Modelos Lineales Generalizados: Un enfoque desde la Estadística Bayesiana

Nerea Pérez Ruiz Jialian Zhou He José Jaén Delgado

Keywords - Regresión Logística, Estadística Bayesiana, Regressión LASSO, Selección de modelos, Convergencia, Predicción, Función vínculo, OpenBugs, Densidad a posteriori, Gibbs Sampling, MCMC

1 Introducción

El área de la Medicina destaca como uno de los principales focos de atención de la predicción estadística, teniendo como objetivo determinar las variables explicativas de diferentes enfermedades para así poder combatirlas eficazmente.

En el presente proyecto estimamos una Regresión Logística empleando tanto la perspectiva Frecuentista como Bayesiana al objeto de analizar rigurosamente la relación entre ataques cardíacos y una serie de regresores que expondremos en la siguiente sección.

2 Descripción de los Datos

Para obtener la información de dolencias del corazón hemos recurrido a una de las páginas web más conocidas dentro de la comunidad Data Science: *Kaggle*. La base de datos con la que trabajaremos cuenta con un total de **918** observaciones, de las cuales aproximadamente un 55% corresponden a personas sin problemas cardíacos, y el 45% restante a pacientes ingresados por afección al corazón.

Por tanto, nuestro estudio se centra en un problema de **predicción de variable binaria**, donde no hace falta recurrir a ténicas de remuestreo tales como el *Random Oversampling* o *Random Undersampling* puesto que los datos están equilibrados, esto es, el número de observaciones de cada categoría no difiere sustancialmente.

La elección del presente tema médico se explica por el hecho de que las enfermedades cardiovasculares son la principal causa de muerte a nivel global (se estiman 17.9 millones de víctimas al año). Aquellas personas en riesgo precisan de una detección rápida y temprana, jugando un papel imprescindible las ténicas cuantitativas de predicción.

A continuación presentamos una breve descripción de las variables con las que trabajaremos para dotar al lector de un mejor contexto:

- Age: Edad del paciente medida en años.
- Sex: Variable binaria indicativa del sexo del paciente (M: Male, F: Female).
- ChestPainType: Tipo de dolor en el pecho (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic).
- RestingBP: Presión sanguínea en reposo medida en mmHg.
- RestingECG: Resultados del electrocardiograma en reposo (Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria).
- Cholesterol: Colesterol sérico medido en mm/dl.
- FastingBS: Nivel de azúcar en ayuno (1: si FastingBS $>$ 120 mg/dl, 0: en caso contrario).
- MaxHR: Pulsaciones máximas (valores numéricos entre 60 y 202).
- Oldpeak: ST depression induced by exercise relative to rest
- ExerciseAngina: Angina inducida por ejercicio (Y: Yes, N: No).
- ST_Slope: Pendiente del pico de ejercico en el segmento ST (Up: upsloping, Flat: flat, Down: downsloping).
- HeartDisease: Variable respuesta que indica si tiene o no problemas cadíacos (1: heart disease, 0: Normal).

La base de datos proviene de combinar observaciones independientes de diferentes países, a saber, Hungría, Estados Unidos y Suiza. Generalizamos así la validez externa de nuestro estudio al tratar con pacientes procedentes de una variedad de Estados, evitando centrarnos demasiado en un punto geográfico concreto.

3 Metodología del Proyecto

A lo largo del presente estudio utilizaremos el software estadístico **R** para llevar a cabo tareas de Data Cleaning, Análisis Exploratorio de Datos y Modelización de la Regresión Logística, al fin de preparar adecuadamente las variables explicativas que servirán de base para la obtención de resultados relevantes. El paquete **R2OpenBUGS** jugará un papel imprescindible a la hora de realizar inferencia bayesiana. Todas las operaciones siguen una lógica estadística que iremos exponiendo a lo largo del proyecto.

Para mejor comprensión de los pasos metodológicos que seguimos recomendamos al lector analizar el código de R adjunto al proyecto, en el cual se exponen resultados que debido a su dimensión no incluimos expresamente en el presente trabajo.

