INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Marketing II
IN5602 [a]

Profesor: Marcel Goić

Curso: IN5602
Semestre: Otoño 2015

[a] Esta es una versión preliminar del apunte, por lo que podrían existir errores. Agradecemos enviar sus sugerencias y comentarios al mail del profesor Marcel Goic mgoic@dii.uchile.cl

Índice general

Parte I

Modelos Probabilísticos

Capítulo 1

Modelos probabilísticos

1.1. Introducción

Usualmente, y en el contexto de Marketing, estamos interesados en estudiar el comportamiento de las personas, de modo de entenderlo y realizar acciones estratégicas en función de los aprendizajes adquiridos. Así, se puedem definir dos tipos de enfoques a usar según distintos supuestos en el comportamiento de los agentes (tomadores de decisiones):

- Enfoque Estructural: Este enfoque asume que los los agentes se comportan de manera racional, tomando decisiones de modo de maximizar sus utilidades. Usualmente aparece cuando hay disponibilidad de largos volumenes de datos.
- Modelos Probabilísticos: Este enfoque asume que los agentes se comportan en base a decisiones aleatorias. Usualmente aparece cuando se tiene información reducida y/o agregada respecto al comportamiento de los agentes en estudio.

En primera instancia, se estudiará el enfoque probabilistico, esto es, el enfoque en cual se asume que los tomadores de decisiones se comportan de manera aleatoria. Dicho enfoque posee una metodología de modelamiento sugerida, que comparten todos los modelos que se verán a lo largo del curso.

La metodología es:

1. Determinar el problema de decisión a estudiar y la información requerida.
2. Identificar el comportamiento a nivel individual.
3. Seleccionar la distribución de probabilidad que caracterice el comportamiento individual $f(x|\theta)$.
4. Escoger la distribución que caracterice la distribución de las características latentes de la población $g(\theta)$.
5. Derivar la distribución agregada del comportamiento de interés.

$$f(x) = \int f(x|\theta)g(\theta)d\theta \tag{1.1}$$

6. Estimar los parámetros.
7. Usar los resultados para resolver el problema y tomar medidas de gestión.

El enfoque de modelos probabilísticos permite abordar una gran cantidad de problemas asociados al Marketing, de entre los cuales se considerarán:

- Timing: Situaciones ligadas a la duración de una determinada conducta de un cliente, como por ejemplo: tiempo de permanencia en una compañía y tiempo de adopción de un cierto producto innovador.
- Conteo: Situaciones ligadas al estudio de llegadas de clientes y contabilización de una determinada conducta, como por ejemplo: número de visitas a un portal web y la cantidad de productos comprados en una tienda de retail.
- Elección: Situaciones asociadas a las decisiones de elección de un determinado cliente, como por ejemplo: clientes que eligen responder una campaña publicitaria y la elección de cambiar o no de canal de televisión.

1.2. Modelos de Duración

En años recientes, las mejoras en las tecnologías de información han dado como resultado un aumento en la disponibilidad de data acerca de los individuos en determinadas situaciones de consumo. Esta tendencia se relaciona íntimamente con el creciente deseo de los gerentes de marketing respecto a utilizar esta data disponible para aprender de manera exhaustiva sobre el comportamiento de los clientes. Muchos analistas tratan de describir y predecir el comportamiento de los consumidores usando variables observables, como lo son variables transaccionales (monto gastado, tienda donde se adquirió un determinado producto, fecha de la compra, etc.) como así también variables que caracterizan a los individuos (edad, nivel socio-económico, estado civil, etc.). A partir de esta información es posible aplicar modelos de regresión lineal o árboles de decisión, con el objetivo de poder proyectar comportamientos o bien comprobar o rebatir hipótesis que previamente se tenían respecto a un escenario determinado.

En este capítulo, se considera un enfoque distinto al anterior, en el cual las decisiones de los individuos se desprenden de un comportamiento **aleatorio**, en que las decisiones no dependen únicamente de variables descriptivas del modelo, sino que también provienen del resultado de un proceso estocástico no observable que opera intrínsecamente en los individuos, es decir, la asunción que el comportamiento se desprende de una distribución de probabilidades que puede variar dependiendo del modelo a estimar y de la complejidad del mismo (alternativamente se puede considerar el enfoque racionalista que considera que los individuos siempre actúan en forma racional, lo que de acuerdo a la experiencia empírica, no se cumple siempre).

Ejemplo 1: Supongamos que un cliente hizo 2 compras el año pasado de nuestro producto. ¿Esto implica inmediatamente que el consumidor mantendrá ese patrón y este año volverá a ese nivel de consumo? ¿O existe alguna posibilidad de que el cliente incremente o disminuya su consumo? ¿Cuál es el proceso que hay detrás?

En lo que sigue, consideraremos 3 tipos de modelos a estimar:

1. **Modelos de duración en tiempo discreto**.
2. **Modelos de duración en tiempo continuo sin dependencia en la duración**.
3. **Modelos de duración en tiempo continuo con dependencia en la duración**.

1.2.1. Modelos de duración en tiempo discreto

Supongamos el siguiente escenario: A través de una propuesta de valor atractiva, adquirimos un cliente. ¿Durante cuántos periodos estará afiliado a la compañía? Se considera que cada periodo se puede cuantificar en términos discretos (días, semanas, meses, años). Algunos ejemplos a considerar:

- Un usuario descarga una aplicación para su teléfono inteligente. ¿Por cuántos meses la utilizará?
- Adquirimos un cliente en un banco. ¿Durante cuántos años permanecerá como cliente?
- Un cliente se suscribe a un plan telefónico o de internet. ¿Por cuántos periodos se mantendrá suscrito?

Modelo Geométrico desplazado

Asumamos que se tiene una cartera de clientes que van abandonando la relación comercial para nunca más retomarla en cualquier periodo definido. De acuerdo a lo descrito en las secciones anteriores, intentaremos describir de manera probabilística la situación.

Supongamos que al final de cada periodo, un cliente decide de manera aleatoria si continúa afiliado a una determinada compañía, esto es, de acuerdo a un proceso de Bernoulli, decide con cierta probabilidad si cancela la relación comercial con l empresa (y con el coplemente respectivo decide su permanencia). Para cada individuo, asumiremos que la probabilidad con la cuál decide **no cambia en el tiempo**, denotándola θ. Finalmente, y como primer approach, asumiremos que dicha probabilidad es de igual forma idéntica a lo largo de individuos distintos (modelo homogéneo).

Sea T la variable aleatoria relativa a la duración de la relación comercial entre el cliente y la compañía, es decir la variable que describe el instante en el cual esta relación se acaba. De acuerdo a la descripción anterior, la variable aleatoria T sigue una distribución **Geométrica Desplazada (sG)** con parámetro θ, es decir, el comportamiento de los individuos puede ser descrito formalmente de acuerdo a la siguiente relación:

1. Probabilidad de que un individuo cualquiera abandone la relación comercial exactamente en el periodo t:
$$P(T = t|\theta) = \theta(1-\theta)^{t-1}$$
2. Probabilidad de que un individuo cualquiera abandone la relación comercial en un periodo posterior al periodo t:
$$P(T > t|\theta) = (1-\theta)^{t}$$

No es muy difícil aplicar un modelamiento a partir de lo anterior para intentar dilucidar de qué forma se debería comportar un determinado grupo de individuos a partir de la data transaccional que se tiene. Veámoslo a partir de un ejemplo práctico:

Ejemplo 2: Consideremos un cohorte inicial de 1000 clientes (indexado por el número 0). Supongamos que año a año, un determinado número de clientes se retira del negocio por razones que se desconocen a priori, pero que asumimos provienen de un proceso estocástico en el que cada cliente en forma independiente toma la decisión de permanecer o abandonar a partir del lanzamiento de una moneda (Bernoulli), esto es, con probabilidad θ abandona y con probabilidad $1-\theta$ permanece en la compañía. La data histórica se presenta a continuación:

Año	# de Clientes	% de Permanencia	% de Retención
0	1000	100 %	-
1	631	63 %	63 %
2	468	47 %	74 %
3	382	38 %	82 %
4	326	33 %	85 %
5	289	29 %	89 %
6	262	26 %	91 %
7	241	24 %	92 %

Entendiéndose el % de Retención como el porcentaje de clientes que se mantuvo en la relación comercial respecto al periodo anterior.

Sin embargo, aún no sabemos cuanto vale θ (es un parámetro poblacional). Dicho valor lo estimaremos mediante el método de máxima verosimilitud, para el cual es necesario determinar la probabilidad de observar lo que efectivamente se está observando (densidad conjunta) asumiendo independencia entre las muestras:

- Densidad de probabilidades conjunta:

$$f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot ... \cdot f(x_n|\theta)$$

- Función de verosimilitud:

$$L(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

A partir de esto y de la data disponible, las contribuciones a la verosimilitud son las siguientes (asumiendo como modelo de comportamiento la distribución geom étrica desplazada)

Año	# de Clientes	# de Abandonos	Pr
0	1000	-	-
1	631	369	$P(T=1\|\theta) = \theta^{369}$
2	468	163	$P(T=2\|\theta) = ((1-\theta)^{(2-1)} \cdot \theta)^{163}$
3	382	86	$P(T=3\|\theta) = ((1-\theta)^{(3-1)} \cdot \theta)^{86}$
4	326	56	$P(T=4\|\theta) = ((1-\theta)^{(4-1)} \cdot \theta)^{56}$
5	289	37	$P(T=5\|\theta) = ((1-\theta)^{(5-1)} \cdot \theta)^{37}$
6	262	27	$P(T=6\|\theta) = ((1-\theta)^{(6-1)} \cdot \theta)^{27}$
7	241	21	$P(T=7\|\theta) = ((1-\theta)^{(7-1)} \cdot \theta)^{21}$
>7	-	-	$P(T>7\|\theta) = ((1-\theta)^{7})^{241}$

Dado que maximizar un producto es complicado, aplicamos logaritmo a lo anterior, de modo de construir la función de log verosimilitud:

- Función de log verosimilitud:

$$\widehat{l}(\theta|x_1, x_2, ..., x_n) = \ln(L(\theta|x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

Con lo anterior, es sencillo maximizar la función de log verosimilitud para un θ desconocido, con lo que se tiene[1]:

$$\widehat{\theta} = 0,226027$$
$$\widehat{l} = -1794,62$$

El modelo antes presentado, si bien permite tomar medidas de gestión a partir de un modelo sencillo, es poco realista (pues se asume que la población posee igual probabilidad de abandono). Una manera de incluir mayor complejidad al modelo y hacerlo más robusto, es asumiendo que la población no es homogénea, sino que existen segmentos de individuos quienes al ser agrupados, presentan un comportamiento similar. La forma más sencilla de modelar esto es asumiendo que la población presenta 2 patrones de comportamiento (2 segmentos), es decir, para un segmento de individuos, la decisión de abandonar o permanecer se identifica a partir de un parámetro θ_1 (del mismo modo que en el caso anterior), y para el otro segmento la decisión se determina a partir de un parámetro θ_2 distinto de θ_1. Formalmente, las relaciones que describen de mejor manera esto son las siguientes:

1. Probabilidad de que un individuo cualquiera abandone la relación comercial exactamente en el periodo t en una población con 2 segmentos:

$$P(T = t|\theta_1, \theta_2, \pi) = \theta_1(1-\theta_1)^{t-1}\pi + \theta_2(1-\theta_2)^{t-1}(1-\pi)$$

2. Probabilidad de que un individuo cualquiera abandone la relación comercial en un periodo posterior al periodo t en una población con 2 segmentos:

$$P(T > t|\theta_1, \theta_2, \pi) = (1-\theta_1)^{t}\pi + (1-\theta_2)^{t}(1-\pi)$$

En el modelo anterior π representa el porcentaje de la población que pertenece al segmento 1, de tal forma que su complemento $1-\pi$ representa el porcentaje de la población que pertenece al segmento 2[2].

[1] Se puede hacer fácilmente en excel a través de la herramienta *solver*. PROPUESTO
[2] Este modelamiento es fácilmente expandible a 2 o más segmentos. PROPUESTO.

Modelo Beta Geométrico desplazado

Los modelos anteriores funcionan bien cuando la población se comporta de manera distinta entre clases latentes, y similar al interior de cada clase latente. Sin embargo, puede ser mucho más realista e interesante el asumir que existe una heterogeneidad continua en la población, es decir que existe un número infinito de segmentos (o al menos tendiente a infinito) de manera de capturar todas las preferencias individuales de cada miembro de la población considerada.

Para estos propósitos, ya no asumiremos que la probabilidad de abandono θ sigue una distribución discreta de Bernoulli (éxito-fracaso), sino que asumiremos que el parámetro proviene de una distribución continua *Beta* de parámetros α y β.

Por tanto, es posible calcular las probabilidades antes presentadas en forma análoga, aplicando el enfoque antes mencionado (probabilidades totales):

1. Probabilidad de que un individuo cualquiera abandone la relación comercial exactamente en el periodo t:

$$P(T = t|\alpha, \beta) = \int_0^1 P(T = t|\theta) B(\theta|\alpha, \beta)\, d\theta$$

   Recordar que:

$$B(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

2. Probabilidad de que un individuo cualquiera abandone la relación comercial en un periodo posterior al periodo t:

$$P(T > t|\alpha, \beta) = \int_0^1 P(T > t|\theta) Beta(\theta|\alpha, \beta)\, \mathrm{d}\theta$$

Al desarrollar la primera integral antes mencionada, y reconociendo las relaciones de la distribución *Beta*, se tiene que:

$$P(T = t|\alpha, \beta) = \frac{B(\alpha+1, \beta+t-1)}{B(\alpha, \beta)}$$

Notar que, se usa indistintamente el $B(\alpha, \beta)$ para hacer alusión tanto a la **función** como a la **distribución** Beta. Bajo ninguna circunstancia dichos objetos son iguales.

Ejemplo 3: Considerando la misma situación que se presentó en el ejemplo anterior (clientes que año a año abandonan la relación comercial), pero ahora asumiendo que existe un comportamiento heterogéneo en la población, es posible reconocer que existe una recursividad en la fórmula del cálculo de la probabilidad de abandono en cada período de la siguiente forma:

$$P(T = t|\alpha, \beta) = \begin{cases} \dfrac{\alpha}{\alpha+\beta} & t = 1 \\ P(T = t-1|\alpha, \beta)\dfrac{\beta+t-2}{\alpha+\beta+t-1} & t > 1 \end{cases}$$

Modelo que al ser evaluado, da el siguiente resultado:

$$\widehat{\alpha} = 0,7041$$
$$\widehat{\beta} = 1,1820$$
$$\widehat{l} = -1680,27$$

Notar que existe una notoria diferencia en cuanto al valor de la log verosimilitud obtenida por el modelo heterogéneo respecto al modelo homogéneo. Si bien, esto indica una mejora del modelo, es necesario realizar la comparación en base a métricas de evaluación mas precisas (AIC; BIC, etc.).

1.2.2. Modelos de duración en tiempo continuo sin dependencia en la duración

Para algunos modelos, el medir el tiempo como si fueran períodos discretos puede ser un buena aproximación de acuerdo a los objetivos del análisis que se desea llevar a cabo.

En otros casos, puede ser en cambio más útil considerar el tiempo como una variable continua, debido a que podría interesar el medir la ocurrencia de un suceso de manera *más exacta*. Algunos casos relativos a este enfoque son:

- Tiempos de respuesta a una campaña promocional de marketing directo.
- Tiempo entre visitas a nuestro website.
- Tiempos entre llamadas en un call center.
- Tiempos de operación en la industria de servicios.

Al igual que en el caso de los modelos en tiempo discreto, lo que interesa es poder implementar un modelo que tenga una forma funcional flexible para ser trabajada y modificada fácilmente, que logre ajustar a la data histórica que se tiene, y que logre además proyectar el comportamiento futuro de los clientes, es decir, que sea un buen modelo predictivo para tomar acciones en función de aquello.

Modelo Exponencial

Supongamos que nos interesa medir el tiempo que pasa desde que se lanza un producto hasta que el consumidor decide adquirirlo. Existen muchos factores externos que determinan esta decisión: exposición a publicidad, número de visitas a la tienda, llamadas recibidas por call center, entre otras. Nuevamente asumiremos que el comportamiento es aleatorio, es decir, que los consumidores deciden el momento en el cuál van a consumira partir de una distribución de probabilidades.

Esto podemos modelarlo a partir de la distribución exponencial[3].

Supongamos la variable aleatoria T definida como el tiempo en que un cliente va a consumir nuestro producto por primera vez. Asumiremos que esta variable está exponencialmente distribuida con una tasa λ. De esta forma, se tiene que la función de distribución acumulada de la variable aleatoria será:

[3]Recordar el curso de Investigación de Operaciones

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}$$

Notar que el término anterior representa la probabilidad que un cliente consuma el producto **antes de** t.

Ahora bien, esto inmediatamente deja en evidencia una limitante a este modelo: para un t muy grande, todos los consumidores van a probar[4], lo cual no es una situación del todo realista. Es necesario, en consecuencia, imponer que existe una fracción de clientes dentro de la muestra considerada que nunca probará el producto y así es posible solucionar la limitante encontrada (2 clases latentes).

1. Segmento que prueba: Tamaño π

$$\lambda = \theta$$
$$\Rightarrow P(T \leq t) = 1 - e^{-\theta t}$$

2. Segmento que no prueba: Tamaño $(1 - \pi)$

$$\lambda = 0$$
$$\Rightarrow P(T \leq t) = 0$$

Luego, la probabilidad total será:

$$P(T \leq t) = P(T \leq t|Prueba)P(Prueba) + P(T \leq t|NoPrueba)P(NoPrueba)$$
$$= (1 - e^{-\theta t})\pi$$

Es importante notar que si bien el modelo describe probabilidades en tiempo continuo, la data aún se presenta y obtiene en tiempo discreto. Incorporando esto, es posible construir la función de log verosimilitud calculando las probabilidades de adopción del producto entre los limites del intervalo temporal definido por el periodo de medición, es decir:

$$P(t_0 \leq T \leq t_1) = F(t_1) - F(t_0)$$

Por lo que la función de log verosimilitud se define como (considerando n periodos discretos para el cálculo):

$$LL(\pi, \theta|data) = N_1 \ln[P(0 \leq T \leq 1) + N_2 \ln[P(1 \leq T \leq 2)] + ... + (N_{panel} - \sum_{i=1}^{n} N_i) \ln[P(T > n)]$$

Adicionalmente, es de interes calcular los valores predichos por el modelo, de modo de realizar predicciones futuras. $F(t)$ representa la probabilidad que un cliente escogido aleatoriamente pruebe el producto en t (tal que $t = 0$ corresponde al instante de lanzamiento del producto. La estimación del futuro se puede hacer a través de la esperanza:

[4] Recordar que $\lim e^{-t} = 0$

$$\mathbb{E}[T(t)] = N_{panel} \cdot \widehat{F}(t)$$

Antes de avanzar, es importante aclarar la distinción de un modelo *sin dependencia en la duración*. Esto se puede explicar con la propiedad fundamental de la distribución exponencial:

Propiedad fundamental: La distribución exponencial no tiene memoria, es decir, poseer información de que un elemento ha sobrevivido un tiempo 's' hasta este momento no modifica la probabilidad de que sobreviva un periodo t más. Es decir la probabilidad de que ocurra un suceso no depende del tiempo en que aún no ha ocurrido. Se puede demostrar matemáticamente:

$$P(T > s+t|T > s) = \frac{P(T > s+t)}{P(T > s)} = \frac{1 - P(T \leq s+t)}{1 - P(\leq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

Modelo Gamma Exponencial

Análogamente a la situación anterior, ahora se asume que el comportamiento de la población es heterogéneo, es decir, que existen diferentes clases de clientes en la población. Esto busca complejizar la suposición que antes hicimos al considerar un grupo de clientes que nunca consume.

