El algoritmo de Metropolis-Hastings: una aplicación de inferencia bayesiana para el sistema previsional argentino

Melina Guardiola [1][0000-0002-6212-3298], Fernanda Villarreal[1][0000-0001-7731-5981] and Milva Geri[2] [0000-0003-3265-3308]

[1] Instituto de Matemática de Bahía Blanca (INMABB), Departamento de Matemática, Universidad Nacional del Sur (UNS)- CONICET, Bahía Blanca, Argentina
[2] Instituto de Investigaciones en Ciencias de la Salud, Departamento de Ciencias de la Salud (UNS-CONICET), Departamento de Matemática, Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina.
guardiol@uns.edu.ar, fvillarreal@uns.edu.ar, mgeri@iiess-conicet.gob.ar

Resumen. El algoritmo de Metropolis-Hastings es un método de Monte Carlo basado en Cadenas de Markov (MCMC) que permite obtener muestras de distribuciones complejas, facilitando la inferencia bayesiana. Es una herramienta fundamental cuando las posteriores condicionales no tienen forma analítica. En este trabajo se presenta una aplicación del algoritmo de Metropolis-Hastings para especificar las distribuciones a posteriori de los parámetros del modelo de regresión logística bayesiana que modela los determinantes de la densidad contributiva del sistema previsional argentino. La implementación del método se realiza utilizando el software estadístico R y la fuente de datos que se utiliza proviene de la Muestra Longitudinal de Empleo Registrado (MLER) del Sistema Integrado Previsional Argentino (SIPA). Entre los resultados se destacan que todas las cadenas convergen y los coeficientes significativos tienen los signos esperados.

Palabras clave: Algoritmo de Metropolis-Hastings, Regresión logística bayesiana, Sistema previsional argentino.

The Metropolis-Hastings Algorithm: A Bayesian Inference Application for the Argentine Pension System

Abstract. The Metropolis-Hastings algorithm is a Markov Chain Monte Carlo (MCMC) method that allows sampling from complex distributions, facilitating Bayesian inference. It is an essential tool when the conditional posteriors lack an analytical form. This work presents an application of the Metropolis-Hastings algorithm to specify the posterior distributions of the parameters in a Bayesian logistic regression model, which captures the determinants of contributory density in the Argentine pension system. The method is implemented using the R

statistical software, and the data source used is the Longitudinal Sample of Registered Employment (MLER) from the Integrated Argentine Pension System (SIPA). Among the results, it is noteworthy that all chains converge, and the significant coefficients exhibit the expected signs.

Keywords: Metropolis-Hastings algorithm - Bayesian logistic regression - Argentine pension system

1 Introducción

El análisis de regresión logística es un método de gran utilidad para modelar relaciones entre una variable dependiente dicotómica y variables independientes que pueden ser de cualquier naturaleza. Para la estimación de los parámetros del modelo de regresión logística (MRL) el enfoque bayesiano resulta muy atractivo y en los últimos años ha crecido su uso permitiendo una fácil interpretación de los parámetros del MLR y la obtención de mejores resultados al trabajar con muestras pequeñas (Johnson et al., 2020). Sin embargo, los métodos para obtener las distribuciones a posterior de estos parámetros requieren de cálculos muy complicados. Por tal motivo, resulta necesario recurrir a métodos que permitan obtener de forma aproximada la distribución posterior de estos parámetros. Estas aproximaciones pueden obtenerse por Métodos de Laplace (Tirney y Kadane, 1986) o métodos basados en Cadenas de Markov Monte Carlo (MCMC) para espacios paramétricos de grandes dimensiones. El algoritmo de Metropolis-Hastings es un método MCMC que permite obtener muestras de distribuciones complejas, facilitando la inferencia bayesiana. En este trabajo se presenta una aplicación del algoritmo de Metropolis-Hastings para especificar las distribuciones a posteriori de los parámetros del modelo de regresión logística bayesiana que modela los determinantes de la densidad contributiva del sistema previsional argentino. La implementación del método se realiza utilizando el software estadístico R y la fuente de datos que se utiliza proviene de la Muestra Longitudinal de Empleo Registrado (MLER) del Sistema Integrado Previsional Argentino (SIPA).

