5

Probit

5.1 Probabilidades de elección

El modelo logit tiene tres limitaciones importantes. (1) No puede representar la variación aleatoria de preferencias. (2) Presenta patrones de sustitución restrictivos debido a la propiedad de IIA. Y (3) no puede utilizarse con datos de panel cuando los factores no observados están correlacionados en el tiempo para cada decisor. Los modelos GEV relajan la segunda de estas restricciones, pero no las otros dos. Los modelos probit abordan las tres limitaciones. Pueden manejar variación de preferencias aleatoria, permiten cualquier patrón de sustitución y son aplicables a datos de panel con errores correlacionados temporalmente.

La única limitación de los modelos probit es que requieren distribuciones normales para todos los componentes no observados de utilidad. En muchos casos, quizá en la mayoría de las situaciones, las distribuciones normales proporcionan una representación adecuada de los componentes aleatorios. Sin embargo, en algunas situaciones las distribuciones normales son inadecuadas y pueden conducir a predicciones perversas. Un ejemplo destacado es el de los coeficientes de precios. Para un modelo probit con variación de preferencias aleatoria, el coeficiente de precio se supone distribuido normalmente en la población. Dado que la distribución normal tiene densidad en ambos lados del cero, el modelo implica necesariamente que algunas personas tienen un coeficiente de precio positivo (preferencia por precios más caros). El uso de una distribución con densidad en un solo lado del cero, como la distribución logarítmica normal (log-normal) es más apropiado y, sin embargo, no puede ser acomodado dentro de un modelo probit. Exceptuando esta restricción, probit es bastante general.

El modelo probit se obtiene bajo el supuesto de que las utilidades no observadas siguen una distribución normal conjunta. La primera formulación de un probit binario a cargo de Thurstone (1927) utilizaba la terminología de estímulos psicológicos, terminología que Marschak (1960) tradujo a términos económicos como utilidad. Hausman y Wise (1978) y Daganzo (1979) dilucidaron la generalidad de la especificación para representar diversos aspectos del comportamiento de elección. Empecemos descomponiendo la utilidad en su parte observada y su parte no observada: $U_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$. Considere el vector compuesto por cada ε_{nj}, etiquetado como $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$. Supondremos que ε_n se distribuye de acuerdo a una distribución normal con un vector de medias cero y una matriz de covarianza Ω. La densidad de ε_n es

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

La covarianza Ω puede depender de variables percibidas por el decisor n, por lo que Ω_n sería la notación más apropiada; sin embargo, omitimos el subíndice en aras de la simplicidad.

La probabilidad elección es

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

$$= \int I\big(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall j \neq i\big)\phi(\varepsilon_n) d\varepsilon_n, \qquad (5.1)$$

donde $I(\cdot)$ es un indicador de si la expresión entre paréntesis es verdadera y la integral es sobre todos los valores de ε_n. Esta integral no tiene una forma cerrada. Debe ser evaluada numéricamente mediante simulación.

Las probabilidades de elección pueden expresarse de otras dos maneras que son útiles para la simulación de la integral. Sea B_{ni} el conjunto de valores de los términos de error que producen que la elección del decisor sea la alternativa i: $B_{ni} = \big\{\varepsilon_n \text{ s.t. } V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall \text{j} \neq \text{i}\big\}$. Entonces

$$P_{ni} = \int_{\varepsilon_n \in B_{ni}} \phi(\varepsilon_n) d\varepsilon_n, \qquad (5.2)$$

que es una integral sobre algunos de los valores de ε_n en lugar de sobre todos los valores posibles, es decir, es sobre los valores de B_{ni}.

Las expresiones (5.1) y (5.2) son integrales J-dimensionales sobre los J errores $\varepsilon_{nj}, j = 1, \dots, J$. Dado que sólo las diferencias de utilidad importan, las probabilidades de elección pueden expresarse de manera equivalente como integrales $(\text{J} - 1)$-dimensionales sobre las diferencias entre errores. Definamos las diferencias respecto a la alternativa i, la alternativa para la que estamos calculando la probabilidad de elección. Definimos $\widetilde{U}_{nji} = U_{nj} - U_{ni}$, $\tilde{V}_{nji} = V_{nj} - V_{ni}$ y $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$. Entonces $P_{ni} = Prob(\widetilde{U}_{nji} < 0 \; \forall j \neq i)$. Es decir, la probabilidad de elegir la alternativa i es la probabilidad de que todas las diferencias de utilidad, cuando se refieren a la alternativa i, sean negativas. Definimos el vector $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \dots, \tilde{\varepsilon}_{nji} \rangle$ donde los puntos "..." representan todas las alternativas excepto i, de manera que $\tilde{\varepsilon}_{ni}$ tiene dimensión $J - 1$. Dado que la diferencia entre dos variables normales es normal, la densidad de las diferencias de error es

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

donde $\widetilde{\Omega}_i$ es la covarianza de $\tilde{\varepsilon}_{ni}$, obtenida a partir de Ω. La probabilidad de elección expresada en diferencias de utilidad es por tanto

$$P_{ni} = \int I\big(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i\big)\phi(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}, \qquad (5.3)$$

que es una integral $(\text{J} - 1)$-dimensional sobre todos los valores posibles de las diferencias de error. Una expresión equivalente es

$$P_{ni} = \int_{\tilde{\varepsilon}_{ni} \in \tilde{B}_{ni}} \phi(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}, \qquad (5.4)$$

donde $\tilde{B}_{ni} = \{\tilde{\varepsilon}_{ni} \text{ s.t. } \tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i\}$, que es una integral $(\mathrm{J}-1)$-dimensional sobre las diferencias de error en $\tilde{B}_{ni}$.

Las expresiones (5.3) y (5.4) utilizan la matriz de covarianza $\widetilde{\Omega}_i$ de las diferencias de error. Hay una manera directa de obtener $\widetilde{\Omega}_i$ a partir de la covarianza de los errores, Ω. Sea M_i la matriz identidad de dimensión $(\mathrm{J}-1)$ con una columna adicional de -1s agregada como columna i-ésima. La columna adicional hace que la matriz tenga dimensiones $(\mathrm{J}-1) \times J$. Por ejemplo, con $J = 4$ alternativas e $i = 3$,

$$M_i = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

Esta matriz puede ser usada para transformar la matriz de covarianza de los errores en la matriz de covarianza de las diferencias entre errores: $\widetilde{\Omega}_i = M_i \Omega M_i'$. Observe que $\widetilde{\Omega}_i$ es de dimensión $(\mathrm{J}-1) \times (J-1)$, mientras que Ω es de dimensión $J \times J$, ya que M_i es $(J-1) \times J$. A modo ilustrativo, considere una situación de tres alternativas con errores $\langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle$ que tiene covarianza

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

Supongamos que calculamos las diferencias de error respecto la alternativa 2. Sabemos por postulación que las diferencias de error $\langle \tilde{\varepsilon}_{n12}, \tilde{\varepsilon}_{n32} \rangle$ tienen covarianza

$$\widetilde{\Omega}_2 = Cov \begin{pmatrix} \varepsilon_{n1} - \varepsilon_{n2} \\ \varepsilon_{n3} - \varepsilon_{n2} \end{pmatrix}$$

$$\begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.$$

Esta matriz de covarianza también se puede obtener a partir de la transformación $\widetilde{\Omega}_2 = M_2 \Omega M_2'$:

$$\widetilde{\Omega}_n = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix}$$

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

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

$$\begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.$$

Como veremos, esta transformación mediante M_i es muy práctica cuando se simulan probabilidades probit.

5.2 Identificación

Tal y como se describió en la sección 2.5, cualquier modelo de elección discreta debe estar normalizado para tener en cuenta el hecho de que el nivel y la escala de la utilidad son irrelevantes. El nivel de utilidad es irrelevante porque se puede añadir una constante a la utilidad de todas las alternativas sin cambiar la alternativa que tiene mayor utilidad: la alternativa con la utilidad más alta antes de añadir la constante sigue siendo la de mayor utilidad después de la adición. Del mismo modo, la escala de la utilidad no importa porque la utilidad de cada alternativa puede ser multiplicada por una constante (positiva) sin cambiar la alternativa que tiene mayor utilidad. En los modelos logit y logit jerárquico la normalización de escala y de nivel se produce de forma automática bajo los supuestos que se realizan relativos a la distribución de los términos de error. Como resultado, para estos modelos no es necesario considerar de forma explícita la normalización. Con modelos probit, sin embargo, la normalización de escala y de nivel no se produce automáticamente. El investigador debe normalizar el modelo directamente.

La normalización del modelo está relacionada con la identificación de parámetros. Un parámetro es *identificado* si se puede estimar, y es *no identificado* si no puede ser estimado. Un ejemplo de un parámetro no identificado es k en la especificación de la utilidad: $U_{nj} = V_{nj} + \mathrm{k} + \varepsilon_{nj}$. Aunque el investigador podría escribir la utilidad de esta forma y podría intentar estimar k para obtener una medida del nivel general de utilidad, eso es imposible. El comportamiento del decisor no se ve afectado por k, por lo que el investigador no puede deducir su valor a partir de las elecciones que los decisores han hecho. Dicho de forma directa, los parámetros que no afectan el comportamiento de los decisores no pueden ser estimados. En un modelo no normalizado, pueden aparecer parámetros no identificados; estos parámetros se refieren a la escala y al nivel de la utilidad, algo que no afecta al comportamiento. Una vez que el modelo se normaliza, estos parámetros desaparecen. La dificultad reside en que no siempre es evidente qué parámetros se refieren a la escala y al nivel de utilidad. En el ejemplo anterior, el hecho de que k es un parámetro no identificado es bastante obvio. En muchos casos, no es en absoluto evidente qué parámetros son identificados. Bunch y Kitamura (1989) han demostrado que los modelos probit que aparecen en varios artículos publicados no están normalizados y contienen parámetros no identificados. El hecho de que ni los autores ni los revisores de estos artículos pudiesen detectar que los modelos no estaban normalizados es un buen testimonio de la complejidad de la cuestión.

A continuación proporciono un procedimiento que siempre puede ser usado para normalizar un modelo probit y asegurar que todos los parámetros son identificados. No es el único procedimiento que puede usarse; véase, por ejemplo, Bunch (1991). En algunos casos, un investigador puede encontrar otros procedimientos de normalización más convenientes. Sin embargo, el procedimiento que facilito siempre se puede utilizar, ya sea por sí mismo o como un control sobre otro procedimiento.

Describo el procedimiento para un modelo de cuatro alternativas. La generalización a más alternativas es obvia. Como de costumbre, la utilidad se expresa como $U_{nj} = V_{nj} + \varepsilon_{nj},\ j = 1, \ldots, 4$. El vector de errores es $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$. Este vector se distribuye normalmente con media cero y una matriz de covarianza que se puede expresar de forma explícita como

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix}, \tag{5.5}$$

donde los puntos se refieren a los elementos correspondientes en la parte superior de la matriz. Tenga en cuenta que hay diez elementos en esta matriz, es decir, diez σs distintas que representan las

varianzas y covarianzas entre los cuatro errores. En general, un modelo con J alternativas tiene $J(J+1)/2$ elementos distintos en la matriz de covarianza de los errores.

Para tener en cuenta el hecho de que el nivel de utilidad es irrelevante, usamos diferencias de utilidad. En mi procedimiento, siempre uso las diferencias respecto a la primera alternativa, ya que simplifica el análisis tal y como veremos. Definimos las diferencias de error como $\tilde{\varepsilon}_{nj1} = \varepsilon_{nj} - \varepsilon_{n1}$ para $j = 2,3,4$ y definimos el vector de las diferencias de error como $\tilde{\varepsilon}_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}, \tilde{\varepsilon}_{n41} \rangle$. Observe que el subíndice 1 en $\tilde{\varepsilon}_{n1}$ significa que las diferencias de error son respecto a la primera alternativa, en lugar de indicar que los errores son de la primera alternativa.

La matriz de covarianza para el vector de las diferencias de error toma la siguiente forma

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

donde las θs se relacionan con las σs originales así:

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

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

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

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

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

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

Computacionalmente, esta matriz se puede obtener utilizando la matriz de transformación M_i definida en la Sección 5.1 como $\widetilde{\Omega}_1 = M_1 \Omega M_1'$.

