RANDOM UTILITY MODELS

Un enfoque más interesante de los modelos de elección discreta, y más cercano a la práctica investigadora de los economistas, requiere replantear la presentación que acabamos de ver en la primera parte.

Los consumidores buscan maximizar su propio interés, que tiene ciertas propiedades de consistencia entre decisiones. La conexión de este principio con un sistema de preferencias estable se remonta a Hicks y Samuelson. Inicialmente las preferencias se consideraban idénticas entre individuos, que respondían al patrón de un consumidor representativo. Las preferencias se representaban mediante una función de utilidad U(**x**), siendo **x** un vector de niveles de consumo de bienes y servicios. El consumidor maximiza esa utilidad bajo ciertas restricciones determinadas por los precios (**p**) y la renta (m), de forma que $\mathbf{px} \leq m$. Esta maximización permite expresar las cantidades consumidas o demandadas en función de los precios y la renta, mediante una función de demanda, $\mathbf{x} = \mathbf{d}(m, \mathbf{p})$. A nivel de mercado podemos observar por tanto la misma relación entre cantidades demandadas, precios y renta, sólo que con un error aleatorio añadido para cubrir las discrepancias con los datos: $\mathbf{x} = \mathbf{d}(m, \mathbf{p}) + \boldsymbol{\varepsilon}$. Estas discrepancias pueden deberse a errores en la medición del vector x, o a errores de los consumidores en la optimización.

En los años 60 la disponibilidad de microdatos aumentó considerablemente, y también el uso de las computadoras. Los económetras se replantearon más cuidadosamente la especificación del comportamiento del consumidor.

En un artículo seminal sobre discriminación psicológica, Thurstone (1927) introdujo la *Ley del juicio comparado*, según la cual a la alternativa i, que tiene un estímulo V_i, se percibe con un error normal como $V_i + \sigma\varepsilon_i$. La probabilidad de elección cuando se comparan dos alternativas debe ser entonces $P_{1,2} = \Phi[(V_1 - V_2)/\sigma]$, que se conoce hoy día como modelo Probit binomial. Cuando el estímulo percibido, $V_i + \varepsilon_i$ se interpreta como niveles de satisfacción, o de utilidad, el modelo se puede interpretar como de elección de naturaleza económica. Marschak (1960) conectó el trabajo de Thurstone con el ámbito de la economía. Marschak estudió las consecuencias que la maximización de utilidades con elementos aleatorios tienen para las probabilidades de elección de las alternativas. Marschak llamó a esto modelo de maximización de utilidades aleatorias (*Random Utility Maximization* model, RUM).

Luce (1959) introduce el axioma de la *Independencia respecto de las alternativas irrelevantes* (*Independence from Irrelevant Alernatives*, IIA). Este axioma simplifica mucho la recolección de datos experimental para estudios de elección, al permitir que las probabilidades de elección multinomiales se infieran de experimentos de naturaleza binomial. El axioma IIA, que volveremos a ver, establece que la razón entre las probabilidades de elección de las alternativas *i* y *j* es la misma para cualquier conjunto de alternativas que incluya a *i* y *j*. Por consiguiente, tanto si *i* y *j* están solas en el conjunto como si son sólo dos posibilidades dentro de un

conjunto mucho más amplio, la razón de sus probabilidades de elección será la misma. Luce demostró que para probabilidades estrictamente positivas, la propiedad IIA implica que la probabilidad de elegir la alternativa i perteneciente al conjunto C es $P_C(i) = w_i \Big/ \sum_{k \in C} w_k$, donde las w_i son las utilidades. Por su parte, Marschak demostró que, bajo ciertas condiciones, IIA implica RUM. La conexión del MNL con el RUM se ha analizado con detalle, y la literatura sobre el tema es extensa. ***McFadden demostró que el modelo de Luce es consistente con el modelo RUM con perturbaciones iid si y sólo si estas perturbaciones aleatorias seguían una distribución de valores extremos de tipo I***. McFadden elaboró una versión del modelo de Luce en el que las utilidades w_i eran una función lineal de una serie de atributos observables, aplicado a un problema de transporte. McFadden bautizó dicho modelo como *modelo logit condicional* (*conditional logit model*), ya que en el caso binomial se convertía en el modelo logístico utilizado en bioestadística, y en el caso multinomial se podía interpretar como la distribución condicional de la demanda dado un conjunto de alternativas. Este modelo pasaría luego a conocerse como *Logit multinomial* (MNL).

Este tipo de modelos se estima por máxima verosimilitud, mediante el empleo de programas informáticos. Las alternativas se caracterizaban por sus atributos "hedónicos", lo que puede considerarse un precedente de los desarrollos en teoría del consumo de Griliches (1961) y Lancaster (1966).

Las variantes anidadas del MNL también tienen origen en trabajos de McFadden de la década de los 70, una vez más desarrolladas en el contexto de investigaciones relacionadas con problemas de transporte (puede verse McFadden, 2001, pp. 354-55 para más detalles).

Los modelos MNL han tenido amplísima aplicación empírica. Pero la propiedad IIA es una restricción fuerte, que limita los casos en los que se puede utilizar dicho modelo (Chipman, 1960, o Debreu, 1960; y la famosa paradoja del autobús rojo y el autobús azul, que veremos más adelante). Para relajar las restricciones del modelo MNL se desarrollaron los modelos MNL anidados, los modelos de valores extremos generalizados (GEV) y el probit multinomial (MNP), *pero ninguno es capaz de representar todos los posibles comportamientos consistentes con el marco RUM*. El logit multinomial mixto (MMNL) es una familia de modelos de elección discreta consistente con el marco RUM y muy flexible (McFadden and Train, 2000). Veremos detenidamente toda esta variedad de modelos. En lo que sigue evitaremos la coletilla "multinomial".

