# Proyecto Inferencia Bayesiana

Vanessa Alcalde, Luciano Garrido, Pablo Martinez Angerosa

27/11/2020

## 1 Introducción

En agosto de 1973, G.C. McDonald y R.C. Schwing [1] utilizaron Regresión Ridge para encontrar un modelo cuyas variables regresoras, compuestas por variables climáticas, socioeconómicas y de polución del aire, logran explicar la tasa de mortalidad de 60 ciudades estadounidenses en el año 1963. Como mencionan los autores, si bien los métodos estadísticos no necesariamente implican un relación de causa y efecto, bajo el supuesto de que esta relación existe, estos métodos proveen una herramienta para entender las contribuciones relativas a una variable de estudio.

En esta investigación utilizamos el método de Bayes empírico para obtener las distribuciones de los coeficientes que explican la tasa de mortalidad sobre la misma base de datos que utilizaron G.C. McDonald y R.C. Schwing [1]. Para esto en una primera instancia se ajusta un modelo de regresión lineal múltiple, donde en el proceso de construcción se eliminaron algunas variables explicativas que no resultaron significativas. Al mismo tiempo los coeficientes $\beta_i$ que se obtienen del modelo lineal construido a partir de la muestra dada de datos pasan a ser los ejes principales de información a priori de los parámetros dándole al análisis el enfoque empírico bayesiano. Junto a esto se realiza un modelo de regresión lineal bayesiano y se comparan los resultados.

## 2 Datos

Los datos corresponden a 60 ciudades de Estados Unidos en el año 1963 y fueron obtenidos de la base de datos correspondientes al artículo original de G.C. McDonald and R.C. Schwing, "Instabilities of Regression Estimates Relating Air Pollution to Mortality" [1]. Cuenta con 16 variables cuantitativas agrupadas en las categorías climática, socioeconómica y de polución del aire. La Tabla 1 muestra la descripción de las variables en la base de datos.

Las variables etiquetadas como HC, NOX y SO pertenecen a la categoría de polución del aire, como se muestra en la Figura 1, HC y NOX presentan una correlación elevada de 0.98 y por lo tanto la información que estas variables adhieren al modelo es simplemente de incertidumbre, por lo cual optamos por excluir la variable explicativa NOX.

Las variables PREC, JANT, JULT y HUMID pertenecen a la categoría de climáticas.

Las variables MORT, OVR65, POPN, EDUC, HOUS, DENS, NONW, WWDRK y POOR pertenecen a la categoría de variables socioeconómicas. Estas variables son importantes para poder medir las diferencias en los estados de salud de las distintas comunidades. La variable explicada MORT representa la tasa de mortalidad cada 100,000 habitantes. Existen correlaciones de 0.7 entre EDUC y WWDRK y entre NONW y POOR. Análisis de multicolinealidad realizados posteriormente en la búsqueda del ajuste del modelo lineal primario no dieron argumentos suficientes para excluir alguna de estas variables.

| | Nombre de variables | Descripción |
|---|---|---|
| 1 | PREC | Promedio anual de precipitación (en pulgadas). |
| 2 | JANT | Promedio de temperatura del mes de Enero (en Farenheit). |
| 3 | JULT | Promedio de temperatura del mes de Julio (en Farenheit). |
| 4 | HUMID | Promedio anual del porcentaje de humedad relativa a las 13 horas. |
| 5 | MORT | Tasa de mortalidad cada 100.000 habitantes. |
| 6 | OVR65 | Porcentaje de población mayor de 65 años en áreas metropolitanas. |
| 7 | POPN | Promedio del tamaño del hogar. |
| 8 | EDUC | Mediana de años de escolarización completos para mayores de 22 años. |
| 9 | HOUS | % de viviendas en buenas condiciones con todos los servicios. |
| 10 | DENS | Población por milla cuadrada en áreas urbanas en 1960. |
| 11 | NONW | % de población no blanca en áreas urbanas en 1960. |
| 12 | WWDRK | % de trabajadores en ocupaciones “no manuales”. |
| 13 | POOR | % de familias con ingresos anuales menores $3000. |
| 14 | HC | Polución potencial relativa de hidrocarbono. |
| 15 | NOX | Polución potencial relativa de óxido nítrico. |
| 16 | SO | Polución potencial relativa de dióxido de azufre. |

Tabla 1: Detalle de las variables en la base de datos.

