OBSERVACIONES ANÓMALAS EN MODELOS DE VARIABLE DEPENDIENTE CUALITATIVA

Tesis Doctoral

Autor: Gregorio R. Serrano García

Directora: Mercedes Gracia Díez

Departamento de Economía Cuantitativa

Facultad de Ciencias Económicas y Empresariales

Universidad Complutense de Madrid

1993

A mis padres, Gregorio y Epifanía

A Eustasia García, *in memoriam*

Agradecimientos

La mayor deuda de gratitud durante la realización de esta Tesis la he contraido con mi directora, Mercedes Gracia. Sin su constante apoyo personal y profesional, nunca hubiese visto la luz este trabajo.

También quisiera expresar mi agradecimiento a Alfonso Novales, y muy especialmente a Emilio Domínguez, Miguel Jerez y Sonia Sotoca por su valiosa ayuda durante el tiempo de realización de este trabajo.

Fuera del terreno profesional, quisiera agradecer el apoyo moral que, en todo momento, he recibido de mis padres y hermano, y la infinita paciencia y dedicación de Auxi.

A todos ellos, y a los compañeros de Departamento que han mostrado algún interés por este trabajo, muchas gracias.

CONTENIDOS

INTRODUCCIÓN

Los modelos de variable dependiente cualitativa o de elección discreta han experimentado, en los últimos tiempos, un importante auge en cuanto a su utilización en la investigación económica. Esto, seguramente, es debido a la mayor disponibilidad de bases de datos microeconómicas y a la importacia creciente del denominado *análisis microeconómico de la macroeconomía*.

En este contexto, se ha desarrollado un conjunto de herramientas fundamentales a la hora de trabajar con cualquier modelo: estimadores óptimos, estadísticos para el contraste de hipótesis e instrumentos de predicción. No obstante, otros aspectos de esta clase de modelos han recibido mucha menos atención, en particular, los relativos a la diagnosis del modelo.

En este trabajo se considera un problema concreto en la interacción modelo-datos: la existencia de observaciones anómalas, que resultan frecuentes en las muestras de corte transversal. Al intentar describir el comportamiento de la muestra mediante un modelo, puede haber un conjunto reducido de observaciones que, debido a su falta de homogeneidad con el resto de la muestra, distorsionen sustancialmente los resultados de la estimación, incluso si se utilizan muestras de gran tamaño. En este trabajo, se supone que dichas observaciones *no se deben a errores en los datos*, sino a que en la muestra hay un grupo de observaciones que proceden de una población diferente que el resto. Por tanto, este trabajo se centra en un tipo muy concreto de errores: los que tienen su origen en el hecho de que entre los datos se encuentra un conjunto de observaciones generadas por un proceso estocástico distinto del que sigue la mayoría de la muestra.

Siguiendo este planteamiento, el primer objetivo del trabajo es mostrar que, contrariamente a lo que se ha propuesto en la mayoría de la literatura anterior, en los modelos de elección cualitativa los análisis que se apoyan en los residuos, o en simples extrapolaciones de los resultados para el modelo lineal general, no resultan adecuados. 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.

El segundo objetivo del trabajo se centra en derivar estadísticos o medidas de influencia para la detección de anomalías en los modelos de variable dependiente cualitativa. Este es el primer paso para, posteriormente, decidir el tratamiento más adecuado que debe darse a las observaciones que se han detectado como anómalas.

El problema de las anomalías en los modelos de variable dependiente cualitativa ha sido tratado con anterioridad: Pregibon (1981), Jennings (1986) y Copas (1988) son algunas referencias. 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 de elección discreta los procedimientos para la detección de anomalías utilizados en los modelos lineales que, en gran medida, se basan en el análisis de residuos y en evaluar el efecto de cada observación en la estimación de los parámetros del modelo. Además, estos trabajos no parten de una definición estadística de dato anómalo, ni analizan las consecuencias que este tipo de observaciones tienen sobre los resultados de la estimación del modelo.

En este trabajo se enfoca el problema de forma diferente, partiendo 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* [Box y Tiao (1968)]. A partir de esta definición, se demuestra que, en los modelos de elección cualitativa, la existencia de anomalías en la muestra afecta a la consistencia del estimador de máxima verosimilitud. Ello se debe a que la presencia de estas observaciones hace que la función de verosimilitud del modelo sea diferente de la habitual.

Este trabajo está organizado como sigue. En el **Capítulo 1** se lleva a cabo una revisión de los modelos de elección cualitativa, con la finalidad de establecer la notación básica y presentar una serie de resultados utilizados en los siguientes capítulos.

En el **Capítulo 2** se aborda el problema de las anomalías en los modelos de elección binaria (MEB). En primer lugar, se exponen los problemas que aparecen en el modelo lineal general para, posteriormente, analizar las particularidades que este tipo de observaciones presentan en los modelos de elección binaria. Estas características propias de los modelos de variable dependiente cualitativa hacen que no sea inmediata la extensión de los planteamientos de diagnosis desarrollados para el modelo lineal general.

En el **Capítulo 3** se trata el problema de la detección de anomalías en los MEB. Los resultados que se desarrollan se encuentran en la línea de *robustecer* la metodología de estimación, tal y como se propone en Box (1980) y Peña y Ruiz-Castillo (1982 y 1984), para el caso de los modelos lineales de regresión. Con este propósito, en este capítulo se derivan estadísticos o medidas de influencia para la detección de anomalías en los MEB. Este es el primer paso para, posteriormente, decidir el tratamiento más adecuado que debe darse a las observaciones que se han detectado como anómalas.

En el **Capítulo 4** se generalizan los resultados de los **Capítulos 2** y **3** para los modelos de elección cualitativa múltiple (MEM) más utilizados en la práctica: el modelo logit multinomial y el modelo probit multinomial. El planteamiento de observaciones anómalas, así como los principales resultados sobre su diagnosis son análogos a los desarrollados para los modelos de elección binaria. En los **Capítulos 2**, **3** y **4** también se ilustran los principales resultados teóricos con experimentos de Monte Carlo.

Por último, en el **Capítulo 5** se aplica la metodología de detección de observaciones anómalas desarrollada en los **Capítulos 2** y **3** a dos muestras de datos reales: la muestra utilizada por Dhillon et al. (1987) en un estudio sobre la elección de tipos de interés fijos frente a tipos variables para préstamos hipotecarios y la muestra utilizada por Pregibon (1981) en un experimento médico. El objetivo de estos análisis se centra en ilustrar la necesidad de analizar los datos empleados en cualquier estudio, antes de pasar a la interpretación de los resultados de estimación obtenidos.

CAPÍTULO 1

MODELOS CON VARIABLE DEPENDIENTE BINARIA

1.1. Introducción

El problema genérico que se plantea en Econometría consiste en explicar el comportamiento o realizar previsiones de una(s) variable(s) endógena(s) a través de un conjunto de variables predeterminadas. Un caso particular de esta situación, consiste en analizar el comportamiento de los individuos cuando tienen que elegir entre un conjunto de alternativas mutuamente excluyentes. En este análisis, la variable dependiente del modelo representa la elección realizada por cada individuo, por lo que es cualitativa, mientras que las variables explicativas recogen las características del decisor y de las alternativas disponibles tal y como éste las percibe. Con estos modelos se trata de explicar las decisiones de los individuos en términos de probabilidad y su aplicación más inmediata es la predicción del comportamiento individual y/o agregado fuera de la muestra.

Los modelos de elección cualitativa se han utilizado ampliamente en aplicaciones biométricas, antes que en aplicaciones económicas. Los biómetras han usado estos modelos para estudiar problemas de estímulo-respuesta como, por ejemplo, el efecto de diferentes dosis de un medicamento en la recuperación o no recuperación de un paciente.

En economía estos modelos se utilizan generalmente para explicar decisiones económicas discretas como, por ejemplo, la participación de los individuos en el mercado de trabajo, la elección de ocupación, la pertenencia a sindicatos, la compra de bienes de consumo duraderos, etc. Por tanto, la motivación y derivación de estos modelos suele realizarse a partir de la teoría de la decisión y, en concreto, de la regla de decisión basada en la utilidad.

Otra característica común de estos trabajos es la hipótesis implícita de que los decisores tienen un comportamiento racional; esto es, sus preferencias son consistentes y transitivas. La consistencia implica que cada decisor, bajo circunstancias idénticas, utilizará

la misma regla de decisión y, en consecuencia, elegirá la misma alternativa. La transitividad implica que si la alternativa A es preferida a la B y ésta es preferida a la C, entonces la alternativa A es preferida a la C.

En la teoría microeconómica del consumidor, o teoría de las preferencias individuales, cada sujeto elige aquella combinación de bienes que le resulta preferible y que satisface su restricción presupuestaria. Esta combinación óptima, se obtiene maximizando la función de utilidad del individuo sujeta a su restricción presupuestaria, lo que permite derivar las funciones de demanda. Sin embargo, en la teoría de la elección discreta, la función de utilidad de cada individuo sólo puede tomar un número reducido de valores, tantos como alternativas disponibles, por lo que dicha función no es diferenciable respecto a las cantidades. Consecuentemente, el concepto de una relación continua entre la cantidad demandada y un conjunto de variables explicativas, carecen de sentido y, por tanto, los modelos teóricos deben basarse directamente en las funciones individuales de utilidad.

En este capítulo se lleva a cabo una revisión de los modelos de elección cualitativa, con la finalidad de establecer la notación básica y presentar una serie de resultados que van a utilizarse en los siguientes capítulos del trabajo.

En la **Sección 1.2** se revisan los modelos de elección binaria, en los que la variable dependiente toma sólo dos valores, correspondientes a las dos alternativas posibles. Se comienza con la derivación de estos modelos y se especifican sus distintas formas funcionales, prestando especial atención a los más utilizados: el modelo lineal de probabilidad, el modelo probit y el modelo logit.

En la **Sección 1.3** se trata la estimación de los modelos de elección binaria. En concreto, se resume el procedimiento de estimación por máxima verosimilitud y la obtención de estimaciones máximo-verosímiles por métodos lineales.

La **Sección 1.4** se dedica al contraste de hipótesis. En primer lugar, se considera el problema de contrastar una hipótesis lineal general y, posteriormente, se trata el caso del contraste de hipótesis de exclusión, utilizando el principio de los multiplicadores de Lagrange. Además, se estudia el problema de la derivación de intervalos y regiones de confianza tanto para los parámetros del modelo como para las probabilidades individuales.

En la **Sección 1.5** se plantea el problema de predicción con los modelos de variable dependiente binaria, prestando especial atención a la predicción de las decisiones agregadas de la población objeto de estudio.

1.2. Modelos de elección binaria

Los modelos de elección binaria (MEB) tienen su origen en la experimentación biomédica, en la que, de una forma natural, es posible codificar los resultados en *éxito* y *fracaso*. En este contexto, se supone la existencia de una variable latente o respuesta no observable, cuyos valores dependen de un conjunto de variables, y que da lugar a una observación binaria codificada normalmente como uno o cero.

En el campo de la economía, los modelos de variable cualitativa se plantean en un contexto de toma de decisiones por parte de los individuos. En esta situación, existe una variable latente para cada alternativa, que puede interpretarse como la utilidad asociada a cada una de ellas según es percibida por el decisor. El individuo elige aquella opción que le reporta una mayor utilidad.

Finalmente, se verá que, tanto el planteamiento de variable latente como el de elección discreta, dan lugar a modelos equivalentes.

1.2.1. Derivación de los modelos de elección binaria

Sea una variable endógena latente y_i^* que puede explicarse a través de una función lineal de un conjunto de k variables exógenas $\boldsymbol{x}_i$ y un vector de parámetros α de dimensión $k \times 1$, más un término de perturbación[1], por lo que dicha relación puede plantearse:

$$y_i^* = \boldsymbol{x}_i^T \alpha + \varepsilon_i \qquad \text{con} \quad E(\varepsilon_i) = 0 \quad \text{y} \quad V(\varepsilon_i) = \sigma^2 \quad \forall i \qquad \textbf{[1.2.1]}$$

de modo que si y_i^* sobrepasa un determinado nivel, se observa un *éxito* en el caso i-ésimo y, por el contrario, si se mantiene por debajo de dicho nivel, se observa un *fracaso*. La interpretación de la variable latente y_i^* depende de la naturaleza exacta del problema objeto de estudio. Por ejemplo, en una situación en que el objetivo sea analizar la decisión de un individuo sobre adquirir o no un bien duradero, y_i^* puede considerarse un indicador de la predisposición de compra, que depende de las características concretas del sujeto.

En esta clase de modelos es necesario definir una función indicador $\mathcal{F}(y_i^*)$ que relaciona la variable latente con la observable denotada por y_i, tal que:

---

[1] Aunque el supuesto de linealidad no es necesario en ningún caso, simplifica considerablemente el análisis.

$$y_i = \mathcal{F}(y_i^*) = \begin{cases} 1 & \text{si} \quad y_i^* \geq 0 \\ 0 & \text{si} \quad y_i^* < 0 \end{cases} \qquad \textbf{[1.2.2]}$$

Aunque en este planteamiento no es relevante la forma funcional de $\mathcal{F}(\cdot)$, conviene tener en cuenta que, en determinados casos, definiciones alternativas producen modelos conceptualmente diferentes. Por otra parte, se elige el valor cero como valor crítico de la elección. Sin embargo, esta elección tampoco es relevante para el análisis siempre que se incluya un término constante entre los regresores. En este caso, un punto de corte distinto de cero quedará reflejado como un cambio de origen.

En este contexto, si $F(\cdot)$ es la función de distribución de ε_i, que se supone independiente de $\boldsymbol{x}_i$, se puede definir la *probabilidad de respuesta* o *de elección* como:

$$\begin{aligned} P_i &= P(y_i = 1 \mid \boldsymbol{x}_i) = P(y_i^* = \boldsymbol{x}_i^T\boldsymbol{\alpha} + \varepsilon_i \geq 0) \\ &= P(\varepsilon_i \geq -\boldsymbol{x}_i^T\boldsymbol{\alpha}) = 1 - F(-\boldsymbol{x}_i^T\boldsymbol{\alpha}) \end{aligned} \qquad \textbf{[1.2.3]}$$

o bien:

$$P(y_i = 0 \mid \boldsymbol{x}_i) = 1 - P_i = F(-\boldsymbol{x}_i^T\boldsymbol{\alpha}) \qquad \textbf{[1.2.4]}$$

Si la función de densidad $f(\cdot)$ asociada a $F(\cdot)$ es simétrica, que es la situación más frecuente, entonces:

$$P_i = F(\boldsymbol{x}_i^T\boldsymbol{\alpha}) \qquad \textbf{[1.2.5]}$$

De hecho, el supuesto de simetría no es necesario, puesto que, sin pérdida de generalidad, la ecuación **[1.2.1]** podría haberse planteado como $y_i^* = \boldsymbol{x}_i^T\boldsymbol{\alpha} - \varepsilon_i$, y siguiendo los pasos expuestos anteriormente, se obtiene que $P_i = F(\boldsymbol{x}_i^T\boldsymbol{\alpha})$ sin requerir ninguna característica particular de $F(\cdot)$.

Obsérvese que la ecuación **[1.2.3]** no está definida de forma única respecto al vector $\boldsymbol{\alpha}$, puesto que:

$$P_i = P\left[\frac{\varepsilon_i}{\lambda} \geq \frac{-\boldsymbol{x}_i^T\boldsymbol{\alpha}}{\lambda}\right] \qquad \forall\ \lambda > 0 \qquad \textbf{[1.2.6]}$$

por lo que, convencionalmente, se emplea una normalización específica para cada función de distribución, lo que se traduce en elegir un λ determinado de forma que la varianza de ε_i sea conocida. De este modo, el modelo latente resulta:

$$y_i^* = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i \qquad \text{con} \quad E(\varepsilon_i) = 0 \quad \text{y} \quad V(\varepsilon_i) = \sigma_0^2 \qquad \forall i \qquad \textbf{[1.2.7]}$$

donde $\boldsymbol{\beta} = \boldsymbol{\alpha}/\lambda$ y σ_0^2 es conocido y depende de la función de distribución que se suponga para ε_i, lo que no afecta a la función indicador ni al modelo observable. Incorporando esta restricción, la ecuación **[1.2.5]** resulta:

$$P_i = F(\boldsymbol{x}_i^T\boldsymbol{\beta}) \qquad \textbf{[1.2.8]}$$

Es importante señalar que este tipo de modelos *no permite* estimar el valor de y_i para un vector $\boldsymbol{x}_i$ dado, sino la probabilidad de que y_i tome un cierto valor, es decir, la probabilidad de que el decisor opte por una de las alternativas que se le presentan o alcance un cierto nivel de respuesta a un estímulo.

A continuación, se desarrolla un planteamiento alternativo de los MEB más ligado a la Teoría Económica y que se debe a McFadden (1973). En este contexto, se supone que, en lugar de una variable latente, el individuo percibe dos variables no observables para el investigador, que indican la utilidad que le reporta cada una de las alternativas. Dicha utilidad se considera determinada por una *componente sistemática* (que es determinista y contiene los componentes más representativos de la misma) y por una componente aleatoria (que refleja las característas no observables de los individuos, como los gustos y los errores de medida en las variables). Entonces, manteniendo el supuesto de que la componente sistemática es lineal en los parámetros, se tiene:

$$\begin{aligned} y_{i0}^* &= \boldsymbol{x}_i^T\boldsymbol{\alpha}_0 + \varepsilon_{i0} \\ y_{i1}^* &= \boldsymbol{x}_i^T\boldsymbol{\alpha}_1 + \varepsilon_{i1} \end{aligned} \qquad \textbf{[1.2.9]}$$

En un contexto de maximización de la utilidad, el decisor elige aquella alternativa con mayor valor de la variable latente por lo que, en este caso, la función indicador que relaciona la variable latente con la observable es:

$$y_i = \mathcal{F}(y_i^*) = \begin{cases} 1 & \text{si} \quad y_{i1}^* \geq y_{i0}^* \\ 0 & \text{si} \quad y_{i1}^* < y_{i0}^* \end{cases} \qquad \textbf{[1.2.10]}$$

y la probabilidad de elección resulta:

$$\begin{aligned} P_i &= P(y_i = 1 \mid \boldsymbol{x}_i) = P(y_{i1}^* \geq y_{i0}^*) \\ &= P(\boldsymbol{x}_i^T\boldsymbol{\alpha}_1 + \varepsilon_{i1} \geq \boldsymbol{x}_i^T\boldsymbol{\alpha}_0 + \varepsilon_{i0}) \\ &= P[\varepsilon_{i0} - \varepsilon_{i1} \leq \boldsymbol{x}_i^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_0)] \end{aligned} \qquad \textbf{[1.2.11]}$$

De nuevo, es necesario imponer restricciones para la identificación del modelo. Esto puede hacerse normalizando, en primer lugar, la varianza de la diferencia entre las perturbaciones. Esta normalización depende de la distribución que se suponga y, en general, puede expresarse:

$$P_i = P\left[\frac{\varepsilon_{i0} - \varepsilon_{i1}}{\lambda} \leq \frac{\boldsymbol{x}_i^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_0)}{\lambda}\right] = P[\varepsilon_i \leq \boldsymbol{x}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)] \qquad \textbf{[1.2.12]}$$

para cualquier $\lambda > 0$, donde $\varepsilon_i = (\varepsilon_{i0} - \varepsilon_{i1})/\lambda$ de forma que $V(\varepsilon_i) = \sigma_0^2$ es conocida.

Por otra parte, todavía existen infinitos vectores $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_0$ tales que su diferencia sea igual a la misma constante, por lo que también se impone la restricción:

$$\boldsymbol{\beta}_0 = \mathbf{0}_k, \quad \boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 \qquad \textbf{[1.2.13]}$$

de modo que sólo puede estimarse de forma única el vector $\boldsymbol{\beta}$. Por tanto, la probabilidad de elección resulta:

$$P_i = P(y_i = 1 \mid \boldsymbol{x}_i) = P(\varepsilon_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}) = F(\boldsymbol{x}_i^T\boldsymbol{\beta}) \qquad \textbf{[1.2.14]}$$

donde $F(\cdot)$ es la función de distribución de ε_i.

Aunque con este planteamiento aparecen dos variables aleatorias, tan sólo es necesario hacer algún supuesto sobre la distribución de ε_i, de forma que los modelos resultantes de la derivación de variable latente y de decisión son equivalentes.

En las dos formas alternativas de derivación de modelos binarios que acaban de exponerse, se ha supuesto que el vector x_i está formado por un conjunto de variables que caracterizan al decisor. Sin embargo, un modelo más general incluiría entre las variables explicativas un conjunto z_{ij} ($j = 0, 1$) de q variables que caracterizan las alternativas disponibles tal y como las percibe el individuo. Este enfoque da lugar a los denominados *modelos de elección condicional* [Luce (1963), McFadden (1973)]. En ellos, el modelo latente es:

$$y_{ij}^{*} = x_i^{T}\alpha_j + z_{ij}^{T}\delta + \varepsilon_{ij} \qquad j = 0, 1 \qquad \textbf{[1.2.15]}$$

La probabilidad de elección, incluyendo restricciones análogas a las de **[1.2.12]** y **[1.2.13]** resulta:

$$P_i = P(\varepsilon_i \leq x_i^{T}\beta + (z_{i1}-z_{i0})^{T}\gamma) \qquad \textbf{[1.2.16]}$$

y denotando por:

$$x_i^{*} = \begin{bmatrix} x_i \\ z_{i1} - z_{i0} \end{bmatrix} \quad \text{y} \quad \beta^{*} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \qquad \textbf{[1.2.17]}$$

el modelo **[1.2.15]** puede reescribirse como un MEB semejante al de la ecuación **[1.2.7]**. Esta generalización no aporta nada al análisis en modelos binarios, pero sí es relevante en una situación de elección múltiple.

1.2.2. Formas funcionales

Una vez derivado conceptualmente el modelo, para definirlo es necesario especificar cómo se distribuye la perturbación o bien la diferencia entre las perturbaciones de cada alternativa, lo que da lugar a los distintos modelos concretos. En este trabajo, tan sólo se suponen tres características generales de la función $F(\cdot)$: i) que es continua y dos veces diferenciable, ii) que es estrictamente creciente en todo su dominio de definición y iii) que su función de densidad asociada es unimodal, esto es, que tiene un sólo óptimo local que, consecuentemente, es el óptimo global.

1.2.2.A. El modelo lineal de probabilidad

El modelo más simple, que se conoce como modelo lineal de probabilidad (MLP), se deriva a partir de la hipóteisis de que ε_i sigue una distribución uniforme entre dos valores prefijados $-L$ y L, con $L > 0$. La probabilidad de elección en este caso resulta:

$$P_i = \begin{cases} 0 & \text{si} \quad x_i^T\beta < -L \\ \int_{-L}^{x_i^T\beta} f(\varepsilon_i)\, d\varepsilon_i = \dfrac{x_i^T\beta + L}{2L} & \text{si} \quad -L \le x_i^T\beta \le L \\ 1 & \text{si} \quad x_i^T\beta > L \end{cases} \qquad \textbf{[1.2.18]}$$

Pese a su simplicidad, este modelo presenta serios inconvenientes teóricos especialmente en los puntos $-L$ y L, donde la primera derivada no es continua. Además, al trabajar con muestras reales, inevitablemente ocurre que algunos individuos elijen una alternativa para la que la probabilidad prevista es cero. Este problema se debe al supuesto de que la función de densidad de ε_i se hace cero a partir de los puntos $-L$ y L[2].

Debido a estos problemas, se han desarrollado modelos que describen de forma más realista las probabilidades de elección. En la práctica, el modelo lineal de probabilídad sólo se utiliza para calcular estimaciones iniciales consistentes de los parámetros, que se emplearán para inicializar procedimientos iterativos de estimación de otros modelos.

1.2.2.B. El modelo probit binario

Una hipótesis lógica sobre las perturbaciones es suponer que son la suma de un número elevado de componentes no observables, que conducen a la decisión. Dado este supuesto, por el teorema central del límite, la distribución de las perturbaciones convergería a una distribución normal.

Concretamente, resulta habitual suponer que ε_{i0} y ε_{i1} siguen distribuciones normales con media nula y varianza finita, por lo que ε_i será también normal con media cero.

---

[2] *Obsérvese que la elección de los límites de la distribución es irrelevante, puesto que no son más que* parámetros de escala, y generalmente se emplea $L = 1/2$.

Incorporando la restricción de identificación dada en **[1.2.12]**, que para el modelo probit es $V(\varepsilon_i) = \sigma_0^2 = 1$, la probabilidad de elección a partir de **[1.2.14]** es:

$$\begin{aligned} P_i &= P(\varepsilon_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}) \\ &= \int_{-\infty}^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\varepsilon_i^2)\, d\varepsilon_i = \int_{-\infty}^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \phi(\varepsilon_i)\, d\varepsilon_i = \\ &= \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}) \end{aligned} \qquad \textbf{[1.2.19]}$$

donde $\Phi(\cdot)$ y $\phi(\cdot)$ denotan, respectivamente, la función de distribución y de densidad de una variable aleatoria normal estándar. Obsérvese que las probabilidades de elección del modelo probit binario no dependen de las varianzas de cada perturbación o su covarianza.

1.2.2.C. El modelo logit binario

Aunque el modelo probit es intuitivamente razonable y el supuesto sobre la distribución de las perturbaciones tiene un fundamento teórico, presenta el inconveniente de que la probabilidad de respuesta no tiene una forma cerrada; es decir, se expresa mediante una integral. Esto hace que, en algunos casos, sea conveniente utilizar una función de distribución que, manteniendo la forma de la normal, sea más sencilla analíticamente. Uno de estos modelos es el *logit binario*.

El modelo logit supone que ε_i sigue una distribución logística o, más precisamente, una *distribución del cuadrado de la secante hiperbólica* ($sech^2$), cuya función de distribución es:

$$F(u) = \frac{\exp[(u-\mu)/\tau]}{1+\exp[(u-\mu)/\tau]}, \qquad -\infty < u < \infty \qquad \textbf{[1.2.20]}$$

y cuya función de densidad asociada es:

$$f(u) = \frac{\exp[(u-\mu)/\tau]}{\tau\{1+\exp[(u-\mu)/\tau]\}^2} \qquad \textbf{[1.2.21]}$$

donde μ es la esperanza de la distribución y la desviación típica es $\pi\tau/\sqrt{3}$. Se conoce como función de distribución *logística estándar* cuando en **[1.2.20]**, $\mu = 0$ y $\tau = 1$, y en este caso se cumple que:

$$f(u) = F(u)\,[1 - F(u)] \qquad \textbf{[1.2.22]}$$

La distribución logística tiene un aspecto semejante a la normal, aunque las colas son más gruesas. Para obtener la distribución logística de ε_i, es necesario suponer que ε_{i0} y ε_{i1} son independientes y se *distribuyen idénticamente valor extremo (distribución del valor extremo tipo I o Gumbel* [Johnson y Kotz (1970)][3]). Como en los casos anteriores, es necesario imponer una normalización en la escala y usar la distribución logística estándar, con lo que se tiene que $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma_0^2 = \pi^2/3$.

Dada esta hipótesis, la probabilidad de elección puede expresarse:

$$\begin{aligned} P_i &= P(\varepsilon_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}) \\ &= \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\boldsymbol{x}_i^T\boldsymbol{\beta})} \qquad \textbf{[1.2.23]} \\ &= \Lambda(\boldsymbol{x}_i^T\boldsymbol{\beta}) \end{aligned}$$

donde, en lo sucesivo, $\Lambda(\cdot)$ y $\lambda(\cdot)$ denotan la función de distribución y de densidad de una variable aleatoria logística estándar.

Es importante señalar que la diferente normalización de escala elegida para cada uno de los tres modelos expuestos, hace que los parámetros estimados no sean directamente comparables. Para relacionar los parámetros del modelo lineal de probabilidad (MLP) con $L = 1/2$ y el probit con $\sigma = 1$, conviene hacer notar que la desviación típica de la distribución uniforme $(-L, L)$ es $L/\sqrt{3}$ y, en nuestro caso, está fijada en $1/2\sqrt{3}$, mientras que es unitaria para el modelo probit. Esta normalización implica que los coeficientes del modelo probit serán $2\sqrt{3}$ veces mayores que los del modelo lineal de probabilidad. Siguiendo un razonamiento análogo, los coeficientes del modelo logit serán $\pi/\sqrt{3}$ veces mayores que los del correspondiente modelo probit normalizado. Esta equivalencia puede tener dos aplicaciones: i) hacer comparables modelos que no lo serían, al tener diferentes escalas los vectores de parámetros, y ii) emplear las estimaciones del MLP transformadas, como condiciones iniciales para estimar los otros modelos por métodos numéricos iterativos.

---

[3] La distribución del valor extremo (EVD) se deriva del límite de la distribución del mayor (o menor) valor de un conjunto de variables aleatorias. La EVD del mayor valor tiene función de distribución $F(u) = \exp\{-\exp[-(u-\mu)/\tau]\}$, y es unimodal y no simétrica.

Finalmente, es importante señalar que, aunque se ha mantenido el supuesto de linealidad de los argumentos en la probabilidad de respuesta, el modelo logit permite introducir de forma sencilla un cierto tipo de no linealidad. Este hecho resulta particularmente interesante para representar *diferentes esquemas de comportamiento de la probabilidad de elección*, sin tener que emplear diferentes funciones de distribución. En general, un modelo univariante dicotómico se puede escribir:

$$P_i = P(y_i = 1) = F(v(\boldsymbol{x}_i, \boldsymbol{\theta})), \qquad i = 1, 2, \ldots, n \qquad \textbf{[1.2.24]}$$

donde $v(\cdot)$ es una función arbitraria elegida por el investigador. Cuando $F(\cdot)$ es la distribución logística y se mantiene la relación lineal entre variables y parámetros, se obtiene lo que McFadden (1978) denomina *logit universal* [Amemiya (1981)].

A partir de **[1.2.23]** y **[1.2.24]**, la probabilidad de elección resulta:

$$P_i = \Lambda^*(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \frac{1}{1 + \exp[-v_i(\boldsymbol{x}_i^T\boldsymbol{\beta})]} \qquad \textbf{[1.2.25]}$$

donde, por ejemplo, si se supone que $v(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \delta_0 + \delta_1(\boldsymbol{x}_i^T\boldsymbol{\beta})$, para $\delta_0 \approx 0.084$ y $\delta_1 \approx 1.702$, la función anterior es aproximadamente igual a la función de distribución normal. También, con valores de los parámetros $\delta_0 \approx 0.012$ y $\delta_1 \approx 0.601$ se consigue una aproximación aceptable a la función de distribución uniforme entre -4 y 4. Estas aproximaciones se muestran en las **Figuras 1.1** y **1.2**.

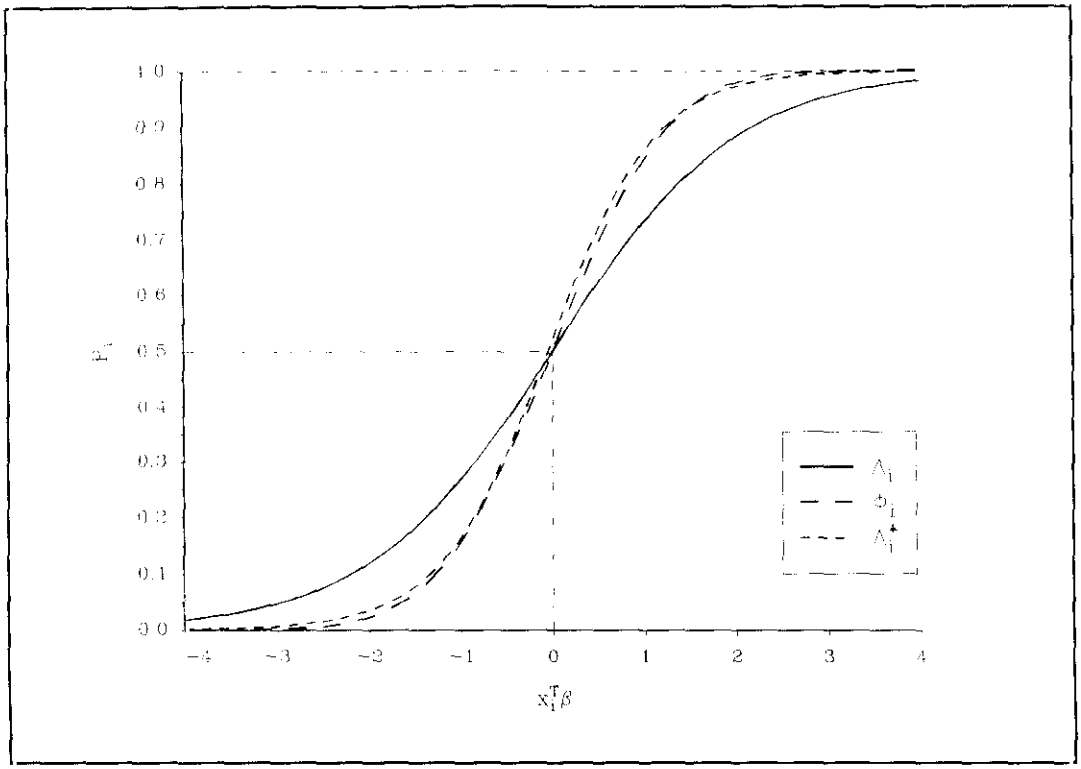

Figura 1.1: Aproximación de la distribución normal mediante un modelo logit generalizado.

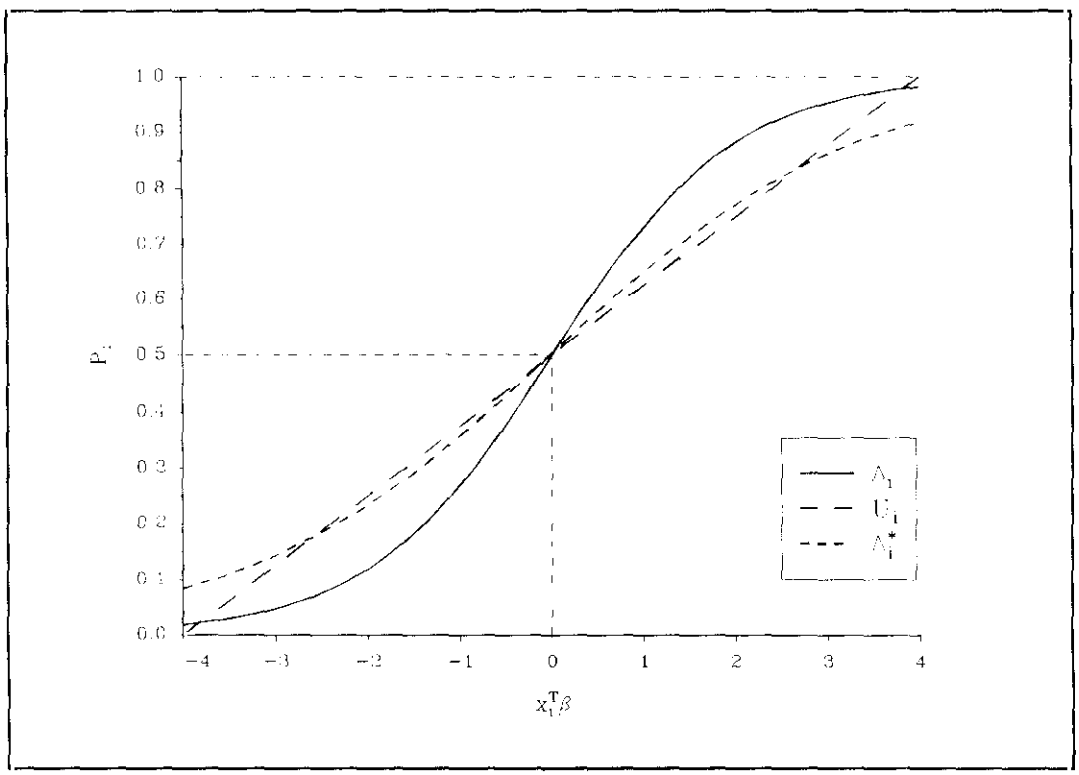

Figura 1.2: Aproximación de la distribución uniforme mediante un modelo logit generalizado.

1.2.2.D. Otros modelos binarios

Aunque los modelos probit y logit son, con diferencia, los más utilizados, se han propuesto otras funciones de distribución para proporcionar diferentes patrones de respuesta de las probabilidades de elección. Aunque este trabajo se centra en los modelos logit y probit, los resultados son válidos para otros modelos, como los que se exponen a continuación.

Otras funciones de distribución que se han considerado en la literatura son las siguientes:

1. *Distribución de Cauchy o del arcotangente* estándar.

$$P_i = \frac{1}{2} + \frac{1}{\pi}\arctan(x_i^T\beta) \qquad \textbf{[1.2.26]}$$

cuya función de densidad asociada es:

$$f(u) = \frac{1}{\pi} \cdot \frac{1}{1 + u^2}, \qquad -\infty < u < \infty \qquad \textbf{[1.2.27]}$$

2. *Distribución de Burr* [Burr (1942)].

$$P_i = 1 - \frac{1}{[1 + (x_i^T\beta)^c]^k}, \qquad c,k > 0,\ x_i^T\beta > 0 \qquad \textbf{[1.2.28]}$$

cuya función de densidad asociada es:

$$f(u) = \frac{cku^{c-1}}{[1 + u^c]^{k+1}}, \qquad c,k > 0,\ u > 0 \qquad \textbf{[1.2.29]}$$

y que se usa, fundamentalmente, para tratar variables aleatorias que toman solamente valores positivos.

3. *Exponencial truncada por la derecha.*

$$P_i = \begin{cases} \exp[-(L - x_i^T\beta)] & si \quad x_i^T\beta < L \\ 1 & si \quad x_i^T\beta \geq L \end{cases} \qquad \textbf{[1.2.30]}$$

4. *Exponencial truncada por la izquierda.*

$$P_i = \begin{cases} 1 - \exp[-(x_i^T\beta - L)] & si \quad x_i^T\beta > L \\ 0 & si \quad x_i^T\beta \leq L \end{cases} \qquad \textbf{[1.2.31]}$$

donde, en **[1.2.30]** y **[1.2.31]**, L es el punto de truncación.

1.3. Estimación de Modelos de Elección Binaria

En general, la estimación de los modelos de respuesta binaria se puede llevar a cabo por máxima verosimilitud, con la excepción del modelo lineal de probabilidad, que puede estimarse consistentemente (aunque no eficientemente) por mínimos cuadrados ordinarios (*MCO*).

En esta sección, como en el resto de este trabajo, el análisis se restringe al caso más frecuente en aplicaciones económicas, que es el de observaciones individuales obtenidas por muestreo aleatorio simple. Un tratamiento de la estimación con datos agrupados, esto es, cuando se dispone de un conjunto importante de observaciones con *idéntico* vector de características, se encuentra en Cox y Snell (1989) o en Amemiya (1981) entre otros. Por otra parte, la formulación y las condiciones de existencia, así como las propiedades del estimador máximo-verosímil (EMV) bajo distintos diseños de muestreo, pueden revisarse en Cosslett (1981a y 1981b) y Manski y McFadden (1981). Entre los esquemas de muestreo no aleatorio simple, el más frecuente en la práctica es el *muestreo basado en la elección* (*choice based sampling*), cuyo primer tratamiento aparece en Manski y Lerman (1977).

1.3.1. Estimación de Máxima Verosimilitud

Dado el modelo en **[1.2.13]**, para una muestra aleatoria de tamaño n, la función de verosimilitud es:

$$\mathcal{L}(\beta \mid X, y) = \prod_{i=1}^{n} P_i^{y_i} (1 - P_i)^{(1-y_i)} \qquad \textbf{[1.3.1]}$$

Definiendo $F_i = F(x_i^T\beta) = P_i$ y tomando logaritmos se tiene:

$$\ell(\beta \mid X, y) = \ln\mathcal{L} = \sum_{i=1}^{n} [y_i \ln F_i + (1 - y_i)\ln(1 - F_i)] \qquad \textbf{[1.3.2]}$$

El *vector gradiente* del logaritmo de la función de verosimilitud es:

$$\nabla \ell = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{f_i}{F_i(1-F_i)}(y_i - F_i)\boldsymbol{x}_i \qquad \textbf{[1.3.3]}$$

donde f_i es la función de densidad asociada a F_i, mientras que el *hessiano* puede escribirse:

$$\nabla^2 \ell = \frac{\partial^2 \ell}{\partial \beta^T \partial \beta} = -\sum_{i=1}^{n} \left[\frac{y_i}{F_i^2} \frac{1-y_i}{(1-F_i)^2} \right] f_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T + \sum_{i=1}^{n} \left[\frac{(y_i - F_i)}{F_i(1-F_i)} \right] f_i' \boldsymbol{x}_i \boldsymbol{x}_i^T \qquad \textbf{[1.3.4]}$$

siendo f'_i la derivada de la función de densidad respecto a $\boldsymbol{x}_i^T \beta$.

Asimismo, la *matriz de información* es la menos esperanza de **[1.3.4]**, esto es:

$$I(\beta) = -E(\nabla^2 \ell) = \sum_{i=1}^{n} \frac{f_i^2}{F_i(1-F_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T \qquad \textbf{[1.3.5]}$$

Bajo condiciones de regularidad no muy restrictivas, y especialmente en los modelos en los que se centra este trabajo, la función de verosimilitud es globalmente cóncava [véase McFadden (1973 y 1983), Amemiya (1985) o Núñez (1990), entre otros]. Por tanto, el estimador máximo verosímil $\hat{\beta}$ de β, si existe, es único. Además es consistente y $\sqrt{n}\,(\hat{\beta} - \beta)$ se distribuye asintóticamente normal con esperanza $\mathbf{0}_k$ y matriz de varianzas covarianzas $[(1/n)\, I(\hat{\beta})]^{-1}$. Los resultados necesarios de teoría asintótica que permiten concluir lo anterior se encuentran en Silvey (1970), Cox y Hinkley (1974) y Amemiya (1985). Un planteamiento general sobre la concavidad de la función de verosimilitud para modelos binarios debida a Nuñez (1990) se presenta en el **Apéndice A.1**.

La maximización de la función de verosimilitud o, lo que es equivalente, la solución de las ecuaciones de las condiciones de primer orden de **[1.3.3]**, debe llevarse a cabo numéricamente. En el **Apéndice A.2** se presenta una revisión breve de los métodos habitualmente empleados para resolver el problema de maximizar una función objetivo sin restricciones.

1.3.2. Estimación de máxima verosimilitud por procedimientos lineales

Siguiendo a Amemiya (1985), un algoritmo de optimización más eficiente computacionalmente que el de Newton, y que permite una interesante interpretación del modelo **[1.2.13]** es el *método de scoring* de Fisher. En este algoritmo se emplea la matriz de información como aproximación al hessiano y el paso en cada iteración viene dado por [véase **Apéndice A.2**]:

$$\hat{\beta}^{\tau+1} = \hat{\beta}^{\tau} + [I(\hat{\beta}^{\tau})]^{-1} \nabla \ell(\hat{\beta}^{\tau}) \qquad \textbf{[1.3.6]}$$

La expresión **[1.3.6]** puede también plantearse:

$$\hat{\beta}^{\tau+1} = [I(\hat{\beta}^{\tau})]^{-1} [\nabla \ell(\hat{\beta}^{\tau}) + I(\hat{\beta}^{\tau}) \hat{\beta}^{\tau}] \qquad \textbf{[1.3.8]}$$

y sustituyendo el gradiente y la matriz de información por las expresiones **[1.3.3]** y **[1.3.5]** resulta:

$$\begin{aligned} \nabla \ell(\hat{\beta}^{\tau}) + I(\hat{\beta}^{\tau}) \hat{\beta}^{\tau} &= \sum_{i=1}^{n} \frac{\hat{f}_i (y_i - \hat{F}_i)}{\hat{F}_i (1 - \hat{F}_i)} \boldsymbol{x}_i + \sum_{i=1}^{n} \frac{\hat{f}_i^2}{\hat{F}_i (1 - \hat{F}_i)} \boldsymbol{x}_i (\boldsymbol{x}_i^T \hat{\beta}^{\tau}) \\ &= \sum_{i=1}^{n} \frac{\hat{f}_i}{\hat{F}_i (1 - \hat{F}_i)} \boldsymbol{x}_i (y_i + \hat{f}_i (\boldsymbol{x}_i^T \hat{\beta}^{\tau}) - \hat{F}_i) \end{aligned} \qquad \textbf{[1.3.9]}$$

donde la virgulita sobre F_i y f_i denota que estas funciones han sido evaluadas en el vector $\hat{\beta}^{\tau}$. Agrupando términos y operando, la ecuación **[1.3.8]** puede escribirse como:

$$\hat{\beta}^{\tau+1} = \left[\sum_{i=1}^{n} \frac{\hat{f}_i^2}{\hat{F}_i (1 - \hat{F}_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T \right]^{-1} \sum_{i=1}^{n} \frac{\hat{f}_i}{\hat{F}_i (1 - \hat{F}_i)} \boldsymbol{x}_i (y_i + \hat{f}_i (\boldsymbol{x}_i^T \hat{\beta}^{\tau}) - \hat{F}_i) \qquad \textbf{[1.3.9]}$$

que puede interpretarse como el estimador por mínimos cuadrados ordinarios (MCO) de β en la regresión:

$$\tilde{y}_i = \tilde{\boldsymbol{x}}_i^T \beta + u_i \qquad \textbf{[1.3.10]}$$

donde

$$\tilde{y}_i = \frac{y_i - \hat{f}_i \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^\tau + \hat{F}_i}{[\hat{F}_i(1 - \hat{F}_i)]^{1/2}} \qquad \textbf{[1.3.11]}$$

y

$$\tilde{\boldsymbol{x}}_i = \frac{\hat{f}_i}{[\hat{F}_i(1 - \hat{F}_i)]^{1/2}} \boldsymbol{x}_i \qquad \textbf{[1.3.12]}$$

Por tanto, un procedimiento de estimación lineal máximo-verosímil para cualquier modelo binario vendría dado por los siguientes pasos:

Paso 0: Elegir un vector $\hat{\boldsymbol{\beta}}^0$ de estimaciones iniciales. Fijar el contador de iteraciones a cero, $\tau = 0$ y elegir una tolerancia para el criterio de parada e_1.

Paso 1: Verificar si se ha producido la convergencia, usando los criterios de parada establecidos. De no ser así, seguir con el paso 2.

Paso 2: Evaluar las funciones de densidad y distribución del modelo a estimar en $\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^\tau$ y transformar la variable y_i y el vector $\boldsymbol{x}_i$ de acuerdo con las expresiones **[1.3.11]** y **[1.3.12]**. Seguir con el paso 3.

Paso 3: Estimar por MCO la regresión **[1.3.10]**. Los estimadores que resultan son idénticos a los obtenidos por el método de scoring para cada iteración. Hacer $\tau = \tau + 1$ y volver al paso 1.