Para ajustar la escala de la utilidad, uno de los elementos de la diagonal se normaliza. Yo fijo el elemento de la parte superior izquierda de $\widetilde{\Omega}_1$, que es la varianza de $\tilde{\varepsilon}_{n21}$, a 1. Esta normalización de la escala nos da la siguiente matriz de covarianza:

(5.6)

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

Las θ^*s se relacionan con las σs originales como sigue:

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

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

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

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

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

Hay cinco elementos en $\tilde{\Omega}_1^*$. Estos son los únicos parámetros identificados del modelo. Este número es menor a los diez elementos que entran en Ω. Cada θ^* es una función de las σs. Puesto que hay cinco θ^*s y diez σs, no es posible resolver para todas las σs a partir de los valores estimados de las θ^*s. Por tanto, no es posible obtener estimaciones para todas las σs.

En general, un modelo con J alternativas y una matriz de covarianza sin restricciones tendrá $[(\mathrm{J}-1)\mathrm{J}/2]-1$ parámetros de covarianza al ser normalizado, en comparación con los $\mathrm{J}(\mathrm{J}+1)/2$ parámetros cuando no se normaliza. Sólo $[(\mathrm{J}-1)\mathrm{J}/2]-1$ parámetros son identificados. Esta reducción en el número de parámetros *no* es una restricción. La reducción en el número de parámetros es una normalización que simplemente elimina los aspectos irrelevantes de la matriz de covarianza original, que son la escala y el nivel de utilidad. Los diez elementos en Ω permiten una varianza y una covarianza debida simplemente a la escala y al nivel de utilidad, que no tiene relevancia en el comportamiento de los decisores. Sólo los cinco elementos de $\tilde{\Omega}_1^*$ contendrán información acerca de la varianza y covarianza de los errores con independencia de la escala y el nivel de utilidad. En este sentido, sólo los cinco parámetros tienen contenido económico y sólo esos cinco parámetros se pueden estimar.

Supongamos ahora que el investigador impone una estructura sobre la matriz de covarianza. Es decir, en lugar de permitir una matriz de covarianza completa para los errores, el investigador considera que los errores siguen un patrón que implica ciertos valores particulares para - o ciertas relaciones entre - los elementos de la matriz de covarianza. El investigador restringe la matriz de covarianza para incorporar este patrón.

La estructura puede adoptar diversas formas, dependiendo de la aplicación. Yai et al . (1997) estiman un modelo probit de elección de ruta donde la covarianza entre dos rutas cualesquiera sólo depende de la longitud de los segmentos de ruta compartidos; esta estructura reduce el número de parámetros de covarianza a uno solo, que captura la relación de la covarianza con la longitud compartida. Bolduc et al. (1996) estiman un modelo de elección de ubicación de médicos donde la covarianza entre ubicaciones es una función de la proximidad entre las propias ubicaciones, usando lo que Bolduc (1992) ha denominado una estructura "generalizada auto-regresiva". Haaijer et al . (1998) imponen una estructura factor-analítica (*factor-analytic structure*) que surge de coeficientes aleatorios de las variables explicativas; este tipo de estructura se describe en detalle en la Sección 5.3. Elrod and Keane (1995) también imponen una estructura factor-analítica, pero que surge a partir de componentes de error en lugar de coeficientes aleatorios per se.

A menudo la estructura impuesta será suficiente para normalizar el modelo. Es decir, las restricciones que el investigador impone a la matriz de covarianza para ajustar sus expectativas acerca de la forma en que los errores se relacionan entre sí, servirán también para normalizar el modelo. Sin embargo, esto no siempre sucede. Los ejemplos citados por Bunch y Kitamura (1989) son casos en que las restricciones que el investigador aplicaba a la matriz de covarianza parecían suficientes para normalizar el modelo, pero en realidad no lo eran.

El procedimiento que he facilitado en el texto anterior se puede utilizar para determinar si las restricciones aplicadas a la matriz de covarianza son suficientes para normalizar el modelo. El investigador especifica Ω con sus restricciones sobre los elementos de la matriz. A continuación, el procedimiento indicado se utiliza para obtener $\tilde{\Omega}_1^*$, que está normalizada para la escala y el nivel. Sabemos que cada elemento de $\tilde{\Omega}_1^*$ es identificado. Si cada uno de los elementos restringidos de Ω puede ser calculado a partir de los elementos de $\tilde{\Omega}_1^*$, entonces las restricciones son suficientes para normalizar el modelo. En este caso, cada parámetro de la Ω restringida es identificado. Por otro lado, si los elementos de Ω no se pueden calcular a partir de los elementos de $\tilde{\Omega}_1^*$, las restricciones no son suficientes para normalizar el modelo y los parámetros de Ω no son identificados.

Para ilustrar este enfoque, supongamos que el investigador está estimando un modelo con cuatro alternativas y asume que la matriz de covarianza de los errores tiene la siguiente forma:

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

Esta matriz de covarianza permite que el primer y el segundo error estén correlacionados, al igual que el error de la tercera y la cuarta alternativas, pero no permite ninguna otra correlación. La correlación entre los pares apropiados es $\rho/(1+\rho)$. Observe que mediante la especificación de los elementos de la diagonal como $1+\rho$, el investigador asegura que la correlación está entre -1 y 1 para cualquier valor de $\rho \geq -\frac{1}{2}$, como se requiere para una correlación. ¿Está este modelo, tal y como se especifica, normalizado para la escala y el nivel? Para responder a esta pregunta, aplicamos el procedimiento descrito. En primer lugar, tomamos las diferencias respecto a la primera alternativa. La matriz de covarianza de las diferencias de error es

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

donde las θs se refieren a las σs originales según las siguientes relaciones:

$$\theta_{22} = 2$$

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

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

$$\theta_{23} = 1,$$

$$\theta_{24} = 1,$$

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

A continuación normalizamos la escala estableciendo el elemento superior izquierdo a 1. La matriz de covarianza normalizada es

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

donde las θ^*s se relacionan con las σs originales por las siguientes fórmulas:

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

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

$$\theta_{23}^* = \frac{1}{2},$$

$$\theta_{24}^* = \frac{1}{2},$$

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

Observe que $\theta_{33}^* = \theta_{44}^* = \theta_{34}^* + \frac{1}{2}$ y que el resto de θ^*s tienen valores fijos. Sólo hay un parámetro en $\tilde{\Omega}_1^*$, tal y como sucedía en Ω. Definimos $\theta = 1 + \rho$. La matriz $\tilde{\Omega}_1^*$ resulta

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

El parámetro ρ original puede ser calculado directamente a partir de θ. Por ejemplo, si θ se estima que es 2.4, entonces la estimación de ρ es $\theta - 1 = 1.4$ y la correlación es $1.4/2.4 = 0.58$. El hecho de que los parámetros que entran en Ω puedan calcularse a partir de los parámetros que entran en la matriz de covarianza normalizada $\tilde{\Omega}_1^*$ significa que el modelo original ya está normalizado para la escala y el nivel de utilidad. Es decir, las restricciones que el investigador ha colocado en Ω también han proporcionado al mismo tiempo la normalización necesaria.

A veces, las restricciones en la matriz de covarianza original pueden parecer suficientes para normalizar el modelo, pero realmente no es así. Aplicar nuestro procedimiento determinará si realmente es el caso. Considere el modelo del ejemplo anterior, pero ahora supongamos que el investigador permite una correlación diferente entre el primer y segundo error, a la permitida entre el tercer y cuarto error. La matriz de covarianza de errores se especifica como

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

La correlación entre el primer y el segundo error es $\rho_1/(1+\rho_1)$, y la correlación entre el tercer y el cuarto error es $\rho_2/(1+\rho_2)$. Podemos obtener $\tilde{\Omega}_1$ para las diferencias de error y luego obtener $\tilde{\Omega}_1^*$ estableciendo el elemento superior izquierdo de $\tilde{\Omega}_1$ a 1. La matriz resultante es

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

donde ahora $\theta = 1 + (\rho_1 + \rho_2)/2$. Los valores de ρ_1 y ρ_2 no se pueden calcular a partir de un valor de θ. Por lo tanto, el modelo original no está normalizado para la escala y el nivel, y los parámetros ρ_1 y ρ_2 no son identificados. Este hecho es algo sorprendente, ya que sólo dos parámetros entran en la matriz de covarianza original Ω. Parecería, a menos que el investigador explícitamente haga la prueba que acabamos de hacer, que restringir la matriz de covarianza para que conste de sólo dos elementos debería ser suficiente para normalizar el modelo. En este caso, sin embargo, no es así.

En el modelo normalizado, sólo aparece el promedio de la ρs: $(\rho_1 + \rho_2)/2$. Es posible calcular la ρ promedio a partir de θ, simplemente como $\theta - 1$. Esto significa que la ρ promedio es identificada, pero no los valores individuales. Cuando $\rho_1 = \rho_2$, como en el ejemplo anterior, el modelo queda normalizado porque cada ρ es igual a la ρ promedio. Sin embargo, como vemos ahora, cualquier modelo con el mismo promedio de ρ es equivalente, después de normalizar la escala y el nivel. Por lo tanto, asumir que $\rho_1 = \rho_2$ es lo mismo que asumir que $\rho_1 = 3\rho_2$ o cualquier otra relación. Lo único que importa para el comportamiento es el promedio de estos parámetros, no sus valores relativos entre ellos. Este hecho es bastante sorprendente y sería difícil darse cuenta del mismo sin el uso de nuestro procedimiento para la normalización.

Ahora que sabemos cómo asegurar que un modelo probit está normalizado para el nivel y la escala, y que por lo tanto contiene únicamente información económicamente relevante, podemos examinar cómo se utiliza el modelo probit para representar distintos tipos de situaciones de elección. Nos fijaremos en tres situaciones en las que los modelos logit presentan limitaciones y mostraremos cómo estas limitaciones se superan con probit. Estas situaciones son las variaciones de preferencia, los patrones de sustitución y las elecciones repetidas a lo largo del tiempo.

5.3 Variaciones de preferencia

Probit se adapta particularmente bien a la incorporación de coeficientes aleatorios, a condición de que los coeficientes se distribuyan normalmente. Hausman y Wise (1978) fueron los primeros, que yo tenga conocimiento, en desarrollar esta formulación. Haaijer et al. (1998) proporcionan una aplicación convincente de su uso. Suponga que la utilidad representativa es lineal en los parámetros y que los coeficientes varían aleatoriamente entre los decisores en lugar de estar fijados como se ha supuesto hasta ahora en este libro. La utilidad es $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, donde β_n es el vector de coeficientes para el decisor n que representa las preferencias de esa persona. Supongamos que β_n presenta una distribución normal en la población con media b y covarianza W: $\beta_n \sim N(b, W)$. El objetivo de la investigación es estimar los parámetros b y W.

La utilidad puede ser reescrita descomponiendo β_n entre su media y las desviaciones de su media: $U_{nj} = b' x_{nj} + \tilde{\beta}_n' x_{nj} + \varepsilon_{nj}$, donde $\tilde{\beta}_n = \beta_n - b$. Los dos últimos términos de la utilidad son aleatorios; denominemos η_{nj} la suma de ambos términos aleatorios para obtener $U_{nj} = b' x_{nj} + \eta_{nj}$. La covarianza de los η_{nj}s depende de W así como de las x_{nj}s, por lo que la covarianza difiere entre decisores.

La covarianza de los términos η_{nj}s se puede describir fácilmente para un modelo con dos alternativas y una variable explicativa. En este caso, la utilidad es

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

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

Supongamos que β_n se distribuye normalmente con media b y varianza σ_β. Supongamos que ε_{n1} y ε_{n2} se distribuyen de forma independiente e idéntica con varianza σ_ε. El supuesto de independencia es para este ejemplo y no es necesario en general. La utilidad se reescribe entonces como

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

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

donde η_{n1} y η_{n2} se distribuyen de acuerdo a una distribución normal conjunta. Cada una tiene una media de cero: $\mathrm{E}(\eta_{nj}) = \mathrm{E}(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = 0$. La covarianza se determina como sigue. La varianza de cada una es $\mathrm{V}(\eta_{nj}) = \mathrm{V}(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = x_{nj}^2\sigma_\beta + \sigma_\varepsilon$. Su covarianza es

$$Cov(\eta_{n1}, \eta_{n2}) = E\left[(\tilde{\beta}_n x_{n1} + \varepsilon_{n1})(\tilde{\beta}_n x_{n2} + \varepsilon_{n2})\right] =$$

$$E\left(\tilde{\beta}_n^2 x_{n1}x_{n2} + \varepsilon_{n1}\varepsilon_{n2} + \varepsilon_{n1}\tilde{\beta}_n x_{n2} + \varepsilon_{n2}\tilde{\beta}_n x_{n1}\right) =$$

$$x_{n1}x_{n2}\sigma_\beta.$$

La matriz de covarianza es

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

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

Se requiere un último paso para la estimación. Recordemos que el comportamiento de los decisores no se ve afectado por una transformación multiplicativa de la utilidad. Por lo tanto, necesitamos establecer la escala de la utilidad. Una normalización conveniente para este caso es $\sigma_\varepsilon = 1$. Bajo esta normalización

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

Los valores de x_{n1} y x_{n2} son observados por el investigador y los parámetros b y σ_β son estimados. De esta forma, el investigador obtiene tanto la media como la varianza del coeficiente aleatorio en la población. La generalización de este caso a más de una variable explicativa y más de dos alternativas es directa.

