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 β_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: Descripción de variables en la base de datos.

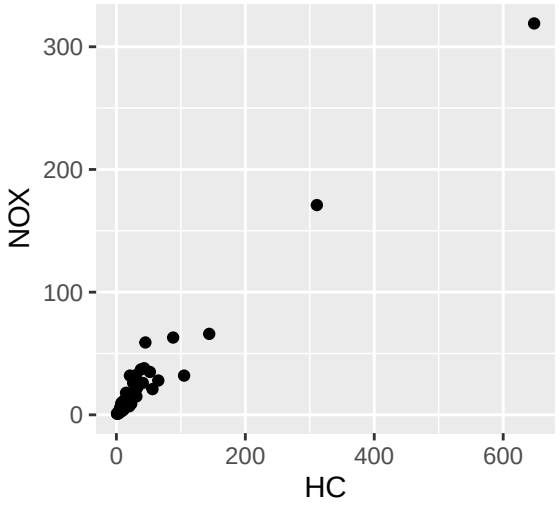

Figura 1: Gráfico de dispersión de HC contra NOX.

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 algunas variables de interés.

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 β_i y el σ^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 para los β_i y el σ^2 presentes en el paquete *stan_glm* de *R* y comparamos el desempeño de ambos modelos bayesianos.

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 β_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 β_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 β_i y σ^2 son variables aleatorias.

La ecuación 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, donde se utiliza los β 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 β_i, pero en esta investigación consideramos que son normales, centradas en los β_i estimados por MCO y con desviaciones típicas basadas en los desvíos de estos parámetros.

3.2 Los modelos

A continuación en la Ecuación (4) 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 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 β_i estimados.

$$
MORT_i = \beta_0 + \beta_1 PREC_i + \beta_2 JANT_i + \beta_3 JULT_i + \beta_4 EDUC_i + \beta_5 NONW_i + \beta_6 SO_i + \varepsilon_i \tag{4}
$$

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

El mismo modelo de la Ecuación (4) 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 β_i y σ^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^2	Varianza	26.19	0.5

Tabla 3: Parámetros de las distribuciones normales previas para el modelo bayesiano empírico.

4 Resultados

Los resultados obtenidos

$$\beta\ \beta$$

$$\beta$$

5 Discusión y conclusiones

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.