Regresión multivariada para respuestas binarias

Basa, Jerónimo[1]

[1]*Facultad de Ingeniería Química*
Director/a: Llop, Pamela
Codirector/a: Bergesio, Andrea

Área: Ciencias Exactas

Palabras claves: Regresión Multivariada, Familias Exponenciales, Modelos logísticos.

INTRODUCCIÓN

Los métodos de regresión son una pieza fundamental en el análisis estadístico, ya que permiten explicar la relación entre una o varias variables (respuesta) y otro conjunto de variables, que llamaremos predictoras. La versión más sencilla que conocemos, es la regresión lineal simple, donde las variables aleatorias X e Y se relacionan siguiendo el siguiente modelo

$$Y = \beta_0 + \beta_1 \mathrm{X} + \varepsilon,$$

donde ε sigue una distribución $N(0, \sigma^2)$ y es independiente de X . Este modelo supone que

$$E(\mathrm{Y}|\mathrm{X}) = \beta_0 + \beta_1 \mathrm{X}. \quad (1)$$

En este caso, la variable respuesta Y es de tipo continua. Un modelo lineal generalizado es una extensión de esta técnica. La diferencia entre estos dos modelos radica en que en el segundo eliminamos este supuesto sobre la variable respuesta, permitiéndole además ser de otra naturaleza; por ejemplo, de conteo, categóricas o binarias. La importancia de esta generalización se debe a que muchos estudios tienen como respuesta variables binarias codificadas como 0 o 1. Por ejemplo, un estudio médico donde se estudia la probabilidad de tener cierta enfermedad cardíaca en función de la edad de un paciente, se utiliza como dato una variable respuesta binaria que vale 1 si el paciente padece la enfermedad cardíaca y 0 sino. Los modelos generalizados se presentaron por primera vez en (Nelder & Wedderburn, 1972). Cuando la variable respuesta sigue una distribución Bernoulli, al modelo se lo conoce como *Modelo Logístico*. La forma en que se deduce es partiendo de la función de probabilidad puntual de la distribución Bernoulli con parámetro π, la cual pertenece a la familia exponencial

$$f(y;\pi) = (1-\pi) exp\left\{ylog\left(\frac{\pi}{1-\pi}\right)\right\}. \quad (2)$$

Recordemos que en este caso, $\pi = P(Y = 1) = E(Y)$. En el caso de modelos logísticos, en lugar de modelar linealmente la esperanza $E(\mathrm{Y}|\mathrm{X})$ como en (1), lo que hacemos es modelar linealmente una función de la misma utilizando lo que se conoce como *función link* g. Más precisamente, de (2) se sigue que la función link para la distribución Bernoulli es

$$g(\pi) = log\left(\frac{\pi}{1-\pi}\right),$$

por lo que, el modelo de regresión logística resultante es

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x, \tag{3}$$

donde $\pi(x) = P(Y = 1|X = x) = E(Y|X = x)$. De este modelo se sigue que la probabilidad de éxito puede ser modelada, a partir de la función logística (y de allí el nombre del modelo) como

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{4}$$

Si bien estos modelos presentan una sola variable predictora, existen estudios que incluyen más de una, esto se conoce como *regresión múltiple*. Por otro lado, cuando aumenta el número de variables respuesta, decimos que se trata de *regresión multivariada*. En particular, el modelo Bernoulli multivariado para respuesta múltiple binaria, presenta una importante complejidad gráfica y computacional. Una forma de tratar dicho problema es considerando el *Modelo Ising*, presentado originalmente en (Ising, 1925). Este es una simplificación que considera sólo interacciones de a pares entre las variables respuestas, despreciando las posibles interacciones de orden superior. Esto hace que su interpretación sea sencilla, motivo por el cual quizás sea ampliamente usado en gran parte de la literatura estadística moderna y uno de los más estudiados. El enfoque principal de Ising es estudiar la posible dependencia condicional entre dos variables respuestas cuando se fijan las demás.

