Capítulo 30

Logit model. De Verhulst (1838) a Mcfadden (2001)

ELENA MARTINEZ RODRIGUEZ
Universidad Complutense de Madrid

Introducción

Definimos la ecuación ó función logística mediante la expresión:

$$P(Y) = \frac{\exp^{Y}}{1+\exp^{Y}}, \tag{1}$$

en la que habitualmente P denota una función de probabilidad, e Y indica una combinación lineal del tipo $Y = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki}$

Así definida, la función logística cumple las propiedades de ser una función monótona creciente, acotada en el intervalo $[0,1]$. Su representación gráfica (figura 1) es una curva de forma sinusoidal

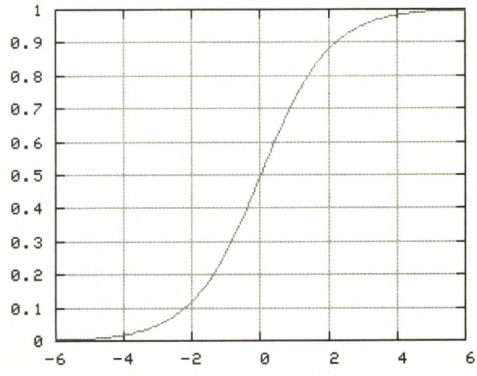

Figura 1: Función Logística

en la que observamos que para valores pequeños de la variable Y la función experimenta un crecimiento lento, que aumenta rápidamente a medida que aumenta la variable (en este tramo se asemeja al crecimiento exponencial), y, finalmente, se ralentiza para valores altos de *Y*, alcanzado su cota máxima situada en el valor 1.

El significado de esta ecuación depende de la definición de las variables. Por ejemplo, en Biología es frecuente que *P* sea la función de probabilidad del suceso dicotómico supervivencia o muerte del organismo observado cuando es sometido a un estímulo continuo, expresado éste mediante un modelo lineal (*Y*)

Siglo XIX: La Ecuación de Verhulst

Según Thompson, el término "curva logística" debe atribuirse a Edward Wright (1558-1615), quien usó este término para referirse a una curva o ecuación logarítmica.

No obstante debemos esperar hasta el siglo XIX para que la función logit o ecuación logística se desarrollara tal y como la conocemos hoy en día. Esta función se aplicó de forma independiente en dos ámbitos bien diferenciados: en química, para explicar reacciones autocatalíticas y en demografía, para explicar el crecimiento de poblaciones. Hay que destacar que la trayectoria en ambas disciplinas fue distinta: su uso en química ha sido ininterrumpido, mientras que en demografía hay un periodo prolongado en el que cae en un injustificado olvido.

Una reacción química es clasificada como autocatalítica o en cadena cuando el producto en sí mismo es el catalizador de la reacción. Son reacciones que se caracterizan porque al principio proceden lentamente debido a la escasa presencia del elemento catalizador, aunque va aumentando progresivamente para, al final, sufrir un retroceso al disminuir la concentración del reactivo. La función logística, tal como hemos visto en el apartado anterior, describe la evolución temporal de este tipo de reacciones. Prueba de ello son los trabajos, fechados en 1883, del químico alemán W. Ostwald.

El otro campo en el que se utilizó la ecuación logística para describir un fenómeno fue la demografía. En concreto, la ecuación logística de crecimiento de una población fue propuesta por P. Verhulst como una posible solución al dilema del crecimiento exponencial de Malthus.

A finales del siglo XVIII el economista inglés Thomas Robert Malthus, publicó la obra *Ensayo sobre el principio de la población* (1798), en la que expuso y defendió sus teorías sobre el crecimiento demográfico, según las cuales la población humana tiende a crecer en progresión geométrica mientras que los medios de subsistencia lo hacen en progresión aritmética, principio que hoy conocemos como "Ley exponencial de crecimiento poblacional". Este modelo podemos representarlo por la siguiente expresión:

$$W(t) = W(0)e^{\beta t}, \qquad (2)$$

siendo $W(t)$ el tamaño de la población en el instante t, $W(0)$ el tamaño inicial y β la tasa constante de crecimiento.

