Revista Colombiana de Estadística
Nº 25 - 26, págs. 13 - 32

REGRESIÓN LOGÍSTICA EN ESTUDIOS EPIDEMIOLÓGICOS DE CASOS Y CONTROLES

Beatriz Melo Villalobos[1] y Siegfried Weber[2]

Resumen[3]

En el contexto de los estudios epidemiológicos el uso del modelo de regresión logística se ha hecho cada vez más común. Esto implica la estimación de los *parámetros y su relación con la estimación de la razón odds, como medida* indirecta del riesgo relativo. El modelo de regresión logística se aplicó a un estudio de casos y controles de la enfermedad de Hodgkin, con datos del Instituto Nacional de cancerología de Santafé de Bogotá, con el propósito de probar hipótesis vigentes acerca de la etiología infecciosa de la enfermedad.

Abstract

In epidemiological studies, the use of the logist regression model has become more and more common. It implies the estimation of the model parameters and their relation with the odds ratio as indirect measure of the relative risk. The logistic regression model it's applied to a case-control study for Hodgkin's disease at the Instituto Nacional de Cancerología in Santafé de Bogotá. The purpose of study was to test current hyphoteses about the infectious etiology of that disease.

---

1 Estudiante de Postgrado de Estadística, Departamento de Matemáticas y Estadística de la U. Nal de Colombia, Santafé de Bogotá

2 Profesor Visitante, Departamento de Matemáticas y Estadística de la U. Nal. de Colombia, Santafé de Bogotá.

3 *El presente artículo se basa, en parte, en la tesis de Magíster presentada por la primera autora bajo la* dirección del segundo autor. La bibliografía se tomó de esta tesis, donde se encuentran las referencias correspondientes.

1. Introducción

En los estudios epidemiológicos, los de casos y controles se ubican dentro de los estudios observacionales, y son comúnmente llamados estudios retrospectivos. Aquí se procede de *efecto a causa*, los individuos con una condición particular o enfermedad (los casos) se seleccionan para ser comparados con una serie de individuos en quienes la condición o enfermedad está ausente (los controles). Los casos y los controles se comparan con respecto a la existencia de exposiciones pasadas que se sospecha son relevantes para el desarrollo de la enfermedad bajo estudio. El requerimiento de un grupo de control es evidente pues proporciona una estimación de la frecuencia de exposición esperada entre los individuos libres de la enfermedad.

Entonces, puesto que el propósito fundamental, en un estudio epidemiológico, es identificar factores de riesgo (exposiciones) asociados al desarrollo de la enfermedad se presentan las medidas relativas de ocurrencia de la enfermedad.

Sea la *variable* aleatoria que se supone como *explicativa* de la enfermedad:

$$X_1 = \begin{cases} 1: & \text{si la exposición está presente} \\ 0: & \text{si la exposición no está presente} \end{cases}$$

Para cada posible valor $x_1 = 0,1$ de X_1 sea:

$$D = \begin{cases} 1: & \text{si ocurre la enfermedad} \\ 0: & \text{si no ocurre la enfermedad} \end{cases}$$

la *variable* de Bernoulli $\sim \beta(1, p(x_1))$ que describe la ocurrencia de la enfermedad donde:

$$p(x_1) = P(D = 1/X_1 = x_1), \text{ para } x_1 = 0,1.$$

Con las notaciones para las *probabilidades conjuntas* π_{jk} y *marginales* π_{j+}, π_{+k}, reunidas en el esquema 1. Se definen:

la *tasa de incidencia de la enfermedad* o *riesgo de la enfermedad* entre los expuestos:

$$p(1) = \frac{\pi_{11}}{\pi_{+1}} = P(D = 1/X_1 = 1),$$

Esquema 1

$\pi_{jk} = P(D = j, X_1 = k)$	$X_1 =$ 0	1	$\pi_{j\cdot} = P(D = j)$
$D =$ 0	π_{00}	π_{01}	$\pi_{0\cdot}$
1	π_{10}	π_{11}	$\pi_{1\cdot}$
$\pi_{+k} = P(X_1 = k)$	π_{+0}	π_{+1}	1

entre los individuos no expuestos:

$$p(0) = \frac{\pi_{10}}{\pi_{+0}} = P(D = 1/X_1 = 0);$$

el riesgo relativo de la ocurrencia de la enfermedad:

$$R(D = 1/X_1) = \frac{p(1)}{p(0)},$$

y de la no ocurrencia de la enfermedad:

$$R(D = 0/X_1) = \frac{1 - p(1)}{1 - p(0)};$$

la *odds de la enfermedad* entre los individuos expuestos:

$$\frac{p(1)}{1 - p(1)},$$

entre los individuos no expuestos:

$$\frac{p(0)}{1 - p(0)};$$

y finalmente la razón odds de la enfermedad entre los individuos expuestos relativa a los individuos no expuestos:

$$\Psi = \Psi(D / X_1) = \frac{p(1)/(1 - p(1))}{p(0)/(1 - p(0))} = \frac{p(1)/p(0)}{(1 - p(1))/(1 - p(0))} = \frac{R(D = 1 / X_1)}{R(D = 0 / X_1)}.$$

De la última expresión para la razón odds se ve que bajo el supuesto de baja probabilidad de ocurrencia de la enfermedad, tanto en los individuos expuestos como en los no expuestos, Ψ se aproxima al riesgo relativo de la enfermedad.

