COMPARACIONES DE MODELOS PARA EL CÁLCULO DE LA DISTRIBUCIÓN PROGNÓSTICA

Mª PILAR ZULUAGA ARIAS
UNIV. COMPLUTENSE DE MADRID

La solución bayesiana a diferentes problemas de predicción lleva consigo el cálculo de la distribución prognóstica. En este trabajo se calcula dicha distribución mediante diferentes modelos paramétricos, comparándose después por dos técnicas. Para todo ello se han confeccionado diversos algoritmos, que escritos en FORTRAN IV, constituyen el software indicado para la resolución, en la práctica, de situaciones reales que se ajustan a los modelos teóricos aquí propuestos.

Keywords: Alpha Distributions, Montecarlo Method, Pronostic Distribution.

1. INTRODUCCIÓN

El problema que aquí se resuelve supone conocidos los siguientes datos:

Y: variable respuesta (continua)

$X = (X^{(1)},...,X^{(n)})$: vector de variables explicativas (continuas)

$D = \{(x_i,y_i),\ i=1,...,n\}$: blanco de datos

El objetivo de los distintos modelos que se proponen es el cálculo de la distribución predictiva p(y/x,D).

El método, más usual, de obtener dicha distribución es el dado por Aitchinson & Dunsmore (ver /1/) que supone un modelo paramétrico:

$$p(y/x,w) \quad \text{con} \quad w \in W$$

y estimando p(w) del banco de datos obtiene:

$$p(y/x,D) = \int_W p(y/x,w)\ p(w/D)\ dw \qquad \textbf{(1.1)}$$

• Mª Pilar Zuluaga Arias, Dept. de Estadística e Inv. Oper. Fac. de Matemáticas - Univ. Complutense de Madrid. 28040 Madrid.
• Article rebut el Juliol de 1986.

2. MODELOS PROPUESTOS

En este trabajo nos vamos a limitar al caso en el que los datos, permitan suponer un modelo paramétrico Normal. En este supuesto se estudian diferentes situacions:

2.1. *MODELO I.*

Supondremos en este caso que sólo se tiene una variable explicativa X y además:

$$p(y/x,w) : N(w_1 + w_2\ x,\ w_3) \qquad \textbf{(2.1)}$$

si ahora tomamos como p(w/x) la a priori no informativa se tiene (ver /1/) que p(y/x,D) evaluada según (1.1) tiene una distribución Student generalizada:

$$p(y/x,D) : St(A_x,\ B_x,\ C_x) \qquad \textbf{(2.2)}$$

Tendríamos, por lo tanto, resuelto el problema cuando se verifica:
$y \simeq w_1\ x + w_2$.

2.2. *MODELO II.*

Supondremos en este caso que el vector X es bidimensional, es decir, $X = (X^{(1)},X^{(2)})$.

El modelo que proponemos está en la linea del dado en /9/ en el contexto de asignación de tratamientos.

Intuitivamente, la forma de construir este modelo se basa en "aprovechar" la información del banco de datos referente a los individuos con características similares a x. De manera teórica esto se concreta en la definición de una colección finita de particiones $C = \{M_1,...,M_t\}$, formada por t particiones de $X^{(2)}$.

Suponiendo ahora, en cada clase de las particiones ($m_{jk} \in M_j$), al igual que se hacía en el MODELO I, un modelo paramétrico Normal, es decir

$$P_{m_{jk}}(y/x,w) : N(w_{1jk} + w_{2jk}x,\ w_{3jk}) \qquad \textbf{(2.3)}$$

y procediendo de igual forma, se concluye que las $P_{M_j}(y/x,D)$ son también Student generalizadas, con lo que dicha distribución predictiva se puede dar como:

$$p(y/x,D) = \sum_{j=1}^{t} p(M_j/x)P_{M_j}(y/x,D) \qquad \textbf{(2.4)}$$

donde $p(M_j/x)$ se puede dar de diferentes formas, por ejemplo si se supone $C = \{M_1, M_2\}$

$$p(M_1/x) = \frac{|x^{(2)} - \mu_{2k'}|}{|x^{(2)} - \mu_{1k}| + |x^{(2)} - \mu_{2k'}|} \qquad \textbf{(2.5)}$$

para $x^{(2)} \in m_{1k} \cap m_{2k'}$, siendo μ_{jk} la media de $X^{(2)}$ en $m_{jk} \in M_j$.

