Documento de Trabajo

9219

"OBSERVACIONES ANOMALAS EN MODELOS DE ELECCION BINARIA"

Mercedes Gracia-Díez

Gregorio R. Serrano

FACULTAD DE CIENCIAS ECONOMICAS Y EMPRESARIALES
UNIVERSIDAD COMPLUTENSE DE MADRID.
Campus de Somosaguas. 28.223 MADRID.

OBSERVACIONES ANOMALAS EN MODELOS DE ELECCION BINARIA

Mercedes Gracia-Díez*
Gregorio R. Serrano*

Departamento de Economía Cuantitativa
Universidad Complutense
Campus de Somosaguas
28223 Madrid
Telf. 394-2370

Junio, 1992

*Queremos agradecer a J. Agulló, T. Aparicio, R. Flores, A. Novales, T. Perez Amaral y A. Treadway sus sugerencias. Todos los errores son nuestros.

RESUMEN

En este trabajo se trata el problema de la existencia de observaciones anómalas en modelos de elección binaria. Se demuestra que la presencia de estas observaciones afecta a la consistencia de los estimadores de máxima verosimilitud. En cuanto a su detección: (i) se muestra que el análisis de residuos no es un instrumento adecuado, debido a la censura de la variable dependiente y (ii) se deriva un estadístico, similar al propuesto por Cook (1977) para modelos lineales, que resulta apropiado para la detección de anomalías en este tipo de modelos. Los resultados teóricos se contrastan con datos simulados.

ABSTRACT

This paper deals with the problem of outliers in binary response models. It is proved that the existence of these observations in the sample affects the consistency of maximum likelihood estimators. Regarding to detection of outliers: (i) it is shown that residual analysis is not a useful diagnostic tool, due to the censoring of the dependent variable and (ii) it is derived a statistic, analogous to the one proposed by Cook (1977) for linear models, which seems to be appropiated for outlier detection in this type of models. The theoretical results are tested using simulated data.

1. Introducción

La presencia de anomalías es frecuente en la estimación de modelos con datos de corte transversal. En estos modelos puede haber un conjunto reducido de observaciones que, debido a su ausencia de homogeneidad con el resto de la muestra, pueden distorsionar sustancialmente los resultados de la estimación, incluso si se utilizan muestras de gran tamaño. En la literatura econométrica, se ha tratado ampliamente este problema para el caso de los modelos lineales de regresión [Box y Tiao (1968), Cook (1977), Peña (1987) y Peña y Ruiz-Castillo (1982 y 1984) son algunas referencias]. Sin embargo, el estudio de la detección y consecuencias de las anomalías en modelos de elección discreta ha recibido menos atención. Para el caso de los modelos de elección binaria (MEB), este problema se trata por primera vez en Pregibon (1981) y, posteriormente, en Jennings (1986), Williams (1987), Copas (1988) y Bedrick y Hill (1990) entre otros; en Lesaffre y Albert (1989) se estudia el caso de los modelos de elección múltiple. Todos estos trabajos analizan básicamente los modelos logit y su planteamiento puede resumirse en los siguientes puntos: (i) no parten de una definición de dato anómalo, considerando como anomalía toda observación cuyo residuo en valor absoluto es "grande" y (ii) adaptan a los modelos logit los procedimientos para la detección de anomalías utilizados en los modelos lineales que, como es conocido, se basan fundamentalmente en el análisis de residuos y en evaluar el efecto de cada observación en la estimación de los parámetros del modelo.

En este trabajo tratamos el problema de forma diferente, ya que partimos de la definición de observación anómala que habitualmente se utiliza en la literatura econométrica: una observación anómala es aquella que no se ha generado por el mismo modelo estocástico que se supone para las restantes observaciones muestrales [ver, por ejemplo, Box y Tiao (1968)]. A partir de esta definición, comenzamos demostrando que en los MEB la existencia de anomalías en la muestra afecta a la consistencia

de los estimadores de máxima verosimilitud (MV). Ello se debe a que la presencia de estas observaciones hace que la función de verosimilitud del modelo sea diferente de la habitual, por lo que, si se ignora este hecho y se maximiza la función de verosimilitud habitual, los parámetros se estimarán inconsistentemente.

Seguidamente, se trata el problema de la detección de anomalías, siendo el objetivo principal del trabajo mostrar que, contrariamente a lo que se ha propuesto en la literatura anterior, en los MEB el análisis de residuos no es un instrumento adecuado. Ello se debe a que sólo se observa una realización dicotómica de la variable dependiente, por lo que el valor de los residuos está acotado y no proporciona información relevante sobre la probabilidad que tiene un dato de ser anómalo. La forma adecuada de detectar si una observación es anómala consiste en evaluar su peso en la estimación MV de los parámetros del modelo. La idea es que si una observación provoca un cambio significativo en el valor de los parámetros estimados, este cambio puede considerarse como una medida del sesgo de estimación que causa dicha observación.

Sobre la base de lo anterior, derivamos un estadístico similar al propuesto por Cook (1977) para modelos lineales, y que mide el efecto de cada observación en el vector de estimaciones MV de un MEB. Dicho estadístico puede calcularse, para cada una de las observaciones muestrales, a partir de la estimación MV del modelo con la muestra completa. Esto es, permite contrastar si cada observación en la muestra es anómala sin tener que volver a estimar el modelo, por el procedimiento no lineal habitual, omitiendo la correspondiente observación. En Pregibon (1981) se deriva un estadístico similar; sin embargo, el estadístico que proponemos en este trabajo puede calcularse fácilmente para cualquier MEB y no sólo para los modelos logit. Para la derivación de este estadístico, nos basamos en que la estimación MV de los parámetros de un MEB, utilizando el algoritmo de

"scoring", puede obtenerse por un procedimiento de mínimos cuadrados ordinarios (MCO) iterativo [Amemiya (1981)].

Por último, presentamos experimentos realizados con datos simulados con el objeto de comprobar empíricamente los resultados teóricos obtenidos. En concreto, tratamos de: (i) evaluar los sesgos en la estimación MV de estos modelos ante la presencia de anomalías y (ii) probar la efectividad del estadístico que proponemos para la detección de éstas.

Es importante señalar que, aunque a lo largo del artículo nos referimos al caso de los modelos probit, el supuesto de normalidad no es, en absoluto, necesario. Todos los resultados que aquí se presentan son aplicables a cualquier MEB, siempre que se suponga una distribución continua y simétrica, como es el caso de los modelos logit. Además, hemos elegido los modelos probit porque dan lugar a expresiones más generales que los logit, ya que la distribución logística permite simplificar, en muchos casos, las expresiones analíticas que utilizamos.

La organización del artículo es la siguiente. En la sección 2 se revisa el procedimiento de estimación por MV de los modelos probit binarios. En la sección 3 se analiza el modo en que la existencia de observaciones anómalas en la muestra afecta a la función de verosimilitud de estos modelos, produciendo inconsistencia en las estimaciones de los parámetros. En la sección 4 se ilustran las razones por las que el análisis de residuos no es un instrumento adecuado para la detección de anomalías en este tipo de modelos, mientras que el estadístico que proponemos para la detección de éstas se deriva en la sección 5. La sección 6 contiene los resultados obtenidos con datos simulados y, finalmente, en la sección 7 se resumen las principales conclusiones.

2. Los modelos probit binarios: derivación y estimación por máxima verosimilitud

Los modelos de elección binaria pueden derivarse a partir de modelos de regresión, en donde la variable dependiente es latente y sólo se observa una realización dicotómica de la misma. La variable observable representa la elección, por parte de los individuos, entre dos posibles alternativas. Consideremos el caso del modelo probit. Sea la ecuación:

$$z_i^* = \mathbf{x}_i'\alpha + \xi_i \qquad i=1,\ldots,n \qquad [1]$$

donde z_i^* es una variable continua no observable, el vector $\mathbf{x}_i$ contiene las observaciones de las k variables explicativas correspondientes al individuo i-ésimo, α es un vector de coeficientes desconocidos y ξ_i son variables aleatorias iid $N(0,\sigma^2)$. Dividiendo la ecuación [1] por σ, se tiene:

$$y_i^* = \mathbf{x}_i'\beta + u_i \qquad i=1,\ldots,n \qquad [2]$$

donde $y_i^* = z_i^*/\sigma$, $\beta = \alpha/\sigma$ y $u_i = \xi_i/\sigma$, de forma que u_i son variables aleatorias iid $N(0,1)$. La variable y_i^* representa el "sentimiento" del individuo i hacia una de las alternativas pero, en la práctica, lo único que se observa es si dicho individuo elige esa alternativa o no. Si denotamos por y_i la realización dicotómica de y_i^*, se puede establecer la siguiente relación[1]:

$$y_i = \begin{cases} 1 \text{ si } y_i^* \geq 0 & \text{(individuo i elige la primera alternativa)} \\ 0 \text{ si } y_i^* < 0 & \text{(individuo i elige la segunda alternativa)} \end{cases}$$

Entonces, si P_i es la probabilidad de que $y_i=1$ y $(1-P_i)$ es la probabilidad de que $y_i=0$, se tiene:

$$P_i = P[y_i=1] = P[y_i^* \geq 0] = P[u_i \geq -\mathbf{x}_i'\beta]$$

o lo que es lo mismo:

$$P_i = \Phi(x_i'\beta) \qquad i=1,\ldots,n \qquad [3]$$

que es el modelo probit, donde Φ es la función de distribución de la normal estándar.

Como es conocido, la estimación óptima de un modelo probit requiere la utilización de un criterio de MV lo que, en este caso, obliga a resolver un problema de estimación no lineal[2]. Para una muestra de n individuos, el logaritmo de la función de verosimilitud del modelo [3], viene dada por [ver, por ejemplo, Amemiya (1981) o Maddala (1983)]:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log\Phi_i + (1-y_i) \log(1 - \Phi_i) \right] \qquad [4]$$

donde Φ_i denota $\Phi(x_i'\beta)$.