En general, los modelos más flexibles requieren de *métodos de simulación* para el cálculo de las probabilidades. Además, se requieren métodos estadísticos paramétricos y no paramétricos, herramientas de diagnóstico para detectar los errores en la especificación de los modelos, y métodos para hacer test de hipótesis.

Estos requerimientos no son exclusivos de los modelos de elección discreta, pero son más difíciles en este caso porque los modelos que se derivan de los RUM son en general *no lineales*. Normalmente, el MNL y sus derivaciones, dependían de métodos de estimación basados en máxima verosimilitud y sus propiedades para grandes muestras. Los programas informáticos permiten trabajar con estos modelos y estos métodos sin grandes problemas. Sin embargo, el uso de estimadores no-paramétricos está cada vez más extendido, junto con métodos *bootstrap* para refinar las aproximaciones asintóticas, método de momentos generalizado y simulación. Veremos algo de estos temas cuando tratemos cada tipo de modelo.

***

Una vez analizado el desarrollo histórico de los modelos vamos a detenernos en un análisis más pormenorizado de los mismos.

Un agente tiene que tomar la decisión de elegir entre distintas alternativas que se le presentan. El conjunto de alternativas debe tener 3 propiedades fundamentales: **1)** las alternativas tienen que ser *mutuamente excluyentes*; **2)** el conjunto de alternativas tiene que ser *exhaustivo*, incluyendo todas las posibilidades; **3)** y el número de alternativas tiene que ser *finito*. La tercera característica es restrictiva, y es la definitoria de los modelos de elección discreta, que se diferencian en ese punto de los modelos de regresión.

¿Cómo se puede modelar esta elección de alternativas? La respuesta tiene el nombre de *random utility models* (RUMs). El origen de estos modelos, como hemos visto, está en el trabajo de Thurstone (1927), que plantea un Probit binario para medir si se perciben diferencias entre niveles de estímulos psicológicos; y Marschak (1960) replantea el problema en términos de niveles de utilidad, en un contexto de maximización de la utilidad, que es el origen de los RUMs.

Los RUMs se plantean de la siguiente forma. Un individuo *n* se enfrenta a *J* alternativas diferentes, cada una de las cuales le proporciona un determinado nivel de utilidad. La utilidad que el individuo *n* obtiene de la alternativa *j* es U_{nj}, donde j=1,…,J. El individuo elige la alternativa con un nivel de utilidad más alto, de forma que prefiere *i* si y sólo si $U_{ni} > U_{nj}$ $\forall j \neq i$. Es interesante que esta utilidad es conocida por el individuo *n*, pero no por el investigador. ¿Qué sabe el investigador? El investigador sólo tiene información sobre una serie de atributos de las alternativas a las que se enfrenta el individuo *n* (x_{nj}) y algunas características observables del propio individuo n. El investigador puede especificar una función de utilidad que relacione dichos atributos con la utilidad (inobservada) del individuo. La función

sería del tipo $V_{nj} = V(x_{nj}, s_n)\ \forall j$, y a esa utilidad se la conoce como *utilidad representativa*.

Dado que V_{nj} es una aproximación a partir de variables observadas, quedando otras variables inobservadas que explican también la utilidad, en general tendremos que $V_{nj} \neq U_{nj}$. Podemos descomponer la utilidad en dos partes:

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

Donde ε_{nj} recoge aquellos factores inobservados que afectan o explican la utilidad y que no están captados por la función V_{nj}. Las características de ε_{nj} dependen de cómo el investigador especifica V_{nj}.

Dado que el ε_{nj} es un término desconocido para el investigador, debe ser tratado como una variable aleatoria. La función de densidad conjunta del vector $\varepsilon_n = (\varepsilon_{n1},\ldots,\varepsilon_{nj})$ se escribe como $f(\varepsilon_n)$. ***Con esa función de densidad asociada al término de error podemos introducir el concepto de probabilidad en el problema de la elección entre alternativas discretas***. La probabilidad de que el individuo *n* elija la alternativa *i* es

$$\begin{aligned} P_{ni} &= \text{Prob}(U_{ni} > U_{nj}\ \forall\ j \neq i) \\ &= \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall\ j \neq i) \\ &= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall\ j \neq i) \end{aligned}$$

Esa probabilidad es una función de distribución (la acumulada de una función de densidad), es decir, la probabilidad de que cada término de error $\varepsilon_{nj} - \varepsilon_{ni}$ sea menor que la cantidad observada $V_{ni} - V_{nj}$. Utilizando la función de densidad $f(\varepsilon_n)$, tendríamos

$$\begin{aligned} P_{ni} &= \text{Prob}\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \ \forall j \neq i\right) \\ &= \int_{\varepsilon} I\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \ \forall j \neq i\right) f(\varepsilon_n)\, d\varepsilon_n \end{aligned}$$

donde I(.) es una función indicador, igual a 1 cuando la expresión entre paréntesis se cumple, y cero cuando no se cumple.

Un ejemplo. Una persona se enfrenta a dos alternativas, viajar al trabajo en coche (alternativa c) o en autobús (alternativa b). La porción observable de su utilidad depende de dos variables, que son el tiempo (T) y el coste monetario (M). Tenemos por tanto que

$$V_c = \alpha T_c + \beta M_c$$

$$V_b = \alpha T_b + \beta M_b$$

Los parámetros pueden ser conocidos por el investigador, o estimados. Supongamos que, para una persona concreta, $V_c = 4$ y $V_b = 3$. La diferencia a favor del coche es de una unidad. Todo dependerá pues de la diferencia en las utilidades inobservadas para esta persona. La probabilidad de que elija el autobús será la probabilidad de que se de que $\varepsilon_b - \varepsilon_c > 1$, y elegirá el coche si $\varepsilon_b - \varepsilon_c < 1$. De forma más explícita:

$$P_c = \text{Prob}(\varepsilon_b - \varepsilon_c < V_c - V_b)$$

y

$$P_b = \text{Prob}(\varepsilon_b - \varepsilon_c > V_c - V_b)$$

$$= \text{Prob}(\varepsilon_c - \varepsilon_b < V_b - V_c)$$

La clave está en cómo se distribuyen los factores inobservados. Imaginemos una serie de individuos con las mismas utilidades observadas, pero que presentan distintos comportamientos (distintas elecciones). Los factores inobservados están siendo diferentes entre ellos. ***La función de densidad $f(\varepsilon_n)$ puede verse como la distribución de la porción inobservada de la utilidad entre esas personas que comparten la misma utilidad observada***.