OBJETIVOS

- Estudiar los modelos lineales generalizados y sus propiedades.
- Realizar estudios de dichos modelos para conjuntos de datos reales.
- Comparar distintos enfoques del modelo y obtener conclusiones sobre su calidad para predecir datos.
- Estudiar el modelo Ising y sus propiedades teóricas.
- Estudiar sus aplicaciones y algoritmos existentes para su cómputo.
- Analizar, vía estudios de simulación, dichos algoritmos.

Título del proyecto: Métodos estadísticos para datos funcionales o de alta dimensión
Instrumento: CAI+D
Año convocatoria: 2016
Organismo financiador: UNL
Director/a: Forzani, Liliana

METODOLOGÍA

La metodología de trabajo consistió de una etapa inicial de formación básica donde se estudiaron los conceptos teóricos, sus alcances y limitaciones, siguiendo las ideas en (Agresti, 2013). Luego se prosiguió a corroborar computacionalmente los resultados estudiados. En esta etapa, utilizando el *software R*, se programaron los códigos necesarios para ajustar el modelo (3) y luego, mediante estudios de simulación, fueron comparados con la función *glm* del *software* R (R Core Team, 2017). Una vez verificada la validez del modelo, se prosiguió la

investigación usando datos reales. Para ello se estudiaron varios problemas: la probabilidad de que un paciente tenga cierta enfermedad cardíaca de acuerdo a la edad del mismo; la dependencia del ancho de los caparazones en cangrejos hembras y su capacidad de tener crías; análisis del accidente de la misión espacial *Challenger* dirigido por la NASA; desarrollo de un nuevo método de enseñanza en estudiantes de economía, entre otros.
En el contexto de modelos con múltiples variables predictoras, se realizaron experimentos computacionales para reducir las dimensiones del problema logrando, luego de la reducción, un modelo más sencillo con menos parámetros a estimar. El primer ejemplo analizado fue un estudio sobre crías de cangrejos hembra, donde a partir de 173 observaciones se estudió la probabilidad de que una hembra tenga un macho para su cría a partir del ancho (variable predictora continua, X) y color de su caparazón (variable predictora categórica con cuatro niveles, C).

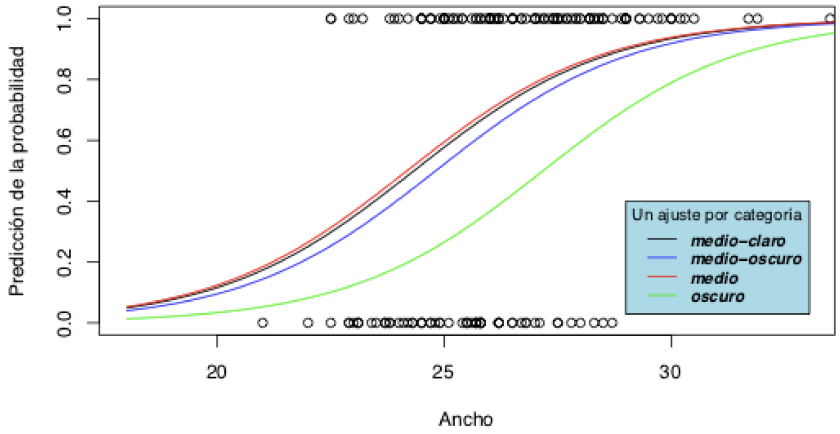

Figura 1. Ejemplo de un estudio sobre las crías en cangrejos hembra a partir del tipo de caparazón. Lo que se observa son cuatro curvas (una por cada tipo de caparazón) que representan el ajuste del modelo logístico sobre el gráfico de dispersión.

La Figura 1 muestra las probabilidades estimadas $\pi(x,c) = P(Y = 1|X = x, C = c)$ para diferentes valores del ancho del caparazón x, y para los cuatro niveles de la variable color del caparazón c. Siguiendo el modelo (4) tenemos que

$$\pi(x,c) = \frac{e^{\beta_0+\beta_1 x+\beta_2 c}}{1+e^{\beta_0+\beta_1 x+\beta_2 c}}.$$

Como puede observarse, la curva verde (caparazón oscuro) se aleja del resto, dando como resultado que, para diferentes anchos, el color si afecta a la probabilidad de tener crías.
En el caso de regresión multivariada, comenzando con el caso bivariado, se utilizaron dos modelos diferentes y, mediante estudios de simulación, se comprobó que ambos modelos conducían a las mismas estimaciones de los parámetros de interés. Uno de ellos fue a partir de las probabilidades condicionales usando los datos, y el otro por medio del paquete MVB de R. Siguiendo con el caso de más de dos variables respuesta, la estrategia fue utilizar los algoritmos del modelo Ising para comparar los efectos del aumento de las dimensiones de los datos y estudiar la posible dependencia entre las variables. Esto se hizo a través del paquete IsingFit.

