

MODELO PROBIT

Como hemos visto, el Logit tiene limitaciones (véase la Tabla 5). La única limitación del Probit es que asume que todos los componentes inobservados de la utilidad siguen una *distribución normal*. Sólo en algunos casos particulares especiales ese supuesto puede llevar a pronósticos erróneos. Un ejemplo serían los coeficientes de los precios en modelos Probit con variación aleatoria en los gustos. Dado que la distribución normal es simétrica en torno a una media cero, el modelo implica necesariamente que, para algunos individuos, el coeficiente de precios es positivo (cuando no debería poder ser). Sería necesario emplear, en casos semejantes, funciones de distribución con densidades sólo a un lado del cero, como la log-normal. Salvo casos como este, el Probit es un modelo muy general.

Tabla 6. *El modelo Probit como superación del Logit.*

Tres casos concretos para los que el Probit supera las limitaciones del Logit:
Variabilidad en los gustos
Patrones de sustitución
Elecciones repetidas a lo largo del tiempo

El modelo Probit se basa en el supuesto de que los componentes inobservados de la utilidad siguen una distribución conjunta normal. En términos más formales:

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

Para cada individuo n y cada alternativa j . Donde la U es la utilidad, que se descompone en una parte observable (V) y otra inobservable (ε). Si tomamos el vector compuesto por las j alternativas para el individuo n , $\varepsilon_n = (\varepsilon_{n1}, \varepsilon_{n2}, \dots, \varepsilon_{nj})$. Ese vector está distribuido normalmente con vector de medias cero y matriz de covarianzas Ω . La función de densidad de ese vector es:

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{J/2} |\Omega|^{1/2}} e^{-\frac{1}{2} \varepsilon_n' \Omega^{-1} \varepsilon_n}$$

La matriz de covarianzas Ω puede ser específica de cada individuo, pero no lo señalamos para no complicar la notación.

La probabilidad de elección de la alternativa i por el individuo n es

$$P_{ni} = \text{Prob}(U_{ni} > U_{nj} \forall j \neq i)$$

$$P_{ni} = \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \forall j \neq i)$$

$$P_{ni} = \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \forall j \neq i) \phi(\varepsilon_n) d\varepsilon_n$$

donde $I(\cdot)$ es un indicador de si lo que hay incluido dentro del paréntesis se cumple, y donde la integral es sobre todos los valores de ε_n . Es importante notar que esa integral no tiene forma pre-definida, y que *debe evaluarse numéricamente en cada caso por medio de simulación*. Precisamente esto hace al Probit computacionalmente más difícil que el Logit.

Las probabilidades de elección pueden expresarse de otras formas convenientes para la simulación de la integral. Puede hacerse la integral, que es una suma, no sobre todos los valores de ε_n , sino sólo sobre un subconjunto B_{ni} de ellos. Se trataría de aquel que conduce a la elección de la alternativa i , es decir, $B_{ni} = [\varepsilon_n \text{ s.a. } V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \forall j \neq i]$. De esta forma la integral queda en

$$P_{ni} = \int_{\varepsilon_n \in B_{ni}} \phi(\varepsilon_n) d\varepsilon_n$$

Las dos integrales suman las J alternativas para el individuo n (son integrales J dimensionales).

Sólo las diferencias en niveles de utilidad son relevantes de cara a la elección. Por tanto, podemos sumar las $J-1$ *diferencias* entre errores. Calculamos la probabilidad de elegir a alternativa i , así que comparamos con ella. Definimos por tanto

$$\tilde{U}_{nji} = U_{nj} - U_{ni}$$

$$\tilde{V}_{nji} = V_{nj} - V_{ni}$$

$$\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$$

y

$$P_{ni} = \text{Prob}(\tilde{U}_{nji} < 0 \quad \forall j \neq i)$$

Según esta definición intuitiva, la probabilidad de elegir i es la probabilidad de que, tras comparar todas las alternativas contra i , la utilidad que proporcione esta sea mayor en todos los casos. Ahora definimos el vector de diferencias entre errores como

$$\tilde{\epsilon}_{ni} = (\tilde{\epsilon}_{n1i}, \tilde{\epsilon}_{n2i}, \dots, \tilde{\epsilon}_{nJi})$$

que tiene dimensión $J-1$, es decir, i no se compara con i . La i del error significa “errores contra la alternativa i ”, es decir, comparados con ella. La diferencia entre dos variables aleatorias normales es otra normal, por lo que la función de densidad de probabilidad de los nuevos errores será

$$\phi(\tilde{\epsilon}_{ni}) = \frac{1}{(2\pi)^{(J-1)/2} |\tilde{\Omega}_i|^{1/2}} e^{-\frac{1}{2} \tilde{\epsilon}_{ni}' \tilde{\Omega}_i^{-1} \tilde{\epsilon}_{ni}}$$

la probabilidad de elección de i , expresada como una integral, pero en términos de diferencias de utilidades, queda ahora como

$$P_{ni} = \int I(\tilde{V}_{nji} + \tilde{\epsilon}_{nji} < 0 \quad \forall j \neq i) \phi(\tilde{\epsilon}_{ni}) d\tilde{\epsilon}_{ni}$$

que es $J-1$ dimensional. De la misma forma:

$$P_{ni} = \int_{\tilde{\epsilon}_{ni} \in \tilde{B}_{ni}} \phi(\tilde{\epsilon}_{ni}) d\tilde{\epsilon}_{ni}$$

donde

$$\tilde{B}_{ni} = \{\tilde{\epsilon}_{ni} \text{ s.a. } \tilde{V}_{nji} + \tilde{\epsilon}_{nji} < 0 \quad \forall j \neq i\}$$