Para la situación de más de una variable explicativa se considera el siguiente modelo.

2. El modelo de regresión logística

Se denota la variable que representa la enfermedad por:

$$D = \begin{cases} 1: & \text{si la enfermedad ocurre} \\ 0: & \text{si la enfermedad no ocurre} \end{cases}$$

Esta se conoce como variable dependiente o variable de respuesta Se supone el vector $X = (X_1, \ldots, X_K)'$, con valores $x = (x_1, \ldots, x_K)'$, como el vector de las variables aleatorias, que se interpretan como *explicativas* de la enfermedad y pueden determinar la presencia de una *exposición* y/o variables de confusión.

Un modelo de regresión logística está basado, entonces, en los siguientes supuestos:

(1). $(D/X = x) \sim \beta(1, p(x))$, con $p(x) = P(D = 1/X = x) = E(D/X = x)$, brevemente $p(x) = P(D = 1/x)$;

(2) $\eta = \eta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K$;

(3) $p(x) = \dfrac{\exp(\eta)}{1 + \exp(\eta)}$ o equivalentemente: $\eta = \text{logit}(p(x)) = \ln \dfrac{p(x)}{1 - p(x)}$.

Es decir, la esencia del modelo de regresión logística es suponer que la función logit de $p(x)$ es una combinación lineal de las variables explicativas, las cuales pueden ser cuantitativas o cualitativas.

En el contexto de los modelos lineales generalizados (GLM) se interpretan:

(1) como supuesto sobre la regla de aleatoriedad de las variables, (2) como supuesto sobre la parte determinística que debe servir como predicción, (3) como enlace entre $\eta(x)$ y la esperanza (condicional) $p(x)$ de la variable de la enfermedad.

Los parámetros introducidos en el numeral (1) se definen aquí de una manera análoga. Particularmente, se obtiene una simple expresión para las razones odds en términos de los coeficientes de regresión logística, para cada variable explicativa dicotómica. Así por ejemplo para X_1:

$$\beta_1 = \text{logit } p(1, x_2, \ldots, x_K) - \text{logit } p(0, x_2, \ldots, x_K)$$

$$= \ln \Psi_1,$$

siendo Ψ_1 la razón odds de la enfermedad correspondiente a la variable X_1.

3. Surgimiento del modelo logístico

Sea D la variable de Bernoulli que describe la ocurrencia de la enfermedad y X el vector explicativo de la enfermedad, con posibles valores x, es decir:

$$(D/X = x) \sim b(1, p(x)) \quad \text{con} \quad p(x) = P(D = 1/x).$$

Entonces, puede expresarse:

$$P(D = d/x = \left(p(x)\right)^d \left(1 - p(x)\right)^{1-d} = \left[\frac{p(x)}{1 - p(x)}\right]^d \left[1 - p(x)\right] = \frac{\exp(\eta d)}{1 + \exp(\eta)},$$

donde $\eta = \eta(x) = \log \frac{p(x)}{1 - p(x)}$ es el log odds, abreviado como antes, también como función logit $p(x)$, $d = 0,1$. Es decir, la función logit de $p(x)$ aparece de una manera "natural".

Para $d = 1$ se recupera, naturalmente:

$$p(x) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)}.$$

4. Estimación de máxima verosimilitud en el modelo de regresión logística

Se observan para cada i = 1, ..., n, las variables muestrales:

$$D_i = \begin{cases} 1: & \text{si ocurre la enfermedad en la i-ésima persona} \\ 0: & \text{si no ocurre la enfermedad en la i-ésima persona} \end{cases}, \text{ y}$$

$X_i = (X_{1i}, \ldots, X_{Ki})'$, el vector de las variables explicativas correspondientes a la i-ésima persona. se supone independencia entre las D_i, dadas $X_i = x_i$, y además los tres supuestos para un modelo logístico del numeral 2 para cada observación. Así sigue la siguiente forma del *logaritmo de la función de verosimilitud*:

$$L = \ln\left[\prod_{i=1}^{n} \left(p(x_i)^{d_i}\left(1 - p(x_i)\right)^{1-d_i}\right)\right]; \text{ por (1) con } d_i = 0,1; \text{ y la independencia;}$$

$$= \sum_{i=1}^{n} \left[d_i \ln \frac{p(x_i)}{1 - p(x_i)} + \ln\left(1 - p(x_i)\right)\right]$$

$$= \sum_{i=1}^{n} \left[d_i \eta_i - \ln\left(1 + \exp(\eta_i)\right)\right]; \text{ por (3) con } \eta_i = \ln \frac{p(x_i)}{1 - p(x_i)}$$

$$= \beta_0\left(\sum_{i=1}^{n} d_i\right) + \sum_{k=1}^{K} \beta_k\left(\sum_{i=1}^{n} d_i x_{ki}\right) - \sum_{i=1}^{n} \ln\left(1 + \exp(\eta_i)\right); \text{ por (2) con}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki}$$