Alfhonse Quetelet, matemático y astrónomo belga (1795-1874), fue uno de los primeros en considerar que el modelo exponencial de crecimiento no era adecuado para explicar la expansión demográfica de un país. Si bien podía reflejar la situación de Estados Unidos a principios del siglo XIX (modelo poblacional observado por Malthus), un país joven y casi vacío, su aplicación a otras sociedades conduciría a valores imposibles. Quetelet estaba

convencido de que una población no podía crecer indefinidamente, sino que existían fuerzas, tanto externas como internas, que tienden a prevenir este crecimiento. Aunque sus aportaciones en dinámica de la población no son destacables, la referencia a este autor es obligada porque fue profesor, en la Universidad de Gante, de Pierre-François Verhulst (1804-1849) con quien trabajó durante un largo periodo y sobre quien tuvo una gran influencia tanto en su vida personal como en su obra. Quetelet fue pionero en la aplicación de la Estadística y la Teoría de la Probabilidad para explicar las regularidades sociales (concepto del "hombre medio"). Su preocupación por el orden social y el papel que le atribuía a la ciencia como un instrumento fundamental para su control, son determinantes en la formación y en la trayectoria académica y científica de su discípulo.

Verhulst abordó el problema del crecimiento de una población adoptando las hipótesis de Quetelet y, por tanto, consideró que es un proceso limitado. Demostró que la tasa de crecimiento de una población está limitada directamente por su propia densidad. Por este motivo añadió al modelo propuesto por Malthus un término adicional que representa la resistencia al crecimiento:

$$W(t) = \beta W(t) - \phi(W(t)). \tag{3}$$

En concreto, $\phi(W(t))$ debía representar las fuerzas que frenan el crecimiento poblacional y que ambos autores consideraban que aumentan con el cuadrado de la tasa de variación de la población. Esta idea se puede expresar mediante la ecuación:

$$W(t) = \beta W(t)(\Omega - W(t)), \tag{4}$$

donde Ω denota el límite máximo o nivel de saturación de la magnitud tamaño poblacional *W*. Observamos en (4) que el crecimiento es proporcional tanto al tamaño poblacional en el momento t, como al crecimiento potencial de la población $(\Omega - W(t))$.

Si expresamos *W (t)* como proporción, tendremos:

$$P(t) = \frac{W(t)}{\Omega} = \beta P(t)[1 - P(t)]. \tag{5}$$

La solución a esta ecuación diferencial no lineal la encontramos en la expresión (6), a la que Verhulst llamó función logística:

$$P(t) = \frac{\exp^{(\alpha+\beta t)}}{1+\exp^{(\alpha+\beta t)}}. \tag{6}$$

De esta forma, el tamaño de una población en el instante *t* se calcularía como:

$$W(t) = \Omega \frac{\exp^{(\alpha+\beta t)}}{1+\exp^{(\alpha+\beta t)}}, \tag{7}$$

ecuación que en Biología se denomina *ecuación de Verhulst,* en honor a su autor, y que describe un modelo de crecimiento autolimitado de una población, conocido también como modelo logístico de crecimiento poblacional.

Como veremos a continuación, Verhulst aplicó su ecuación para calcular el límite superior de la población de algunos países europeos, entre ellos Bélgica, situando el nivel de saturación en 9,5 millones de habitantes. La población belga en 1994 era de 10,118 millones

de habitantes; si descontamos el efecto de la inmigración (no contemplado en el modelo (7)) parece que la predicción de Verhulst es bastante acertada.