El problema está en hallar la matriz de covarianzas de las diferencias entre errores ($\tilde{\Omega}_i$). Afortunadamente es posible hallarla a partir de la matriz de covarianzas de los errores (Ω).

Si M_i es una matriz identidad de dimensiones $J-1$ con una columna extra de -1 en el lugar i -ésimo, la matriz queda con dimensiones $(J-1) \times J$, y tenemos que

$$\tilde{\Omega}_i = M_i \Omega M_i'$$

Siendo las dimensiones, por orden: $(J-1) \times (J-1) = [(J-1) \times J] \times [J \times J] \times [J \times (J-1)]$.

Un ejemplo. Un individuo se enfrenta a tres alternativas, de entre las que tiene que elegir una. Los errores son $(\varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3})$, y la matriz de covarianzas es

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}$$

ahora calculamos las diferencias respecto de la alternativa 2, lo que nos lleva a unas diferencias entre errores de $(\tilde{\varepsilon}_{n12}, \tilde{\varepsilon}_{n32})$, con una matriz de covarianzas

$$\begin{aligned} \tilde{\Omega}_2 &= \text{Cov} \begin{pmatrix} \varepsilon_{n1} - \varepsilon_{n2} \\ \varepsilon_{n3} - \varepsilon_{n2} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix} \\ &= \begin{pmatrix} \theta_{11} & \theta_{13} \\ \theta_{13} & \theta_{33} \end{pmatrix} \end{aligned}$$

Esa matriz puede obtenerse también aplicando la fórmula $\tilde{\Omega}_2 = M_2 \Omega M_2'$:

$$\begin{aligned}
\tilde{\Omega}_h &= \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} & -\sigma_{22} + \sigma_{23} & -\sigma_{23} + \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \sigma_{12} - \sigma_{12} + \sigma_{22} & -\sigma_{12} + \sigma_{22} + \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} + \sigma_{22} - \sigma_{23} & \sigma_{22} - \sigma_{23} - \sigma_{23} + \sigma_{33} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}
\end{aligned}$$

El modelo Probit presenta un *problema de identificación*. Como es bien sabido, las funciones de utilidad son ordinales. Ni el nivel ni la escala de la utilidad debe afectar al problema de elección. Multiplicar las funciones de utilidad de todas las alternativas por un mismo número (positivo), escalar, no cambia el orden de preferencias, ni debería cambiar las probabilidades relativas. Lo mismo si alteramos el nivel de todas las funciones de utilidad mediante la suma de una misma constante. En los modelos Logit y Logit anidado, los supuestos sobre la distribución de los errores garantizan una normalización de nivel y escala, automática. Esto quiere decir que cualquier constante no estimable (parámetro no identificable) presente es eliminada. En el modelo Probit esto no pasa automáticamente, por lo que hay que normalizar los modelos.

Un parámetro está identificado si se puede estimar. Si tenemos una función de utilidad del tipo $U_{nj} = V_{nj} + k + \varepsilon_{nj}$, el parámetro k no está identificado. La elección del individuo no se ve afectada por la presencia de esa constante, como hemos visto, por lo que no es posible inferir la presencia de la misma. Si no se normaliza el modelo, esos parámetros no identificados quedan ahí.

¿Dónde está el problema? No siempre está claro qué parámetros afectan sólo a la escala y el nivel de la utilidad y qué otros están afectando a la elección. Bunch y Kitamura (1989) muestran cómo en muchos modelos Probit de artículos publicados, que no han sido debidamente normalizados, hay parámetros no identificados, es decir, puede haber parámetros relevantes para la elección que quedan sin identificar. Un ejemplo son las covarianzas. En un modelo de 4 alternativas las covarianzas (σ) son diez. Sin embargo, las covarianzas de las diferencias entre errores (θ) son sólo cinco. Pero no es necesario conocer las 10 covarianzas originales para estimar las cinco derivadas de las diferencias. Esa reducción de parámetros a estimar no es una restricción, es simplemente una normalización. La matriz de covarianzas original contiene aspectos irrelevantes

para el problema de elección, y al ser irrelevantes están afectando únicamente a la escala y nivel de utilidad. Sólo los parámetros estimables contienen información independiente de la reordenación de niveles de utilidad, y sólo ellos tienen contenido económico. En general, un modelo con J alternativas tendrá $J(J+1)/2$ parámetros antes de normalizar, y $[(J-1)J/2]-1$ parámetros de covarianza después de la normalización. Sólo estos están identificados.