Por tanto, el modelo heterogéneo ahora considerará que la tasa de prueba λ se distribuye $Gamma$ en la población:

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

Dónde r es un parámetro de forma y α es un parámetro de escala.

Al incorporar la heterogeneidad mencionada, la probabilidad que un cliente adquiera un producto antes de un tiempo t es la siguiente:

$$\begin{aligned} P(T \leq t) &= \int_0^\infty P(T \leq t|\lambda) g(\lambda)\, d\lambda \\ &= 1 - \left(\frac{\alpha}{\alpha + t}\right)^r \end{aligned}$$

Este modelo lo llamaremos *Gamma Exponencial*.

1.2.3. Modelos de duración en tiempo continuo con dependencia en la duración

Otra de las grandes limitantes del modelo *Exponencial* es que posee pérdida de memoria, es decir la probabilidad de adopción de un cliente no cambia a medida que pasa el tiempo. Se necesita incorporar esta distinción, es decir, la probabilidad de que un evento ocurra dado que hasta este momento no ha ocurrido. Esto último se conoce como *tasa de riesgo* o *hazard rate*:

$$h(t) = \frac{f(t)}{1 - F(t)}$$

Gráficamente, la tasa de riesgo se comporta de la siguiente manera 1.1:

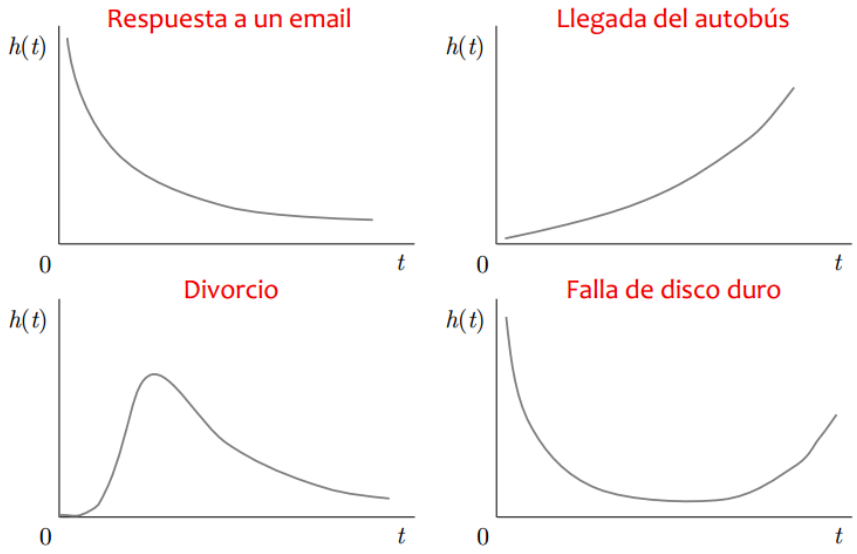

Figura 1.1: Ejemplos de tasas de riesgo

En el primer caso, la intuición es que si una persona no ha respondido a un e-mail, cada vez es menos probable que lo responda, pues en general las personas tienden a ignorar los correos con una antigüedad superior a un par de días. En la llegada de un bus - si bien en ramos pasados se ha modelado con una exponencial, es decir, sin memoria - se asume que a medida que más se demora en llegar al paradero, cada vez la espera debe ser menor, pues tarde o temprano este deberá llegar. Los otros análisis quedan propuestos, pero la intuición es fácil de comprender.

A partir de la tasa de riesgo, se puede definir unívocamente la distribución de una variable aleatoria no negativa a través de la siguiente integral:

$$F(t) = 1 - \exp\left(-\int_0^t h(u)\,du\right)$$

Este concepto será útil para definir los modelos de duración en tiempo continuo en que la duración sí es un factor relevante.

Modelo Weibull

A pesar de las generalizaciones de las funciones de tasas de riesgo para generar modelos de tiempo de ocurrencia, nos enfocaremos en la distribución Weibull debido a que es fácil de trabajar y entrega una fórmula cerrada muy similar a la de la distribución exponencial. Se tiene que para la misma variable aleatoria T que se definió en la sección anterior, la probabilidad de ocurrencia de que un cliente pruebe nuestro producto en un tiempo inferior a t será:

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t^c}$$

Y la tasa de riesgo asociada a esta distribución:

$$h(t) = c\lambda t^{c-1}$$

El primer parámetro λ que compone la fórmula lo interpretamos como un parámetro de escala, mientras que el parámetro c le llamamos parámetro de forma. Es importante notar que para $c = 1$, la distribución se convierte en la distribución exponencial, por lo que se puede decir que la distribución Weibull es una generalización de la exponencial. Notar además que para $c = 1$, la tasa de riesgo es constante, lo que es consistente con la propiedad de pérdida de memoria de la distribución exponencial.

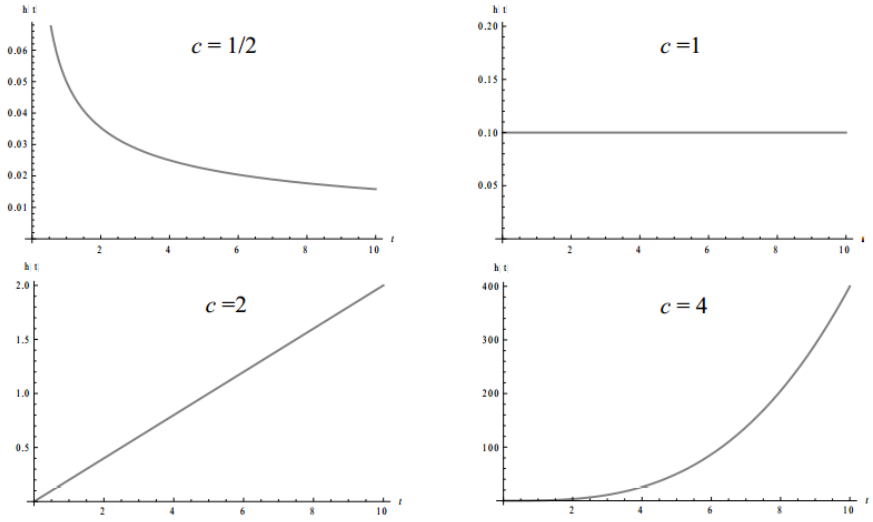

Figura 1.2: Ejemplos de tasas de riesgo para distintos valores de *c*.

En la distribución de Weibull generalizada por tanto, la propiedad de pérdida de memoria no aplica como en el caso de la exponencial, es decir, la probabilidad de ocurrencia varía a medida que pasa el tiempo:

$$P(T > s + t|T > s) = \frac{P(T > s + t)}{P(T > s)} = \frac{1 - P(T \leq s + t)}{1 - P(\leq s)} = \frac{e^{-\lambda(s+t)^c}}{e^{-\lambda s^c}}$$

Modelo Gamma Weibull

Una de las propiedades interesantes de la distribución Weibull, es que es sencillo introducir heterogeneidad sobre los parámetros, y de esa forma capturar los distintos posibles comportamientos de la población.

Al igual que en el modelo Gamma-Exponencial, asumiremos que el parámetro de escala λ está distribuido $Gamma(\alpha, r)$ en la población. La probabilidad de ocurrencia del consumo de los clientes se puede modelar por tanto de la siguiente forma:

$$\begin{aligned} F(t) = P(T \leq t) &= -\int_0^\infty \frac{(1 - e^{-\lambda t^c})\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} \, d\lambda \\ &= 1 - \left(\frac{\alpha}{\alpha + t^c}\right)^r \end{aligned}$$

1.3. Modelos de conteo

Permiten modelar cuantas veces los consumidores incurrirán en un comportamiento determinado en un período de tiempo (ejemplo: problema exposición publicitaria).

Algunas medidas de efectividad son:

- **Alcance:** Proporción de la población expuesta al evento al menos una vez durante el período: $1 - P(X_t = 0)$
- **Frecuencia promedio:** número promedio de exposiciones en el período entre aquellos que han experimentado el evento (por ejemplo, ver la valla publicitaria)

$$\frac{\mathbb{E}(X_t)}{1 - P(X_t = 0)}$$

- **Puntos de rating brutos (GRPs):** número promedio de exposiciones por cada 100 personas.

$$100 \cdot \mathbb{E}(X_t)$$

El fenómeno que se quiere estudiar es el número de veces que cada individuo ve la valla publicitaria. Para ello, se define el modelo individual *Poisson*

$$P(N_t = m|\lambda) = \frac{(\lambda t)^m e^{-\lambda t}}{m!} \tag{1.2}$$

lo cual corresponde a la probabilidad de que el número de exposiciones sea m en un intervalo de largo t.

Al igual que en los modelos anteriores, es posible incluir heterogeneidad asumiendo que el parámetro λ distribuye de acuerdo a una determinada distribución. Suponiendo que dicha distribución es *Gamma*

$$g(\lambda|\alpha, r) = \frac{\alpha^r \lambda^{r-1} e^{\alpha\lambda}}{\Gamma(r)} \tag{1.3}$$

Usando el modelo individual en 1.2 y la distribución en 1.3, se puede estimar la probabilidad de un número de exposiciones

$$\begin{aligned} P(N_t = m) &= \int_0^\infty P(N_t = m|\lambda) g(\lambda) d\lambda \\ &= \int_0^\infty \frac{(\lambda t)^m e^{-\lambda t}}{m} \cdot \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} d\lambda \\ &= \left(\frac{\alpha}{\alpha + t}\right)^r \cdot \left(\frac{t}{\alpha + t}\right)^m \cdot \frac{\Gamma(r + m)}{\Gamma(r) m!} \end{aligned} \tag{1.4}$$

1.4. Modelos de elección

Permiten modelar la probabilidad de que los individuos elijan un determinado comportamiendo, como por ejemplo, compra en una visita a una tienda, respuesta a una campaña de marketing directo, uso de un producto, etc.

Consideremos como variable de interés la probabilidad de que un individuo perteneciente a un segmento responda positivamente a una campaña de marketing. En el enfoque tradicional, se realiza una segmentación de clientes en grupos homogéneos, se envía mensajes a muestras aleatorias de cada segmento y se implementa un campaña en segmentos con tasa de respuesta (TR) sobre cierto corte, por ejemplo, $TR > \frac{\text{Costo de envio}}{\text{Margen unitario}}$.

Sin embargo, es posible incorporar un enfoque de modelos probabilísticos de manera de abordar el problema. Si se considera la probabilidad de responder de manera positiva que tiene un segmento s en particular, p_s, es posible intepretar de manera sensilla la cantidas de respuestas obtenidas. Recordando que, la suma de experimentos de Bernoulli corresponde a una variable aleatoria Binomial, es posible interpretar X_s, la cantidad de respuestas obtenidas de un total de m_s enviadas, como una variable aleatoria $Bin(m_s, p_s)$, luego

$$P(X_s = x_s | m_s, p_s) = \binom{m_s}{x_s} p_s^{x_s} (1 - p_s)^{m_s - x_s} \tag{1.5}$$

donde m_s es la población del segmento s y p_s es la probabilidad de respuesta del segmento s.

Luego se introduce heterogeneidad a través de la distribución $B(\alpha, \beta)$:

$$\begin{aligned} P(X_s = x_s) &= \int_0^1 P(X_s = x_s | m_s, p_s) \cdot g(p_s | \alpha, \beta) dp_s \\ &= \int_0^1 \binom{m_s}{x_s} p_s^{x_s} (1 - p_s)^{m_s - x_s} \cdot \frac{p_s^{\alpha - 1} (1 - p_s)^{\beta - 1}}{B(\alpha, \beta)} dp_s \\ &= \binom{m_s}{x_s} \frac{B(\alpha + x, \beta + m_s - x_s)}{B(\alpha, \beta)} \end{aligned} \tag{1.6}$$

1.5. Esperanzas Condicionales

Permiten tomar decisiones a nivel desagregado. Por ejemplo, en el modelo de elección que la $P(X_s = x_s)$ quedaba definida en 1.6. Una pregunta válida que nos podríamos hacer es cúal es la tasa de respuesta de un segmento s determinado. Intuitivamente debería estar entre la tasa de respuesta esperada de la población y la observada, es decir,

$$\mathbb{E}(\theta_s | m_s, x_s) = \gamma \frac{\alpha}{\alpha + \beta} + (1 - \gamma) \frac{x_s}{m_s} \tag{1.7}$$

donde $\mathbb{E}(Beta(\alpha, \beta)) = \frac{\alpha}{\alpha + \beta}$.

Recordemos que la distribución de θ condicionado a un número de respuestas recibidas, por Bayes, es

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta} \tag{1.8}$$

donde $g(\theta)$ es la distribución del parámetro, definida a priori, y $f(x|\theta)$ es la distribución de la probabilidad de la data dado los parámetros, es decir, la función de verosimilitud. De esto se deduce

$$g(\theta_s|x_s) \sim B(\alpha + x_s, \beta + m_s - x_s) \tag{1.9}$$

Teniendo clara la distribución condicionada es fácil deducir la esperanza

$$\begin{aligned}
\mathbb{E}(\theta_s|x_s) &= \frac{\alpha + x_s}{\alpha + \beta + m_s} \\
&= \frac{\alpha}{\alpha + \beta + m_s} + \frac{x_s}{\alpha + \beta + m_s} \\
&= \frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha + \beta}{\alpha + \beta + m_s} + \frac{x_s}{m_s} \cdot \frac{m_s}{\alpha + \beta + m_s} \\
&= \gamma \frac{\alpha}{\alpha + \beta} + (1 - \gamma)\frac{x_s}{m_s}
\end{aligned} \tag{1.10}$$

Esta última igualdad se encuentra al hacer el reemplazo $\gamma = \frac{\alpha+\beta}{\alpha+\beta+m_s}$, la cual coincide con 1.7, resultado que esperábamos encontrar.

Se podría replantear la regla de decisión, y enviar catálogos a los segmentos s tales que

$$\mathbb{E}(\theta_s|x_s) = \frac{\alpha + x_s}{\alpha + \beta + m_s} > \frac{\text{costo de envio}}{\text{margen unitario}}$$

1.6. Variables explicativas

1.6.1. Variables explicativas en modelos de duración en tiempo continuo sin dependencia de la duración

En secciones anteriores, hemos expuesto modelos que intentan explicar y predecir el tiempo en que los individuos realizarán una determinada acción (e.g: tiempo de prueba de un producto), considerando que el comportamiento de los agentes se debe netamente a factores aleatorios. En esta sección se incorporará heterogeneidad observable a un modelo de duración en tiempo continuo sin dependencia en la duración. Entendemos por heterogeneidad observable, aquellos factores observables (que están en los datos) intrínsecos a los individuos que los hacen distintos y, por ejemplo: sexo, edad, entre otras.

Sea T_i la variable aleatoria que describe el instante en que el individuo i realiza una determinada acción. Modelaremos dicha variable aleatoria con una distribución exponencial de parámetro λ_i:

$$\mathbb{P}(T_i < t_i|\lambda_i) = 1 - e^{-\lambda_i t_i}$$

Cabe destacar que, dada la naturaleza de los datos, el comportamiento descrito se realizará de manera desagregada (dependencia de i en el parámetro), es decir, dado que existe información

individual para cada individuo, es posible estimar el parámetro de cada uno de éstos (no así en los casos agregados vistos anteriormente).

Sea x_i el vector que contiene las variables explicativas pertinentes del individuo i. Modelaremos la tasa de llegada de i, de la manera siguiente:

$$\lambda_i = \exp(\beta_0 + \beta' x_i) = \lambda_0 \exp(\beta' x_i)$$

Donde β corresponde al vector de coeficientes asociados a las variables explicativas en cuestión.

La inclusión de la exponencial se debe a que, por a razones de convergencia e interpretación, la tasa de respuesta individual debe ser positiva. De esta forma, se puede capturar el efecto marginal de las variables demográficas sin restricción de signos, esto es, será posible obtener valores de β negativos.

Modelo sin Heterogeneidad no observable

La probabilidad que un individuo i realice un evento determinado antes del tiempo t_i, incluyendo su información observable, es:

$$\begin{aligned}\mathbb{P}(T_i < t_i|\beta, \lambda_0) &= 1 - e^{-\lambda_i t_i} \\ &= 1 - e^{-\lambda_0 \exp(\beta' x_i) t_i}\end{aligned}$$

Con lo cual (considerando instantes de tiempo t_i^- y t_i^+ para discretizar el tiempo, un panel de N individuos y un vector de parámetros $\theta = (\beta, \lambda_0)$), la log verosimilitud del problema resulta:

$$\begin{aligned}LL(\theta) &= \sum_{i=1}^{N} \ln(\mathbb{P}(t_i^- < T_i < t_i^+|\beta, \lambda_0)) \\ &= \sum_{i=1}^{N} \ln((\mathbb{P}(T_i < t_i^+|\beta, \lambda_0) - \mathbb{P}(T_i < t_i^-|\beta, \lambda_0)) \\ &= \sum_{i=1}^{N} \ln\left((1 - e^{-\lambda_0 \exp(\beta' x_i) t_i^+}) - (1 - e^{-\lambda_0 \exp(\beta' x_i) t_i^-})\right) \\ &= \sum_{i=1}^{N} \ln\left(e^{-\lambda_0 \exp(\beta' x_i) t_i^-} - e^{-\lambda_0 \exp(\beta' x_i) t_i^+}\right)\end{aligned}$$

Modelo con Heterogeneidad no observable

Para introducir heterogeneidad no observable en el modelo, se dejará el parámetro λ_0 distribuyendo de manera continua en la población según una ley $\Gamma(\alpha, r)$, pues de esta forma, es posible mezclar tanto la heterogeneidad no observable, como la observable.

