Trabajo Final de Estadística Bayesiana

Daniel Felipe Villa Rengifo

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2022

Contenido

1 Introducción

La estructura del modelo bayesiano permite capturar las relaciones de dependencia que existe entre los atributos de los datos que se estudien, por medio de una distribución "posterior", por lo cual en este artículo veremos una aplicación en la rama de la salud, específicamente en casos de cáncer de mama, esto aplicando inferencia bayesiana por medio de un modelo logístico binario, para ver las probabilidades según un conjunto de datos conocidos de que X mujer (hablando desde la parte biológica, abstrayendo todo concepto de género diferente) pueda ser diagnosticada con cáncer de mama (sin discriminación por tipo de cáncer).

Más adelante veremos por medio de diagnósticos (la mayoría, por no decir todos) visuales, para un mejor entendimiento (cada uno de ellos explicados posteriormente) del modelo y si este nos ayuda a predecir el problema en cuestión, además veremos un cambio en las variables de respuesta por lo que veremos otros modelos del mismo tipo para una comparación entre ellos por medio del BF(Factor de Bayes).

2 Objetivo

Proponer un modelo por el cual se pueda predecir según los datos del examen ambulatorio si la mujer examinada presenta cáncer de mama.

2.1. Preguntas de interés

Generamos algunas preguntas, las cuales al finalizar este articulo las responderemos

- ¿Será la densidad del seno según la escala de $BI - RADS$ una variable de peso para el modelo logístico?

- ¿Los antecedentes familiares serán razán suficiente para determinar signos de cáncer de mama?

- ¿Mayor IMC implica mayores casos de cáncer de mama?

3 Antecedentes

En base al software estadístico R versión 4.2.0, con la ayuda de los paquetes:

- tidyverse
- magrittr
- janitor
- readr
- BayesFactor
- bayesplot
- brms
- RStan

con las propiedades de RStan para la construcción del modelo logístico; *cabe destacar que donde se corrieron los modelos no contaban con un gran poder computacional, por lo que los tiempos de ejecución eran muy largos.*

La base de datos fue tomada del Instituto Nacional del Cáncer (INC), donde recopilaron información de 20.000 mamografías digitales y 20.000 de proyección de película realizadas entre enero de 2005 y diciembre de 2008 de mujeres incluidas en el Consorcio de Vigilancia del Cáncer de Mama en las cuales encontramos las siguientes variables:

Nota: *La base de datos contenía muchas más variables pero decidimos tomar las más representativas, según un asesor externo que guió el proceso de seleccionar las columnas con mayor peso para el desarrollo de este trabajo practico.*

Si necesita conocer más:
"Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). You can learn more about the BCSC at: http://www.bcsc-research.org/."

DIGITAL MAMMOGRAPHY DATASET DOCUMENTATION

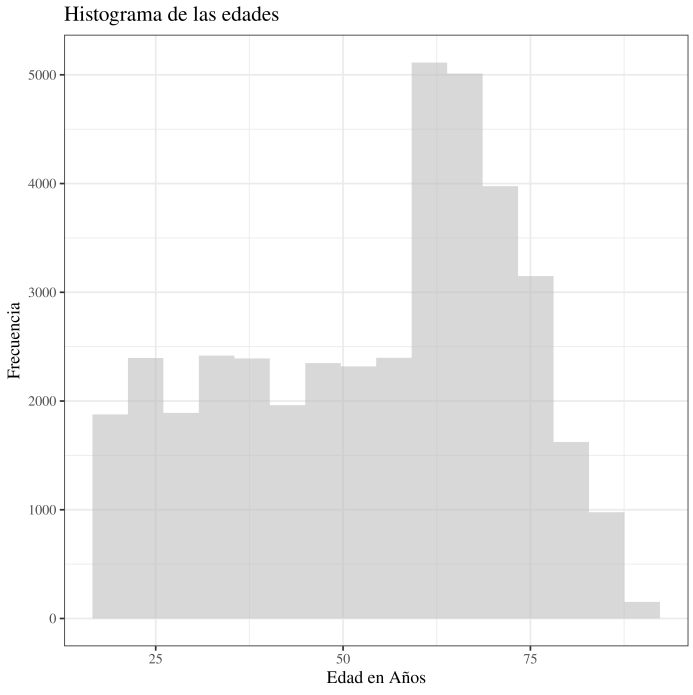
Nombre de la Variable	**Descripción**	**Codificación**
Edad	Edad de la paciente en años en el momento de la mamografía	Numérico
Evaluación	Valoración del radiólogo basada en la Escala BI-RADS (Hallazgos de posibles masas o tumores)	0 = Necesita imagenes adiocionales 1 = Negativo 2 = Hallazgo(s) Benigno(s) 3 = Probablemente Benigno 4 = Anomalía sospechosa 5 = Probablemente maligno
Desisdad	Densidad mamaria en escala BI-RADS de la paciente registrada en el momento de la radigrafía	1 = Casi totalmente Graso 2 = Densidades fibroglandulares dispersas 3 = Heterogéneamente denso 4 = Extremadamente denso
Antecendentes Familiares	antecedentes familiares de cáncer de mama en un familiar de primer grado	0 = no 1 = sí 9 = falta
IMC	índice de masa corporal en el momento de la mamografía	Numérico
Cáncer	*indicador binario de diagnóstico de cáncer en la entrega de resultados del examen ambulatorio*	*0 = Negativo* *1 = Positivo*

