VI Congreso Galego de Estatística e Investigación de Operacións
Vigo 5-7 de Novembro de 2003

ESTUDIO COMPARATIVO DE MODELOS FLEXIBLES DE DISCRIMINACIÓN DE LA CIRROSIS EN PACIENTES CON AFECTACIÓN HEPÁTICA

Isabel Fuentes Santos[1], Mónica López Ratón[1], Carmen Cadarso Suárez[1]
[1]Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela

1. INTRODUCCIÓN

Este trabajo es un estudio comparativo de los distintos modelos de discriminación de la enfermedad cirrosis en los pacientes aquejados de alguna dolencia hepática. Para ello se tomó como referencia una base de datos constituida por los historiales clínicos de 184 pacientes del Hospital Clínico Universitario en el servicio de Medicina Interna de Santiago de Compostela. Las variables de interés recogidas en dicha base se pueden agrupar en sociodemográficas (sexo y edad), de hábito (gramos de alcohol en sangre (grs), antigüedad en el hábito (ant) ,...) y clínicas (por ejemplo, la bilirrubina, los leucocitos, etc)

Hasta el momento los estudios que se habían realizado estaban basados en la aplicación de los Modelos Lineales Generalizados (GLM) de regresión logística (McCullagh-Nelder, 1989) de respuesta binaria (cirrosis si / no), considerando, pues, que todos los efectos eran paramétricos lineales. Primeramente se realizaba una regresión simple para detectar qué variables eran significativas , es decir, qué variables podían participar en el modelo de regresión logística múltiple y al ir introduciendo cada una de estas variables, se construía dicho modelo multivariante. De entre todas las variables, resultaban significativas las siguientes: la hemoglobina, el nivel plaquetario, la albúmina, la bilirrubina, la protombina, IGG e IGA y el modelo definitivo que se obtenía revelaba que las variables discriminadoras más importantes son la protombina, el nivel plaquetario, IGA, y la bilirrubina, además de plasmar también una interacción entre las variables plaquetas y albúmina. Este modelo tiene la ventaja de ser muy sencillo desde un punto de vista funcional, pero presenta el grave inconveniente de ser demasiado rígido al trabajar sobre una familia paramétrica determinada y esto puede estar ocultando la relación de cada efecto continuo con la variable respuesta.

En este trabajo se trata de aplicar herramientas más flexibles, concretamente los modelos de regresión spline (de Boor, 1976) y los Modelos Aditivos Generalizados (GAM) (Hastie-Tibshirani, 1990) que en vez de considerar todos los efectos paramétricos lineales (como en el caso de los GLM) permiten expresar los efectos mediante funciones no paramétricas suaves.

Así, mediante la aplicación de estas técnicas se construirá un modelo de regresión multivariante que compararemos con el modelo obtenido mediante los Modelos Lineales Generalizados (GLM).

2. DESCRIPCIÓN DEL ESTUDIO

En una primera etapa se realiza un estudio a nivel exploratorio, analizando la forma funcional de cada una de las variables que intervienen en el modelo, es decir, si actúan con efectos paramétricos lineales o si por el contrario, dichos efectos se expresan mediante funciones no paramétricas suaves. En el primer caso se aplicarán los Modelos Lineales Generalizados (GLM) de regresión logística (McCullagh-Nelder,1989) que no son más que modelos paramétricos generales de regresión multivariante con respuesta transformada Y; así, por ejemplo, si suponemos que se tienen p covariables (en nuestro caso 16) el modelo puede escribirse así:

$$\eta_X = g(\mu_X) = g\left[E\left[Y/X_1, X_2, ..., X_p\right]\right] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

con β el vector de parámetros y g una función monótona e invertible denominada función link

La estructura en S-Plus de dicho modelo es la siguiente:

Modelo.lm <- glm (Y~variable, family = binomial, data = biopsia, na.action= na.omit)

E implementaremos esta sentencia para cada una de las variables que asumamos como lineales.

