Proyecto Inferencia Bayesiana

Vanessa Alcalde, Luciano Garrido, Pablo Martinez Angerosa

27/11/2020

Introducción

En agosto de 1973, G.C. McDonald and R.C. Schwing [1] utilizaron Regresión Ridge múltiple para encontrar un modelo estable cuyos variables regresoras compuestas por variables climáticas, socioeconómicas y de polución del aire logran explicar la tasa de mortalidad de 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. 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 previa 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.

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," Technometrics, vol. 15, pp. 463-482, 1973. [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 100000 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 para excluir alguna de estas variables.

Tabla 1: Descripción de variables en la base de datos.

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

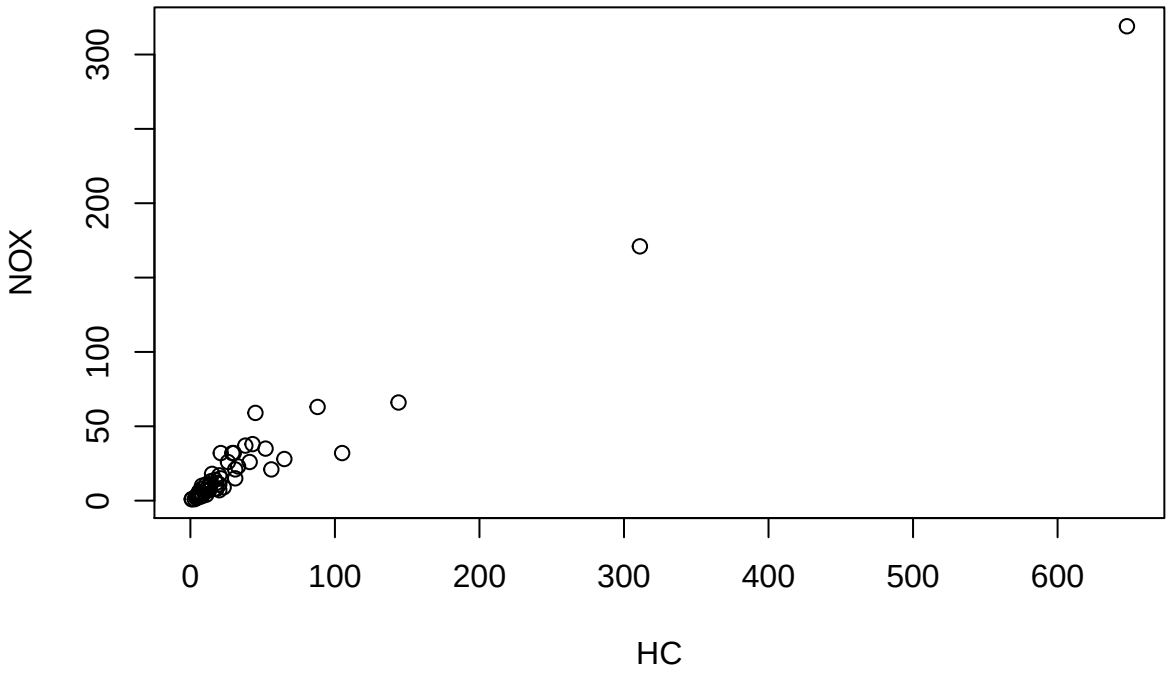

Figura 1: Gráfico de dispersión de HC contra NOX

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

```
# resumen <- utils::read.table('summary.txt', sep='\t',
#                              header=TRUE,
#                              row.names=NULL, stringsAsFactors=FALSE,
#                              dec = ",")
# data.table::setDT(resumen)
#
# base::print(xtable::xtable(data.table::setDT(resumen),
#             caption = "Resumen de algunas variables de interés",
#             type="html"),
#             caption.placement = "bottom")
```

Tabla 2: Resumen descriptivo de las variables.

Variable	Media	Min	Max
PREC	37.37	10.00	60.00
JANT	33.98	12.00	67.00
JULT	74.58	63.00	85.00
HUMID	57.67	38.00	73.00
MORT	940.40	790.70	1113.20
OVR65	8.79	5.60	11.80
POPN	3.26	2.92	3.53
EDUC	10.97	9.00	12.30
HOUS	80.91	66.80	90.70
DENS	3876	1441	9699
NONW	11.87	0.80	38.50
WWDRK	46.08	33.80	59.70
POOR	14.37	9.40	26.40
HC	37.85	1.00	648.00
NOX	22.65	1.00	319.00
SO	53.77	1.00	278.00

Métodos

El modelo resultante por modelos lineales es:

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

Donde $\varepsilon_i \sim N(0, \sigma^2)$.