El vector gradiente de esta función es:

$$g = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \Phi_i}{\Phi_i (1 - \Phi_i)} \phi_i x_i \qquad [5]$$

donde ϕ_i denota la derivada de Φ_i con respecto a β; esto es, el valor de la función de densidad normal estándar en $x_i'\beta$. Mientras que la matriz hessiana puede escribirse como:

$$G = \frac{\partial^2 \ell}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \left[\frac{y_i}{\Phi_i^2} \quad \frac{1 - y_i}{(1 - \Phi_i)^2} \right] \phi_i^2 x_i x_i' +$$

$$+ \sum_{i=1}^{n} \left[\frac{y_i - \Phi_i}{\Phi_i (1 - \Phi_i)} \right] \phi_i' x_i x_i' \qquad [6]$$

donde ϕ_i' es la derivada de ϕ_i. Dado que $Ey_i = \Phi_i$, a partir de [6] la matriz de información se reduce a:

$$I = - E \frac{\partial^2 \ell}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \frac{\phi_i^2}{\Phi_i (1 - \Phi_i)} x_i x_i' \qquad [7]$$

Una posible forma de resolver este problema de optimización es mediante el algoritmo de "scoring", cuya iteración $\tau+1$ puede escribirse como:

$$\hat{\beta}_{\tau+1} = \hat{\beta}_\tau + \hat{I}_\tau^{-1} \hat{g}_\tau \qquad [8]$$

donde por $\hat{I}_\tau^{-1}$ y $\hat{g}_\tau$ se denota respectivamente la inversa de la matriz de información y el vector gradiente evaluados en $\hat{\beta}_\tau$.

Los estimadores resultantes son consistentes, asintóticamente eficientes y tienen una distribución asintótica normal cuya matriz de covarianzas viene dada por la inversa de [7]. Por lo tanto, dicha matriz de covarianzas puede estimarse evaluando la inversa de [7] para la estimación MV de β obtenida en el proceso iterativo.

En el Apéndice se demuestra [ver también Amemiya (1981)] que la estimación MV del modelo [3], utilizando el algoritmo de "scoring", puede obtenerse por un procedimiento lineal iterativo.
Proposición 1: la estimación de $\beta_{\tau+1}$ que se obtiene a partir de [8] coincide con la estimación MCO de $\beta_{\tau+1}$ en el siguiente modelo de regresión:

$$\tilde{y}_{i\tau} = \tilde{x}_{i\tau}' \beta_{\tau+1} + \tilde{v}_i \qquad [9]$$

con las variables transformadas:

$$\tilde{y}_{i\tau} = \frac{y_i - \hat{\Phi}_{i\tau} + \hat{\phi}_{i\tau} x_i' \hat{\beta}_\tau}{[\hat{\Phi}_{i\tau}(1 - \hat{\Phi}_{i\tau})]^{1/2}} \qquad [10]$$

$$\tilde{x}_{i\tau} = \frac{\hat{\phi}_{i\tau} x_i}{[\hat{\Phi}_{i\tau}(1 - \hat{\Phi}_{i\tau})]^{1/2}} \qquad [11]$$

donde $\hat{\Phi}_{i\tau}$ y $\hat{\phi}_{i\tau}$ denotan respectivamente las funciones Φ_i y ϕ_i evaluadas en $\hat{\beta}_\tau$ y la var$(\tilde{v}_i) = 1$ para todo i.

Según esta proposición, una vez alcanzada la convergencia en el modelo [9], las estimaciones resultantes de β son numéricamente idénticas que las MV. Por otra parte, definiendo: $\tilde{X}_\tau = (\tilde{x}_{1\tau}, \ldots, \tilde{x}_{n\tau})'$ de orden (nxk), la matriz de covarianzas de $\hat{\beta}_\tau$ puede estimarse fácilmente por $(\tilde{X}_\tau' \tilde{X}_\tau)^{-1}$. Obsérvese que la expresión de esta matriz coincide con la inversa de la matriz de información en [7].

3. Efecto de la existencia de observaciones anómalas en la estimación de modelos probit

El objetivo de esta sección es analizar las posibles consecuencias de la existencia de observaciones anómalas en los resultados de la estimación MV de un modelo probit.

Consideremos la ecuación [2] utilizada para derivar el modelo probit. Esta ecuación es un modelo lineal de regresión en el que la variable dependiente y_i^* no es observable, la varianza de las perturbaciones es conocida e igual a 1 y las restantes hipótesis habituales del modelo se cumplen. En particular, la ecuación [2] establece que las variables y_i^* se han generado por el mismo modelo estocástico; esto es: y_i^* se distribuye iid $N(\mathbf{x}_i'\boldsymbol{\beta},1)$. Entonces, teniendo en cuenta que una observación anómala puede definirse como aquella que no se ha generado por el mismo experimento aleatorio que las restantes observaciones muestrales [Box y Tiao (1968)], un valor de y_i^* será anómalo si no se ha generado por el modelo [2]. Obsérvese que el hecho de que y_i^* sea una variable latente no significa que teóricamente no pueda presentar anomalías[3].

Según esto, una forma de modelizar la presencia de observaciones anómalas en el modelo [2] es suponer que, aunque las perturbaciones u_i se distribuyen iid N(0,1), existe una pequeña proporción desconocida ϵ de perturbaciones que siguen una distribución también normal, con media 0 y varianza h^2, donde h>1 [ver, por ejemplo, Box y Tiao (1968 y 1973) y Peña y Ruiz-Castillo (1982 y 1984)]. Esto es, se supone que las variables y_i^* en [2] provienen, bien de una distribución $N(\mathbf{x}_i'\boldsymbol{\beta},1)$, o bien de una $N(\mathbf{x}_i'\boldsymbol{\beta},h^2)$, con proporciones $(1-\epsilon)$ y ϵ respectivamente. En Box y Tiao (1968) se demuestra que, bajo estas condiciones, las perturbaciones en [2] pueden considerarse iid con una función de distribución que es una combinación lineal de dos distribuciones normales independientes y que depende de ϵ y h, tal que:

$$F(u_i) = (1-\epsilon)\Phi(u_i|0,1) + \epsilon\Phi_h(u_i|0,h^2) \qquad [12]$$

donde, lo mismo que en la sección anterior, Φ denota la función de distribución de la normal estándar y Φ_h la función de distribución de una normal con media 0 y varianza h^2. Esto es, si $\epsilon=0$ se tiene que $F(u_i) = \Phi(u_i|0,1)$, por lo que se mantiene la hipótesis de normalidad con parámetros 0 y 1 de u_i. Pero ante la presencia de un porcentaje ϵ de observaciones anómalas, la distribución de u_i es la indicada en [12], que no es normal. Además en este caso, para todo i:

$$E(u_i) = 0$$
$$var(u_i) = 1 + \epsilon(h^2-1)$$

por lo que la varianza de la distribución $F(u_i)$ es mayor que la unidad.

La implicación fundamental de que las perturbaciones en el modelo [2] se distribuyan como en [12] es que se produce un cambio en la forma funcional que determina P_i, ya que:

$$P_i = P[y_i=1] = P[y_i^*\geq 0] = P[u_i \geq -x_i'\beta] = F(x_i'\beta) \equiv F_i$$

por lo que, a partir de [12]:

$$P_i = F_i = (1-\epsilon)\Phi_i + \epsilon\Phi_{hi} = \Phi_i + \epsilon(\Phi_{hi}-\Phi_i) \qquad [13]$$

donde Φ_i y Φ_{hi} denotan $\Phi(x_i'\beta)$ y $\Phi_h(x_i'\beta)$ respectivamente.

Obsérvese que ante la presencia de este tipo de anomalías, la especificación correcta en la determinación de P_i viene dada por la ecuación [13], que establece que $P_i=F_i$, donde F_i es igual a Φ_i más un término adicional cuya magnitud depende de h y ϵ.

En la Figura 1 se representan las funciones Φ_i y F_i. Dado que Φ_{hi} es una función de distribución normal con media 0 y varianza mayor que 1, esta función se encontrará por encima de Φ_i para valores de $x_i'\beta<0$ y por debajo de Φ_i para valores de $x_i'\beta>0$. Por lo tanto, se tiene que:

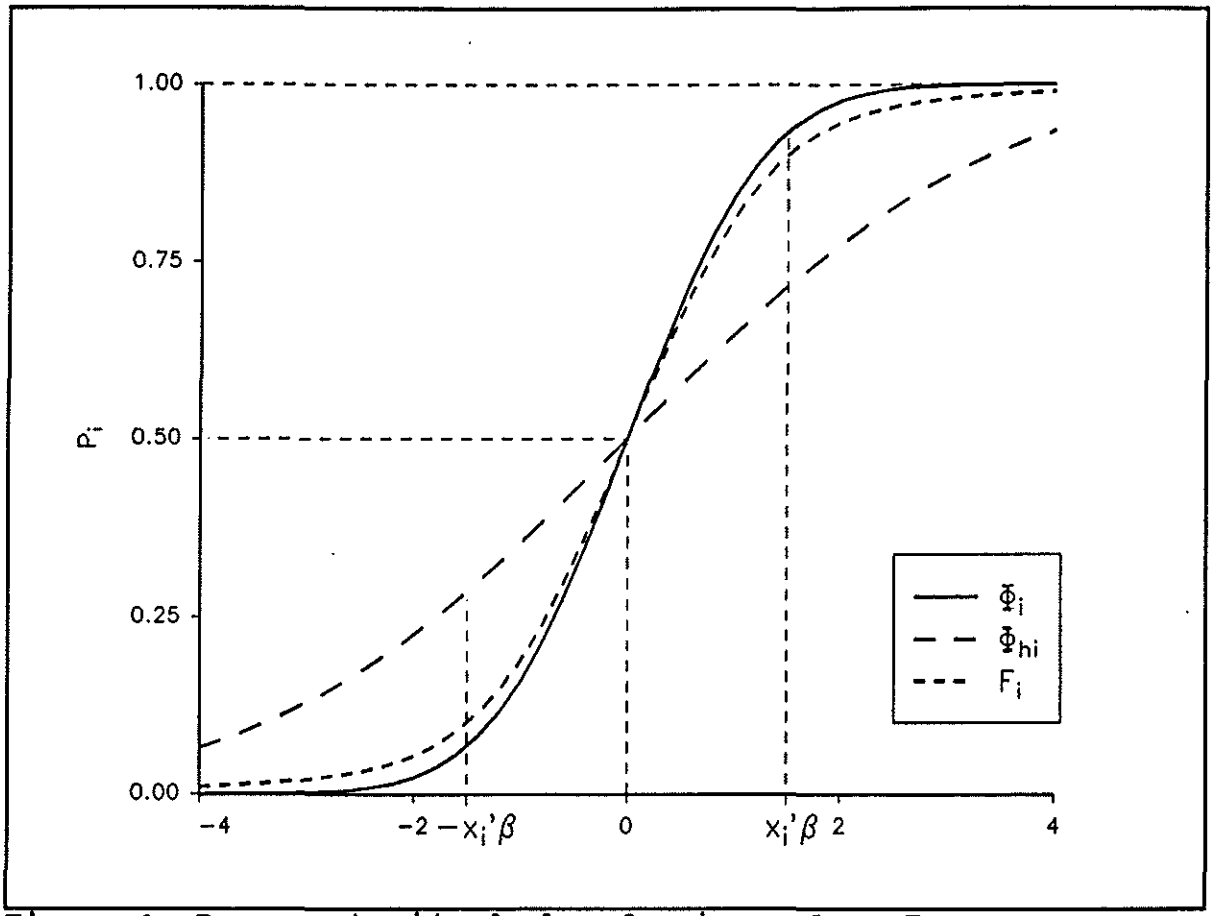

Figura 1: Representación de las funciones Φ_i y F_i

$$\text{si} \quad \begin{cases} x_i'\beta<0 \Rightarrow \Phi_{hi}>\Phi_i \Rightarrow F_i>\Phi_i & \text{para} \quad P_i<1/2 \\ x_i'\beta=0 \Rightarrow \Phi_{hi}=\Phi_i \Rightarrow F_i=\Phi_i & \text{para} \quad P_i=1/2 \\ x_i'\beta>0 \Rightarrow \Phi_{hi}<\Phi_i \Rightarrow F_i<\Phi_i & \text{para} \quad P_i>1/2 \end{cases}$$

Luego, F_i tendrá la forma que se indica en la Figura 1 y la discrepancia entre F_i y Φ_i dependerá del valor de los parámetros h y ϵ. A partir de [13] se tiene:

$$\frac{\partial(F_i-\Phi_i)}{\partial\epsilon} = \Phi_{hi} - \Phi_i$$

$$\frac{\partial(F_i-\Phi_i)}{\partial h} = \epsilon \frac{\partial\Phi_{hi}}{\partial h} = \epsilon \frac{\partial\Phi(x_i'\beta/h)}{\partial h} = \epsilon\, \phi\left[\frac{x_i'\beta}{h}\right]\left[\frac{-x_i'\beta}{h^2}\right]$$

por lo que ambas derivadas son mayores que 0 para valores de $x_i'\beta<0$ y menores que 0 para valores $x_i'\beta>0$. Por lo tanto, cuanto mayor sea el valor de ϵ y/o h, mayor será en términos absolutos la diferencia entre F_i y Φ_i.

Según estos resultados, el logaritmo de la función de verosimilitud correspondiente al modelo [13] es:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log F_i + (1-y_i) \log(1 - F_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log[\Phi_i+\epsilon(\Phi_{hi}-\Phi_i)] + (1-y_i)\log[1-\Phi_i-\epsilon(\Phi_{hi}-\Phi_i)] \right] \qquad [14]$$

donde sólo si $\epsilon=0$ esta expresión coincide con el logaritmo de la función de verosimilitud del modelo probit que figura en [4]. Pero si hay un porcentaje ϵ de observaciones con varianza mayor que 1, la expresión en [14] es la función que debería maximizarse para obtener las estimaciones MV del vector β. El problema es que esta función depende de los parámetros ϵ y h que, por lo general, no son conocidos. El desarrollo de un procedimiento de estimación adecuado para este caso es algo que no nos proponemos en el presente estudio[4]. No obstante, el objetivo de la ecuación [14]

es mostrar el tipo de error de especificación que se comete si se ignora la presencia de observaciones anómalas en un modelo probit y se utiliza Φ_i en lugar de F_i para calcular la verosimilitud de cada observación.

La consecuencia inmediata de este error de especificación en la función de verosimilitud del modelo es que el vector β y las probabilidades P_i se estimarán inconsistentemente. En Godfrey (1988, cap.6) se encuentra una demostración de esta afirmación. En ese trabajo se trata el problema de la heteroscedasticidad en los MEB y se demuestra que la existencia de dos grupos de observaciones con distinta varianza es equivalente a un error de especificación en la forma funcional del modelo, lo que produce inconsistencia en la estimación MV de los parámetros. El correspondiente sesgo asintótico no puede evaluarse ya que, en nuestro caso, depende de ϵ y h. Sin embargo, como hemos comprobado anteriormente, el sesgo será mayor cuanto mayor sea el número de observaciones anómalas en la muestra y cuanto mayor sea la varianza de la distribución que ha generado estas observaciones. Nótese que esta es la principal diferencia que presentan los MEB con respecto a los modelos lineales donde, a pesar de la presencia de este tipo de anomalías, los estimadores MCO siguen siendo insesgados, aunque no eficientes.

Es importante señalar que los resultados anteriores se basan en que, si existen anomalías en la muestra, la función de distribución de u_i viene dada por la expresión [12]. No obstante, este supuesto se hace por simplicidad, ya que otros supuestos alternativos sobre la generación de anomalías por el lado de su varianza, conducirían a errores de especificación del mismo tipo en la función de verosimilitud. En particular, el análisis anterior puede extenderse fácilmente al caso en que las anomalías se consideren generadas por distribuciones normales con distintas varianzas.

Un segundo tipo de observaciones anómalas podría deberse a que una proporción de las variables y_i^* se haya generado por una

distribución con distinta media que las restantes. Esta situación podría modelizarse suponiendo que, aunque las variables y_i^* se distribuyen iid $N(\mathbf{x}_i'\beta,1)$, existe un porcentaje desconocido ϵ de estas variables que siguen una distribución iid $N(\mathbf{x}_i'\gamma,1)$. Obviamente, si la proporción ϵ es grande, estaríamos ante la presencia de un cambio estructural; esto es, la población a analizar se compondría de dos grupos de individuos distintos entre sí. Sin embargo, el caso que consideramos en este trabajo es cuando ϵ es pequeño y sólo se trata de unos pocos individuos atípicos. En estas circunstancias, la función de distribución de y_i^* vendrá dada por:

$$G(y_i^*) = (1-\epsilon)\Phi(y_i^*|\mathbf{x}_i'\beta,1) + \epsilon\Phi(y_i^*|\mathbf{x}_i'\gamma,1)$$

por lo que es inmediato que:

$$\begin{aligned} P_i = P[y_i^*\geq 0] &= (1-\epsilon)P[u_i\geq -\mathbf{x}_i'\beta] + \epsilon P[u_i\geq -\mathbf{x}_i'\gamma] \\ &= \Phi(\mathbf{x}_i'\beta) + \epsilon[\Phi(\mathbf{x}_i'\gamma)-\Phi(\mathbf{x}_i'\beta)] \equiv \Psi_i \end{aligned} \qquad [15]$$

De manera similar al caso anterior, ignorar el segundo término del lado derecho de la ecuación [15], conduce a un error de especificación en la determinación de P_i. De nuevo, si se ignora este término y se utiliza Φ_i en lugar de Ψ_i para calcular la verosimilitud de cada observación, los parámetros del modelo se estimarán de forma inconsistente. Obsérvese que, según [15], para un $\mathbf{x}_i$ dado, el hecho de que la función Ψ_i esté por encima o por debajo de Φ_i dependerá del valor de los coeficientes en el vector γ. En cualquier caso, la discrepancia entre ambas funciones será mayor cuanto mayor sea ϵ y/o la diferencia entre los componentes de γ y β.

4. El problema de la detección de anomalías en los modelos probit

En la práctica, no se conoce a priori si existen observaciones anómalas en la muestra, ni mucho menos, cuál es la distribución que las ha generado. Por lo tanto, la forma habitual de detectar la presencia de estas observaciones es mediante la inspección de los datos muestrales, de manera que un dato se considera anómalo si es poco probable que se haya generado por la distribución que se supone para las restantes observaciones.

En un modelo de regresión lineal pueden existir básicamente dos tipos de anomalías, tal y como se ilustra en la Figura 2 [Peña y Ruiz-Castillo (1982 y 1984)]. En la regresión de z_i sobre x_i, el punto A puede considerarse un dato anómalo, ya que responde a un valor de z_i muy superior al de la media de las restantes observaciones muestrales. Este punto desplazaría hacia arriba la recta estimada y además su residuo será grande. Por otra parte, el punto B también puede considerarse anómalo, ya que tanto el valor de z_i como de x_i están muy por encima de sus valores medios. Sin embargo, aunque este punto afectaría gravemente a la pendiente de la recta estimada, su residuo puede ser muy pequeño en valor absoluto. Según esto, la inspección de residuos es un instrumento de análisis importante para la detección de anomalías, aunque no suficiente, ya que sólo sirve para detectar las del tipo A. En Belsley et al. (1980) se presentan distintos procedimientos de diagnosis del modelo basados fundamentalmente en medir el efecto de cada observación sobre los coeficientes estimados y la matriz de covarianzas y que, por tanto, permiten detectar si una observación es anómala aunque su residuo sea pequeño.