Los resultados de sus investigaciones sobre el crecimiento demográfico vieron la luz a través de varios artículos. En 1838 publica en *Correspondance Mathématique et Physique* (editado por A. Quetelet) «*Notice sur la loi que la population suit dans son accroissement*», en el que expone la esencia de su teoría y muestra cómo su modelo describe fielmente el crecimiento de las poblaciones de Bélgica, Rusia y Francia, utilizando para ello datos anteriores a 1833. No obstante, en este artículo no menciona cómo deduce la ecuación de crecimiento ni tampoco se refiere a ella como función logística.

En 1845 publica, en *Nouveaux Mémoires de l'Académie Royale des Sciences de Belgique*, un segundo artículo en el que por primera vez introduce el término "logístico" para referirse a la ecuación de crecimiento poblacional (7), proporcionando más detalles sobre sus propiedades, incluso llega a representar la curva logística junto con la curva exponencial. A nivel aplicado, este artículo es el más completo ya que estima los parámetros α y β de la ecuación (7)[1], lo que le permite realizar una predicción para el tamaño máximo de la población (Ω) de Bélgica y de Francia. Para Bélgica estima ese límite máximo en 6,6 millones y para Francia en 40 millones.

El último artículo referido al modelo de crecimiento poblacional lo publica en 1849, también en *Nouveaux Mémoires de l'Académie Royale des Sciences de Belgique*. El objetivo de este artículo es publicar las notables correcciones en las estimaciones de los parámetros del modelo, que permiten una nueva predicción del nivel máximo de la población en el caso de Bélgica, situándolo en 9,5 millones.

Tras la muerte de Verhulst (1849), J. B. Liagre, amigo del matemático, repite, utilizando datos más actualizados, la estimación del límite máximo de crecimiento de la población de Bélgica, que recoge en la segunda edición de su libro *Calcul des probabilités et théorie des erreurs* (1879), llegando, en este caso, a la cifra de13,7 millones. Con la excepción de este trabajo, la obra de Verhulst permanece casi 70 años en el olvido.

Primera mitad del siglo XX: Redescubrimiento y desarrollo de la función Logit

La primera vez, a excepción del libro de Liagre, que se menciona la obra de Verhulst es en 1918, tras más de 60 años de silencio. En este año, el matemático francés Gustave Du Pasquier publicó un artículo en el que se limitaba a exponer distintas teorías matemáticas formuladas sobre el comportamiento poblacional. En concreto, las teorías de Halley, Moivre, Euler y Verhulst. El artículo no tuvo repercusión.

En 1920 el biólogo americano Raymond Pearl, en colaboración con el matemático L. J. Reed, redescubre la función logística como un simple modelo de crecimiento demográfico. Hablamos de redescubrimiento porque, a diferencia de Du Pasquier que tan sólo la menciona, Pearl la aplica en varios de sus trabajos.

Raymond Pearl es sin duda uno de los biólogos más importantes del siglo XX, a la vez que uno de los científicos más prolíferos de esta época, como atestiguan sus 712 artículos y más de 17 libros, escritos sobre temas tan diversos como genética, alcohol, tabaco, fertilidad, longevidad, alimentos y precios, estadística o crecimiento de la población, entre otros

[1] Para la estimación de los parámetros, tanto para el caso de la población belga como de la francesa, considera únicamente valores conocidos de la población en tres años distintos.

muchos. Después de su trabajo doctoral, Pearl comenzó a investigar sobre la aplicación de los métodos estadísticos en biología, en colaboración con Karl Pearson. Ambos se asociaron en 1906 para editar la revista *Biométrica*, asociación que duró hasta 1910. A partir de 1919 su interés se centra en temas sobre crecimiento poblacional, llegando a proyectar el crecimiento de la población de Estados Unidos hasta 1940. Para realizar esta proyección, Pearl y sus colaboradores se basan en datos de la población de este país registrados desde 1790 a 1910, y utilizan la ecuación (7) de crecimiento poblacional de Verhulst, en la que, al igual que el matemático belga, estiman los parámetros a partir de los datos poblacionales conocidos en tres periodos.

