Classical Discrete Choice Theory

James J. Heckman
University of Chicago

Econ 312, Spring 2022

THE UNIVERSITY OF CHICAGO

- Consider how to solve P3 forecasting a new policy never previously experienced.
- Suppose we want to forecast demand for a new good. We observe consumption data on old goods $x_1 \ldots x_I$. (Each good could represent a transportation mode, for example, or an occupation choice.)
- Assume people choose a good that yields highest utility. When we have a new good, we need a way of putting it on a basis with the old.
- Earliest literature on discrete choice was developed in psychometrics where researchers were concerned with modeling choice behavior (Thurstone).
- These are also models of counterfactual utilities.

Two dominant modeling approaches

i. Luce Model (1953) $\Longleftrightarrow$ McFadden Conditional logit model
ii. Thurstone-Quandt Model (1929, 1930s). (Multivariate probit/normal model)
iii. We start with parametric models but will relax, principles are general
iv. At the same time, some specifications are widely used and you should know them

Two Basic Frameworks

- GEV models
  - Includes conditional logit
  - widely used in economics
  - easy to compute
  - identifiability of parameters understood
  - very restrictive substitution possibilities among goods
  - restrictive heterogeneity
  - imposes arbitrary preference shocks
- Quandt-Thurstone Model
  - very general substitution possibilities
  - allows for more general forms of heterogeneity
  - more difficult to compute
  - identifiability less easily established
  - does not necessarily rely on preference shocks

Luce Model/McFadden Conditional Logit Model

- References: Manski and McFadden, Chapter 5 (posted on McFadden's website), Yellot paper
- Notation:
  - X: universe of objects of choice
  - S: universe of attributes of persons
  - B: feasible choice set ($x \in B \subseteq X$)

Luce Model/McFadden Conditional Logit Model

- Behavior rule mapping attributes into choices: h

$$h(B, S) = x$$

- We might assume that there is a distribution of choice rules.
- h might be random because
  a. in observation we lose some information governing choices (unobserved characteristics of choice and person)
  b. there can be random variation in choices due to unmeasured psychological factors
  c. the arrival of information ex ante vs ex post
- Define $P(x|S, B) = \Pr\{h \epsilon H \ni h(S, B) = x\}$
- Probability that an individual drawn randomly from the population with attributes S and alternative set B chooses x.

Luce Axioms

- Maintain some restrictions on $P(x|S,B)$ and derive implications for the functional form of P.
- Axiom #1: "Independence of Irrelevant Alternatives"

$$\begin{array}{rcl} x,y & \in & B \qquad s \in S \\ \dfrac{P(x|s,\{xy\})}{P(y|s,\{xy\})} & = & \dfrac{P(x|s,B)}{P(y|s,B)} \end{array}$$

B = larger choice set

Luce Axioms

- Example: Suppose choice is career decision and individual is choosing to be
  - an economist (E)
  - a fireman (F)
  - a policeman (P)

$$\frac{\Pr(E|s,\{EF\})}{\Pr(F|s,\{EF\})} = \frac{\Pr(E|s,\{EFP\})}{\Pr(F|s,\{EFP\})} \longleftarrow$$

would think that introducing 3rd alternative might increase ratio

Luce Axioms

- Another example: Red bus-Blue bus
- Choices:
  - take car C
  - red bus RB
  - blue bus BB

Luce Axioms

- Axiom #2

  $\Pr(y|s, B) > 0 \quad \forall y \in B$ (i.e. eliminate 0 probability choices)

Implications of above axioms

- Define $P_{xy} = P(x|s, \{xy\})$
- Assume $P_{xx} = \frac{1}{2}$

$$P(y|s,B) = \frac{P_{yx}}{P_{xy}} P(x|s,B) \text{ by IIA axiom}$$

$$\sum_{y \in B} P(y|s,B) = 1 \Longrightarrow P(x|s,B) = \frac{1}{\sum_{y \in B} \frac{P_{yx}}{P_{xy}}}$$

Implications of above axioms

- Furthermore,

$$
\begin{aligned}
P(y|s,B) &= \frac{P_{yz}}{P_{zy}} P(z|s,B) \\
P(x|s,B) &= \frac{P_{xz}}{P_{zx}} P(z|s,B) \\
P(y|s,B) &= \frac{P_{yx}}{P_{xy}} P(x|s,B) \\
\frac{P_{yx}}{P_{xy}} &= \frac{P(y|s,B)}{P(x|s,B)} = \frac{\frac{P_{yz}}{P_{zy}}}{\frac{P_{xz}}{P_{zx}}}
\end{aligned}
$$

Implications of above axioms

- Define

$$\tilde{v}(s,x,z) = \ln\left(\frac{P_{xz}}{P_{zx}}\right) \qquad \tilde{v}(s,y,z) = \ln\left(\frac{P_{yz}}{P_{zy}}\right)$$

$$\Longrightarrow \quad \frac{P_{yx}}{P_{xy}} = \frac{e^{\tilde{v}(s,y,z)}}{e^{\tilde{v}(s,x,z)}}$$

- Axiom #3: Separability Assumption

$$\tilde{v}(s,x,z) = v(s,x) - v(s,z) \quad \longleftarrow \quad v(s,z) \text{ can be interpreted as a utility indicator of representative tastes}$$

- Then

$$P(x|s,B) = \frac{1}{\sum_{y \in B} \frac{P_{yx}}{P_{xy}}} = \frac{1}{\sum_{y \in B} \frac{e^{v(s,y)-v(s,z)}}{e^{v(s,x)-v(s,z)}}}$$

$$P(x|s,B) = \frac{e^{v(s,x)}}{\sum_{y \in B} e^{v(s,y)}} \longleftarrow \text{Get logistic from from Luce Axioms}$$

- Now link model to familiar models in economics.
- Marshak (1959) established link between Luce Model and random utility models (Rum's).

Random Utility Models: Thurstone (1927, 1930s)

- Assume utility from choosing alternative j is

$$u_j = v\left(s, x_j\right) + \varepsilon(s, x_j)$$

- $v\left(s, x_j\right)$is a nonstochastic function and $\varepsilon(s, x_j)$ is stochastic, reflecting idiosyncratic tastes.

Random Utility Models: Thurstone (1927, 1930s)

- $\Pr(j \text{ is maximal in set } B) = \Pr\left(u(s, x_j) \geq u(s, x_l)\right) \quad \forall l \neq j$

$$
\begin{aligned}
&= \Pr\left(v(s, x_j) + \varepsilon(s, x_j) \geq v(s, x_l) + \varepsilon(s, x_l)\right) \quad \forall l \neq j \\
&= \Pr\left(v(s, x_j) - v(s, x_l) \geq \varepsilon(s, x_l) - \varepsilon(s, x_j)\right) \quad \forall l \neq j
\end{aligned}
$$

- Specify a cdf $F(\varepsilon_1, ..., \varepsilon_N)$
- Then

$$\begin{aligned}
\Pr(v_j - v_I \quad &\geq \quad \varepsilon_I - \varepsilon_j \quad \forall I \neq j) \\
&= \quad \Pr(v_j - v_I + \varepsilon_j \geq \varepsilon_I \quad \forall I \neq j) \\
&= \quad \int_{-\infty}^{\infty} F_j(\, v_j - v_1 + \varepsilon_j, ..., v_j - v_{j-1} + \varepsilon_j, \varepsilon_j, ..., v_j - v_J + \varepsilon_j) d\varepsilon_j
\end{aligned}$$

(Prove)

- If ε is iid, then

$$F(\varepsilon_1, ..., \varepsilon_n) = \prod_{i=1}^{n} F_i(\varepsilon_i)$$

- So $\Pr(v_j - v_l \geq \varepsilon_l - \varepsilon_j \quad \forall l \neq j)$

$$\int_{-\infty}^{\infty} \left[\prod_{\substack{i=1 \\ i \neq j}}^{n} F_i(v_j - v_i + \varepsilon_j) \right] f_j(\varepsilon_j) d\varepsilon_j$$

Binary Example ($N = 2$)

$$P(1 \mid s, B) = \int_{-\infty}^{\infty} \int_{-\infty}^{v_1 - v_2 + \varepsilon_1} f_1(\varepsilon_1, \varepsilon_2)\, d\varepsilon_1\, d\varepsilon_2$$

- If $\varepsilon_1, \varepsilon_2$ are normal then $\varepsilon_1 - \varepsilon_2$ is normal, so $\Pr(v_1 - v_2 \geq \varepsilon_1 - \varepsilon_2)$ is normal.
- If $\varepsilon_1, \varepsilon_2$ are Weibull then $\varepsilon_1 - \varepsilon_2$ is logistic

$$\varepsilon \sim \text{Weibull} \Longrightarrow \Pr(\varepsilon < c) = e^{-e^{-c+\alpha}}$$

- Also called double exponential" or "Type I extreme value"
- ε Weibull

$$\begin{aligned}\Pr(v_1 + \varepsilon_1 > v_2 + \varepsilon_2) &= \Pr(v_1 - v_2 > \varepsilon_2 - \varepsilon_1) \\ &= \Omega(v_1 - v_2) = \frac{e^{v_1 - v_2}}{1 + e^{v_1 - v_2}} = \frac{e^{v_1}}{e^{v_1} + e^{v_2}}\end{aligned}$$

- Result: Assuming that the errors follow a Weibull distribution yields same logit model derived from the Luce Axioms.
- This link was established by Marshak (1959)
- Turns out that Weibull is sufficient but not necessary.
- Some other distributions for ε generate a logit.
- Yellot (1977) showed that if we require "*invariance under uniform expansions of the choice set*" then only double exponential gives logit.
- Example: Suppose choice set is {coffee, tea, mild}, then "invariance" requires that probabilities stay the same if we double the choice set (i.e., 2 coffees, 2 teas, 2 milks).
- This is a form of homotheticity.