Para obtener las estimaciones de máxima verosimilitud de $\beta_0, \beta_1, \ldots, \beta_K$, se calculan las "*ecuaciones normales*":

(a) $$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = n_1 - \sum_{i=1}^{n} p(x_i) \stackrel{!}{=} 0, \text{ siendo } n_1 = \sum_{i=1}^{n} d_i,$$

(b) $$\frac{\partial L}{\partial \beta_k} = \sum_{i=1}^{n} d_i x_{ki} - \sum_{i=1}^{n} p(x_i)\, x_{ki} \stackrel{!}{=} 0, \text{ para } k = 1, \ldots, K.$$

La existencia única de las estimaciones se demostrará en el siguiente numeral. Las estimaciones de los coeficientes de regresión logística implican estimaciones para los parámetros de los numerales 1 y 2. Particularmente sigue, si X_1 es una variable dicotómica:

$$\hat{\beta}_1 = \ln \hat{\Psi}_1.$$

5. Existencia de las estimaciones de máxima verosimilitud

Usando en el numeral 4, $x_{01} = 1$ para cada $i = 1, \ldots, n$, se pueden reescribir las ecuaciones normales (a) y (b) en una sola fórmula:

$$\frac{\partial L}{\partial \beta_k} = \sum_{i=1}^{n} d_i x_{ki} - \sum_{i=1}^{n} p(x_i)\, x_{ki} = 0 \quad \text{para } k = 0, 1, \ldots, K.$$

Estas ecuaciones pueden ser escritas como:

$$\begin{bmatrix} \frac{\partial L}{\partial \beta_0} \\ \vdots \\ \frac{\partial L}{\partial \beta_K} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{1n} \\ \vdots & & \\ x_{K1} & \cdots & x_{Kn} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} - \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{1n} \\ \vdots & & \\ x_{K1} & \cdots & x_{Kn} \end{bmatrix} \begin{bmatrix} p(x_1) \\ \vdots \\ p(x_n) \end{bmatrix} \overset{!}{=} \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

abreviando como:

$$\frac{\partial L}{\partial \beta} = x' \cdot d - x' \cdot m = 0;$$

siendo:

$\frac{\partial L}{\partial \beta}$ el vector de las primeras derivadas parciales;

x' la matriz de diseño;

d el vector de las observaciones $d_i = 0,1$;

m el vector de las esperanzas (condicionales) $m_i = p(x_i) = P(D_i = 1/x_i) = E(D_i/x_i)$.

Se necesita calcular ahora las segundas derivadas parciales:

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_j} = - \sum_{i=1}^{n} \left[\frac{\partial p(x_i)}{\partial \beta_j} \right] \cdot x_{ki} = - \sum_{i=1}^{n} \left[\frac{\exp(\eta_i)}{\left(1 + \exp(\eta_i)\right)^2} \cdot x_{ji} \right] \cdot x_{ki}$$

$$= - \sum_{i=1}^{n} p(x_i) \cdot \left(1 - p(x_i)\right) \cdot x_{1i} \cdot x_{ki} = - \sum_{i=1}^{n} v_i \cdot x_{ki} \cdot x_{1i}; \; v_i = p(x_i)\left(1 - p(x_i)\right).$$

Estas segundas derivadas se reúnen en la matriz:

$$\frac{\partial^2 L}{\partial \beta^2} = - x' \cdot v \cdot x \quad \text{que sea definida negativa,} \quad v = \begin{bmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_n \end{bmatrix},$$

donde $v_i = p(x_i)\,((1 - p(x_i)) = V(D_i/X_i = x_i)$ son las varianzas condicionales de D_i, dado $X_i = x_i$.

Por lo tanto existen las estimaciones de máxima verosimilitud $\hat{\beta}_k$ de los parámetros β_k.

6. Solución aproximada de la estimación de máxima verosimilitud

Se calculan las estimaciones $\hat{\beta}_k$ mediante el método de aproximación de Newton-Raphson, como se explica a continuación.