De este modo, La probabilidad que un individuo i realice un evento determinado antes del tiempo t_i es:

$$\begin{aligned}\mathbb{P}(T_i < t_i|\alpha, r, \beta) &= \int_0^\infty \mathbb{P}(T_i < t_i|\alpha, r, \beta, \lambda_0) f(\lambda_0) d\lambda_0 \\ &= \int_0^\infty (1 - e^{-\lambda_0 \exp(\beta' x_i) t_i}) \left(\frac{\alpha^r \lambda_0^{r-1} e^{-\alpha\lambda_0}}{\Gamma(r)}\right) d\lambda_0 \\ &= 1 - \frac{\alpha^r}{\Gamma(r)} \int_0^\infty \lambda_0^{r-1} e^{-\lambda_0(\alpha + \exp(\beta' z_i) t_i)} d\lambda_0\end{aligned}$$

Luego, basta con multiplicar y dividir por $(\alpha + \exp(\beta' z_i) t_i)^r$ Para obtener como resultado:

$$\begin{aligned}\mathbb{P}(T_i < t_i|\alpha, r, \beta) &= 1 - \left(\frac{\alpha}{\alpha + \exp(\beta' z_i) t_i}\right)^r \int_0^\infty f(\lambda_0|\alpha + \exp(\beta' z_i) t_i, r) d\lambda_0 \\ &= 1 - \left(\frac{\alpha}{\alpha + \exp(\beta' z_i) t_i}\right)^r\end{aligned}$$

Finalmente, la función de log verosimilitud resulta, con $\theta = (\beta, \alpha, r)$:

$$\begin{aligned}LL(\theta) &= \sum_{i=1}^N \ln(\mathbb{P}(t_i^- < T_i < t_i^+|\alpha, r, \beta)) \\ &= \sum_{i=1}^N \ln\left(\left(\frac{\alpha}{\alpha + \exp(\beta' z_i) t_i^-}\right)^r - \left(\frac{\alpha}{\alpha + \exp(\beta' z_i) t_i^+}\right)^r\right)\end{aligned}$$

1.6.2. Variables explicativas en modelos de duración en tiempo continuo con dependencia de la duración

Cuando el tiempo en que ocurre un determinado suceso posee dependencia en la duración, el procedimiento es análogo que en el caso sin dicha dependencia, pero considerando que T_i distribuye según una ley Weibull.

$$\mathbb{P}(T_i < t_i|\lambda_i, c) = 1 - e^{-\lambda_i t_i^c}$$

Modelo sin Heterogeneidad no observable

La probabilidad que un individuo i realize un evento determinado antes del tiempo t_i, incluyendo su información observable, es:

$$\begin{aligned}\mathbb{P}(T_i < t_i|\beta, \lambda_0, c) &= 1 - e^{-\lambda_i t_i} \\ &= 1 - e^{-\lambda_0 \exp(\beta' x_i) t_i^c}\end{aligned}$$

Por lo que, la fucnión de log verosimilitud toma la siguiente forma:

$$\begin{aligned}
LL(\theta) &= \sum_{i=1}^{N} \ln(\mathbb{P}(t_i^- < T_i < t_i^+ | \beta, \lambda_0, c)) \\
&= \sum_{i=1}^{N} \ln((\mathbb{P}(T_i < t_i^+ | \beta, \lambda_0, c) - \mathbb{P}(T_i < t_i^- | \beta, \lambda_0, c)) \\
&= \sum_{i=1}^{N} \ln\left((1 - e^{-\lambda_0 \exp(\beta' x_i)(t_i^+)^c}) - (1 - e^{-\lambda_0 \exp(\beta' x_i)(t_i^-)^c})\right) \\
&= \sum_{i=1}^{N} \ln\left(e^{-\lambda_0 \exp(\beta' x_i)(t_i^-)^c} - e^{-\lambda_0 \exp(\beta' x_i)(t_i^+)^c}\right)
\end{aligned}$$

Modelo con Heterogeneidad no observable

De manera análoga al caso anterior, se introduce heterogeneidad no observable mediante el parámetro λ_0 según una distribución $\Gamma(\alpha, r)$. Luego:

$$\begin{aligned}
\mathbb{P}(T_i < t_i | \beta, \alpha, r, c) &= \int_0^\infty \mathbb{P}(T_i < t_i | \beta, \lambda_0, c) f(\lambda_0 | \alpha, r) d\lambda_0 \\
&= \int_0^\infty \left(1 - e^{-\lambda_0 \exp(\beta' z_i) t_i^c}\right) \frac{\alpha^r \lambda_0^{r-1} e^{-\alpha\lambda_0}}{\Gamma(r)} d\lambda_0 \\
&= 1 - \frac{\alpha^r}{\Gamma(r)} \int_0^\infty \lambda_0^{r-1} e^{-\lambda_0(\alpha + \exp(\beta' z_i) t_i^c)} d\lambda_0
\end{aligned}$$

Luego, basta con multiplicar y dividir por $(\alpha + \exp(\beta' z_i) t_i^c)^r$ Para obtener como resultado:

$$\mathbb{P}(T_i < t_i | \beta, \alpha, r, c) = 1 - \left(\frac{\alpha}{\alpha + \exp(\beta' z_i) t_i^c}\right)^r$$

Notar que, cuando $\beta = 0$ se obtiene el modelo Gamma-Weibull usual.

De esta forma, la log verosimilitud es:

$$\begin{aligned}
LL(\theta) &= \sum_{i=1}^{N} \ln(\mathbb{P}(t_i^- < T_i < t_i^+ | \beta, \alpha, r, c) \\
&= \sum_{i=1}^{N} \ln\left(\left(\frac{\alpha}{\alpha + \exp(\beta' z_i)(t_i^-)^c}\right)^r - \left(\frac{\alpha}{\alpha + \exp(\beta' z_i)(t_i^+)^c}\right)^r\right)
\end{aligned}$$

1.6.3. Caso Modelo de Conteo: KhakiChinos.com

Khaki Chinos, INC, es una compañía de ventas de ropa por catálogo con presencia en internet. La empresa posee información respecto al comportamiento de compras de los clientes registrados en su página web, sin embargo, desconoce los patrones de visita de los usuarios en general.

Para estudiar el patrón de visitas, la empresa compró un panel de $N = 2728$ usuarios de internet con al menos una visita a la tienda de ropa. El set de datos entrega el número de visitas y_i de cada individuo i y las siguientes variables demográficas agrupadas en el vector x_i:

- $\log(I_i)$ representando el logaritmo del ingreso del individuo i.
- La variable binaria G_i igual a 0 si el individuo i es mujer y 1 si no.
- $\log(E_i)$ representando el logaritmo de la edad del individuo i.
- El tamaño del hogar del individuo i, dado por S_i.

Modelo de Regresión de Poisson

Se puede describir el número de de visitas de cada individuo mediante una distribución de Poisson. Para esto, sea Y_i la variable aleatoria que cuenta el número de veces que el individuo i visita el sitio en una unidad de tiempo.

A nivel individual, se asume que Y_i se distribuye Poisson con media λ_i:

$$\mathbb{P}(Y_i = y_i|\lambda_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}$$

Notar que, a diferencia de los modelos anteriormente planteados, en este caso se cuenta con data desagregada.

Para incluir variables demográficas, se puede asumir que estas ayudan a explicar las medias individuales λ_i:

$$\lambda_i = \lambda_0 exp(\beta' x_i)$$

Con esto, el Modelo de Regresión de Poisson toma la siguiente forma:

$$\mathbb{P}(Y_i = y_i|\lambda_0, \beta) = \frac{(\lambda_0 e^{\beta' x_i})^y e^{-\lambda_0 e^{\beta' x_i}}}{y!}$$

Siendo la log-verosimilitud a maximizar:

$$LL(\theta) = \sum_{i=1}^{N} \ln(\mathbb{P}(Y_i = y|\lambda_i))$$

Donde $\theta = (\lambda_0, \beta)$ corresponde a los parámetros a estimar.

Modelo de Regresión NBD

Una transición natural es agregar heterogeneidad no observable al modelo anteriormente descrito. Para esto, se asumirá que a nivel individual $Y_i \sim Poisson(\lambda_i)$ y $\lambda_0 \sim \Gamma(\alpha, r)$. De esta forma, se mantiene la heterogeneidad observable dada por las variables demográficas y, adicionalmente, se agrega una componente no observable distribuyendo de manera continua a lo largo de la población, que incorpora efectos aleatorios.

Así:

$$\mathbb{P}(Y_i = y_i|\alpha, r) = \int_0^\infty \mathbb{P}(Y_i = y_i|\lambda_0) f(\lambda_0|\alpha, r) d\lambda_0$$

Desarrollando el termino al interior de la integral:

$$\begin{aligned}\mathbb{P}(Y_i = y_i|\lambda_0) f(\lambda_0) &= \left(\frac{(\lambda_0 e^{\beta' x_i})^{y_i} e^{-\lambda_0 e^{\beta' x_i}}}{y_i!}\right)\left(\frac{\alpha^r \lambda_0^{r-1} e^{-\alpha\lambda_0}}{\Gamma(r)}\right) \\ &= \left(\frac{\alpha^r (e^{\beta' x_i})^{y_i}}{y_i!\Gamma(r)}\right) \lambda_0^{r+y_i-1} e^{-\lambda_0(\alpha+e^{\beta' x_i})}\end{aligned}$$

Reconociendo términos, es fácil ver que para recuperar una densidad Gamma, es necesario multiplicar y dividir por $\frac{(\alpha+e^{\beta' x_i})^{r+y_i}}{\Gamma(r+y_i)}$

$$\begin{aligned}\mathbb{P}(Y_i = y_i|\lambda_0) f(\lambda_0) &= \left(\frac{\alpha^r \Gamma(r+y)(e^{\beta' x_i})^{y_i}}{y_i!\Gamma(r)(\alpha+e^{\beta' x_i})^{r+y_i}}\right) \frac{(\alpha+e^{\beta' x_i})^{r+y_i} \lambda_0^{r+y_i-1} e^{-\lambda_0(\alpha+e^{\beta' x_i})}}{\Gamma(r+y_i)} \\ &= \frac{\Gamma(r+y)}{\Gamma(r) y_i!}\left(\frac{\alpha}{\alpha+e^{\beta' x_i}}\right)^r \left(\frac{e^{\beta' x_i}}{\alpha+e^{\beta' x_i}}\right)^{y_i} f(\lambda_0|r+y_i, \alpha+e^{\beta' x_i})\end{aligned}$$

Finalmente, al integrar sobre todos los valores de λ_0 se obtiene:

$$\mathbb{P}(Y_i = y_i|\alpha, r) = \frac{\Gamma(r+y)}{\Gamma(r) y_i!}\left(\frac{\alpha}{\alpha+e^{\beta' x_i}}\right)^r \left(\frac{e^{\beta' x_i}}{\alpha+e^{\beta' x_i}}\right)^{y_i}$$

Notar que, cuando $\beta = 0$ se recupera el modelo NBD tradicional.

1.7. Modelos integrados

Permiten modelar fenómenos complejos que incorporan más de uno de los modelos básicos planteados anteriormente.

Supongamos el caso de itemes no reportados. Supondremos que la cantidad de items comprados sigue una distribución de *Poisson* y la elección para escoger cuántos items declarar sigue una distribución *Binomial*. La heterogeneidad se incluye con una distribución *Gamma* para la tasa del modelo de conteo y con una distribución *Beta* para la probabilidad de declaración. Entonces:

$$
\begin{aligned}
P(X=k) &= \sum_{n=0}^{\infty} P(X=k|N=n)\cdot P(N=n) \\
&= \sum_{n=0}^{\infty} \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\cdot\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}d\theta\cdot\int_0^{\infty}\frac{\lambda^n e^{-n}}{n!}\cdot\frac{\alpha^r\lambda^{r-1}e^{\alpha\lambda}}{\Gamma(r)}d\lambda \\
&= \frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+1}\right)\frac{\Gamma(a+x)}{\Gamma(a)}\cdot\frac{\Gamma(a+b)}{\Gamma(a+b+x)}\cdot{}_2F_1\left(r+x,b;a+b+x;\frac{1}{\alpha+1}\right) \qquad (1.11)
\end{aligned}
$$

(1.12)

El último termino es la función *Hypergeométrica Gaussiana*. Esta función queda definida como

$$
{}_2F_1 \equiv \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{j=0}^{\infty}\frac{\Gamma(a+j)\Gamma(b+j)z^j}{\Gamma(c+j)j!} \qquad (1.13)
$$

Como su cálculo puede ser complicado, puede usarse la siguiente recursión:

$$
{}_2F_1\left(a,b;c;z\right) = \sum_{j=0}^{\infty}u_j \approx \sum_{j=0}^{M}u_j \qquad (1.14)
$$

donde

$$
\begin{aligned}
u_0 &= 1 \\
\frac{u_j}{u_{j-1}} &= \frac{(a+j-1)(b+j-1)}{(c+j-1)j}z \qquad \forall j \geq 1
\end{aligned}
$$

1.8. Customer lifetime value caso contractual

Database Marketing posee dos elementos escenciales, tiempo de permanencia con la firma e intensidad de compra mientras el cliente está en la firma. Para el caso determinista se define *Life Time Value (CLV)* como

$$
CLV = \sum_{t=0}^{T} m\cdot\frac{r^t}{(1+d)^t} \qquad (1.15)
$$

donde m es el flujo neto por período (si el cliente está activo); r es la tasa de retención; d es la tasa de descuento; y T es el horizonte de evaluación.

Para el caso estocástico, sean $\mathbb{E}(v(t))$ valor esperado de los flujos del cliente en el instante t (asumiendo que está activo); $S(t)$ la probabilidad que el cliente siga activo en el instante t; y $d(t)$ el factor de descuento que refleja el valor presente del dinero recibido en el instante t. El cálculo de CLV es

$$
\mathbb{E}(CLV) = \int_0^{\infty} E(v(t))S(t)d(t)dt \qquad (1.16)
$$

Esta definición es inútil a menos que operacionalicemos $\mathbb{E}(v(t))$, $S(t)$ y $d(t)$ para la situación particular.

Es importante distinguir entre situaciones contractuales y no contractuales:

- **Contractual**: Observamos cuando un cliente deja de estar activo. Ejemplo: suscripción a una revista, plan VTR, etc.
- **No Contractual**: No observamos cuando un cliente deja de estarlo.

El desafío de lo mercados contractuales: ¿Cómo diferenciamos aquellos clientes que han terminado su relación con la firma, de aquellos que simplemente están en un largo período de inactividad?

También se debe distinguir según la oportunidad de hacer la transacción:

- **Discreta**: La acción puede realizarse en un número discreto de ocasiones.
- **Continua**: La acción puede realizarse en cualquier momento. Ejemplo, transacción con una tarjeta de crédito.

1.8.1. Modelo contractual a tiempo discreto

En el mercado de las revistas, típicamente el 30 % renueva al final de su primera subscripción, pero ese número salta al 50 % para la segunda renovación y hasta el 75 % para subscriptores de mayor antigüedad (Fielding, Michael (2005), "Get Circulation Going: DM Redesign Increases Renewal Rates for Magazines", Marketing News, September 1, 9-10).

Al evaluar las tasas de retención de una base de clientes es necesario considerar las diferencias entre los cohortes y proyectar los comportamientos más allá de los que observamos.

Explicaciones alternativas (y complementarias) para el incremento de las tasas de retención: Dinámicas a nivel individual (incremento de lealtad) y un cambio en la mezcla de la composición de la población.

Ejemplo: Supongamos que analizamos un cohorte de 10.000 clientes que en promedio gastan $100 por período y que corresponden a dos tipos de clientes:

- **Segmento 1**: Un tercio de los clientes tiene una tasa de retención (constante en el tiempo) de 0.9
- **Segmento 2**: Dos tercios de los clientes tienen una tasa de retención anual de 0.5

Año	# Clientes activos			Tasa de retención		
	Seg-1	Seg-2	Total	Seg-1	Seg-2	Total
1	3.333	6.667	10			
2	3	3.334	6.334	0.9	0.5	0.633
3	2.7	1.667	4.367	0.9	0.5	0.689
4	2.43	0.834	3.264	0.9	0.5	0.747
5	2.187	0.417	2.604	0.9	0.5	0.798

Cuadro 1.1: Rol de la heterogeneidad

En el Cuadro 1.1 la tasa de retención agregada (en rojo en la tabla) es decreciente aún cuando a nivel individual las retenciones son constantes en el tiempo.

El valor residual de un cliente activo del cohorte, si pertenece al segmento 1 es

$$\mathbb{E}(RLV) = \sum_{t=1}^{\infty} \$100 \cdot \frac{0.9^t}{(1+0.1)^{t-1}} = \$495 \tag{1.17}$$

Si el cliente pertenece al segmento 2:

$$\mathbb{E}(RLV) = \sum_{t=1}^{\infty} \$100 \cdot \frac{0.5^t}{(1+0.1)^{t-1}} = \$92 \tag{1.18}$$

Sin embargo, la regla de Bayes nos permite mostrar que, condicional en estar activo, un cliente es más probable que tenga una alta tasa de retención.

$$\begin{aligned} P(\text{seg-1}|\text{renovar 4 veces}) &= \frac{P(\text{renovar 4 veces}|\text{seg-1})P(\text{seg-1})}{P(\text{renovar 4 veces})} \\ &= \frac{0.9^4 \cdot 0.333}{0.9^4 \cdot 0.333 + 0.5^4 \cdot 0.6667} \\ &= 0.84 \end{aligned}$$

Luego, el *Lifetime Value Residual* viene dado por:

$$\mathbb{E}(RLV) = 0.84 \cdot \$495 + (1 - 0.84) \cdot \$92 = \$430$$

En mercados contractuales, ¿cuánto perdemos si no consideramos la heterogeneidad? Veamos el ejemplo que hemos usado. Si tomamos en cuenta la tasa de retención agregada, el valor de la base de clientes es $4.945.049. En cambio, si distinguimos por segmentos, este valor asciende a $7.940.992.

En estudios con bases de datos reales muestran que el error en el CLV se eleva hasta el 50 %. El impacto sobre el CLV de aumentar las tasas de retención de hasta un 50 %.

Para calcular CLV tenemos que hacerlo condicional en la duración. Veamos primero el caso continuo.

