Vol. 13 (núm. 1) pp. 79-92
Enero-Junio de 2022

Ciencia en Desarrollo
Revista de la Facultad de Ciencias
Vicerrectoría en Investigación y Extensión - VIE
Centro de Investigaciones y Extensión de la Facultad de Ciencias - CIEC

https://doi.org/10.19053/01217488.v13.n1.2022.13655

Comparación de algunos métodos para estimar el modelo de riesgos proporcionales de Cox para datos con censura a intervalo

Comparison of Some Methods to Estimate the Cox Proportional Hazards Model for Interval-Censored Data

Olga A. Bustos-Giraldo[1], Mario C. Jaramillo-Elorza[2] y [3]Carlos M. Lopera-Gómez

Resumen

Los datos con censura a intervalo son comunes en varias áreas del conocimiento, tales como: epidemiología, finanzas, demografía, medicina, entre otras. Ocurren cuando el evento de interés, el tiempo de falla, no se observa exactamente, sino que se encuentra dentro de algún intervalo del tiempo de observación. Con frecuencia en esta situación se realiza una imputación de los datos que no se conocen exactamente. Algunos de los métodos de imputación múltiple propuestos en la literatura son el algoritmo PMDA (Poor Man's Data Augmentation) y el algoritmo ANDA (Asymptotic Normal Data Augmentation), los cuales permiten estimar los parámetros del modelo de riesgos proporcionales de Cox utilizando métodos clásicos de estimación. También existen métodos alternativos para realizar estas estimaciones, como el algoritmo ICM (Iterative Convex Minorant) y un enfoque Bayesiano, que no realizan imputación de los datos con censura a intervalo.

En este trabajo se realizó una comparación vía simulación del desempeño de los estimadores de los parámetros del modelo de Cox producidos por cada uno de los métodos anteriormente mencionados. Los resultados evidenciaron que en términos generales los métodos ICM y el enfoque Bayesiano presentan valores de probabilidad de cobertura más altos y errores cuadráticos medios más bajos, además al aumentar el tamaño de la muestra estos valores mejoran notablemente comparados con los métodos PMDA y ANDA. En estos últimos no se evidenciaron diferencias considerables entre los resultados. Finalmente, se realizó una aplicación con datos reales asociados a un estudio de mastitis en ganado lechero.

***Palabras clave*:** Métodos de imputación múltiple, Censura a intervalo, Enfoque bayesiano, Algoritmo ICM (Iterative Convex Minorant), Modelo de riesgos proporcionales de Cox.

Abstract

Interval censored data is common in several areas of knowledge, such as: epidemiology, finance, demography, medicine, among others. They occur when the event of interest, the failure time, is not observed exactly, but is within some interval of the observation time. Often in this situation an imputation is made of the data that is not exactly known. Some methods of multiple imputation proposed in the literature are the PMDA (Poor Man's Data Augmentation) algorithm and the ANDA (Asymptotic Normal Data Augmentation) algorithm, which allow estimating the parameters of the Cox proportional hazards model using classical estimation methods. There are also alternative methods to make these estimations such as the ICM (Iterative Convex Minorant) algorithm and a Bayesian approach, which do not impute the data with interval censoring.

In this work, a comparison was made via simulation of the performance of the estimators of the Cox model parameters produced by each of the aforementioned methods. The results showed that in general terms the ICM methods and the Bayesian approach present higher coverage probability values and lower mean square errors, in addition when increasing the sample size these values significantly improve compared to the PMDA and ANDA multiple imputation methods. In the latter, there were no significant differences between the results. Finally, an application was made with real data associated with a study of mastitis in milk cattle.

***Keywords*:** Multiple imputation methods, Interval-censored, Bayesian approach, ICM (Iterative Convex Minorant) algorithm, Cox proportional hazards model.

Recepción: 12-Nov-2021
Aceptación: 15-Dic-2021

[1]Magíster en Ciencias-Estadística, Universidad Nacional de Colombia, Medellín, Colombia.
Correo electrónico: oabustos@unal.edu.co
[2]Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia.
Correo electrónico: mcjarami@unal.edu.co
[3]Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia.
Correo electrónico: cmlopera@unal.edu.co

1 Introducción

Los datos con censura a intervalo ocurren comúnmente en muchos campos, tales como: demografía, epidemiología y estudios médicos. En tales estudios, los participantes se someten periódicamente a observaciones o exámenes y los tiempos de falla no son observados exactamente, pero se conoce que se encuentran dentro de algún intervalo [1]. Esto último es un problema ya que convencionalmente se ha usado el límite inferior, el punto medio o el límite superior del intervalo de inspección como tiempo de falla, esto es conocido en la literatura como imputación simple y ha sido bastante utilizado por su simpleza frente a otros métodos [2]. Sin embargo, estos métodos presentan problemas de sesgo de los estimadores de la función de supervivencia, especialmente cuando los intervalos son de gran tamaño, o son de diferentes longitudes [3, 4]. El uso de los métodos de imputación múltiple incorporando variables auxiliares, es decir, información auxiliar acerca del tiempo de falla para las observaciones con censura a intervalo, pueden mejorar la eficiencia de los estimadores y reducir el efecto de visitas perdidas en comparación con enfoques más simples [3].

Algunos métodos de imputación múltiple para estimar el modelo de riesgos proporcionales de Cox para datos con censura a intervalo fueron estudiados inicialmente por [5], donde proponen los algoritmos denominados PMDA (Poor Man's Data Augmentation) y ANDA (Asymptotic Normal Data Augmentation) para datos con censura a intervalo, basados en imputación múltiple para el modelo de regresión de Cox, y demostró que los resultados de las estimaciones de los coeficientes de regresión y el error estándar asociado suministra una alternativa prometedora para el estimador de máxima verosimilitud no paramétrico (NPMLE).

Un método alternativo de estimación de los parámetros del modelo de riesgos proporcionales de Cox para datos con censura a intervalo, que no involucra imputación de los tiempos de falla no observados, fue propuesto por [6] y consiste en un algoritmo ICM (Iterative Convex Minorant). En [7] se reformuló el algoritmo ICM como un método de proyección gradiente generalizado el cual conduce a una extensión natural al modelo de Cox. Los enfoques Bayesianos para el análisis de datos con censura a intervalo nos presentan otra opción para abordar el problema de estimación. En [1] proponen un novedoso y eficiente enfoque Bayesiano para el análisis de los datos con censura general bajo el modelo de riesgos proporcionales de Cox.

En este trabajo, se comparará vía simulación el desempeño de los estimadores de los parámetros del modelo de riesgos proporcionales de Cox, producidos por el algoritmo ICM, por un enfoque Bayesiano y luego de aplicar los métodos de imputación múltiple PMDA y ANDA, para datos con censura a intervalo. La comparación se realizará utilizando el error cuadrático medio (ECM) de las respectivas estimaciones de los parámetros de interés y la probabilidad de cobertura empírica obtenida bajo un nivel de confianza nominal del 95%.