Abreviando: el vector $g(\beta) = \frac{\partial L}{\partial \beta}$, la matriz $G(\beta) = \frac{\partial^2 L}{\partial \beta^2}$, sigue según Taylor:

$$L(\beta) \simeq L\left(\beta^{(r)}\right) + g'\left(\beta^{(r)}\right) \cdot \left(\beta - \beta^{(r)}\right) + \frac{1}{2}\left(\beta - \beta^{(r)}\right)' \cdot G\left(\beta^{(r)}\right) \cdot \left(\beta - \beta^{(r)}\right).$$

siendo $\beta^{(r)}$ algún vector fijo (aproximado a $\hat{\beta}$).

Derivando con respecto a β (como vector), sigue:

$$g(\beta) \simeq g\left(\beta^{(r)}\right) + G\left(\beta^{(r)}\right) \cdot \left(\beta - \beta^{(r)}\right).$$

Para la (ML -) estimación $\hat{\beta}$ de β vale:

$$0 = g(\hat{\beta}) \simeq g\left(\beta^{(r)}\right) + G\left(\beta^{(r)}\right) \cdot \left(\hat{\beta} - \beta^{(r)}\right), \text{ es decir:}$$

$$\hat{\beta} \simeq \beta^{(r)} - G^{-1}\left(\beta^{(r)}\right) \cdot g\left(\beta^{(r)}\right).$$

De aquí se obtiene como $(r + 1)$-ésima aproximación de $\hat{\beta}$:

$$\beta^{(r+1)} = \beta^{(r)} - G^{-1}\left(\beta^{(r)}\right) \cdot g\left(\beta^{(r)}\right) = \beta^{(r)} + \left(x' \cdot v^{(r)} \cdot x\right)^{-1} \cdot x' \cdot \left(d - m^{(r)}\right),$$

siendo,

$$m^{(r)} = \left(m_1^{(r)}, \ldots, m_n^{(r)}\right)' \text{ con } m_i^{(r)} = p^{(r)}(x_i) = \frac{\exp \eta_i^{(r)}}{1 + \exp \eta_i^{(r)}},$$

$$\eta_i^{(r)} = \beta_0^{(r)} + \beta_1^{(r)} x_{1i} + \ldots + \beta_K^{(r)} x_{Ki};$$

$$v^{(r)} = \begin{bmatrix} v_1^{(r)} & & 0 \\ & \ddots & \\ 0 & & v_n^{(r)} \end{bmatrix} \text{ con } v_i^{(r)} = m_i^{(r)}\left(1 - m_i^{(r)}\right),$$

y suponiendo que $\left(x' \cdot v^{(r)} \cdot x\right)^{-1}$ existe.

Como paso inicial se puede tomar, por ejemplo:

$$\beta_0^{(0)} = \cdots = \beta_K^{(0)} = 0, \text{ que implica: } m_i^{(0)} = \frac{1}{2},\ v_i^{(0)} = \frac{1}{4}.$$

7. Modelo de regresión logística para un diseño de casos y controles

Sean

$$S_i = \begin{cases} 1: & \text{si el i-ésimo individuo está incluido en la muestra} \\ 0: & \text{si el i-ésimo individuo no está incluido en la muestra} \end{cases}$$

las variables que caracterizan cómo se escoge la muestra, y sean:

$$\nu_0 = P\left(S_i = 1 \mid D_i = 0, X_i = x_i\right)$$
$$\nu_1 = P\left(S_i = 1 \mid D_i = 1, X_i = x_i\right),$$

las probabilidades (en principio *no* conocidas!) para que sea incluido en la muestra el i-ésimo individuo con valor x_i, como control ($D_i = 0$) y como caso ($D_i = 1$), e independientes de i y de x_i.

Entonces, aplicando el *teorema de Bayes* en la siguiente forma:

$$P(A \mid B \cap C) = \frac{P(A \cap B \cap C)}{P(A \cap B \cap C) + P(\overline{A} \cap B \cap C)} = \frac{P(B \mid A \cap C) \cdot P(A \mid C) \cdot P(C)}{P(B \mid A \cap C) \cdot P(A \mid C) \cdot P(C) + P(B \mid \overline{A} \cap C) \cdot P(\overline{A} \mid C) \cdot P(C)},$$

se obtiene:

$$p^*(x_i) = P(D_i = 1 \mid S_i = 1, X_i = x_i)$$

$$= \frac{P(S_i = 1 \mid D_i = 1, X_i = x_i) \cdot P(D_i = 1 \mid X_i = x_i)}{P(S_i=1 \mid D_i=1, X_i=x_i)P(D_i=1 \mid X_i=x_i) + P(S_i=1 \mid D_i=0, X_i=x_i)P(D_i=0 \mid X_i=x_i)}$$

$$= \frac{\nu_1 \cdot p(x_i)}{\nu_1 \cdot p(x_i) + \nu_0 \cdot (1 - p(x_i))}.$$

De aquí se sigue:

$$\frac{p^*(x_i)}{1 - p^*(x_i)} = \frac{\nu_1}{\nu_0} \cdot \frac{p(x_i)}{1 - p(x_i)}.$$

Es decir, la *odds en un estudio-caso control* (en el lado izquierdo, marcado con un *) es igual a la odds en la población (en el lado derecho considerado hasta ahora), multiplicado por la razón (ν_1/ν_0) de las proporciones teóricas según las cuales se incluye un individuo en la muestra.

