UNIVERSIDAD AUTÓNOMA CHAPINGO

DIVISIÓN DE CIENCIAS FORESTALES

INTRODUCCIÓN A LA TEORIA BAYESIANA

Regresión logística con enfoque bayesiano para los datos de ocupación de la ENOE

Alumna:
Vásquez González Arely Yazmin

Instructor:
Dr. Rodríguez Yam Gabriel Arcángel

Junio 2022

INTRODUCCIÓN

La estadística bayesiana ha surgido como una alternativa a la estadística clásica o frecuentista, para resolver problemas estadísticos como la estimación de los parámetros, pruebas de hipótesis o predicción. Este enfoque está basado en el teorema de Bayes, y su principal diferencia del enfoque frecuentista es que incorpora información externa al estudio que se esté realizando, de forma que si se conoce la probabilidad de que ocurra un suceso, este valor será modificado cuando disponga de esa información. Bajo este enfoque también es posible obtener modelos de regresión, por ejemplo, la regresión logística, este tipo de regresión nos permite modelar relaciones entre una variable dependiente dicotómica y variables independientes. Cabe resaltar que generalmente se utilizan métodos numéricos para calcular por decir, la posterior marginal o las densidades predictivas para cada uno de los parámetros del modelo, sin embargo, existen diversos programas que nos ayudan a hacer estos cálculos, en el trabajo desarrollado se utilizó el software OpenBugs.
Ahora el conjunto de datos que se usó para ajustar el modelo, fueron obtenidos de la Encuesta Nacional de Ocupación y Empleo (ENOE), el cual es la principal fuente de información sobre el mercado laboral mexicano la subocupación y la desocupación. Los datos son acerca de si una persona esta ocupada o no, de acuerdo con otras variables tales como el sexo, la edad, la escolaridad y si pertenece a una área rural o urbana.

OBJETIVOS GENERALES

Aplicar los temas revisados en el curso de Introducción a la Estadística Bayesiana

OBJETIVOS ESPECIFICOS

Realizar una regresión logística con enfoque bayesiano, para ajustar los datos de ocupación.

MARCO TEÓRICO

La teoría bayesiana se fundamenta desde un punto de vista subjetivo de la probabilidad, en donde los parámetros son considerados como variables aleatorias, las inferencias se basan en la distribución del parámetro de interés condicionado a x, denominada distribución a posteriori, al utilizar el teorema de Bayes nos permite actualizar la información sobre los parámetros de la distribución. En este enfoque, es esencial tener la distribución a priori y también la función de verosimilitud de los datos dado el valor del parámetro.
Un modelo estadistico bayesiano consiste basicamente de un modelo estadístico parámetrico $f(x|\theta)$ y de una distribución a priori $\pi(\theta)$ de paramétro θ. Donde la distribción conjunta de x y θ esta dada por

$$f(x, \theta) = \ell(\theta|x)\pi(\theta)$$

Ya que $f(\theta|x) = f(x,\theta)/f(x)$, se obtiene que la distribución a posteriori de θ esta dada por

$$\pi(\theta|x) \propto \ell(\theta|x)\pi(\theta)$$

Por otra parte los modelos de regresión logística son usados para modelar relaciones entre una variable respuesta dicotómica en función de una o varias variables explicativas.
La función de distribución logistica esta dada por

$$\pi_i = \frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)}$$

Datos

Los datos que se utilizaron tiene como variable respuesta 1 si una persona esta ocupada y 0 si la persona no esta ocupada, y como variables explicativas el sexo tomando el valor de 1 si es hombre y 0 si es mujer; la edad del entrevistado; años de escolaridad del entrevistado y por último se tiene la variable rururb que toma el valor de 1 si es de una área urbana o 0 si es de una área rural. En total se cuenta con 27639 datos.

