ESTADÍSTICA ESPAÑOLA
Vol. 37, Núm.140, 1995, págs. 321 a 348

Estadísticos para la detección de observaciones anómalas en modelos de elección binaria: una aplicación con datos reales

por
GREGORIO R. SERRANO(1)
Departamento de Economía Cuantitativa
Facultad de Ciencias Económicas
Universidad Complutense de Madrid
Campus de Somosaguas

RESUMEN

Este trabajo trata el problema de la detección de observaciones anómalas en modelos de elección binaria. Partiendo del estadístico propuesto en Gracia-Díez y Serrano (1994) que mide la influencia individual de cada observación sobre el vector de parámetros estimado, se derivan otros estadísticos que evalúan la influencia individual y de grupos de observaciones sobre i) el vector de probabilidades estimadas e ii) subconjuntos de parámetros y combinaciones lineales de los mismos. También, se generaliza el método de Peña y Yohai (1995) para la detección de observaciones enmascaradas en modelos lineales al caso de los modelos de elección binaria. Finalmente, se propone una estrategia de diagnosis para la detección de anomalías en

(1) En este trabajo se presentan algunos resultados de mi Tesis doctoral. Quiero expresar mi agradecimiento a Mercedes Gracia-Díez, Daniel Peña y a un evaluador anónimo por sus interesantes comentarios. Cualquier error es responsabilidad del autor.

este tipo de modelos. Esta estrategia se ilustra mediante su aplicación al modelo probit estimado por Dhillon et. al (1987).

Palabras clave: anomalías en modelos de elección binaria, detección, problema de enmascaramiento, estrategia de diagnosis.

1. INTRODUCCIÓN

El objetivo de este trabajo es contribuir al estudio del problema de la existencia de observaciones anómalas en modelos de elección binaria (MEB). En Gracia-Díez y Serrano (1994) se trata el problema de la detección de anomalías en este tipo de modelos y se muestra que, contrariamente a lo que se ha propuesto en la literatura anterior [ver Pregibon (1981), Jennings (1986), Williams (1987), Copas (1988) y Bedrick y Hill (1990) entre otros], el análisis de residuos no es una herramienta adecuada. Además, se propone un estadístico, computacionalmente eficiente, que mide la influencia de cada observación sobre la estimación del vector de parámetros de un MEB.

El presente trabajo se encuentra en la misma línea de investigación, abordando los siguientes aspectos:

1. Se deriva un nuevo estadístico que complementa al anterior y que se basa en medir la influencia de cada observación en el vector de probabilidades estimadas. Este estadístico es relevante en esta clase de modelos donde, debido a la no linealidad, el efecto de cada observación sobre las probabilidades estimadas no sólo depende del cambio en el vector de parámetros, sino también del valor que toman las probabilidades estimadas para cada observación individual.

2. Se generalizan los dos estadísticos de influencia individual (sobre el vector de parámetros y sobre el vector de probabilidades estimadas) para evaluar la influencia de grupos de observaciones. También, se derivan expresiones particulares para medir el peso de un grupo de observaciones sobre subconjuntos de parámetros y combinaciones lineales de los mismos.

3. Se trata el problema del enmascaramiento. El enmascaramiento se produce cuando grupos de observaciones anómalas disimulan el efecto individual de cada una de ellas, no pudiendo detectarse su presencia mediante estadísticos de influencia individual. Para ello, se adapta al caso de los MEB el método propuesto por Peña y Yohai (1995) para modelos lineales. Este método también resulta computacionalmente eficiente en este caso.

Seguidamente y dado que, al igual que ocurre en los modelos lineales de regresión, los estadísticos propuestos no siguen una distribución conocida, se sugiere

una estrategia de detección de anomalías en los MEB. Esta estrategia se basa fundamentalmente en la utilización de los estadísticos anteriores con un orden preestablecido y en la comparación en términos relativos de los valores que toman dichos estadísticos. Por último, la estrategia de detección propuesta se ilustra mediante su aplicación a la muestra de datos utilizada por Dhillon et. al. (1987), donde se estima un modelo probit para determinar la elección por parte de los individuos entre tipos de interés fijos o variables a la hora de contratar préstamos hipotecarios. La detección de anomalías en esta muestra y su posterior eliminación da lugar a resultados distintos a los obtenidos en el artículo original, donde se ignora la existencia de anomalías. Este ejemplo pone de relieve los sesgos que puede ocasionar la presencia de estas observaciones en la muestra.

El trabajo está organizado como sigue. En la sección 2 se derivan los estadísticos de influencia individual para evaluar el efecto de cada observación sobre el vector de parámetros estimados y sobre el vector de probabilidades estimadas. En la sección 3 se extienden los resultados anteriores a estadísticos para grupos de observaciones y a medidas de influencia sobre distintos subconjuntos y combinaciones lineales del vector de parámetros. En la sección 4 se generaliza el método de Peña y Yohai (1995) para la detección de observaciones enmascaradas en los MEB. En la sección 5 se presenta una estrategia general de diagnosis para la detección de observaciones anómalas en este tipo de modelos. En la sección 6 se ilustra la aplicación de estas técnicas a la diagnosis del modelo propuesto por Dhillon et. al. (1987). Por último, la sección 7 contiene las principales conclusiones y posibles extensiones del trabajo.

2. ESTADÍSTICOS DE INFLUENCIA: OBSERVACIONES INDIVIDUALES

En un MEB, la probabilidad de que un individuo elija una de las dos alternativas viene determinada por la expresión [Amemiya (1981)]:

$$P_i = P(y_i = 1) = F(x_i^T \beta) \quad i = 1,2,\ldots,n \qquad [1]$$

donde y_i es una variable binaria que toma valores cero o uno según la alternativa elegida por el individuo i-ésimo, $\boldsymbol{x}_i$ es el vector de k características relativas al individuo i, β es el correspondiente vector de parámetros y $F(\cdot)$ es una función de distribución, que usualmente es la normal estándar o la logística estándar.

En Gracia-Díez y Serrano (1994) se deriva un estadístico con objeto de evaluar el efecto de cada observación sobre la estimación MV de β en el modelo **[1]**. Este estadístico se basa en la distancia propuesta por Cook (1977) para modelos lineales, y puede escribirse:

$$\hat{c}_i = (\hat{\beta}-\hat{\beta}_{(i)})^T \hat{I}(\hat{\beta})(\hat{\beta}-\hat{\beta}_{(i)}) \quad i = 1,\ldots,n \qquad [2]$$

donde $\hat{\beta}_{(i)}$ denota la estimación MV de β eliminando la observación *i*-ésima e $\hat{I}(\hat{\beta})$ es la matriz de información del modelo evaluada en $\hat{\beta}$.

La evaluación eficiente del estadístico **[2]** requiere utilizar el hecho de que dada una estimación inicial $\hat{\beta}^{\tau}$, la estimación MV de β puede calcularse por el método de *scoring*, y que una iteración por dicho método puede obtenerse aplicando mínimos cuadrados ordinarios al modelo lineal [Amemiya (1981 y 1985)]:

$$\tilde{y}_i = \tilde{x}_i^T \beta + u_i \qquad [3]$$

donde:

$$\tilde{y}_i = \frac{y_i + \hat{f}_i x_i^T \hat{\beta}^{\tau} - \hat{F}_i}{[\hat{F}_i(1-\hat{F}_i)]^{1/2}} \qquad [4]$$

$$\tilde{x}_i = \frac{\hat{f}_i}{[\hat{F}_i(1-\hat{F}_i)]^{1/2}} x_i \qquad [5]$$

$$\hat{F}_i = F(x_i^T \hat{\beta}^{\tau}) \qquad \hat{f}_i = f(x_i^T \hat{\beta}^{\tau}) \qquad [6]$$

Suponiendo que la convergencia se produce en la iteración τ, es decir, $\hat{\beta}^{\tau} = \hat{\beta}$, es la estimación MV de β, una expresión computacionalmente eficiente del estadístico **[2]** viene dada por:

$$\hat{c}_i = \frac{e_i^{*2} \tilde{x}_i^T (\tilde{X}^T \tilde{X})^{-1} \tilde{x}_i}{[1-\tilde{x}_i^T (\tilde{X}^T \tilde{X})^{-1} \tilde{x}_i]^2} = \frac{e_i^{*2} \tilde{h}_i}{(1-\tilde{h}_i)^2} \qquad [7]$$

donde $e_i = y_i - \hat{P}_i$ es el residuo asociado a la observación *i*-ésima de un MEB y e_i^* es el residuo estandarizado definido por:

$$\frac{e_i}{[\hat{P}_i(1-\hat{P}_i)]^{1/2}} = \tilde{y}_i - \tilde{x}_i^T \hat{\beta} \qquad [8]$$

$\tilde{X}$ es una matriz con las *n* filas definidas por $\tilde{x}_i^T$, $\tilde{h}_{ij} = \tilde{x}_i^T (\tilde{X}^T \tilde{X})^{-1} \tilde{x}_j$ y específicamente $\tilde{h}_{ii} \equiv \tilde{h}_i$. Además, se cumple $\hat{I}(\hat{\beta}) = (\tilde{X}^T \tilde{X})$.

