Ciencias Matemáticas, Vol. 30, No. 2, Pag. 65-71, 2016
Recibido 09-2016

Aplicación del Criterio BMA de selección de modelos bajo enfoque Bayesiano a un problema de Sensibilidad Alérgica Ocupacional en Panaderías de La Habana.
Aplication of BMA Criteria of models selection under Bayesian approach to an allergic sensitivity occupational study on bakers of Havana.

Marelys Crespo [1], Vivian Sistachs [1]

Resumen La incertidumbre inherente a un modelo, generalmente no es considerada al seleccionar y construir un modelo estadístico, lo que afecta la interpretación predictiva y sobreestima las inferencias sobre cualquier resultado de interés. El Promedio Bayesiano de Modelos (BMA, por sus siglas en inglés "Bayesian Model Averaging") es una alternativa viable que incorpora la incertidumbre. El presente estudio ilustra la aplicación a un problema de salud pública, relacionado a la sensibilidad alérgica ocupacional en panaderías de La Habana para determinar el "mejor" modelo en el contexto de la regresión logística binaria utilizando la estrategia desarrollada en [1]. Además, se muestra la implementación de la aplicación en el software R.

Abstract The uncertainty inherent to a model, generally is not considered on the selection and construction moment of a statistic model, this takes an impact on the predictive interpretation and overestimates the inferences on any result of interest. The Bayesian Model Averaging (BMA) is a viable alternative that includes uncertainty. This paper illustrates the application of a public health problem: The allergic sensitivity an occupational study on bakers of Havana to determine the "best" model and the factors that have an influence on such model, using the strategy developed by [1]. We also show the implementation on R software of the BMA criteria.

Palabras Clave
Regresión Logística — Selección de modelos — BMA

[1] *Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba, m.crespo@matcom.uh.cu, vivian@matcom.uh.cu*

Introducción

La incertidumbre de un modelo, generalmente, no es considerada al seleccionar y construir un modelo estadístico adecuado. Los procedimientos convencionales, comunmente formulan y asumen- guiados por una serie de pruebas de significancia- un único modelo como el correcto para realizar inferencias, sin embargo, ignorar la incertidumbre inherente a un modelo afecta su interpretación predictiva y sobreestima las inferencias sobre cualquier estadístico de interés.

Los métodos Bayesianos, al cuantificar la incertidumbre inicial y posterior de un modelo, junto al desarrollo de los métodos MCMC (*Monte Carlo Markov Chain*) y a la mejora de los algoritmos computacionales para su implementación, han ampliado e intensificado, sustancialmente, su desarrollo y alcance, transformándose rápidamente en una alternativa viable para tratar la incertidumbre de una extensa cantidad de modelos.

El promedio bayesiano de modelos (BMA, por sus siglas en inglés "Bayesian Model Averaging"), es una alternativa a la selección Bayesiana de modelos que incorpora la incertidumbre de un modelo, al considerar una mezcla de una amplia variedad de modelos, es un mecanismo coherente para evaluar, dentro de un marco probabilístico, la incertidumbre de un modelo y la relacionada a sus parámetros.

Por otro lado, la regresión logística es una de las técnicas estadísticas más conocidas y utilizadas para modelar una variable de respuesta categórica en función de variables independientes. Se aplica en campos tan distintos como epidemiología, ecología, sociología, medicina por mencionar algunos.

El presente estudio tiene como objetivo encontrar un modelo de regresión logística bajo el enfoque Bayesiano utilizando el criterio BMA, y su implementación en el software R, para detectar el nivel de influencia de determinados ácaros que puedan ser detonantes de enfermedades alérgicas en los panaderos.

Los contenidos de este artículo han sido organizados en tres secciones. En la primera sección se exponen los resultados del estudio de la regresión logística binaria, haciendo énfasis en el enfoque Bayesiano. En la segunda sección se presenta el estudio del criterio BMA. Se discuten los aspectos a tener en cuenta para su implementación y se muestra la estrategia propuesta en [1] para el software estadístico R. En la última sección, se presenta la aplicación al problema en particular, se resaltan las salidas del criterio BMA, cómo determinar el "mejor" modelo y la importancia de cada una de las variables incluidas en dicho modelo. También, se muestra que la estrategia tiene muy buenos resultados a partir de su capacidad predictiva. Finalmente, en las conclusiones se exponen los resultados fundamentales obtenidos.