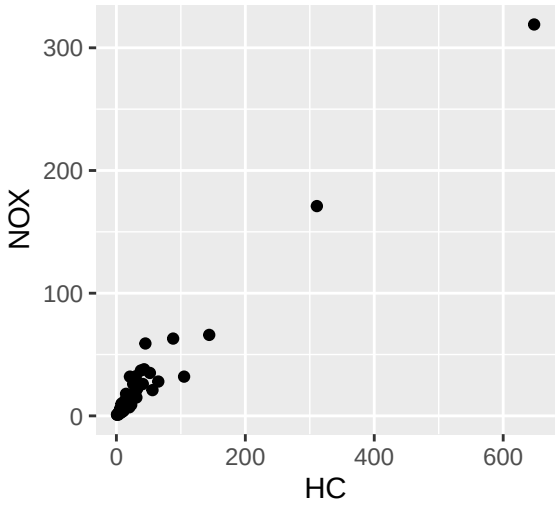


Figura 1: Gráfico de dispersión de HC contra NOX.

En la Tabla 2 se muestra la media, el mínimo y el máximo para las 16 variables cuantitativas.

| | Variable | Media | Mínimo | Máximo |
|---|---|---|---|---|
| 1 | PREC | 37.37 | 10.00 | 60.00 |
| 2 | JANT | 33.98 | 12.00 | 67.00 |
| 3 | JULT | 74.58 | 63.00 | 85.00 |
| 4 | HUMID | 57.67 | 38.00 | 73.00 |
| 5 | MORT | 940.40 | 790.70 | 1113.20 |
| 6 | OVR65 | 8.79 | 5.60 | 11.80 |
| 7 | POPN | 3.26 | 2.92 | 3.53 |
| 8 | EDUC | 10.97 | 9.00 | 12.30 |
| 9 | HOUS | 80.91 | 66.80 | 90.70 |
| 10 | DENS | 3876 | 1441 | 9699 |
| 11 | NONW | 11.87 | 0.80 | 38.50 |
| 12 | WWDRK | 46.08 | 33.80 | 59.70 |
| 13 | POOR | 14.37 | 9.40 | 26.40 |
| 14 | HC | 37.85 | 1.00 | 648.00 |
| 15 | NOX | 22.65 | 1.00 | 319.00 |
| 16 | SO | 53.77 | 1.00 | 278.00 |

Tabla 2: Resumen de todas las variables.

# 3 Métodos

Los tres modelos utilizados en esta investigación incluyen regresión lineal múltiple, modelo bayesiano y modelo bayesiano empírico. El proceso consistió primero en obtener un modelo de regresión lineal múltiple que cumpla con los supuestos teóricos y que esté compuesto por un conjunto de variables significativas. Posteriormente utilizamos la información obtenida de las estimaciones puntuales de los $\beta_i$ y el $\sigma^2$ como información a priori para realizar un modelo bayesiano empírico con las mismas variables resultantes del modelo de regresión. Para finalizar realizamos un modelo bayesiano utilizando como distribuciones a priori las default presentes en el paquete *rstanarm* de *R* para que este realice un ajuste por escala de las posterioris y comparamos el desempeño de todos los modelos.

## 3.1 Metodología

Un modelo de regresión lineal múltiple es un modelo lineal en los parámetros en el cual la variable de respuesta, $Y$, es determinada por un conjunto de variables independientes, las variables explicativas. Se busca el hiperplano que mejor ajuste a los datos.

$$Y_i = \beta^T x_i + \varepsilon_i \tag{1}$$

Asumiendo las hipótesis de Gauss-Markov los $\varepsilon_i \sim N(0, \sigma^2)$ y son incorrelacionados. Los $\beta_i$ y las $x_i$ son considerados constantes y por ende la variable explicada de la Ecuación (1) distribuye $y_i \sim normal(\beta^T x_i, \sigma^2)$. Las estimaciones puntuales de los $\beta_i$ que minimizan el error cuadrático medio del modelo se obtienen por el método de mínimos cuadrados ordinarios (MCO).