2.3. MODELO III.

Supondremos en este caso $X = (X^{(1)}, X^{(2)})$ y además otra variable Z (que denominamos perturbadora) la cual es conocida para los individuos del banco de datos, pero no lo es para futuros individuos.

Tendremos ahora como banco de datos $D = \{(x_i, z_i, y_i), i=1,...,n\}$ del que obtenemos:

1. Aunque para individuos futuros no conozcamos z si podemos dar la distribución diagnóstica: p(z/x,D), la cual se puede calcular según distintos métodos (ver /1/, /2/, /3/, /7/, /8/).

2. De manera análoga al MODELO II, una colección C formada por s particiones del banco de datos según $X^{(2)}$ y Z.

3. Para cada x, considerando una a priori $p(M_j/x)$ tal que j=1,...,s podemos dar

$$p(y/x,z,D) = \sum_{j=1}^{s} p(M_j/x) p_{M_j}(y/x,z,D) \qquad \textbf{(2.6)}$$

4. El cálculo de la distribución prognóstica será según

$$p(y/x,D) = \sum_{z \in Z} p(y/x,z,D)\ p(z/x,D) \qquad \textbf{(2.7)}$$

3. RESOLUCION DE UN CASO REAL

En este apartado se resuelve un problema real aplicando los MODELOS propuestos.

3.1. Planteamiento

Sea Y la variable peso de un recién nacido. Sean $X^{(1)}$ y $X^{(2)}$ variables obtenidas mediante ecografía del feto, concretamente

$X^{(1)}$ = Diámetro transverso abdominal

$X^{(2)}$ = Longitud del fémur

Además, una vez nacidos se sabe si los niños son distróficos (z=1) o normales (z=2), es decir, si presentan lesiones orgánicas debidas a problemas de nutrición durante el embarazo. Se parte de un banco de 80 niños y se busca la distribución predictiva del peso en función de dichas variables. Aplicaremos los tres modelos propuestos anteriormente:

3.2. Aplicación del MODELO I

En este modelo se supone una única variable explicativa, por ello se elegira de entre $X^{(1)}$ y $X^{(2)}$ la más correlacionada linealmente con Y, que es $X^{(1)}$:$r(y,z^{(1)}) = 0.9056$. Esto permite suponer

$$p(y/x^{(1)},w) : N(w_1 + w_2x^{(1)}, w_3)$$

y por lo tanto

$$p(y/x^{(1)}, D) : St(A_{x^{(1)}}, B_{x^{(1)}}, C_{x^{(1)}})$$

siendo estos parámetros:

$$\left.\begin{aligned} A_{x^{(1)}} &= n - 2 = 78 \\ B_{x^{(1)}} &= \bar{y} - \bar{x}^{(1)}\hat{w}_1 + \hat{w}_1x^{(1)} = -2562.9 + 557.7\, x^{(1)} \\ C_{x^{(1)}} &= \frac{v}{(n-2)} + \frac{v}{(n-2)n} + \frac{v}{(n-2)S(x^{(1)},x^{(1)})}(x^{(1)} - \bar{x}^{(1)})^2 = \\ &= 89461.3 + 808.5\,(x^{(1)} - 10.3)^2 \end{aligned}\right\} \quad \textbf{(3.1)}$$

donde

$$\hat{w}_1 = \frac{S(x^{(1)}, y)}{S(x^{(1)},x^{(1)})}$$

con

$$S(x^{(1)}, x^{(1)}) = \sum_i (x_i^{(1)} - \bar{x}^{(1)})^2$$

$$S(x^{(1)}, y) = \sum_i (x_i^{(1)} - \bar{x}^{(1)})(y_i - \bar{y})$$

$$v = S(y,y) - \frac{S(x^{(1)},y)^2}{S(x^{(1)},x^{(1)})}$$

siendo $\bar{x}^{(1)}$ e $\bar{y}$ las medias respectivas de $x^{(1)}$ e Y en D.

