ECONOMETRICS II (ECO 2401)

Victor Aguirregabiria

Winter 2018

TOPIC 3: MULTINOMIAL CHOICE MODELS

1. Introduction

2. Nonparametric model

3. Random Utility Models

- Definition; - Common Specification and Normalizations; - Choice Probabilities; - Some Theorems

4. Logit Model

5. Nested Logit Model

6. Random Coefficients Logit Model

7. Monte Carlo Simulation

8. Simulation-Based Estimation

1. INTRODUCTION

• Economics deals with agents' choices.

• Many important economic decisions can be described as discrete choices within a finite number of choice alternatives.

- Consumer choice of store, or brand, or product variety;

- Occupational choice; Migration decisions; School / university choice;

- Firms' decisions of where to locate plants / stores; which products to sell;

- Commuters's choice of transportation mode: car, bus, subway, bicycle, walk, mixed.

...

INTRODUCTION [2]

• Let $\mathcal{J} = \{0, 1, ..., J\}$ be the set of choice alternatives that the agent faces. We index choice alternatives by j.

• Let $Y \in \mathcal{J}$ be the variable that represents the actual choice of an individual. Let X be a vector of exogenous variables such as individual characteristics, and attributes of each choice alternative.

• Using a sample of $\{Y, X\}$ we are interested in learning how X affects Y.

- How prices or other product attributes affect consumer demand;

- How neighborhood amenities and housing prices affect people decisions of where to live;

...

Stylized description of model and estimation

• The model can be described as:

$$Y = h\left(X,\ \varepsilon,\ \beta\right)$$

- ε is a vector of unobservables;

- β is a vector of parameters;

- $h(.)$ is a function that maps (X, ε, β) into the choice set $\mathcal{J}$.

• Define the **Conditional Choice Probability (CCP) function** as the probability distribution of Y conditional on X. For any pair (j, x):

$$P(j \mid x) \equiv \mathsf{Pr}\left(Y = j \mid X = x\right)$$

Note that $P(j \mid x) = \mathbb{E}\left(1\{Y = j\} \mid X = x\right)$.

Stylized description of model and estimation [2]

- Suppose that:

$$\varepsilon \sim \text{ independent of } X \text{ with CDF } \; F(\varepsilon; \Omega)$$

where Ω represents the unknown parameters in this distribution function.

- Model $Y = h\,(X,\ \varepsilon,\ \beta)$ and distribution $F(\varepsilon; \Omega)$ imply a CCP function:

$$P(j \mid x,\ \beta, \Omega) = \int 1\,\{h\,(x,\ \varepsilon,\ \beta) = j\}\ dF(\varepsilon;\ \Omega)$$

where $1\{.\}$ is the indicator function.

Stylized description of model and estimation [3]

• The researcher observes a random sample of N agents, indexed by n, with information on $\{y_n, x_n : n = 1, 2, ..., N\}$. She is interested in the estimation of the parameters $\theta \equiv (\beta, \Omega)$.

• The (conditional) log-likelihood function for this model and data is,

$$\ell_N(\theta) = \sum_{n=1}^{N} \ln \Pr(Y = y_n \mid X = x_n;\ \theta) = \sum_{n=1}^{N} \ln P(y_n \mid x_n,\ \theta)$$

• The MLE is:

$$\widehat{\theta} = \arg\max_{\theta}\ \ell_N(\theta)$$

Predictions / Counterfactual analysis

• Given the estimated model, we can make predictions and counterfactual analysis.

• Let (x^*, θ^*) be a value of (X, θ) that is different in some of its components to the observed/estimated value $(x_n,\widehat{\theta})$; e.g., a change in an attribute of a choice alternative; a change in agents' characteristics; shutting down the effect of a variable ($\theta_k = 0$); removing some choice alternatives; etc.

• We can compare estimated and counterfactual CCPs,

$$P(j|x_n, \widehat{\theta}) \text{ and } P(j|x^*, \theta^*)$$

• This is a helpful exercise for policy analysis or managerial decisions.

2. NONPARAMETRIC MODEL

• Suppose that X has also a discrete & finite support, $X \in \mathcal{X} \equiv \{1, 2, ..., M\}$. Consider a fully nonparametric specification of the CCPs.

• **Nonparametric model:** The vector of parameters is the vector of $M * (J + 1)$ CCPs, $\mathbf{P} = \{P(j|x) : (j, x) \in \mathcal{J} \times \mathcal{X}\}$, with the only restriction that $\sum_{j \in \mathcal{J}} P(j \mid x) = 1$ for any value of x.

• Let $N_{jx} \equiv \sum_{n=1}^{N} 1\{y_n = j \ ; \ x_n = x\}$ be the number of observations in the sample where we observe $(x_n, y_n) = (x, j)$.

NONPARAMETRIC MODEL [2]

• The log-likelihood function is:

$$\begin{aligned}
\ell_N(\mathbf{P}) &= \sum_{n=1}^{N} \ln P(y_n \mid x_n) \\
&= \sum_{n=1}^{N} \left[\sum_{(j,x)\in\mathcal{J}\times\mathcal{X}} 1\{y_n = j \ ; \ x_n = x\} \ \ln P(j \mid x) \right] \\
&= \sum_{(j,x)\in\mathcal{J}\times\mathcal{X}} N_{jx} \ \ln P(j \mid x)
\end{aligned}$$

• Taking into account the restrictions $\sum_{j=0}^{J} P(j \mid x) = 1$ for any value of x, the f.o.c. for the Lagrange problem:

$$\frac{N_{jx}}{P(j \mid x)} - \lambda = 0 \qquad \text{or} \qquad N_{jx} = \lambda \ P(j \mid x)$$

NONPARAMETRIC MODEL [2]

• Then, $\sum_{i \in \mathcal{J}} N_{ix} = \sum_{i \in \mathcal{J}} \lambda \, P(i|\, x) = \lambda$. Therefore, the MLE of the CCP is:

$$\widehat{P}(j \mid x) = \frac{N_{jx}}{\sum_{i \in \mathcal{J}} N_{ix}}$$

• The MLE of this Nonparametric model is just the frequency estimator of the CCPs.

• As usual, this MLE is **consistent, asymptotically normal, and efficient** given the minimal restrictions in this nonparametric model.

Some Limitations of this Nonparametric model

[1] When x is continuous: curse of dimensionality in the speed of asymptotic convergence.

[2] Suppose that some of the X variables are characteristics of the choice alternative (not of the individuals): $\{X_j : j = 1, 2, ..., J\}$. For instance, the price of product j. Suppose that all the individuals face the same choice set (with the same values of X), such that variables X do not have variation over n. Therefore, the nonparametric CCPs $P(j|X)$ do not provide any information about how $P_j(j|X)$ changes when X_j changes (keeping $X_i : i \neq j$ constant) or when $X_i : i \neq j$ changes, e.g., no information about demand price elasticities.

3. RANDOM UTILITY MODELS

• An agent should choose one alternative from a choice set with J mutually exclusive alternatives $\mathcal{J} \equiv \{0, 1, ..., J\}$.

• We use n to index agents, i or j to index alternatives, and k to index explanatory variables.

• Let $Y_n \in \mathcal{J}$ be the random variable that represents the choice of agent n.

Assumption of Utility Maximization: The agent makes this choice to maximize her payoff or utility.

$$Y_n = \arg\max_{j \in \mathcal{J}} \; U_n(j)$$

where $U_n(j)$ is the utility or payoff for agent n of choosing alternative j.

Principle of Revealed Preference

• Suppose that we observe an agent n making choices under different choice sets $\mathcal{J}$: i.e., $y_n(\mathcal{J}_1)$, $y_n(\mathcal{J}_2)$, ..., $y_n(\mathcal{J}_T)$.

• Under the Assumption of Utility Maximization, the **agent's choices reveal information on her preferences.**

• It turns out that, under some assumptions on the structure of $U_n(j)$, we can also identify how $U_n(j)$ varies with characteristics X_j (e.g., price) without the need of multiple choice sets $\mathcal{J}_t$.

• This a powerful principle in Econometrics and it is behind the estimation of demand or supply functions.

RANDOM UTILITY MODELS (2)

• In a Random Utility Model (RUM) the specification of $U_n(j)$ is:

$$U_n(j) = u_n(j, X_{jn}) + \varepsilon_{jn}$$

where:

- X_{jn} is a $K \times 1$ vector of characteristics of agent n and/or choice alternative j that are observable to the researcher;

- $u_n(.)$ is a real-valued function;

- $\varepsilon_n = \{\varepsilon_{0n}, \varepsilon_{1n}, ..., \varepsilon_{J,n}\}$ represents unobservable variables to the researcher, but observable to the agent and therefore affecting her choice.

RANDOM UTILITY MODELS (3)

- A common specification of the a RUM is:

$$U_n(j) = \widetilde{X}_j \; \beta_n + W_{jn} \; \delta_n + Z_n \; \gamma_j + \varepsilon_{jn}$$

- $\widetilde{X}_j$ is a $1 \times K_x$ vector of characteristics of alternative j (e.g., price);

- Z_n is a $1 \times K_z$ vector of observable attributes of the agent (e.g., income);

- W_{jn} is a $1 \times K_w$ vector of characteristics that vary across individuals (e.g., commuting time to work using transportation mode j);

- γ_j is a $K_z \times 1$ vector of parameters.

- β_n and δ_n are $K_x \times 1$ and $K_w \times 1$ vectors, respectively, that represent the marginal utility of each product attribute.

RANDOM COEFFICIENTS RUMs

• We can distinguish two types of models according to the specification of the coefficients β_n and δ_n.

• **Models without random coefficients.** Either (β_n, δ_n) are constant parameters (i.e., $\beta_n = \beta$ and $\delta_n = \delta$ for any n) or they are deterministic functions of observable agent's observable Z_n.

• In the later case, the terms $\widetilde{X}_j \, \beta_n + W_{jn} \, \delta_n$ are equivalent to $\widetilde{W}_{jn} \, \widetilde{\delta}$ where $\widetilde{W}_{jn}$ includes products of characteristics $\widetilde{X}_j$ and attributes Z_n.

• **Models with random coefficients.** β_n and/or δ_n depend on unobservable random variables for the researcher.

EXAMPLE 1: Choice of Transportation Mode to Work

- $Y \in \{$ Walking, Bike, Bus, Subway, Car, Bundles of the previous $\}$
- $\widetilde{X}_j =$ (Price per mile)
- $W_{jn} =$ (Commuting time using mode j)
- $Z_n =$ (Income, Age, Gender, etc)

EXAMPLE 2: Choice (Demand) of Differentiated Product (Laptops)

- $Y \in \{$ every laptop product available in the market $\}$

- $\widetilde{X}_j =$ (Price, Brand, CPU speed, Screen size, Weight, Color, RAM, HD size, etc)

- β_n contains the marginal utilities of each product attribute (for individual n);

- $Z_n =$ (Income, Age, Gender, etc).

- $W_{jn} =$ Indicator of n has bought brand j before.

SOME NORMALIZATIONS

• For constants a_n and $b_n > 0$, function $a_n + b_n\ U_n(j)$ is a **positive affine transformation** of utility $U_n(j)$.

• Any positive affine transformation of the utility function generates the same (utility maximizing) behavior for agent n.