Otra derivación alternativa del algoritmo propuesto, puede llevarse a cabo [Amemiya (1981, 1985)] del siguiente modo. Sea el modelo no lineal:

$$y_i = F(\boldsymbol{x}_i^T \boldsymbol{\beta}) + u_i \qquad \textbf{[1.3.13]}$$

donde u_i es una variable binaria que toma los valores $1\text{-}P_i$ con probabilidad P_i y $-P_i$ con probabilidad $1\text{-}P_i$, de forma que:

$$E(u_i) = 0 \quad \text{y} \quad V(u_i) = P_i(1 - P_i) \qquad \textbf{[1.3.14]}$$

Si se lleva a cabo una aproximación lineal del modelo mediante una expansión por Taylor de $F(\boldsymbol{x}_i^T \boldsymbol{\beta}^{\tau+1})$ en torno a un vector de condiciones iniciales $\hat{\boldsymbol{\beta}}^\tau$, se obtiene que:

$$F(x_i^T\beta^{\tau+1}) = F(x_i^T\hat{\beta}^{\tau}) + f(x_i^T\hat{\beta}^{\tau})x_i^T(\beta^{\tau+1} - \hat{\beta}^{\tau}) + R_i \qquad \textbf{[1.3.15]}$$

Teniendo en cuenta que $R_i \to 0$ en probabilidad si $\hat{\beta}^{\tau}$ es una estimación consistente de β, sustituyendo el modelo **[1.3.13]** en la ecuación **[1.3.15]**, la aproximación puede escribirse:

$$y_i - F(x_i^T\hat{\beta}^{\tau}) + f(x_i^T\hat{\beta}^{\tau})(x_i^T\hat{\beta}^{\tau}) = f(x_i^T\hat{\beta}^{\tau})\,x_i^T\beta^{\tau+1} + u_i \qquad \textbf{[1.3.16]}$$

El modelo en **[1.3.16]** es lineal con perturbaciones heterocedásticas, cuya varianza es $F_i(1\text{-}F_i)$. Por lo tanto, la estimación eficiente del mismo por mínimos cuadrados ponderados, coincide con la estimación MCO del modelo **[1.3.10]**.

1.4. Contraste de hipótesis

En esta sección se presentan un conjunto de resultados sobre el contraste de hipótesis en los modelos de elección binaria. En primer lugar, se trata el caso de hipótesis lineales, y posteriormente, en un contexto más general, se utiliza el principio de Multiplicadores de Lagrange para plantear contrastes para hipótesis nulas del tipo H_0: $\alpha = \mathbf{0}_q$; esto es, de exclusión de algún tipo de variables, para cualquier función $F(\cdot)$ que cumpla las condiciones de regularidad.

1.4.1. Contraste de restricciones lineales

Se considera el contraste de una hipótesis general lineal de la forma:

$$H_0: R\beta = r \qquad \textbf{[1.4.1]}$$

donde R es una matriz de constantes conocidas $m \times k$, r es un vector $m \times 1$ también conocido y además se cumple que $m \leq k$ y que rango$(R) = m$. En lo que sigue, se distingue entre el caso en que $m = 1$ y $m > 1$.

Siguiendo a Amemiya (1985), es sencillo comprobar que si $m = 1$, bajo la hipótesis nula:

$$\frac{R\hat{\beta} - r}{\sqrt{R\, V(\hat{\beta})^{-1} R^T}} \rightarrow N(0, 1) \qquad \textbf{[1.4.2]}$$

donde $V(\hat{\beta})$ es una estimación consistente de la matriz de covarianzas del vector de parámetros estimado, que generalmente será la inversa de la matriz de información. También se puede utilizar la distribución t_{n-k} para llevar a cabo el contraste aunque, al utilizar muestras grandes, hecho en el que se apoyan los resultados de convergencia, resulta más apropiado el empleo de la distribución normal [Amemiya (1981)].

Un caso particular del que se acaba de exponer es el contraste de significación de un parámetro individual. En ese caso, el estadístico de contraste se reduce a:

$$\frac{\hat{\beta}_j}{\sqrt{V(\hat{\beta}_j)}} \rightarrow N(0, 1) \qquad \textbf{[1.4.3]}$$

donde $V(\hat{\beta}_j)$ es el elemento j-ésimo de la diagonal principal de $V(\hat{\beta})$.

Cuando $m > 1$, puede emplearse alguno de los conocidos contrastes de Wald (WT), Razón de Verosimilitudes (LRT) o Multiplicadores de Lagrange (LMT) [Engle (1983)]. El contraste de Wald requiere un estimador no restringido consistente y asintóticamente normal, así como una estimación consistente de la matriz de covarianzas[4]. El estadístico correspondiente puede escribirse:

$$WT = (R\hat{\beta} - r)^T [R\, V(\hat{\beta})\, R^T]^{-1} (R\hat{\beta} - r) \rightarrow \chi^2_m \qquad \textbf{[1.4.4]}$$

Obsérvese que si $m = 1$, el estadístico de Wald se reduce al cuadrado del estadístico que figura en **[1.4.2]**.

El estadístico de contraste de razón de verosimilitudes se formula como:

$$LRT = 2[\ell(\hat{\beta}) - \ell(\hat{\beta}_R)] \rightarrow \chi^2_m \qquad \textbf{[1.4.5]}$$

donde $\ell(\hat{\beta})$ y $\ell(\hat{\beta}_R)$ denotan, respectivamente, el logaritmo de la función de verosimilitud evaluada en el máximo y en el estimador bajo la hipótesis nula. Este estadístico está asociado a la estimación máximo-verosímil y para su cálculo es necesario evaluar tanto el máximo de la función de verosimilitud del modelo sin restricciones como el de la función de verosimilitud bajo la hipótesis nula, lo que generalmente supone obtener ambos estimadores. Por esta razón, este estadístico no suele emplearse en la práctica para el contraste de restricciones lineales generales.

El contraste de multiplicadores de Lagrange también está ligado a la estimación máximo-verosímil y requiere evaluar el gradiente del logaritmo de la función de verosimilitud bajo la hipótesis nula, así como una estimación consistente de la matriz de covarianzas. El estadístico de contraste puede escribirse [Engle (1983)]:

$$LMT = \nabla\ell(\hat{\beta}_R)^T I(\hat{\beta}_R)^{-1} \nabla\ell(\hat{\beta}_R) \rightarrow \chi^2_m \qquad \textbf{[1.4.6]}$$

---

[4] Aunque es habitual usar el estimador máximo-verosímil y la inversa de la matriz de información, no es necesario para la aplicación de este contraste [Amemiya (1981)].

1.4.2. Contraste general de hipótesis de exclusión basado en el principio de multiplicadores de Lagrange

En este apartado se analiza el contraste de hipótesis paramétricas generales que se pueden plantear como H_0: $\alpha = \mathbf{0}_q$ [Godfrey (1988, cap. 6)]. Este tipo de contrastes de exclusión es muy flexible si se introducen relaciones no lineales entre las variables. Sea un modelo más general que los que se han tratado en el apartado anterior, en el que la componente sistemática es una función genérica $v(z_i^T\alpha, x_i^T\beta)$ donde z_i es un vector de q variables exógenas y α es un vector $q\times 1$ de parámetros desconocidos. La función $v(\cdot)$ es tal que verifica:

$$v(0, x_i^T\beta) = x_i^T\beta \qquad \textbf{[1.4.7]}$$

siendo continua y diferenciable con primeras derivadas continuas. Bajo el supuesto anterior, siguiendo los pasos dados en las ecuaciones **[1.2.1]**-**[1.2.5]** se tiene que:

$$P_i = F[v(z_i^T\alpha, x_i^T\beta)] \qquad \textbf{[1.4.8]}$$

Aplicando el principio de los *multiplicadores de Lagrange* (LM) [Engle (1983)], la adecuación del modelo se puede analizar contrastando el conjunto de restricciones H_0: $\alpha = \mathbf{0}_q$. La construcción del estadístico de contraste requiere evaluar el gradiente del logaritmo de la función de verosimilitud bajo la hipótesis nula y el estimador máximo-verosímil bajo la hipótesis nula. Para simplificar la notación definimos: i) $\theta^T \equiv (\alpha^T, \beta^T)$; ii) $v_i(\theta) \equiv v(z_i^T\alpha, x_i^T\beta)$; iii) $v_{ir}(\theta) \equiv \partial v_i(\theta)/\partial\theta_r$ donde θ_r es el componente r-ésimo del vector θ, $r = 1, \ldots, k+q$; iv) $\nabla v_i(\theta) \equiv \partial v_i(\theta)/\partial\theta$; v) $F_i(\theta) \equiv F[v_i(\theta)]$. El EMV restringido se denota por $\hat{\theta} = (\theta_q^T, \hat{\beta}^T)^T$.

El vector gradiente del logaritmo de la función objetivo resulta en este caso:

$$\nabla \ell = \frac{\partial \ln \mathcal{L}}{\partial\theta} = \sum_{i=1}^{n} \frac{f_i}{F_i(1-F_i)}(y_i - F_i)\nabla v_i(\theta) \qquad \textbf{[1.4.9]}$$

Davidson y MacKinnon (1984) argumentan que el cálculo del hessiano puede ser complicado en este contexto y utilizan la forma OPG[5] equivalente, de modo que estiman la matriz de información como $W(\hat{\theta})^T W(\hat{\theta})$, donde $W(\hat{\theta})$ es una matriz $n\times(k+q)$ cuyas filas son los vectores gradiente de cada observación. El estadístico puede calcularse como:

---

[5] *Outer Product of the Gradient*: producto exterior del gradiente.

$LM_1 = nR^2_n$, donde R^2_n es el R^2 *no centrado* de la regresión de un vector de unos sobre $W(\hat{\theta})$ y sigue una distribución χ^2_q. Una versión equivalente del test es:

$$F_1 = \left[\frac{R_n^2}{1 - R_n^2}\right]\left[\frac{n-k-q}{q}\right] = \left[\frac{LM_1}{q}\right]\left[\frac{n-k-q}{n(1-R_n^2)}\right] \qquad \textbf{[1.4.10]}$$

que, bajo la hipótesis nula, sigue una distribución centrada $F_{q,\,n-k-q}$. La evidencia para otros modelos econométricos [Godfrey (1988, cap. 3)], es que este estadístico no tiene un buen comportamiento en muestras pequeñas. La alternativa es, claramente, utilizar la matriz de información en lugar de su estimación con el equivalente OPG. La matriz de información, en este caso, resulta:

$$I(\theta) = -E(\nabla^2 \mathcal{L}) = \sum_{i=1}^{n} \frac{f_i^2}{F_i(1-F_i)} \nabla v_i(\theta) \nabla v_i(\theta)^T \qquad \textbf{[1.4.11]}$$

por lo que el estadístico de contraste [Engle (1983)] es:

$$LM_2 = \nabla \ell(\hat{\theta})^T [I(\hat{\theta})]^{-1} \nabla \ell(\hat{\theta}) \qquad \textbf{[1.4.12]}$$

que puede interpretarse como la suma explicada en la regresión de los residuos estandarizados:

$$s_i(\hat{\theta}) = \frac{y_i - F_i(\hat{\theta})}{[F_i(\hat{\theta})(1 - F_i(\hat{\theta}))]^{1/2}} \qquad \textbf{[1.4.13]}$$

sobre:

$$S_i(\hat{\theta}) = \frac{f_i(\hat{\theta})}{[F_i(\hat{\theta})(1 - F_i(\hat{\theta}))]^{1/2}} \nabla v_i(\hat{\theta})^T \qquad \textbf{[1.4.14]}$$

con lo que el estadístico en **[1.4.6]** puede escribirse:

$$LM_2 = s(\hat{\theta})^T S(\hat{\theta}) [S(\hat{\theta})^T S(\hat{\theta})]^{-1} S(\hat{\theta})^T s(\hat{\theta}) \qquad \textbf{[1.4.15]}$$

y es análogo a los estadísticos derivados en Gourieroux et al. (1987) que se basan en el concepto de *residuos generalizados*.

La hipótesis H_0: $\alpha = \mathbf{0}_q$ se puede contrastar comparando el valor del estadístico con el valor crítico de una distribución χ^2_q. El estadístico LM_2, tiene una versión equivalente que se distribuye como una $F_{q,n-k-q}$ y análogo a F_1 cuya expresión es:

$$F_2 = \left[\frac{LM_2}{q}\right] \Big/ \left[\frac{n-k-q}{n(1-R_n^2)}\right] \qquad \textbf{[1.4.16]}$$

Por otra parte, para contrastar la hipótesis de omisión de variables en un MEB, se puede utilizar el siguiente planteamiento, donde se considera:

$$v_i(\theta) = x_i^T\beta + z_i^T\alpha \qquad \textbf{[1.4.17]}$$

como expresión del modelo alternativo. Como en el caso del modelo lineal general (MLG), el EMV restringido $\hat{\beta}$ no será consistente para β si $\alpha \neq \mathbf{0}_q$. El problema es que, mientras que las expresiones de la inconsistencia del estimador son sencillas de derivar para el modelo de regresión lineal general, en el caso de los modelos de elección binaria se debe recurrir a aproximaciones basadas en el supuesto de que los elementos de α están próximos a cero. Godfrey (1988) considera aproximaciones al vector esperanza de la distribución de $\sqrt{n}(\hat{\theta} - \theta)$ bajo una secuencia de alternativas locales H_n: $\alpha = \delta/\sqrt{n}$, $\delta^T\delta < \infty$.

Sea $d_2(\theta) = \partial\ell(\theta)/\partial\beta$, $D_{21}(\theta) = \partial^2\ell(\theta)/\partial\beta\partial\alpha^T$ y $D_{22}(\theta) = \partial^2\ell(\theta)/\partial\beta\partial\beta^T$. El EMV restringido satisface $d_2(\hat{\theta}) = \mathbf{0}_q$ y, bajo H_n, se tiene:

$$d_2(\hat{\theta}) \approx d_2(\theta) - D_{21}(\theta) + D_{22}(\theta)(\hat{\beta} - \beta) \qquad \textbf{[1.4.18]}$$

y de lo anterior, despejando, resulta:

$$\sqrt{n}(\hat{\beta} - \beta) \approx -\sqrt{n}[D_{22}(\theta)]^{-1}d_2(\theta) + [D_{22}(\theta)]^{-1}D_{21}(\theta)\delta \qquad \textbf{[1.4.19]}$$

El primer elemento del lado derecho de **[1.4.19]** es asintóticamente normal con esperanza nula y matriz de covarianzas finita, mientras que el segundo término converge en probabilidad a un vector con elementos finitos, no todos nulos. Por tanto, el vector esperanza de la distribución asintótica de $\sqrt{n}(\hat{\beta} - \beta)$ es plim $[D_{22}(\theta)]^{-1}D_{21}(\theta)\delta$. Así, el efecto aproximado (*inconsistencia local*[6]) debido a la omisión de las variables de Z puede estimarse:

---

[6] El término inconsistencia local no es riguroso. El EMV de β es consistente bajo H_n con $(\hat{\beta} - \beta)$ $O_p(n^{-1})$.

$$\frac{1}{\sqrt{n}}[D_{22}(\hat{\theta})]^{-1}D_{21}(\hat{\theta})\delta = [D_{22}(\hat{\theta})]^{-1}D_{21}(\hat{\theta})\alpha \qquad \textbf{[1.4.20]}$$

La expresión **[1.4.20]** puede combinarse con las expresiones **[1.4.13]**-**[1.4.14]** para obtener una simplificación para los MEB. Para estos modelos, la inconsistencia local de $\hat{\beta}$ debido a variables omitidas es:

$$\eta(\hat{\beta}) = [X^T\Omega X]^{-1}X^T\Omega Z\alpha \qquad \textbf{[1.4.21]}$$

donde X y Z son las matrices de datos cuya i-ésima fila es x_i^T y z_i^T respectivamente, y Ω es una matriz diagonal $n \times n$ cuyo elemento característico es:

$$\{\Omega_{ii}\} = \frac{f_i^2}{[F_i(1-F_i)]} \qquad \textbf{[1.4.22]}$$

De la expresión **[1.4.21]**, se deduce claramente que $\eta(\hat{\beta})$ es el estimador por *mínimos cuadrados ponderados de la regresión de* $Z\alpha$ *sobre* X, *con matriz de ponderación* Ω.

1.4.3. Intervalos de confianza para las probabilidades estimadas

Partiendo de los estadísticos expuestos en el **Apartado 1.4.1**, es sencillo derivar intervalos de confianza para parámetros individuales, así como regiones de confianza para el conjunto de parámetros del modelo. Sin embargo, otro aspecto interesante de la inferencia, que puede presentar alguna dificultad adicional, es el cálculo de intervalos de confianza para P_i. Este asunto se aborda a continuación.

Puesto que el estimador máximo-verosímil de β sigue una distribución asintótica normal multivariante con esperanza β y matriz de covarianzas $V(\hat{\beta})$, la variable $x^T\hat{\beta}$ sigue una distribución:

$$N(x^T\beta, x^TV(\hat{\beta})x) \qquad \textbf{[1.4.23]}$$

Bajo la especificación lineal de la parte sistemática del modelo que se ha seguido anteriormente, para un modelo general se tiene que $F^{-1}(P_i) = x^T\beta$, y de ahí:

$$\frac{x^T\hat{\beta} - F^{-1}(P_i)}{(x^T V(\hat{\beta})x)^{1/2}} \qquad [1.4.24]$$

es un estadístico pivote para P_i, que sigue una distribución normal estándar.

Definiendo η_α como el percentil de la distribución normal para un nivel de significación α, se puede escribir:

$$P\left[-\eta_{\alpha/2} \leq \frac{x^T\hat{\beta} - F^{-1}(P_i)}{(x^T V(\hat{\beta})x)^{1/2}} \leq \eta_{\alpha/2}\right] = 1 - \alpha \qquad [1.4.25]$$

y despejando:

$$P\left[x^T\hat{\beta} - \eta_{\alpha/2}(x^T V(\hat{\beta})x)^{1/2} \leq F^{-1}(P_i) \leq x^T\hat{\beta} + \eta_{\alpha/2}(x^T V(\hat{\beta})x)^{1/2}\right] = 1 - \alpha \qquad [1.4.26]$$

Teniendo en cuenta las propiedades de $F(\cdot)$ y, en particular, que es estrictamente creciente en su dominio, el intervalo de confianza a un nivel $1 - \alpha$ para P_i resulta:

$$\left\{F[x^T\hat{\beta} - \eta_{\alpha/2}(x^T V(\hat{\beta})x)^{1/2}],\ F[x^T\hat{\beta} + \eta_{\alpha/2}(x^T V(\hat{\beta})x)^{1/2}]\right\} \qquad [1.4.27]$$

1.5. Previsión con modelos de variable dependiente binaria

Una de las principales razones para la construcción de modelos de variable dependiente cualitativa es su utilización para la previsión de las decisiones agregadas de una población objeto de estudio. En esta sección se presentan los principales métodos para realizar previsión agregada. Un análisis más extenso de estas técnicas puede encontrarse en Ben-Akiva y Lerman (1985, cap. 6), Daganzo (1979, caps. 4 y 5) y Cramer (1991, cap. 5).

En esta sección, no se hace especial diferencia entre modelos binarios y multinomiales, debido a que las técnicas que se describen son de aplicación general. No obstante, cuando existan diferencias relevantes se harán notar de forma explícita.

1.5.1. El problema de la previsión agregada

Se supone que el número de decisores en la población, denotado por N, es conocido. Si se conoce el vector de características $\boldsymbol{x}_i$ para todos los individuos de la población, calcular una previsión del número de individuos que optarían por la alternativa j es, al menos conceptualmente, sencillo. Esta previsión, que se conoce como *demanda agregada*, sería:

$$N_j = \sum_{i=1}^{N} P_{ij} \qquad \textbf{[1.5.1]}$$

donde N_j es el número total de individuos que eligen la alternativa j, P_{ij} es la probabilidad de que el individuo i elija la alternativa j y P_j es la probabilidad de que un individuo genérico con vector de características $\boldsymbol{x}$ elija la alternativa j-ésima.

Nótese que N_j es el valor esperado del número de individuos que elegirían la alternativa j-ésima, y es un estimador insesgado y consistente del verdadero valor, esto es, de los individuos que efectivamente eligen j. Una forma más conveniente de expresar **[1.5.1]** es formularla en términos relativos, es decir, estimar la proporción de individuos que elegirían j. Esta proporción se denomina genéricamente *participación*:

$$w_j = \frac{1}{N}\sum_{i=1}^{N} P_{ij} = E_x(P_j) \qquad \textbf{[1.5.2]}$$

El problema del planteamiento anterior es que es poco realista, puesto que muy raramente se puede conocer el vector de características para toda la población y, aunque así fuera, el esfuerzo computacional para calcular previsiones sobre poblaciones grandes podría ser muy elevado. Partiendo de **[1.5.2]**, si se conoce la distribución de las características $\boldsymbol{x}_i$ en la población $g(\boldsymbol{x})$, puede escribirse que:

$$w_j = \int_{\boldsymbol{x}} P_j\, g(\boldsymbol{x})\, d\boldsymbol{x} = E_x(P_j) \qquad \textbf{[1.5.3]}$$

donde P_j es una función de $\boldsymbol{x}$. Generalmente, $g(\boldsymbol{x})$ es desconocida y, aunque no lo fuese, un tratamiento general de la expresión **[1.5.3]** podría ser sumamente complejo dependiendo de las formas concretas de las funciones dentro de la integral.

El propósito de los métodos de previsión agregada es, por tanto, reducir la cantidad de datos y cálculos necesarios para realizar la previsión objetivo.

Aunque hasta el momento sólo se ha expuesto el tratamiento general para realizar previsiones de la demanda total (N_j) o de la participación (w_j), la forma de tratar cualquier otra magnitud de interés asociada a la alternativa j que, en general, se puede denotar como $T_j(\boldsymbol{x},\boldsymbol{\beta})$, consiste en calcular o aproximar $E_x[T_j(\boldsymbol{x},\boldsymbol{\beta})]$. Bajo el supuesto de que se conoce el vector $\boldsymbol{x}$ para todos los individuos de la población, esto puede llevarse a cabo calculando:

$$E_x[T_j(\boldsymbol{x}, \boldsymbol{\beta})] \approx \frac{1}{N}\sum_{i=1}^{N} T_j(\boldsymbol{x}_i, \boldsymbol{\beta}) \qquad \textbf{[1.5.4]}$$

o, de forma más precisa:

$$E_x[T_j(\boldsymbol{x}, \boldsymbol{\beta})] = \int_{\boldsymbol{x}} T_j(\boldsymbol{x}, \boldsymbol{\beta})\, g(\boldsymbol{x})\, d\boldsymbol{x} \qquad \textbf{[1.5.5]}$$

y como puede observarse, el tratamiento de la proporción es un caso particular en el que:

$$T_j(x, \beta) = P_j \qquad [1.5.6]$$

Una medida usualmente empleada en la evaluación de los efectos que tienen cambios en una variable explicativa sobre la variable dependiente, es la elasticidad. Dicha medida tiene la ventaja de estar normalizada por la dimensión de las variables. La elasticidad de la probabilidad de elección de la opción j ante variaciones en la variable k es, en general:

$$\eta_{jk} = \frac{x_k}{P_j}\frac{\partial P_j}{\partial x_k} = x_k \frac{\partial \ln P_j}{\partial x_k} \qquad [1.5.7]$$

Para derivar casos particulares será necesario diferenciar entre el caso en que la variable esté asociada a la alternativa j o a otras alternativas (*elasticidad cruzada*). Para modelos binarios como los descritos en las secciones anteriores, las elasticidades resultan:

$$\begin{aligned} \eta_{1k} &= \beta_k x_k \frac{f(x^T\beta)}{F(x^T\beta)} \\ \eta_{0k} &= -\beta_k x_k \frac{f(x^T\beta)}{1 - F(x^T\beta)} \end{aligned} \qquad [1.5.8]$$

En los casos particulares de los modelos logit y probit, la expresión anterior puede escribirse:

$$\begin{aligned} \eta_{1k} &= \beta_k x_k (1 - P_1) \\ \eta_{0k} &= -\beta_k x_k P_1 \end{aligned} \qquad [1.5.9]$$

para el modelo logit, y:

$$\begin{aligned} \eta_{1k} &= \beta_k x_k \frac{\phi(x^T\beta)}{\Phi(x^T\beta)} \\ \eta_{0k} &= -\beta_k x_k \frac{\phi(x^T\beta)}{1 - \Phi(x^T\beta)} \end{aligned} \qquad [1.5.10]$$

para el modelo probit.

La elasticidad agregada de cada alternativa, esto es, la variación porcentual agregada de la probabilidad de elección de dicha alternativa ante un cambio en la variable

k-ésima, puede calcularse particularizando las expresiones **[1.5.4]** ó **[1.5.5]** en función de la información disponible, haciendo: $T_j(\boldsymbol{x}, \beta) = \eta_{jk}$.

1.5.2. Métodos de previsión agregada

Siguiendo a Ben-Akiva y Lerman (1985), los métodos de previsión agregada se fundamentan en la realización de hipótesis simplificadoras sobre el modelo de elección, la población, o ambos. Los métodos básicos que pueden considerarse son los siguientes:

- *Enumeración muestral*: Se utiliza una muestra representativa de la población y se extrapolan los resultados. La muestra no tiene por qué ser aleatoria, pudiendo emplearse muestras estratificadas endógenas o exógenas.

- *Clasificación por características*: Se divide la población en G subgrupos homogéneos y para la previsión se utiliza un *individuo medio* representativo de cada grupo. La previsión agregada se obtiene como la media ponderada de las obtenidas para cada segmento, utilizando como ponderaciones los pesos de cada estrato en la población.

- *Diferenciales estadísticas:* Se aproxima la distribución de las características en la población por sus momentos y se utilizan dichos momentos en la aproximación de las previsiones agregadas. Generalmente, sólo se consideran los primeros y segundos momentos respecto a la media de la distribución.

- *Integración explícita*: Se intenta evaluar de forma aproximada la expresión **[1.5.3]**. Este método requiere hacer algún supuesto sobre la distribución de las características individuales. Puesto que habitualmente $\boldsymbol{x}$ incluye variables continuas y discretas, esa integral debe tomarse en sentido amplio.

En lo que sigue, se presentan los dos primeros métodos, puesto que son los más utilizados y requieren menos hipótesis para su desarrollo. Se presta especial atención al estimador de la proporción de individuos que eligen cada alternativa (w_j), aunque es posible generalizarlo a cualquier otra medida de interés identificando: $T_j(\boldsymbol{x}_i, \hat{\beta}) = \hat{P}_{ij}$.

1.5.2.A. Método de enumeración muestral

La forma más simple de aplicar este método consiste en utilizar una muestra aleatoria representativa de la población objeto de estudio. La proporción de individuos que optan por la alternativa j puede estimarse mediante:

$$\hat{w}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{P}_{ij} \qquad \textbf{[1.5.13]}$$

donde n es el tamaño muestral.

Este método puede usarse también cuando la muestra es estratificada (endógena o exógenamente); esto es, cuando diferentes grupos de la población están representados en la muestra en proporciones distintas de las que poseen en la población. En este caso, el método se aplica primero utilizando la expresión **[1.5.13]** para los individuos de cada grupo y, posteriormente, se obtiene la estimación global calculando una media ponderada de las estimaciones intra-grupo:

$$\hat{w}_j = \sum_{k=1}^{K} \frac{N_k}{N}\left[\frac{1}{n_k}\sum_{i=1}^{n_k} \hat{P}_{ij}\right] \qquad \textbf{[1.5.14]}$$

donde N_k es el tamaño del grupo k en la población y n_k es el tamaño de ese mismo grupo en la muestra. Obviamente, se pueden obtener estimaciones de la participación de cada estrato utilizando el componente entre corchetes de la expresión **[1.5.14]**.

Este procedimiento es el más empleado en la práctica por su relativa sencillez y economía de cálculo, siendo además inmediata la obtención de previsiones para grupos de población. Las previsiones obtenidas son consistentes si las estimaciones de los parámetros son consistentes, aunque la varianza de la previsión estará sujeta a dos fuentes de error: el debido al muestreo y el asociado a la estimación. Puede demostrarse [Cramer (1991, cap. 5)] que la varianza del estimador **[1.5.13]** es:

$$V(\hat{w}_j) = \left[\frac{\partial \hat{w}_j}{\partial \beta}\right] V(\hat{\beta}) \left[\frac{\partial \hat{w}_j}{\partial \beta}\right]^T \qquad \textbf{[1.5.15]}$$

y para cualquier magnitud de interés, como la elasticidad, la varianza se obtiene de forma similar.

1.5.2.B. Método de clasificación por características

El método de clasificación por características se utiliza cuando no se dispone de una muestra representativa de la población, o cuando se intenta realizar previsión para un grupo de población escasamente representado (o no representado) en la muestra disponible. Como se verá a continuación, y pese a lo atractivo que resulta desde un punto de vista intuitivo, este método es inconsistente, aunque la inconsistencia puede reducirse a niveles poco significativos si se toman las precauciones adecuadas.

Este procedimiento es una extensión lógica del método del individuo medio, por lo que primero se expone éste, basado en construir un individuo *representativo* de la población o del grupo de población objeto de interés, denotado por $\bar{x}$, y obtener estimaciones poblacionales evaluando las magnitudes de interés para este individuo medio. La participación resulta:

$$\tilde{w}_j = P_j(\bar{x}) = F(\bar{x}^T\hat{\beta}_j) \qquad \textbf{[1.5.16]}$$

y en general:

$$E_x[T_j(x, \beta)] \approx T_j(\bar{x}, \hat{\beta}) \qquad \textbf{[1.5.17]}$$

La inconsistencia se debe a que, para una función no lineal, la media de la variable dependiente para un rango de valores de las independientes no es necesariamente igual a la función evaluada en la media de las variables independientes.

La inconsistencia es proporcional al rango de variación de x, por lo que es un procedimiento escasamente útil para tratar poblaciones en conjunto. Sin embargo, si la población se estratifica en grupos homogéneos y se agregan las previsiones de cada grupo, la inconsistencia disminuye y puede ser una buena aproximación. Esta es la idea sobre la que se basa el procedimiento de clasificación por características.

Formalmente el método de previsión por clasificación por características sigue las siguientes etapas:

Paso 1: Se particiona la población en G grupos excluyentes y exhaustivos, cada uno de los cuales corresponde a un rango de variación del vector de características definido por el conjunto $\{X_g\}$, cumpliendo que:

$$\bigcup_{g=1}^{G} X_g = X \qquad \textbf{[1.5.18]}$$

donde X denota el espacio de las variables explicativas. A veces puede ser necesario recurrir a métodos multivariantes, como el análisis de conglomerados, para determinar los conjuntos $\{X_g\}$, aunque también pueden ser fijados por el investigador.

Paso 2: Para cada grupo, obtener el individuo representativo $\bar{x}_g$ y el tamaño del grupo en la población que, en algunas ocasiones, puede ser necesario estimar.

Paso 3: El estimador de la proporción de individuos que eligen la alternativa j con este método resulta:

$$\tilde{w}_j = \sum_{g=1}^{G} \frac{N_g}{N} P_j(\bar{x}_g) \qquad \textbf{[1.5.19]}$$

donde N_g es el número de individuos en el g-ésimo grupo, N es el tamaño de la población y $P_j(\bar{x})$ es la probabilidad de que el *individuo medio* elija la alternativa j.

La clave de este procedimiento reside en la partición de la población en segmentos. Cuantos más estratos se definan, menor será la inconsistencia del estimador. Por otra parte, conviene poner de relieve que el criterio que debe considerarse a la hora de definir los grupos no es el de homogeneidad del vector $\boldsymbol{x}_i$, sino que la homogeneidad debe exigirse en la componente sistemática [Ben-Akiva y Lerman (1985, pag. 138)].

A la hora de aplicar este método, también hay que considerar que la clasificación utilizando todas las variables explicativas puede conducir a una infinidad de grupos, lo que impondría un coste de proceso innecesario. Con objeto de tener un número de grupos razonable, es recomendable realizar la agrupación basándola en las variables *principales*; esto es, variables que tengan el mayor peso en la probabilidad de elección individual (las derivadas de las probabilidades de elección o las elasticidades serían los indicadores a considerar) ignorando el resto. También conviene tener en cuenta que los grupos con escasa representación en la población contribuyen poco a *refinar* la estimación, por lo que se puede prescindir de ellos.

CAPÍTULO 2

OBSERVACIONES ANÓMALAS EN MODELOS DE ELECCIÓN BINARIA: PLANTEAMIENTO Y CONSECUENCIAS

2.1. Introducción

Citando a Hocking (1983), debe admitirse que "... el ajuste de ecuaciones a datos observados [frente a datos procedentes de experimentos cuidadosamente diseñados] es, en el mejor de los casos, un asunto peligroso ...". Inevitablemente, existe una distancia entre un modelo y la realidad; una cosa es especificar un modelo y otra que ese modelo represente adecuadamente los datos. En este contexto surge un conjunto de problemas en la interacción modelo-datos como, por ejemplo, errores numéricos, muestreo inadecuado, cifras erróneas o defectos de codificación o, incluso, que el modelo mismo sea una mala aproximación a los hechos que pretende explicar:

- Los *errores numéricos* aparecen en cualquier análisis, debido a la necesidad de trabajar con representaciones numéricas de precisión finita. De todos modos, este tipo de error no provoca problemas serios si se toman las debidas precauciones, muy especialmente en los algoritmos numéricos que se emplean para realizar la tareas de cálculo, desde la inversión de una matriz a los métodos de estimación no lineal.

- Respecto a la cuestión del *diseño muestral*, lo deseable, como sugiere Snee (1983), sería contar siempre con muestras que representen adecuadamente la población en estudio, basadas en un diseño experimental previo. Desafortunadamente, en economía, casi siempre se dispone de una muestra *dada* en cuyo diseño no ha participado el investigador, por lo que los resultados estarán siempre *condicionados* a la muestra disponible.

- Los *errores en las cifras* surgen más a menudo con algunos tipos de datos que con otros. Un modelo de series temporales que utiliza datos de Contabilidad Nacional y un número moderado de observaciones, es difícil que contenga esta clase de errores si la muestra se ha revisado con cierta atención. Sin embargo, una muestra de corte transversal de gran tamaño, obtenida enviando cuestionarios a individuos, puede tener bastantes errores: algunos encuestados interpretarán mal ciertas preguntas, otros darán información incorrecta de forma deliberada, habrá errores de transcripción a soporte magnético, etc.

- La *especificación del modelo* es también un problema sustancial y, potencialmente, una nueva fuente de errores para el análisis. El supuesto de linealidad en los parámetros, el conjunto de las variables explicativas o la forma funcional elegida pueden no ser adecuados. En este aspecto, sólo un profundo conocimiento del problema que se desea tratar y el uso, en la medida de lo posible, de herramientas de diagnosis pueden ayudar a modelizar correctamente.

En este trabajo se considera un problema concreto en la interacción modelo-datos: la existencia de observaciones anómalas, que resultan frecuentes en las muestras de corte transversal. Al intentar describir el comportamiento de la muestra mediante un modelo, puede haber un conjunto reducido de observaciones que, debido a su falta de homogeneidad con el resto de la muestra, distorsionen sustancialmente los resultados de la estimación, incluso si se utilizan muestras de gran tamaño. En este trabajo, se supone que dichas observaciones *no se deben a errores en los datos*, sino a que en la muestra hay un grupo de datos que proceden de una población diferente que el resto. Por tanto, en lo que sigue, el análisis se centra en un tipo muy concreto de errores: los que tienen su origen en el hecho de que entre los datos se encuentra un conjunto (usualmente pequeño) de observaciones generadas por un proceso estocástico distinto del que sigue la mayoría de la muestra.

Un ejemplo sencillo del problema que origina la presencia de observaciones anómalas, puede plantearse del siguiente modo [Krasker et al. (1983)]: mediante una encuesta, se obtiene una muestra de tamaño n de cierta población de individuos, con el objeto de estimar la esperanza de alguna característica de dicha población, que sigue una distribución con esperanza μ y desviación típica σ. Sin embargo, con una probabilidad ω aparecen observaciones procedentes de una población distinta, con esperanza $\mu+\delta$ y desviación típica $k\sigma$. En estas circunstancias, el error cuadrático medio de la media muestral $\bar{x}_n$ es: $[(1-\omega-\omega k^2)+\delta^2\omega(1-\omega)]\sigma^2/n$. Sin pérdida de generalidad, si se supone que $\sigma=1$, $\delta=1$, $k=2$ y $\omega=0.05$, el error cuadrático medio es $0.0025 + 1.20/n$ por

lo que, en este caso, no hay una gran ventaja en usar muestras mayores de 1000 observaciones y los recursos se podrían destinar a mejorar la calidad de los datos.

En este capítulo se aborda el problema de las anomalías en los MEB. Este tema ha sido tratado con anterioridad: Pregibon (1981), Jennings (1986) y Copas (1988) son algunas referencias. Sin embargo, estos trabajos no parten de una definición estadística de dato anómalo, ni analizan las consecuencias que este tipo de observaciones tienen sobre los resultados de estimación del modelo. En las páginas siguientes se abordan estos aspectos y se demuestra que en los MEB la presencia de observaciones procedentes de una población diferente que las restantes afecta a la consistencia del estimador MV.

El capítulo está organizado como sigue. En la **Sección 2.2** se considera el caso del *modelo lineal general* (MLG), donde se analizan los tipos de observaciones anómalas que pueden aparecer y se lleva a cabo una breve revisión de los métodos utilizados para su tratamiento. El objetivo de esta sección es presentar el problema de las anomalías en el MLG, como punto de partida para abordar el mismo problema en los MEB.

A continuación, en la **Sección 2.3**, se plantea el problema de observaciones anómalas en los modelos de elección binaria y se analizan las consecuencias que puede tener su presencia en el estimador de máxima verosimilitud del modelo. En concreto, se muestra que ante la presencia de anomalías, el estimador de MV es inconsistente.

Por último, en la **Sección 2.4** se ilustran numéricamente los resultados generales de la **Sección 2.3** usando datos simulados. Para ello, se consideran los dos modelos que se emplean más frecuentemente en la práctica: el modelo probit y el modelo logit.

2.2. El problema de observaciones anómalas en el modelo lineal general

En esta sección se define, en primer lugar, lo que se entiende por observación anómala. Seguidamente, se ilustran de forma muy general los problemas que se plantean en el modelo lineal general (MLG) cuando aparecen observaciones anómalas en la muestra y, por último, se presenta un breve resumen de los métodos para su tratamiento.

2.2.1. Observaciones anómalas en el modelo lineal general

Siguiendo a Box y Tiao (1968), una observación anómala puede definirse como aquella que *no se ha generado por el mismo experimento aleatorio que las restantes observaciones muestrales*. Por tanto, la diferencia en el proceso generador de los datos, suponiendo que la hipótesis sobre la forma funcional de su distribución sea la correcta, puede tener básicamente dos fuentes de procedencia: i) distintas varianzas en el término de perturbación (y por lo tanto, en la variable dependiente) y ii) distintas esperanzas de la variable dependiente, aunque no en el término de perturbación. Por supuesto, se puede contemplar el caso en el que la distribución de las observaciones anómalas es diferente, tanto en sus primeros momentos como en la forma funcional, pero esto complicaría innecesariamente el análisis.

Partiendo de la definición anterior, en un modelo de regresión lineal pueden existir dos tipos de anomalías[7], tal y como se ilustra en la **Figura 2.1** [Peña y Ruiz-Castillo (1982 y 1984)] para el caso de una variable explicativa. En la regresión de z_i sobre x_i, el punto A puede considerarse un dato anómalo, ya que corresponde a un valor de z_i muy alejado de la media de las restantes observaciones muestrales y por tanto, es poco probable que su presencia se deba al mismo mecanismo que ha generado los restantes datos. Tal y como se indica en la figura, la presencia de un punto como éste desplazaría hacia arriba la recta estimada por MCO y su residuo sería 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 alejados de sus valores medios. Sin embargo, aunque este punto afectaría gravemente a la pendiente y a la constante de la recta estimada, su residuo

[7] Estamos ignorando conscientemente la posibilidad de que existan otros errores en los datos.

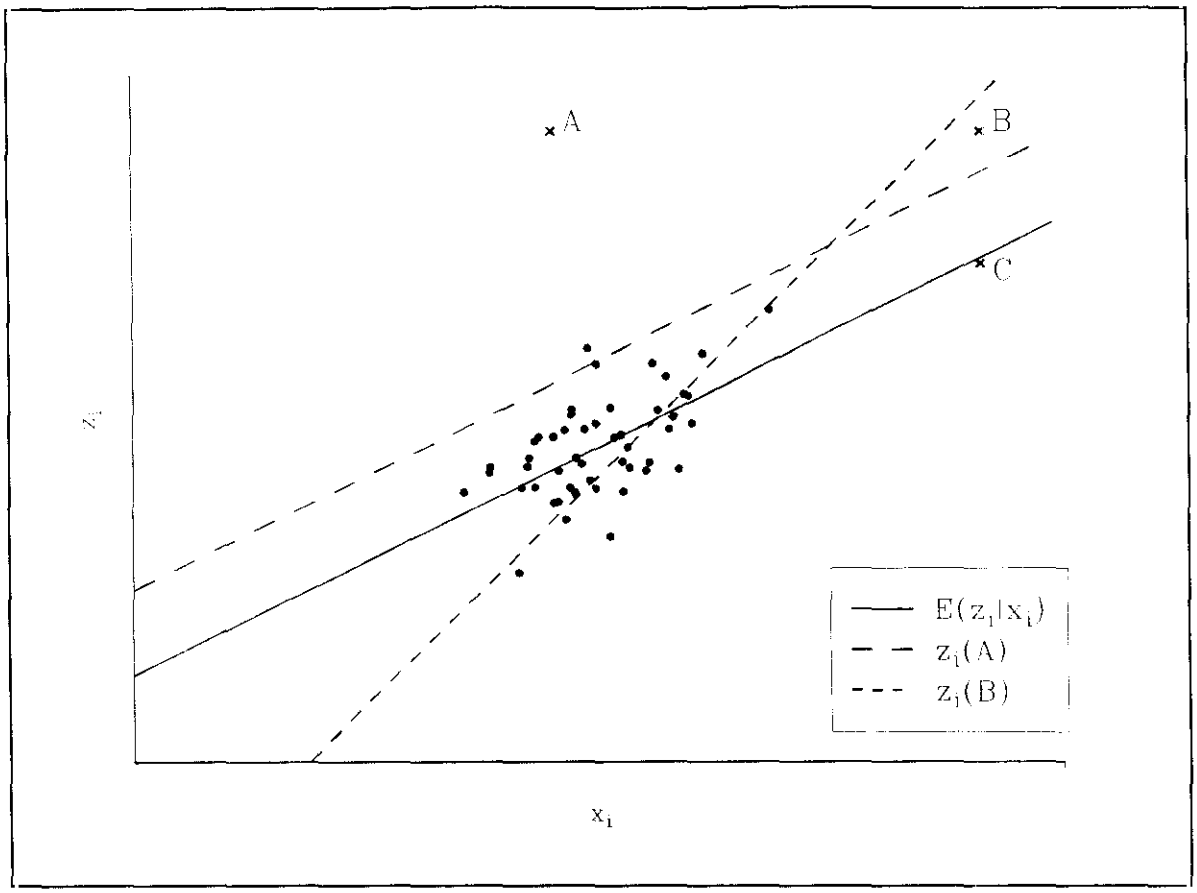

Figura 2.1: Dos ejemplos clásicos de anomalías.

sería muy pequeño en valor absoluto (mucho menor que el de otras observaciones no anómalas). Por tanto, 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 este sentido, la utilización de estadísticos que midan el peso de cada observación o grupo de observaciones sobre los coeficientes estimados [como, por ejemplo, el propuesto por Cook (1977)], es también un elemento fundamental en la detección de anomalías.

Ahora se considera el punto C. Este caso parece semejante al punto B, pero al contrario que en éste, es muy probable que la observación C haya sido generada por el mecanismo que relaciona z_i con x_i, aunque para un valor extremo en el espacio de las X, sobre cuya distribución no se hace ningún supuesto. Esta clase de puntos no pueden ser considerados anómalos sobre la base de la definición que estamos empleando, puesto que no se puede descartar que correspondan al mecanismo aleatorio implícito, aunque tienen un elevado peso en la estimación por MCO del modelo. De hecho, esta clase de puntos puede contener información muy relevante para la estimación, pero también puede hacer que la varianza de los parámetros estimada sea significativamente inferior a la que se habría obtenido sin la presencia de los mismos.

En la **Figura 2.2** se presentan algunas configuraciones de interés que pueden aparecer en los datos. El caso (A) es una situación extrema en la que un sólo punto, como el C, determina completamente la recta estimada. En este caso, la muestra es poco informativa, pero un punto alejado del resto puede hacer que los coeficientes estimados sean significativos cuando, en realidad, no existe una relación.

El caso (B) representa una situación en la que debido a la presencia de dos observaciones anómalas que se compensan, la recta estimada no sufriría variación. En este caso se observarían dos residuos grandes y de signo contrario. Al utilizar estadísticos que evalúan el efecto de cada observación sobre los coeficientes estimados [como por ejemplo el de Cook (1977)], ambas observaciones resultarían anómalas, puesto que la eliminación de una de ellas nos llevaría a la presencia de una anomalía como la del punto A de la **Figura 2.1**. Sin embargo, conjuntamente ambas anomalías no presentan problemas sobre los coeficientes estimados, aunque tendrán un efecto importante sobre la varianza estimada de dichos coeficientes.

En (C) y (D) se ilustran situaciones donde la detección de las anomalías es considerablemente más complicada que en los casos anteriores. En ambos casos, los residuos asociados a los puntos B y B' son pequeños en valor absoluto pero, además, la eliminación de uno sólo de dichos puntos no provoca un cambio significativo de la recta estimada, por lo que no se detectarían como anómalos a partir de un estadístico que evalúe el efecto de cada observación por separado. En estos casos, se dice que aparece un problema de *enmascaramiento*.

Para terminar este apartado, es importante señalar que en la literatura no existe una terminología de uso generalizado para referirse al tipo de problema que nos ocupa, por lo que es conveniente aclarar algunas cuestiones de léxico. Como se ha indicado anteriormente, en este trabajo se entiende por *observación anómala* o *atípica* aquella que ha sido generada por un proceso estocástico distinto del que se supone para la mayor parte de la muestra, mientras que se utiliza el término de *observación influyente* para designar a toda observación que tenga un *efecto significativo* sobre la estimación del modelo.