Para pareja de datos $(y,x^{(1)})$ del banco de datos se obtiene:

$$p(y/x^{(1)},D) = \frac{1}{\beta(\frac{1}{2},\frac{1}{2}A_{x^{(1)}})\,(A_{x^{(1)}}C_{x^{(1)}})^{1/2}\left\{1+\frac{(y-B_{x^{(1)}})^2}{A_{x^{(1)}}C_{x^{(1)}}}\right\}^{\frac{(A_{x^{(1)}}+1)}{2}}} \qquad \textbf{(3.2)}$$

3.3. Aplicación del MODELO II

Debido al tamaño muestral (n=80) se van a considerar dos particiones M_1 y M_2 del banco de datos según $X^{(2)}$. M_1 tiene dos clases (m_{11},m_{12}) y M_2 tres (m_{21},m_{22},m_{23}). Se tiene así lo siguiente:

- Los coeficientes de correlación lineal de Y y $X^{(1)}$ permiten suponer un modelo paramétrico Normal en cada m_{jk} , con lo que $P_{M_j}(y/x,D)$ serán Student generalizadas que se pueden calcular según (3.1) considerando como banco de datos en cada clase los elementos de dicha clase.
- Cada valor del banco de datos $x^{(2)}$ pertenece a una de las clases de la partición M_1 y a otra de M_2 , por lo tanto para cada $(y, x^{(1)}, x^{(2)})$ se calcularán $P_{M_1}(y/x^{(1)},D)$ y $P_{M_2}(y/x^{(1)},D)$ de manera análoga a (3.2) según lo siguiente.

$M_1 = \{ m_{11}, m_{12}\} = \{(\leq 7] , (> 7)\}$

$M_2 = \{ m_{21}, m_{22}, m_{23}\} = \{(\leq 6.75] , (6.75, 7.25] , (> 7.25)\}$

	$r(x^{(1)},y)$	n	$A_x(1)$	$A_x(1)$	$C_x(1)$
m_{11}	0.9222	36	34	$-2512.6 + 544.7x^{(1)}$	$105622.6 + 1535.4(x^{(1)} - 9.8)^2$
m_{12}	0.8182	44	42	$-1731.2 + 486.3x^{(1)}$	$69966.8 + 2780.8(x^{(1)} - 10.7)^2$
m_{21}	0.9457	22	20	$-2696.2 + 562.4x^{(1)}$	$103644.7 + 1869.2(x^{(1)} - 9.4)^2$
m_{22}	0.5875	28	26	$-167.3 + 322.8x^{(1)}$	$75222.4 + 7606.2(x^{(1)} - 10.4)^2$
m_{23}	0.8298	30	28	$-1346.9 + 457.2x^{(1)}$	$60191.5 + 3367.2(x^{(1)} - 10.9)^2$

- Por último aplicando (2.4) se obtiene la distribución prognóstica.

3.4. Aplicación del MODELO III.

Dado el bajo tamaño de algunas clases al dividir M_1 y M_2 según las categorías de z, se utilizará solamente $X^{(1)}$, con lo cual suponiendo un modelo paramétrico Normal para niños distróficos y otro para niños normales se tiene:

- La distribución de p(y/x,z,D) es una Student generalizada, cuyos parámetros son:

	$r(x^{(1)},y)$	n	$A_x(1)$	$B_x(1)$	$C_x(1)$
z = 1	0.8034	66	64	$-1850.5 + 493.3x^{(1)}$	$90851.7 + 2149.2(x^{(1)} - 10.6)^2$
z = 2	0.9600	14	12	$-2533.1 + 540.2x^{(1)}$	$64309.9 + 2069.1(x^{(1)} - 8.8)^2$

- La distribución diagnóstica la calculamos mediante regresión logística (ver /8/), resultando:

$$p(z = 1 \;/\; x^{(1)},D) = \frac{\exp(-17.873 + 1.941\ x^{(1)})}{1 + \exp(-17.873 + 1.941\ x^{(1)})}$$

- Aplicando, ahora (2.7) se tiene la distribución prognóstica.

4. COMPARACION DE MODELOS

Los tres modelos dados en el apartado 2. obtienen la distribución prognóstica, para elegir uno de ellos proponemos dos técnicas, las cuales aplicaremos al ejemplo del apartado 3.