Distintas especificaciones de $f(\varepsilon_n)$, es decir, distintos supuestos sobre la distribución de la utilidad inobservada, nos lleva a distintos modelos de elección discreta. ***Algunas versiones de la integral anterior tienen forma cerrada, es decir, solución analítica, y otras no. En los modelos Logit, Logit anidado y Logit ordenado la integral tiene forma cerrada, mientras que en el Probit y Logit Mixto no, y se requiere un cálculo numérico por simulación para hallar esa suma***.

Tabla 3. *Diferencias esenciales entre los modelos*

Modelo	**Logit**	**GEV models**	**Probit**	**Mixed Logit**	**Mixed Probit**
***f*(εₙ)**	ε_n iid según una distribución de valores extremos $\forall i \Rightarrow \varepsilon_n$ están incorrelacionados entre alternativas (también en secuencia), y con una misma varianza	ε_n siguen una distribución de valores extremos generalizada ⇒ ε_n están correlacionados entre alternativas alternativas	$\varepsilon_n = (\varepsilon_{n1}, \varepsilon_{n2},..., \varepsilon_{nj}) \sim N(0,\Omega) \Rightarrow \varepsilon_n$ siguen una distribución conjunta normal (también secuencialmente, dado el caso), y Ω permite cualquier patrón de correlaciones y heterocedasticidad	ε_n siguen cualquier distribución, y se dividen en dos partes, una iid de valores extremos, y otra que contiene toda la correlación y heterocedasticidad con cualquier distribución	ε_n siguen cualquier distribución, y se dividen en dos partes, una iid normal, y otra que contiene toda la correlación y heterocedasticidad con cualquier distribución
Resumen	Cada elección es totalmente independiente de las demás	Se permiten diversos patrones de dependencia entre alternativas	Es un modelo mucho más flexible que los basados en valores extremos	Contiene como casos particulares a todos los demás modelos	Contiene como casos particulares a todos los demás modelos, pero puede ser más fácil de estimar que el mixed Logit en ocasiones
Simulación	No se requiere (las probabilidades tienen forma analítica cerrada)	No se requiere (las probabilidades tienen forma analítica cerrada)	Se requiere simulación (simulación completa)	Se requiere simulación (partición conveniente del error)	Se requiere simulación (las probabilidades vienen dadas por una integral)

La elección del agente viene determinada por factores observables (**x**) y no observables (ε), y la función y = h(**x**, ε) se conoce como *función de comportamiento*. Dado que ε no es observable, la elección del individuo no es determinística y no se puede predecir con exactitud. Tenemos que trabajar con probabilidades. La probabilidad de que se de un resultado concreto es la probabilidad de que los factores inobservados sean tales que la función de comportamiento determine ese resultado. Por tanto: P(y/x) = Prob(ε s.a. h(x, ε)=y). A partir de esa expresión podemos construir una función indicador I[y = h(x, ε)] que toma un valor 1 cuando lo que está entre paréntesis se cumple, y 0 en otro caso. La probabilidad es simplemente el valor esperado de este indicador, es decir,

$$P(y/x) = Prob(I[h(x, \varepsilon)=y] = 1)$$

$$= \int I[h(x, \varepsilon)=y]f(\varepsilon)d\varepsilon$$

Se lee como "la suma de los casos en los que se elige la opción multiplicado por la probabilidad de que se den los valores de ε que explican que se haya elegido dicha opción".

Esa integral debe calcularse, y se dan varios casos: integral con solución analítica (probabilidad con una fórmula con forma cerrada, obtenida a partir del cálculo analítico de esa integral), integral sin solución analítica (calculada mediante simulación), o integral que puede descomponerse en una parte con solución analítica y otra que debe estimarse simulando.

Un ejemplo de integral con forma cerrada es el Logit. Para simplificar, si la elección de una determinada alternativa depende de que ésta proporcione una utilidad positiva, de forma que el modelo de comportamiento es $U = \beta'x + \varepsilon$, la probabilidad de que la alternativa sea elegida será

$$\begin{aligned} P &= \int_{\varepsilon} I(\beta' x + \varepsilon > 0) f(\varepsilon) d\varepsilon \\ &= \int_{\varepsilon} I(\varepsilon > -\beta' x) f(\varepsilon) d\varepsilon \\ &= \int_{\varepsilon=-\beta' x}^{\infty} f(\varepsilon) d\varepsilon \end{aligned}$$

Si ε sigue una distribución logística, la función de densidad será $f(\varepsilon_n) = e^{-\varepsilon}/(1+e^{-\varepsilon})$, a la que corresponde una función de distribución $F(\varepsilon) = 1/(1+e^{-\varepsilon})$. Por tanto, tendremos que

$$P = 1 - F(-\beta'x) = 1 - 1/(1+e^{\beta'x}) = e^{\beta'x}/(1+e^{\beta'x})$$

A veces la integral no tiene forma cerrada, es decir, solución analítica o matemática directa. En principio habría que introducir fuertes restricciones sobre *h* y sobre la distribución de los términos de error *f* para que la integral tenga solución en forma cerrada, pero estas restricciones convierten al modelo en irrealista e inadecuado para muchas situaciones. La simulación permite resolver cualquier caso. La clave está en que la integración (suma) a lo largo de una función de densidad es una forma de promediar. Si tenemos una integral de la forma $t = \int t(\varepsilon) f(\varepsilon) d\varepsilon$, donde $t(\varepsilon)$ es un estadístico basado en la variable ε cuya densidad es $f(\varepsilon)$. Esa integral es el valor esperado de t para todos los valores posibles de ε. El modelo Probit calcula las probabilidades asociadas a cada alternativa mediante este procedimiento basado en simulación.