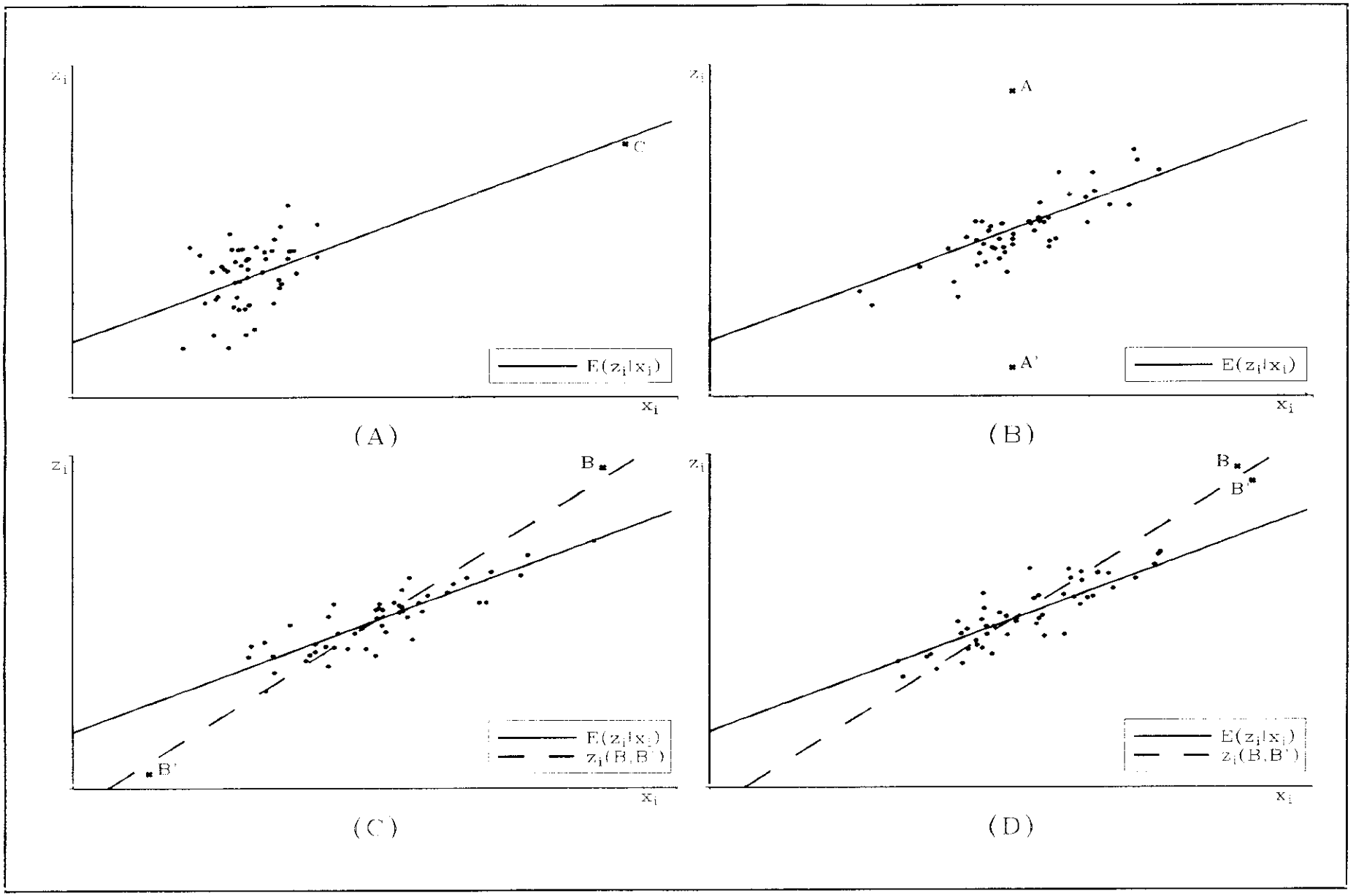

Figura 2.2: Ejemplos de configuraciones de datos problemáticas.

De acuerdo con estos criterios, los puntos A y B de la **Figura 2.1** son anómalos y además influyentes; por el contrario, el punto C de la misma figura es influyente, puesto que puede tener un efecto importante en la matriz de covarianzas de los coeficientes estimados (aunque no en el valor de dichos coeficientes), pero no puede considerarse anómalo. A puntos como el C se les denomina *observaciones extremas*. De hecho, una observación extrema es una anomalía en el espacio de las variables explicativas (ese valor del vector x_i es poco probable), mientras que lo que denominamos observación anómala lo es respecto a la distribución del término de perturbación o de la variable dependiente.

2.2.2. Métodos de tratamiento

En la literatura econométrica se ha estudiado ampliamente el problema del tratamiento de observaciones anómalas para los modelos lineales de regresión estática [Belsley et al. (1980), Box y Tiao (1968), Cook (1977), Krasker et al. (1983), Weisberg (1983) son algunas referencias]. Principalmente los desarrollos pueden agruparse en cuatro clases:

- *Diagnosis y detección* [Belsley et al. (1980), Cook (1977)]. En esta línea se trata de desarrollar estadísticos que ayuden a decidir si los supuestos básicos del modelo son aceptables y, por tanto, permiten cuestionar la homogeneidad de la muestra. Para el caso de observaciones anómalas, el método más extendido y usado desde el nacimiento de las técnicas de regresión es el análisis de residuos. No obstante, en los últimos años se ha producido un considerable aumento de estadísticos para la detección de observaciones influyentes en modelos lineales.

  Básicamente, hay dos tipos de medidas para caracterizar las observaciones influyentes: i) las de distancia de las variables explicativas x_i al centro de gravedad del espacio de las X (valores extremos) y ii) medidas del efecto que tiene una observación, o un grupo de ellas, sobre los aspectos más relevantes del modelo que, generalmente, son los parámetros estimados y/o las previsiones.

- *Análisis de influencia* [Cook y Weisberg (1982)]. La idea general del análisis de influencia consiste en estudiar los cambios en el modelo estimado, o en otros aspectos del análisis, cuando se introduce una perturbación en algunos de los elementos que componen el modelo (variables explicativas, término de error, variable explicada). Mientras que los diagnósticos se usan para encontrar problemas con un modelo y unos datos dados, el análisis de influencia se basa

en suponer que el modelo es correcto y estudiar la sensibilidad de un estimador particular bajo un esquema de perturbación dado.

Básicamente, el análisis parte de un modelo sencillo como, por ejemplo:

$$y = X\beta + \delta z + \varepsilon \qquad [2.2.1]$$

donde z es un vector unitario. A partir de este planteamiento, se derivan expresiones analíticas que describen el comportamiento de los diferentes componentes del modelo (estimaciones, previsiones, etc.) cuando δ es distinto de cero.

- *Estimación robusta y de influencia acotada* [Huber (1981), Krasker et al. (1983)]. El desarrollo de métodos robustos de estimación se basa en ponderar las observaciones en proporción inversa a su peso en la estimación del modelo base. No se hace ningún supuesto sobre la procedencia de las anomalías, pero se intenta limitar su efecto en la estimación.

  Los métodos de influencia acotada son semejantes a los robustos, aunque basan sus ponderaciones en medidas del efecto de cada observación sobre la estimación, como por ejemplo el estadístico de Cook [Cook (1977)]. La diferencia entre estos métodos frente a los métodos robustos se basa en que limitan el efecto de cada observación sobre un aspecto bien definido del problema, mientras que los métodos robustos generalmente basan las funciones de ponderación en el tamaño de los residuos.

- *Transformaciones de los datos* [Atkinson (1982)]. El principio general de esta metodología consiste en transformar las variables de modo que el modelo resultante cumpla las hipótesis de partida. Un ejemplo puede plantearse del siguiente modo [Box y Tiao (1968)]: sea un modelo de regresión lineal donde las perturbaciones, debido a la presencia de anomalías, no siguen una distribución normal. En este caso, se podría buscar una transformación de la familia Box-Cox de modo que, una vez transformada la variable dependiente y/o las explicativas, se pueda mantener el supuesto de normalidad en la perturbación. No obstante, este planteamiento no está exento de problemas: i) puede ocurrir que una sola observación anómala dicte la transformación y ii) debido a un planteamiento erróneo del modelo, al imponer alguna transformación *a priori* surgen observaciones aparentemente anómalas.

Muy ligada al primer grupo (métodos de diagnosis y detección), aparece la idea de *robustecer* la metodología de estimación [Box (1980), Peña y Ruiz-Castillo (1982 y 1984)], que se basa en analizar detalladamente los datos disponibles, tanto *a priori* como una vez llevado a cabo el proceso de estimación por algún método convencional apropiado. Esta metodología se basa en el uso de estadísticos de diagnóstico e influencia y, posteriormente, en tomar de decisiones sobre qué hacer con cada observación o grupo de observaciones para las que se ha comprobado algún efecto serio sobre los resultados del modelo. Metodológicamente, esta idea resulta atractiva, ya que no implica el mecanicismo asociado a otros planteamientos. En esta línea se desarrollan los resultados de este trabajo.

2.3. Anomalías en modelos de elección binaria

El problema de la detección y el análisis de las consecuencias de las observaciones anómalas e influyentes en los modelos de elección discreta ha recibido menos atención en la literatura que en los modelos lineales. Los resultados existentes son generalizaciones de los obtenidos para el modelo lineal de regresión. Para el caso de los MEB, el problema de la detección de observaciones influyentes se trata por primera vez en Pregibon (1981), que plantea estadísticos generales de diagnosis en modelos logit. Posteriormente, Jennings (1986) cuestiona algunos aspectos del trabajo de Pregibon y amplia algunos resultados, Copas (1988) trata el problema de anomalías debidas a errores de codificación de los datos y Bedrick y Hill (1990) enfocan el problema desde el punto de vista del análisis de influencia; en la misma línea, en Lesaffre y Albert (1989) se estudia el caso de los modelos de elección múltiple. Por otra parte, Cook y Weisberg (1980) generalizan algunos resultados de Cook (1977) para los modelos lineales generalizados (GLM), Williams (1987) también presenta resultados sobre diagnosis para los GLM, Green (1984) desarrolla alternativas de estimación lineales y de estimación robusta y resistente para el caso de los GLM, alguno de cuyos casos particulares incluye modelos de elección binaria. Por último, Aranda-Ordaz (1981) y Guerrero y Johnson (1982) tratan la transformación de variables para modelos con datos binarios agrupados.

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 MEB los procedimientos para la detección de anomalías utilizados en los modelos lineales que, en gran medida, se basan en el análisis de residuos y en evaluar el efecto de cada observación en la estimación de los parámetros del modelo. Sin embargo, no tienen en cuenta las particularidades de los MEB, que hacen que dichos métodos no sean directamente aplicables.

En este trabajo se trata el problema de forma diferente, partiendo de la definición de observación anómala que habitualmente se utiliza en la literatura econométrica y que se ha introducido anteriormente: *una observación anómala es aquella que no se ha generado por el mismo modelo estocástico que se supone para las restantes observaciones muestrales* [Box y Tiao (1968)]. A partir de esta definición, se demuestra que, en los modelos de elección binaria, la existencia de anomalías en la muestra afecta a la consistencia del estimador de máxima verosimilitud. Ello se debe a que la presencia de

estas observaciones hace que la función de verosimilitud del modelo sea diferente de la habitual.

2.3.1. Planteamiento del problema

Consideremos la ecuación **[1.2.7]** utilizada para derivar un MEB. 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 σ_0^2 y se cumplen las restantes hipótesis habituales del modelo. En particular, la ecuación **[1.2.7]** establece que las variables y_i^* se han generado por el mismo modelo estocástico; esto es, las y_i^* se distribuyen independientemente, con $E(y_i^*) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ y $V(y_i^*) = \sigma_0^2$. De hecho, puede suponerse que y_i^* sigue la misma distribución que ε_i. Entonces, teniendo en cuenta la definición de observación anómala que se acaba de formular, un valor de y_i^* será anómalo si no se ha generado por **[1.2.7]**. Obsérvese que el hecho de que y_i^* sea una variable latente no significa que teóricamente no pueda presentar comportamientos anómalos[8].

Según esto, pueden considerarse dos tipos de anomalías en la variable y_i^*: aquellas generadas por una distribución con distinta varianza que las restantes observaciones muestrales y aquellas generadas por una distribución con distinta media. A continuación se estudian ambos casos.

2.3.1.A. Anomalías generadas por el lado de la varianza

La primer forma de modelizar la presencia de observaciones anómalas en el modelo **[1.2.7]** es suponer que, aunque las perturbaciones ε_i se distribuyen i.i.d. $F(\cdot)$, existe una pequeña proporción desconocida ω de perturbaciones que siguen la misma distribución, con esperanza nula y varianza $\sigma_0^2h^2$, 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 ε_i en **[1.2.7]** provienen o de una distribución $F(\cdot \mid \sigma_0^2)$ o de una $F(\cdot \mid \sigma_0^2h^2)$ con proporciones $(1-\omega)$ y ω respectivamente. En Box y Tiao (1968) se demuestra que, bajo estas condiciones y para el caso en que $F(\cdot)$ es la distribución normal, las perturbaciones en **[1.2.7]** pueden considerarse i.i.d. con una función de distribución que es una combinación lineal de dos

[8] 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.

funciones de distribución normales y que depende de ω y h. En este trabajo, se extiende ese planteamiento al caso genérico, de forma que se supone que la distribución de las perturbaciones es:

$$G(\varepsilon_i) = (1-\omega)F(\varepsilon_i \mid 0, \sigma_0^2) + \omega F(\varepsilon_i \mid 0, \sigma_0^2 h^2) \qquad \textbf{[2.3.1]}$$

donde $F(\cdot)$ denota la función de distribución con varianza normalizada conocida σ_0^2 y $F_h(\cdot)$ la función de distribución con varianza $\sigma_0^2 h^2$. Esto es, si $\omega = 0$ se tiene que $G(\varepsilon_i) = F(\varepsilon_i \mid 0, \sigma_0^2)$, por lo que se mantiene la hipótesis sobre la distribución de ε_i. Pero ante la presencia de un porcentaje ω de observaciones anómalas, la distribución de ε_i es la indicada en **[2.3.1]**. Además en este caso, para todo i:

$$\begin{aligned} E(\varepsilon_i) &= 0 \\ V(\varepsilon_i) &= \sigma_0^2 [1 + \omega(h^2 - 1)] \end{aligned} \qquad \textbf{[2.3.2]}$$

por lo que la varianza de la distribución $G(\varepsilon_i)$ es mayor que σ_0^2.

La implicación fundamental de que las perturbaciones en el modelo **[1.2.7]** se distribuyan como en **[2.3.1]** es que se produce un cambio en la forma funcional que determina P_i, por lo que, a partir de **[2.3.1]**:

$$P_i = G_i = (1-\omega) F_i + \omega F_{hi} = F_i + \omega (F_{hi} - F_i) \qquad \textbf{[2.3.3]}$$

donde F_i y F_{hi} denotan $F(\boldsymbol{x}_i^T\boldsymbol{\beta})$ y $F_h(\boldsymbol{x}_i^T\boldsymbol{\beta}) = F(\boldsymbol{x}_i^T\boldsymbol{\beta}/h)$, respectivamente.

Obsérvese que, si $\omega = 0$, entonces $P_i = F_i$, que es el MEB de la ecuación **[1.2.8]**. Sin embargo, ante la presencia de este tipo de anomalías, la especificación correcta en la determinación de P_i viene dada por la ecuación **[2.3.3]**, que establece que $P_i = G_i$, donde G_i es igual a F_i más un término adicional cuya magnitud depende de h y ω.

En la **Figura 2.3** se representan las funciones F_i y G_i para el caso particular en que $F(\cdot)$ es la distribución normal estándar ($\Phi(\cdot)$), $h^2 = 7$ y $\omega = 0.15$. Dado que F_{hi} es una función de distribución normal con media nula y varianza mayor que la unidad, esta función se encontrará por encima de F_i para valores de $\boldsymbol{x}_i^T\boldsymbol{\beta} < 0$ y por debajo de F_i para valores de $\boldsymbol{x}_i^T\boldsymbol{\beta} > 0$, mientras que $G(\cdot)$, al ser una combinación lineal convexa de ambas, se encuentra entre $F(\cdot)$ y $F_h(\cdot)$.

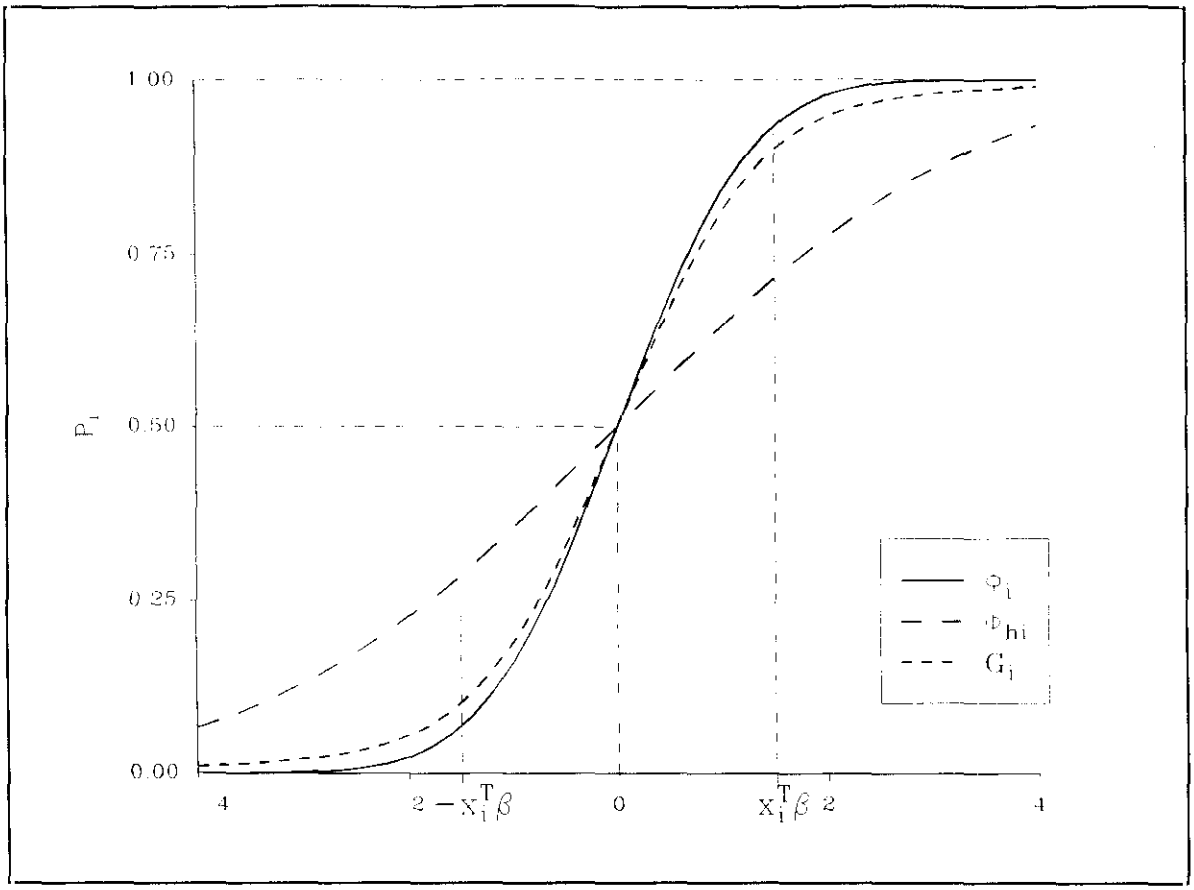

Figura 2.3: Representación de las funciones Φ_i, Φ_{hi} y G_i.

Por tanto, se tiene que:

$$\text{si} \begin{cases} x_i^T\beta < 0 \Rightarrow \Phi_{hi} > \Phi_i \Rightarrow G_i > \Phi_i \quad \text{para} \quad P_i < 1/2 \\ x_i^T\beta = 0 \Rightarrow \Phi_{hi} = \Phi_i \Rightarrow G_i = \Phi_i \quad \text{para} \quad P_i = 1/2 \\ x_i^T\beta > 0 \Rightarrow \Phi_{hi} < \Phi_i \Rightarrow G_i < \Phi_i \quad \text{para} \quad P_i > 1/2 \end{cases} \qquad \textbf{[2.3.4]}$$

Luego G_i tendrá la forma que se indica en la **Figura 2.3** y la discrepancia entre F_i y G_i dependerá del valor de los parámetros h y ω. Para el caso general, a partir de **[2.3.3]** se tiene:

$$\frac{\partial(G_i-F_i)}{\partial\omega} = F_{hi}-F_i$$
$$\frac{\partial(G_i-F_i)}{\partial h} = \omega\frac{\partial F_{hi}}{\partial h} = \omega\frac{\partial F(x_i^T\beta/h)}{\partial h} = \omega f\left[\frac{x_i^T\beta}{h}\right]\left[-\frac{x_i^T\beta}{h^2}\right] \qquad \textbf{[2.3.5]}$$

por lo que ambas derivadas son positivas cuando $x_i^T\beta < 0$ y negativas para valores $x_i^T\beta > 0$. Por lo tanto, cuanto mayor sea el valor de ω y/o h, mayor será, en términos absolutos, la diferencia entre G_i y F_i.

Según estos resultados, el logaritmo de la función de verosimilitud correspondiente al modelo **[2.3.3]** es:

$$\ell = \sum_{i=1}^{n}\{y_i\ln[F_i+\omega(F_{hi}-F_i)] + (1-y_i)\ln[1-F_i-\omega(F_{hi}-F_i)]\} \qquad \textbf{[2.3.6]}$$

de forma que sólo si $\omega = 0$, **[2.3.6]** coincide con el logaritmo de la función de verosimilitud del modelo **[1.3.2]**. En cambio, si existe un porcentaje ω de observaciones con varianza mayor que σ_0^2, la expresión **[2.3.6]** es la función objetivo que debería maximizarse para obtener las estimaciones MV del vector β. El problema es que esta función depende de los parámetros ω y h que, por lo general, son desconocidos.

Por otra parte, la maximización de **[2.3.6]** por los procedimientos expuestos en la **Sección 1.3** presenta considerables problemas. En concreto, si $F(\cdot)$ es la distribución normal, la función de verosimilitud no está acotada [Day (1969), Quandt y Ramsey (1978)]. El desarrollo de un procedimiento de estimación adecuado para este caso es algo que no se plantea en este trabajo, aunque el camino a seguir sería la utilización de alguna variante especializada del *algoritmo EM*, que se describe en el **Apéndice A.2**. No obstante, el objetivo de la ecuación **[2.3.6]** es mostrar el tipo de error de especificación que se comete si se ignora la presencia de observaciones anómalas en un MEB y se utiliza F_i en lugar de G_i para calcular la verosimilitud de cada observación.

La consecuencia inmediata de este error de especificación en la función de verosimilitud, es que el vector β y las probabilidades P_i se estimarán inconsistentemente. Este sesgo asintótico no puede evaluarse de forma general ya que depende de ω y h, aunque sí es posible evaluar la *inconsistencia local* como se desarrolla en el **Apartado 2.3.2**. Sin embargo, como indican las ecuaciones **[2.3.5]**, es claro que 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 dichas observaciones.

La inconsistencia en la estimación de los parámetros como consecuencia de la presencia de este tipo de anomalías demuestra una diferencia sustancial de los modelos de elección binaria con respecto al modelo lineal general donde, a pesar de la presencia de anomalías generadas por el lado de la varianza, los estimadores por MCO siguen siendo insesgados y consistentes, aunque no eficientes [véase, por ejemplo, Peña y Ruíz-Castillo (1982) y (1984)].

Es importante señalar que este resultado se basa en que, si existen anomalías en la muestra, la función de distribución de ε_i viene dada por la expresión **[2.3.1]**. Esta expresión es una forma de modelizar la presencia de observaciones anómalas, bajo el supuesto de que todas ellas provienen de la misma distribución con media nula y varianza mayor que σ_0^2. 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 un conjunto de distribuciones $F(\cdot)$ con distintas varianzas, todas ellas mayores que σ_0^2. En lo sucesivo, se mantiene el supuesto simplificador, aunque no restrictivo, de que sólo hay dos grupos de observaciones en la muestra.

2.3.1.B. Anomalías generadas por el lado de la media

Un segundo tipo de observaciones anómalas puede 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^* siguen una distribución $F(\cdot)$, existe un porcentaje desconocido ω de estas variables que, aunque también se distribuye $F(\cdot)$, su esperanza es $E(y_i^*) = \boldsymbol{x}_i^T\boldsymbol{\gamma}$, donde $\boldsymbol{\gamma} \neq \boldsymbol{\beta}$. Obviamente, si la proporción ω es grande, se podría decir que existe un cambio estructural; esto es, que la población a analizar se compone de dos grupos de individuos distintos entre sí. Sin embargo, el caso que se considera en este trabajo es cuando ω es pequeño y sólo se trata de unos pocos individuos atípicos. En estas circunstancias, la función de distribución de y_i^* vendrá dada por:

$$H(y_i^*) = (1-\omega)\,F(y_i^* \mid \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_0^2) + \omega\,F(y_i^* \mid \boldsymbol{x}_i^T\boldsymbol{\gamma}, \sigma_0^2) \qquad \textbf{[2.3.7]}$$

por lo que es inmediato que:

$$P_i = P[y_i^* \geq 0] = F(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \omega\,[F(\boldsymbol{x}_i^T\boldsymbol{\gamma}) - F(\boldsymbol{x}_i^T\boldsymbol{\beta})] \equiv H_i \qquad \textbf{[2.3.8]}$$

De manera similar al caso anterior, ignorar el segundo término del lado derecho de la ecuación **[2.3.8]**, conduce a un error de especificación en la determinación de P_i. De nuevo, si se ignora este término y se utiliza F_i en lugar de H_i para calcular la verosimilitud de cada observación, los parámetros del modelo se estimarán de forma inconsistente. Obsérvese que, según **[2.3.8]**, para un $\boldsymbol{x}_i$ dado, el que la función H_i esté por encima o por debajo de F_i, dependerá del valor de los coeficientes en el vector $\boldsymbol{\gamma}$. En cualquier caso, la discrepancia entre ambas funciones será mayor cuanto mayor sea ω y/o la diferencia entre los componentes de $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$. Conviene tener presente que dicha diferencia no será tanto en magnitud (puesto que ambos están normalizados y operan sobre las mismas variables), sino en el ángulo que forman.

De igual forma que en el supuesto de diferentes varianzas, este análisis puede extenderse fácilmente al caso de que las anomalías se consideren generadas por un conjunto de distribuciones $F(\cdot)$ con distintas esperanzas. Este supuesto más general conduciría a errores de especificación en la función de verosimilitud del mismo tipo que los vistos hasta ahora.

2.3.2. Inconsistencia del estimador máximo-verosímil

Para demostrar la inconsistencia del EMV ante la presencia de observaciones anómalas de acuerdo con los esquemas planteados en el apartado anterior, se utiliza la expresión de inconsistencia local de la ecuación **[1.4.21]**. Para ello, previamente hay que parametrizar la presencia de anomalías siguiendo el esquema general de errores de especificación presentado en el **Apartado 1.4.2**.

Para el caso de anomalías por el lado de la varianza, se puede seguir un desarrollo similar al que se presenta en Godfrey (1988) para un esquema general de heterocedasticidad, donde:

$$V(\varepsilon_i) = \sigma_i^2 = \sigma_0^2 h_i^2, \qquad i = 1, 2, \ldots, n \qquad \textbf{[2.3.9]}$$

de forma que, imponiendo la restricción de normalización de **[1.2.6]**, resulta:

$$P_i = P\left[\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_i} < \varepsilon_i\right] \qquad \textbf{[2.3.10]}$$

donde h_i es desconocido. Sin embargo, se puede suponer que:

$$h_i = h(\boldsymbol{z}_i^T\boldsymbol{\alpha}) \qquad \text{tal que} \quad h(0) = 1 \quad \text{y} \quad h'(0) \neq 0 \qquad \textbf{[2.3.11]}$$

y, además, puede suponerse que: $h(\cdot) = \exp(\cdot)$ [Davidson y MacKinnon (1984)]. Bajo estas circunstancias se plantea la hipótesis nula H_0: $\boldsymbol{\alpha} = \mathbf{0}_q$, lo que hace posible evaluar la inconsistencia local del EMV de $\boldsymbol{\beta}$ para *pequeños* alejamientos de la hipótesis nula.

Manteniendo la notación de la **Sección 1.4**, a partir de **[2.3.10]** se tiene que: $v_i(\boldsymbol{\theta}) = \boldsymbol{x}_i^T\boldsymbol{\beta}/h(\boldsymbol{z}_i^T\boldsymbol{\alpha})$, con lo que el problema queda reducido a uno de omisión de las variables $\boldsymbol{z}_i$. Realizando una aproximación de Taylor de primer orden y suponiendo que los componentes de $\boldsymbol{\alpha}$ están próximos a cero, resulta:

$$v_i(\boldsymbol{\theta}) = \frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h(\boldsymbol{z}_i^T\boldsymbol{\alpha})} \approx \boldsymbol{x}_i^T\boldsymbol{\beta} - h'(0)\,(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{z}_i^T\boldsymbol{\alpha} \qquad \textbf{[2.3.12]}$$

donde se ha tenido en cuenta que $h(0) = 1$. De **[2.3.12]** se puede eliminar el término $h'(0)$ [Godfrey (1988)] puesto que es irrelevante para contrastar H_0: $\boldsymbol{\alpha} = \mathbf{0}_q$ y puede usarse la aproximación de la expresión de inconsistencia local que aparece en **[1.4.21]**.

Por otra parte, se puede suponer, sin pérdida de generalidad, que las n_1 primeras observaciones se han generado con el modelo **[1.2.7]** y que, para las restantes, la única diferencia es que $V(\varepsilon_i) = \sigma_0^2 h$. Entonces, se puede plantear que:

$$h_i = 1 + \alpha z_i, \qquad i = 1, 2, \ldots, n \qquad \textbf{[2.3.13]}$$

donde α es un escalar y

$$z_i = \begin{cases} 0 & \text{si} \quad i = 1, 2, \ldots, n_1 \\ 1 & \text{si} \quad i = n_1 + 1, \ldots, n \end{cases} \qquad \textbf{[2.3.14]}$$

Esta especificación corresponde a una situación donde $V(\varepsilon_i) = \sigma_0^2$ para las primeras n_1 observaciones, $V(\varepsilon_i) = \sigma_0^2 h^2$ para las restantes y los elementos de $\boldsymbol{\beta}$ se estiman imponiendo la normalización del primer grupo. El parámetro α puede interpretarse como h - 1, resultando que, a partir de **[2.3.12]**, la variable omitida para obtener la especifica-

ción bajo la hipótesis nula es $(x_i^T\beta)z_i$, que es un vector $n\times 1$ con los primeros n_1 elementos iguales a cero y los restantes iguales a $x_i^T\beta$. Sustituyendo este vector por Z y α por h-1 en la expresión **[1.4.21]** con Ω definida en **[1.4.22]**, se obtiene la expresión de inconsistencia local del EMV de β bajo la hipótesis nula.

Davidson y MacKinnon (1984) presentan resultados de Monte Carlo en los que se muestra que, con el problema de variables omitidas, la variante LM_2 de **[1.4.15]** es preferible para llevar a cabo los contrastes. La expresión del término adicional que se requiere para evaluar la expresión en **[1.4.14]** es:

$$\nabla v_i(\hat{\theta})^T = \begin{bmatrix} \partial v_i(\hat{\theta})/\partial \alpha \\ \partial v_i(\hat{\theta})/\partial \beta \end{bmatrix} = \begin{bmatrix} -(x_i^T\hat{\beta})z_i \\ x_i \end{bmatrix} \qquad \textbf{[2.3.15]}$$

por lo que es inmediata la aplicación de las expresiones **[1.4.13]** y **[1.4.14]** para evaluar el estadístico de **[1.4.15]**.

De un modo similar se puede parametrizar la situación en que las observaciones anómalas proceden de diferentes vectores de parámetros. En este caso se tiene que:

$$v_i(\theta) = x_i^T\beta + z_i^T\alpha \qquad \textbf{[2.3.16]}$$

donde ahora:

$$z_i = \begin{cases} \mathbf{0}_k & \text{si} \quad i = 1,2,\ldots,n_1 \\ x_i & \text{si} \quad i = n_1+1,\ldots,n \end{cases} \qquad \textbf{[2.3.17]}$$

y el vector α puede interpretarse como la diferencia entre los vectores β y γ de **[2.2.7]**. En este caso, también es posible evaluar la expresión de inconsistencia local del EMV, no siendo necesaria la aproximación de Taylor, y derivar expresiones del estadístico LM_2 para llevar a cabo el contraste.

2.3.3. Sensibilidad de los modelos

En este apartado se analiza la sensibilidad del estimador máximo-verosímil ante la presencia de observaciones atípicas en el caso de los dos modelos de elección binaria más utilizados: el modelo probit y el modelo logit.

En lo que sigue se utiliza la definición de residuo habitual en econometría [véase Pregibon (1981), Jennings (1986) y Cox y Snell (1989), por ejemplo]; esto es:

$$e_i = y_i - E(y_i \mid \boldsymbol{x}_i) = y_i - P_i \qquad \textbf{[2.3.18]}$$

Partiendo de **[2.3.28]**, el gradiente de la función de verosimilitud en **[1.3.3]** puede escribirse:

$$\nabla \ell = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} w_i e_i \boldsymbol{x}_i \qquad \textbf{[2.3.19]}$$

donde:

$$w_i = \frac{f_i}{F_i(1 - F_i)} \qquad \textbf{[2.3.20]}$$

El estimador máximo-verosímil se define igualando la expresión **[2.3.19]** a $\mathbf{0}_k$, la cual presenta una gran similitud con el estimador por mínimos cuadrados ponderados en un modelo lineal. En la **Figura 2.4** se representan las ponderaciones w_i de los modelos probit y logit para un rango dado de valores de $\boldsymbol{x}_i^T\boldsymbol{\beta}$.

Obsérvese que a partir de las expresiones **[2.3.20]** y **[1.2.22]**, para el modelo logit se tiene que: $w_i = 1 \ \forall i$; esto es, todas las observaciones tienen idéntico peso en la función de verosimilitud. Por el contrario, en el caso del modelo probit binario, w_i es una función convexa con un mínimo en cero, lo que implica que las observaciones con valores extremos de $\boldsymbol{x}_i^T\boldsymbol{\beta}$ tienen las mayores ponderaciones en las condiciones de primer orden de la función de verosimilitud.

La propiedad anterior puede interpretarse como que el modelo logit será más robusto que el probit ante valores extremos en el espacio de las X. Dicha propiedad es un reflejo de la forma de la distribución normal, donde las colas son más finas que en la logística, por lo que valores escasamente probables tienen un efecto más importante en la estimación.

Otra forma de ilustrar este resultado consiste en despejar w_i en función de e_i en **[2.3.20]**, que para el modelo probit resulta:

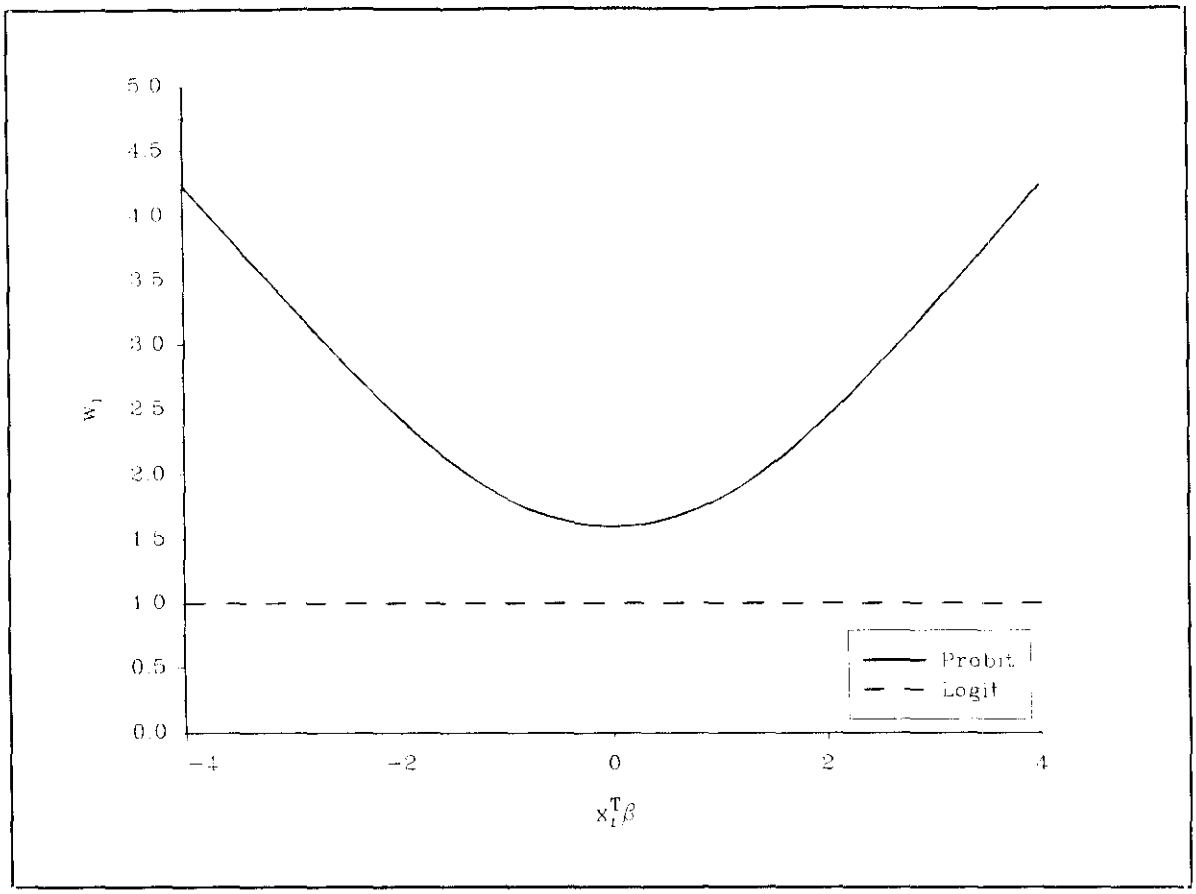

Figura 2.4: Representación de w_i para los modelos probit y logit.

$$w_i = \frac{\phi\left[\Phi^{-1}(1/2 + |\,1/2 - |e_i|\,|)\right]}{|e_i|\,(1 - |e_i|)} \qquad \textbf{[2.3.21]}$$

Nuevamente, la función w_i es convexa, lo que significa que valores con residuos grandes, esto es, próximos a la unidad en valor absoluto, tendrán un peso en la estimación superior al de otras observaciones. Ambas interpretaciones son equivalentes puesto que, a diferencia de un modelo lineal, en los MEB, sólo pueden producir residuos elevados en valor absoluto observaciones tales que $\boldsymbol{x}_i^T\boldsymbol{\beta}$ sea elevado, es decir, valores en las colas.

2.4. Resultados con datos simulados

En esta sección se ilustran los resultados teóricos de las secciones anteriores utilizando datos simulados. En concreto, se analizan los sesgos en la estimación MV de modelos probit y logit derivados de la existencia de anomalías en la muestra.

2.4.1. Planteamiento de los modelos

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

$$y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \textbf{[2.4.1]}$$

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* \geq 0 \\ 0 & \text{si} \quad y_i^* < 0 \end{cases} \qquad \textbf{[2.4.2]}$$

La variable explicativa se ha generado, en todos los casos de observaciones no anómalas, como una normal, $x_i \sim iid\ N(0,1)$ y el vector de parámetros es: $\beta = (-0.65,\ 1)^T$. Para este vector de parámetros, y usando tanto la distribución logística como la normal para ε_i, la proporción de unos en la muestra es aproximadamente del 25%.

A partir de este mecanismo, se han creado muestras donde se incluye un porcentaje ω de observaciones y_i^* generadas por la misma distribución que para el resto, pero con momentos distintos de los que se acaba de señalar. En particular, se consideran los siguientes casos, cuyos planteamientos teóricos se han discutido en la **Sección 2.3**:

Caso 1: Un porcentaje ω de observaciones y_i^* en la muestra proviene de una distribución con la misma media que las restantes observaciones, pero con $h^2 = 7$.

Caso 2: Un porcentaje ω de observaciones y_i^* proviene de la misma distribución con varianza igual a σ_0^2, pero con distinta media que las restantes observaciones. En concreto, se ha incluido una proporción ω de observaciones tales que

$E(y_i^*) = x_i^T\gamma$, donde $\gamma=(1, -0.5)^T$. Obsérvese que se ha considerado un caso extremo, en el que las componentes del vector γ son muy diferentes a las del vector β, formando un ángulo de 178° aproximadamente.

Caso 3: Un porcentaje ω de observaciones están caracterizadas por $E(y_i^*) = x_i^T\delta$, e idéntica varianza que las demás, donde $\delta = (-0.5, 0.5)^T$ 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 (forman un ángulo de 14° aproximadamente) pero, es de esperar que, para las observaciones anómalas, una proporción importante de los valores de x_i sean mucho mayores que los de las restantes observaciones.

Estos tres esquemas pretenden reflejar situaciones que pueden producirse con cierta frecuencia en el análisis de datos de corte transversal. El caso 1 se basa en la idea de que la heterocedasticidad aparece frecuentemente en datos de sección cruzada. Los casos 2 y 3 pueden interpretarse como originados por las técnicas de muestreo, básicamente el muestreo estratificado [Azorín y Sanchez-Crespo (1986)], técnica con la que se pueden estar incluyendo en la muestra elementos de subpoblaciones distintas entre sí; por un lado, respecto del comportamiento, aunque no respecto a sus variables características (caso 2) y por otro, respecto de sus variables características aunque homogéneas en su comportamiento (caso 3).

2.4.2. Aspectos técnicos de la simulación

Todos los cálculos, en esta sección y en el resto del trabajo, se han llevado a cabo utilizando el paquete matemático GAUSS 386 versión 2.2.

Un primer aspecto relevante en estos experimentos, es la técnica para generar variables aleatorias. Esta cuestión se ha resuelto de diferente modo para cada una de las distribuciones utilizadas. En el caso de la normal, se han utilizado las funciones internas del paquete estadístico después de comprobar la adecuación de dichas funciones (independencia de diferentes muestras generadas consecutivamente y ajuste a la normal). Para generar variables aleatorias logísticas se usa la transformación integral [Arnáiz (1978)], de modo que se generan valores de z_i a partir de una distribución uniforme (0,1) usando la función interna de GAUSS, y el valor de la variable aleatoria se obtiene evaluando $\Lambda^{-1}(z_i)$. Una interesante discusión sobre estos aspectos del análisis econométrico puede encontrarse en Quandt (1983).

La estimación de los modelos se ha realizado por el método de máxima verosimilitud por procedimientos lineales, descrito en el **Apartado 1.3.2**. En este proceso se utiliza un sólo criterio de parada basado en los parámetros, que puede formularse:

$$\sum_{j=1}^{k} |\hat{\beta}_j^{\tau+1} - \hat{\beta}_j^{\tau}| \leq 10^{-3} \qquad \textbf{[2.4.3]}$$

En este caso, no se utilizan criterios relativos debido a la homogeneidad en la escala de las variables.

Todos los experimentos se han realizado con muestras de tamaño $n = 200$ y se han replicado 500 veces. Los resultados que se presentan en las tablas son valores medios, para las diferentes replicaciones, de los siguientes estadísticos:

- Las estimaciones de los parámetros β_1 y β_2 y la desviación típica de β_2, puesto que es la de mayor interés en cualquier aplicación empírica.

- El *error cuadrático medio* (*ECM*) definido como:

$$ECM = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta) \qquad \textbf{[2.4.4]}$$

  que es una medida de la precisión de la estimación.

- La *suma de residuos al cuadrado* (*SSR*) definida como:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{P}_i)^2 \qquad \textbf{[2.4.5]}$$

  como medida alternativa de la precisión de la estimación al ECM.

2.4.3. Resultados de la simulación

En las **Tablas 2.1** a **2.6** se presentan los resultados de la estimación MV para cada uno de los tres casos expuestos en el **Apartado 2.4.1**, para los modelos probit y logit. En la primera fila de las tablas, figuran las medias de las estimaciones muestrales de los parámetros con $\omega = 0.0$; esto es, sin anomalías en la muestra.

A partir de estas tablas, un primer resultado que requiere algún comentario es el sesgo positivo y sistemático en las estimaciones de los parámetros con las muestras sin

observaciones anómalas y que se debe a que el estimador MV es *sesgado* en este tipo de modelos. Cox y Hinkley (1984, pag. 309) derivan expresiones generales para el sesgo basadas en evaluar los términos de tercer orden de la aproximación de Taylor que se emplea para maximizar la función de verosimilitud [véase **Apéndice A.2**] y como ejemplo, Copas (1988, pag. 230) obtiene la expresión particular para un modelo logit con una sola variable explicativa, que es:

$$\frac{1}{2}\,\frac{\sum x_i^3 P_i(1-P_i)(2P_i-1)}{\left(\sum x_i^2 P_i(1-P_i)\right)^2} \qquad \textbf{[2.4.6]}$$

y que tiene el mismo signo que el parámetro β, por lo que generalmente estará sobreestimado en términos absolutos. Utilizando una aproximación de P_i alrededor de $x_i = 0$ y suponiendo valores pequeños de β, se sugiere que el sesgo puede aproximarse por 0.034β.

Centrando la atención en el problema de observaciones anómalas, los resultados más importantes de la simulación, pueden resumirse en los siguientes puntos:

- En los tres casos considerados, tanto en el modelo probit como en el logit, el valor de los parámetros estimados se aleja de su valor teórico a medida que aumenta el porcentaje de anomalías en la muestra, como cabría esperar a partir de las conclusiones de la **Sección 2.3**. Como consecuencia de este sesgo de estimación, el ECM y la SSR presentan valores más altos cuanto mayor es ω, con la única excepción del caso 3, tanto para el modelo probit como para el logit. La explicación de este hecho es que al tratarse de observaciones extremas, pero generadas con parámetros similares a las observaciones no anómalas, los errores de previsión son menores.

- Además, tal y como cabría esperar, las desviaciones típicas apenas cambian en los casos 1 y 2 puesto que las observaciones son homogéneas y la variación sólo se debe al factor de ponderación de **[2.3.20]**. Por el contrario, para el caso 3, con observaciones extremas, la desviación típica sufre un fuerte cambio, ya que los valores extremos de x_i provocan una importante disminución en la matriz de varianzas estimada a partir de la inversa de **[1.3.5]**.

- Los resultados para el caso 1 figuran en las **Tablas 2.1** y **2.4** para el modelo probit y logit, respectivamente. Obsérvese que los sesgos en la estimación de β_1 y β_2 son menores que en los otros casos. Esto es debido a que, aunque la

varianza de las observaciones anómalas es igual a 7, estas observaciones están aleatoriamente distribuidas alrededor de la recta teórica $x_i^T\beta$, por lo que se produce una cierta *compensación* asimétrica. Dicha asimetría se debe a que, no todas las observaciones anómalas son percibidas por el modelo. 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, se obtienen unos cambios considerablemente mayores que los presentados en las **Tablas 2.1** y **2.4**.

- El caso 2, donde existen observaciones muestrales con una media muy distinta de las restantes, parece especialmente grave (**Tablas 2.2** y **2.5**). Obsérvese que, solamente con un 5% de este tipo de anomalías en la muestra, los sesgos de estimación en los dos parámetros son muy elevados, siendo casi tan altos como los detectados en el caso 1 cuando $\omega = 30\%$. En la práctica, lógicamente se desconoce el origen de las anomalías existentes; sin embargo, este análisis muestra que es erróneo pensar que un número reducido de anomalías no pueda tener efectos apreciables en la estimación del modelo, ya que esto depende tanto del tipo como de la magnitud de las mismas.

- Los resultados de las simulaciones para el caso 3 se muestran en las **Tablas 2.3** y **2.6**. Este caso trata de reflejar una situación habitual en la práctica, donde para los individuos anómalos, no sólo 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. 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.