Some Important Properties of the Weibull

- Developed 1928 (Fisher & Tippet showed it's one of 3 possible limiting distributions for the maximum of a sequence of random variables)
- Closed under maximization (i.e. max of n Weibulls is a Weibull)

$$\Pr(\max_i \varepsilon_i \leq c) = \prod_i e^{-e^{-(c+\alpha_i)}} = e^{-\sum_i e^{-c} e^{\alpha_i}} = e^{-e^{-c}\sum_i e^{\alpha_i}} = e^{-e^{-c+\ln \sum_i e^{\alpha_i}}}$$

- Difference between two Weibulls is a logit
- Under Luce axioms (on R.U.M. with Weibull assumption)

$$\Pr(j \mid s, B) = \frac{e^{v(s,x_j)}}{\sum_{l=1}^{N} e^{v(s,x_l)}}$$

- Now reconsider *the forecasting problem, Problem P3:*
- Let x_j =set of characteristics associated with choice j
- Usually, it is assumed that $v(s, x_j) = \theta(s)' x_j$
- This is a matter of convenience, not at all essential here
- Need to know functional form of V
- Dependence of θ on s reflects fact that individuals differ in their evaluation of characteristics.

- Get

$$\Pr(j \mid s, B) = \frac{e^{\theta(s)' x_j}}{\sum_{l=1}^{N} e^{\theta(s)' x_l}}$$

- Likelihood

$$\max_{\theta(s)} \prod_{i=1}^{N} \left(\frac{\left[e^{\theta(s)x_1}\right]^{D_{1_i}}}{\sum_{i=1}^{N} e^{\theta(s)x_i}} \right) \left(\frac{\left[e^{\theta(s)x_2}\right]^{D_{1_i}}}{\sum_{i=1}^{N} e^{\theta(s)x_i}} \right) \cdots \left(\frac{\left[e^{\theta(s)x_N}\right]^{D_{N_i}}}{\sum_{i=1}^{N} e^{\theta}(s)X_i} \right)$$

- This assumes independence across the N observations

- If a new good has different values of the same set of characteristics, get probabilities by

$$B' = \{B, N+1\}$$

$$P(N+1 \mid B', s) = \frac{e^{\theta(s)' x_{N+1}}}{\sum_{l=1}^{N+1} e^{\theta(s)' x_l}}$$

- Here N is the number of goods

Debreu (1960) criticism of Luce Model

- "Red Bus - Blue Bus Problem"
- Suppose $N+1^{th}$ alternative is identical to the first

$$\Pr(\text{choose 1 or } N+1 \mid s, B') = \frac{2e^{\theta(s)'x_{N+1}}}{\sum_{l=1}^{N+1} e^{\theta(s)'x_l}}$$

- $\Longrightarrow$ Introduction of identical good changes probability of riding a bus.
  - not an attractive result
  - comes from need to make iid assumption on new alternative

Why?

Debreu (1960) criticism of Luce Model: Some Alternative Assumptions

1. Could let $v_i = \ln(\theta(s)' x_i)$

$$\Pr(j \mid s, B) = \frac{\theta(s)' x_j}{\sum_{l=1}^{N+1} \theta(s)' x_l}$$

If we also imposed $\sum_{l=1}^{N} \theta(s)' x_l = 1$, we would get linear probability model but this could violate IIA.

Debreu (1960) criticism of Luce Model: Some Alternative Assumptions

1. Could consider model of form

$$\Pr(j \mid s, B) = \frac{e^{\theta^j(s)x^i}}{\sum_{l=1}^{N} e^{\theta^l(s)x_l}}$$

but here we have lost our forecasting ability (cannot predict demand for a new good).

2. This is called the multinomial logit model in statistics
3. Universal Logit Model

$$\Pr(i \mid s, x_1, \ldots, x_N) = \frac{e^{\varphi_i(x_1,\ldots,x_N)\beta(s)}}{\sum_{l=1}^{N} e^{\varphi_l(x_1,\ldots,x_N)\beta(s)}}$$

Here we lose IIA and forecasting (Bernstein Polynomial Expansion).

More General Models

1. Goal: We want a probabilistic choice model that
   1. has a flexible functional form
   2. is computationally practical
   3. allows for flexibility in representing substitution patterns among choices
   4. is consistent with a random utility model (RUM) $\Longrightarrow$ has a structural interpretation

How do you verify that a candidate PCS is consistent with a RUM?

1. Goal:
   a. Either start with a R.U.M.

$$u^i = v(s, x^i) + \varepsilon(s, x^i)$$

   and solve integral for

$$\Pr(u^i > u^l,\ \forall l \neq i) = \Pr(i = \arg\max_l \left(v^l + \varepsilon^l\right))$$

   or

   b. start with a candidate PCS and verify that it is consistent with a R.U.M. (easier)
2. McFadden provides sufficient conditions
3. See discussion of Daley-Zachary-Williams theorem

Link to Airum Models

Daly-Zachary-Williams Theorem

- Daly-Zachary (1976) and Williams (1977) provide a set of conditions that makes it easy to derive a PCS from a RUM with a class of models ("generalized extreme value" (GEV) models)
- Define $G : G(Y_1, \ldots, Y_J)$
- If G satisfies the following
  1. nonnegative defined on $Y_1, \ldots, Y_J \geq 0$
  2. homogeneous degree one in its arguments
  3. $\lim_{Y_i \to \infty} G(Y_1, \ldots, Y_i, \ldots, Y_J) \to \infty, \ \forall i = 1, \ldots, J$
-

$$\frac{\partial^k G}{\partial Y_1 \cdots \partial Y_k} \quad \text{is} \quad \begin{array}{l} \text{nonnegative if } k \text{ odd} \\ \text{nonpositive if even} \end{array} \tag{1}$$

- Then for a R.U.M. with $u_i = v_i + \varepsilon_i$ and

$$F(\varepsilon_1, \ldots, \varepsilon_J) = \exp\left\{-G\left(e^{-\varepsilon_1}, \ldots, e^{-\varepsilon_J}\right)\right\}$$

- This cdf has Weibull marginals but allows for more dependence among ε's.
- The PCS is given by

$$P_i = \frac{\partial \ln G}{\partial v_i} = \frac{e^{v_i} G_i\left(e^{v_1}, \ldots, e^{v_J}\right)}{G\left(e^{v_1}, \ldots, e^{v_J}\right)}$$

- Note: McFadden shows that under certain conditions on the form of the indirect utility function (satisfies AIRUM form), the DZW result can be seen as a form of Roy's identity.

- Let's apply this result

$$\begin{aligned} \text{cdf } F(\varepsilon_1, \ldots, \varepsilon_J) &= e^{-e^{-\varepsilon_1}} \cdots e^{-e^{-\varepsilon_J}} \longleftarrow \text{product of iid} \\ &\quad \text{Weibulls} \\ &= e^{-\sum_{j=1}^{J} e^{-\varepsilon_j}} \end{aligned}$$

- Can verify that $G\left(e^{v_1}, \ldots, e^{v_J}\right) = \sum_{j=1}^{J} e^{v_i}$ satisfies DZW conditions

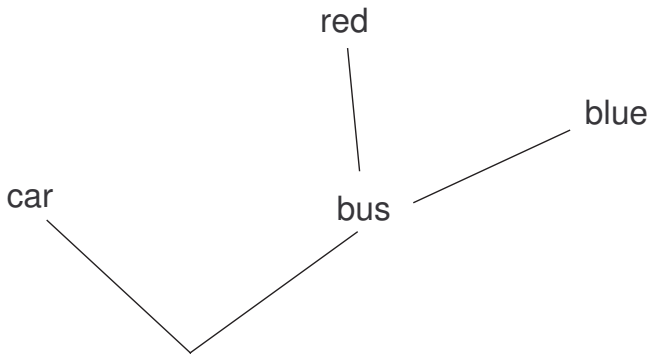
$$P(j) = \frac{\partial \ln G}{\partial v_i} = \frac{e^{v_j}}{\sum_{l=1}^{J} e^{v_l}} = \text{conditional logit model (CLM)}$$

- Another GEV model
- Nested logit model (addresses to a limited extent the IIA criticism)
- Let

$$G\left(e^{v_1}, \ldots, e^{v_J}\right) = \sum_{m=1}^{M} a_m \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}}\right]^{1-\sigma_m}$$

- σ_m like an elasticity of substitution.