4 Selección del Modelo

Si bien el número de pacientes no es considerablemente alto, la base de datos es especialmente rica en el número de variables explicativas de problemas cardíacos. Consecuentemente, empleamos un método de regularización para seleccionar aquellas variables verdaderamente importantes para nuestro estudio. Proponemos así una *Regresión Lasso* o Regresión Logística con *Regularización L1*, que en la práctica elimina los regresores nimios. Nos centramos en el valor de los coeficientes de dicha Regresión Lasso, sabiendo que cuando éstos no son estadísticamente significativos no se incluirán en el modelo final. Los coeficientes resuelven el siguiente problema de optimización:

$$\hat{w} = \arg\max_{w} \left\{ \ln \prod_{i=1}^{n} \Pr(Y_i = y_i | X_i, w) - \lambda \left|\left|w\right|\right|_1 \right\}$$

A la luz de los resultados obtenidos de la Regresión Lasso concluimos que todas las variables con significativas a excepción de RestingBP y RestingECG. Acto seguido estimamos una regresión logística removiendo dichas variables y analizamos la relevancia del resto de variables, continuando con el proceso de selección.

Es posible apreciar que tanto Age como MaxHR no son significativas, por lo tanto, creamos un nuevo modelo logístico sin estos predictores. Finalmente, se obtiene un modelo en el cual todas las variables son significativas menos el intercepto. Optaremos proceder con este modelo para realizar el resto del trabajo.

5 Predicción Frecuentista y Bayesiana

5.1 Estadística Frecuentista

Estimando la Regresión Logística mediante Máxima Verosimilitud con todas las variables menos las descartadas es fácil adoptar la perspectiva frecuentista, confiando únicamente en los datos disponibles. Maximimizamos la función de log-verosimilitud:

$$\mathcal{L}(w) = \ln \prod_{i=1}^{n} \Pr(Y_i = y_i | X_i, w)$$

La Regresión Logística se caracteriza por incluir una *función vínculo* al modelo de regresión lineal, concretamente denominada función sigmoide, la cual fuerza que los valores ajustados o predicciones finales se encuentren acotados en el intervalo [0, 1]. En el contexto de predicciones en términos probabilísticos este rasgo se presenta del todo razonable, superando considerables limitaciones del Modelo de Probabilidad Lineal.

$$\Pr(Y_i = 1 | X_i, w) = \frac{1}{1 + e^{-w^\intercal h(x_i)}}$$

Ajustando el modelo a los datos podemos predecir que un hombre asintomático con un colesterol de 223, nivel de azúcar en ayuno mayor a 120mg/dl, con una angina inducida por ejercicio, oldpeak de 0.6 y un ST_Slope normal tendrá una probabilidad del 98% de tener problemas cardíacos.

5.2 Estadística Bayesiana

Estimamos una nueva Regresión Logística aplicando la Estadística Bayesiana al fin de comparar los resultados con la inferencia frecuentista. Para ello nos hemos apoyado en el programa OpenBugs, dado que la inferencia generalmente no es conjugada y un muestreo directo requiere de ímprobos esfuerzos.

Configuramos nuestros a prioris de tal manera que sigan una distribución normal con una varianza considerable, no siendo muy informativos ya que como estadísticos no contamos con suficiente conocimiento sobre el tema tratado. No obstante, sería interesante consultar a los médicos y expertos de las enfermedades cardíacas, proponiendo

a prioris más significativos e incluso pudiendo consistir en una mixtura para combinar diferentes opiniones.

Como se ha mencionado anteriormente, la elección de un a priori apropiado no es trivial, en tanto en cuanto no está tan clara la relación de conjunción entre distribuciones. Por ello, nos servimos del algoritmo probabilista *Gibbs sampling*, aplicando métodos Markov Chain Monte Carlo (MCMC).

La primera consecuencia del uso del algoritmo Gibbs Sampling es la obtención del valor de los coeficientes de forma iterativa, siendo necesario el estudio de la convergencia de los mismos. De forma esquemática y muy general, presentamos el funcionamiento del muestreo de Gibbs:

Algorithm 1: Algoritmo de Gibbs Sampling

Input: Datos y variables aletarorias explicativas

Output: Distribuciones a posteriori $f(\beta|\text{datos})$

Inicialización: $x^{(0)} \in \mathbb{R}^D$

for $i = 0$ *to* $n-1$ **do**

$$x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_D^{(i)})$$
$$x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_D^{(i)})$$
$$\vdots$$
$$x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j-1}^{(i)}, \ldots, x_D^{(i)})$$
$$\vdots$$
$$x_D^{(i+1)} \sim p(x_D|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{D-1}^{(i+1)})$$

return $\{\mathrm{x}^{(i)}\}_{i=0}^{N-1}$

end

Tras determinar el valor del *burn-in* adecuado, obtenemos un comportamiento ideal en términos de autocorrelación en la convergencia de los coeficientes. Con el burn-in descartamos varias iteraciones al inicio de la MCMC para garantizar la máxima aleatorización posible, a saber, los valores de los coeficientes en iteraciones próximas no son dependientes entre sí. Todo ello es apreciable en los gráficos que expondremos a continuación, donde ningún patrón concreto puede ser identificado, confirmando la validez de nuestro procedimiento. Además, se puede observar que la esperanza de los predictores dados los datos se estabiliza a partir de las 5,000 iteraciones en todos los casos. Debido a la cantidad de parámetros estimados y la semejanza entre todos ellos presentamos una selección de gráficos de algunos coeficientes.