En la Sección 2 se describen brevemente los datos de vida censurados a intervalos, el modelo de riesgos proporcionales de Cox y los métodos de estimación basados en el algoritmo ICM (Iterative Convex Minorant) y en un enfoque Bayesiano. La Sección 3 describe los métodos de imputación múltiple PMDA y ANDA. Un estudio de simulación y el análisis de sus resultados es presentado en la Sección 4. La Sección 5 presenta una aplicación en datos reales asociados a un estudio de mastitis en ganado lechero. Finalmente, en la Sección 6 se presentan las conclusiones del trabajo.

2 Conceptos básicos

2.1 Datos de vida censurados a intervalo

Una de las características de los datos de tiempos de vida es la censura. Para las observaciones con censura a intervalo, sólo se conoce un intervalo, dentro del cual la falla ha ocurrido. Los tiempos de falla exactos y la censura a la derecha pueden ser registrados como un caso especial de los tiempos de falla con censura a intervalo, en tales casos, el intervalo se reduce a un simple punto en los tiempos exactos y para la censura a la derecha el intervalo consiste del límite inferior y el límite superior se toma como infinito, es decir, un número muy grande, ya que el evento de interés no se presenta.

Suponga que los tiempos de vida para n individuos independientes consisten de n intervalos (uno por individuo), dados por

$$(L_1, R_1], \ldots, (L_n, R_n] \tag{1}$$

Los datos de vida censurados por intervalo que incluyen al menos un intervalo $(L, R]$ con ambos L y R pertenencientes a $(0, \infty)$, son usualmente referidos como datos con censura a intervalo general o tipo II [6, 8, 9]. Es decir, los datos con censura a intervalo tipo II son datos con censura que incluyen algunos intervalos finitos que no contienen el cero. Otra manera de representar una observación con censura a intervalo tipo II es usar:

$$\begin{aligned} \delta_1 &= I(T \leq L), \\ \delta_2 &= I(L < T \leq R), \\ \delta_3 &= I(T > R). \end{aligned} \tag{2}$$

donde $0 \leq L_i \leq T_i \leq R_i \leq \infty$ y $L_i < R_i$ para todo i, es decir, una observación censurada a intervalo por individuo, donde T_i es el tiempo de falla no observado del individuo i.

Asumiendo que cada sujeto es observado dos veces, donde L y R son dos variables aleatorias que satisfacen $L < R$ con probabilidad 1, es decir, las observaciones son censuradas a intervalo. Note que, si $L = R$ se tienen datos exactos y por lo tanto no existe la necesidad de la imputación. Este trabajo se enfocará estrictamente en observaciones censuradas a intervalo. Sea T una variable aleatoria que representa el tiempo de falla de un individuo, δ_1 es la indicadora de una censura a la izquierda, es decir, el evento de interés sucedió antes de inicio del proceso de observación, δ_2 es la indicadora de una censura a intervalo, es decir, el evento de interés sucedió dentro de un intervalo de dos observaciones consecutivas y δ_3 es la indicadora de una censura a la derecha, es decir el evento de interés no sucedió hasta el final del proceso de observación. Este tipo de datos surgen de estudios longitudinales con seguimientos periódicos.

2.2 Modelo de riesgos proporcionales de Cox

En los modelos de riesgos proporcionales de Cox, la función de riesgo depende en general del tiempo y un conjunto de variables explicatorias, las cuales pueden ser por ejemplo un indicador de tratamiento, la edad y el género. Un análisis de regresión suministra una evaluación de los efectos de las variables explicatorias en el tiempo de falla. El modelo de riesgos proporcionales propuesto por [10], separa estos componentes especificando que el riesgo en el tiempo t para un individuo cuyo vector de variables explicatorias es x, está dado por:

$$h(t|x) = h_0(t)\exp(x'\beta) \tag{3}$$

donde $h_0(t)$ es la función de riesgo de la distribución base F_0 y β es un vector de coeficientes de regresión. El segundo término está escrito en forma exponencial porque debe ser positivo, $x'\beta$ es llamado el predictor lineal. El modelo en (3) implica que la razón de riesgos para dos individuos depende únicamente de la diferencia entre sus predictores lineales en cualquier tiempo. Este modelo especifica que las variables explicatorias actúan multiplicativamente en la función de riesgo.

Bajo el modelo de riesgos proporcionales la función de supervivencia condicional de T dado x tiene la forma

$$\begin{aligned} S(t|x) &= \exp[-H_0(t)\exp(x'\beta)] \\ &= [S_0(t)]^{\exp(x'\beta)}, \end{aligned} \tag{4}$$

donde

$$H_0(t) = \int_0^t h_0(s)ds$$

y

$$S_0(t) = \exp\left[-\int_0^t h_0(s)ds\right]$$

son la función de riesgo acumulada base y la función de supervivencia base.

La función de riesgo acumulada condicional de T dado x tiene la forma

$$H(t|x) = H_0(t)\exp(x'\beta). \tag{5}$$

El modelo de riesgos proporcionales es el modelo de regresión más usado en el análisis de tiempos de falla.

2.3 Estimación del modelo de Cox para datos con censura a intervalo usando el algoritmo ICM

Para entender el algoritmo ICM, se debe revisar la proyección gradiente generalizada (GGP), un esquema de optimización general [11, 12]. Específicamente, suponga que se quiere maximizar una función $f(x)$ en un conjunto cerrado convexo X. Denote ∇f la primera derivada de f, H una matriz simétrica definida positiva. La GGP actualiza su estimación actual $x^{(i)}$ por

$$x^{(i+1)} = \text{Proj}\left[x^{(i)} + \alpha^{(i)} {H^{(i)}}^{-1} \nabla f(x^{(i)}), H^{(i)}, X\right]$$

donde, $\alpha^{(i)} > 0$ y Proj es la operación proyección definida por

$$\text{Proj}(x_0, H, X) := \arg\min_{x \in X} \left\{ (x - x_0)' H (x - x_0) \right\}.$$

La iteración $x^{(i)}$ en la convergencia se toma como una solución del problema de maximización.

Observe que en el modelo de Cox, la ecuación (4) se puede escribir en términos de las funciones de distribución, así:

$$1 - F(t|x) = [1 - F_0(t)]^{\exp(x'\beta)}$$

donde F_0 es la distribución de base desconocida, β es el vector de coeficientes de regresión y las n observaciones son: $(L_1, R_1, X_1), \ldots, (L_n, R_n, X_n)$.