• Therefore, we need to make some normalization assumptions on the parameters in the utility function $U_n(j) = \widetilde{X}_j\ \beta_n + W_{jn}\ \delta_n + Z_n\ \gamma_j + \varepsilon_{jn}$, such that we can identify the parameters in the utility function.

SOME NORMALIZATIONS [2]

• A necessary condition to identify a parameter is that **a marginal change in the parameter implies a change in the optimal choice of some agents in the population** (such that some CCPs change).

• Other necessary condition is that it is **not possible to completely offset the effect on all CCPs of a marginal change in the parameter by making a marginal change in other parameter.**

• Consider a model with 3 choice alternatives. $Y_n = 2$ iff:

$$\varepsilon_{0n} - \varepsilon_{2n} \leq (\widetilde{X}_2 - \widetilde{X}_0)\beta_n + (W_{2n} - W_{0n})\delta_n + Z_n(\gamma_2 - \gamma_0)$$

$$\varepsilon_{1n} - \varepsilon_{2n} \leq (\widetilde{X}_2 - \widetilde{X}_1)\beta_n + (W_{2n} - W_{1n})\delta_n + Z_n(\gamma_2 - \gamma_1)$$

SOME STANDARD NORMALIZATIONS [3]

• Some standard normalization assumptions are:

• (1) No constant terms (i.e., no a_n that does not depend on j): no constant term in $\widetilde{X}_j \ \beta_n$ or $W_{jn} \ \delta_n$;

• (2) If model includes $\widetilde{X}_j \beta$, then Z_n does not include constant term.

• (3) $\gamma_0 = 0$. [If all the $\gamma_j' s$ are additively transformed by the same constant, the optimal choice does not change].

• (4) $Var(\varepsilon_{1n} - \varepsilon_{0n}) = 1$ [If we multiply all the differences $\varepsilon_{jn} - \varepsilon_{0n}$ by the same constant, the optimal choice does not change].

Choice Probabilities in RUM

- The CCP for alternative j is (omitting agent subindex n)

$$\begin{aligned} P(j\ |x) &= \text{Pr}\Big(\ u(j,x)+\varepsilon_j \geq u(i,x)+\varepsilon_i \quad \text{for any } i \neq j\ \Big) \\ &= \text{Pr}\Big(\ \varepsilon_i \leq \varepsilon_j + u(j,x) - u(i,x) \quad \text{for any } i \neq j\ \Big) \\ &= \int_{-\infty}^{+\infty} F_{\varepsilon_{-j}|\varepsilon_j}\Big(\varepsilon_j + u(j,x) - u(i,x) \quad \text{for any } i \neq j\ \Big)\ f_{\varepsilon_j}(\varepsilon_j)\ d\varepsilon_j \end{aligned}$$

where f_{ε_j} is the marginal density of ε_j, and $F_{\varepsilon_{-j}|\varepsilon_j}$ is the CDF of $\varepsilon_{-j} \equiv \{\varepsilon_i : i \neq j\}$ conditional on ε_j.

- Integral of dimension J. Only for some specifications of the CDF $F(\varepsilon)$ has a closed form expression.

Maximum Likelihood estimation

• Let θ be the vector of parameters of the model, and $P(j|x,\theta)$ the CCPs according to the model. The log-likelihood function is:

$$\begin{aligned}\ell_N(\theta) &= \sum_{n=1}^{N} \ln \Pr(Y = y_n \mid X = x_n;\ \beta) \\ &= \sum_{n=1}^{N} \sum_{j \in \mathcal{J}} 1\{y_n = j\}\ \ln P(j|x_n,\theta)\end{aligned}$$

• The f.o.c. or likelihood equations:

$$\frac{\partial \ell_N(\theta)}{\partial \theta} = \sum_{n=1}^{N} \sum_{j \in \mathcal{J}} 1\{y_n = j\} \frac{1}{P(j|x_n,\theta)} \frac{\partial P(j|x_n,\theta)}{\partial \theta} = 0$$

Maximum Likelihood estimation [2]

• Now, for any x_n, $\sum_{j\in\mathcal{J}} P(j|x_n,\theta) = 1$, and therefore $\sum_{j\in\mathcal{J}} \frac{\partial P(j|x_n,\theta)}{\partial\theta} = 0$. This implies that we can write the likelihood equations as:

$$\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} \frac{\partial P(j|x_n,\theta)}{\partial\theta}\left[1\{y_n = j\}\frac{1}{P(j|x_n,\theta)} - 1\right] = 0$$

Or

$$\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} \frac{\partial P(j|x_n,\theta)}{\partial\theta}\frac{1}{P(j|x_n,\theta)}\left[1\{y_n = j\} - P(j|x_n,\theta)\right] = 0$$

• This expression of the MLE has a Method of Moments interpretation.

MLE as Method of Moments Estimator of Regression-like model

- By definition of CCPs, we have that

$$P(j \mid x_n) = \mathsf{Pr}\left(y_n = j \mid x_n\right) = \mathbb{E}\left(\, \mathsf{1}\{y_n = j\} \mid x_n\right)$$

Therefore,

$$\mathsf{1}\{y_n = j\} = P(j \mid x_n) + v_{jn}$$

where, by construction, the error term v_{jn} is such that $\mathsf{E}\Big(v_{jn} \mid x_n\Big) = 0$.

- This system of equations can be seen as a regression-like representation of a multinomial choice model.

MLE as MME of Regression-like model [2]

• At the true parameters θ, the following moment conditions should hold: for any j and any function $h_j(x_n)$:

$$\mathbb{E}\left(h_j(x_n)\ [1\{y_n = j\} - P(j|x_n, \theta)]\right) = 0$$

• MME is based on sample counterpart of population moment conditions.

$$\frac{1}{N}\sum_{n=1}^{N} h_j(x_n)\ [1\{y_n = j\} - P(j|x_n, \theta)] = 0$$

• The MLE provides the optimal K moment conditions to estimate the vector of K parameters θ:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{j\in\mathcal{J}} \frac{\partial P(j|x_n, \theta)}{\partial \theta}\frac{1}{P(j|x_n, \theta)}\,[1\{y_n = j\} - P(j|x_n, \theta)] = 0$$

4. MULTINOMIAL LOGIT MODEL [without random coefficients]

• We have $U_n(j) = X_{jn}\,\beta + \varepsilon_{jn}$ where

$$\varepsilon_{jn} \text{ are i.i.d. over } (n, j) \text{ Type 1 Extreme Value}$$

Type 1 Extreme Value is also called Gumbel distribution.

• For any j, we have that the CDF

$$F(\varepsilon_j) = \exp\left\{-\exp\left\{-\varepsilon_j\right\}\right\}$$

and the PDF is $f(\varepsilon_j) = \exp\left\{-\varepsilon_j - \exp\left\{-\varepsilon_j\right\}\right\}$. The PDF is asymmetric.

• The difference of two independent Type 1 Extreme Value variables has a Logistic distribution with CDF,

$$F^*(\varepsilon_j - \varepsilon_i) = \frac{\exp\left\{\varepsilon_j - \varepsilon_i\right\}}{1 + \exp\left\{\varepsilon_j - \varepsilon_i\right\}}$$

• Under this assumption on the distribution of ε, we have the following form for the CCPs:

$$
\begin{aligned}
P(j) &= \int_{-\infty}^{+\infty} F_{\varepsilon_{-j}|\varepsilon_j}\left(\varepsilon_j + u_j - u_i \quad \text{for any } i \neq j\right) f_{\varepsilon_j}(\varepsilon_j)\, d\varepsilon_j \\
&= \int_{-\infty}^{+\infty} f(\varepsilon_j) \left[\prod_{i \neq j} F(\varepsilon_j + u_j - u_i)\right] d\varepsilon_j \\
&= \int_{-\infty}^{+\infty} \prod_{i \in \mathcal{J}} \exp\{-\varepsilon_j\} \exp\{-\exp\{-\varepsilon_j - u_j + u_i\}\}\, d\varepsilon_j \\
&= \int_{-\infty}^{+\infty} \exp\{-\varepsilon_j\} \exp\left\{-\exp\{-\varepsilon_j - u_j\}\left[\textstyle\sum_{i \in \mathcal{J}} \exp\{u_i\}\right]\right\} d\varepsilon_j
\end{aligned}
$$

• Define $S \equiv \sum_{i \in \mathcal{J}} \exp\{u_i\}$, and make the change in variable, $v = \varepsilon_j + u_j - \ln S$

$$
\begin{aligned}
P(j) &= \int_{-\infty}^{+\infty} \exp\{-v + u_j - \ln S\} \ \exp\{-\exp\{-v\}\} \ dv \\
&= \exp\{u_j - \ln S\} \left[\int_{-\infty}^{+\infty} \exp\{-v\} \ \exp\{-\exp\{-v\}\} \ dv \right] \\
&= \frac{\exp\{u_j\}}{\exp\{\ln S\}} \ 1 \\
&= \frac{\exp\{u_j\}}{\sum_{k=0}^{J} \exp\{u_k\}}
\end{aligned}
$$

ML ESTIMATION OF LOGIT MODEL

• The closed form expression for CCPs is very convenient for the estimation of the model.

• Consider the logit model $U_n(j) = X_{jn}\ \beta + \varepsilon_{jn}$, and the random sample $\{y_n, x_n : i = 1, 2, ..., N\}$. The log-likelihood function is:

$$\begin{aligned} \ell_N(\beta) &= \sum_{n=1}^{N} \ln \Pr(Y = y_n \mid X = x_n;\ \beta) \\ &= \sum_{n=1}^{N} \sum_{j \in \mathcal{J}} 1\{y_n = j\}\ \ln \left[\frac{\exp\{x_{jn}\ \beta\}}{\sum_{i \in \mathcal{J}} \exp\{x_{in}\ \beta\}} \right] \end{aligned}$$

• This log-likelihood function if **globally concave in** β. Furthermore, the gradient and Hessian of this function have simple closed form expressions. Therefore, the numerical computation of the MLE can be implemented in a simple way using Newton's method.

ML ESTIMATION OF LOGIT MODEL (2)

- You can verify that in a Logit model,

$$\frac{\partial P(j)}{\partial u_j} = P(j)\,[1 - P(j)]$$

Taking this into account, we can show that:

$$\frac{\partial \ln P(j \mid x_n,\ \beta)}{\partial \beta} = \frac{\partial P(j)}{\partial u_j}\frac{1}{P(j)}\frac{\partial u_{jn}}{\partial \beta} = x_{jn} - m(x_n, \beta)$$

where $m(x_n, \beta) \equiv \sum_{i \in \mathcal{J}} x_{in}\ P(i|x_n, \beta)$. And the likelihood equation equations are:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=0}^{J}\Big[x_{jn} - m(x_n, \beta)\Big]\,[1\{y_n = j\} - P(j|x_n, \beta)]\right) = 0$$

- But it is clear that $\sum_{j \in \mathcal{J}} m(x_n, \beta) \left[1\{y_n = j\} - P(j|x_n, \beta)\right] = 0$ because $\sum_{j \in \mathcal{J}} 1\{y_n = j\} = \sum_{j \in \mathcal{J}} P(j|x_n, \beta) = 1$.

- Therefore,

$$\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{j=0}^{J} x_{jn} \left[1\{y_n = j\} - P(j|x_n, \beta)\right] \right) = 0$$

Luce Theorem (1959)

Consider a RUM with utilities $u_j + \varepsilon_j$. Consider the following two axioms.