DESARROLLO

En regresión logistica, la variable respuesta y_i sigue una distribución Bernoulli, tomando el valor de 1 con probabilidad de π_i y 0 con probabilidad $1-\pi_i$. El modelo de regresión logística para datos binarios esta dada por:

$$y_i|\pi_i \sim Ber(\pi_i)$$
$$\pi_i = \frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)} \quad i = 1,..,n$$

donde:
$x_i^t = (x_{i1}, x_{i2}, ..., x_{ik})$ es un vector con los valores de k variables explicativas y $\beta = (\beta_1, ..., \beta_k)$ es un vector k coeficientes de regresión.
La función de verosimilitus está dada por:

$$L(\beta|y,X) = \prod_{i=1}^{n} \left(\frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)}\right)^{y_i} \left(\frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)}\right)^{1-y_i}$$

Entonces la distribución a posteriori de los parámetros β es la siguiente

$$\pi(\beta|y,|X) \propto \pi(\beta)\prod_{i=1}^{n} \left(\frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)}\right)^{y_i} \left(\frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)}\right)^{1-y_i}$$

donde $\pi(\beta)$ es la distribución a priori.

En el caso de no tener información apriori sobre los parámetros se puede usar una distribución a priori no informativa.

Para obtener las estimaciones de los parámetros, se implementó el muestreador Gibss, que sigue los siguientes pasos:

1. Se escogen valores iniciales de $\beta^{(0)} = (\beta_1^{(0)}, ..., \beta_q^{(0)})$
2. Se obtiene un nuevo valor de $\beta^{(j)}$ de $\beta^{(j-1)}$ por sucesivos valores generados
   Generando $\beta_1^{(j)}$ de $\pi(\beta_1|\beta_2^{(j-1)}, .., \beta_q^{(j-1)})$
   $\vdots$
   Generando $\beta_q^{(j)}$ de $\pi(\beta_q|\beta_1^{(j-1)}, .., \beta_q^{(j-1)})$
3. Se aumenta el contador de j a $j+1$ y luego se retorna al paso 2.

Este proceso define una cadena de Markov homogénea debido a que cada valor simulado depende solo del valor simulado anterior y no de otros.

Podemos observar que la distribución a posteriori no tiene una forma conocida por lo que es necesario usar métodos MCMC, en nuestro caso para llevar a cabo esta regresión se uso el programa OpenBugs el cual emplea el muestreador Gibbs; este software incluye un sistema que determina un esquema MCMC apropiado basado en el muestreador de Gibbs para analizar el modelo especificado.

Primero se tiene que escribir el modelo deseado en un "txt" para luego cargarlo al programa, asimismo es importante tener en cuenta que los datos deben estar en forma de lista en un documento aparte, si se quiere especificar los valores iniciales es necesario tener otro "txt" especificando dichos valores.
Para los datos considerados el modelo que se especificó fue el siguiente:

```
model
{
for ( i in 1: N ) {
ocupado[i] ~ dbern(p[i])
logit(p[i]) <-  b0+ b1* sexo [i] + b2*edad [i]+b3*aesc[i]+b4* rururb[i]
}
b0 ~ dnorm (0.0 , 0.0001)
b1 ~ dnorm (0.0 , 0.0001)
b2 ~ dnorm (0.0 , 0.0001)
b3 ~ dnorm (0.0 , 0.0001)
b4 ~ dnorm (0.0, 0.0001)
}
```

Los valores iniciales para todos los parámetros se especifico como 0.

Se usaron distribuciones a priori no informativas, tratando a todos los valores como igualmente plausibles. En OpenBugs se puede aproximar mediante:
$b_i \sim dnorm(0.0, 0.0001)$
En OpenBugs el segundo parámetro de la normal es la precisión $(1/\sigma^2)$, que esto

equivaldría a poner una varianza de 10000 en un parámetrización tipica de la normal.

RESULTADOS

Los resultados de las estimaciones obtenidas se resumen en la siguiente tabla:

	mean	sd	MC-error	val2.5pc	median	val97.5pc	start	sample
b0	-2.855	0.1386	0.01133	-2.897	-2.866	-2.833	1	2000
b1.sexo	1.098	0.02498	0.001518	1.08	1.099	1.117	1	2000
b2.edad	0.02415	7.364E-4	4.488E-5	0.02374	0.02419	0.02463	1	2000
b3.años_ esc	0.1837	0.00652	4.064E-4	0.182	0.1841	0.1861	1	2000
b4.rururb	-0.297	0.05742	0.005349	-0.3189	-0.2921	-0.2695	1	2000

La gráfica de la distribución a posterior para cada parámetro se presenta en la figura 1.