Por otro lado, los modelos bayesianos son capaces de sintetizar la información de la muestra y una creencia a priori, no muestral, utilizando el Teorema de Bayes. La creencia a priori de los parámetros que se quieren estimar se expresan a través de una distribución de probabilidad, llamada distribución a priori. Los parámetros a estimar, a diferencia del enfoque clásico, ya no son una estimación puntual sino que tienen un comportamiento de distribución dentro de una medida de probabilidad.

En la Ecuación (1) dentro de un enfoque bayesiano, los $\beta_i$ y $\sigma^2$ son variables aleatorias.

La distribución conjunta resultante de la Ecuación (1) es

$$
\begin{aligned}
& p\left(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2\right) \\
& = \prod_{i=1}^{n} p\left(y_i \mid x_i, \beta, \sigma^2\right) \\
& = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2\right\}
\end{aligned} \tag{2}
$$

Si el vector de $\beta \sim Normal\ multivariada(\beta_0, \Sigma_0)$ obtenemos que la posteriori normal conjugada es

$$
\begin{aligned}
& p\left(\beta \mid y, \mathbf{X}, \sigma^2\right) \\
& \propto p\left(y \mid \mathbf{X}, \beta, \sigma^2\right) \times p(\beta) \\
& \propto \exp\left\{-\tfrac{1}{2}\left(-2\beta^T \mathbf{X}^T y/\sigma^2 + \beta^T \mathbf{X}^T \mathbf{X}\beta/\sigma^2\right) - \tfrac{1}{2}\left(-2\beta^T \Sigma_0^{-1}\beta_0 + \beta^T \Sigma_0^{-1}\beta\right)\right\} \\
& = \exp\left\{\beta^T\left(\Sigma_0^{-1}\beta_0 + \mathbf{X}^T y/\sigma^2\right) - \tfrac{1}{2}\beta^T\left(\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}/\sigma^2\right)\beta\right\}
\end{aligned} \tag{3}
$$

Un desafio importante de la estadística bayesiana es definir la información necesaria para construir la distribución a priori. Incluso algunas veces ni siquiera existe información previa o precisa que se pueda considerar como una creencia de los parámetros.

Un posible enfoque a esta problemática es la utilización de un modelo bayesiano empírico y utilizar los $\beta$ obtenidos por mínimos cuadrados para centrar las distribuciones a priori en base a estos parámetros puntuales estimados.

Los métodos empíricos de Bayes son procedimientos de inferencia estadística en los cuales la creencia a priori se construye a partir de los datos. Si bien, Kass y Wasserman (1995) en [2] sugieren que esta distribución no puede considerarse una previa real, la cantidad de información de $y$ que se utiliza no es de un margen considerable.

Otro posible problema de este enfoque es que no se conoce las distribuciones de los $\beta_i$, pero en esta investigación consideramos que son normales, centradas en los $\beta_i$ estimados por MCO y con desviaciones típicas basadas en los desvíos de estos parámetros.

Para tener una medida de comparación entre distintos modelos se utilizó el criterio de predicción RMSE descrito en la Ecuación (4) donde $y_i$ representa las predicciones y $f_i$ los valores reales. El mejor desempeño se asocia a un menor RMSE.

$$
RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_i\right)^2} \tag{4}
$$

Para obtener las simulaciones de las cadenas se utilizó el paquete *stanarm* de *R*, el cual genera 4 cadenas con 2000 simulaciones cada una para obtener simulaciones Markov MonteCarlo (MCMC) de la posteriori. Las primeras 1000 cadenas son desechadas para evitar la dependencia con los valores iniciales y que se estabilice la cadena. Dos indicadores que se utilizan para monitorear un indicio de convergencia son el $\widehat{R}$ y el número de muestras efectivas, $\widehat{n}_{eff}$. Las medidas $\widehat{R} < 1.1$ y $\widehat{n}_{eff} > 5m$, entendiendo $m$ como el número de cadenas, se consideran un indicio de convergencia.