¿Qué ocurre si se impone una estructura a las covarianzas originales (σ)? Esto es como asumir un determinado patrón de comportamiento de los errores, que fija determinados valores o relaciones entre valores de los elementos de la matriz de covarianzas. En ese caso, usualmente, el número de parámetros de covarianza (σ) se reduce fuertemente, y es suficiente con esto para normalizar el modelo. Pero a veces no es así (véanse los ejemplos de Bunch and Kitamura, 1989). Es fácil comprobar si las restricciones sobre la matriz de covarianzas son suficientes para normalizar el modelo. Tenemos por un lado la matriz Ω con las restricciones sobre sus elementos, y por otro lado la matriz $\tilde{\Omega}^*$, normalizada para escala y nivel. Sabemos que cada elemento de ésta está perfectamente identificado. Por tanto, si cada uno de los elementos restringidos de Ω se pueden calcular a partir de los elementos identificados de $\tilde{\Omega}^*$ las restricciones habrán normalizado el modelo. Si los elementos de Ω no pueden calcularse a partir de los de $\tilde{\Omega}^*$ las restricciones no serán suficientes y los parámetros de la Ω restringida no estarán identificados.

Un ejemplo. Un caso de elección entre cuatro alternativas. La utilidad, como de costumbre, se expresa como suma de un vector de componentes conocidas y un término de error: $U_{nj} = V_{nj} + \varepsilon_{nj}$, para $j=1,2,3,4$. El vector de errores es $\varepsilon'_n = (\varepsilon_{n1}, \dots, \varepsilon_{n4})$, estando normalmente distribuido con media cero y una matriz de covarianzas:

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix}$$

Hay diez elementos σ distintos en esa matriz, representando las varianzas y covarianzas entre los cuatro errores. Como regla general, un modelo con J alternativas de elección tiene $J(J+1)/2$ elementos diferentes en la matriz $J \times J$ de covarianzas de los errores.

Dado que el nivel de la utilidad es irrelevante, tomamos diferencias con respecto a la primera alternativa: $\tilde{\varepsilon}_{nj} = \varepsilon_{nj} - \varepsilon_{n1}$ para $j=2,3,4$, lo que nos lleva a un vector de

diferencias en los errores: $\tilde{\epsilon}_{n1} = (\tilde{\epsilon}_{n21}, \tilde{\epsilon}_{n31}, \tilde{\epsilon}_{n41})$. Este vector tiene asociada su propia matriz de covarianzas:

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{23} & \theta_{33} & \theta_{34} \\ \theta_{24} & \theta_{34} & \theta_{44} \end{pmatrix} = M_1 \Omega M_1'$$

Y la relación de los elementos de esa matriz con los de la matriz de covarianzas normal es

$$\theta_{22} = \sigma_{22} + \sigma_{11} - 2\sigma_{12}$$

$$\theta_{33} = \sigma_{33} + \sigma_{11} - 2\sigma_{13}$$

$$\theta_{44} = \sigma_{44} + \sigma_{11} - 2\sigma_{14}$$

$$\theta_{23} = \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}$$

$$\theta_{24} = \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}$$

$$\theta_{34} = \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}$$

Normalizamos para la escala igualando a 1 uno de los elementos de la diagonal principal. Por ejemplo, $\theta_{22} = 1$.

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \theta_{23}^* & \theta_{33}^* & \theta_{34}^* \\ \theta_{24}^* & \theta_{34}^* & \theta_{44}^* \end{pmatrix}$$

Las equivalencias entre esos nuevos parámetros y las varianzas y covarianzas originales (σ) es

$$\theta_{33}^* = \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}$$

$$\theta_{44}^* = \frac{\sigma_{44} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}$$

$$\theta_{23}^* = \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}$$

$$\theta_{24}^* = \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}$$

$$\theta_{34}^* = \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}$$

Hay sólo cinco elementos en la matriz $\tilde{\Omega}_1^*$ (de los diez que había en la matriz original Ω), que son los que están realmente identificados, y se pueden estimar. Las varianzas y covarianzas originales (σ) no son estimables a partir de los θ^* .

Si imponemos ahora sobre la matriz original una determinada estructura, habrá que comprobar si ello es suficiente para normalizar el modelo. Supongamos la siguiente estructura para Ω :

$$\Omega = \begin{pmatrix} 1+\rho & \rho & 0 & 0 \\ \rho & 1+\rho & 0 & 0 \\ 0 & 0 & 1+\rho & \rho \\ 0 & 0 & \rho & 1+\rho \end{pmatrix}$$

Se deduce que los errores primero y segundo, por un lado, y tercero y cuarto, por otro lado, pueden estar correlacionados. No puede haber otras correlaciones. Las varianzas son $1+\rho$, y las correlaciones $\rho/(1+\rho)$, que necesariamente se mueven entre -1 y 1. Ahora nos preguntamos si ese modelo queda, con esa estructura, normalizado para escala y nivel.