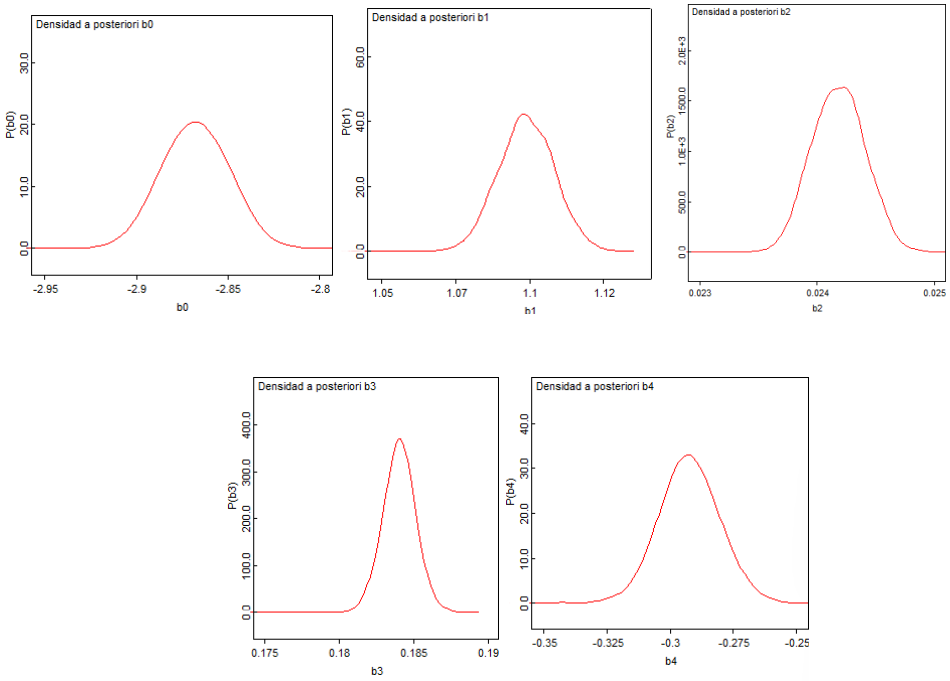


Figure 1: *Distribución a posteriori para* β_0, β_1, β_2,β_3 *y* β_4 *respectivamente.*

El coeficiente que corresponde a la variable sexo nos indica que para un hombre en promedio será por arriba de 2.998 (exp(1.098)) más propenso a estar ocupado es decir a tener empleo que para una mujer, en tanto se matienen fijos las demás variables.
Para el coeficiente de la varible edad nos indica que para una persona que su edad aumente en una unidad será en promedio por arriba de 1.024444 (exp(0.02415)) más propenso a tener empleo que para aquellos que no incremente su edad, ma-

tieniendo fijos los demás variables.

Para el coeficiente de la varible años de escolaridad nos indica que para una persona que aumenta en una unidad su escolaridad será en promedio por arriba de 1.201655 (exp(0.1837)) más propenso a tener empleo que para aquellos que no incremente su escolariadad, matieniendo fijos los demás variables.

Por otro lado, el logaritmo de odds de que una persona tenga empleo esta negativamente relacionado con la puntuación obtenida en el examen de lectura (coeficiente parcial = -0.02617)

Un análisis para el diagnóstico de convergencia es usando la traza, en la figura 2 se pueden ver los gráficos correspondiente, podemos notar que la serie de tiempo presenta estacionalidad indicando que la convergencia en todas las variables se alcanzó.