4 Análisis Descriptivo

En este capítulo veremos el comportamiento y algunas tablas de cada variable de interés para el tratamiento del modelo logístico, por medio de un análisis descriptivo, es decir, veremos nuestras por medio de gráficos para un mejor entendimiento; ya para el caso de las variables categóricas veremos su representación el tablas por medio de porcentajes.

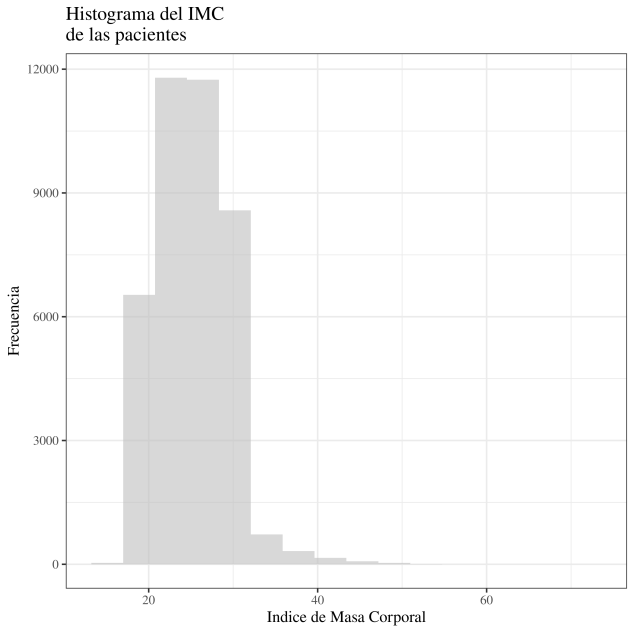
4.1. Edad

La variable edad tiene un comportamiento algo interesante, si observamos las edades menores a 60 años vemos un comportamiento algo uniforme, para las edades entre los 60 a los 90 años (edad máxima de la base de datos) tiene un comportamiento decreciente, con un pico muy alto desde los 60 hasta los 70 años, esto podría hacernos pensar que las mujeres hasta los 60 años no se preocupan por este tipo de exámenes hasta que llegan edades avanzadas donde sí es importante, es como el caso de los hombres y los examenes de prostata.



4.2. IMC

Para el caso del IMC, es algo estable a pesar de que su distribución sea algo aglomerada entre los rangos más comunes, ya que esto representa una población sana con bajas tendencias de obesidad o desnutrición, los rangos comunes van desde los $18,5$ y los 30 puntos.



4.3. Evaluación BI-RADS

En la evaluación de la encontrar algún tipo de tumor, la mayor población pertenece a tener masas benignas seguida de tener una probabilidad de 50/50, esto nos dice que muchas de las mujeres asistían seguramente porque presentaban alguna anomalía en el seno.

Evaluación	**Freq**
0	0.09
1	0.43
2	0.23
3	0.10
4	0.10
5	0.05

4.4. Densidad del Seno

La proporción de mujeres en cómo se distribuye la densidad del seno vemos que la mayoría presentaba alguna combinación entre la adiposidad y fibras glandulares duras. Esto se debe a que entre más denso sea el seno más riesgos (médicamente comprobado) de contraer cáncer de mama.