Seguidamente se deriva un estadístico alternativo al anterior, que se basa en medir la influencia de cada observación sobre el vector de probabilidades estimadas. En los modelos de regresión lineales resulta idéntico medir la influencia de una observación en el vector de parámetros estimados que en el vector $\hat{y}$ de valores ajustados [Peña (1987)]. Sin embargo, en los MEB este resultado no se mantiene, por lo que es necesario derivar un estadístico específico que mida la influencia de la observación *i*-ésima sobre el vector de probabilidades estimadas ($\hat{P}$). Utilizando una expansión de Taylor de primer orden, para el caso en que se elimina la *i*-ésima observación, la probabilidad estimada se puede aproximar de la siguiente forma:

$$F(x_i^T \hat{\beta}_{(i)}) \approx F(x_i^T \hat{\beta}) + f(x_i^T \hat{\beta}) x_i^T (\hat{\beta}_{(i)} - \hat{\beta}) \qquad [9]$$

A partir de **[9]**, la probabilidad estimada para la observación *j*-ésima al eliminar la observación *i*-ésima puede calcularse como:

$$F(x_j^T \hat{\beta}_{(i)}) \approx F(x_j^T \hat{\beta}) + f(x_j^T \hat{\beta}) x_j^T (\hat{\beta}_{(i)} - \hat{\beta}) \qquad [10]$$

y la diferencia entre las probabilidades estimadas para toda la muestra resulta:

$$\hat{F} - \hat{F}_{(i)} = -\hat{\psi} X (\hat{\beta}_{(i)} - \hat{\beta}) = -\hat{\Psi}^{1/2} \tilde{X} (\hat{\beta}_{(i)} - \hat{\beta}) \qquad [11]$$

donde $\hat{\psi}$ y $\hat{\Psi}$ son matrices diagonales de dimensión *n* con elemento genérico $\hat{f}_i$ y $\hat{F}_i(1-\hat{F}_i)$ respectivamente.

Por lo tanto, una medida de influencia sobre las probabilidades estimadas es:

$$\begin{aligned} \hat{c}_i(P) &= (\hat{F} - \hat{F}_{(i)})^T (\hat{F} - \hat{F}_{(i)}) \\ &= \frac{e_i^{2*} \bar{h}_i}{(1 - \bar{h}_i)^2} \end{aligned} \qquad [12]$$

donde, en general:

$$\bar{h}_{ij} = \tilde{x}_i^T (\tilde{X}^T \tilde{X})^{-1} (\tilde{X} \hat{\Psi} \tilde{X}) (\tilde{X}^T \tilde{X})^{-1} \tilde{x}_j \quad \text{y} \quad \bar{h}_i \equiv \bar{h}_{ii} \qquad [13]$$

La diferencia entre los estadísticos **[7]** y **[12]** se encuentra en la matriz $\hat{\Psi}$ de la expresión **[13]**. La presencia de esta matriz se debe a que en un modelo no lineal, el cambio en los argumentos de la función no tiene por qué ser igual a la variación en la función. Más concretamente, lo que indica la expresión **[11]** es que el cambio en las probabilidades estimadas no sólo depende del cambio en los parámetros,

sino de la situación inicial de dicha probabilidad, siendo mayor el cambio cuanto más próxima se encuentre a 0.5. Es importante señalar que el empleo del estadístico **[12]** es especialmente importante en situaciones en las que el objetivo sea realizar previsiones agregadas para poblaciones grandes.

Es posible realizar una comparación entre los valores de ambos estadísticos. El cociente entre $\hat{c}_i$ (P) y $\hat{c}_i$ está acotado entre el menor y el mayor autovalor de la matriz $\hat{\Psi}$ [Rao (1973)], es decir:

$$\lambda_{min}(\hat{\Psi}) \leq \frac{(\hat{\beta}-\hat{\beta}_{(i)})^T \tilde{X}^T \hat{\Psi} \tilde{X}(\hat{\beta}-\hat{\beta}_{(i)})}{(\hat{\beta}-\hat{\beta}_{(i)})^T \tilde{X}^T \tilde{X}(\hat{\beta}-\hat{\beta}_{(i)})} \leq \lambda_{max}(\hat{\Psi}) \qquad [14]$$

Dado que $\hat{\Psi}$ es una matriz diagonal con elemento característico $P_i(1-P_i)$, ocurre que $\lambda_{min}(\hat{\Psi}) \geq 0$ y $\lambda_{max}(\hat{\Psi}) \leq 1/4$. Por lo tanto, la razón entre los estadísticos resulta:

$$0 \leq \frac{\hat{c}_i(P)}{\hat{c}_i} \leq \frac{1}{4} \qquad [15]$$

El resultado anterior indica que el estadístico propuesto siempre tiene un valor inferior a $\hat{c}_i$. Este cociente alcanza su valor máximo para observaciones que, teniendo una escasa influencia en el vector de parámetros estimados, presentan una influencia importante en el vector de probabilidades estimadas, aunque generalmente ambos presentan un comportamiento monótono. Además, partiendo de la expresión **[12]** es evidente que el estadístico $\hat{c}_i$ (P) está acotado entre 0 y *n*, lo que puede facilitar el análisis de su distribución empírica.

En Gracia-Díez y Serrano (1994) se presentan simulaciones para evaluar el funcionamiento del estadístico $\hat{c}_i$. Respecto al estadístico $\hat{c}_i$ (P) también se han realizado experimentos con datos simulados con el objeto de comparar el funcionamiento de ambos estadísticos. Aunque en este trabajo no se presentan todas las simulaciones realizadas, es importante señalar que:

– Ambos estadísticos presentan un comportamiento análogo en cuanto a que detectan observaciones anómalas de forma similar.

– En general, $\hat{c}_i$ (*P*) es más sensible a clasificar como anómalas observaciones que no lo son en muestras sin observaciones anómalas. Sin embargo, cuando existen observaciones anómalas, $\hat{c}_i$ (P) tiende a clasificar como anómalas menos observaciones que $\hat{c}_i$, obteniendo tasas de acierto similares (observaciones que son realmente anómalas dentro del conjunto de las clasificadas como anómalas).

Por lo tanto, parece que $\hat{c}_i$ (P) tiene un mejor comportamiento que $\hat{c}_i$ en el sentido de que tiende a eliminar menos observaciones buenas.

A modo de ilustración, en las tablas 2.1 y 2.2 se presenta un resumen de los resultados obtenidos mediante datos simulados. En ambos casos se ha utilizado un modelo probit con constante y una sola variable explicativa del que se generaron 500 repeticiones con muestras de tamaño 200. Cada una de estas muestras incluye una proporción conocida de observaciones anómalas en la muestra (denotada por *w*). Se han considerado dos fuentes de procedencia de las observaciones anómalas: el caso I que supone observaciones con idéntico vector de parámetros pero varianza muy superior al de las observaciones generadas por el modelo considerado, y el caso II en el que se generaron observaciones extremas de la variable explicativa.

TABLA 2.1

RESULTADOS CON LOS ESTADÍSTICOS $\hat{c}_i$ Y $\hat{c}_i$ (*P*). CASO I

	$\hat{c}_i$		$\hat{c}_i$ (*P*)	
w%	Buenas	Bien previstas	Buenas	Bien previstas
0	181.1	157.0	182.9	156.9
5	181.3	155.6	182.8	155.4
10	181.5	153.7	182.8	153.4
15	181.8	152.2	182.9	152.1
20	181.8	150.9	182.7	150.7
25	181.7	149.2	182.3	148.9
30	181.3	147.6	181.9	147.5

TABLA 2.2

RESULTADOS CON LOS ESTADÍSTICOS $\hat{c}_i$ Y $\hat{c}_i$ (*P*). CASO II

	$\hat{c}_i$		$\hat{c}_i$ (*P*)	
w%	Buenas	Bien previstas	Buenas	Bien previstas
0	181.4	156.4	182.9	156.2
5	182.9	159.0	184.2	159.0
10	183.7	160.9	185.0	160.8
15	184.3	162.5	185.6	162.6
20	185.2	163.9	186.3	164.0
25	185.6	165.3	186.4	165.3
30	185.9	166.7	186.6	166.8

Dado que $\hat{c}_i$ y $\hat{c}_i$ (P) no siguen una distribución conocida, para poder llevar a cabo la simulación se utilizan valores críticos que clasifican como anomalías un porcentaje fijo α de observaciones en muestras en las que no existen observaciones de esta clase. Los resultados de las tablas 2.1 y 2.2 corresponden a un valor de α igual a 10%. En las tablas se presenta el número de observaciones consideradas como buenas por cada uno de los estadísticos y el número de observaciones dentro de las consideradas como buenas que el modelo es capaz de prever correctamente tras la eliminación de las observaciones que superaban el valor crítico prefijado.

– Como se puede apreciar, en ambos casos $\hat{c}_i$ (P) retiene un número ligeramente mayor de observaciones en la muestra, mientras que el número de las correctamente clasificadas es prácticmente igual para ambos estadísticos.

