Tecnicas para la estimacion de modelos MEV con muestreo de alternativas

Ricardo Hurtubia, Gunnar Flötteröd, Michel Bierlaire

Universidad de Chile - Modelos avanzados de demanda

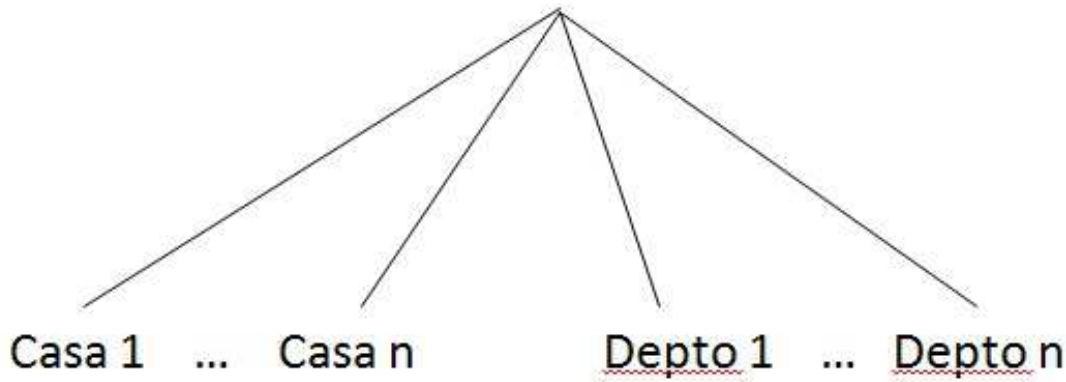
Santiago, 20 de octubre, 2010

Motivación

- Los modelos de eleccion discreta con gran numero de alternativas deben ser estimados utilizando una muestra de alternativas
- En el caso de un modelo tipo Logit (MNL), es posible obtener parametros consistentes (McFadden, 1978)
- Modelos más complejos (otra estructura de correlacion de los errores) no pueden ser directamente estimados

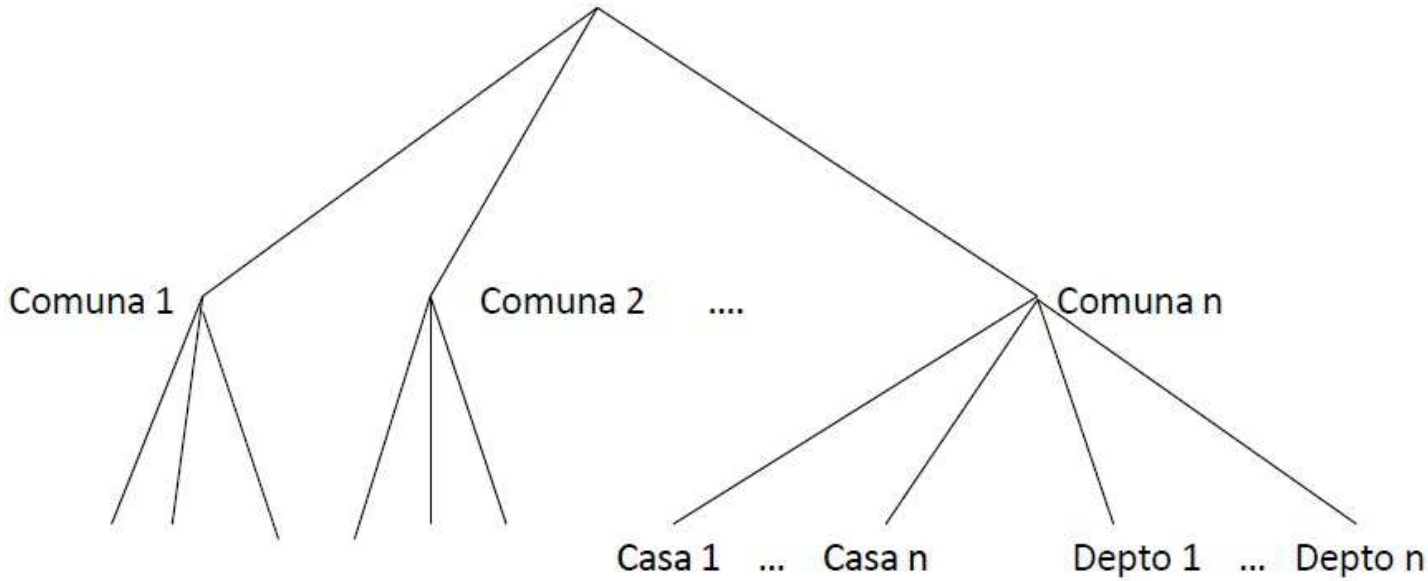
Motivación

- Ejemplo: modelación de la elección de localización residencial (logit)