Se postulan los siguientes supuestos:

1. La tasa de retención a nivel individual es $1 - \theta$

$$S(t|\theta) = (1-\theta)^t, \qquad t = 1, 2, 3, ... \tag{1.19}$$

2. La heterogeneidad en θ es capturada por una distribución *Beta*

$$f(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}}{(1-\theta)^{\beta-1}} B(\alpha, \beta) \tag{1.20}$$

La probabilidad de que el cliente siga activo en t

$$S(t|\alpha, \beta) = \int_0^1 S(t|\theta) f(\theta|\alpha, \beta) d\theta \qquad = \frac{B(\alpha, \beta + t)}{B(\alpha, \beta)}, \qquad t = 1, 2, 3, ...$$

Consideremos un cliente que ha estado activo por n períodos

$$\begin{aligned}\mathbb{E}(RLV(d|\text{activo } n \text{ períodos})) &= \sum_{t=n}^{\infty} E(v(t)) \cdot \frac{S(t|t>n-1)}{(1+d)^{t-n}} \\ &= \bar{v} \sum_{t=n}^{\infty} \frac{S(t|t>n-1)}{(1+d)^{t-n}} \qquad \text{Asumiendo flujos constantes} \end{aligned} \tag{1.21}$$

DERL es el valor esperado residual del cliente (condicional en la antigüedad). Para el caso de la distribución geométrica desplazada

$$\begin{aligned} DERL(d|\theta, \text{activo } n \text{ períodos}) &= \sum_{t=n}^{\infty} \frac{S(t)}{S(n-1)} \cdot \left(\frac{1}{1+d}\right)^{t-n} \\ &= \frac{(1-\theta)(1+d)}{d+\theta} \end{aligned} \tag{1.22}$$

Cuando la tasa de abandono θ no es observable, debemos encontrar la distribución de esta variable en la población. Para ello usamos la regla de Bayes para calcular la distribución posterior condicional en la antigüedad.

$$\begin{aligned} DERL(d|\alpha, \beta, \text{activo } n \text{ perídos}) &= \int_0^1 \frac{(1-\theta)(1+d)}{d+\theta} \cdot \frac{S(n-1|\theta) f(\theta|\alpha,\beta)}{S(n-1|\alpha,\beta)} d\theta \\ &= \left(\frac{\beta+n+1}{\alpha+\beta+n-1}\right) \cdot {}_2F_1\left(1, \beta+n; \alpha+\beta+n; \frac{1}{1+d}\right) \end{aligned} \tag{1.23}$$

1.8.2. Modelo contractual a tiempo continuo

Algunos supuestos son:

1. La duración de la relación de un cliente con la firma está caracterizada por una distribución exponencial, con densidad y función de sobrevivencia dadas por:

$$f(t|\lambda) = \lambda e^{-\lambda t} \tag{1.24}$$

$$S(t|\lambda) = e^{-\lambda t} \tag{1.25}$$

2. La heterogeneidad en λ es capturada por una distribución $Gamma$

$$g(\lambda|r,\alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} \tag{1.26}$$

Entonces, la probabilidad de seguir activo en t es

$$\begin{aligned} S(t|r,\alpha) &= \int_0^{\infty} S(t|\lambda) g(\lambda|r,\alpha) d\lambda \\ &= \left(\frac{\alpha}{\alpha+t}\right) \end{aligned} \tag{1.27}$$

El valor esperado de *Residual Lifetime Value*

$$\begin{aligned}\mathbb{E}(RLV(\delta|\text{activo en } s)) &= \int_0^\infty E(v(t))S(t|t>s)\delta(t)dt \\ &= \bar{v}\cdot DERL(\delta|r,\alpha,\text{activo e } s)\end{aligned} \tag{1.28}$$

donde

$$DERL(\delta|r\alpha,\text{activo en } s) = (\alpha+s)^r\delta\Psi(r;r;(\alpha+s)\delta) \tag{1.29}$$

Ψ es la *función hiper geométrica confluyente del segundo tipo*

Parte II

Modelos Estructurales

Capítulo 2

Introducción a Modelos Estructurales

2.1. Introducción

En esencia, un modelo econométrico *estructural* es aquel que deriva relaciones estimables estadísticamente a partir de supuestos bien definidos de comportamiento de los agentes que deciden respecto a las cantidades observables. En contraposición a los modelos estructurales están los modelos de *forma reducida* donde los modelos simplemente describen la variabilidad de alguna medida de interés en base a un conjunto de variables observables exógenas.

La disciplina económica suele llamar modelo estructurales a los resultantes de asumir que los consumidores maximizan una utilidad subyacente y que las firmas maximizan su rentabilidad esperada. Desde el marketing, consideramos también en la definición en aquellos que postulan hipotesis alternativas de comportamiento incluyendo así una variedad de teorías de comportamiento que nutren la disciplina tales como teoría de prospectos, contabilidad mental, elección sobre conjuntos de consideración, etc. Como discutiremos más adelante, no existe un modelo estructural puro y la línea que los separa de los modelos de forma reducida es ciertamente difusa. Incluiremos en nuestra discusión de modelos estructurales a cualquiera que considere alguna historia de comportamiento que permita añadir interpretabilidad a los parametros del modelo.

Ejemplo 1: Supogamos que un analista busca estudiar como el precio en la región i (p_i) se ve afectado por la presencia o no de competencia. Si además de los precios observamos la cantidad de clientes en la region (POP_i), el ingreso per capita en la región (INC_i) y un una indicatriz CMP_i que toma el valor 1 si en la región correspondiente presenta competencia (0 en caso contrario). Entonces, un modelo de forma reducida sencillo para estudiar el problema viene dado por:

$$p_i = \beta_0 + \beta_1 POP_i + +\beta_2 INC_i + \beta_3 CMP_i + \varepsilon_i$$

Bajo este enfoque, podemos usar técnicas de regresión tradicionales para estimar β_3 que en principio indicaría el impacto de la competencia en el nivel de precios. Sin embargo, la presencia de competencia en un determinado mercado depende también del nivel de precios. Si los precios en una región son altos, la rentabilidad esperada por entrar tambien es alta motivando a potenciales competidores a participar. En consecuencia, un modelo como el planteado podría subestimar el efecto de la competencia.

Un modelo estructural buscaría derivar relaciones estimables a partir de supuestos básicos del comportamiento de la firma. Por ejemplo, podriamos asumir que cada firma decide conjuntamente la entrada/salida de un mercado y los precios a cobrar de modo de maximizar la rentabilidad

esperada. □

Ejemplo 2: Supongamos buscamos describir la productividad de los miembros de la fuerza de venta medida como número de unidades vendidad q.

$$q = f(X, \beta) + \varepsilon$$

La especificación del termino de error ε puede por si solo permitirnos dar una interpretación estructural a los estimadores. Si simplemente asumimos un error normalmente distribuido, entonces corresponderá simplemente a un ruido blanco y la regresión simplemente nos indicará a través de los parámetros β como las variables X en promedio afectan las ventas q. Por el contrario, si asumimos que el termino ε considera además del ruido una componente no observable positiva asociada a la breacha de productividad de los miembros menos eficientes de la fuerza de venta, entonces la regresión describira la frontera eficiente de ventas. Esto puede hacerse por ejemplo especificando que $\varepsilon = \epsilon - \xi$ donde ϵ está normalmente distribuida centrada en cero, pero ξ proviene de una normal truncada en los numeros positivos (este enfoque se le suele llamar de regresión estocástica de frontera). □

El gran desafío de la aplicación de modelos econométricos a problemas comerciales es enriquecer el conocimiento respecto a cómo se comportan los agentes relevantes del negocio, para así tomar decisiones más consistentes y más rentables. Desde este punto de vista, apuntamos a modelos que describan la lógica que determina el comportamiento de los clientes y firmas más allá de simples correlaciones estadísticas entre las variables observables. En general, son varias las ventajas de usar modelos estructurales por sobre modelos de forma reducida:

1. *La capacidad de contar una mejor historia del comportamiento de los agentes.* Esto se expresa por la capacidad de interpretación directa a los parámetros del modelo. Mientras los parámetros asociados a enfoques de regresión tradicionales típicamente nos indican la magnitud en que en promedio varia alguna magnitud de interés ante variaciones de otra, los parámetros de un modelo estructural nos indican entre otros la valoracion relativa de un atributo en la funcion de utilidad, los precios de referencia de un producto o la aversión al riesgo de un tomador de decisión. La provisión de una historia de comportamiento más completa no se deriva exclusivamente de la interpretación directa de los parámetros del modelo si no que tambien de la capacidad de derivar métricas complementarias tales como elasticidades y excedentes de consumidores. Más aún, podemos proyectar el comportamiento para calcular probabilidades y frecuencias de compra, participaciones de mercado, etc.

2. *La generación de estimaciones consistentes con las expectativas de los analistas.* Frecuentemente, al analizar los datos queremos dejar la mayor libertad posible al modelo para *dejar que la data hable*. Este enfoque puede tener valor y ser recomendable en estudios exploratorios, pero para tomar decisiones necestiamos estimaciones robustas y usar tanta información como sea posible. Las teorías usadas para derivar modelos econométricos estructurales suelen estar soportadas tanto por estudios experimentales como por amplia evidencia empírica en múltiples dominios. Por lo tanto, al incorporar teoria estamos implicitamente usando información que ha demostrado consistentemente su validez.

   Ejemplo 3: Supongamos que queremos proponer un modelo que describa la participación de mercado de las distintas marcas en una industria. Si usamos un enfoque de regresión en que simplemente disponemos los shares al lado izquierdo y una forma funcional flexible al lado derecho, el modelo resultante podría predecir participaciones fuera del rango [0,1],

que dificilmente pueden justificarse. Por el contrario si adscribimos al axioma de eleccion de Luce (1959) que indica que la probabilidad de eleccion en un determinado conjunto depende del ratio entre una medida de atraccion de la alternativa con respecto al atractivo total del conjunto, forzamos a que las participaciones siempre esten en el rango deseado. □

Ejemplo 4: La teoría económica predice que en general, las cantidades demandadas decrecen ante aumentos en su precio. Sin embargo, en muchas situaciones prácticas la disponibilidad de datos al nivel de agregación requerido es limitado dificultando la estimación de esta relación inversa entre precio y demanda. En estas situaciones no es raro que un modelo flexible prediga que la demanda crece en función del precio. Agregar estructura nos permite limitar la busqueda solo entre aquellos modelos que son consistentes con la premisa que las demandas decrecen en el precio. □

3. *Evaluación de impacto de modificación de políticas* Una de las herramientas fundamentales de la función comercial es la generación de planes comerciales que buscan proponer un diseño del conjunto producto, plaza, precio y promoción que genere el mayor valor para el cliente y la captura del mayor excedente por parte de la firma. El rol de los modelos econométricos es estudiar el impacto que tendrían distintas estrategias en el comportamiento del consumidor. En esencia, plan de marketing propone un cambio en las reglas del juego que han generado la data que observamos y por tanto necesitamos apuntar a estimar los elementos más básicos del comportamiento que se mantendran invariantes ante modificacion de productos, precios, canales de distribución, etc. En este grupo tenemos, valoraciones por atributos de productos, costo de transporte, aversión al riesgo, entre otros, que no pueden ser estimados a menos que derivemos el modelo a partir de teorías indivduales de comportamiento. En otras palabras, la derivación de modelos de demanda a partir de teorias comportamiento nos permiten evaluar contrafactuales que apoyan el diseño de propuestas de valor efectivas.

   La necesidad de evaluar contrafactuales usando elementos fundamentales que no se vean afectados por cambios en los sistemas fue inicialmente discutido por Robert Lucas (1976) en la famosa crítica que lleva su nombre. En el contexto de la predicción de efectos macroeconomicos, Lucas postuló que cualquier cambio en las politicas variaran sistematicamente la estructura de los modelos y por tanto debemos apuntar a describir parametros profundos que gobiernan comportamiento individual.

   Ejemplo 5: Consideren un retailer que vende múltiples productos a traves de dos canales, las salas de venta tradicionales y un sitio web con despacho directo. El retailer esta evaluando la posibilidad de re-asignar el conjunto de productos que vende a traves de cada canal para aumentar la rentabilidad del negocio. Para apoyar esta decisión, parece evidente que el simple analisis de las ventas de cada producto en cada canal no nos ayudara a predecir como dichos productos se venderian en el otro canal o cómo se afectaría la venta si un producto deja de venderse en algunos de los canales. Para hacer este ejercicio necesariamente necesitaremos investigar primitivas más fundamentales del comportamiento como preferencias intrínsecas por canal para cada categoría y patrones de sustitucion entre las alternativas disponibles dentro del canal y con respecto al otro canal. □

   Ejemplo 6: En muchas industrias como la de vestuario de moda o de articulos tecnológicos, hay una alta variabilidad de la oferta con constante entradas y salidas de diferentes versiones de los productos dificultando la proyección del desempeño de cada variante en el

tiempo. Mientras el surtido de producto varía con frecuencia, hay parámetros de la demanda pueden perdurar por varias temporadas tales como la elasticidad al precio, crecimiento de la categoría, factores estacionales y de substitución/complementariedad de atributos. Un enfoque estructural apunta precisamente a la estimación de estos parámetros estables. □

4. *Testear aplicabilidad de teoría.* Al usar un enfoque estructural, nos forzamos a pensar detalladamente respecto al problema y explicitar cada una de los supuestos de comportamiento. Las especificaciones alternativas de modelos de forma reducida simplemente corresponden a formas funcionales diferentes y por tanto no son informativas respecto a lógica en que deciden los agentes. Por otra parte, dos modelos estructurales diferentes provienen de supuestos de comportamiento diferentes y por tanto cuando uno de ellos ajusta mejor a la data nos indica que hay una teoría de comportamiento es más plausible que la otra en el dominio de aplicación del modelo. Así, los modelos estructurales no solo se nutren de teoría sino que tambien ayudan a su desarrollo.

Las ventajas antes descritas no implican que siempre debieran preferirse modelos estructurales por sobre los de forma reducida. Como hemos descrito, los modelos de forma reducida suelen proveer suficiente flexibilidad para dejar que sea la data la que hable, lo que puede ser particularmente util en análisis exploratorios del caso bajo estudio. Además, muchas veces la inclusión de más estructura en el modelo implica rutinas de estimación más sofisticada siendo con frecuencia altamente intensivas computacionalmente.

Es importante destacar que no existe un modelo puramente estructural. Todo modelo requiere en algún momento suponer alguna forma funcional flexible sin fundamento teórico sólido. Por ejemplo, podemos asumir que los consumidores al elegir un producto están maximizando una utilidad subyacente, pero ¿cómo describimos dicha función de utilidad? ¿qué variables explicativas usamos y cual forma funcional escogemos? Ciertamente la especificidad de las teorías disponibles no alcanza a responder a estas preguntas y debemos por lo tanto escoger en base a la intuición y empiricamente entre aquellas que generen mejor ajuste y/o capacidad de pronóstico. De esta forma, un buen modelo debe balancear adecuadamente el uso de la teoría con la simpleza y flexibilidad del modelo.

Para ser convincente, un modelo estructural debe al menos (i) entregar suficiente flexibilidad para aprender de la data, (ii) derivar las ecuaciones de comportamiento de razonables respectos de los agentes involucrados y (iii) incorporar explicitamente en la descripcion la naturaleza no experimental de la data.

Observación: En nuestra discusión, hemos hecho la distinción entre modelos probabilisticos y modelos estructurales. Aunque los modelos probabilisticos proveen una historia de comportamiento de los agentes, los supuestos básicos usados para derivarlos no se sustentan en ninguna *teoría* de comportamiento. Por ejemplo, en modelos de duración en tiempo discreto solemos suponer que los clientes dejan de estar activos con cierta probabilidad. Más que una teoria de comportamiento esto es simplemente una descripción probabilistica de un fenomeno. En determinadas situaciones, especialmente en casos en que no disponemos una descripcion rica del ambiente en que se los agentes toman sus decisiones, nos conformamos con esta descripcion agregada del comportamiento. El enfoque estructural sobre el que ahondaremos en esta parte resulta particularmente util cuando tenemos suficiente información para investigar las motivaciones profundas de las elecciones. Al definir un modelo estructural, tanto las teorias de comportamiento como la descripcion probabilistica del sistema son fuentes válidas de estructura. Sin embargo, consideraremos como modelo econométrico estructural a aquellos que se nutren de ambas fuentes.

2.2. Modelos Estructurales en Marketing

El desarrollo de modelos estructurales se ha gestado en varias areas del conocimiento tales como Economía, Transportes, Logística, Finanzas y Marketing. Entre estas áreas, la del marketing se ha constituido en un terreno particularmente fertil para el desarrollo y adopcion del enfoque estructural. Identificamos al menos cuatro motivos por los cuales la adición de estructura en los modelos econométricos son particularmente útiles para el análisis de problemas comerciales:

1. *Disponibilidad de Data*. Gran parte de la data que registran las compañías dan cuenta de las interacciones entre clientes y firma como son ocasiones de compra, visitas a sitios web corporativos o llamadas a los call center. De esta forma, un conjunto importante de la data disponible dentro de las organizaciones son informativos respecto a procesos claves de la funcion comercial. Así, los requerimientos de datos impuestos por los modelos estructurales están inmediatamente satisfechos por procesos operacionales.

2. *Atractivo de la Evaluacion de la Intervención de sistemas* En la función comercial, casi por definición buscamos perturbar los sistemas para mejorar la oferta de valor cambiando precios, proponiendo nuevos diseños de productos, redefiniendo la cadena logistica, etc.). De esta forma necesitamos disponer de modelos que describan la reacción de los consumidores ante dichos cambios del ambiente competitivo lo que, de acuerdo a la *critica de Lucas*, solo puede hacerse con un modelo estructural.

3. *Importancia de Heterogeneidad*. En Marketing buscamos hacer inferencia desagregada a nivel de cliente o segmento para poder diseñar versiones especializadas del marketing mix que sea atractivo para segmentos especificos de clientes. Como los modelos estructurales requieren especificar los supuestos de comportamiento a nivel individual, la generacion de estimaciones desagregadas suele derivarse directamente.

4. *Pragmatismo en la aceptación de teorías.* Como hemos argumentado, una de las ventajas de los modelos estructurales es que nos permite testear si una determinada teoría de comportamiento aplica a una situación. A diferencia de otras disciplinas, en marketing hay una tradición de una revisión continua de las fuerzas que moldean el comportamiento de las personas y por tanto el enfoque de modelos estructurales entrega una herramienta alternativa a la verificación experimental de nuevas teorias.

2.3. Taxonomía de Modelos Estrucuturales

Metodológicamente, es útil generar un clasificación de los tipos de modelos estructurales existentes en la literatura. Como hemos consignado, uno de los costos de la inclusión de teoría en modelos econométricos es la mayor complejidad en las rutinas de estimación. Es esta complejidad la que dificulta la generación un mecanismo único que permita estimar modelos generales y por tanto nos vemos forzados a usar metodologías específicas dependiendo de la naturaleza del problema. En nuestra discusión basaremos nuestra clasificación en la evaluación de cuatro factores.

1. *Nivel de agregación de la Data*. Hemos propuesto que un modelo estructural debe basarse en una descripción detallada de los supuestos de los tomadores de decisión a nivel individual. Por lo tanto la disponibilidad de data a nivel individual como la decision de compra

de cada uno de los individuos de un panel de consumidores, nos habilita para, imponiendo las restricciones de identificacion necesaria, estimar los parametros de comportamiento de manera mas o menos directa. Sin embargo, en ciertas situaciones solo se dispone de informacion agregada, como participaciones de mercado o datos agregados de venta. En estos casos, la identificación de parametros de comportamiento requiere además de una descripcion del mecanismo mediante el cual se agregan las decisiones individuales. Este mecanismo tipicamente considera la especificacion de un modelo de heterogeneidad describiendo como se distribuyen los parametros entre los clientes la que se integra sobre la poblacion para generar las metricas agregadas. Esto es precisamente lo propuesto por el metodo BLP (a partir de Berry, Levinsohn y Pakes quienes primero propusieron el metodo en 1995) que describe un metodo que basado en un modelo logit permite estimar ofertas y demandas de un modelo oligopolico con informacion agregada. Por simplicidad, en esta versión nos concentraremos en modelos estimables directamente sobre data desagregada a nivel individual.

2. *Temporalidad de las Decisiones.* Dependiendo de la amplitud temporal considerada por los agendes al evaluar las alternativas de decisión distinguiremos entre problemas estáticos y dinamicos. Basicamente, si consideramos que las acciones que observamos resultan de una evaluación completa del horizonte, entonces hablaremos de problemas *dinámicos*. En caso contrarios diremos que el problema es *estático*. La distinción es importante desde un punto de vista metodologico. Si el tomador de decisiones basa sus decisiones exclusivamente mirando el pasado, entonces estas decisiones pueden caracterizarse directamente mediante condiciones de optimalidad sencillas. Por el contrario, si el tomador de decisión además evalua las repercusiones (inciertas) que sus acciones de hoy podrian tener en su bienestar futuro, entonces necesitamos caracterizar las politicas optimas a traves de ecuaciones de Bellman que incorporen explicitamente la naturaleza multiperiodo del problema. En este caso, para encontrar la politica óptima del problema se requiere usar tecnicas como programación dinámica estocastica o control óptimo aumentando de manera importante la complejidad computacional de la estimacion.