– El mejor comportamiento de $\hat{c}_i$ (P) comparado con $\hat{c}_i$ es más evidente en el caso II que en el caso I, aunque la ventaja de $\hat{c}_i$ (P) queda diluída cuando el número de anomalías aumenta.

3. ESTADÍSTICOS DE INFLUENCIA: GRUPOS DE OBSERVACIONES Y OTROS CASOS PARTICULARES

Como se ha expuesto en el apartado anterior, a partir de la linealización del modelo binario en **[3]**, se pueden derivar estadísticos de influencia individual semejantes a los desarrollados para el modelo lineal general [Atkinson (1985)]. Siguiendo en esta línea, es posible particularizar los resultados anteriores a situaciones en las que se eliminan conjuntos de observaciones, así como evaluar el efecto de dichas observaciones para un subconjunto de parámetros del modelo.

A partir de **[3]**, una iteración por el algoritmo de *scoring* eliminando la información de las observaciones del conjunto *I* puede escribirse como:

$$\hat{\beta}_{(I)} = \hat{\beta} + \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T\left[I_p - \tilde{X}_I\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T\right]^{-1}\left(\tilde{X}_I\hat{\beta} - \tilde{y}_I\right) \qquad [16]$$

donde $\tilde{X}$ e $\tilde{y}$ están formadas por las observaciones transformadas como en **[4]**-**[5]** y la inversa de la matriz de información es:

$$\left(\tilde{X}_{(I)}^T\tilde{X}_{(I)}\right)^{-1} = \left(\tilde{X}^T\tilde{X}\right)^{-1} + \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T\left[I_p - \tilde{X}_I\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T\right]^{-1}\tilde{X}_I\left(\tilde{X}^T\tilde{X}\right)^{-1} \qquad [17]$$

A partir de **[16]**, la influencia de un conjunto de observaciones puede evaluarse utilizando un estadístico similar a **[7]**, que resulta:

$$\begin{aligned} \hat{c}_I &= (\hat{\beta}_{(I)}-\hat{\beta})^T\left(\tilde{X}^T\tilde{X}\right)(\hat{\beta}_{(I)}-\hat{\beta}) \\ &= e_I^{*T}(I_p-\tilde{H}_I)^{-1}\tilde{H}_I(I_p-\tilde{H}_I)^{-1}e_I^* \end{aligned} \qquad [18]$$

donde $\tilde{H}_I = \tilde{X}_I\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T$.

De manera análoga al caso de la influencia de observaciones individuales, es posible obtener un estadístico alternativo al de la expresión **[18]**, que se base en medir el cambio en el vector de probabilidades estimadas, en lugar de evaluar la diferencia entre los vectores de parámetros. Esta medida es:

$$\begin{aligned} \hat{c}_I(P) &= (\hat{\beta}_{(I)}-\hat{\beta})^T(\tilde{X}^T\hat{\Psi}\tilde{X})(\hat{\beta}_{(I)}-\hat{\beta}) \\ &= e_I^{*T}(I_p-\tilde{H}_I)^{-1}\overline{H}_I(I_p-\tilde{H}_I)^{-1}e_I^* \end{aligned} \qquad [19]$$

donde $\overline{H}_I = \tilde{X}_I\left(\tilde{X}^T\tilde{X}\right)^{-1}(\tilde{X}^T\hat{\Psi}X)\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}_I^T$.

Por último, se derivan expresiones particulares para evaluar la influencia de grupos de observaciones sobre un subconjunto de parámetros del modelo. El interés se centra en *m* elementos que se corresponden con las filas del vector $\theta = R^T\beta$, donde R^T es una matriz m×k de constantes conocidas con rango(R) = $m \leq k$, y la matriz de varianzas de la estimación máximo-verosímil de θ es $R^T[I(\hat{\beta})]^{-1}R$. Una medida de influencia análoga al estadístico **[18]** aplicada a una combinación lineal de los parámetros originales es:

$$\begin{aligned} \hat{c}_I(\theta) &= (\hat{\theta}_{(I)}-\hat{\theta})^T\left[R^T\left(\tilde{X}^T\tilde{X}\right)^{-1}R\right]^{-1}(\hat{\theta}_{(I)}-\hat{\theta}) \\ &= e_I^{*T}(I_p-\tilde{H}_I)^{-1}\tilde{X}_I\tilde{N}\,\tilde{X}_I^T(I_p-\tilde{H}_I)^{-1}e_I^* \end{aligned} \qquad [20]$$

donde:

$$\tilde{N} = \left(\tilde{X}^T\tilde{X}\right)^{-1}R\left[R^T\left(\tilde{X}^T\tilde{X}\right)^{-1}R\right]^{-1}R^T\left(\tilde{X}^T\tilde{X}\right)^{-1} \qquad [21]$$

que es sencillo de calcular puesto que $\tilde{N}$ no depende de la observación eliminada. En particular, el efecto de eliminar una sola observación sobre θ puede escribirse:

$$\hat{c}_i(\theta) = \frac{e_i^{*2}\tilde{x}_i^T\tilde{N}\,\tilde{x}_i}{(1-\tilde{h}_i)^2} \qquad [22]$$

El estadístico en **[20]** es especialmente interesante, debido a que en los MEB es frecuente la presencia de un cierto número de variables cualitativas entre las variables explicativas. Por lo tanto, en muchas ocasiones, es importante medir el efecto de un grupo de observaciones sobre un subconjunto de parámetros. Sin pérdida de generalidad, se puede suponer que los parámetros de interés son los últimos *m* componentes del vector β. En ese caso:

$$R^T\beta = (0_{mx(k-m)} \mid I_m)\beta = (\beta_{k-m+1},\ldots,\beta_k)^T \qquad [23]$$

y el estadístico **[20]** puede escribirse:

$$\hat{c}_I(\theta) = \hat{c}_I - e_I^{*T}(I_p-\tilde{H}_I)^{-1}\tilde{G}_I(I_p-\tilde{H}_I)^{-1}e_I^* \qquad [24]$$

donde $\tilde{G}_I = \tilde{X}_2(\tilde{X}_2^T\tilde{X}_2)^{-1}\tilde{X}_2$ y $\tilde{X}_2$ es la submatriz formada por las últimas *m* columnas de $\tilde{X}$. Cuando sólo se elimina una observación, la expresión anterior queda simplificada a:

$$\hat{c}_i(\theta) = \frac{e_i^{*2}(\tilde{h}_i-\tilde{g}_i)}{(1-\tilde{h}_i)^2} \qquad [25]$$

donde g_i es el elemento *i*-ésimo de la diagonal principal de G_I.

El estadístico **[25]** puede particularizarse para el caso en que se desee medir el efecto de una observación en la estimación de un parámetro β_j del vector β. Denotando por v_j el elemento *j*-ésimo de la diagonal principal de la matriz $\hat{I}(\hat{\beta})^{-1}$, es inmediato que el estadístico:

$$\hat{c}_i(\beta_j) = \frac{(\hat{\beta}_{(i)j}-\hat{\beta}_j)^2}{v_j} \qquad [26]$$

proporciona una medida del desplazamiento que experimenta la estimación del coeficiente β_j cuando se elimina de la muestra la observación *i*-ésima.

4. EL PROBLEMA DE ENMASCARAMIENTO

El problema de enmascaramiento se produce cuando la muestra incluye un grupo de observaciones tales que su influencia conjunta disimula el efecto individual

de cada una de ellas, provocando que éste no sea detectado mediante el uso de los estadísticos que analizan una observación cada vez. Esta clase de grupos de observaciones pueden presentar patrones muy diferentes, aunque el efecto de su presencia es que los estadísticos de influencia individual no son capaces de detectarlas.

Los estadísticos para grupos de observaciones, derivados en la sección 3, pueden utilizarse para analizar la influencia de cualquier conjunto de observaciones. Sin embargo, la dificultad estriba en determinar *eficientemente* los grupos cuya influencia se pretende medir, entendiendo por *eficiente* cualquier método que, proporcionando los resultados perseguidos, no requiera la exploración exhaustiva de todos los grupos de distintos tamaños posibles de observaciones.

En esta sección, se extiende al caso de los MEB la estrategia desarrollada por Peña y Yohai (1995) para tratar este problema en los modelos lineales. Dicho método se basa en un análisis de los autovalores de una *matriz de influencia M*, que se define como la matriz de covarianzas no centrada de un conjunto de vectores que miden el efecto sobre el ajuste que tiene la eliminación de cada observación. En el modelo lineal general, el elemento característico de la matriz de influencia M viene dado por $m_{ij} = (\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(j)})$, donde $\hat{y}_{(i)}$ e $\hat{y}_{(j)}$ son los vectores de valores ajustados en el modelo lineal cuando se eliminan las observaciones i-ésima y j-ésima respectivamente. Este método ha sido probado en diferentes conjuntos de datos y funciona adecuadamente. Además, es computacionalmente eficiente comparado con otros procedimientos propuestos en la literatura para modelos lineales(2).

