Asociación de la Sociedad Española de Hipertensión
Liga Española para la lucha contra la Hipertensión Arterial

LA REGRESION LOGISTICA (II)

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

Artículo en formato PDF

Febrero 2001

Regresión logística (I)

Interacción y confusión

El empleo de técnicas de regresión sirve para dos objetivos:

1. **Estimar la relación entre dos variables** teniendo en cuenta la presencia de otros factores
2. **Construir un modelo que permita predecir el valor de la variable dependiente** (en regresión logística la probabilidad del suceso) para unos valores determinados de un conjunto de variables pronóstico

Cuando el objetivo es estimar la relación o asociación entre dos variables, los modelos de regresión permiten considerar que puede haber otros factores que modifiquen esa relación.

Así, por ejemplo, si se está estudiando la posible relación, como factor de riesgo, entre el síndrome de apnea nocturna y la probabilidad de padecer hipertensión, dicha relación puede ser diferente si se tiene en cuenta otras variables como pueden ser la edad, el sexo o el índice de masa corporal. Por ello en un modelo de regresión logística podrían ser incluidas como variables independientes, además del dato de apnea. En la ecuación obtenida al considerar como variables independientes *APNEA*, *EDAD*, *SEXO*, *IMC*, el *exp(coeficiente de la ecuación para APNEA)* nos determina el odds ratio debido a la apena, **ajustado o controlado para el resto de los factores**.

A las variables que, además del factor de interés (en el ejemplo *EDAD*, *SEXO*, *IMC*), se introducen en el modelo, se las denomina en la literatura de diferentes formas: variables control, variables extrañas, covariantes, o **factores de confusión**.

Interacción

Cuando la relación entre el factor en estudio y la variable dependiente se modifica según el valor de una tercera estamos hablando de interacción. Así en nuestro ejemplo, supongamos que la probabilidad de padecer *HTA* cuando se tiene síndrome de apnea aumenta con la edad. En este caso decimos que existe interacción entre las variables *EDAD* y *APNEA*.

Si nos fijamos sólo en el exponente del modelo logístico, en el caso de no considerar interacción éste será:

$$-b_0 - b_1 \cdot APNEA - b_2 \cdot EDAD$$

Si deseamos considerar la presencia de interacción entre APNEA y EDAD el modelo cambia:

$$-b_0 - b_1 \cdot APNEA - b_2 \cdot EDAD - b_3 \cdot APNEA \times EDAD$$

Si la variable *APNEA* es dicotómica (valores 0 y 1) la relación entre *HTA* y *APNEA* vendrá cuantificada por b1 en el primer modelo mientras que en el segundo

$$-\left(b_1 + b_3 \cdot EDAD\right) \cdot APNEA$$

es decir que ahora esa relación se modifica en función del valor de la *EDAD*.

Algunas precauciones

La amplia disponibilidad de potentes programas que permiten el acceso a sofisticadas pruebas estadísticas puede conducir a la utilización inadecuada y mecánica de éstas. En particular los modelos de regresión requieren de quien los construye un mínimo de comprensión de la filosofía subyacente, así como no sólo el conocimiento de las ventajas, sino también de los problemas y debilidades de éstas técnicas. La utilización de procedimientos matemáticos a menudo nos convece de que estamos introduciendo "objetividad" en los resultados y ello es así en cierta medida, pero también lleva aparejada una gran carga de subjetividad, donde se incluye desde la misma elección de un modelo matemático determinado, hasta la selección de las variables en él contenidas.

Una de la primeras consideraciones que hay que hacer es que la relación entre la variable independiente y la probabilidad del suceso no cambie de sentido, ya que en ese caso no nos sirve el modelo logístico. Esto es algo que habitualmente no ocurre en los estudios clínicos, pero por ello es más fácil pasarlo por alto cuando se presenta.
Un ejemplo muy claro de esa situación se da si estamos evaluando la probabilidad de nacimiento un niño con bajo peso (de riesgo) en función de la edad de la madre. Hasta una edad esa probabilidad puede aumentar a medida que la edad de la madre disminuye (madres muy jóvenes) y a partir de una edad (madres muy mayores) la probabilidad puede aumentar a medida que lo hace la edad de la madre. En este caso el modelo logístico no sería adecuado.