Axiom 1. Let i,j be two choice alternatives. And let $\mathcal{J}$ and $\mathcal{J}'$ be two different choice sets that include i and j. Then,

$$\frac{P(j|\mathcal{J})}{P(i|\mathcal{J})} = \frac{P(j|\mathcal{J}')}{P(i|\mathcal{J}')}$$

Axiom 2. For any $\mathcal{J}$ and any $j \in \mathcal{J}$, $P(j|\mathcal{J}) > 0$.

Under Axioms 1 and 2, the form of the CCPs is:

$$P(j|\mathcal{J}) = \frac{w(u_j)}{\sum_{i \in \mathcal{J}} w(u_i)}$$

with $w(.) > 0$, strictly increasing, and $w(u_j)/w(u_j) = w(u_j - u_i)$.

Independence of Irrelevant Alternatives (IIA)

• The logit model imposes the restriction that the ratio between the probabilities of two alternatives, say j and i, depends ONLY on the utilities of these alternatives, and not on utilities of other alternatives:

$$\frac{P(j)}{P(i)} = \frac{\exp\left\{u_j\right\}}{\exp\left\{u_i\right\}}$$

• Therefore, if we change the choice set $\mathcal{J}$, by adding or/and removing alternatives, the ratios between probabilities should not change. For any two choice sets $\mathcal{J}$ and $\mathcal{J}'$ (that include j and i as alternatives), we have that:

$$\frac{P(j|\mathcal{J})}{P(i|\mathcal{J})} = \frac{P(j|\mathcal{J}')}{P(i|\mathcal{J}')}$$

• This property, though reasonable in a deterministic environment, it generates unrealistic predictions in a RUM.

IIA in Deterministic and in Stochastic Decision Environments

• Let a, b, c be three choice alternatives.

Deterministic choice.

- Suppose $\mathcal{J} = \{a, b\}$.

$P(a|\mathcal{J})/P(b|\mathcal{J})$ is either ∞ (if $a \succ b$), or 0 (if $b \succ a$), or 1 (if $a \sim b$).

- Suppose $\mathcal{J}' = \{a, b, c\}$ with $a \sim c$.

It seems reasonable that $P(a)/P(c) = 1$.

This is consistent with IIA: if $a \succ b$, $P(a|\mathcal{J})/P(b|\mathcal{J}) = 0.5/0 = \infty$; if $b \succ a$, $P(a|\mathcal{J})/P(b|\mathcal{J}) = 0$; and $a \sim b$, $P(a|\mathcal{J})/P(b|\mathcal{J}) = (1/3)/(1/3) = 1$.

Stochastic choice.

- Suppose $\mathcal{J} = \{a, b\}$ with $P(a|\mathcal{J}) = 0.9999$ and $P(b|\mathcal{J}) = 0.0001$, such that $P(a|\mathcal{J})/P(b|\mathcal{J}) = 9999$ [very close to the determinsitic case].

- Suppose $\mathcal{J}' = \{a, b, c\}$ with $a \sim c$ [identical choice alternatives].

It seems reasonable that $P(a|\mathcal{J}') = P(c|\mathcal{J}') = 0.9999/2$, and $P(b|\mathcal{J}') = 0.0001$.

This is NOT consistent with IIA because now $P(a|\mathcal{J}')/P(b|\mathcal{J}') = 9999/2 \neq 9999 = P(a|\mathcal{J})/P(b|\mathcal{J})$.

IIA: Example

• Consider consumers deciding which car model to purchase. The set of available models in year 2014 (i.e., choice set $\mathcal{J}_{2014}$) includes model "Lux" that is a luxury car; and model "Econ", that is a very modest and unexpensive car. In year 2014, their market shares are:

$$P(Lux \mid \mathcal{J}_{2014}) = 0.10 \quad ; \quad P(Econ \mid \mathcal{J}_{2014}) = 0.40;$$

$$\frac{P(Lux \mid \mathcal{J}_{2014})}{P(Econ \mid \mathcal{J}_{2014})} = \frac{1}{4}$$

• In 2015, the new luxury model "NewLux" (very similar to Lux) appears in the market. The logit model predicts that:

$$P(Lux \mid \mathcal{J}_{2015}) = 0.10 * (1 - P_{NewLux})$$

$$P(Econ \mid \mathcal{J}_{2015}) = 0.40 * (1 - P_{NewLux})$$

IIA: Example (2)

• For instance, if $P_{NewLux} = 10\%$, then $P(Lux \mid \mathcal{J}_{2015}) = 9\%$ and $P(Econ \mid \mathcal{J}_{2015}) = 36\%$, what seems unrealistic.

• The IIA is an implication of the Logit property that the differences $\varepsilon_j - \varepsilon_i$ are i.i.d. across any pair of choices.

IIA and Average Partial Effects

• In most applications we are interested in the estimation of Average Partial Effects (APE). In a discrete choice model, define $APE_{k,j}(x)$ as the APE of variable k in choice alternative j when the explanatory variables are x.

$$APE_{k,j}(x) = \frac{\partial P(j|x)}{\partial X_k}$$

• Does the MNL impose a restrictive / unrealisitc structure on these APEs?

• Yes, especially when we are interested in APE of changes in product characteristics, X_j.

IIA and Average Partial Effects [2]

• Remember that in the MNL:

$$P_{jn} = \frac{\exp\{X_j\beta + Z_n\gamma_j\}}{\sum_{i=0}^{J} \exp\{X_i\beta + Z_n\gamma_i\}}$$

• Consider the APE on P_j of a change in the X_i for $i \neq j$ (i.e., effect on demand of product j of a change in the price of product i):

$$\frac{\partial P_{jn}}{\partial X_i} = -\beta \; P_{jn}P_{in}$$

The effect is proportional to P_{jn} and P_{in}.

• Two products j with the same P_j are affected exactly in the same way by an increase in the price of product i. This is very restrictive.

IIA and Average Partial Effects [3]

• Consider now the partial effect of a change in a characteristic of individual n.

$$\frac{\partial P_{jn}}{\partial Z_n} = P_{jn} \left(\gamma_j - \sum_{i=0}^{J} \gamma_i \, P_{in} \right)$$

• This is not a particularly restrictive APE.

• As in a binary choice model, this APE goes to zero when $P_{jn} \longrightarrow 0$ and when $P_{jn} \longrightarrow 1$. It depends on the value of γ_j relative to the other γ's, and these parameters are unrestricted.

Solutions to Independence of Irrelevant Alternatives

• Different models have been proposed to deal with this limitation of the Logit model:

(1) Multinomial probit;

(2) Nested Logits;

(3) Random coefficients logit

5. NESTED LOGIT MODEL

• Suppose that the set $\mathcal{J}$ chocie alternative can be partitioned into G (mutually exclusive) groups of alternatives, that we index by g. Let $\mathcal{J}_g$ be the set of alternatives in group g such that:

$$\mathcal{J} = \bigcup_{g=1}^{G} \mathcal{J}_g$$

The idea is that alternatives within a group share some unobserved features that make them closer substitutes that alternatives in different groups.

• Suppose that random variables ε_j of the RUM has the following structure:

$$\varepsilon_j = \varepsilon_g^{(1)} + \sigma_g \, \varepsilon_{j|g}^{(2)}$$

- $\varepsilon_1^{(1)}, \varepsilon_2^{(1)}, ..., \varepsilon_G^{(1)}$ are i.i.d. EV type 1.

- For any g, $\{\varepsilon^{(2)}_{j|g} : j \in \mathcal{J}_g\}$ are i.i.d. EV type 1.

- $\varepsilon^{(1)\prime}s$ and $\varepsilon^{(2)\prime}s$ are independent.

NESTED LOGIT MODEL [2]

- This model has the following CCPs: $P_j = P_g^{(1)} \; P_{j|g}^{(2)}$ with

$$P_{j|g}^{(2)} = \frac{\exp\left\{\frac{u_j}{\sigma_g}\right\}}{\sum\limits_{i \in \mathcal{J}_g} \exp\left\{\frac{u_i}{\sigma_g}\right\}} \; ; \; P_g^{(1)} = \frac{\exp\{v_g\}}{\sum\limits_{g'=1}^{G} \exp\left\{v_{g'}\right\}}$$

and v_g is the $E(\max_{j \in \mathcal{J}_g}\{u_j + \sigma_g \; \varepsilon_{j|g}^{(2)}\})$ that has the following form:

$$v_g = \sigma_g \; \ln\left(\sum_{j \in \mathcal{J}_g} \exp\left\{\frac{u_j}{\sigma_g}\right\}\right)$$

NESTED LOGIT MODEL [3]

• A different way to represent the Nested Logit is the following.

• Consider the RUM $Y = \arg\max_{j \in \mathcal{J}} \{u_j + \varepsilon_j\}$ with the the G groups and where $\boldsymbol{\varepsilon} = (\varepsilon_0, \varepsilon_1, ..., \varepsilon_J)$ has a **Generalized Extreme Vaue (GEV) distribution**:

$$F(\boldsymbol{\varepsilon}) = \exp\left\{ - \sum_{g=1}^{G} \left[\sum\nolimits_{j \in \mathcal{J}_g} \exp\left(-\frac{\varepsilon_j}{\sigma_g} \right) \right]^{\sigma_g} \right\}$$

where σ_1, σ_2, ..., σ_R are positive parameters.

• Then, the CCPs are: $P_j = P_g^{(1)} \; P_{j|g}^{(2)}$

$$P_{j|g}^{(2)} = \frac{\exp\left\{\frac{u_j}{\sigma_g}\right\}}{\sum_{i\in\mathcal{J}_g}\exp\left\{\frac{u_i}{\sigma_g}\right\}} \quad ; \quad P_g^{(1)} = \frac{\exp\{v_g\}}{\sum_{g'=1}^{G}\exp\left\{v_{g'}\right\}}$$

with $v_g = \sigma_g \ln\left(\sum_{j\in\mathcal{J}_g}\exp\left\{\frac{u_j}{\sigma_g}\right\}\right)$.

NESTED LOGIT MODEL [4]

• The NL has an interpretation as a sequential decision model. Let $Y_n^{(1)} \in \{1, 2, ..., G\}$ represent agent n's choice of group. And let $Y_n^{(2)}$ represent the choice of specific alternative.

• The model implies that:

$$\Pr(Y_n^{(1)} = g \mid X_n) = P_g^{(1)}(X_n)$$

and

$$\Pr(Y_n^{(2)} = j \mid X_n,\ Y_n^{(1)} = g) = P_{j|g}^{(2)}(X_n)$$

• Therefore, the likelihood function of the model, $l(\theta) = \sum_{n=1}^{N} \ln \Pr(Y_n | X_n, \theta)$

can be written as the sum of two likehoods: $l^{(1)}(\theta) + l^{(2)}(\theta)$

$$
\begin{aligned}
l(\theta) &= \sum_{n=1}^{N} \sum_{g=1}^{G} 1\{y_n^{(1)} = g\} \ln P_g^{(1)}(X_n, \theta) \\
&+ \sum_{n=1}^{N} \sum_{j \in \mathcal{J}_{y_n^{(1)}}} 1\{y_n^{(2)} = j\} \ln P^{(2)}_{j|y_n^{(1)}}(X_n, \theta)
\end{aligned}
$$

NESTED LOGIT MODEL [5]

• The Nested Logit maintains the property of IIA for alternatives within the same group but not for alternatives in different groups.