El log de la verosimilitud es [13]:

$$L(F_0, \beta) = \sum_{j=l}^{n} \log \Big\{ [1 - F_0(L_j -)]^{\exp(x_j'\beta)} - [1 - F_0(R_j)]^{\exp(x_j'\beta)} \Big\}. \quad (6)$$

L se maximiza para obtener el NPMLE de los coeficientes de regresión junto con la distribución de base, $\widehat{F_0}$. El enfoque a este problema de maximización es extender el ICM propuesto por [7]. Extender el ICM es sencillo considerando los coeficientes de regresión como otro componente de parámetros, además de la función de distribución base. Denote por $\nabla_1 L(F_0, \beta) = \partial L(F_0, \beta)/\partial F_0$ y $\nabla_2 L(F_0, \beta) = \partial L(F_0, \beta)/\partial \beta$ las primeras derivadas, $G_1(F_0, \beta)$ y $G_2(F_0, \beta)$ son las correspondientes matrices diagonales de las segundas derivadas. El ICM extendido itera como

$$F^{(i+1)} = \text{Proj}\Big\{ F_0^{(i)} + \alpha_j G_1(F_0^{(i)}, \beta^{(i)})^{-1} \nabla_1 L(F_0^{(i)}, \beta^{(i)}), G_1(F_0^{(i)}, \beta^{(i)}),\ R \Big\} \quad (7)$$

$$\beta^{(i+1)} = \beta^{(i)} + \alpha_j G_2(F_0^{(i)}, \beta^{(i)})^{-1} \nabla_2 L(F_0^{(i)}, \beta^{(i)}) \quad (8)$$

donde α_j y Proj fueron definidos anteriormente. La proyección se toma para asegurar que la estimación de F_0 sea una función de distribución apropiada lo que requiere que $\widehat{F_0}$ sea no decreciente y esté entre 0 y 1. Ya que no existen restricciones para β no se realiza ninguna proyección.

En [7] se propone maximizar la verosimilitud conjuntamente con F_0 y β.

Este método no suministra información necesaria para calcular los intervalos de confianza para los parámetros de regresión, esta característica es un gran inconveniente para este método. Se usó bootstrap para calcular los intervalos de confianza de los parámetros.

2.4 Enfoque Bayesiano

Siguiendo a [1], suponga que existen n sujetos independientes en el estudio. Por cada sujeto j denote por T_j el tiempo de falla de interés y x_j el vector de variables explicatorias. T_j no es observado exactamente pero se conoce que cae en un intervalo observado $(L_j, R_j]$. Denote por $F(t|x)$ la función de distribución acumulada del tiempo de falla de interés dado el vector de variables explicatorias x. Por lo tanto, para el modelo de riesgos proporcionales de Cox, $F(t|x) = 1 - \exp[-\Delta_0(t)\exp(x'\beta)]$, donde $\Delta_0(t)$ denota la función de riesgo acumulada base. La verosimilitud basada en los datos observados $(L_j, R_j, x_j)_{j=1}^n$ es

$$L = \prod_{j=1}^{n} [F(R_j|x_j) - F(L_j|x_j)]. \quad (9)$$

El intervalo observado $(L_j, R_j]$ para el tiempo de falla T_j toma $(0, R_j]$ en el caso de la censura a la izquierda y (L_j, ∞) en el caso de la censura a la derecha. Para diferenciar los tipos de censura la

función de verosimilitud está dada por:

$$\begin{aligned} L &= \prod_{j=1}^{n} F(R_j|x_j)^{\delta_{j1}} \times \\ &\quad \prod_{j=1}^{n} [F(R_j|x_j) - F(L_j|x_j)]^{\delta_{j2}} \times \\ &\quad \prod_{j=1}^{n} [1 - F(L_j|x_j)]^{\delta_{j3}} \end{aligned} \tag{10}$$

donde δ_{j1}, δ_{j2}, δ_{j3} son las indicadoras de censura para el sujeto j, denotando censura a la izquierda, intervalo y derecha, respectivamente, y $\delta_{j1} + \delta_{j2} + \delta_{j3} = 1$. Este enfoque está basado en la función de verosimilitud dada en (11).

Es mejor usar la función de verosimilitud porque no requiere especificar ningún supuesto de la distribución del proceso de observación. Cuando la distribución del proceso de observación no contiene los parámetros de interés β y Δ_0, la función de verosimilitud hace una inferencia eficiente de β y Δ_0 [14].

El modelo de la función de riesgo acumulada base, $\Delta_0(t)$, está basado en una combinación lineal de *splines* monótonos [15, 16, 17]. Esta estrategia ha sido efectivamente usada para modelar funciones no decrecientes desconocidas en otros modelos de supervivencia, tales como modelar la distribución acumulada base transformada del modelo probit [18] y la odds base [19] o el logaritmo de la odds base [20] en el modelo de odds proporcionales. La función de riesgo acumulada base es modelada así,

$$\Delta_0(t) = \sum_{l=1}^{k} \gamma_l I_l(t|d) \tag{11}$$

donde cada $I_l(.|d)$ es una función base de *splines* monótonos con grado d, cada una de ellas es una función no decreciente de 0 a 1, y las γ_l's son coeficientes no negativos de los *splines* con el fin de asegurar que $\Delta_0(t)$ sea una función no decreciente, d controla el suavizamiento de los *splines*, tomando primero las funciones por tramos lineales, luego las funciones cuadráticas seguido de las funciones cúbicas, etc. Los *splines* monótonos son llamados *I-splines* (o integrados) porque ellos son funciones integradas de *M-splines* [15].

Los métodos Bayesianos por lo general requieren muestrear todos los parámetros desconocidos y las variables latentes de sus distribuciones posteriores formadas por la combinación de la función de verosimilitud y la distribución apriori. Con el fin de facilitar el cálculo de la distribución posterior en [1] se propone un aumento de datos tomando ventaja de la relación del modelo de riesgos proporcionales y un proceso de Poisson latente no homogéneo.

Se especifican las distribuciones apriori de los parámetros desconocidos β y $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$. Se asignaron distribuciones apriori exponenciales, $\exp(\lambda)$, para los γ_l's y se asignó una distribución apriori Gamma $G(a_\lambda, b_\lambda)$ al hiperparámetro λ, con media a_λ / b_λ y varianza a_λ / b_λ^2. Esta especificación de la apriori es favorable desde una perspectiva computacional porque conduce a formas conjugadas para cada una de las distribuciones condicionales posteriores de los γ_l's y de λ. Este enfoque trata a λ como aleatoria y asigna a este hiperparámetro una distribución apriori Gamma, con el fin de permitir un ajuste automático con mucho menos esfuerzo computacional. Para β_r, $r = 1, 2, \ldots, p$ se asignaron distribuciones apriori normales, $\pi(\beta_r) = N(\mu_r, \sigma_r^2)$, esto conduce a una distribución posterior condicional log-concava para cada β_r, que se puede muestrear fácilmente usando el muestreo de rechazo adaptado (ARS) [21]. Los valores especificados para $a_\lambda = b_\lambda = 1$ y $\sigma_j^2 = 100$.

