo los criterios anteriormente mencionados fue el 4, pues era el modelo más parsimonioso, con todas las variables siendo significativas, tenia los indicadores más pequeños y sus tasas de error eran las mejores de los 4 modelos, tanto para el conjunto de entrenamiento como para el conjunto de prueba. Los coeficientes y *odd ratios* del modelo se muestran en la Tabla 3.

	Coeficiente	*odd ratio*
Intercepto	-6.7723	
`tobacco`	1.2896	3.631
`typea`	0.0526	1.054
`age`	0.0517	1.053
`tobacco*typea`	-0.0191	0.981
`ldl*famhistAbsent`	0.06320	1.065
`ldl*famhistPresent`	0.26730	1.306

Cuadro 3: Coeficientes y *odd ratios* del Modelo 4.

4. Ajuste del modelo de Regresión logística bayesiano

Los modelos de esta sección fueron ajustados usando el paquete de `R` `rjags`. Para el primer modelo propusimos como variables explicativas solo las variables continuas `tobacco`, `ldl`, `typea` y `age`, para las cuales usamos distribuciones iniciales normales no informativas (normal estándar). Para el ajuste del modelo usamos 3 cadenas y descartamos las primeras 1000 iteraciones.

En la Figura 6 podemos ver que las cadenas para los coeficientes de las variables, a excepción de alpha (el intercepto) se ven bastante estables antes de la iteración 6000 y las tres convergen.

En la Figura 8 comprobamos que no tenemos problemas de convergencia en ninguna variable y que de hecho, con menos interacciones tampoco hay problemas.

Revisando el resumen de este modelo notamos que los intervalos de confianza estimados para cada coeficiente no contienen el 0, es decir, las cuatro variables explicativas (y el intercepto) son significativas.

Los valores de las tasas de error de clasificación de este modelo se muestran en la Tabla 4.

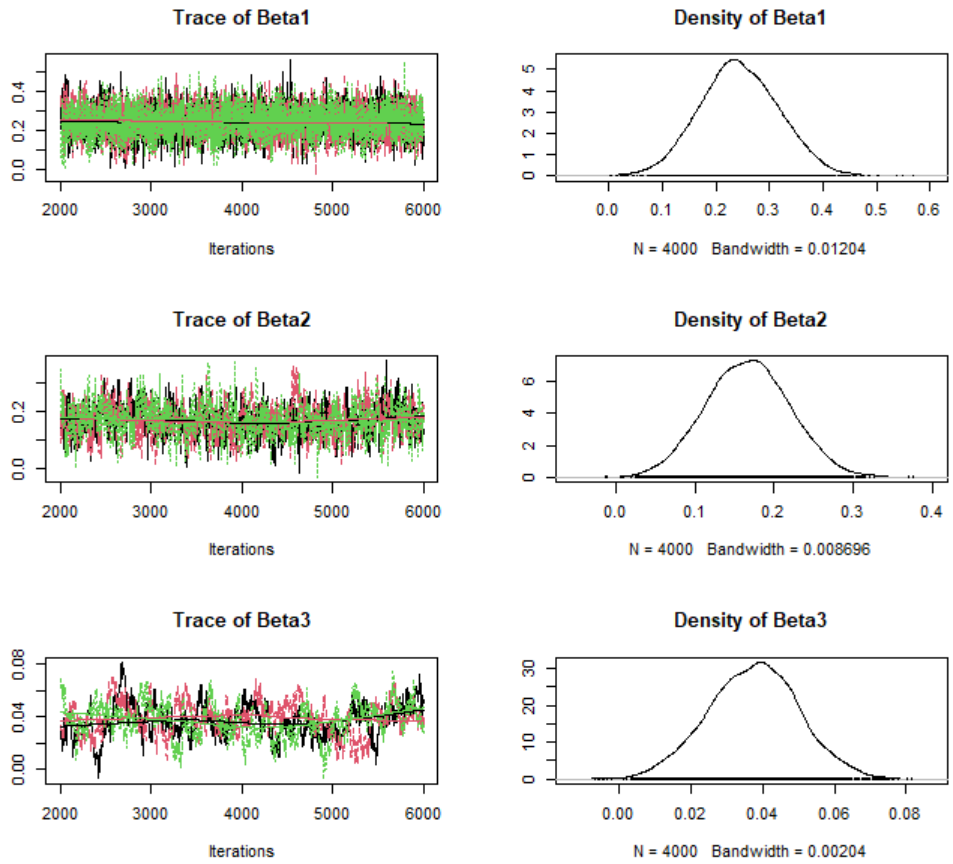

Figura 6: *trace-plots* de las cadenas para el modelo 1 y densidades a posterior de los parámetros.

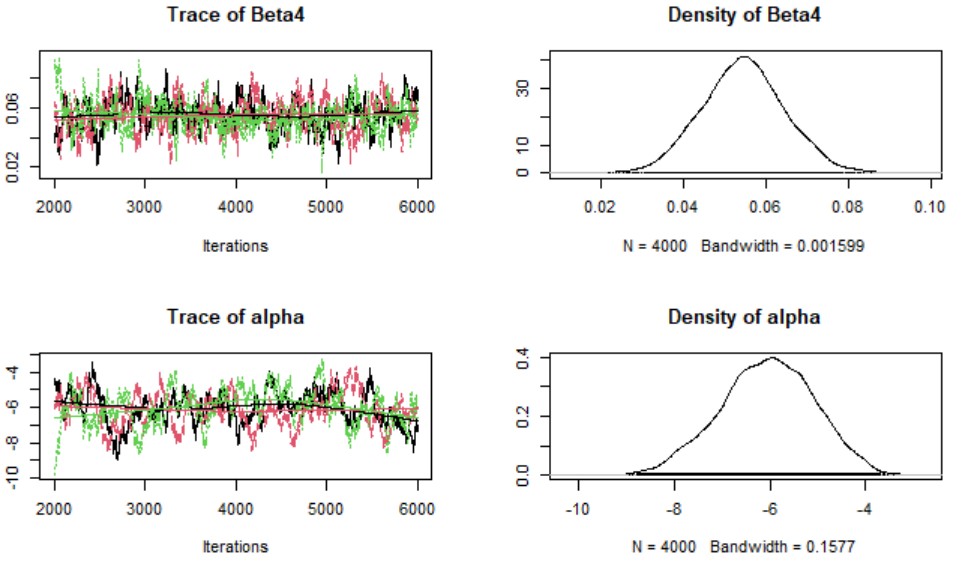

Figura 7: *trace-plots* de las cadenas para el modelo 1 y densidades a posterior de los parámetros.