- Idea: divide goods into branches
- First choose branch, then good within branch



- Will allow for correlation between errors (this is role of σ))
-

$$\begin{aligned} B_m &\subseteq \{1, \ldots, J\} \\ \bigcup_{m=1} B_m &= B \end{aligned}$$

is a single branch—need not have all choices on all branches

- Note: if $\sigma = 0$, get usual MNL form
- Calculate equation for

$$
\begin{aligned}
p_i &= \frac{\partial \ln G}{\partial v_i} = \frac{\partial \ln \left[\sum_{m=1}^{m} a_m \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}} \right]^{1-\sigma_m} \right]}{\partial v_i} \\
&= \frac{\sum_{m \ni i \in B_m} a_m \left(e^{\frac{v_i}{1-\sigma_m}} \right) \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}} \right]^{-\sigma_m} \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}} \right]^{-1} \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}} \right]}{\sum_{m=1}^{m} a_m \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}} \right]^{1-\sigma_m}} \\
&= \sum_{m=1}^{m} P(i \mid B_m) P(B_m)
\end{aligned}
$$

- Where

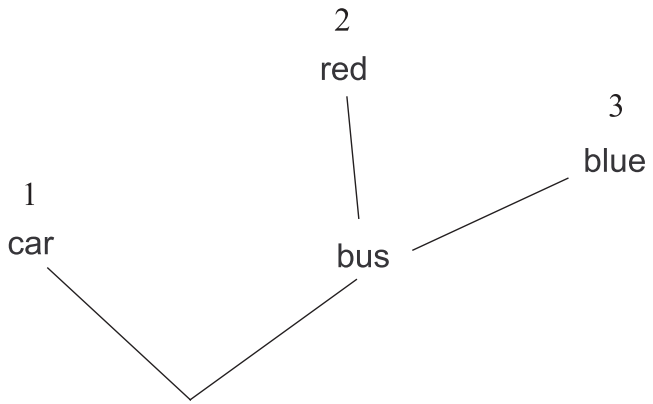
$$P(i \mid B_m) = \frac{e^{\frac{v_i}{1-\sigma_m}}}{\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}}} \text{ if } i \in B_m, \ 0 \text{ otherwise}$$

$$P(B_m) = \frac{a_m \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}}\right]^{1-\sigma_m}}{\sum_{m=1}^{m} a_m \left[\sum_{i \in B_m} e^{\frac{v_i}{1-\sigma_m}}\right]^{1-\sigma_m}}$$

- Note: If $P(B_m) = 1$ get logit form
- Nested logit requires that analyst make choices about nesting structure
- **Problem:** Prove this

- How does nested logit solve red bus/blue bus problem?
- Suppose

$$G = Y_1 + \left[Y_2^{\frac{1}{1-\sigma}} + Y_3^{\frac{1}{1-\sigma}} \right]^{1-\sigma} \qquad Y_i = e^{v_i}$$



$$P(1 \mid \{123\}) = \frac{\partial \ln G}{\partial v_i} = \frac{e^{v_1}}{e^{v_1} + \left[e^{\frac{v_2}{1-\sigma}} + e^{\frac{v_3}{1-\sigma}}\right]^{1-\sigma}}$$

$$P(2 \mid \{123\}) = \frac{\partial \ln G}{\partial v_i} = \frac{e^{\frac{v_2}{1-\sigma}}\left[e^{\frac{v_2}{1-\sigma}} + e^{\frac{v_3}{1-\sigma}}\right]^{-\sigma}}{e^{v_1} + \left[e^{\frac{v_2}{1-\sigma}} + e^{\frac{v_3}{1-\sigma}}\right]^{1-\sigma}}$$

- As $v_3 \to -\infty$

$$P(1 \mid \{123\}) = \frac{e^{v_1}}{e^{v_1} + e^{v_2}} \quad \text{(get logistic)}$$

What Role Does σ Play?

- σ is the degree of substitutability parameter
- Recall
$$F(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\{-G(e^{-\varepsilon_1}, e^{-\varepsilon_2}, e^{-\varepsilon_3})\}$$
- Here
$$\sigma = \frac{cov(\varepsilon_2, \varepsilon_3)}{\sqrt{var\ \varepsilon_2\ var\ \varepsilon_3}} = \text{correlation coefficient}$$
- Thus we require $-1 \leq \sigma \leq 1$, but turns out we also need to require $\sigma > 0$ for DZW conditions to be satisfied. This is unfortunate because it does not allow ε's to be negatively correlated.
- Can show that
$$\lim_{\sigma \to 1} P(1 \mid \{123\}) = \frac{e^{v_1}}{e^{v_1} + \max(e^{v_2}, e^{v_3})} \text{ (L'Hôpital's Rule)}$$

- If $v_2 = v_3$, then

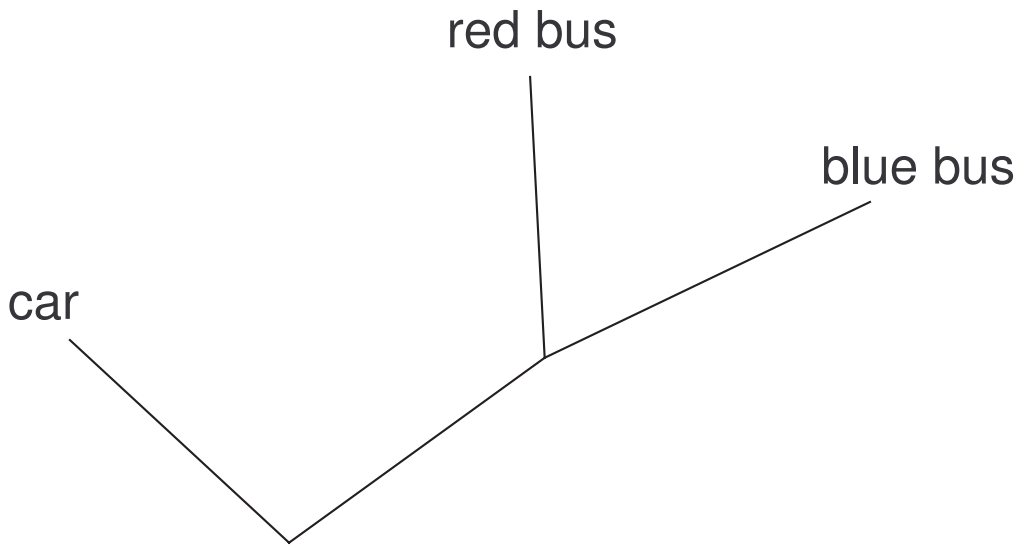
$$P(2 \mid \{123\}) = \frac{e^{\frac{v_2}{1-\sigma}}\left[2e^{\frac{v_2}{1-\sigma}}\right]^{-\sigma}}{e^{v_1} + \left[2e^{\frac{v_2}{1-\sigma}}\right]^{1-\sigma}}$$

$$= 2^{-\sigma}\frac{e^{v_2}}{e^{v_1} + \left(e^{v_2}\right)\left(2^{1-\sigma}\right)}$$

$$\lim_{\sigma \to 1} = 2^{-1}\frac{e^{v_2}}{e^{v_1} + e^{v_2}} \text{ when } v_1 = v_2$$

↗ introduce 3rd identical alternative and cut the probability of choosing 2 in half

- Solves red-bus/blue-bus problem
- Probability cut in half with two identical alternatives



- σ is a measure of similarity between red and blue bus.
- When σ close to one, the conditional choice probability selects with high probability the alternative.