Para construir una matriz análoga $\tilde{M}$ para detectar conjuntos de observaciones influyentes en MEB, tan sólo es necesario aplicar el procedimiento propuesto por Peña y Yohai (1995) a la ecuación **[3]**. Si se denota por $\tilde{y}_j$ el valor ajustado para lo observación j mediante la ecuación **[3]**, y por $\tilde{y}_{(i)j}$ el valor ajustado cuando se elimina la observación i-ésima, entonces:

$$\tilde{y}_j \hat{} - \tilde{y}_{(i)j} \hat{} = \frac{\dot{e}_i \tilde{h}_{ij}}{1 - \tilde{h}_i} \qquad [27]$$

(2) Rousseeuw y van Zomeren (1990) proponen una estrategia completamente distinta, basada en la búsqueda de elipsoides de confianza de volumen mínimo. La idea básica es caracterizar un elipsoide tal que, minimizando el volumen, deje fuera a un número reducido de observaciones. Aunque es un planteamiento atractivo, el mayor inconveniente se debe a que resulta muy costoso en términos de cálculo, ya que para tener la seguridad de que se ha encontrado el elipsoide óptimo es necesario llevar a cabo una búsqueda exhaustiva en el espacio de variables explicativas [Véase la discusión que sigue al trabajo citado].

Por lo tanto, definiendo $\hat{\tilde{y}}_{(i)} = \left(\hat{\tilde{y}}_{(i)1} \cdots \hat{\tilde{y}}_{(i)n}\right)^T$, el vector $\tilde{t}_i = \hat{\tilde{y}} - \hat{\tilde{y}}_{(i)}$ resume el efecto sobre el ajuste de eliminar la observación *i*.

Entonces, para detectar conjuntos de posibles observaciones influyentes con efecto similar u opuesto, se puede analizar la matriz de covarianzas no centrada de $\tilde{t}_i$. Si se define la matriz $\tilde{T}$ de dimensión *n*×*n*, cuyas columnas son los vectores $\tilde{t}_i$, se puede definir la matriz de influencia $\tilde{M} = \tilde{T}^T\tilde{T}$, que tiene dimensión *n*×*n*, y con elemento genérico:

$$\tilde{m}_{ij} = \tilde{t}_i^T \tilde{t}_j = \frac{e_i^* e_j^* \tilde{h}_{ij}}{(1-\tilde{h}_i)(1-\tilde{h}_j)} \qquad [28]$$

Obsérvese que la diagonal principal de $\tilde{M}$ está formada por el estadístico **[7]** para cada observación. También conviene destacar que la medida $\hat{c}_l$ de **[18]** puede aproximarse utilizando $\hat{c}_l \approx \sum_{i \in l} \sum_{j \in l} \tilde{m}_{ij}$.

Teniendo en cuenta que el rango de la matriz de influencia es *k*, el procedimiento para identificar conjuntos de observaciones influyentes puede llevarse a cabo en la práctica mediante los siguientes pasos:

Paso 1: Calcular los autovectores correspondientes a los *k* autovalores no nulos de la matriz de influencia $\tilde{M}$.

Paso 2: Utilizando los autovectores asociados a los *m* autovalores mayores, seleccionar los pares de conjuntos de observaciones I_j^1 y I_j^2, j = 1,...,*m* ≤ *k*, incluyendo en cada uno de ellos las observaciones cuyo componente del autovector sea grande y positivo (efecto similiar en el ajuste) o negativo (efecto opuesto en el ajuste) respectivamente.

Paso 3: Empleando los estadísticos para evaluar la influencia de grupos de observaciones la sección 3, determinar los grupos de observaciones influyentes.

De manera análoga a como se ha llevado a cabo para el estadístico **[7]**, es posible obtener expresiones alternativas, que se basen en medir el cambio en el vector de probabilidades estimadas con objeto de detectar observaciones con problemas de enmascaramiento. En particular, la matriz de influencia alternativa tiene como elemento genérico en este caso:

$$\bar{m}_{ij} = \frac{e_i^* e_j^* \bar{h}_{ij}}{(1-\tilde{h}_i)(1-\tilde{h}_j)} \qquad [29]$$

donde $\bar{h}_{ij}$ está definido en **[13]**.

5. DETECCIÓN DE OBSERVACIONES INFLUYENTES EN LOS MEB

A partir de los estadísticos expuestos, se puede desarrollar una estrategia de diagnosis de observaciones anómalas para los modelos de elección binaria, que se puede resumir de la siguiente forma.

En una primera etapa, es necesario utilizar instrumentos *a priori* como los desarrollados para los modelos lineales, y que se basan en medir la distancia de cada observación al centro del espacio de las variables explicativas [Peña (1987), Belsley et al. (1981)], así como el estadístico $h_i = x_i^T (X^T X)^{-1} x_i$, para evaluar la dispersión de las observaciones muestrales, lo que proporciona información sobre potenciales observaciones extremas. Aunque valores moderadamente altos de estos estadísticos no son concluyentes, permiten fijar la atención sobre algunas observaciones en fases posteriores. Hay que tener en cuenta que estos estadísticos sólo son válidos para variables continuas. Esto supone una limitación importante ya que, trabajando con modelos de elección discreta, resulta frecuente la presencia de variables exógenas cualitativas.

Una vez estimado el modelo, también es conveniente utilizar el estadístico $\tilde{h}_i = \tilde{x}_i^T \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{x}_i$ que utiliza las variables transformadas en lugar de las originales. Aunque éste tiene la misma interpretación que h_i, puede contener información diferente debido a la no linealidad del modelo. El empleo de $\tilde{h}_i$ sigue los mismos criterios; esto es, localizar observaciones extremas y potencialmente influyentes.

El siguiente paso es calcular los estadísticos $\hat{c}_i$ y $\hat{c}_i(P)$ definidos en **[7]** y en **[12]**. Valores elevados de alguno de ellos indicarán la presencia de observaciones influyentes. En este sentido, la principal dificultad estriba en que al igual que ocurre en los modelos lineales, no existen formas concluyentes de evaluar valores críticos para dicho estadístico. No obstante, sí se pueden obtener valores indicativos utilizando las tablas de la distribución χ^2_k [Gracia-Díez y Serrano (1994)]. Además, lo más importante es la comparación de efectos relativos, por lo que resulta recomendable realizar comparaciones de las estimaciones obtenidas con toda la muestra y las que resultan de eliminar un conjunto, usualmente pequeño, formado por aquellas observaciones con un valor elevado del estadístico respecto a la media del mismo para toda la muestra.

Una vez determinado un conjunto de observaciones individualmente influyentes, conviene analizar la posibilidad de que existan observaciones *enmascaradas* utilizando la adaptación a los MEB del procedimiento propuesto por Peña y Yohai (1995) que se ha descrito en la sección 4, cuya influencia se mide a través del estadístico $\hat{c}_I$ en **[18]**. Una vez más y lo mismo que para el caso de los modelos lineales, no se pueden ofrecer valores críticos concluyentes para dichos estadísticos, por lo que, en parte, las decisiones finales sobre los grupos de observaciones

problemáticas dependen fundamentalmente del criterio del analista, así como de las posibles causas de la presencia de observaciones anómalas.

Por último, es importante señalar que el conjunto de instrumentos de detección que se acaba de proponer revela la presencia de observaciones influyentes en la muestra, aunque ninguno de ellos indica el posible origen de las mismas. Bajo determinados supuestos sobre la fuente de procedencia de las observaciones anómalas, en Serrano (1993) se derivan estadísticos que permiten contrastar el origen de las mismas. Estos contrastes pueden utilizarse para, una vez detectadas las observaciones influyentes, determinar la fuente que causa la anomalía, suponiendo que ésta sea única. No obstante, las primeras pruebas realizadas parecen indicar una baja potencia de estos estadísticos. En los experimentos de Montecarlo realizados, no resultaron concluyentes sobre el tipo de anomalía en ningún caso, aunque sí resultaban suficientemente potentes como para contrastar la hipótesis nula de que las observaciones seleccionadas como potencialmente anómalas no habían sido generadas por el mismo proceso que las restantes.

6. APLICACIÓN: UN MODELO DE ELECCIÓN DE TIPO DE INTERÉS FIJO FRENTE A TIPO DE INTERÉS VARIABLE

En una nota en el *Journal of Money, Credit and Banking*, Dhillon et al. (1987) estudian el conjunto de características personales y financieras que determinan la elección por parte de los individuos de un tipo de interés fijo frente a uno variable a la hora de contratar sus préstamos hipotecarios. El artículo utiliza un modelo probit para determinar los principales factores que influyen en la decisión. El interés principal del estudio se centra en contrastar las dos posturas que dominan los planteamientos teóricos sobre el tema. La primera, pone de relieve la independencia de las características personales del prestatario en la elección del tipo, dados los precios y los términos del contrato(3). La segunda, supone información asimétrica; esto es, dadas las condiciones del mercado, los prestatarios pueden favorecerse no revelando sus características personales a la hora de firmar el contrato(4).

(3). La información es simétrica, y el efecto de las características personales ya se encuentra incluido en los términos del contrato.

(4). La información asimétrica supone que existen características personales que, de conocerse, podrían perjudicar al prestatario.

6.1. Descripción del modelo

El modelo básico utilizado por Dhillon et al. (1987) relaciona la probabilidad de que un individuo elija, dadas sus características personales y las condiciones del mercado, un tipo de interés variable para un préstamo hipotecario.