• In the example of the demand of cars: the new car will have a stronger substitution effect within its own group, e.g., luxury cars.

TWO-STEP ESTIMATION OF NL

• Note that $l(\theta) = l^{(1)}(\theta) + l^{(2)}(\theta)$ where:

$l^{(1)}(\theta)$ is the **between-group** likelihood function for the choice variable $Y_n^{(1)}$ conditional on X_n

$l^{(2)}(\theta)$ is the **within-group** likelihood function for the choice variable $Y_n^{(2)}$ conditional on X_n and $Y_n^{(1)}$.

• We can estimate a combination of the parameters in θ by maximizing $l^{(1)}(\theta)$, and other combination of parameters θ by maximizing $l^{(2)}(\theta)$.

• This two-step procedure is not statistically efficient but it is computationally very convenient because each step consists of a standard MNL estimation (i.e., globally concave likelihood function).

TWO-STEP ESTIMATION OF NL [2]

• **Step 1:** Maximization of within-group likelihood function $l^{(2)}(\theta)$ with probabilities:

$$P_{j|g,n} = \frac{\exp\{X_j\ \beta_g + Z_n\ \gamma_{j,g}\}}{\sum_{i \in \mathcal{J}_g} \exp\{X_i\ \beta_g + Z_n\ \gamma_{i,g}\}}$$

where the estimated parameters are: $\beta_g \equiv \frac{\beta}{\sigma_g}$ and $\gamma_{j,g} \equiv \frac{\gamma_j}{\sigma_g}$ (where one of the $\gamma' s$ within each group is notmalized to zero).

TWO-STEP ESTIMATION OF NL [3]

• **Step 2:** Construct the estimated **inclusive values:**

$$\widehat{I}_{gn} = \ln\left(\sum_{j\in\mathcal{J}_g} \exp\{X_j\ \widehat{\beta}_g + Z_n\ \widehat{\gamma}_{j,g}\}\right)$$

And maximization of betwithin-group likelihood function $l^{(1)}(\theta)$ with probabilities:

$$P_{j|g,n} = \frac{\exp\{\sigma_g\ \widehat{I}_{gn}\}}{\sum_{g'=1}^{G} \exp\{\sigma_{g'}\ \widehat{I}_{g'n}\}}$$

- The estimated parameters are σ_g, with one of these parameters normalized to 1.

EFFICIENT ESTIMATION OF NL

• Given this consistent two-step estimator, we can construct an efficient estimator, and a valid variance-covariance matric by doing one Newton or BHHH iteration in the estimation of the full likelihood function:

$$\widehat{\theta}_{eff} = \widehat{\theta}_{2step} - \left[\frac{\partial^2 l(\widehat{\theta}_{2step})}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial l(\widehat{\theta}_{2step})}{\partial\theta}$$

6. RANDOM COEFFICIENTS LOGIT (MIXED LOGIT)

• In the standard RCLogit we have that:

$$U_{jn} = X_{jn}\,\boldsymbol{\beta}_n + \varepsilon_{jn}$$

where:

- ε_{jn} are i.i.d. over (n, j) Type 1 Extreme Value;

- $\boldsymbol{\beta}_n$ is i.i.d. over $n \sim N(\mathbf{b}, \boldsymbol{\Omega})$;

- $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\beta}_n$ are independent.

• We can also represent $\boldsymbol{\beta}_n$ as

$$\boldsymbol{\beta}_n = \mathbf{b} + \mathbf{W}\ \mathbf{v}_n$$

where $\mathbf{W}$ is a $K \times K$ lower triangular matrix that is the Cholesky's decomposition of $\boldsymbol{\Omega}$ (i.e., $\mathbf{W}\mathbf{W}' = \boldsymbol{\Omega}$) and $\mathbf{v}_n = (v_{1n}, v_{2n}, ..., v_{Kn})'$ is a vector of independent standard normals.

RC LOGIT

• We can write

$$\begin{aligned} U_{jn} &= X_{jn}\,\mathbf{b} + [X_{jn}\,\mathbf{W}]\,\mathbf{v}_n + \varepsilon_{jn} \\ &= X_{jn}\,\mathbf{b} + \left[\sum_{k=1}^{K}\left(\sum_{k'=k}^{K} X_{k'jn}\,w_{k'k}\right)v_{kn}\right] + \varepsilon_{jn} \end{aligned}$$

• The parameters of the model are $\mathbf{b}$ and $\mathbf{W}$.

• The RCLogit can be generalized to a allow for a nonparametric specification of the distribution of $\mathbf{v}_n$.

• Fox, Kim, Ryan, and Bajari (JoE, 2012) show that the nonparametric RCLogit is identified.

RC LOGIT - CCPs

• To obtain CCPs, we should integrate over $\boldsymbol{\varepsilon}_n$ and $\mathbf{v}_n$ the optimal decision $\{Y_n = j\} \Leftrightarrow \{\ X_{jn}\ \mathbf{b} + [X_{jn}\ \mathbf{W}]\ \mathbf{v}_n + \varepsilon_{jn} \geq X_{in}\ \mathbf{b} + [X_{in}\ \mathbf{W}]\ \mathbf{v}_n + \varepsilon_{in}$ for any $i \neq j\}$:

$$P(j\ |X_n) \ = \ \int \frac{\exp\left\{X_{jn}\ \mathbf{b} + [X_{jn}\ \mathbf{W}]\ \mathbf{v}\right\}}{\sum_{i=0}^{J} \exp\left\{X_{in}\ \mathbf{b} + [X_{in}\ \mathbf{W}]\ \mathbf{v}\right\}} \prod_{k=1}^{K} \phi(v_k)\ dv_k$$

• It requires numerical integration over the distribution of the K random variables $\{v_k\}$.

RC LOGIT and IIA

- Consider the effect on P_j of a marginal change in the attributes of product $i \neq j$: $\dfrac{\partial P_j}{\partial X_i}$.

- In the Logit model, this effect is the same for every choice alternaitve j:

$$\frac{\partial P_j}{\partial X_i} = -\mathbf{b}\ P_j\ P_i$$

- In RC Logit, this effect is:

$$\frac{\partial P_j}{\partial X_i} = -\int [\mathbf{b} + \mathbf{W}\ \mathbf{v}]\ \Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\ f(\mathbf{v})\ d\mathbf{v}$$

and $\Lambda_j(\mathbf{v}) = \exp\left\{X_j\ \mathbf{b} + [X_j\ \mathbf{W}]\ \mathbf{v}\right\} / \sum_{i=0}^{J} \exp\{X_i\ \mathbf{b} + [X_i\ \mathbf{W}]\ \mathbf{v}\}$.

RC LOGIT and IIA [2]

• The effect $\frac{\partial P_j}{\partial X_i} = -\int [\mathbf{b} + \mathbf{W}\ \mathbf{v}]\ \Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\ f(\mathbf{v})\ d\mathbf{v}$ depends on $\mathbb{E}_{\mathbf{v}}\left(\Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\right)$ that is equal to $Cov\left(\Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\right) + P_j\ P_i$.

• Therefore,

$$\left.\frac{\partial P_j}{\partial X_i}\right|_{RCLogit} - \left.\frac{\partial P_j}{\partial X_i}\right|_{Logit} = -\mathbf{b}\ Cov\left(\Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\right)$$

• This covariance depends on the distance between the vectors X_j and X_i.

- When $\left\|X_j - X_i\right\|$ is small, low values of $\Lambda_j(\mathbf{v})$ are associated with low $\Lambda_i(\mathbf{v})$, and $Cov\left(\Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\right) > 0$

- When $\left\|X_j - X_i\right\|$ is large, $Cov\left(\Lambda_j(\mathbf{v})\ \Lambda_i(\mathbf{v})\right)$ can be zero o even negative.

RC LOGIT - MLE

• Given a random sample, the log-likelihood function $\ell_N(\mathbf{b},\mathbf{W})$ is:

$$\sum_{n=1}^{N}\sum_{j=0}^{J} 1\{y_n = j\} \ \ln\left[\int \frac{\exp\left\{X_{jn}\ \mathbf{b} + [X_{jn}\ \mathbf{W}]\ \mathbf{v}\right\}}{\sum_{i=0}^{J}\exp\left\{X_{in}\ \mathbf{b} + [X_{in}\ \mathbf{W}]\ \mathbf{v}\right\}} \prod_{k=1}^{K}\phi(v_k)dv_k\right]$$

• The MLE is the value of $(\mathbf{b},\mathbf{W})$ that maximizes $\ell_N(\mathbf{b},\mathbf{W})$. This MLE has the standard good properties:

- MLE is CAN and AE.

- $\ell_N(\mathbf{b},\mathbf{W})$ is twice continuously differentiable in $(\mathbf{b},\mathbf{W})$: we can use gradient methods (e.g., Newton, BHHH) to search for the MLE.

- $\ell_N(\mathbf{b},\mathbf{W})$ is not globally concave but is concave in $\mathbf{b}$ given $\mathbf{W}$.

RC LOGIT - MLE [2]

• The main issue in the implementation of the MLE of the RCLogit is the computation of the CCPs by solving the multiple integration problem.

• We can use Monte Carlo simulation methods to approximate CCPs. However, we need to take into account how the approximation error affect the properties of our estimators.

7. MONTE CARLO SIMULATION

• Monte Carlo simulation is a general method to approximate multiple-dimensional integrals. It is used in any scientific application, empirical or theoretical, that requires the computation of multiple-dimensional integrals.

• Let $\mathbf{v} = (v_1, v_2, ..., v_K)$ be a vector of continuous random variables with joint density $\phi(\mathbf{v})$ that is continuous over the compact support $\mathcal{V}$.

• Let P be a parameter that is defined as:

$$P = \int h(\mathbf{v})\ \phi(\mathbf{v})\ d\mathbf{v}$$

where $h(\mathbf{v})$ is a known function. Suppose there is NOT a closed-form expression for this integral.

Fundamental Theorem of Sampling

• Let v be a scalar random variable with CDF $F(v)$ that is continuous and strictly increasing on the the support $\mathcal{V}$. Then:

(1) There exists an inverse function $F^{-1}(.)$ (the Quantile function) such that $v = F^{-1}(u)$.

(2) The random variable $u = F(v)$ has a distribution $U\,[0, 1]$.

This implies that if $\{u_1, u_2, ..., u_R\}$ are R i.i.d. random draws from a $U\,[0, 1]$, then $\left\{F^{-1}(u_1),\ F^{-1}(u_2), ...,\ F^{-1}(u_R)\right\}$ are R i.i.d. random draws from the distribution F.

Proof:

• (1) F is strictly increasing on $\mathcal{V}$. Therefore, by the inverse function theorem, there exists an inverse function $F^{-1}(.)$ such that for any $(v, u) \in \mathcal{V} \times [0, 1]$, $u = F(v) \Leftrightarrow v = F^{-1}(u)$.

• (2) Define the random variable u, $u = F(v)$. The CDF of u evaluated at an arbitrary $u_0 \in [0, 1]$ is:

$$\begin{aligned} \Pr(u \leq u_0) &= \Pr\left(F^{-1}(u) \leq F^{-1}(u_0)\right) \\ &= \Pr\left(v \leq F^{-1}(u_0)\right) = F(F^{-1}(u_0)) = u_0 \end{aligned}$$

So, u has a $U[0, 1]$ distribution.