3 Métodos de imputación múltiple

Los datos con censura a intervalo son realmente datos incompletos, y no exactamente datos faltantes, sin embargo, es posible realizar el proceso de imputación de este tipo de datos a través de los tiempos observados. La imputación múltiple a diferencia de la imputación simple reemplaza cada valor no observado por dos o más valores probables lo que conduce a múltiples conjuntos de datos imputados cada uno de los cuales es analizado separadamente por un método estándar, los análisis de estos conjuntos de datos toman en cuenta la variabilidad debida a los valores no observados de los datos originales [22].

3.1 Algoritmo PMDA para los datos con censura a intervalo

En [5] se utilizó el algoritmo PMDA dado en [23] con el fin de imputar los tiempos de supervivencia de datos con censura a intervalo, esto implica crear conjuntos de datos múltiples imputados usando el siguiente algoritmo iterativo.

El superíndice (i) representa la i-ésima iteración, el subíndice k da cuenta del k-ésimo conjunto de datos imputados y el subíndice j corresponde al número de observaciones.

Para un conjunto de datos arbitrario, se define (L_j, R_j) como la j-ésima observación con censura a intervalo, y sea X una matriz que representa las variables explicatorias del modelo de Cox, con X'_j la j-ésima fila de esta matriz, $j = 1, \ldots, n$. Entonces, los parámetros a ser estimados incluyen al vector de coeficientes de regresión β y a la función de supervivencia base S_0.

a. Para comenzar, suponga que las estimaciones actuales del coeficiente de regresión y la supervivencia base son $\widehat{\beta}^{(i)}$ y $\widehat{S}_0^{(i)}$, respectivamente.

b. Genere m bases de datos de posibles observaciones con censura a la derecha. Para ello, el k-ésimo conjunto de datos ($k = 1, \ldots, m$) se obtiene así:

   Para cada observación (L_j, R_j, X'_j), $j = 1, \ldots, n$: si $R_j < \infty$, obtenga Z_j de la distribución $(\widehat{S}_0^{(i)})^{\exp(X'_j \widehat{\beta}^{(i)})}$, condicional en que $(L_j < Z_j \leq R_j)$ y haga $T_{kj} = Z_j$ y $\delta_{kj} = 1$. Si $R_j = \infty$, haga $T_{kj} = L_j$ y $\delta_{kj} = 0$.

c. Use cada base de datos del paso anterior para ajustar un modelo de Cox y obtener las estimaciones $\widehat{\beta}_k^{(i)}$ y $\widehat{\Sigma}_k^{(i)}$, $k = 1, \ldots, m$.

d. Basado en $(T_{kj}, \delta_{kj}, X'_{kj})$ y $\widehat{\beta}_{(k)}^{(i)}$, calcule el estimador de Breslow de la supervivencia base $\widehat{S}_{0,(k)}^{(i)}$ para $k = 1, \ldots, m$.

e. Haga

$$\widehat{\beta}^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\beta}_{(k)}^{(i)}$$

$$\widehat{S}_0^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \widehat{S}_{0,(k)}^{(i)}$$

$$\widehat{\Sigma}^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\Sigma}_{(k)}^{(i)} + \frac{m+1}{m(m-1)} S^* \qquad (12)$$

con $S^* = \sum_{k=1}^{m} (\widehat{\beta}_{(k)}^{(i)} - \widehat{\beta}^{(i+1)})(\widehat{\beta}_{(k)}^{(i)} - \widehat{\beta}^{(i+1)})'$.

f. Verifique que $\left| \widehat{\beta}^{(i+1)} - \widehat{\beta}^{(i)} \right|$ es suficientemente pequeño (la estimación $i+1$ convergió). Si no, haga $i+1$ igual a i, vuelva al paso a. y repita el procedimiento hasta que la estimación converja.

Los valores de $\widehat{\beta}^{(i+1)}$, $\widehat{\Sigma}^{(i+1)}$ y $\widehat{S}_0^{(i+1)}$ donde se obtiene la convergencia son las estimaciones finales.

3.2 Algoritmo ANDA para datos con censura a intervalo

El algoritmo ANDA se implementa modificando el algoritmo PMDA, así:

- A partir de la primera iteración, en el paso e. se aproxima la posterior del coeficiente de regresión como una mezcla de distribuciones normales,

$$g^{(i+1)}(\beta) = \frac{1}{m} \sum_{k=1}^{m} g^{(i)}(\widehat{\beta}_{(k)}^{(i)}, \widehat{\Sigma}_{(k)}^{(i)}), \qquad (13)$$

  donde $g^{(i)}$ es una función de densidad de probabilidad normal multivariada, con media $\widehat{\beta}_{(k)}^{(i)}$ y matriz de varianzas–covarianzas $\widehat{\Sigma}_{(k)}^{(i)}$.

- A partir de la segunda iteración, en el paso b. primero se muestrea m veces de $g^{(i)}(\beta)$ para obtener $\beta_k^{(i)}$, $k = 1, 2, .., m$. Luego, para cada k y cada observación con censura a intervalo (L_j, R_j, X_j), se obtiene Z_j de la distribución $(\widehat{S}_0^{(i)})^{\exp(X'_j \beta_k^{(i)})}$, condicional en que $(L_j < Z_j \leq R_j)$ y conserva las observaciones con censura a la derecha.

- Los otros pasos son los mismos que en el método PMDA.

4 Estudio de simulación

Se desarrolla un estudio de simulación con el fin de comparar los resultados en la estimación de los parámetros de un modelo de Cox producidos por el Algoritmo ICM, el enfoque Bayesiano y al aplicar los métodos de imputación múltiple PMDA y ANDA en los datos con censura a intervalo. El protocolo de simulación usado en este trabajo siguió el propuesto por [1], donde los datos corresponden a individuos que se someten periódicamente a observaciones o exámenes.

Los pasos del estudio de simulación son los siguientes:

1. Generación de los datos. Se generaron 500 bases de datos con tres tamaños de muestra: $50, 100, 200$, del siguiente modelo:

$$F(t|x_1,x_2) = 1 - \exp[-\Lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)] \quad (14)$$

donde x_1 es una variable aleatoria Bernoulli(0.5) y x_2 es una variable aleatoria Normal$(0, 0.5^2)$ y la función de distribución acumulada base se tomó $\Lambda_0(t) = t + \log(1+t)$. Cada individuo tiene un número aleatorio de observaciones, correspondientes a visitas periódicas, las cuales son aleatorias para cada sujeto (el sujeto puede no asistir a una visita programada), pero para cada uno se garantiza al menos una visita. El número de visitas está determinado por 1 mas una variable aleatoria Poisson con media igual a 2. Los tiempos de las visitas se determinan con una distribución exponencial con media igual a 1. El intervalo observado es determinado por dos observaciones adyacentes dentro del cual se encuentra el tiempo de falla verdadero.