Puede ocurrir que nos encontremos ante una integral que admite una solución en parte analítica y en parte simulada. Si el término de error puede descomponerse en dos partes ε_1 y ε_2, de forma que $f(\varepsilon) = f(\varepsilon_1, \varepsilon_2)$, la distribución conjunta puede

expresarse como el producto de la marginal y una densidad condicional, de forma que $f(\varepsilon_1, \varepsilon_2) = f(\varepsilon_2/\varepsilon_1)\, f(\varepsilon_1)$. Con esa descomposición tendríamos

$$
\begin{aligned}
P(y/x) &= \int_{\varepsilon} I\left[h(x,\varepsilon) = y\right] f(\varepsilon)\, d\varepsilon \\
&= \int_{\varepsilon_1} \left[\int_{\varepsilon_2} I\left(h(x,\varepsilon_1,\varepsilon_2) = y\right) f(\varepsilon_2 / \varepsilon_1)\, d\varepsilon_2 \right] f(\varepsilon_1)\, d\varepsilon_1 \\
&= \int_{\varepsilon_1} \left[\int_{\varepsilon_2} I\left(x,\varepsilon_1,\varepsilon_2\right) f(\varepsilon_2 / \varepsilon_1)\, d\varepsilon_2 \right] f(\varepsilon_1)\, d\varepsilon_1
\end{aligned}
$$

Puede existir una solución analítica para la integral $g(\varepsilon_1) = \int_{\varepsilon} I(x,\varepsilon_1, \varepsilon_2) f(\varepsilon_2/\varepsilon_1) d\varepsilon_2$, que es sobre ε_2 condicionado a ε_1, de forma que $P = \int_{\varepsilon} g(\varepsilon_1) f(\varepsilon_1) d\varepsilon_1$ puede calcularse por simulación, si no tiene solución analítica.

El procedimiento de cálculo de la simulación es sencillo: se toman numerosas muestras aleatorias de ε a partir de su distribución *f*(ε), se calcula el valor de t(ε) para cada ε tomado, y se promedian los resultados. Esta media simulada es una estimación insesgada de la media verdadera. Cuanto mayor es el número de muestras aleatorias, más se aproximará la media calculada mediante simulación a la verdadera. Este procedimiento de simular una media es común a casi todos los métodos de simulación, y se llama "convenient error partitioning" (*partición conveniente del error*). El modelo Logit mixto emplea este método de cálculo. Otros ejemplos son el modelo Probit binario para datos de panel de Gourieroux y Monfort (1993) y el modelo de respuestas ordenadas de Bhat (1999).

Aunque las distintas especificaciones de *f*(ε) generan distintos modelos, todos comparten una serie de características o propiedades comunes, esencialmente dos: 1) sólo las diferencias en niveles de utilidad tienen relevancia, no los niveles en sí; 2) y la escala de la utilidad es arbitraria. Lo primero conduce a la necesidad de normalizar las constantes y la función de densidad. Lo segundo a la necesidad de normalizar la escala de la utilidad haciendo lo propio con las varianzas.

1.

Es obvio que si añadimos una misma constante a las utilidades de todas las alternativas las diferencias entre éstas no se alteran. Lo importante es que sólo son estimables (sólo están identificados) los parámetros que afectan a dichas diferencias. Los que explican los niveles de utilidad no son identificables. Este importante corolario de la primera propiedad tiene varias implicaciones interesantes.

La utilidad observada para el individuo *n* y la alternativa *j* se representa como $V_{nj} = x'_{nj}\beta$, pero se puede añadir una constante específica para cada alternativa, de forma que $V_{nj} = x'_{nj}\beta + k_j$. Esa constante representa el efecto *medio* sobre la utilidad de todos los factores que no están incluidos en el modelo. Cuando aparece esta constante en la especificación, la variable aleatoria ε_{nj} tiene media cero, por construcción, ya que esa variable recoge las variables no explicitadas, y su efecto medio ya está representado por la constante. Dado que sólo las diferencias en utilidad son relevantes, sólo las diferencias entre constantes lo son también. En tanto que estas diferencias sean las mismas, distintos niveles para estas constantes no implican ningún cambio en el modelo. A la hora de estimar, sólo las diferencias son inidentificables, por lo que hay que normalizar el nivel absoluto de las constantes, normalmente igualando una de ellas a cero.

Un ejemplo. Siguiendo con el caso del coche y el autobús, dos modelos exactamente iguales serían

$$U_c = V_c + \varepsilon_c = \alpha T_c + \beta M_c + k^0_c + \varepsilon_c$$

$$U_b = V_b + \varepsilon_b = \alpha T_b + \beta M_b + k^0_b + \varepsilon_b$$

y

$$U_c = V_c + \varepsilon_c = \alpha T_c + \beta M_c + k^1_c + \varepsilon_c$$

$$U_b = V_b + \varepsilon_b = \alpha T_b + \beta M_b + k^1_b + \varepsilon_b$$

Siempre que $k^1_b - k^1_c = d = k^0_b - k^0_c$. Además, podemos hacer $k_c = 0$ y estimar directamente

$$U_c = \alpha T_c + \beta M_c + \varepsilon_c$$

$$U_b = \alpha T_b + \beta M_b + k_b + \varepsilon_b$$

Siendo ahora $k_b = d$.