Naturalmente, las *razones odds* Ψ son iguales, ya que el factor ν_1/ν_0 se cancela. Como consecuencia, para los supuestos del modelo de regresión logística, sigue:

$$(3)^* \quad \eta^*(x_i) = \ln\left(\frac{p^*(x_i)}{1 - p^*(x_i)}\right) = \ln\left(\frac{\nu_1}{\nu_0}\right) + \ln\left(\frac{p(x_i)}{1 - p(x_i)}\right) = \ln\left(\frac{\nu_1}{\nu_0}\right) + \eta(x_i);$$

$$(2)^* \quad \eta^*(x_i) = \beta_0^* + \sum_{k=1}^{K} \beta_k x_{ki}, \text{ donde}$$

$\beta_1, \ldots, \beta_K$ son los mismos coeficientes para $\eta(x_i)$ de (2), mientras:

$$\beta_0^* = \ln\left(\frac{\nu_1}{\nu_0}\right) + \beta_0.$$

Para estudios caso-control, β_0^* es el parámetro que se puede estimar. Generalmente, no se puede estimar β_0, ν_0, ν_1.

ν_1/ν_0 es una *proporción teórica* (que no se conoce) de la cual sólo se sabe que es pequeña para enfermedades no tan frecuentes. Mientras, $n_1/(n - n_1)$ es una *proporción muestral* (que se fija y que, en principio, no tiene nada que ver con la anterior

8. Diseño de un estudio de casos y controles de la enfermedad de Hodgkin

Ahora se describe la aplicación del modelo de regresión logística de un estudio de casos y controles de la enfermedad de Hodgkin (EH), realizado en Santafé de Bogotá. Este estudio, basado en registro hospitalario, se apoyó en la hipótesis de que la EH está asociada con la exposición (anterior) a agentes infecciosos los cuales a su vez están relacionados con bajas condiciones socioeconómicas.

Estudios anteriores han reportado, para países en desarrollo, un comportamiento bimodal por edad, con un primer pico en la infancia. Este hecho sumado a las hipótesis vigentes llevó a la determinación de tomar como población los menores de 16 años. Además se restringió la población a estratos socioeconómicos bajos puesto que no era posible obtener casos de otros estratos.

Para una potencia de 0.8 y un nivel de significancia de 0.1 se adquirió una muestra de 91 casos y 182 controles. Los casos fueron pacientes menores de 16 años, registrados en el Instituto Nacional de Cancerología (I.N.C.), y a quienes se le diagnosticó, primariamente EH, durante el período de enero de 1984 a diciembre de 1990. Los controles fueron menores de 16 años registrados en el I.N.C., durante el mismo período de los casos y diagnosticados con otra enfermedad (excluyendo linfomas).

Mediante las historias clínicas de cada uno de los casos y de cada uno de los controles, se obtuvo información para las siguientes características investigadas:

- Edad, sexo, lugar de procedencia, orden de nacimiento, número de hermanos, edad de la madre en el momento del parto, subtipo histológico.
- Socioeconómicas compuestas por: ámbito de la vivienda (rural o urbana), tipo de vivienda, servicios básicos de la vivienda, escolaridad de los padres.
- Antecedentes infecciosos: varicela, sarampión, tosferina, tuberculosis (TBC), faringoamigdalitis, malaria.

Antecedentes de vacunación.

Finalmente para establecer asociaciones se tomaron como factores de riesgo o exposición los investigados en estudios foráneos y sugeridos como po-

sibles agentes etiológicos de la enfermedad. Particularmente, los antecedentes de infección y características ambientales.

9. Algunos resultados del análisis de regresión logística

Con el propósito de encontrar modelos explicativos para la EH mediante el procedimiento CATMOD del SAS, utilizando la posibilidad (para las variables categóricas con más de 2 categorías) de entrar cada categoría como una variable dicotómica, para obtener una estimación $\hat{\beta}_k$ por categoría, se recodificaron las variables candidatizadas a entrar en el modelo como explicativas de la enfermedad; pero siempre sobre la base de su significado en el contexto de la enfermedad que se estaba investigando, así:

La variable EDAD (EDA) se agrupó en las siguientes 3 categorías: 1 - 4 (años cumplidos, 5 - 9, 10 - 15. A cada categoría se asocia una variable dicotómica: EDA1 = '1' si EDA = 1 - 4 y '0' en otro, EDA2 = '1' si EDA = 5 - 9 y '0' en otro, EDA3 = '1' si EDA = 10 -15 y '0' en otro.

En el mismo orden de sus categorías se hizo la dicotomización de las demás variables que se consideran en el presente artículo.