4.1. Comparación I.

La primera técnica sugerida se basa en definir una utilidad del modelo propuesto. Definimos esta utilidad a partir de la logarítmica debido a las buenas propiedades que esta verifica (ver /4/), así se define:

$$U = A\left(\frac{1}{n}\sum_{i} \ln p(y_i/x_i,D)\right) + B \qquad (4.1)$$

teniendo en cuenta:

- Como el logaritmo neperiano podría ir a $-\infty$, fijamos una cota ε, de forma que si algún $p(y_i/x_i,D)$ fuese menor que ε se sustituye por la cota.
- A y B se fijan de forma que la utilidad de una $U(0, \frac{1}{\varepsilon})$ valga cero, es decir, $A = -1 / \ln \varepsilon$, $B = 1$

Ejemplo

Con esta técnica se tiene para el ejemplo 3 que fijado $\varepsilon = 0.0001$ se tiene:

(MODELO I) = A(-7.10) + B = -0.3827

(MODELO II) = A(-6.95) + B = 0.3963

(MODELO III) = A(-7.31) + B = 0.3650

hay que hacer notar que ningún $p(y_i/x_i,D)$ de ningún modelo se alcanza la cota fijada.

A la vista de lo anterior concluimos, que según esta técnica para este ejemplo el modelo más adecuado es II.

4.2. Comparación II.

Otra forma que proponemos para comparar estos modelos, es aplicar los contrastes desde una perspectiva bayesiana (ver /6/) usando la distribución alfa. Ello permite ver a cual de las distribuciones prognósticas, obtenidas según cada MODELO, se ajustan mejor los datos. En concreto se trata de ver si los valores $\{v_1,\ldots,v_n\}$ pueden ser considerados como una muestra aleatoria simple de una uniforme, donde $v_i = F_i(y_i)$ y F_i es la función de distribución que corresponde a $p(y/x_i,D)$.

La distribución alfa (con la misma moda) es:

$$p(v/\theta) = \theta \exp\{-(1-\theta)\ln(1 - |1 - 2v|)\}, \quad \theta > 0$$

que es la uniforme si y sólo si $\theta = 1$.

La distribución a posteriori (con a priori la de referencia) es:

$$p(\theta/t): \quad \gamma(n,t) \text{ con } t = -\sum_i \ln(1 - |1 - 2v_i|)$$

y por lo tanto (ver /5/)

$$p(\ln \theta/t) \simeq N(\mu_t, \sigma_t) \text{ con } \mu_t = \ln \frac{n}{t} - \frac{1}{2n}$$

$$\sigma_t = \frac{1}{\sqrt{n}}$$

Por los razonamientos dados en /6/ compararemos los tres modelos mediante el coeficiente $H = \left|\frac{\mu_t}{\sigma_t}\right|$, con lo cual el mejor modelo será aquel con menor H.

Ejemplo

Para aplicar esta técnica al ejemplo 3. haremos lo siguiente:

- Evaluar la función de distribución en cada y_i, según los tres MODELOS. Exponemos la forma de evaluar la función de distribución en el MODELO I (análogo en los otros) según técnicas de Montecarlo.

  - Sea $p = F(y)$ el valor de la función de distribución en y, para una $St(A_{x(1)}, B_{x(1)}, C_{x(1)})$.

  - Generación de N valores $(St_j, j=1,...,N)$ de una $St(A_{x(1)}, B_{x(1)}, C_{x(1)})$. Cada uno de estos valores se obtiene a partir de un valor de la t de Student con $A_{x(1)}$ grados de libertad $(t_{A_{x(1)}})$, teniendo en cuenta:

$$St(A_{x(1)}, B_{x(1)}, C_{x(1)}) = \sqrt{C_{x(1)}}\, t_{A_{x(1)}} + B$$

A su vez cada $t_{A_{x(1)}}$ se obtiene como

$$t_{A_{x(1)}} = N(0,1) \Big/ \sqrt{\frac{x^2 A_{x(1)}}{A_{x(1)}}}$$

donde las Normales se obtienen a partir de números pseudoaleatorios de la U(0,1) mediante el teorema Central del Límite.