Examples of Quantile functions: Logistic distribution

- $v \sim Logistic$. The CDF is $F(v) = \dfrac{\exp(v)}{1 + \exp(v)}$.

- The inverse CDF (i.e., Quantile function) of the Logistic is:

$$F^{-1}(u) = \ln\left(\frac{u}{1-u}\right)$$

- Then, we can get a random draw from the Logistic by getting a draw u from $U\,[0, 1]$ and then apply transformation:

$$v = \ln\left(\frac{u}{1-u}\right)$$

Drawing from Multivariate Normal

• If v is Standard Normal with CDF Φ and $u \sim U[0,1]$, then $v = \Phi^{-1}(u)$. There are very efficient procedures to calculate the inverse function Φ^{-1}.

• Let $\mathbf{v} = (v_1, v_2, ..., v_K)$ be a vector of Normal random variables $\sim N(\mathbf{m},\mathbf{\Omega})$. Then, we can write:

$$\mathbf{v} = \mathbf{m} + \mathbf{W}\ \mathbf{v}^*$$

where:

- $\mathbf{v}^* = (v_1, v_2, ..., v_K)$ is a vector of i.i.d. standard normals;

- $\mathbf{W}$ is a lower triangular matrix obtained as the Cholesky decomposition of $\mathbf{\Omega}$, i.e., $\mathbf{W}\ \mathbf{W}' = \mathbf{\Omega}$.

• We can get a random draw of $\mathbf{v} \sim N(\mathbf{m},\mathbf{\Omega})$ by taking K independent random draws from $U[0,1]$, $(u_1, u_2, ..., u_K)$, and the applying the transformation:

$$\mathbf{v} = \mathbf{m} + \mathbf{W} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_K) \end{pmatrix}$$

where Φ^{-1} is the inverse of the CDF of the standard normal.

• As we will see below, in the context of simulation-based estimation, if in our econometric model we are interested in the estimation of parameters $\mathbf{m}$ or/and $\mathbf{W}$, we need to keep our simulations $\Phi^{-1}(u_j)$ constant over the estimation procedure, while the value of $\mathbf{m}$ or/and $\mathbf{W}$ varies during the gradient search for the estimator.

Frequency Simulator

- Remember that P is a parameter defined as:

$$P = \mathbb{E}_{\phi}\left(h(\mathbf{v})\right) = \int h(\mathbf{v})\ \phi(\mathbf{v})\ d\mathbf{v}$$

For instance, $h(\mathbf{v}) = 1\{v_1 \leq \beta_1,\ v_2 \leq \beta_2,\ ...,\ v_K \leq \beta_K\}$.

- Let $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_R\}$ be R independent random draws from the CDF Φ. Then, the Frequency Simulator of P is defined as:

$$\widetilde{P}_R = \frac{1}{R}\sum_{r=1}^{R} h(\mathbf{v}_r)$$

- The simulation error (approximation error) is: $e_R = \widetilde{P}_R - P$.

Properties of the Frequency Simulator [Same as a "sample mean"]

- (1) Unbiased: $\mathbb{E}\left(\widetilde{P}_R\right) = P$.

- (2) Variance: $Var\left(\widetilde{P}_R\right) = \dfrac{Var(h(\mathbf{v}))}{R}$.

- (3) Consistent: As R goes to infinity, $\widetilde{P}_R \to P$.

- (4) $\sqrt{R}$ asymptotically normal: $\sqrt{R}\dfrac{\widetilde{P}_R - P}{\sqrt{Var(h(\mathbf{v}))}} \to N(0,1)$.

Properties of the Frequency Simulator (2)

• In general, it is possible to obtain simulators more precise (with lower variance) than the frequency simulator [e.g., Importance Sampling]

• Other limitation of the FS is that if the $h(.)$ is discontinuous or non-differentiable with respect to some parameters [e.g., $h(\mathbf{v}) = \mathbf{1}\{v_1 \leq \beta_1, ..., v_K \leq \beta_K\}$], then the simulator is also discontinuous and non-differentiable.

• This has important implications in the estimation of dicrete choice models. In some Simulated-Based estimators that use the frequency simulator of CCPs are such that:

- Criterion function is a step function of the parameters: numerical optimization problems;

- Estimator may not be root-N asymptotical normal.

Simulation-Based Estimation using Frequency Simulator

• Consider the RUM with utilities $U_{jn} = X_{jn}[\mathbf{b} + \mathbf{W}\ \mathbf{v}_n] + \varepsilon_{jn}$ such that:

$$P_{jn}(\boldsymbol{\theta}) = \int 1\left\{\varepsilon_{in} \leq \varepsilon_{jn} + (X_{jn} - X_{in})[\mathbf{b} + \mathbf{W}\mathbf{v}_n]\right\}\ f(\boldsymbol{\varepsilon}_n, \mathbf{v}_n)\ d\boldsymbol{\varepsilon}_n d\mathbf{v}_n$$

• A **Simulated-MLE** of $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{W})$ **based on the Frequency Simulator of the CCPs** is the value of $\boldsymbol{\theta}$ that maximizes the Simulated log-likelihood:

$$\ell^{(R)}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{j=0}^{J} 1\{y_n = j\}\ \ln \widetilde{P}_{jn}^{(R)}(\boldsymbol{\theta})$$

where $\widetilde{P}_{jn}^{(R)}(\boldsymbol{\theta})$ is the frequency simulator of $P_{jn}(\boldsymbol{\theta})$.

Simulation-Based Estimation using Frequency Simulator [2]

• The frequency simulator $\widetilde{P}_{jn}^{(R)}(\boldsymbol{\theta})$ is:

$$\widetilde{P}_{jn}^{(R)}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R} 1\left\{\varepsilon_{in}^{(r)} \leq \varepsilon_{jn}^{(r)} + (X_{jn} - X_{in})[\mathbf{b} + \mathbf{W}\ \mathbf{v}_n^{(r)}]\right\}$$

where and $\{\boldsymbol{\varepsilon}_n^{(r)}, \mathbf{v}_n^{(r)} : r = 1, 2, ..., R\}$ are R i.i.d. draws from the distribution $f(\boldsymbol{\varepsilon}_n, \mathbf{v}_n)$.

• This was the SMLE proposed in a seminal paper by Lerman and Manski (1981) for the Multinomial Probit, i.e., $\boldsymbol{\varepsilon}_n \sim N(0, \Omega)$ and $\mathbf{W} = \mathbf{0}$.

Sim-Based Estimation using Frequency Simulator [3]

• This estimator very poor statistical and computational properties of this estimator. There are several related issues.

[1] For choice alternatives with low $P_{jn}(\theta)$ we have that $\widetilde{P}_{jn}^{(R)}(\theta) = 0$, unless R is very large. The log-likelihood becomes minus infinite, even at the true θ.

[2] R should be very large to have an estimator with decent properties.

[3] $\ell^{(R)}(\theta)$ is a step function. Standard gradient methods do not work.

[4] For fixed R, the estimator is not consistent, it is not asymptotical normal. Poor small sample properties.

Solutions to the Problems of SMLE with Frequency Simulator

• There are several approaches / methods that overcome the limitations of the SMLE with the Frequency simulator.

[1] **RC Logit model**: Simulating $\mathbf{v}'s$ but analytical formula for integration over $\varepsilon's$ such that the simulator is a smooth function of parameters and always > 0.

[2] **Importance-sampling simulators**: Like GHK for the Probit model. Improves precision, is always > 0, and smooth function of parameters.

[3] **Simulated Method of Moments** and **Simulated Scores** [+ smooth and >0 simulator] that are root-N consistent and asymptotically normal estimators even when R is fixed and small (e.g., $R = 1$).

Solution using RC Logit

• In the RC Logit we take into account that ε_n is independent of $\mathbf{v}_n$ with Extreme Value distirbution, such that we have closed form expressions of probs conditional on $\mathbf{v}_n$, and these probs a re smooth functions.

$$P_{jn}(\theta) = \int \frac{\exp\left\{X_{jn}\ \mathbf{b} + [X_{jn}\ \mathbf{W}]\ \mathbf{v}_n\right\}}{\sum_{i=0}^{J} \exp\left\{X_{in}\ \mathbf{b} + [X_{in}\ \mathbf{W}]\ \mathbf{v}_n\right\}} \phi(\mathbf{v}_n) d\mathbf{v}_n$$

• Therefore, we can use the simulator:

$$\widetilde{P}_{jn}^{\Lambda,R}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \frac{\exp\left\{X_{jn}\ \mathbf{b} + [X_{jn}\ \mathbf{W}]\ \mathbf{v}_n^{(r)}\right\}}{\sum_{i=0}^{J} \exp\left\{X_{in}\ \mathbf{b} + [X_{in}\ \mathbf{W}]\ \mathbf{v}_n^{(r)}\right\}}$$

where $\{\mathbf{v}_n^{(r)} : r = 1, 2, ..., R\}$ are R i.i.d. draws from the distribution $f(\mathbf{v}_n)$.

Solution using RC Logit [2]

• The simulator $\widetilde{P}_{jn}^{\Lambda,R}(\theta)$ has several important advantages over the frequency simulator $\widetilde{P}_{jn}^{R}(\theta)$.

- $\widetilde{P}_{jn}^{\Lambda,R}(\theta)$ is continusously differentiable in θ.

- It is always > 0 and < 1 for any value of R, even for $R = 1$.

- Variance of the simulation error is subtantially smaller than for the frequency simulator [and the ratio of their variances increases exponentially with J].

Importance sampling simulation (IS)

• Let ϕ^* be a density function different to ϕ. Density ϕ^* is denoted the Importance Sampling density. By definition of P, we have that:

$$\begin{aligned} P &= \mathbb{E}_{\phi}\left(h(\mathbf{v})\right) = \int h(\mathbf{v})\ \phi(\mathbf{v})\ d\mathbf{v} \\ &= \int h(\mathbf{v})\ \frac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}\ \phi^*(\mathbf{v})\ d\mathbf{v} \\ &= \mathbb{E}_{\phi^*}\left(h(\mathbf{v})\ \frac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}\right) \end{aligned}$$

• Let $\left\{\mathbf{v}_1^*, \mathbf{v}_2^*, ..., \mathbf{v}_R^*\right\}$ be R independent random draws from ϕ^*. Then, the **Importance Sampling Simulator** (based on ϕ^*) of P is:

$$\widetilde{P}_R = \frac{1}{R}\sum_{r=1}^{R} h(\mathbf{v}_r)\ \frac{\phi(\mathbf{v}_r)}{\phi^*(\mathbf{v}_r)}$$

Properties of IS

- (1) Unbiased: $\mathbb{E}\left(\widetilde{P}_R\right) = P$.

- (2) Variance: $Var\left(\widetilde{P}_R\right) = \dfrac{Var\left(h(\mathbf{v})\,\frac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}\right)}{R}$.

- (3) Consistent: As R goes to infinity, $\widetilde{P}_R \to P$.

- (4) Asymptotically normal: $\sqrt{R}\dfrac{\widetilde{P}_R - P}{\sqrt{Var\left(h(\mathbf{v})\,\frac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}\right)}} \to N(0,1)$.

Relative variances of FS and IS

- $Var\left(\widetilde{P}_{FS}\right) = \dfrac{Var(h(\mathbf{v}))}{R}$ and $Var\left(\widetilde{P}_{IS}\right) = \dfrac{Var\left(h(\mathbf{v})\,\frac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}\right)}{R}$.