- Comparando los diferentes resultados entre los modelos probit y logit se observa, tal y como se sugiere en el **Apartado 2.3.2**, que el segundo es más robusto que el primero ante los distintos esquemas de anomalías, aunque esa diferencia es menor de lo que cabía esperar a la vista de la **Figura 2.4**.

- Por último, un aspecto también relevante es que, a pesar de que los porcentajes de observaciones anómalas en las tablas pueden tomar valores *demasiado grandes*, como el 30%, esto no debe interpretarse como que realmente puede existir un 30% de observaciones anómalas en la muestra sin que el investigador lo haya notado sino, más bien, como una medida del efecto que podrían producir esa proporción de observaciones anómalas con los parámetros dados o un número inferior con parámetros característicos más *extremos*.

Tabla 2.1. Estimaciones MV con anomalías en la muestra: modelo probit, caso 1.

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6519	1.0070	0.1449	0.0338	29.23
2.5	-0.6415	0.9945	0.1440	0.0360	29.99
5.0	-0.6276	0.9699	0.1414	0.0374	30.16
7.5	-0.6090	0.9331	0.1380	0.0416	30.86
10.0	-0.5985	0.9193	0.1361	0.0426	31.66
12.5	-0.5759	0.8901	0.1334	0.0502	31.75
15.0	-0.5647	0.8600	0.1308	0.0557	32.41
17.5	-0.5495	0.8459	0.1295	0.0653	32.75
20.0	-0.5368	0.8190	0.1267	0.0723	32.97
22.5	-0.5182	0.7994	0.1249	0.0849	33.5
25.0	-0.5185	0.7814	0.1237	0.0941	33.99
27.5	-0.4902	0.7510	0.1212	0.1158	34.48
30.0	-0.4829	0.7370	0.1199	0.1226	34.81

Tabla 2.2. Estimaciones MV con anomalías en la muestra: modelo probit, caso 2.

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6594	1.0280	0.1473	0.0353	29.26
2.5	-0.5740	0.9064	0.1347	0.0467	31.24
5.0	-0.5056	0.8306	0.1271	0.0770	33.13
7.5	-0.4516	0.7644	0.1215	0.1199	34.71
10.0	-0.3978	0.6916	0.1156	0.1811	36.29
12.5	-0.3488	0.6431	0.1118	0.2379	37.90
15.0	-0.2994	0.5959	0.1085	0.3071	39.28
17.5	-0.2553	0.5598	0.1056	0.3666	40.57
20.0	-0.2144	0.5138	0.1030	0.4430	41.81
22.5	-0.1815	0.4757	0.1009	0.5111	42.62
25.0	-0.1508	0.4438	0.0995	0.5742	43.75
27.5	-0.1105	0.4157	0.0980	0.6477	44.72
30.0	-0.0763	0.3845	0.0969	0.7234	45.38

Tabla 2.3. Estimaciones MV con anomalías en la muestra: modelo probit, caso 3.

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6658	1.0300	0.1470	0.0384	29.15
2.5	-0.6580	0.9918	0.1454	0.0444	28.73
5.0	-0.6575	0.9567	0.1431	0.0497	28.20
7.5	-0.6503	0.9226	0.1403	0.0572	27.84
10.0	-0.6407	0.8711	0.1342	0.0695	27.68
12.5	-0.6386	0.8625	0.1333	0.0702	26.88
15.0	-0.6448	0.8323	0.1298	0.0804	26.32
17.5	-0.6296	0.7993	0.1246	0.0913	26.06
20.0	-0.6402	0.7803	0.1219	0.0999	25.44
22.5	-0.6292	0.7589	0.1185	0.1069	24.88
25.0	-0.6323	0.7360	0.1134	0.1155	24.37
27.5	-0.6294	0.7253	0.1112	0.1253	23.78
30.0	-0.6189	0.6954	0.1050	0.1390	23.56

Tabla 2.4. Estimaciones MV con anomalías en la muestra: modelo logit, caso 1.

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6540	1.0140	0.1940	0.0634	38.40
2.5	-0.6505	1.0020	0.1938	0.0622	38.73
5.0	-0.6252	0.9897	0.1924	0.0663	39.02
7.5	-0.6264	0.9513	0.1894	0.0667	39.35
10.0	-0.5893	0.9343	0.1869	0.0714	39.55
12.5	-0.5978	0.9111	0.1863	0.0741	39.89
15.0	-0.5838	0.9036	0.1850	0.0762	40.39
17.5	-0.5696	0.8770	0.1827	0.0839	40.20
20.0	-0.5574	0.8657	0.1815	0.0855	40.64
22.5	-0.5409	0.8253	0.1793	0.1043	41.08
25.0	-0.5328	0.8129	0.1777	0.1082	41.30
27.5	-0.5234	0.8148	0.1781	0.1057	41.59
30.0	-0.5111	0.7996	0.1774	0.1150	41.71

Tabla 2.5. Estimaciones MV con anomalías en la muestra: modelo logit, caso 2

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6635	1.0360	0.1960	0.0733	38.54
2.5	-0.6003	0.9543	0.1887	0.0695	39.35
5.0	-0.5489	0.9043	0.1840	0.0761	40.24
7.5	-0.5134	0.8482	0.1805	0.1016	41.42
10.0	-0.4482	0.7874	0.1752	0.1398	42.21
12.5	-0.405	0.7407	0.1711	0.1751	42.95
15.0	-0.3556	0.6970	0.1682	0.2260	43.64
17.5	-0.3061	0.6520	0.1655	0.2840	44.28
20.0	-0.274	0.6022	0.1625	0.3456	45.05
22.5	-0.2411	0.5714	0.1603	0.3932	45.54
25.0	-0.1995	0.5413	0.1584	0.4582	46.20
27.5	-0.149	0.4944	0.1557	0.5474	46.59
30.0	-0.1077	0.4683	0.1545	0.6181	47.24

Tabla 2.6. Estimaciones MV con anomalías en la muestra: modelo logit, caso 3

$\omega\%$	$\hat{\beta}_1$	$\hat{\beta}_2$	$dt(\hat{\beta}_2)$	ECM	SSR
0.0	-0.6611	1.0230	0.1952	0.0681	38.37
2.5	-0.6473	0.9407	0.1839	0.0785	38.23
5.0	-0.6506	0.8734	0.1712	0.0855	37.94
7.5	-0.6718	0.8243	0.1603	0.0973	37.37
10.0	-0.6627	0.8016	0.1520	0.1033	36.88
12.5	-0.6507	0.7467	0.1396	0.1220	36.87
15.0	-0.6539	0.7093	0.1295	0.1350	36.60
17.5	-0.6627	0.7059	0.1250	0.1406	35.87
20.0	-0.6413	0.6776	0.1172	0.1564	35.71
22.5	-0.6668	0.6478	0.1091	0.1728	35.50
25.0	-0.6628	0.6452	0.1057	0.1764	34.84
27.5	-0.6445	0.6254	0.1006	0.1866	34.72
30.0	-0.6633	0.6195	0.0971	0.1901	34.12

Por último, se ilustra 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 2.5** contiene las probabilidades estimadas con el modelo probit en el caso 3 para valores de $\omega = 0.0\%$, 15% y 30%. Obsérvese que estos sesgos también pueden ser considerablemente altos. Además, las probabilidades estimadas son especialmente importantes para realizar previsión agregada, lo puede dar lugar a errores de previsión muy elevados.

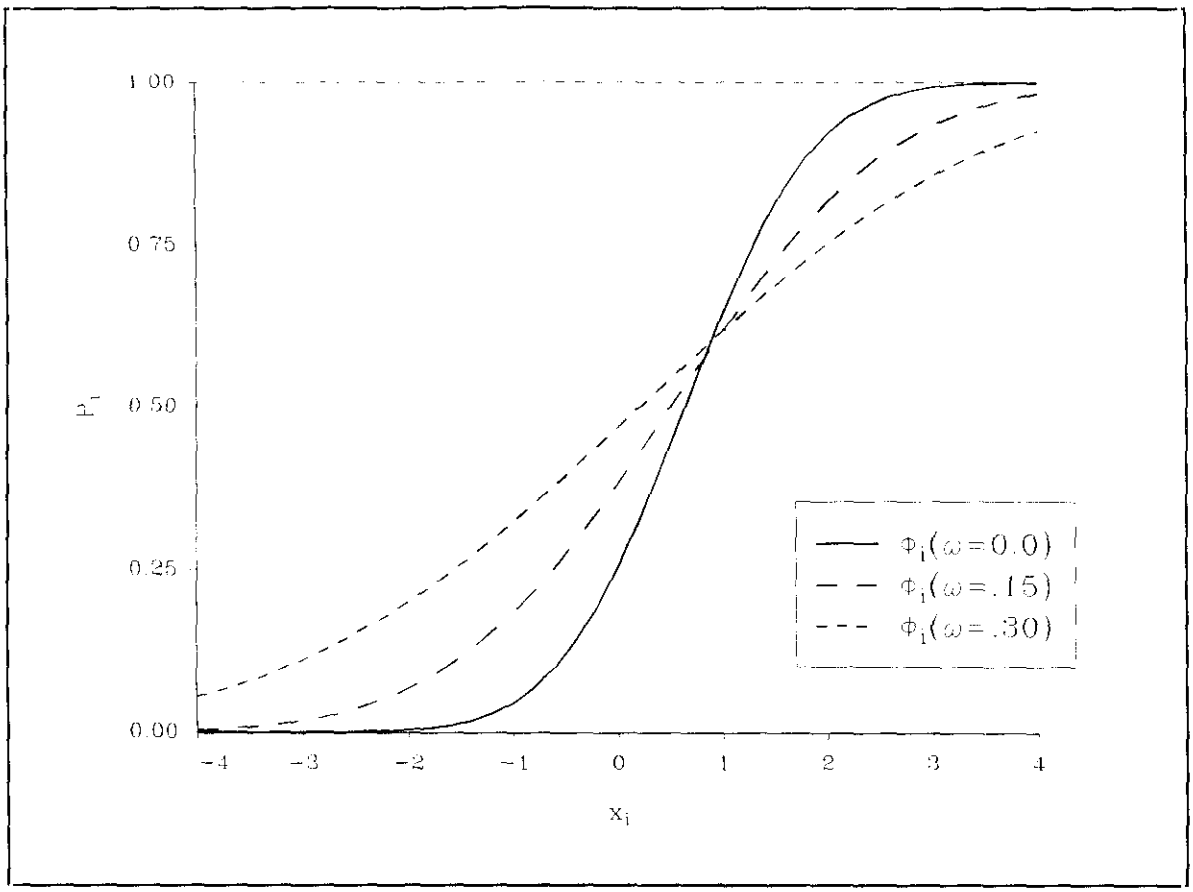

Figura 2.5: Ilustración del cálculo de las probabilidades estimadas en el Caso 3.

CAPÍTULO 3

OBSERVACIONES ANÓMALAS EN MODELOS DE ELECCIÓN BINARIA: DETECCIÓN

3.1. Introducción

En este capítulo se trata el problema de la detección de anomalías en los MEB. Como se ha indicado en el **Apartado 2.2.2**, los resultados que se desarrollan se encuentran en la línea de *robustecer* la metodología de estimación, tal y como se propone en Box (1980) y Peña y Ruiz-Castillo (1982 y 1984) para el caso de los modelos lineales de regresión. Con este propósito, en este capítulo se derivan estadísticos o medidas de influencia para la detección de anomalías en los MEB. Este es el primer paso para, posteriormente, decidir el tratamiento más adecuado que debe darse a las observaciones que se han detectado como anómalas.

Siguiendo este planteamiento, el primer objetivo de este trabajo es mostrar que, contrariamente a lo que se ha propuesto en la mayoría de la literatura anterior, en los MEB los análisis que se apoyan en los residuos, o en simples extrapolaciones de los resultados para el modelo lineal general, no resultan adecuados. 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.

En la **Sección 3.2** se presenta una amplia batería de estadísticos generalmente utilizados en la detección de observaciones anómalas en el modelo lineal general. Esto sirve como fundamento de los principales resultados del capítulo.

En la **Sección 3.3** se exponen las particularidades que presentan los MEB a la hora de detectar observaciones anómalas e influyentes, y que hacen que no sea suficiente el empleo de simples extrapolaciones de los resultados de la sección anterior.

En la **Sección 3.4** se deriva un estadístico de influencia aplicable a los MEB y se discute su aplicabilidad, así como sus diferencias respecto a los expuestos en la **Sección 3.2**. Además, se plantea la metodología que sería deseable aplicar a un MEB a fin de determinar la presencia de observaciones influyentes.

Por último, en la **Sección 3.5** se ilustran los resultados de la sección anterior aplicándolos a un conjunto de datos simulados, como los que se utilizaron en la **Sección 2.4**, con objeto de validar los estadísticos de la **Sección 3.4**.

3.2. Instrumentos de detección en el MLG

En la **Sección 2.3** se han presentado los problemas derivados de la presencia de observaciones anómalas en los modelos de elección binaria. Antes de entrar en los mecanismos de detección de anomalías específicos para los MEB, vamos a exponer los principales resultados para el MLG, puesto que los resultados de la **Sección 3.4** se basan, parcialmente, en los de ésta.

En términos generales, existen dos medidas estadísticas básicas que, individualmente y combinadas, permiten caracterizar la presencia de observaciones anómalas. En primer lugar, las denominadas medidas *a priori*, que señalan los vectores de observaciones de variables explicativas alejados del centro del espacio muestral de las X. En segundo lugar, las medidas *a posteriori*, que ofrecen información sobre el efecto de cada observación (o grupo de observaciones) sobre los resultados de estimación relevantes del modelo, principalmente sobre los coeficientes estimados. Este segundo grupo de estadísticos son los denominados *estadísticos de influencia*.

3.2.1. Instrumentos de diagnóstico *a priori*

Sea el modelo lineal de regresión:

$$y = X\beta + \varepsilon \qquad [3.2.1]$$

donde y es la variable endógena continua, X es la matriz de variables exógenas, β es el vector de parámetros desconocidos y, por último, ε es un vector de perturbaciones.

Sea $H = X(X^TX)^{-1}X^T$, la matriz de proyección en el modelo lineal general. La matriz H es idempotente y semidefinida positiva con rango(H) = rango(X) = k, siendo su elemento característico:

$$h_{ij} = \{H_{ij}\} = x_i^T(X^TX)^{-1}x_j \qquad [3.2.2]$$

La matriz H puede utilizarse para analizar el efecto que cada observación tiene sobre la variable y_i estimada. Dado que $\hat{y} = Hy$, para una observación concreta se puede plantear:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = \sum_{j \neq i} h_{ij} y_j + h_{ii} y_i \qquad \textbf{[3.2.3]}$$

La interpretación de la expresión anterior es que h_i ($\equiv h_{ii}$) mide el efecto que tiene la y_i observada sobre la y_i ajustada.

También se puede interpretar h_i como una medida directa de distancia. Si se denota por $\tilde{X}$ la matriz X en diferencias respecto a las medias de cada columna, se puede definir la distancia de cada vector de observación $\boldsymbol{x}_i$ al centro de las X como:

$$d_i = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \left[\frac{1}{n} \tilde{X}^T \tilde{X} \right]^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \qquad \textbf{[3.2.4]}$$

de modo que es posible demostrar que [Peña (1987)]:

$$h_i = \frac{1}{n} (1 + d_i) \qquad \textbf{[3.2.5]}$$

esto es, h_i es una transformación monótona de la distancia **[3.2.4]** de cada observación al centro del espacio muestral de las X.

Es fácil demostrar que h_i tiene las dos propiedades siguientes:

$$\begin{gathered} 1/n \leq h_i \leq 1 \\ \sum h_i = k \end{gathered} \qquad \textbf{[3.2.6]}$$

por lo tanto, si una matriz X está perfectamente equilibrada, esto es, todas sus observaciones tienen el mismo peso, se tendría que: $h_i = k/n \ \forall i$. De hecho, tal y como argumentan Belsley et al. (1980), cuando h_i toma valores superiores a $2k/n$ la observación en cuestión requiere mayor atención y para valores superiores a $3k/n$ se puede afirmar que la observación es extrema respecto de las restantes observaciones en X.

En Belsley et al. (1980) se sugiere el empleo de la matriz ampliada $Z = [\boldsymbol{y}\ X]$ de forma semejante a como se ha expuesto para la matriz X, con el fin de detectar observaciones alejadas del centro de gravedad de Z. Este punto de vista añade la consideración de $\boldsymbol{y}$ como fuente potencial de anomalías, como es el caso de observaciones para las que $\boldsymbol{x}_i$ no presenta ningún problema pero y_i está muy alejado del centro de las $\boldsymbol{y}$. Las medidas de distancia **[3.2.4]** o **[3.2.5]** se pueden adaptar de forma inmediata a este caso, con tan sólo sustituir X por Z.

Adicionalmente, se aconseja el empleo del *estadístico de Wilks* [Belsley et al. (1980, pag. 26)] para contrastar la diferencia de medias entre dos poblaciones. En el problema que nos ocupa, una población estaría definida por la observación i y la otra por el resto de los datos. El estadístico de Wilks puede plantearse:

$$W(\tilde{z}_i) = \left[\frac{n}{n-1}\right](1-h_i)\left[1+\frac{\tilde{e}_i^2}{n-k-1}\right]^{-1} \qquad \textbf{[3.2.7]}$$

donde $\tilde{z}_i$ es la fila i de la matriz Z centrada respecto a la media y $\tilde{e}_i$ es el *residuo estudentizado* definido por:

$$\tilde{e}_i = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1-h_i}} \qquad \textbf{[3.2.8]}$$

donde $s_{(i)}$ es la desviación típica residual estimada omitiendo la observación i. Suponiendo que $\tilde{Z}$ está formada por n muestras independientes de una distribución normal k variante (se está excluyendo el término constante)[9], es posible demostrar que:

$$\frac{n-k}{k}\frac{1-W(\tilde{z}_i)}{W(\tilde{z}_i)} \rightarrow F_{k,n-k} \qquad \textbf{[3.2.9]}$$

La particularización del estadístico de Wilks para la matriz X resulta:

$$W(\tilde{x}_i) = \frac{n}{n-1}(1-h_i) \qquad \textbf{[3.2.10]}$$

y bajo supuestos equivalentes a los realizados anteriormente, es decir, que las filas de $\tilde{X}$ son muestras aleatorias de una normal k-1 variante (se está excluyendo el término constante), se puede demostrar que:

$$\frac{n-k-1}{k-1}\frac{1-W(\tilde{x}_i)}{W(\tilde{x}_i)} \rightarrow F_{k-1,n-k-1} \qquad \textbf{[3.2.11]}$$

Aunque los supuestos realizados para derivar las distribuciones de **[3.2.9]** y **[3.2.11]** resultan un tanto restrictivos, la distribución es útil para obtener, al menos, indicaciones de los valores críticos de los estadísticos propuestos.

---

[9] Para derivar la distribución del estadístico se parte del supuesto de que ambas poblaciones cuya diferencia de medias se intenta contrastar siguen una distribución normal, lo que obviamente, no es posible para el término constante.

Para terminar este apartado, conviene poner de relieve que los estadísticos derivados no pueden considerarse, en ningún caso, concluyentes sobre la posible anormalidad de una observación en el sentido establecido en el **Capítulo 2**. Sin embargo, sí presentan información relevante sobre la rareza relativa de la observación comparada con las restantes, tanto de X como de y, lo que es un claro indicativo de que es necesario prestar mayor atención al dato concreto.

3.2.2. Estadísticos de influencia

La mayoría de los estadísticos que se presentan en este apartado miden el efecto que tiene la eliminación de una observación (o un grupo de observaciones) sobre los resultados de la estimación del modelo, principalmente, sobre el vector de coeficientes y su matriz de covarianzas.

3.2.2.A. Algunos resultados previos

En primer lugar, se trata el problema de cómo evaluar eficientemente los elementos de un modelo estimado por MCO cuando se elimina de la muestra un conjunto de observaciones denotado por I. De esta forma, se utiliza el subíndice I entre paréntesis (I) para indicar que a una matriz le faltan las filas pertenecientes al conjunto I, y el subíndice I sin paréntesis para indicar que la matriz está formada exclusivamente por las filas del conjunto I. Un primer resultado matemático, que va a utilizarse posteriormente, es el denominado *lema de inversión de matrices* que puede formularse:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \qquad \textbf{[3.2.12]}$$

donde A es una matriz $m \times m$, B es $m \times k$, C es $k \times k$, D es $k \times m$ y, además, A y C son no singulares.

En el modelo lineal general es evidente que:

$$\begin{aligned} X_{(I)}^T X_{(I)} &= X^T X - X_I^T X_I \\ X_{(I)}^T y_{(I)} &= X^T y - X_I^T y_I \end{aligned} \qquad \textbf{[3.2.13]}$$

Con el fin de aplicar el lema de inversión de matrices a las expresiones anteriores, se hace:

$$A = X^T X \qquad B = X_I^T$$
$$C = I_p \qquad D = X_I \qquad \textbf{[3.2.14]}$$

donde p denota el número de filas de X_I.

Aplicando **[3.2.12]** a la primera igualdad de **[3.2.13]**, es inmediato que:

$$(X_{(I)}^T X_{(I)})^{-1} = (X^T X)^{-1} + (X^T X)^{-1} X_I^T \left[I_p - X_I (X^T X)^{-1} X_I^T \right]^{-1} X_I (X^T X)^{-1} \qquad \textbf{[3.2.15]}$$

Sustituyendo la expresión **[3.2.15]** en el estimador mínimo cuadrático de β sin el conjunto I, y despejando, es sencillo demostrar que:

$$\hat{\beta}_{(I)} = \hat{\beta} + (X^T X)^{-1} X_I \left[I_p - X_I (X^T X)^{-1} X_I^T \right]^{-1} (X_I \hat{\beta} - y_I) \qquad \textbf{[3.2.16]}$$

Expresión que permite calcular la estimación MCO de β omitiendo el conjunto I de observaciones a partir de la estimación con las n observaciones muestrales.

También son de interés las siguientes definiciones. Se denomina vector de *residuos previstos* al vector de residuos calculado a partir de estimaciones de β sin un conjunto de observaciones, esto es:

$$e_{(I)} = y - X\hat{\beta}_{(I)} \qquad \textbf{[3.2.17]}$$

En particular, $e_{I(I)}$ es el vector de residuos previstos para el grupo de observaciones del conjunto I.

La matriz de proyección H, para el conjunto de observaciones omitidas resulta:

$$H_I = X_I (X^T X)^{-1} X_I^T \qquad \textbf{[3.2.18]}$$

y la varianza residual estimada, sin las observaciones pertenecientes a I, se puede calcular como:

$$(n - k - p) s_{(I)}^2 = (n - k) s^2 - e_{I(I)}^T (I_p - H_I) e_{I(I)} \qquad \textbf{[3.2.19]}$$

donde $s^2 = e^T e / n\text{-}k$.

3.2.2.B. Estadísticos de influencia: observaciones individuales

Los estadísticos de influencia y herramientas de diagnosis que se presentan a continuación, se apoyan principalmente en los trabajos de Belsley et al. (1980), Krasker et al. (1983), Peña (1987) y Atkinson (1985), así como en los artículos de Cook (1977) y Cook y Weisberg (1980). En este apartado se considera el caso de eliminar una observación cada vez, es decir, se presentan estadísticos que miden el efecto de una observación en los resultados de estimación del modelo. La exposición que se presenta no es exhaustiva, debido a la amplia batería de estadísticos que es posible derivar. Para una revisión más completa se puede consultar Belsley et al. (1980) o Cook y Weisberg (1982).

Se debe comenzar señalando que, en la diagnosis de un modelo, los residuos *grandes* han sido considerados como una indicación de problemas en la especificación del mismo, puesto que indican una discrepancia entre el valor observado y estimado de la variable endógena. De hecho, un mínimo análisis de diagnosis de un modelo consiste en la inspección gráfica de los *residuos estandarizados*, que se pueden formular en términos de h_i como:

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_i}} \qquad \textbf{[3.2.20]}$$

aunque algunos autores [Krasker et al. (1983)] prefieren utilizar los *residuos estudentizados*, que emplean la desviación típica residual estimada sin la observación i:

$$\tilde{e}_i = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1 - h_i}} \qquad \textbf{[3.2.21]}$$

Estos residuos siguen una distribución t cuando la perturbación ε_i sigue una distribución normal. Obsérvese que el empleo de la desviación típica residual estimada sin la observación i, refuerza el efecto de *anomalía* si el residuo correspondiente es grande, pero no tendrá ninguna ventaja si el residuo es pequeño, puesto que la estimación de σ apenas cambiará.

Además del análisis de residuos, se puede plantear el estudio del efecto que produce la observación i en los valores ajustados. Más concretamente, se trata de medir cómo se ve afectada la previsión del i-ésimo valor de la variable endógena cuando ésta ha sido omitida en la estimación. Dicha medida es el residuo previsto definido en **[3.2.17]** que, para la observación eliminada, puede escribirse:

$$e_{i(i)} = y_i - \hat{y}_{i(i)} = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} \qquad \textbf{[3.2.22]}$$

Dado que y_i e $\hat{y}_{i(i)}$ son independientes, la varianza de $e_{i(i)}$ es [Atkinson (1985)]:

$$\sigma^2[\,1 + \boldsymbol{x}_i^T (X_{(i)}^T X_{(i)})^{-1} \boldsymbol{x}_i] = \sigma^2 \frac{1}{1 - h_i} \qquad \textbf{[3.2.23]}$$

Para estimar la varianza de la perturbación, Atkinson (1985) sugiere utilizar el estimador sin la observación i, de forma que una medida de la discrepancia entre el valor observado y el previsto, o lo que es lo mismo, el *residuo previsto estudentizado* para la i-ésima observación resulta:

$$\tilde{e}_{i(i)} = \frac{e_{i(i)}}{s_{(i)}\sqrt{1/(1 - h_i)}} = e_i^* \frac{s}{s_{(i)}} \qquad \textbf{[3.2.24]}$$

que sigue una distribución t_{n-k-1}, bajo el supuesto de normalidad para ε_i. Una interpretación interesante de **[3.2.24]** puede hacerse considerando que el residuo previsto estudentizado es el residuo estandarizado corregido por el ratio de desviaciones típicas estimadas con y sin la observación en cuestión. Este residuo es el que Belsley et al. (1980) denominan *RSTUDENT*.

Sin embargo, el análisis de residuos tiene un interés limitado, puesto que la forma más adecuada de analizar la influencia de una observación sobre las estimaciones de un modelo de regresión se basa en comparar los resultados de la estimación con la muestra completa y sin la observación objeto de interés. Esta comparación puede centrarse tanto en la y ajustada como en los coeficientes estimados o en su matriz de varianzas y covarianzas, aunque como se verá más adelante, los dos primeros ofrecen la misma información.

Siguiendo a Belsley et al. (1980), una medida elemental se basa en comparar el vector de coeficientes estimados con y sin la observación i-ésima. El estadístico correspondiente se denomina *DFBETA* y se obtiene de forma inmediata a partir de **[3.2.16]**:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(X^T X)^{-1} \boldsymbol{x}_i e_i}{1 - h_i} \qquad \textbf{[3.2.25]}$$

Alternativamente, se puede evaluar el efecto que tiene la observación i sobre el valor ajustado de la observación j-ésima. Multiplicando la expresión anterior por $\boldsymbol{x}_j^T$ resulta:

$$\hat{y}_j - \hat{y}_{j(i)} = \frac{h_{ij} e_i}{1 - h_i} \qquad \textbf{[3.2.26]}$$

Un paso lógico a partir de **[3.2.25]** consiste en obtener una medida escalar, para lo cual se normaliza la expresión por la matriz de varianzas-covarianzas y el número de coeficientes en β, lo que da lugar a:

$$c_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X)(\hat{\beta} - \hat{\beta}_{(i)})}{k s^2} \qquad \textbf{[3.2.27]}$$

que es el estadístico propuesto por Cook (1977).

Una forma más conveniente para el cálculo de c_i puede derivarse sustituyendo la expresión **[3.2.16]** en **[3.2.27]**, resultando:

$$c_i = \frac{e_i^2 h_i}{k s^2 (1 - h_i)^2} = \frac{e_i^{*2} h_i}{k (1 - h_i)} \qquad \textbf{[3.2.28]}$$

que refleja la relación entre c_i y los residuos estandarizados de **[3.2.20]**. Una derivación alternativa para c_i se obtiene evaluando la diferencia entre los vectores de valores ajustados con y sin la observación i:

$$\begin{aligned} c_i &= \frac{[X(\hat{\beta}_{(i)} - \hat{\beta})]^T [X(\hat{\beta}_{(i)} - \hat{\beta})]}{k s^2} \\ &= \frac{(\hat{y}_{(i)} - \hat{y})^T (\hat{y}_{(i)} - \hat{y})}{k s^2} \end{aligned} \qquad \textbf{[3.2.29]}$$

que puede interpretarse como una medida del efecto que se produce en el vector de valores ajustados ante la ausencia de la i-ésima observación.

Atkinson (1985) sugiere el empleo del *estadístico de Cook modificado*, similar al *DFFITS* de Belsley et al. (1980), que puede escribirse:

$$\tilde{c}_i = \left[\frac{n - k}{k} \frac{h_i}{1 - h_i} \right]^{1/2} |\tilde{e}_{i(i)}| \qquad \textbf{[3.2.30]}$$

Las diferencias entre el estadístico **[3.2.30]** frente al de **[3.2.28]**, además de usar la raíz cuadrada positiva, son las siguientes: i) para el caso en el que todas las observaciones de la matriz X tienen idéntico peso ($h_i = k/n$), resulta que $\tilde{c}_i = |\tilde{e}_{i(i)}|$, por lo que la distancia

entre las estimaciones se debe al vector $\boldsymbol{y}$, y ii) el empleo de $s_{(i)}$ en lugar de s como estimación de σ, así como utilizar $e_{i(i)}$ en vez de e_i como en el estadístico original, refuerza el efecto de la anomalía si el residuo es grande. La ventaja de estas modificaciones es que el estadístico resultante concede más peso a las observaciones anómalas que c_i.

El estadístico de Cook formulado en **[3.2.27]** tiene una clara interpretación geométrica: la magnitud de la distancia entre las estimaciones de β puede evaluarse comparando c_i con las probabilidades de la distribución $F_{k,n-k}$ centrada. Por ejemplo, si $c_i = F_{k,n-k}(0.5)$, significa que la eliminación de la observación i desplaza la estimación mínimo cuadrática de β a la frontera de la elipse de confianza del 50%, lo que resulta un cambio apreciable. Cook (1977) sugiere que, idealmente, ninguna observación debería provocar un desplazamiento superior al 10%.

Otro aspecto relevante en la estimación de un modelo, es la matriz de varianzas de los coeficientes, donde también quedarán reflejados los efectos de la presencia de observaciones anómalas en la muestra. Más concretamente, una medida de influencia podría obtenerse evaluando el cambio en la región de confianza del 1-α cuando se elimina una observación. El volumen del elipsoide de confianza es proporcional al determinante de la matriz de covarianzas, lo que puede expresarse como:

$$E \propto \left[\frac{s^{2k}}{|X^T X|} \right]^{1/2} \qquad \textbf{[3.2.31]}$$

y cuando se elimina la i-ésima observación, el volumen resulta:

$$E_{(i)} \propto \left[\frac{s_{(i)}^{2k}}{|X_{(i)}^T X_{(i)}|} \right]^{1/2} \qquad \textbf{[3.2.32]}$$

Belsley et al. (1980) denominan *COVRATIO* al estadístico:

$$COVRATIO_i = \left[\frac{E_{(i)}}{E} \right]^2 = \frac{s_{(i)}^{2k}}{s^{2k}} \frac{1}{1 - h_i} \qquad \textbf{[3.2.33]}$$

y tomando el logaritmo de *COVRATIO* se obtiene una medida que toma el valor nulo cuando el volumen no se ha visto afectado por la eliminación de i. Evaluando el estadístico **[3.2.33]** para situaciones concretas, Belsley et al. (1980, pag. 22) determinan unos valores críticos para *COVRATIO*, considerando que una observación tiene un efecto apreciable si el estadístico cae fuera del intervalo $1 \pm 3k/n$. Cook y Weisberg (1980) utilizan el

logaritmo del *COVRATIO*, pero ajustado por los valores de la distribución F de la definición de la región de confianza.

Por último, Peña (1987) presenta una medida global de la robustez del modelo que denomina *coeficiente de robustez*, y que puede definirse:

$$B^2 = \frac{\sum e_i^2}{\sum e_{i(i)}^2} = \left[\frac{1}{n-(k+1)} \frac{e_i^2}{s^2(1-h_i)^2} \right]^{-1} \qquad \textbf{[3.2.34]}$$

Es claro que B^2 está acotado entre cero y uno, de forma que tomará valores próximos a la unidad cuando los valores ajustados $\hat{y}_{i(i)}$ estén próximos a $\hat{y}_i$, y se aproximará a cero cuanto mayor sea la diferencia que presenten, es decir, cuando más afectados estén por la eliminación de la observaciones.

3.2.2.C. Estadísticos de diagnóstico: grupos y otras extensiones

En general, no es complicado extender los estadísticos presentados en el apartado anterior al caso en que se desea medir el efecto que se produce cuando se eliminan grupos de observaciones. La dificultad estriba en que los criterios de selección de grupos no son únicos ni concluyentes. En el peor de los casos, sería necesario decidir un tamaño máximo de grupo e ir probando con todos los grupos posibles de tamaño menor o igual al máximo lo que, obviamente, sólo es posible en muestras relativamente pequeñas y para tamaños máximos reducidos. En este apartado, se generalizan los estadísticos del **Apartado 3.2.2.B** para el caso en que se eliminan grupos de observaciones y en el **Apartado 3.2.2.D** se presentan algunos resultados para tratar el problema del enmascaramiento.

Uno de los mayores inconvenientes al tratar con grupos de observaciones, es que algunas de las magnitudes escalares expuestas para una sola observación se convierten en vectoriales o matriciales; esto ocurre, por ejemplo, con el estadístico h_i. Para transformar una matriz como H_I en **[3.2.18]** en una medida escalar, Cook y Weisberg (1980, 1982) proponen emplear la traza; otra alternativa [Atkinson (1985)] sería emplear el determinante. Esta multiplicidad de definiciones refleja el hecho de que la idea de peso en la estimación no es tan clara cuando se trata con grupos de observaciones.

La extensión de los estadísticos de influencia del apartado anterior a casos de eliminación de grupos de variables es inmediata, utilizando los resultados del **Aparta-**

do 3.2.2.A. En particular, a partir de **[3.2.16]** la influencia de un conjunto de observaciones puede evaluarse utilizando:

$$c_I = \frac{(\hat{\beta}_{(I)} - \hat{\beta})(X^T X)(\hat{\beta}_{(I)} - \hat{\beta})}{k s^2} = \frac{e_I^T (I_p - H_I)^{-1} H_I (I_p - H_I)^{-1} e_I}{k s^2} \qquad \textbf{[3.2.35]}$$

estadístico similar al de las ecuaciones **[3.2.27]** y **[3.2.30]**. Si en lugar de s^2 se emplea $s^2_{(I)}$ y los residuos previstos $e_{I(I)}$, se obtendría un estadístico modificado similar al de **[3.2.30]**.

Un modo alternativo de calcular c_I de forma aproximada puede derivarse a partir de la matriz de influencia M [Peña y Yohai (1991)], que se obtiene evaluando el efecto conjunto de la eliminación simultánea de las observaciones i-ésima y j-ésima de la muestra y cuyo elemento genérico es:

$$m_{ij} = (\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(j)}) = \frac{e_i e_j h_{ij}}{k s^2 (1 - h_{ii})(1 - h_{jj})} \qquad \textbf{[3.2.36]}$$

donde $\hat{y}_{(i)}$ e $\hat{y}_{(j)}$ son los vectores de valores ajustados eliminado las observaciones i y j respectivamente y h_{ij} son los elementos de la matriz H definidos en **[3.2.2]**. Nótese que la diagonal principal de M está formada por el estadístico de Cook de **[3.2.28]** para cada observación.

Basándose en el comportamiento de la función de influencia teórica, Peña y Yohai (1991) sugieren que una forma aproximada de evaluar c_I es:

$$c_I \approx \sum_{i \in I} \sum_{j \in I} m_{ij} \qquad \textbf{[3.2.37]}$$

Otra extensión inmediata, consiste en particularizar las expresiones anteriores para parámetros individuales o conjuntos de parámetros. De hecho, el análisis de influencia en conjuntos de parámetros es una de las extensiones más interesantes, sobre todo cuando el conjunto de variables puede dividirse en grupos de forma natural. Por ejemplo, en un modelo lineal, donde los coeficientes de las variables son de mayor interés que el término constante, o en modelos de elección discreta, donde puede distinguirse entre variables características del individuo y variables características de la alternativa.

La particularización de las expresiones generales anteriores, puede plantearse de la siguiente forma [Cook y Weisberg (1980), Atkinson (1985)]. Se puede suponer que el conjunto de parámetros de interés son m elementos que se corresponden con las filas del vector $\theta = R^T\beta$, donde R^T es una matriz $m \times k$ de constantes conocidas, con rango$(R) = m \leq k$. La matriz de varianzas de la estimación por mínimos cuadrados de θ es $\sigma^2 R^T(X^TX)^{-1}R$. Por tanto, una medida de influencia análoga al estadístico en **[3.2.35]** para este caso es:

$$c_I(\theta) = \frac{(\hat{\theta}_{(I)} - \hat{\theta})^T \left[R^T(X^TX)^{-1}R\right]^{-1}(\hat{\theta}_{(I)} - \hat{\theta})}{m s^2} \qquad \textbf{[3.2.38]}$$

que también puede expresarse:

$$c_I(\theta) = \frac{e_I^T(I_p - H_I)^{-1}X_I N X_I^T(I_p - H_I)^{-1}e_I}{m s^2} \qquad \textbf{[3.2.39]}$$

donde:

$$N = (X^TX)^{-1}R\left[R^T(X^TX)^{-1}R\right]^{-1}R^T(X^TX)^{-1} \qquad \textbf{[3.2.40]}$$

El efecto de eliminar una sola observación sobre θ puede escribirse:

$$c_i(\theta) = \frac{e_i^{*2}\, x_i^T N x_i}{m(1 - h_i)} \qquad \textbf{[3.2.41]}$$

que es sencillo de calcular, puesto que N no depende de la observación eliminada.

Un caso particular interesante, se produce cuando se analiza un grupo de parámetros, por ejemplo, los últimos m componentes del vector β. En ese caso:

$$R^T\beta = (0_{m\times(k-m)} \quad I_m)\,\beta = (\beta_{k-m+1}, \ldots, \beta_k)^T \qquad \textbf{[3.2.42]}$$

y el estadístico de **[3.2.39]** puede escribirse [Atkinson (1985)]:

$$c_I(\theta) = \frac{k}{m}c_I - \frac{e_I^T(I_p - H_I)^{-1}G_I(I_p - H_I)^{-1}e_I}{m s^2} \qquad \textbf{[3.2.43]}$$

donde $G_I = X_2(X_2^TX_2)^{-1}X_2$ es la matriz de proyección sobre las últimas m variables del modelo. Cuando sólo se elimina una observación, la expresión **[3.2.43]** puede reducirse a:

$$c_i(\theta) = \frac{e_i^2 (h_i - g_i)}{m s^2 (1 - h_i)^2} \qquad \textbf{[3.2.44]}$$

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

Para el caso de un modelo de regresión, el estadístico para todos los parámetros excepto la constante, queda:

$$c_i(\theta) = \frac{e_i^{*2} (h_i - 1/n)}{(k-1)(1-h_i)} \qquad \textbf{[3.2.45]}$$

Finalmente, es posible particularizar la mayoría de los estadísticos planteados para analizar la influencia sobre un sólo parámetro. En este caso particular $R^T(X^TX)^{-1}R = v_j$; esto es, el elemento j-ésimo de la diagonal principal de la matriz de varianzas de β. El estadístico de Cook resulta:

$$c_i(\beta_j) = \frac{(\hat{\beta}_{(i)j} - \hat{\beta}_j)^2}{s^2 v_j} \qquad \textbf{[3.2.46]}$$

Naturalmente, estas formulaciones pueden modificarse en el sentido de la expresión **[3.2.30]**, introduciendo el estimador de la varianza residual con la observación omitida y diferentes definiciones de residuo.

3.2.2.D. Algunos tratamientos para el problema del enmascaramiento

El problema de enmascaramiento se produce 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 bien distintos, como se ilustraba en la **Figura 2.2**.

Los estadísticos para grupos, derivados en el **Apartado 3.2.2.C**, pueden utilizarse para analizar la influencia de cualquier conjunto de observaciones. 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. A continuación se exponen brevemente dos estrategias para tratar el problema siguiendo los planteamientos de Peña y Yohai (1991) y Rousseeuw y van Zomeren (1990).

El método propuesto por Peña y Yohai (1991) se basa en la matriz M definida en **[3.2.39]** y se justifica mediante un argumento heurístico. El planteamiento práctico es el siguiente:

Paso 1: Calcular los autovectores correspondientes a los k autovalores no nulos de la matriz de influencia M.

Paso 2: Utilizando los autovectores asociados a los m mayores autovalores, seleccionar los pares de conjuntos de observaciones I_j^1 y I_j^2, $j = 1,\ldots,m \leq k$, incluyendo en cada uno de ellos las observaciones cuyo componente del autovector sea grande y positivo o negativo, respectivamente.

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

En Peña y Yohai (1991) se aplica el método a diversos conjuntos de datos ya utilizados en la literatura y se pone en evidencia que este método permite seleccionar eficientemente grupos de observaciones influyentes que pasan desapercibidos al emplear estadísticos de influencia individual y con un coste computacional muy inferior al de otros métodos propuestos en la literatura. Conviene poner de relive que el coste de cálculo no es alto puesto que: i) existen algoritmos eficientes específicos para evaluar los mayores autovalores de un matriz real simétrica, por lo que no es necesario evaluarlos todos[10], y ii) no es necesario almacenar en la memoria del ordenador toda las matrices M y H, puesto que los elementos m_{ij} pueden ser evaluados a medida que sean necesitados si las limitaciones de espacio lo exigen.

Rousseeuw y van Zomeren (1990) plantean 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

[10] Una revisión de algoritmos para el cálculo de autovalores y autovectores para distintas matrices puede verse en Ralston y Rabinowitz (1978, cap. 10) y código eficiente para realizar los cálculos puede encontrarse en Smith et al. (1974).

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 artículo].

3.3. El problema de la detección de anomalías en MEB: el análisis de residuos

A pesar del planteamiento teórico desarrollado en la **Sección 2.3**, en la práctica, no se sabe *a priori* si existen observaciones anómalas en la muestra, ni mucho menos cuál es la distribución que las ha generado. Por lo tanto, al igual que en los modelos lineales, 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 haya sido generado por la distribución que se supone para las restantes observaciones. Sin embargo, el análisis para la detección de anomalías en los MEB presenta algunas peculiaridades respecto a los modelos lineales de regresión. Ello se debe a que en un MEB sólo se observa una realización dicotómica de la variable dependiente y_i^*.

El objetivo de esta sección es mostrar que en el proceso de detección de observaciones anómalas en los MEB, los residuos no juegan el mismo papel que en los modelos lineales. Ello se debe a que en un modelo de variable de elección cualitativa el valor de los residuos está acotado como consecuencia de la censura que presenta la variable y_i^*, de la que sólo se sabe si es mayor o menor que cero.

Definiendo el residuo correspondiente a la observación i-ésima como la diferencia entre y_i y $\hat{P}_i$, tal y como se ha indicado en el **Apartado 2.3.2**:

$$e_i = y_i - \hat{P}_i \qquad \textbf{[3.3.1]}$$

donde $E(e_i) = 0$ y $V(e_i) = P_i(1 - P_i)$. A partir de **[3.3.1]** es obvio que e_i está acotado entre $(-1,1)$[11] pudiendo tomar, para cada observación, solamente dos valores: $(1 - \hat{P}_i)$ y $-\hat{P}_i$.

En Pregibon (1981), donde se trata por primera vez el problema de la detección de observaciones anómalas en los MEB, se considera como anomalía toda observación que, una vez estimado el modelo, presenta un residuo e_i próximo a la unidad en valor absoluto. En consecuencia, en el citado trabajo se propone el análisis de residuos como un elemento de diagnosis para la detección de valores anómalos y se analizan los efectos en los coeficientes estimados así como algunos estadísticos basados en los residuos e_i estandarizados. En el contexto del trabajo citado, esto tiene bastante sentido, ya que se trata,

[11] En el modelo lineal de probabilidad el intervalo es cerrado puesto que puede haber residuos iguales a uno en valor absoluto. Sin embargo, dado que este modelo se utiliza escasamente en la práctica, no se considera este caso particular.

principalmente, de observaciones agrupadas y modelos lineales generalizados, en los que los residuos no plantean las mismas dificultades que en los MEB.

Posteriormente este análisis se extiende, entre otros, por Williams (1987), que plantea estadísticos de diagnóstico generales para modelos lineales generalizados a partir de los resultados de Pregibon (1981) y utiliza un enfoque de análisis de influencia. Copas (1988) analiza los modelos binarios bajo el supuesto de que aparecen errores en los datos, como por ejemplo, de codificación. Bedrick y Hill (1990) plantean un enfoque alternativo para modelos logit, pero también basado en el análisis de influyencia. El principal inconveniente del análisis de influencia se encuentra en que se plantean situaciones excesivamente *puras*, que, en muchos casos, no reflejan la realidad de los datos, donde lo frecuente es que aparezcan grupos de observaciones *diferentes* de la mayoría de la muestra y que pueden alterar apreciablemente los resultados del análisis.