Densidad	**Freq**
1	0.07
2	0.76
3	0.15
4	0.02

4.5. Antecedentes Familiares

Solo un pequeño porcentaje de mujeres tienen familiares de primer grado con esta clase de enfermedad terminal, pero también vemos como el 17 % de las pacientes no sabe o no responde si sabe que sus familiares de primer grado presentaron cáncer de mama.

Antecedentes Familiares	**Freq**
0	0.58
1	0.25
9	0.17

4.6. Diagnostico de Cáncer

Para el caos del cáncer es el menos favorable ya que los pacientes diagnosticados con cáncer solo representan el 3 % de la población, esto de antemano nos dirá que los parámetros no serán nada significativos ya que solo la base de datos contiene muy pocos datos de personas con cáncer de mama, nuestra distribución posterior tendrá el mayor peso en $cáncer = 0$.

Cancer	**Freq**
0	0.97
1	0.03

5 Modelo

Utilizaremos un modelo de regresión logística con enfoque bayesiano para estimar la probabilidad de ser diagnosticado con cáncer en función de la edad, el IMC, antecedentes familiares, densidad del seno, entre otros.

Podemos pensar que tenemos una variable de respuesta definida como:

$$\mathrm{Y} = \begin{cases} 1 & \text{Positivo para Cáncer} \\ 0 & \text{Negativo para Cáncer} \end{cases}$$

y nos interesa modelar $\pi = P(Y = 1)$ (también conocida como probabilidad de éxito) en función de 5 variables explicativas:

- Edad
- Evaluación $BI - RADS$
- Densidad del Seno
- Antecedentes Familiares
- Índice de Masa Corporal (IMC)

Una regresión logística es un modelo que relaciona a $logit(\pi)$ como una combinación lineal de los predictores. La ecuación matemática de nuestro modelo es:

$$log(\tfrac{\pi}{1-\pi}) = \beta_0 + \textstyle\sum_{i=1}^{5} \beta_i X_i, \quad i = 1, 2, 3, 4, 5$$

Donde $X_i,\ i = 2, 3, 4$ son variables categóricas con nieveles que explicamos en la tabla anterior, las otras dos son numéricas, por lo que a cada uno se le dará el primer nivel como guía, para le calculo de los coeficientes.

Si representamos el lado derecho de la ecuación del modelo con η , la expresión puede reordenarse para expresar nuestra probabilidad de interés, π , en función del predictor lineal η.

$$\pi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}$$

Dado que somos personas bayesianas que extraen muestras de **posteriors**, necesitamos especificar una prior para los parámetros, así como una función de verosimilitud antes de realizar nuestra tarea. En esta ocasión, vamos a utilizar las prior por defecto del sistema de RStan, ya que al no definir las priors del modelo, el asume que necesitamos parámetros nada informativos, de esa manera tiene una prior llamadas *Flat* en las cuales se basa en una distribución uniforme.

La probabilidad es el producto de los ensayos de Bernoulli

$$\prod_{i=1}^{n} p_i^y (1-p_i)^{1-y_i} \text{ donde } p_i = P(Y=1) \text{ y } y_i = 1$$
si da positiva la prueba y $y_i = 0$ resultados negativos para cáncer.

5.1. Modelo con Brms-Package

```
model2 <- brm(formula = cancer_c ~ age_c + assess_c + famhx_c + density_c
          +bmi_c,
          data=df,
         family = bernoulli(link = "logit"),
         warmup = 3000,
         iter = 15000,
         chains = 3,
         inits= "0",
         cores=2,
         seed = 123)
```

Aquí vemos que definimos un modelo *Bernoulli* con 15,000 iteraciones y un burning de 3,000 iteraciones por cadena, dandonos un total de 36,000 iteraciones totales. Para la segunda parte, veremos el codigo definido en RStan por el cual definimos nuestras priors y toda la configuración que Stan necesita.