1. Regresión Logística Binaria

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. En muchas áreas de la medicina como Salud Pública o Medicina Clínica, se aplican los métodos de la regresión logística; particularmente la Regresión Logística Binaria debido a que la variable respuesta es dicótomica.

El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados [9] [11] [8] y su objetivo es encontrar el modelo con un mejor ajuste y el menor número de parámetros [11] [3]. El tratamiento que se da en la literatura a este modelo es desde dos puntos de vista: el clásico y el Bayesiano.

1.1 Modelo de Regresión Logística

Sea Y una variable respuesta o dependiente, que toma valores 1 o 0, lo que indica presencia o ausencia de la característica o suceso de interés, y k variables independientes, $\mathrm{X} = (X_1, ..., X_k)^t$. El modelo del suceso es $\mathrm{Y} = \mathrm{P}(x) + \varepsilon$, donde ε es el término de error, $\mathrm{P}(x)$ es la probabilidad de que $\mathrm{Y} = 1$ para el valor observado $x(x_1, ..., x_k)^t$ de las variables independientes, y su modelo de regresión logística es:

$$\begin{aligned} \mathrm{P}(x) &= \mathrm{P}(\mathrm{Y} = 1 \big| \mathrm{X} = x) \\ &= \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)} \end{aligned} \quad (1)$$

De ahí que $1 - \mathrm{P}(x)$ indicará la probabilidad de que Y tome el valor 0.

Al aplicar una transformación *logit* a la ecuación 1 es posible linealizarla, como se muestra a continuación:

$$logit(\mathrm{P}(x)) = \ln\Big(\frac{\mathrm{P}(x)}{1 - \mathrm{P}(x)}\Big) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

1.2 Modelo Bayesiano de Regresión Logística

Bajo el enfoque Bayesiano hay tres componentes asociadas a la estimación de los parámetros: la distribución *a priori*, la función de verosimilitud y la distribución *a posteriori*. Las cuales están relacionadas por el Teorema de Bayes:

$$\mathrm{P}(M_i \big| y) \propto f(y \big| \theta_i)\mathrm{P}(\theta_i)$$

A continuación se explican brevemente las tres componentes.

Función de verosimilitud:

Dada la probabilidad de un suceso, la verosimilitud del $i-$*ésimo* sujeto es *Bernoulli*.

$$\mathrm{L}(y \big| \theta) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1 - y_i}$$

donde $\pi_i = \mathrm{P}(x_i)$ y $\theta = (\beta_0, \beta_1, ..., \beta_k)$.

Distribución *a priori*

Se pueden utilizar dos tipos de distribuciones *a priori*: informativas y no informativas. Las distribuciones *a priori* informativas se utilizan si se conoce información previa sobre el valor de los parámetros desconocidos. En otro caso, se utilizan las distribuciones no informativas.

Distribución *a posteriori*

La distribución *a posteriori* se obtiene multiplicando la distribución *a priori* de los parámetros por la función de verosimilitud,

$$\begin{aligned} f(\theta \big| y) &= \prod_{i=1}^{n} \Big(\frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})} \Big)^{y_i} \cdot \\ &\cdot \Big(1 - \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})} \Big)^{1 - y_i} \mathrm{I}(\theta)^{\frac{1}{2}}, \end{aligned}$$

donde $\theta = (\beta_0, \beta_1, ..., \beta_k)$ e $\mathrm{I}(\theta)$ es la función de información de Fisher.

1.2.1 Estimación Bayesiana de los parámetros

A diferencia del enfoque clásico que estima los parámetros β_j del modelo utilizando el método de máxima verosimilitud, en el enfoque Bayesiano estos coeficientes se obtienen directamente de la distribución de probabilidad *a posteriori* de los parámetros desconocidos.

En modelos simples podría ser fácil encontrar la distribución marginal, sin embargo, cuando la dimensión se incrementa, también aumenta la dificultad de los cálculos. La limitación principal para la implementación del enfoque Bayesiano es precisamente, que la obtención de la distribución *a posteriori* a menudo requiere de la integración de funciones de alta dimensión, por ello, los métodos MCMC proporcionan una solución viable.