Respecto al empleo de los residuos en modelos de elección binaria, Jennings (1986) critica el trabajo de Pregibon (1981), señalando los puntos siguientes: i) los residuos e_i no son comparables a los residuos MCO, ya que cada e_i depende de 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, ii) eliminar de la muestra datos con un residuo próximo a la unidad 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, nuestro punto de vista es que si una observación presenta un residuo e_i próximo a uno en valor absoluto, simplemente quiere decir que la $P(y_i=j \mid \boldsymbol{x}_i) < \alpha$, donde $j = 0, 1$ 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, 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 atendiendo exclusivamente a que el residuo correspondiente se encuentre cercano a uno 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 3.1**, donde se considera el caso del modelo probit. La parte inferior de la figura contiene la nube de puntos $(y_i^*, \boldsymbol{x}_i^T)$ (en este caso, $\boldsymbol{x}_i = (1\ x_i)^T$) asociada al modelo **[1.2.7]** y la correspondiente recta teórica. En la parte superior de la figura se han trasladado al eje de abscisas los valores

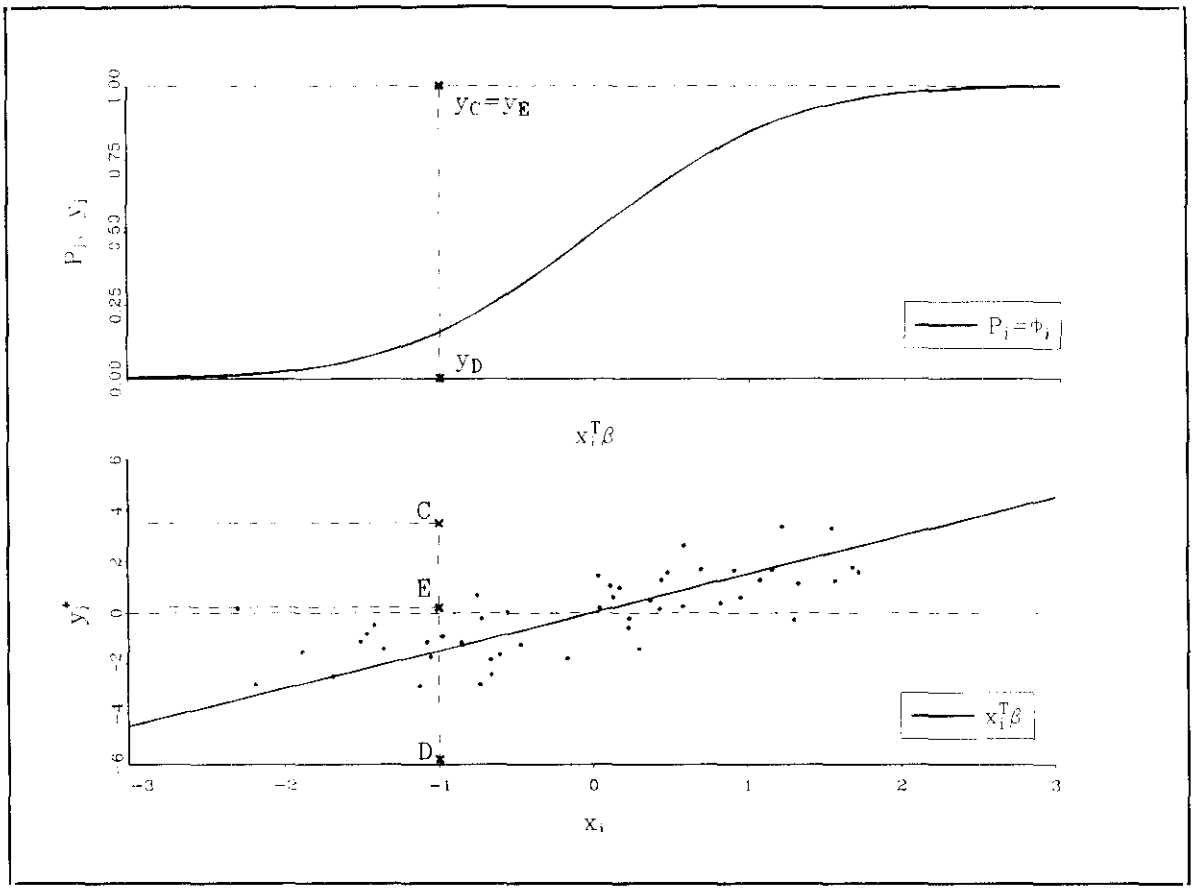

Figura 3.1: Ejemplos de anomalías en un modelo probit.

de la recta teórica $x_i^T\beta$, 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 $x_i^T\beta$, mientras que los valores de y_i responden a la relación **[1.2.2]**. El problema es que la muestra disponible para la estimación del modelo está formada solamente por los pares (y_i, x_i^T), por lo que *no se observa* la nube de puntos de la parte inferior de la figura. Entonces, pueden darse las siguientes situaciones:

- 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.

- 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, sería muy próximo a cero, puesto que, en este caso, $y_D = 0$ al ser $y_D^* < 0$.

Con este ejemplo, se trata de ilustrar que dos observaciones igualmente anómalas pueden presentar, dependiendo del signo de la variable no observable y_i^*, un residuo próximo a uno o a cero. 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:

- Que una observación i tenga un residuo próximo a cero 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 muy pequeño y ser realmente anómalas, como es el caso del punto **D** en la **Figura 3.1**. Nótese que en un modelo lineal también puede haber observaciones anómalas con un residuo próximo a cero, pero dichas anomalías no son del tipo **D**, que podría detectarse fácilmente por su residuo si se observase y_D^*.

- Que una observación tenga un residuo próximo a la unidad en valor absoluto puede ser un indicio de que se trata 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 3.1** presentaría un residuo e_i positivo y próximo a uno (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.

En definitiva, los residuos resultantes de la estimación de un modelo probit, o de cualquier otro MEB, no son informativos sobre la probabilidad que tiene cada observación de ser anómala. Como se acaba de ilustrar, residuos próximos en valor absoluto a uno o a cero 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 2.1**), pero donde un residuo grande sí presenta evidencia de que el correspondiente dato puede ser anómalo.

Por último, hay que señalar que, aunque el análisis de residuos no sea el instrumento adecuado para la detección de anómalos en los MEB, dicho análisis puede resultar de interés para detectar otros problemas. En una muestra dada, cabe esperar que un porcentaje pequeño de observaciones presente un residuo próximo a la unidad en valor absoluto, sean o no anómalas. 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 $\boldsymbol{x}_i$ no son relevantes para explicar la variable y_i^* y, por tanto, las probabilidades P_i y, ii) a la existencia, al menos, de dos grupos distintos de individuos en la muestra (la noción de *cambio estructural*), que debe identificarse y modelizarse de la forma más adecuada.

3.4. Procedimientos de detección de observaciones anómalas en los MEB

De la exposición en la sección anterior, se deduce que la forma adecuada para detectar si una observación es anómala en un MEB, debe ser mediante un estadístico que mida el efecto de esa observación sobre la estimación MV de los parámetros del modelo. Como se ha demostrado en el **Capítulo 2**, 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.

En esta sección, se deriva un conjunto de estadísticos que miden la influencia de observaciones individuales o grupos de observaciones sobre los dos aspectos fundamentales del modelo: los parámetros y las probabilidades estimadas. La estrategia general de estas derivaciones utiliza el modelo linealizado **[1.3.10]** y el estimador máximo-verosímil por procedimientos lineales **[1.3.9]**, y se aplican los planteamientos generales de los estadísticos para el modelo lineal general de la **Sección 3.4** teniendo en cuenta las particularidades de los MEB, que hacen que no todos los resultados anteriores sean de aplicación inmediata.

En concreto, el procedimiento de estimación máximo-verosímil lineal permite evaluar, con un coste de cálculo reducido, las estimaciones de los parámetros cuando se elimina un conjunto de observaciones, imprescindible para poder utilizar los estadísticos cuando la muestra con la que se trabaja es de gran tamaño.

3.4.1. Estadísticos para la detección de observaciones anómalas en los modelos de elección binaria

El estadístico que proponemos a continuación es una adaptación del presentado en **[3.2.27]** y mide el efecto de cada observación sobre la estimación MV de β en el modelo **[1.2.7]**. Para su derivación, se puede utilizar el Teorema 4.30 de White (1984, pag. 70):

Sea un vector de variables aleatorias $\boldsymbol{\Theta}$ de dimensión k. Si $\boldsymbol{\Theta} \rightarrow N_k(\mathbf{0},\Sigma)$, donde Σ es la matriz $(k \times k)$ de covarianzas asintótica de $\boldsymbol{\Theta}$, y existe una matriz $\hat{\Sigma}$ simétrica y definida positiva tal que plim $\hat{\Sigma} = \Sigma$, entonces $\boldsymbol{\Theta}^T \hat{\Sigma}^{-1} \boldsymbol{\Theta} \rightarrow \chi^2_k$.

Por tanto, teniendo en cuenta que: i) el vector $\hat{\beta}$ de estimaciones MV sigue una distribución asintótica normal con media β y cuya matriz de covarianzas puede estimarse con la inversa de la matriz de información $I(\hat{\beta}) \equiv \hat{I}$ en **[1.3.5]**, y ii) el procedimiento MV garantiza que plim $I(\hat{\beta})^{-1} = I(\beta)^{-1}$, se tiene que:

$$(\hat{\beta} - \beta)^T \hat{I} (\hat{\beta} - \beta) \rightarrow \chi^2_k \qquad \textbf{[3.4.1]}$$

Entonces, denotando por $\hat{\beta}_{(i)}$ la estimación 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:

$$\hat{c}_i = (\hat{\beta} - \hat{\beta}_{(i)})^T \hat{I} (\hat{\beta} - \hat{\beta}_{(i)}) \qquad i = 1,\ldots,n \qquad \textbf{[3.4.2]}$$

Este estadístico, que es semejante al que aparece en **[3.2.27]**, 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 $\hat{c}_i$ 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 $\hat{c}_i = 0.57$ y $k = 2$, se puede decir que la eliminación de la observación i-ésima desplaza la estimación de β hasta el borde de la región de confianza de nivel 25% centrada en $\hat{\beta}$. 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, 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 el análisis 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.

Para evaluar eficientemente el estadístico de **[3.4.2]**, es necesario emplear las expresiones desarrolladas en el **Apartado 3.2.2.A** para el estimador por mínimos cuadrados ordinarios cuando se elimina una observación de la muestra. En este caso, dichas expresiones deben aplicarse al modelo linealizado de **[1.3.10]**. Así, dada una estimación MV de β en la iteración τ y una vez transformadas las variables, resulta:

$$\hat{\beta}_{(i)} = \hat{\beta} + (\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i[1 - \tilde{x}_i(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i]^{-1}(\tilde{x}_i^T\hat{\beta} - \tilde{y}_i) \qquad \textbf{[3.4.3]}$$

donde, tal y como se ha definido en la **Sección 1.3**:

$$\tilde{y}_i = \frac{y_i + \hat{f}_i(x_i^T\hat{\beta}^r) - \hat{F}_i}{[\hat{F}_i(1-\hat{F}_i)]^{1/2}} \qquad \textbf{[3.4.4]}$$

$$\tilde{x}_i = \frac{\hat{f}_i}{[\hat{F}_i(1-\hat{F}_i)]^{1/2}}x_i \qquad \textbf{[3.4.5]}$$

$$\hat{F}_i = F(x_i^T\hat{\beta}^r) \qquad \hat{f}_i = f(x_i^T\hat{\beta}^r) \qquad \textbf{[3.4.6]}$$

siendo $\tilde{X}$ una matriz cuyas filas vienen dadas por la expresión **[3.4.5]**.

Sustituyendo **[3.4.3]** en **[3.4.2]** y teniendo en cuenta que:

$$\hat{I} = \tilde{X}^T\tilde{X}$$

$$e_i^* = \frac{e_i}{[\hat{P}_i(1-\hat{P}_i)]^{1/2}} = \tilde{y}_i - \tilde{x}_i^T\hat{\beta} \qquad \textbf{[3.4.7]}$$

donde e_i^* es el residuo estandarizado de un modelo de elección binaria, el estadístico **[3.4.2]** puede escribirse:

$$\hat{c}_i = \frac{e_i^{*2}\tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i}{[1 - \tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i]^2} = \frac{e_i^{*2}\tilde{h}_i}{(1-\tilde{h}_i)^2} \qquad \textbf{[3.4.8]}$$

que es una expresión similar a **[3.2.28]**, donde, en general, $\tilde{h}_{ij} = \tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_j^T$ y $\tilde{h}_{ii} \equiv \tilde{h}_i$.

A pesar de la similitud con los estadísticos de influencia derivados para el modelo lineal general en la **Sección 3.2**, el estadístico $\hat{c}_i$ presenta un conjunto de particularidades que pueden resumirse en tres puntos:

- En primer lugar, el efecto de la i-ésima observación no ha sido completamente eliminado al utilizar la expresión **[3.4.3]**, puesto que las variables han sido transformadas con información que depende de dicha observación. Así, aunque

esta información no tenga un efecto importante, tampoco el resultado es idéntico al de **[3.2.18]**. En este caso, la expresión **[3.4.3]** puede interpretarse como un paso por el algoritmo de *scoring* en el que no se utiliza información de la observación i-ésima. Es posible eliminar completamente el efecto de la observación en cuestión iterando hasta la convergencia, aunque en este caso, el coste de cálculo del estadístico puede ser demasiado elevado.

- Para el modelo lineal general, se planteaba una discusión sobre qué residuos utilizar en los estadísticos. Como se veía, numerosos autores han sugerido el uso del residuo estudentizado. Aunque en este caso también es posible definir los residuos previstos y estandarizarlos, la naturaleza diferente del modo en que se interpretan los residuos hace que la discusión sea bastante menos fructífera. Por un lado, debido a que cada residuo tiene distinta varianza y, por otro, debido a que la evaluación de las funciones $F(\cdot)$ y $f(\cdot)$ para $\hat{\beta}_{(i)}$ puede ser excesivamente costosa.

- También para el modelo lineal se argumentaba que, en algunos casos, es preferible emplear estimadores de la desviación típica residual que no incluyesen el residuo i-ésimo. En los MEB, esta discusión es improcedente, puesto que la varianza de las perturbaciones está predeterminada, debido a las restricciones de identificación expuestas en la **Sección 1.2**.

Es importante señalar que un estadístico similar al de la ecuación **[3.4.8]** se propone en Pregibon (1981). En este caso, para su derivación se utiliza el algoritmo de Newton, en lugar del algoritmo de *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 **[1.3.4]** se reduce a: $-\Sigma\Lambda_i(1-\Lambda_i)\boldsymbol{x}_i^T\boldsymbol{x}_i$, siendo Λ_i la función de distribución logística evaluada en $\boldsymbol{x}_i^T\boldsymbol{\beta}$, el estadístico análogo a $\hat{c}_i$ resultante tiene una expresión sencilla. Sin embargo, esto no ocurre, por ejemplo, con el modelo probit, donde la utilización del hessiano de **[1.3.4]**, complicaría innecesariamente la expresión del estadístico. En este sentido, la expresión en **[3.4.8]** es considerablemente 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 $\hat{c}_i$.

Otra ventaja respecto al planteamiento de Pregibon (1981) es la mayor simplicidad de cálculo en el procedimiento lineal iterativo que se deriva del empleo del algoritmo de *scoring* frente al de Newton. Green (1984) plantea un esquema de estimación basado en

el algoritmo de *scoring* para aplicarlo a la familia de modelos lineales generalizados[12], aunque no explota las características especiales de los modelos binarios.

En el **Apartado 3.2.2.B**, al plantear medidas de influencia para una observación, se ha comprobado que, en el MLG utilizando el estadístico c_i, es idéntico medir la influencia para el vector de parámetros estimados que para el vector $\hat{y}$. Para los modelos binarios esto no es cierto, por lo que es necesario derivar un estadístico específico que mida la influencia sobre $\hat{P}$, el vector de probabilidades estimadas. Utilizando la expansión de Taylor de primer orden en **[1.3.15]** para el caso en que se elimina la i-ésima observación, la probabilidad estimada se puede aproximar de la siguiente forma:

$$F(x_i^T\hat{\beta}_{(i)}) \approx F(x_i^T\hat{\beta}) + f(x_i^T\hat{\beta})x_i^T(\hat{\beta}_{(i)} - \hat{\beta}) \qquad \textbf{[3.4.9]}$$

A partir de **[3.4.9]**, la diferencia en la probabilidad estimada para la observación j al eliminar la observación i-ésima puede calcularse como:

$$F(x_j^T\hat{\beta}_{(i)}) \approx F(x_j^T\hat{\beta}) + f(x_j^T\hat{\beta})x_j^T(\hat{\beta}_{(i)} - \hat{\beta}) \qquad \textbf{[3.4.10]}$$

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

$$\hat{F} - \hat{F}_{(i)} = -\hat{\psi}X(\hat{\beta}_{(i)} - \hat{\beta}) = -\hat{\Psi}^{1/2}\tilde{X}(\hat{\beta}_{(i)} - \hat{\beta}) \qquad \textbf{[3.4.11]}$$

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

$$\begin{aligned} \hat{C}_i(P) &= (\hat{F} - \hat{F}_{(i)})^T(\hat{F} - \hat{F}_{(i)}) \\ &= \frac{e_i^{2*}\bar{h}_i}{(1-\bar{h}_i)^2} \end{aligned} \qquad \textbf{[3.4.12]}$$

donde, en general:

$$\bar{h}_{ij} = \tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}(\tilde{X}\hat{\Psi}\tilde{X})(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_j^T \qquad \text{y} \qquad \bar{h}_i \equiv \bar{h}_{ii} \qquad \textbf{[3.4.13]}$$

---

[12] Green (1984) denomina a este algoritmo *Mínimos Cuadrados Ponderados Iterativos*. Estrictamente, sugiere la palabra *reponderados* (*reweighted*), pero parece que esa noción ya se encuentra implícita en la palabra *iterativos*.

La diferencia entre el estadístico **[3.4.12]** y **[3.4.8]** se encuentra en la matriz $\hat{\Psi}$, y la interpretación es que, en un modelo no lineal, el cambio en los argumentos de la función no tiene por qué ser igual que la variación en la función. Más concretamente, lo que indica la expresión **[3.4.13]** 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 **[3.4.14]** es especialmente importante en situaciones en las que el objetivo sea realizar previsiones agregadas para poblaciones grandes.

La conclusión de este apartado es que el conjunto de particularidades reseñadas, tanto referentes a los estadísticos que miden el efecto de observaciones individuales, como sobre el comportamiento general de los MEB en lo referente a observaciones anómalas, hacen que la extrapolación de resultados para modelos lineales o modelos lineales generalizados sea claramente insuficiente para arrojar alguna luz sobre el problema que nos ocupa. La principal deficiencia de la mayoría de la literatura existente sobre el problema de las observaciones anómalas en los MEB, es que trata del mismo modo todo el conjunto de modelos lineales generalizados, sin considerar los casos especiales, y esta situación es la que hace más difícil la aplicación de los métodos a los datos objeto de análisis.

3.4.2. Estadísticos de influencia: grupos de observaciones y otros casos particulares

Como se ha expuesto en el **Apartado 3.4.1**, a partir de la linealización del modelo binario **[1.3.10]**, se pueden derivar estadísticos de influencia individual semejantes a los desarrollados para el MLG en la **Sección 3.2**. 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 los parámetros del modelo.

Una iteración por el algoritmo de *scoring* eliminando la información de las observaciones del conjunto I puede plantearse como:

$$\hat{\beta}_{(I)} = \hat{\beta} + (\tilde{X}^T\tilde{X})^{-1}X_I\left[I_p - \tilde{X}_I(\tilde{X}^T\tilde{X})^{-1}\tilde{X}_I^T\right]^{-1}(\tilde{X}_I\hat{\beta} - \tilde{y}_I) \qquad \textbf{[3.4.14]}$$

donde $\tilde{X}$ e $\tilde{y}$ están formadas por las observaciones transformadas como en **[3.4.4]**-**[3.4.5]**. La inversa de la matriz de información resulta:

$$(\tilde{X}_{(I)}^T\tilde{X}_{(I)})^{-1} = (\tilde{X}^T\tilde{X})^{-1} + (\tilde{X}^T\tilde{X})^{-1}\tilde{X}_I^T\left[I_p - \tilde{X}_I(\tilde{X}^T\tilde{X})^{-1}\tilde{X}_I^T\right]^{-1}\tilde{X}_I(\tilde{X}^T\tilde{X})^{-1} \qquad \textbf{[3.4.15]}$$

En las expresiones **[3.4.14]** y **[3.4.15]**, no se ha eliminado completamente el efecto del conjunto de observaciones en I y se podría iterar hasta la convergencia. No obstante, si las observaciones eliminadas no son influyentes, dicho proceso iterativo no introduciría variaciones importantes. Por el contrario, si el efecto del conjunto es relevante, la iteración posterior reforzaría el efecto de anomalía de la observación. Dado que para muestras de gran tamaño, las consideraciones de tiempo de cálculo son importantes, no resulta necesario continuar el proceso iterativo, excepto, tal vez, en situaciones dudosas. De hecho, no iterar hace más robusto el estadístico ante la posibilidad de rechazar observaciones no anómalas aunque, en cambio, reduce la probabilidad de detectar una observación anómala.

A partir de **[3.4.14]**, la influencia de un conjunto de observaciones puede evaluarse utilizando un estadístico similar a **[3.2.35]**, que puede plantearse:

$$\begin{aligned}\hat{c}_I &= (\hat{\beta}_{(I)} - \hat{\beta})(\tilde{X}^T\tilde{X})(\hat{\beta}_{(I)} - \hat{\beta}) \\ &= e_I^{*T}(I_p - \tilde{H}_I)^{-1}\tilde{H}_I(I_p - \tilde{H}_I)^{-1}e_I^*\end{aligned} \qquad \textbf{[3.4.16]}$$

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

También en los MEB se puede definir una matriz de influencia semejante a la matriz M en **[3.2.39]**, que puede emplearse para evaluar de forma aproximada el estadístico **[3.4.16]**, así como para detectar observaciones que presentan problemas de enmascaramiento. En este caso, el elemento genérico resulta:

$$\tilde{m}_{ij} = \frac{e_i^* e_j^* \tilde{h}_{ij}}{(1 - \tilde{h}_{ii})(1 - \tilde{h}_{jj})} \qquad \textbf{[3.4.17]}$$

que debe interpretarse en términos de los valores ajustados de la variable endógena $\tilde{y}_i$ en el MEB linealizado. Nótese que la diagonal principal de $\tilde{M}$ está formada por el estadístico de **[3.4.8]** para cada observación.

De manera análoga a como se planteó para el estadístico **[3.4.8]**, es posible obtener expresiones alternativas para **[3.4.16]** y **[3.4.17]** basadas en medir el cambio en las probabilidades estimadas, en lugar de evaluar la diferencia entre los vectores de parámetros. En particular, la matriz de influencia alternativa tiene como elemento genérico:

$$\bar{m}_{ij} = \frac{e_i e_j \bar{h}_{ij}}{(1 - \bar{h}_{ii})(1 - \bar{h}_{jj})} \qquad \textbf{[3.4.19]}$$

donde $\bar{h}_{ij}$ está definido en **[3.4.14]**.

Como se puso de relieve en la **Sección 3.2**, es conveniente derivar expresiones particulares para evaluar la influencia de grupos de observaciones sobre un subconjunto de los 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 $m \times k$ de constantes conocidas, con rango$(R) = m \leq k$. 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 en **[3.4.16]** aplicada a una combinación lineal de los parámetros originales es:

$$\begin{aligned} \hat{c}_I(\theta) &= (\hat{\theta}_{(I)} - \hat{\theta})^T \left[R^T(\tilde{X}^T\tilde{X})^{-1}R\right]^{-1}(\hat{\theta}_{(I)} - \hat{\theta}) \\ &= e_I^{*T}(I_p - \tilde{H}_I)^{-1}\tilde{X}_I \tilde{N} \tilde{X}_I^T (I_p - \tilde{H}_I)^{-1} e_I^* \end{aligned} \qquad \textbf{[3.4.20]}$$

donde:

$$\tilde{N} = (\tilde{X}^T\tilde{X})^{-1}R\left[R^T(\tilde{X}^T\tilde{X})^{-1}R\right]^{-1}R^T(\tilde{X}^T\tilde{X})^{-1} \qquad \textbf{[3.4.21]}$$

En particular, el efecto de eliminar una sola observación sobre θ puede escribirse:

$$\hat{c}_i(\theta) = \frac{e_i^{*2}\tilde{x}_i^T \tilde{N} \tilde{x}_i}{(1 - \tilde{h}_i)^2} \qquad \textbf{[3.4.22]}$$

que es sencillo de calcular puesto que $\tilde{N}$ no depende de la observación eliminada.

En los MEB es frecuente la presencia de un cierto número de variables cualitativas entre las variables exógenas, por lo que, en muchas ocasiones es interesante 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:

$$R^T\beta = (0_{m\times(k-m)} \mid I_m)\beta = (\beta_{k-m+1}, \ldots, \beta_k)^T \qquad \textbf{[3.4.23]}$$

El estadístico **[3.4.20]** puede escribirse:

$$\hat{c}_I(\boldsymbol{\theta}) = \hat{c}_I - \boldsymbol{e}_I^{*T}(I_p - \tilde{H}_I)^{-1}\tilde{G}_I(I_p - \tilde{H}_I)^{-1}\boldsymbol{e}_I^{*} \qquad \textbf{[3.4.24]}$$

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:

$$\hat{c}_i(\boldsymbol{\theta}) = \frac{e_i^{*2}(\tilde{h}_i - \tilde{g}_i)}{(1 - \tilde{h}_i)^2} \qquad \textbf{[3.4.25]}$$

donde $\tilde{g}_i$, es el elemento i-ésimo de la diagonal principal de $\tilde{G}_I$.

El estadístico **[3.4.8]** 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 $\boldsymbol{\beta}$. Denotando por v_j el elemento j-ésimo de la diagonal principal de la matriz $I(\hat{\boldsymbol{\beta}})^{-1}$, es inmediato que el estadístico:

$$\hat{c}_i(\beta_j) = \frac{(\hat{\beta}_{(i)j} - \hat{\beta}_j)^2}{v_j} \qquad \textbf{[3.4.26]}$$

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

3.4.3. Detección de observaciones influyentes en MEB

A partir de la exposición de las **Secciones 3.2** y **3.3**, así como los estadísticos planteados en la **Sección 3.4**, se puede desarrollar una estrategia de diagnosis de observaciones anómalas para los modelos de elección binaria, que puede plantearse de la siguiente forma.

En una primera etapa, es necesario utilizar los instrumentos *a priori* del **Apartado 3.2.2.A**, así como el estadístico h_i en **[3.2.2]** para evaluar la dispersión de las observaciones muestrales, lo que proporciona información sobre potenciales observaciones extremas. Alternativamente, puede usarse el *estadístico de Wilks*. 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 $\tilde{h}_i = \tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i$ que utiliza la 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, lógicamente, calcular el estadístico $\hat{c}_i$ definido en **[3.4.8]**. Valores elevados de $\hat{c}_i$ indicarán la presencia de posibles observaciones anómalas. En este sentido, la principal dificultad estriba en que no existen formas concluyentes de evaluar valores críticos para dicho estadístico. No obstante, sí se puede obtener valores indicativos utilizando las tablas de la distribución χ^2_k. 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 eliminando 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 el procedimiento de Peña y Yohai (1991) descrito en el **Apartado 3.2.3**. Una vez más, no se pueden ofrecer valores críticos concluyentes y, en parte, depende del criterio del analista la decisión sobre el efecto de cada grupo, medido a través del estadístico $\hat{c}_I$ en **[3.4.16]**, así como las posibles causas de la presencia de observaciones anómalas para tomar las decisiones finales sobre los grupos de observaciones problemáticas.

Otros estadísticos que se desarrollaron para el modelo lineal no tienen especial interés aquí. Por ejemplo, en el **Apartado 3.2.2** se puso de relieve la importancia de analizar el efecto sobre la matriz de varianzas de los parámetros, para lo cual se sugería emplear el estadístico *COVRATIO* en **[3.2.33]**. Para los MEB, se puede demostrar que dicho estadístico resulta $1/(1 - \tilde{h}_i)$ y, por tanto, proporciona la misma información que se obtiene al analizar $\tilde{h}_i$.

Por último, es importante señalar que el conjunto de instrumentos de detección planteado anteriormente revela la presencia de observaciones influyentes en la muestra, aunque ninguno de ellos hace indicación del posible origen de las mismas. Una forma de obtener alguna información al respecto, es utilizando los contrastes desarrollados en el **Apartado 2.3.2**.

Los contrastes mencionados pueden utilizarse para, una vez detectadas observaciones influyentes, determinar la fuente de anomalía, suponiendo que ésta sea única. No obstante, en los experimentos de Monte Carlo realizados, no resultaron concluyentes en ningún caso sobre el tipo de anomalía, 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.

3.5. Resultados con datos simulados

En esta sección se pretende ilustrar el funcionamiento de los estadísticos propuestos en la **Sección 3.4** en algunos aspectos concretos. En particular, se evaluan tres elementos de la metodología planteada utilizando los modelos y los casos de anomalías de la **Sección 2.4**. En primer lugar, se plantea si el criterio de selección de puntos de corte para $\hat{c}_i$ basado en comparar el cociente entre el estadístico y su valor medio para la muestra permite diferenciar las observaciones anómalas. En segundo lugar, se analiza el punto crítico que hace mínimo el *ECM* de las estimaciones definido en **[2.4.4]** en comparación con los valores críticos indicativos que se obtendrían de una distribución χ^2_k. Por último, se analiza la detección de observaciones que presentan efecto *enmascaramiento* utilizando el método de Peña y Yohai (1991) aplicado a la matriz $\tilde{M}$ definida en **[3.4.17]**. Sólo se presentan resultados para el modelo probit puesto que, de forma similar a como ocurría en la **Sección 2.4**, los resultados para el modelo logit son casi iguales a los del probit.

Los modelos utilizados, los casos analizados y los aspectos técnicos relevantes de esta sección ya fueron descritos en la **Sección 2.4**. El modelo considerado es un MEB con una sola variable explicativa y término constante:

$$y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \textbf{[3.5.1]}$$

Las observaciones se han generado mediante el siguiente mecanismo:

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* \geq 0 \\ 0 & \text{si} \quad y_i^* < 0 \end{cases} \qquad \textbf{[3.5.2]}$$

La variable explicativa se ha generado, en todos los casos de observaciones no anómalas, como una normal, $x_i \sim iid\ N(0,1)$ y el vector de parámetros es: $\beta = (-0.65,\ 1)^T$. Los casos de observaciones anómalas considerados son:

Caso 1: Un porcentaje ω de observaciones y_i^* en la muestra proviene de una distribución con la misma media que las restantes observaciones, pero con $h^2 = 7$.

Caso 2: Un porcentaje ω de observaciones y_i^* proviene de la misma distribución con varianza igual a σ_0^2, pero con distinta media que las restantes observaciones. Una proporción ω de observaciones son tales que $E(y_i^*) = x_i^T\gamma$, donde $\gamma = (1,\ -0.5)^T$.

Caso 3: Un porcentaje ω de observaciones están caracterizadas por $E(y_i^*) = x_i^T\delta$, e idéntica varianza que las demás, donde $\delta = (-0.5, 0.5)^T$ y, además, para estas observaciones, $x_i \sim iid\ N(5,1)$.

El propósito de la primera simulación es comprobar hasta qué punto las observaciones anómalas pueden diferenciarse de las no anómalas a partir del estadístico **[3.4.8]**. En las **Tablas 3.1-3.3** aparecen valores medios, para 500 replicaciones, de los siguientes estadísticos: la media del estadístico $\hat{c}_i$ para todas las observaciones de la muestra, la media del estadístico $\hat{c}_i$ para las observaciones no anómalas ($\hat{c}_i(B)$), la media del estadístico $\hat{c}_i$ para las observaciones anómalas ($\hat{c}_i(M)$) y el ratio $\hat{c}_i(M)/\hat{c}_i(B)$. La primera columna de cada tabla indica el porcentaje de observaciones anómalas presente en las muestras.

A la vista de las citadas tablas, cabe hacer los siguientes comentarios:

- El valor del estadístico de influencia resulta muy homogéneo con independencia del nivel de anomalías presentes en la muestra, tomando un valor de aproximadamente k/n. Por el contrario, la media de $\hat{c}_i$ para las observaciones *buenas* disminuye, indicando que estas observaciones tienen menos efecto en la estimación a medida que aumenta ω. A medida que crece el número de observaciones anómalas, el valor del estadístico disminuye, siendo un claro indicativo de que se produce un efecto de enmascaramiento: cuantas más observaciones anómalas se encuentran presentes, menor será la influencia de cada una de ellas por separado. No obstante, el valor medio del estadístico para estas observaciones se encuentra siempre por encima de la media para las no anómalas.

- Teniendo en cuenta la última columna de cada tabla, un punto crítico mínimo para el estadístico de influencia se encontraría, aproximadamente, entre dos y cinco veces el valor medio para la muestra, que en estos casos, estaría entre 0.02 y 0.05. El valor crítico de la distribución χ^2_k con dos grados de libertad a un nivel de confianza del 10% es, aproximadamente, 0.2.

- A la hora de aplicar este criterio, debe tenerse en cuenta que será más válido cuantas menos observaciones influyentes se encuentren en la muestra, ya que es una medida de influencia individual. Sin embargo, aparecen problemas de enmascaramiento no será tan útil. Por otra parte, también es necesario considerar el número de observaciones que sobrepasan el nivel crítico: si el número es excesivo, habrá que elegir como influyentes aquellas observaciones

con valor más elevado puesto que, si se eliminan observaciones no anómalas, se está eliminando información relevante, produciendo sesgos importantes en la estimación.

Tabla 3.1. Valores medios del estadístico $\hat{c}_i$: modelo probit, caso 1.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(B)/\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0101	0.0101	--	--	--
2.5	0.0102	0.0098	0.0262	2.68	2.57
5.0	0.0104	0.0094	0.0289	3.07	2.78
7.5	0.0104	0.0090	0.0276	3.07	2.66
10.0	0.0104	0.0087	0.0255	2.92	2.45
12.5	0.0105	0.0085	0.0249	2.93	2.36
15.0	0.0106	0.0084	0.0232	2.78	2.19
17.5	0.0106	0.0082	0.0220	2.69	2.07
20.0	0.0106	0.0080	0.0210	2.63	1.98
22.5	0.0106	0.0078	0.0204	2.63	1.92
25.0	0.0107	0.0077	0.0198	2.59	1.85
27.5	0.0106	0.0076	0.0184	2.42	1.74
30.0	0.0106	0.0075	0.0181	2.42	1.70

Tabla 3.2. Valores medios del estadístico $\hat{c}_i$: modelo probit, caso 2.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(B)/\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0101	0.0101	--	--	--
2.5	0.0107	0.0088	0.0862	9.79	8.03
5.0	0.0110	0.0080	0.0669	8.34	6.10
7.5	0.0112	0.0075	0.0578	7.75	5.15
10.0	0.0113	0.0072	0.0479	6.66	4.25
12.5	0.0112	0.0069	0.0412	5.94	3.67
15.0	0.0112	0.0068	0.0357	5.22	3.20
17.5	0.0111	0.0068	0.0313	4.60	2.82
20.0	0.0110	0.0068	0.0280	4.13	2.54
22.5	0.0109	0.0068	0.0251	3.66	2.29
25.0	0.0108	0.0069	0.0226	3.27	2.09
27.5	0.0108	0.0070	0.0208	2.96	1.92
30.0	0.0107	0.0072	0.0188	2.62	1.76

Tabla 3.3. Valores medios del estadístico $\hat{c}_i$: modelo probit, caso 3.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(B)/\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0102	0.0102	--	--	--
2.5	0.0118	0.0101	0.0757	7.49	6.44
5.0	0.0135	0.0098	0.0842	8.57	6.22
7.5	0.0146	0.0097	0.0748	7.73	5.14
10.0	0.0149	0.0095	0.0638	6.73	4.28
12.5	0.0154	0.0095	0.0562	5.89	3.66
15.0	0.0160	0.0093	0.0539	5.78	3.37
17.5	0.0158	0.0093	0.0465	5.02	2.95
20.0	0.0158	0.0093	0.0420	4.53	2.65
22.5	0.0165	0.0090	0.0423	4.71	2.57
25.0	0.0158	0.0089	0.0368	4.14	2.32
27.5	0.0155	0.0090	0.0329	3.69	2.12
30.0	0.0149	0.0089	0.0291	3.29	1.95

El siguiente grupo de simulaciones también se ha diseñado para evaluar criterios de selección óptima de puntos de corte del estadístico de influencia $\hat{c}_i$. En este caso, se ha investigado el valor crítico tal que, eliminando aquellas observaciones con $\hat{c}_i > \hat{c}_i^*$, el error cuadrático medio de la estimación definido en **[2.4.4]** sea mínimo.

Para cada una de las muestras generadas se calcula $\hat{c}_i$ y se realiza una búsqueda por el *método de la rejilla* para determinar el valor de $\hat{c}_i$ que minimiza el ECM de las estimaciones sin las observaciones que superan el valor crítico. En las **Tablas 3.4-3.6** se presentan los resultados para el modelo probit. En las columnas aparecen, para cada valor de ω, las medias de las 500 replicaciones de los siguientes estadísticos: el valor crítico de $\hat{c}_i$ que minimiza el ECM de las estimaciones ($\hat{c}_i^*$), el ECM una vez eliminadas las observaciones cuyo valor de $\hat{c}_i$ supera $\hat{c}_i^*$ y entre paréntesis, el ECM sin eliminar observaciones (tomado de las tablas de la **Sección 2.4**), el número de observaciones detectadas como anómalas y, en la última columna, el número de observaciones realmente anómalas entre las detectadas. Teniendo en cuenta los resultados obtenidos, se pueden hacer los siguientes comentarios:

- El principal resultado es que, como puede apreciarse, el punto de corte óptimo es muy sensible al tipo de anomalía que aparece en la muestra. La conclusión de esto es que, con datos reales, no es posible determinar un punto de corte de aplicación general, por lo que es necesaria la intervención del analista que deberá probar diferentes puntos críticos considerando el número de observaciones eliminadas así como los efectos individuales y los conjuntos.

- Los peores resultados en cuanto al número de observaciones anómalas detectadas, se producen en el caso 3 (**Tabla 3.6**). Esto es debido a que estas observaciones anómalas se deben a valores extremos de la variable explicativa, y el modo de llevar a cabo la detección es utilizando los estadísticos *a priori* en **[3.2.2]** y el estadístico $\tilde{h}_i$. No obstante, el ECM, una vez eliminadas las observaciones con $\hat{c}_i > \hat{c}_i^*$ mejora apreciablemente en todos los casos.

- El enmascaramiento hace que el valor crítico del estadístico disminuya a medida que aumenta el número de observaciones anómalas, de modo que se eliminan más observaciones, aunque la proporción de observaciones anómalas eliminadas se mantiene, aproximadamente, constante.

Tabla 3.4. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el estadístico $\hat{c}_i$: modelo probit, caso 1.

$\omega\%$	$\hat{c}_i^*$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	0.2665	0.0298 (0.034)	0.85	--
5.0	0.2257	0.0192 (0.037)	1.93	0.40
10.0	0.1864	0.0175 (0.043)	2.93	1.03
15.0	0.1556	0.0159 (0.056)	4.11	1.81
20.0	0.1174	0.0141 (0.070)	5.97	3.10
25.0	0.0871	0.0155 (0.094)	7.74	4.31
30.0	0.0692	0.0162 (0.120)	9.80	5.89

Tabla 3.5. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el estadístico $\hat{c}_i$: modelo probit, caso 2.

$\omega\%$	$\hat{c}_i^*$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	0.2439	0.0294 (0.035)	1.07	--
5.0	0.1175	0.0155 (0.077)	5.66	2.29
10.0	0.0502	0.0290 (0.180)	12.40	6.32
15.0	0.0252	0.0543 (0.307)	21.06	11.27
20.0	0.0166	0.0969 (0.443)	27.40	15.91
25.0	0.0142	0.1615 (0.574)	30.48	19.18
30.0	0.0129	0.2531 (0.723)	32.91	21.32

Tabla 3.6. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el estadístico $\hat{c}_i$: modelo probit, caso 3.

$\omega\%$	$\hat{c}_i^*$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	0.2438	0.0298 (0.038)	1.10	--
5.0	0.3778	0.0250 (0.050)	1.76	0.31
10.0	0.3465	0.0229 (0.069)	2.49	0.66
15.0	0.3016	0.0203 (0.080)	3.66	0.97
20.0	0.2650	0.0256 (0.100)	5.02	1.41
25.0	0.3001	0.0325 (0.115)	5.58	1.71
30.0	0.1741	0.0372 (0.139)	6.50	2.15

El último conjunto de simulaciones trata de ilustrar la detección de observaciones anómalas cuando existe un problema de enmascaramiento, esto es, cuando la presencia de un cierto número de observaciones anómalas hace que los efectos individuales de cada una de ellas queden *disimulados*. Para ello, se utiliza el método de Peña y Yohai (1991) aplicado a la matriz de influencia definida en **[3.4.17]**.

La principal dificultad a la hora de instrumentar el método, es que requiere la intervención del investigador para determinar los grupos de observaciones potencialmente influyentes, lo que era inapropiado para realizar un volumen importante de simulaciones. Por ello, se recurrió a la aplicación de la técnica de forma *no estricta*, y se considera potencialmente anómala cualquier observación tal que su componente asociado a cualquiera de los dos autovectores no nulos fuese superior a 0.15. Naturalmente, esta forma de utilización ofrece resultados bastante peores de los que se obtendrían aplicándola correctamente a una muestra real, pero es suficiente para ilustrar el buen funcionamiento del mecanismo cuando se aplica a los MEB.

En las **Tablas 3.7-3.9** se presentan las medias, para 500 replicaciones y para distintos valores de ω, de los siguientes estadísticos: el valor del estadístico de influencia $\hat{c}_I$ para el grupo de observaciones tales que algún componente asociado de los dos autovectores principales es superior a 0.15, el ECM una vez eliminadas dichas observaciones y, entre paréntesis, el ECM sin eliminarlas, el número de observaciones eliminadas y, por último, el número de observaciones anómalas entre las eliminadas. A la vista de los resultados cabe hacer los siguientes comentarios:

- Se puede afirmar que el método, aún utilizado inadecuadamente, funciona bien en situaciones en las que efectivamente se produce enmascaramiento; esto es, cuando ω es elevado. En particular, en los casos 1 y 2, el ECM se reduce a la mitad para los valores más elevados de la proporción de anomalías (25% y 30%).

- Cuando la proporción de anomalías es pequeña o nula, el mal uso del método hace eliminar observaciones no anómalas, provocando un ECM superior al que se obtendría manteniéndolas en la muestra.

- Dado que los autovectores están normalizados, sistemáticamente se elimina un número de observaciones (aproximadamente 12), con independencia de que sean o no anómalas. El problema es que, cuando son observaciones *buenas*, estos

datos están en las colas, y se produce el efecto, ya comentado, de pérdida de información relevante.

Tabla 3.7. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el procedimiento de Peña y Yohai (1992): modelo probit, caso 1.

$\omega\%$	$\hat{c}_I$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	10.55	0.3463 (0.034)	14.43	--
5.0	10.47	0.2552 (0.037)	13.73	1.68
10.0	10.87	0.2074 (0.043)	13.58	3.29
15.0	10.57	0.1274 (0.056)	13.28	4.55
20.0	11.06	0.0922 (0.070)	13.44	5.98
25.0	11.16	0.0869 (0.094)	13.68	6.36
30.0	10.55	0.0672 (0.120)	13.02	7.36

Tabla 3.8. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el procedimiento de Peña y Yohai (1992): modelo probit, caso 2.

$\omega\%$	$\hat{c}_I$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	10.70	0.3577 (0.035)	14.34	--
5.0	11.05	0.1095 (0.077)	12.89	3.77
10.0	11.03	0.0673 (0.180)	11.80	6.12
15.0	11.10	0.1052 (0.307)	12.64	7.55
20.0	11.93	0.1819 (0.443)	13.20	9.15
25.0	10.13	0.2895 (0.574)	11.35	8.86
30.0	9.96	0.3948 (0.723)	12.46	10.03

Tabla 3.9. Error cuadrático medio de las estimaciones MV eliminando las anomalías de la muestra utilizando el procedimiento de Peña y Yohai (1992): modelo probit, caso 3.

$\omega\%$	$\hat{c}_I$	$ECM_{(I)}$ (ECM)	N° OBS	N° ANOBS
0.0	10.57	0.3660 (0.038)	13.90	--
5.0	11.27	0.2613 (0.050)	12.86	0.36
10.0	11.72	0.2168 (0.069)	13.21	0.75
15.0	11.00	0.2177 (0.080)	12.77	0.97
20.0	10.86	0.1990 (0.100)	11.77	1.31
25.0	10.70	0.1626 (0.115)	12.48	1.63
30.0	11.32	0.1458 (0.139)	12.77	2.44

CAPÍTULO 4

OBSERVACIONES ANÓMALAS EN MODELOS DE ELECCIÓN MÚLTIPLE

4.1. Introducción

En este capítulo se generalizan los resultados de los **Capítulos 2** y **3** para los modelos de elección cualitativa múltiple (MEM) más utilizados en la práctica: el modelo logit multinomial y el modelo probit multinomial. Como se ilustra más adelante, tanto el planteamiento del problema de observaciones anómalas, como los principales desarrollos para su tratamiento son análogos al de los modelos de elección binaria, por lo se utiliza gran parte de las discusiones de capítulos anteriores.

Sin entrar en consideraciones sobre hasta qué punto el planteamiento de la función de utilidad refleja adecuadamente el comportamiento individual, el empleo de los MEM en economía parte del supuesto de que el decisor elige aquella alternativa del conjunto factible que maximiza la utilidad esperada. Aunque es posible derivar estos modelos bajo otros planteamientos, en particular el modelo logit, en este capítulo se mantiene el supuesto de maximización de la utilidad.

El modelo logit, debido a la forma cerrada de las probabilidades de elección es el más utilizado en aplicaciones prácticas, aunque tiene algunas limitaciones para representar decisiones en las que puede haber una importante sustituibilidad entre alternativas. Por el contrario, el modelo probit es menos utilizado, sobre todo porque es más complejo y costoso, y sólo compensa cuando las propiedades del logit lo hacen inadecuado.

En la **Sección 4.2** se derivan los modelos de elección cualitativa múltiple a partir de la teoría de la utilidad y se analizan las restricciones necesarias para su estimación así como algunas propiedades, en particular, la de *independencia de las alternativas irrelevantes* y las limitaciones en modelizar efectos individuales.

A continuación, en la **Sección 4.3**, se plantea la estimación de estos modelos por máxima verosimilitud y se desarrolla un método máximo-verosímil por procedimientos lineales análogo al derivado para los MEB en el **Capítulo 1**. Debido a que los resultados sobre la distribución asintótica de los estimadores máximo-verosímiles son análogos a los de los MEB, no se derivan expresiones específicas para los modelos tratados en este capítulo, pues es suficiente con aplicar las de las **Secciones 1.3-1.5** utilizando la nueva notación.

En la **Sección 4.4**, se extiende el planteamiento de observaciones anómalas del **Capítulo 2** y se derivan estadísticos adecuados para detectar la presencia de observaciones anómalas en modelos de elección múltiple. El cálculo de estos estadísticos puede llevarse a cabo de una forma computacionalmente eficiente gracias al procedimiento de estimación máximo-verosímil por procedimientos lineales que se propone en la sección anterior.

Por último, en la **Sección 4.5**, se ilustran, con datos simulados, los principales resultados de las secciones anteriores aplicados al modelo logit multinomial.

4.2. Modelos de variable dependiente cualitativa múltiple

4.2.1. Planteamiento de los modelos a partir de la teoría de la utilidad

En la **Sección 1.2** se plantearon los modelos de elección binaria bajo el supuesto de la existencia de una variable dependiente no observable que indicaba la preferencia del individuo por cada una de las alternativas que se le presentaban. En esta línea, se derivan los modelos de esta sección. Un planteamiento alternativo para el modelo logit, más ligado al problema de clasificación, puede encontrarse en Maddala (1983, pags. 34-37).