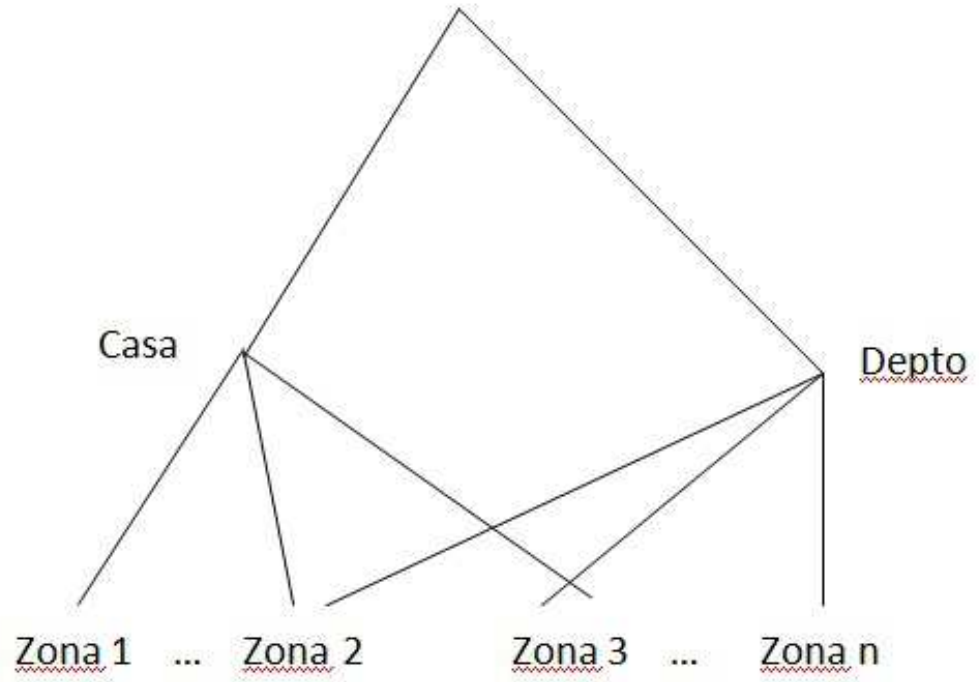
Motivación

- Ejemplo: modelación de la elección de localización residencial (logit jerárquico)



Motivación

- Ejemplo: modelación de la elección de localización residencial (Cross-nested logit)



Motivación

- Guevara y Ben-Akiva (2010) proponen un metodo para estimar modelos MEV con una muestra de alternativas. La estimacion es consistente para muestras grandes, pero muestras pequeñas generan parametros sesgados
- El sesgo puede ser reducido utilizando diferentes tecnicas: bootstrapping o importance sampling

Outline

1. Muestreo de alternativas para logit
2. Modelos MEV
3. Muestreo de alternativas para modelos MEV
4. Metodos para estimación con sesgo reducido
   - 4.1 Bootstrapping
   - 4.2 Importance Sampling
5. Conclusiones

Muestreo de alternativas para MNL

- Probabilidad de elección con set completo de alternativas

$$P(i) = \frac{e^{V_{ni}}}{\sum_{j \in C_n} e^{V_{nj}}}$$

- Probabilidad de elección con una muestra D_n (McFadden, 1978):

$$P(i|D_n) = \frac{e^{\mu V_{ni} + \ln \pi(D_n|i)}}{\sum_{j \in D_n} e^{\mu V_{nj} + \ln \pi(D_n|j)}}$$