Se tomaron varias configuraciones para los valores de los parámetros: $\beta_1 = 1$ y $\beta_2 = 0$, $\beta_1 = 0$ y $\beta_2 = 1$, $\beta_1 = 1$ y $\beta_2 = 1$ y $\beta_1 = 0$ y $\beta_2 = 0$. De acuerdo a este esquema de generación de datos, la tasa de censura a la derecha se encuentra entre el 9% y el 21% para todas las configuraciones de los valores de los parámetros.

2. Implementación de los métodos de estimación de los parámetros del modelo de riesgos proporcionales de Cox, los cuales son: Los métodos PMDA y ANDA a través del paquete `MIICD`, el algoritmo ICM a través del paquete `intcox` y el enfoque Bayesiano a través del paquete `ICBayes`.

4.1 Resultados

En las Tablas 1, 2 y 3, se presentan los resultados de cada uno de los métodos de estimación antes descritos, basado en 500 bases de datos simuladas con tamaños de muestra 50, 100, y 200, respectivamente. En cada tabla, se presentan los parámetros estimados del modelo de Cox, junto con el error cuadrático medio estimado (ECM) y la probabilidad de cobertura empírica (CR).

Tabla 1. Resultados de los métodos de estimación del modelo de Cox para un tamaño de muestra $n = 50$.

(β_1,β_2)	Estim.	PMDA	ANDA	ICM	Bayes
(1,0)	$\widehat{\beta}_1$	0.520	0.510	1.077	0.930
	ECM	0.251	0.263	0.076	0.065
	CR	0.122	0.142	0.940	0.996
	$\widehat{\beta}_2$	–1.007	–0.005	–0.052	0.000
	ECM	0.020	0.022	0.050	0.042
	CR	0.904	0.922	0.986	1.000
(1,1)	$\widehat{\beta}_1$	0.568	0.551	1.567	1.234
	ECM	0.212	0.230	0.490	0.155
	CR	0.260	0.284	0.774	0.964
	$\widehat{\beta}_2$	0.548	0.551	1.413	1.181
	ECM	0.258	0.258	0.416	0.177
	CR	0.532	0.598	0.882	0.992
(0,0)	$\widehat{\beta}_1$	0.243	0.233	0.474	0.309
	ECM	0.079	0.073	0.269	0.131
	CR	0.564	0.590	0.426	0.976
	$\widehat{\beta}_2$	0.142	0.144	0.180	0.224
	ECM	0.046	0.046	0.085	0.097
	CR	0.832	0.830	0.894	0.992
(0,1)	$\widehat{\beta}_1$	0.373	0.362	0.821	0.629
	ECM	0.163	0.155	0.743	0.446
	CR	0.292	0.364	0.126	0.734
	$\widehat{\beta}_2$	0.712	0.711	1.358	1.226
	ECM	0.122	0.128	0.227	0.115
	CR	0.537	0.564	0.836	0.998

Tabla 2. Resultados de los métodos de estimación del modelo de Cox para un tamaño de muestra $n = 100$.

(β_1,β_2)	Estim.	PMDA	ANDA	ICM	Bayes
(1,0)	$\widehat{\beta}_1$	0.604	0.600	1.376	1.174
	ECM	0.165	0.170	0.186	0.072
	CR	0.008	0.018	0.608	0.970
	$\widehat{\beta}_2$	–1.026	–0.028	–0.084	–0.045
	ECM	0.011	0.010	0.039	0.026
	CR	0.732	0.760	0.988	0.998
(1,1)	$\widehat{\beta}_1$	0.435	0.428	0.922	0.786
	ECM	0.328	0.336	0.035	0.071
	CR	0.000	0.000	0.940	0.964
	$\widehat{\beta}_2$	0.590	0.591	1.045	0.994
	ECM	0.178	0.177	0.028	0.023
	CR	0.006	0.006	0.958	0.994
(0,0)	$\widehat{\beta}_1$	0.036	0.040	0.211	0.075
	ECM	0.012	0.012	0.069	0.024
	CR	0.704	0.726	0.754	0.998
	$\widehat{\beta}_2$	–0.014	–0.011	–0.009	–0.018
	ECM	0.009	0.009	0.022	0.019
	CR	0.730	0.706	0.956	0.996
(0,1)	$\widehat{\beta}_1$	–0.047	–0.052	0.097	–0.052
	ECM	0.015	0.015	0.038	0.023
	CR	0.674	0.676	0.910	1.000
	$\widehat{\beta}_2$	0.690	0.685	0.885	0.937
	ECM	0.114	0.116	0.045	0.032
	CR	0.172	0.178	0.994	1.000

Tabla 3. Resultados de los métodos de estimación del modelo de Cox para un tamaño de muestra $n = 200$.

(β_1,β_2)	Estim.	PMDA	ANDA	ICM	Bayes
(1,0)	$\widehat{\beta}_1$	0.460	0.460	1.033	0.850
	ECM	0.296	0.296	0.024	0.039
	CR	0.000	0.000	0.938	0.966
	$\widehat{\beta}_2$	–1.031	–0.032	–0.052	–0.036
	ECM	0.006	0.006	0.015	0.012
	CR	0.504	0.538	0.994	0.998
(1,1)	$\widehat{\beta}_1$	0.059	0.585	1.224	1.066
	ECM	0.176	0.178	0.067	0.018
	CR	0.000	0.000	0.626	0.996
	$\widehat{\beta}_2$	0.574	0.576	1.032	1.036
	ECM	0.189	0.188	0.020	0.017
	CR	0.000	0.000	0.958	0.998
(0,0)	$\widehat{\beta}_1$	–0.056	–0.057	0.125	–0.062
	ECM	0.009	0.009	0.030	0.012
	CR	0.428	0.418	0.828	0.998
	$\widehat{\beta}_2$	–0.2	–0.19	–0.28	–0.259
	ECM	0.046	0.042	0.088	0.076
	CR	0.036	0.040	1.000	0.822
(0,1)	$\widehat{\beta}_1$	–0.067	–0.069	0.098	0.176
	ECM	0.010	0.010	0.024	0.040
	CR	0.402	0.410	0.868	0.948
	$\widehat{\beta}_2$	0.654	0.652	0.924	0.990
	ECM	0.126	0.128	0.022	0.014
	CR	0.000	0.002	0.990	0.998

En la Figura 1, se muestran las probabilidades de cobertura (CR) de los métodos estudiados para cada uno de los escenarios (cada fila corresponde a una combinación de parámetros distinta).

En la Figura 2, se muestran los errores cuadráticos medios (ECM) de los métodos estudiados para cada uno de los escenarios (cada fila corresponde a una combinación de parámetros distinta).