We Can Expand Logit to Accommodate Multiple Nesting Levels

$$G = \sum_{q=1}^{Q} a_q \left\{ \sum_{m \in Q_q} a_m \left[\sum_{i \in B_m} y_i^{\frac{1}{1-\sigma_m}} \right]^{1-\sigma_m} \right\} \text{ 3 levels}$$

- Example: Two Choices
  1. Neighborhood (m)
  2. Transportation mode (t)
  3. $P(m)$: choice of neighborhood
  4. $P(i \mid B_m)$: probability of choosing i^{th} mode, given neighborhood m

1. Not all modes available in all neighborhoods

$$P_{m,t} = \frac{e^{\frac{v(m,t)}{1-\sigma_m}} \left[\sum_{t=1}^{T_m} e^{\frac{v(m,t)}{1-\sigma_m}}\right]^{-\sigma_m}}{\sum_{j=1}^{m} \left[\sum_{t=1}^{T_j} e^{\frac{v(m,t)}{1-\sigma_m}}\right]^{1-\sigma_m}}$$

$$P_{t|m} = \frac{e^{\frac{v(m,t)}{1-\sigma_m}}}{\sum_{t=1}^{T_m} e^{\frac{v(m,t)}{1-\sigma_m}}}$$

$$P_m = \frac{\left[\sum_{t=1}^{T_m} e^{\frac{v(m,t)}{1-\sigma_m}}\right]^{1-\sigma_m}}{\sum_{j=1}^{m} \left[\sum_{t=1}^{T_j} e^{\frac{v(m,t)}{1-\sigma_m}}\right]^{1-\sigma_m}} = P(B_m)$$

- Standard type of utility function that people might use

$$v(m,t) = z_t'\gamma + x_{mt}'\beta + y_m'\alpha$$

- z_t' is transportation mode characteristics, x_{mt}' is interactions and y_m' is neighborhood characteristics.
- Then

$$P_{t|m} = \frac{e^{\frac{(z_t'\gamma + x_{mt}'\beta)}{1-\sigma_m}}}{\left[\sum_{t=1}^{T_m} e^{\frac{(z_t'\gamma + x_{mt}'\beta)}{1-\sigma_m}}\right]}$$

$$P_m = \frac{e^{y_m'\alpha}\left[\sum_{t=1}^{T_m} e^{\frac{(z_t'\gamma + x_{mt}'\beta)}{1-\sigma_m}}\right]^{1-\sigma_m}}{\sum_{j=1}^{m} e^{y_m'\alpha}\left[\sum_{t=1}^{T_m} e^{\frac{(z_t'\gamma + x_{mt}'\beta)}{1-\sigma_j}}\right]^{1-\sigma_j}}$$

- Estimation (in two steps) (see Amemiya, Chapter 9)
- Let

$$I_m = \sum_{t=1}^{T_m} e^{\frac{(z_t'\gamma + x_{mt}'\beta)}{1-\sigma_m}}$$

1. Within each neighborhood, get $\frac{\widehat{\gamma}}{1-\sigma_m}$ and $\frac{\widehat{\beta}}{1-\sigma_m}$ by logit
2. Form $\widehat{I_m}$
3. Then estimate by MLE

$$\frac{e^{y_m'\alpha+(1-\sigma_m)\ln \widehat{I_m}}}{\sum_{j=1}^{m} e^{y_m'\alpha+(1-\sigma_j)\ln \widehat{I_j}}} \qquad \text{get } \widehat{\alpha}, \widehat{\sigma}_m$$

- Assume $\sigma_m = \sigma_j \ \forall j, m$ or at least need some restrictions across multiple neighborhoods?
- Note: $\widehat{I_m}$ is an estimated regressor ("Durbin problem")
- Need to correct standard errors

Multinomial Probit Models

1. Also known as:
   1. Thurstone Model V (1929; 1930)
   2. Thurstone-Quandt Model
   3. Developed by Domencich-McFadden (1978) (on reading list)

$$
\begin{aligned}
u_i &= v_i + \eta_i \quad i = 1, ..., J \\
v_i &= Z_i\beta \quad \text{(linear in parameters form)} \\
u_i &= Z_i\beta + \eta_i
\end{aligned}
$$

MNL	**MNP**
$(i)\beta$ fixed	$(i)\beta$ random coefficient $\beta \sim N\left(\bar{\beta}, \Sigma_\beta\right)$
$(ii)\eta_i$ iid	$(ii)\beta$ independent of $\eta \quad \eta \sim (0, \Sigma_\eta)$,

- Allow gen. forms of correlation between errors
- Digression $\rightarrow$

Random Coefficient Model

$$Y = \beta + U$$

β, X, U are all iid mutually independent random variables. Suppose

$$\begin{aligned} &\beta \perp\!\!\!\perp (X, U) \quad E(\beta) = \bar{\beta} < \infty \\ &X \perp\!\!\!\perp (\beta, U) \\ &E(Y|X) \\ =&X\bar{\beta} + \underbrace{X(\beta - \bar{\beta}) + U}_{\text{error term}} \end{aligned}$$

Digression, Continued

OLS identifies $\bar{\beta}$ under rank condition.

$$E(X(\beta - \bar{\beta}) + U) = 0$$

For observation i:

$$var(X_i(\beta - \bar{beta}) + U_i) = X_i' \sum_{\beta} X_i + \sigma_U^2.$$

Notice i. Regress squarely OLS residuals ($W = Y - X\hat{\beta}$) on quadratic terms in X_i.

Prove it can be used to identify σ_U^2 and $\sum_{\beta}$.

$$u_i = Z_i\bar{\beta} + Z_i\left(\beta - \bar{\beta}\right) + \eta_i$$

- $\left(\beta - \bar{\beta}\right) = \varepsilon$ and $Z_i\left(\beta - \bar{\beta}\right) + \eta_i$ is a composite heteroskedastic error term.
- β random $=$ taste heterogeneity,
- η_i can interpret as unobserved attributes of goods
- Main advantage of MNP over MNL is that it allows for general error covariance structure.
- Note: To make computation easier, users sometimes set $\Sigma_\beta = 0$ (fixed coefficient version)
  - allowing for β random
  - permits random taste variation
  - allows for possibility that different persons value 2 characteristics differently

How do we solve the forecasting problem?

- Suppose that we have 2 goods and add a 3rd

$$\begin{aligned}\Pr\left(1 \text{ chosen}\right) &= \Pr\left(u^1 - u^2 \geq 0\right) \\ &= \Pr 1\left(\left(Z^1 - Z^2\right)\bar{\beta} \geq \omega^2 - \omega^1\right)\end{aligned}$$

- Define η_1, η_2, η_3 as random choice specific shocks independent of Z^1, Z^2 and Z^3.
- $(\beta - \bar{\beta})$ arises from variability in slope coefficients.

- Define:

$$\omega^1 = Z^1\left(\beta - \bar{\beta}\right) + \eta^1, \quad \omega^2 = Z^2\left(\beta - \bar{\beta}\right) + \eta^2$$

$$= \int_{-\infty}^{\frac{\left(z^1 - z^2\right)\bar{\beta}}{\left[\sigma_{11}+\sigma_{22}-2\sigma_{12}+(z^2-z^1)\Sigma_\beta(z^2-z^1)'\right]^{1/2}}} \frac{1}{\sqrt{2\pi}} e^{-t/2} dt$$

- Pr(1 chosen).
- Now add a 3rd good

$$u^3 = Z^3\bar{\beta} + Z^3\left(\beta - \bar{\beta}\right) + \eta^3.$$

- **Problem**: η^3 comes out of the blue: We don't know correlation of η^3 with other errors.
- Suppose that $\eta^3 = 0$ (*i.e.* only preference slope heterogeneity). Then

$$\begin{aligned}
\Pr(1 \text{ chosen}) &= \int_{-\infty}^{a}\int_{-\infty}^{b} B.V.N.\ dt_1 dt_2 \\
\text{when } a &= \frac{\left(Z^1 - Z^2\right)\bar{\beta}}{\left[\sigma_{11} + \sigma_{22} - 2\sigma_{12} + \left(Z^2 - Z^1\right)\Sigma_\beta\left(Z^2 - Z^1\right)'\right]^{1/2}} \\
\text{and } b &= \frac{\left(Z^1 - Z^3\right)\bar{\beta}}{\left[\sigma_{11} + \left(Z^3 - Z^1\right)\Sigma_\beta\left(Z^3 - Z^1\right)'\right]^{1/2}}
\end{aligned}$$

- We could also solve the forecasting problem if we make an assumption like $\eta^2 = \eta^3$.
- We solve red-bus//blue-bus problem if $\eta^2 = \eta^1 = 0$ and $z^3 = z^2$.

$$\Pr\left(1 \text{ chosen}\right) = \Pr\left(u^1 - u^2 \geq 0, u^1 - u^3 \geq 0\right)$$

- but $u^1 - u^2 \geq 0 \wedge u^1 - u^3 \geq 0$ are the same event.
- $\therefore$adding a third choice does not change the choice of 1.