5.4 Patrones de sustitución y fallo de la IIA

Probit puede representar cualquier patrón de sustitución. Las probabilidades probit no exhiben la propiedad de IIA que da lugar a la sustitución proporcional de logit. Diferentes matrices de covarianza Ω proporcionan diferentes patrones de sustitución y, mediante la estimación de la matriz de covarianza, el investigador determina el patrón de sustitución que es más adecuado para sus datos.

Es posible estimar una matriz de covarianza completa o, alternativamente, el investigador puede imponer cierta estructura en la matriz de covarianza para representar fuentes particulares de no independencia. Esta estructura suele reducir el número de parámetros y facilita su interpretación. Consideraremos en primer lugar la situación en la que el investigador estima una matriz de covarianza completa y posteriormente una situación en la que el investigador impone una estructura sobre la matriz de covarianza.

Covarianza Completa: patrones de sustitución no restringidos

Para simplificar la notación, considere un modelo probit con cuatro alternativas. Una matriz de covarianza completa para los componentes no observados de utilidad toma la forma de Ω en (5.5). Cuando normalizamos la escala y el nivel, la matriz de covarianza se convierte en $\widetilde{\Omega}_1^*$ en (5.6). Los elementos de $\widetilde{\Omega}_1^*$ son estimados. Los valores estimados pueden representar cualquier tipo de patrón de sustitución; es importante destacar que la normalización de la escala y el nivel no restringe los patrones de sustitución. La normalización sólo elimina los aspectos de Ω que son irrelevantes para el comportamiento.

Observe, sin embargo, que los valores estimados de las θ^*s no proporcionan esencialmente ninguna información interpretable en ellas mismas (Horowitz, 1991). Por ejemplo, supongamos que estimamos que θ_{33}^* es mayor que θ_{44}^*. Puede ser tentador interpretar este resultado como una indicación de que la varianza en la utilidad no observada de la tercera alternativa es mayor que la de la cuarta alternativa, es decir, que $\sigma_{33} > \sigma_{44}$. Sin embargo, esta interpretación no es correcta. Es perfectamente posible que $\theta_{33}^* > \theta_{44}^*$ y sin embargo $\sigma_{44} > \sigma_{33}$, si la covarianza σ_{14} es suficientemente mayor a σ_{13}. Del mismo modo, supongamos que se estima que el valor θ_{23}^* es negativo. Esto no significa que la utilidad no observada de la segunda alternativa se correlacione negativamente con la utilidad no observada de la tercera alternativa (es decir, $\sigma_{23} < 0$). Es posible que σ_{23} sea positiva y sin embargo σ_{12} y σ_{13} sean suficientemente grandes para hacer θ_{23}^* sea negativa. El punto a destacar aquí es que la estimación de una matriz de covarianza completa permite que el modelo represente cualquier patrón de sustitución, pero hace que los parámetros estimados sean esencialmente imposibles de interpretar.

Covarianza estructurada: patrones de sustitución restringidos

Al imponer estructura en la matriz de covarianza, los parámetros estimados por lo general se vuelven más interpretables. La estructura es una restricción en la matriz de covarianza y, como tal, reduce la capacidad del modelo de representar diversos patrones de sustitución. Sin embargo, si la estructura es correcta (es decir, representa realmente el comportamiento de los decisores), entonces el verdadero patrón de sustitución podrá ser representado por la matriz de covarianza restringida.

La estructura es necesariamente dependiente de la situación: una estructura adecuada para una matriz de covarianza depende de las características específicas de la situación de elección que se está modelando. En la sección 5.2 se describen varios estudios que utilizan diferentes tipos de estructura. Como ejemplo de cómo se puede imponer estructura a la matriz de covarianza y por lo tanto a los patrones de sustitución, considere la elección que un comprador de vivienda hace entre diferentes tipos de hipotecas. Supongamos que el comprador puede escoger entre cuatro hipotecas de cuatro entidades financieras diferentes: una con un tipo de interés fijo y tres con tipos variables. Supongamos que la parte no observada de la utilidad de esta decisión se compone de dos partes: la preocupación del

comprador de vivienda por el riesgo de incremento de los tipos de interés, etiquetada r_n, y que es común a todos los préstamos de tipo variable, y todos los demás factores no observados, etiquetados colectivamente η_{nj}. El componente no observado de utilidad es por lo tanto

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

donde $d_j = 1$ para los préstamos a tipo variable y 0 para el préstamo a tipo fijo, y donde el signo negativo indica que la utilidad disminuye a medida que la preocupación por el riesgo aumenta. Supongamos que r_n se distribuye normalmente sobre los compradores de vivienda con varianza σ y que η_{nj} $\forall j$ es normal iid con media cero y varianza ω. La matriz de covarianza para $\varepsilon_n = \langle \varepsilon_{n1}, \dots, \varepsilon_{n4} \rangle$ es

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

El modelo necesita ser normalizado para la escala pero, como veremos, ya está normalizado para el nivel. La covarianza de las diferencias de error es

$$\Omega = \begin{pmatrix} \sigma & \sigma & \sigma \\ \cdot & \sigma & \sigma \\ \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 2 & 1 & 1 \\ \cdot & 2 & 1 \\ \cdot & \cdot & 2 \end{pmatrix}.$$

Esta matriz no tiene menos parámetros que Ω. Es decir, el modelo ya estaba normalizado para el nivel. Para normalizar la escala, fijamos $\sigma + 2\omega = 1$. La matriz de covarianza se convierte en

$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta & \theta \\ \cdot & 1 & \theta \\ \cdot & \cdot & 1 \end{pmatrix},$$

donde $\theta = (\sigma + \omega)/(\sigma + 2\omega)$. Los valores de σ y ω no se pueden calcular a partir de θ. Sin embargo, el parámetro θ proporciona información sobre la varianza en la utilidad debida a la preocupación por el riesgo respecto a la varianza debida a todos los demás factores no observados. Por ejemplo, supongamos que se estima θ en 0.75. Esta estimación puede ser interpretada como una indicación de que la varianza en la utilidad atribuible a la preocupación sobre el riesgo es dos veces la varianza en la utilidad atribuible a todos los demás factores:

$$\theta = 0.75,$$

$$\frac{\sigma + \omega}{\sigma + 2\omega} = 0.75,$$

$$\sigma + \omega = 0.75\sigma + 1.5\omega,$$

$$0.25\sigma = 0.5\omega,$$

$$\sigma = 2\omega,$$

Dicho de forma equivalente, $\hat{\theta} = 0.75$ significa que la preocupación por el riesgo representa dos tercios de la varianza en el componente no observado de la utilidad.

Dado que el modelo original ya estaba normalizado para el nivel, el modelo podría ser estimado sin reformular la matriz de covarianza en términos de las diferencias de error. La normalización de la escala podría lograrse simplemente estableciendo $\omega = 1$ en la Ω original. Usando este procedimiento, el parámetro σ es estimado directamente. Su valor en relación a 1 indica la varianza debida a la preocupación por el riesgo en relación a la varianza debida a la percepción sobre la facilidad de tratar con cada entidad financiera. Una estimación de $\hat{\theta} = 0.75$ corresponde a una estimación de $\hat{\sigma} = 2$.

5.5 Datos de panel

El modelo probit para elecciones repetidas es similar al probit para una elección por decisor. La única diferencia es que la dimensión de la matriz de covarianza de los errores se ve expandida. Considere un decisor que se enfrenta a una elección entre J alternativas en cada uno de los T períodos de tiempo o situaciones de elección. Las alternativas pueden cambiar a lo largo del tiempo, y J y T pueden diferir para diferentes decisores; sin embargo, suprimimos la notación para estas posibilidades. La utilidad que el decisor n obtiene de la alternativa j en el período T es $U_{njt} = V_{njt} + \varepsilon_{njt}$. En general, sería de esperar que ε_{njt} estuviese correlacionado en el tiempo así como respecto a otras alternativas, dado que los factores no observados por el investigador pueden persistir a lo largo del tiempo. Denotemos el vector de errores para todas las alternativas en todos los períodos de tiempo como $\varepsilon_n' = \langle \varepsilon_{n11}, \dots, \varepsilon_{nJ1}, \varepsilon_{n12}, \dots, \varepsilon_{nJ2}, \dots, \varepsilon_{n1T}, \dots, \varepsilon_{nJT} \rangle$. La matriz de covarianza para este vector se denomina Ω, que tiene dimensiones JT × JT.

Considere una secuencia de alternativas concreta, una alternativa para cada período de tiempo, $\mathbf{i} = \{i_1, \dots, i_T\}$. La probabilidad de que el decisor haga esta secuencia de elecciones es

$$P_{n\boldsymbol{i}} = Prob(U_{ni_tt} > U_{njt} \; \forall j \neq i_t, \forall t)$$

$$= Prob(V_{ni_tt} + \varepsilon_{ni_tt} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t)$$

$$= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n,$$

donde $B_n = \{\varepsilon_n \; s.t. V_{ni_tt} + \varepsilon_{ni_tt} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t\}$ y $\phi(\varepsilon_n)$ es la densidad normal conjunta con media cero y covarianza Ω. En comparación con la probabilidad probit para una situación de una única elección, la integral simplemente se ha ampliado hasta JT dimensiones en lugar de J.

A menudo es más conveniente trabajar con diferencias de utilidad. La probabilidad de la secuencia $\mathbf{i}$ es la probabilidad de que las diferencias de utilidad sean negativas para cada alternativa en cada período de tiempo, cuando las diferencias se calculan respecto a la alternativa identificada por $\mathbf{i}$ para ese período de tiempo:

$$P_{ni} = Prob(\tilde{U}_{nji_tt} < 0 \; \forall j \neq i_t, \forall t)$$

$$= \int_{\tilde{\varepsilon}_n \in \tilde{B}_n} \phi(\tilde{\varepsilon}_n) d\tilde{\varepsilon}_n,$$

donde $\tilde{U}_{nji_tt} = U_{njt} - U_{ni_tt}$; $\tilde{\varepsilon}_n' = \langle (\varepsilon_{n11} - \varepsilon_{ni_11}), \dots, (\varepsilon_{nJ1} - \varepsilon_{ni_11}), \dots, (\varepsilon_{n1T} - \varepsilon_{ni_TT}), \dots, (\varepsilon_{nJT} - \varepsilon_{ni_TT}) \rangle$ con cada "…" refiriéndose a todas las alternativas excepto i, y $\tilde{B}_n$ es el conjunto de $\tilde{\varepsilon}_n$s para las que $\tilde{U}_{nji_tt} < 0 \; \forall j \neq i_t, \forall t$. Esta es una integral $(J-1)$T-dimensional. La densidad $\phi(\tilde{\varepsilon}_n)$ se distribuye

normalmente con matriz de covarianza obtenida a partir de Ω. La simulación de la probabilidad de elección es la misma que para situaciones con una elección por decisor, descrita en la Sección 5.6, pero con una dimensión mayor tanto en la matriz de covarianza como en la integral. Borsch-Supan et al. (1991) proporcionan un ejemplo de un probit multinomial sobre datos de panel que permite covarianza en el tiempo y entre alternativas.

Para elecciones binarias, tales como si una persona compra un producto en particular en cada período de tiempo o si trabaja en un puesto de trabajo remunerado cada mes, el modelo probit se simplifica considerablemente (Gourieroux y Monfort, 1993). La utilidad neta de tomar la acción (por ejemplo, trabajar) en el período t es $\mathrm{U}_{nt} = \mathrm{V}_{nt}+\varepsilon_{nt}$, y la persona realiza la acción si $\mathrm{U}_{nt} > 0$. Esta utilidad se llama utilidad neta, ya que es la diferencia entre la utilidad de tomar la acción y no tomarla. Como tal, ya está expresada en términos de diferencias. Los errores están correlacionados en el tiempo, y la matriz de covarianza para $\varepsilon_{n1}, \ldots, \varepsilon_{nT}$ es Ω, que es de dimensiones $\mathrm{T} \times \mathrm{T}$.