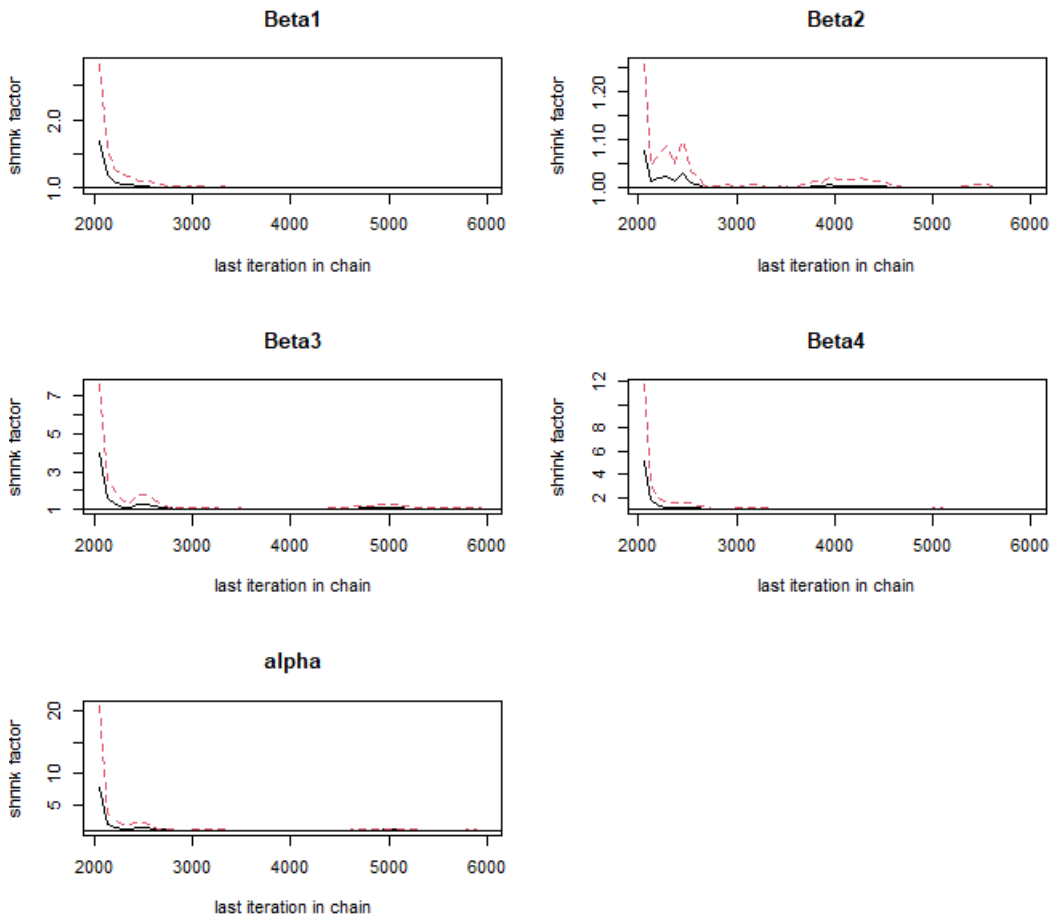

Figura 8: Diagnóstico de convergencia para los coeficientes del modelo 1.

Para el segundo modelo, propusimos como variables explicativas las variables continuas (`tobacco`, `ldl`, `typea`, `age`) y la variable categoritca (`famhist`), para las cuales usamos distribuciones iniciales normales no informativas (normal estándar). Para el ajuste del modelo usamos 3 cadenas y descartamos las primeras 1000 iteraciones.

Las Figuras 9 y 10, presentamos las cadenas de nuestras β_i donde $i = 1, \ldots, 5$, de las cuales son los coeficientes de nuestras variables en el siguiente orden `tobacco`, `ldl`, `famhist`,`typea` y `age`. Las cadenas de β_1, β_2 y β_3, podemos apreciar que convergen de manera adecuada, para las 3 cadenas siguientes aunque se ve que convergen, todavia faltarían unas interacciones.

