COMPARACIÓN DE UN MODELO LINEAL FRECUENTISTA Y UNO BAYESIANO USANDO SIMULACIÓN DE MONTE CARLO

Jesús Escobar Corpas
Estadística Facultad de Ciencias
Universidad Nacional de Colombia
jeaescobarco@unal.edu.co

INTRODUCCIÓN

En el mudo de la estadística han existido durante la historia dos mundos diferentes para estudiar el comportamiento de los datos y extraer la información que estos pueden brindar, para así poder hacer modelamientos que expliquen el comportamiento de algún suceso y de esta manera poder inferir en posibles eventualidades futuras.
Estos dos mundos son la estadística clásica o estadística frecuentista y la estadística bayesiana, aunque las dos metodologías tienen el mismo fin, extraer información importante que ayude a inferir sobre eventualidades futuras a partir de una base de datos simple, las dos extraen esta información de forma diferente. El caso de la estadística frecuentista se basa en que existen unos parámetros fijos cuyo valor es desconocido pero que pueden ser estimados a partir de los datos, algo muy opuesto a lo que piensa la estadística bayesiana, ya que esta considera que los parámetros son valores aleatorios y que se pueden encontrar mediante el concepto de la ley de los grandes números.
Por esto se puede decir que su principal diferencia es el concepto de probabilidad, para la estadística frecuentista es un concepto objetivo y para la estadísta bayesiana es un concepto subjetivo.

M. FRECUENTISTA

El método frecuentista que utilizaremos en este estudio es el método de regresión lineal simple (lm). Este método ajusta el modelo de una variable respuesta estimando los valores del intercepto o β_0 , pendiente o β_1 y σ sigma, esto lo hace a partir de la base de datos y mediante mínimos cuadrados.

$$\text{Modelo} \quad Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$\text{donde} \quad E(e_i) = 0 \,;\, Var(e_i) = \sigma^2 \quad ; i=1,2,3,..,n$$

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

donde $\widehat{\beta}_0$, $\widehat{\beta}_1$ son la estimación de los parametros reales β_0 , β_1 además $\widehat{Y}_i$ es la estimación de la variable respuesta Y_i.

M. BAYESIANO

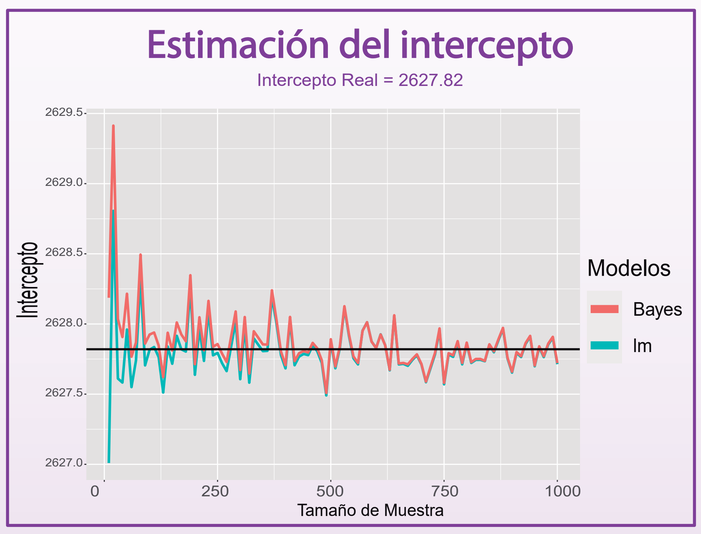
El método bayesiano que utilizaremos en este estudio es el método de Monte Carlo por cadenas de Markov. (MCMCregress). Este método permite ajustar un modelo lineal bayesiano conjugado. Por defecto trabaja con distribuciones a priori no informativas. Esta función asume que la distribución generadora de datos es normal y la distribución de los parámetros es Normal-Gamma.

EST. DE SIMULACIÓN

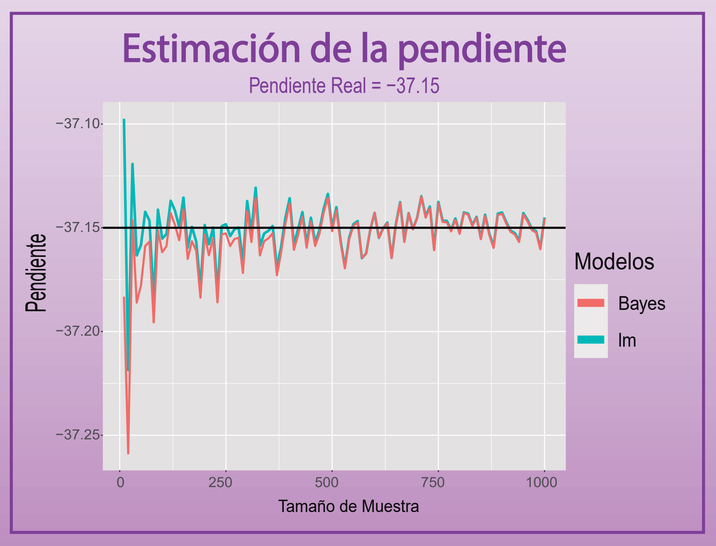
El estudio consiste en hacer una comparción del modelo (lm) y el modelo (MCMCregress) con el fin de determinar cuál de los dos puede estimar mejor los parámetros (β_0 , β_1 y σ). Para ello se realizará el estudio mediante simulación de monte carlo, asumiendo como conocidos los valores de los parametros (β_0 = 2626.8, β_1 =-37.15 y σ^2 = 9244.59) los cuales fueron obtenidos del ejemplo 2.1 (Datos de propelente), del capítulo 2 del libro Montgomery, D. Peck, E. y Vining, G. (2005).