La muestra está compuesta por 78 clientes de un banco hipotecario de Louisiana (EE.UU.)(5). Los préstamos fueron concedidos durante el período comprendido entre enero de 1983 y febrero de 1984. Del total de observaciones, 46 eligieron un tipo de interés fijo y 32 un tipo de interés variable no acotado. Todos los préstamos a interés fijo tenían un plazo de vencimiento de 30 años. Las variables disponibles aparecen definidas en la Tabla 6.1 y un listado completo de los datos utilizados se incluye en el Apéndice A.

TABLA 6.1

VARIABLES EN EL MODELO DE ELECCIÓN DE TIPOS DE INTERÉS

Variable dependiente	
ADJ	Ficticia, el individuo elige tipo de interés variable, 1 = Sí.
Variables exógenas de condiciones de mercado y características del contrato	
FI	Tipo de interés fijo.
MAR	Margen sobre el tipo de interés variable.
YLD	Diferencia entre el tipo de interés del Tesoro a 10 años menos el de 1 año.
PTS	Ratio entre el tipo de interés fijo y variable.
MAT	Ratio entre los vencimientos de los préstamos hipotecarios con tipo variable y fijo.
Variables exógenas de características personales	
BA	Edad del prestatario.
BS	Años de escolarización del prestatario.
FTB	Ficticia, el prestatario compra vivienda por primera vez, 1 = Sí.
CB	Ficticia, existe un co-prestatario, 1 = Sí.
MC	Ficticia, el prestatario está casado, 1 = Casado.
SE	Ficticia, el prestatario trabaja por cuenta propia, 1 = Sí.
MOB	Movilidad: años en la dirección actual.
Variables exógenas de características económicas	
NW	Riqueza neta del prestatario.
LA	Activos líquidos.
STL	Compromisos del prestatario a corto plazo.

(5). Los datos empleados son los que acompañan, en soporte magnético, al libro de Lott y Ray (1992). Aunque estos autores afirman que se trata de los datos utilizados por Dhillon et al. (1987), los resultados son semejantes a los obtenidos en el artículo original, aunque no idénticos, lo que puede atribuirse a los distintos algoritmos de estimación utilizados.

Los autores especifican un modelo probit no restringido utilizando todas las variables disponibles (Modelo 3) y dos versiones restringidas del mismo. El Modelo 1 excluye las variables LA y STL, con el fin de contrastar la significación de dichas variables económicas personales, mientras que el Modelo 2 excluye las variables de características personales, con el propósito de contrastar la hipótesis de información asimétrica. Una variable fundamental en el trabajo es la prima de riesgo, medida por la diferencia de los tipos del Tesoro a diez y un año (YLD).

6.2. Resultados empíricos con los modelos originales

En la Tabla 6.2 aparecen los resultados de la estimación de los tres modelos considerados por los autores del trabajo. Para ello se ha utilizado el método de máxima verosimilitud por procedimientos lineales expuesto la sección 2. Debajo del coeficiente asociado a cada variable figura la desviación típica estimada. En las tres últimas filas de la tabla se ofrecen, para cada modelo, el logaritmo de la función de verosimilitud, el valor del estadístico de contraste de razón de verosimilitudes (bajo el modelo restringido) y el número condición de la matriz de varianzas-covarianzas estimada(6).

Siguiendo a Dhillon et al. (1987), y a la vista de los resultados de la Tabla 6.2, se pueden hacer las siguientes consideraciones:

– Las variables de precio resultan claramente significativas y tienen los signos que cabría esperar, con la excepción de la prima por riesgo (YLD) y el ratio de vencimientos (MAT), que no son individualmente significativas en el modelo no restringido.

– Las variables personales no son significativas individualmente en ninguno de los modelos que las incluyen, aunque llevando a cabo un contraste de razón de verosimilitudes entre los Modelos 2 y 3, sí resultan conjuntamente significativas(7), lo que presenta cierta evidencia, aunque en ningún caso concluyente, a favor de la hipótesis de información asimétrica.

– Las variables de características económicas (LA y STL) no resultan significativas en ningún caso, ni individual, ni conjuntamente.

– El número condición de las matrices de varianzas-covarianzas es muy elevado en todos los casos. Esto es un indicativo claro de que la estimación está mal condicionada y, por tanto, pequeños cambios en la muestra pueden inducir variaciones importantes en los coeficientes estimados.

(6) Ratio entre el mayor y el menor autovalor de la matriz de varianzas-covarianzas.

(7) El valor tabular de la distribución χ^2_7 es de 12.0 al 90% y de 14.1 al 95% de confianza.

TABLA 6.2

ESTIMACIÓN DE LOS MODELOS ORIGINALES DE DHILLON ET AL (1987)

Variable	Modelo 1	Modelo 2	Modelo 3
Constante	-3.4855	-1.8774	-3.1077
	(5.2870)	(4.2249)	(5.8775)
FI (Tipo fijo)	0.9786	0.4987	1.0081
	(0.3911)	(0.2772)	(0.4107)
MAR (Margen)	-0.6268	-0.4310	-0.7052
	(0.2588)	(0.1736)	(0.2723)
YLD (r_{10}-r_1)	-2.2381	-2.3840	-2.5251
	(1.4310)	(1.0880)	(1.5881)
PTS (Puntos)	-0.7226	-0.2999	-0.8303
	(0.3753)	(0.2415)	(0.3977)
MAT (Ratio de vencimientos)	-1.1366	-0.0592	-1.1644
	(0.8927)	(0.6147)	(0.8946)
BA (Edad)	-0.0031	--	-0.0040
	(0.0390)		(0.0429)
BS (Estudios)	-0.1094	--	-0.1083
	(0.0967)		(0.0998)
FTB (Primera compra de	0.2398	--	0.1434
	(0.5208)		(0.5583)
CB (Co-prestatario)	-0.8061	--	-1.0666
	(0.6044)		(0.6922)
MC (Casado)	-1.0358	--	-1.0586
	(0.6557)		(0.6728)
SE (Cuenta propia)	-0.5906	--	-1.1275
	(1.2238)		(1.5598)
MOB (Movilidad)	-0.0882	--	-0.0930
	(0.0521)		(0.0550)
NW (Riqueza neta)	0.1349	0.0838	0.1288
	(0.0901)	(0.0422)	(0.1053)
LA (Activos líquidos)	--	--	0.0146
			(0.0350)
STL (Compromisos a c.	--	--	0.0161
			(0.0283)
$\ln \ell$	-31.53	-39.21	-30.73
LRT	1.60	16.96	
Nº condición	1.7e+6	1.4e+05	2.4e+06

En la Tabla 6.3 se muestran las elasticidades de la probabilidad de elegir un tipo de interés variable respecto de las variables continuas del Modelo 3. Como puede observarse, la variable que puede producir mayores cambios en la decisión es el tipo de interés fijo. Un incremento del 1% puede llegar a producir un importante

aumento en la probabilidad de elegir el tipo de interés variable (de hasta un 48%). No obstante, hay que tener en cuenta que el cambio en la decisión del individuo sólo se produce cuando la probabilidad de elección de éste se encuentra próxima a 0.5, por lo que sería necesario un cambio sustancial del tipo de interés fijo para inducir cambios en la decisión.

TABLA 6.3

ELASTICIDADES, PARA EL MODELO 3, DE LA PROBABILIDAD DE ELEGIR UN TIPO DE INTERÉS VARIABLE RESPECTO DE LAS VARIABLES CONTINUAS

Variable	Media	D.t.	Min	Max
FI	17.23	13.29	0.00	47.86
MAR	-2.49	2.62	-9.95	0.07
YLD	-5.41	4.26	-15.56	-0.00
PTS	-1.69	1.67	-8.21	0.00
MAT	-1.55	1.23	-5.07	-0.00
NW	0.30	0.38	-0.00	2.30
BA	-0.19	0.16	-0.89	-0.00
BS	-2.21	1.81	-7.92	-0.00
MOB	-0.65	1.38	-7.95	-0.00
STL	0.23	0.31	0.00	1.83
LA	0.09	0.24	0.00	1.80

Con estos resultados, los autores concluyen que "... en general, las características individuales del prestatario tienen una influencia débil en el tipo de préstamo elegido. Hay una tendencia a que algunas clases de prestatarios, [...] tengan preferencia por los tipos de interés variable. Esto es consistente con la hipótesis de información asimétrica" Dhillon et al. (1987, pág. 265).

6.3. Detección de observaciones anómalas

En este apartado se aplica la metodología propuesta en la sección 5 para detectar observaciones anómalas e influyentes. La Tabla 6.4 contiene los valores de los estadísticos correspondientes al conjunto de instrumentos de diagnosis propuesto en las secciones 2, 3 y 4 para el conjunto de observaciones que presentan valores apreciables en alguno de ellos. Por columnas, la Tabla 6.4 contiene: el número de la observación, el estadístico de distancia h_i, el estadístico $\tilde{h}_i$ de distancia para las variables transformadas con las expresiones **[4]**-**[5]**, el residuo, los estadísticos de influencia individual **[7]** y **[12]** y, por último, los componentes de los autovectores asociados a los dos mayores autovalores de la matriz de influencia $\tilde{M}$ definida en **[28]**.