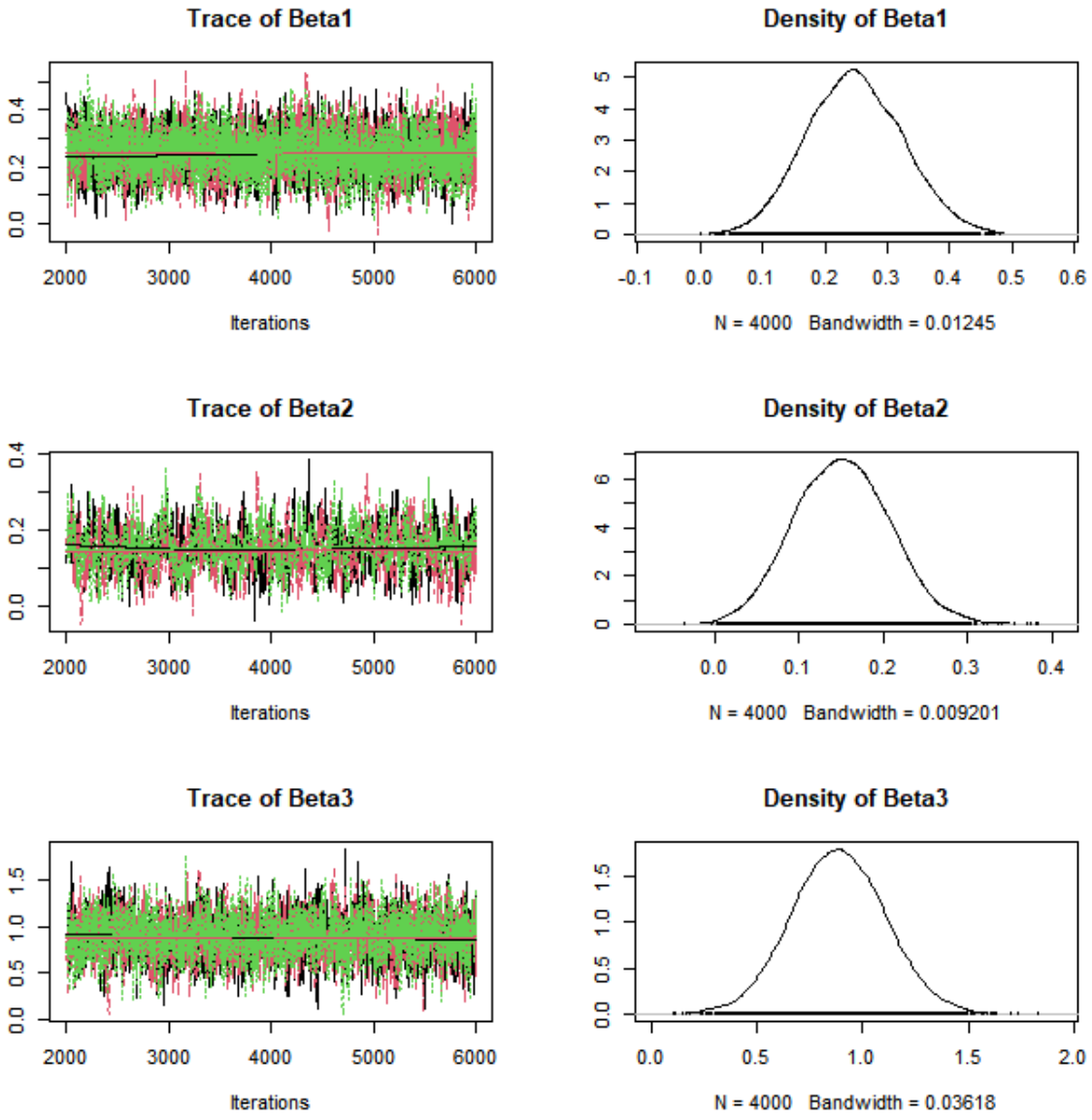

Figura 9: *trace-plots* de las cadenas para el modelo 2 y densidades a posterior de los parámetros.

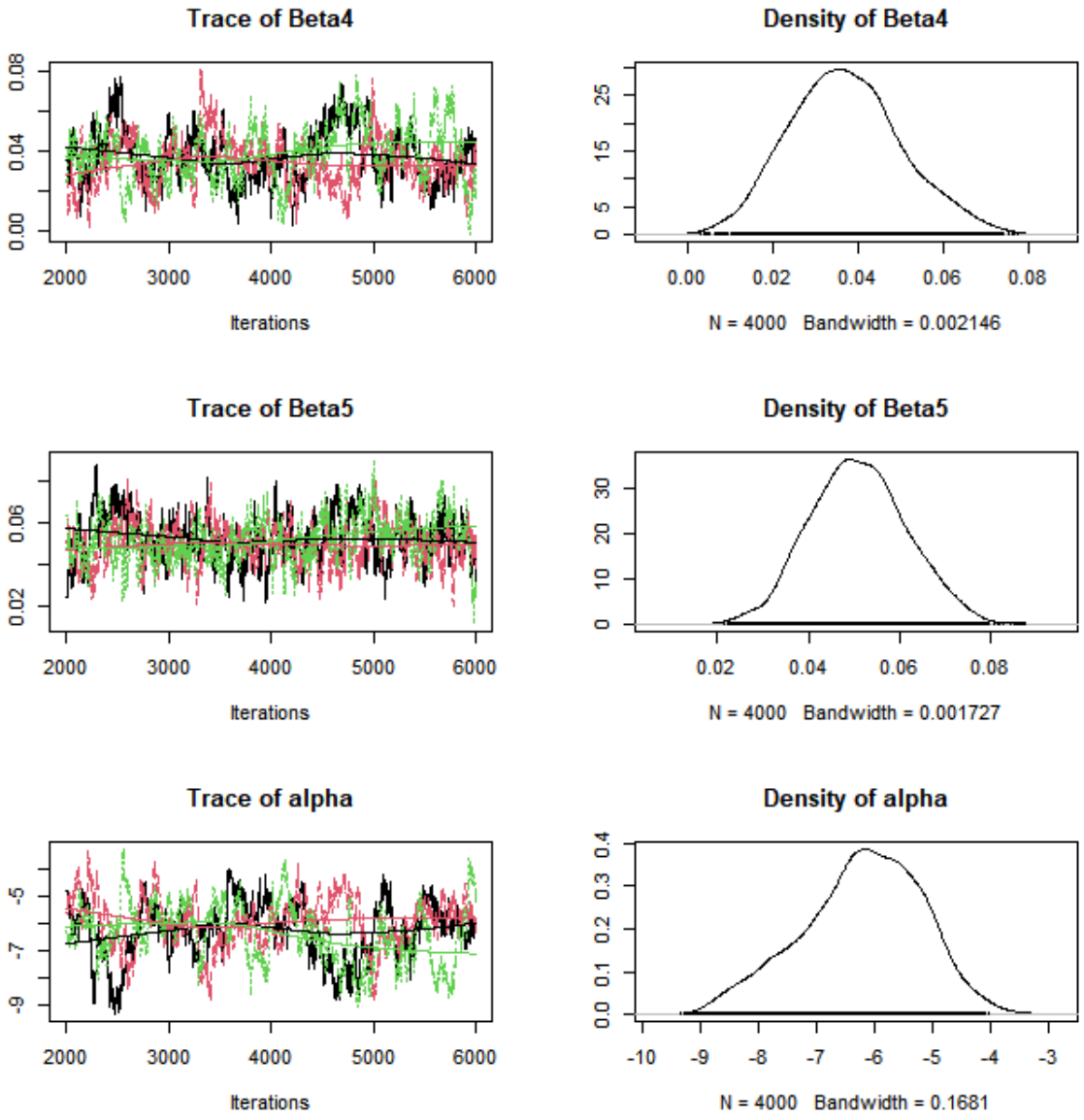

Figura 10: *trace-plots* de las cadenas para el modelo 2 y densidades a posterior de los parámetros.

En la Figura 11, podemos ver que convergen todas las β antes de las 6000 interacciones, e incluso excepto β_1 β_2 y β_5, podemos decir que convergen antes de las 4000 interacciones.