Ahora se supone que el individuo i puede elegir entre un conjunto de m alternativas. De la elección de la alternativa j, el individuo obtendrá una cierta utilidad, que se denota por U_{ij}, de forma que $\boldsymbol{U}_i^T = [U_{i1}, \ldots, U_{im}]$ es el vector formado por las utilidades que el individuo i-ésimo obtendría de cada posible alternativa en caso de elegirla. Dicha utilidad dependerá de un conjunto de factores observables, entre los cuales se encuentran fundamentalmente los atributos de la opción y las características del individuo que puede medir el investigador, así como un conjunto de componentes no obsevables ε_{ij}. Denotando por $\boldsymbol{z}_{ij}$ el vector de las características relevantes de la opción j percibidas por i y por $\boldsymbol{x}_i$ el vector de las características personales relevantes del individuo i-ésimo que puede observar el investigador, en general, se puede escribir que:

$$U_{ij} = v(\boldsymbol{z}_{ij}, \boldsymbol{x}_i, \boldsymbol{\theta}) + \varepsilon_{ij} = v_{ij} + \varepsilon_{ij} \qquad \forall j \qquad \textbf{[4.2.1]}$$

donde a v_{ij} se le denomina *utilidad observable* o *representativa.* Adicionalmente, el investigador conoce (o supone) la forma funcional de v_{ij} que depende de un vector de parámetros $\boldsymbol{\theta}$. Naturalmente, el sujeto elegirá aquella opción que le reporte una utilidad mayor, es decir, elige la alternativa k, si y solo si[13]:

$$U_{ik} \geq U_{ij} \qquad \forall j \mid j \neq k \qquad \textbf{[4.2.2]}$$

En este contexto, ε_{ij} se considera un componente aleatorio no observable que no depende de $\boldsymbol{z}_{ij}$ ni de $\boldsymbol{x}_i$. Puesto que U_{ij} no es observable, no se puede prever perfectamente la elección del individuo. No obstante, sí se puede estimar una medida de la utilidad del

[13] Aunque en la comparación se utiliza el signo mayor o igual, el individuo elige una y sólo una de ellas.

individuo (v_{ij}) y con ella inferir su decisión; esto es, estimar la probabilidad de que el individuo elija la alternativa j.

Además de las variables en x_i y z_{ij}, se observa la alternativa elegida por el sujeto. Más formalmente, se define una función indicador $\mathcal{F}(U_i)$ que traduce la utilidad no observable del individuo para cada una de las alternativas en un vector y_i de dimensión $m \times 1$ con elemento genérico y_{ij}, de modo que:

$$y_i = \mathcal{F}(U_i) = \begin{cases} y_{ik} = 1 \quad si \quad U_{ik} \geq U_{ij} \quad \forall j \mid k \neq j \\ y_{ij} = 0 \quad \forall j \neq k \end{cases} \qquad \textbf{[4.2.3]}$$

y $\sum_j y_{ij} = 1$. Es decir, el individuo sólo elige una de las alternativas que se le presentan.

La función $\mathcal{F}(\cdot)$ transforma el mecanismo de decisión en un vector de ceros con un uno en la posición correspondiente a la alternativa elegida por el individuo. La elección de $\mathcal{F}(\cdot)$ implica hacer un supuesto sobre el modo en que el individuo toma la decisión, y en esta situación, distintas hipótesis pueden dar lugar a diferentes modelos o a interpretaciones alternativas de una misma forma funcional del modelo [Maddala (1983) y Ben-Akiva y Lerman (1985)].

Obsérvese que, según el planteamiento que se ha seguido hasta ahora, la elección por parte del individuo es *determinista*, es decir, la probabilidad de elegir la alternativa k es uno o cero, dependiendo de que se cumpla o no la condición **[4.2.2]**.

Para caracterizar totalmente el problema falta definir un punto importante. Sea un conjunto de individuos que se enfrentan al mismo conjunto de alternativas y que poseen idénticos valores de la parte de utilidad observable. El investigador, no obstante, observará diferentes elecciones por parte de los distintos individuos debidas al componente aleatorio de **[4.2.1]**. Entonces, definiendo y_j como el número de individuos en la muestra que han elegido la opción j-ésima, esto es: $y_j = \sum_i y_{ij}$, se puede definir la *probabilidad de elección* o *de respuesta* de la alternativa j por el individuo i, que se denota P_{ij}, como el límite de la proporción de individuos que, para idénticos valores de utilidad observable, elegirían la alternativa j cuando el número de sujetos investigados tiende a infinito. Por tanto, la probabilidad de elección se define en términos de lo que el investigador observa, y no en función del comportamiento individual.

Esta última definición es fundamental en la interpretación de lo que sigue, puesto que el sujeto *no tiene probabilidades de elegir*, el sujeto elige. Por el contrario, lo que el investigador hace es estimar la probabilidad (límite de proporción) de que el individuo prefiera una cierta alternativa, condicionado a la parte de la utilidad que puede observar.

Dada la definición anterior, la probabilidad de elección será la probabilidad de que se mantenga la condición **[4.2.2]** dados los componentes observados de la utilidad ($v_{ij} = v(z_{ij}, x_i, \theta)$), esto es:

$$\begin{aligned} P_{ik} &= P(U_{ik} \geq U_{ij},\ \forall j \mid k \neq j) \\ &= P(v_{ik} + \varepsilon_{ik} \geq v_{ij} + \varepsilon_{ij},\ \forall j \mid k \neq j) \\ &= P(\varepsilon_{ij} - \varepsilon_{ik} \leq v_{ik} - v_{ij},\ \forall j \mid k \neq j) \\ &= F_k(v_{ik} - v_{i1}, \ldots, v_{ik} - v_{im}) \end{aligned} \qquad \textbf{[4.2.4]}$$

donde $F_k(\cdot)$ es la función de distribución marginal de ε_{ij}-ε_{ik} $\forall j \neq k$. En general, $F(\cdot)$ será una función de distribución conjunta que se evalúa para m-1 alternativas y la última probabilidad se calcula teniendo en cuenta que las probabilidades de elección deben sumar la unidad. La interpretación de la expresión **[4.2.4]** es que la probabilidad de elección de la alternativa k-ésima es una función que depende de la diferencia entre las utilidades observables derivadas de la alternativa k y las restantes.

Finalmente, con respecto a la forma funcional de v_{ij}, el planteamiento más usual, siguido en este tabajo, es una especificación lineal para la utilidad representativa:

$$v_{ij} = v(\boldsymbol{z}_{ij}, \boldsymbol{x}_i, \boldsymbol{\theta}_j) = \boldsymbol{z}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j \qquad \textbf{[4.2.5]}$$

4.2.2. El modelo logit multinomial

De forma similar a como se planteó en el **Capítulo 1**, para definir totalmente el problema, bastaría con elegir una distribución adecuada para ε_{ij}-ε_{ik}, aunque es más interesante derivar alguna distribución para ε_{ij}.

Se supone una muestra de n individuos, cada uno de los cuales se enfrenta a un conjunto idéntico de m alternativas, y en la que cada uno de ellos tiene idéntica utilidad representativa v_{ij} para todas las alternativas. En este caso particular, la probabilidad de elección dada en **[4.2.4]** se reduce a:

$$P(\varepsilon_{ij} \leq \varepsilon_{ik},\ \forall j \mid k \neq j) = P(\varepsilon_{ik} = \max(\varepsilon_{i1},\ldots,\varepsilon_{ik},\ldots,\varepsilon_{im})) \qquad \textbf{[4.2.6]}$$

donde, para un m suficientemente grande, y para cualquier hipótesis sobre la distribución de ε_{ij}, $\varepsilon_{ik} = \max(\varepsilon_{i1},\ldots,\varepsilon_{im})$ se distribuirá asintóticamente de acuerdo con la *distribución del valor extremo* (EVD). Bajo hipótesis distribucionales más fuertes para ε_{ij}, por ejemplo, normalidad, la convergencia de la distribución ε_{ik} a la EVD es mucho más rápida.

Por tanto, se supone que ε_{ij} $\forall j$, se *distribuy idéntica e independientemente* (iid) *EVD*. Dada esta distribución de la componente no observable y partiendo de **[4.2.4]**, mediante un sencillo cambio de variable, se obtiene que:

$$P_{ik} = \frac{\exp(v_{ik})}{\sum_{j=1}^{m} \exp(v_{ij})} \qquad \forall k \qquad \textbf{[4.2.7]}$$

que no es más que la *función de distribución logística.*

La función en **[4.2.7]** está acotada entre cero y uno, y es continua y derivable $\forall\ v_{ij} \in \mathbb{R}$. Además, si $v_{ij} \to -\infty \Rightarrow P_{ij} \to 0$, y por el contrario, si $v_{ij} \to \infty \Rightarrow P_{ij} \to 1$; es decir, cuanta menos (más) utilidad representativa obtenga el individuo i de la alternativa j, menor (mayor) probabilidad de elección tendrá dicha alternativa. Por último, es obvio que **[4.2.7]** también cumple que $\Sigma_j P_{ij} = 1$.

De forma análoga a como se expuso para el modelo logit binario en el **Apartado 1.2.2.C**, transformaciones adecuadas de v_{ij}, representadas por $G(v_{ij})$, permiten que la forma funcional dada en **[4.2.7]** se aproxime arbitrariamente a otras funciones de distribución. Esto es lo que se denomina el modelo *logit universal.*

4.2.3. El modelo probit multinomial

En el modelo logit del **Apartado 4.2.2** se ha supuesto que los componentes no observables de la utilidad ε_{ij} se distribuyen iid EVD. En esta sección se modifica dicha hipótesis manteniendo los restantes elementos del modelo. Ahora se supone que los componentes ε_{ij} de $\boldsymbol{\varepsilon}_i$ siguen una distribución normal multivariante con vector de esperanzas $\mathbf{0}_m$. Esto es:

$$\varepsilon_i \sim N(\mathbf{0}_m, \Sigma_i) \qquad [4.2.8]$$

con una función de densidad conjunta:

$$f(\varepsilon_i) = \phi(\varepsilon_i) = (2\pi)^{-\frac{1}{2}m} |\Sigma_i|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\varepsilon_i^T \Sigma_i^{-1} \varepsilon_i) \qquad [4.2.9]$$

y funciones de densidad marginal:

$$f_j(\varepsilon_{ij}) = \phi_j(\varepsilon_{ij}) = (2\pi\sigma_{ij}^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_{ij}}{\sigma_{ij}}\right)^2\right] \qquad [4.2.10]$$

donde m es el número de alternativas y $\sigma_{ij}^2 = \{\Sigma_{ijj}\}$ son los elementos de la diagonal principal de Σ_i, mientras que las covarianzas se denotan σ_{ijk}. La diferencia fundamental estriba en que ahora, con la distribución normal conjunta, los ε_{ij} pueden tener covarianzas no nulas.

A pesar de que el modelo probit da lugar a representaciones más generales que el logit, presenta mayores dificultades de estimación que lo hacen poco adecuado cuando el número de alternativas es grande. Bajo el supuesto de que el individuo elige aquella alternativa que le reporta la mayor utilidad, se tiene que:

$$P_{ik} = P(\varepsilon_{ij} - \varepsilon_{ik} \leq v_{ik} - v_{ij}, \ \forall j \mid k \neq j) \qquad [4.2.11]$$

que es la expresión en **[4.2.4]**. Para desarrollar una expresión equivalente a **[4.2.7]** es necesario especificar:

$$P_{ik} = \int_{-\infty}^{\infty} d\varepsilon_{ik} \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{il}} d\varepsilon_{il} \cdot \ldots \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{ik-1}} d\varepsilon_{ik-1} \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{ik+1}} d\varepsilon_{ik+1} \cdot \ldots \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{im}} \phi(\varepsilon_i) d\varepsilon_m \qquad [4.2.12]$$

que es la probabilidad de elección en el modelo probit.

Alternativamente, la probabilidad de elección se podría haber derivado analizando la distribución conjunta de las m-1 variables aleatorias ε_{ij}-ε_{ik}, que también siguen una distribución normal multivariante. Entonces, P_{im} puede obtenerse a partir del hecho de que la suma de las probabilidades de elección es igual a la unidad. En ese caso, es necesario evaluar una integral múltiple de orden m-1, que no simplifica el problema computacional cuando m es superior a tres.

Obsérvese que si se compara con la forma funcional de las probabilidades de elección del modelo logit en **[4.2.7]**, la ecuación **[4.2.12]** es bastante desalentadora, principalmente por lo costoso en términos de cálculo que supone la evaluación de una integral de orden m-1 para cada una de las m-1 probabilidades de elección y cada uno de los n individuos. La conclusión de lo anterior es que, pese a lo atractivo del planteamiento general del probit, sólo resulta de utilidad práctica cuando el número de alternativas es relativamente pequeño.

Debido a dicha dificultad de evaluación se han derivado métodos que permiten calcular expresiones como **[4.2.12]** a un coste reducido. Lo más utilizados son [véase Daganzo (1979)]: i) integración numérica, ii) método de simulación y iii) aproximación de Clark. Naturalmente, el coste en que se incurre es la pérdida de precisión que, en algunos casos, puede ser considerable.

4.2.4. Especificación de la utilidad observada y condiciones de identificación

Como se ha indicado anteriormente, se supone una utilidad representativa lineal en los parámetros, que se puede escribir como:

$$v_{ij} = z_{ij}^T \alpha + x_i^T \beta_j = w_{ij}^T \theta_j \qquad \textbf{[4.2.13]}$$

donde $w_{ij}^{\mathrm{T}} = [z_{ij}^{\mathrm{T}}, x_i^{\mathrm{T}}]$ y $\theta_j^T = [\alpha^T, \beta_j^T]$.

Es importante tener en cuenta que a partir de **[4.2.7]** y **[4.2.12]** las probabilidades de elección dependen de las diferencias entre las utilidades representativas, y no de los valores que tomen éstas. Este hecho tiene una clara implicación: no todos los parámetros específicos para cada alternativa pueden ser estimados. Suponiendo que el decisor dispone de m alternativas, de **[4.2.4]** se tiene que P_{im} depende de $v_{ij} - v_{im}$, $\forall j \neq m$, y sustituyendo **[4.2.13]** en esta expresión resulta:

$$\begin{aligned} v_{ij} - v_{im} &= z_{ij}^T\alpha + x_i^T\beta_j - z_{im}^T\alpha - x_i^T\beta_m \\ &= (z_{ij} - z_{im})^T\alpha + x_i^T(\beta_j - \beta_m) \\ &= (z_{ij} - z_{im})^T\alpha + x_i^T\beta_j^* \qquad \forall\, j \neq m \end{aligned} \qquad \textbf{[4.2.14]}$$

Por tanto, es obvio que existen infinitos conjuntos de pares de vectores β_j y β_m que verifican que sus diferencias sean iguales a β_j^*. Por ello, convencionalmente se impone la normalización:

$$\begin{aligned} \beta_j^* &= \beta_j - \beta_m \qquad j \neq m \\ \beta_m^* &= \mathbf{0} \end{aligned} \qquad \textbf{[4.2.15]}$$

donde los m-1 vectores β_j^* son los que pueden estimarse.

Otra cuestión importante que debe resaltarse a partir de **[4.2.4]** es que, puesto que en la determinación de la probabilidad de elección sólo influyen las diferencias en la utilidad representativa, *las variables que son constantes entre alternativas no influyen en la probabilidad de elección*. Por lo tanto, también hay que aplicar una normalización para las variables propias de cada alternativa.

El resultado de este conjunto de restricciones de identificación es que las utilidades observables serán:

$$\begin{aligned} v_{ij}^* &= (z_{ij} - z_{im})^T\alpha + x_i^T\beta_j^* \qquad \forall\, j \neq m \\ v_{im}^* &= 0 \end{aligned} \qquad \textbf{[4.2.16]}$$

y las probabilidades de elección resultan, para el modelo logit:

$$\begin{aligned} P_{ik} &= \frac{\exp(v_{ik})}{\sum_j \exp(v_{ij})} = \frac{\exp(v_{ik}^*)}{1 + \sum_{j|j\neq m} \exp(v_{ij}^*)} \qquad \forall\, k \neq m \\ P_{im} &= \frac{1}{1 + \sum_{j|j\neq m} \exp(v_{ij})} \end{aligned} \qquad \textbf{[4.2.17]}$$

y para el probit:

$$P_{ik} = \int_{-\infty}^{\infty} d\varepsilon_{ik} \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{i1}} d\varepsilon_{i1} \cdot \ldots \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{ik-1}} d\varepsilon_{ik-1} \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{ik+1}} d\varepsilon_{ik+1} \cdot \ldots \cdot \int_{-\infty}^{\varepsilon_{ik}+v_{ik}-v_{im}} \phi(\varepsilon_i) d\varepsilon_m \qquad \textbf{[4.2.18]}$$

Por último, conviene hacer notar que al igual que ocurre en los modelos de elección binaria, las probabilidades de elección para cada individuo no se ven alteradas por cambios de escala en la utilidad. Los modelos:

$$U_{ij} = \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j + \varepsilon_{ij} \qquad \textbf{[4.2.19]}$$

y

$$U_{ij}/\lambda = \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j/\lambda + \varepsilon_{ij}/\lambda \qquad \lambda > 0 \qquad \textbf{[4.2.20]}$$

son observacionalmente equivalentes. Por tanto, sólo podemos obtener estimaciones de los parámetros $\boldsymbol{\theta}$ hasta un factor de escala, y es necesario imponer una normalización suponiendo que var(ε_{ij}) $\forall i,j$ es conocida, y los parámetros estimados lo son hasta un factor de escala σ, que es la desviación típica de ε_{ij} $\forall i,j$.

4.2.5. Otros aspectos de los modelos multinomiales

En esta sección se discuten brevemente dos aspectos importantes a la hora de modelizar situaciones reales utilizando modelos de elección múltiple. En particular, se analiza la *propiedad de independencia de las alternativas irrelevantes*, que es una consecuencia del supuesto de independencia entre las perturbaciones en un modelo logit y las dificultades de modelizar la *variación en gustos*.

4.2.5.A. La propiedad de Independencia de las Alternativas Irrelevantes (IIAP)

Se dice que se cumple la *IIAP*[14] para un modelo, cuando el ratio de las probabilidades de elección entre dos alternativas del conjunto disponible para el decisor, es constante e independiente de las restantes alternativas.

La formulación del modelo logit en **[4.2.7]** verifica esta propiedad puesto que:

$$\frac{P_{ik}}{P_{il}} = \frac{\exp(v_{ik}) \,/\, \sum_j \exp(v_{ij})}{\exp(v_{il}) \,/\, \sum_j \exp(v_{ij})} = \frac{\exp(v_{ik})}{\exp(v_{il})} = \exp(v_{ik} - v_{il}) \qquad \forall\, k,l \quad \textbf{[4.2.21]}$$

Como puede verse en **[4.2.21]**, el ratio de las probabilidades sólo depende de la *diferencia* entre las componentes observables de la utilidad de las alternativas k y l, con independencia de cuántas alternativas se encuentran disponibles o de las características de las mismas. Esta propiedad no es necesariamente negativa, puesto que permite representar algunas situaciones reales, aunque también resulta inapropiada en otras ocasiones.

Un ejemplo clásico del problema de la IIAP es el conocido como el del autobús rojo y el azul. Sea un viajero que tiene la posibilidad de trasladarse en automóvil (*Auto*) o en una línea de autobús con vehículos rojos (*BusR*), y que para ambas alternativas la utilidad representativa es idéntica. En esta situación: $P_{i\,Auto} = P_{i\,BusR} = 1/2$. Si se pone en circulación una nueva línea idéntica en todo a la roja excepto porque los vehículos son azules (*BusA*), debería ocurrir que $P_{i\,BusR}/P_{i\,BusA} = 1$. De acuerdo con **[4.2.21]**, en el modelo logit **[4.2.7]**, el ratio $P_{i\,Auto}/P_{i\,BusR} = 1$ debe ser constante con independencia de que haya o no otras alternativas, este ratio continuará siendo la unidad. El único valor que cumple las condiciones anteriores es: $P_{i\,Auto} = P_{i\,BusR} = P_{i\,BusA} = 1/3$, lo que obviamente no es verosímil. En buena lógica, lo que ocurriría es que: $P_{i\,Auto} = 1/2$, $P_{i\,BusR} = 1/4$ y $P_{i\,BusA} = 1/4$. Es decir, la probabilidad de ir en automóvil no debería verse afectada por la introducción de una nueva línea de autobús, aunque la línea inicial (la roja) sí se vería afectada debido a la (supuesta) indiferencia de los individuos hacia los colores de los autobuses. En este caso, las probabilidades de elección estarían infraestimadas para el automóvil, pero sobreestimadas para los autobuses.

---

[14] *Independence of Irrelevant Alternatives Property*: Propiedad de Independencia de las Alternativas Irrelevantes

Por el contrario, una situación en la que la IIAP es útil puede plantearse como sigue. En una situación de elección de modo de transporte, el conjunto de alternativas al que se enfrenta cada individuo puede ser infinitamente grande (además de los clásicos automóvil y autobús, podemos incluir bicicleta, burro, patinete, etc.). Si el interés se centra en un par de medios de locomoción (automóvil y autobús, por ejemplo), se pueden agrupar todos los restantes en una sola opción o simplemente excluir de la muestra a aquellos sujetos que no hubiesen elegido ni automóvil ni autobús. La IIAP, en este caso, permite estimar correctamente las probabilidades de elección de interés [Train (1986)].

Otra aplicación interesante de la IIAP consiste en estimar probabilidades para alternativas que no están disponibles en el momento de la estimación. Si se cree verosímil el cumplimiento de la IIAP, el modelo que se estima para el cojunto de alternativas vigentes puede ser empleado para estimar probabilidades de elección de alternativas cuyas características puedan suponerse (o conocerse), aunque dichas alternativas no existan. Por ejemplo, en el caso de la demanda de automóviles. Si se ha estimado la probabilidad de elección para una serie de marcas y modelos concretos, se puede prever la probabilidad de elección de un nuevo modelo usando las características que presentará o, de otro modo, evaluar la aceptación potencial de diferentes modelos utilizando las características que podrían ser incluídas en el vehículo.

La primera solución para el problema de la IIAP en el caso de que no sea realista, consiste en utilizar otra forma funcional para $F(\cdot)$, como por ejemplo la normal multivariante, que no posee la IIAP. Sin embargo, se pueden plantear soluciones manteniendo la forma logística para la probabilidad de elección. En el ejemplo de los autobuses, es sencillo demostrar que el problema se soluciona especificando la utilidad representativa:

$$\begin{aligned} G(v_{i\ Auto}) &= v_{i\ Auto} \\ G(v_{ij}) &= \mu + v_{ij} \qquad j = BusR,\ BusA \end{aligned} \qquad \textbf{[4.2.22]}$$

con $\mu = \ln 1/2$. En general, el factor de corrección μ es desconocido y es necesario incluir, entre la variables exógenas, una constante propia para cada alternativa. La inclusión de esta constante supone que en la fase de estimación, se obtienen estimaciones de los factores de corrección necesarios para evitar (al menos en parte) los desajustes provocados por la IIAP. Esta inclusión, en cambio, hace que no se puedan estimar probabilidades para alternativas no incluidas en la muestra, a no ser que se conozca su factor de ajuste.

Por otra parte, el modelo probit no presenta el problema de independencia de alternativas irrelevantes. Esto es debido a que se permiten covarianzas no nulas en la matriz de varianzas de los componentes no observables de la utilidad. De hecho, es posible imponer restricciones sobre la correlación entre las perturbaciones. Esto permite, además, contrastar la validez del modelo respecto de otro estimado sin la estructura impuesta *a priori*.

Sea, nuevamente, una situación de elección de modo de transporte. Las alternativas que se presentan al individuo son: automóvil propio (1), metro (2) y autobús (3). Posiblemente, las alternativas (2) y (3) son altamente sustitutivas, y la elección entre ellas es independiente de (1). La situación que previsiblemente se dará si desaparece el metro, es que los individuos vayan en autobús y viceversa, y en menor medida recurrirán al coche.

Considerando que la situación anterior es una representación plausible de la realidad, la matriz de varianzas covarianzas restringida quedará:

$$\Omega_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{bmatrix} \qquad \textbf{[4.2.23]}$$

siendo posible contrastar si el modelo con la restricción de covarianzas representa los datos de igual forma que el no restringido.

4.2.5.B. Variación en gustos

Otro punto importante en este tipo de modelos es que la inclusión de efectos individuales, lo que puede entenderse como modelizar la variación en gustos, tiene ciertas limitaciones en el modelo logit.

El supuesto subyacente es que una cierta característica de las alternativas no es igualmente valorada por los individuos, lo que se puede ilustrar con el siguiente ejemplo. Supongamos que la utilidad observable de la elección de modo de transporte al lugar de trabajo depende del coste de la alternativa (C_j) y de otro conjunto de características. Es claro que un mismo coste no es valorado de igual modo por diferentes individuos. Por lo tanto, la utilidad observable puede formularse:

$$V_{ij} = \alpha_i C_j + \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j \qquad \textbf{[4.2.24]}$$

donde α_i es un parámetro específico para el sujeto i. En este caso, bajo un planteamiento determinista, se puede suponer que la valoración del coste será inversamente proporcional a la renta del individuo i (R_i):

$$\alpha_i = \frac{\alpha^*}{R_i} \qquad \textbf{[4.2.25]}$$

y sustituyendo en **[4.2.24]**:

$$V_{ij} = \alpha^* \left(\frac{C_j}{R_i} \right) + \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j \qquad \textbf{[4.2.26]}$$

donde C_j/R_i puede interpretarse como la interacción coste-renta.

En el modelo logit, el tratamiento de factores individuales da lugar a un problema si éstos dependen de variables no observables o poseen algún componente aleatorio. Supongamos que en **[4.2.25]**:

$$\alpha_i = \frac{\alpha^*}{R_i} + \zeta_i \qquad \textbf{[4.2.27]}$$

donde ζ_i es una variable aleatoria con $E(\zeta_i) = 0$ y varianza finita. En esta situación, **[4.2.25]** queda:

$$V_{ij} = \alpha^* \left(\frac{C_j}{R_i} \right) + \zeta_i C_j + \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j \qquad \textbf{[4.2.28]}$$

de forma que la utilidad puede escribirse como:

$$\begin{aligned} U_{ij} &= \alpha^* (C_j/R_i) + \zeta_i C_j + \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j + \varepsilon_{ij} \\ &= \alpha^* (C_j/R_i) + \boldsymbol{w}_{ij}^T \boldsymbol{\theta}_j + \varepsilon_{ij}^* \end{aligned} \qquad \textbf{[4.2.29]}$$

donde $\varepsilon_{ij}^* = \zeta_i C_j + \varepsilon_{ij}$ ya no se distribuye idéntica e independientemente EVD, como se requiere para el modelo que estamos tratando. De hecho, puesto que ζ_i es igual entre alternativas, la $\text{cov}(\varepsilon_{ij}^*, \varepsilon_{ik}^*) \neq 0$, $\forall\, j,k$. Pero además, como C_j varía entre alternativas, ahora $\text{var}(\varepsilon_{ij}^*) \neq \text{var}(\varepsilon_{ik}^*)$, $\forall\, j \neq k$.

Lo que sugiere la discusión anterior es que el supuesto habitualmente utilizado de que los componentes no observables de la utilidad se distribuyen independientemente es bastante restrictivo. Su principal deficiencia reside en que no permite la presencia de correlación entre alternativas (sustituibilidad), situación bastante realista y cuya consecuencia más clara es la IIAP.

Nuevamente, el modelo probit no presenta este tipo de problemas, ya que permite covarianzas no nulas entre los componentes no observables, haciéndolo más flexible que el modelo logit.

4.3. Estimación de los modelos de elección múltiple

4.3.1. *Estimación de máxima verosimilitud*

El tratamiento más extendido y generalmente satisfactorio para la estimación del vector de parámetros θ en un MEM, es la estimación máximo-verosímil. La función de verosimilitud para la observación i-ésima condicionada a la información disponible es:

$$\mathcal{L}_i = \mathcal{L}_i(\theta \mid W_i, y_i) = \prod_{j=1}^{m} P_{ij}^{y_{ij}} = P_{i1}^0 P_{i2}^0 \cdot \ldots \cdot P_{ik}^1 \cdot \ldots \cdot P_{im}^0 = P_{ik} \qquad [4.3.1]$$

Esto es, como cada individuo elige una sola alternativa (sólo un elemento de y_i es distinto de cero) se puede definir $\mathcal{L}_i = P_{ik}$ como la probabilidad de elección (estimada por el observador) de que el individuo elija la alternativa por la que realmente opta.

Por lo tanto, la función de verosimilitud muestral resulta:

$$\mathcal{L} = \mathcal{L}(\theta \mid W, Y) = \prod_{i=1}^{n} \mathcal{L}_i = \prod_{i=1}^{n} \prod_{j=1}^{n} P_{ij}^{y_{ij}} \qquad [4.3.2]$$

y tomando logaritmos

$$\ell = \ln \mathcal{L}(\theta \mid W, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \ln P_{ij} \qquad [4.3.3]$$

que es la función a maximizar, donde P_{ij} depende de θ y de los datos según la especificación de v_{ij} y de la función de distribución de los componentes no observables. Conviene resaltar que puesto que y_{ij} es cero para las alternativas no elegidas, **[4.2.18]** es la suma, para todos los individuos, del logaritmo de la probabilidad de la alternativa elegida.

A continuación, se desarrollan las expresiones para los modelos logit y probit que se han tratado en los apartados anteriores. Supongamos n individuos y m alternativas. Para cada individuo se dispone de m vectores z_{ij} de dimensión $k_1 \times 1$ de atributos de las opciones, y un vector x_i de dimensión $k_2 \times 1$ de características personales. Además, también se observa el correspondiente vector de decisión $y_i^T = [y_{i1}, \ldots, y_{im}]$ formado por m-1 ceros y un uno en la posición de la alternativa elegida por el individuo.

Los vectores de parámetros a estimar serán: i) el vector α de dimensión $k_1 \times 1$ correspondiente a las variables en z_{ij}, y ii) $\beta^T = [\beta_1^T, \beta_2^T, \ldots, \beta_{m-1}^T]$ de dimensión $(m\text{-}1)\cdot k_2 \times 1$, formado por los vectores de parámetros de cada alternativa asociados a x_i, y donde se ha impuesto la normalización **[4.2.16]**. El vector del conjunto de parámetros es $\theta^T = [\alpha^T, \beta^T]$ de dimensión $k \times 1$, donde $k = k_1 + (m\text{-}1)\cdot k_2$.

El vector w_{ij} de variables se puede definir del siguiente modo. Sea $\tilde{x}_{ij}$ un vector $(m\text{-}1)\cdot k_2 \times 1$, particionado en bloques de k_2 elementos que contiene en su j-ésimo bloque el vector x_i, y en el resto vectores de k_2 ceros; esto es, $\tilde{x}_{ij}^T = [\mathbf{0}_{k_2}^T, \mathbf{0}_{k_2}^T, \ldots, x_i^T, \ldots, \mathbf{0}_{k_2}^T]$, de modo que se cumple que $\tilde{x}_{ij}^T \beta = x_i^T \beta_j$, $j = 1, \ldots, m-1$. Entonces, se define $w_{ij}^T = [z_{ij}^T, \tilde{x}_{ij}^T]$, de forma que, sujeto a la normalización **[4.2.16]**, se cumple que: $w_{ij}^T \theta = z_{ij}^T \alpha + x_i^T \beta_j$.

Entonces, se define la matriz W_i de dimensión $k \times m\text{-}1$:

$$W_i^T = \begin{bmatrix} w_{i1}^T \\ w_{i2}^T \\ \vdots \\ w_{im-1}^T \end{bmatrix} = \begin{bmatrix} z_{i1,1} & \cdots & z_{i1,k_1} & x_{i1} & \cdots & x_{ik_2} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ z_{i2,1} & \cdots & z_{i2,k_1} & 0 & \cdots & 0 & x_{i1} & \cdots & x_{ik_2} & \cdots & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & & \vdots & & & \vdots & & \\ z_{im-1,1} & \cdots & z_{im-1,k_1} & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & x_{i1} & \cdots & x_{ik_2} \end{bmatrix} \qquad \textbf{[4.3.4]}$$

donde $z_{ij,1}, \ldots, z_{ij,k_1}$ son las variables que describen cómo el individuo i percibe las k_1 características de la alternativa j, mientras que $x_{i1}, \ldots, x_{ik_2}$ denotan las k_2 características personales del individuo i-ésimo.

Sea $P_i^T = F_i^T = [P_{i1}, P_{i2}, \ldots, P_{im-1}] = [F_{i1}, F_{i2}, \ldots, F_{im-1}]$ el vector de probabilidades de elección de las alternativas del individuo i. También se denota por:

$$f_{ij} = \begin{bmatrix} \dfrac{\partial F_{ij}}{\partial (w_{i1}^T \theta)} \\ \vdots \\ \dfrac{\partial F_{ij}}{\partial (w_{im-1}^T \theta)} \end{bmatrix} = \begin{bmatrix} f_{ij}^1 \\ \vdots \\ f_{ij}^{m-1} \end{bmatrix} \qquad \text{y} \qquad f_i = [f_{i1} \cdots f_{im-1}] \qquad \textbf{[4.3.5]}$$

A partir de lo anterior, es claro que:

$$\frac{\partial F_{ij}}{\partial \boldsymbol{\theta}} = \sum_{l=1}^{m-1} \frac{\partial F_{ij}}{\partial (\boldsymbol{w}_{il}^T \boldsymbol{\theta})} \boldsymbol{w}_{il} = W_i \boldsymbol{f}_{ij} \qquad \textbf{[4.3.6]}$$

y por lo tanto:

$$\frac{\partial F_i}{\partial \boldsymbol{\theta}} = W_i f_i \qquad \textbf{[4.3.7]}$$

El logaritmo de la función de verosimilitud en **[4.3.3]**, teniendo en cuenta las restricciones de identificación en **[4.2.16]**, se puede escribir:

$$\ell = \sum_{i=1}^{n} \sum_{j=1}^{m-1} y_{ij} \ln F_{ij} \qquad \textbf{[4.3.8]}$$

y el gradiente resulta:

$$\nabla \ell = \sum_i \sum_j y_{ij} \frac{1}{F_{ij}} W_i \boldsymbol{f}_{ij} = \sum_i W_i f_i D_i^{-1} \mathbf{y}_i \qquad \textbf{[4.3.9]}$$

donde $D_i = \text{diag}(F_{i1}, \ldots, F_{im-1})$. Definiendo la matriz de covarianzas de $\mathbf{y}_i$ como:

$$A_i = E[(\mathbf{y}_i - \boldsymbol{P}_i)(\mathbf{y}_i - \boldsymbol{P}_i)^T] = D_i - \boldsymbol{F}_i \boldsymbol{F}_i^T \qquad \textbf{[4.3.10]}$$

y utilizando el lema de inversión de matrices de **[3.2.12]** puede demostrarse que $D_i^{-1}\mathbf{y}_i = A_i^{-1}(\mathbf{y}_i - \boldsymbol{F}_i)$, por lo que otro modo de expresar el gradiente en **[4.3.9]** es:

$$\nabla \ell = \sum_i W_i f_i D_i^{-1} \mathbf{y}_i = \sum_i W_i f_i A_i^{-1} (\mathbf{y}_i - \boldsymbol{F}_i) \qquad \textbf{[4.3.11]}$$

Finalmente, usando la equivalencia asintótica:

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] = E\left[\frac{\partial \ell}{\partial \boldsymbol{\theta}} \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\right] \qquad \textbf{[4.3.12]}$$

la matriz de información puede escribirse:

$$I(\boldsymbol{\theta}) = \sum_i W_i f_i A_i^{-1} f_i^T W_i^T \qquad \textbf{[4.3.13]}$$

Para el modelo logit multinomial, las expresiones anteriores pueden simplificarse considerablemente, puesto que en este caso particular $A_i = f_i$. El logaritmo de la función de verosimilitud de **[4.3.8]** se puede escribir como:

$$\begin{aligned}
\ell &= \sum_i \sum_j y_{ij} \ln P_{ij} \\
&= \sum_i \sum_j y_{ij} [v_{ij} - \ln \sum_j \exp(v_{ij})] \\
&= \sum_i \sum_j y_{ij} \boldsymbol{w}_{ij}^T \boldsymbol{\theta} - \sum_i \sum_j y_{ij} \ln [\sum_j \exp(\boldsymbol{w}_{ij}^T \boldsymbol{\theta})] \\
&= \sum_i \boldsymbol{y}_i^T \boldsymbol{W}_i^T \boldsymbol{\theta} - \sum_i \sum_j y_{ij} \ln [\sum_j \exp(\boldsymbol{w}_{ij}^T \boldsymbol{\theta})] \\
&= \sum_i \boldsymbol{y}_i^T \boldsymbol{W}_i^T \boldsymbol{\theta} - \sum_i \ln [\sum_j \exp(\boldsymbol{w}_{ij}^T \boldsymbol{\theta})]
\end{aligned} \qquad \textbf{[4.3.14]}$$

donde, en la última igualdad, se ha usado el hecho de que cada individuo sólo elige una alternativa. El vector gradiente resulta:

$$\begin{aligned}
\nabla \ell = \frac{\partial \ell}{\partial \boldsymbol{\theta}} &= \sum_i \boldsymbol{W}_i \boldsymbol{y}_i - \sum_i \sum_j \frac{w_{ij} \exp(\boldsymbol{w}_{ij}^T \boldsymbol{\theta})}{\sum_j \exp(\boldsymbol{w}_{ij}^T \boldsymbol{\theta})} \\
&= \sum_i \boldsymbol{W}_i \boldsymbol{y}_i - \sum_i \boldsymbol{W}_i \boldsymbol{P}_i \\
&= \sum_i \boldsymbol{W}_i (\boldsymbol{y}_i - \boldsymbol{P}_i)
\end{aligned} \qquad \textbf{[4.3.15]}$$

y el hessiano:

$$\nabla^2 \ell = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^T} = \sum_i \boldsymbol{W}_i \boldsymbol{P}_i (\boldsymbol{y}_i - \boldsymbol{P}_i)^T \boldsymbol{W}_i^T \qquad \textbf{[4.3.16]}$$

Por último, la matriz de información en el modelo logit multinomial es:

$$I(\boldsymbol{\theta}) = \sum_i \boldsymbol{W}_i \boldsymbol{A}_i \boldsymbol{W}_i^T \qquad \textbf{[4.3.17]}$$

La función de verosimilitud de estos modelos es globalmente cóncava en general, por lo que la maximización de **[4.3.3]** puede realizarse por alguno de los métodos usuales de optimización numérica expuestos en el **Apéndice A.2**. Debido a las propiedades de la función de verosimilitud, el algoritmo Newton-Raphson o el de *scoring* son elecciones adecuadas [véase Daganzo (1979) y Ben-Akiva y Lerman (1985)].

Bajo condiciones de regularidad no muy restrictivas, los parámetros estimados siguen una distribución asintótica normal y, en general, son aplicables todos los resultados de las **Secciones 1.3**, **1.4** y **1.5** para modelos binarios. Esto es, todos los contrastes de resitricciones lineales, así como el principio de multiplicadores de Lagrange son de aplicación inmediata, con la única precaución de observar la dimensión diferente de las variables explicativas, que ahora forman una matriz $k \times m$-1 para cada observación.

4.3.2. Estimación de máxima verosimilitud por procedimientos lineales

Un resultado importante para poder derivar los estadísticos de influencia para modelos binarios de la **Sección 3.4** era la posiblidad de estimar el modelo por máxima verosimilitud, pero con el procedimiento lineal derivado en el **Apartado 1.3.2**. En esta sección se plantea un problema similar y se obtienen expresiones para poder estimar un modelo logit multinomial mediante procedimientos lineales.

La iteración por el procedimiento de *scoring* de **[A.2.7]**-**[A.2.8]** puede escribirse:

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\tau+1} &= \hat{\boldsymbol{\theta}}^{\tau} + [\hat{\boldsymbol{I}}^{\tau}]^{-1} \nabla \ell \\ &= [\hat{\boldsymbol{I}}^{\tau}]^{-1} (\nabla \ell + \hat{\boldsymbol{I}}^{\tau} \hat{\boldsymbol{\theta}}^{\tau}) \end{aligned} \qquad \textbf{[4.3.18]}$$

A partir de las expresiones del gradiente y la matriz de información en **[4.3.11]** y **[4.3.13]** respectivamente se tiene que:

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\tau+1} &= \left[\sum_i \boldsymbol{W}_i \hat{\boldsymbol{f}}_i \hat{\boldsymbol{A}}_i^{-1} \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T\right]^{-1} \left[\sum_i \boldsymbol{W}_i \hat{\boldsymbol{f}}_i \hat{\boldsymbol{A}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{P}}_i) + \left(\sum_i \boldsymbol{W}_i \hat{\boldsymbol{f}}_i \hat{\boldsymbol{A}}_i^{-1} \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T\right) \hat{\boldsymbol{\theta}}^{\tau}\right] \\ &= \left[\sum_i \boldsymbol{W}_i \hat{\boldsymbol{f}}_i \hat{\boldsymbol{A}}_i^{-1} \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T\right]^{-1} \left[\sum_i \boldsymbol{W}_i \hat{\boldsymbol{f}}_i \hat{\boldsymbol{A}}_i^{-1} \left(\boldsymbol{y}_i - \hat{\boldsymbol{P}}_i + \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T \hat{\boldsymbol{\theta}}^{\tau}\right)\right] \end{aligned} \qquad \textbf{[4.3.19]}$$

donde la virgulita denota que las expresiones están evaluadas en $\hat{\boldsymbol{\theta}}^{\tau}$. La expresión anterior puede interpretarse como el estimador por *mínimos cuadrados generalizados* (MCG) del modelo de m-1 ecuaciones en forma reducida:

$$\boldsymbol{y}_i - \hat{\boldsymbol{P}}_i + \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T \hat{\boldsymbol{\theta}}^{\tau} = \hat{\boldsymbol{f}}_i^T \boldsymbol{W}_i^T \boldsymbol{\theta}^{\tau+1} + \boldsymbol{u}_i \qquad \textbf{[4.3.20]}$$

con $E(\boldsymbol{u}_i) = \boldsymbol{0}_m$ y $V(\boldsymbol{u}_i) = \boldsymbol{A}_i$.

Otra derivación alternativa del algoritmo propuesto puede hacerse del siguiente modo. Sea el modelo no lineal:

$$y_i = F(W_i^T\theta) + u_i \qquad \textbf{[4.3.21]}$$

donde u_i es un vector de variables binarias que toma los valores $1\text{-}P_{ij}$ con probabilidad P_{ij} y $-P_{ij}$ con probabilidad $1\text{-}P_{ij}$, de forma que:

$$E(u_i) = 0_m \quad \text{y} \quad V(u_i) = \text{diag}(P_i) - P_i P_i^T = A_i \qquad \textbf{[4.3.22]}$$

Si se lleva a cabo una aproximación lineal del modelo **[4.3.21]** mediante una expansión por Taylor de $F(W_i^T\hat{\theta}^{\tau+1})$ alrededor de un vector de condiciones iniciales $\hat{\theta}^{\tau}$, se obtiene:

$$F_i = \hat{F}_i + \left[\frac{\partial F_i}{\partial \hat{\theta}^{\tau}}\right]^T (\hat{\theta}^{\tau+1} - \hat{\theta}^{\tau}) + R_i \qquad \textbf{[4.3.23]}$$

Teniendo en cuenta que $R_i \to 0$ en probabilidad si $\hat{\theta}^{\tau+1}$ es una estimación consistente de θ. Sustituyendo en la ecuación **[4.3.21]** y despejando, la aproximación puede escribirse:

$$y_i - \hat{P}_i + \hat{f}_i^T W_i^T \hat{\theta}^{\tau} = \hat{f}_i^T W_i^T \theta^{\tau+1} + u_i \qquad \textbf{[4.3.24]}$$

que es la expresión en **[4.3.20]**.

Una vez más, para el modelo logit multinomial las expresiones anteriores pueden simplificarse considerablemente. En particular, en la expresión **[4.3.24]** no hay más que sustituir $\hat{f}_i$ por A_i y el paso de *scoring* de **[4.3.19]** resulta:

$$\hat{\theta}^{\tau+1} = \left[\sum_i W_i \hat{A}_i W_i^T\right]^{-1} \left[\sum_i W_i \left(y_i - \hat{P}_i + \hat{A}_i W_i^T \hat{\theta}^{\tau}\right)\right] \qquad \textbf{[4.3.25]}$$

4.4. Observaciones anómalas en modelos multinomiales

En esta sección se extienden los resultados previos de este trabajo, desarrollados para los modelos binomiales, a los dos modelos de elección múltiple más utilizados. Desde un punto de vista general, esta extensión no requiere planteamientos adicionales, y el problema básico consiste en trabajar con un vector de variables dependientes de mayor dimensión. Este aumento de dimensión se traduce en una mayor complejidad analítica al derivar instrumentos de diagnóstico apropiados, aunque las cuestiones conceptuales sobre el tratamiento de observaciones anómalas en modelos de variable dependiente cualitativa se mantienen.

En particular, los residuos no son un instrumento apropiado y es necesario derivar estadísticos análogos a los de la **Sección 3.4** que midan el efecto de una observación o conjunto de observaciones sobre el vector de parámetros del modelo. En este punto, el problema de dimensión hace que ahora sea necesario definir una medida escalar del efecto de las observaciones sobre las probabilidades estimadas para derivar un estadístico análogo al $\hat{c}_i$ en **[3.4.8]**.

También será necesario desagregar los instrumentos de diagnóstico *a priori*, puesto que, en general, en los modelos multinomiales pueden aparecer dos tipos de variables: las características individuales y las características de las alternativas, por lo que será necesario investigar separadamente, individuos extremos y alternativas *extrañas* respecto al conjunto. Esto será especialmente importante para aquellas variables que no sean medidas *objetivas*, sino indicadores de la percepción que tiene el individuo sobre los atributos de la opción.

No obstante, los instrumentos y la forma de utilizarlos, es análoga a la expuesta para los modelos de elección binaria, por lo que nos remitimos a la línea argumental y metodológica desarrollada en dicha sección.

4.4.1. Observaciones anómalas en modelos de elección discreta múltiple: planteamiento

De forma análoga a lo expuesto para modelos binarios, los problemas derivados de la presencia de observaciones anómalas en las muestra se ha tratado poco y, en general, bajo planteamientos metodológicos distintos de los desarrollados en este trabajo. Los trabajos anteriores se han centrado en un marco, aparentemente más amplio, como son los modelos lineales generalizados o se han basado en un enfoque de análisis de influencia. Entre estos trabajos se pueden citar Lesaffre y Albert (1989) que enfocan el problema desde el punto de vista del análisis de influencia y se estudia el caso de los modelos de elección múltiple. Por otra parte, Cook y Weisberg (1980) generalizan algunos resultados de Cook (1977) para los modelos lineales generalizados (GLM), Williams (1987) también presenta resultados sobre diagnosis para los GLM, Green (1984) desarrolla alternativas de estimación lineales y de estimación robusta y resistente para el caso de los GLM, alguno de cuyos casos particulares incluye modelos de elección múltiple.

Como se planteó anteriormente para modelos binarios, en este trabajo se trata el problema partiendo de la definición de observación anómala que se ha introducido anteriormente: *una observación anómala es aquella que no se ha generado por el mismo modelo estocástico que se supone para las restantes observaciones muestrales* [Box y Tiao (1968)].

