LA REGRESION LOGISTICA (I)

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Enero 2001

Regresión logística (II)

Introducción

No cabe ninguna duda que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización.

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

De todos es sabido que este tipo de situaciones se aborda mediante técnicas de regresión. Sin embargo, la metodología de la regresión lineal no es aplicable ya que ahora la variable respuesta sólo presenta dos valores (nos centraremos en el caso dicotómico), como puede ser presencia/ausencia de hipertensión.

Si clasificamos el valor de la variable respuesta como 0 cuando no se presenta el suceso (ausencia de hipertensión) y con el valor 1 cuando sí está presente (paciente hipertenso), y buscamos cuantificar la posible relación entre la presencia de hipertensión y, por ejemplo, la cantidad media de sal consumida al día como posible factor de riesgo, podríamos caer en la tentación de utilizar una regresión lineal:

$$Hipertensión = a + b \cdot [Consumo\_sal]$$

y estimar, a partir de nuestros datos, por el procedimiento habitual de mínimos cuadrados, los coeficientes *a* y *b* de la ecuación. Sin embargo, y aunque esto es posible matemáticamente, nos conduce a la obtención de resultados absurdos, ya que cuando se calcule la función obtenida para diferentes valores de consumo de sal se obtendrá resultados que, en general, serán diferentes de 0 y 1, los únicos realmente posibles en este caso, ya que esa restricción no se impone en la regresión lineal, en la que la respuesta puede en principio tomar cualquier valor.

Si utilizamos cómo variable dependiente la probabilidad ***p*** de que un paciente padezca hipertensión y construimos la siguiente función:

$$\ln \frac{p}{1-p}$$

ahora sí tenemos una variable que puede tomar cualquier valor, por lo que podemos plantearnos el buscar para ella una ecuación de regresión tradicional:

$$\ln \frac{p}{1-p} = a + b \cdot [consumo\_sal]$$

que se puede convertir con una pequeña manipulación algebraica en

$$\Pr.HTA = \frac{1}{1+e^{(-a-b\cdot[consumo\_sal])}}$$

Y este es precisamente el tipo de ecuación que se conoce como modelo logístico, donde el número de factores puede ser más de uno, así en el exponente que figura en el denominador de la ecuación podríamos tener:

$$b1.\text{consumo\_sal} + b2.\text{edad} + b3.\text{sexo} + b4.\text{fumador}$$

Los coeficientes del modelo logístico como cuantificadores de riesgo

Una de las características que hacen tan interesante la regresión logística es la relación que éstos guardan con un parámetro de cuantificación de riesgo conocido en la literatura como "**odds ratio**" (aunque puede tener traducción al castellano, renunciamos a ello para evitar confusión ya que siempre se utiliza la terminología inglesa).

El odds asociado a un suceso es el cociente entre la probabilidad de que ocurra frente a la probabilidad de que no ocurra:

$$odds = \frac{p}{1-p}$$

siendo ***p*** la probabilidad del suceso. Así, por ejemplo, podemos calcular el odds de presencia de hipertensión cuando el consumo diario de sal es igual o superior a una cierta cantidad, que en realidad determina cuántas veces es más probable que haya hipertensión a que no la haya en esa situación. Igualmente podríamos calcular el odds de presencia de hipertensión cuando el consumo de sal es inferior a esa cantidad. Si dividimos el primer odds entre el segundo, hemos calculado un cociente de odds, esto es un odds ratio, que de alguna manera cuantifica cuánto más probable es la aparición de hipertensión cuando se consume mucha sal (primer odds) respecto a cuando se consume poca. La noción que se está midiendo es parecida a la que encontramos en lo que se denomina **riesgo relativo** que corresponde al cociente de la probabilidad de que aparezca un suceso (hipertensión) cuando está presente el factor (consumo elevado de sal) respecto a cuando no lo está. De hecho cuando la prevalencia del suceso es baja (< 20 %) el valor del odds ratio y el riesgo relativo es muy parecido, pero no es así cuando el suceso es bastante común, hecho que a menudo se ignora y será objeto de un comentario más extenso en un nuevo artículo.

Si en la ecuación de regresión tenemos un factor dicotómico, como puede ser por ejemplo si el sujeto es no fumador, el coeficiente ***b*** de la ecuación para ese factor está directamente relacionado con el odds ratio ***OR*** de ser fumador respecto a no serlo

$$OR = \exp(b)$$

es decir que *exp(b)* es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes.