El objetivo de esta sección es mostrar que en el proceso de detección de observaciones anómalas en un modelo probit (o cualquier MEB en general), los residuos no juegan el mismo papel que en los modelos lineales. Ello se debe a que en un modelo probit el valor de los residuos está acotado como consecuencia

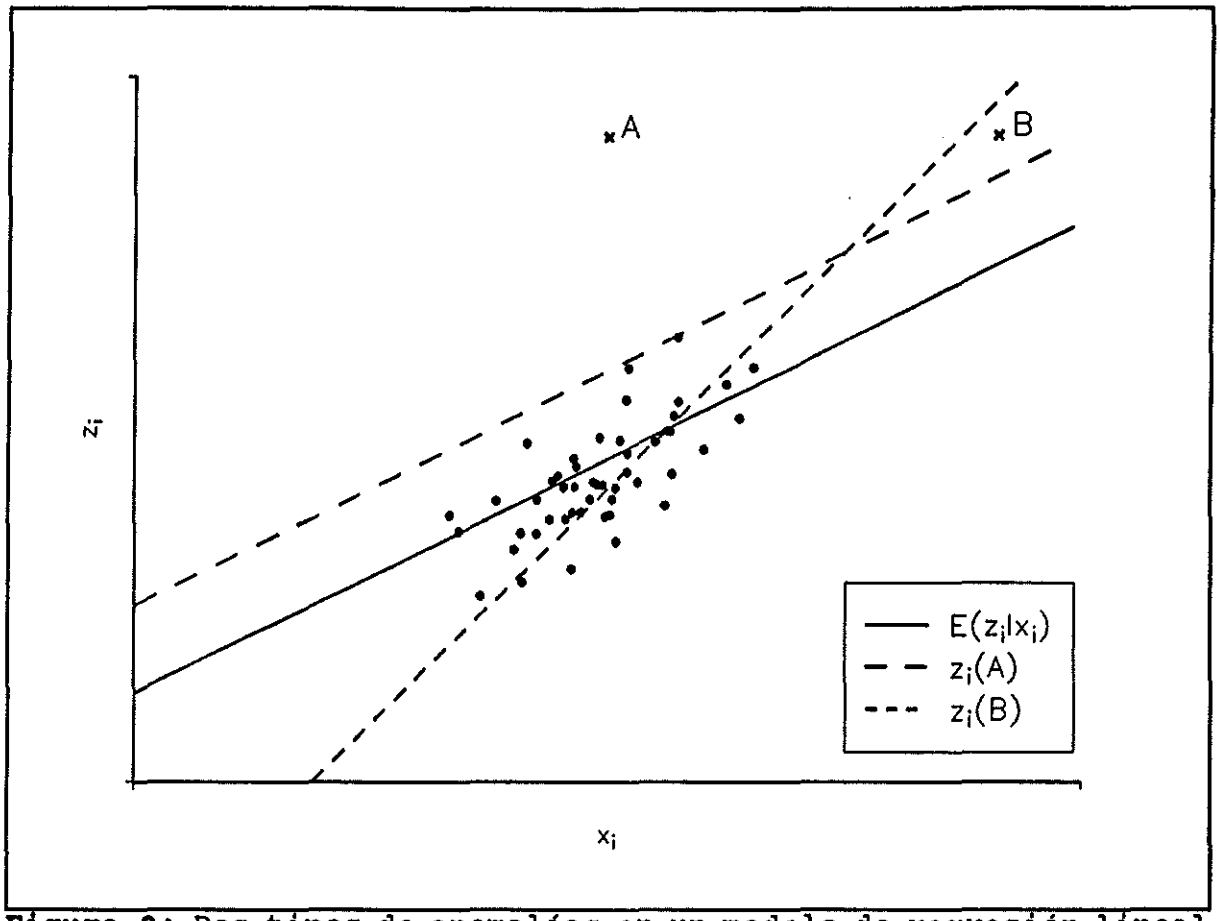

Figura 2: Dos tipos de anomalías en un modelo de regresión lineal

de la censura que presenta la variable y_i^*, de la que sólo se sabe si es mayor o menor que 0. Definiendo el residuo correspondiente a la observación i-ésima como la diferencia entre y_i y $\hat{P}_i$ [ver Pregibon (1981), Jennings (1986) y Cox y Snell (1989), por ejemplo]:

$$e_i = y_i - \hat{P}_i \qquad [16]$$

donde $E(e_i) = 0$ y $var(e_i) = \hat{P}_i(1- \hat{P}_i)$, se tiene que e_i está acotado entre (-1,1) pudiendo tomar, para cada observación, solamente dos valores: $(1-\hat{P}_i)$ y $(-\hat{P}_i)$.

El problema de la detección de observaciones anómalas en los MEB se trata por primera vez en Pregibon (1981), donde se considera como anomalía toda observación que, una vez estimado el modelo, presenta un residuo e_i próximo a 1 en valor absoluto. En consecuencia, en el citado trabajo se propone el análisis de residuos como un elemento de diagnosis para la detección de anómalos y se desarrollan una serie de estadísticos basados en los residuos e_i estandarizados. Posteriormente, este análisis se extiende en Williams (1987), Copas (1988) y Bedrick y Hill (1990) entre otros. Por el contrario, Jennings (1986) critica el trabajo de Pregibon (1981), señalando los puntos siguientes: (1) los residuos e_i no son comparables a los residuos MCO, ya que cada e_i depende de $\mathbf{x}_i$ a través de P_i, por lo que cada residuo tiene una distribución única y los residuos estandarizados no siguen una distribución normal y (2) eliminar de la muestra datos con un residuo próximo a 1 en valor absoluto equivale a truncar la muestra por una sola cola, con el consiguiente sesgo en la estimación de los parámetros. En este sentido, para Jennings "... las anomalías son necesarias".

En relación con la discusión anterior, es importante señalar que si una observación presenta un residuo e_i próximo a 1 en valor absoluto, simplemente quiere decir que la $P[y_i=s|\mathbf{x}_i] < \alpha$, donde s=0,1 y α es pequeño; esto es, que el valor que toma y_i en la muestra es poco probable. Pero no quiere decir que se trate

necesariamente de una observación anómala, ya que el hecho de que un valor de y_i sea poco probable no implica que y_i^* no haya sido generada por el modelo considerado, sino que puede ocurrir que y_i^* se encuentre en las colas de la distribución. Por consiguiente, estamos de acuerdo con Jennings en que no deben eliminarse de la muestra las observaciones con un residuo cercano a 1 en valor absoluto, pero no porque "las anomalías sean necesarias", sino porque es posible que estas observaciones no sean anómalas.

La afirmación anterior puede ilustrarse mediante la Figura 3. La parte inferior de la Figura contiene la nube de puntos $(y_i^*, \mathbf{x}_i')$ [donde, en este caso, $\mathbf{x}_i'=(1\ x_i)$] asociada al modelo [2] y la correspondiente recta teórica. En la parte superior de la Figura se han transladado al eje de abscisas los valores de la recta teórica $\mathbf{x}_i'\boldsymbol{\beta}$, mientras que en el eje de ordenadas se representan las probabilidades teóricas P_i y los valores observados de y_i. Las probabilidades P_i se obtienen evaluando la función de distribución normal estándar en $\mathbf{x}_i'\boldsymbol{\beta}$, mientras que los valores de y_i responden a la siguiente relación: $y_i=1$ si $y_i^*\geq 0$ e $y_i=0$ si $y_i^*<0$. El problema es que la muestra disponible para la estimación del modelo consiste solamente en los puntos $(y_i, \mathbf{x}_i')$, por lo que no se observa la nube de puntos de la parte inferior de la Figura. Entonces, pueden darse las siguientes situaciones:

(a) Consideremos el punto C en la parte inferior de la Figura, correspondiente a un valor negativo de x_i y a un valor muy grande y positivo de y_i^*. En el caso de un modelo lineal, donde observásemos y_C^*, este punto presentaría un residuo grande y positivo, por lo que consideraríamos con una probabilidad alta que se trata de un dato anómalo. Sin embargo, en el caso de un modelo probit, la realización de $y_C^*>0$ es $y_C=1$, por lo que el correspondiente residuo e_i será positivo y con un valor próximo a 1.

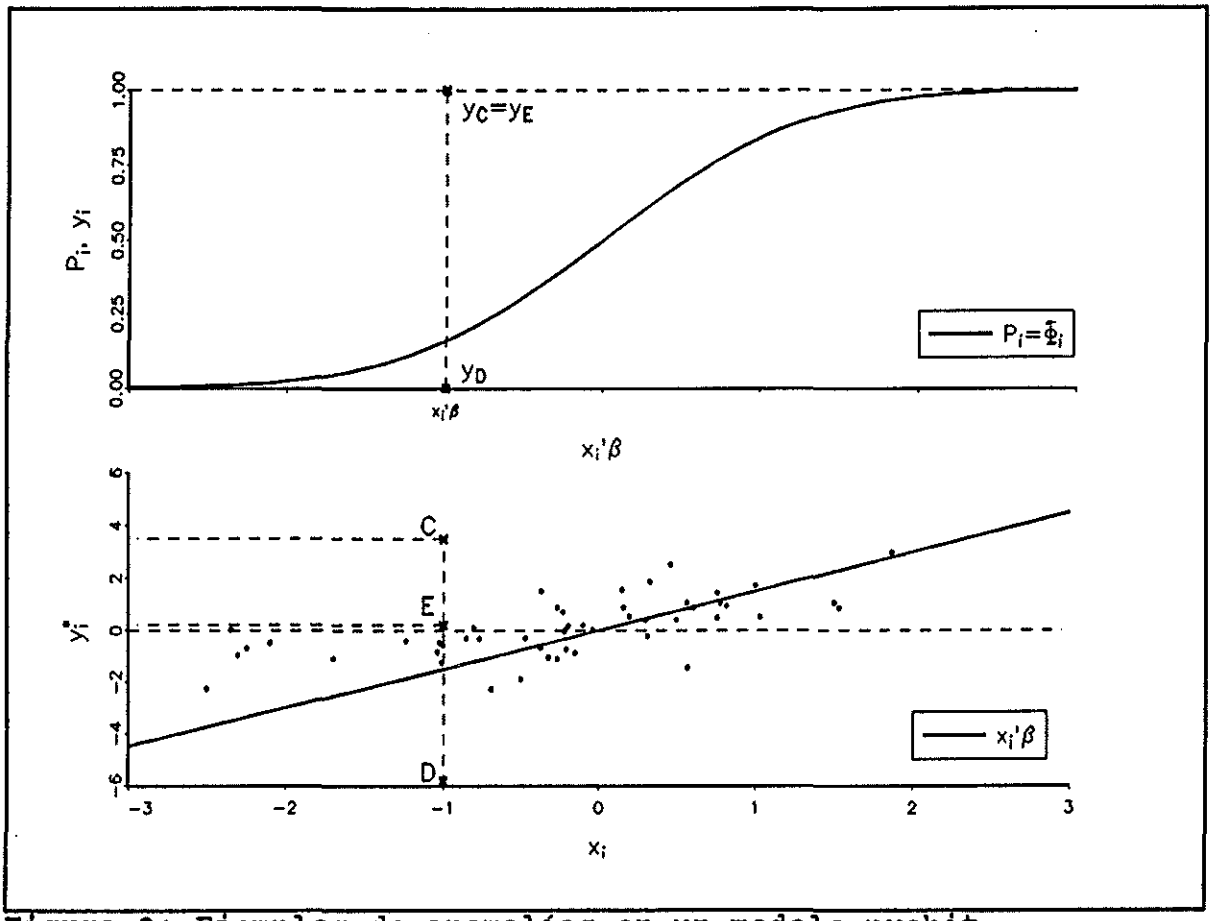

Figura 3: Ejemplos de anomalías en un modelo probit.