3. *Naturaleza de las Variables de Decisión*. Si las variables sobre las que deciden los agentes son continuas (gasto, montos de inversion, unidades compradas, etc.), hablaremos de un modelo de decisión continuo. Si las variables sobre las que deciden los agentes don discretos (si visita o no visita la tienda, si elige la marca A o marca B, etc.), hablaremos de un modelo de decisión discreto. La distinción es relevante en cuanto las soluciones de un problema de decisión continua puede caracterizarse directamente mediante condiciones de Karush-Kuhn-Tucker mientras que las soluciones de un problema de decisión discreta requieren una enumeración del valor de las alternativas.

4. *Identidad de los Agentes.* Los modelos estructurales pueden usarse para estudiar tanto el comportamiento de los clientes o de las otras firmas en el mercado. El area que estudia el comportamiento de las firmas ha tenido una gran desarrollo en los ultimos años y se conoce como *Organización Industrial Empirica*. En esta version, concentraremos la discusion en el estudio de los clientes por dos motivos principales: la disponibilidad de data de comportamiento de cliente y la simpleza de las nociones de equilibrio requeridas para describir a los clientes. Mientras cada cliente suele tener poco poder de mercado por si mismo, las acciones de marketing de las firmas competidoras tipicamente pueden modificar de manera importante las condiciones del mercado. Así, la descripción de las decisiones de las firmas

conlleva desafíos metodologicos importantes como la inclusión de nociones sofisticadas de equilibrio para internalizar que las decisiones de las firmas resultan tanto de mirar las respuestas esperadas de los clientes como las reacciones estrategicas de los competidores.

Metodologicamente es útil también distinguir los métodos de estimación de los modelos. La literatura reconoce dos grandes enfoques para estimar modelos estructurales como los aquí presentados: Método de los Momentos Generalizados (GMM) y Método de la Máxima Versosimilitud. Dada su eficiencia estadística (en el sentido que usa toda la información disponibles), en esta primera versión usaremos solo el método de la máxima verosimilitud. En lo que sigue nos enfocaremos la discusión al estudio del comportamiento de clientes, en problemas estáticos (o con dinámica limitada a la incorporación del pasado) y con data desagregada. Partiremos describiendo brevemente modelos de decisión continuos para luego iniciar una discusión más extensa en modelos de decisión discreta que tienen una tradición más larga en marketing.

Capítulo 3

Logit

3.1. Modelos de Elección Discreta

Nota Editorial: Aunque conceptualmente esta sección podría ser un capitulo entero antes de *logit* y *probit*, dada su extensión la presentación en un capitulo aparte no se justifica. Una posibilidad es combinar con *Modelos de Elección Continua* en un capitulo llamado *Modelos de Elección Continua y Discreta*. □

Un modelo de elección discreta consiste básicamente en situaciones en que la naturaleza de las variables de decisión a las que se enfrenta el tomador de *decisión* son discretas. Para ilustrar la intuición de la diferencia con respecto a modelos de decisión continua es útil pensar que mientras estos últimos buscan describir decisiones de el *cuanto*, los modelos de elección discreta se concentran en el *cuál*. La distinción además relevante desde un punto de vista metodológico. A diferencia de los modelos de elección continua en que la optimalidad de la elección queda bien descrita por condiciones de primer orden, al enfrentar decisiones discretas caracterizaremos la optimalidad por enumeración. Ejemplos típicos en que la decisión a evaluar es de naturaleza discreta incluye la elección de una marca por sobre otra en la góndola de un supermercado, la decisión de visitar o no a una tienda, la elección del color de una prenda de vestir, de un canal de venta y la elección de las firmas respecto a entrar o no entrar a un mercado.

Para que un problema de elección discreta este bien definido necesitamos además de variables de decisión discretas, que el conjunto de alternativas presente las siguientes tres características:

1. EXHAUSTIVAS: El conjunto sobre el que los tomadores de decisión eligen deben incluir todas las alternativas posibles. En otras palabras, cualquiera sea la decisión observada, debe estar incluida en el conjunto de elección. Esta condición es poco restrictiva ya que siempre es posible incluir en el set de alternativas la posibilidad "ninguna de las anteriores" o similar que por definición incluya toda las otras posibilidades no consideradas en conjunto. Sin embargo, esta estrategia debe usarse con precaución. Por ejemplo, al estudiar la elección de marca en una categoría en que observamos que los clientes no siempre compran alguna de las marcas disponibles podríamos incluir la alternativa de no compra en el conjunto de elección. Si la proporción de no compras es alta en nuestra muestra, la inclusión de la alternativa de no compra podría limitar la habilidad del modelo de aprender respecto a como los clientes eligen entre marcas. En este caso, podría convenir concentrarse en la elección de la marca condicional en haber hecho una compra en la categoría.

2. MUTUAMENTE EXCLUYENTES: El conjunto de decisión debe definirse de modo que en cada

ocasión el tomador de decisión seleccione solo una de las alternativas disponibles. Esto es, la elección de una alternativa implica necesariamente la no elección de cualquiera de las alternativas restante. Aunque aparentemente restrictiva, la definición de conjunto de elección puede acomodarse para generar conjuntos mutuamente excluyente. Por ejemplo, consideremos un modelo para describir la elección de los clientes entre la *tienda física tradicional* o la *tienda virtual*. Si simplemente una alternativa de elección por cada canal, entonces excluimos la posibilidad que un mismo cliente más de un canal en un mismo periodo. Para incorporar esta posibilidad debiéramos redefinir las alternativas agregando la opción de *tienda tradicional y virtual*.

3. FINITAS: El conjunto debe contener un conjunto finito de alternativas. Esta condición es importante por dos motivos técnicos. Primero, un conjunto finito facilita la evaluación de la optimalidad de las decisiones y segundo, facilita la definición de probabilidades de elección. Existen situaciones que la decisión teóricamente permite infinitas posibilidades, pero que en la practica se concentran en un numero reducido de alternativas y por tanto quedan bien representadas por un modelo de elección discreta. Por ejemplo, podemos usar el numero de cajas de cereal compradas por los clientes en cada visita al supermercado. Aunque teóricamente los clientes siempre podrían comprar una unidad adicional, el problema queda bien descrito considerando solo las alternativas de 0,1,2,3 o más de 3 cajas.

El comportamiento observado de los agentes es que alternativa eligieron en cada oportunidad y por tanto los modelos de elección discreta se enfocan en describir la probabilidad de elección de cada alternativa. Aunque frecuentemente nos encontraremos con situaciones en que solo observamos una decisión por agente, a continuación describiremos el caso de panel en que observamos múltiples agentes tomando decisiones en múltiples períodos.

Un modelo estructural para describir la probabilidad de elegir cada alternativa necesita especificar el mecanismo que usan los agentes para decidir entre las alternativas. Partiremos asumiendo que en cada oportunidad de compra t, el tomador de decisión n eligen la alternativa i que le reporta mayor utilidad u_{nit}. Aunque el tomador de decisión necesariamente necesita conocer la utilidad que deriva de cada una de las alternativa, desde la perspectiva del analista solo observamos algunas características del ambiente de decisión y del tomador de decisión a partir de las cuales podemos intentar aproximar la utilidad del tomador de decisión a través de una función $v_{nit}(x_{nit}, \theta)$ donde x_{nit} son las características observables del problema y θ el vector de parámetros que buscamos estimar y que describen la relación de dichas características con la utilidad.

Ejemplo: Supongamos que queremos describir la elección del medio de pago que usan los usuarios de una tienda determinado, el que permite pagar en efectivo o con alguna tarjeta bancaria. El analista observa 3 variables que intuye pueden ser relevantes en la elección del medio de pago: el genero del cliente ($F_n = 1$ si cliente es de género femenino), su nivel de ingresos (I_n) y el monto de la transacción (M_{nt}). Son precisamente estas características las que estarían incluidas en la matriz que hemos llamado x_{nit}. A partir de esta información pueden plantearse múltiples modelos para describir v_{nit} (asumiremos que $i = 0$ corresponde al caso de pago con efectivo mientras que $i = 1$ al de pago con tarjeta).

- *Modelo Lineal Homogeneo*: Aquí, la utilidad para ambas alternativas crece linealmente con las variables observables. En este caso, los parámetros son los mismos para todos los tomadores de decisión y por tanto el vector de parámetros viene dado por $\theta = (\alpha_0, \alpha_1, \beta, \gamma, \delta)$

$$v_{nit} = \alpha_i + \beta F_n + \gamma I_n + \delta M_{nt}$$

- *Modelo Lineal Heterogéneo:* Aquí, la utilidad para ambas alternativas también crecen linealmente con las variables observables, pero ahora los parámetros varían por alternativa y por agente y por tanto el vector de parámetros viene dado por $\theta = (\{\alpha_{1n}\}_{n=1}^{N}, \beta_0, \beta_1, \gamma_0, \gamma_1, \{\delta_n\}_{i=1}^{N})$[1]

$$v_{nit} = \alpha_{in} + \beta_i F_n + \gamma_i I_n + \delta_n M_{nt}$$

  La definición que los interceptos dependen del cliente n simplemente nos indica que cada cliente tiene una preferencia intrínseca por cada medio de pago. Del mismo modo, estamos imponiendo que la influencia que tiene el monto en el atractivo que tiene cada alternativa depende del cliente. Por ejemplo, mientras para algunos clientes el monto de la transacción puede jugar un rol importante en la decisión del medio de pago, para otros este efecto podría no ser relevante. Por ultimo la dependencia de la alternativa en los parámetros asociados a genero e ingreso podrían usarse para por ejemplo situaciones en que el nivel de ingreso afecta el atractivo de un medio de pago pero no del otro (la intuición para el género es análoga).

Por supuesto, también podemos postular modelos no lineales u otras especificaciones de la heterogeneidad. Por ejemplo que la influencia del ingreso varíe por medio de pago, pero que el efecto del genero sea constante entre las alternativas. Descubrir la especificación que mejor describe el problema es precisamente la tarea del analista □.

Observación: En el ejemplo hemos introducido brevemente el concepto de heterogeneidad. Sin embargo, para facilitar la exposición de los temas básicos, en primera instancia nos concentraremos en modelos sin heterogeneidad. En marketing los modelos que incluyen heterogeneidad en las preferencias son tan importantes que postergaremos su discusión en una capitulo separado.

En la práctica, aún en situaciones en que observamos con detalle el ambiente de decisión, no podremos describir con exactitud todas las factores que gobiernan el comportamiento de los agentes. Por lo tanto, definiremos ε_{nit} como el error (aditivo) que cometemos al aproximar u_{nit} a través de v_{nit}.

$$u_{nit} = v_{nit} + \varepsilon_{nit}$$

Así, descomponemos la utilidad de cada alternativa en una componente sistemática (u observable o explicable) v_{nit} y en una componente aleatoria (o no observable o inexplicable) ε_{nit}. Como veremos, la tarea de modelamiento del problema involucra tanto la especificación de la componente sistemática como de la aleatoria.

La componente básica para estimar estadísticamente un modelo de elección discreta es la especificación de la probabilidad de elección de cada alternativa. Sea P_{nit} la probabilidad que el agente n escoja la alternativa i en la oportunidad de compra t. El supuesto de maximización de utilidades implica que P_{nit} puede escribirse como:

$$\begin{aligned} P_{nit} &= \Pr(u_{nit} > u_{njt}, \forall j \neq i) \\ &= \Pr(v_{nit} + \varepsilon_{nit} > v_{nit} + \varepsilon_{njt}, \forall i \neq j) \\ &= \int \mathbf{1}(\varepsilon_{njt} - \varepsilon_{nit} > v_{nit} - v_{njt}) f(\varepsilon_{nt}) d\varepsilon_{nt} \end{aligned}$$

donde $\mathbf{1}(\cdot)$ toma el valor 1 si se cumple el argumento y el valor 0 en caso contrario. En esta expresión, $\varepsilon_{nt} = (\varepsilon_{n1t}, \varepsilon_{n2t}, \ldots \varepsilon_{nIt})$ es el vector de las componentes aleatorias de la elección

[1] Como veremos, para identificar el problema necesitamos imponer que $\alpha_{0n} = 0 \; \forall n = 1, ..., N$

del agente n en la oportunidad t y $f(\cdot)$ la función de densidad que describe su comportamiento probabilístico. La elección de la distribución de la componente aleatoria es importante en cuanto impone restricciones a los patrones de comportamientos que pueden ser capturados por el modelo. Concentraremos nuestra atención en los casos en que ε_{nit} se distribuye valor extremo que da origen al modelo *logit* y normal que da origen al modelo *probit*.

3.2. Modelo Logit

El modelo logit resulta de asumir que cada ε_{nit} es independientemente distribuido de acuerdo a una distribución gumbel o de valor extremo tipo I.

$$F(\varepsilon_{nit}) = e^{-e^{-\varepsilon_{nit}}} \qquad f(\varepsilon_{nit}) = e^{-\varepsilon_{nit}} e^{-e^{-\varepsilon_{nit}}} \tag{3.1}$$

Aplicando esta definición, podemos demostrar que la probabilidad de elección en un modelo logit corresponde a una formaula cerrada sencilla (para el detalle de la derivación ver apéndice):

$$\begin{aligned} P_{nit} &= \int \Pr(\varepsilon_{njt} < v_{nit} - v_{njt} + \varepsilon_{nit}, \forall j \neq i \mid \varepsilon_{nit}) f(\varepsilon_{nit}) d\varepsilon_{nit} \\ &= \int \left(\prod_{j \neq i} e^{-e^{-(v_{nit} - v_{njt} + \varepsilon_{nit})}} \right) e^{-\varepsilon_{nit}} e^{-e^{-\varepsilon_{nit}}} d\varepsilon_{nit} \\ &= \frac{e^{v_{nit}}}{\sum_j e^{v_{njt}}} \end{aligned}$$

En algunos libros de texto se justifica esta expresión simplemente como una regresión logística, esto es una transformación lineal para normalizar la utilidad de modo de interpretarla directamente como una probabilidad de elección en el rango [0,1]. Aunque válido, resulta útil entender que en efecto dicha expresión puede derivarse a partir de supuestos de maximización de utilidades.

Para ganar algo de intuición respecto a la expresión de la probabilidad de elección, es útil graficarla con respecto a la utilidad derivada por cada alternativa. Por ejemplo, supongamos que tenemos una decisión binaria que por ejemplo corresponde a decisión de comprar o no comprar un producto. En este caso, la probabilidad de comprar el producto crece *sigmoidalmente* con la utilidad derivada de la compra. Esto es, al graficar la probabilidad de compra con respecto a la utilidad derivada obtenemos una curva S como muestra la Figura 1. En la figura, hemos agregado también la curva de la probabilidad de elección en el caso en que en vez de asumir que el error se distribuye valor extremo como demanda el modelo logit, asumimos que el error está normalmente distribuido como tradicionalmente hacemos en otros modelos econométricos.

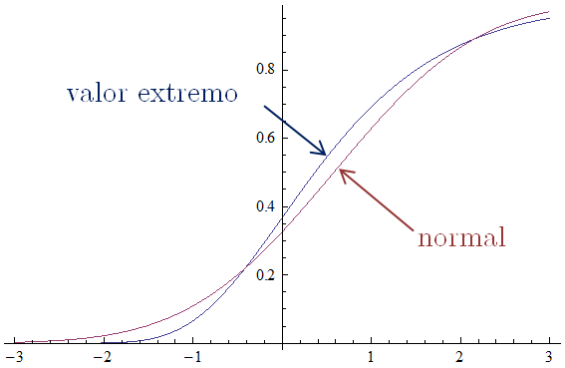

Figura 1: Probabilidad de elección

La disposición de una fórmula cerrada para la probabilidad de elección facilita el cálculo de múltiples métricas asociadas que permiten complementar el análisis. Supongamos que la utilidad de una alternativa viene dada por $v_{nit} = v(x_{nit}, \theta)$, entonces podemos calcular

- Como varía la probabilidad de elegir la alternativa i al variar alguna componente de la utilidad de la misma alternativa.

$$\frac{dP_{nit}}{dx_{nit}} = \frac{\partial v_{nit}}{\partial x_{nit}} \cdot P_{nit}(1 - P_{nit})$$

- Como varía la probabilidad de elegir la alternativa i al variar alguna componente de la utilidad otra alternativa.

$$\frac{dP_{nit}}{dx_{njt}} = \frac{\partial v_{njt}}{\partial x_{njt}} \cdot P_{nit} \cdot P_{njt}$$

- Elasticidad de la probabilidad de elegir la alternativa i con respecto alguna componente de la utilidad de la misma alternativa.

$$e_{ix_{nit}} = \frac{\partial P_{nit}}{\partial x_{nit}} \cdot \frac{x_{nit}}{P_{nit}} = \frac{\partial v_{nit}}{\partial x_{nit}} x_{nit}(1 - P_{nit})$$

- Elasticidad de la probabilidad de elegir la alternativa i con respecto alguna componente de la utilidad otra alternativa.

$$e_{ix_{njt}} = \frac{\partial P_{nit}}{\partial x_{njt}} \cdot \frac{x_{njt}}{P_{nit}} = \frac{\partial v_{njt}}{\partial x_{njt}} x_{njt} P_{njt}$$

Recuerden que unas de las motivaciones para el uso de modelos estructurales es la posibilidad de analizar contrafactuales, esto es ver que pasaría con el mercado si hay cambio en alguna variable de control interesante. Por ejemplo que pasa con las participaciones de mercado si sube el precio de una alternativa, si se aumenta la frecuencia publicitaria, etc. Las métricas recien presentadas permiten precisamente hacer dichas evaluaciones de manera directa.

3.2.1. Propiedades del modelo Logit

El modelo logit es bastante flexible para acomodar una amplia variedad de situaciones. En efecto, distintas especificaciones de las funciones de utilidades de las alternativas permiten describir múltiples fenómenos asociados a la elección. Sin embargo, es importante reconocer que los supuestos subyacentes al logit imponen importantes restricciones a como describimos la lógica en que los agentes evaluan las alternativas y escogen entre ellas.

Para fijar ideas, resulta útil pensar qué restricciones impone asumir que las componentes no observables de la utilidad son todas independientes entre ellas. El supuesto de independencia nos obliga a imponer que cualquier relación entre las utilidades de dos alternativas debe necesariamente capturarse a través de variables observables. Del mismo modo, las utilidades que derivamos por dos alternativas en ocasiones de elección diferentes solo pueden describirse a través de elementos que podamos observar a lo largo del tiempo. Para entender mejor como estas limitaciones se materializan en la formulación del modelo, discutiremos formalmente tres características del modelo logit: la existencia de patrones de substitución proporcional, la incapacidad de capturar tanto heterogeneidad aleatoria en las preferencias como componentes dinámicas no observables.