Tomamos diferencias con respecto a la primera alternativa, nos queda:

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{23} & \theta_{33} & \theta_{34} \\ \theta_{24} & \theta_{34} & \theta_{44} \end{pmatrix}$$

donde

$$\theta_{22} = 2$$

$$\theta_{33} = 2 + 2\rho$$

$$\theta_{44} = 2 + 2\rho$$

$$\theta_{23} = 1$$

$$\theta_{24} = 1$$

$$\theta_{34} = 1 + 2\rho$$

Normalizamos para la *escala* haciendo el elemento $\theta_{22} = 1$, y nos queda la matriz:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \theta_{23}^* & \theta_{33}^* & \theta_{34}^* \\ \theta_{24}^* & \theta_{34}^* & \theta_{44}^* \end{pmatrix}$$

donde tenemos las siguientes correspondencias:

$$\theta_{33}^* = 1 + \rho$$

$$\theta_{44}^* = 1 + \rho$$

$$\theta_{23}^* = 1/2$$

$$\theta_{24}^* = 1/2$$

$$\theta_{34}^* = 1/2 + \rho$$

Por tanto, tanto en $\tilde{\Omega}_1^*$ como en Ω hay un solo parámetro. Basta definir $\theta = 1 + \rho$. Entonces nos queda:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \theta & \theta - \frac{1}{2} \\ \frac{1}{2} & \theta - \frac{1}{2} & \theta \end{pmatrix}$$

Por tanto, todos los parámetros de la matriz Ω (que es uno) pueden calcularse a partir de los parámetros identificados, que son los contenidos en la matriz $\tilde{\Omega}_1^*$ (uno también). Se deduce que el modelo original, con su patrón de comportamiento para las covarianzas originales, estaba normalizado para escala y nivel.

Esto no siempre es así. Si, por ejemplo, la estructura asumida para las varianzas y covarianzas hubiera sido otra, como esta:

$$\Omega = \begin{pmatrix} 1+\rho_1 & \rho_1 & 0 & 0 \\ \rho_1 & 1+\rho_1 & 0 & 0 \\ 0 & 0 & 1+\rho_2 & \rho_2 \\ 0 & 0 & \rho_2 & 1+\rho_2 \end{pmatrix}$$

Se permite una correlación entre el primer y segundo error y el tercero y el cuarto, pero diferente en uno y otro caso ($\rho_1/(1+\rho_1)$ y $\rho_2/(1+\rho_2)$ respectivamente). La matriz de varianzas y covarianzas de las diferencias entre errores normalizada es

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \theta & \theta - \frac{1}{2} \\ \frac{1}{2} & \theta - \frac{1}{2} & \theta \end{pmatrix}$$

Si bien ahora tendremos que $\theta = 1 + (\rho_1 + \rho_2)/2$. Como puede observarse, los valores ρ_1 y ρ_2 no pueden calcularse a partir del θ estimado (aunque sí su media), por lo que dichos parámetros no están identificados y el modelo no está normalizado para escala y nivel. Puede parecer extraño porque en la matriz de covarianzas original (restringida) sólo había dos parámetros, y uno podría pensar que reduciendo de diez a dos parámetros el modelo debería quedar inmediatamente normalizado, pero no es así, como vemos. Lo único que está identificado es la media $(\rho_1 + \rho_2)/2$, y sólo cuando $\rho_1 = \rho_2$ la media coincide con los valores individuales de los parámetros y el modelo está identificado. Obsérvese que todo modelo con la misma media de esos dos parámetros sería equivalente, y que sus valores concretos no importan.

Con la normalización para escala y nivel nos aseguramos de que sólo la información con significado económico queda en el modelo.

Veremos ahora tres casos concretos para los que el Probit supera las limitaciones del Logit: 1) Variabilidad en los gustos; 2) Patrones de sustitución; y 3) Elecciones repetidas a lo largo del tiempo.

1.

El modelo Probit incorpora con facilidad *variabilidad en los gustos* entre decisores, lo que se concreta en coeficientes aleatorios, siempre que estos sigan una distribución normal.

Supongamos ahora que la utilidad representativa es lineal en los parámetros, y que los parámetros varían aleatoriamente *entre individuos* (variedad en los gustos), en vez de ser fijos (iguales para todos los individuos) como veníamos suponiendo hasta ahora. En un caso como este la utilidad será:

$$U_{nj} = \beta'_n x_{nj} + \varepsilon_{nj}$$

Donde β_n es el *vector* de coeficientes para el individuo n , y representa sus gustos personales, no necesariamente iguales a los de otro individuo. Si β_n sigue una distribución normal en la población, con media b y covarianza W , tendremos que $\beta_n \sim N(b, W)$. Hay que estimar pues b y W .