1.3 Evaluación del ajuste del modelo

Como el objetivo de seleccionar un modelo es predecir valores de la variable respuesta, una forma coherente de juzgar la capacidad de un modelo es precisamente, evaluar qué tan bien predice futuras observaciones. [10]

Para realizar este análisis, se propone particionar el conjunto de datos en dos subconjuntos, el primero se llamará conjunto de prueba, que se emplea para la modelación y el segundo llamado conjunto de validación, que se utilizará para evaluar la capacidad predictiva del modelo. La partición es implementada mediante un submuestreo aleatorio.

El *error de clasificación* es una medida muy utilizada para determinar la capacidad predictiva para datos binarios, es el porcentaje de respuestas mal clasificadas. La probabilidad predictiva del modelo ajustado se calcula y el grupo de variables independientes se predice de acuerdo a la siguiente regla. En este caso el punto de corte es: $c = 0{,}5$.

$$r_i = \begin{cases} 1 & \text{si} \quad |y_i - \hat{\pi}_i| > 0{,}5 \\ 0{,}5 & \text{si} \quad |y_i - \hat{\pi}_i| = 0{,}5 \\ 0 & \text{si} \quad |y_i - \hat{\pi}_i| < 0{,}5 \end{cases}$$

1.4 Selección Bayesiana de modelos

En el enfoque bayesiano, al igual que en el frecuentista, no se puede asegurar encontrar el modelo verdadero, por lo que para elegir el "mejor" modelo que aproxima los datos, es necesario tener un método de selección.

De ahí que se presenten dos problemas fundamentales: la búsqueda de modelos y los criterios para la selección de modelos; ambos problemas se encuentran integrados en el Promedio Bayesiano de Modelos (BMA), el cual elimina las deficiencias inherentes en la selección de modelos determinista combinando la información de todos los modelos en lugar de usar solo uno.

Por lo anterior, en el presente estudio se hace uso del criterio BMA para la selección de modelos.

2. Promedio Bayesiano de Modelos (BMA)

El trabajar con Factores de Bayes implica, la elección de un modelo (o posiblemente varios modelos), sin embargo, si el modelo elegido es solo uno de muchas posibilidades, se corre el riesgo de ignorar la incertidumbre asociada al modelo [2], por lo que, se debe considerar todos los modelos e inferencias que pudieran originar. Una solución a la selección Bayesiana de modelos que incorpora más que ignorar la incertidumbre de un modelo, es la combinación de modelos.

Este enfoque, conocido como BMA [6] [2] [10], se centra en la distribución de una cantidad de interés Δ (una observación futura del mismo proceso que genera el conjunto de datos y), y bajo el modelo completo para los datos, inducido por las distribuciones iniciales, se sigue que la distribución predictiva Bayesiana de Δ es:

$$\mathrm{P}(\Delta|y) = \sum_{i=1}^{k} \mathrm{P}(\Delta|y, M_i)\mathrm{P}(M_i|y) \tag{2}$$

donde M_i es el modelo $i-$ésimo del espacio de modelos M, y

$$\mathrm{P}(\Delta|y, M_i) = \int_{\Theta} \mathrm{P}(\Delta|\theta_i, M_i)\mathrm{P}(\theta_i|M_i)d\theta_i$$

El desarrollo general para el BMA permitió un amplio trabajo de la incertidumbre en la selección de modelos empleando la búsqueda determinística y estocástica y el uso de Factores de Bayes para la comparación de modelos.

Se han desarrollado varios algoritmos eficientes y aproximaciones para la mezcla de modelos bajo independencia en las variables. La idea principal es que cuando las variables son independientes, el promedio a través de diferentes modelos se simplifica de manera considerable porque se puede obtener la muestra en el espacio de variables en lugar del espacio de modelos.

2.1 Implementación del BMA

Para la correcta aplicación del BMA se deben precisar los siguientes aspectos fundamentales:

- Especificación de distribuciones *a priori* para modelos y parámetros.
- Búsqueda de modelos en M a partir de los datos.
- Cálculo de probabilidades *a posteriori* de los modelos.

2.1.1 Especificación de distribuciones.

Los modelos Bayesianos requieren la especificación de la distribución *a priori* para los parámetros del modelo. Estas pueden ser informativas o no informativas.