TABLA 6.4

ESTADÍSTICOS DE DIAGNÓSTICO PARA LAS OBSERVACIONES MÁS SIGNIFICATIVAS

i	h_i	$\tilde{h}_i$	e_i	$\hat{c}_i$	$\hat{c}_i\,(P)$	λ_i^1	λ_i^2
5	0.2006	0.3072	0.5854	0.9040	0.1877	0.0460	-0.0010
14	0.3737	0.6298	0.6990	10.6700	2.0910	-0.9762	0.0118
15	0.1491	0.5046	0.6070	3.1750	0.6207	-0.0025	-0.0075
22	0.1905	0.2090	-0.8115	1.4380	0.2280	0.0339	0.0001
23	0.1531	0.2305	-0.5074	0.4009	0.0887	0.0094	0.0007
24	0.1531	0.2305	-0.5074	0.4009	0.0887	0.0094	0.0007
25	0.1531	0.2305	-0.5074	0.4009	0.0887	0.0094	0.0007
26	0.1905	0.2090	-0.8115	1.4380	0.2280	0.0339	0.0001
35	0.0987	0.3097	-0.5198	0.7037	0.1162	-0.0264	-0.0004
37	0.8889	0.9675	-0.2322	277.2000	49.1000	-0.0103	-0.9998
45	0.3373	0.0315	-0.0008	0.0000	0.0000	0.0001	0.0000
46	0.3579	0.0570	-0.0020	0.0001	0.0000	0.0015	0.0000
53	0.1101	0.3102	-0.3445	0.3426	0.0643	-0.0076	0.0001
55	0.0604	0.1820	-0.8855	2.1030	0.2963	0.0194	0.0053
58	0.1353	0.4689	-0.6062	2.5580	0.4804	0.0160	0.0070
59	0.4819	0.5944	-0.1261	0.5211	0.0636	0.0338	0.0013
61	0.1650	0.3417	-0.3911	0.5066	0.0938	0.1577	0.0021
62	0.1456	0.4145	-0.4997	1.2070	0.2240	-0.0379	0.0000
63	0.3541	0.0832	-0.0042	0.0004	0.0001	0.0028	0.0000
64	0.0637	0.2432	-0.4694	0.3757	0.0715	0.0250	0.0021
67	0.1159	0.4599	-0.5995	2.3600	0.4387	0.0022	0.0027
68	0.0976	0.1798	0.8874	2.1060	0.3470	0.0815	0.0047
69	0.0996	0.3206	0.6719	1.4220	0.2577	0.0102	-0.0023
71	0.1208	0.1890	0.7418	0.8257	0.1454	0.0478	0.0037
76	0.1211	0.2318	0.8253	1.8560	0.3052	-0.0180	0.0036
77	0.1890	0.4327	0.4205	0.9757	0.2028	-0.0072	0.0010
78	0.1482	0.3224	0.3790	0.4286	0.0790	0.0343	0.0010

A la vista de la Tabla 6.4 se pueden hacer los siguientes comentarios:

– El estadístico h_i calculado para las variables continuas del modelo sugiere que las observaciones 14 y 37 son extremas, muy especialmente esta última. Una vez estimado el modelo por el procedimiento MV lineal y transformadas las variables, el estadístico $\tilde{h}_i$ confirma que dichas observaciones son extremas en el espacio de las X.

– Atendiendo al estadístico de influencia para cada observación $\hat{c}_i$, puede comprobarse que la observación número 37 toma un valor extremadamente elevado. También la observación 14 tiene una influencia alta, aunque considerablemente menor que la 37. Nótese, que el valor de la distribución χ^2_{15} para una probabilidad del 10% es de 8.6. Por otra parte, teniendo en cuenta que la media del estadístico $\hat{c}_i$ para la muestra completa se encuentra por encima de cuatro, y la misma media,

eliminando la observación 37, es aproximadamente 1.5, parece claro que las observaciones 14 y 37 con altamente influyentes. De manera análoga a $\hat{c}_i$, $\hat{c}_i(P)$ también presenta valores extremos para dichas observaciones.

– Como se pone de relieve en Gracia-Díez y Serrano (1994), los residuos en los modelos de variable dependiente binaria no son una indicación de la posible anormalidad de una observación. Como muestra la Tabla 6.4, las observaciones cuyos residuos son más elevados no presentan ninguna evidencia de anormalidad cuando se presta atención a los estadísticos de influencia(8).

– Según estos resultados, las observaciones 14 y 37 pueden calificarse como influyentes. El valor del estadístico de influencia para el conjunto de ambas observaciones resulta ser 176.7. Además, dado el carácter extremo de ambas, y que la muestra es de tamaño reducido, la decisión adecuada es eliminarlas en el proceso de estimación.

– Para analizar el posible efecto de enmascaramiento provocado por estas observaciones, se ha utilizado el procedimiento de Peña y Yohai (1995) aplicado a la matriz $\tilde{M}$ definida en **[28]**. En las dos últimas columnas de la Tabla 6.4 aparecen los correspondientes componentes de los autovectores asociados a los dos mayores autovalores de $\tilde{M}$. Como puede apreciarse, ambos autovectores están claramente dominados por las observaciones 14 y 37, respectivamente. También se puede comprobar que la observación 61, que no aparecía como potencialmente influyente considerando los estadísticos anteriores, tiene un componente asociado relativamente elevado en el primer autovector (aproximadamente el doble que el valor que toma el siguiente componente), por lo que también se incluye en el grupo de observaciones influyentes. El valor del estadístico de influencia para las tres observaciones resulta ser de 170.4.

– Dado el reducido tamaño muestral, parece aconsejable no eliminar más observaciones sin contar con información adicional sobre el diseño de la muestra. Aunque, se han considerado algunas otras observaciones(9) como potencialmente influyentes, los resultados del estadístico de influencia conjunta, así como el análisis de la muestra, no permitían concluir que realmente lo fueran, por lo que es preferible mantenerlas en la muestra.

En la Tabla 6.5 se presenta la estimación de los tres modelos de la Tabla 6.2, eliminando el efecto de las observaciones 14, 37 y 61. A efectos de comparación,

(8) Aunque no se incluyen en la tabla, los residuos estandarizados tampoco presentaban valores especialmente elevados. Tan sólo un pequeño porcentaje de las observaciones tenía un valor superior a dos en valor absoluto y ningún residuo sobrepasaba tres en valor absoluto.

(9) En particular, se llevaron a cabo pruebas incluyendo como influyentes, además de las mencionadas 14, 37 y 61, las observaciones 68 y 69.

en la última columna de la tabla se incluye el Modelo 3 estimado hasta la convergencia sin dichas observaciones (Modelo 3*).

TABLA 6.5

ESTIMACIÓN DE LOS TRES MODELOS CONSIDERADOS ELIMINANDO EL EFECTO DE LAS OBSERVACIONES 14, 37 Y 61

Variable	Modelo 1	Modelo 2	Modelo 3	Modelo 3*
Constante	-4.5046	-1.1074	-4.7255	-5.5329
	(5.4612)	(4.2720)	(6.0367)	(6.2820)
FI (Tipo fijo)	0.8075	0.5402	0.7735	1.1339
	(0.3966)	(0.2826)	(0.4197)	(0.4488)
MAR (Margen)	-0.3098	-0.5000	-0.2779	-0.2064
	(0.3021)	(0.1864)	(0.3163)	(0.3082)
YLD (r_{10}-r_1)	-1.4383	-2.9296	-0.6247	-0.5817
	(1.6207)	(1.2077)	(1.8326)	(1.9498)
PTS (Puntos)	-0.7358	-0.3719	-0.6362	-1.1011
	(0.3867)	(0.2704)	(0.4105)	(0.4857)
MAT (Ratio de vencimientos)	0.1122	-0.1849	0.3134	-0.2943
	(1.0311)	(0.6404)	(1.0958)	(1.1719)
BA (Edad)	-0.2141	--	-0.1958	-0.0602
	(0.0737)		(0.0849)	(0.0664)
BS (Estudios)	-0.0123	--	-0.0514	-0.2311
	(0.0398)		(0.0468)	(0.1156)
FTB (Primera compra de vivienda)	-0.0866	--	-0.1120	-1.4168
	(0.0971)		(0.1011)	(0.9045)
CB (Co-prestatario)	0.0021	--	-0.3887	-1.8438
	(0.5338)		(0.5908)	(0.9051)
MC (Casado)	-0.8557	--	-1.3870	-0.4495
	(0.6268)		(0.7348)	(0.7130)
SE (Cuenta propia)	-0.6201	--	-0.3346	-4.2153
	(0.6802)		(0.7104)	(2.4431)
MOB (Movilidad)	-0.4544	--	-3.6371	-0.7308
	(1.2332)		(1.9868)	(0.2861)
NW (Riqueza neta)	0.1867	0.0816	0.1974	0.6143
	(0.0951)	(0.0430)	(0.1154)	(0.2305)
LA (Activos líquidos)	--	--	0.1506	0.0699
			(0.0879)	(0.0762)
STL (Compromisos a c. p.)	--	--	0.0462	0.0614
			(0.0306)	(0.0372)
$\ln \ell$	-42.80	-39.41	-55.46	-22.39
LRT	25.32	32.10		
Nº condición	1.7e+6	1.4e+5	2.2e+6	1.9e+6

Las principales diferencias con respecto a las estimaciones resumidas en la Tabla 6.2 son:

– Al contrario que con los modelos de la Tabla 6.2, realizando un contraste de razón de verosimilitudes, se puede rechazar la hipótesis de que las variables financieras personales (LA y STL) son no significativas. Adicionalmente, se confirma menor aversión al riesgo de los individuos más ricos (coeficientes positivos y significativos de LA y NW).