- Therefore, if the ratios $\dfrac{\phi(\mathbf{v})}{\phi^*(\mathbf{v})}$ are smaller than 1 for values of $\mathbf{v}$ with large $(h(\mathbf{v}) - P)^2$, then the ISS will have lower variance than the FS.

- For the ISS to have a lower variance than FS, the ISS density ϕ^* should "over sample" (relative to ϕ) those regions in the support of $\mathbf{v}$ where $(h(\mathbf{v}) - P)^2$ is large.

Simulation of Multinomial Probit probabilities: GHK Simulator

• Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_J)$ be a vector of Normal random variables with vector of means $\mathbf{0}$ and variance-covariance $\boldsymbol{\Omega}$. Let $\mathbf{c} = (c_1, c_2, ..., c_J)$ be a vector of constants. Consider the following probability:

$$\begin{aligned} P &= \mathsf{Pr}\left(\varepsilon_1 \leq c_1,\, \varepsilon_2 \leq c_2,\, ...,\, \varepsilon_J \leq c_J\right) \\ &= \int \mathsf{1}\left\{\varepsilon_1 \leq c_1,\, \varepsilon_2 \leq c_2,\, ...,\, \varepsilon_J \leq c_J\right\}\, \phi(\mathbf{v}; \boldsymbol{\Omega})\, d\mathbf{v} \end{aligned}$$

• These probabilities appear in a Multinomial Probit model.

• The Geweke-Hajivassiliou-Keane (GHK) simulator is a very efficient simulator of these probabilities. It is also continuously differentiable in the argument (vector of parameters) $\mathbf{c}$.

GHK Simulator (2)

• Let $\mathbf{W} = \{w_{ij}\}$ be a **lower triangular matrix** that comes from the Cholesky decomposition of $\mathbf{\Omega}$. Then, $\boldsymbol{\varepsilon} = \mathbf{W}\ \mathbf{z}$, where $\mathbf{z}$ is a vector of J independent standard normals, such that:

$$\begin{aligned}
\varepsilon_1 &= w_{11}\ z_1 \\
\varepsilon_2 &= w_{21}\ z_1 + w_{22}\ z_2 \\
\vdots &\quad\ \vdots \\
\varepsilon_J &= w_{J1}\ z_1 + w_{J2}\ z_2 + ... + w_{JJ}\ z_J
\end{aligned}$$

and

$$\begin{aligned}
\Big\{\varepsilon_j \leq c_j\Big\} &= \Big\{w_{j1}\ z_1 + w_{j2}\ z_2 + ... + w_{jj}\ z_j \leq c_j\Big\} \\
&= \Big\{z_j \leq \widetilde{c}_j - \widetilde{w}_{j1}\ z_1 - ... - \widetilde{w}_{j,J}\ z_J\Big\}
\end{aligned}$$

with $\widetilde{c}_j = \dfrac{c_j}{w_{jj}}$ and $\widetilde{b}_{ji} = \dfrac{w_{ji}}{w_{jj}}$ for $i < j$

GHK Simulator (3)

- Therefore,

$$P = \int 1\left\{z_j \leq \widetilde{c}_j - \widetilde{w}_{j1}\ z_1 - ... - \widetilde{w}_{j,j-1}\ z_{j-1} \text{ for any } j\right\}\ \phi(\mathbf{z})\ d\mathbf{z}$$

GHK Simulator (4)

- Consider the following IS density, $f^*(\mathbf{z}^*)$

  [1] $z_1^* \equiv \{z_1 | z_1 \leq \widetilde{c}_1\}$ is a random draw from the standard normal right truncated at $\widetilde{c}_1$.

  [2] Given z_1^*, then $z_2^* \equiv \{z_2 | z_2 \leq \widetilde{c}_2 - \widetilde{w}_{21}\ z_1^*\}$ is a random draw from the standard normal right truncated at $\widetilde{c}_2 - \widetilde{w}_{21}\ z_1^*$.

  ...

  [j] Given $(z_1^*, .., z_{j-1}^*)$, then $z_j^* \equiv \{z_j | z_j \leq \widetilde{c}_j - \widetilde{w}_{j1}\ z_1^* - ... - \widetilde{w}_{j,j-1}\ z_{j-1}^*\}$ is a random draw from the standard normal right truncated at $\widetilde{c}_j - \widetilde{w}_{j1}\ z_1^* - ... - \widetilde{w}_{j,j-1}\ z_{j-1}^*$.

GHK Simulator (5)

• What is the form of the IS density $f^*(\mathbf{z}^*)$? Note that the density of a random variable z^* that is a right-truncated normal at c is $\dfrac{\phi(z^*)}{1-\Phi(c)}$, where here ϕ and Φ represent the pdf and cdf of the standard normal, respectively. Then, by definition:

$$
\begin{aligned}
& f^*(\mathbf{z}^*) \\
= \; & \frac{\phi(z_1^*)\ \phi(z_2^*)\ ...\ \phi(z_J^*)}{[1-\Phi(\widetilde{c}_1)]\left[1-\Phi(\widetilde{c}_2-\widetilde{w}_{21}\ z_1^*)\right]...\left[1-\Phi(\widetilde{c}_J-\widetilde{w}_{J1}z_1^*-...-\widetilde{w}_{J,J}z_J^*)\right]} \\
= \; & \frac{\prod_{j=1}^{J}\phi(z_1^*)\ \phi(z_2^*)\ ...\ \phi(z_J^*)}{\prod_{j=1}^{J}\left[1-\Phi(\widetilde{c}_j-\widetilde{w}_{j1}z_1^*-...-\widetilde{w}_{j,J}z_J^*)\right]}
\end{aligned}
$$

GHK Simulator (5)

• The GHK simulator of P is the ISS that uses IS density $f^*(\mathbf{z}^*)$. Let $\{\mathbf{z}_r^* : r = 1, 2, ..., R\}$ be R independent random draws from the IS density $f^*(\mathbf{z}^*)$. Then,

$$\widetilde{P}_{GHK}^{(R)} = \frac{1}{R}\sum_{r=1}^{R} 1\left\{z_{jr}^* \leq \widetilde{c}_j - \widetilde{w}_{j1}\ z_1^* - ... - \widetilde{w}_{j,j-1}\ z_{j-1}^* \text{ for any } j\right\} \frac{\phi(\mathbf{z}_r^*)}{f^*(\mathbf{z}_r^*)}$$

Note that:

- $\dfrac{\phi(\mathbf{z}_r^*)}{f^*(\mathbf{z}_r^*)} = \prod_{j=1}^{J}\left[1 - \Phi(\widetilde{c}_j - \widetilde{w}_{j1}z_1^* - ... - \widetilde{w}_{j,j-1}z_{J-1}^*)\right]$

- By construction of the $\mathbf{z}_r^*$ simulations, the indicator of $\left\{z_{jr}^* \leq \widetilde{c}_j - \widetilde{w}_{j1}\ z_1^* - ... - \widetilde{w}_{j,j}\right.$ is always 1.

- Therefore,

$$\widetilde{P}_R^{GHK} = \frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{J}\left[1-\Phi(\widetilde{c}_j - \widetilde{w}_{j1}z_1^* - ... - \widetilde{w}_{j,j-1}z_{J-1}^*)\right]$$

Properties of GHK Simulator

[1] It is unbiased, consistent, asymptotically normal.

[2] It has substantially lower variance than the FS. In some standard settings the ratios of variances can the of the order of 100 or even 1000.

[3] $\widetilde{P}_R^{GHK}(\mathbf{c}, \mathbf{\Omega})$ is continuously differentiable in the parameters $(\mathbf{c}, \mathbf{\Omega})$.

[4] $\widetilde{P}_R^{GHK}$ is always strictly greater than 0 and lower than 1.

[5] It is simple to get random draws from a truncated standard normal. If z^* that is a right-truncated normal at c then its CDF is $F(z^*) = \dfrac{\Phi(z^*) - \Phi(c)}{1 - \Phi(c)}$, such that given $u \sim U[0, 1]$,

$$z^* = F^{-1}(u) = \Phi^{-1}\left(\Phi(c) + [1 - \Phi(c)]\ u\right)$$

8. SIMULATION-BASED ESTIMATION (SBE)

8.1. Refreshing Estimation and Asymptotic Theory

8.2. SBE: Conditions on the Simulators

8.3. Simulated Based Estimators: SMM and SMLE

8.4. Asymptotic Properties

8.1 REFRESHING ESTIMATION & ASYMPTOTIC THEORY

• Consider a discrete choice model with CCPs $P(j|x_n, \theta)$, where θ is a $K \times 1$ the vector of parameters. Let θ_0 be the true value of θ in the population under study. Let $\{y_n, x_n : n = 1, 2, ..., N\}$ be a random sample from the population.

• The model implies the following moment conditions:

$$\mathbb{E}\left(\sum_{j=1}^{J} z_{jn} \left[1\{y_n = j\} - P(j|x_n, \theta_0)\right]\right) = 0$$

where z_{jn} is a $K \times 1$ vector of functions of x_n, e.g., $z_{jn} = (x'_{jn}, \sum_{i \neq j} x'_{in}, [x_{jn} * x_{jn}]')$.

• The population likelihood equations, is particular example of these moment conditions. In this case, $z_{jn} = \frac{\partial \ln P(j|x_n,\theta_0)}{\partial \theta'}$:

$$\mathbb{E}\left(\sum_{j=1}^{J} \frac{\partial \ln P(j|x_n,\theta_0)}{\partial \theta'} \left[1\left\{y_n = j\right\} - P(j|x_n,\theta_0)\right]\right) = 0$$

• We can represent these moment conditions in a compact form as:

$$\mathbb{E}\left(\mathbf{z}_n \ \left[\mathbf{y}_n - \mathbf{P}_n(\theta_0)\right]\right) = 0$$

where:

$\mathbf{z}_n$ is the $K \times J$ matrix $(z_{1n}, z_{2n}, ..., z_{Jn})$;

$\mathbf{y}_n$ is the $J \times 1$ vector $\left(1\left\{y_n = 1\right\}, 1\left\{y_n = 2\right\}, ..., 1\left\{y_n = J\right\}\right)'$;

$\mathbf{P}_n(\theta)$ is the $J \times 1$ vector $(P(1|x_n, \theta), P(2|x_n, \theta), ..., P(J|x_n, \theta))'$.

Identification Assumption. θ_0 is the unique value in the parameter space Θ that solves the system of equations $\mathbb{E}\left(\mathbf{z}_n \ \left[\mathbf{y}_n - \mathbf{P}_n(\theta_0)\right]\right) = 0$.