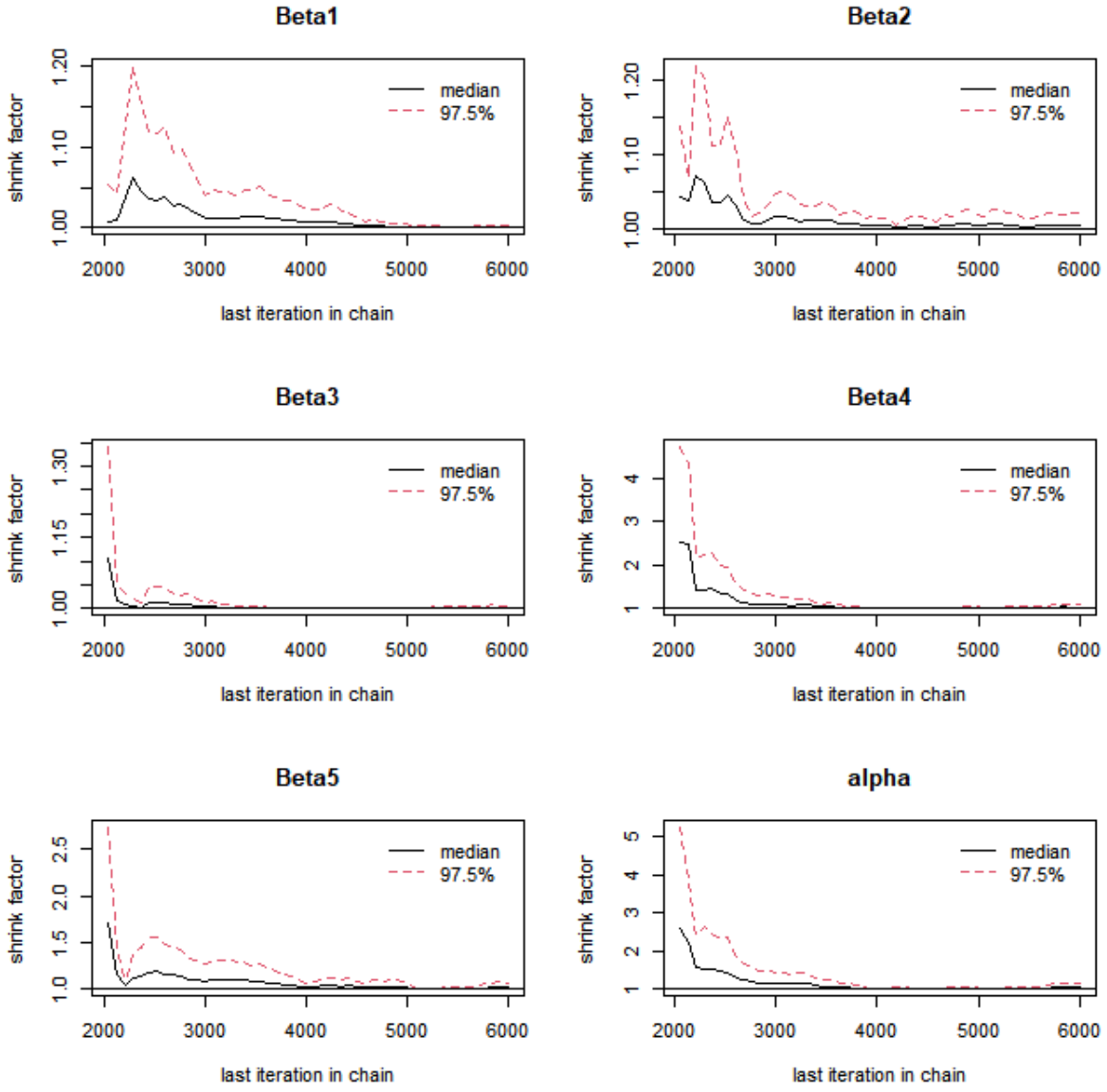

Figura 11: Diagnóstico de convergencia para los coeficientes del modelo 2.

Como se menciono en la Figura 11, β_1, β_2 y β_5, le faltarían interacciones para asegurar que convergen adecuadamente y no hacen algun tipo de salto, para esto se descartaran del modelo 4000 interacciones. Por lo que en las Figuras 12 y 13, se muestran las cadenas y densidades a posterior donde ya todas las cadenas se ven mejores y se arreglan los problemas que se observaban cuando solo se descartaban 1000 interacciones.