CONCLUSIONES

En los experimentos realizados con datos reales, se encontró que el modelo logístico es un modelo adecuado para la representación y predicción futura de los datos. Además, los métodos numéricos de Newton-Raphson dieron buenas aproximaciones de los estimadores del modelo. Para los casos de datos que presentan observaciones dentro de un conjunto reducido de valores posibles, se encontró que es posible agrupar los datos en bloques, disminuyendo los grados de libertad del problema, resultando en estimaciones idénticas a la del modelo desagrupado. En la Figura 2 vemos el ejemplo sobre estimar la variable respuesta *Probabilidad de una enfermedad cardíaca* a partir de la variable predictora *Edad* en donde los 100 datos recolectados para el estudio se dividieron en 8 grupos. En el caso de variables categóricas, se concluyó que no podemos analizar sus categorías de manera individual ya que se pueden obtener conclusiones erróneas (como el ejemplo de los cangrejos detallado en la Sección metodología). En estos casos, lo que hacemos es analizar el impacto de la variable completa dentro del modelo. Lo mismo ocurre al considerar las interacciones entre variables categóricas.

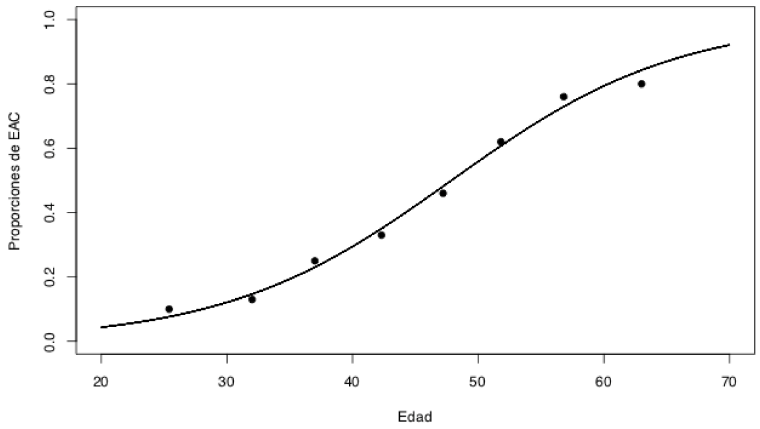

Figura 2. Ejemplo gráfico de la probabilidad de Enfermedad de Arteria Coronaria (EAC) en función de la edad donde se agrupan los datos por bloques. La curva de ajuste obtenida al resolver el modelo logístico sigue la tendencia de los datos.

Para el caso bivariado, se encontró un modelo conjunto, detallado en (Islam et al, 2013), que fue estudiado mediante simulación, encontrándose buenas aproximaciones de los estimadores para tamaños de muestra cercanos a 500. En el caso de modelos sin variables predictoras, se encontró una ventaja para la estimación de los parámetros, pudiendo estos ser ajustados usando solo las probabilidades condicionales calculadas a partir de los datos. Comparado con lo obtenido al usar la función MVB de R, las estimaciones son las mismas.

BIBLIOGRAFÍA BÁSICA

Agresti, A., 2013. *Categorical data analysis*. Wiley Series in Probability and Statistics. Wiley.

Ising, E., 1925. *Beitrag zur theorie des ferromagnetismus*. Zeitschrift fur Physik, 31:253-258.

Islam, M. A., Alzaid, A. A., Chowdhury, R. I., 2013. *A generalized bivariate Bernoulli model with covariate dependence*. Journal of Applied Statistics, 40(5):1064-1075.

Nelder, J. A., Wedderburn, R. W. M., 1972. *Generalized linear models*. Journal of the Royal Statistical Society, 135(3):370-384.

R Core Team 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.