Las prioris por defecto en *stanarm* fueron diseñadas para ser levemente informativas. Según la documentación en muchos casos, sino en la mayoría, los valores por defecto van a obtener un buen desempeño, aunque no siempre suceda. La forma en la que este paquete funciona es a través del ajuste de la escalde las prioris. La priori auxiliar, el desvío estándar, por defecto tiene una configuración de $exponencial(1)$. La priori para los coeficientes son normales centradas en 0 y con una escala (desviación estádar) de 2.5.

## 3.2 Los modelos

A continuación en la Ecuación (5) se muestra el modelo resultante por modelos lineales. Todas las variables de la base original fueron analizadas en modelos preliminares pero muchas resultaron no significativas por lo que no se incluyen en el modelo final. Las variables PREC, NONW y SO resultaron significativas a un 0.1%

y JANT, JULT y EDUC al 5%. Se puede ver en el modelo final que quedan variables de las tres categorías principales en las que fueron organizadas en la base.

Este modelo es significativo globalmente al 95% de confianza y logra una variablidad explicada de la $y_i$ expresada mediante el $R^2$ de un 0.8076 y un $R^2$ ajustado de un 0.7836 (el cual penaliza por la cantidad de regresores), logrando un buen despeño.

Se decidió sacar de la base de datos cinco observaciones que resultaron influyentes o atípicas luego de los análisis correspondientes y esto se corroboró dado que al excluirlas de la base y volver a ajustar el modelo se apreciaba un cambio sustancial en el $R^2$ y el valor de los $\beta_i$ estimados.

$$MORT_i = \beta_0 + \beta_1 PREC_i + \beta_2 JANT_i + \beta_3 JULT_i + \beta_4 EDUC_i + \beta_5 NONW_i + \beta_6 SO_i + \varepsilon_i \qquad (5)$$

Donde $\varepsilon_i \sim N(0, \sigma^2)$, $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$.

El mismo modelo de la Ecuación (5) se utiliza para el enfoque bayesiano y el bayesiano empírico. En la Tabla 3 se muestran el resumen de las medias y las desviaciones típicas de las distribuciones a priori normales utilizadas para los parámetros $\beta_i$ y $\sigma^2$ del modelo bayesiano empírico. Estos parámetros de las distribuciones a priori están basados en los resultados estimados por MCO del modelo de regresión lineal.

| | Parámetro | Varible asociada | Media | Desviación Típica |
|---|---|---|---|---|
| 1 | beta0 | intercepto | 1122.791036 | 101.12442 |
| 2 | beta1 | PREC | 2.239360 | 0.52209 |
| 3 | beta2 | JANT | -1.037841 | 0.48114 |
| 4 | beta3 | JULT | -2.102458 | 0.98114 |
| 5 | beta4 | EDUC | -12.412463 | 5.36332 |
| 6 | beta5 | NONW | 3.891093 | 0.63640 |
| 7 | beta6 | SO | 0.302782 | 0.06265 |
| 8 | sigma | Desvío | 26.19 | 0.5 |

Tabla 3: Parametros de las distribuciones normales previas para el modelo bayesiano empirico.

# 4 Resultados

En esta sección mostramos los resultados obtenidos para ambos modelos bayesianos, hacemos la comprobación de ambos y posteriormente comparamos el despeño de los tres modelos ajustados.

En la Tabla 4 se muestran los resultados obtenidos de los indicadores de convergencia, $\hat{R}$ y $\hat{n}_{eff}$ para ambos modelos bayesianos. Como se puede observar los indicadores sugieren que no hubo problemas de convergencia para ambos modelos.

| | Parámetro MB | Rhat | n_eff | Parámetro MBE | Rhat | n_eff |
|---|---|---|---|---|---|---|
| 1 | Intercepto | 1.0 | 3100 | Intercepto | 1.0 | 3519 |
| 2 | PREC | 1.0 | 2901 | PREC | 1.0 | 3230 |
| 3 | JANT | 1.0 | 3618 | JANT | 1.0 | 3653 |
| 4 | JULT | 1.0 | 3004 | JULT | 1.0 | 3533 |
| 5 | EDUC | 1.0 | 3376 | EDUC | 1.0 | 3478 |
| 6 | NONW | 1.0 | 3199 | NONW | 1.0 | 2608 |
| 7 | SO | 1.0 | 3322 | SO | 1.0 | 3371 |
| 8 | sigma | 1.0 | 2891 | sigma | 1.0 | 4128 |

Tabla 4: Indicadores de convergencia para los modelos bayesianos.