- La generalización de este resultado a modelos MEV no es directa

Modelos MEV

- Función generadora $G(e^{V_1}, ..., e^{V_J}) \rightarrow$ define la estructura de correlacion de los errores
- Probabilidad de elección (similar a un Logit):

$$P_n(i) = \frac{e^{V_{in} + \ln G_i}}{\sum_{j \in C_n} e^{V_{jn} + \ln G_j}}$$

- donde $G_i = \frac{\partial G(e^{V_{1n}}, e^{V_{2n}}, ..., e^{V_{Jn}})}{\partial e^{V_{in}}}$

Modelos MEV

- Diferentes funciones G generan distintos tipos de modelos:
  - Logit: $G = \sum_{j \in C_n} e^{\mu V_{jn}}$
  - Nested Logit: $G = \sum_{m=1}^{M} \left(\sum_{j \in C_{mn}} e^{\mu_m V_{jn}} \right)^{\frac{\mu}{\mu_m}}$
  - Cross-nested Logit: $G = \sum_{m=1}^{M} \left(\sum_{j \in C_{mn}} \left(\alpha_{jm} e^{V_{in}} \right)^{\mu_m} \right)^{\frac{\mu}{\mu_m}}$

Muestro de alternativas para modelos MEV

- Probabilidad de elección considerando una muestra D_n (Bierlaire, Bolduc and McFadden, 2008):

$$P_n(i|D_n) = \frac{e^{V_{in}+\ln G_i+\ln \pi(D_n|i)}}{\sum_{j\in D_n} e^{V_{jn}+\ln G_j+\ln \pi(D_n|j)}}$$

- En muchos casos (NL, CNL) $\ln G_i$ dependerá de todos los elementos dentro del set completo de alternativas C_n

Muestro de alternativas para modelos MEV

- En el caso de un Nested Logit:

$$\ln G_{in} = \left(\frac{\mu}{\mu_{m(i)}} - 1 \right) \left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) + \ln \mu + (\mu_{m(i)} - 1) V_{in}$$

- la logsuma depende de $C_{m(i)n}$

Muestro de alternativas para modelos MEV

- Aproximación de la logsuma (Guevara and Ben-Akiva, 2010):

$$\left(\ln \sum_{j \in C_{m(i)n}} e^{\mu_{m(i)} V_{jn}} \right) \approx \left(\ln \sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}} \right)$$

- con $w_{jn} = \frac{\widetilde{n}_{jn}}{E_n(j)}$

(numero de veces que j es seleccionada / n° esperado de veces que j es seleccionada)

Muestro de alternativas para modelos MEV

- La muestra D_{mn} incluye la alternativa elegida y un conjunto de alternativas escogidas aleatoriamente dentro del nido
- Estimación a través de maxima log-verosimilitud, usando la siguiente probabilidad

$$P_n(i|D_n) = \frac{e^{V_{in}+\ln G_i(D_{m(i)n})+\ln \frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j\in D_n} e^{V_{jn}+\ln G_j(D_{m(j)n})+\ln \frac{|C_{m(j)}|}{|D_{m(j)n}|}}}$$

- donde $\ln G_i(D_{m(i)n})$ utiliza la logsuma aproximada

Muestro de alternativas para modelos MEV

- La logsuma aproximada genera parametros insesgados cuando $D_n = C_n$

$\rightarrow$ cuando la muestra es relativamente pequeña, se obtienen parametros sesgados (incluso cuando w_{jn} es calculado utilizando las probabilidades de eleccion reales reales)

- Posibles metodos para "mejorar" la aproximacion de la logsuma
  - Correccion del sesgo usando Bootstrapping
  - Muestreo por importancia de los elementos de la logsuma

Bootstrapping

- Técnica basada en simulacion para la inferencia estadísitica de las propiedades de un estimador
- Ejemplo simple?