El estudio se inicia con tamaño de muestra n=10 y se aumenta a esa misma razón hasta llegar a mil y en cada una de las escalas se hacen 10.000 simulaciones. Luego se promedian cada uno de los valores estimados por cada simulación en los diferentes modelos.
Inicialmente se hará el estudio con un modelo bayesiano no informativo y se comparará con el modelo lm, luego se hará el mismo estudio pero con modelo bayesino informativo y nuevamente se volverá a comparar con el modelo lm.
Lo que se espera es que a medida que aumenta el numero de datos se pueda hacer una mejor estimación de los parámetros en los diferentes modelos, de esto surgen alguna preguntas como:
¿habrá un modelo que necesite menos datos para estimar muy bien los parámetros? ¿hay uno de ellos que estime mejor que el otro sin importar el tamaño de muestra? ¿son iguales?.

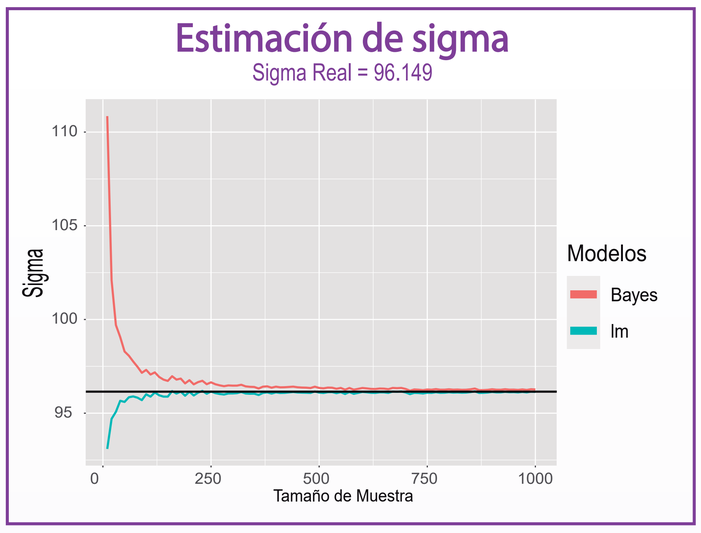
RESULTADOS



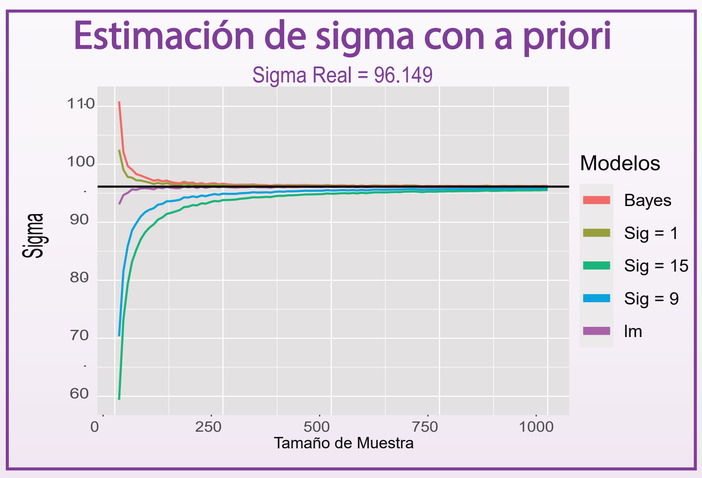
En este resultado se puede apreciar como los dos modelos a medida que aumenta la muestras mejoran la precisión de la estimación del parámetro intercepto. Algo llamativo es que tienen un comportamiento muy similar, aunque al inicio parece funciona mejor el modelo lm, al final son idénticos los dos, al punto de sobreponerse.



En esta estimación se observa un comportamiento muy parecido en los modelos al visto en la estimación del intercepto, es decir, a media que aumenta el tamaño de muestra aumenta la precisión en la estimación del parámetro.



este resultado muestra que los modelos estiman el parámetro sigma de manera opuesta, ambos tienden al valor real pero el modelo lm subestima el valor y MCMCregress lo sobreestima durante todo el estudio.



En este estudio se le asignaron unos valores a priori al parámetro sigma del modelo MCMCregress, los resultados obtenidos nos muestran que con un valor a priori uno (1) para el parámetro sigma, el modelo mejora la precisión para estimar el valor real del parámetro, pero si se asignan valores a priori más grandes, el modelo subestima el valor real del parámetro muy por debajo.

CONCLUSIONES

Basado en los resultados que se obtuvieron en el estudio realizado mediante simulación de Monte Carlo, podemos decir que los valores estimados para los parámetros mediante los modelos lm y MCMCregress son muy acertados si se comparan con el valor real, además, que son muy similares entre ellos independientemente del parámetro que se quiera estimar, excepto con el parámetro sigma ya que en este parámetro muestran una tendencia opuesta.
Por otra parte, cuando se trabajó con un modelo MCMCregress informativo, es decir, se asignaron valores a priori a los parámetros; sólo se reflejaron cambios cuando se asignaron valores a priori al parámetro sigma, este cambio fue visible cuando se estimó este mismo parámetro en el modelo de estudio, ya que en ocasiones mejoró la estimación y en otras por el contrario empeoró. Sin embargo los otros parámetros no mostraron cambio alguno ya que se mantenían iguales al modelo sin valor a priori, por tal motivo se desistió mostrarlos en gráficas.

Referencias: Devore, Jay L. probabilidad y estadística para ingeniería y ciencias. Séptima edición.ISBN-13:978-970-686-831-2, ISBN-10:970-686-831-3
Montgomery, D., Peck, E. y Vining, G. (2005). Introducción lanálisis de regresión lineal, Cecsa, tercera edición.