Cuando la variable es numérica, como puede ser por ejemplo la edad, o el índice de masa corporal, es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro, permaneciendo constantes el resto de variables. Así el odds ratio que supone pasar de la edad *X1* a la edad *X2*, siendo *b* el coeficiente correspondiente a la edad en el modelo logístico es:

$$OR = \exp[b \cdot (X2 - X1)]$$

Nótese que se trata de un modelo en el que el aumento o disminución del riesgo al pasar de un valor a otro del factor es proporcional al cambio, es decir a la diferencia entre los dos valores, pero no al punto de partida, quiere esto decir que el cambio en el riesgo, con el modelo logístico, es el mismo cuando pasamos de 40 a 50 años que cuando pasamos de 80 a 90.

Cuando el coeficiente *b* de la variable es positivo obtendremos un odds ratio mayor que 1 y corresponde por tanto a un factor de riesgo. Por el contrario, si *b* es negativo el odds ratio será menor que 1 y se trata de un factor de protección.

Las variables cualitativas en el modelo logístico

Puesto que la metodología empleada para la estimación del modelo logístico se basa en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, es incorrecto que en él intervengan variables cualitativas, ya sean nominales u ordinales.

La asignación de un número a cada categoría no resuelve el problema ya que si tenemos, por ejemplo, la variable ejercicio físico con tres posibles respuestas: sedentario, realiza ejercicio esporádicamente, realiza ejercicio frecuentemente, y le asignamos los valores 0, 1, 2, significa a efectos del modelo, que efectuar ejercicio físico frecuentemente es dos veces mayor que solo hacerlo esporádicamente, lo cual no tienen ningún sentido. Más absurdo sería si se trata, a diferencia de ésta, de una variable nominal, sin ninguna relación de orden entre las respuestas, como puede ser el estado civil.

La solución a este problema es crear tantas variables dicotómicas como número de respuestas − 1. Estas nuevas variables, artificialmente creadas, reciben en la literatura anglosajona el nombre de "*dummy*", traduciéndose en español con diferentes denominaciones como pueden ser **variables internas**, **indicadoras**, o **variables diseño**.

Así por ejemplo si la variable en cuestión recoge datos de tabaquismo con las siguientes respuestas: *Nunca fumó, Ex−fumador, Actualmente fuma menos de 10 cigarrillos diarios, Actualmente fuma 10 o más cigarrillos diarios*, tenemos 4 posibles respuestas por lo que construiremos 3 variables internas dicotómicas (valores 0,1), existiendo diferentes posibilidades de codificación, que conducen a diferentes interpretaciones, y siendo la más habitual la siguiente:

	I1	**I2**	**I3**
Nunca fumó	0	0	0
Ex− fumador	1	0	0
Menos de 10 cigarrillos diarios	0	1	0
10 o más cigarrillos diarios	0	0	1

En este tipo de codificación el coeficiente de la ecuación de regresión para cada variable diseño (siempre transformado con la función exponencial), se corresponde al odds ratio de esa categoría con respecto al nivel de referencia (la primera respuesta), en nuestro ejemplo cuantifica cómo cambia el riesgo respecto a no haber fumado nunca.

Existen otras posibilidades entre las que se destaca con un ejemplo para una variable cualitativa de tres respuestas:

	I1	**I2**
Respuesta 1	0	0
Respuesta 2	1	0
Respuesta 3	1	1

Con esta codificación cada coeficiente se interpreta como una media del cambio del riesgo al pasar de una categoría a la siguiente.

En el caso una categoría que NO pueda ser considerada de forma natural como nivel de referencia, como por ejemplo el grupo sanguíneo, un posible sistema de clasificación es:

	I1	**I2**
Respuesta 1	−1	−1
Respuesta 2	1	0
Respuesta 3	0	1

donde cada coeficiente de las variables indicadoras tiene una interpretación directa como cambio en el riesgo con respecto a la media de las tres respuestas.

Consejos sobre cómo presentar los resultados de una regresión logística

Es habitual presentar los resultados de la regresión logística en una tabla en la que aparecerá para cada variable el valor del coeficiente; su error estándar; un parámetro, denominado de ***chi² Wald***, que permite contrastar si el coeficiente es significativamente diferente de 0 y el valor de *p* para ese contraste; así como los odds ratio de cada variable, junto con su intervalo de confianza para el 95 % de seguridad.