Utilizamos el mismo modelo desde un enfoque bayesiano. Como información a priori de cada uno de los β_i y σ^2 utilizamos las estimaciones puntuales obtenidas desde modelos lineales en particular para los β_i tomamos distribuciones normales cuya media son las estimaciones puntuales y su correspondiente error estándar como desviación típica.

Resultados

Los resultados obtenidos

```
# datos <- utils::read.table('polucion.txt', sep='\t',
#                            header=TRUE,
#                            row.names=NULL, stringsAsFactors=FALSE,
#                            dec = ",")
#
#
# datosML <- datos[-c(6, 28, 32, 37, 2),]
#
# modelo2 <- rstanarm::stan_glm(MORT ~ PREC + JANT +
#                                 JULT + EDUC + NONW + SO,
#                               data = datosML,
#                               family = gaussian(link = "identity"),
#                               prior_intercept = rstanarm::normal(1122.79104, 101.12442),
#                               prior = rstanarm::normal(base::c(2.23936, -1.03784, -2.10246,
#                                                                -12.41246, 3.89109, 0.30278),
#                                                        base::c(0.52209, 0.48114, 0.98114,
#                                                                5.36332, 0.63640, 0.06265)),
#                               prior_aux = rstanarm::normal(26.19,0.5),
#                               seed = 12345)
#
# base::summary(modelo2)
# rstanarm::prior_summary(modelo2)
#
# # Estimates:
# #   mean   sd     10%    50%    90%
# # (Intercept) 1121.4   56.2 1051.6 1120.7 1195.0
# # PREC           2.2    0.3    1.9    2.2    2.6
# # JANT          -1.0    0.3   -1.4   -1.0   -0.7
# # JULT          -2.1    0.6   -2.8   -2.1   -1.4
# # EDUC         -12.3    3.1  -16.2  -12.4   -8.4
# # NONW           3.9    0.3    3.5    3.9    4.3
# # SO             0.3    0.0    0.3    0.3    0.4
# # sigma         19.5    1.4   17.8   19.4   21.2
#
#
# #prior_aux: normal
# # Estimates:
# #   mean   sd     10%    50%    90%
# # (Intercept) 1123.2   67.4 1036.6 1122.2 1210.6
# # PREC           2.2    0.4    1.8    2.2    2.7
# # JANT          -1.0    0.3   -1.5   -1.0   -0.6
# # JULT          -2.1    0.7   -3.0   -2.1   -1.2
# # EDUC         -12.4    3.7  -17.1  -12.4   -7.7
# # NONW           3.9    0.4    3.4    3.9    4.5
# # SO             0.3    0.0    0.2    0.3    0.4
# # sigma         27.2    0.8   26.3   27.0   28.3
#
#
# modelo2
# modelo2$coefficients
# modelo2$stanfit
#
#
```

```
# ###Diagnostico de las Cadenas
# bayesplot::mcmc_trace(modelo2)
# bayesplot::mcmc_trace(modelo2,pars = "(Intercept)")
# bayesplot::mcmc_trace(modelo2,pars = "PREC")
# bayesplot::mcmc_trace(modelo2,pars = "JANT")
# bayesplot::mcmc_trace(modelo2,pars = "JULT")
# bayesplot::mcmc_trace(modelo2,pars = "EDUC")
# bayesplot::mcmc_trace(modelo2,pars = "NONW")
# bayesplot::mcmc_trace(modelo2,pars = "SO")
# bayesplot::mcmc_trace(modelo2,pars = "sigma")
#
# ##Intervalos de credibilidad del 95%
# bayesplot::mcmc_intervals(modelo2, prob = 0.95)
# bayesplot::mcmc_intervals(modelo2 ,pars=c("PREC",
#                                           "JANT",
#                                           "JULT",
#                                           "EDUC",
#                                           "NONW",
#                                           "SO"), prob = 0.95)
#
# bayesplot::mcmc_intervals(modelo2 ,pars=c("SO"), prob = 0.95)
#
#
# bayesplot::mcmc_hist(modelo2)#Histograma de los coeficientes
# bayesplot::mcmc_hist_by_chain(modelo2)#Histograma de los coeficientes por cadena
# bayesplot::mcmc_dens(modelo2)#Estimacion de densidad de los coeficientes
#
# # Diagnostico de Modelo
# ##Predictivas posteriores
# pred2<-rstanarm::posterior_predict(modelo2,draws = 100)
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "mean")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "median")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "var")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "IQR")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "min")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "max")
# bayesplot::pp_check(modelo2, plotfun = "stat", stat = "q75")
#
# bayesplot::ppc_dens_overlay(datosML$MORT,pred2)
```

```
{r, fig.cap ="Gráfico...",fig.pos="H"} # bayesplot::ppc_dens_overlay(dat
#
```

Referencias

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

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