4.2 Análisis de los resultados de simulación

De acuerdo a los resultados obtenidos en el estudio de simulación se puede observar que las estimaciones resultantes del algoritmo ICM y del enfoque Bayesiano presentan en la mayoría de las configuraciones de los parámetros del modelo de Cox mejores probabilidades de cobertura que los métodos de imputación PMDA y ANDA. En relación al error cuadrático medio, se puede apreciar que en algunas configuraciones de los parámetros los métodos de estimación basados en imputación a través de los algoritmos PMDA y ANDA presentan valores más bajos que los otros dos métodos.

Al aumentar el tamaño de la muestra los métodos de estimación bajo estudio tienden a disminuir el error cuadrático medio en la mayoría de los escenarios de parámetros considerados. Con respecto a la probabilidad de cobertura, se observa que al aumentar el tamaño de la muestra el método basado en un enfoque Bayesiano produce valores altos en la mayoría de las configuraciones, seguido del método de estimación basado en el algoritmo ICM (especialmente para muestras mayores a 50), mientras que los métodos de imputación PMDA y ANDA producen probabilidades de cobertura considerablemente bajas.

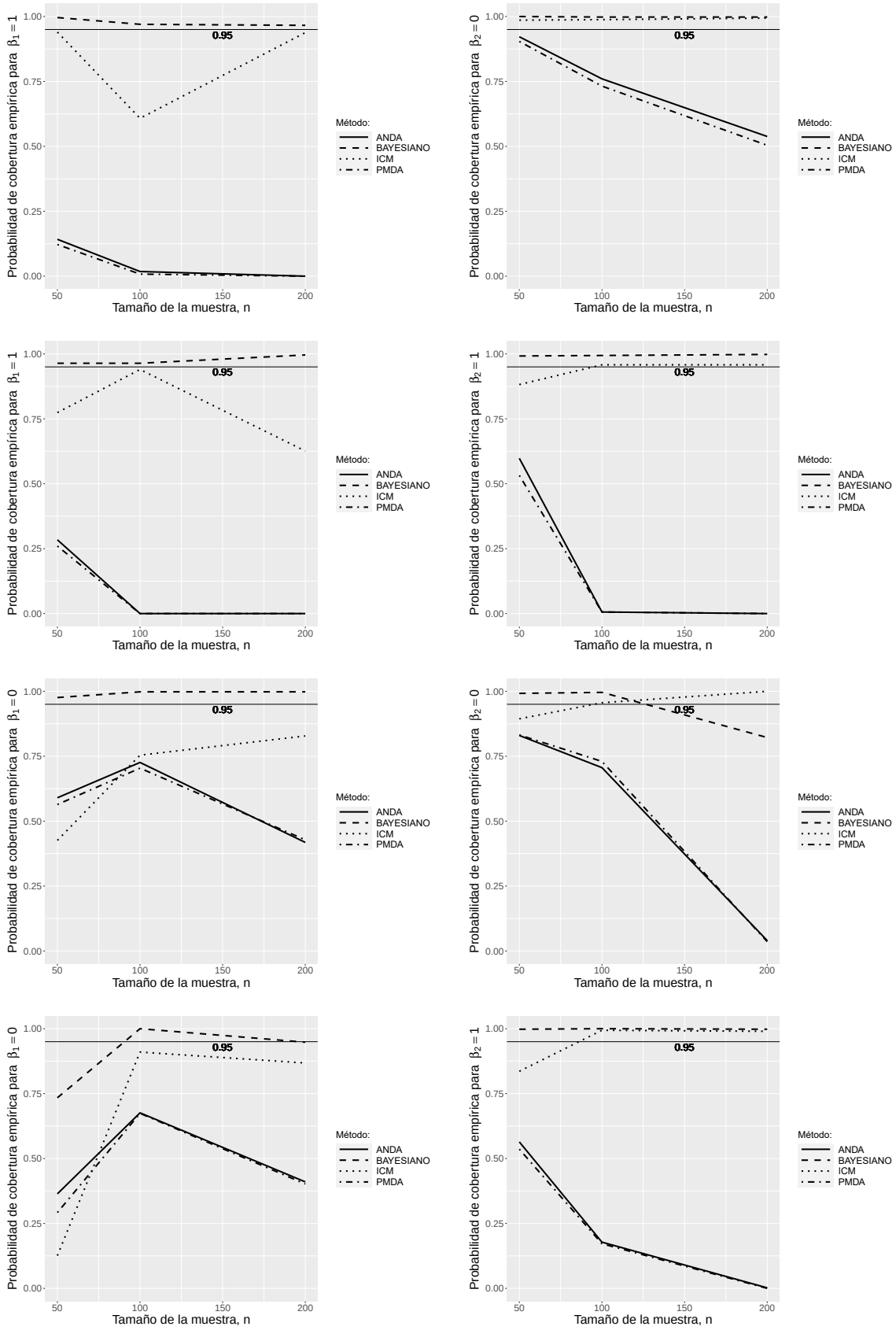

Figura 1. Probabilidades de cobertura alcanzadas por los métodos estudiados a un nivel de confianza nominal del 95%