2 Algoritmo de Metropolis-Hastings

El algoritmo de Metropolis-Hastings es una técnica versátil que se usa en muchos campos más allá de la inferencia bayesiana, incluyendo la optimización estocástica, la simulación de sistemas físicos, la reconstrucción de imágenes, la bioinformática, los modelos económicos y la teoría de redes. Su capacidad para generar muestras de distribuciones complejas y explorar eficientemente espacios de parámetros hace que sea útil en cualquier área donde se necesiten muestreo o simulación en presencia de incertidumbre o distribuciones difíciles de manejar de manera directa. A continuación, se presentan los pasos del algoritmo:

1. Establecer un valor inicial θ^0 para los parámetros del modelo.
2. Para s = 1, 2, ….., S (S es el número total de muestras que se desea generar):

2.1 generar un valor θ^* a partir de una distribución propuesta $g_s(\theta^*|\theta^{s-1})$

2.2 calcular la probabilidad de aceptación

$$\alpha(\theta^*, \theta^{s-1}) = min\left\{\frac{p(\theta^*|y)g_s(\theta^{s-1}|\theta^*)}{p(\theta^{s-1}|y)g_s(\theta^*|\theta^{s-1})}, 1\right\}$$

2.3 $\theta^s = \begin{cases} \theta^* \text{ con probabilidad } \alpha\ (\theta^*, \theta^{s-1}) \\ \theta^{s-1} \text{ caso contrario} \end{cases}$

se genera un valor u de una distribución uniforme $U(0,1)$ y si $u \leq \alpha\ (\theta^*, \theta^{s-1})$ se acepta θ^* si no se mantiene el valor anterior.

3. Se repiten estos pasos hasta que se obtienen las S muestras deseadas.

Este algoritmo es un pilar fundamental porque proporciona un método para generar muestras de distribuciones posteriores que son difíciles de obtener directamente, facilitando la inferencia y la toma de decisiones estadísticas en modelos complejos.

3 Aplicación y resultados

Se presenta la aplicación del algoritmo de Metropolis Hastings para especificar las distribuciones a posterior de los parámetros del modelo de regresión logística bayesiano para el sistema previsional argentino. Los datos surgen de la Muestra Longitudinal de Empleo Registrado (MLER) publicada por el Ministerio de Trabajo Empleo y Seguridad Social (MTEySS). La base es representativa de la población de empleados en relación de dependencia del sector privado y contiene información de historias laborales de unos 600 mil trabajadores entre enero de 1996 y diciembre de 2021, (última actualización:2021). Se construyó la matriz de datos cuya unidad de observación fue el trabajador/a que cumple la edad mínima jubilatoria en 2022. Para cada trabajador se computa el número de meses de aporte en la ventana de observación: esta variable asume valores que van desde 1 mes de aporte a 312 meses de aporte. Siguiendo a Rofman & Oliveri (2012), se considera que aquellos trabajadores con al menos un 80% del total de meses de la ventana (250 meses) tendrán una densidad contributiva suficiente para cumplir con las condiciones de elegibilidad para acceder a una jubilación. De tal manera, nuestra variable dependiente Y_i asume valor 1 si el individuo i tiene al menos 250 meses de aporte y 0 en caso contrario. Como variables independientes se consideraron: la edad de ingreso al mercado laboral, el género, el salario mediano en dólares percibido durante la ventana de observación, el sector de actividad de las empresas donde trabajó (categoría múltiple), el tamaño de las empresas en las que trabajó (categoría múltiple). Para el ajuste del modelo se utiliza el software R comparando el algoritmo de Metropolis Hasting generado por la función “MCMClogit” (Martin et al., 2011) con el algoritmo diseñado paso a paso. Se realizaron 25000 iteraciones siendo descartadas las primeras 5000. Entre los resultados preliminares se encuentra que los varones presentan mayor probabilidad que las mujeres de completar aportes; cuanto más se demora la persona en aportar por primera vez, menor será su probabilidad de completar aportes y cuanto mayor haya sido el ingreso mediano en dólares percibido por la persona en sus distintas relaciones laborales, mayor es la probabilidad de completar aportes. A su vez,

las personas que trabajan en cualquier sector de actividad que no sea el industrial, presentan menor probabilidad de completar aportes (con excepción de los sectores antes mencionados que no difieren significativamente del industrial). Las personas que trabajaron durante todo el período en empresas chicas o medianas, presentan menor probabilidad de completar aportes que aquellas personas que trabajaron siempre en empresas grandes de más de 200 empleados. La Figura 1 muestra los gráficos de las trazas para cada parámetro del modelo bajo el supuesto de prior impropia uniforme.