– La importancia de las variables personales, utilizando el mismo contraste que con los datos originales, puede considerarse superior. Algunas variables pasan a ser individualmente significativas, en concreto, BA, BS y CB.

– Los cambios en los coeficientes estimados para el Modelo 2 son muy pequeños, lo que hace suponer que la fuente de las anomalías procede de variables de características personales. Este hecho es lógico ya que, dado el lapso de tiempo con que se tomaron los datos, no cabe esperar que el mercado sufriera variaciones sustanciales.

– Debido a los cambios en los coeficientes, las elasticidades estimadas también han cambiado. Como puede apreciarse en la Tabla 6.6, las elasticidades de las variables que ahora son significativas son claramente superiores a las que aparecían en la Tabla 6.2.

Tabla 6.6

ELASTICIDADES, PARA EL MODELO 3, DE LA PROBABILIDAD DE ELEGIR UN TIPO DE INTERÉS VARIABLE RESPECTO DE LAS VARIABLES CONTINUAS UNA VEZ ELIMINADO EL EFECTO DE LAS OBSERVACIONES 14, 37 Y 61

Variable	Media	D.t.	Min	Max
FI	14.28	13.59	0.00	56.80
MAR	-1.01	1.22	-5.75	0.04
YLD	-1.43	1.33	-5.84	-0.00
PTS	-1.30	1.35	-5.63	0.00
MAT	0.44	0.42	0.00	2.01
NW	0.47	0.72	-0.00	4.00
BA	-2.72	3.11	-16.30	-0.00
BS	-2.41	2.30	-11.84	-0.00
MOB	-2.16	5.48	-31.05	-0.00
STL	0.62	0.81	0.00	3.75
LA	0.62	1.30	0.00	9.14

A la vista de estos resultados, se puede concluir que, a diferencia del artículo original, las variables personales sí resultan ser relevantes a la hora de explicar la elección de tipo de interés y, por tanto, estos resultados apoyan claramente la hipótesis de información asimétrica. La menor aversión al riesgo de los individuos

más ricos, sugerida por Dhillon et al. (1987) también queda contrastada, así como la mayor aversión de los individuos de más edad. Tal y como los autores del trabajo original concluyen, hay algunos tipos de individuos que prefieren más claramente los tipos de interés variables: aquellas familias con co-prestatarios, las parejas casadas y los individuos con elevada movilidad.

7. CONCLUSIONES Y POSIBLES EXTENSIONES

En este trabajo se presenta un conjunto de extensiones de los resultados propuestos en Gracia-Díez y Serrano (1994). En concreto: i) se deriva un estadístico que evalúa la influencia de cada observación sobre las probabilidades estimadas de los MEB, ii) se generalizan los estadísticos de influencia individual para evaluar el efecto de grupos de observaciones sobre las probabilidades estimadas, el vector de parámetros del modelo y subconjuntos del mismo vector, e iii) se trata el problema de enmascaramiento extendiendo a los MEB un procedimiento debido a Peña y Yohai (1995) para modelos lineales. Además, se propone una estrategia de diagnosis para abordar el problema de observaciones anómalas que se aplica a una muestra de datos reales analizada en Dhillon et al. (1987).

En la aplicación empírica de la sección 6 se pone de relieve que, de la misma forma que ocurre en otros modelos, en los MEB es necesario aplicar un proceso de diagnosis para determinar la posible falta de robustez de los resultados obtenidos debido a la presencia de observaciones anómalas. Los procedimientos presentados se han aplicado con éxito a los datos del ejemplo y han permitido detectar observaciones que alteraban los resultados de la estimación, hasta el punto que algunas conclusiones del trabajo original pueden ser rebatidas y otras reforzadas.

En este trabajo no se presentan resultados exhaustivos sobre el rendimiento de los estadísticos propuestos, ni sobre el procedimiento para la detección de observaciones enmascaradas. Sin embargo, en Gracia-Díez y Serrano (1994) y, en mayor medida, en Serrano (1993), se presentan experimentos de simulación para validar buena parte de los instrumentos de diagnosis propuestos en este trabajo.

Una primera extensión relevante de este planteamiento, es el análisis de valores críticos de los estadísticos de influencia propuestos. Si bien parece difícil determinar sus distribuciones teóricas, una labor de gran utilidad puede ser el desarrollo de distribuciones empíricas que permitan clasificar las observaciones de forma *objetiva*, sin requerir el análisis pormenorizado del investigador. Esta idea también es aplicable al procedimiento de Peña y Yohai (1995) para determinar grupos de observaciones influyentes.

Una segunda extensión, que también puede ser importante, consistiría en un análisis exhaustivo de los estadísticos LM propuestos en Serrano (1993) para

contrastar hipótesis sobre el origen de las anomalías. A la vista de los resultados obtenidos, el principal problema es su falta de potencia para discriminar entre las posibles fuentes de anomalías aunque, no obstante, su funcionamiento es aceptable para contrastar si determinados conjuntos de observaciones son efectivamente anómalas.

REFERENCIAS

AMEMIYA, T (19. 81).«Qualitative Response Models: A Survey», *Journal of Economic Literature*, XIX, 1483-1536.

AMEMIYA, T. (1985). «Advanced Econometrics», Oxford, *Basil Blackwell Ltd.*

ATKINSON, A.C. (1985). «Plots, Transformations and Regression», New York, Oxford *University Press.*

BEDRICK, E. J. y HILL. J. R.(1990). «Outlier Tests for Logistic Regression, a Conditional Approach», *Biometrika*, 77, 4, 815-827.

BELSLEY, D. A., KUH, E. y WELSCH, R. E (1981). «Regression Diagnostics. Identifying Influential Data and Sources of Collinearity», New York, *John Wiley & Sons.*

COOK, R. D. (1977). «Detection of Influential Observation in Linear Regression» *Technometrics*, 19, 1, 15-18.

COPAS, J. B. (1988). «Binary Regression Models for Contaminated Data», *Journal of the Royal Statistical Society*, B, 50, 2, 225-265.

DHILLON, U. S., SHILLING, J. D. y SIRMANS. C. F. (1987). «Choosing between Fixed and Adjustable Rate Mortgages», *Journal of Money, Credit and Banking*, 19, 1, 260-267.

GRACIA-DÍEZ, M. y SERRANO. G. R. (1994). «Observaciones Anómalas en Modelos de Elección Binaria», *Estadística Española*, 36, 135, 115-141.

JENNINGS, D. E. (1986). «Outliers and Residual Distributions in Logistic Regression», *Journal of the American Statistical Association*, 81, 396, 987-990.

LOTT, W. F. y RAY. S. C. (1992), «Applied Econometrics: Problems with Data Sets», *The Dryden Press.*

PEÑA, D. (1987). «Observaciones Influyentes en Modelos Econométricos», *Investigaciones Económicas*, XI, 1, 3-24.

PEÑA, D. y YOHAI, V. J.(1995). «The Detection of Influential Subsets in Linear Regression using an Influence Matrix», *Journal of the Royal Statistical Society*, B, 57, 2, próxima aparición.

PREGIBON, D. (1981). «Logistic Regression Diagnostics», *The Annals of Statistics*, 9, 4, 705-724.

RAO, C. R. (1973). «Linear Statistical Inference and its Applications», 2nd. ed., *John Wiley.*

ROUSSEEUW, P. J. y ZOMEREN, B. C. VAN (1990). «Unmasking Multivariate Outliers and Leverage Points», con discusión. *Journal of de American Statistical Association*, 85, 411, 633-651.

SERRANO, G. R. (1993). «Observaciones anómalas en modelos de variable dependiente cualitativa». *Tesis doctoral. Universidad Complutense de Madrid.*

WILLIAMS, D. A. (1987). «Generalized Linear Model Diagnostics: The Deviance and Single Case Deletion», *Applied Statistics*, 36, 2, 181-191.

Apéndice A

Datos de Dhillon et al. (1987), tomados del soporte magnético que acompaña el libro de Lott y Ray (1992). Las definiciones se encuentran en la Tabla 6.1.