5.2. Modelo Con Stan

```
    functions {
}
data {
  int<lower=1> N;  // total number of observations
  int Y[N];  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  int prior_only;  // should the likelihood be ignored?
}
transformed data {
  int Kc = K - 1;
  matrix[N, Kc] Xc;  // centered version of X without an intercept
  vector[Kc] means_X;  // column means of X before centering
  for (i in 2:K) {
    means_X[i - 1] = mean(X[, i]);
    Xc[, i - 1] = X[, i] - means_X[i - 1];
  }
}
parameters {
  vector[Kc] b;  // population-level effects
  real Intercept;  // temporary intercept for centered predictors
}
transformed parameters {
  real lprior = 0;  // prior contributions to the log posterior
  lprior += student_t_lpdf(Intercept | 3, 0, 2.5);
}
model {
  // likelihood including constants
  if (!prior_only) {
    target += bernoulli_logit_glm_lpmf(Y | Xc, Intercept, b);
  }
  // priors including constants
  target += lprior;
}
generated quantities {
  // actual population-level intercept
  real b_Intercept = Intercept - dot_product(means_X, b);
}
```

6 Resultados

Antes de cualquier cosa analizaremos la tabla de resumen del modelo:

Parametro	**mean**	**se_mean**	**sd**	**2.5 %**	**25 %**	**50 %**	**75 %**	**97.5 %**	**n_eff**	**Rhat**
b_Intercept	-2.81	0.00	0.37	-3.56	-3.06	-2.80	-2.56	-2.11	24693	1
b_age_c	-0.03	0.00	0.00	-0.03	-0.03	-0.03	-0.02	-0.02	58347	1
b_assess_c1	-1.07	0.00	0.10	-1.26	-1.13	-1.07	-1.00	-0.88	21333	1
b_assess_c2	-0.71	0.00	0.10	-0.90	-0.78	-0.71	-0.64	-0.52	21055	1
b_assess_c3	-0.46	0.00	0.10	-0.65	-0.52	-0.46	-0.39	-0.26	21854	1
b_assess_c4	-0.31	0.00	0.10	-0.50	-0.37	-0.31	-0.24	-0.12	21441	1
b_assess_c5	-0.48	0.00	0.12	-0.73	-0.57	-0.48	-0.40	-0.24	27003	1
b_famhx_c1	0.16	0.00	0.07	0.03	0.12	0.16	0.21	0.30	39451	1
b_famhx_c9	0.06	0.00	0.08	-0.08	0.01	0.06	0.11	0.21	38102	1
b_density_c2	1.20	0.00	0.29	0.67	1.00	1.18	1.38	1.80	21183	1
b_density_c3	0.64	0.00	0.31	0.06	0.43	0.63	0.84	1.28	21740	1
b_density_c4	-0.80	0.00	0.85	-2.69	-1.30	-0.71	-0.20	0.64	30663	1
b_bmi_c	0.00	0.00	0.01	-0.01	0.00	0.00	0.01	0.02	52375	1

Evaluaremos dos cosas:

1. Las cadenas: de esta manera sabremos si hubo una convergencia en las cadenas y no encontraremos dificultad al explorar la posterior.

2. Modelo: necesitamos evaluar cuan bueno es nuestro poder de predicción y si hay un buen ajuste de los parámetros al mismo.

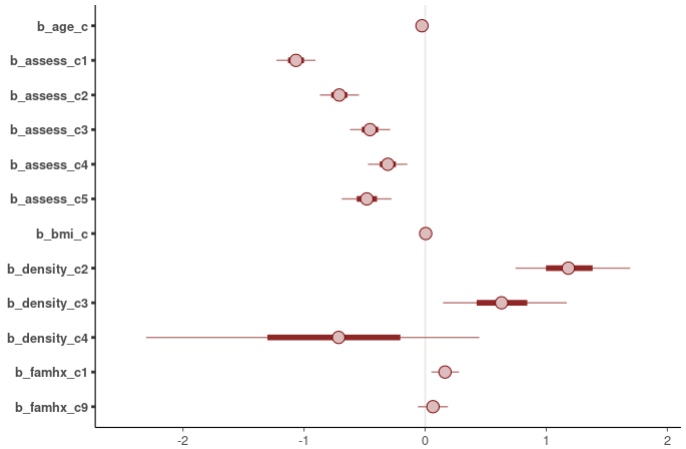
6.1. Metodos de Evaluación

- HDI
- $\hat{R}$
- N_{eff}/N
- Trace plot
- ACF
- Energy plot

6.2. HDI

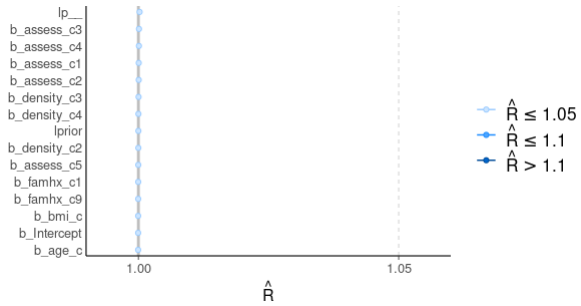
El primer gráfico nos presenta, que la mayoría de parámetros son poco significativos, esto era de esperarse ya que la proporciones de poblaciones son menores al 5 %, esto nos hace pensar que los cambios en los datos serán muy pocos por lo que....veremos mejor los HDI en grupo para sacar una mejor conclusión.