La utilidad puede escribirse de otra forma, descomponiendo β_n en dos partes: su media (b) y las desviaciones respecto de su media ($\tilde{\beta}$). En ese caso tendríamos:

$$U_{nj} = b' x_{nj} + \tilde{\beta}'_n x_{nj} + \varepsilon_{nj} = b' x_{nj} + \eta_{nj}$$

donde $\tilde{\beta}_n = b - \beta_n$ y donde $\eta_{nj} = \tilde{\beta}'_n x_{nj} + \varepsilon_{nj}$. La covarianza de los η_{nj} depende de W así como de los x_{nj} , lo que implica que la covarianza difiere entre decisores (n).

Imaginemos un caso en el que el decisor tiene ante sí dos alternativas, con una sola variable explicativa. La utilidad para cada alternativa sería en este caso:

$$U_{n1} = \beta_n x_{n1} + \varepsilon_{n1}$$

$$U_{n2} = \beta_n x_{n2} + \varepsilon_{n2}$$

Asumimos que β_n está distribuido normalmente con media b y varianza σ_β . Suponemos que ε_{n1} y ε_{n2} son *iid* con varianza σ_ε , aunque el supuesto de independencia no es necesario en general. Podemos reescribir las utilidades como

$$U_{n1} = bx_{n1} + \eta_{n1}$$

$$U_{n2} = bx_{n2} + \eta_{n2}$$

Donde η_{n1} y η_{n2} están normalmente distribuidos, con media cero, varianza, covarianza y matriz de covarianzas bajo estas formas:

$$E(\eta_{nj}) = E(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = 0$$

$$V(\eta_{nj}) = V(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = x_{nj}^2 \sigma_\beta + \sigma_\varepsilon$$

$$\begin{aligned} \text{Cov}(\eta_{n1}, \eta_{n2}) &= E[(\tilde{\beta}_n x_{n1} + \varepsilon_{n1})(\tilde{\beta}_n x_{n2} + \varepsilon_{n2})] \\ &= E(\tilde{\beta}_n^2 x_{n1} x_{n2} + \varepsilon_{n1} \varepsilon_{n2} + \varepsilon_{n1} \tilde{\beta}_n x_{n2} + \varepsilon_{n2} \tilde{\beta}_n x_{n1}) \\ &= x_{n1} x_{n2} \sigma_\beta \end{aligned}$$

$$\begin{aligned} \Omega &= \begin{pmatrix} x_{n1}^2 \sigma_\beta + \sigma_\varepsilon & x_{n1} x_{n2} \sigma_\beta \\ x_{n1} x_{n2} \sigma_\beta & x_{n2}^2 \sigma_\beta + \sigma_\varepsilon \end{pmatrix} \\ &= \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1} x_{n2} \\ x_{n1} x_{n2} & x_{n2}^2 \end{pmatrix} + \sigma_\varepsilon \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

Queda normalizar para la escala. Si hacemos $\sigma_\varepsilon = 1$, la matriz de covarianzas queda reducida a

$$\Omega = \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1} x_{n2} \\ x_{n1} x_{n2} & x_{n2}^2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Ahora tenemos unas variables x_{n1} y x_{n2} que son observadas por el investigador, y los parámetros b y σ_β deben estimarse.

El mismo procedimiento se aplica a más de una variable explicativa y más de dos alternativas.

2.

La segunda de las ventajas del Probit es que no exige que se cumpla la propiedad IIA (independencia de las alternativas irrelevantes), por lo que cualquier patrón de sustitución puede representarse mediante este modelo. Dependiendo de cómo sea la matriz de covarianzas Ω de los componentes inobservados de la utilidad, y al estimar ésta, se descubre el patrón de sustitución que se ajusta a los datos. Otra posibilidad es imponer una estructura a dicha matriz de covarianzas para representar diversas formas de no independencia (correlación entre los errores), reduciéndose el número de parámetros a estimar. Veremos las dos posibilidades.

Si se estima la matriz Ω para un caso de 4 alternativas, partiremos de una matriz del tipo:

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix}$$

que normalizada para escala y nivel queda como

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \theta_{23}^* & \theta_{33}^* & \theta_{34}^* \\ \theta_{24}^* & \theta_{34}^* & \theta_{44}^* \end{pmatrix}$$

cuyos elementos se estiman. Esos valores estimados pueden representar cualquier patrón de sustitución, y la normalización no reduce en absoluto las posibilidades. Otro hecho relevante es que los θ^* estimados no proporcionan ninguna información directamente interpretable. Por ejemplo, si θ_{33}^* fuera mayor que θ_{44}^* podríamos sentirnos inclinados a deducir que la varianza de la utilidad inobservada de la tercera alternativa es mayor que la de la cuarta, es decir, que $\sigma_{33} > \sigma_{44}$. Pero es fácil ver que esto puede ser falso, ya que aún siendo $\theta_{33}^* > \theta_{44}^*$

podría darse que $\sigma_{33} < \sigma_{44}$ siempre que $\sigma_{14} > \sigma_{13}$ con una diferencia suficientemente grande¹.