Ejemplo de presentación de una regresión logística:

Término	Coef.	Err.est.	chi²	p	Nivel signif.
Indepen.	−1.2168	0.9557	1.621	0.2029	NO
Edad	−0.0465	0.0374	1.545	0.2138	NO
Raza *			* 5.684	0.0583	casi(p < 0.1)
Raza 1	1.0735	0.5151	4.343	0.0372	p < 0.05
Raza 2	0.8154	0.4453	3.353	0.0671	casi(p < 0.1)
Fumador	0.8072	0.4044	3.983	0.0460	p < 0.05
HT	1.4352	0.6483	4.902	0.0268	p < 0.05
UI	0.6576	0.4666	1.986	0.1587	NO
LWD	0.8421	0.4055	4.312	0.0379	p < 0.05
PTD	1.2817	0.4621	7.692	0.0055	p < 0.01

Variable	Odds ratio	OR inf.95%	OR sup.95%
Edad	0.95	0.89	1.03
Raza 1	2.93	1.07	8.03
Raza 2	2.26	0.94	5.41
Fumador	2.24	1.01	4.95
HT	4.20	1.18	14.97
UI	1.93	0.77	4.82
LWD	2.32	1.05	5.14
PTD	3.60	1.46	8.91

Bondad del ajuste

Siempre que se construye un modelo de regresión es fundamental, antes de pasar a extraer conclusiones, el corroborar que el modelo calculado se ajusta efectivamente a los datos usados para estimarlo.

En el caso de la regresión logística una idea bastante intuitiva es calcular la probabilidad de aparición del suceso, presencia de hipertensión en nuestro caso, para todos los pacientes de la muestra. Si el ajuste es bueno, es de esperar que un valor alto de probabilidad se asocie con presencia real de hipertensión, y viceversa, si el valor de esa probabilidad calculada es bajo, cabe esperar también ausencia de hipertensión.

Esta idea intuitiva se lleva a cabo formalmente mediante la prueba conocida como de Hosmer−Lemeshow (1989), que básicamente consiste en dividir el recorrido de la probabilidad en deciles de riesgo (esto es probabilidad de hipertensión ≤ 0.1, ≤ 0.2, y así hasta ≤ 1) y calcular tanto la distribución de hipertensos, como no hipertensos prevista por la ecuación y los valores realmente observados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de chi².

En la presentación final de los datos de regresión logística debiera figurar siempre algún tipo de prueba de bondad de ajuste y las conclusiones comentadas que de ella se deducen, pues en el caso de la prueba Hosmer−Lemeshow es más ilustrativo que el propio resultado del contraste, los valores de la distribución obtenida.

Bibliografía seleccionada

Nuestra recomendación para quien desee iniciarse en el tema de la regresión logística con un libro de amena lectura es sin lugar a dudas la primera referencia. Las segunda es mucho más técnica y no apta para quien no tenga un buen nivel de estadística. Respecto a la tercera, es también bastante técnica, y como su título indica trata, con calidad y rigor, no sólo la regresión logística sino otros temas de análisis multivariante.

- **Excursión a la regresión logística en ciencias de la salud.**
  Luis Carlos Silva Ayçaguer
  Ed. Díaz de Santos
  Madrid 1995
- **Applied Logistic Regression**
  David W. Hosmer
  Stanley Lemeshow
  Ed.John Wiley
  New York 1989
- **Métodos multivariantes en bioestadística**
  Víctor Abraira Santos
  Alberto Pérez de Vargas Luque
  Ed. Centro de Estudios Ramón Areces
  Madrid 1996

Direcciones de interés

- MEDLINE: Referencias "logistic regression" and hypertension
- MEDLINE: Referencias "logistic regression"
- BMJ: Referencias "logistic regression" and hypertension
- BMJ: Referencias "logistic regression"
- BMJ. Statistics Note: The odds ratio (HTML y PDF)
- Unidad de Bioestadística Clínica del Hospital Ramón y Cajal que mantiene el Dr. Víctor Abraira
- Cálculadora on−line de regresión logística

Indice de artículos Regresión logística (II) Principio de la página