A partir de esta definición, y utilizando los resultados para los modelos binarios es sencillo demostrar que la existencia de anomalías en la muestra afecta a la consistencia del estimador de máxima verosimilitud. Ello se debe a que la presencia de estas observaciones hace que la función de verosimilitud del modelo sea diferente de la habitual.

El planteamiento básico sigue las líneas de la **Sección 2.3**. Considerando que una proporción de observaciones ha sido generada por un proceso diferente a las restantes observaciones del modelo, se puede suponer que las probabilidades de elección de cada individuo están formadas por una combinación lineal de funciones de distribución. Por una parte, de la *verdadera*, y por otra, de la distribución de las observaciones anómalas. Esta nueva combinación lineal debe verificar que la suma de las probabilidades de elección sea la unidad.

Bajo este planteamiento, se pueden formular probabilidades de elección análogas a las de los **Apartados 2.3.1.A** y **2.3.1.B**, y la función de verosimilitud en **[4.3.1]** tendría ambos tipos de componentes. Nuevamente, los desarrollos del **Apartado 2.3.2**, sobre la

inconsistencia del estimador máximo-verosímil son de aplicación inmediata a este caso, por lo que no se repiten aquí.

Se pueden considerar dos tipos de anomalías en la variable y_{ij}^*: aquellas generadas por una distribución con distinta varianza que las restantes observaciones muestrales y aquellas generadas por una distribución con distinta media.

Con los MEM, este planteamiento puede dar lugar a casos particulares diferentes de los modelos binarios, por ejemplo, si se suponen distintas varianzas para diferentes componentes del vector de perturbaciones para cada observación. Sobre este aspecto cabe decir que, si bien es una posibilidad, la interpretación de la fuente de anomalías hace que sean poco plausibles y, por otra parte, en modelos como el logit multinomial, es necesario que todos los elementos del vector de perturbaciones sigan la misma distribución. Por lo tanto, estos casos se pueden traducir en nuevos errores de especificación que afectarán a *todos* los parámetros del modelo (no sólo a los de la alternativa para la que se produzcan), por lo que el efecto de observación anómala puede mantenerse.

4.4.2. Estadísticos de detección de observaciones anómalas en los modelos de elección múltiple

Nuestro primer objetivo es desarrollar un estadístico que permita medir el efecto de eliminar una observación cada vez sobre los coeficientes estimados. Para ello, se parte del estimador máximo-verosímil por procedimientos lineales en **[4.3.19]**, con el fin de obtener, de forma computacionalmente eficiente, estimaciones de β cuando se elimina una observación. Sobre este punto, conviene destacar que, aunque se han derivado otros estimadores de máxima verosimilitud por procedimientos lineales, como en Amemiya (1985), la parametrización empleada para derivar el estimador en **[4.3.19]** permite llevar a cabo las siguientes fases de este análisis con mucha mayor sencillez. En concreto, Amemiya (1985), deriva los diferentes bloques de los elementos que componen el estimador a partir de bloques de parámetros.

En primer lugar, es necesario desarrollar expresiones del estimador que permitan obtener estimaciones de β cuando se elimina una observación, eficientemente. Para ello, se introduce primero la siguiente notación:

$$\tilde{W}_i = \hat{f}_i W_i \qquad \textbf{[4.4.1]}$$

donde W_i y $\hat{f}_i$ están definidas en **[4.3.4]** y **[4.3.5]**, respectivamente. Además, sea $A = \text{diag}(A_1,..., A_n)$ la matriz diagonal por bloques de dimensión $(m\text{-}1)n \times (m\text{-}1)n$ y $\tilde{W}^T = [\tilde{W}_1^T, ..., \tilde{W}_n^T]$.

Partiendo de las definiciones anteriores, y dada una condición inicial del estimador $\hat{\theta}^r$, el estimador en **[4.3.19]** en notación matricial puede escribirse como:

$$\hat{\theta} = [\tilde{W}^T A^{-1} \tilde{W}]^{-1} \tilde{W}^T A^{-1} \tilde{Y} \qquad \textbf{[4.4.2]}$$

donde $\tilde{Y}^T = [\tilde{y}_1, ..., \tilde{y}_n]$ y:

$$\tilde{y}_i = y_i - \hat{P}_i + \tilde{W}_i \hat{\theta}^r \qquad \textbf{[4.4.3]}$$

En este contexto, la eliminación de una observación es equivalente a eliminar las m-1 filas asociadas de la matriz $\tilde{W}$. Así, denotando con el subíndice (i) aquellas matrices de las que se ha eliminado la *observación i*, se tienen las siguientes igualdades:

$$(\tilde{W}^T A^{-1} \tilde{W})_{(i)} = \tilde{W}^T A^{-1} \tilde{W} - \tilde{W}_i^T A_i^{-1} \tilde{W}_i \qquad \textbf{[4.4.4]}$$

$$(\tilde{W}^T A^{-1} \tilde{Y})_{(i)} = \tilde{W}^T A^{-1} \tilde{Y} - \tilde{W}_i^T A_i^{-1} \tilde{y}_i \qquad \textbf{[4.4.5]}$$

Aplicando el lema de inversión de matrices dado en **[3.2.12]** a **[4.4.4]** resulta que:

$$\begin{aligned}(\tilde{W}^T A^{-1} \tilde{W})_{(i)}^{-1} &= (\tilde{W}^T A^{-1} \tilde{W})^{-1} \\ &+ (\tilde{W}^T A^{-1} \tilde{W})^{-1} \tilde{W}_i^T [A_i - \tilde{W}_i (\tilde{W}^T A^{-1} \tilde{W})^{-1} \tilde{W}_i^T]^{-1} \tilde{W}_i (\tilde{W}^T A^{-1} \tilde{W})^{-1}\end{aligned} \qquad \textbf{[4.4.6]}$$

Posmultiplicando **[4.4.6]** por la expresión en **[4.4.5]** y despejando, queda:

$$\hat{\theta}_{(i)}^{-1} = \hat{\theta} - (\tilde{W}^T A^{-1} \tilde{W})^{-1} \tilde{W}_i^T [A_i - \tilde{W}_i (\tilde{W}^T A^{-1} \tilde{W})^{-1} \tilde{W}_i^T]^{-1} (\tilde{y}_i - \tilde{W}^T \hat{\theta}) \qquad \textbf{[4.4.7]}$$

Bajo los mismos planteamientos que en la **Sección 3.4**, una primera medida del efecto que tiene la i-ésima observación sobre el vector de coeficientes estimados resulta:

$$\hat{c}_i = (\hat{\theta}_{(i)} - \hat{\theta})^T [\tilde{W}^T A^{-1} \tilde{W}] (\hat{\theta}_{(i)} - \hat{\theta}) \qquad \textbf{[4.4.8]}$$

Sustituyendo la expresión **[4.4.7]** en **[4.4.8]** se obtiene una expresión, más sencilla de calcular, para el estadístico de influencia en **[4.4.8]**:

$$\hat{c}_i = \tilde{e}_{(i)}^T (A_i - N_i) N_i (A_i - N_i) \tilde{e}_{(i)} \qquad \textbf{[4.4.9]}$$

donde $N_i = \tilde{W}_i (\tilde{W}^T A^{-1} \tilde{W}) \tilde{W}_i^T$.

Dada la normalidad asintótica del estimador máximo-verosímil, la interpretación en términos de regiones de confianza del vector de parámetros estimados que se expuso para los modelos binarios sigue siendo útil para seleccionar puntos de corte indicativos.

Una última extensión de estos estadísticos se encuentra, naturalemente, en evaluar el efecto que tiene una observación anómala sobre un conjunto de parámetros. Este caso es especialmente importante en este contexto, donde el vector de parámetros esta formado tanto por subconjuntos de parámetros: los propios de cada alternativa y los asociados a las características de las distintas opciones. De forma semejante a como se hizo anteriormente, en primer lugar se considera una medida de influencia para el vector $\eta = R\theta$.

Una vez obtenidas las estimaciones máximo-verosímiles de θ, resulta evidente que la matriz de varianzas-covarianzas de $\hat{\eta}$ es:

$$R(\tilde{W}^T \hat{A}^{-1} \tilde{W})^{-1} R^T \qquad \textbf{[4.4.10]}$$

Por tanto, un estadístico análogo a **[4.4.8]** puede formularse:

$$\hat{c}_i(\eta) = (\hat{\theta}_{(i)} - \hat{\theta})^T R^T [R(\tilde{W}^T A^{-1} \tilde{W})^{-1} R^T]^{-1} R(\hat{\theta}_{(i)} - \hat{\theta}) \qquad \textbf{[4.4.11]}$$

Sustituyendo la expresión **[4.4.7]** en **[4.4.11]** y particularizando para disintas elecciones de R se obtiene el conjunto de estadísticos especializados de interés análogos a los del **Capítulo 3**.

4.5. Resultados con datos simulados para el modelo logit múltiple

Con el propósito de ilustrar que la presencia de observaciones anómalas induce el mismo tipo de problemas en los modelos multinomiales, así como la analogía del estadístico de detección de observaciones anómalas **[4.4.9]** respecto del derivado en **[3.4.8]** para modelos binarios, se ha realizado un subconjunto de las simulaciones que se presentaron en los **Capítulos 2** y **3** para modelos binarios.

4.5.1. Planteamiento de los modelos

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

$$\begin{aligned} y_{i1}^{*} &= \beta_{10} + \beta_{11} x_i + \varepsilon_{i1} \\ y_{i2}^{*} &= \beta_{20} + \beta_{21} x_i + \varepsilon_{i2} \end{aligned} \qquad \textbf{[4.5.1]}$$

y se observa un vector de variables binarias:

$$\begin{aligned} y_{i1} &= 1 \quad \text{si} \quad y_{i1}^{*} \geq y_{ij}^{*} \quad j=2,3 \\ y_{i2} &= 1 \quad \text{si} \quad y_{i2}^{*} \geq y_{ij}^{*} \quad j=1,3 \\ y_{i3} &= 1 \quad \text{demás casos} \end{aligned} \qquad \textbf{[4.5.2]}$$

La variable explicativa se ha generado, en todos los casos de observaciones no anómalas, como una normal, $x_i \sim iid\ N(0,1)$ y los vectores de parámetros son: $\beta_1 = (0.5,\ 2.5)^T$ y $\beta_2 = (-0.5,\ 0.5)^T$. Para generar las perturbaciones distribuidas EVD de cada ecuación, se ha empleado la transformación integral, de modo análogo a como se hizo para las perturbaciones logísticas en la **Sección 2.4**. Dado que ahora se trata con modelos de elección múltiple, las muestras son de tamaño 500. Cada experimento se replicó 500 veces.

A partir de este mecanismo, se han creado muestras donde se incluye un porcentaje ω de observaciones y_i^{*} generadas por la misma distribución que para el resto, pero con momentos distintos de los que se acaban de señalar. En particular, se consideran los siguientes casos, cuyos planteamientos teóricos se han discutido en la **Sección 2.3**:

Caso 1: Un porcentaje ω de observaciones y_i^* en la muestra proviene de una distribución con la misma media que las restantes observaciones, pero con la varianza multiplicada por un factor $h^2 = 7$.

Caso 2: Un porcentaje ω de observaciones y_i^* proviene de la misma distribución con varianza igual a σ_0^2, pero con distinta media que las restantes observaciones. En concreto, se ha incluido una proporción ω de observaciones tales que $E(y_{ij}^*) = \boldsymbol{x}_i^T \boldsymbol{\gamma}_j$, $j = 1, 2$, donde $\boldsymbol{\gamma}_1 = (0.2, 1.0)^T$ y $\boldsymbol{\gamma}_2 = (-1.0, 0.2)^T$. Obsérvese que se ha considerado un caso extremo, en el que las componentes de los vectores $\boldsymbol{\gamma}$ son muy diferentes a las de los vectores $\boldsymbol{\beta}$.

Caso 3: Un porcentaje ω de observaciones están caracterizadas por $E(y_{ij}^*) = \boldsymbol{x}_i^T \boldsymbol{\delta}_j$, $j = 1, 2$, e idéntica varianza que las demás, donde $\boldsymbol{\delta}_1 = (-0.5, 1.0)^T$ y $\boldsymbol{\delta}_2 = (0.3, -1.0)^T$. Para las observaciones anómalas, $x_i \sim iid\ N(5,1)$. Esto es, las componentes de los vectores $\boldsymbol{\delta}$ son parecidas a las de $\boldsymbol{\beta}$ pero, es de esperar que, para las observaciones anómalas, una proporción importante de los valores de x_i sean mucho mayores que los de las restantes observaciones.

La interpretación de estos tres esquemas, expuesta en la **Sección 2.4**, es la siguiente. El caso 1 se basa en la idea de que la heterocedasticidad aparece frecuentemente en datos de sección cruzada. Los casos 2 y 3, pueden interpretarse como originados por las técnicas de muestreo, básicamente el muestreo estratificado [Azorín y Sanchez-Crespo (1986)], técnica con la que se puede estar incluyendo en la muestra elementos de subpoblaciones distintas entre sí; por un lado, respecto al comportamiento, aunque no respecto a sus variables características (caso 2) y por otro, respecto a sus variables características aunque homogéneas en su comportamiento (caso 3).

4.5.2. Resultados de la simulación

En las **Tablas 4.1-4.3** se presentan los resultados de la estimación, utilizando el método de máxima verosimilitud del **Apartado 4.3.2** para diferentes proporciones de observaciones anómalas. En las tablas figuran, para distintas proporciones de observaciones anómalas en la muestra, los coeficientes estimados, las desviaciones típicas estimadas de los coeficientes asociados a las variables explicativas, el error cuadrático medio definido en **[2.4.4]** y la suma de cuadrados de residuos de cada ecuación definida en **[2.4.5]**.

Tabla 4.1. Estimaciones MV con anomalías en la muestra: modelo logit multinomial, caso 1

$\omega\%$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$dt(\hat{\beta}_{11})$	$dt(\hat{\beta}_{21})$	ECM	SSR1	SSR2	SSR3
0.0	0.5104	2.5264	-0.5023	0.4993	0.2265	0.1907	0.1380	69.21	62.10	78.72
2.5	0.4928	2.4560	-0.4970	0.5037	0.2205	0.1878	0.1415	70.49	62.85	79.54
5.0	0.4673	2.4072	-0.4889	0.4815	0.2170	0.1848	0.1521	71.20	63.95	80.50
7.5	0.4690	2.3128	-0.4749	0.5008	0.2095	0.1828	0.1588	73.45	64.81	81.11
10.0	0.4603	2.2648	-0.4677	0.4879	0.2058	0.1803	0.1739	74.29	65.56	81.75
12.5	0.4444	2.1828	-0.4590	0.4629	0.1998	0.1765	0.2145	75.95	66.95	83.03
15.0	0.4291	2.1433	-0.4603	0.4527	0.1969	0.1745	0.2513	76.87	67.54	83.84
17.5	0.4360	2.0856	-0.4225	0.4724	0.1927	0.1725	0.2923	78.55	69.07	84.34
20.0	0.4206	2.0356	-0.4278	0.4570	0.1889	0.1701	0.3374	79.68	69.64	85.12
22.5	0.4102	1.9777	-0.4143	0.4369	0.1852	0.1676	0.3965	81.13	71.12	86.37
25.0	0.3909	1.9204	-0.4138	0.4225	0.1812	0.1655	0.4624	82.53	72.00	87.58
27.5	0.3886	1.8866	-0.4082	0.4237	0.1788	0.1638	0.5058	83.45	72.46	87.91
30.0	0.3753	1.8374	-0.4061	0.4174	0.1757	0.1626	0.5734	85.08	73.43	88.79

Tabla 4.2. Estimaciones MV con anomalías en la muestra: modelo logit multinomial, caso 2

$\omega\%$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$dt(\hat{\beta}_{11})$	$dt(\hat{\beta}_{21})$	ECM	SSR1	SSR2	SSR3
0.0	0.5072	2.5280	-0.4999	0.5027	0.2265	0.1903	0.1441	69.15	62.30	78.96
2.5	0.4891	2.4446	-0.5207	0.4852	0.2200	0.1877	0.1338	70.57	62.22	79.66
5.0	0.4899	2.4072	-0.5346	0.4922	0.2169	0.1875	0.1483	71.14	61.73	79.99
7.5	0.4738	2.3102	-0.5525	0.4640	0.2094	0.1848	0.1643	72.89	62.02	81.27
10.0	0.4572	2.2519	-0.5736	0.4400	0.2050	0.1835	0.1960	73.86	61.94	82.13
12.5	0.4515	2.2059	-0.5911	0.4378	0.2017	0.1825	0.2214	74.99	61.54	82.64
15.0	0.4464	2.1513	-0.6071	0.4351	0.1972	0.1816	0.2650	76.16	61.37	83.29
17.5	0.4268	2.0837	-0.6228	0.4067	0.1923	0.1789	0.3147	77.30	61.66	84.48
20.0	0.4142	2.0442	-0.6389	0.3976	0.1895	0.1786	0.3559	78.26	61.47	85.21
22.5	0.4127	1.9865	-0.6492	0.3928	0.1852	0.1769	0.4147	79.58	61.49	85.89
25.0	0.4023	1.9423	-0.6646	0.3846	0.1822	0.1762	0.4698	80.89	61.39	86.85
27.5	0.3973	1.9050	-0.6756	0.3825	0.1795	0.1758	0.5222	81.73	61.41	87.48
30.0	0.3781	1.8474	-0.7043	0.3456	0.1756	0.1737	0.6108	83.01	61.05	88.72

Tabla 4.3. Estimaciones MV con anomalías en la muestra: modelo logit multinomial, caso 3

$\omega\%$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$dt(\hat{\beta}_{11})$	$dt(\hat{\beta}_{21})$	ECM	SSR1	SSR2	SSR3
0.0	0.5035	2.5224	-0.4993	0.5093	0.2261	0.1907	0.1475	69.38	62.29	78.77
2.5	0.4908	2.4911	-0.5196	0.4910	0.2266	0.1919	0.1478	67.36	60.48	76.95
5.0	0.4739	2.4357	-0.5364	0.4744	0.2250	0.1922	0.1776	66.24	58.93	75.46
7.5	0.4759	2.3995	-0.5437	0.4594	0.2254	0.1945	0.1920	64.72	57.39	73.62
10.0	0.4673	2.3686	-0.5555	0.4487	0.2256	0.1955	0.1857	63.01	55.72	71.67
12.5	0.4584	2.3005	-0.5533	0.4278	0.2233	0.1953	0.2284	61.54	54.73	69.99
15.0	0.4442	2.2407	-0.5748	0.4103	0.2213	0.1970	0.2865	60.41	53.12	68.56
17.5	0.4386	2.1839	-0.5895	0.3768	0.2200	0.1979	0.3477	58.64	51.47	66.58
20.0	0.4485	2.1536	-0.5917	0.3803	0.2204	0.2002	0.3606	57.34	49.89	64.60
22.5	0.4184	2.0952	-0.6093	0.3543	0.2189	0.2004	0.4115	55.67	48.57	63.14
25.0	0.4229	2.0711	-0.6245	0.3464	0.2199	0.2038	0.4596	54.14	46.56	61.06
27.5	0.4059	2.0135	-0.6446	0.3373	0.2180	0.2052	0.5206	52.53	44.92	59.25
30.0	0.3848	1.9430	-0.6641	0.2964	0.2149	0.2052	0.6176	51.16	43.73	57.79

A la vista de estos resultados, es posible hacer las siguientes consideraciones:

- De forma análoga a lo que ocurría para los modelos binarios, el estimador máximo-verosímil es sesgado. Esto puede apreciarse en la primera fila de cada tabla, donde no hay observaciones anómalas en la muestra.

- La presencia de observaciones anómalas en la muestra, de forma semejante a como ocurría para los modelos binarios, induce importantes sesgos en los coeficientes estimados, más que en las desviaciones típicas estimadas. Como puede observarse, en todos los casos, los sesgos dependen positivamente de la proporción de observaciones anómalas, confirmando la inconsistencia del estimador máximo-verosímil ante la presencia de este problema.

- A diferencia de lo que ocurría con los modelos binarios, donde los sesgos de estimación eran bastante diferentes según el tipo de anomalía presente, en las tablas anteriores se puede observar que la magnitud de los cambios es semejante, con independencia del tipo de anomalía presente. También el ECM tiene un comportamiento homogéneo en todos los casos. Esto, posiblemente, se puede

atribuir a la mayor complejidad de los modelos, que hace que el efecto de las observaciones anómalas se reparta entre todos los coeficientes del modelo.

- Como puede observarse, la suma de cuadrados de residuos es más pequeña en el caso 3 que en los demás casos. La explicación de este hecho, como sucedía en los modelos binarios se puede atribuir a que, las observaciones extremas en este caso, dado que los parámetros son parecidos a los del *modelo verdadero*, son más *fáciles* de prever.

En las **Tablas 4.4-4.6** se presentan los valores medios del estadístico de influencia individual para toda la muestra, así como las medias para las observaciones no anómalas y anómalas. En las dos últimas columnas figuran los ratios entre la media del estadístico $\hat{c}_i$ para las observaciones no anómalas y anómalas y los ratios entre la media para las observaciones anómalas y la muestra completa.

Tabla 4.4. Valores medios del estadístico $\hat{c}_i$: modelo logit multinomial, caso 1.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i(B)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0080	0.0080	--	--	--
2.5	0.0081	0.0078	0.0190	2.4243	2.3377
5.0	0.0082	0.0076	0.0183	2.3909	2.2355
7.5	0.0082	0.0074	0.0174	2.3402	2.1238
10.0	0.0082	0.0072	0.0169	2.3428	2.0654
12.5	0.0082	0.0071	0.0158	2.2131	1.9197
15.0	0.0082	0.0070	0.0155	2.2277	1.8812
17.5	0.0082	0.0068	0.0150	2.2044	1.8189
20.0	0.0082	0.0067	0.0144	2.1599	1.7532
22.5	0.0083	0.0065	0.0142	2.1692	1.7158
25.0	0.0083	0.0064	0.0138	2.1428	1.6667
27.5	0.0083	0.0063	0.0134	2.1080	1.6143
30.0	0.0083	0.0063	0.0130	2.0679	1.5662

Tabla 4.5. Valores medios del estadístico $\hat{c}_i$: modelo logit multinomial, caso 2.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i(B)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0081	0.0081	--	--	--
2.5	0.0081	0.0079	0.0158	2.0024	1.9515
5.0	0.0081	0.0077	0.0146	1.8898	1.8093
7.5	0.0081	0.0076	0.0146	1.9269	1.8001
10.0	0.0081	0.0075	0.0138	1.8428	1.6996
12.5	0.0081	0.0074	0.0133	1.8099	1.6423
15.0	0.0081	0.0073	0.0132	1.8178	1.6192
17.5	0.0082	0.0071	0.0129	1.8016	1.5789
20.0	0.0082	0.0070	0.0126	1.7951	1.5488
22.5	0.0082	0.0070	0.0123	1.7586	1.5012
25.0	0.0082	0.0069	0.0121	1.7574	1.4776
27.5	0.0081	0.0068	0.0116	1.6992	1.4244
30.0	0.0082	0.0067	0.0115	1.7069	1.4083

Tabla 4.6. Valores medios del estadístico $\hat{c}_i$: modelo logit multinomial, caso 3.

$\omega\%$	$\hat{c}_i$	$\hat{c}_i(B)$	$\hat{c}_i(M)$	$\hat{c}_i(M)/\hat{c}_i(B)$	$\hat{c}_i(M)/\hat{c}_i$
0.0	0.0081	0.0081	--	--	--
2.5	0.0083	0.0082	0.0134	1.6398	1.6130
5.0	0.0086	0.0083	0.0141	1.7030	1.6452
7.5	0.0088	0.0084	0.0143	1.7024	1.6161
10.0	0.0091	0.0085	0.0143	1.6844	1.5765
12.5	0.0093	0.0087	0.0135	1.5555	1.4537
15.0	0.0097	0.0087	0.0152	1.7452	1.5697
17.5	0.0097	0.0089	0.0136	1.5290	1.3987
20.0	0.0101	0.0090	0.0144	1.6055	1.4321
22.5	0.0102	0.0092	0.0138	1.4998	1.3476
25.0	0.0105	0.0093	0.0140	1.5148	1.3421
27.5	0.0106	0.0095	0.0135	1.4241	1.2749
30.0	0.0108	0.0097	0.0134	1.3866	1.2425

En este grupo de simulaciones, los resultados son análogos a los obtenidos para los modelos binarios. A la vista de las tablas, es posible hacer las siguientes consideraciones:

- Dado que se trata de un estadístico de influencia individual, la mayor diferencia entre el valor medio para las observaciones anómalas y no anómalas se produce para valores bajos de ω. Para proporciones elevadas de observaciones anómalas, el efecto individual de cada una de ellas queda enmascarado por las restantes y, por tanto, la media del estadístico individual es inferior, aunque siempre por encima de la media para las observaciones no anómalas.

- Como ya ocurría en los modelos binarios, en el caso 3 la diferencia entre el estadístico individual para observaciones anómalas y no anómalas es inferior que en los demás casos, debido a que estas observaciones son extremas en el espacio de variables explicativas, pero dado que sus vectores de parámetros son muy semejantes a las observaciones no anómalas, la influencia es limitada.

- A la hora de aplicar este criterio, debe tenerse en cuenta que será tanto más válido cuanto menos observaciones influyentes se encuentren en la muestra, ya que es una medida de influencia individual, y si aparecen problemas de enmascaramiento no será tan útil. Por otra parte, también es necesario considerar el número de observaciones que sobrepasan el nivel crítico: si el número es excesivo, habrá que elegir como influyentes aquellas observaciones con valor más elevado puesto que, si se eliminan observaciones no anómalas, se está eliminando información relevante, por lo que se pueden llegar a introducir sesgos importantes en la estimación.

CAPÍTULO 5

APLICACIONES CON DATOS REALES

5.1. Introducción

En este último capítulo se aplica la metodología de detección de observaciones anómalas desarrollada en los **Capítulos 2** y **3** a dos muestras de datos reales. Aunque las simulaciones realizadas ilustran bien el funcionamiento de los estadísticos propuestos, los modelos planteados no reflejan exactamente situaciones reales, donde lo frecuente será encontrar pocas observaciones que condicionan los resultados de la estimación y cuya fuente de procedencia no se suele conocer.

En la **Sección 5.2** se analiza la muestra utilizada por Dhillon et al. (1987), en un estudio sobre la elección de tipos de interés fijos frente a tipos variables para préstamos hipotecarios. El objetivo de este primer ejemplo es ilustrar la necesidad de analizar los datos empleados en cualquier estudio y, en particular, en los modelos de variable dependiente cualitativa en busca de observaciones tales que su sola presencia pueda condicionar los resultados de un análisis.

En la **Sección 5.3** se aplican los planteamientos metodológicos desarrollados anteriormente al mismo conjunto de datos utilizado en el trabajo de Pregibon (1981). El interés principal se centra en comprobar que, como se ha argumentado en el **Capítulo 2**, los residuos son un instrumento muy limitado a la hora de detectar observaciones influyentes y que, además, los estadísticos de influencia individual pueden resultar insuficientes para detectar situaciones de enmascaramiento. Además, se ilustra el empleo del método de Peña y Yohai (1991) aplicado a modelos de variable dependiente cualitativa.

5.2. Elección de tipo de interés fijo frente a variable

En una nota en el *Journal of Money, Credit and Banking*, Dhillon et al. (1987) plantean el estudio de las características personales y financieras que hacen que los individuos elijan tipos de interés fijos o variables 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[15]; 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[16].

5.2.1. Planteamiento 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.

Los datos utilizados 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), la variable PR (ratio entre los pagos con interés fijo y variable) no aparece en el archivo. Esto no plantea ningún inconveniente serio, puesto que los resultados son semejantes a los obtenidos en el artículo original, aunque la omisión de esta variable puede explicar las diferencias numéricas obtenidas, en concreto en la estimación del término constante.

La muestra está compuesta por 78 clientes de un banco hipotecario de Louisiana (EE.UU.). Los préstamos fueron concedidos durante el período que va desde enero de 1983 a febrero de 1984. Del total de observaciones, 46 eligieron un tipo de interés fijo y

---

[15] La información es simétrica, y el efecto de las características personales ya se encuentra incluido en los términos del contrato.

[16] La información asimétrica supone que existen características personales que, de conocerse, podrían perjudicar al prestatario.

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 5.1** y un listado completo de los datos utilizados se incluye en el **Apéndice A.3**.

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, que se mide por la diferencia de los tipos del Tesoro a diez y un año.

Tabla 5.1. 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.

5.2.2. Resultados empíricos con los modelos originales

En la **Tabla 5.2** 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 verosimilititud por procedimientos lineales expuesto el **Apartado 1.3.2**. 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 definido en **[1.3.2]**, el valor del estadístico de contraste de razón de verosimilitudes en **[1.4.5]** (bajo el modelo restringido) y el número condición de la matriz de varianzas-covarianzas estimada[17].

Siguiendo a Dhillon et al. (1987), y a la vista de los resultados de la **Tabla 5.2** 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.

- Las variables personales no son significativas indidividualmente 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[18], lo que presenta una evidencia 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 a variaciones importantes en los coeficientes estimados.

---

[17] Ratio entre el mayor y el menor autovalor de la matriz de varianzas-covarianzas.

[18] El valor tabular de la distribución χ^2_7 es de 12.0 al 90% y de 14.1 al 95%. El valor del estadístico de contraste es 16.8.

Tabla 5.2. 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 vivienda)	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. p.)	--	--	0.0161
			(0.0283)
ln ℓ	-31.53	-39.21	-30.73
LRT	1.60	16.96	
N° condición	1.7e+6	1.4e+05	2.4e+06

En la **Tabla 5.3** aparecen las elasticidades, calculadas mediante las expresiones **[1.5.10]**, 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 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 5.3. 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, ... tienen preferencia por los tipos de interés variable. Esto es consistente con la hipótesis de información asimétrica" Dhillon et al. (1987, pág. 265).

5.2.3. Detección de observaciones anómalas

En este apartado se aplica la metodología desarrollada en la **Sección 3.4** para detetectar observaciones anómalas e influyentes. En la **Tabla 5.4** se presentan un conjunto

de instrumentos de diagnosis desarrollados en el **Capítulo 3** para un conjunto de observaciones tales que presentaban valores apreciables en alguno de ellos.

Por columnas, la **Tabla 5.4** contiene: el número de la observación, el estadístico de distancia definido en **[3.2.2]**, el estadístico $\tilde{h}_i$ de distancia para las variables transformadas con las expresiones **[3.4.4]**-**[3.4.5]**, el residuo definido en **[3.3.1]**, el estadístico de influencia individual de **[3.4.8]** y, por último, los componentes de los autovectores asociados a los dos mayores autovalores de la matriz de influencia $\tilde{M}$ definida en **[3.4.17]**.

Tabla 5.4. Estadísticos de diagnóstico para las observaciones más significativas.

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

A la vista de la **Tabla 5.4** se pueden hacer los siguientes comentarios:

- El estadístico h_i en **[3.2.2]** 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 del

Apartado 1.3.2 y transformadas las variables, el estadístico $\tilde{h}_i$ confirma que existen algunas observaciones extremas en el espacio de las X.

- Atendiendo al estadístico para cada observación, puede comprobarse que la 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, aunque este punto crítico puede considerarse elevado. Obsérvese también que la media del estadístico $\hat{c}_i$ para la muestra completa se encuentra por encima de cuatro, aunque esa media está muy afectada debido al valor extremo de la observación 37. Utilizando la media de las observaciones, eliminando la 37, un valor crítico bajo es, aproximadamente, 1.5. Por último, un punto crítico mínimo se encuentra en $2k/n$, pero a la vista de la tabla, puede resultar en la eliminación de un número excesivo de observaciones.

- Como se puso de relieve en los **Capítulos 2** y **3**, los residuos en los modelos de variable dependiente binaria no son una indicación de la posible anormalidad de una observación. Como muestra la **Tabla 5.4**, las observaciones cuyos residuos son más elevados no presentan ninguna evidencia de anormalidad cuando se presta atención a los estadísticos de influencia[19].

- 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 la estimación.

- Para analizar el posible efecto de enmascaramiento provocado por estas observaciones, se ha utilizado el procedimiento de Peña y Yohai (1991) aplicado a la matriz $\tilde{M}$ definida en **[3.4.17]**. En las dos últimas columnas de la **Tabla 5.4** 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

[19] 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.

potencialmente influyente considerando los estadísticos anteriores, tiene un componente asociado relativamente elevado en el primer autovector (aproximadamente el doble que la siguiente observación), 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. Adicionalmente, se consideraron algunas otras observaciones[20] 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 5.5** aparece la estimación de los tres modelos de la **Tabla 5.2**, eliminando el efecto de las observaciones 14, 37 y 61. A efectos de comparación, en la última columna de la tabla se incluye el Modelo 3 estimado hasta la convergencia sin dichas observaciones (Modelo 3*).

Las principales diferencias con respecto a las estimaciones resumidas en la **Tabla 5.2** son:

- Al contrario que con los modelos de la **Tabla 5.2**, 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.

[20] En particular, se llevaron a cabo pruebas incluyendo como influyentes, además de las mencionadas 14, 37 y 61, las observaciones 68 y 69.

Tabla 5.5. 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 ℓ	-42.80	-39.41	-55.46	-22.39
LRT	25.32	32.10		
N^o condición	1.7e+6	1.4e+5	2.2e+6	1.9e+6

- 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 en 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 5.6**, las elasticadades de las variables que ahora son significativas son claramente superiores a las que aparecían en la **Tabla 5.2**.

Tabla 5.6. 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

Finalmente, 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 ellos 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.

5.3. Análisis de los datos de Pregibon (1981)

En Pregibon (1981) se presenta, como ejemplo, el análisis de unos datos sobre vaso-constricción en la piel de los dedos. Los datos proceden de Finney (1947), aunque aparecen listados en la Tabla 1 del artículo, de donde se han tomado. En el **Apéndice A.3** se incluye el listado completo de estas observaciones.

Las variable endógena (VC) es binaria, y está codificada como uno si el individuo presentó vaso-constricción en la piel de los dedos y cero en caso contrario. Las variables exógenas son la tasa (*tasa*) y el volumen (*vol*) de aire inspirado durante una fase pasajera de vaso-constricción de la piel de los dedos. Pregibon (1981) estima un modelo logit con las variables en logaritmos. Utilizando el algoritmo de máxima verosimilitud por procedimientos lineales desarrollado en el **Apartado 1.3.2**, el modelo estimado, numéricamente idéntico al del artículo original, resulta ser:

$$\ln \frac{\hat{P}_i}{1-\hat{P}_i} = \underset{}{-2.875} + \underset{(1.319)}{4.562 \ln tasa_i} + \underset{(1.835)}{5.179 \ln vol_i}$$

$$(1.862)$$

En la **Tabla 5.7** se muestran los estadísticos de diagnóstico para todas las observaciones de la muestra. Por columnas, la **Tabla 5.7** contiene: el número de la observación, el estadístico de distancia definido en **[3.2.2]**, el estadístico $\tilde{h}_i$ de distancia para las variables transformadas con las expresiones **[3.4.4]**-**[3.4.5]**, el residuo definido en **[3.3.1]**, el estadístico de influencia individual de **[3.4.8]** y, por último, los componentes de los autovectores asociados a los dos mayores autovalores de la matriz de influencia $\tilde{M}$ definida en **[3.4.17]**.

Tabla 5.7. Estadísticos de diagnosis para todas las observaciones en la muestra.

i	h_i	$\tilde{h}_i$	e_i	$\hat{c}_i$	λ_i^1	λ_i^2
1	0.1456	0.0927	0.0464	0.0055	-0.0797	-0.0132
2	0.1412	0.0429	0.0179	0.0009	-0.0279	-0.0085
3	0.0391	0.0612	0.0787	0.0059	-0.0289	-0.0208
4	0.0101	0.0867	0.9252	1.2870	0.2137	0.6863
5	0.0453	0.1158	0.2184	0.0414	0.0374	-0.0143
6	0.0550	0.1524	0.2705	0.0787	0.0958	-0.0048
7	0.0296	0.0076	-0.0011	0.0000	-0.0005	-0.0016
8	0.0120	0.0559	-0.5097	0.0652	0.1132	-0.1152
9	0.0047	0.0342	-0.0087	0.0003	0.0000	-0.0102
10	0.0255	0.0072	-0.0009	0.0000	0.0000	-0.0014
11	0.0187	0.0097	-0.0014	0.0000	-0.0002	-0.0020
12	0.0521	0.1481	-0.2047	0.0525	-0.1351	-0.1014
13	0.0512	0.1628	-0.3753	0.1395	-0.2096	-0.1185
14	0.0423	0.0551	0.0615	0.0040	-0.0306	-0.0189
15	0.0590	0.1336	0.1592	0.0337	0.0413	-0.0257
16	0.0801	0.0402	0.0242	0.0011	-0.0259	-0.0109
17	0.1407	0.0172	0.0050	0.0001	-0.0071	-0.0036
18	0.0051	0.0954	0.8941	0.9845	0.0712	0.6072
19	0.0257	0.1315	-0.5346	0.2003	0.4290	-0.1442
20	0.0513	0.0525	0.0547	0.0034	-0.0434	-0.0168
21	0.0747	0.0373	-0.0114	0.0005	-0.0108	-0.0119
22	0.0032	0.1015	-0.1495	0.0221	0.0085	-0.0911
23	0.0134	0.0761	-0.5120	0.0935	0.2433	-0.1278
24	0.0210	0.0717	-0.6519	0.1558	0.3898	-0.0892
25	0.0380	0.0587	0.1007	0.0074	-0.0666	-0.0185
26	0.0235	0.0548	-0.0248	0.0016	-0.0120	-0.0232
27	0.0422	0.0661	0.1173	0.0101	-0.0948	-0.0161
28	0.0139	0.0647	-0.4469	0.0597	0.0221	-0.1190
29	0.0355	0.1682	0.4465	0.1961	-0.4408	0.1233
30	0.0380	0.0507	-0.0097	0.0006	0.0066	-0.0119
31	0.0825	0.2459	0.2776	0.1661	-0.4463	0.0432
32	0.4203	0.0000	-0.0000	0.0000	0.0000	-0.0000
33	0.0151	0.0510	-0.5926	0.0824	0.1302	-0.0922
34	0.0245	0.0601	0.2288	0.0202	-0.0501	-0.0041
35	0.0230	0.0552	0.2260	0.0180	-0.0725	-0.0028
36	0.0484	0.1177	0.1890	0.0352	0.0326	-0.0200
37	0.0139	0.0647	-0.4469	0.0597	0.0221	-0.1190
38	0.0171	0.1000	-0.1919	0.0293	-0.0458	-0.0995
39	0.0176	0.0531	0.3322	0.0295	-0.1384	0.0351

A la vista de los estadísticos de la **Tabla 5.7** se pueden hacer la siguientes afirmaciones:

- Considerando el estadístico $\hat{c}_i$ las observaciones 4 y 18 resultan claramente influyentes, tanto de forma individual como conjunta, puesto que ambas exceden el valor crítico de la distribución χ^2_3 que, para un 10% de confianza, es 0.584.

El valor medio del estadístico resulta ser 0.1. El estadístico de influencia conjunta resulta ser 5.6.

- Las observación 32, así como las 1, 2 y 17 que podían ser potencialmente influyentes, dado que se encuentran alejadas del centro del espacio de las X, no revelaron nada concluyente a la vista de $\hat{c}_i$ y de los componenentes asociados de los autovectores.

- Al analizar la existencia de un posible efecto de enmascaramiento, se observa que el segundo autovector está claramente dominado por las observaciones 4 y 18, ambas influyendo en el mismo sentido (igual signo del componente asociado del autovector). Sin embargo, en el primer autovector se observan dos grupos de observaciones dominantes, el formado por la 19 y la 24, ambas con signo positivo, y el formado por la 29 y 31, ambas con signo negativo. Los estadísticos de influencia conjunta resultaron ser 1.26 para el par {31, 29} y 0.85 para el par {24, 19}. Por tanto, parece que, al menos el primer grupo es anómalo (un valor elevado del estadístico de influencia conjunta) y está enmascarado por las observaciones 4 y 18 que tienen el mismo sentido de influencia (mismo signo en los componentes de los autovectores).

- En su trabajo, Pregibon (1981) sólo detecta las observaciones 4 y 18, puesto que sólo éstas muestran tanto un residuo como un estadístico individual elevado. No obstante, el grupo {4, 18, 29 , 31} acumula una influencia de 9.48, siendo el conjunto de mayor influencia. Nuevamente, y a falta de información adicional, se puede eliminar este conjunto de observaciones.

En la **Tabla 5.8** se muestran las estimaciones del modelo eliminando el efecto del grupo de observaciones de la primera fila de la tabla. En la fila inferior, aparece el estadístico de influencia conjunta.

Tabla 5.8. Estimaciones del modelo de Pregibon eliminando distintos conjuntos de observaciones y estadísticos de influencia conjunta.

$I =$	$\{\varnothing\}$	{4, 18}	{29, 31}	{19, 24}	{4, 18, 29, 31}
Const	-2.8754	-5.8591	-3.4866	-2.3026	-6.7427
Tasa	4.5617	8.1523	5.2717	4.0575	9.1817
Vol	5.1793	9.0431	4.7612	5.3940	8.6468
$\hat{C}_I$	--	5.5680	1.2620	0.8459	9.4840

Como puede observarse, el efecto del grupo {29, 31} es muy pequeño, especialmente si se compara con el efecto de los restantes. El grupo {4, 18} es el de mayor influencia, pero cuando se añade el par {29, 31} el cambio resulta aún más apreciable.

Como conclusión de esta sección cabe decir que, para detectar observaciones anómalas, no es suficiente con el análisis de residuos y los estadísticos de influencia individal, sino que, además, es necesario analizar la influencia de grupos de observaciones. Para seleccionar eficientemente dichos grupos de observaciones, el método propuesto por Peña y Yohai (1991), aplicado a una matriz de influencia propia de los modelos de elección cualitativa, mantiene la validez del planteamiento aplicado a los modelos de regresión lineal.

CONCLUSIONES

Conclusiones generales

La principal conclusión que se puede obtener de este trabajo es que, como se ha pueste de relieve, la presencia de observaciones anómalas en modelos de elección cualitativa tiene como consecuencia la inconsistencia del estimador máximo-verosímil. La detección de observaciones anómalas en dicha clase de modelos presenta diferencias sustanciales con respecto a los modelos lineales, o a los lineales generalizados, que hacen necesario derivar una metodología propia para abordar el problema.

En este trabajo se ha analizado el efecto de las observaciones anómalas bajo planteamientos paramétricos y se han derivado instrumentos de detección apropiados para modelos de elección discreta. Las principales aportaciones de este trabajo pueden resumirse en los siguientes puntos:

- El estimador máximo-verosímil en los modelos de variable dependiente cualitativa, ante la presencia de observaciones anómalas, es inconsistente. Se demuestra que la inconsistencia depende tanto de la proporción de observaciones anómalas existentes como de los parámetros que caracterizan la distribución de dichos datos.

- Los residuos, y en general, la mera extrapolación de los planteamientos para el modelo lineal general, no son un instrumento apropiado. Se han derivado estadísticos específicos que miden el efecto de una observación o conjunto de observaciones sobre el vector de parámetros del modelo. En general, es posible adaptar los instrumentos utilizados para el modelo lineal general o los modelos lineales generalizados, pero estas extensiones debe llevarse a cabo considerando las particularidades de los modelos de variable dependiente cualitativa.

- Se han derivado estadísticos de influencia para su aplicación tanto a modelos de elección binaria como múltiple. Además, se han particularizado dichos estadísticos para el análisis de influencia sobre grupos de parámetros o conjuntos de observaciones.

- Utilizando experimentos de Monte Carlo, se han comprobado los efectos de las observaciones anómalas que sugerían los desarrollos teóricos y se han validado los estadísticos propuestos, poniendo de relieve sus limitaciones cuando aparecen efectos de enmascaramiento.

- Definiendo las matrices de influencia adecuadas, se ha extendido el plantemiento de Peña y Yohai (1991) para tratar el enmascaramiento en modelos de elección binaria. También se ha comprobado su funcionamiento, tanto con datos simulados como con datos reales.

- Se ha desarrollado un algoritmo de máxima verosimilitud por procedimientos lineales para modelos de elección cualitativa múltiple. Este algoritmo, semejante al de Amemiya (1985), permite estimar eficientemente los modelos eliminando conjuntos de observaciones, lo que hace posible la derivación de estadísticos de influencia para esta clase de modelos.

Finalmente, con el conjunto de instrumentos desarrollados, se ha planteado una metodología de diagnosis y detección aplicada a dos muestras de datos.

Extensiones

Una primera extensión posible, es la derivación de una versión del algoritmo EM particularizada para la estimación de modelos con anomalías. La idea básica es separar los conjuntos de observaciones anómalas y no anómalas utilizando los estadísticos de influencia y estimar los parámetros característicos de ambos conjuntos mediante un procedimiento iterativo que contemplara ambos pasos.

Una segunda extensión importante se centra en un análisis exhaustivo de los estadísticos LM propuestos para contrastar las hipótesis correspondientes sobre la procedencia 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 funcionan bien para contrastar si determinados conjuntos de observaciones lo son.

Una última extensión relevante es el análisis de valores críticos de los estadísticos de influencia propuestos. Si bien parece difícil determinar sus distribuciones teóricas, una extensión de gran utilidad puede ser el desarrollo de distribuciones empíricas que permitan la clasificación de observaciones de una forma *objetiva*, sin requerir el análisis pormenori-

zado del investigador. Esta idea también es aplicable al procedimiento de Peña y Yohai (1991) para determinar grupos de observaciones influyentes.

APÉNDICES

A.1. Concavidad Global de las Funciones de Verosimilitud de los MEB

Es importante analizar la concavidad de las funciones de verosimilitud de los modelos binarios y, especialmente, de los modelos probit y logit por dos razones: i) para una función de verosimilitud estrictamente cóncava, si existe un EMV, este será único, y ii) desde el punto de vista del procedimiento iterativo no restringido que se utilice, a pesar del incumplimiento de alguna de las hipótesis realizadas sobre la función de verosimilitud, la concavidad asegura la convergencia de los algoritmos.

Aunque la concavidad de la función de verosimilitud para probit y logit binarios ha sido ampliamente demostrado en la literatura [Amemiya (1985) es un ejemplo], en un artículo reciente Núñez (1990) plantea una clase general de modelos de elección binaria y deriva condiciones bajo las cuales la función de verosimilitud será estrictamente cóncava. A continuación se presentan los principales resultados de dicho trabajo.