Colinealidad

Otro problema que se puede presentar en los modelos de regresión, no sólo logísticos, es que la variables que intervienen estén muy correlacionadas, lo que conduce a un modelo desprovisto de sentido y por lo tanto a unos valores de los coeficientes no interpretables. A esta situación, de variables independientes correlacionadas, se la denomina colinealidad.

Para entenderlo supongamos el caso extremo, en el que se introduce en el modelo dos veces la misma variable, tendríamos entonces el siguiente término

$$\exp(-b_0 - b_1 \cdot X - b_2 \cdot X)$$

o lo que es lo mismo

$$\exp\left[-b_0 - \left(b_1 + b_2\right) \cdot X\right]$$

Donde la suma b1+b2 admite infinitas posibilidades a la hora de dividir en dos sumandos el valor de un coeficiente, por lo que la estimación obtenida de b1 y b2 no tiene realmente ningún sentido.

Un ejemplo de esta situación se podría dar si incluimos en la ecuación variables como la hemoglobina y el hematocrito que está altamente correlacionadas.

Tamaño de muestra

Como regla "de andar por casa" podemos considerar necesario disponer de al menos 10 . (k + 1) casos para estimar un modelo con k variables independientes; es decir, al menos 10 casos por cada variable que interviene en el modelo, considerando también la variable dependiente (la probabilidad del suceso).

Conviene llamar la atención respecto a que las cualitativas intervienen como c − 1 variables en el modelo, al construir a partir de ellas las correspondientes variables internas.

Selección de modelos

Al estar hablando de modelos que pueden ser multivariantes, un aspecto de interés es cómo seleccionar el mejor conjunto de variables independientes a incluir en el modelo.

La definición de mejor modelo depende del tipo y el objetivo del estudio. En un modelo con finalidad predictiva se considerará como mejor modelo áquel que produce predicciones más fiables, mientras que en un modelo que pretende estimar la relación entre dos variables (corrigiendo el efecto de otras, como se vió más arriba), se considerará mejor áquel con el que se consigue una estimación más precisa del coeficiente de la variable de interés. Esto se olvida a menudo y sin embargo conduce a estrategias de modelado completamente direfentes. Así en el segundo caso un covariante con coeficiente estadísticamente significativo pero cuya inclusión en la ecuación no modifica el valor del coeficiente de la variable de interés, será excluído de la ecuación, ya que no se trata de un factor de confusión: la relación entre la variable de interés y la probabilidad no se modifica si se tiene en cuenta esa variable. Sin embargo si lo que se busca un modelo predicitivo sí que se incluirá en la ecuación pues ahora lo que buscamos es predicciones más fiables.

Otra consideración que hay que hacer siempre que se analizan datos es distinguir entre diferencias numéricas, diferencias estadísticamente significatifvas y diferencias clínicamente relevantes. No siempre coinciden los tres conceptos.

Lo primero que habrá que plantear es el **modelo máximo**, o lo que es lo mismo el número máximo de variables independientes que pueden ser incluidas en la ecuación, considerando también las interacciones si fuera conveniente.

Aunque existen diferentes procedimientos para escoger el modelo sólo hay tres mecanismos básicos para ello: empezar con una sola variable independiente e ir añadiendo nuevas variables según un criterio prefijado (procedimiento hacia adelante), o bien empezar con el modelo máximo e ir eliminando de él variables según un criterio prefijado (procedimiento hacia atrás). El tercer método, denominado en la literatura "***stepwise***" , combina los dos anteriores y en cada paso se puede tanto añadir una variable como eliminar otra que ya estaba en la ecuación.

En el caso de la regresión logística el criterio para decidir en cada paso si escogemos un nuevo modelo frente al actual viene dado por el **logaritmo del cociente de verosimilitudes de los modelos**.
La **función de verosimilitud** de un modelo es una medida de cuán compatible es éste con los datos realmente observados. Si al añadir una nueva variable al modelo no mejora la verosimilitud de forma apreciable, en sentido estadístico, ésta variable no se incluye en la ecuación.

Para evaluar la significación estadística de una variable concreta dentro del modelo, nos fijaremos en el valor de chi² (estadístico de Wald) correspondiente al coeficiente de la variable y en su nivel de probabilidad

Enlaces de interés

- Logistic regression models used in medical research are poorly presented. BMJ 1996;313: 628
- Prognostic models: clinically useful or quickly forgotten?. Wyatt JC, Altman, DG. BMJ 1995;311:1539−1541

Indice de artículos Regresión logística (I) Arriba