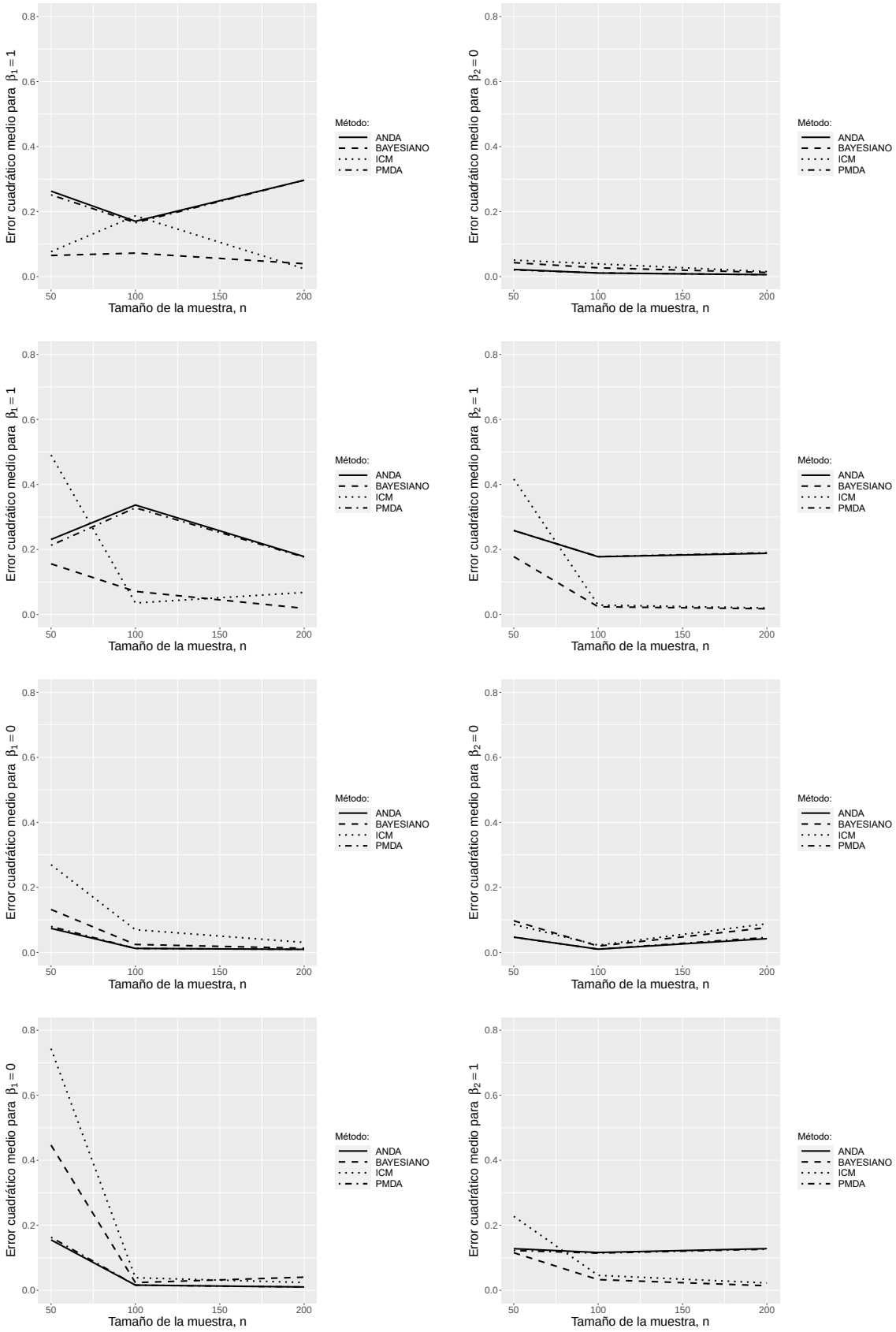

Figura 2. Errores cuadráticos medios alcanzados por los métodos estudiados

Para los métodos de imputación PMDA y ANDA se observa que la estimación de la variable continua presenta valores de probabilidad de cobertura más altos y errores cuadráticos medios más bajos que en la variable discreta. En el método basado en un enfoque Bayesiano no se aprecian diferencias apreciables entre las probabilidades de cobertura ni entre los errores cuadráticos medios obtenidos con los dos tipos de variables. En el método basado en el algoritmo ICM los valores más altos de probabilidad de cobertura y los valores más bajos de error cuadrático medio se presentan en la variable continua.

No se observan diferencias en los esquemas de imputación múltiple PMDA y ANDA, pues sus resultados son muy similares en todas las configuraciones de los parámetros.

Los métodos de imputación PMDA y ANDA presentan los mayores tiempos de simulación (1 a 3 días), seguido del método basado en un enfoque Bayesiano (30 a 40 minutos), mientras que los menores tiempos de simulación fueron obtenidos por el método de estimación basado en el algoritmo ICM (3 a 4 minutos).

5 Aplicación a una base de datos real

Se aplican los métodos de estimación del algoritmo ICM y el enfoque Bayesiano para un modelo de riesgos proporcionales de Cox a una base de datos real con censura a intervalo y se analiza el desempeño de las estimaciones de los parámetros. Estos dos métodos fueron seleccionados, ya que mostraron un mejor desempeño en el estudio de simulación presentado en la Sección 4.

Datos de mastitis en ganado lechero.

La mastitis en el ganado lechero es la inflamación de la ubre y la más importante enfermedad en el sector lechero del mundo occidental. La mastitis reduce la producción y la calidad de la leche. En [24] se llevó a cabo un estudio de mastitis donde se incluyeron 100 vacas desde el momento del parto (se asume que están libres de la infección). Fueron examinadas mensualmente a nivel de cuarto de ubre para detectar infecciones bacterianas. Dado que los cuartos de ubre están separados, un cuarto puede estar infectado mientras que el resto de los cuartos permanecen libres de infección. Las vacas fueron examinadas hasta el final del período de lactancia que dura aproximadamente entre 300 a 350 días. Algunas vacas se perdieron el seguimiento debido por ejemplo, al sacrificio. Debido al seguimiento mensual aproximado (excepto en julio/agosto en el cual sólo se planeó una visita por falta de personal), y a que en estas condiciones, es imposible observar exactamente los tiempos en los que ocurren los eventos de interés (infección de mastitis), los datos son censurados a intervalo. Los datos con censura a la derecha se presentan cuando no ocurre la infección antes del final del período de lactancia o por el tiempo perdido en el seguimiento. En este caso se tiene un porcentaje de datos con censura a derecha del 20.7%. Se registraron dos variables explicatorias, la primera (x_1) es la posición del cuarto de ubre (delantera o trasera) y la segunda (x_2) es el número de partos, con los siguientes niveles: (1) un parto, (2-4) 2 a 4 partos y (>4) más de 4 partos. Ambas variables han sido sugeridas en la literatura que impactan la incidencia de la mastitis [25, 26].

En la Tabla 4 se muestran los resultados obtenidos aplicando el método de estimación basado en un enfoque Bayesiano para los datos de mastitis.

Tabla 4. Parámetros estimados $\widehat{\beta}$ y sus respectivos intervalos de credibilidad (ICr) del 95% para $\exp(\widehat{\beta})$ de la base de datos sobre mastitis en ganado lechero aplicando el enfoque Bayesiano.

Variable	$\widehat{\beta}$	$\exp(\widehat{\beta})$	ICr para $\exp(\widehat{\beta})$
x_1 : trasera	0.11	1.12	(0.89, 1.37)
x_2 : 2-4	0.07	1.07	(0.84, 1.37)
x_2 : > 4	0.95	2.58	(1.82, 3.56)

De acuerdo a los resultados se tiene lo siguiente:

- El riesgo de que una vaca contraiga mastitis es 1.12 veces más alto si la posición de la ubre es trasera en relación a si la posición de la ubre es delantera (nivel de referencia).

- El riesgo de que una vaca que ha tenido entre 2-4 partos contraiga mastitis es 1.07 veces mayor que si la vaca ha tenido 1 parto (nivel de referencia).

- El riesgo de que una vaca que ha tenido más de 4 partos contraiga mastitis es 2.58 veces mayor que si la vaca ha tenido 1 parto (nivel de referencia).

La Tabla 5 presenta los resultados obtenidos aplicando el algoritmo ICM para los datos de mastitis.

Tabla 5. Parámetros estimados $\widehat{\beta}$ y sus respectivos intervalos de confianza (IC) aprox. del 95% para $\exp(\widehat{\beta})$ de la base de datos sobre mastitis en ganado lechero aplicando el algoritmo ICM.