Agrupando por variable las categorías, tanto las evaluaciones como la densidad del seno en sus primeras categorías, es decir, en donde supone unos resultados de tranquiladad, es donde es más significativo, además mientras la categoría de aumentar las probabilidades de cáncer de mama



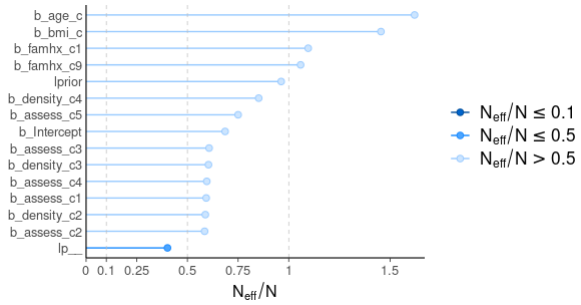
6.3. $\hat{R}$

El $\hat{R}$ nos ayudará a ver que tan buena fue la convergencia en nuestras iteraciones, lo estándar es que sean muy parecidas a 1; esto se cumple por lo cual podremos decir poco a poco que el modelo si tuvo una convergencia óptima, aunque eso no quiere decir que el modelo no puede mejorar.



6.4. N_{eff}/N

Aqui ya entramos en algo diferente, como se puede notar, esto es un cociente del tamaño de muestra efectivo versus el tamaño total de las iteraciones, para así poder notar si son efectivas; no hay algo teórico detrás de esto, pero en la practica es preocupante valor menores a 0,1, lo más común es encontrar valores mayores o iguales a 1.

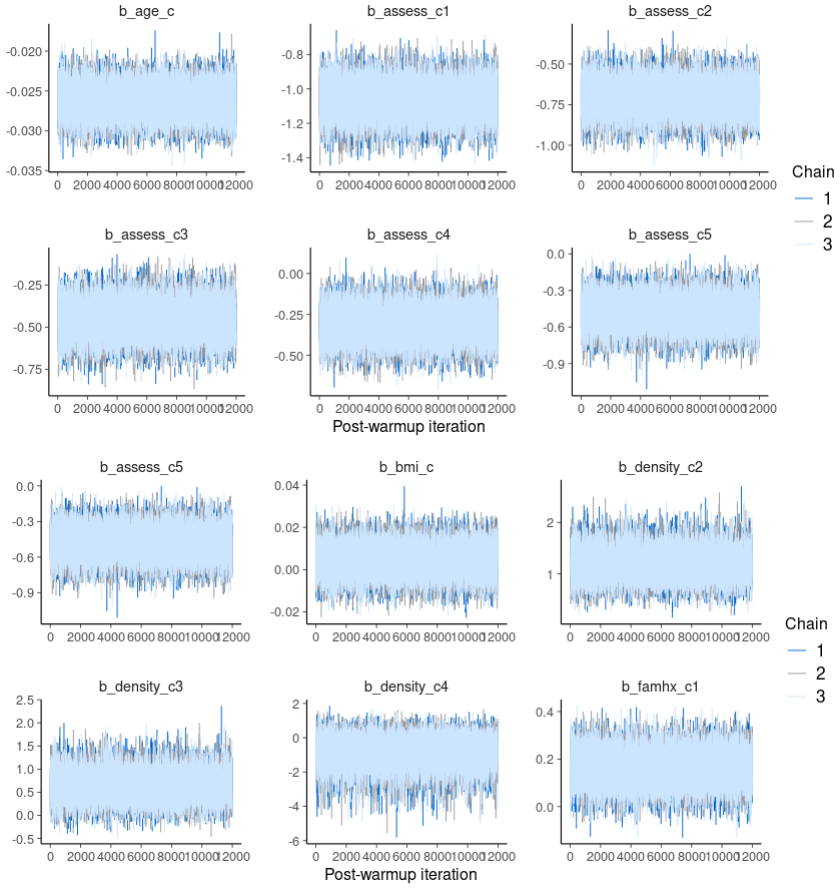


Así podemos concluir que las variables que sobrepasen el valor de 1 tienden a tener un mejor peso dentro del modelo