Mientras que en el caso de que los efectos no sean lineales, se trabajará con los modelos de regresión spline (de Boor, 1976) y con los Modelos Aditivos Generalizados (GAM) (Hastie-Tibshirani, 1990) caracterizándose éstos últimos por ser una extensión paramétrica de los GLM y en los que las covariables son continuas. La regresión spline(aquí utilizaremos los "natural spline") se basa en una clase especial de funciones lineales paramétricas que aportan una mayor flexibilidad, se trata de dividir el rango de la covariable en k+1 regiones disjuntas (k es el número de puntos de corte o knots interiores en la covariable y nos indica el grado de flexibilidad que introducimos en el modelo) y en cada región ajustar la respuesta por regresión polinómica. Aquí la estructura es la siguiente:

Modelo.gam <- gam (Y ~ ns (variable, df), family = binomial, data = biopsia, na.action = na.omit)

En la sentencia anterior ns denota la técnica de los "natural spline" y df el número de grados de libertad, es decir, la flexibilidad del modelo. En consonancia con esto, cuando tengamos una variable con efecto no lineal se le aplicará el modelo anterior probando con distintos grados de libertad y mediante la tabla anova se deducirá cuál es el ajuste adecuado, es decir, cuál es el mejor valor para df.

A continuación se muestran los resultados obtenidos:

Primeramente en la tabla 1 se tiene el modelo válido para cada una de las variables que recoge el efecto de las mismas (lineal o no) y luego se observa gráficamente cuál es la forma funcional de las variables que no participan en el modelo con efecto lineal.

Tabla 1: efecto de las variables.

VARIABLE	**MODELO**	**DF**
EDAD	glm	
SEXO	gam	
ANTIQÜEDAD DEL HÁBITO	glm	
GRAMOS DE ALCOHOL	glm	
HEPATITIS	gam	
HEMOGLOBINA	glm	
LEUCOCITOS	gam	3
PLAQUETAS	glm	
VOLUMEN CORCUSPULAR MEDIO	gam	3
ALBÚMINA	gam	4
BILIRRUBINA	glm / gam (*)	4
PROTOMBINA	glm	
GOT	gam	3
GGPT	gam	8
IGG	glm	
IGA	gam	3

(*) La variable bilirrubina, al eliminar datos aislados pasa de tener efecto no lineal a lineal, por tanto vamos a estudiar su comportamiento en el modelo en ambos casos.

Veamos, gráficamente, el efecto de las variables no lineales, así como las categóricas:

Figura 1: Efecto de la variable sexo:

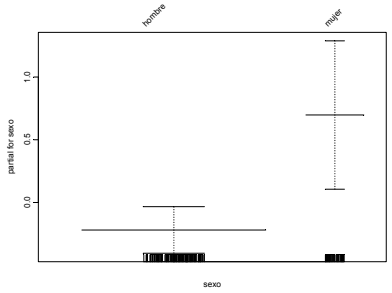
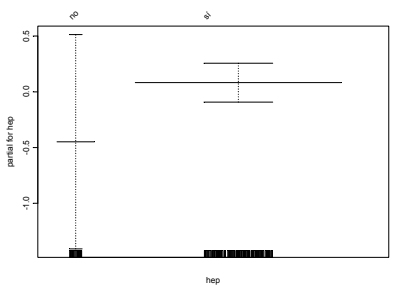


Figura 2: Efecto de la variable hepatitis:



Observando los intervalos de confianza del error en los modelos con variables categóricas, podemos deducir si éstas son significativas o no. Así, vamos a descartar la variable hepatitis dado que los intervalos se solapan. Por el contrario, la variable sexo si parece significativa.

Figura 3: Efecto de la variable leucocitos:

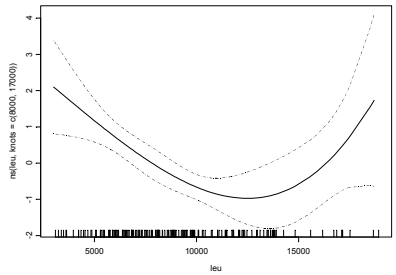


Figura 4: Efecto de la variable volumen corpuscular medio:

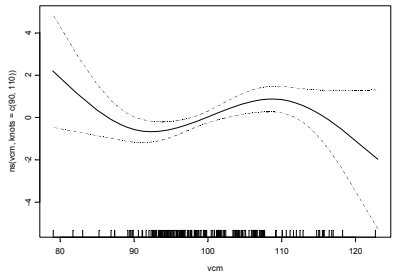


Figura 5: Efecto de la variable albúmina:

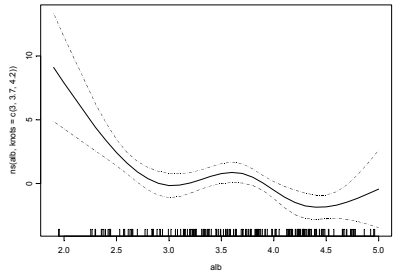


Figura 6: Efecto de la variable bilirrubina:

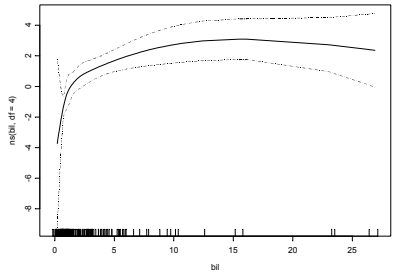


Figura 7: Efecto de la variable got:

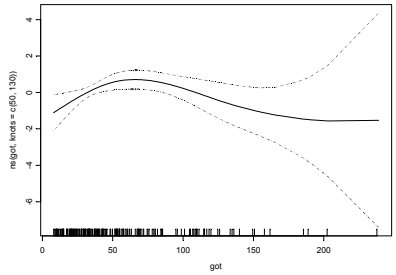


Figura 8: Efecto de la variable ggpt:

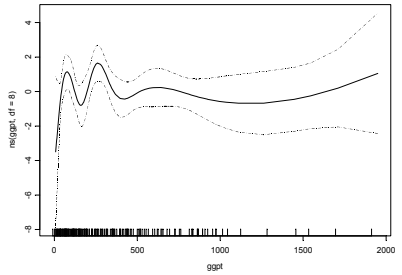
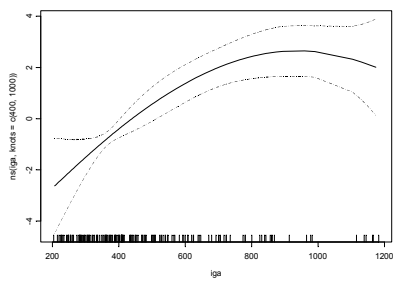


Figura 9: Efecto de la variable iga:



En una segunda y última etapa, se analiza qué variables son significativas, es decir, que variables podrían participar en el modelo de regresión logística múltiple; para las variables con efecto lineal se deduce la significación a partir del p-valor y para las variables no lineales se estudia su importancia mediante la disminución en términos de la "deviance" (ver tabla 2)

Tabla 2: variables lineales en el modelo.

VARIABLE	P- valor	SIGNIFICATIVA
EDAD	0.3361	NO
ANTIQÜEDAD DEL HÁBITO	0.1435	NO
GRAMOS DE ALCOHOL	0.04632	SI
HEMOGLOBINA	1,59E-04	SI
PLAQUETAS	6,39E-05	SI
BILIRRUBINA	3,04E-07	SI
PROTOMBINA	0	SI
IGG	1,42E-08	SI

Tabla 3 : variables no lineales

Variable	Null Deviance	Residual Deviance	SIGNIFICA
Leu	179,2467	160,5228	SI
Vcm	186,4957	176,4374	SI
Alb	186,4957	133,3819	SI
Got	183,6318	175,3956	SI
GGTP	181,9567	165,6774	SI
IGA	177,4159	131,6368	SI

En consonancia con los resultados obtenidos, se toman como variables "candidatas" a intervenir en nuestro modelo las siguientes: los gramos de alcohol, los leucocitos, el nivel plaquetario, el volumen corpuscular medio, la albúmina, la bilirrubina, la protombina y grasas como GOT, GGTP, IGG e IGA, y una vez determinadas dichas variables ya estamos en condiciones de construir el modelo que nos ocupa. Para ello se trata de ir introduciendo paso a paso cada una de las variables en el modelo y al mismo tiempo, ir analizando la importancia de cada nueva variable introducida, mediante el test " Chi cuadrado" (que va comparando ambos modelos, el modelo con y sin dicha variable), proceso que se refeja claramente en la figura 10. Una vez terminado el procedimiento anterior, se obtiene un modelo multivariante que detecta como variables discriminadoras más importantes la albúmina, el nivel plaquetario y la protombina sin posibles interacciones entre las mismas.

Si se compara este modelo con el que ya se tenía se encuentran notables diferencias, en resumen, se puede decir que las variables comunes a los mismos son el nivel plaquetario y la protombina (ambas con efecto lineal) y mientras que le modelo GLM reflejaba la importancia de la bilirrubina e IGA, el modelo actual detecta la significación de la albúmina (con efecto no paramétrico suave).

Figura 10: esquema de la construcción del modelo multivariante.