Una secuencia de elecciones binarias se representa más fácilmente por un conjunto de T variables indicadoras (*dummy*): $d_{nt} = 1$ si la persona n ha tomado la acción en el período t y $d_{nt} = -1$ en caso contrario. La probabilidad de que se produzca la secuencia de elecciones $d_n = d_{n1}, \ldots, d_{nT}$ es

$$P_{nd_n} = Prob(U_{nt}d_{nt} > 0 \; \forall t)$$

$$= Prob(V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t)$$

$$= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n,$$

donde B_n es el conjunto de ε_ns para las que $V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t$ y $\phi(\varepsilon_n)$ es la densidad normal conjunta con covarianza Ω.

Es posible aplicar cierta estructura a la covarianza de los errores a lo largo del tiempo. Supongamos que en el caso binario, por ejemplo, el error consiste en una parte que fija específica del decisor, reflejando en qué medida es proclive a tomar la acción, y una parte que varía en el tiempo para cada decisor: $\varepsilon_{nt} = \eta_n+\mu_{nt}$, donde μ_{nt} es iid a lo largo del tiempo y entre personas, con una densidad normal estándar, y η_n es iid entre personas con una densidad de probabilidad 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 diferentes t y s es $\mathrm{Cov}(\varepsilon_{nt}, \varepsilon_{ns}) = \mathrm{E}(\eta_n+\mu_{nt})(\eta_n+\mu_{ns}) = \sigma$. Por tanto, la matriz de covarianza toma la forma

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

Sólo un parámetro, σ, entra en la matriz de covarianza. Su valor indica la varianza en la utilidad no observada entre individuos (la varianza de η_n) en relación con la varianza a lo largo del tiempo para cada individuo (la varianza de μ_{nt}). A menudo, este parámetro recibe el nombre de *varianza entre-sujetos relativa a la varianza intra-sujetos* (*cross-subject variance relative to the within-subject variance*).

Bajo esta estructura aplicada a los errores, las probabilidades de elección se pueden simular fácilmente utilizando los conceptos de partición conveniente del error, tratados en la sección 1.2. Condicionada a η_n, la probabilidad de no tomar la acción en el período t es $\mathrm{Prob}(V_{nt} + \eta_n + \mu_{nt} < 0) = \mathrm{Prob}(\mu_{nt} <$

$-(V_{nt}+\eta_n)\big) = \Phi(-(V_{nt}+\eta_n))$, donde $\Phi(\cdot)$ es la distribución normal acumulativa. La mayoría de los paquetes de software estadístico comerciales incluyen rutinas para calcular esta función. La probabilidad de tomar la acción condicionada a η_n, es $1-\Phi\big(-(V_{nt}+\eta_n)\big) = \Phi(V_{nt}+\eta_n)$. La probabilidad de la secuencia de elecciones d_n, condicionada a η_n, es por lo tanto $\prod_t \Phi((V_{nt}+\eta_n)\mathrm{d}_{nt})$, que podemos etiquetar como $\mathrm{H}_{nd_n}(\eta_n)$.

Hasta ahora hemos condicionado a η_n, cuando en realidad η_n es aleatoria. La probabilidad *no condicionada* es la integral de la probabilidad condicionada $\mathrm{H}_{nd_n}(\eta_n)$ sobre todos los valores posibles de η_n:

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

donde $\phi(\eta_n)$ es la densidad normal con media cero y varianza σ. Esta probabilidad se puede simular muy fácilmente de la siguiente manera:

1. Extraiga un valor al azar de una densidad normal estándar utilizando un generador de números aleatorios. Multiplique el valor extraído por $\sqrt{\sigma}$, de manera que se convierta en una extracción de η_n, de una densidad normal con varianza σ.
2. Para esta extracción de η_n, calcule $\mathrm{H}_{nd_n}(\eta_n)$.
3. Repita los pasos 1-2 numerosas veces y promedie los resultados. Este promedio es una aproximación simulada de P_{nd_n}.

Este simulador es mucho más fácil de calcular que los simuladores probit generales que se describen en la siguiente sección. La posibilidad de utilizarlo surge de la estructura que impusimos al modelo, es decir, de imponer que la dependencia temporal de los factores no observados podía ser capturada en su totalidad por un componente aleatorio η_n que se mantiene constante en el tiempo para cada persona. Gourieroux y Monfort (1993) proporcionan un ejemplo de la utilización de este simulador con un modelo probit de este tipo.

La utilidad representativa en un período de tiempo puede incluir variables exógenas para otros períodos de tiempo, tal y como ya hemos comentado respecto a los modelos logit sobre datos de panel (sección 3.3.3). Es decir, V_{nt} puede incluir variables exógenas que se refieren a períodos distintos de t. Por ejemplo, una respuesta diferida a cambios de precios se puede representar mediante la inclusión de los precios en períodos anteriores en la V del período actual. Una conducta anticipatoria (por la cual, por ejemplo, una persona compra un producto ahora porque anticipa correctamente que el precio se incrementará en el futuro) puede ser representada incluyendo los precios previstos en períodos futuros en la V del período actual.

Introducir una variable dependiente diferida es posible, pero introduce dos dificultades que el investigador debe abordar. En primer lugar, dado que los errores están correlacionados en el tiempo, la elección en un período está correlacionada con los errores en períodos posteriores. Como resultado, la inclusión de una variable dependiente diferida sin ajustar convenientemente el procedimiento de estimación da como resultado estimaciones inconsistentes. Este problema es análogo al del análisis de regresión, donde el estimador de mínimos cuadrados ordinarios es inconsistente cuando se incluye una variable dependiente diferida y los errores están correlacionados en serie. Para estimar un probit consistentemente en esta situación, el investigador debe determinar la distribución de cada ε_{nt} condicionada al valor de las variables dependientes diferidas. La probabilidad de elección se basa en esta distribución condicionada en lugar de basarse en la distribución no condicionada $\phi(\cdot)$ que se utilizó anteriormente. En segundo lugar, el investigador a menudo no puede observar las elecciones de los decisores desde la primera elección que estos tuvieron a su disposición. Por ejemplo, un investigador que estudia los patrones de empleo tal vez

observe la situación laboral de una persona durante un período de tiempo (por ejemplo, 1998-2001), pero por lo general no va a observar la situación laboral de la persona desde la primera vez que esa persona podría haber tenido un trabajo (algo que podría ser muy anterior a 1998). En este caso, la probabilidad para el primer período que el investigador observa depende de las decisiones de la persona en los períodos anteriores que el investigador no observa. El investigador debe determinar una forma de representar la primera probabilidad de elección que permita una estimación consistente teniendo en cuenta la información perdida de las elecciones anteriores. Esto se conoce como el *problema de las condiciones iniciales* (*initial conditions problem*) de los modelos de elección dinámicos. Ambas cuestiones, así como los posibles enfoques para tratar con ellas, han sido abordadas por Heckman (1981b, 1981a) y Heckman y Singer (1986). Debido a su complejidad, no describo los procedimientos aquí y remito al lector interesado - y valiente - a que lea estos artículos.

Papatla y Krishnamurthi (1992) evitan estos problemas en su modelo probit con variables dependientes diferidas, al asumir que los factores no observados son independientes en el tiempo. Como ya comentamos en relación con el modelo logit para datos de panel (Sección 3.3.3), las variables dependientes diferidas no se correlacionan con los errores actuales cuando los errores son independientes en el tiempo y por lo tanto se pueden introducir sin inducir inconsistencia. Por supuesto, este procedimiento sólo es apropiada si el supuesto de errores independientes en el tiempo es realmente cierto, en lugar de ser simplemente un supuesto.

5.6 Simulación de las probabilidades de elección

Las probabilidades probit no tienen una expresión cerrada por lo que deben aproximarse numéricamente. Para ello, se han empleado varios procedimientos sin simulación que pueden ser efectivos en ciertas circunstancias.

Los métodos de cuadratura numérica aproximan la integral mediante una función ponderada de puntos de evaluación especialmente elegidos. Una buena explicación de estos procedimientos la proporciona Geweke (1996). Ejemplos de su uso para probit los podemos encontrar en Butler y Moffitt (1982) y Guilkey y Murphy (1993). La cuadratura numérica opera de forma efectiva cuando la dimensión de la integral es pequeña, pero no sucede así con dimensiones altas. Se puede utilizar para modelos probit si el número de alternativas (o, con datos de panel, el número de alternativas por el número de períodos de tiempo) no es mayor a cuatro o cinco. También se puede utilizar si el investigador ha especificado una estructura de componentes de error con no más de cuatro o cinco términos. Sin embargo, no es eficaz para modelos probit generales. E incluso para integración con pocas dimensiones, la simulación es a menudo más fácil

Otro procedimiento sin simulación que ha sido sugerido es el algoritmo de Clark, introducido por Daganzo et al. (1977). Este algoritmo utiliza el hecho, mostrado por Clark (1961), de que el máximo de varias variables distribuidas normalmente es en sí mismo aproximadamente una distribución normal. Desafortunadamente, la aproximación puede ser muy imprecisa en algunas circunstancias (como se muestra por Horowitz et al., 1982) y el grado de precisión es difícil de evaluar para un contexto determinado.

Por último, la simulación ha demostrado ser muy general y útil para aproximar probabilidades probit. Se han propuesto numerosos simuladores para los modelos probit, un resumen lo proporciona Hajivassiliou et al. (1996). En la sección anterior he descrito un simulador que es apropiado para un modelo probit que tiene una estructura particularmente conveniente: un probit binario sobre datos de panel donde la dependencia del tiempo es capturada por un factor aleatorio. En esta sección, describo tres simuladores que son aplicables para probits de todo tipo: simulador aceptación-rechazo, simulador aceptación-rechazo suavizado y GHK. El simulador GHK es, con mucha diferencia, el simulador probit más utilizado, por razones que se tratan a continuación. Los otros dos métodos son valiosos

pedagógicamente. También tienen relevancia más allá de los modelos probit y pueden aplicarse en prácticamente cualquier situación. Pueden ser muy útiles cuando el investigador está desarrollando sus propios modelos en lugar de emplear modelos probit o cualquier otro modelo descrito en este libro.

5.6.1 Simulador por aceptación-rechazo

El método de aceptación-rechazo (*accept-reject,* AR) es el simulador más directo. Considere la simulación de P_{ni}. Se extraen valores al azar de los términos aleatorios a partir de sus distribuciones. Para cada valor extraído, el investigador determina si esos valores de los errores, al ser combinados con las variables observadas que afronta la persona n, darían lugar a que la alternativa i fuese la elegida. Si es así, el valor extraído se califica como una *aceptación*. Si el valor extraído resultaría en la elección de otra alternativa, el valor es un *rechazo*. La probabilidad simulada es la proporción de valores extraídos que son aceptados. Este procedimiento se puede aplicar a cualquier modelo de elección con cualquier distribución de los términos aleatorios. Se propuso originalmente para probits (Manski y Lerman, 1981), y por ello facilitamos los detalles del enfoque adoptado por este método en términos del modelo probit. Su uso para otros modelos es obvio.

Utilizamos la expresión (5.1) para las probabilidades probit:

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

donde $\mathrm{I}(\cdot)$ es una función indicadora de si la expresión entre paréntesis es verdadera y $\phi(\varepsilon_n)$ es la densidad normal conjunta con media cero y covarianza Ω. El simulador AR de esta integral se calcula como sigue:

1. Haga una extracción de valores al azar para el vector J-dimensional de errores ε_n, a partir de una densidad normal con media cero y covarianza Ω. Etiquete el vector de valores extraído como ε_n^r con $\mathrm{r} = 1$ y los elementos de la extracción como $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.
2. Utilizando estos valores para los errores, calcule la utilidad que cada alternativa obtiene con estos errores. Es decir, calcule $\mathrm{U}_{nj}^r = \mathrm{V}_{nj} + \varepsilon_{nj}^r\ \forall j$.
3. Determine si la utilidad de la alternativa i es mayor a la de todas las otras alternativas. Es decir, calcule $\mathrm{I}^r = 1$ si $\mathrm{U}_{ni}^r > \mathrm{U}_{nj}^r \forall j \neq i$, lo que indicaría una aceptación, y $\mathrm{I}^r = 0$ en cualquier otro caso, indicando un rechazo.
4. Repita los pasos 1-3 múltiples veces. Etiquete el número de repeticiones (incluyendo la primera) como R, de modo que r toma valores de 1 a R.
5. La probabilidad simulada es la proporción de extracciones que son aceptaciones: $\breve{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r$.