Distribuciones informativas

Las distribuciones *a priori* informativas representan el conocimiento sobre los parámetros. Se pueden construir u obtener de estudios empíricos previos o del conocimiento del investigador experto en el problema de interés, estas pueden obtenerse por medio de estimación puntual o por intervalos y deben estar en forma de distribución.

Distribuciones no informativas

En ocasiones no se dispone de información *a priori*, esta carencia podría utilizarse si se encontrara una distribución de probabilidad $\pi(\theta)$ que no contenga información acerca de θ, en el sentido de no favorecer ningún valor de θ por encima de todos los lógicamente posibles. Este tipo de distribuciones recibe el nombre de no informativas.

En ocasiones ocurre que las distribuciones *a posteriori* resultantes son impropias o las distribuciones *a priori* no son invariantes bajo reparametrización. En [4] se propone una aproximación en la que se evita el tener que considerar la invarianza. Este método, llamado de Jeffreys, consiste en tomar una densidad *a priori* proporcional a la raíz cuadrada del determinante de la matriz de información de Fisher. La distribución resultante es invariante bajo transformaciones y relativamente fácil de calcular.

2.1.2 Búsqueda de modelos en M a partir de los datos

La búsqueda de buenos modelos es también una parte integral del BMA, sin embargo, esto no es obvio en la ecuación (2), la cual simplemente promedia el espacio de modelos completo. Se ha comprobado que cuando la cantidad de modelos es grande la implementación no siempre devuelve resultados factibles.

Para implementar la metodología del BMA en un espacio reducido de modelos, se buscan solo los modelos que mejor se ajustan a los datos [7]. Para ello existen dos métodos: la búsqueda determinística y la búsqueda estocástica en el espacio de modelos.

Los esquemas de búsqueda determinísticos sugeridos en el BMA son el método de la Ventana de Occam [7] y el algoritmo de *leaps and bounds* [12].

2.2 Estrategia de Selección de Modelos en Regresión Logística bajo Enfoque Bayesiano

Para el estudio del BMA en el modelo de Regresión Logística aplicado a problemas de salud se sigue la estrategia desarrollada en [1]:

1. Fijar la probabilidad *a priori* para los parámetros del modelo. (Se utiliza una distribución *a priori* no informativa de Jeffreys.)

2. Fijar la probabilidad *a priori* para los modelos. (En este caso se utiliza una distribución uniforme.)

3. Definir el espacio de modelos. (En este caso se utiliza una búsqueda determinística, particularmente, se emplea el algoritmo *leaps and bounds*.)

4. Realizar la selección de modelos usando el BMA.

La estrategia fue implementada en R y se explica a continuación.

2.2.1 Implementación del BMA en R

R es un lenguaje y entorno de programación, cuya característica principal es que forma un entorno de análisis estadístico para la manipulación y el análisis de datos, así como la creación de gráficos. R puede considerarse como otra implementación del lenguaje de programación S-PLUS, con la particularidad de que es un software GNU, General Public License.

El entorno incluye un intérprete del lenguaje R y paquetes para aplicaciones estadísticas concretas. El lenguaje es orientado a objetos, interpretado a alto nivel y tiene una sintaxis dirigida al manejo de datos estadísticos. Desde la página oficial de R (www.r-project.org) es posible descargar el archivo de instalación que permite una fácil, práctica y muy rápida puesta en marcha del software, y funciona en una amplia variedad de plataformas.

Para la implementación de la selección de modelos en el paquete estadístico R, se hace uso de la librería BMA, esta permite aplicar la selección a modelos lineales, a modelos lineales generalizados y a modelos de supervivencia, además incluye funciones que permiten mostrar los resultados gráficamente.

La función necesaria para la implementación del BMA a modelos de regresión logística es big.glm en la cual es necesario definir la fórmula (modelo), especificar la familia de la distribución de la variable dependiente Y y la base donde están contenidos los datos.

La Figura 1 muestra el diagrama de la implementación del BMA.