6.5. Trace plot

Aquí podemos ver las cadenas de markov en series temporales, por lo que el *trace plot* me ayudará a observar sí no solo hubo divergencias, sino que también los movimientos por medio de lineas o picos para obtener los valores mínimos y máximos de cada iteración y así observar si su convergencia se centra en la media estimada o al pasar por las iteracciones si se alejan del valor estimado.

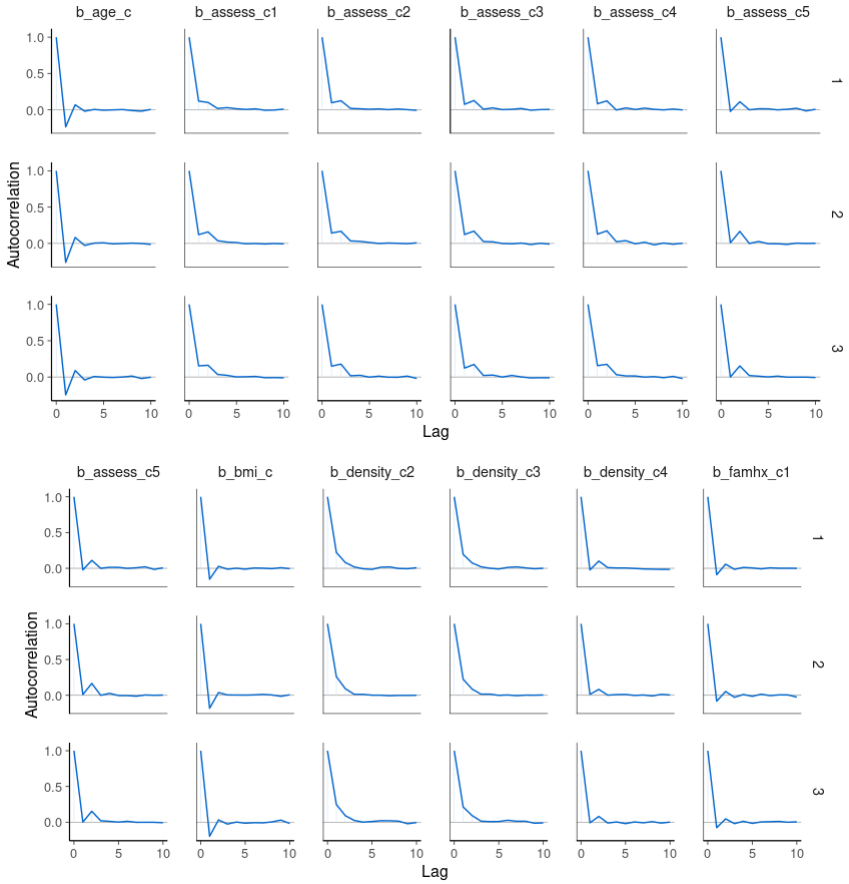
Los parámetros nos muestran que cumplen en ser estacionarios ya que no vemos divergencias y tampoco vemos valores extremos mantenidos, sino solo de vez en cuando, esto puede ser a que el modelo se demora en converger (gráfico anterior).



6.6. ACF

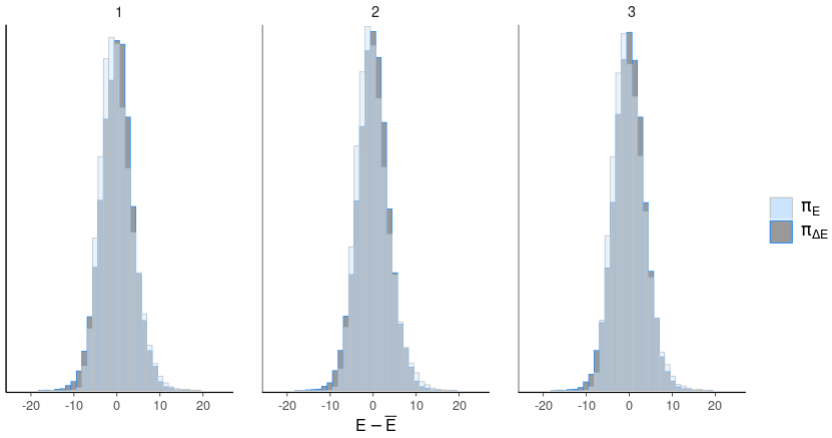
Nota: Tambien llamado gráfico de autocorrelación