La integral $\int \mathrm{I}(\cdot)\phi(\varepsilon_n)d\varepsilon$ se aproxima por el promedio $\frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r(\cdot)$ para extracciones de valores de $\phi(\cdot)$. Obviamente, $\breve{\mathrm{P}}_{ni}$ es un estimador no sesgado de P_{ni}: $\mathrm{E}\left(\breve{\mathrm{P}}_{ni}\right) = \frac{1}{R}\sum E[\mathrm{I}^r(\cdot)] = \frac{1}{R}\sum \mathrm{P}_{ni} = \mathrm{P}_{ni}$, donde la expectativa es entre diferentes conjuntos de R extracciones. La varianza de $\breve{\mathrm{P}}_{ni}$ entre diferentes conjuntos de extracciones disminuye a medida que el número de extracciones se eleva. El simulador es a menudo llamado el "simulador de frecuencia cruda" ("*crude frequency simulator"),* ya que es la frecuencia de veces que la extracción de valores de los errores produce que la alternativa especificada sea la elegida. La palabra "crudo" distingue este simulador del simulador de frecuencia suavizado que describimos en la siguiente sección.

El primer paso del simulador AR para un modelo probit es extraer un valor al azar de una densidad normal conjunta. La siguiente pregunta surge de inmediato: ¿cómo se obtienen estas extracciones? El

procedimiento más sencillo es el descrito en la Sección 9.2.5, que utiliza el factor Choleski. La matriz de covarianza de los errores es Ω. Un factor Choleski de Ω es una matriz triangular inferior L tal que $\mathrm{LL}' = \Omega$. Este factor es llamado en ocasiones la raíz cuadrada generalizada de Ω. La mayoría de los paquetes de software estadístico contienen rutinas para calcular el factor Choleski de cualquier matriz simétrica. Supongamos ahora que η es un vector de J normales estándar iid, tal que $\eta \sim \mathrm{N}(0, \mathrm{I})$, donde I es la matriz identidad. Este vector se puede obtener extrayendo J valores de un generador de números aleatorios de una distribución normal estándar, agrupándolos luego en un vector. Podemos construir un vector ε que se distribuya $\mathrm{N}(0, \Omega)$ usando el factor Choleski para transformar η. Concretamente, calculamos $\varepsilon = \mathrm{L}\eta$. Dado que la suma de normales es normal, ε se distribuye normalmente. Y como η tiene media cero, también ε tiene media cero. La covarianza de ε es $Cov(\varepsilon) = \mathrm{E}(\,\varepsilon\,\varepsilon') = \mathrm{E}(\mathrm{L}\eta(\mathrm{L}\eta)') = \mathrm{E}(\mathrm{L}\eta\eta'\mathrm{L}') = \mathrm{LE}(\eta\eta')\mathrm{L}' = \mathrm{LIL}' = \mathrm{LL}' = \Omega$.

Utilizando el factor Choleski L de Ω, el primer paso del simulador AR se descompone en dos sub-etapas:

1A. Extraiga J valores al azar de una densidad normal estándar, utilizando un generador de números aleatorios. Agrupe estos valores en un vector y etiquete el vector como η^r.

1B. Calcule $\varepsilon_n^r = L\eta^r$.

Posteriormente, utilizando ε_n^r, calcule la utilidad de cada alternativa y verifique si la alternativa i es la que tiene mayor utilidad.

El procedimiento que hemos descrito funciona empleando utilidades y la expresión (5.1), que es una integral J-dimensional. El procedimiento se puede aplicar de forma análoga a las diferencias de utilidad, lo que reduce la dimensión de la integral a $\mathrm{J} - 1$. Como se ha indicado en (5.3), las probabilidades de elección se pueden expresar en términos de diferencias de utilidad:

$$P_{ni} = \int I\big(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \;\forall j \neq i\big)\boldsymbol{\phi}(\tilde{\varepsilon}_{ni})d\tilde{\varepsilon}_{ni},$$

donde $\phi(\tilde{\varepsilon}_{ni})$ es la densidad normal conjunta con media cero y covarianza $\widetilde{\Omega}_i = M_i \Omega M_i'$. Esta integral puede ser simulada con métodos AR a través de los siguientes pasos:

1. Extraiga un valor $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$ de la siguiente manera:
    a. Extraiga $\mathrm{J} - 1$ valores de una densidad normal estándar utilizando un generador de números aleatorios. Agrupe estos valores en un vector y etiquete el vector como η^r.
    b. Calcule $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$, donde L_i es el factor Choleski de $\widetilde{\Omega}_i$.
2. Utilizando estos valores de los errores, calcule la diferencia de utilidad para cada alternativa, respecto a la utilidad de la alternativa i. Es decir, calcule $\widetilde{\mathrm{U}}_{nji}^r = V_{nj} - V_{ni} + \tilde{\varepsilon}_{nji}^r \;\forall j \neq i$.
3. Determine si cada diferencia de utilidad es negativa. Es decir, calcule $I^r = 1$ si $\widetilde{\mathrm{U}}_{nji}^r < 0 \;\forall j \neq i$, lo que indicaría una aceptación, y $I^r = 0$ en caso contrario, lo que indicaría un rechazo.
4. Repita los pasos 1 a 3 R veces.
5. La probabilidad simulada es el número de aceptaciones dividido por el número de repeticiones: $\widetilde{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r$.

Usar diferencias de utilidad es ligeramente más rápido computacionalmente que usar las utilidades en sí mismas, ya que reducimos una dimensión. Sin embargo, a menudo es más fácil conceptualmente trabajar con utilidades.

Como acabamos de indicar, el simulador AR es muy general. Puede ser aplicado a cualquier modelo para el cual podamos extraer valores al azar de sus términos aleatorios y para el que sea posible determinar el comportamiento que el decisor adoptaría con esos valores. Asimismo, es un simulador muy intuitivo,

lo que representa una ventaja desde el punto de vista de programación, ya que la depuración se convierte en una tarea relativamente fácil. Sin embargo, el simulador AR tiene varias desventajas, especialmente cuando se utiliza en el contexto de la estimación de máxima verosimilitud.

Recordemos que la función log-verosimilitud es $LL = \sum_n \sum_j d_{nj} log P_{nj}$, donde $d_{nj} = 1$ si n eligió j y 0 en caso contrario. Cuando las probabilidades no se pueden calcular con exactitud, como en el caso del modelo probit, se utiliza la función log-verosimilitud simulada en lugar de la log-verosimilitud real, reemplazando las verdaderas probabilidades por las probabilidades simuladas: $SLL = \sum_n \sum_j d_{nj} log \check{P}_{nj}$. El valor de los parámetros que maximiza la SLL recibe el nombre de estimador de máxima verosimilitud simulada (*máximum simulated likelihood estimator*, MSLE). Es, con diferencia, el procedimiento de estimación basada en simulación más utilizado. Sus propiedades se describen en el capítulo 8. Desafortunadamente, usar el simulador AR en la SLL puede ser problemático.

Los problemas son dos. En primer lugar, $\check{P}_{ni}$ puede ser cero para cualquier número finito de extracciones R. Es decir, es posible que cada uno de los R valores extraídos de los términos de error dé como resultado un rechazo, de manera que la probabilidad simulada resulte cero. Los valores cero para $\check{P}_{ni}$ son problemáticos debido a que calculamos el logaritmo de $\check{P}_{ni}$ cuando entra en la función de verosimilitud y el logaritmo de cero es indeterminado. SLL no puede calcularse si la probabilidad simulada es cero para algún decisor de la muestra.

La ocurrencia de una probabilidad simulada igual a cero es particularmente probable cuando la verdadera probabilidad es baja. A menudo, al menos un decisor de la muestra habrá hecho una elección que tenga baja probabilidad. Por ejemplo, cuando los decisores se enfrentan a numerosas alternativas (como miles de marcas y modelos en la elección de un automóvil), cada alternativa tiene baja probabilidad. Con elecciones repetidas, la probabilidad de cualquier secuencia concreta de elecciones puede ser extremadamente pequeña; por ejemplo, si la probabilidad de elegir una alternativa concreta es 0.25 en cada uno de los 10 períodos de tiempo en los que se repite una elección, la probabilidad de la secuencia consistente en repetir 10 veces la misma elección es $(0.25)^{10}$, que es menor a 0.000001.

Además de este problema, la SLL se debe calcular en cada paso del proceso de búsqueda de su máximo. Algunos de los valores de los parámetros para los que necesitaremos calcular la SLL durante el proceso de maximización pueden estar muy lejos de los verdaderos valores. Durante el proceso, pueden aparecer probabilidades bajas incluso aunque éstas no existan en los valores que maximizan la SLL.

Siempre podemos obtener probabilidades simuladas no nulas al realizar suficientes extracciones. Sin embargo, si el investigador continúa extrayendo valores hasta obtener al menos una aceptación para cada decisor, el número de extracciones se convierte en una función de las probabilidades. El proceso de simulación deja de ser independiente del proceso de elección que se está modelando y las propiedades del estimador pasan a ser más complejas.

Existe una segunda dificultad en el uso del simulador AR para obtener el MSLE. Las probabilidades simuladas no son funciones suaves en relación a los parámetros, es decir, no son dos veces diferenciables. Como se explica en el capítulo 8, los procedimientos numéricos que se utilizan para localizar el máximo de la función log-verosimilitud se basan en las primeras derivadas y, en ocasiones, en las segundas derivadas, de las probabilidades de elección. Si no existen estas derivadas, o no se dirigen hacia el máximo, el procedimiento numérico no será efectivo.

La probabilidad AR simulada es una función escalonada, tal como se representa en la figura 5.1. $\check{P}_{ni}$ es la proporción de valores extraídos para los que la alternativa i tiene la utilidad mayor. Un cambio infinitesimalmente pequeño en un parámetro por lo general no va a producir que un valor extraído pase de ser una aceptación a un rechazo, y viceversa. Si U_{ni}^r está por debajo de U_{nj}^r para algunas alternativas j en un nivel determinado de los parámetros, también lo estará para un cambio infinitesimalmente

pequeño en cualquier parámetro. Así que, por lo general, $\check{P}_{ni}$ es constante respecto a pequeños cambios en los parámetros. Sus derivadas respecto a los parámetros son cero en este rango. Si los parámetros cambian de tal manera que un rechazo se convierte en una aceptación, entonces $\check{P}_{ni}$ se incrementa en una cantidad discreta, que va desde M/R a $(M+1)/R$, donde M es el número de aceptaciones contabilizadas en los valores originales de los parámetros. $\check{P}_{ni}$ es constante (pendiente cero) hasta que una aceptación se convierta en un rechazo, o viceversa, momento en el que $\check{P}_{ni}$ salta en una cantidad $1/R$. Su pendiente en ese momento no está definida. Por lo tanto, la primera derivada de $\check{P}_{ni}$ con respecto a los parámetros o es cero o es indefinida.

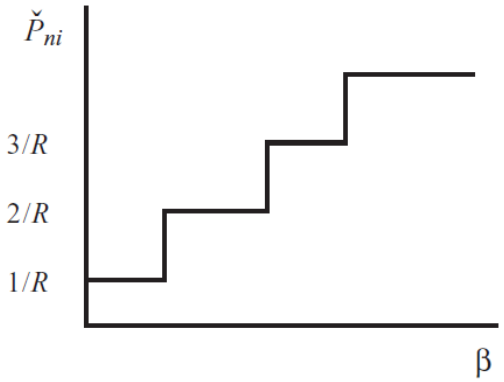

Figura 5.1. El simulador AR es una función escalonada en los parámetros.

Este hecho dificulta los procedimientos numéricos que se utilizan para localizar el máximo de la SLL. Como se trata en el Capítulo 8, los procedimientos de maximización utilizan el gradiente calculado en unos valores de prueba de los parámetros para determinar la dirección en la cual moverse para encontrar los parámetros con mayor SLL. Siendo la pendiente $\check{P}_{ni}$ para cada n cero o indefinida, el gradiente de la SLL es cero o indefinido. Este gradiente no proporciona ninguna ayuda en la búsqueda del máximo.

Este problema realmente no es tan drástico como parece. El gradiente de la SLL se puede aproximar como el cambio producido en la SLL para un cambio no infinitesimalmente pequeño en los parámetros. De esta forma, los parámetros se cambian en una cantidad que es lo suficientemente grande como para producir cambios entre aceptaciones y rechazos para, por lo menos, algunas de las observaciones. El gradiente aproximado, que puede ser llamado un gradiente de arco (*arc gradient*), se calcula como la cantidad en que ha cambiado la SLL dividida por el cambio en los parámetros. Siendo precisos: para un vector de parámetros β de longitud K, la derivada de la SLL respecto al parámetro k-ésimo se calcula como $(SLL^1 - SLL^0)/(\beta_k^1 - \beta_k^0)$, donde SLL^0 se calcula con los parámetros β originales con el elemento k-ésimo igual a β_k^0 y SLL^1 se calcula en β_k^1 con todos los demás parámetros iguales a sus valores originales. El gradiente de arco calculado de esta manera no es cero o indefinido, y proporciona información sobre la dirección de incremento. Sin embargo, la experiencia indica que aun así, la probabilidad simulada AR es difícil de usar.