TIPO DE VIVIENDA (TIP)	❒ pieza
	❒ choza o rancho
	❒ casalote o casa
	❒ apartamento
NUMERO DE SERVICIOS PÚBLICOS QUE CUENTA LA VIVIENDA (SER)	❒ ninguno
	❒ uno
	❒ dos
	❒ tres
ANTECEDENTES DE INFECCIÓN EXPLÍCITOS (ANTINF)	❒ ninguno
	❒ por lo menos uno

Las siguientes variables de dos categorías fueron codificadas con valores '1' y '0', donde se escogió como '0' el valor que debe servir como *referencia*:

INMUNIZACIONES/ESQUEMA DE VACUNAS (INM)	❒ incompletas (1) ❒ completas (0)
ÁMBITO DE LA VIVIENDA (AVI)	❒ rural (1) ❒ urbano (0)
SEXO (SEX)	❒ masculino (1) ❒ femenino (0)
PACIENTE (PA1)	❒ caso (1) ❒ control (0)

Para cada variable dicotómica, digamos X_1 codificada con valores 1 y 0, el procedimiento CATMOD presenta en su salida la estimación $\hat{\beta}(X_1 = 0)$ correspondiente a la categoría más pequeña 0, de X_1. La estimación para la categoría $X_1 = 1$ se obtiene como $\hat{\beta}(X_1 = 1) = -\hat{\beta}(X_1 = 0)$, ya que CATMOD exige que la suma de los dos parámetros correspondientes a las categorías de una variable sea igual a cero. En cuanto a la variable de respuesta, aquí PA1 (pacientes casos Hodgkin y controles mezclados), que se codificó con los valores 1 (para caso) y 0 (para control), CATMOD escoge como "perfil 1" el valor más pequeño 0 y "perfil 2" el valor 1 y, en consecuencia, se basa en el modelo logístico para la probabilidad de (PA1 = 0), dadas las variables explicativas. Teniendo en cuanta este proceso interno de CATMOD, se obtiene la estimación de la odds-ratio Ψ_1 de la variable X_1 tomando el valor 0 como referencia, según la siguiente fórmula.:

$$\begin{aligned} \ln \hat{\Psi}_1 &= \text{logit}\,\hat{p}(1, x_2, \ldots, x_K) - \text{logit}\,\hat{p}(0, x_2, \ldots, x_K) \\ &= -\text{logit}\left[1 - \hat{p}(1, x_2, \ldots, x_K)\right] + \text{logit}\left[1 - \hat{p}(0, x_2, \ldots, x_K)\right] \\ &= -\hat{\beta}(X_1 = 1) + \hat{\beta}(X_1 = 0) = 2\hat{\beta}(X_1 = 0) \end{aligned}$$

Para la selección de modelos se tomaron los dos criterios:

- *Parcial,* (pruebas univariadas para cada hipótesis: $\beta_k = 0$, correspondientes a las variables del modelo, con sus estadísticas de prueba X^2).

- *Global* (prueba para la bondad del ajuste del modelo, en el sentido de una prueba para la hipótesis del modelo logístico en cuestión, con -2 ln (razón de verosimilitud) como estadística de prueba).

Según el criterio parcial, se esperan p-valores pequeños para cada prueba univariada. Mientras, según el criterio global, se necesita un p-valor grande (por lo menos mayor que 10%).

Las tablas 9.1 y 9.2 presentan las estimaciones de los parámetros para tres de los modelos escogidos, así como errores estándar de cada estimación y los correspondientes intervalos al 95% de confianza.

En el modelo 1 de la tabla 9.1, se eliminaron las variables con p-valores de las pruebas parciales mayores que 10%, según el criterio parcial. En el modelo 2 se eliminó además de la variable SER3 por el p-valor de la prueba parcial mayor que el 10%; así se llegó al modelo 3, para el cual todos los p-valores de las pruebas parciales ya son menores que 5% y para el cual el p-valor de la prueba de ajuste es igual a 34%, es decir mayor que 10%.

En resumen, el modelo 3 puede ser considerado como un buen modelo logístico, según los dos criterios, parcial y global. Mientras el modelo 1 no cumplió con el criterio parcial y el modelo 2 no cumplió con los dos criterios (tabla 9.1).

A continuación se sigue, por lo tanto, analizando el modelo 3.

Reemplazando en (2) y (3), del numeral 2, los parámetros β_k por sus estimaciones $\hat{\beta}_k$ de la tabla 9.1, se obtiene como estimación para la probabilidad de no enfermarse, dadas las variables del modelo 3.

$$\begin{aligned}
&\text{logit } \hat{P}(PA1 = 0/INT, SEX, EDA1, EDA2, ANTINF) \\
&= \text{logit } (1 - \hat{P}(INT, SEX, EDA1, EDA2, ANTINF) \\
&= \hat{\beta}_1(INT) + \hat{\beta}_2(SEX=0) - \hat{\beta}_2(SEX=1) + \hat{\beta}_3(EDA1=0) - \hat{\beta}_3(EDA1=1) + \hat{\beta}_4(EDA2=0) - \hat{\beta}_4(EDA2=1) \\
&\quad + \hat{\beta}_5(ANTINF=0) - \hat{\beta}_5(ANTINF=1).
\end{aligned}$$

Recuérdese que CATMOD toma PA1 = 0 como primer nivel de categoría de respuesta PA1.