Para poder haber llegado a esta fase, necesitábamos probar que los parámetros eran estacionarios.
Este gráfico nos permite ver cuales fueron las dificultades que tuvo el modelo vs la autocorrelación entre los parámetros, es decir, entre más suave este la gráfica, será de mayor soporte para una definición de un bueno modelo.



6.7. Energy plot

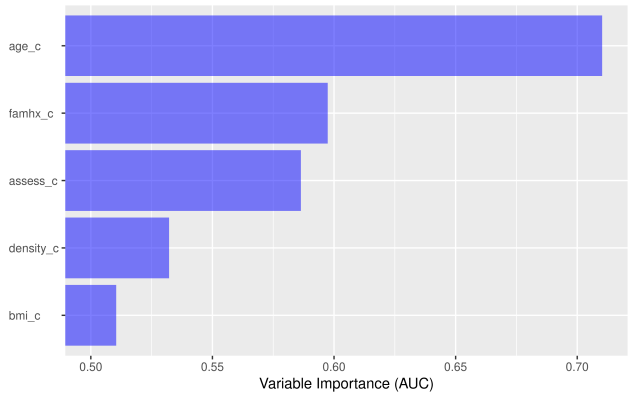
Las energías exploradas por una cadena de Markov hamiltoniana pueden utilizarse para visualizar tanto la densidad de transición de energía, $\pi(E|q)$ y la distribución de energía marginal,$\pi(E)$. Cuando estas distribuciones están bien ajustadas, la cadena de Markov hamiltoniana debería funcionar con solidez, pero si la densidad de transiciones de energía es significativamente más estrecha que la distribución de energía marginal, entonces la cadena puede no ser capaz de explorar completamente explorar las colas de la distribución objetivo.



Vemos como la distribución marginal deja que la transición de energía sea aceptable, ya que esta se ajusta bien a los datos, de esta manera vemos como podemos explorar las colas de nuestra distribución posterior.

6.8. Predicciones

Este gráfico nos muestra los niveles de importancia de las variables dentro del modelo, por medio de un valor AUC



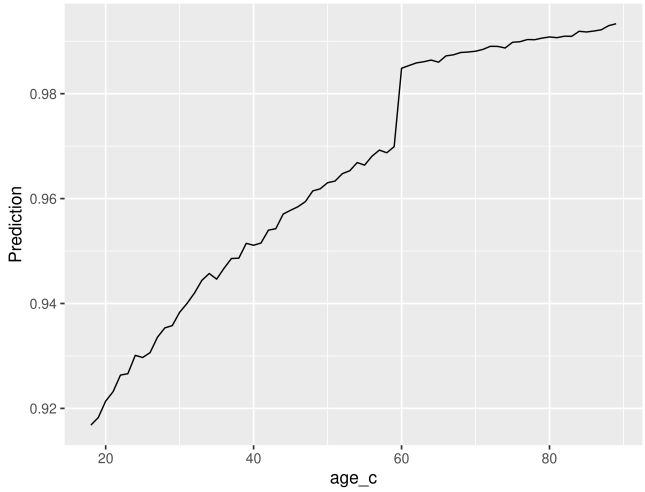
Gracias a este valor tenemos que hay rangos de variables significativas para el modelo, de esta manera sería bueno plantearse si la transformación de algunas variables afecta las predicciones, por el momento solo sabremos que las variables má

6.8.1. probabilidad de predicción

En esta sección mostraré algunos gráficos de predicción, donde veremos la según la probabilidad de éxito máxima, dejando estáticas las demás variables como son los niveles de predicciones por años.