ESTIMATION. The estimator $\widehat{\theta}_N$ is the value θ that solves the system of sample moment conditions:

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{z}_n\ [\mathbf{y}_n - \mathbf{P}_n(\theta)] = 0$$

• **Example: MLE.** When $z_{jn} = \frac{\partial \ln P(j|x_n,\theta)}{\partial \theta'}$, and $\mathbf{z}_n = \frac{\partial \ln \mathbf{P}_n(\theta)'}{\partial \theta}$, we have that the sample moment conditions above define the MLE:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \ln \mathbf{P}_n(\theta)'}{\partial \theta}\ [\mathbf{y}_n - \mathbf{P}_n(\theta)] = 0$$

• **Example: MM**

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{z}_n\ [\mathbf{y}_n - \mathbf{P}_n(\theta)] = 0$$

CONSISTENCY. Suppose that: (a) $\mathbf{P}_n(\theta)$ is continuously differentiable in θ; (b) for any $\theta \in \Theta$, we have that $Var(\mathbf{z}_n\,[\mathbf{y}_n - \mathbf{P}_n(\theta)])$ is finite; (c) Θ is a compact set; and (d) θ_0 is the unique value in the parameter space Θ that solves the system of equations $\mathbb{E}(\mathbf{z}_n\,[\mathbf{y}_n - \mathbf{P}_n(\theta)]) = 0$. Then, as $N \to \infty$,

(i) $\frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n\,[\mathbf{y}_n - \mathbf{P}_n(\theta)]$ converges in probability uniformly in $\theta \in \Theta$ to $\mathbb{E}(\mathbf{z}_n\,[\mathbf{y}_n - \mathbf{P}_n(\theta)])$;

(ii) $\widehat{\theta}_N \to_p \theta_0$.

ASYMPTOTIC NORMALITY. Let $g_n(\theta) \equiv \mathbf{z}_n\,[\mathbf{y}_n - \mathbf{P}_n(\theta)]$. Using a Taylor expansion of $\frac{1}{N}\sum_{n=1}^{N} g_n(\widehat{\theta}_N) = 0$ around $\theta = \theta_0$:

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right) = \left[\frac{1}{N}\sum_{n=1}^{N}\frac{\partial g_n(\theta_0)}{\partial \theta'}\right]^{-1}\left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N} g_n(\theta_0)\right] + o(1)$$

ASYMPTOTIC NORMALITY [Cont.] Then, under standard regularity conditions, we have that $\frac{1}{N}\sum_{n=1}^{N}\frac{\partial g_n(\theta_0)}{\partial \theta'} \rightarrow_p \mathbb{E}\left(\frac{\partial g_n(\theta_0)}{\partial \theta'}\right) \equiv G_0$, and $\frac{1}{\sqrt{N}}\sum_{n=1}^{N} g_n(\theta_0) \rightarrow_d N(0,\Omega_0)$ with $\Omega_0 = \mathbb{E}\left(g_n(\theta_0)\ g_n(\theta_0)'\right)$. By Slutsky's Theorem:

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right) \rightarrow_d N(0, G_0^{-1}\ \Omega_0\ \left(G_0^{-1}\right)')$$

- In our Discrete Choice Model we have that:

$$G_0 = \mathbb{E}\left(\mathbf{z}_n\ \frac{\partial \mathbf{P}_n(\theta_0)}{\partial \theta'}\right)$$

$$\Omega_0 = \mathbb{E}\left(\mathbf{z}_n\ \mathbf{\Sigma}_n^{\mathbf{P}}\ \mathbf{z}_n'\right)$$

where $\mathbf{\Sigma}_n^{\mathbf{P}}$ is $J \times J$ matrix where the element (j,j) in the main diagonal is $P(j|x_n,\theta_0)\ [1 - P(j|x_n,\theta_0)]$, and the element (j,i) out of the main diagonal is $-P(j|x_n,\theta_0)\ P(i|x_n,\theta_0)$.

8.2. SIMULATION-BASED ESTIMATION

• Given a model defined by the moment conditions $\mathbb{E}(\ g_n(\theta_0)\ ) = 0$, a Simulated-Based-Estimator is the value $\widehat{\theta}_{N,R}$ that solves the moment conditions:

$$\frac{1}{N}\sum_{n=1}^{N} \widetilde{g}_n^R(\theta) = \frac{1}{N}\sum_{n=1}^{N} \widetilde{\mathbf{z}}_n^R(\theta)\ \left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right] = 0$$

• $\widetilde{\mathbf{P}}_n^R(\theta)$ is the $J \times 1$ vector of simulators ($\widetilde{P}_R(1|x_n,\theta)$, $\widetilde{P}_R(2|x_n,\theta)$, ..., $\widetilde{P}_R(J|x_n,\theta)$)$'$.

• Note that for the SMLE, $\mathbf{z}_n(\theta) = \dfrac{\partial \ln \mathbf{P}_n(\theta)'}{\partial\theta}$. Therefore, we also need a simulator for $\mathbf{z}_n(\theta)$, say $\widetilde{\mathbf{z}}_n^R(\theta)$.

• Note also that if we use the same simulator, $\widetilde{\mathbf{P}}_n^R(\theta)$, for $\mathbf{P}_n(\theta)$ and for $\dfrac{\partial \ln \mathbf{P}_n(\theta)'}{\partial \theta}$:

(1) one of the two simulators is biased for finite R;

(2) the simulation errors in $\widetilde{\mathbf{P}}_n^R(\theta)$ and $\widetilde{\mathbf{z}}_n^R(\theta)$ will be correlated.

SIMULATED BASED ESTIMATORS: SMM, SMLE, S-Scores

Simulated method of Moments (SMM). $\widehat{\theta}_{N,R}$ that solves:

$$\sum_{n=1}^{N} \mathbf{z}_n \left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right] = 0 \qquad \text{with } \widetilde{\mathbf{P}}_n^R(\theta) \text{ unbiased sim. of } \mathbf{P}_n(\theta)$$

Simulated maximum likelihood (SML). $\widehat{\theta}_{N,R}$ that solves:

$$\sum_{n=1}^{N} \frac{\partial \ln \widetilde{\mathbf{P}}_n^R(\theta)'}{\partial \theta} \left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right] = 0 \text{ with } \widetilde{\mathbf{P}}_n^R(\theta) \text{ unb. im. of } \mathbf{P}_n(\theta)$$

Simulated Scores (SScores). $\widehat{\theta}_{N,R}$ that solves:

$$\frac{1}{N}\sum_{n=1}^{N} \widetilde{\mathbf{s}}_n^R(\theta) \left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right] = 0$$

$\widetilde{\mathbf{s}}_n^R(\theta)$ and $\widetilde{\mathbf{P}}_n^R(\theta)$ unb. & independent sims of $\frac{\partial \ln \mathbf{P}_n(\theta)'}{\partial \theta}$ and $\mathbf{P}_n(\theta)$, resp.

Conditions on Simulator [McFadden, ECMA 1989] particularized to the RC Logit

[1] The random draws $\mathbf{v}_n^{(r)} = \{v_{nk}^{(r)} : k = 1, 2, ..., K\}$ from the Standard normal are indepently distributed over k, n, and r. Each observation n (and each random coefficient v_{nk}) has its own R independent random draws.

[2] These random draws are made at the beginning of the estimation procedure and they are kept fixed during the implementation of the algorithm that searches for $\widehat{\theta}_{N,R}$. That is, the same set of random draws is used to construct simulators for different values of θ.

• If new drawings were made at each [Newton] iteration of the gradient algorithm, they would introduce new randomness at each step and it would not be possible to obtain numerical convergence of the algorithm, and the asymptotic properties of the estimator would not hold: McFadden (Econometrica, 1989).

• Note that some components in θ are parameters in the distribution of the random coefficients β_n, i.e., parameters $\mathbf{b}$ and $\mathbf{W}$. The values of these parameters change during our search for the estimator $\widehat{\theta}_{N,R}$. Therefore, we cannot keep constant the random draws from the distribution of the random coefficients β_n

• However, we can always keep constant the random draws from the distribution of the standard normals in $\mathbf{v}_n$. [Or more generally, the random draws from a $U[0,1]$ that we can use to construct draws from any distribution].

Conditions on Simulator [Cont.]

[3] The simulator $\widetilde{P^R}(j|x_n,\theta)$ is continuously differentiable in θ, and it is always within $(0,1)$.

[4] For any value (j,x_n,θ), the simulator $\widetilde{P^R}(j|x_n,\theta)$ is unbiased, and as R goes to infinity, it is consistent, and asymptotically normal:

$$\mathbb{E}\left[\widetilde{P^R}(j|x_n,\theta)\right] = P(j|x_n,\theta)$$

$$\text{As } R\to\infty, \quad \widetilde{P^R}(j|x_n,\theta) \to_p P(j|x_n,\theta)$$

$$\text{As } R\to\infty, \quad \sqrt{R}\left[\widetilde{\mathbf{P}}_n^R(\theta)-\mathbf{P}_n(\theta)\right] \to_p N(0,\widetilde{V}(x_n,\theta))$$

where $\widetilde{V}(x_n,\theta)$ is the variance matrix of the J simulation errors.

8.3. ASYMPTOTIC PROPERTIES

• There are two types of asymptotics we can consider for SB Estimators.

- As $N \to \infty$ and R is fixed.

- As $N \to \infty$ and $R \to \infty$.

• Asymptotics as $N \to \infty$ and R is fixed are particularly interesting because they fully take into account how simulation error affects the asymptotic bias and variance of SB Estimators.

• We start presenting asymptotic results as $N \to \infty$ and R is fixed.

A useful decomposition

• For the derivation of the asymptotic results of SBEs, it is helpful to consider the following decomposition of the conditions that define the estimator:

$$\frac{1}{N}\sum_{n=1}^{N}\widetilde{g}_n^R(\theta) \quad =$$

$$= \frac{1}{N}\sum_{n=1}^{N} g_n(\theta) \quad \textbf{[A] Without Sim. error}$$

$$+\frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) - g_n(\theta)\right] \quad \textbf{[B] Simulation Bias}$$

$$+\frac{1}{N}\sum_{n=1}^{N}\left[\widetilde{g}_n^R(\theta) - \mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right)\right] \quad \textbf{[C] Simulation Noise}$$

where $\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right)$ represents the expectation over the simulated random draws $\{v_{nr}\}$ but conditional on the observed data (y_n, x_n).

Term [A]: Standard Moment Conditions [Without Sim. Error]

$$\frac{1}{N}\sum_{n=1}^{N} g_n(\theta) \quad \textbf{[A] Without Sim. error}$$

• Under standard regularity conditions, we have that: $N^{-1}\sum_{n=1}^{N} g_n(\widehat{\theta}) = 0$ implies:

$$0 = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} g_n(\theta_0) + \left[\frac{1}{N}\sum_{n=1}^{N}\frac{\partial g_n(\theta_0)}{\partial \theta'}\right]\sqrt{N}\left(\widehat{\theta} - \theta_0\right) + o(1)$$

- $\frac{1}{\sqrt{N}}\sum_{n=1}^{N} g_n(\theta_0)$ converges in distribution to $N(0, \Omega_0)$.
- $\frac{1}{N}\sum_{n=1}^{N}\frac{\partial g_n(\theta_0)}{\partial \theta'}$ converges in probability to $\mathbb{E}\left(\frac{\partial g_n(\theta_0)}{\partial \theta'}\right)$.