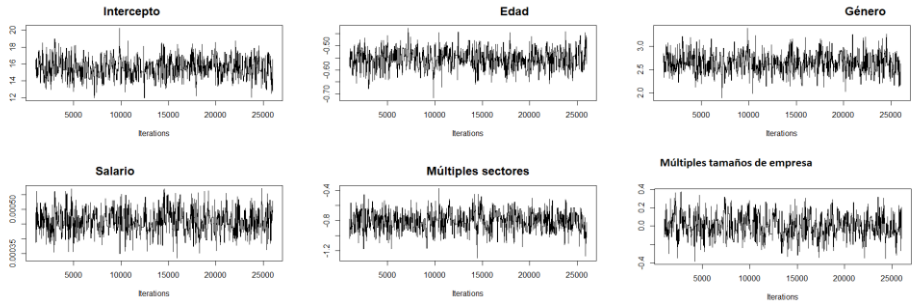

Fig. 1. Gráficos de trazas para cada parámetro del modelo

Los gráficos indican que la media y varianza son constantes a lo largo de las iteraciones, lo cual es signo de que las series son estacionarias y se logra la convergencia (Plummer et al. 2006). A este resultado de convergencia se le suma la prueba de igualdad de medias de Geweke (1992) que considera como muestras la primera y última parte de la cadena de Markov (primer 10% y último 50%). Si las dos muestras se han obtenido de una distribución estacionaria, las medias son iguales y el estadístico de Geweke tiene una distribución normal estándar asintótica. Por lo tanto, valores de Z que caen dentro de los extremos de la distribución normal estándar indicarían que la cadena converge. Esta prueba refuerza los resultados hallados al observar las gráficas de trazas.

4 Conclusiones

En este caso de aplicación, todas las cadenas convergen y los coeficientes significativos tienen los signos esperados. Haber podido mostrar el desarrollo del algoritmo comparándolo con el paquete propuesto en R nos permite comprender mejor su funcionamiento. Cuando el tamaño de muestra es grande, en este caso mayor a 6000 observaciones, los resultados del modelo de regresión logística bajo un enfoque bayesiano, independientemente de la distribución a priori propuesta, arrojan resultados muy similares a los estimadores por máxima verosimilitud obtenidos con un enfoque frecuentista tradicional. Desde una perspectiva más amplia, estos resultados poseen implicancias relevantes para la integración del enfoque bayesiano en esquemas de optimización estocástica y simulación. La posibilidad de incorporar la incertidumbre paramétrica en modelos complejos mediante simulaciones a partir de la distribución posterior convierte al marco bayesiano en una herramienta particularmente adecuada para alimentar

algoritmos de simulación Monte Carlo o técnicas de optimización bayesiana. En este sentido, el presente trabajo constituye un aporte metodológico que puede ser extendido a contextos de toma de decisiones en ambientes inciertos, donde la actualización secuencial del conocimiento y la propagación de la incertidumbre son elementos centrales.

Referencias

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculation of posterior moments, Bayesian Statistics, vol.4, Eds. Bernardo J.M., Berger J.O., David A.P and Smith A.F, Clarendon Press, Oxford.

Johnson, A., Ott, M.Q., & Dogocu, M. (2022). Bayes Rules! An Introduction to Applied Bayesian Modeling. Chapman & Hall /CRC Texts in Statistical Science.

Martin, A.D., Quinn, K.M., & Park, J.H. (2011), MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 42, 91–21. https://doi.org/10.18637/jss.v042.i09

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006). Output Analysis and Diagnostics for MCMC (CODA). R NEWS. 6(1), 7-11. https://cran.r-project.org/doc/Rnews/Rnews2006-1.pdf

Rofman, R. & Olivieri, M.L. (2012). Un repaso sobre las políticas de protección social y la distribución del ingreso en Argentina. Económica, 58, 97–128.

Tierney, L. and Kadane, J.B. (1986) Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association, 81, 82-86.