- Estimación de p por T donde:

$$T = \frac{1}{N} \sum_{j=1}^{N} \eta_j \quad \text{con} \quad \eta_j = \begin{cases} 1 & \text{si } St_j \leq y \\ 0 & \text{si } St_j > y \end{cases}$$

este estimador verifica $E(T) = p$ y $V(T) = \frac{p(1-p)}{N}$

- Elección del número N, acotando el error por la desigualdad de Tchebicheff:

Estimamos la varianza de T con una muestra piloto de tamaño arbitrario (por ejemplo 100):

$$\hat{p} = \frac{1}{100} \sum_{j=1}^{100} \eta_j \quad \text{con lo que } \hat{V}(T) = \frac{\hat{p}(1-\hat{p})}{N}$$

para ε y k constantes la desigualdad de Tchebicheff indica:

$$P_r \left\{ | T - p | > k\sigma \right\} \leq \frac{1}{k^2}$$

si hacemos $k\sigma = \varepsilon$ entonces $k^2 = \frac{\varepsilon^2}{\sigma^2}$

con lo que $Pr \left\{ | T - p | > \varepsilon \right\} \leq \frac{p(1-p)}{\varepsilon^2 N} < \text{COTA}$

Por lo tanto para ε y COTA fijados por nosotros tomaremos:

$$N > \left\{ p(1-p) \right\} \Big/ \left\{ \varepsilon^2 \, \text{COTA} \right\}$$

- Los resultados para el ejemplo 3 son:

	t	n	μ_t	σ_t	H
MODELO I	1473.6	80	-2.9197	0.1118	26.11
MODELO II	107.5	80	-0.3018	0.1118	2.69
MODELO III	1473.6	80	-2.9197	0.1118	26.11

NOTA: La estimación de la función de distribución puede evaluarse por cualquier otra técnica apropiada.

hay que notar que los modelos I y III coinciden por alcanzarse para todos los elementos la cota fijada para $1 - |1 - 2v_i|$, mientras que no se alcanza para ninguno en el MODELO II.

Se concluye, por lo tanto que en este caso real según esta técnica de evaluación de los modelos, es más apropiado el MODELO II para el cálculo de la distribución prognóstica.

5. DISCUSION

El objetivo de este trabajo es la comparación de distintos modelos propuestos para el cálculo de la distribución prognóstica. Para ello se sugieren dos técnicas de comparación, las cuales se aplican a un ejemplo concreto, resultando en ambos casos mejor el mismo modelo.

Se han realizado distintos algoritmos, que escritos en FORTRAN IV constituyen un software apropiado para solucionar estos problemas.

6. REFERENCIAS

/1/ AITCHINSON, J. Y DUNSMORE; I.R.: "Statistical Prediction Analysis". Cambridge: University Press. (1975)

/2/ ANDERSON, J.A.: "Separate sample logistic discrimination". Biometrika 59, pp. 19-35. (1972).

/3/ BERMUDEZ, J.D.: "Modelos de clasificación regulares". Tesis Doctoral. Universidad de Valencia. (1984).

/4/ BERNARDO, J.M.: "Expected information as expected utility". Ann. Statist. 7, pp. 686-690. (1979).

/5/ BERNARDO, J.M.: "Bioestadística. Una perspectiva bayesiana". Ed. Vicens-Vives. (1981).

/6/ BERNARDO, J.M.: "Contraste de modelos probabilísticos desde una perspectiva Bayesiana". Trab. Estadística, 33, pp.16-30. (1982).