La Tabla 5 muestra los intervalos de credibilidad de las distribuciones posteriori de los parámetros para ambos modelos en un 95%. Los números reflejan que no hay grandes diferencias y todos los intervalos se intersectan en su mayoría.

| | Parámetro MB | 2.5% | 50% | 97.5% | Parámetro MBE | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|---|
| 1 | Intercepto | 918.84 | 1125.75 | 1324.80 | Intercepto | 988.61 | 1123.88 | 1255.68 |
| 2 | PREC | 1.22 | 2.23 | 3.26 | PREC | 1.59 | 2.23 | 2.92 |
| 3 | JANT | -2.03 | -1.04 | -0.05 | JANT | -1.69 | -1.04 | -0.40 |
| 4 | JULT | -4.06 | -2.10 | -0.11 | JULT | -3.40 | -2.10 | -0.77 |
| 5 | EDUC | -23.13 | -12.50 | -1.81 | EDUC | -19.67 | -12.52 | -5.41 |
| 6 | NONW | 2.65 | 3.89 | 5.18 | NONW | 3.06 | 3.88 | 4.70 |
| 7 | SO | 0.18 | 0.30 | 0.43 | SO | 0.22 | 0.30 | 0.39 |
| 8 | sigma | 22.07 | 26.40 | 32.55 | sigma | 26.21 | 26.52 | 27.30 |

Tabla 5: Intervalos de credibilidad para los parametros de los modelos bayesianos.

En la Figura 2 se muestran las distribuciones de los estadísticos para las distintas simulaciones en ambos modelos bayesianos. Se puede ver que ambos tienen un buen desempeño en estadísticos con distintos criterios como la mediana, el rango intercuartílico y el máximo. El desempeño entre los dos modelos es muy similar.