De aquí siguen, particularmente, estimaciones para las razones odds.

Por ejemplo la razón odds de la enfermedad de Hodgkin relativa al antecedente de infección, ajustado por todas las demás variables se estima por:

$$\hat{\Psi}_5 = \hat{\Psi}(ANTINF) = \exp(2\,\hat{\beta}_5)) = (\exp(0.5161))^2 = 2.81\ .$$

lo que significa que el riesgo de enfermar de Hodgkin es de 2.81 veces más grande relativo al riesgo de enfermar de otro cáncer (o de otra enfermedad) cuando se está expuesto a agentes infecciosos.

Un intervalo del 95% de confianza para β_5 es (0.14, 0.89) y por lo tanto para Ψ_5 es igual a $((\exp(0.14))^2, (\exp(0.89))^2) = (1.32, 5.97)$, que no contiene el 1. Es decir la estimación $\hat{\Psi}_5 = 2.81$ puede considerarse como estadísticamente mayor que 1.

Hay que corroborar los resultados respecto a las variables dicotomizadas de la variable EDA: El valor $\hat{\Psi}_3 = 0.38$ se interpreta como estimación aproximada del riesgo relativo de enfermar de Hodgkin para los menores de 5 años comparado con los que están entre los 5 y los 15 años. El riesgo relativo estimado para los de 5 a 15 años, comparado con los menores de 5 años, resulta aproximadamente igual a $1/0.38 = 2.63$.

Comparando este último valor con el riesgo relativo estimado aproximado $\hat{\Psi}_4 = 2.69$ para los que tienen entre 5 y 9 años comparado con los demás, permite concluir que el riesgo relativo realmente tiene un pico significativo para la edad entre 5 y 9 años.

Los demás resultados se encuentran reunidos en la tabla 9.2.

Tabla 9.1

ESTIMACIONES DE MÁXIMA VEROSIMILITUD DE LOS COEFICIENTES β_K PARA TRES MODELOS QUE RELACIONAN DISTINTAS VARIABLES CON LA ENFERMEDAD DE HODGKIN

VARIABLE	MODELO 1			MODELO 2			MODELO 3		
	$\hat{\beta}_K$	ERROR ESTÁNDAR		$\hat{\beta}_K$	ERROR ESTÁNDAR		$\hat{\beta}_K$	ERROR ESTÁNDAR	
INT	-2.1186	0.5842	***	0.8064	0.3088	**	0.5270	0.2394	*
SEX	0.7016	0.1620	***	0.7137	0.1556	***	0.7176	0.1550	***
EDA1	-0.3383	0.1978	*	-0.5136	0.1877	**	-0.4865	0.1861	**
EDA2	0.6692	0.1806	***	0.4906	0.1674	**	0.4943	0.1665	**
ANTINF	0.5635	0.2085	**	0.5058	0.1935	**	0.5161	0.1924	**
SER1	-0.2972	0.3362							
SER2	0.2002	0.3534							
SER3	-0.4655	0.2754	•	-0.3404	0.2321				
AVI	0.0387	0.3050							
INM	-0.0654	0.2312							
TIP1	0.0333	0.2000							
TIP2	0.4186	0.2770							
P-VALOR CORRESPONDIENTE A LA RAZÓN DE VEROSIMILITUD	18.7%			7.2%			34.3%		

INT : INTERCEPTO

*** : p-valor menor que 0.1%

** : p-valor entre 0.1% y 1%

* : p-valor entre 1% y 5%

• : p-valor entre 5% y 10%; los otros p-valores son mayores que 10%

Tabla 9.2

RAZONES ODDS ESTIMADAS E INTERVALOS DE CONFIANZA

VARIABLE	MODELO 1			MODELO 2			MODELO 3		
	$\hat{\Psi}$	L.I.	L.S.	$\hat{\Psi}$	L.I.	L.S.	$\hat{\Psi}$	L.I.	L.S.
INT	0.01	0.001	0.14	5.02	1.50	16.83	2.87	1.12	7.33
SEX	4.04	2.16	7.68	4.17	2.26	7.67	4.20	2.29	7.71
EDA1	0.51	0.23	1.10 (1)	0.36	0.17	0.75	0.38	0.18	0.78
EDA2	3.81	1.88	7.74	2.67	1.38	5.14	2.69	1.40	5.16
ANTINF	3.09	1.36	6.99	2.75	1.29	5.87	2.81	1.32	5.97
SER1	0.55	0.15	2.06 (1)						
SER2	1.49	0.37	5.96 (1)						
SER3	0.39	0.13	1.16 (1)	0.93	0.20	1.26 (1)			
AVI	1.08	0.33	3.57 (1)						
INM	0.88	0.35	2.17 (1)						
TIP1	1.07	0.49	2.34 (1)						
TIP2	2.31	0.78	6.84 (1)						

INT : Intercepto

L.S. : Límite Superior

L.I. : Límite Inferior

(1) : El intervalo de confianza del 95% contiene el 1, por lo tanto la estimación de la razón odds correspondiente no puede considerarse significativamente diferente de 1.