Estimation Methods for MNP Models

- Models tend to be difficult to estimate because of high dimensional integrals.
- Integrals need to be evaluated at each stage of estimating the likelihood.
- Simulation provides a means of estimating $P_{ij} = \Pr(i \text{chooses} j)$

Computation and Estimation

Link to Appendix

THE UNIVERSITY OF CHICAGO

Appendix

THE UNIVERSITY OF CHICAGO

Airum Models

Notes on McFadden Chapter/Integrating Discrete Continuous (see Heckman, 1974b, 1978, change notation)

- Notation:
  - I: enumeration of discrete alternatives
  - x: divisible goods
  - w: attributes of discrete choices
  - r: price of x
  - q_i: price of good i
  - y: income
  - y: $rx + q_i$
  - $\widetilde{u}$: $\widetilde{x} \times \omega \times I \rightarrow [0, 1]$ utility
- Define indirect utility function

$$v(y - q, r, w_i, i, \widetilde{u}) = \max_x \widetilde{u}\,(x, w_i, i \mid rx \leq y - q_i)$$

- Maximize out over continuous goods so we are left with discrete goods

Assumptions

- We assume v has usual properties of an indirect utility function
- Continuous, twice differentiable, homogeneous degree 0 in $(y, q-r)$, quasiconcave in r, $\frac{dv}{d(y-q)} > 0$)
- Then get

$$x(y-q, r, w_i, i; \widetilde{u}) = \frac{\frac{-\partial v}{\partial r}}{\frac{\partial v}{\partial y}}. \text{ (Roy's Identity)}$$

Assumptions

- For discrete alternative, we also get something like Roy's Identity

$$\delta_j = D(j \mid B, s; \widetilde{u}) = \frac{\frac{-\partial v^*}{\partial q_j}}{\frac{\partial v^*}{\partial y}}$$

where

$$\begin{aligned} v^*(y - q_B, r, w_R, B; \widetilde{u}) &= \max_{i \in B} v(y - q_i, r, w_i, i; \widetilde{u}) \\ \delta_j &= \begin{cases} 1 \\ 0 \end{cases} \text{ if } j \in B \; v_j \geq v_k \; \forall k \end{aligned}$$

- If IU assumptions satisfied, can write relationship between the probability of choosing ji and the utility function as

$$P(j \mid B, s) = E_{u|s} D(j \mid B, s; \widetilde{u})$$

- We seek sufficient conditions on preferences u such that we can integrate out over characteristics and come up with probabilities
- McFadden Shows that v takes AIRUM form

$$v(y - q, r, w_i, i; \widetilde{u}) = \frac{y - q - \alpha(r, w_i, i, \widetilde{u})}{\beta(r)}$$

where $y > q + \alpha$ $\quad$ α, β homogeneous of degree one wrt r

- Then

$$\begin{aligned}\overline{v} &= E_{u|s} \max_{i \in B} v(y - q_i, r, w_i, i; \widetilde{u}) \\ &= \frac{1}{\beta(r)} \left[y + \max_{i \in B} \left(E_{u|s} \left(-q_i - \alpha(r, w, i; \widetilde{u}) \right) \right) \right]\end{aligned}$$

and

$$P(j) = E_{u|s} D(j \mid B, s) = \frac{-\frac{\partial \overline{v}}{\partial q_j}}{\frac{\partial \overline{v}}{\partial y}}$$

- $\overline{v}$ is a utility function yielding the PCS
- Demand distribution can be analyzed as if it were generated by a population with common tastes, with each representative consumer having fractional consumption rates for the discrete alternative.

- Let

$$\widetilde{G}(q_B, r, w_B, B, s) = E_{u|s} \max_{i \in B} [-q_i - \alpha(r, w_B, i; \widetilde{u})] \qquad (*)$$

"Social surplus function"

- Then

$$P(j \mid B, s) = \frac{-\partial \widetilde{G}(q_B, r, w_B, B, s)}{\partial q_j} \qquad (**)$$

under SS conditions given in Mcfadden's chapter

- I.e., choice probabilities given by the gradient of the SS function.

Return to main text

THE UNIVERSITY OF CHICAGO

Problem of Identification and Normalization in the MNP Model

- Reference: David Bunch (1979), "Estimability In the Multinominal Probit Model" in *Transportation Research*
- Domencich and McFadden
- Let

$$Z\bar{\beta} = \begin{pmatrix} Z_1 \cdot \bar{\beta} \\ \vdots \\ Z_J \cdot \bar{\beta} \end{pmatrix} \quad \tilde{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_J \end{pmatrix} \quad \begin{array}{l} J \text{ alternatives} \\ K \text{ characteristics} \\ \beta \text{ random} \quad \beta \sim N(\beta, \Sigma_\beta) \end{array} \tag{2}$$

Problem of Identification and Normalization in the MNP Model

- Pr (alternative j selected):

$$= \Pr\left(u_j > u_i\right) \quad \forall i \neq j$$

$$= \int_{u_j=-\infty}^{\infty} \int_{u_i=-\infty}^{u_j} \int_{u_J=-\infty}^{u_j} \Phi\left(u \mid V_\mu, \Sigma_\mu\right) du_J du_I du_j$$

where $\Phi\left(u \mid V_\mu, \Sigma_\mu\right)$ is *pdf*

(Φ is J-dimensional MVN density with mean V_μ, Σ_μ)

- Note: Unlike the MVL, no closed form expression for the integral.
- The integrals often evaluated using simulation methods (we will work an example).

How many parameters are there?

- $\bar{\beta}$: K parameters
- Σ_β: $K \times K$ symmetric matrix $\frac{K^2-K}{2} + K = \frac{K(K+1)}{2}$
- Σ_η: $\frac{J(J+1)}{2}$
- Note: When a person chooses j, all we know is relative utility, not absolute utility.
- This suggests that not all parameters in the model will be identified.
- Requires normalizations.

Digression on Identification

- What does it mean to say a parameter is not identified in a model?
- Model with one parameterization is observationally equivalent to another model with a different parameterization

- Example: Binary Probit Model (fixed β)

$$\begin{aligned}
\Pr(D=1 \mid Z) &= \Pr(v_1 + \varepsilon_1 > v_2 + \varepsilon_2) \\
&= \Pr(x\beta + \varepsilon_1 > x_2\beta + \varepsilon_2) \\
&= \Pr((x_1 - x_2)\beta > \varepsilon_2 - \varepsilon_1) \\
&= \Pr\left(\frac{(x_1 - x_2)\beta}{\sigma} > \frac{\varepsilon_2 - \varepsilon_1}{\sigma}\right) \\
&= \Phi\left(\frac{\tilde{x}\beta}{\sigma}\right) \quad \bar{x} = x_1 - x_2
\end{aligned}$$

- $\Phi\left(\frac{\tilde{x}\beta}{\sigma}\right)$ is observationally equivalent to $\Phi\left(\frac{\tilde{x}\beta^*}{\sigma^*}\right)$ for $\frac{\beta}{\sigma} = \frac{\beta^*}{\sigma^*}$.

- β not separably identified relative to σ but ratio is identified:

$$\begin{aligned} \Phi\left(\frac{\tilde{x}\beta}{\sigma}\right) &= \Phi\left(\frac{\tilde{x}\beta^*}{\sigma^*}\right) \\ \Phi^{-1} \cdot \Phi\left(\frac{\tilde{x}\beta}{\sigma}\right) &= \Phi^{-1}\Phi\left(\frac{\tilde{x}\beta^*}{\sigma^*}\right) \\ \Rightarrow \quad \frac{\beta}{\sigma} &= \frac{\beta^*}{\sigma^*} \end{aligned}$$

- Set $\{b : b = \beta \cdot \delta,\ \delta$ any positive scalar$\}$ is identified (say "β is identified up to scale and sign is identified").

Identification in the MVP model

$$\Pr\left(j \text{ selected} \mid V_\mu, \Sigma_\mu\right) = \Pr\left(u_i - u_j < 0 \quad \forall i \neq j\right)$$

$$\text{Define } \Delta_j = \begin{pmatrix} 1 & 0 & .. & -1 & .. & 0 \\ 0 & 1 & .. & -1 & .. & 0 \\ \vdots & & & \vdots & & \vdots \\ 0 & .. & .. & -1 & 0 & 1 \end{pmatrix}_{(J-1)\times J} \quad \text{(contrast matrix)}$$

$$\Delta_j \tilde{u} = \begin{pmatrix} u' - u^j \\ \vdots \\ u^J - u^j \end{pmatrix}$$

Identification in the MVP model

$$\begin{aligned}\Pr\left(j \text{ selected} \mid V_\mu, \Sigma_\mu\right) &= \Pr\left(\Delta_j \tilde{u} < 0 \mid V_\mu, \Sigma_\mu\right) \\ &= \Phi\left(0 \mid V_Z, \Sigma_Z\right)\end{aligned}$$

- Where
  1. V_Z is the mean of $\Delta_j \tilde{u} = \Delta_j \tilde{Z} \bar{\beta}$
  2. Σ_Z is the variance of $\Delta_j \tilde{Z} \Sigma_\beta \tilde{Z}' \Delta_j' + \Delta_j \Sigma_\eta \Delta_j'$
  3. V_Z is $(J-1) \times 1$
  4. Σ_Z: $(J-1) \times (J-1)$
- We reduce dimensions of the integral by one.

- This says that all of the information exists in the contrasts.
- Can't identify all the components because we only observe the contrasts.
- Now define $\tilde{\Delta}_j$ as Δ_j with Jth column removed and choose J as the reference alternative with corresponding Δ_J.
- Then can verify that

$$\Delta_j = \tilde{\Delta}_j \cdot \Delta_J$$

- For example, with three goods:

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

- $\tilde{\Delta}_j$, ($j = 2$, Δ_J, ($J = 3$, Δ_j, ($j = 2$, 3rd column included)

3rd column reference alt.)

removed)

- Therefore, we can write

$$\begin{aligned} V_Z &= \Delta_j \tilde{Z} \bar{\beta} \\ \Sigma_Z &= \Delta_j \tilde{Z} \Sigma_\beta \tilde{Z}' \Delta_j' + \tilde{\Delta}_j \Delta_J \Sigma_\eta \Delta_J' \tilde{\Delta}_j' \end{aligned}$$

- where $C_J = \Delta_J \Sigma_\eta \Delta_J'$ and $(J-1) \times (J-1)$ has $\frac{(J-1)^2-(J-1)}{2} + (J+1)$ parameters $= \frac{J(J-1)}{2}$ total.
- Since original model can always be expressed in terms of a model with $(\beta, \Sigma_\beta, C_J)$, it follows that some of the parameters in the original model are not identified.