Patrones de sustitución

Los patrones de substitución derivados de un modelo logit son bastante peculiares y aunque desde un punto de vista econométrico puede resultar beneficioso, desde el punto de vista de la investigación de teorías de comportamiento suele ser considerado como bastante restrictivo. Entenderemos por patrones de substitución a la forma en que cambia la probabilidad de elección de alguna alternativa cuando se modifica el atractivo de otra alternativa. Para entender la naturaleza de los patrones de substitución del modelo logit es util calcular el ratio de las probabilidades de elección de dos alternativas cualquiera i y j.

$$\frac{P_{ni}}{P_{nj}} = e^{v_{ni} - v_{nj}}$$

Este ratio solo depende de las utilidades observables de las dos alternativas consideradas lo que indica que la probabilidad relativa de elegir la alternativa i sobre la alternativa j no depende de que otras alternativas existan ni de los atributos que ellas tengan. Por ejemplo, si agregamos una alternativa al conjunto de elección, el ratio de probabilidades de las alternativas existentes se mantendrá constante independiente de las características de la nueva alternativa. Nos referiremos a esta característica como *independencia de alternativas irrelevantes* o *IIA*.

Para ejemplificar consideremos una botillería que ofrece dos variedades de vino, uno blanco y otro tinto. Supongamos ademas que estas dos alternativas tienen la misma participación de mercado, esto es la mitad de los clientes de la botillería compra vino blanco y la otra mitad compra vino tinto. En este caso, las utilidades sistemáticas debieran ser similares y por tanto el ratio de probabilidades de elección de vino blanco sobre vino tinto debiera acercarse a 1. Motivado por un mayor margen de los vinos tintos, el administrador de la botillería decide incorporar una nueva variedad de vino tinto. Intuitivamente esperaríamos que, como la nueva variedad de vino tinto es sustituto más cercano al tinto existente, la participación de mercado de este debiera decrecer más que la de vino blanco. Sin embargo la propiedad de IIA impone que este ratio se mantiene constante. En otras palabras, la introducción de una nueva alternativa disminuirá la participación

de todas la otras alternativas independiente de las similitudes que tengan. Esta última observación puede corroborarse calculando la elasticidad de sustitución $E_{iz_{nj}}$ que determina como cambia la probabilidad de consumir la alternativa i ante un cambio en un atributo z_{nj} de la alternativa j

$$E_{ix_{n_j}} = -\frac{\partial v_{nj}}{\partial x_{nj}} x_{nj} P_{nj}, \forall i \neq j$$

Notamos en esta expresión que la expresión no depende de i por lo que es constante para todas las alternativas de elección. Luego, si ocurre una mejora en los atributos de una alternativa la probabilidad de elección de las demás disminuye en el mismo porcentaje independiente de la similitud entre alternativas. Nos referiremos a esta caracteristica como *patrones de sustitución proporcionales*.

Una ventaja de los patrones de substitución del modelo logit es que permite que los parámetros del modelo sean estimados consistentemente en base a un subconjunto de las alternativas. Esto es particularmente útil en ambientes de decisión de marketing donde típicamente nos encontramos con centenas de productos que potencialmente pueden constituir alternativas de elección en una situación de compra. De esta forma, para estimar un modelo logit podemos seleccionar conjuntos reducidos de alternativas que capturan los elementos esenciales de la elección e ignorar que pasa con todas las otras alternativas.

Incapacidad de estimar componentes aleatorias

La investigación de las diferencias entre las preferencias de los distintos clientes es un tema fundamental para el desarrollo de planes comerciales exitosos. Tradicionalmente distinguimos dos tipos de heterogeneidad de acuerdo a la capacidad de observación del analista. Por un lado tenemos el estudio de heterogeneidad observable que indica como las preferencias de los tomadores de decisiones varían de acuerdo a sus características medibles. Este tipo de heterogeneidad nos permite por ejemplo estudiar diferencias en las preferencias entre hombres y mujeres, por edad o por niveles de ingreso. Sin embargo una proporción importante de las diferencias de las preferencias no es atribuible a características observables como las recién descritas. Por ejemplo, dos hermanos del mismo género de edades similares viviendo en el mismo hogar pueden tener preferencias completamente diferentes respecto a sabores de yogurt.

El resultado fundamental en esta sección indica que un modelo logit permite estudiar variaciones de preferencias asociadas a componentes observables, pero no a componentes no observables. Para ilustrar este resultado, supongamos un tomadores de decisión caracterizados por la siguiente función de utilidad:

$$u_{nit} = \alpha_i + \beta_n p_{it} + \varepsilon_{nit}$$

Es decir, la utilidad de cada alternativa tiene una componente base que es constante entre los tomadores de decisión y una penalización por precio p_{it} al que se enfrenta el tomador de decisión. Al indexar β_n por agente estamos explícitamente permitiendo que algunos tomadores decisión sean más sensibles al precio que otros. Supongamos que postulamos que el coeficiente precio viene dado por la siguiente ecuación de regresión.

$$\beta_n = \lambda_0 + \lambda_1 I_n + \mu_n$$

donde λ_0 captura la sensibilidad base al precio, I_n el nivel de ingreso del agente n y λ_1 el coeficiente que indica como dichos niveles de ingresos afectan la sensibilidad al precio. Por último μ_n es un valor aleatorio que captura todas las otras componentes que modifican la sensibilidad al precio más allá del nivel base y los ingresos.

$$\begin{aligned} u_{nit} &= \alpha_i + (\lambda_0 + \lambda_1 I_n + \mu_n)p_{it} + \varepsilon_{nit} \\ &= \alpha_i + \lambda_0 p_{it} + \lambda_1 p_{it} I_n + \xi_{nit} \end{aligned}$$

donde $\xi_{nit} = \mu_n p_{it} + \varepsilon_{nit}$. De esta expresión debiera ser claro que la inclusión de heterogeneidad observable puede ser capturada bajo un enfoque logit. En efecto, los parámetros α_i, λ_0 y λ_1 dan cuenta respectivamente del nivel de utilidad base por alternativa, de la penalización por precio y de como dicha penalización se ve modificada por el nivel de ingresos. Lamentablemente, la variación aleatoria μ_n no puede ser incluida ya que su inclusión necesariamente implica que las componentes errores ξ_{nit} no están idénticamente distribuidas. En efecto, se puede mostrar que $\mathbb{V}\text{ar}(\xi_{nit}, \xi_{njt}) = \mathbb{V}\text{ar}(\mu_n)p_{it}^2$ que evidentemente varía entre alternativas. Más aún, también se puede mostrar que $\mathbb{C}\text{ov}(\xi_{nit}, \xi_{njt}) = \mathbb{V}\text{ar}(\mu_n)p_{it}p_{jt} \neq 0$ violando también el supuesto de independencia.

Es importante notar que la incapacidad de capturar aleatoriedad aplica también a componentes dinámicas. Esto es, al observar compras repetidas en el tiempo, el modelo logit no permite capturar que hay componentes no observables que varíen en el tiempo. Por ejemplo no podemos incorporar que, debido a factores externos no observables, en algunos periodos algunas alternativas son más atractivas para todos los agentes decidiendo en dichos periodos. Al igual que en el ejemplo anterior, incluir estas variaciones viola los supuestos de distribuciones independientes e idénticamente distribuidas para las componentes no observables.

3.2.2. Estimación

Para estimar el modelo, necesitamos escribir la verosimilitud del problema. La componente fundamental para la construcción de la verosimilitud es la descripción de la probabilidad de elección P_{nit}. Para el caso del modelo logit, como la expresión de la probabilidad de elección corresponde a una formula analítica cerrada, la construcción de la verosimilitud es directa. Si la componente determinística de la utilidad viene dada por $v_{nit}(x_{nit}, \theta)$ y si y_{nit} es una variable que toma valor 1 si el tomador de decisión n escoge alternativa i en oportunidad t, entonces la verosimilitud viene dada por:

$$L(\theta) = \prod_n \prod_i \prod_t (P_{nit})^{y_{nit}} = \prod_n \prod_i \prod_t \left(\frac{e^{v_{nit}(x_{nit},\theta)}}{\sum_j e^{v_{njt}(x_{njt},\theta)}} \right)^{y_{nit}}$$

La que podemos maximizar directamente usando rutinas estándares de programación convexa. Computacionalmente, suele ser más conveniente trabajar con la log-verosimilitud en vez de la verosimilitud. Esto porque la multiplicación de probabilidades genera muy rápidamente valores que computacionalmente son indistinguibles de cero. Recordar que el valor de los valores óptimos son invariantes a transformaciones monótonas como la del logaritmo.

$$\begin{aligned} LL(\theta) &= \sum_n \sum_i \sum_t y_{nit} \ln \left(\frac{e^{v_{nit}(x_{nit},\theta)}}{\sum_j e^{v_{njt}(x_{njt},\theta)}} \right) \\ &= \sum_n \sum_i \sum_t y_{nit} v_{nit}(x_{nit},\theta) - \sum_n \sum_i \sum_t \ln \left(\sum_j e^{v_{njt}(x_{njt},\theta)} \right) \end{aligned}$$

Como hemos indicado, esta función objetivo puede ser ingresada directamente a cualquier rutina de optimización para encontrar los estimadores máximo verosímiles. Para muchas instancias prácticas, es conveniente contar además con las derivadas de la log-verosimilitud de modo de encontrar eficientemente direcciones de máximo ascenso o evaluar si el punto es estacionario o no. Afortunadamente, para la mayoría de las especificaciones del modelo logit, estas derivadas también son fáciles de obtener. Por ejemplo, si la componente sistemática de la utilidad viene dada por $v_{nit}(x_{nit},\theta) = x'_{nit}\theta$ entonces

$$\frac{\partial LL(\theta)}{\partial \theta} = \sum_n \sum_i \sum_t \left(y_{nit} - \frac{e^{x'_{nit}\theta}}{\sum_j e^{x'_{njt}\theta}} \right) x_{nit}$$

Del mismo modo, podemos calcular segundas derivadas que resultan útiles para el cálculo de errores estándares de los parámetros.

Evaluación del modelo

Nota Editorial: Dado que la mayoría de las métricas de evaluación son transversales al modelo, sería conveniente disponer de un único capitulo dedicado a este tema □

Al igual que en otros modelos econométricos, una de las componentes fundamentales del análisis es la evaluación de la calidad del modelo. La variedad de métricas disponibles para la evaluación es muy amplia y la mayoría son transversales a cualquier modelo. Categorizaremos las herramientos de evaluación en tres grupos: bondad de ajuste, capacidad de pronóstico y test de hipótesis.

1. BONDAD DE AJUSTE: Las métricas de bondad de ajuste básicamente nos indican que tan bien el modelo ajusta a la data. En el contexto de modelos de regresión, solemos analizar el estadístico R^2 que mide la proporción de la variabilidad de la variable dependiente que puede ser explicado por la variación de las variables independientes. En el contexto de modelos de elección discreta basaremos la evaluación en el valor de la verosimilitud usando alguno o varios de los siguientes indicadores:

   - ρ de McFadden. Este índice está en el rango [0,1] e informalmente, se suele interpretar como el coeficiente de determinación (R^2) en el sentido que un valor cercano a 0 indica un mal ajuste y un valor cercano a 1 indica un buen ajuste. Sin embargo, es importante notar que no puede decirse que ρ mida la variabilidad explicada por el modelo como hace el coeficiente de determinación

$$\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)}$$

- Criterio de información de Akaike(AIC) y Bayesiano (BIC): Uno de las limitaciones del ρ de McFadden es que solo permite comparar modelos con el mismo numero de parámetros. Los dos indicadores más usados para comparar modelos con distintos números de parámetros son AIC y BIC en que se penaliza la verosimilitud por el numero de parámetros para capturar el hecho que al incluir nuevos parámetros la verosimilitud necesariamente crecerá. La diferencia entre AIC y BIC es que el primero tiene una penalización constante por numero de parámetros mientras que la penalización del segundo depende de la cantidad de data disponible. Si la log verosimilitud de un modelo con n observaciones y k parámetros es LL entonces AIC y BIC vienen dados por:

$$AIC = -1LL(\hat{\theta}) + 2k \qquad BIC = -2LL(\hat{\theta}) + k\ln(n)$$

2. CAPACIDAD DE PRONÓSTICO: Un modelo que explique muy bien la data puede correr el riesgo de sobreajustar. Esto es que no permita describir el fenómeno más allá de los datos con que se calibran. Para medir la capacidad de pronóstico se suele dividir la data en un subconjunto de calibración en que estimamos el modelo y otro de validación en que comparamos las realizaciones con lo pronosticado usando las estimaciones del subconjunto de calibración. Supongamos que estamos interesados en evaluar la capacidad de pronostico de un indicador f_{ni} que puede corresponder a las elecciones mismas, participaciones de mercado o cualquier otra. Si $\hat{f}_{ni}$ es el pronostico del modelo entonces solemos usar el *mean absolute error* (MAE) o el *mean absolute percentage error* (MAPE)

$$MAE = \frac{1}{N}\sum_n\sum_i |f_{ni} - \hat{f}_{ni}| \qquad MAPE = \frac{1}{N}\sum_n\sum_i \left|\frac{f_{ni} - \hat{f}_{ni}}{f_{ni}}\right|$$

3. TEST DE HIPÓTESIS: La evaluación de hipótesis, también puede contribuir a diagnosticar un modelo. Por ejemplo, al agregar una variable explicativa, nos gustaría evaluar si el coeficiente correspondiente es significativamente diferente de 0, lo que podemos hacer directamente a través de la construcción de intervalos de confianza o su estadístico t equivalente (recordar que la varianza de del estimador máximo verosímil puede obtenerse usando el inverso del Hessiano). En ocasiones también estaremos interesados en testear hipótesis más complejas para lo que recurrimos a test de ratios de verosimilitud. Supongamos por ejemplo que tenemos un modelo en que los coeficientes asociado a *display* difieren por marca para incorporar la posibilidad que algunas de ellas sean más efectivas en su comunicación en sala. El tests de ratios de verosimilitud nos permite por ejemplo testear si estos coeficientes son iguales o si efectivamente difieren entre marcas. Si la hipótesis nula puede expresarse como k restricciones sobre los parámetros, entonces podemos estimar un modelo A irrestricto y otro B restringido y calculamos el estadístico $LR = 2(LLA - LLB)$, que se distribuye χ^2 con k grados de libertad.

Apéndice: Derivación probabilidad de elección modelo logit

Por definición

$$P_{nit} = \Pr(\varepsilon_{njt} < v_{nit} - v_{njt} + \varepsilon_{nit}, \forall j \neq i)$$

Fijando el valor de ε_{nit}, la probabilidad anterior no es más que una multiplicación de funciones distribución de variables aleatorias valor extremo. Por lo tanto podemos condicionar en ε_{nit} y luego integrar respecto a los valores que puede tomar. Para simplificar la notación, sea $s = \varepsilon_{nit}$

$$\begin{aligned} P_{nit} &= \int_{-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-(s+v_{ni}-v_{nj})}} \right) e^{-s} e^{-e^{-s}} ds \\ &= \int_{-\infty}^{\infty} \left(\prod_{j} e^{-e^{-(s+v_{ni}-v_{nj})}} \right) e^{-s} ds \\ &= \int_{-\infty}^{\infty} \exp\left(-e^{s} \sum_{j} e^{-(v_{ni}-v_{nj})} \right) e^{-s} ds \end{aligned}$$

Para resolver la integral podemos recurrir a un cambio de variables $t = e^{-s}$ y $dt = e^{-s} ds$. Con esto

$$\begin{aligned} P_{nit} &= \int_{-\infty}^{0} -e^{-t \sum_j e^{-(v_{ni}-v_{nj})}} dt \\ &= -\frac{e^{-(v_{ni}-v_{nj})}}{\sum_j e^{-(v_{ni}-v_{nj})}} \Bigg|_{0}^{\infty} \\ &= \frac{e^{v_{ni}}}{\sum_j e^{v_{nj}}} \end{aligned}$$

□

Capítulo 4

Probit

4.1. Definición

Al introducir modelos de elección discreta, postulamos que los tomadores de decisiones disponían de una función de utilidad subyacente que descomponíamos en una componente determinística y otra aleatoria. Más aún, discutimos que el modelo que describe la probabilidad de elegir cada una de las alternativas quedaba directamente determinada por la distribución que asumiéramos para la componente aleatoria de la utilidad. Aunque una especificación de errores normales centrados en cero tiene una larga tradición en modelos econométricos, por simplicidad optamos iniciar la discusión con modelos *logit* derivados de asumir que la componente aleatoria de la utilidad se distribuía valor extremo tipo I. En este capitulo volveremos al caso de componentes aleatorias normales que dan origen al modelo *probit*. Formalmente, un modelo probit resulta de los siguientes supuestos de comportamiento:

$$u_{ni} = v_{ni} + \varepsilon_{ni} \qquad \varepsilon_n \sim N(0, \Sigma) \tag{4.1}$$

La normalidad de los errores provee bastante flexibilidad para acomodar una amplia variedad de estructuras de las preferencias. Como veremos en la discusión que sigue, un modelo con errores normales permite acomodar factores sistemáticos no observables en la utilidad. Una de las pocas limitaciones de un modelo *probit* viene de la normalidad dichos factores. Por ejemplo, si queremos incorporar el efecto que tiene el precio en la utilidad como una componente aleatoria, entonces las colas de la distribución normal implicara una probabilidad positiva de que algunos clientes aumenten la utilidad de una alternativa si aumenta el precio de esta. Formalmente, el supuesto de la normalidad de la componente aleatoria de la utilidad implica que su función de densidad viene dada por:

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{I/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\varepsilon_n' \Sigma^{-1} \varepsilon_n} \tag{4.2}$$

Esta expresión no es más que la versión multivariada de la bien conocida densidad de la distribución $N(0, \sigma^2)$. La matriz Σ corresponde a la matriz varianza-covarianza de los errores. Por tratarse de una distribución normal, la matriz Σ es simétrica y de dimensión $I \times I$, donde I es el número de alternativas disponibles para el tomador de decisión. Por ejemplo, si hay tres alternativas disponibles, la matriz Σ tomaría la siguiente forma:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_{22} & \sigma_{23} \\ \cdot & \cdot & \sigma_{33} \end{bmatrix} \tag{4.3}$$

Los coeficientes en la diagonal dan cuenta de la variabilidad de la componente aleatoria de la utilidad. Así por ejemplo, si σ_{ii} tiene un valor alto indica que hay una fracción importante de la utilidad de la alternativa i que no es capturada por el modelo de la componente sistemática. Los coeficientes fuera de la diagonal dan cuenta de la correlación de las componentes no observables de cada una de las alternativas. De este modo, si σ_{ij} tiene un valor positivo alto indica que existe un elemento no observable importante que afecta simultáneamente las alternativas i y j.

Como vimos en el desarrollo del modelo *logit*, una componente fundamental para estimar un modelo de elección discreta es la derivación de una expresión para la probabilidad que cada agente elija cada alternativa en cada ocasión. Para el modelo *probit*, la probabilidad que el individuo n elija la alternativa i viene dada por:

$$\begin{aligned} P_{ni} &= \Pr(v_{ni} + \varepsilon_{ni} > v_{nj} + \varepsilon_{nj} \quad \forall j \neq i) \\ &= \int \mathbf{1}(v_{ni} + \varepsilon_{ni} > v_{nj} + \varepsilon_{nj} \quad \forall j \neq i)\phi(\varepsilon_n)d\varepsilon_n \end{aligned} \tag{4.4}$$

Intuitivamente, simplemente calculamos el volumen bajo la densidad $\phi(\varepsilon_n)$ en la región en que los errores son tales que la alternativa i es aquella que reporta mayor utilidad al individuo n. A diferencia del modelo logit, la integral sobre la densidad $\phi(\cdot)$ no tiene primitiva analítica y por tanto no disponemos de una formula cerrada para P_{ni}.