Lo mismo ocurre con las variables socio-demográficas que entran en el modelo, es decir, atributos que no varían entre alternativas, como características del

individuo decisor. La única forma de introducir estas variables es especificándolas de forma que creen diferencias entre las utilidades de las alternativas.

Un ejemplo. En el caso de la elección de autobús o coche para ir a trabajar, podemos pensar que el nivel de renta afecta a la utilidad, y además que no afecta igual a un medio de transporte y otro ($\theta^0_c \neq \theta^0_b$, ambos positivos). Esto segundo permite introducir la variable renta (Y) en el modelo:

$$U_c = \alpha T_c + \beta M_c + \theta^0_c Y + \varepsilon_c$$

$$U_b = \alpha T_b + \beta M_b + k_b + \theta^0_b Y + \varepsilon_b$$

Normalizando uno de los parámetros a cero ($\theta^0_c=0$) tendremos

$$U_c = \alpha T_c + \beta M_c + Y + \varepsilon_c$$

$$U_b = \alpha T_b + \beta M_b + k_b + \theta_b Y + \varepsilon_b$$

Donde ahora $\theta_b = \theta^0_b - \theta^0_c$. Es importante señalar que cuando las variables demográficas interaccionan con los atributos de las alternativas propiamente dichas no es necesario normalizar.

La tercera propiedad relacionada con el hecho de que sólo las diferencias en utilidad cuentan tiene que ver con el número de términos de error independientes. Recordamos que la probabilidad de que el individuo n optara por la alternativa i era

$$\begin{aligned} P_{ni} &= \mathrm{Prob}\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i\right) \\ &= \int_{\varepsilon} I\left(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i\right) f(\varepsilon_n) d\varepsilon_n \end{aligned}$$

que es una integral j–dimensional. Sin embargo tenemos j-1 diferencias entre errores, de forma que $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$, lo que nos permite emplear una integral j-1 dimensional:

$$P_{ni} = \text{Prob}\left(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \ \forall j \neq i\right)$$
$$= \int_{\varepsilon} I\left(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \ \forall j \neq i\right) g(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}$$

La función g(.) es la función de densidad *de las diferencias entre errores*, y no coincide con la función de densidad *f*(.) de los errores originales, aunque ambas funciones están relacionadas. El problema es que ***hay un número infinito de densidades para los errores originales compatibles con la función de densidad de las diferencias, debido a que el vector de aquellos tienen una dimensión mayor. Esto explica que la función $f(\varepsilon_n)$ deba ser normalizada por el investigador***.

Esto puede hacerse de muy diversas formas. En el modelo Logit, por ejemplo, las restricciones que la distribución asume son suficientes para normalizar el modelo automáticamente. En el modelo Probit es normal especificar directamente en términos de diferencias en los errores, es decir, trabajar con g(.) sin referencias a *f*(.).

2.

El segundo principio común a todos los modelos de elección discreta es que *la escala de la utilidad es irrelevante*. Si multiplicamos la utilidad de todas las alternativas por una misma constante el orden de las mismas no se ve alterado. Dicho de otra forma, los modelos $U^0_{nj} = V_{nj} + \varepsilon_{nj} \ \forall j$, es equivalente a $U^1_{nj} = \lambda V_{nj} + \lambda\varepsilon_{nj} \ \forall j$ para cualquier $\lambda>0$. Para tener en cuenta esto el investigador tiene que normalizar la *escala* de la utilidad. La forma de hacerlo es normalizando la varianza de los términos de error, pues ambos están ligados: cuando la utilidad se multiplica por λ la varianza de cada término de error se multiplica por λ^2, dado que $\text{var}(\lambda\varepsilon_{nj}) = \lambda^2 \text{var}(\varepsilon_{nj})$.

Si los términos de error son *independientes e idénticamente distribuidos* (*iid*) la varianza será la misma para todos, y sólo hay que igualarla a un valor conveniente (por ejemplo, a uno). Si el modelo es lineal en los parámetros esta normalización altera la interpretación de los coeficientes. En efecto, para un modelo tal que

$$U^0_{nj} = x'_{nj}\beta + \varepsilon^0_{nj} \qquad \text{var}(\varepsilon^0_{nj}) = \sigma^2$$

Que equivale a su versión normalizada, que sería

$$U^1_{nj} = x'_{nj}\,(\beta/\sigma) + \varepsilon^1_{nj} \qquad \text{var}(\varepsilon^1_{nj}) = 1$$

En un modelo Logit, normalmente, no se normaliza a 1, sino a $\pi^2/6 \approx 1{,}6$, con lo que tendremos que

$$U_{nj} = x'_{nj}\,(\beta/\sigma)(1{,}6)^{1/2} + \varepsilon_{nj} \qquad \text{var}(\varepsilon_{nj}) = \pi^2/6$$

En modelos Probit independientes se emplea para normalizar, en cambio, el 1, por lo que al comparar resultados entre una estimación a partir de unos mismos datos y dos modelos distintos (Logit y Probit, por ejemplo) hay que tener muy en cuenta esto. Los coeficientes del Logit serán $(1{,}6)^{1/2}$ veces mayores que los del Probit sólo por efecto de la normalización. Pero también puede ocurrir que estimemos un mismo modelo (Logit o Probit) con dos bases de datos distintas. Las diferencias en el nivel de los coeficientes estimados reflejará las diferencias en las varianzas de los factores inobservados: dado que los coeficientes están divididos por la desviación estándar de esos factores (después de normalizar), allí donde la varianza sea mayor los coeficientes serán menores.