Bootstrapping

- Método aplicado a la aproximación de la logsuma
  1. Estimación inicial utilizando aproximacion de la logsuma
  2. Re-muestro de los elementos en la muestra
  3. Cálculo de la logsuma con las nuevas muestras
  4. Cálculo del sesgo
  5. Re-estimación corrigiendo el sesgo

Bootstrapping

- Estimador "bootstrap" del sesgo:

$$\rho_{mn} = \frac{1}{B} \sum_{b}^{B} \left(\ln \sum_{j \in D_{mn}^{b}} w_{jn} e^{\mu_m^0 V_{jn}(\beta^0)} \right) - \left(\ln \sum_{j \in D_{mn}} w_{jn} e^{\mu_m^0 V_{jn}(\beta^0)} \right)$$

- donde
  - β^0, μ^0 : set de parametros de la estimacion inicial
  - D_{mn}^{b}: set de alternativas cada instancia de re-muestreo (b)
  - B : número de instancias de re-muestreo

Bootstrapping

- Estimacion via maxima log-verosimilitud usando la siguiente probabilidad:

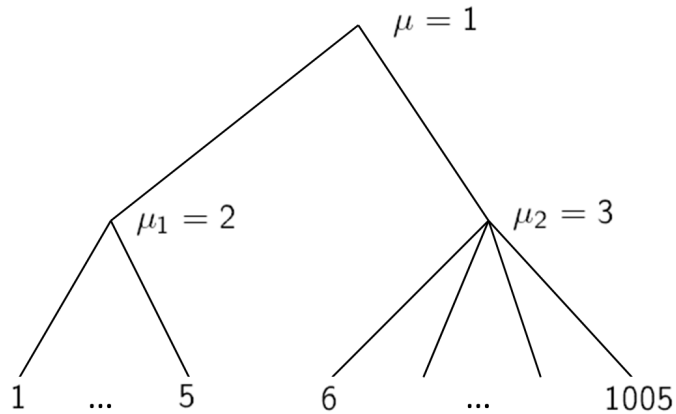
$$P_n(i|D_n) = \frac{e^{V_{in} + \ln \hat{G}_i(D_{m(i)n}) + \ln \frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j \in D_n} e^{V_{jn} + \ln \hat{G}_j(D_{m(j)n}) + \ln \frac{|C_{m(j)}|}{|D_{m(j)n}|}}}$$

donde:

$$\ln \hat{G}_i(D_{m(i)n}) = \left(\frac{\mu}{\mu_{m(i)}} - 1\right)\left(\left(\ln \sum_{j \in D_{m(i)n}} w_{jn} e^{\mu_{m(i)} V_{jn}}\right) - \rho_{m(i)n}\right) + \ln \mu + (\mu_{m(i)} - 1) V_{in}$$

Bootstrapping: Experimento

- Nested logit:



- Utilidades: $V_{in} = \beta_a a_{in} + \beta_b b_{in}$
- Atributos: $a_{in}, b_{in} \sim U(-1, 1)$
- Parametros "reales" $\beta_a = 1$, $\beta_b = 1$, $\mu_1 = 2$, $\mu_2 = 3$
- Muestreo de alternativas solo dentro del nido 2

Bootstrapping: Resultados

- Simualción de Monte Carlo usanlo la logsuma aproximada (tamaño de la muestra = 5, probabilidades reales para w_{jn}):

parameter	average value	std-error	true value	t-test
β_a	0.855	0.082	1	1.773
β_b	0.843	0.068	1	2.288 *
μ_1	2.569	0.581	2	0.978
μ_2	3.622	0.272	3	2.290 *

* Parámetros sesgados

Bootstrapping: Resultados

- Resultados despues de aplicar bootstrapping (tamaño de la muestra = 5):

parameter	average value	std-error	true value	t-test
β_a	0.953	0.079	1	0.595
β_b	0.957	0.079	1	0.548
μ_1	2.264	0.517	2	0.511
μ_2	3.224	0.229	3	0.974

- reducción significativa del sesgo

Importance sampling

- El sesgo puede ser reducido si tenemos una mejor muestra de alternativas para la logsuma
- Las alternativas incluidas en la muestra de la logsuma no tienen por que ser las mismas del choice set
- Metodo:
  1. Inicialmente, muestra aleatoria para las alternativas en la loguma
  2. Estimación utilizando logsuma aproximada $\rightarrow \beta^0, \mu^0$
  3. Importance sampling: nueva muestra de alternativas para la logsuma siguiendo $P(\beta^0, \mu^0)$
  4. Re-estimación

Importance sampling

- Primera estimacion:

$$P_n(i|D_n) = \frac{e^{V_{in} + \ln G_i(L_{m(i)n}) + \ln \frac{|C_{m(i)}|}{|D_{m(i)n}|}}}{\sum_{j \in D_n} e^{V_{jn} + \ln G_i(L_{m(j)n}) + \ln \frac{|C_{m(j)}|}{|D_{m(j)n}|}}}$$