5.6.2 Simuladores AR suavizados

Una forma de mitigar las dificultades del simulador AR es reemplazar el indicador AR 0-1 por una función suave y estrictamente positiva. La simulación comienza del mismo modo que con un simulador AR, extrayendo valores al azar de los términos aleatorios y calculando la utilidad de cada alternativa para cada valor extraído: U_{nj}^r. Pero en lugar de determinar si la alternativa i tiene la mayor utilidad (es decir, en lugar de calcular la función indicadora I^r), las utilidades simuladas $U_{nj}^r \forall j$ se introducen en una función. Puede utilizarse cualquier función para simular P_{ni} siempre y cuando se incremente cuando U_{ni}^r

se incremente y disminuya cuando el resto de utilidades U_{nj}^r se incrementen, sea estrictamente positiva y tenga definidas la primera y segunda derivadas respecto a $U_{nj}^r \forall j$. Una función particularmente adecuada es la función logit, como sugirió McFadden (1989). El uso de esta función da lugar al simulador AR suavizado-logit (*logit-smoothed AR simulator*).

El simulador se implementa mediante los pasos siguientes, que son los mismos del simulador AR excepto el paso 3:

1. Haga una extracción de valores al azar para el vector J-dimensional de errores ε_n, a partir de una densidad normal con media cero y covarianza Ω. Etiquete el vector de valores extraído como ε_n^r con $\mathrm{r} = 1$ y los elementos de la extracción como $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.
2. Utilizando estos valores para los errores, calcule la utilidad que cada alternativa obtiene con estos errores. Es decir, calcule $\mathrm{U}_{nj}^r = \mathrm{V}_{nj} + \varepsilon_{nj}^r \ \forall j$.
3. Introduzca estas utilidades en la fórmula logit. Es decir, calcule

$$S^r = \frac{e^{U_{ni}^r/\lambda}}{\sum_j e^{U_{nj}^r/\lambda}},$$

   donde $\lambda > 0$ es un factor de escala especificado por el investigador y que se trata en el siguiente texto.
4. Repita los pasos 1-3 muchas veces. Etiquete el número de repeticiones (incluyendo la primera) como R, de modo que r toma valores de 1 a R.
5. La probabilidad simulada es el promedio de los valores de la fórmula logit: $\breve{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{S}^r$.

Dado que $S^r > 0$ para cualquier valor finito de U_{nj}^r, la probabilidad simulada es estrictamente positiva para cualquier extracción de valores de error. Se incrementa cuando U_{ni}^r se incrementa y disminuye cuando $\mathrm{U}_{nj}^r \ j \neq i$ se incrementa. Es suave (dos veces diferenciable), dado que la propia fórmula logit es suave.

El simulador AR suavizado-logit se puede aplicar a cualquier modelo de elección, simplemente simulando las utilidades bajo la correspondiente hipótesis relativa a las distribuciones de los errores e insertando posteriormente las utilidades en la fórmula logit. Cuando este simulador se aplica al modelo probit, Ben-Akiva y Bolduc (1996) lo han denominado "probit logit-kernel" (*logit-kernel probit*).

El factor de escala λ determina el grado de suavizado. A medida que $\lambda \to 0$, S^r se acerca a la función indicadora I^r. La figura 5.2 ilustra esta circunstancia para un caso con dos alternativas. Para una extracción dada de ε_n^r, se calcula la utilidad de las dos alternativas. Considere la probabilidad simulada para la alternativa 1. Empleando AR, la función indicadora 0-1 es cero si U_{n1}^r está por debajo de U_{n2}^r y uno si U_{n1}^r excede U_{n2}^r. Empleando el suavizado-logit, la función escalonada se sustituye por una curva sigmoidea suave. El factor λ determina el grado de proximidad de la sigmoidea con la función indicadora 0-1. Bajar λ aumenta la escala de las utilidades cuando entran en la función logit (ya que las utilidades se dividen por λ). Incrementar la escala de utilidad aumenta la diferencia absoluta entre las dos utilidades. La fórmula logit da probabilidades que están más cerca de cero o uno cuando la diferencia de utilidades es mayor. Por lo tanto, la función logit suavizada S^r se vuelve más cercana a la función escalonada a medida que λ se aproxima más a cero.

El investigador necesita establecer el valor de λ. Un valor de λ menor hace del logit suave una mejor aproximación de la función indicadora. Sin embargo, este hecho es un arma de doble filo: si el logit suave se aproxima a la función indicadora demasiado bien, las dificultades numéricas del uso del simulador AR no suavizado simplemente se reproducirán en el simulador logit suavizado. Lo que quiere el investigador es establecer una λ lo suficientemente baja como para obtener una buena aproximación,

pero no tan baja como para reintroducir dificultades numéricas. Existen pocas recomendaciones a dar sobre el nivel apropiado de λ. Tal vez el mejor enfoque posible para el investigador es experimentar con diferentes λs. Para experimentar, los mismos valores extraídos de ε_n deberían ser usados con cada posible λ, de manera que aseguremos que las diferencias en los resultados son debidos al cambio en la λ y no a diferencias en los propios valores extraídos.

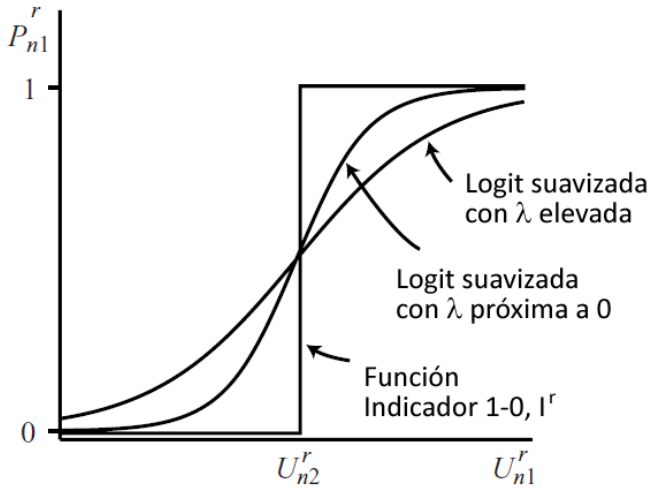

Figura 5.2. AR suavizado

McFadden (1989) describe otras funciones de suavizado. Para todas ellas, el investigador debe especificar el grado de suavizado. Una ventaja del logit suavizado es su simplicidad. Además, veremos en el capítulo 6 que el logit suavizado aplicado a un modelo probit o a cualquier otro modelo, constituye un tipo de especificación logit mixta. Es decir, en lugar de ver el logit suavizado como una aproximación que no tiene relación alguna con el modelo de comportamiento (sólo tiene un objetivo numérico), podemos verlo como el resultado de un tipo particular de estructura de error dentro del propio modelo de comportamiento. Según esta interpretación, la fórmula logit aplicada a las utilidades simuladas no es una aproximación, sino que en realidad representa el verdadero modelo.

5.6.3 Simulador GHK

El simulador probit más utilizado se denomina GHK, en referencia a los autores Geweke (1989, 1991), Hajivassiliou (tal y como se informa en Hajivassiliou y McFadden, 1998) y Keane (1990, 1994), quien desarrolló el procedimiento. En una comparación entre numerosos simuladores probit, Hajivassiliou et al. (1996) encontraron que el simulador GHK era el más preciso en las situaciones de elección que se examinaron. Geweke et al. (1994) encontraron que el simulador GHK funciona mejor que el AR suavizado. La experiencia ha confirmado su utilidad y exactitud relativa (por ejemplo, Borsch-Supan y Hajivassiliou, 1993).

El simulador GHK opera con diferencias de utilidades. La simulación de la probabilidad P_{ni} comienza restando la utilidad de la alternativa i de la utilidad de cada una del resto de las alternativas. Es importante destacar que se resta la utilidad de una alternativa diferente dependiendo de qué probabilidad se está simulando: para P_{ni}, U_{ni} es la utilidad restada de las otras utilidades, mientras que para P_{nj}, se resta U_{nj}. Este hecho es crítico para la aplicación del procedimiento.

Voy a explicar el procedimiento GHK en primer lugar para un caso de tres alternativas, ya que esta situación se puede representar gráficamente en dos dimensiones para las diferencias de utilidad. A continuación describiré el procedimiento en general, para cualquier número de alternativas. Bolduc

(1993, 1999) proporciona una excelente descripción alternativa del procedimiento, junto con los métodos para simular las derivadas analíticas de las probabilidades probit. Keane (1994) proporciona una descripción de la utilización de GHK para probabilidades de transición.

Tres alternativas

Empezamos con una especificación del modelo de comportamiento en las utilidades: $\mathrm{U}_{nj} = \mathrm{V}_{nj} + \varepsilon_{nj}, j = 1,2,3$. Suponemos el vector $\varepsilon'_n = \langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle \sim N(0, \Omega)$. Asumimos que el investigador ha normalizado el modelo para la escala y el nivel, por lo que los parámetros que entran en Ω están identificados. Asimismo, Ω puede ser una función paramétrica de los datos, así como incluir variación aleatoria de las preferencias, aunque no mostramos esta dependencia en nuestra notación.

Supongamos que queremos simular la probabilidad de la primera alternativa, P_{n1}. Podemos reformular el modelo en diferencias de utilidad substrayendo la utilidad de la alternativa 1:

$$U_{nj} - U_{n1} = \left(V_{nj} - V_{n1}\right) + \left(\varepsilon_{nj} - \varepsilon_{n1}\right),$$

$$\widetilde{U}_{nj1} = \tilde{V}_{nj1} + \tilde{\varepsilon}_{nj1},$$

para $j = 2,3$. El vector $\varepsilon'_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31} \rangle$ se distribuye $N(0, \widetilde{\Omega}_1)$,donde $\widetilde{\Omega}_1$ se obtiene de Ω.

Hacemos una transformación más para hacer el modelo más conveniente para la simulación. Para ello, sea L_1 el factor Choleski de $\widetilde{\Omega}_1$. Dado que $\widetilde{\Omega}_1$ es 2×2 en la situación de nuestro ejemplo, L_1 es una matriz triangular inferior que toma la forma

$$L_1 = \begin{pmatrix} C_{aa} & 0 \\ C_{ab} & C_{bb} \end{pmatrix}.$$

Usando este factor Choleski, las diferencias de error originales, que están correlacionadas, pueden reescribirse como funciones lineales de normales estándar *no correlacionadas*:

$$\tilde{\varepsilon}_{n21} = c_{aa}\eta_1,$$

$$\tilde{\varepsilon}_{n31} = c_{ab}\eta_1 + c_{bb}\eta_2,$$

donde η_1 y η_2 son $N(0,1)$ iid. Las diferencias de error $\tilde{\varepsilon}_{n21}$ y $\tilde{\varepsilon}_{n31}$ están correlacionadas porque ambas diferencias dependen de η_1. Expresando las diferencias de error de esta forma, las diferencias de utilidad pueden ser escritas como

$$\widetilde{U}_{n21} = \tilde{V}_{n21} + c_{aa}\eta_1,$$

$$\widetilde{U}_{n31} = \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2,$$

La probabilidad de la alternativa 1 es $\mathrm{P}_{n1} = \mathrm{Prob}\left(\widetilde{\mathrm{U}}_{n21} < 0 \; y \; \widetilde{\mathrm{U}}_{n31} < 0\right) = \mathrm{Prob}(\widetilde{\mathrm{V}}_{n21} + \tilde{\varepsilon}_{n21} < 0 \; y \; \widetilde{\mathrm{V}}_{n31} + \tilde{\varepsilon}_{n31} < 0)$. Esta probabilidad es difícil de evaluar numéricamente en términos de los $\tilde{\varepsilon}$s, porque están correlacionados. Sin embargo, utilizando la transformación basada en el factor Choleski, la probabilidad se puede escribir de forma que involucre términos aleatorios independientes. La

probabilidad se convierte en una función de la distribución normal acumulativa estándar unidimensional:

$$P_{n1} = Prob(\tilde{V}_{n21} + c_{aa}\eta_1 < 0 \; y \; \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0)$$

$$= Prob\left(\tilde{V}_{n21} + c_{aa}\eta_1 < 0\right) \times Prob\left(\tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0 \,\middle|\, \tilde{V}_{n21} + c_{aa}\eta_1 < 0\right)$$

$$= Prob\left(\eta_1 < -\tilde{V}_{n21}/c_{aa}\right) \times Prob\left(\eta_2 < -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb} \,\middle|\, \eta_1 < -\tilde{V}_{n21}/c_{aa}\right)$$

$$= \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\left(\tilde{V}_{n31} + c_{ab}\eta_1\right)}{c_{bb}}\right) \bar{\phi}(\eta_1) d\eta_1$$

donde $\Phi(\cdot)$ es la distribución normal estándar acumulativa evaluada en el punto indicado entre paréntesis y $\bar{\phi}(\cdot)$ es la densidad normal truncada[i]. El primer factor $\Phi\left(-\tilde{V}_{n21}/c_{aa}\right)$ es fácil de calcular: simplemente es la distribución normal acumulativa estándar evaluada en $-\tilde{V}_{n21}/c_{aa}$. Los paquetes informáticos de estadística contienen rutinas rápidas para la distribución normal acumulativa. El segundo factor es una integral. Como sabemos, las computadoras no pueden integrar, por lo que utilizamos la simulación para aproximar las integrales. Este es el corazón del proceso GHK: usar la simulación para aproximar la integral en P_{n1}.