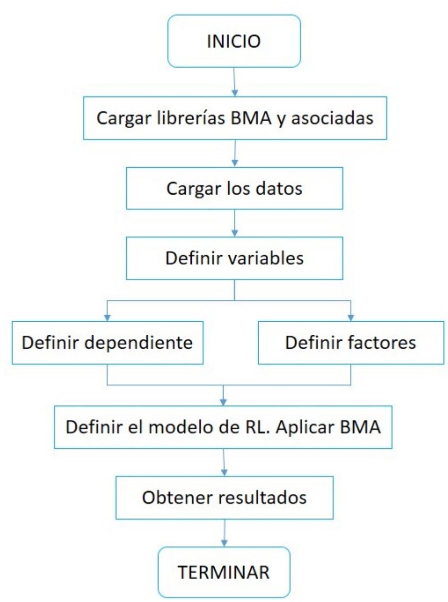

Figura 1. Implementación del BMA en R

3. Sensibilidad Alérgica Ocupacional en panaderías de La Habana

La actividad de los panaderos es una de las profesiones más vinculadas al desarrollo del asma ocupacional, reportándose una fuerte asociación entre la exposición al polvo de harina y la sensibilización a los alérgenos. En Cuba, actualmente no se conoce el riesgo al que están expuestos los trabajadores que interactúan con los ácaros de la levadura, el trigo y la soya en polvo.

Todo lo anterior motivó un estudio para determinar factores de riesgo para la salud de los panaderos en La Habana.

Por indicación del médico especialista, las variables analizadas se muestran en el Cuadro 1, ya que se considera que influyen en la sensibilidad alérgica.

Las variables que representan la sensibilidad a la soya en polvo (*soyap*), al trigo ($trigo$) y a la levadura (lev) son propias del ambiente de las panaderías. Mientras que las restantes variables relacionadas a ácaros (*DP*, $blomia$, *DS*, *siro*, $farinae$, $tyro$ y $lepi$) se encuentran presente en el medio ambiente en general. La variable respuesta es *sens*, que toma valor 1 en presencia de la alergia y 0 en ausencia.

Variable	*Descripción*	*Tipo de variable*
edad	Edad	Numérica
sexo	Género (1-M, 2-F)	Dicotómica
tiexp	Tiempo de exposición (años)	Numérica
fuma	Fuma (1-Si, 2-No)	Dicotómica
soyap	Sensibilidad a la soya en polvo	Categórica
trigo	Sensibilidad al trigo	Categórica
lev	Sensibilidad a la levadura	Categórica
DP	Sensibilidad al ácaro DP	Categórica
blomia	Sensibilidad a la blomia	Categórica
DS	Sensibilidad al ácaro DS	Categórica
siro	Sensibilidad al siro	Categórica
farinae	Sensibilidad al farinae	Categórica
tyro	Sensibilidad al tyrophagus	Categórica
lepi	Sensibilidad al lepidoglyphus	Categórica
sens	Sensibilidad alérgica (1-Si, 0-No)	Dicotómica

Cuadro 1. Descripción de las variables.

Para realizar el estudio se dividió el conjunto total de datos en dos subconjuntos, el conjunto de prueba que contenía 90 observaciones (el 75 % del total), y un conjunto de validación que contenía las 30 observaciones restantes, que fueron seleccionadas mediante un submuestreo aleatorio.

Al aplicar el criterio BMA en el software estadístico R al conjunto de prueba, se obtuvo el Cuadro 2, en el cual se muestran seis columnas. La primera con el nombre del intercepto (término independiente) y las variables estudiadas; a su derecha, las tres próximas columnas (P!, EV y SD) indican respectivamente, la probabilidad *a posteriori* en porciento de cada variable de estar en el modelo ideal (se puede interpretar como la importancia de cada variable), el valor esperado y la desviación estándar *a posteriori* de cada coeficiente.

	P!	EV	SD	Mod 1	Mod 2
Intercep	100	-1.4020	0.4410	-1.3730	-1.5759
edad	0.0	0.0	0.0	.	.
sexo	0.0	0.0	0.0	.	.
tiexp	0.0	0.0	0.0	.	.
fuma	0.0	0.0	0.0	.	.
soyap	0.0	0.0	0.0	.	.
trigo	0.0	0.0	0.0	.	.
lev	85.7	5.3889	492.9554	6.2892	.
DP	0.0	0.0	0.0	.	.
blomia	0.0	0.0	0.0	.	.
DS	14.3	0.1163	0.3009	.	0.8123
siro	0.0	0.0	0.0	.	.
farinae	100.0	0.9034	0.2126	0.9003	0.9218
tyro	0.0	0.0	0.0	.	.
lepi	0.0	0.0	0.0	.	.
nVar				2	2
BIC				-338.6085	-335.0299
post prob				0.857	0.143

Cuadro 2. Salida del BMA.