Si los errores son heterocedásticos, es decir, con distintas varianzas para cada individuo, o grupo de individuos (regiones, períodos, bases de datos… véase Swait y Louviere (1993)), es obvio que habrá que normalizar para cada grupo separadamente. Si se estima de forma conjunta, fijar la escala global de utilidad supone normalizar la varianza en un área, y estimar la varianza de la otra, relativa a la primera.

Un ejemplo. Si, por ejemplo, hay dos muestras para estudiar el medio de transporte al trabajo, una para Madrid y otra para Barcelona, con varianzas distintas, tendremos

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon^M_{nj} \qquad \forall n \text{ en Madrid}$$

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon^B_{nj} \qquad \forall n \text{ en Barcelona}$$

Donde la $\varepsilon^M_{nj} \neq \varepsilon^B_{nj}$. Podemos hacer $k = \text{var}(\varepsilon^B_{nj})/\text{var}(\varepsilon^M_{nj})$, y esta k es conocida como el *parámetro de escala*. Si dividimos la utilidad de los viajeros de Barcelona por $k^{1/2}$, lo que no afecta el orden, como sabemos. El modelo quedaría:

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon^M_{nj} \qquad \forall n \text{ en Madrid}$$

$$U_{nj} = (\alpha k^{-1/2}) T_{nj} + (\beta k^{-1/2}) M_{nj} + \varepsilon^B_{nj} \quad \forall n \text{ en Barcelona}$$

La varianza de los ε_{nj} será ahora la misma para ambas ciudades, ya que se cumplirá que

$$var(\varepsilon^{B}_{nj}\, k^{-1/2}) = k^{-1}\, var(\varepsilon^{B}_{nj}) = [var(\varepsilon^{M}_{nj})/var(\varepsilon^{B}_{nj})]\, var(\varepsilon^{B}_{nj}) = var(\varepsilon^{M}_{nj})$$

En el modelo se estiman simultáneamente α, β y k.

Si los errores no son independientes, es decir, están correlados entre alternativas, la normalización de escala se hace más difícil. Ya no basta con normalizar el error de una alternativa para fijar la escala global de las utilidades, o de las diferencias entre utilidades. En términos de cuatro alternativas, la utilidad sería $U_{nj} = V_{nj} + \varepsilon_{nj}$, j=1,…,4. El vector de errores es $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{n4})$ tiene media cero y una matriz de covarianzas:

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

Dado que sólo las diferencias de utilidad son relevantes, este modelo equivale totalmente a otro en el que a todas las utilidades se les resta alguna de las cuatro. Si restamos la primera, el nuevo modelo será $\tilde{U}_{njl} = \tilde{V}_{njl} - \tilde{\varepsilon}_{njl}$ para j=2,3,4, donde, como es obvio, $\tilde{U}_{njl} = U_{nj} - U_{n1}$, $\tilde{V}_{njl} = V_{nj} - V_{n1}$, siendo el vector de diferencias en los errores $\tilde{\varepsilon}_{n1} = \left[(\varepsilon_{n2} - \varepsilon_{n1}),(\varepsilon_{n3} - \varepsilon_{n1}),(\varepsilon_{n4} - \varepsilon_{n1})\right]$. La varianza de cada diferencia entre errores *depende de las varianzas y covarianzas de los errores originales*. Cuando dichas covarianzas no son cero (errores no independientes), normalizar la varianza de los errores originales no supone normalizar la de las diferencias entre errores.

Por ejemplo, la varianza de la diferencia entre el primer y segundo error es $var(\tilde{\varepsilon}_{n21}) = var(\varepsilon_{n2} - \varepsilon_{n1}) = var(\varepsilon_{n1}) + var(\varepsilon_{n2}) - 2\, cov(\varepsilon_{n1}, \varepsilon_{n2}) = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. La covarianza entre la diferencia entre el primer y segundo error, y la diferencia entre el primer y tercer error sería: $cov(\tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}) = E(\varepsilon_{n2} - \varepsilon_{n1})(\varepsilon_{n3} - \varepsilon_{n1}) = E(\varepsilon_{n2}\varepsilon_{n3} - \varepsilon_{n2}\varepsilon_{n1} - \varepsilon_{n3}\varepsilon_{n1} + \varepsilon_{n1}\varepsilon_{n1}) = \sigma_{23} - \sigma_{21} - \sigma_{31} + \sigma_{11}$. Repitiendo estos mismos cálculos para el resto de varianzas y covarianzas tendremos:

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma_{11}+\sigma_{22}-2\sigma_{12} & \sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13} & \sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14} \\ \sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13} & \sigma_{11}+\sigma_{33}-2\sigma_{13} & \sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14} \\ \sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14} & \sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14} & \sigma_{11}+\sigma_{44}-2\sigma_{14} \end{pmatrix}$$

Fijar un valor para una de las varianzas de los errores originales (por ejemplo, σ_{11} = k), como puede verse, no sería suficiente para fijar el valor de la varianza de una de las diferencias entre errores. ¿Cómo fijar la escala de la utilidad cuando los errores no son *iid*? La respuesta es asignar un valor arbitrario a la varianza de una diferencia de errores, por ejemplo, $\tilde{\varepsilon}_{n21}$=1. La matriz de covarianzas para las diferencias entre errores quedaría

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & (\sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13})/m & (\sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14})/m \\ (\sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13})/m & (\sigma_{11}+\sigma_{33}-2\sigma_{13})/m & (\sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14})/m \\ (\sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14})/m & (\sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14})/m & (\sigma_{11}+\sigma_{44}-2\sigma_{14})/m \end{pmatrix}$$

donde m = $\sigma_{11} + \sigma_{22} - 2\sigma_{12}$. La utilidad queda dividida por $(\sigma_{11} + \sigma_{22} - 2\sigma_{12})^{1/2}$ con esa normalización de la varianza de las diferencias entre errores.