Examinemos esta integral más de cerca. Es una integral sobre una normal truncada, es decir, sobre η_1 hasta $-\tilde{V}_{n21}/c_{aa}$. La simulación se realiza como sigue. Extraiga un valor al azar de η_1 de una densidad normal estándar truncada por encima de $-\tilde{V}_{n21}/c_{aa}$. Para este valor, calcule el factor $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb})$. Repita este proceso para muchas extracciones y promedie los resultados. Este promedio será una aproximación simulada de $\int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}\right)\bar{\phi}(\eta_1)d\eta_1$. La probabilidad simulada se obtiene entonces multiplicando esta media por el valor de $\Phi(-\tilde{V}_{n21}/c_{aa})$, que se calcula exactamente. ¡Bastante simple!

No obstante, se nos plantea la siguiente cuestión: ¿cómo extraemos un valor al azar de una distribución normal truncada? Describiremos cómo extraer valores al azar de distribuciones univariadas truncadas en la Sección 9.2.4. Llegados a este punto, si el lector lo desea, puede consultar esta sección antes de continuar. Pero básicamente, el proceso consiste en extraer un valor al azar de una distribución uniforme estándar y etiquetarlo μ. Posteriormente se calcula $\eta = \Phi^{-1}(\mu\Phi(-\tilde{V}_{n21}/c_{aa}))$. El η resultante es una extracción de un valor al azar de una densidad normal truncada por encima de $-\tilde{V}_{n21}/c_{aa}$.

Ahora podemos poner todo esto junto para mostrar explícitamente los pasos concretos que se utilizan para el simulador GHK en nuestro caso de tres alternativas. La probabilidad de la alternativa 1 es

$$P_{n1} = \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\left(\tilde{V}_{n31} + c_{ab}\eta_1\right)}{c_{bb}}\right) \bar{\phi}(\eta_1) d\eta_1$$

Esta probabilidad se simula de la siguiente manera:

1. Calcule $k = \Phi(-\tilde{V}_{n21}/c_{aa})$.
2. Extraiga un valor de η_1, etiquetado como η_1^r, de una distribución normal estándar truncada en $-\tilde{V}_{n21}/c_{aa}$. Esto se logra de la siguiente manera:
    a. Extraiga un valor de una distribución uniforme estándar μ^r.

[i] Para ser precisos $\bar{\phi}(\eta_1) = \phi(\eta_1)/\Phi\left(-\tilde{V}_{n21}/c_{aa}\right)$ para $-\infty < \eta_1 < -V_{n21}/c_{aa}$, y =0 en otro caso.

b. Calcule $\eta_1^r = \Phi^{-1}(\mu^r \Phi(-\tilde{V}_{n21}/c_{aa}))$.

3. Calcule $g^r = \Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$.
4. La probabilidad simulada de este valor es $\breve{P}_{n1}^r = k \times g^r$.
5. Repita los pasos 1- 4 R veces y promedie los resultados. Este promedio es la probabilidad simulada: $\breve{P}_{n1} = (1/R)\sum P_{n1}^r$.

Una representación gráfica puede resultar útil. La figura 5.3 muestra la probabilidad de la alternativa 1 en el espacio de los errores independientes η_1 y η_2. El eje x es el valor de η_1 y el eje y es el valor de η_2. La línea etiquetada como A indica la zona en la que η_1 es igual a $-\tilde{V}_{n21}/c_{aa}$. La condición de que η_1 esté por debajo de $-\tilde{V}_{n21}/c_{aa}$ se cumple en la zona de rayas a la izquierda de la línea A. La línea etiquetada como B indica donde $\eta_2 = -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}$. Tenga en cuenta que la intersección en el eje y se produce donde $\eta_1 = 0$, de modo que $\eta_2 = -\tilde{V}_{n31}/c_{bb}$ en este punto. La pendiente de la línea es $-c_{ab}/c_{bb}$. La condición de que $\eta_2 < -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}$ se satisface debajo de la línea B. El área sombreada es donde η_1 está a la izquierda de la línea A y η_2 está por debajo de la línea B. Por tanto, la probabilidad de que se escoja la alternativa 1 es la masa de densidad de probabilidad en el área sombreada.

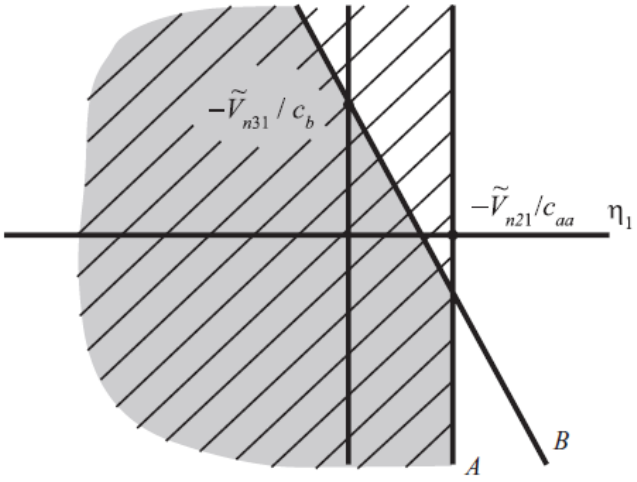

Figura 5.3. Probabilidad de la alternativa 1.

La probabilidad (es decir, la masa sombreada) es el producto de la masa de densidad de la zona rayada por la proporción de esta masa rayada que está por debajo de la línea B. El área rayada tiene masa $\Phi(-\tilde{V}_{n21}/c_{aa})$. Esto es fácil de calcular. Para cualquier valor dado de η_1, la porción de la masa rayada que está por debajo de la línea B también es fácil de calcular. Por ejemplo, en la figura 5.4, cuando η_1 toma el valor η_1^r, la probabilidad de que η_2 esté por debajo de la línea B es la proporción de la masa de la línea C que está por debajo de la línea B. Esta proporción es simplemente $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$. Por tanto, la proporción de la masa rayada que está por debajo de la línea B es el promedio de $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ sobre todos los valores de η_1 que están a la izquierda de la línea A. Este promedio se simula mediante la extracción de valores de η_1 a la izquierda de la línea A, calculando $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ para cada valor extraído y promediando los resultados. La probabilidad es el resultado de este promedio por la masa de la zona rayada, $\Phi(-\tilde{V}_{n21}/c_{aa})$.

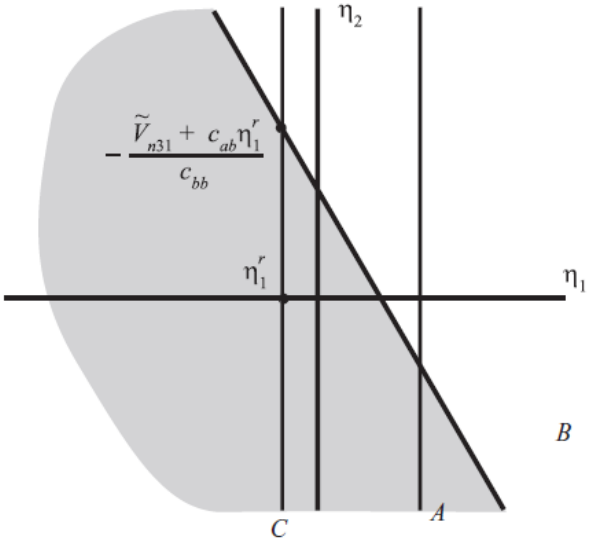

Figura 5.4. Probabilidad de que η_2 esté en el rango correcto, dado η_1^r.

Modelo general

Ahora podemos describir el simulador GHK en términos generales de forma rápida, ya que la lógica básica detrás del modelo ya ha sido expuesta. Esta expresión sucinta sirve para reforzar la idea de que el simulador GHK es realmente más simple de lo que puede parecer a primera vista

La utilidad se expresa como

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \qquad j = 1, \dots, J,$$

$$\varepsilon_n' = \langle \varepsilon_{n1}, \dots, \varepsilon_{nJ} \rangle, \qquad \varepsilon_n : J \times 1,$$

$$\varepsilon_n \sim N(0, \Omega).$$

Transformamos a diferencias de utilidad respecto a la alternativa i:

$$\tilde{U}_{nji} = \tilde{V}_{nji} + \tilde{\varepsilon}_{nji}, \qquad j \neq i,$$

$$\varepsilon_{ni}' = \langle \tilde{\varepsilon}_{n1}, \dots, \tilde{\varepsilon}_{nJ} \rangle, \qquad donde \dots es\ sobre\ toda\ alternativa\ excepto\ i,$$

$$\tilde{\varepsilon}_{ni} : (J-1) \times 1,$$

$$\tilde{\varepsilon}_{ni} \sim N\left(0, \tilde{\Omega}_i\right),$$

donde $\tilde{\Omega}_i$ se obtiene de Ω.

Re-expresamos los errores como una transformación Choleski de normales estándar iid:

$$L_i \ s.t. \ L_i L_i' = \tilde{\Omega}_i,$$

$$L_i = \begin{pmatrix} c_{11} & 0 & \dots & \dots & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

A continuación, agrupando utilidades $\widetilde{\mathrm{U}}'_{ni} = \left(\widetilde{U}_{n1i}, \dots, \widetilde{U}_{nJi}\right)$, obtenemos la forma vectorial del modelo,

$$\widetilde{U}_{ni} = \tilde{V}_{ni} + L_i \eta_n,$$

donde $\eta'_n = \langle \eta_{1n}, \dots, \eta_{J-1,n} \rangle$, es un vector de normales estándar iid: $\eta_{nj} \sim N(0,1)\ \forall j$. Escrito de forma explícita, el modelo es

$$\widetilde{U}_{n1i} = \tilde{V}_{n1i} + c_{11}\eta_1,$$

$$\widetilde{U}_{n2i} = \tilde{V}_{n2i} + c_{21}\eta_1 + c_{22}\eta_2,$$

$$\widetilde{U}_{n3i} = \tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3,$$

y así sucesivamente. Las probabilidades de elección son

$$\mathrm{P_{ni}} = Prob(\widetilde{\mathrm{U}}_{\mathrm{nji}} < 0\ \forall j \neq i)$$

$$= Prob\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times Prob\left(\eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}} \middle| \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times Prob\left(\eta_3 < \frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2\right)}{c_{33}} \middle| \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\ y\ \eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\right)$$

$$\times \cdots.$$

El simulador GHK se calcula como sigue:

1. Calculamos

$$Prob\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right) = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right).$$

2. Extraemos un valor al azar de η_1 , etiquetado como η_1^r, de una distribución normal estándar truncada en $-\widetilde{\mathrm{V}}_{\mathrm{n1i}}/\mathrm{c}_{11}$. Este valor se obtiene de la siguiente manera:
    a. Extraemos un valor al azar de una distribución uniforme estándar μ_1^r.
    b. Calculamos $\eta_1^{\mathrm{r}} = \Phi^{-1}(\mu_1^{\mathrm{r}}\Phi(-\widetilde{\mathrm{V}}_{\mathrm{n1i}}/\mathrm{c}_{11}))$.
3. Calculamos

$$Prob\left(\eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\middle|\, \eta_1 = \eta_1^r\right) = \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right)$$

4. Extraemos un valor al azar de η_2, etiquetado como η_2^r , de una distribución normal estándar truncada en $-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)/c_{22}$. Este valor se obtiene de la siguiente manera:
    a. Extraemos un valor al azar de una distribución uniforme estándar μ_2^r.
    b. Calculamos $\eta_2^r = \Phi^{-1}(\mu_2^r \Phi(-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)/c_{22}))$.
5. Calculamos

$$Prob\left(\eta_3 < \frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2\right)}{c_{33}}\middle|\, \eta_1 = \eta_1^r, \eta_2 = \eta_2^r\right)$$
$$= \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right)$$