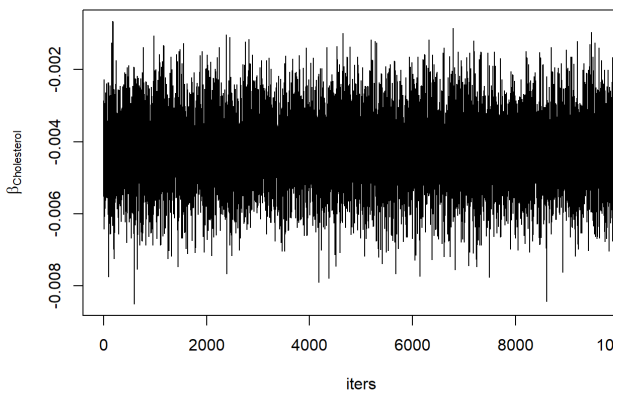


Figure 1: Comportamiento Iteraciones

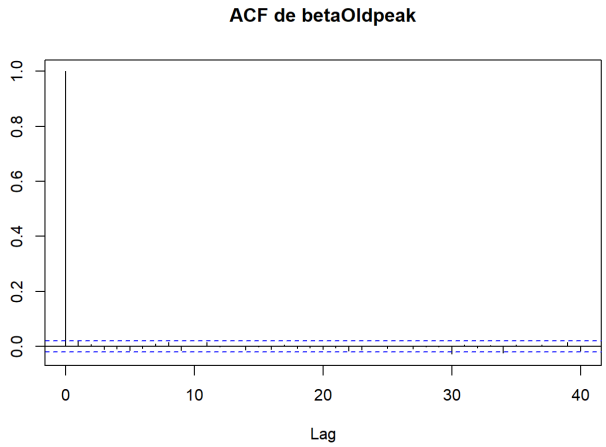


Figure 2: Función de Autocorrelación

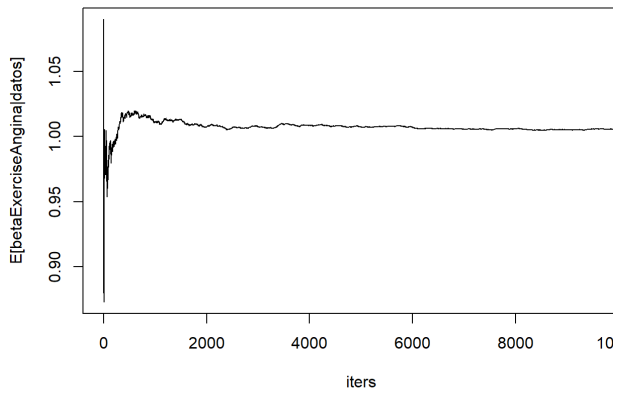


Figure 3: Convergencia para *Exercise Angina*

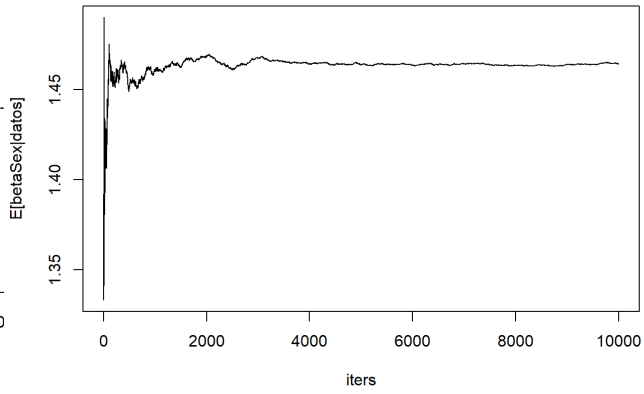


Figure 4: Convergencia para *Sex*

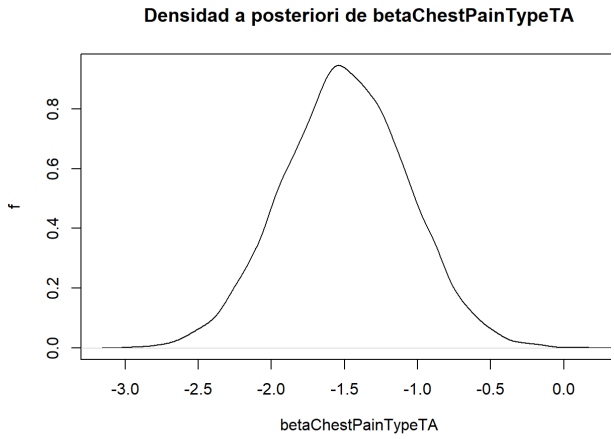
Centrándonos ahora en los resultados de las densidades a posteriori de los predictores, es evidente que se asemejan bastante a las conclusiones de la inferencia frecuentista, tanto en su significancia estadística como en los propios valores. Por ejemplo, de acuerdo con los métodos frecuentistas el intercepto no era significativo, y en la inferencia bayesiana el **intervalo de credibilidad** para el mismo apunta a que hay un 95% de probabilidad de que el cero esté contenido en él. Existe evidencia pues para rechazar que sea significativo.

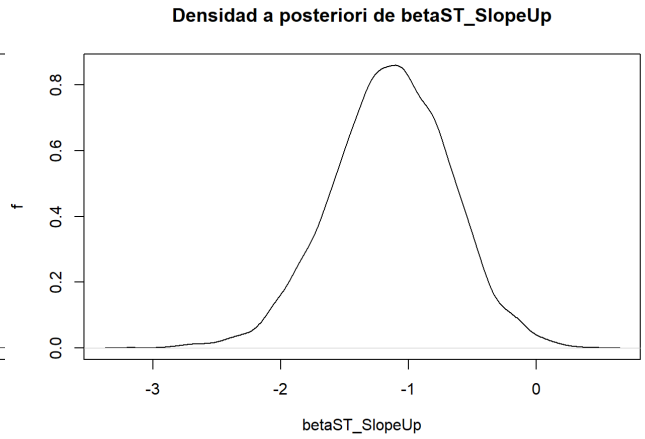
La similitud no acaba en la constante del modelo de regresión, sino que se extiende al resto de coeficientes. Según la Estadística Frecuentista los coeficientes son significativos

al estar contenidos en un **intervalo de confianza** con nivel de significación al 5%. La interpretación de dichos intervalos es cuanto menos intricada, siendo preferible los ya mencionados intervalos de credibilidad.

Acabamos de ver cómo las distribuciones a posteriori afianzan la hipótesis de que todas las variables relevantes obtenidas para el modelo frecuentista también lo son para el bayesiano.