Las columnas restantes representan los dos modelos seleccionados por el criterio BMA, ordenados de forma decreciente de acuerdo a su probabilidad *a posteriori*, por lo que el "mejor" modelo es el que aparece bajo el encabezado "Mod 1" (modelo 1). También se observa cuales son las variables que están incluidas y el valor de los coeficientes en cada modelo.

Las últimas filas muestran el número de variables que contiene cada modelo (nVar), el valor del Criterio de Información Bayesiano (BIC) y la probabilidad *a posteriori* del modelo (post prob).

En la Figura 2 se puede apreciar la gráfica de las distribuciones *a posteriori* de los coeficientes para las variables incluidas en el modelo 1, las cuales tiene forma de campana, lo que representa que tienen una alta probabilidad de inclusión en el modelo ideal. En los anexos se muestra el gráfico de la probabilidad de inclusión de las variables restantes.

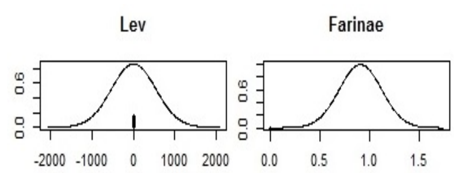

Figura 2. Probabilidad de inclusión de las variablesdel modelo 1.

En la Figura 3 se puede observar de manera gráfica la selección de modelos. El color azul indica que el signo del parámetro de las variables en los modelos analizados es positivo, en el caso de aparecer el color rojo indicaría un valor negativo. El ancho de las columnas representa la importancia de los modelos. Todo lo anterior se relaciona con lo obtenido en el Cuadro 2.

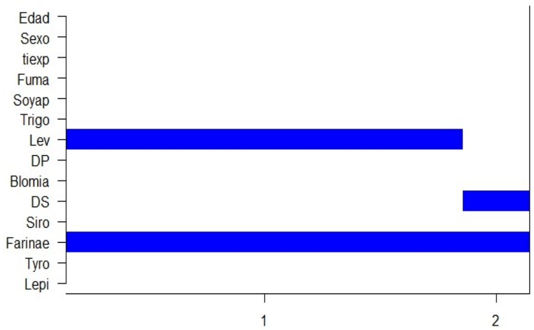

Figura 3. Selección de modelos.

Por todo lo anterior, se seleccionó el modelo 1 ya que tiene una probabilidad *a posteriori* de 0,857 e incluye las variables lev y $farinae$ que tienen una importancia del 85,7 y 100 % respectivamente. De ahí que el modelo quede especificado por la siguiente ecuación 3:

$$P(Y=1) = \frac{\exp\{\beta_0 + \beta_1(lev) + \beta_2(farinae)\}}{1+\exp\{\beta_0 + \beta_1(lev) + \beta_2(farinae)\}}, \quad (3)$$

donde $\beta_0 = -1{,}373$, $\beta_1 = 6{,}2892$ y $\beta_2 = 0{,}9003$.

Sin embargo, a pesar de que el modelo seleccionado tiene una alta probabilidad de ser el ideal, el especialista plantea que estas no son las únicas variables importantes para estar en el modelo; lo cual indica que la selección de modelos usando el criterio BMA no está excenta de que el proceso de medición o la forma en cómo se obtuvieron los datos esté relacionado con el resultado de dicha selección.

Con el objetivo de analizar la calidad del modelo, primeramente se calculó el error cuadrático medio (ECM = 0,0566), el cual verifica que los datos de validación tenían muy poca desviación respecto al modelo.

Para evaluar su capacidad predictiva se analizó el porciento de datos bien clasificados a través del conjunto de validación. El Cuadro 3 muestra que el 93 % fueron bien clasificados por el modelo, lo que es una prueba de que el BMA devolvió excelentes resultados:

Bien clasificados	Mal clasificados
93.3 %	6.7 %

Cuadro 3. Clasificación de los datos de validación.