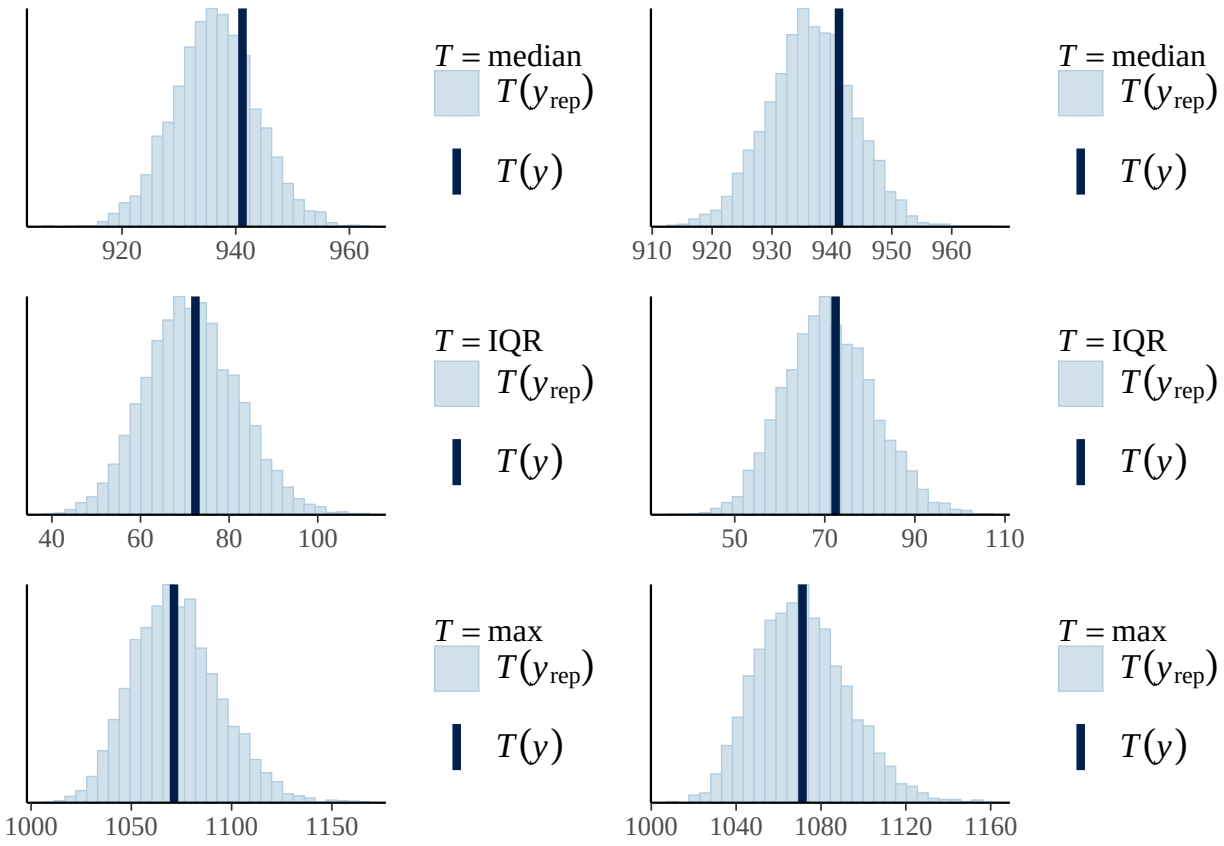


Figura 2: Gráfico comparativo de estimadores, en la izquierda se muestran la mediana, el rango intercuartílico y el máximo para el modelo de Bayes y en la derecha los mismos estimadores para el modelo de Bayes empírico.

La Figura 3 muestra las distribuciones simuladas de la distribución predictiva posteriori para ambos modelos, estas son contrastadas con la distribución empírica de la variable explicada, la tasa de mortalidad. Se puede

apreciar que existe una mayor concentración de las simulaciones para el modelo bayesiano empírico, pero igualmente ambos modelos muestran buenos resultados.

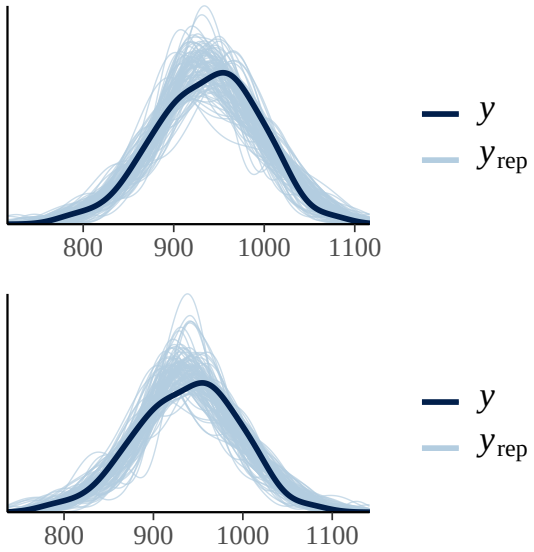


Figura 3: Gráfico comparativo de las predictivas posteriori, en la imagen superior se muestra el modelo de Bayes y en la inferior el modelo de Bayes empírico.

En la Tabla 6 utilizamos el RMSE como medida predictiva para establecer una medida de comparación entre los modelos. Se puede observar que entre los modelos bayesianos hay una diferencia muy pequeña de error en favor del modelo bayesiano empírico pero podemos concluir que ambos logran un mismo desempeño predictivo. Sin embargo, comparando ambos con el resultado obtenido por modelos lineales se aprecia una diferencia considerable en la exactitud de predicción en favor de los modelos bayesianos.

| | Modelo | RMSE |
|---|---|---|
| 1 | Lineal | 51.79169 |
| 2 | Bayesiano | 24.56636 |
| 3 | Bayesiano Empírico | 24.47284 |

Tabla 6: RMSE para todos los modelos.