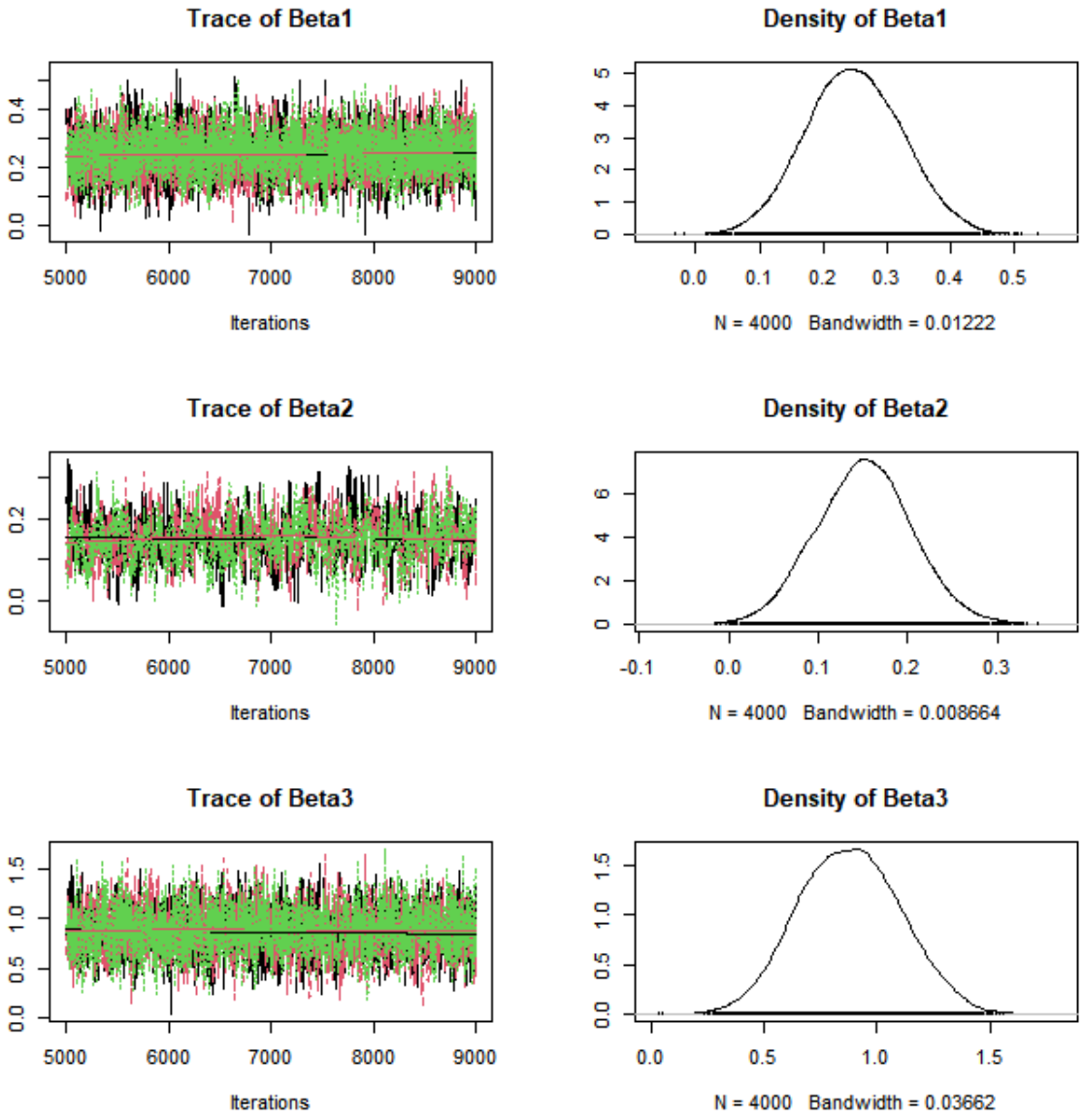

Figura 12: *trace-plots* de las cadenas para el modelo 2 y densidades a posterior de los parámetros.

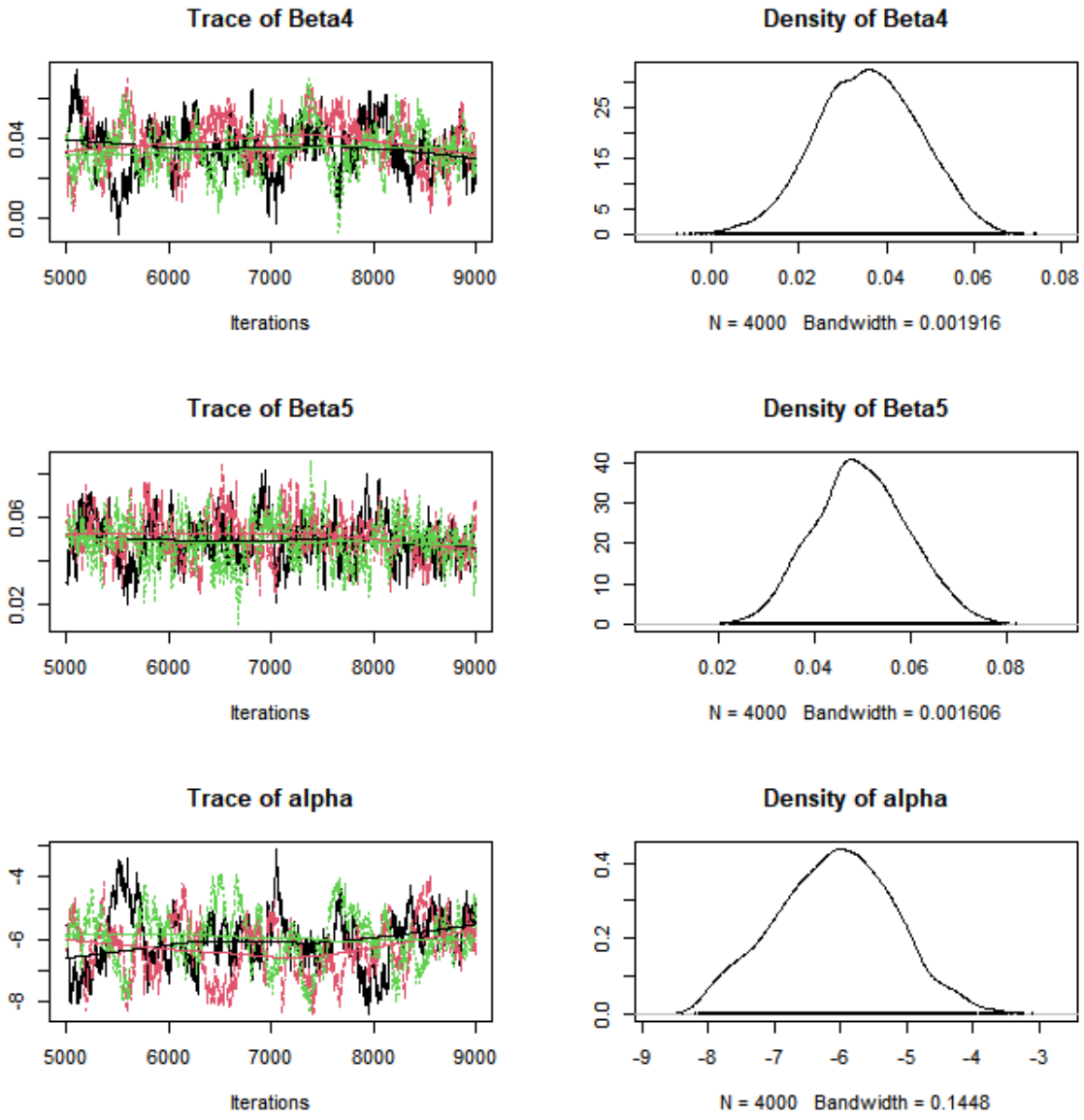

Figura 13: *trace-plots* de las cadenas para el modelo 2 y densidades a posterior de los parámetros.

En la Figura 14, ya se puede observar que convergen todos los coeficientes e incluso antes de las 7000 interacciones, como se meciona se resuelve el problema que surge anteriormente.