Cuando los errores originales son *iid*, las covarianzas son cero, y normalizando los errores originales automáticamente normaliza las diferencias entre errores. En efecto, si los errores son iid tendremos que $\sigma_{ii} = \sigma_{jj}$ y $\sigma_{ij} = 0$ para $i \neq j$. Por tanto, si σ_{11} = k al normalizar, la varianza de la diferencia entre errores se convierte en $\sigma_{11} + \sigma_{22} - 2\sigma_{12}$ = k + k – 0 = 2k, y las varianzas de dichas diferencias quedan normalizadas también.

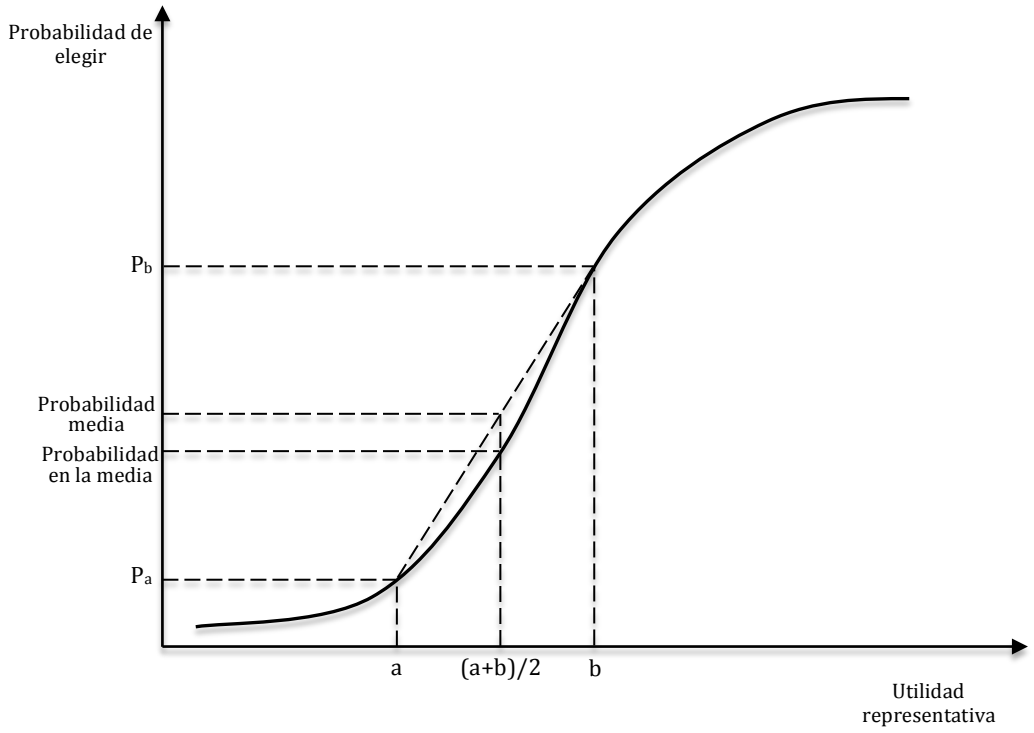
La normalización reduce el número de parámetros estimables de la matriz de covarianzas. No todas las varianzas y covarianzas de los errores originales se podrán estimar. De 10 elementos en la matriz original para 4 alternativas pasaríamos a 6 elementos en la matriz de covarianzas para las diferencias entre errores, de entre los que 1 es una constante. Un modelo con j alternativas llevaría a j(j-1)/2 – 1 parámetros estimables después de la normalización de las varianzas de las diferencias entre errores. La matriz resultante para el caso de 4 alternativas sería del tipo:

$$\tilde{\Omega}_1^* = \begin{pmatrix} k & \omega_{ab} & \omega_{ac} \\ \omega_{ab} & \omega_{bb} & \omega_{bc} \\ \omega_{ac} & \omega_{bc} & \omega_{cc} \end{pmatrix}$$

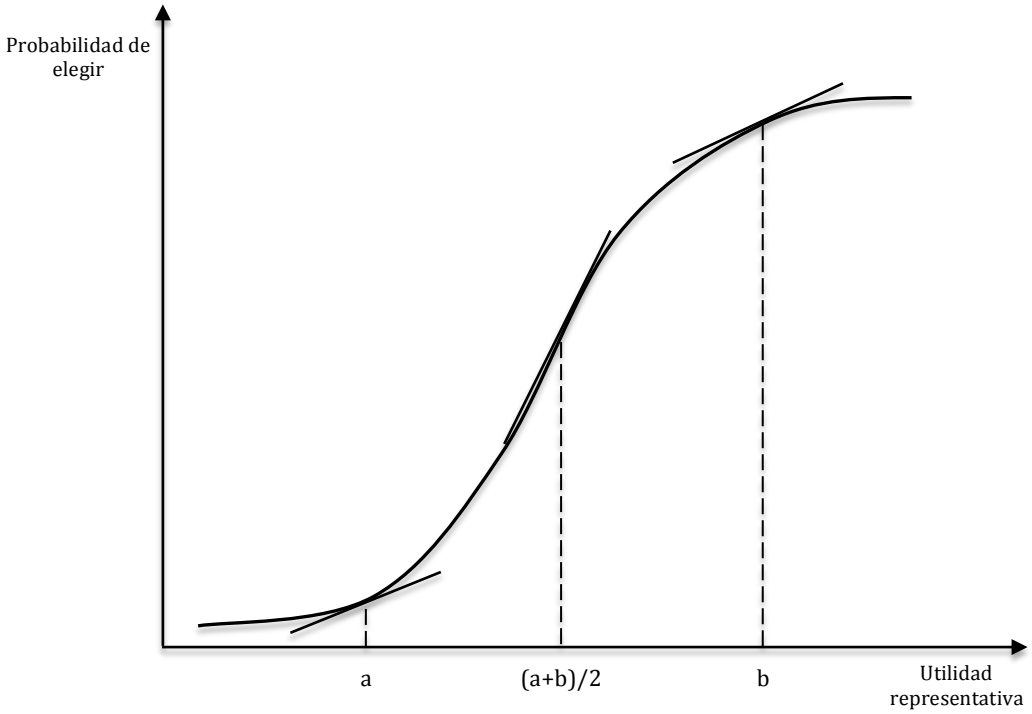
La normalización también afecta a cómo interpretamos el modelo. El parámetro ω_{bb} es la varianza de la diferencia entre los errores de la alternativa 1 y 3 relativa a la varianza de la diferencia entre los errores de las alternativas 1 y 2. Para acabar de complicar la cosa, las varianzas de las diferencias entre errores reflejan también la covarianza entre ellos.