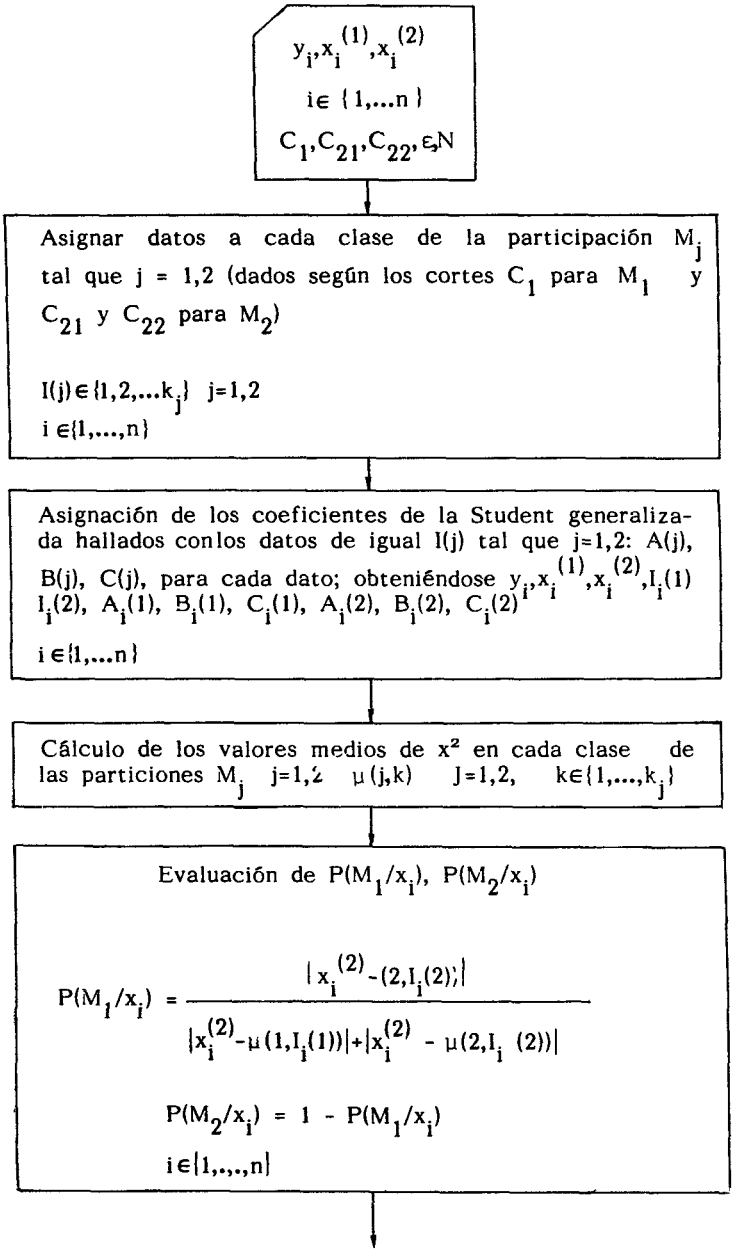
/7/ BERNARDO, J.M.: "Diagnóstico automático en Medicina". Estadíst. Española. (accepted). (1986).

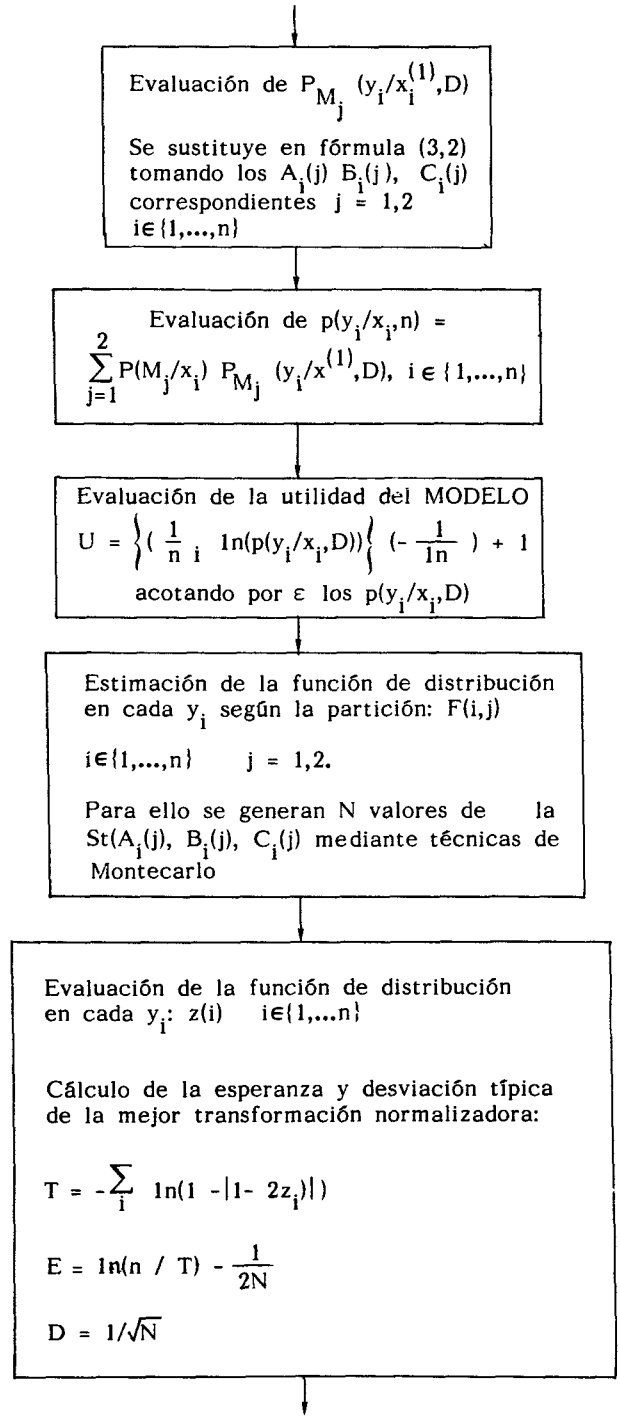
/8/ COX, D.R.: "The analysis of binary data". London. Methuen. (1970).