La idea ahora está en conocer de qué manera afectan a la probabilidad de tener problemas cardíacos. Para ello nos fijaremos en el signo que acompaña a cada parámetro de los distintos predictores. Si éste es negativo, un aumento de la variable a la que hace referencia provocará una disminución de la probabilidad del grupo de interés (padecer problemas de corazón). Por el contrario, si aumenta una variable con un parámetro positivo, su probabilidad también lo hará. De esta forma, predictores como la pendiente del pico de ejercicio en el segmento ST para *upsloping*, disminuyen la probabilidad de pertenecer al grupo de interés cuando aumenten o bien presenten esta característica (tener un determinado dolor de pecho por ejemplo, frente a ser asintomático), ya que su coeficiente correspondiente es de signo negativo. Por otro lado, en cuanto a las variables como el sexo, se espera que los hombres sean más proclives a sufrir problemas cardíacos (para dos personas con el resto de predictores iguales), al ser positivo el parámetro que acompaña a este predictor.





De forma inesperada, el predictor colesterol presenta un coeficiente de signo negativo, lo cual resulta contraintuitivo. Desde un punto de vista de la inferencia causal, seguramente se trate de una variable endógena, y por ello correlacionada con el error. Aplicando métodos econométricos sería posible corregir este efecto mediante la

introducción de variables instrumentales. No obstante, no sería posible proceder con el consabido método de Mínimos Cuadrados Bietápicos o *Two-Stage Least Squares* (TSLS), ya que nuestro modelo no es lineal en las betas. Acudiríamos a la estimación por *Generalized Method of Moments* o Método Generalizado de Momentos, pero queda fuera del alcance de este proyecto.

Presentamos a continuación una breve exposición de los resultados a los que hemos estado haciendo alusión a lo largo del proyecto.