Los modelos Logit y Logit anidado *imponen una fuerte serie de restricciones sobre los términos de error, y por tanto sobre su matriz de covarianzas*, por lo que no es necesario normalizar. Los modelos Probit y Logit mixto, en cambio, requieren de normalización, lo que complica la especificación e interpretación de esos modelos.

Los modelos de elección discreta son no lineales en las variables explicativas, lo que tiene implicaciones relativas a la posibilidad de agregar. En un modelo lineal puede hacerse la estimación con datos individuales, e insertar datos agregados para las variables explicativas en el modelo estimado para obtener un valor de la variable dependiente también agregado. Si h_n es el gasto en vivienda de la persona n, e y_n es la renta de esa persona, y si el modelo es $h_n = \alpha + \beta y_n$, es suficiente con insertar la renta media de una población y* para obtener h*, el gasto medio en vivienda. Si el modelo no es lineal la relación no es tan simple. El problema está en que en ese caso el modelo evaluado para valores medios de las explicativas no coincide con la media de los resultados del modelo para los valores individuales.



Con el tema del cálculo de las elasticidades el error puede ser incluso mayor de lo que acabamos de ver



La derivada de la probabilidad de elegir ante un cambio infinitesimal en la utilidad representativa es muy pequeña, tanto en a como en b. Por tanto, la media de las derivadas es también pequeña. Sin embargo, la derivada en el punto medio tiene una pendiente muy fuerte. El error puede ser muy grande (de hasta 2 o tres veces la magnitud, según Talvitie, 1976).

Se pueden obtener variables dependientes estimadas para valores agregados de forma consistente utilizando dos tipos de procedimiento: **1)** mediante *segmentación*; **2)** y mediante *enumeración de la muestra* (*sample enumeration*).

1.

El procedimiento de *segmentación* funciona bien cuando el número de variables explicativas es pequeño, y toman pocos valores (por ejemplo, cuando son variables categóricas). El número total de "tipos" de individuos es limitado. Imaginemos que el nivel de educación y el género explican la utilidad en un determinado problema (el que sea). La educación tiene cuatro posibles alternativas: no acabó la enseñanza media; la acabó pero no fue a la universidad; fue a la universidad pero no se graduó; se graduó. Hay 8 tipos de individuos, según 4 tipos de educación y 2 sexos. Las probabilidades de elegir una opción no varía entre individuos, sino entre cada uno de esos 8 grupos. Ahora, si el investigador sabe cuántos individuos hay dentro de cada uno de esos 8 *segmentos*, se puede calcular cualquier resultado agregado sumando ponderadamente las probabilidades de los 8 segmentos. El número de individuos que eligen la alternativa *i* será:

$$\hat{N}_i = \sum_{s=1}^{8} w_s P_{si}$$

donde P_{si} es la probabilidad de que un individuo en el segmento *s* elija la alternativa *i*, y w_s es el número de decisores en el segmento *s*.

2.

La enumeración de muestra se aplica a casos más generales. Consiste en sumar o promediar las probabilidades de elegir de cada individuo en la muestra. P_{ni} es la probabilidad para el individuo *n* de elegir *i*, y si tenemos una muestra de N individuos, siendo w_n la inversa de la probabilidad de que el individuo *n* caiga en la muestra, o el número de individuos similares a *n* en la población (el tamaño relativo del *segmento* al que pertenece, si se permite la expresión). Si la muestra que seleccionó los individuos para los que hay datos es aleatoria, todos los w_n son iguales. Una estimación consistente del número de individuos en la población que escogen *i* sería la suma ponderada de las probabilidades individuales:

$$\hat{N}_i = \sum_{n} w_n P_{ni}$$

La probabilidad media es $\hat{N}_i / N$. Si queremos calcular elasticidades o derivadas medias, las calculamos para los individuos en la muestra y después sumamos ponderando.

Como resumen, la siguiente tabla sintetiza este conjunto de propiedades y características compartidas por los modelos de elección discreta.

Tabla 4. *Propiedades comunes de los modelos de elección discreta*

Sólo las diferencias en niveles de utilidad tienen relevancia			**La escala de la utilidad es arbitraria**		
Constantes específicas para cada alternativa	Variables sociodemográficas	Número de términos de error independientes	Normalización con errores *iid*	Normalización con errores heterocedásticos	Normalización con errores correlados (no independientes)
Normalizar (igualar a cero) una de ellas	Normalizar (igualar a cero) el parámetro para una de las alternativas	Especificar sobre las diferencias de errores	Normalizar (igualar a un número) la varianza de los términos de error	Normalizar la varianza para un grupo y estimar la varianza del otro relativa al primero (parámetro de escala)	Hay que normalizar las varianzas de las diferencias entre errores, y no las varianzas de los errores originales