How many parameters not identified?

- Original model:

$$K + \frac{K(K+1)}{2} + \frac{J(J+1)}{2}$$

- Now:

$$K + \frac{K(K+1)}{2} + \frac{J(J-1)}{2}, \qquad \frac{J^2 + J - (J^2 - J)}{2}$$

$$= J \text{ not identified}$$

- Turns out that one additional parameter not identified.
- Total: $J + 1$
- **Note**: Evaluation of $\Phi(0 \mid kv_Z, k^2\Sigma_Z)$ $k > 0$ gives same result as evaluating $\Phi(0 \mid v_Z, \Sigma_Z)$ can eliminate one more parameter by suitable choice of k.

Illustration

$$J = 3 \qquad \Sigma_\eta = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

$$C_2 = \Delta_2 \Sigma_\eta \Delta_2' = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \cdot \Sigma_\eta \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}'$$

$$= \begin{pmatrix} \sigma_{11} & -2\sigma_{21} & +\sigma_{22}, & \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{22} \\ \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{22}, & \sigma_{33} & -2\sigma_{31} & +\sigma_{22} \end{pmatrix}$$

Illustration

$$C_2 = \tilde{\Delta}_2 \Delta_3 \Sigma_\eta \Delta_3' \Delta_2' = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \cdot$$
$$\begin{pmatrix} \sigma_{11} & -2\sigma_{21} & +\sigma_{33}, & \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{33} \\ \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{33} & \sigma_{22} & -2\sigma_{32} & \sigma_{33} \end{pmatrix} \cdot$$
$$\begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$$

Normalization Approach of Albreit, Lerman, and Manski (1978)

- **Note:** Need $J+1$ restrictions on VCV matrix.
- Fix J parameters by setting last row and last column of Σ_η to 0
- Fix scale by constraining diagonal elements of Σ_η so that trace $\frac{\Sigma_\varepsilon}{J}$ equals variance of a standard Weibull. (To compare estimates with MNL and independent probit)

Variety of Simulation Methods

- Simulated method of moments
- Method of simulated scores
- Simulated maximum likelihood

References:

- Lerman and Manski (1981), *Structural Analysis* (online at McFadden's website)
- McFadden (1989), *Econometrica*
- Ruud (1982), *Journal of Econometrics*
- Hajivassilon and McFadden (1990)
- Hajivassilon and Ruud (Ch. 20), *Handbook of Econometrics*
- Stern (1992), *Econometrica*
- Stern (1997), Survey in JEL
- Bayesian MCMC (Chib *et al.* on reading list)

Model

$$
\begin{aligned}
u_j &= Z_j\beta + \eta_j \quad \text{with } \beta \text{ fixed, } \eta_j \sim N(0,\Omega),\ J \text{ choices} \\
P_{ij} &= \text{prob } i \text{ chooses } j \\
Y_{ij} &= 1 \text{ if } i \text{ chooses } j,\ 0 \text{ else}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{N}\prod_{j=1}^{J}(P_{ij})^{Y_{ij}} \\
\log \mathcal{L} &= \sum_{i=1}^{N}\sum_{j=1}^{J} Y_{ij}\log P_{ij}
\end{aligned}
$$

Simulation Algorithm

i. For given β, Ω generate Monte Carlo draws (R of them)

$$u_j^r, \ j = 1...J, \ r = 1...R$$

ii. Let $\tilde{P}_k = \frac{1}{R} \sum_{r=1}^{R} 1(u_k^r = \max\{u_1^r, ..., u_J^r\})$ where $\tilde{P}_k$ is a "frequency simulator" of $\Pr(k \text{ chosen}; \beta, \Omega)$

iii. Maximize $\sum_{i=1}^{N} \log \tilde{P}_{ik}$ over alternative values for β, Ω

- **Note:** Lerman and Manski found that this procedure performs poorly and requires a large number of draws, particularly when P is close to 0 or 1.

$$var\left(\frac{1}{R}\sum 1(\cdot)\right) = \frac{1}{R^2}\sum_{i=1}^{R} var1(\cdot) \text{ with } var1(\cdot) \text{ at true values}$$

- McFadden (1989) provided some key insights into how to improve the simulation method. He showed that simulation is viable even for a small number of draws provided that:
  a. an unbiased simulator is used
  b. functions to be simulated appear linearly in the conditions defining the estimator
  c. same set of random draws is used to simulate the model at different parameter values
- **Note**: Condition (b) is violated for crude frequency method which had $\log \tilde{P}_{ik}$

Simulated Method of Moments (McFadden, Econometrica, 1989)

$$u_{ij} = Z_{ij}\beta = Z_{ij}\bar{\beta} + Z_{ij}\varepsilon_i \quad \beta = \bar{\beta} + \varepsilon_i$$

(earlier model with only preference heterogeneity)

$$P_{ij}(\gamma) = P(i \text{ chooses } j \mid w_i, \gamma) \quad (w_i \text{ are regressors})$$

- Define $Y_{ij} = 1$ if i chooses j, $Y_{ij} = 0$ otherwise.

$$\log \mathcal{L} = \frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \ln P_{ij}(\gamma) \quad N_0 = NJ$$

$$\frac{\partial \log \mathcal{L}}{\partial \gamma} = \frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \left[\frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}(\gamma)} \right] = 0 \tag{3}$$

Simulated Method of Moments (McFadden, Econometrica, 1989)

$$\sqrt{N_0}(\hat{\gamma}_{MLE} - \gamma_0) \sim N\left(0, I_f^{-1}\right)$$

$$\hat{I}_f = \frac{1}{N_0} \sum_{i=1}^{N} \left[\sum_{j=1}^{J} Y_{ij} \left[\frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}(\gamma)} \right] \right] \left[\sum_{j=1}^{J} Y_{ij} \left[\frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}(\gamma)} \right] \right]' .$$

(outer product of score vector)

- Now use the fact that $\sum_{j=1}^{J} P_{ij}(\gamma) = 1$

$$\Rightarrow \sum_{j=1}^{J} \frac{\partial P_{ij}}{\partial \gamma} = 0 \quad \Rightarrow \sum_{j=1}^{J} \frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}} P_{ij} = 0$$

- Rewrite 3 as

$$\frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} (Y_{ij} - P_{ij}) \frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}} = 0$$

- **Note**: $E(Y_{ij}) = P_{ij}$.

- Letting $\varepsilon_{ij} = Y_{ij} - P_{ij}$, and $Z_{ij} = \frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}}$, we have

$$\frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} \varepsilon_{ij} Z_{ij} = \frac{\frac{\partial P_{ij}}{\partial \gamma}}{P_{ij}}$$

- like a moment condition using Z_{ij} as the instrument but so far P_{ij} still a $J-1$ dimensional intergral.

Simulation Algorithm

- **Model**
  $u_{ij} = Z_{ij}\bar{\beta} + Z_{ij} \cdot \varepsilon_i$ J choices, K characteristics
  $u_{ij} : 1 \times 1$ $Z_{ij} : 1 \times K$ $\bar{\beta} : K \times 1$
  $Z_{ij} : 1 \times K$ $\varepsilon_i : K \times 1$
- Rewrite as
  $\tilde{u}_i = \tilde{Z}_i\bar{\beta} + \tilde{Z}_i\Gamma\tilde{e}_i$ where $\Gamma\Gamma' = \Sigma_\varepsilon$ (Cholesky decomposition),
  $\tilde{e}_i \sim N(0, I_k), \varepsilon_i = \Gamma\tilde{e}_i$
  $\tilde{u}_i : J \times 1$ $\tilde{Z}_i : J \times K$ $\bar{\beta} : K \times 1$
  $\tilde{Z}_i : J \times K$ $\Gamma : K \times X$ $\tilde{e}_i : K \times 1$

THE UNIVERSITY OF CHICAGO

- **Step (i)**. Generate $\tilde{e}_i$ for each i such that $\tilde{e}_i$ are iid across persons and distributed $N(0, I_k)$. In total, generate

$$N\,(\text{sample size}) \cdot K\,(\text{vector length}) \cdot R$$

(number of Monte Carlo draws)

- **Step (ii)**. Fix matrix Γ and obtain $\eta_{ij} = Z_{ij}\Gamma\tilde{e}_i$, where $Z_{ij} : 1 \times K$; $\Gamma : K \times K$; $\tilde{e}_i$ is $K \times 1$.
- Form vector $\begin{pmatrix} Z_{i1}\Gamma\tilde{e}_i \\ Z_{i2}\Gamma\tilde{e}_i \\ \vdots \\ Z_{iJ}\Gamma\tilde{e}_i \end{pmatrix}$ for each person.