Conclusiones

Al aplicar el BMA en el contexto de la Regresión Logística bajo el enfoque Bayesiano, se obtuvo un modelo con probabilidad *a posteriori* de 0.86 y calidad de clasificación del 93 % de los datos de validación. Por lo que se determinaron dos factores de riesgo importantes para la sensibilidad alérgica de los panaderos en La Habana: la presencia de reactividad cutánea a los ácaro levadura y farinae.

Los resultados del estudio se obtuvieron mediante la implementación en el software libre R de la estrategia desarrollada en [1] aplicada al problema estudiado.

Finalmente, se desea resaltar el buen comportamiento del criterio BMA para la selección de modelos en Regresión Logística Binaria bajo el enfoque Bayesiano.

Código Fuente

A continuación se muestra el código programado en el software R para la aplicación del BMA al problema estudiado.

```
library(MASS);
library(splines);
library(survival);
library(leaps);
library(BMA);
datos<-read.table('datos.txt',header=T)
y<-datos$sint
x<-data.frame(datos[,-15])
mod=bic.glm(x,y,glm.family='binomial')
summary(mod)
plot(mod,mfrow=c(3,3))
imageplot.bma(mod)
```

Anexos

En la Figura 4 se puede apreciar la gráfica de las distribuciones *a posteriori* de los coeficientes para las variables que no fueron selecionadas por el criterio BMA para el modelo 1.

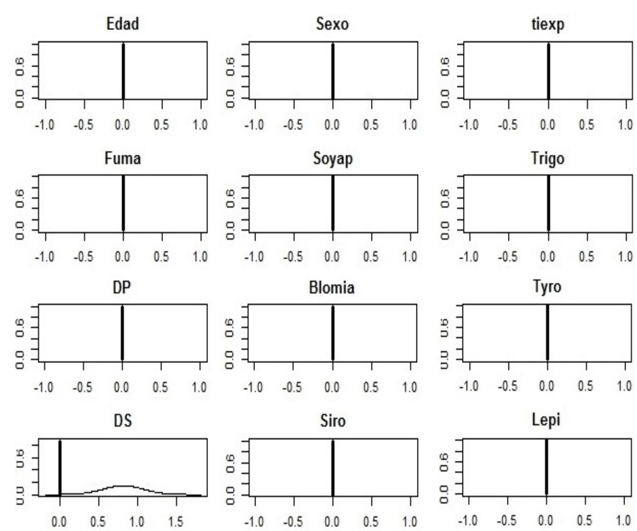

Figura 4. Probabilidad de inclusión de las variables no seleccionadas.

Referencias

[1] Díaz, L., *Selección de modelos en regresión logística binaria bajo paradigma bayesiano.* Tesis de doctorado UH, 2016.

[2] Draper, D., *Assessment and propagation of model uncertainty (with discusssion).* Journal of the Royal Statistical Society, 1995.

[3] Hosmer, D. W., Lemeshow, S. y Sturdivant, R. X. *Applied Logistic Regression*. (3rd edition), John Wiley, 2013.

[4] Jeffreys H, *Theory of Probability*, (3rd edition). Cambridge, MA; New York: Oxford University Press, 1961.

[5] Kass, R. E. y Raftery, A. E. *Bayes Factors*. Journals of the American Statistical Association, 1995.

[6] Leamer, E. E., *Specification Searches*. Wiley, New York, 1978.

[7] Madigan, D. y Raftery, A. E., *Model Selection and accounting for model uncertainty in graphical models using Occamś window.* Journal of the American Statistical Association, 1994.

[8] Montgomery, D., Peck, E. y Vining, G. *Introducción al Análisis de Regresión Lineal*. CECSA (3ra edición), 2006.

[9] Nelder, J. A. y Wedderburn, R. W. *Generalized Linear Models*. Journal of the Royal Statistical Society, 1972.

[10] Raftery, A. E., Madigan, D. y Volinsky, C., *Accounting for model uncertainty in survival analysis improves predictive performance*. Bayesian statistics, 1996.

[11] Sistachs, V. *Un estudio del modelo de regresión logística binario bajo el paradigma bayesiano*. Tesis de doctorado UH, 2005.

[12] Volinsky, C., Madigan, D., Raftery, A. E. y Kronmal, R.A.,*Bayesian model averaging in proportional hazard models: assessing the risk of a stroke.* Applied Statistics, 1997.