(b) Consideremos ahora el punto D, donde para el mismo valor negativo de x_i, el valor de y_i^* es negativo y está a la misma distancia de la recta teórica que el punto C. Igual que antes,si observásemos y_D^*, este punto presentaría un residuo grande aunque negativo. No obstante, el residuo e_i que obtenemos, una vez estimado el modelo probit, sería muy próximo a 0, puesto que, en este caso, $y_D=0$ al ser $y_D^*<0$.

Con este ejemplo, hemos tratado de ilustrar que dos observaciones igualmente anómalas pueden presentar, dependiendo del signo de la variable no observable y_i^*, un residuo próximo a 1 ó a 0. Sin ignorar que, en estos modelos, una anomalía del tipo C es, por lo general, más "peligrosa" que una del tipo D, se pueden extraer las siguientes conclusiones:

(1) Que una observación i tenga un residuo próximo a 0 no implica que no se trate de una observación anómala. En el caso más simple de una sola variable explicativa, observaciones con $x_i<0$ e $y_i^*<0$ ó $x_i>0$ e $y_i^*>0$ pueden presentar un residuo e_i muy pequeño y ser realmente anómalas, como es el caso del punto D en la Figura 3. Nótese que en un modelo lineal también puede haber observaciones anómalas con un residuo próximo a 0, pero dichas anomalías no son del tipo del punto D, que podría detectarse fácilmente por su residuo si se observase y_D^*.

(2) Que una observación tenga un residuo próximo a 1 en valor absoluto puede ser un indicio de que se trate de una observación anómala, pero no tiene que ser necesariamente así. Por ejemplo, una vez estimado el modelo probit, el punto E de la Figura 3 presentaría un residuo e_i positivo y próximo a 1 (de hecho, idéntico al del punto C) aunque, en este caso, si se observase y_E^* no habría razón para pensar que se trata de un dato anómalo; esto es, que no haya sido generado por una distribución $N(\mathbf{x}_i'\boldsymbol{\beta},1)$.

En definitiva, los residuos resultantes de la estimación de un modelo probit no son en absoluto informativos sobre la

probabilidad que tiene cada observación de ser anómala. Como acabamos de ilustrar, residuos próximos en valor absoluto a 1 ó a 0 pueden corresponder tanto a observaciones anómalas como a observaciones generadas por el modelo considerado. Esta es la mayor diferencia que presentan estos modelos respecto a los modelos lineales, donde el análisis de residuos no permite detectar todo tipo de anomalías (sólo las del tipo A de la Figura 2), pero donde un residuo grande sí presenta evidencia de que el correspondiente dato puede ser anómalo.

Para terminar esta sección, queremos señalar que, aunque el análisis de residuos no sea el instrumento adecuado para la detección de anómalos en los modelos probit, dicho análisis puede resultar de interés para detectar posibles errores de especificación. En una muestra dada, cabe esperar que un porcentaje pequeño de observaciones presente un residuo próximo a 1 en valor absoluto, sean o no anómalas, pero si este porcentaje es elevado puede deberse a una de las siguientes causas: (i) a un error de especificación, en el sentido de que las variables en $\mathbf{x}_i$ no son relevantes para explicar la variable y_i^* y, por tanto, las probabilidades P_i y (ii) a la existencia de, al menos, dos grupos distintos de individuos en la muestra ("cambio estructural"), lo que debe identificarse y modelizarse de la forma más adecuada.

5. Un estadístico para la detección de observaciones anómalas en los modelos probit

De la exposición en la sección anterior, se deduce que la forma adecuada para detectar si una observación es anómala en un modelo probit, debe ser mediante un estadístico que mida el efecto de esa observación en la estimación MV de los parámetros del modelo. Como se ha demostrado en la sección 3, la existencia de observaciones anómalas genera un error de especificación en la función de verosimilitud del modelo, que puede conducir a sesgos en la estimación de los parámetros. Por lo tanto, si el efecto de una observación en el valor de los coeficientes estimados es grande, dicho efecto puede considerarse como una medida del sesgo de estimación provocado por la presencia de esa observación en la muestra.

El estadístico que proponemos a continuación mide el efecto de cada observación en la determinación del vector $\hat{\beta}$, donde $\hat{\beta}$ denota la estimación MV de β. La distribución de este estadístico se basa en la normalidad asintótica del estimador MV de un modelo probit y para su derivación utilizamos el Teorema 4.30 de White (1984, pag. 70): Sea un vector de variables aleatorias θ de dimensión k. Si $\theta \overset{a}{\sim} N_k(0,\Sigma)$, donde Σ es la matriz (kxk) de covarianzas asintótica de θ, y existe una matriz $\hat{\Sigma}$ simetrica y definida positiva tal que plim $\hat{\Sigma} = \Sigma$, entonces $\theta'\hat{\Sigma}^{-1}\theta \overset{a}{\sim} \chi^2_k$.

Por lo tanto, teniendo en cuenta que: (i) el vector $\hat{\beta}$ de estimaciones MV sigue una distribución asintótica normal con media β y con matriz de covarianzas que es la inversa de la matriz de información I en [7] y (ii) el procedimiento MV garantiza que plim $\hat{I}^{-1} = I^{-1}$, donde $\hat{I}^{-1}$ denota I^{-1} evaluada en $\hat{\beta}$, se tiene que:

$$(\hat{\beta}-\beta)'\hat{I}\ (\hat{\beta}-\beta) \overset{a}{\sim} \chi^2_k$$

Entonces, si denotamos por $\hat{\beta}_{(i)}$ la estimacion MV de β eliminando la observación i-ésima, una medida de la distancia entre $\hat{\beta}$ y $\hat{\beta}_{(i)}$ vendrá dada por el estadístico:

$$D_i = (\hat{\beta}-\hat{\beta}_{(i)})'\hat{I}\ (\hat{\beta}-\hat{\beta}_{(i)}) \qquad i=1,\ldots,n \qquad [17]$$

Este estadístico, que es similar al estadístico propuesto por Cook (1977) para los modelos de regresión lineales [ver también Cook y Weisberg (1980), Peña (1977) y Peña y Ruiz-Castillo (1982 y 1984)], proporciona una medida de la distancia entre $\hat{\beta}$ y $\hat{\beta}_{(i)}$ en términos de niveles de significación. Esto es, a partir del valor de D_i y de la tabulación de la distribución χ^2_k, puede determinarse en qué medida la eliminación del punto i desplaza el vector de coeficientes estimados dentro de la región de confianza de β, calculada sobre $\hat{\beta}$, a un nivel de significación determinado. Por ejemplo, si D_i=0.57 y k=2, diremos que la eliminación de la observación i-ésima desplaza la estimación de β hasta el borde de la región de confianza de nivel 25% alrededor de $\hat{\beta}$. Cook (1977) sugiere, sobre la base de experimentos realizados en modelos lineales, que es deseable que cada $\hat{\beta}_{(i)}$ se encuentre dentro de la región de confianza de nivel 10%. Sin embargo, nosotros pensamos que lo importante de un estadístico de este tipo, no es la elección del nivel de significación para el que se realiza el contraste, sino la detección de las observaciones para las que el estadístico toma un valor más alto en términos relativos. Como es obvio, éstas serán las observaciones con una probabilidad más alta de ser anómalas.

El problema del estadístico D_i, según su expresión en [17], es que para su cálculo es necesario conocer $\hat{\beta}_{(i)}$, lo que exigiría estimar de nuevo el modelo, omitiendo una observación, cada vez que se desea contrastar si esa observación es anómala. Por lo tanto, para medir la influencia de cada una de las observaciones muestrales, sería necesario estimar n veces el modelo por un procedimiento no lineal. Obviamente, si el tamaño de la muestra

es grande, el cálculo de este estadístico puede resultar muy costoso desde el punto de vista computacional.

Para solucionar este problema, utilizamos el resultado de la Proposición 1 de la sección 2. Dado que el modelo [9] se estima por MCO, aplicando el lema de inversión de matrices a:

$$\hat{\beta}_{(i)} = (\tilde{X}'\tilde{X} - \tilde{x}_i\tilde{x}_i')^{-1} (\tilde{X}'\tilde{y} - \tilde{x}_i\tilde{y}_i)$$

se deriva fácilmente que:

$$\hat{\beta}_{(i)} = \hat{\beta} + (\tilde{X}'\tilde{X})^{-1} \tilde{x}_i [1 - \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i]^{-1} (\tilde{x}_i'\hat{\beta} - \tilde{y}_i) \qquad [18]$$

donde $\tilde{X}$, $\tilde{x}_i$ y $\tilde{y}_i$ denotan respectivamente la matriz $\tilde{X}_\tau$, el vector $\tilde{x}_{i\tau}$ y la variable $\tilde{y}_{i\tau}$ evaluados en $\hat{\beta}$, mientras que $\tilde{y}$ es un vector (nx1) cuyas componentes son las variables $\tilde{y}_i$. Obsérvese que esta expresión, aplicada recursivamente, permite también obtener la estimación MV de β cuando se elimina un grupo de observaciones.