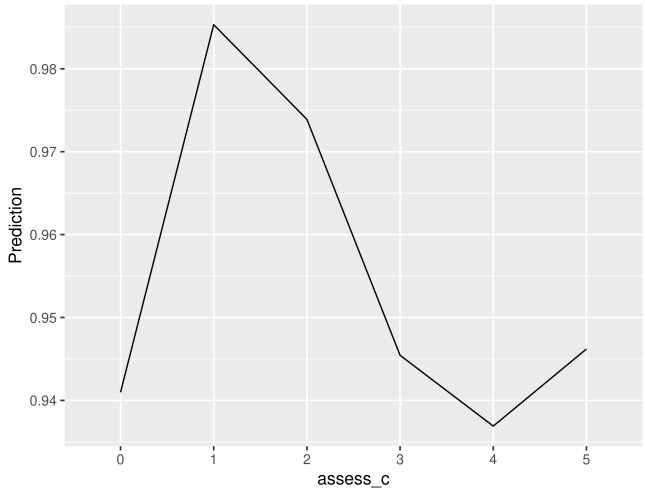
Aquí vemos como aumenta la escala de predicción, este método sirve para evaluar los rangos de las variables de respuestas y sus efectos en Y, es decir, que entre más aumente la edad, más chance o probabilidad tendrá el modelo entrenado en dar una respuesta sin un sesgo.

estos gráficos están hechos para saber sí todos los valores que recorré X tienen el mismo peso, gracias a la estadística bayesiana poseemos la capacidad de hacer esto.



En el caso de una variable categórica vemos como cada categória tiene un nivel diferente, es más probable dar una respuesta sin tanta incertidumbre en las evaluaciones de con $BI - RADS = 1, 2$ que tener una respuesta con muy poco nivel de aceptación en el $BI - RADS = 4$.

Los factores que determinan esta gráfica son en gran parte los pesos de los porcentajes en la base de datos, claro, esto se debe a que el modelo en cuestión tiene más formas de entrenarse y aprender del comportamiento de diferentes usuarios.



7 Conclusiones

Podemos decir de ultimas que el modelo se ajusta no como esperábamos pero si al orden de los datos, es decir, que aunque sea poco los parámetros estimados son proporcionales a los datos dados; además en todos los analisís los resultados eran estándar por lo que se puede decir que es un modelo que informaría solamente el 3 % de las veces con diagnostico positivo de cáncer.

Otra cosa a resaltar es que las cadenas convergieron de buena manera, a pesar que cada una empezó en un lugar distinto, pero casí siempre se pudo ver que llegaban al mismo punto por lo que la posterior aunque su tiempo de convergencia es más lento (no peor) escarba lo suficiente las colas de la distribución para podernos decir que no hay datos fuera de los que están al rededor de la media, con una varianza muy chica en la mayoria de los casos.

Una recomendación es probar el modelo con otras variables de la base de datos, pero es más recomendable buscar otros de predicción para la solución de este problema, ya que no es nada robusto el sistema del modelo bayesiano "logit"binario, por el momento diremos que los resultados fueron muy escuetos, se puede decir que falto tanto poder computacional, ya que las iteraciones requieren ram y tiempos de ejecución arbitraria (dependiendo del modelo) por eso decimos que de alguna manera si es significativo el IMC, tambien los antecedentes familiares son más influyentes que las evaluaciones $BI - RADS$ pero apesar de eso, el dato por excelencia es la edad de las mujeres mayor, ya que presentan una mayor probabilidad de dar positivo para cáncer de mama.

8 Bibliografía

[1] **Visual MCMC diagnostics using the bayesplot package** *https://cran.r-project.org/web/packages /bayesplot/vignettes/visual-mcmc-diagnostics.html* (Diagnostics MCMC).

[2] **Interpreting ACF or Auto-correlation plot** *https://medium.com/analytics-vidhya/interpreting-acf-or-auto-correlation-plot-d12e9051cd14* (Interpretation ACF).

[3] **Logistic regression with PyMC3Â¶** *https://goldinlocks.github.io/Bayesian-logistic-regression-with-pymc3/* (COnstrución de un modelo logístico computacional).

[4] **Bayesian Model Evaluation** *https://ccrgpages.rit.edu/ whelan/courses/* $2017_1 sp_S TAT_4 89/notes_m odels.pdf (Evaluación\ de\ un\ modelo\ Bayesiano) [pages\ 7\text{-}9].$

[5] ***Chapter 8 Model Diagnostics*** $https://bookdown.org/marklhc/notes_b ookdown/markov - chain - monte - carlo.html (Evaluación\ de\ un\ modelo\ Bayesiano) [Chapter\ 8].$