4.2. Patrones de substitución

Una de las grandes ventajas de un modelo *probit* es su flexibilidad para capturar una amplia variedad de patrones de comportamiento. En efecto, un modelo *probit* no impone restricciones en los patrones de substitución más allá de la simetría propia de la distribución normal lo que posibilita al analista explorar el esquema que mejor se ajusta a la data. En este sentido, es útil compararlo con el modelo *logit* que, aunque provee una formula analítica cerrada para la probabilidad de cada elección, impone la propiedad de substitución proporcional (o de independencia de alternativas irrelevantes). El modelo *probit* no tiene esta propiedad y por tanto el aumento de la probabilidad de elección de una alternativa puede tener impactos diferentes en las probabilidades de elección de las alternativas remanentes. Esto permitiría por ejemplo identificar pares de alternativas que son mejores substitutos (complementos) más allá de las comonalidades que podrían existir en las componentes determinísticas de su utilidad.

A continuación discutiremos como el modelo *probit* puede ser usado para representar algunas situaciones de elección discreta.

4.2.1. Variación aleatorias en preferencias

Una de las componentes más importantes en el diseño de un plan comercial exitoso es la identificación de como las preferencias de los potenciales clientes se distribuyen en la población. Identificando estas variaciones, podemos encontrar las propuestas de valor que resulten más atractivas

para cada grupo de clientes. En un modelo *probit*, podemos asumir que los parámetros que definen la componente determinística son heterogéneos en la población sin perder los supuestos básicos que definen el modelo. Por simplicidad, supongamos que la componente determinística de la utilidad es lineal:

$$u_{ni} = \beta_n' x_{ni} + \varepsilon_{ni} \qquad \varepsilon_n \sim N(0, \Sigma) \tag{4.5}$$

Notar que a diferencia de los modelos anteriores, ahora hemos asumido que cada tomador de decisión n tiene su propio set de parámetros β_n que describen sus preferencias por las alternativas disponibles. Para completar el modelo necesitamos especificar una distribución de β_n en la población. Para mantener la estructura del modelo asumiremos normalidad: $\beta_n \sim N(b, \sigma_\beta^2)$. Dado que la suma de dos variables aleatorias normales se distribuye normal, es fácil ver que el modelo es equivalente a

$$u_{ni} = b' x_{ni} + \eta_{ni} \qquad \eta_n \sim N(0, \hat{\Sigma}) \tag{4.6}$$

Las componentes de la matriz de varianza-covarianza resultante $\hat{\Sigma}$ pueden trazarse directamente a las componentes de la matriz Σ original como lo indica el siguiente ejemplo:

Ejemplo: Consideremos un modelo de elección con dos alternativas y un modelo lineal con una única variable para describir la componente sistemática de la utilidad. En este caso, las utilidades por cada alternativa vienen dadas por:

$$u_{n1} = \beta_n x_{n1} + \varepsilon_{n1}$$
$$u_{n2} = \beta_n x_{n2} + \varepsilon_{n2}$$

donde ε_{n1} y ε_{n2} son términos independientes e identicamente distribuidos con varianza σ_ε. Si asumimos que el parámetro β_n se distribuye normal con media b y varianza σ_β, entonces podemos re-escribir las utilidades como:

$$u_{n1} = b x_{n1} + \eta_{n1}$$
$$u_{n2} = b x_{n2} + \eta_{n2}$$

donde η_{n1} y η_{n2} están normalmente distribuidas. Cada una tiene esperanza cero: $\mathbb{E}(\eta_{ni}) = \mathbb{E}(\beta_n x_{ni} + \varepsilon_{ni}) = 0$, varianza igual a $\mathbb{V}\text{ar}(\eta_{ni}) = \mathbb{V}\text{ar}(\beta_n x_{ni} + \varepsilon_{ni}) = x_{ni}^2 \sigma_\beta^2 + \sigma_\varepsilon$ y covarianzas $\mathbb{C}\text{ov}(\eta_{n1}, \eta_{n2}) = x_{n1} x_{n2} \sigma_\beta$. Así, la matriz de covarianza viene dada por:

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

El siguiente paso es estimar. Recordando que el comportamiento no es afectado por transformaciones multiplicativas de la utilidad, es necesario escalar esta matriz. Lo recomendable es fijar $\sigma_\varepsilon = 1$, obteniendo así

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

4.2.2. Dependencia en el tiempo

Hemos discutido que bajo un modelo *probit* podemos estudiar relaciones no observables entre las alternativas de elección. En las bases disponibles para la función comercial, las observaciones suelen estar indexadas temporalmente generando estructuras de panel que permiten estudiar aspectos interesantes de los agentes. Discutiremos a continuación como usar un modelo *probit* para explorar no solo relación entre las utilidades de alternativas sino que también del comportamiento de las utilidades de las alternativas en el tiempo. Al igual que en la sección anterior, buscamos encontrar patrones temporales en las componentes no observables de la utilidad, ya que las variaciones en la componente observable puede ser fácilmente estudiada incluyendo variables observables que describan la evolución temporal del sistema. Por ejemplo, si creemos que la utilidad de una de las alternativas es creciente en el tiempo, basta incluir el tiempo t entre las variables independientes en la descripción de la utilidad de la alternativa. En general, debiéramos esperar que las utilidades estén correlacionadas tanto en el tiempo como entre las alternativas ya que los factores que no son observados por el analista suelen ser persistentes en el tiempo. Eventualmente un modelo probit también podría ayudar a identificar shocks en que hay variaciones instantáneas (o de unos pocos periodos) en las utilidades de varias de las alternativas.

Supongamos que observamos un panel de N clientes que deciden respecto de I alternativas en T periodos y que la utilidad del producto que el agente n deriva sobre la alternativa i en el periodo t viene dada por:

$$u_{nit} = v_{nit} + \varepsilon_{nit} \qquad \left[\varepsilon_{n11}, ..., \varepsilon_{nI1}, \varepsilon_{n12}, ..., \varepsilon_{nI2} ..., \varepsilon_{n1T}, ..., \varepsilon_{nIT}\right] \sim N(0, \Sigma) \tag{4.7}$$

La matriz de covarianza Σ tiene dimensión $IT \times IT$ (como veremos, no todas las componentes son identificables y tendremos que imponer ciertas restricciones).

Para paneles típicos, T es grande y generan matrices de varianza covarianza muy grandes. Por ejemplo, si tenemos datos semanales de compras de 5 marcas por un periodo de 2 años, nos enfrentaremos con una matriz de varianza (sin normalizar) con $5 \times 104 = 520$ filas y 520 columnas, lo que nos generaría no solo un modelo difícil de estimar numéricamente si no que también difícil de interpretar. Así, para usar un modelo *probit* con dependencia en el tiempo, típicamente agregaremos estructura al modelo. Por ejemplo podemos restringir nuestro análisis a grupos de periodos que podría ser el caso de las decisiones antes y después de una intervención en el sistema (e.g. antes y después del lanzamiento de una campaña publicitaria).

Ejemplo: Supongamos un caso de elección binaria, el error está compuesto por una componente sistemática específica del tomador de decisión, y otra que es variable en el tiempo.

$$\varepsilon_{nt} = \eta_n + \mu_{nt} \tag{4.8}$$

Si asumimos que η_n está distribuida $N(0, \sigma)$ y μ_{nt} $N(0, 1)$, entonces la varianza y covarianza son

$$\mathbb{V}\text{ar}(\varepsilon_{nt}) = \mathbb{V}\text{ar}(\eta_n + \mu_{nt}) = \sigma + 1 \tag{4.9}$$

$$\mathbb{C}\text{ov}(\varepsilon_{nt}, \varepsilon_{ns}) = \mathbb{E}((\eta_n + \mu_{nt})(\eta_n + \mu_{ns})) = \sigma \tag{4.10}$$

La matriz Σ, por lo tanto, es

$$\Sigma = \begin{bmatrix} \sigma+1 & \sigma & \cdots & \sigma \\ \sigma & \sigma+1 & \cdots & \sigma \\ \vdots & \vdots & \ddots & \vdots \\ \sigma & \sigma & \cdots & \sigma+1 \end{bmatrix} \tag{4.11}$$

4.3. Identificación

Para estimar un modelo probit, junto con los parámetros de la componente sistemática de la utilidad necesitamos estimar los coeficientes la matriz Σ. Por tratarse de una distribución normal, la matriz Σ es simétrica y por tanto en principio se deben estimar $\frac{I(I+1)}{2}$ de sus componentes. Sin embargo dicho problema no es identificable y necesitamos imponer restricciones adicionales. La intuición detrás de esta falta de identificación resulta de asumir que las utilidades subyacentes que maximizan los individuos son monótonas y homotéticas. En otras palabras, podemos agregar un valor constante a las utilidades de cada una de las alternativas o escalarlas en cualquier proporción y la identidad de la alternativa de mayor utilidad no cambia. En general, si tenemos I alternativas, solo podemos identificar $\frac{I(I-1)}{2} - 1$ parámetros. A continuación discutiremos dos enfoques para generar restricciones que hagan el problema identificable.

4.3.1. Normalización de las funciones de utilidad

Motivados en las propiedades de la función de utilidad, este enfoque consiste en imponer directamente restricciones de escala y locación. Este enfoque es completamente general y permite además garantizar identificación con un procedimiento estándar que puede incluso automatizarse. Formalmente el proceso consiste en imponer dos restricciones:

1. FIJAR LOCACIÓN: Como el valor absoluto de las utilidades es irrelevante, podemos fijar arbitrariamente el punto de referencia sobre el cual interpretaremos las utilidades. De esta forma, tomaremos la utilidad de una de las alternativas como referencia y re-definiremos las utilidades como las diferencias con respecto a la alternativa de referencia.

2. FIJAR ESCALA: Como la escala de las utilidades es irrelevante, podemos fijarla asignando un valor arbitrario a cualquiera de las componentes de la matriz de varianza covarianza. Típicamente impondremos que la primera componente de la diagonal tome el valor 1.

Ejemplo: Consideremos la normalización de una matriz Σ resultante de un problema de elección discreto de 4 alternativas.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{bmatrix} \tag{4.12}$$

El primer paso en la normalización es considerar diferencias de utilidades con respecto a una alternativa de referencia, la que por simplicidad escogeremos como la primera de la lista. Al fijar

esta utilidad y tomar las diferencias, hemos reducido la dimension del vector errores, resultando en una matriz de varianza-covarianza $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}_{i,j=1}^{3}$ cuyas componentes vienen dadas por:

$$\begin{aligned}
\hat{\sigma}_{22} &= \sigma_{22} + \sigma_{11} - 2\sigma_{12} \\
\hat{\sigma}_{33} &= \sigma_{33} + \sigma_{11} - 2\sigma_{13} \\
\hat{\sigma}_{44} &= \sigma_{22} + \sigma_{11} - 2\sigma_{14} \\
\hat{\sigma}_{23} &= \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} \\
\hat{\sigma}_{24} &= \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14} \\
\hat{\sigma}_{34} &= \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}
\end{aligned}$$

El segundo paso en la normalización es fijar en 1 (o cualquier otro real positivo) una de las componentes de la diagonal de la matriz de varianza covarianza para precisar la escala de la función de utilidad. Por simplicidad escogemos la primera componente de la diagonal. Para hacerla 1 basta con dividir toda la matriz por dicha componente, resultando en una matriz de varianza-covarianza $\tilde{\Sigma} = \{\tilde{\sigma}_{i,j}\}_{i,j=1}^{3}$ cuyas componentes vienen dadas por:

$$\begin{aligned}
\tilde{\sigma}_{33} &= \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\tilde{\sigma}_{44} &= \frac{\sigma_{22} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\tilde{\sigma}_{23} &= \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\tilde{\sigma}_{24} &= \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\tilde{\sigma}_{34} &= \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}
\end{aligned}$$

La matriz resultante $\tilde{\Sigma}$ es identificable. En ella es importante trazar sus componentes originales de la matriz sigma porque nos ayudan a darle interpretación a los resultados obtenidos en la estimación.

4.3.2. Incorporación de restricciones estructurales

Aunque completamente general, la normalización descrita en la sección anterior, muchas veces puede ser algo inconveniente en cuanto los parámetros estimados no tienen interpretación directa. Un enfoque que permite interpretar directamente los parámetros se obtiene al imponer estructura sobre la matriz de varianza-covarianza a partir de supuestos de comportamiento. Por ejemplo, podemos imponer que las componentes aleatorias de algunos pares de alternativas no están correlacionadas o que algún grupo de alternativas tiene la misma variabilidad de la componente no observable. El cuadro 4.1 ejemplifica algunas de las estructuras de varianza-covarianza comúnmente usadas en la literatura.

Otros modelos usados en la literatura y que están implementadas en aplicaciones comerciales incluyen estructuras de bandas, Huynh-Feldt, autoregresivo heterogéneo y simetría compuesta. Como en otros aspectos de la modelación, la elección de la estructura a elegir para la matriz de varianza-covarianza dependerá de las hipótesis de comportamiento que tengamos a la mano y la dificultad numérica de estimar el modelo resultante.

ESTRUCTURA	DESCRIPCIÓN	EJEMPLO
Solo Componentes de Varianza	Las componentes aleatorias de todas las alternativas son independientes entre si.	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \cdot & \sigma_2^2 & 0 & 0 \\ \cdot & \cdot & \sigma_3^2 & 0 \\ \cdot & \cdot & \cdot & \sigma_4^2 \end{bmatrix}$
Autoregresivo	Las varianzas son homogéneas y las correlaciones disminuyen exponencialmente con la distancia.	$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \cdot & 1 & \rho & \rho^2 \\ \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & 1 \end{bmatrix}$
Toeplitz Heterogéneo	Generaliza el modelo autoregresivo permitiendo que las correlaciones disminuyan con un patrón diferente al exponencial y que las varianzas en la diagonal sean heterogéneas.	$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \cdot & \sigma_1^2 & \rho_1 & \rho_2 \\ \cdot & \cdot & \sigma_2^2 & \rho_1 \\ \cdot & \cdot & \cdot & \sigma_3^2 \end{bmatrix}$
Simetría Compuesta	Varianzas y Covarianzas homogéneas (pero distintas entre ellas). Aunque aparentemente arbitraria estructuras de varianza-covarianza como esta aparecen frecuentemente en modelos de regresión y análisis de diseños experimentales.	$\begin{bmatrix} \sigma^2 & \rho & \rho & \rho \\ \cdot & \sigma^2 & \rho & \rho \\ \cdot & \cdot & \sigma^2 & \rho \\ \cdot & \cdot & \cdot & \sigma^2 \end{bmatrix}$

Cuadro 4.1: Tabla 1

Ejemplo: Suponga un modelo de elección en una categoría con 4 marcas de las cuales las 2 primeras son del tipo A (e.g. marca regular) y las dos ultimas son del tipo B (e.g marca premium). En esta situación podríamos combinar algunos de los elementos antes expuestos para decir que las componentes aleatorias son independientes por tipo y para cada tipo tenemos varianzas y covarianzas homogéneas.

$$\Sigma = \begin{bmatrix} \sigma_A^2 & \rho_A & 0 & 0 \\ \cdot & \sigma_A^2 & 0 & 0 \\ \cdot & \cdot & \sigma_B^2 & \rho_B \\ \cdot & \cdot & \cdot & \sigma_B^2 \end{bmatrix} \tag{4.13}$$

Al agregar estructura a la matriz Σ, al reducir el numero de parámetros, típicamente alcanzamos identificación. Sin embargo, para cada estructura tenemos que verificar formalmente que el modelo es identificable. Para eso, podemos usar el procedimiento general de identificación descrito en la sección anterior. Para eso, el analista debe especificar la matriz Σ con la estructura deseada y calcular la matriz $\tilde{\Sigma}$ normalizada por locación y escala. Finalmente si todos los parámetros de Σ pueden ser calculados a partir de los parámetros identificables de $\tilde{\Sigma}$ entonces la estructura propuesta para Σ es identificable y podemos estimarla directamente.

4.4. Estimación

Para estimar el modelo, podríamos proceder tratando de maximizar la log-verosimilitud del modelo como hicimos para el modelo *logit*. Si suponemos que la utilidad de cada tomador de decisión n deriva por cada alternativa i depende de un vector de parámetros θ, entonces, el estimador máximo-verosímil viene dado por:

$$\hat{\theta} = \arg\max LL(\theta) = \arg\max \sum_n \sum_i y_{ni} \ln(P_{ni}(\theta)) \tag{4.14}$$

Lamentablemente, para el modelo probit no disponemos de una expresión cerrada para calcular la probabilidad de elección P_{ni} ya que requiere integrar sobre la densidad de una distribución normal la que no tiene primitiva conocida. Para estimar el modelo entonces aproximamos P_{ni} numéricamente en lo que llamaremos el método de máxima verosimilitud simulada (SML). Sea $\{\varepsilon_n^r\}_{r=1}^R$ una muestra de R vectores aleatorios independientes e idénticamente distribuidos $N(0, \Sigma)$. Entonces, la probabilidad de elección puede aproximarse como:

$$P_{ni} = \int \mathbf{1}(v_{ni} + \varepsilon_{ni} > v_{nj} + \varepsilon_{nj},\ \forall j \neq i)\phi(\varepsilon_n)d\varepsilon_n \approx \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}(v_{ni} + \varepsilon_{ni}^r > v_{nj} + \varepsilon_{nj}^r,\ \forall j \neq i) \tag{4.15}$$

La aproximación numérica de la log-verosimiltud es llamada por algún método numérico de optimización (e.g. método de Raphson-Newton o de descenso de gradiente) los que cada vez que requiere evaluar la función objetivo en algún punto del espacio de parámetros, genera R valores normales multivariados y se calcula la proporción de veces en que cada alternativa genera la mayor utilidad. Con esto el método de optimización puede elegir un nuevo punto del espacio de parámetros con mejor verosimilitud. Notar que la verosimilitud es maximizada con respecto a θ que incluye tanto los parámetros de que definen la componente determinística de la utilidad como los de la matriz de varianza-covarianza de la componente aleatoria.

Algunos aspectos técnicos son importantes de discutir respecto al método de la máxima verosimilitud simulada.

- Al aumentar el valor de R, la aproximación numérica de P_{ni} puede ser arbitrariamente precisa. En efecto, la aproximación numérica es simplemente la media que es un estimador consistente que converge a tasa $\sqrt{R}$.
- Lamentablemente, para un valor fijo de R, los estimadores derivados de maximizar la log-verosimilitud simulada no son consistentes por lo que al aumentar el numero de observaciones no garantiza que el estimador se aproximará al valor verdadero del parámetro. Esto se explica porque a pesar de que la media es un estimador consistente de P_{ni}, al tomar logaritmo estamos aplicando una transformación no lineal damos un mayor peso a los errores negativos que a los positivos y por tanto ellos no se cancelan sesgando al estimador.
- La elección de R es importante para el buen desempeño del estimador para lo que se debe balancear la reducción del sesgo y el costo computacional de aumentar el tamaño muestral.
- Existen alternativas para estimar modelos *probits* como son el métodos de los momentos simulados o método de los scores simulados que aunque consistentes pueden ser ineficientes desde un punto de vista estadístico (MMS) y computacional (MSS).