ADJ	FI	MAR	YLD	PTS	MAT	BA	BS	FTB	CB	MC	SE	MOB	NW	LA	STL
0	13.62	1.50	1.38	2.33	1.50	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.62	1.50	1.38	2.33	1.50	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.62	1.50	1.38	2.33	1.50	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.62	1.50	1.38	2.33	1.50	38	22	1	0	0	0	1	7.56	8.91	3.89
0	14.00	5.50	1.38	1.75	1.00	41	16	1	0	0	1	4	7.82	12.50	50.93
0	14.00	4.75	1.38	1.75	1.00	41	16	1	0	0	1	4	8.01	17.74	50.44
0	14.00	4.75	1.38	1.75	1.00	41	16	1	0	0	1	4	8.01	17.74	50.44
0	13.62	1.50	1.38	2.33	1.50	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.50	2.40	1.59	1.00	1.00	44	16	1	0	1	1	11	17.86	17.12	31.46
0	13.75	2.44	1.45	2.00	0.67	34	19	1	0	0	1	2	9.10	6.18	40.48
0	14.00	2.45	1.64	1.00	1.00	44	16	1	0	1	0	2	2.42	5.01	28.81
0	14.00	2.45	1.64	1.00	1.00	44	16	1	0	1	0	2	2.42	5.01	28.81
0	13.50	2.40	1.59	1.00	1.00	44	16	1	0	1	1	11	17.86	17.12	31.46
0	14.00	0.35	1.64	1.25	0.67	57	17	1	1	1	0	28	5.62	16.84	22.53
0	13.90	3.04	1.50	2.03	1.00	42	20	1	0	1	0	8	12.40	0.00	0.00
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
0	13.75	2.33	1.45	2.50	1.00	38	22	1	0	0	0	1	7.56	8.91	3.69
1	13.50	2.40	1.59	1.00	1.00	44	16	1	0	1	1	11	17.86	17.12	31.46
1	13.88	0.35	2.04	0.83	1.00	39	21	1	1	1	0	2	4.26	1.20	25.80
1	13.88	0.35	2.04	0.83	1.00	39	21	1	1	1	0	2	4.26	1.20	25.80
1	13.88	0.35	2.04	0.83	1.00	39	21	1	1	1	0	2	4.26	1.20	25.80
1	13.50	2.40	1.59	1.00	1.00	44	16	1	0	1	1	11	17.86	17.12	31.46
1	13.50	3.86	1.60	0.74	0.42	53	16	1	1	1	0	17	1.98	7.05	0.30
1	12.38	2.73	1.40	1.66	0.85	32	18	1	0	1	0	4	1.11	3.60	0.59
1	12.13	3.36	1.60	1.66	0.85	24	17	0	0	0	0	1	0.12	0.28	1.07
1	12.25	3.36	1.60	1.66	0.85	43	16	1	0	1	0	6	0.88	3.44	9.35
1	12.38	3.36	1.60	1.66	0.85	30	13	0	1	1	0	1	0.36	2.34	11.56
1	12.38	3.36	1.60	1.66	0.85	25	16	0	0	0	0	3	0.46	1.37	0.00
1	12.25	3.36	1.60	1.66	0.85	26	14	1	1	1	0	5	0.57	0.75	15.32
1	12.40	3.36	1.60	1.66	0.85	32	12	0	1	0	0	3	0.35	0.69	27.91
1	12.50	2.10	1.77	0.00	1.00	27	13	1	0	1	0	1	0.61	0.17	7.03
1	13.00	3.61	1.69	1.81	1.00	27	17	1	0	0	0	4	0.73	0.25	12.56
1	13.25	3.61	1.69	4.34	1.00	25	16	1	1	1	0	1	13.57	93.49	86.35
1	12.25	2.60	1.59	2.55	0.93	24	16	0	1	1	0	1	0.48	2.01	8.08
1	13.00	2.40	1.59	2.00	1.00	34	9	0	1	1	0	2	0.17	0.44	0.34

ADJ	FI	MAR	YLD	PTS	MAT	BA	BS	FTB	CB	MC	SE	MOB	NW	LA	STL
1	12.50	2.60	1.59	1.27	0.93	44	12	0	0	1	0	9	0.46	2.10	3.04
1	12.50	2.60	1.59	2.55	0.93	30	18	0	0	0	0	2	0.42	1.55	18.12
1	12.50	2.60	1.59	1.27	0.93	35	24	0	1	1	0	1	3.20	27.58	0.00
1	13.00	3.86	1.60	1.48	1.69	34	17	1	0	1	0	2	3.43	1.22	26.82
1	12.50	2.60	1.59	2.55	0.93	55	14	1	0	1	0	6	1.68	5.71	0.13
1	13.25	3.86	1.60	1.48	1.27	65	6	0	1	1	0	10	0.07	0.21	1.23
1	12.50	2.60	1.59	1.09	0.93	27	18	0	0	0	0	27	0.19	0.48	0.52
1	12.75	3.86	1.60	1.48	0.85	31	20	1	1	1	0	2	0.72	1.07	11.52
1	12.13	3.36	1.60	1.66	0.85	36	16	0	0	0	0	1	0.37	1.08	8.62
1	12.75	3.86	1.60	1.48	0.85	27	16	0	0	0	0	2	0.21	0.97	9.18
1	12.25	2.73	1.40	1.24	0.85	31	12	0	1	1	0	1	0.42	3.03	7.31
1	12.75	2.60	1.59	0.76	0.93	31	15	1	1	1	0	1	1.00	0.25	2.40
1	13.25	2.08	1.50	0.97	1.42	45	14	1	0	1	0	5	0.79	1.32	14.94
1	13.90	3.04	1.50	2.03	1.00	37	12	0	0	1	0	1	0.26	0.70	1.91
1	12.25	2.60	1.59	0.69	0.93	37	14	1	0	1	0	1	0.75	2.33	0.80
1	12.75	2.08	1.50	0.49	0.95	32	16	0	0	0	0	1	0.11	0.40	9.81
1	13.90	3.04	1.50	2.03	1.00	41	18	1	1	1	0	2	0.88	1.22	0.00
1	12.60	3.36	1.60	1.66	0.85	31	16	0	1	1	0	1	0.60	2.12	3.92
1	14.00	2.45	1.64	1.00	1.00	36	25	0	0	0	0	1	0.44	0.71	0.20
1	13.70	2.08	1.50	0.97	2.38	43	16	1	0	0	0	18	0.80	0.09	19.45
1	13.80	3.04	1.50	2.03	1.00	38	16	0	0	0	0	3	0.24	0.98	10.06
1	13.75	1.04	1.45	0.67	1.00	48	17	1	1	1	0	17	2.66	7.80	13.88
1	13.62	1.50	1.38	2.33	1.50	27	14	1	0	1	0	1	1.24	1.29	3.33
1	14.00	2.40	1.59	1.50	1.00	26	11	0	1	1	0	26	0.32	0.39	16.61
1	13.00	2.40	1.59	2.00	1.00	39	12	0	0	0	0	2	0.12	0.35	4.88
1	13.37	0.35	2.04	1.67	1.00	31	12	1	1	1	0	3	0.41	0.08	9.29
1	13.50	0.35	2.04	1.67	1.50	34	12	0	1	1	0	2	0.27	0.54	8.26
1	14.00	0.35	2.04	1.67	1.50	36	16	1	1	0	0	1	3.53	1.15	12.25
0	11.77	1.90	1.88	0.46	1.13	31	12	1	0	1	0	1	0.44	1.17	10.68
0	11.76	1.75	1.74	0.45	1.11	39	16	0	0	0	0	2	0.31	1.44	7.05
0	14.00	1.66	1.74	0.50	1.50	33	12	1	0	1	0	1	0.44	0.66	0.00
0	12.84	0.85	2.03	0.00	1.20	30	12	0	1	1	0	1	0.36	1.34	12.93
0	13.75	-0.90	1.45	1.00	1.00	24	17	0	0	0	0	1	-0.06	0.46	23.88
0	12.50	0.95	1.77	0.67	1.00	30	12	0	0	1	0	1	0.18	0.49	6.66
0	12.50	-0.25	1.77	1.00	1.00	35	12	0	0	1	0	1	0.25	0.93	4.56
0	13.75	1.04	1.45	0.67	1.00	25	15	1	1	1	0	1	0.71	0.20	27.23
0	13.75	0.35	2.04	1.67	1.00	31	16	0	1	1	0	1	0.12	0.36	19.39
0	14.50	2.10	1.77	0.00	1.00	24	17	0	1	1	0	3	0.34	1.98	4.60
0	14.00	1.10	1.74	0.00	1.50	25	15	0	1	1	0	2	0.09	0.51	14.54

STATISTICS FOR DETECTION OF OUTLIERS IN BINARY CHOICE MODELS: AN APPLICATION TO A DATA SET

SUMMARY

This paper considers the problem of outliers in binary response models. Based on the statistic proposed by Gracia-Díez y Serrano (1994) which measures the influence of each observation on the estimated parameter vector, we derive other statistics in order to measure the influence of each observation as well as the influence of a group of observations on i) the vector of estimated probabilities and ii) subsets and linear combinations of the parameters in the model. Also, the method proposed by Peña y Yohai (1995) to deal with the masking problem in linear models has been generalised to the case of binary choice models. Lastly, we propose a diagnostic strategy to detect outliers in this type of models. The application of this strategy is illustrated by estimating the probit model used by Dhillon et. al (1987).

Key Words: outliers in binary choice models, detection, masking problem, diagnostic strategy.