De la misma forma, un $\theta_{23}^* < 0$ no significa que $\sigma_{23} < 0$, ya que siendo ésta positiva, si σ_{12} y σ_{13} son suficientemente grandes podría ocurrir que θ_{23}^* sea negativa².

Ahora consideremos la posibilidad de imponer una estructura a la matriz de covarianzas. Esto sí restringe los patrones de sustitución posibles, dado que la estructura es lógicamente un conjunto de restricciones sobre la matriz de covarianzas. La restricción que introducimos, si es correcta, acorde con el comportamiento del decisor, debe revelar el patrón de sustitución también correcto.

Un ejemplo. A un comprador de vivienda se le presentan cuatro alternativas de hipoteca: una tiene un tipo fijo, y las otras tres tipos variables. Ahora imaginemos que la parte no observada de su función de utilidad tiene dos partes: la preocupación de cada individuo por el riesgo de que los tipos de interés puedan subir (r_n), que afecta a todas las hipotecas de interés variable; y los demás factores inobservables (η_{nj}). La suma queda entonces como:

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

$$\varepsilon_{nj} = -r_n d_j + \eta_{nj}$$

donde $d_j = 1$ si la hipoteca es de interés variable, y cero para la de interés fijo, con signo negativo por decrecer la utilidad conforme la preocupación crece. Ahora supondremos que r_n está generada normalmente, con una varianza σ ; por otro lado, η_{nj} es *iid*, normal, con media cero y varianza ω . Bajo estas condiciones la matriz de covarianzas para el vector $\varepsilon_n = (\varepsilon_{n1}, \dots, \varepsilon_{n4})$ es:

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma & \sigma & \sigma \\ 0 & \sigma & \sigma & \sigma \\ 0 & \sigma & \sigma & \sigma \end{pmatrix} + \omega \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

¹ Recordemos que $\theta_{33} = \sigma_{33} + \sigma_{11} - 2\sigma_{13}$ y $\theta_{44} = \sigma_{44} + \sigma_{11} - 2\sigma_{14}$.

² Recordemos que $\theta_{23} = \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}$.

Hay que normalizar el modelo para ajustar la escala, pero ya está ajustado para nivel. La matriz de covarianzas para las diferencias entre errores queda:

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma & \sigma & \sigma \\ \sigma & \sigma & \sigma \\ \sigma & \sigma & \sigma \end{pmatrix} + \omega \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Dado que el número de parámetros no se ha reducido, el modelo ya estaba ajustado para nivel. Normalizamos para escala, haciendo $\sigma + 2\omega = 1$, después de lo cual nos queda:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta & \theta \\ \theta & 1 & \theta \\ \theta & \theta & 1 \end{pmatrix}$$

donde $\theta = (\sigma + \omega) / (\sigma + 2\omega)$. Es obvio que los valores de σ y ω no se pueden obtener a partir del θ estimado, pero éste sí nos da alguna información interesante. En efecto, podemos saber a partir de él el ratio entre las varianzas de la preocupación por los tipos relativa a los demás factores. Si $\theta = 0,75$ podemos deducir que la varianza atribuible a la preocupación por el riesgo de subida de tipos es el doble que la varianza de la utilidad inobservada debida a los demás factores. En efecto, si

$$\theta = (\sigma + \omega) / (\sigma + 2\omega) = 0,75$$

$$\sigma = 2\omega$$

Dicho de otra forma, la varianza por la preocupación por la subida de tipos es 2/3 del total de la varianza en la utilidad inobservada.

3.

La tercera gran ventaja del Probit sobre el Logit es que permite trabajar con paneles de datos, es decir, con procesos de elección que se repiten varias veces en el tiempo para un mismo agente decisor. La única diferencia entre el Probit con panel de datos y el Probit normal es que la matriz de covarianzas se ve aumentada. Si las alternativas pasan de J a ser $J \times T$, para T períodos de tiempo o procesos de elección (pudiendo ser J y T distintos para cada agente), la función de utilidad del decisor n sería:

$$U_{njt} = V_{njt} + \varepsilon_{njt}$$

Donde ε_{njt} puede presentar correlaciones no sólo entre alternativas, sino también a lo largo del tiempo. El vector de errores sería $\varepsilon_{njt} = (\varepsilon_{n11}, \dots, \varepsilon_{nJ1}, \varepsilon_{n12}, \dots, \varepsilon_{nJ2}, \dots, \varepsilon_{n1T}, \dots, \varepsilon_{nJT})$, y la matriz de covarianzas Ω tendrá dimensiones $JT \times JT$.