J. S. Cramer defiende, en el capítulo 9 de su libro *Logit Models. From Economiccs and Other Fields,* que Pearl y sus colaboradores llegaron de forma independiente a la formulación de la ecuación logit recogida en la expresión (7), y que utilizan en varias de sus publicaciones, ya que, aunque conocían la aplicación de la curva logística para explicar la reacciones químicas en cadena, en 1920 (primera fecha de publicación de una serie de artículos fundamentados en el modelo logístico de crecimiento) no conocían el modelo propuesto por Verhulst. En el siguiente trabajo publicado en 1922 aparece la primera y escueta referencia, como pie de página, al matemático Verhulst. Referencia que se repite, con algo más de detalle, en la publicación de 1923.

En 1925 encontramos dos aportaciones destacadas al desarrollo de la función logística. Por una parte, el estadístico G. U. Yule, que conoce la aportación de Verhulst gracias a los trabajos de Pearl, dedica una parte del artículo publicado en este año a comentar el modelo de crecimiento de Verhulst, en concreto, uno de los apéndices. Es también Yule quien reestablece el término de función logística, que no aparece en ninguno de los papeles de Pearl y Reed ni en el artículo de Du Pasquierer. Por otra parte, el biólogo Alfred J. Lotka obtuvo en el ámbito de la ecología la ecuación logística para explicar el crecimiento demográfico de una comunidad. Este autor la denomina "ley del crecimiento poblacional".

El término "Logit Model" fue acuñado por Joseph Berkson en 1944. La formación inicial de este autor es en Ciencias Físicas, disciplina en la que se familiariza con la función logística, publicando, junto con Reed (1929), un artículo referente a las funciones que modelan las reacciones autocatalíticas. Posteriormente se interesó por la Estadística, en concreto, por la aplicación de los métodos estadísticos en el campo de la Biología (Biometría). Quizá una de sus aportaciones más conocidas sea el llamado *modelo del error de Berkson*, en el que considera, en contraposición al modelo clásico de regresión, que el error es independiente de la variable observada. Igualmente se le reconoce como un de los autores que más utilizaron la función logística, usando por primera vez la denominación de Logit Model, en el artículo *Application to the Logistic Function to Bio-assay*. En este artículo el autor explica que usa el término Logit para la expresión

$$\log\left(\frac{p_i}{1-p_i}\right) = \log it(P), \tag{8}$$

siguiendo a Ittner Bliss (1934), quien llamó Probit a una función análoga, aunque lineal en variable independiente, cuando asume como función de probabilidad la curva Normal en lugar de una función logística. Podemos considerar, por tanto, que es Berkson quien introduce el uso del Modelo Logit en la Estadística, aunque la generalización de su uso no fue ni fácil ni rápida.

Segunda mitad del siglo XX: Modelos de elección discreta de McFadden

La etapa de desarrollo del Modelo Logit y su aplicación en distintos campos no es sencilla. Existe, desde su introducción por Berkson, una "rivalidad" entre los partidarios de la utilización de Modelos Logit y de Modelos Probit, tal y como se puede comprobar en muchas publicaciones, sobre todo durante las décadas de los 60 y 70. A esta rivalidad contribuye el propio Berkson, mediante una serie de artículos que abarcan desde 1944 hasta 1980. Los partidarios del Modelo Logit argumentan la flexibilidad en su interpretación, ya que ésta depende de la definición de las variables, y su simplicidad de cálculo y aplicación, que se ve reforzada en los 80 con la revolución informática. Sus detractores, en cambio, utilizan esta falta de interpretación específica para designar al Modelo Logit como una simple herramienta estadística, llegando a afirmar que carece de base teórica sólida para su aplicación, ya que, a diferencia del Modelo Probit, no tiene asociada una distribución de probabilidad reconocida (Aitchis y Brown, 1957).