Desde otra perspectiva, en la Tabla 7 se observa como las estimaciones para todos los parámetros, tanto las puntuales de modelos lineales como las medias en las que se centran las distribuciones obtenidas de las posterioris por modelos bayesianos y sus desvíos correspondientes, se encuentran en entornos cercanos sin notarse diferencias considerables en las estimaciones propuestas por cada modelo. Observando los desvíos correspondientes se ve una mayor concentración en las medias en el modelo bayesiano empírico.

# 5 Discusión y conclusiones

En esta investigación se utilizaron modelos estadísticos de regresión lineal múltiple, bayesianos y bayesianos empírico con una misma base de datos. Los tres modelos estimados muestran consistencia entre sí, ya que tanto los parámetros puntuales para los $\beta_i$ obtenidos por modelos lineales y las medias de las distribuciones de los parámetros a posteriori por modelos bayesianos, se mueven en intervalos muy similares.

Aunque el modelo bayesiano fue obtenido mediante una búsqueda de las distribuciones posterioris, mediante ajuste por escala,

| | Variable | ML | Estimación | Error Estándar | MB | Media | Desvío | MBE | Media | Desvío |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Intercepto | beta0 | 1122.791036 | 101.12442 | beta0 | 1123.3 | 101.7 | beta0 | 1123.4 | 67.1 |
| 2 | PREC | beta1 | 2.239360 | 0.52209 | beta1 | 2.2 | 0.5 | beta1 | 2.2 | 0.3 |
| 3 | JANT | beta2 | -1.037841 | 0.48114 | beta2 | -1.0 | 0.5 | beta2 | -1.0 | 0.3 |
| 4 | JULT | beta3 | -2.102458 | 0.98114 | beta3 | -2.1 | 1.0 | beta3 | -2.1 | 0.7 |
| 5 | EDUC | beta4 | -12.412463 | 5.36332 | beta4 | -12.5 | 5.5 | beta4 | -12.5 | 3.7 |
| 6 | NONW | beta5 | 3.891093 | 0.63640 | beta5 | 3.9 | 0.6 | beta5 | 3.9 | 0.4 |
| 7 | SO | beta6 | 0.302782 | 0.06265 | beta6 | 0.3 | 0.1 | beta6 | 0.3 | 0.0 |
| 8 | Desvío | sigma | 26.19 | 0.5 | sigma | 26.6 | 2.7 | sigma | 26.6 | 0.3 |

Tabla 7: Media y desvio de los parametros a posteriori de los modelos bayesianos y estimaciones para modelos lineales.

sin ninguna información previa de la muestra, este logra obtener una similitud muy grande con las posterioris encontradas por el modelo bayesiano empírico, cuya información a priori estaba centrada en información de la muestra, particularmente de los parámetros obtenidos por MCO. Al mismo tiempo, el modelo bayesiano logra una similitud con el modelo lineal. De igual manera las distribuciones obtenidas por el modelo bayesiano empírico no se alejan tanto de la información a priori establecida. Esto nos da a interpretar que hay una validación de las variables y las estimaciones obtenidas.

En cuanto a un enfoque de selección de uno de los tres modelos, el modelo bayesiano empírico es el que logra un mejor desempeño en los distintos casos evaluados. Desde una óptica de predicción el modelo empírico bayesiano logra una pequeña diferencia a favor sobre el modelo bayesiano, y una diferencia considerable sobre el modelo lineal. En cuanto a lo observado, el modelo bayesiano empírico demuestra una mayor concentración de las simulaciones de la distribución predictiva con respecto a la variable explicada en comparativa con el modelo bayesiano.

A futuro, sería interesante comparar los resultados de esta investigación para datos actualizados con las mismas variables y comprobar si se siguen obteniendo los mismos resultados de desempeño y el mismo conjunto de variables significativas para los modelos propuestos.

# Referencias

[1] G.C. McDonald y R.C. Schwing. 1973. *Instabilities of Regression Estimates Relating Air Pollution to Mortality.* Technometrics, vol. 15, pp. 463-482.

[2] P.D. Hoff. 2009. *A First Course in Bayesian Statistical Methods.* Springer.