Ahora pensemos en una secuencia de alternativas $\mathbf{i} = (i_1, \dots, i_T)$. La probabilidad de que el agente decisor tome esas T posibilidades, de entre las $J \times T$ posibles, tiene una probabilidad

$$\begin{aligned} P_{ni} &= \Pr \text{ob} \left(U_{ni,t} > U_{njt} \quad \forall j \neq i_t, \forall t \right) \\ &= \Pr \text{ob} \left(V_{ni,t} + \varepsilon_{ni,t} > V_{njt} + \varepsilon_{njt} \quad \forall j \neq i_t, \forall t \right) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n \end{aligned}$$

donde $B_n = (\varepsilon_n \text{ s.t. } V_{ni,t} + \varepsilon_{ni,t} > V_{njt} + \varepsilon_{njt} \quad \forall j \neq i_t, \forall t)$ y $\phi(\varepsilon_n)$ es la función de densidad conjunta normal, con media cero y matriz de covarianzas Ω . Es muy similar al caso de un solo período, pero aquí la integral se expande a $J \times T$ dimensiones en lugar de J .

Si expresamos la utilidad en términos de diferencias podemos decir que la probabilidad de que se de el vector \mathbf{i} , o la secuencia de elecciones \mathbf{i} , es la probabilidad de que las diferencias entre utilidades sea negativa para cada alternativa en cada período de tiempo, cuando dichas diferencias se calculan respecto de la correspondiente alternativa de \mathbf{i} para cada momento:

$$P_{ni} = \text{Pr ob}(\tilde{U}_{ni,t} < 0 \quad \forall j \neq i_t, \forall t) \\ = \int_{\tilde{\epsilon}_n \in \tilde{B}_n} \phi(\tilde{\epsilon}_n) d\tilde{\epsilon}_n$$

donde

$$\tilde{U}_{nji,t} = U_{njt} - U_{ni,t}$$

quedando el vector de diferencias entre errores como

$$\tilde{\epsilon}'_n = \{(\epsilon_{n11} - \epsilon_{ni1}), \dots, (\epsilon_{nJ1} - \epsilon_{ni1}), \dots, (\epsilon_{n1T} - \epsilon_{niT}), \dots, (\epsilon_{nJT} - \epsilon_{niT})\}$$

La matriz \tilde{B}_n es el conjunto de $\tilde{\epsilon}_n$ (diferencias entre errores) para los que $\tilde{U}_{nji,t} < 0 \quad \forall j \neq i_t, \forall t$. La integral acaba con dimensiones $(J-1)T$. La función de densidad $\phi(\tilde{\epsilon}_n)$ es conjuntamente normal y la matriz de covarianzas se deriva de Ω .

La simulación es similar a la del caso en que $T=1$, sólo que la matriz de covarianzas y la integral tienen mayor dimensión. Cuando las posibilidades son binarias, es decir, cuando $J=2$, el modelo Probit se simplifica considerablemente (para el caso multinomial véase Borsch-Supan et al. 1991; para el binario Gouieroux y Monfort, 1993). Este tipo de casos es muy común. Por ejemplo, cuando un individuo se enfrenta a la posibilidad de trabajar o no, durante varios períodos. No hay que calcular diferencias en este caso, y la función de utilidad queda reducida a la forma:

$$U_{nt} = V_{nt} + \epsilon_{nt}$$

La persona en cuestión toma la decisión de actuar cuando $U_{nt} > 0$, que se llama utilidad neta, pues es la diferencia de utilidad entre tomar o no tomar la acción. Los errores pueden tener correlación en el tiempo, y la matriz de covarianzas para $\epsilon_{n1}, \dots, \epsilon_{nT}$ es Ω , de dimensiones $T \times T$.

Esta secuencia de alternativas binarias se representa más cómodamente mediante una serie de variables *dummy* que adoptan valor 1 cuando la persona n decide

actuar en el período t ($d_{nt}=1$) y un valor -1 en si decide no actuar ($d_{nt}=-1$). La probabilidad de que se de la secuencia de decisiones $d_n = d_{n1}, \dots, d_{nT}$ es

$$\begin{aligned} P_{nd_n} &= \text{Prob}(U_{nt}d_{nt} > 0 \quad \forall t) \\ &= \text{Prob}(V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \quad \forall t) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n \end{aligned}$$

donde B_n es el conjunto de ε_n para los que $V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \quad \forall t$, y $\phi(\varepsilon_n)$ es una función de densidad normal conjunta con covarianzas Ω . Se puede imponer una estructura sobre esta matriz, afectando a las correlaciones entre errores en el tiempo.