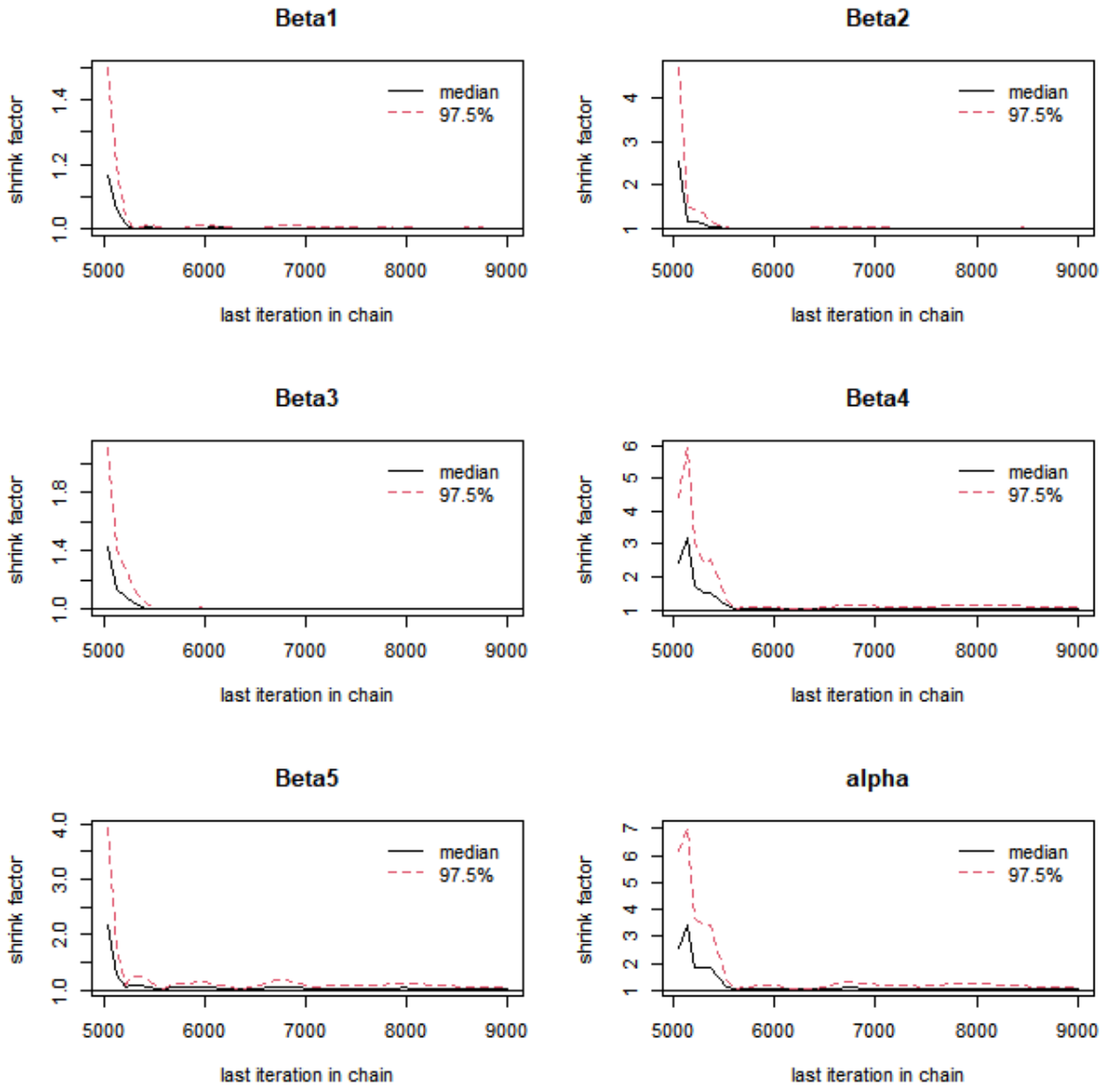

Figura 14: Diagnóstico de convergencia para los coeficientes del modelo 2.

Finalmente, para el tercer modelo propusimos como variables explicativas 4 variables continuas (tobacco, ldl, typea y age), la variable categórica (famhist), y 2 interacciones (tobacco*typea y ldl*famhistPresent) para las cuales usamos distribuciones iniciales normales no informativas (normal estándar). Para el ajuste del modelo usamos 3 cadenas y descartamos las primeras 1000 iteraciones.Las Figuras 15 y 16, presentan las cadenas de nuestras β_i.