6. Y así sucesivamente para todas las alternativas exceptuando i.
7. La probabilidad simulada para esta r-ésima extracción de valores de η_1, η_2,… se calcula como

$$\check{P}_{ni}^r = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right)$$

$$\times \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right)$$

$$\times \cdots.$$

8. Repetimos los pasos 1-7 muchas veces, para $r = 1, \ldots, R$.
9. La probabilidad simulada es

$$\check{P}_{in} = \frac{1}{R}\sum_r \check{P}_{in}^r$$

Simulador GHK con estimación de máxima verosimilitud

Hay varias cuestiones que deben abordarse al utilizar el simulador GHK en una estimación de máxima verosimilitud. En primer lugar, en la función log-verosimilitud utilizamos la probabilidad de la alternativa elegida por el decisor. Dado que diferentes decisores eligen diferentes alternativas, P_{ni} debe calcularse para diferentes is. El simulador GHK usa diferencias de utilidad respecto a la alternativa para la que se calcula la probabilidad y, por lo tanto, es necesario considerar diferentes diferencias de utilidad para los decisores que eligieron distintas alternativas. En segundo lugar, para una persona que eligió la alternativa i, el simulador GHK utiliza la matriz de covarianza $\widetilde{\Omega}_i$, mientras que para una persona que

eligió la alternativa j, se utiliza la matriz $\widetilde{\Omega}_j$. Ambas matrices se obtienen de la misma matriz de covarianza Ω de los errores originales. Debemos asegurar que los parámetros en $\widetilde{\Omega}_i$ son consistentes con los de $\widetilde{\Omega}_j$, en el sentido de que ambas matrices se obtienen de una Ω común. En tercer lugar, tenemos que asegurar que los parámetros que se estiman por máxima verosimilitud implican el uso de matrices de covarianza $\widetilde{\Omega}_j$ $\forall j$ que son definidas positivas, como debe ser cualquier matriz de covarianza. En cuarto lugar, como siempre, debemos asegurarnos de que el modelo está normalizado para la escala y el nivel de utilidad, por lo que los parámetros son identificados.

Los investigadores utilizan diversos procedimientos para abordar estas cuestiones. Voy a describir el procedimiento que yo uso.

Para asegurar que el modelo es identificado, parto de la matriz de covarianza de las diferencias de utilidad escaladas, con las diferencias calculadas respecto a la primera alternativa. Esta es la matriz $\widetilde{\Omega}_1$, que es $(J-1)\times(J-1)$. Para asegurar que la matriz de covarianza es definida positiva, parametrizo el modelo en términos del factor Choleski de $\widetilde{\Omega}_1$. Es decir, empiezo con una matriz triangular inferior que es $(J-1)\times(J-1)$ y cuyo elemento superior izquierdo es 1:

$$L_1 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Los elementos c_{kl} de este factor Choleski son los parámetros que se estiman en el modelo. Cualquier matriz que sea resultado del producto de una matriz de rango completo triangular inferior por sí misma es definida positiva. De esta forma, usando los elementos de L_1 como parámetros, puedo estar seguro de que $\widetilde{\Omega}_1$ es definida positiva para cualquier valor estimado de estos parámetros.

La matriz Ω para los J errores no diferenciados se crea a partir de L_1. Yo creo un factor Cholesky $J\times J$ para Ω mediante la adición de una fila de ceros en la parte superior de L_1 y una columna de ceros a la izquierda. La matriz resultante es

$$L = \begin{pmatrix} 0 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ 0 & c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

A continuación Ω se calcula como LL'. Teniendo esta Ω, puedo obtener $\widetilde{\Omega}_j$ para cualquier j. Observe que la matriz Ω construida de esta manera es totalmente general (es decir , permite cualquier patrón de sustitución), ya que utiliza todos los parámetros de la matriz $\widetilde{\Omega}_1$ normalizada.

La utilidad se expresa en forma de vector agrupado por alternativas: $U_n = V_n + \varepsilon_n$, $\varepsilon_n \sim N(0,\Omega)$. Considere una persona que ha elegido la alternativa i. Para la función log-verosimilitud, queremos calcular P_{ni}. Recordemos la matriz M_i que introdujimos en la sección 5.1. Las diferencias de utilidad se calculan usando esta matriz: $\widetilde{U}_{ni} = M_i U_n$, $\tilde{V}_{ni} = M_i V_n$ y $\tilde{\varepsilon}_{ni} = M_i \varepsilon_n$. La covarianza de las diferencias de error $\tilde{\varepsilon}_{ni}$ se calcula como $\widetilde{\Omega}_i = M_i \Omega M_i'$. Se toma el factor Choleski de $\widetilde{\Omega}_i$ y se etiqueta como $\mathrm{L_i}$. (Observe que la matriz L_1 obtenida aquí será necesariamente la misma L_1 que se utilizó al principio para parametrizar el modelo). La utilidad de la persona se expresa como: $\widetilde{U}_{ni} = \tilde{V}_{ni} + \mathrm{L_i}\eta_n$, donde η_n es un $(J-1)$-vector de normales estándar iid. El simulador GHK se aplica a esta expresión.

Este procedimiento satisface todos nuestros requerimientos. El modelo está necesariamente normalizado para la escala y el nivel, ya que lo parametrizamos en términos del factor Choleski L_1 de la covarianza de las *diferencias* de error *escaladas*, $\tilde{\Omega}_1$. Cada $\tilde{\Omega}_i$ es consistente con cada $\tilde{\Omega}_j$ para $j \neq i$, porque ambos se obtienen de la misma Ω (que está construida a partir de L_1) . Cada $\tilde{\Omega}_i$ es definida positiva para cualquier valor de los parámetros, ya que los parámetros son los elementos de L_1. Como se dijo anteriormente, cualquier matriz que sea el resultado del producto de una matriz triangular inferior multiplicada por sí misma es definida positiva, por lo que $\tilde{\Omega}_1 = LL'$ es definida positiva. Y cada una de las otras matrices $\tilde{\Omega}_j$s , para $j = 2, \ldots, J$, también es definida positiva, ya que se construyen para ser consistentes con Ω_1, que es definida positiva.

GHK como muestreo por importancia

Como describí en el caso de tres alternativas, el simulador GHK proporciona una aproximación simulada de la integral

$$\int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\left(\tilde{V}_{n31} + c_{ab}\eta_1\right)}{c_{bb}}\right)\phi(\eta_1)d\eta_1$$

El simulador GHK puede interpretarse de una forma alternativa que a menudo es útil. El muestreo por importancia (*importance sampling*) es una manera de transformar una integral para que sea más conveniente para la simulación. El procedimiento se describe en la Sección 9.2.7, por lo que el lector puede encontrar útil avanzar hasta ese apartado para leer la descripción. El muestreo por importancia puede resumirse de la siguiente manera. Considere cualquier integral $\bar{t} = \int t(\varepsilon)g(\varepsilon)d\varepsilon$ sobre una densidad g. Supongamos que existe otra densidad de la que es fácil extraer valores al azar. Etiquetamos esta otra densidad $f(\varepsilon)$. La densidad g se denomina densidad objetivo y f densidad generadora. La integral puede rescribirse como $\bar{t} = \int [t(\varepsilon)g(\varepsilon)/f(\varepsilon)]f(\varepsilon)d\varepsilon$. Esta integral se simula extrayendo valores de f, calculando $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ para cada valor y promediando los resultados. Este procedimiento se denomina muestreo por importancia porque cada extracción de f se pondera por g/f cuando se calcula el promedio de t; el peso g/f es la "importancia" del valor extraído de f. Este procedimiento es ventajoso si (1) resulta más fácil extraer valores al azar de f que de g , y/o (2) el simulador basado en $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ tiene mejores propiedades (por ejemplo , suavidad) que el simulador basado en $t(\varepsilon)$.

El simulador GHK puede considerarse que hace este tipo de transformación y, por lo tanto, puede ser visto como un tipo de muestreo por importancia. Sea ser un vector η de $J-1$ normales estándar iid. La probabilidad de elección se puede expresar como

$$P_{ni} = \int I(\eta \in B)g(\eta)d\eta, \qquad (5.7)$$

donde $B = \{\eta \text{ s.t. } \tilde{U}_{nji} < 0 \ \forall j \neq i\}$ es el conjunto de ηs que producen que la alternativa i sea la elegida; $g(\eta) = \phi(\eta_1)\cdots\phi(\eta_{J-1})$ es la densidad , donde ϕ es la densidad normal estándar; y las utilidades son

$$\tilde{U}_{n1i} = \tilde{V}_{n1i} + c_{11}\eta_1,$$
$$\tilde{U}_{n2i} = \tilde{V}_{n2i} + c_{21}\eta_1 + c_{22}\eta_2,$$
$$\tilde{U}_{n3i} = \tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3,$$

y así sucesivamente.

La forma más directa para simular esta probabilidad es extraer valores de η, calcular $I(\eta \in B)$ para cada valor, y promediar los resultados. Este es el simulador AR. Este simulador tiene las desafortunadas propiedades de que puede ser cero y no es suave.

Para el simulador GHK extraemos η de una densidad diferente, no de g(η). Recordemos que para el simulador GHK extraemos el valor η_1 de una densidad normal estándar truncada en $-\tilde{V}_{n1i}/c_{11}$. La densidad de esta normal truncada es $\phi(\eta_1)/\Phi(-\tilde{V}_{n1i}/c_{11})$, es decir, la densidad normal estándar normalizada por la probabilidad total bajo el punto de truncamiento. Se obtienen extracciones de η_2, η_3 y así sucesivamente, de densidades truncadas pero con diferentes puntos de truncamiento. Cada una de estas densidades truncadas toma la forma $\phi(\eta_j)/\Phi(\cdot)$ para algún punto de truncamiento en el denominador. Por tanto, la densidad de la que extraemos valores para el simulador GHK es

$$f(\eta) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\tilde{V}_{n1i}/c_{11})} \times \frac{\phi(\eta_2)}{\Phi(-(\tilde{V}_{n2i}+c_{21}\eta_1)/c_{22})} \times \cdots & para\ \eta \in B \\ 0 & para\ \eta \notin B \end{cases} \tag{5.8}$$

Tenga en cuenta que sólo extraemos valores que sean consistentes con la persona que elige la alternativa i (dado que extraemos valores de las distribuciones correctamente truncadas). Por lo tanto, $f(\eta) = 0$ para $\eta \notin B$.

Recuerde que para una extracción de η dentro del simulador GHK, calculamos:

$$\check{P}_{in}(\eta) = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right) \times \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\right) \times \cdots. \tag{5.9}$$

Observe que esta expresión es el denominador de f(η) para η ∈ B, dada en la ecuación (5.8). Usando este hecho, podemos reescribir la densidad f(η) como

$$f(\eta) = \begin{cases} g(\eta)/\check{P}_{ni}(\eta) & para\ \eta \in B \\ 0 & para\ \eta \notin B \end{cases}$$

Con esta expresión para f(η), podemos probar que para el simulador GHK, $\check{P}_{in}(\eta)$ es un estimador no sesgado de $P_{ni}(\eta)$.

$$E(\check{P}_{in}(\eta)) = \int \check{P}_{in}(\eta) f(\eta) d\eta$$

$$= \int_{\eta \in B} \check{P}_{in}(\eta) \frac{g(\eta)}{\check{P}_{in}(\eta)} d\eta \quad por\ (5.6.3)$$

$$= \int_{\eta \in B} g(\eta) d\eta$$

$$= \int I(\eta \in B) g(\eta) d\eta$$

$$= P_{in}.$$

La interpretación del simulador GHK como un muestreo por importancia también se obtiene a partir de esta expresión:

$$P_{in} = \int I(\eta \in B) g(\eta) d\eta$$

$$= \int I(\eta \in B) g(\eta) \frac{f(\eta)}{f(\eta)} d\eta$$

$$= \int I(\eta \in B) \frac{g(\eta)}{g(\eta)/\check{P}_{in}(\eta)} f(\eta) d\eta \quad por\ (5.6.3)$$

$$= \int I(\eta \in B) \check{P}_{in}(\eta) f(\eta) d\eta$$

$$= \int \check{P}_{in}(\eta) f(\eta) d\eta$$

donde la última igualdad se debe a que $f(\eta) > 0$ sólo cuando $\eta \in B$. El procedimiento GHK extrae valores de $f(\eta)$, calcula $\check{P}_{in}(\eta)$ para cada valor extraído y promedia los resultados. Básicamente, GHK reemplaza la función indicadora $0 - 1\ I(\eta \in B)$ por una $\check{P}_{in}(\eta)$ suave, y hace el cambio correspondiente en la densidad de $g(\eta)$ a $f(\eta)$.