Sustituyendo [18] en [17] y teniendo en cuenta que $\hat{I} = (\tilde{X}'\tilde{X})$ y que $\tilde{y}_i - \tilde{x}_i'\hat{\beta} = w_i$, donde w_i es el residuo e_i definido en [16] estandarizado: $w_i = e_i/[\hat{\Phi}_i(1-\hat{\Phi}_i)]^{1/2}$, se obtiene:

$$D_i = \frac{w_i^2 \, \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i}{[1 - \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i]^2} \qquad i=1,\ldots,n \qquad [19]$$

Nótese que la expresión [19] permite calcular el estadístico D_i para cada una de las n observaciones muestrales, a partir de la estimación MV de β con la muestra completa, sin tener que llevar a cabo ninguna estimación auxiliar.

El estadístico en [17] 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_{jj} el

elemento j-ésimo de la diagonal principal de la matriz $\hat{I}^{-1}$, es inmediato que el estadístico:

$$d_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{(v_{jj})^{1/2}}$$

proporciona una medida del desplazamiento que experimenta la estimación del coeficiente β_j cuando se elimina de la muestra la observación i-ésima. En este caso, la magnitud de dicho desplazamiento debe determinarse a partir de la distribución t_{n-k}.

Un estadístico similar al de la ecuación [19] se propone en Pregibon (1981). En este caso, para su derivación se utiliza el algoritmo de Newton, en lugar del algoritmo "scoring", en el proceso de estimación MV. Dado que el citado trabajo se restringe al caso particular de los modelos logit y que con la distribución logística la expresión de la matriz hessiana en [6] se reduce a: $-\Sigma\Lambda_i(1-\Lambda_i)\mathbf{x}_i\mathbf{x}_i'$, siendo Λ_i la función de distribución logística evaluada en $\mathbf{x}_i'\boldsymbol{\beta}$, el estadístico D_i resultante tiene una expresión sencilla. Sin embargo, esto no ocurre con los modelos probit, donde la utilización del hessiano (ver expresión [6]), complicaría innecesariamente la expresión del estadístico. En este sentido, el estadístico en [19] es más general, ya que puede aplicarse a cualquier MEB. Además, la matriz de información evaluada en el óptimo, y no el hessiano, es la matriz que teóricamente debe utilizarse para estimar la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$ en el cálculo de D_i. Otro motivo adicional para utilizar el algoritmo de "scoring" en lugar del de Newton en la estimación de los modelos probit, es que numéricamente la matriz **I** es siempre definida positiva, lo que no está garantizado con la matriz hessiana. Este hecho puede comprobarse fácilmente a partir de la expresión [6], donde **G** se calcula como la diferencia de dos matrices definidas positivas.

Es importante señalar que, aunque en esta sección nos hemos centrado en el estadístico de Cook, existen otras medidas del efecto de una observación en la estimación de un modelo lineal, que pueden también adaptarse fácilmente a los MEB. Estas medidas se basan fundamentalmente en el cambio que provoca una observación en el tamaño de la región de confianza del vector β. Por ejemplo, en Belsley et al. (1980) se propone el estadístico $COVRATIO_i$, que se calcula como el cociente entre los determinantes de las matrices de covarianzas de $\hat{\beta}$ sin y con la observación i-ésima. En los MEB, este estadístico se calcularía a partir de la matriz de información, mediante el ratio: $|\hat{I}_{(i)}|/|\hat{I}|$. En Cook y Weisberg (1980) se proponen otras medidas similares.

Por último, de la misma forma que en los modelos lineales, puede utilizarse el valor $h_i = \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i$, que es proporcional a la distancia de Mahalanobis, como una medida de la distancia entre el punto i y el centro de gravedad de todas las observaciones muestrales [ver Cook (1977), Peña (1987) y Peña y Ruiz Castillo (1982 y 1984) para el caso de los modelos lineales]. Esta medida puede resultar de utilidad para detectar si existen puntos en la muestra para los que las variables explicativas presenten valores muy diferentes del resto de las observaciones.

6. Resultados con datos simulados

En esta sección ilustramos los resultados teóricos de las secciones anteriores utilizando datos simulados. En concreto, tratamos de evaluar: (i) los sesgos en la estimación MV de un modelo probit, derivados de la existencia de anomalías en la muestra y (ii) el comportamiento del estadístico en [19] para la detección de dichas anomalías.

Consideramos un modelo probit con una sola variable explicativa y término constante, en el que las observaciones se han generado mediante el siguiente mecanismo:

$$y_i^* = \beta_1 + \beta_2 x_i + u_i$$

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* \geq 0 \\ 0 & \text{si} \quad y_i^* < 0 \end{cases}$$

$$P_i = P[y_i^* \geq 0] = \Phi(\beta_1 + \beta_2 x_i)$$

donde:

$$x_i \sim \text{iid } N(0,1),\ u_i \sim \text{iid } N(0,1) \text{ y } \beta' = (-0.65, 1)$$

siendo $y_i^* \geq 0 \Rightarrow y_i = 1$ para, aproximadamente, el 25% de las observaciones en la muestra.

A partir de este mecanismo, se han creado muestras donde se incluye un porcentaje ϵ de observaciones y_i^* generadas por una distribución también normal, pero con distintos momentos que los que acabamos de señalar. En particular, consideramos los siguientes casos ya discutidos en la sección 3:

Caso 1: Un porcentaje ϵ de observaciones y_i^* en la muestra proviene de una distribución normal con la misma media que las restantes observaciones, pero con varianza $h^2=5$. Esto es, las anomalías se han generado por la distribución: $y_i^* \sim \text{iid } N(x_i'\beta, 5)$.

Caso 2a: Un porcentaje ϵ de observaciones y_i^* proviene de una distribución normal con varianza igual a 1, pero con distinta media que las restantes observaciones. En concreto, hemos incluido una proporción ϵ de observaciones $y_i^* \sim iid\ N(\mathbf{x}_i'\gamma,1)$, donde $\gamma'=(1,-0.5)$. Obsérvese que se ha considerado un caso extremo, en el que las componentes del vector γ son muy diferentes a las del vector β.

Caso 2b: Un porcentaje ϵ de observaciones $y_i^* \sim iid\ N(\mathbf{x}_i'\delta,1)$, donde $\delta'=(-0.5,1.5)$ y, además, para estas observaciones, $x_i \sim iid\ N(5,1)$. Esto es, las componentes del vector δ no son muy distintas de las del vector β pero, es de esperar que, para las anomalías, los valores de x_i sean mucho mayores que los de las restantes observaciones.

En la Tabla 1 se presentan los resultados de la estimación MV, con muestras de 200 observaciones y 400 repeticiones, para cada uno de estos tres casos. Estas estimaciones se han obtenido mediante la linealización del algoritmo de scoring de la ecuación [9]. En la primera línea de la Tabla, figuran las medias de las estimaciones muestrales de los parámetros con ϵ=0.00; esto es, sin anomalías en la muestra. También se incluyen las medias del error cuadrático medio (ECM) y de la suma residual (SSR).

Los resultados más importantes de esta Tabla, pueden resumirse en los siguientes puntos:

(i) En los tres casos considerados, el valor de los parámetros estimados se aleja de su valor teórico a medida que aumenta el porcentaje de anomalías en la muestra. Como consecuencia de este sesgo de estimación, el ECM y la SSR presentan valores más altos cuanto mayor es ϵ. Sin embargo, las desviaciones estándar de los coeficientes estimados no se ven prácticamente alteradas.

(ii) En el caso 1, los sesgos en la estimación de β_1 y β_2 son menores que en los otros casos. Esto se debe a que, aunque

Tabla 1. Estimaciones MV con anomalias en la muestra

	ϵ(%)	$\hat{\beta}_1$	$\hat{\beta}_2$	ECM	SSR
	0.00	-0.66 (0.12)	1.02 (0.12)	0.04	29.35
CASO 1	0.05	-0.62 (0.11)	0.98 (0.11)	0.04	30.13
	0.10	-0.59 (0.11)	0.92 (0.11)	0.05	31.17
	0.15	-0.56 (0.10)	0.88 (0.10)	0.05	31.88
	0.20	-0.55 (0.10)	0.85 (0.10)	0.06	32.48
	0.30	-0.50 (0.10)	0.79 (0.10)	0.09	33.99
CASO 2a	0.05	-0.50 (0.10)	0.82 (0.10)	0.08	33.16
	0.10	-0.40 (0.10)	0.69 (0.10)	0.17	36.40
	0.15	-0.31 (0.09)	0.59 (0.09)	0.30	39.13
	0.20	-0.22 (0.09)	0.51 (0.09)	0.44	41.71
	0.30	-0.08 (0.09)	0.38 (0.09)	0.71	45.30
CASO 2b	0.05	-0.55 (0.11)	0.96 (0.11)	0.04	31.29
	0.10	-0.45 (0.10)	0.90 (0.10)	0.08	33.11
	0.15	-0.36 (0.10)	0.87 (0.10)	0.13	34.41
	0.20	-0.27 (0.10)	0.85 (0.10)	0.18	35.41
	0.30	-0.10 (0.09)	0.83 (0.09)	0.34	36.66

Notas:

(1) Modelo verdadero: $y_i^* \sim$ iid $N(x_i'\beta,1)$, $\beta = (-0.65,1)$
Caso 1: un % ϵ de observaciones $y_i^* \sim$ iid $N(x_i'\beta,5)$
Caso 2a: un % ϵ de observaciones $y_i^* \sim$ iid $N(x_i'\gamma,1)$ con $\gamma = (1.0,-0.5)$
Caso 2b: un % ϵ de observaciones $y_i^* \sim$ iid $N(x_i'\delta,1)$ con $\delta = (-0.5,1.5)$ y $x_i \sim$ iid $N(5,1)$

(2) n = 200, 400 repeticiones. $\hat{\beta}_1$ y $\hat{\beta}_2$ son las medias de las estimaciones muestrales de los parámetros y entre paréntesis figuran las medias de sus desviaciones típicas estimadas. ECM es la media del error cuadrático medio definido como $(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$ y SSR es la media de la suma residual.

la varianza de las observaciones anómalas es igual a 5, estas observaciones están aleatoriamente distribuidas alrededor de la recta teórica $\mathbf{x}_i'\boldsymbol{\beta}$, por lo que valores grandes positivos y negativos de y_i^* se "compensan" (sería, por ejemplo, el caso de los puntos C y D de la Figura 3). En este sentido, cabe resaltar que no ocurriría lo mismo si, por ejemplo, en una muestra dada, los valores anómalos de y_i^* fuesen sistemáticamente positivos. En muestras generadas por el mismo mecanismo, pero donde se ha tomado el valor absoluto de las perturbaciones correspondientes a las observaciones anómalas, forzando a que haya muchas más anomalías de y_i^* positivas que negativas, hemos obtenido que para ϵ=0.15: $\hat{\boldsymbol{\beta}}'$=(-0.27,0.90), por lo que, como era de esperar, el sesgo en la estimación de la ordenada en el origen es considerablemente mayor.