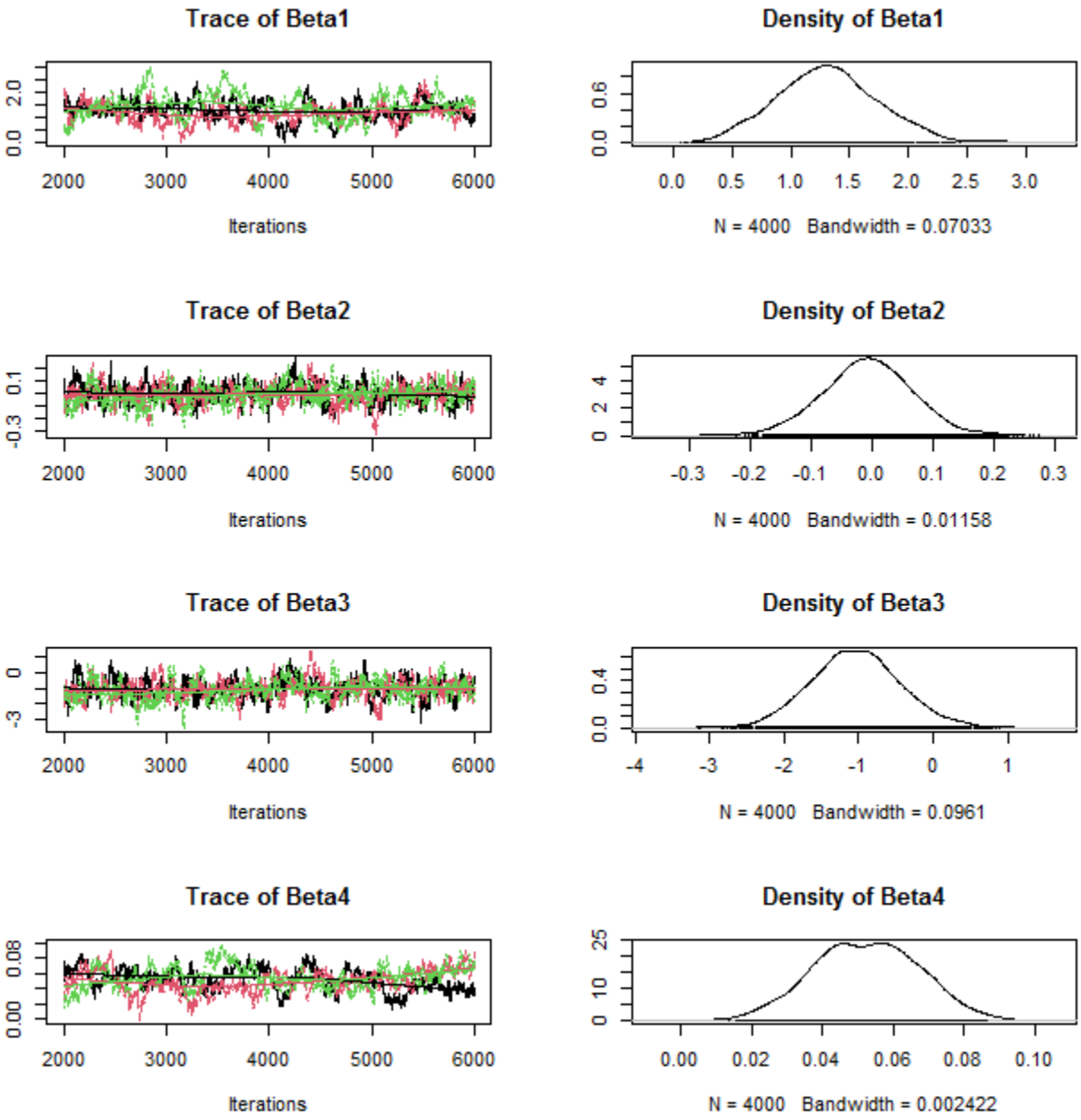

Figura 15: *trace-plots* de las cadenas para el modelo 3 y densidades a posterior de los parámetros.

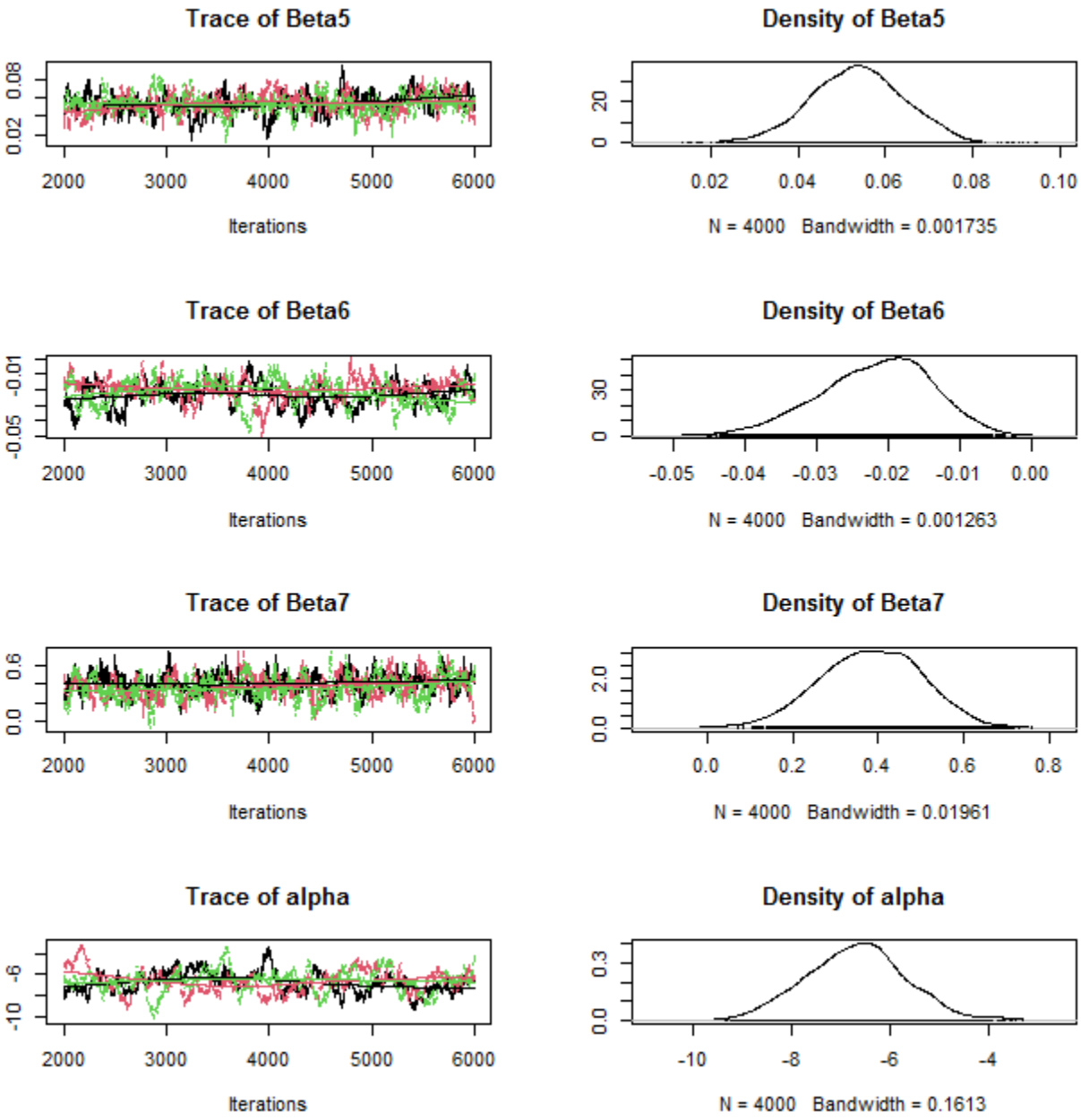

Figura 16: *trace-plots* de las cadenas para el modelo 3 y densidades a posterior de los parámetros.