- In the discrete choice model:

$$G_0 = \mathbb{E}\left(\mathbf{z}_n \, \frac{\partial \mathbf{P}_n(\theta_0)}{\partial \theta'}\right) \quad \text{and} \quad \Omega_0 = \mathbb{E}\left(\mathbf{z}_n \, \boldsymbol{\Sigma}_n^{\mathbf{P}} \, \mathbf{z}_n'\right)$$

Term [B]: Simulation Bias

$$\frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) - g_n(\theta)\right] \quad \textbf{[B] Simulation Bias}$$

- In our model, for the **Simulated Method of Moments:**

$$\begin{aligned}
\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) &= \mathbb{E}_v\left(\mathbf{z}_n\left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right]\right) \\
&= \mathbf{z}_n\left[\mathbf{y}_n - \mathbb{E}_v\left(\widetilde{\mathbf{P}}_n^R(\theta)\right)\right] \\
&= \mathbf{z}_n\left[\mathbf{y}_n - \mathbf{P}_n(\theta)\right] = g_n(\theta)
\end{aligned}$$

- Therefore, for the SMM with unbiased simulator of CCPs, the Simulation Bias term is exactly zero for any value of θ, N, and R.

Term [B]: Simulation Bias

- For **Simulated Maximum Likelihood:**

$$\begin{aligned}
\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) &= \mathbb{E}_v\left(\frac{\partial \ln \widetilde{\mathbf{P}}_n^R(\theta)'}{\partial \theta}\left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right]\right) \\
&= \mathbb{E}_v\left(\left[\frac{\partial \ln \mathbf{P}_n(\theta)'}{\partial \theta} + \widetilde{\delta}_n^R(\theta)\right]\left[\mathbf{y}_n - \mathbf{P}_n(\theta) - \widetilde{e}_n^R(\theta)\right]\right) \\
&= g_n(\theta) + \mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\right)[\mathbf{y}_n - \mathbf{P}_n(\theta)] - \mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right) \\
&\neq g_n(\theta)
\end{aligned}$$

$\widetilde{\delta}_n^R(\theta)$ is the $K \times J$ matrix of simulation errors in $\frac{\partial \ln \widetilde{\mathbf{P}}_n^R(\theta)'}{\partial \theta}$;

$\widetilde{e}_n^R(\theta)$ is the the $J \times 1$ vector of simulation errors in $\widetilde{\mathbf{P}}_n^R(\theta)$.

Term [B]: Simulation Bias [SML]

• For SML, $\frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{E}_u\left(\widetilde{g}_n^R(\theta)\right) - g_n(\theta)\right] \neq 0$ because both $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\right) \neq 0$ and $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right) \neq 0$.

• $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\right) \neq 0$ because an unbiased simulator of the CCP typically implies a bias simulator of the derivative of the log-CCP:

$$\frac{\partial \ln P(j|x_n,\theta)}{\partial\theta} = \frac{\partial P(j|x_n,\theta)}{\partial\theta}\ \frac{1}{P(j|x_n,\theta)}$$

Note that simulation that enters additively in the simulator of $P(j|x_n,\theta)$, however enters in the denominator $\frac{1}{P(j|x_n,\theta)}$ in the simulator of $\frac{\partial \ln P(j|x_n,\theta)}{\partial\theta}$.

• $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right) \neq 0$ because simulation error in CCPs is correlated with simulation error in the derivatives of the log-CCPs.

Term [B]: Simulation Bias [SMLE]

• Importantly, this Simulation Bias does not go to zero as the sample size goes to infinity

$$p \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left[\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) - g_n(\theta)\right] = \mathbb{E}\left[\left(\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\right)[\mathbf{y}_n - \mathbf{P}_n(\theta)]\right)\right]$$

$$-\mathbb{E}\left[\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right)\right]$$

$$= -\mathbb{E}\left[\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right)\right] \neq 0$$

The first term is zero at $\theta = \theta_0$, but the second term is not zero.

• Therefore, **SML is inconsistent as $N \to \infty$ and R is fixed.** Consistency of the SML requires that as $N \to \infty$ the number of simulations R also goes to infinity.

Term [B]: Simulation Bias [Method of Simulated Scores]

- For **Simulated Scores:**

$$\begin{aligned}
\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) &= \mathbb{E}_v\left(\widetilde{\mathbf{s}}_n^R(\theta)\ \left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right]\right) \\
&= \mathbb{E}_v\left(\left[\mathbf{s}_n(\theta) + \widetilde{\delta}_n^R(\theta)\right]\left[\mathbf{y}_n - \mathbf{P}_n(\theta) - \widetilde{e}_n^R(\theta)\right]\right) \\
&= g_n(\theta)
\end{aligned}$$

Because, now the simulator $\widetilde{\mathbf{s}}_n^R(\theta)$ is such that $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\right) = 0$ and $\mathbb{E}_v\left(\widetilde{\delta}_n^R(\theta)\ \widetilde{e}_n^R(\theta)\right)$ 0.

Term [C]: Simulation Noise

$$\frac{1}{N}\sum_{n=1}^{N}\left[\widetilde{g}_n^R(\theta) - \mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right)\right] \quad \textbf{[C] Simulation Noise}$$

• For the **Simulated Method of Moments** we have showed that $\mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) = g_n(\theta)$. Therefore,

$$\begin{aligned}\widetilde{g}_n^R(\theta) - \mathbb{E}_v\left(\widetilde{g}_n^R(\theta)\right) &= \mathbf{z}_n\left[\mathbf{y}_n - \widetilde{\mathbf{P}}_n^R(\theta)\right] - \mathbf{z}_n\left[\mathbf{y}_n - \mathbf{P}_n(\theta)\right] \\ &= -\mathbf{z}_n\,\widetilde{\mathbf{e}}_n^R(\theta)\end{aligned}$$

• Given the properties of our simulator, we have that the vector of $K \times 1$ random variables $\mathbf{z}_n\,\widetilde{\mathbf{e}}_n^R(\theta)$ is such that, for any n and θ:

$$\mathbb{E}\left(\mathbf{z}_n\,\widetilde{\mathbf{e}}_n^R(\theta)\right) = 0$$

and conditional on $\mathbf{x}_n$, $\mathbb{E}(\widetilde{\mathbf{e}}_n^R(\theta)|\mathbf{x}_n) = 0$.

Asymptotic Distribution of the SMM

• Using a Taylor approximation around $\theta = \theta_0$ of the moment conditions of the SMM, and taking into account that $\widehat{\theta}_{N,R} \rightarrow_p \theta_0$, we have that:

$$\sqrt{N}\left(\widehat{\theta}_{N,R} - \theta_0\right) =$$

$$\left[\frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n \frac{\partial \widetilde{\mathbf{P}}_n^R(\theta_0)}{\partial \theta'}\right]^{-1} \left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \mathbf{z}_n \left[\mathbf{y}_n - \mathbf{P}_n(\theta_0) - \widetilde{\mathbf{e}}_n^R(\theta_0)\right]\right] + o(1)$$

• As $N \rightarrow \infty$ with R fixed, we have that:

$$\left[\frac{1}{N}\sum_{n=1}^{N} \mathbf{z}_n \frac{\partial \widetilde{\mathbf{P}}_n^R(\theta_0)}{\partial \theta'}\right] \rightarrow_p \widetilde{G}_R \equiv \mathbb{E}\left(\mathbf{z}_n \frac{\partial \widetilde{\mathbf{P}}_n^R(\theta_0)}{\partial \theta'}\right)$$

- Also:

$$\left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{z}_n\left[\mathbf{y}_n-\mathbf{P}_n(\theta_0)\right]\right]\ \rightarrow_d\ \ N\left(0\ ,\ \Omega_0\right)$$

$$\text{where:}\quad \Omega_0\ \equiv\ \mathbb{E}\left(\mathbf{z}_n\ \mathbf{\Sigma}_n^{\mathbf{P}}\ \mathbf{z}_n'\right)$$

- And:

$$\left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{z}_n\ \widetilde{\mathbf{e}}_n^R(\theta_0)\right]\ \rightarrow_d\ \ N\left(0\ ,\ \widetilde{\Omega}_R\right)$$

$$\text{where:}\quad \widetilde{\Omega}_R\ \equiv\ \mathbb{E}\left(\mathbf{z}_n\ \frac{\widetilde{V}(x_n,\theta_0)}{R}\ \mathbf{z}_n'\right)$$

where remember that $\dfrac{\widetilde{V}(x_n,\theta_0)}{R}$ is the variance matrix of the simulation errors in $\widetilde{\mathbf{P}}_n^R(\theta_0)$.

• And the terms $\left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{z}_n\left[\mathbf{y}_n-\mathbf{P}_n(\theta_0)\right]\right]$ and $\left[\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{z}_n\,\widetilde{\mathbf{e}}_n^R(\theta_0)\right]$ are independent due to the conditional mean independence of the simulation error, i.e., $\mathbb{E}(\widetilde{\mathbf{e}}_n^R(\theta)|\mathbf{x}_n)=0$.

• Therefore, applying Slutsky's Theorem, we have that the asymptotic distribution of the SSM estimator as $N\to\infty$ with R fixed, is:

$$\sqrt{N}\left(\widehat{\theta}_N-\theta_0\right)\to_d N\left(0\ ,\ \widetilde{G}_R^{-1}\ \left[\Omega_0+\widetilde{\Omega}_R\right]\ \left(\widetilde{G}_R^{-1}\right)'\right)$$

• As R goes to infinity, $\widetilde{G}_R\to G_0$, and $\widetilde{\Omega}_R\to 0$, such that the SMM estimator becomes equivalent to the MM estimator without simulation. But in any empirical application, with finite number of simulation R, the simulation error introduces additional noise that increases the variance of the estimator. This is fully taken into account by the expression of the asymptotic variance above.

WILLIAMS-DALY-ZACHARY (WDZ) THEOREM

• Consider the RUM $U(j) = u_j + \varepsilon_j$ where $\varepsilon = \{\varepsilon_j : j = 0, 1, ..., J\}$ has a CDF $F(\varepsilon)$ that is continuously differentiable over the whole Euclidean space $\mathbb{R}^{J+1}$.

• Define the **Social Surplus function** (McFadden, 1981)

$$S(\mathbf{u}) = \int \max_{j \in A} \left\{ u_j + \varepsilon_j \right\} \, dF(\varepsilon)$$

• Then,

$$\frac{\partial S(\mathbf{u})}{\partial u_j} = P(j)$$

Proof: Exercise. Hint: Note that $\partial \max_{j \in A} \left\{ u_j + \varepsilon_j \right\} / \partial u_j$ is equal to $1\{$ $u_j + \varepsilon_j \geq u_i + \varepsilon_i$ for any $i \neq j$ $\}$.

• Note that this result is like the discrete choice version of Roy's Theorem in Consumer Demand: The derivative of the indirect utility function with respect to price is equal to the demand.

HOTZ-MILLER PROPOSITION

• Consider the RUM $U(j) = u_j + \varepsilon_j$ where $\varepsilon = \{\varepsilon_j : j = 0, 1, ..., J\}$ has a CDF $F(\varepsilon)$ that is continuously differentiable over the whole Euclidean space $\mathbb{R}^{J+1}$.

• Define the vector of utility differences $\mathbf{u} = \{u_j - u_0 : j > 0\}$ and the vector of CCPs $\mathbf{P} = \{P(j) : j > 0\}$.

• Given the CDF $F(\varepsilon)$, the definition of CCPs provides a mapping from the vector of utility differences $\mathbf{u}$ into the vector of CCPs, $\mathbf{P}$.

$$\mathbf{P} = G(\mathbf{u})$$

Hotz and Miller show that this mapping is invertible:

$$\mathbf{u} = G^{-1}(\mathbf{P})$$

There is a unique vector of utility differences $\mathbf{u}$ that can rationalize (generate as optimal choices) a vector of CCPs $\mathbf{P}$. Revealed Preference.