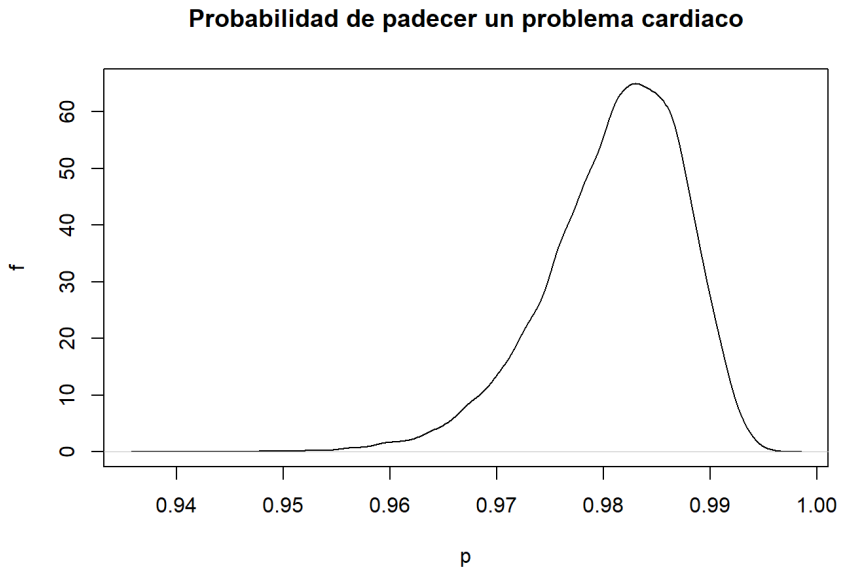
Table 1: Resultados de la Regresión Logística

	Estadística Frecuentista	Estadística Bayesiana
Coeficientes	Heart Disease	Heart Disease
β_0	-0.481	-0.396
	(0.562)	(0.597)
Sex	1.454***	1.464***
	(0.278)	(0.279)
Cholesterol	-0.004***	-0.00426***
	(0.001)	(0.00103)
FastingBS	1.193***	1.217***
	(0.271)	(0.277)
Oldpeak	0.410***	0.414***
	(0.115)	(0.119)
ExerciseAngina	0.991***	1.005***
	(0.235)	(0.238)
ChestPainTypeATA	-1.878***	-1.918***
	(0.322)	(0.327)
ChestPainTypeNAP	-1.706***	-1.743***
	(0.260)	(0.262)
ChestPainTypeTA	-1.458***	-1.492***
	(0.424)	(0.426)
ST_SlopeFlat	1.443***	1.411***
	(0.425)	(0.450)
ST_SlopeUp	-1.060*	-1.141*
	(0.443)	(0.466)

Nota: Variable Dependiente debajo del tipo de inferencia estadística adoptada

Observamos cómo los coeficientes obtenidos con el método bayesiano se asemejan bastante a los del método frecuentista, en buena parte explicado por el hecho de que los a prioris no son informativos.

Finalmente, la predicción bayesiana arroja resultados similares a la frecuentista, como era de esperar. Para las mismas características descritas en la subsección anterior, la probabilidad esperada de padecer un ataque cardíaco es del 98.1%. Adjuntamos un gráfico ilustrativo de la predicción:



6 Conclusión

Los resultados obtenidos con los métodos frecuentista y bayesiano son bastante similares al no dar mucha información los a prioris escogidos. Por simplicidad se podría optar por estimar el modelo de Regresión Logística empleando la inferencia frecuentista, no precisando así del algoritmo del muestreo de Gibbs. Sin embargo, los resultados obtenidos en el método bayesiano son más fáciles de interpretar, empleando intervalos de credibilidad en vez de intervalos de confianza.

Para mayor precisión y rigurosidad sería deseable contar con opiniones de expertos para así introducir a prioris informativos y comparar verdaderamente los resultados de la Estadística Frecuentista y Bayesiana. Asimismo, al objeto de llevar a cabo una correcta inferencia causal, los métodos econométricos contribuyen a superar la ausencia de ortogonalidad de los regresores con el término de error, proponiendo la estimación GMM con variables instrumentales para corregir el signo de predictores como el colesterol.

Apéndice: Diagrama de Causalidad en OpenBugs

Hemos utilizado tanto R como el programa OpenBugs para la realización del trabajo. La modelización en el último software estadístico mediante diagramas de causalidad facilita la interpretación gráfica de las relaciones entre variables y constantes relevantes en el presente estudio. El paquete R2OPENBUGS nos ha permitido una implementación directa en R de los algoritmos necesarios para llevar a cabo correctamente el muestreo de Gibbs, pero con el fin de obtener una visión global e intuitiva de las interconexiones de los componentes básicos de nuestro modelo hemos estimado conveniente recurrir a OpenBugs como ejercicio de completitud del proyecto.