La utilización del Logit Model en distintas ciencias ocurre a un ritmo desigual e, incluso, de forma independiente. Los primeros avances de este modelo ocurrieron en Estadística y en Epidemiología.

En Epidemiología los estudios de casos y controles, basados en ratios de probabilidad, propician la rápida incorporación del modelo logit. De hecho, ya en los años 50, J. Cornfield utilizó la regresión logística para el cálculo de los *odds ratio* como valores aproximados del riesgo relativo.

En estadística el principal difusor de la regresión logística fue D. R. Cox con la publicación en 1970 de su libro *The Análisis of Binaria Data.* La importancia de los trabajos de este autor reside tanto en la utilización que hace del Modelo Logit para formalizar fenómenos binarios (sólo dos posibles concreciones), en la difusión del modelo en aplicaciones econométricas, como en la introducción de transformaciones del modelo, como por ejemplo la creación del Modelo Logit Mixto, como un modelo alternativo que puede situarse entre el Logit (mantiene su simplicidad) y el Probit (comparte su flexibilidad).

El empuje definitivo para el reconocimiento de Logit Model lo encontramos en los trabajos del economista americano McFadden, quien vincula este modelo a la Teoría de la Elección Discreta, abriendo un nuevo campo de trabajo que le hizo merecedor del Premio Nóbel de Economía en el año 2000. Los trabajos de McFadden se remontan a 1973, cuando trabajaba en California como consultor en un proyecto público en materia de transporte. Este economista fue pionero en usar el modelo logit para representar las preferencias de los individuos.

La importancia de los modelos de elección discreta radica en que permiten la modelización de variables cualitativas, característica que exige la codificación de la variable como paso previo a la modelización. En este proceso, los distintos estados de la variable se transforman en códigos o valores susceptibles de ser tratados utilizando técnicas de regresión.

McFadden planteó inicialmente el caso en el que los individuos se enfrentan a procesos de decisión dicotómicos, es decir, en los que únicamente hay dos posibles alternativas que representa de la forma y = 1 o y = 0.

El argumento principal de su tesis era que cada individuo tiene una función de utilidad U_i asociada a cada una de las alternativas (y=0; y=1). Esta función de utilidad puede dividirse en una componente sistemática V_i, que recoge el efecto de las variables explicativas (atributos observables), y una componente aleatoria ε_i, que recoge los efectos que tanto las variables no

relevantes individualmente como del azar pueden tener sobre la utilidad. Según esto, la función de utilidad de un individuo se representa por la expresión

$$U_i = V_i + \varepsilon_i . \tag{9}$$

Si aceptamos la hipótesis de linealidad para la componente sistemática, la función de utilidad para cada estado posible de la elección se formula en los siguientes términos:

$$\begin{array}{l} \text{para } y = 0 \rightarrow U_0 = \alpha_0 + \beta_0 X + \varepsilon_0 \\ \text{para } y = 1 \rightarrow U_1 = \alpha_1 + \beta_1 X + \varepsilon_1 \end{array}, \tag{10}$$

donde α y β son parámetros y ε_0 y ε_1 se supone son independientes e idénticamente distribuidos (i.i.d.), según el modelo de valores extremo tipo I de Gumbel.

Si el comportamiento del individuo obedece al principio económico de maximización, elegirá la alternativa que le proporcione la máxima utilidad. En esta situación, la probabilidad de que el individuo elija la alternativa representada como y=1 será:

$$\begin{array}{c} P(y=1) = P(U_0 < U_1) = P\left(\varepsilon_0 - \varepsilon_1 < (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)X\right) = \\ = F\left[(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)X\right] \end{array}, \tag{11}$$

siendo F la función de distribución.

Llegado este punto, el problema se centra en el modelo de probabilidad de F. McFadden demostró que en las condiciones descritas (ε_1 y ε_0 i.i.d.) la variable diferencia de términos error $(\varepsilon_0 - \varepsilon_1)$ sigue un modelo de distribución de probabilidad logística.