(iii) El caso 2a, donde existen observaciones muestrales con una media muy distinta de las restantes, parece especialmente grave. Obsérvese que, solamente con un 5% de este tipo de anomalías en la muestra, los sesgos de estimación en los dos parámetros son ya muy elevados, siendo casi tan altos como los detectados en el caso 1 cuando ϵ=0.30. En la práctica, obviamente desconocemos el origen de las anomalías existentes; sin embargo, este análisis muestra que es erróneo pensar que un número reducido de anomalías tenga efectos despreciables en la estimación del modelo, ya que esto depende del tipo de anomalías de que se trate.

(iv) El caso 2b trata de reflejar una situación habitual en la práctica, donde para los individuos anómalos, no solo la variable dependiente del modelo proviene de otra distribución, sino que también las variables explicativas toman valores muy distintos que para el resto de la muestra. Un ejemplo típico de esta situación puede encontrarse cuando se trabaja con la encuesta de Central de Balances, debido a la heterogeneidad de empresas que contiene. Obsérvese que aunque, en el caso que consideramos, el vector de parámetros δ no es muy distinto del que se ha utilizado para generar las restantes observaciones, los

sesgos en la estimación de los parámetros son importantes, especialmente el de la ordenada en el origen.

Por último, ilustramos cómo los sesgos en la estimación de los parámetros del modelo se traducen en que las probabilidades P_i también se estiman inconsistentemente. La Figura 4 contiene las probabilidades estimadas en el caso 2b para valores de ϵ=0.00, 0.15 y 0.30. Obsérvese que estos sesgos también pueden ser considerablemente altos.

En la segunda parte del experimento, tratamos de evaluar cómo se comporta el estadístico D_i, propuesto en [19], para la detección de anomalías. En la Tabla 2, se presentan los resultados obtenidos con muestras de 200 observaciones, que contienen un porcentaje ϵ=0.15 de anomalías, y donde se han llevado a cabo 100 repeticiones. Para cada uno de los tres casos que estamos considerando, esta Tabla contiene la media de las estimaciones de los parámetros y del ECM en las siguientes situaciones: con anomalías en la muestra (situación que denotamos por "A") y una vez que se han eliminado las anomalías para las que el estadístico D_i toma un valor superior a 0.02 (situación que denotamos por "A(-)"). Dado que se conoce cuáles son las observaciones anomálas en cada muestra, dicho estadístico sólo se ha calculado para estas observaciones.

Nótese que, en todos los casos, los sesgos en la estimación de los parámetros y el ECM se reducen considerablemente cuando se eliminan de la muestra las anomalías detectadas por el estadístico. Sin embargo, es importante señalar que el valor crítico de 0.02, que corresponde a un nivel de confianza del 1% de la distribución χ^2, se ha elegido arbitrariamente, por lo que estos resultados podrían mejorarse, llevando a cabo un análisis de sensibilidad para distintos valores críticos del estadístico en cada caso.

Respecto al valor de los residuos e_i, definidos en [16], hay que señalar que estos residuos son altos (superiores a 0.5 en

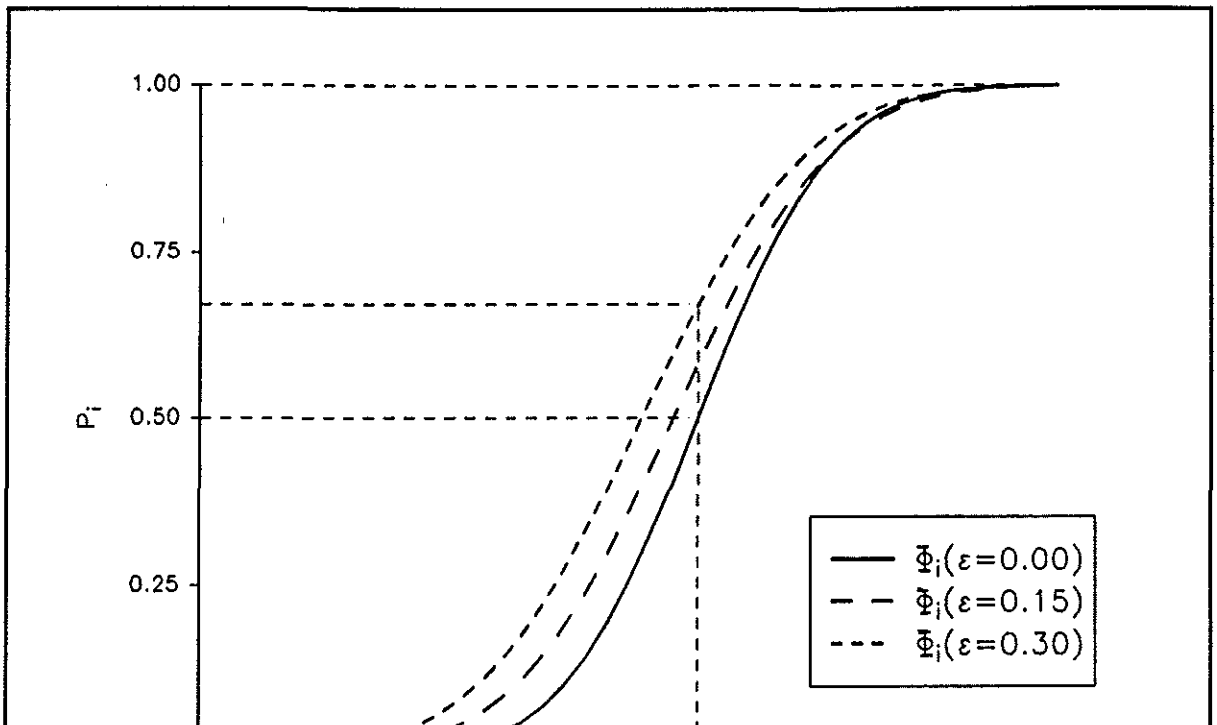

Figura 4: Ilustración del cálculo de las probabilidades estimadas en el Caso 2b

Tabla 2. Detección de anomalías mediante el estadístico D_i en [19]

	$\hat{\beta}_1$	$\hat{\beta}_2$	ECM
CASO 1			
A	-0.56	0.88	0.06
A(-)	-0.64	1.06	0.05
CASO 2a			
A	-0.30	0.59	0.30
A(-)	-0.49	0.99	0.06
CASO 2b			
A	-0.34	0.88	0.13
A(-)	-0.49	1.06	0.06

Notas:

(1) En el modelo generador de los datos: $\beta'=(-0.65,1)$. Para la definición de los casos, ver nota (1) en Tabla 1.

(2) Porcentaje de anomalías: $\epsilon = 0.15$.
Para cada caso, se presentan los resultados:
"A": con anomalías en la muestra
"A(-)": eliminando las anomalías para las que el estadístico D_i toma un valor mayor que 0.02

(3) n = 200, 100 repeticiones, $\hat{\beta}_1$ y $\hat{\beta}_2$ son las medias de las estimaciones muestrales de los parámetros y ECM es la media del error cuadrático medio definido como $(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$

valor absoluto en todos los casos) para las observaciones que ha detectado el estadístico, y que sabemos que son anómalas. Sin embargo, también hay un porcentaje considerable de observaciones generadas por el modelo considerado, que tienen un residuo superior a 0.5 en valor absoluto, e incluso muy próximo a 1 ó a -1, y que sabemos que no son anómalas. Lo mismo ocurre con los valores e_i estandarizados, aunque éstos no estén acotados. Este resultado aporta evidencia a favor de lo que hemos intentado mostrar en la sección 4: el residuo (estandarizado o no) no aporta información sobre la probabilidad que tiene un dato de ser anómalo.

Finalmente, queremos mencionar que los resultados de esta sección no resultan prácticamente alterados si se utilizan muestras de 400 observaciones y/o si los valores de x_i se generan por una distribución uniforme, con la misma esperanza y varianza que la distribución normal que se ha utilizado.

7. Conclusiones

Los resultados más importantes de este trabajo pueden resumirse en los siguientes puntos:

(i) Hemos demostrado que la existencia de observaciones anómalas en la muestra produce inconsistencias en la estimación MV de los MEB. También, a partir de un análisis realizado con datos simulados, mostramos que, dependiendo de la naturaleza de las anomalías, estos sesgos pueden ser muy altos, incluso si la muestra es grande y contiene un porcentaje reducido de este tipo de observaciones.

(ii) Para la detección de anomalías en estos modelos, el análisis de residuos no es un instrumento adecuado, ya que su valor está acotado entre -1 y 1. Además, residuos próximos a 0 ó a cualquiera de estos límites pueden corresponder tanto a observaciones anómalas como a observaciones generadas por el modelo subyacente.