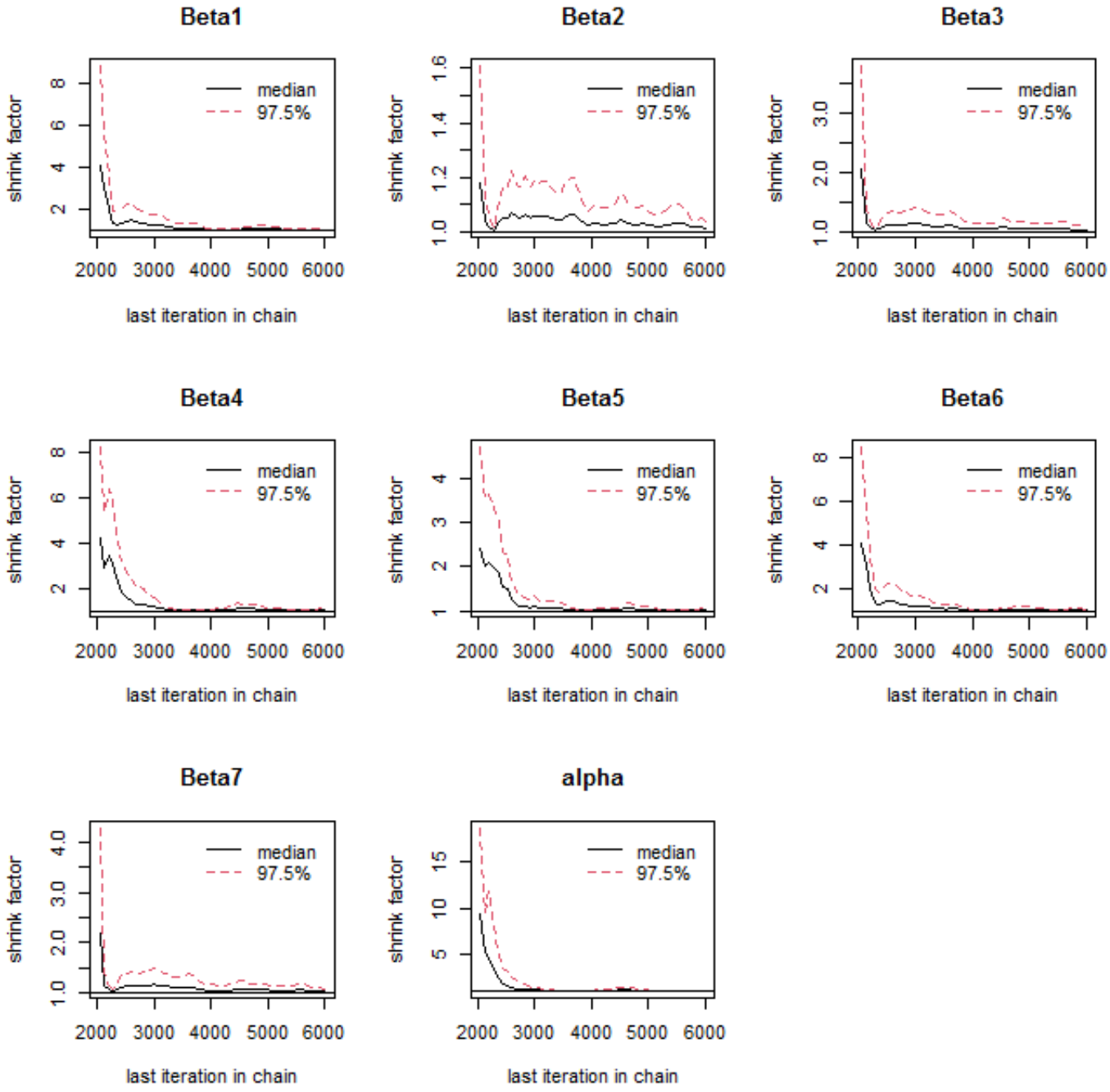

Figura 17: Diagnóstico de convergencia para los coeficientes del modelo 3.

Modelo	Error global	Error para chd=0	Error para chd=1
Modelo 1	28.79	15.89	53.12
Modelo 2	25.5	15.2	45.0
Modelo 3	24.68	13.25	46.25

Cuadro 4: Errores de clasificación aparentes para cada modelo bayesiano.

Aunque el modelo 3 cuenta con un error de clasificación menor a los otros dos modelos previos, notamos que los coeficientes β_2 y β_3 no son significativos (sus intervalos del 95 % de confianza contienen al cero). Además, solo se mejora la tasa de error de clasificación en menos del 1 % con respecto al modelo 2. Dado que en el modelo 2 todos los coeficientes resultan significativos optamos por elegir este modelo como nuestro modelo final.

Los coeficientes estimados con el modelo 2 y los límites del intervalo del 95 % de cada uno se muestran en la Tabla 5.

Variable	Media	Cuantil 2.5 %	Cuantil 97.5 %
`tobacco`	0.248	0.1022	0.4012
`ldl`	0.1521	0.0464	0.2582
`famhist==Present`	0.888	0.4496	1.3282
`typea`	0.037	0.0127	0.0649
`age`	0.051	0.0311	0.0723
Intercepto	-6.224	-8.4299	-4.3884

Cuadro 5: Valores estimados de los coeficientes del modelo 2.

5. Conclusiones

- No se observaron muchas diferencias en el ajuste del modelo Bayesiano y el Clásico, lo que se esperaba dado que usamos prioris no informativas, pero no se selecciono el mismo modelo, esto es, el modelo bayesiano y el frecuentista no consideran las mismas variables explicativas.

- El hecho de que variables que se esperaba que influyeran en la variable respuesta como obesidad y sbp no fueran significativas se puede explicar por el hecho de que el estudio fue retrospectivo. No podemos asegurar que la obesidad o la presión sanguínea de una persona no influyan en su predisposición a la enfermedad pero estos datos no pudieron captarlo, ya que quienes reportaban la presencia de la enfermedad, en muchos casos, ya habian realizado cambios a su estilo de vida, controlando su peso y presión sanguínea.

- Las variables que más ayudan a clasificar a la variable respuesta en el modelo logístico son: tobacco y ldl*famhist, mientras que en el modelo bayesiano son: tobacco y famhist. Esto significa que lo que más influye en la presencia de la enfermedad es la cantidad de tabaco que una persona fuma y el tener historial familiar de esta enfermedad.