Por ejemplo, en un caso binario el error puede tener dos partes, una específica de cada individuo, que refleje su propensión personal a tomar la decisión de actuar, y otra que varía con el tiempo, de forma que: $\varepsilon_{nt} = \eta_n + \mu_{nt}$, donde μ_{nt} es *iid* en el tiempo y entre personas, con una función de densidad normal; y donde η_n es *iid* entre personas con una función de densidad normal con media cero y varianza σ . La varianza del error en cada período es $V(\varepsilon_{nt}) = V(\eta_n + \mu_{nt}) = \sigma + 1$. La covarianza entre los errores en dos períodos distintos t y s es $\text{Cov}(\varepsilon_{nt}, \varepsilon_{ns}) = E(\eta_n + \mu_{nt})(\eta_n + \mu_{ns}) = \sigma$. La matriz de covarianzas adopta entonces la forma:

$$\Omega = \begin{pmatrix} \sigma+1 & \sigma & \dots & \dots & \sigma \\ \sigma & \sigma+1 & \sigma & \dots & \sigma \\ \dots & \dots & & \dots & \dots \\ \sigma & \dots & \dots & \sigma & \sigma+1 \end{pmatrix}$$

Como puede verse, sólo un parámetro entra en la matriz de covarianzas, y su valor, σ , es la varianza de η_n (la utilidad inobservada entre individuos) relativa a μ_{nt} (la utilidad inobservada para cada individuo a lo largo del tiempo). A menudo se llama a ese cociente la *cross-subject variance relative to the within-subject variance*.

Las probabilidades con esta estructura tienen forma de integral, como sabemos que ocurre con el Probit siempre. Esta integral se puede preparar mediante *partición conveniente del error* en dos partes, una con solución cerrada y otra que requiere simulación para el cálculo.

La probabilidad de *no* tomar la decisión (u optar por la alternativa en cuestión) en el período t , condicionada a η_n es $\text{Prob}(V_{nt} + \eta_n + \mu_{nt} < 0) = \text{Prob}(\mu_{nt} < - (V_{nt} + \eta_n)) = \Phi(- (V_{nt} + \eta_n))$, donde $\Phi(\cdot)$ es la función de distribución normal estándar. Por tanto, la probabilidad de tomar la decisión (u optar por la alternativa), condicionada a η_n , es $1 - \Phi(- (V_{nt} + \eta_n)) = \Phi(V_{nt} + \eta_n)$. Ahora, el siguiente paso: la probabilidad de la secuencia de decisiones d_n , condicionada a η_n , que será $\Pi_t \Phi((V_{nt} + \eta_n)d_{nt})$, que puede expresarse como $H_{ndn}(\eta_n)$.

Hemos condicionado a η_n , cuando ésta es una variable aleatoria. Por tanto, la probabilidad no condicionada es la integral de la probabilidad condicionada $H_{ndn}(\eta_n)$ multiplicada por todos los posibles valores de η_n :

$$P_{nd_n} = \int H_{nd_n}(\eta_n) \phi(\eta_n) d\eta_n$$

donde $\phi(\eta_n)$ es la función de densidad normal con media cero y varianza σ . Esa propiedad requiere de simulación para su cálculo, lo que puede hacerse mediante varios pasos sencillos: se toma una muestra de una función de densidad normal mediante un generador de números aleatorios, y se multiplica la muestra por $\sigma^{1/2}$ para convertirla en una muestra de η_n , variable aleatoria de una función de densidad normal con varianza σ ; para esta muestra de η_n calculamos $H_{ndn}(\eta_n)$; y repetimos los dos primeros pasos muchas veces, para promediar los resultados, siendo ese promedio una aproximación simulada a P_{ndn} .

El simulador que acabamos de describir es muy simple, y esa simpleza se debe a la estructura que hemos impuesto a la matriz de covarianzas, sustentada en el supuesto de que la dependencia temporal de los factores inobservados queda totalmente capturada por un componente aleatorio, η_n , que permanece constante a lo largo del tiempo para cada persona. Gourieroux y Monfort (1993) trabajan con este modelo. Sin esos supuestos simplificadores los simuladores se hacen mucho más complejos (véase Train, 2003, pp. 118 y siguientes).

Una complicación adicional, que no hemos analizado, se presenta cuando en la utilidad representativa de un período determinado entran variables de otros períodos, como precios de períodos anteriores o futuros. Esto complica el análisis en dos sentidos. Primero, los errores empiezan a estar correlacionados en el tiempo, por lo que la elección en un momento determinado está correlacionada con los errores de otros períodos (una correlación serial). Esto exige determinar la distribución de cada ε_{nt} condicionado al valor las variables dependientes de otros períodos. La probabilidad de elegir está basada en esa distribución condicionada

en lugar de la incondicionada $\phi(.)$ que acabamos de usar. Segundo, a veces el investigador no observa a los individuos desde su primera elección, sino en algún momento posterior. Por ejemplo, la situación de desempleo de una persona, pero no desde que tenía edad para formar parte de población activa. En un caso así, la probabilidad de elección en el primer período observado depende necesariamente de las elecciones pasadas, no observadas. Dicha primera probabilidad de elección observada debe representarse adecuadamente para poder obtenerse una estimación consistente. Este es el famoso *problema de las condiciones iniciales* en los problemas de elección dinámica. Ambos problemas han sido tratados por Heckman (1981a y 1981b) y Heckman y Singer (1986), aunque el tema tiene gran complejidad. Papatla y Krishnamurthi (1992) evitan el problema asumiendo, simplemente, que los factores inobservados son independientes en el tiempo.