(iii) El procedimiento adecuado para detectar si una observación es anómala en los MEB es midiendo el efecto de esa observación en la estimación MV de los parámetros del modelo. El estadístico derivado en la sección 5 cumple con este objetivo y puede calcularse fácilmente, para cada una de las observaciones, a partir de la estimación MV del modelo con toda la muestra. Experimentos realizados con datos simulados sugieren que este estadístico es efectivo en la detección de observaciones anómalas. No obstante, un tema que queda pendiente es la realización de un análisis sobre la potencia de dicho estadístico ante distintas condiciones, como por ejemplo: número de variables explicativas en el modelo, valor crítico y tamaño muestral.

Para terminar, es importante señalar que, aunque en los experimentos realizados con datos simulados se eliminan de la muestra las observaciones detectadas como anómalas, no pretendemos decir que esta sea la forma correcta de proceder en

todos los casos. Aún quedan muchas preguntas por contestar y quizá una de las más importantes sea: ¿que hacemos con las anomalias?. Nuestra postura es que depende de cuál sea el porcentaje de anomalías detectado, si es un porcentaje alto y existe la sospecha de que estamos ante la presencia de un "cambio estructural" o si, por el contrario, se trata solamente de unos pocos individuos atípicos. Si se trata de unos pocos individuos atípicos, pensamos que deben eliminarse de la muestra, a fin de evitar los sesgos que éstas provocan, pero también hay que contemplar la posibilidad de que el modelo esté mal especificado. En cualquier caso, para tomar una decisión en la linea de "robustificar" la metodología de estimación habitual de estos modelos, es necesario llevar a cabo un análisis de detección de anomalías en el modelo inicialmente especificado. Otra posibilidad sería abordar el problema de utilizar técnicas de estimación robusta o de influencia acotada ante la presencia de anomalías. Esto exigiría desarrollar procedimientos de estimación para formas funcionales del modelo como las especificadas en las ecuaciones [13] y [15], teniendo en cuenta que el porcentaje de anomalías en la muestra no es conocido.

Apéndice.

Demostración de la Proposición 1.

A partir del modelo [3], puede establecerse que:

$$y_i = \Phi(x_i'\beta) + v_i \qquad [A.1]$$

donde v_i es una variable binaria que toma los valores $1-P_i$ con probabilidad P_i y $-P_i$ con probabilidad $1-P_i$, de forma que:

$$E(v_i) = 0$$
$$var(v_i) = P_i(1 - P_i) = \Phi_i(1 - \Phi_i) \qquad [A.2]$$

Si se lleva a cabo una aproximación lineal del modelo [A.1], mediante una expansión por Taylor de $\Phi(x_i'\beta_{\tau+1})$ alrededor de un vector de condiciones iniciales $\hat{\beta}_\tau$, se tiene que:

$$\Phi(x_i'\beta_{\tau+1}) = \Phi(x_i'\hat{\beta}_\tau) + \phi(x_i'\hat{\beta}_\tau)x_i'(\beta_{\tau+1}-\hat{\beta}_\tau) + R_i \qquad [A.3]$$

Teniendo en cuenta que $R_i \to 0$ en probabilidad si $\hat{\beta}_\tau$ es una estimación consistente de β, sustituyendo [A.3] en [A.1], esta aproximación puede escribirse:

$$y_i - \Phi(x_i'\hat{\beta}_\tau) + \phi(x_i'\hat{\beta}_\tau)x_i'\hat{\beta}_\tau = \phi(x_i'\hat{\beta}_\tau)x_i'\beta_{\tau+1} + v_i$$

que en nuestra notación abreviada es:

$$y_i - \hat{\Phi}_{i\tau} + \hat{\phi}_{i\tau}x_i'\hat{\beta}_\tau = \hat{\phi}_{i\tau}x_i'\beta_{\tau+1} + v_i \qquad [A.4]$$

donde $\hat{\Phi}_{i\tau}$ y $\hat{\phi}_{i\tau}$ denotan respectivamente las funciones Φ_i y ϕ_i evaluadas en $\hat{\beta}_\tau$.

El modelo en [A.4] es lineal con perturbaciones heteroscedásticas, cuya varianza viene dada en [A.2]. Por lo tanto, una estimación del modelo por MC ponderados se obtiene aplicando MCO a:

$$\tilde{y}_{i\tau} = \tilde{x}_{i\tau}'\beta_{\tau+1} + \tilde{v}_i \qquad [A.5]$$

que es el modelo en [9] con las variables $\tilde{y}_{i\tau}$ y $\tilde{x}_{i\tau}$ definidas en [10] y [11] respectivamente.

Por otra parte, el algoritmo de scoring en [8] puede reescribirse como:

$$\hat{\beta}_{\tau+1} = \hat{I}_\tau^{-1}(\hat{g}_\tau + \hat{I}_\tau\hat{\beta}_\tau) \qquad [A.6]$$

Utilizando las expresiones en [5] y [7] para **g** e **I** respectivamente, se tiene que:

$$\hat{g}_\tau + \hat{I}_\tau\hat{\beta}_\tau = \sum_{i=1}^{n} \frac{1}{\hat{\Phi}_{i\tau}(1 - \hat{\Phi}_{i\tau})} \hat{\phi}_{i\tau}x_i \,(y_i - \hat{\Phi}_{i\tau} + \hat{\phi}_{i\tau}x_i'\hat{\beta}_\tau) \qquad [A.7]$$

por lo que, sustituyendo la inversa de [7] y [A.7] en [A.6], resulta:

$$\hat{\beta}_{\tau+1} = \left[\sum_{i=1}^{n} \frac{\hat{\phi}_{i\tau}^2}{\hat{\Phi}_{i\tau}(1-\hat{\Phi}_{i\tau})} x_i x_i' \right]^{-1} \cdot$$

$$\cdot \left[\sum_{i=1}^{n} \frac{1}{\hat{\Phi}_{i\tau}(1-\hat{\Phi}_{i\tau})} \hat{\phi}_{i\tau}x_i \,(y_i - \hat{\Phi}_{i\tau} + \hat{\phi}_{i\tau}x_i'\hat{\beta}_\tau) \right]$$

expresión que coincide con el estimador MCO del modelo en [A.5].

NOTAS.

(1) Dado que y_i^* no es observable, por conveniencia se establece un valor crítico igual a cero.

(2) Sólo si la muestra está formada por un número suficiente de observaciones repetidas o datos agrupados resulta posible obtener estimaciones consistentes y asintóticamente eficientes mediante un procedimiento de estimación lineal por MC ponderados [Amemiya (1981) y Maddala (1983)].

(3) Por ejemplo, si y_i^* representa la predisposición que tiene el individuo i-ésimo a adquirir un automóvil de lujo, que se supone depende única y positivamente de su renta, entonces un valor anómalo de y_i^* sería el de un individuo con renta muy alta que "odia" los coches de lujo o el de un individuo con renta muy baja que gana un automóvil de lujo en una rifa. En ambos casos, el valor de y_i^* es anómalo porque no se ha generado por el modelo considerado.

(4) En Quandt y Ramsey (1978) se trata el problema de estimación de modelos lineales cuyas perturbaciones se distribuyen como en [12]. En el caso de los modelos lineales, si el parámetro h no es conocido, la correspondiente función de verosimilitud no está acotada, por lo que, para obtener consistencia y normalidad asintótica, se propone un método de estimación basado en la función generatriz de momentos. Sin embargo, este método no es directamente aplicable a nuestro caso. En Quandt (1983) se abordan algunos aspectos relativos a este problema.

REFERENCIAS.

Amemiya T. (1981) "Qualitative Response Models: A Survey", Journal of Economic Literature, XIX, 1483-1536.

Bedrick E.J. y J.R. Hill (1990) "Outlier Tests for Logistic Regression: a Conditional Approach", Biometrika, 77, 4, 815-827.

Belsley D.A., E. Kuh y R.E. Welsch (1980) Regression Diagnostics, John Wiley.

Box G.E.P. y G.C. Tiao (1968) "A Bayesian Approach to some Outlier Problems", Biometrika, 55, 1, 119-129.

Box G.E.P. y G.C. Tiao (1973) Bayesian Inference in Statistical Analysis, Reading Mass: Addison-Wesley.

Cook R.D. (1977) "Detection of Influential Observation in Linear Regression", Technometrics, 19, 1, 15-18.

Cook R.D. y Weisberg (1980) "Characterization of an Empirical Influence Function for Detecting Influential Cases in Regression", Technometrics, 22, 4, 495-508.

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

Cox D.R. y E.J. Snell (1989) Analysis of Binary Data, Monographs on Statistics and Applied Probability, 32, Chapman & Hall, segunda edición.

Godfrey L.G. (1988) Misspecification Tests in Econometrics, Cambridge University Press.

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

Lesaffre E. y A. Albert (1989) "Multiple-group Logistic Regression Diagnostics", Applied Statistics, 38, 3, 425-440.

Maddala G.S. (1983) Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press.

Peña D. (1987) "Observaciones Influyentes en Modelos Econométricos", Investigaciones Económicas, XI, 1, 3-24

Peña D. y J. Ruiz-Castillo (1982) "Métodos Robustos de Construcción de Modelos de Regresión. Una Aplicación al Sector de la Vivienda, Estadística Española, 97, 47-76.

Peña D. y J. Ruiz-Castillo (1984) "Robust Methods of Building Regression Models. An Application to the Housing Sector", Journal of Bussines and Economic Statistics, 2, 1, 10-20.

Pregibon D. (1981) "Logistic Regression Diagnostics", The Annals of Statistics, 9, 4, 705-724.

Quandt R.E. (1983) "Computational Problems and Methods" en Handbook of Econometrics, vol. 1, Z. Griliches y M.D Intriligator Eds., North Holand Publishing Co.

Quandt R.E. y J. Ramsey (1978) "Estimating Mixtures of Normal Distributions and Switching Regressions, Journal of the American Statistical Association, 73, 364, 730-752.

White H. (1984) Asymptotic Theory for Econometricians, Academic Press.

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