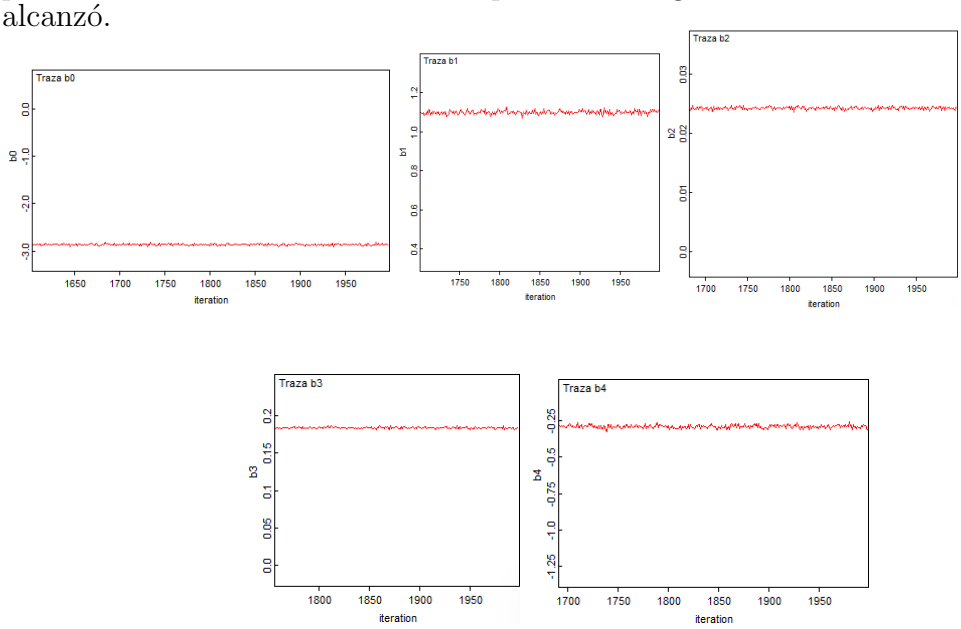
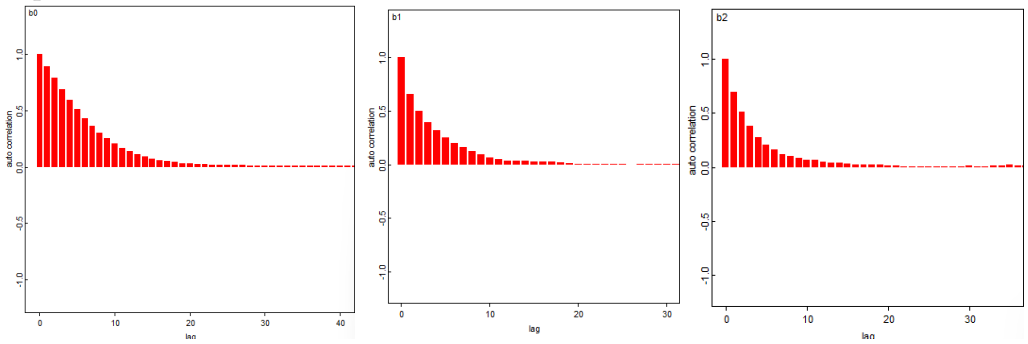


Figure 2: *Gráfica de la traza para* β_0, β_1, β_2,β_3 *y* β_4 *respectivamente.*

Otra forma de diagnosticar la convergencia es observando el gráfico de autocorrelaciones que se muestra en la figura 3, se nota que existe convergencia en todas las variables, y es más se puede observar que la convergencia es relativamente rápida.