- **Step (iii)**. Fix $\bar{\beta}$ and generate

$$\tilde{u}_{ij} = Z_{ij}\bar{\beta} + \eta_{ij} \quad \forall i.$$

- **Step (iv)**. Find relative frequency that ith person chooses alternative j across Monte Carlo draws

$$\tilde{P}_{ij}(\gamma) = \frac{1}{R}\sum_{r=1}^{R} 1(\tilde{u}_{ij} > \tilde{u}_{im}; \quad \forall m \neq j)$$

- where $\tilde{P}_{ij}(\gamma)$ is the "simulator" for $P_{ij}(\gamma)$. Stack to get $\tilde{P}_i(\gamma)$.

- **Step (v)**. To get $\tilde{P}_i(\gamma)$ for different values of γ, repeat steps (ii) through (iv) using the same r.v.'s $\tilde{e}_i$ generated in step (i).
- **Step (vi)**. Define

$$w_{ij} = \frac{\frac{\partial P_{ij}(\gamma)}{\partial \gamma}}{P_{ij}}$$

- and w_{ij} as corresponding stacked vector simulator for w_{ij} can be obtained by a numerical derivative

$$\frac{\partial P_{ij}(\gamma)}{\partial \gamma} = \frac{P_{ij}(\gamma + h_{lm}) - P_{ij}(\gamma - h_{lm})}{2h}$$

where $m = 1...J$, lm = vector with 1 in mth place.

Solve Moment Condition

- Apply Gauss-Newton Method, iterate to convergence

$$\gamma_1 = \gamma_0 + \{\frac{1}{N}\sum w_i(\gamma)\{y_i - \tilde{P}_i(\gamma_0)\}\{y_i - \tilde{P}_i(\gamma_0)\}^{-1}$$
$$\cdot \frac{1}{N}\sum_{i=1}^{N} w_i(\gamma_0)\{y_i - \tilde{P}_i(\gamma_0)\}$$

Digression on Gauss-Newton

- Suppose problem is

$$S = \min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left[y_i - f_i(\beta) \right]^2 \quad \text{(nonlinear least squares).}$$

$$f_i(\beta) = f_i\left(\hat{\beta}_1\right) + \left.\frac{\partial f_i}{\partial \beta}\right|_{\beta_1} \left(\beta - \hat{\beta}_1\right) + \ldots$$

by Taylor expansion around initial guess $\hat{\beta}_1$

$$= f_i\left(\hat{\beta}_1\right) + \left.\frac{\partial f}{\partial \beta}\right|_{\hat{\beta}_1} \left(\beta - \hat{\beta}_1\right) + \ldots$$

(terms ignored)

- Substitution gives

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} [y_i - f_i\left(\hat{\beta}_1\right) - \left.\frac{\partial f}{\partial \beta}\right|_{\hat{\beta}_1} \left(\beta - \hat{\beta}_1\right)]^2$$

- Solve for $\hat{\beta}_2$ to get

$$\hat{\beta}_2 = \beta_1 + \left[\sum_{i=1}^{N} \left.\frac{\partial f_i}{\partial \beta}\right|_{\hat{\beta}_1} \left.\frac{\partial f_i}{\partial \beta}\right|'_{\hat{\beta}_1}\right]^{-1} \left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}_1}$$

- Repeat until convergence (problem if matrix is singular).

Disadvantages of Simulation Methods

1. $\tilde{P}_{ij}$ is not smooth due to indicator function (causes difficulties in deriving asymptotic distribution; need to use methods developed by Pakes and Pollard (1989) for nondifferentiable functions). Smoothed SMOM methods developed by Stern, Hajivassiliou, and Ruud.
2. $\tilde{P}_{ij}$ cannot be 0 (causes problems in denominator when close to 0)
3. Simulating small P_{ij} may require large number of draws

Disadvantages of Simulation Methods

- Refinement: "Smoothed Simulated Method of Moments" replaces indicator with a smooth function (Stern (1992), *Econometrica*).

$$\int \text{ instead of}$$

$$\tilde{P}_{ij}(\gamma) = \frac{1}{R}\sum_{r=1}^{R} R\left(\Phi\left(\tilde{u}_{ij} - \tilde{u}_{im}\right)\right)$$

How does simulation affect the asymptotic distribution?

- Without simulation get

$$\sqrt{\eta}\left(\hat{\gamma}_{mme}-\gamma_0\right) \sim N\left(0, \left[\operatorname*{p\,lim}_{N\to\infty} \frac{1}{N}\sum w_i\left(y_i-P_i\right)' w_i'\right]^{-1}\right)$$

- with simulation, the variance is slightly hgher due to simulation error

$$\sqrt{n}\left(\hat{\gamma}_{msm}-\gamma_0\right) \sim N\left(0, \operatorname*{plim}_{N\to\infty} C^{-1}\{1+\frac{1}{\eta}\}\right)$$

How does simulation affect the asymptotic distribution?

- where

$$C = -\frac{1}{N}\sum_{i=1}^{N} w_i \left(y_i - P_i\right)\left(y_i - P_i\right)' w_{i'}$$

- as $N \to \infty$, $\quad \sqrt{\eta}\left(\gamma_{msm} - \gamma\right) \sim N\left(0, C^{-1}\right).$
- **Note:** Method does not require that number of draws go to infinity.

Choice-Based Sampling (See Heckman in New Palgrave)

- **Reference**:
  - Chs. 1-2 of Manski and McFadden volume
  - Manski and Lerman (1978 *Econometrica*)
  - Amemiya
- **Examples:**
  1. Suppose we gather data on transportation mode choice at the
     - train station
     - subway station
     - car checkpoints (toll booths etc.)

- We observe characteristics on populations conditioned on the choice that they made (this type of sampling commonly arises)
  2. Evaluating effects of a social program; have data on particpants and non-participants; usually participants oversampled relative to frequency in the population.
- Distinguish between exogenous stratification and endogenous stratification, the latter of which is choice-based. (But a special type of endogenous stratification)
- Oversampling in high population areas (as is commonly done to reduce sampling costs or to increase representation of some groups) could be exogenous stratification (depending on phenomenon being studied).

Notation:

- Let $P_i = P(i \mid Z)$ in a random sample P_i^* in a choice-based sample (CBS)
- Under CBS, sampling is assumed to be random within i partitions of the data

$$P(Z \mid i) = P^*(Z \mid i) \text{ but } P(Z) \neq P^*(Z)$$

- Suppose that we want to recover $P(i \mid Z)$ from choice-based data
- We observe

$P^*(i \mid Z)$ (assume Z are discrete conditioning cells)

$P^*(Z)$

$P^*(i)$

By Bayes' Rule

$$P(A,B) = \frac{P(A,B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P^*(i \mid Z) = \frac{P^*(Z \mid i) \cdot P^*(i)}{P^*(Z)}$$

$$P(i \mid Z) = \frac{P(Z \mid i) \cdot P(i)}{P(Z)}$$

By Bayes' Rule

$$
\begin{aligned}
P(i \mid Z) &= \frac{\left[\frac{P^*(i|Z)\cdot P^*(Z)}{P^*(i)}\right] P(i)}{P(Z)} \\
P(Z) &= \sum_j P(Z \mid j) P(i) \\
P(Z \mid j) &= P^*(Z \mid j) \\
P^*(Z \mid j) &= \frac{P^*(j \mid Z) \cdot P^*(Z)}{P^*(j)}
\end{aligned}
$$

$$
P(i \mid Z) = \frac{\frac{P^*(i|Z)P^*(Z)}{P^*(i)} P(i)}{\sum_j \frac{P^*(j|Z)P^*(Z)}{P^*(j)} P(j)} = \frac{P^*(i \mid Z) \frac{P(i)}{P^*(i)}}{\sum_j P^*(j \mid Z) \frac{P(j)}{P^*(j)}}
$$

- To recover $P(i \mid Z)$ from choice-based sampled data, you need to know $P(j)$, $P^*(j)$ $\forall j$. $P^*(j)$ can be estimated from sample, but $P(j)$ requires outside information. Need weights $\frac{P(i)}{P^*(i)}$.
- **Note:** Problem set asks you to consider how CBS biases the coefficients and intercept in a logit model. (Can show that bias only in the constant)

Application and Extension: Berry, Levinsohn, and Pakes (1995)

- Develop equilibrium model and estimation techniques for analyzing demand and supply in differentiated product markets
- Use to study automobile industry
- Goal is to estimate parameters of both the demand and cost functions incorporating own and cross price elasticities and elasticities with respect to product attributes (car horse power, MPG, air conditioning, size,...) using only aggregate product level data supplemented with data on consumer characteristic distributions (income distribution from CBS)
- Want to allow for flexible substitution patterns

Key assumptions

i. joint distribution of observed and unobserved product and consumer characteristics

ii. price taking for consumers, Nash eq assumptions on producers in oligopolistic, differentiated products market.