Definición D.A.1.1. Clase Ψ de funciones de densidad. Sea $f(\cdot)$ la función de densidad de una variable aleatoria de tipo continuo. Se dice que $f \in \Psi$ si y sólo si se verifican las siguientes condiciones:

A. $f'(x) > 0, \forall\, x < 0$, y

B. $f(x) = f(-x), \forall\, x \in \mathbb{R}$.

Partiendo de la definición anterior, las principales características de las funciones que pertenecen a dicha clase pueden resumirse en:

1) Son funciones continuas y estrictamente positivas sobre $\mathbb{R}$.
2) Si $f \in \Psi$, $f'(x) < 0, \forall\, x > 0$.
3) Alcanzan su máximo en $x=0$ y éste es único.
4) Si $f \in \Psi$, se cumple que

Algunas densidades pertenecientes a la clase definida son la normal, la logística o la Cauchy.

$$\int_{-\infty}^{0} f(x)\,dx = \int_{0}^{\infty} f(x)\,dx = \frac{1}{2} \qquad \textbf{[A.1.1]}$$

Lema L.A.1.1. Sea el MEB definido en **[1.2.13]**. Si F(·) es la función de distribución correspondiente a una función de densidad $f(\cdot)$ tal que $f \in \Psi$, se tiene que si ln $F(\cdot)$ es estrictamente cóncava, entonces ln L también lo es.

Lema L.A.1.2. Sea $f(\cdot)$ una función de densidad de la clase Ψ, y sea $F(\cdot)$ su función de distribución asociada. Entonces:

A. ln $F(\cdot)$ es estrictamente cóncava en $\mathbb{R}^+$.

B. Si además existe $f'(0)$, entonces ln $F(\cdot)$ es estrictamente cóncava en $\mathbb{R}^+ \cup \{0\}$.

Teorema T.A.1.1. Condición suficiente de concavidad estricta de ln $L(\theta)$. Sea un MEB como el planteado en **[1.2.13]**, y sea $f(\cdot)$ la función de densidad correspondiente a la distribución $F(\cdot)$. Consideremos la función $g(\cdot)$ definida por:

$$g(z) = \frac{F(z)}{f(z)}, \qquad \forall\, z \in \mathbb{R} \qquad \textbf{[A.1.2]}$$

si se cumplen las siguientes condiciones:

A. $f \in \Psi$.

B. Existe una función $h(\cdot)$ continua, tal que:

1) $f(z) = c \exp[\int h(z)\,dz]$ donde c es una constante, y

2) $h(z)\cdot g(z) < 1, \forall\, z \in \mathbb{R}^-$.

entonces la función de verosimilitud ln $L(\cdot)$ es estrictamente cóncava.

La aplicación inmediata de estos resultados es la demostración de la concavidad en el caso del modelo logit y el modelo probit, y la única dificultad estriba en encontrar la función $h(\cdot)$ apropiada. Para el primero de ellos se puede comprobar que:

$$f(z) = \exp\left[\int \frac{1-e^z}{1+e^z}\,dz\right] \qquad \textbf{[A.1.3]}$$

por lo que $h(z) = (1-e^z)/(1+e^z)$, $c = 1$ y además $h(z)g(z) = 1 - e^z < 1, \forall\, z \in \mathbb{R}^-$.

Para el modelo probit se tiene que:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\int -z\, dz \right) \qquad \textbf{[A.1.4]}$$

donde en este caso $h(z) = -z$ y $c = 1/\sqrt{2\pi}$. Además:

$$h(z)g(z) = -z \int_{-\infty}^{z} \exp\left[\frac{z^2 - t^2}{2} \right] dt =$$

$$= \exp(z^2/2) \int_{-\infty}^{z} -z \exp(-t^2/2)\, dt < \qquad \textbf{[A.1.5]}$$

$$< \exp(z^2/2) \int_{-\infty}^{z} -t \exp(-t^2/2)\, dt = 1$$

Para una función de verosimilitud obtenida a partir de una distribución general será necesario demostrar que el hessiano es una matriz *definida negativa* que es condición suficiente [Bazaraa y Shetty (1979)]. A partir de la expresión **[1.3.4]**, el hessiano de la función de verosimilitud se puede escribir como:

$$\nabla^2 \ell = -\sum_{i=1}^{n} \frac{f_i[(y_i - 2y_iF_i + F_i^2)f_i + (y_i - F_i)F_i(1 - F_i)x_i^T\beta]}{[F_i(1-F_i)]^2} \qquad \textbf{[A.1.6]}$$

y siguiendo a Amemiya (1985), es necesario demostrar la positividad del numerador de **[A.1.6]**, que puede plantearse como una función:

$$g(y,z) \equiv (y - 2yF + F^2)f + (y - F)F(1 - F)z \qquad \textbf{[A.1.7]}$$

y es necesario comprobar que es estrictamente positiva para $y = 0$, 1 y cualquier valor de z.

A.2. Notas sobre Métodos Numéricos de Optimización No Restringida

El método de estimación es un aspecto fundamental de cualquier investigación sobre modelos econométricos, sobre todo cuando los modelos implicados son no lineales. Las funciones de verosimilitud de los modelos tratados en este trabajo son cóncavas, lo que evita el posible problema de óptimos parciales. Pese a estas buenas propiedades, seleccionar un método de optimización eficiente sigue siendo un aspecto central del problema [ver Bunch (1988)].

En este apéndice se presentan un conjunto de técnicas generales de optimización y se analiza con especial atención su aplicación a problemas de máxima verosimilitud. Para un análisis en profundidad de estos temas se pueden consultar: Dennis y Schnabel (1983) o Gill et al. (1981), y para revisiones de los métodos Goldfeld y Quandt (1972) y Quandt (1983).

A.2.1. Planteamiento del problema

Sea el problema:

$$\max f(x_1,\ldots,x_n) \qquad \textbf{[A.2.1]}$$

respecto al vector $x = (x_1,\ldots,x_n)^T$, donde f: $\mathbf{R}^n \rightarrow \mathbf{R}$ es un campo escalar que, en general, supondremos dos veces diferenciable, con derivadas continuas, aunque este supuesto no siempre es necesario. El procedimiento analítico general para resolver el problema, consiste en plantear y resolver el sistema de ecuaciones que viene dado por sus *condiciones de primer orden*:

$$\frac{\partial f(x^*)}{\partial x} = \mathbf{0}_n \qquad \textbf{[A.2.2]}$$

Una vez determinado el conjunto de vectores $x^* \in \mathbf{R}^n$ que satisfacen estas condiciones, debe comprobarse cuáles de estos puntos satisfacen también las *condiciones de segundo orden*, esto es, que la matriz hessiana $H(x^*)$ sea definida negativa.

El *enfoque analítico* para resolver problemas de optimización, se caracteriza por la generalidad de su planteamientos, así como un elevado rigor matemático. Sin embargo,

muchos problemas prácticos no pueden resolverse analíticamente. En el extremo opuesto, los *métodos numéricos* son poco generales, ya que cada algoritmo está especializado en unos pocos casos concretos y su fundamento matemático es poco riguroso; a cambio, pueden resolver numerosos problemas prácticos.

En este Apéndice, trataremos algunos aspectos de la resolución numérica de programas matemáticos. Para ello, se presentan distintas técnicas que permiten calcular una sucesión de valores x^0 x^1,...,x^k que, idealmente, convergerá a la solución óptima del problema.

Comenzaremos definiendo un esquema que contiene todos los elementos que se encuentran en los métodos basados en direcciones de búsqueda, que son los de uso más frecuente en aplicaciones econométricas. Para la definición matemática de este algoritmo−tipo, emplearemos la siguiente notación:

x^*: Óptimo del problema.
f^k: Valor de la función objetivo alcanzado en la k−ésima iteración.
x^k: Estimación de x^* en la k−ésima iteración.
p^k: Dirección de búsqueda del óptimo utilizada en la k−ésima iteración.
g^k: Gradiente de $f(x)$ evaluado en x^k.
G^k: Hessiano de $f(x)$ evaluado en x^k.

Asimismo, todos los algoritmos que discutiremos comparten una serie de elementos comunes: i) se dispone de unas *condiciones iniciales*, esto es, un valor inicial del vector de variables x^0 para iniciar el proceso de cálculo, ii) un criterio para determinar en qué dirección del espacio $\mathbb{R}^n$ se encuentran soluciones mejores que la actual (*vector de paso*), iii) un criterio para determinar cuánto hay que avanzar en la dirección del vector de paso (*longitud del paso*), y por último, iv) un *criterio de convergencia* que permita determinar si la solución actual cumple las condiciones de primer orden con el grado de precisión requerido. Debido a la existencia de estos elementos comunes, todos los algoritmos que vamos a tratar pueden describirse dentro del siguiente esquema de cálculo:

Paso 0: Inicialización: Situar el contador de iteraciones k en cero. Seleccionar arbitrariamente una estimación inicial del óptimo x^0[21], el máximo número de iteraciones admisibles K y una tolerancia para los criterios de parada e.

[21] La elección de x^0 puede ser un problema serio en sí mismo, puesto que generalmente la convergencia es más rápida cuanto más próximo se encuentre x^0 al máximo. Por otra parte, en problemas de estimación, es conveniente (cuando no necesario) que x^0 sea una estimación consistente, aunque esto es irrelevante a efectos computacionales.

Paso 1: Comprobar si se cumplen las condiciones de convergencia en la iteración actual. Si se cumplen, se toma x^k como aproximación suficiente a x^* y finaliza el proceso. Si no se cumplen, se sigue con el paso 2.

Paso 2: Determinar una dirección de búsqueda p^k.

Paso 3: Determinar una longitud de paso α^k.

Paso 4: $x^{k+1} = x^k + \alpha^k p^k$.

Paso 5: Hacer $k = k + 1$. Si $k \leq K$, volver a comenzar una iteración en el paso 1, en otro caso, detener el proceso.

Por otra parte, consideraremos que la característica esencial de un algoritmo es la forma en la que se genera la secuencia de vectores p^k. Los demás aspectos, aún siendo importantes, no parecen suficientemente sustanciales como para que una variación en las mismas caracterice un nuevo algoritmo.

La elección de un algoritmo debe hacerse desde distintos puntos de vista que suelen implicar un intercambio de ventajas e inconvenientes, ya que no existe un algoritmo que sea el mejor desde todos los puntos de vista y para todos los problemas que pueden plantearse. Dos aspectos fundamentales a la hora de elegir un algoritmo son: i) su *robustez*, esto es, el grado hasta el que el algoritmo en cuestión sea capaz de dar una estimación de x^k del verdadero máximo x^* tal que $\|x^k - x^*\| < e$, para algún e dado y positivo; ii) su *coste computacional*, que se supone, de algún modo, proporcional al número de iteraciones y evaluaciones de la función objetivo, memoria de ordenador, tiempo de cálculo, operaciones de lectura/escritura, etc y iii) sus propiedades específicas para resolver la familia de problemas que nos interesan.

A.2.2. Criterios de convergencia

El criterio de convergencia es un elemento arbitrario del algoritmo de forma que, si se cumple la condición que lo caracteriza, se considera que el algoritmo ha alcanzado un solución satisfactoria. Sea x^{k+1} la estimación actual de x^* obtenida mediante un algoritmo cualquiera y sea una tolerancia arbitraria $e > 0$. En estas circunstancias, algunos posibles criterios de terminación son:

A. $| \alpha^k | \leq e$, esto es, detener el algoritmo si el paso es *pequeño*. Un criterio semejante al anterior puede plantearse de la siguiente manera:

$$\left[\frac{1}{n} \sum_{i=1}^{n} (x_i^{k+1} - x_i^k)^2 \right]^{1/2} \leq e \qquad \textbf{[A.2.3]}$$

B. $| f^{k+1} - f^k | \leq e$, intuitivamente, detener el algoritmo cuando la mejora en la función objetivo es *pequeña*.

C. $\| g^{k+1} \| \leq e$, que es equivalente a una comprobación aproximada del cumplimiento de las condiciones necesarias de primer orden.

D. $| g_i^{k+1} | \leq e \ (i = 1, \ldots, n)$, que supone una comprobación de las condiciones necesarias más estricta que con el criterio C.

Existen otros criterios para decidir si el proceso de cálculo debe detenerse en la iteración actual. Sin embargo, todos ellos se basan en las mismas ideas: i) comprobar el cumplimiento aproximado de las condiciones de primer orden o bien ii) detener el proceso iterativo si el paso es pequeño o iii) si la función objetivo mejora poco.

Todos estos criterios son válidos aunque tienen el defecto de que son sensibles a la métrica en que está definido el problema. Por ello, si un mismo algoritmo va a utilizarse para resolver problemas definidos en distintos espacios métricos, resulta difícil determinar un valor adecuado de e. Para resolver esta dificultad, conviene usar un criterio de convergencia adimensional.

Otra consideración relevante es la velocidad a la que converge el algoritmo, que generalmente vendrá determinada por la elección del modo de cálculo de $\boldsymbol{p}^k$. Se dice que un algoritmo es *cuadráticamente convergente* si alcanza el máximo de una función cuadrática acotada en una iteración. Asimismo, se dice que un algoritmo es *linealmente convergente* si alcanza el óptimo de una función lineal acotada en una iteración. Por último, se dice que una algoritmo posee *convergencia superlineal* si, aplicado a una función cuadrática acotada, converge al óptimo en un número finito de iteraciones. Para una discusión más formal sobre las propiedades de convergencia puede verse Bazaraa y Shetty (1979) y Dennis y Schnabel (1983).

Las nociones de convergencia cuadrática y superlineal son importantes, puesto que a menudo se puede suponer que la función objetivo es aproximadamente cuadrática (al menos en un entorno de óptimo) de forma que $f(x)$ queda bien aproximada por un desarrollo de Taylor de segundo orden. Un problema adicional que surge en problemas de

estimación es los que, si el resto de la aproximación no converge a cero cuando el tamaño de la muestra empleada aumenta arbitrariamente, la estimación de los parámetros en el óptimo puede estar sesgada [Cox y Hinkley (1974)].

A.2.3. Criterios para determinar la longitud de paso

En la práctica, existen muchos criterios para determinar la longitud del paso α^k en una iteración cualquiera. La mayor parte de estos métodos intenta garantizar que, una vez realizada la iteración, no se produzca un empeoramiento en el valor de la función objetivo. Una discusión de estos procedimientos puede encontrarse en Bazaraa y Shetty (1979) y Dennis y Schnabel (1983).

En realidad, sólo hay un criterio que tenga una base objetiva, el de la *longitud de paso óptima.* La idea en que se basa el procedimiento es sencilla. Una vez elegido el vector de desplazamiento $\boldsymbol{p}^k$, el valor de la función objetivo que se alcanzará después del paso es función de una sola variable:

$$f(\boldsymbol{x}^{k+1}) = f(\boldsymbol{x}^k + \alpha^k \boldsymbol{p}^k) \qquad \textbf{[A.2.4]}$$

donde sólo se desconoce α^k, que será el valor positivo que produzca un aumento mayor de la función objetivo. Por tanto, es posible calcular de forma óptima la longitud del desplazamiento utilizando un algoritmo eficiente de optimización para problemas de una sóla variable como, por ejemplo, el de Fibonacci [ver Gill et al. (1981)].

A.2.4. Métodos tipo Newton

El punto de partida de los métodos tipo Newton, consiste en aproximar $f(\boldsymbol{x})$ por un desarrollo en serie de Taylor de segundo orden alrededor de $\boldsymbol{x}^k$ y optimizar la función resultante, esto es, resolver el problema:

$$Max \quad f(\boldsymbol{x}^k) + \boldsymbol{g}^{kT}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{1}{2}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)^T \boldsymbol{G}^k (\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) \qquad \textbf{[A.2.5]}$$

Aplicando la condición necesaria de primer orden a la aproximación **[A.2.14]** y despejando $\boldsymbol{x}^{k+1}$, el paso de Newton-Raphson resulta:

$$x^{k+1} = x^k - [G(x^k)]^{-1} g(x^k) \qquad \textbf{[A.2.6]}$$

Este criterio proporciona, por tanto, el máximo aumento de valor de la aproximación de segundo orden de la función en un entorno de x^k si G^k es una matriz estrictamente definida negativa.

La expresión anterior es la base de los denominados algoritmos *tipo Newton*, cuyo objetivo es reducir el coste computacional manteniendo las propiedades de convergencia del algoritmo de Newton en el que se basan. Para ello, sustituyen el hessiano en **[A.2.16]** por aproximaciones adecuadas. La expresión general de estos algoritmos es:

$$x^{k+1} = x^k - \alpha^k [H^k]^{-1} g(x^k) \qquad \textbf{[A.2.7]}$$

donde la dirección de búsqueda es $p^k = -[H^k]^{-1} g(x^k)$.

Algunas elecciones habituales de H y α en problemas de estimación son: i) la matriz de información de los parámetros y $\alpha^k = 1$, lo que produce el algoritmo de *scoring*:

$$-H^k = -E\left[\frac{\partial^2 \ell}{\partial \theta \, \partial \theta^T}\right]_{\theta^k} = I(\hat{\theta}^k) \qquad \textbf{[A.2.8]}$$

donde ℓ es, normalmente, el logaritmo de la función de verosimilitud objetivo y θ el vector de parámetros a estimar, ii) utilizar la forma del producto exterior del gradiente:

$$H^k = \left[\frac{\partial \ell}{\partial \theta}\right]^T \left[\frac{\partial \ell}{\partial \theta}\right]_{\theta^k} \qquad \textbf{[A.2.9]}$$

y $\alpha^k = 1$ que da lugar al *algoritmo Gauss–Newton*, que se suele emplear en problemas en los que la función objetivo es una suma de cuadrados; o bien iii) hacer $H = -I_n$ y $\alpha^k = 1$ que es el *método del gradiente o máximo ascenso*[22], cuyas propiedades de convergencia lo hacen adecuado para iniciar cualquier proceso iterativo.

El algoritmo de Newton-Raphson posee excelentes propiedades de convergencia *local*; esto es, si se parte de un punto relativamente cercano al óptimo y además requiere que H sea definida negativa. De no cumplirse este último requisito, las iteraciones podrían representar alejamientos del objetivo. Para evitar este inconveniente se han sugerido diferentes procedimientos, pero el más simple de utilizar y que ofrece resultados óptimos

[22] El método del gradiente también puede derivarse directamente optimizando un desarrollo en serie de primer orden en un entorno del vector de prueba actual, añadiendo una restricción de normalización arbitraria para acotar el resultado.

consiste en perturbar los elementos de la diagonal principal de H con la cantidad mínima que asegure que la matriz es definida negativa. Este planteamiento da lugar al *algoritmo de ascenso cuadrático* [Goldfeld, Quandt y Trotter (1966)].

A.2.5. Métodos *quasi−Newton*

La idea fundamental tras estos métodos es similar a los expuestos en el apartado anterior: mantener las buenas propiedades del algoritmo de Newton pero reduciendo el coste de cálculo en cada iteración y garantizando el condicionamiento en signo de los autovalores de H^k. Para conseguir esto, los métodos *quasi−Newton* optan por actualizar (en vez de calcular como los tipo Newton) la matriz H en cada iteración, utilizando para ello información de las primeras derivadas y los resultados de la iteración anterior. Este planteamiento da lugar a una amplia familia de algoritmos de optimización eficientes, con elevadas tasas de convergencia y computacionalmente menos costosos que los expuestos más arriba. Las características de esta clase de algoritmos pueden resumirse en: i) en cada iteración, la búsqueda es unidimensional, ii) el algoritmo sólo calcula valores de la función objetivo y primeras derivadas, y iii) en la iteración k−ésima, sólo emplea información calculada en las iteraciones k y $k-1$.

El desarrollo de los métodos quasi−Newton parte de considerar la expansión del gradiente alrededor de x^k en la dirección $s^k = x^{k+1} - x^k = \alpha^k p^k$:

$$g(x^k + s^k) \qquad \textbf{[A.2.10]}$$

La curvatura de F en la dirección s^k está dada por $s^k G^k s^k$, que puede ser aproximada usando información de primer orden:

$$s^{kT} G^k s^k \approx (g(x^k + s^k) - g^k)^T s^k \qquad \textbf{[A.2.11]}$$

Esta relación será exacta para un modelo cuadrático, y se puede suponer que la aproximación es adecuada en un entorno del máximo.

Al comienzo de la iteración k por un método quasi−Newton se dispone de un *hessiano aproximado* B^k y, normalmente, se comienza con $B^0 = I_n$ o con un Hessiano calculado, si es posible. Si se considera B^k como el hessiano de función cuadrática, la dirección de búsqueda es la solución al sistema:

$$B^k p^k = g^k \qquad \textbf{[A.2.12]}$$

Una vez se ha calculado x^{k+1}, se obtiene una nueva aproximación al hessiano actualizando B^k, de acuerdo con la expresión:

$$B^{k+1} = B^k + U^k \qquad \textbf{[A.2.13]}$$

donde U^k es la matriz de actualización que se calcula a partir de la información de las primeras derivadas. Además, el hessiano actualizado debería aproximar la curvatura de F en la dirección del vector de paso s^k, es decir, cumplir lo que se denomina *condición quasi−Newton*:

$$B^{k+1}s^k = y^k \qquad \textbf{[A.2.14]}$$

donde $y^k = g(x^{k+1}) - g(x^k)$. La selección apropiada de U^k requiere imponer que la nueva actualización sea simétrica y cumpla la condición quasi−Newton. Además, la matriz resultante debería ser definida negativa. Diferentes elecciones de U^k dan lugar a distintos algoritmos de una misma familia.

Si se hace $U_k = uv^T$ con $u = y^k - B^k s^k$ y se imponen las restricciones expuestas, se puede obtener una familia de actualizaciones denominada *Powell−Symmetric−Broyden* [Gill et al. (1981)]. Haciendo $v = y^k - B^k s^k$ se obtiene la actualización *Davidon-Fletcher-Powell (DFP)*:

$$B^{k+1} = B^k - \frac{1}{s^{kT}B^k s^k} B^k s^k s^{kT} B^k + \frac{1}{y^{kT}s^k} y^k y^{kT} + (s^{kT}B^k s^k) w^k w^k \qquad \textbf{[A.2.15]}$$

donde:

$$w^k = \frac{1}{y^{kT}s^k} y^k - \frac{1}{s^{kT}B^k s^k} B^k s^k \qquad \textbf{[A.2.16]}$$

En general, se considera que la mejor de toda la familia de actualizaciones es el denominado *método Broyden−Fletcher−Goldfarb−Shanno (BFGS)*, cuya actualización es:

$$B^{k+1} = B^k - \frac{1}{s^{kT}B^k s^k} B^k s^k s^{kT} B^k + \frac{1}{y^{kT}s^k} y^k y^{kT} \qquad \textbf{[A.2.17]}$$

Si además del empleo de alguna de estas actualizaciones, se determina la longitud de paso de forma óptima, esto es, se elige α^k tal que para x^k y p^k dados se cumpla:

$$\frac{\partial F(x^k + \alpha^k p^k)}{\partial \alpha^k} = 0 \qquad \textbf{[A.2.18]}$$

los algoritmos expuestos generan direcciones conjugadas y se asegura que la actualización del hessiano es definida negativa.

Una variación de los métodos *quasi*−Newton es el algoritmo *BHHH* [Brendt et al. (1974)], que utiliza técnicas de actualización como las expuestas, pero aplicadas la matriz de información del lugar del hessiano. A pesar de que fue concebido para problemas de estimación, hay evidencias de que su rendimiento general no es superior al de otros métodos tipo Newton o quasi−Newton, y no es muy empleado en la práctica.

A.2.6. Métodos que no emplean derivadas

En principio, los métodos de optimización que no requieren usar derivadas son atractivos ya que: i) en ocasiones no se conocen las expresiones analíticas de las derivadas de la función objetivo y ii) el cálculo de derivadas es una tarea costosa en términos de tiempo de cálculo. Por otra parte, estos métodos suelen tener garantizada su convergencia a algún máximo, con independencia de la concavidad de la función objetivo. Sus inconvenientes principales son dos: lentitud en la convergencia, incluso en la proximidad del óptimo, y elevado coste computacional, al tener que evaluar la función objetivo un elevado número de veces.

Una clase de algoritmos sin derivadas, emplea la noción de búsqueda en una rejilla de puntos. Un procedimiento sencillo se obtiene comenzando en algún punto x^0 y evaluar la función en x^0 y en los $2n$ puntos de la rejilla dada por $x^0 \pm hv_i$, donde v_i $(i = 1,\ldots,n)$ es un vector con un uno en la posición i−ésima y ceros en el resto, y h es la anchura de la rejilla. Se pasa de x^0 a un x^1 tal que $f(x^1) = \sup f(x^0 \pm hv_i)$. El procedimiento se repite partiendo de x^1 hasta que no se obtiene mejora. En ese caso, se reduce la anchura de la rejilla h y se continúa hasta alcanzar un valor de h especificado de antemano que será la precisión con la que se obtiene el óptimo.

Un método de búsqueda alternativo y más eficiente es el de Hooke y Jeeves (1961). Este algoritmo emplea dos tipos de movimientos: i) *búsquedas exploratorias*, que se realizan en direcciones paralelas a los ejes de coordenadas, y ii) *búsquedas patrón*, que se hacen en una dirección dada por una combinación lineal de las direcciones de las búsquedas exploratorias anteriores. Si una búsqueda exploratoria y la siguiente búsqueda

patrón resultan en una mejora de la función objetivo, se aceptan; de lo contrario, se hace un movimiento en la dirección exploratoria. En general, se comienza con un valor fijado de h y se continúa hasta que se ha reducido lo suficiente, aunque una modificación con la que se gana eficiencia consisten en calcular h en cada iteración.

Un problema serio con el tipo de algoritmos que sólo modifican una variable en cada paso es que pueden no converger todas las variables simultáneamente. Aunque en general funcionan bien si se detiene el proceso cuando la mejora en la función objetivo es pequeña, el gradiente puede ser no nulo a lo largo del recorrido del algoritmo e incluso es posible que itere indefinidamente.

Por otra parte, también es posible combinar los algoritmos expuestos en el apartado anterior con un criterio de aproximación numérica, que permite llevar a cabo el proceso de optimización sin un conocimiento explícito de la forma funcional de las primeras y segundas derivadas de la función objetivo.

Por ejemplo, para el caso de funciones de una sóla variable, se puede obtener una buena aproximación del valor de la primera derivada utilizando la siguiente expresión:

$$\hat{f}'(x) = \frac{\dfrac{f(x+\Delta x)-f(x)}{\Delta x} - \dfrac{f(x-\Delta x)-f(x)}{\Delta x}}{2} \qquad \textbf{[A.2.19]}$$

que corresponde al método de *aproximación por diferencias centrales*. Aplicando un procedimiento similar, también podría obtenerse una aproximación numérica a los valores de la segunda derivada.

El uso de estas técnicas tiene la ventaja indudable de que permite trabajar sin conocer la forma funcional exacta de las derivadas de la función objetivo. Sin embargo, siempre tienen un cierto coste en términos de precisión.

A.2.7. Un algoritmo especializado: El algoritmo EM

Un algoritmo especialmente efectivo en problemas con datos incompletos (muestras censuradas o truncadas) o con variables no observables, es el denominado *algoritmo EM*,

cuyas propiedades básicas se exponen en Dempster et al. (1977) y algunas extensiones en Ruud (1991).

El problema de datos incompletos puede plantearse suponiendo una variable aleatoria x con función de densidad $f(x \mid \theta)$, además se supone que existe una relación con una variable y observable. El problema se produce debido a que una observación de y no identifica de forma única a la x correspondiente, aunque puede estimarse la probabilidad de que la observación y haya sido generada por un conjunto de x. Las observaciones de y se suponen generadas por:

$$g(y \mid \theta) = \int_{X} f(x \mid \theta)\, dx \qquad \textbf{[A.2.20]}$$

Un ejemplo sencillo es el modelo de *regresión cambiante (switching regression)* con una estructura:

$$\begin{aligned} y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + u_{1i} \quad \text{con probabilidad } \pi \\ y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta}_2 + u_{2i} \quad \text{con probabilidad } 1-\pi \end{aligned} \qquad \textbf{[A.2.21]}$$

donde $\boldsymbol{x}_i$ son variables exógenas, que pueden ser no observables, y_i es la variable endógena, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ son vectores de parámetros desconocidos y u_{i1}, u_{i2} son las perturbaciones habituales. La probabilidad π es desconocida y tampoco se conoce a que régimen pertenecen las observaciones.

Los pasos fundamentales del algoritmo EM son el *paso E (Esperanza)* y el *paso M (Maximización)*, que se llevan a cabo en cada iteración y que pueden resumirse en:

Paso E: Dado un valor del vector de parámetros $\boldsymbol{\theta}^k$ y los datos observados $\boldsymbol{y}$, obtener estimaciones de $\boldsymbol{x}^k$ mediante $E(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^k)$, que puede ser una mezcla de funciones discretas y continuas. Las expresiones a utilizar dependerán del problema concreto que se trate.

Paso M: Usando los valores estimados $\boldsymbol{x}^k$, maximizar la función de verosimilitud del problema completo (como si se observaran las $\boldsymbol{x}$) para obtener $\boldsymbol{\theta}^{k+1}$. Esta etapa se llevará a cabo utilizando algún algoritmo de los expuestos anteriormente.

Si la secuencia de estimaciones obtenidas converge, ese punto es un punto estacionario, generalmente un máximo local, de la función objetivo.

A.3. Datos de los ejemplos del Capítulo 5

Tabla A.1. 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 5.1**.

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
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
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

ADJ	FI	MAR	YLD	PTS	MAT	BA	BS	FTB	CB	MC	SE	MOB	NW	LA	STL
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

Tabla A.2. Datos de Pregibon (1981) tomados de la Tabla 1, pag. 709, del artículo original. Las definiciones se encuentran en la **Sección 5.2**.

VC	TASA	VOL
1	0.825	3.70
1	1.090	3.50
1	2.500	1.25
1	1.500	0.75
1	3.200	0.80
1	3.500	0.70
0	0.750	0.60
0	1.700	1.10
0	0.750	0.90
0	0.450	0.90
0	0.570	0.80
0	2.750	0.55
0	3.000	0.60
1	2.330	1.40
1	3.750	0.75
1	1.640	2.30
1	1.600	3.20
1	1.415	0.85
0	1.060	1.70
1	1.800	1.80
0	2.000	0.40
0	1.360	0.95
0	1.350	1.35
0	1.360	1.50
1	1.780	1.60
0	1.500	0.60
1	1.500	1.80
0	1.900	0.95
1	0.950	1.90
0	0.400	1.60
1	0.750	2.70
0	0.030	2.35
0	1.830	1.10
1	2.200	1.10
1	2.000	1.20
1	3.330	0.80
0	1.900	0.95
0	1.900	0.75
1	1.625	1.30

REFERENCIAS

Altman, E. I., R. B. Avery, R. A. Eisenbeis y J. F. Sinkey (1981). *Application of classification techniques in business, banking and finances*. Greenwich C.T., JAI Press.

Amemiya, T. (1974). "The Nonlinear Two-Stage Least-Squares Estimator", *Journal of Econometrics*, 2, 105-110.

Amemiya, T. (1978a). "On a Two-Step Estimation of a Multivariate Logit Model", *Journal of Econometrics*, 8, 13-21.

Amemiya, T. (1978b). "The Estimation of a Simultaneous Equation Generalized Probit Model", *Econometrica*, 46, 1193-1205.

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

Amemiya, T. (1983). "Non-Linear Regression Models", en Z. Griliches y M. D. Intriligator, *Handbook of Econometrics, vol. I.*

Amemiya, T. (1985). *Advanced Econometrics*, Oxford, Basil Blackwell Ltd.

Aranda-Ordaz, F. J. (1981). "On Two Families of Transformations to Additivity for Binary Response Data", *Biometrika*, 68, 357-363.

Arnaiz, G. (1978). *Introducción a la Estadística Teórica*, 3ª ed., Valladolid, Lex Nova.

Atkinson, A. C. (1982). "Regression Diagnostics, Transformations and Constructed Variables" con discusión, *Journal of the Royal Statistical Society*, B, 44, 1, 1-36.

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

Azorín, F. y J. L. Sanchez-Crespo (1986). *Métodos y Aplicaciones del Muestreo*, Madrid, Alianza Editorial, S.A.

Bazaraa, M. S. y C. M. Shetty (1979). *Nonlinear Programming: Theory and Algorithms*, New York, John Wiley & Sons, Inc.

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. Identifying Influential Data and Sources of Collineartiy*, New York, John Wiley & Sons.

Ben-Akiva, M. y S. R. Lerman (1985). *Discrete Choice Analysis. Theory and Application to Travel Demand*, Cambridge, Mass., M.I.T. Press.

Berman, G. (1979). "Lattice Approximations to the Minima of Functions of Several Variables", *Journal of the Association of Computing Machinery*, 16, 286-294.

Bierman, H. y W. H. Hausman (1970). "The credit granting decision", *Management Science*, 16, B519-B532.

Box, G. E. P. (1980). "Sampling and Bayes's Inference in Scientific Modelling and Robustness" con discusión, *Journal of the Royal Statistical Society*, A, 383-430.

Box, G. E. P. y D. R. Cox (1964). "An Analysis of Transformations" con discusión, *Journal of the Royal Statistical Society*, B, 26, 211-252.

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.

Boyes, W. J., D. L. Hoffman y A. S. Low (1989). "An econometric Analysis of the bank credit scoring problem", *Journal of Econometrics*, 41, 3-14.

Brendt, E. R., B. H. Hall, R. E. Hall, y J. A. Hausman (1974). "Estimation and Inference in Non-linear Structural Models", *Annals of Economic and Social Measurement*, 3, 4, 653-665.

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, N.J., Prentice-Hall.

Bunch, D. S. (1988). "A Comparison of Algorithms for Maximum Likelihood Estimation of Choice Models", *Journal of Econometrics*, 38, 145-167.

Burr, I. W. (1942). "Cumulative Frequency Functions", *Annals of Mathematical Statistics*, 13, 215-232.

Chambers, E. A. y D. R. Cox (1967). "Discrimination between Alternative Binary Response Models", *Biometrika*, 54, 3 y 4, 573-578.

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

Cook, R. D. (1979). "Influential Observations in Linear Regresion", *Journal of the American Statistical Association*, 74, 365, 169-174.

Cook, R. D. (1986). "Assessment of Local Influence", *Journal of the Royal Statistical Society*, B, 48, 133-169.

Cook, R. D., N. Holschuh y S. Weisberg (1982). "A Note on an Alternative Outlier Model", *Journal of the Royal Statistical Society*, B, 44, 3, 370-376.

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

Cook, R. D. y S. Weisberg (1982).

Copas, J. B. (1983). "Regression, Prediction and Shrinkage", *Journal of the Royal Statistical Society*, B, 45, 3, 311-354.

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

Cosslett, S. R. (1981a). "Maximum Likelihood Estimator for Choice-Based Samples", *Econometrica*, 49, 5, 1289-1316.

Cosslett, S. R. (1981b). "Efficient Estimation of Discrete-Choice Models". En Manski, C. F. y D. L. McFadden (eds.), *Structrual Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass., M.I.T. Press.

Cosslett, S. R. (1983). "Distribution-free Maximum Likelihood Estimator of the Binary Choice Model". *Econometrica*, 51, 765-782.

Cox, D. R. y D. V. Hinkley (1974). *Theoretical Statistics*, London, Chapman and Hall Inc.

Cox, D. R. y E. J. Snell (1989). *Analysis of Binary Data*, 2nd. edition, London, Chapman and Hall Inc.

Cramer, J. S. (1991). *The Logit Model for Economists*, London, Chapman and Hall Inc.

Daganzo, C. (1979). *Multinomial Probit. The Theory and its Application to Demand Forecasting*, New York, Academic Press.

Dagenais, M. G. (1974). "Multiple Regression Analysis with Incomplete Observations, from a Bayesian Viewpoint", en Fienberg, S. E. y A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics*, North-Holland.

Davidson, R. y J. G. MacKinnon (1984). "Convenient specification tests for logit and probit models". *Journal of Econometrics*, 25, 241-262.

Day, N. E. (1969). "Estimating the Components of a Mixture of Normal Distributions", *Biometrika*, 56, 3, 463-474.

Dempster, A. P., N. M. Laird y D. P. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, B, 39, 1-38.

Dennis, J. E. y R. B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, Prentice-Hall.

Dhillon, U. S., J. D. Shilling y C. F. Sirmans (1987). "Choosing between Fixed and Adjustable Rate Mortgages", *Journal of Money, Credit and Banking*, 19, 1, 260-267.

Dufour, J. M. (1982). "Recursive Stability Analysis of Linear Regression Relationships", *Journal of Econometrics*, 19, 31-76.

Engle, R. F. (1983). "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics", en Z. Griliches y M. D. Intriligator (Eds.), *Handbook of Econometrics, vol. II*, 775-826. Amsterdam, North Holland.

Gill, P. E., W. Murray y M. H. Wright (1981). *Practical Optimization*. London, Academic Press Ltd.

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

Goldfeld, S. M. y R. E. Quandt (1972). *Nonlinear Methods in Econometrics*. Amsterdam, North-Holland.

Goldfeld, S. M., R. E. Quandt y H. F. Trotter (1966). "Maximization by Quadratic Hill-Climbing", *Econometrica*, 34, 541-551.

Goldberger, A. S. (1983). "Abnormal Selection Bias". En S. Karlin, T. Amemiya y L. A. Goodman (eds.), *Studies in Econometrics, Time-Series and Multivariate Analysis*. New York, Academic Press.

Gourieroux, C., A. Monfort y A. Trognon (1984). "Pseudo Maximum Likelihood Methods, Theory", *Econometrica*, 52, 3, 681-700.

Gourieroux, C., A. Monfort y A. Trognon (1984). "Pseudo Maximum Likelihood Methods, Applications to Poisson Models", *Econometrica*, 52, 3, 701-720.

Gourieroux, C., A. Monfort, E. Renault y A. Trognon (1987). "Generalised Residuals", *Journal of Econometrics*, 34, 5-32.

Guerrero, V. M. y R. A. Johnson (1982). "Use of the Box-Cox Transformation with Binary Response Models", *Biometrika*, 69, 2, 309-314.

Gracia Díez, M. y G. R. Serrano (1991). "Algunos Aspectos sobre el Análisis Empírico de *Credit Scoring*", *Estadística Española*, 34, 130, 261-283.

Gracia Díez, M. y G. R. Serrano (1992). "Observaciones Anómalas en Modelos de Elección Binaria", *Documento de Trabajo, Dpto. Economía Cuantitativa, UCM.*

Green, P. J. (1984). "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives" con discusión, *Journal of the Royal Statistical Society*, B, 46, 2, 149-192.

Hampel, F. R. (1974). "The Influence Curve and its Role in Robust Estimation", *Journal of the American Statistical Association*, 69, 383-394.

Hausman, J. A. (1978). "Specification Tests in Econometrics". *Econometrica*, 46, 6, 1251-1271.

Hausman, J. A. y D. L. McFadden (1984). "Specification Tests for the Multinomial Logit Model", *Econometrica*, 52, 5, 1219-1240.

Hausman, J. A. y D. A. Wise (1977). "Social experimentation, truncated distributions and efficient estimation", *Econometrica*, 45, 319-339.

Hausman, J. A. y D. A. Wise (1978). "A Conditional Probit Model for Qualitative Choice, Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences", *Econometrica*, 46, 2, 403-426.

Heckman, J. J. (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models". *Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. J. (1978). "Dummy Endogenous Variables in a Simultaneous Equation System", *Econometrica*, 46, 6, 931-959.

Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error", *Econometrica*, 47, 153-161.

Heckman, J. J. (1982). "Statistical Models for the Analysis of Dicrete Panel Data". En Manski, C. F. y D. L. McFadden, Eds., *Structrual Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass., M.I.T. Press.

Hendry, D. F. (1984). "Monte Carlo Experimentation in Econometrics", en Z. Griliches y M. D. Intriligator, *Handbook of Econometrics, vol. II.*

Hocking, R. R. (1983). "Developements in Linear Regression Methodology: 1959-1982" con discusión, *Technometrics*, 25, 3, 219-249.

Hooke, R. y T. A. Jeeves (1961). "Direct Search Solution of Numerical and Statistical Problems", *Journal of the Association of Computing Machinery*, 8, 212-221.

Huber, P. J. (1981). *Robust Statistics*, New York, John Wiley & Sons.

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

Johnson N. L. y S. Kotz (1970). *Distributions in Statistics: Continuous Univariate Distributions, vols. 1 y 2.* New York, John Wiley & Sons.

Johnson N. L. y S. Kotz (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* New York, John Wiley & Sons.

Jones, P. N. y G. J. McLachlan (1990). "Maximum Likelihood Estimation from Grouped and Truncated Data with Finite Normal Mixture Models", *Applied Statistics*, 39, 2, 273-312.

Keane, M. P. (1992). "A Note on Identification in the Multinomial Probit Model", *Journal of Business & Economic Statistics*, 10, 2, 193-200.

Krasker, W. S., E. Kuh y R. E. Welsch (1983). "Estimation for Dirty Data and Flawed Models". En Z. Griliches y M. D. Intriligator, *Handbook of Econometrics, vol. I.*

Lee, L. F. (1979a). "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables", *Econometrica*, 47, 4, 977-996.

Lee, L. F. (1979b). "On Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis", *Economics Letters*, 4, 243-249.

Lee, L. F. (1984). "Non-Parametric Testing of Discrete Panel Data Models", *Journal of Econometrics*, 34, 147-177.

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

Lo, A. W. (1986). "Logit versus discriminant analysis: a specification test and application to corporate bankruptcies", *Journal of Econometrics*, 31, 151-178.

Lo, A. W. y A. C. MacKinlay (1989). "The Size and Power of the Variance Ratio Test in Finite Samples. A Monte Carlo Investigation", *Journal of Econometrics*, 40, 203-238.

Lott, W. F. y S. C. Ray (1992), *Applied Econometrics: Problems with Data Sets*, The Dryden Press.

Luce, R. D. y P. Suppes (1963). "Preference, Utility and Subjective Probability". En Luce, R. D., R. R. Bush y E. Galanter (eds.), *Handbook of Mathematical Psychology, vol. 3*. New York, Wiley.

McCullagh, P. y J. A. Nelder (1983). *Generalized Linear Models*, London: Chapman and Hall, Inc.

McFadden, D. L. (1973). "Conditional Logit Analysis of Qualitative Choice Behavior", En P. Zarembka (ed.), *Frontiers in Econometrics*. New York, Academic.

McFadden, D. L. (1976). "A Comment on Discriminant Analysis versus Logit Analysis", *Annals of Economic and Social Measurement*, 5, 511-523.

McFadden, D. L. (1978). "Modelling the Choice of Residential Location". En A. Karlquist et al. (eds.), *Spatial Interaction Theory and Residential Location*, 75-96. Amsterdam, North-Holland.

McFadden, D. L. (1980). "Econometric Models for Probabilistic Choice Among Products", *Journal of Business*, 53, 3, 513-529.

McFadden, D. L. (1981). "Econometric Models of Probabilistic Choice". En Manski, C. F. y D. L. McFadden, Eds., *Structrual Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass., M.I.T. Press.

McFadden, D. L. (1983). "Econometric Analysis of Qualitative Response Models". En Z. Griliches y M. D. Intriligator, *Handbook of Econometrics, vol. II.*

McFadden, D. L. (1987). "Regression-Based Specification Tests for the Multinomial Logit Model", *Journal of Econometrics*, 34, 63-82.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, Mass., M.I.T. Press.

Manski, C. F. (1975). "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228.

Manski, C. F. (1983). "Closest empirical Distribution Estimation", *Econometrica*, 51, 2,

Manski, C. F. y S. Lerman (1977). "The Estimation of Choice Probabilities from Choice-Based Samples", *Econometrica*, 45, 8, 1977-1988.

Manski, C. F. y D. L. McFadden, Eds. (1981). *Structrual Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass., M.I.T. Press.

Manski, C. F. y D. L. McFadden (1981). "Alternative Estimators and Sample Designs for Discrete Choice Analysis". En Manski, C. F. y D. L. McFadden, Eds., *Structrual Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass., M.I.T. Press.

Manski, C. F. y T. S. Thompson (1986). "Operational Characteristics of Maximum Score Estimation", *Journal of Econometrics*, 32, 85-108.

Manski, C. F. y T. S. Thompson (1989). "Estimation of Best Predictors of Binary Response", *Journal of Econometrics*, 40, 97-123.

Marquardt, D. W. (1963). "An Algorithm for Least Squares Estimation of Nonlinear Parameters", *Journal of the Society for Industrial Applied Mathematics*, 2, 431-441.

Meng C. L. y P. Schmidt (1985). "On the Cost of Partial Observability in the Bivariate Probit Model", *International Economic Review*, 26, 71-85.

Morimune, K. (1979). "Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis", *Econometrica*, 47, 4, 957-975.

Nelder, J. A. y R. W. M. Wedderburn (1972). "Generalized Linear Models", *Journal of the Royal Statistical Society*, A, 135, 370-384.

Nuñez, J. J. (1990). "Una clase de modelos lineales binarios de regresión cualitativa", *Estadística Española*, 32, 124, 389-400.

Pagan, A. y F. Vella (1989). "Diagnostic Tests for Models Based on Individual Data: a Survey", *Journal of Applied Econometrics*, 4, S29-S59.

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

Peña, D. (1990). "Influential Observations in Time Series", *Journal of Business and Economic Statistics*, 8, 2, 235-241.

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 & Economic Statistics*, 2, 1, 10-20.

Peña, D. y V. J. Yohai (1991). "The Detection of Influential Subsets in Linear Regression using an Influence Matrix", mimeo.

Poirier, D. J. (1978). "Partial Observability in Bivariate Probit Models". *Journal of Econometrics*, 12, 209-217.

Powell, M. J. D. (1964). "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives", *Computer Journal*, 7, 155-162.

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. I*, 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" con discusión, *Journal of the American Statistical Association*, 73, 364, 730-752.

Ralston, A. y P. Rabinowitz (1978). *A First Course in Numerical Analysis*, McGraw-Hill.

Reiss, P. C. (1990). "Detecting Multiple outliers with an application to R&D Productivity", *Journal of Econometrics*, 43, 293-315.

Rousseeuw, P. J. y B. C. van Zomeren (1990). "Unmasking Multivariate Outliers and Leverage Points", con discusión. *Journal of de American Statistical Association*, 85, 411, 633-651.

Ruud, P. A. (1991). "Extensions of Estimation Methods Using the EM Algorithm", *Journal of Econometrics*, 49, 305-341.

Silvey, S. D. (1970). *Statistical Inference*. Harmondsworth, Penguin.

Smith, B. T., J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema y C. B. Moler (1974). *Maxtix Eigensystem Routines. EISPACK Guide*, Springer-Verlag.

Snee, R. D. (1983). Discusión del papel de Hocking (1983).

Srinivasan, V. e Y. H. Kim (1987). "Credit Granting: A Comparative Analysis of Classification Procedures", *Journal of Finance*, 42, 665-683.

Thomas, W. y R. D. Cook (1990). "Assessing Influence on Predictions from Generalized Linear Models", *Technometrics*, 32, 1, 59-65.

Train, K. (1986). *Qualitative Choice Analysis. Theory, Econometrics and an Application to Automobile Demand*. Cambridge, Mass., M.I.T. Press.

Vuong, Q. H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses", *Econometrica*, 57, 2, 307-333.

Weisberg, S. (1983). Discusión del papel de Hocking (1983).

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.

Wills, H. (1987). "A note on Specification Tests for the Multinomial Logit Model", *Journal of Econometrics*, 34, 263-274.

Windmeijer, F. A. G. (1991). "A Goodnes-of-Fit Test in the Multinomial Logit Model Based on Weighted Squared Residuals", mimeo.

INDICE DE AUTORES