10. Conclusiones y discusión

Se encontró que el modelo 3 con las variables sexo, edad y antecedentes de infección explican estadísticamente bien la variable de interés: enfermarse de Hodgkin. Los resultados del análisis estadístico fueron consistentes con el comportamiento biológico sospechado, como se resume a continuación.

Se corroboró la hipótesis vigente de que un agente infeccioso juega un papel importante en la epidemiología de la EH. Esta afirmación está apoyada no solamente por las diferencias estadísticamente significativas que presentan los coeficientes de regresión, sino por diferentes aspectos que fueron completados tanto en el diseño del estudio como en el manejo y el análisis mismo de los datos.

Este trabajo reportó, al igual que muchos otros, diferencias por sexo, la distribución muestral dio una razón de 2.8 (con predominio para el sexo masculino). La estimación de la razón odds fue de 4.2, lo que corroboró en términos de riesgos relativos lo que la estadística descriptiva ya sugería.

La mayor prevalencia para la edad se encontró en el grupo de 5 a 9 años. Naturalmente no se pudo confirmar el comportamiento bimodal por edad por el diseño restringido a menores de 16 años, pero la proporción en la infancia parece ser considerable, correspondió al 22% del total. También se conservó la prevalencia encontrada por Correa and O'Conor (1971), para los subtipos histológicos de celularidad y deplesión linfocitaria 50% aproximadamente.

La hipótesis de relación con una amigdalectomía finalmente no pudo ser analizada, pues para que sea confiable la información se requiere de un examen físico al paciente para verificar si tiene amígdalas y ésto no pudo ser realizado puesto que la información se obtuvo de fuente secundaria.

Esta investigación no destaca solamente la confirmación de la hipótesis de trabajo, quedan interrogantes que resolver, particularmente en torno a dos aspectos: 1) si existe relación entre las condiciones climáticas dadas por la posición geográfica de donde vive el paciente, 2) con qué se relaciona el subtipo histológico, pues se ha observado que en países tropicales y desaventajados económicamente hay un alto porcentaje de casos con subtipos histológicos de pobre pronóstico, éste es un comportamiento completamente opuesto al de países más industrializados, donde predomina el subtipo histológico de mejor pronóstico.

Agradecimiento:

A la doctora Margarita Ronderos, epidemióloga del I.N.C., quén llevó nuestra atención a esta investigación y nos ayudó en el suministro de los datos y por sus comentarios.

Bibliografía

Anderson S. et al., 1980, *Statitiscal Methods for Comparative Studies*, John Wiley & Sons.

Breslow N. E. and N. E. Day., 1980, *Statistical Methods in Cancer Research* Volume 1, IARC Scientific Publications Nº 32.

Casagrande J. T. and M. C. Pike, 1978, *An Improved Aproximate Formula fo Calculation Sample Sizes for Comparing Two Binomial Distribution. Biometrics*, 34: 483 - 486.

Correa Pelayo and O'Conor T. Gregory, 1971, *Epidemiologic Patterns o Hodgkin's Disease*, Int. J. Epidemiol, 8: 192 - 201.

Kirchhoff V. Louis, Alfred S. Evans, Karen E. McClelland, Renato P. S. Carvalhc and Claudio S. Pannuti, 1980, *A Case-Control Study of Hodgkin's Disease ir Brazil*, Am. J. Epidemiology, 112. 595 - 608.

Linet S. Martha and Ron Brookmeyer, 1987, *Use of Cancer Control in Case-Control Cancer Studies*, Am. J. Epidemiology, 125: 1 - 10.

Mueller E. Nancy, (Preprint 1990), *Hodgkin's Desease in Cancer Epidemiology and Prevention*, Oxford University Press, Second Edition, Chapter 41.

Schlesselman James J., 1982, *Case-Control Studies*, Oxford University Press, New York.

Smith and Their, 1988, *Fisiopatología*, Segunda edición.

Weber S., 1991, *Estadística*, (partes III, IV), Preprint, material de clase.

West D. W. et al, 1984, *Differences in Risk Estimations from a Hospital and c Population-Based Case-Control Study*, Int. J. Epidemiology, 13: 235 - 239.