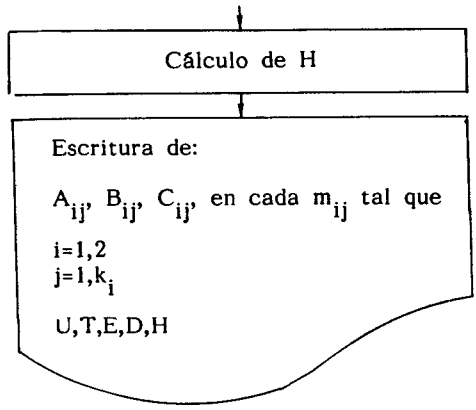
/9/ GARCIA-CARRASCO, P.: "Modelos para la asignación de tratamiento". Trabajos de Estadística (accepted). (1986).

7. APENDICE

Organigrama del MODELO II, con cálculo de la utilidad y H.


$y_i, x_i^{(1)}, x_i^{(2)}$
$i \in \{1,...n\}$
$C_1, C_{21}, C_{22}, \varepsilon N$
Asignar datos a cada clase de la participación M_j tal que j = 1,2 (dados según los cortes C_1 para M_1 y C_{21} y C_{22} para M_2)
$I(j) \in \{1,2,...k_j\}$ j=1,2
$i \in \{1,...,n\}$
Asignación de los coeficientes de la Student generalizada hallados conlos datos de igual I(j) tal que j=1,2: A(j), B(j), C(j), para cada dato; obteniéndose $y_i, x_i^{(1)}, x_i^{(2)}, I_i(1)$ $I_i(2)$, $A_i(1)$, $B_i(1)$, $C_i(1)$, $A_i(2)$, $B_i(2)$, $C_i(2)$
$i \in \{1,...n\}$
Cálculo de los valores medios de x^2 en cada clase de las particiones M_j j=1,2 $\mu(j,k)$ J=1,2, $k \in \{1,...,k_j\}$
Evaluación de $P(M_1/x_i)$, $P(M_2/x_i)$
$P(M_1/x_i) = \frac{|x_i^{(2)} - (2, I_i(2))|}{|x_i^{(2)} - \mu(1, I_i(1))| + |x_i^{(2)} - \mu(2, I_i(2))|}$
$P(M_2/x_i) = 1 - P(M_1/x_i)$
$i \in \{1,.,.,n\}$


Evaluación de $P_{M_j}(y_i/x_i^{(1)},D)$
Se sustituye en fórmula (3,2) tomando los $A_i(j)$ $B_i(j)$, $C_i(j)$ correspondientes $j = 1,2$ $i \in \{1,...,n\}$
Evaluación de $p(y_i/x_i,n) = \sum_{j=1}^{2} P(M_j/x_i)\, P_{M_j}(y_i/x^{(1)},D)$, $i \in \{1,...,n\}$
Evaluación de la utilidad del MODELO
$U = \left\{ \left(\frac{1}{n} \sum_i \ln(p(y_i/x_i,D)) \right\} \right. \left(- \frac{1}{\ln} \right) + 1$
acotando por ε los $p(y_i/x_i,D)$
Estimación de la función de distribución en cada y_i según la partición: F(i,j)
$i \in \{1,...,n\}$ $j = 1,2.$
Para ello se generan N valores de la $St(A_i(j), B_i(j), C_i(j)$ mediante técnicas de Montecarlo
Evaluación de la función de distribución en cada y_i: $z(i)$ $i \in \{1,...n\}$
Cálculo de la esperanza y desviación típica de la mejor transformación normalizadora:
$T = -\sum_i \ln(1 - |1 - 2z_i|)$
$E = \ln(n / T) - \frac{1}{2N}$
$D = 1/\sqrt{N}$


Cálculo de H
Escritura de:
A_{ij}, B_{ij}, C_{ij}, en cada m_{ij} tal que
i=1,2
j=1,k_i
U,T,E,D,H