Las innovaciones de McFadden proporcionaron al Modelo Logit una sólida base teórica, al tiempo que le dotó de un significado propio como modelo de elección discreta.

Por último, señalar que los últimos avances en el campo de la regresión tienden a unificar la teoría existente en cuanto a Modelos Probit y Logit con los modelos lineales generalizados basados en la distribución Normal y el análisis de la varianza. La tendencia en el siglo XXI no es la competición o exclusión entre modelos de regresión, sino una búsqueda del modelo más adecuado a las características del fenómeno observado, sean cuantitativas o cualitativas, proponiendo, incluso, una mixtura entre ellos.

Bibliografía

AITCHISON, J.; BROWN, J. A. (1957). *The lognormal Distribution.* Number 5 in University of Cambridge, Department of Applied Economics Monographs. Cambridge University Press, Cambridge.

BERKSON, J. (1944). "Application to the Logistic Function to Bio-assay". Journal of The American Statistical Association, 39, 357-365.

BLISS, C. I. (1934). "The method of probit". Science, 79, 38-39.

CRAMER, J. S. (2002). *Logit Models. From Economics and Other Fields.* Cambridge University Press, Cambridge.

CRAMER, J. S. (2003). *The Origins of Logistic regresion.* Tinbergn Institute Working Paper, nº 2002, 119/4

CORNFIELD, J. (1951). "A method of estimating comparative rates from clinical data". Journal of the National Cancer Institute, 11, 1269-1275.

COX, D. R. (1958). "The regresion analysis of binary sequences". Journal of the Royal Statistical Society, Series B, 20, 215-242.

COX, D. R. (1958). *The Analysis of Binaria Data.* Chapman and Hall. London

DU PASQUIER, L.-G. (1918). "Esquisse d'une nouvelle théorie de la population". Vierteljahrssshrift der Naturforschenden Gesellschaft, 63, 236-249.

KINGSLAND, S. E. (1985). *Modeling Nature.* The University of Chicago Press, Chicago.

LANDAV, D.; M LAZARSFELD, P. F. (1976). "Quetelet, Adolphe". Enciclopedia Internacional de las Ciencias Sociales. Aguilar. Madrid.

LIAGRE, J. B. (1879). *Calcul des probabilités et théorie des erreurs.* Bruselas: Muquardt. 2ª edición.

LOTKA, A. J. (1925). *Elementos de la Biología Física.* Williams y Wilkins publicaciones, Baltimore.

MALTHUS, T. E. (1798). *An Essay on the Principle of Population.* London.

MCFADDEN, D. (1974). "Condicional Logit análisis of Qualitative Choice Behavior". En Zerenbka (ed.). *Fronties in Econometrics.* New York.

MCFADDEN, D. (2001). "Economic choice". American Economic Review, 91, 352-357. (Discurso de aceptación del Premio Nobel)

PEARL, R. L.; REED, L. J. (1920). "On the rate of growth of the population of the United status since 1870 and its mathematical representation". Proceedings of the National Academy of Sciences, 6, 275-288.

PEARL R. (1925). *La biología del crecimiento de la población.* Knopf, New York.

REED, L. J.; BERKSON, J. (1929). "The application of the logistic function to experimental data". Journal of Physical Chemistry, 33, 760-779.

STIGLER, S. M. (1986). *The History of Statistics.* Cambridge, Mass. Harvard University Press.

VERHULST, P-F. (1838). *Notice sur la loi que la population suit dans son accroissement. Correspondance Mathématique et Physique.* Publicado por A. Quetelet, 10, 113.

VERHULST, P-F. (1845). "La Loi d'Accoissemrnt de la Population". Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-lettres de Belgique, 18, 1-59.

VERHULST, P-F. (1847). "La Loi d'Accoissemrnt de la Population". Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-lettres de Belgique, 20, 1-32.

YULE, G. U. (1925). "The growth of population and the factors which control". Journal of Royal Statistical Society, 138, 1-59.