- donde $L_{m(i)n}$ es la muestra aleatoria de alternativas en la logsuma ($|L_{m(i)n}| = |D_{m(i)n}|$)
- De esta estimacion obtenemos β^0, μ^0

Importance sampling

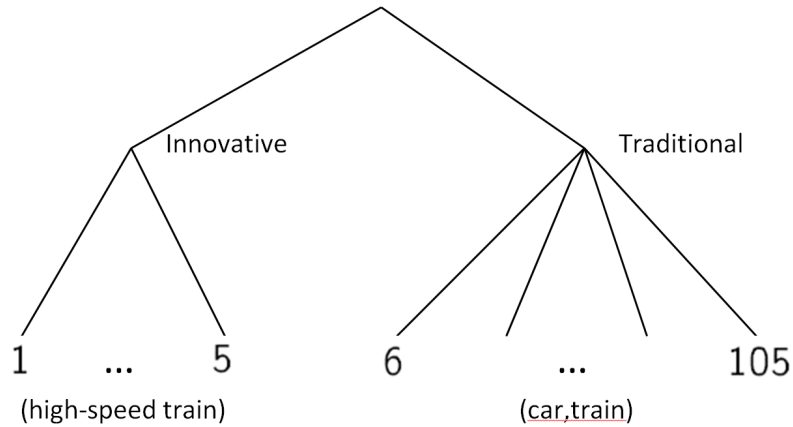
- La muestra para la logsuma es generada nuevamente, con la siguiente probabilidad de selección

$$g_n(i|m) = \frac{e^{V_{ni}(\beta^0,\mu^0)}}{\sum_{j \in C_m} e^{V_{nj}(\beta^0,\mu^0)}}$$

- Los elementos en la muestra del choice set se mantienen
- Una nueva estimacion es realizada

Importance sampling: Experimento

- Datos sinteticos generados a partir de una encuesta (SP/PD) para evaluar un tren de alta velocidad en Suiza



- $V_{hs} = \beta_{cost} C_{hs} + \beta_{time\_T} TT_{hs} + \beta_{headway} HE_{hs}$
- $V_{car} = \beta_{cost} C_{car} + \beta_{time\_C} TT_{car}$
- $V_{train} = \beta_{cost} C_{train} + \beta_{time\_T} TT_{train} + \beta_{headway} HE_{train}$

Importance sampling:Resultados

- Simualción de Monte Carlo usando la logsuma aproximada (tamaño de la muestra = 5, probabilidades reales para w_{jn}):

parameter	average value	std-error	true value	t-test
β_{cost}	-1.253	0.152	-0.849	2.666 *
β_{time_C}	-2.958	0.359	-1.760	3.388 *
β_{time_T}	-2.708	0.306	-1.840	2.835 *
$\beta_{headway}$	-0.967	0.217	-0.496	2.165 *
μ_1(innovative)	1.220	0.160	2	4.873 *
μ_2(traditional)	3.146	0.368	4	2.318 *

* Parametros sesgados

Importance sampling:Resultados

- Resultados al utilizar importance sampling (sample size = 5):

parameter	average value	std-error	true value	t-test
β_{cost}	-0.930	0.135	-0.849	0.560
β_{time_C}	-1.997	0.321	-1.760	0.736
β_{time_T}	-2.008	0.314	-1.840	0.535
$\beta_{headway}$	-0.592	0.143	-0.496	0.672
μ_1(innovative)	1.766	0.359	2	0.652
μ_2(traditional)	3.503	0.430	4	1.155

- reducción significativa del sesgo

Conclusiones

- Dos metodos para estimar modelos MEV con sesgo reducido
- Bootstrapping reduce el sesgo de cualquier estimador (en este caso la logsuma aprox.)
  - La calidad final de los resultados dependerá de la calidad del estimador original
- Importance sampling para los elementos de la logsuma permite estimar parametros insesgados
  - Diferentes muestras de alternativas en la logsuma y en el choice set
- Trabajo a futuro:
  - Comprobación con otros tipos de modelo (e.g. Cross-nested logit)
  - Estimación con datos reales (localizacion residencial)

Gracias!