Variable	$\widehat{\beta}$	$\exp(\widehat{\beta})$	IC aprox. para $\exp(\widehat{\beta})$
x_1 : trasera	0.12	1.13	(0.91, 1.43)
x_2 : 2-4	0.08	1.08	(0.85, 1.42)
x_2 : > 4	0.92	2.50	(1.88, 3.61)

De acuerdo a los resultados se tiene lo siguiente:

- El riesgo de que una vaca contraiga mastitis es 1.13 veces más alto si la posición de la ubre es trasera en relación a si la posición de la ubre es delantera (nivel de referencia).
- El riesgo de que una vaca que ha tenido entre 2-4 partos contraiga mastitis es 1.08 veces mayor que si la vaca ha tenido 1 parto (nivel de referencia).
- El riesgo de que una vaca que ha tenido más de 4 partos contraiga mastitis es 2.5 veces mayor que si la vaca ha tenido 1 parto (nivel de referencia).

A pesar de que el algoritmo ICM no suministra información necesaria para calcular intervalos de confianza para los parámetros de regresión se construyeron intervalos de confianza aproximados del 95% basados en remuestreo. El número de remuestras usado fue $R = 10000$.

6 Conclusiones

- En términos generales el algoritmo ICM y el enfoque Bayesiano presentan valores de probabilidad de cobertura más altos y errores cuadráticos medios más bajos que los obtenidos de los métodos de imputación PMDA y ANDA.
- Los métodos de imputación múltiple requieren aumentar las bases de datos para mejorar la probabilidad de cobertura pero para esto se necesitan tiempos de simulación demasiado grandes.
- A medida que aumenta el tamaño de la muestra todos los métodos estudiados basados o no en imputación, tienden a mejorar tanto las probabilidades de cobertura como los errores cuadráticos medios.
- Los métodos de imputación múltiple PMDA y ANDA no presentan diferencias considerables en sus resultados de desempeño.
- En todos los métodos estudiados, se observa un mejor desempeño en la estimación del parámetro asociado a la variable continua, comparado con la estimación del parámetro asociado a la variable discreta.
- En la aplicación a una base de datos real se puede observar que los valores de los parámetros obtenidos son muy similares tanto en el algoritmo ICM como del enfoque Bayesiano, lo cual es un resultado esperado, ya que en el estudio de simulación estos métodos presentaron resultados y desempeños muy parecidos.

Agradecimientos

Queremos agradecer al profesor Luis A. Escobar, de Louisiana State University, USA, por sus acertadas observaciones en este trabajo.

Referencias

[1] X. Lin, B. Cai, L. Wang y Z. Zhang, "A Bayesian proportional hazards model for general interval-censored data", *Lifetime Data Analysis*, vol. 21, pp. 470-490, 2015.

[2] E. Strapasson, "A simulation study to compare imputation methods to handle grouped survival data", *Revista Brasileira de Biometría*, vol. 27, pp. 210-224, 2009.

[3] C. H. Hsu, J. M. Taylor, S. Murray y D. Commenges, "Survival analysis using

auxiliary variables via non-parametric multiple imputation", *Statistics in Medicine*, vol. 25, pp. 3503-3517, 2006.

[4] M. C. Jaramillo-Elorza y C. M. Lopera-Gómez, "Estudio del efecto de la imputación de fallas en la estimación de la curva de supervivencia bajo censura a intervalo", *Ciencia en Desarrollo*, vol. 8(1), pp. 21-28, 2017.

[5] W. Pan, "A multiple imputation approach to Cox regression with interval-censored data", *Biometrics*, vol. 56, pp. 199-203, 2000.

[6] P. Groeneboom y J. A. Wellner, *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Springer, 1992.

[7] W. Pan, "Extending the iterative convex minorant algorithm to the Cox model for interval-censored data", *Journal of Computational and Graphical Statistics*, vol. 8, pp. 109-120, 1999.

[8] J. Huang y J. A. Wellner, *Interval censored survival data: A review of recent progress*, Springer, 1997.

[9] J. Sun, *Interval censoring. Encyclopedia of Biostatistics*, Wiley, 1998.

[10] D. Cox y D. Oakes, "Regression models and life tables", *Journal of the Royal Statistical Society*, vol. 34, pp. 187-220, 1972.

[11] O. Mangasarian, "Algorithms of nonlinear mathematical programming", *CS730 Course Notes, Computer Sciences Department, University of Wisconsin, Madison, WI*, 1996.

[12] D. P. Bertsekas, "Projected newton methods for optimization problems with simple constraints", *SIAM Journal on Control and Optimization*, vol. 20, pp. 221-246, 1982.

[13] D. M. Finkelstein, "A proportional hazards model for interval-censored failure time data", *Biometrics*, vol. 42, pp. 845-854, 1986.

[14] J. Sun, *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, 2007.

[15] J. O. Ramsay, "Monotone regression *splines* in action", *Statistical Science*, vol. 3, pp. 425-461, 1988.

[16] P. Joly, D. Commenges y L. Letenneur, "A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia", *Biometrics*, vol. 54, pp. 185-194, 1998.

[17] B. Cai, X. Lin y L. Wang, "Bayesian proportional hazards model for current status data with monotone *splines*", *Computational Statistics and Data Analysis*, vol. 55, pp. 2644-2651, 2011.

[18] X. Lin y L. Wang, "A semiparametric probit model for case 2 interval-censored failure time data", *Statistics in Medicine*, vol. 29, pp. 972-981, 2010.

[19] X. Lin y L. Wang, "Bayesian proportional odds models for analyzing current status data: Univariate, clustered, and multivariate", *Communications in Statistics-Simulation and Computation*, vol. 40, pp. 1171-1181, 2011.

[20] L. Wang y D. B. Dunson, "Semiparametric Bayes proportional odds models for current status data with underreporting", *Biometrics*, vol. 67, pp. 1111-1118, 2011.

[21] W. R. Gilks y P. Wild, "Adaptive rejection sampling for Gibbs sampling", *Applied Statistics*, vol. 41, pp. 337-348, 1992.

[22] M. A. Tanner y W. H. Wong, "The calculation of posterior distributions by data augmentation", *Journal of the American Statistical Association*, vol. 82, pp. 528-540, 1987.

[23] G. C. Wei y M. A. Tanner, "Applications of multiple imputation to the analysis of censored regression data", *Biometrics*, vol. 47, pp. 1297-1309, 1991.

[24] K. Goethals, B. Ampe, D. Berkvens, H. Laevens, P. Janssen y L. Duchateau, "Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma

frailty model", *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 14, pp. 1-14, 2009.

[25] J. Weller, A. Saran y Y. Zeliger, "Genetic and environmental relationships among somatic cell count, bacterial infection, and clinical mastitis", *Journal of Dairy Science*, vol. 75, pp. 2532-2540, 1992.

[26] R. Adkinson, K. Ingawa, D. Blouin y S. Nickerson, "Distribution of clinical mastitis among quarters of the bovine udder", *Journal of Dairy Science*, vol. 76, pp. 3453-3459, 1993.