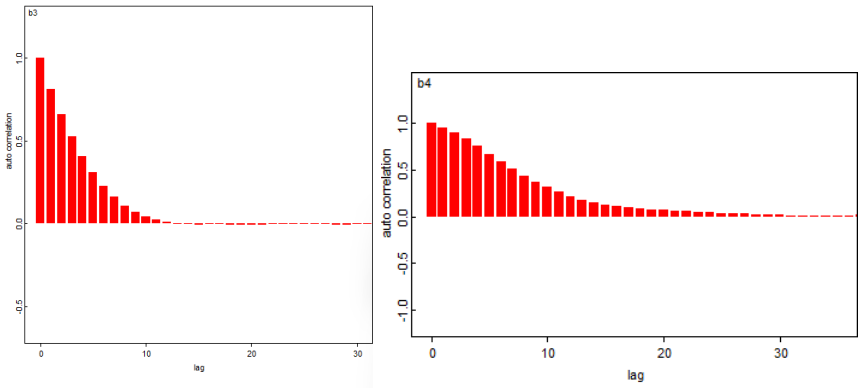


Figure 3: *Gráfica de autocorrelaciones para* β_0, β_1, β_2,β_3 *y* β_4 *respectivamente.*

Otro de los graficos que nos presenta OpenBugs es el de los cuartiles que se presentan en la figura 4.

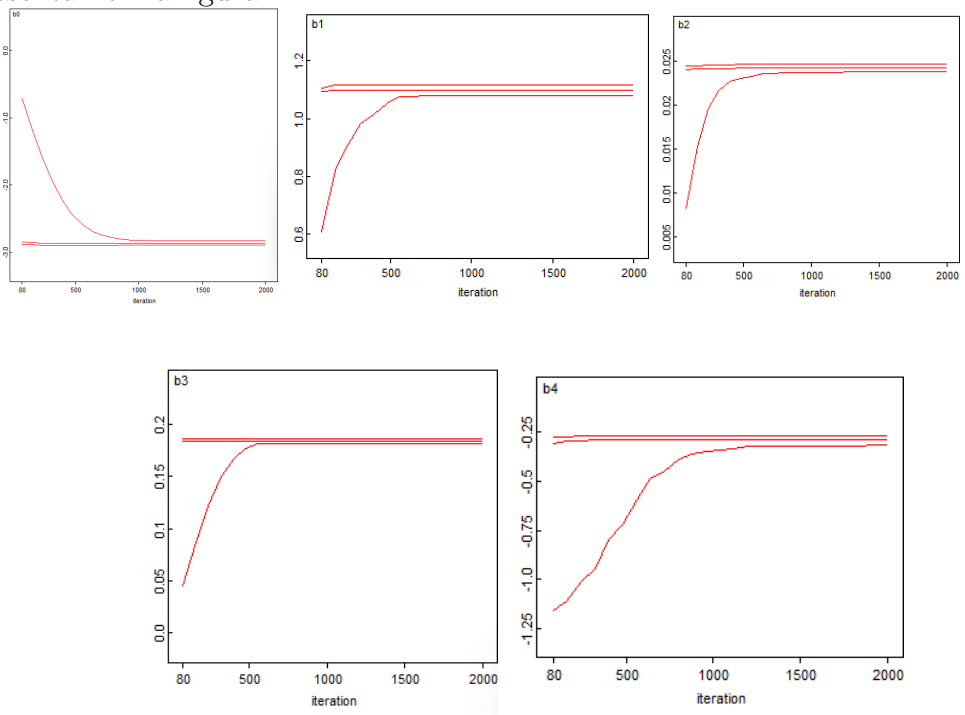


Figure 4: *Gráfica de los cuartiles para* β_0, β_1, β_2,β_3 *y* β_4 *respectivamente.*

CONCLUSIÓN

Se logró obtener una regresión logistica usando el enfoque bayesiano, los resultados obtenidos son muy similares usando una regresión logistica clásica. Podemos decir que se cumplieron los objetivos propuestos.

BIBLIOGRAFIA

CONGDON, P. (s.f.). Bayesian Statistical Modelling.

ENOE. (s.f.). Obtenido de https://www.inegi.org.mx/programas/enoe/15ymas/#Publicaciones

González, L. D. (2015). ENFOQUE BAYESIANO DEL MODELO DE REGRESION LOGISTICA USANDO CADENAS DE MARKOV MONTE CARLO. REVISTA INVESTIGACION OPERACIONAL, 36(02).

Juan Carlos Correa Morales, C. J. (2018). INTRODUCCIÓN A LA ESTADÍSTICA BAYESIANA. Medellín, Colombia: Fondo Editorial ITM.