Hemos planteado que podemos aproximar la probabilidad de elegir cada alternativa como la proporción de veces que dicha alternativa reporta la mayor utilidad en una muestra de tamaño R en un enfoque que llamaremos *frecuencias crudas*. Este enfoque tiene dos limitaciones importantes. Primero, las estimaciones de las probabilidades de elección no son continuas en los parámetros lo que dificulta la aplicación de rutinas de optimización como los métodos del gradiente o de Newton-Raphson. Segundo, para muestras finitas el enfoque de frecuencias crudas puede sugerir probabilidades de elección nulas para algunas de las alternativas, especialmente aquellas con menor componentes determinísticas de la utilidad. Probabilidades de elección nulas son problemáticas porque el logaritmo no esta definido en cero. Ambas dificultades pueden ser corregidas reemplazando el estimador de frecuencias crudas por uno de de frecuencias *suavizadas* donde reemplazamos la función indicatriz $\mathbf{1}(\cdot)$ por una función logística que da una probabilidad positiva a todas las alternativas.

$$P_{ni} \approx \frac{1}{R} \sum_{r=1}^{R} \frac{\exp((v_{ni} + \varepsilon_{ni}^r)/\lambda)}{\sum_j \exp((v_{nj} + \varepsilon_{nj}^r)/\lambda)} \tag{4.16}$$

donde $\lambda > 0$ es un parámetro que permite controlar la suavidad de la curva (entre más cercano a 0 sea el parámetro, la curva más se aproximará a la función indicatriz).

Aunque conceptualmente sencillos, los métodos recién descritos no necesariamente son los más usados en aplicaciones recientes. Uno de los métodos más usados en la actualidad es el de GHK (Geweke, Hajivassiliou y Keane) cuya derivación esta más allá de los alcances de este apunte.

Capítulo 5

Mixed Logit

5.1. Probabilidad de elección

Como hemos visto en secciones anteriores, en el modelo logit los tomadores de decisiones eligen la alternativa i que representa la mayor utilidad frente a sus pares, siendo uno de los supuestos fundamentales de aquella formulación la suposición de la componente aleatoria siguiendo una distribución valor extremo. Dado lo anterior, se obtiene

$$P_{ni} = \frac{e^{v_{ni}(\beta)}}{\sum_j e^{v_{nj}(\beta)}}$$

Para la fomulación del modelo *mixed* logit, asumiremos que los parámetros siguen una distribución con densidad $f(\beta)$ a lo largo de la población. Por lo tanto, para obtener la probabilidad incondicional P_{ni}, se deberá integrar la probabilidad de elección del modelo logit estandar a lo largo de todos los posibles valores de β

$$P_{ni} = \int \left(\frac{e^{v_{ni}(\beta)}}{\sum_j e^{v_{nj}(\beta)}} \right) f(\beta) d\beta$$

Donde $v_{ni}(\beta)$ es la porción observable de la utilidad del individuo n al elegir la alternativa i (que, claramente, depende del valor de β). Si se asume linealidad en la función de utilidad, ie, $v_{ni}(\beta) = \beta' x_{ni}$, la formulación toma la siguiente forma

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{ni}}} \right) f(\beta) d\beta$$

Notar que el modelo logit estándar es un caso particular del *mixed* logit cuando la distribución de β es degenerada. Esto es, cuando

$$f(\beta) = \begin{cases} 1 & \text{si } \beta = b \\ 0 & \text{si } \beta \neq b \end{cases}$$

Por lo que se recupera la probabilidad

$$P_{ni} = \frac{e^{b' x_{ni}}}{\sum_j e^{b' x_{ni}}}$$

Otro aspecto a considerar, es que la distribución de β no posee restricción alguna (a priori), y en consecuencia, puede tomar configuraciones discretas o continuas según la naturaleza del problema modelado.

Suponiendo que β toma M posibles valores discretos $b_1, \ldots, b_M$ con probabilidad s_m, el *mixed* logit se reduce a un modelo de clases latentes con probabilidad de elección

$$P_{ni} = \sum_{m=1}^{M} s_m \left(\frac{e^{b'_m x_{ni}}}{\sum_j e^{b'_m x_{ni}}} \right)$$

Por otro lado, se podría asumir que β proviene de una distribución normal con media μ y varianza Σ, dando como resultado una probabilidad de elección con la siguiente estructura

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{ni}}} \right) \phi(\beta|\mu, \Sigma) d\beta$$

Cabe destacar que la elección de la distribución a usar por parte del modelador, va acorde a las expectativas del mismo respecto al comportamiento estudiado, y en ese sentido, el modelo mixto entrega alta versatilidad a distintos configuraciones.

Interpretación 1: Coeficientes aleatorios

El tomador de decisiones enfrenta su elección entre J alternativas, siendo la utilidad del individuo n al elegir j

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

Donde x_{nj} son variables observables relativas al tomador de decisiones, β_n es un vector de coeficientes por individuo que refleja los *gustos* de n, y ε_{nj} es un término aleatorio iid valor extremo. El vector de coeficientes que refleja los gustos de los individuos varia a lo largo de la población con densidad $f(\beta)$.

Suponiendo que los agentes son maximizadores de utilidad, se tendrá que n eligirá i si y solo si $U_{ni} > U_{nj}\ \forall j \neq i$, por lo que, condicional al valor de β_n, la probabilidad de elección se reduce al modelo logit estandar

$$P_{ni}|\beta_n = \frac{e^{\beta'_n x_{ni}}}{\sum_j e^{\beta'_n x_{ni}}}$$

Sin embargo, dado que el valor de β_n distribuye a lo largo de la población, para poder encontrar el valor de P_{ni} es necesario integrar respecto a todos los posibles valores que puede tomar β_n, esto es

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{ni}}} \right) f(\beta) d\beta$$

Interpretación 2: Componentes del error

El modelo mix logit puede ser interpretado de acuerdo a las componentes de los errores, los cuales crean correlación entre las utilidades para las diferentes opciones. La utilidad es

$$U_{nj} = \alpha' x_{nj} + \mu_n' z_{nj} + \varepsilon_{nj}$$

Donde x_{nj} y z_{nj} son vectores de variables observables relacionados a la alternativa j, α es un vector de coeficientes fijos, μ es un vector aleatorio con esperanza 0, y ε_{nj} es un término iid que distribuye valor extremo.

Dado lo anterior, el término aleatorio de la utilidad viene dado por $\eta_{nj} = \mu_n' z_{nj} + \varepsilon_{nj}$, el cual puede poseer correlación dependiendo de la especificación de z_{nj}. Si z_{nj} no es idénticamente cero, se tiene que

$$\begin{aligned} Cov(\eta_{ni}, \eta_{nj}) &= Cov(\mu_n' z_{ni} + \varepsilon_{ni}, \mu_n' z_{nj} + \varepsilon_{nj}) \\ &= Cov(\mu_n' z_{ni}, \mu_n' z_{nj}) + Cov(\mu_n' z_{ni}, \varepsilon_{nj}) + Cov(\varepsilon_{ni}, \mu_n' z_{nj}) + Cov(\varepsilon_{ni}, \varepsilon_{nj}) \\ &= z_{ni}' Var(\mu_n') z_{nj} \\ &= z_{ni}' W z_{nj} \end{aligned}$$

Notar que, existe correlación incluso cuando las componentes de los errores son independientes, en cuyo caso W es una matriz diagonal.

La interpretación de coeficientes aleatorios y la de componentes del error son equivalentes. Bajo el enfoque de coeficientes aleatorios, la utilidad es $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, por lo que, descomponiendo el vector β_n en una componente con su media α y otra con su desviación μ_n se obtiene que $U_{nj} = \alpha' x_{nj} + \mu_n' x_{nj} + \varepsilon_{nj}$. Finalmente, haciendo $x_{nj} = z_{nj}$ se obtiene la equivalencia.

5.2. Patrones de sustitución

El enfoque mix logit no posee independencia de alternativas irrelevantes (IIA). Esto pues el ratio P_{ni}/P_{nj} depende de todas las alternativas disponibles (notar que, a diferencia del modelo logit estandar, los términos del denominador no se cancelan debido a que estan dentro de la integral).

Por otro lado, el modelo mix logit tampoco posee la propiedad de patrones de sustitución proporcionales, puesto que

$$\begin{aligned} E_{ni,x_{nj}^m} &= \frac{\partial P_{ni}}{\partial x_{nj}^m} \cdot \frac{x_{nj}^m}{P_{ni}} \\ &= \frac{x_{nj}^m}{P_{ni}} \int \frac{\partial}{\partial x_{nj}^m} \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{ni}}} \right) f(\beta) d\beta \\ &= -\frac{x_{nj}^m}{P_{ni}} \int \frac{e^{\beta' x_{ni}} e^{\beta' x_{nj}}}{\left(\sum_j e^{\beta' x_{ni}} \right)^2} \beta^m f(\beta) d\beta \\ &= -\frac{x_{nj}^m}{P_{ni}} \int \beta^m P_{ni}(\beta) P_{nj}(\beta) f(\beta) d\beta \end{aligned}$$

Notar que la elasticidad depende de la correlación que exista entre $P_{ni}(\beta)$ y $P_{nj}(\beta)$ a lo largo d los valores de β.

5.3. Estimación

En la mayoría de los casos, para el modelo *mixed* logit no existe una formula cerrada para la probabilidad de elección P_{ni}. Es por esto que para su estimación, es necesario recurrir a métodos de simulación que aproximen de manera adecuada dicha probabilidad.

Al igual que el modelo probit, el *mixed* logit puede ser estimado usando métodos numéricos, como por ejemplo, el método de máxima verosimilitud simulada (SML).

Parte III

Apéndices Técnicos

Capítulo 6

Métodos de estimación y evaluación de modelos

6.1. Método de máxima verosimilitud

Existen dos métodos generales para estimar parámetros de un modelo. Uno es estimación por *Mínimos cuadrados* (LSE, por sus siglas en inglés) y el otro *Máxima verosimilitud*. Este último método, si bien requiere supuestos importantes de distribución, cuenta con propiedades deseables, tales como:

- **Suficiencia**: Toda la información de interés sobre el parámetro la toma en cuenta este estimador.
- **Consistencia**: El verdadero valor del parámetro que genera la data se recupera asintóticamente, es decir, para datas con muestras lo suficientemente grande.
- **Eficiencia**: Asintóticamente la varianza del parámetro es cero.
- **Parametrización invariante**: Se obtiene la misma solución independiente de la parametrización usada.

Además, muchos métodos de inferencia son desarrollados en base a *Máxima verosimilitud*, tales como *métodos bayesianos, modelos con efectos aleatorios, modelos de selección de criterios* como *Akaike information criterion* y *Bayesian information criteria*.

Consideremos un vector de datos $y = (y_1, y_2, ..., y_m)$, una muestra aleatoria de una población. El objetivo es idenficar los parámetros que más probablemente hayan generado la muestra. Cada población es identificada con una distribución, la cual tiene asociada los parámetros que se buscan.

Denotemos $f(y|w)$ la función de densidad, que refleja la probabilidad de observar y dado el parámetro w. Si asumimos que las observaciones y_i son estadísticamente independientes, la función $f(y|w)$ podemos expresarla como una multiplicación de las observaciones individuales,

$$f(y|w) = f_1(y_1|w)f_2(y_2|w)\cdots f_m(y_m|w) \tag{6.1}$$

Dado un set de valores de parámetros, $f(y|w)$ mostrará qué data es más probable que otra. Desafortunadamente, el problema al que uno se enfrenta en realidad es al revés: Dado la data observada y un modelo definido, queremos encontrar una función de probabilidad que con mayor

seguridad haya producido la data. Para resolver este problema definamos la función de verosimilitud como

$$L(w|y) = f(y|w) \tag{6.2}$$

Esta es una función de w dado y, por lo que $f(y|w)$ y $L(w|y)$ son dos funciones definidas en ejes distintos, por lo tanto, no son directamente comparables.

Una vez definida la función, lo que se busca es el valor w tal que la máximice. Para ello, necesitamos que el óptimo exista y sea único. Asumiendo que se cumple esto, obtenemos el valor estimado usando la función log-verosimilitud, puesto que se comporta mejor computacionalmente. El w que encontremos no será distinto si usamos esta función porque la dos funciones se relacionan monótonamente. Si asumimos que la función $lnL(w|y) = LL(w|y)$ es diferenciable y w existe, se deben satisfacer las siguientes condiciones

$$\frac{\partial LL(w|y)}{\partial w_i} = 0 \qquad \forall i \tag{6.3}$$

$$\frac{\partial^2 LL(w|y)}{\partial w_i^2} < 0 \qquad \forall i \tag{6.4}$$

En la práctica, sin embargo, es común que no encontremos una solución analítica, especialmente cuando el modelo involucra muchos parámetros y la función de probabilidad es altamente no lineal. En ese caso es mejor estimar numéricamente usando algoritmos de optimización no lineales. Por lo general, estos métodos realizan la búsqueda en subconjuntos más pequeños, y de forma iterativa, modificando los parámetros que se obtienen de la iteración anterior. Estos algoritmos tienen como criterio de parada, ya sea un número máximo de iteraciones o un mínimo cambio que debe existir entre una iteración y otra.

A veces pueden no garantizar que el único set de parámetros que maximiza la log-verosimilitud sea encontrado. El algoritmo puede detenerse en un sub-óptimo local. Desafortunadamente, no existe una solución general para el problema de máximo local, pero si se han desarrollado una serie de procedimientos que evitan este problema.

En general, el método de máxima verosimilitud se prefiere sobre mínimos cuadrados, a menos que la densidad de probabilidad sea desconocida o difícil de obtener. En otros casos los resultados encontrados por los dos métodos pueden coincidir. Esto ocurre cuando las observaciones son independientes, y se encuentran normalmente distribuidas con varianza constante.

Ejemplo 1: Consideremos el caso simple de una observación y un parámetro. Además, los datos representarán el número de eventos en 10 lazamientos Bernoulli (monedas por ejemplo) de parámetro w. Suponiendo que $y = 7$ tenemos que

$$\begin{aligned} L(w|y=7) &= f(y=7|w) \\ &= \frac{10!}{7!3!} w^7 (1-w)^3 \qquad 0 \le w \le 1 \end{aligned}$$

En la Figura 6.1 se grafica la función de verosimilitud.

Tomando logaritmo y derivando,

$$\begin{aligned} LL(w|y=7) &= \ln(10!) - \ln(7!) - \ln(3!) + 7\ln(w) + 3\ln(1-w) \\ \frac{dLL(w|y=7)}{dw} &= \frac{7}{w} - \frac{3}{1-w} \end{aligned}$$

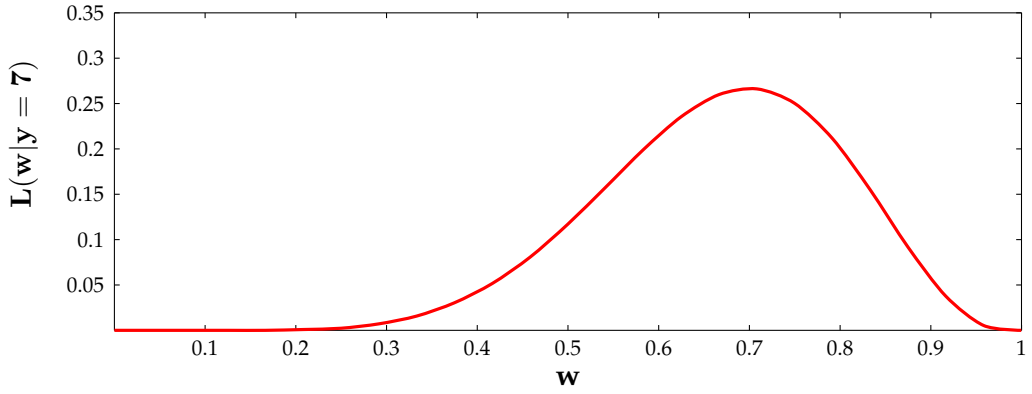

Figura 6.1: Función de verosimilitud con $y = 7$

lo cual a igualar a cero se encuentra $w = 0.7$. Finalmente se puede verificar que es un óptimo,

$$\begin{aligned}\frac{d^2LL(w|y=7)}{dw^2} &= -\frac{7}{w^2} - \frac{3}{(1-w)^2} \\ &= -47.62 < 0\end{aligned}$$

Ejemplo 2: Sea $y = y_1, y_2, ..., y_n$ una muestra aleatoria de una distribución normal con media μ y varianza σ^2. Para obtener los estimadores de máxima verosimilitud de μ y σ^2, primero debemos notar que y_i son variables aleatorias continuas independientes, por lo que $L(\mu, \sigma^2|y)$ es la multiplicación de las densidades de probabilidad.

$$\begin{aligned}L(\mu, \sigma|y) &= f(y|\mu, \sigma^2) \\ &= f(y_1|\mu, \sigma^2)f(y_2|\mu, \sigma^2)\cdots f(y_n|\mu, \sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)\end{aligned}$$

Tomando logaritmo,

$$LL(\mu, \sigma^2|y) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

Los estimadores de máxima verosimilitud de μ y σ^2 son los valores que maximizan $LL(\mu, \sigma^2|y)$. Si tomamos derivadas respecto a μ y σ^2, obtenemos

$$\frac{\partial LL(\mu, \sigma^2|y)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)$$

y

$$\frac{\partial LL(\mu, \sigma^2|y)}{\partial \sigma^2} = -\frac{2}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2$$

Si igualamos estas derivadas a cero simultáneamente, de la primera ecuación obtenemos

$$\sum_{i=1}^{n} y_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Al sustituir $\bar{y}$ por $\hat{\mu}$ en la segunda ecuación y despejar $\hat{\sigma^2}$, llegamos a

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Por consiguiente, $\bar{y}$ y $\hat{\sigma^2}$ son los estimadores de máxima verosimilitud de μ y σ^2 respectivamente. Cabe destacar que $\bar{y}$ es insesgado para μ, mientras que $\hat{\sigma^2}$ no lo es, pero que se puede ajustar fácilmente al estimador insesgado.

6.2. Métricas de ajuste

- R^2, coeficiente de determinación. $R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y_i})^2}$, con $\hat{y_i}$ predicción, $\bar{y}$ media e y observación real. Permite ver la varianza explicada por el modelo.
- MAE, $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}|$. Permite ver si el modelo es bueno o no.
- MAPE, $MAPE = \frac{1}{n}\sum_i \frac{|y_i - \hat{y}|}{y_i}$. Es análogo a MAE.

6.3. Test de bondad de ajuste

Permite testear si el modelo es suficientemente bueno. Se plantea H_0 : el modelo probabilística describe bien la data que observamos. Se rechaza la hipótesis nula si:

$$\chi^2 = \sum_n \frac{(y_i - \hat{y}}{\hat{y})} \geq \chi^2_{N_1,\alpha} \tag{6.5}$$

donde $P(\chi^2 \geq \chi^2_{N-1,\alpha}) = \alpha$.

6.4. Test de ratio de verosimilitud

Permite ver si vale la pena complejizar un modelo. Para ello, compara un modelo A con uno B anidado (es decir, que imponiendo restricciones sobre los parámetro de A se puede obtener B).

H_0: el modelo A es mejor que el B.

Se rechaza la hipótesis nula si:

$$LR = 2 \cdot (LL_A - LL_B) \leq \chi^2_{N,\alpha}$$