Notation

- ζ : individual characteristics

x (observed)
ξ (unobserved) : product characteristics
p (price)

$$u_{ij} = u\left(\zeta_i, p_j, x_j, \theta\right) : \text{utility if person } i \text{ chooses } j$$

(Cobb-Douglas assumption here)

$$j = 0, 1, ..., J$$

$0 =$ not purchasing any

Notation

- Define

$$A_j = \{\zeta : u_{ij}(\zeta, p_j, x_j, \xi_j ; \theta) \geq u(\zeta, p_r, x_r, \xi_r ; 0) \quad r = 0, ..., J\},$$

- the set of ζ that induces choice of good j. This is defined over individual characteristics which may be observed or unobserved.

Market Share

$$s_j(p, x, \xi; \theta) = \int_{\zeta \epsilon A_j} f(\zeta)\, d\zeta; \ (s \text{ is vector of market shares})$$

- Special functional form:

$$u_{ij} = u(\zeta_i, p_j, x_j, \xi_j; \theta) = x_j\beta - \alpha p_j + \xi_j + \epsilon_{ij} = \delta_j + \epsilon_{ij}$$

Market Share

- $\delta_j = x_j\beta - \alpha p_j + \xi_j =$ mean utility from good j
- $\bar{\xi}_j$ is mean across consumers of unobserved component of utility
- ϵ_{ij} are the only elements representing consumer characteristics
- **Special Case:**

$$\begin{aligned} \xi_j &= 0 \quad \text{(no unobserved characteristic)} \\ & \epsilon_{ij} \text{ iid over } i,\ j,\ \text{independent of } x_j \end{aligned}$$

- Then share

$$s_j = \int_{-\infty}^{\infty} \Pi_{q \neq j} F\left(\delta_j - \delta_q + \epsilon\right) f\left(\epsilon\right) d\epsilon$$

- Unidimensional integral; has closed form solution under extreme value.

Why is assumption that utility is additively separable and iid in consumer and product characteristics highly restrictive?

a. Implies that all substitution effects depend only on the δs (since there is a unique vector of market shares associated with each δ vector). Therefore, conditional on market shares, substitution patterns don't depend on characteristics of the product. Example: if Mercedes and Yugo have some market share then they must have the same δs and some cross derivative with respect to any 3rd car (BMW).

$$\frac{\partial s_i}{\partial p_k} = \int \Pi_{q \neq k} F\left(\delta_k - \delta_q + \epsilon\right) F'\left(\delta_k - \delta_q + \epsilon\right) \frac{\partial \delta_k}{\partial p_k} f\left(\epsilon\right) d\epsilon$$

(same if δs same)

Why is assumption that utility is additively separable and iid in consumer and product characteristics highly restrictive?

- Two products with same market share have same own price derivatives (not good because you expect product markup to depend on more than market share)
- Also assumes that individuals value product characteristics in same way (no preference heterogeneity)

Alternative Model (Random Coefficients Versions)

$$
\begin{aligned}
u_{ij} &= x_j\bar{\beta} - \alpha p_j + \xi_j + \sum_k \sigma_k x_{jk}\nu_{ik} + \epsilon_{ij} \\
\beta_k &= \bar{\beta}_k + \sigma_k \nu_k \\
E\left(\nu_{ik}\right) &= 1
\end{aligned}
$$

- Could still assume ϵ_{ij} has iid extreme value.

Model Actually Used

- Impose alternative functional form assumption because they want to incorporate prior info on distribution of relevant consumer characteristics and on interactions between consumer and product characteristics.

$$u_{ij} = (y - p_j)^{\alpha} \, G\left(x_i, \xi_j, \nu_i\right) e^{\epsilon(i,j)}$$

- Assume G is log linear

$$\tilde{u}_{ij} = \log u_{ij} = \alpha \log\left(y_i - p_j\right) + x_j \bar{\beta} + \xi_j + \sum_k \sigma_k x_{jk} \nu_{ik} + \epsilon_{ij}$$

Model Actually Used

- No good utility:

$$= \alpha \log y_i + \xi_0 + \sigma_0 \nu_{io} + \epsilon_{io}$$

- **Note:** Prices are likely to be correlated with unobserved product attributes, ξ, which leads to an endogeneity problem. (ξ may represent things like style, prestige, reputation, etc.)

$$\text{quantity demanded, } q_j = M_{s_j}(x, \xi, p; \theta) \text{ (share)}$$

- ξ enters nonlinearly, so we need to use some transformation to be able to apply instrumental variables (Principle of Replacement Function).

Cost Side

- Multiproduct firms $1, \ldots, F$. Each produces subset τ_F of J possible products. Cost of producing good assumed to be independent of output levels and log linear in vector of cost characteristics (W_j, ω_j).

$$\ln mc_j = W_j\gamma + \omega_j \Rightarrow \Pi_f = \sum_{j \in \tau_F} (p_j - mc_j)\, ms_j\,(p, x, \xi; \theta)$$

Nash Assumption

- Firm chooses prices that maximize profit taking as given attributes of its products and the prices and attributes of its competitor's products. P_j satisfies

$$s_j\,(p, x, \xi; \theta) + \sum_{r \epsilon \tau_F} (p_r - mc_r)\, \frac{\partial s_r\,(p, x, \xi; \theta)}{\partial p_j} = 0$$

Cost Side

- Define

$$\Delta_{jr} = \left\{ \begin{array}{c} \frac{-\partial s_r}{\partial p_j} \\ 0 \end{array} \right\} \text{ if } r \text{ and } j \text{ produced by same firm}$$

$$\begin{aligned} \Rightarrow \quad & s\left(p, x, \xi; \theta\right) - \Delta\left(p, x, \xi; \theta\right)\left[p - mc\right] = 0 \\ \Rightarrow \quad & p = mc + \Delta\left(p, x, \xi; \theta\right)^{-1} s\left(p, x, \xi; \theta\right) \ \text{(market)} \end{aligned}$$

- Market term depends on parameters of demand system and on equilibrium price vector

$$p = mc + \Delta\left(p, x, \xi; \theta\right)^{-1} s\left(p, x, \xi; \theta\right)$$

- Mark-up depends only on the parameters of the demand system and equilibrium price vector.
- Since p is a function of w, $b\left(p, x, \xi; \theta\right)$ also a function of ω (unobs cost determinants)
- Let

$$\begin{aligned} \ln mc_j &= W_j\gamma + \omega_j \\ &\Rightarrow \quad p \exp\left\{W\gamma + \omega\right\} + b\left(p, x, \xi; \theta\right) \\ \ln\left(p - b\left(p, x, \xi, \theta\right)\right) &= W\gamma + \omega \quad \text{(pricing equation)} \end{aligned}$$

Estimation

- Need instruments for both demand and pricing equations. *i.e.* need variables correlated with (x, ω) uncorrelated with ξ and ω. Let

$$Z = (X, W) \quad (p, q \text{ not included in } Z)$$

- Assume

$$\begin{aligned} E(\xi_j \mid Z_j) &= E(\omega_j \mid Z) = 0 \\ E((\xi_j, \omega_j)(\xi_j, \omega_j) \mid Z) &= \Omega(Z_j) \quad (\text{finite almost every } Z_j) \end{aligned}$$

- Note that demand for any product is a function of characteristics of all products, so don't have any exclusion restrictions.

Data

- J vectors $(x_j, \omega_j, p_j, q_j)$
- n : number of households sampled
- s^n : vector of sampled market shares
- Assume that a true θ_0 population abides by models.
- Decision Rules
- s^n converges to s^0 (multinomial)
- $\sqrt{n}(s^n - s^0) = O_p(1)$

- Assume we could calculate

$$\left\{\varepsilon_j\left(\theta, s, p_0\right), \omega_j\left(\theta, s, p_0\right)\right\}_{j=1}^{J}$$

  for alternative values of θ
- They show that any choice of
  a. observed vector of market shares, s
  b. distribution of consumer characteristics, P
  c. parameter of model
- implies a unique sequence of estimates for the two unobserved characteristics

$$\xi_j(\theta, s, P), \omega_j\left(\theta, s, P\right)$$

- Then any function of Z must be uncorrelated with the vector

$$\left\{\xi\left(\theta, s^0, P_0\right), \omega\left(\theta, s^0, P_0\right)\right\}_{j=1}^{J} \text{ when } \theta = \theta_0$$

- Can use GMM
- Note: Conditional moment restriction implies infinite number of unconditional restrictions

$$\min \frac{1}{J} \sum H_j(Z) \begin{pmatrix} \xi_j\left(\theta, s^0, P_0\right) \\ w_j\left(\theta, s^0, P_0\right) \end{pmatrix}$$

Compuation in Logit Case

- Logit ϵ_{ij} is Weibull.

$$
\begin{aligned}
\delta_j &= x_j\beta - \alpha p_j + \varepsilon_j \\
u_{ij} &= x_j\beta - \alpha p_j + \varepsilon_j + \epsilon_{ij} \\
s_j(p, x, \varepsilon) &= \frac{e^{\delta_j}}{1 + \sum_{j=1}^{J} e^{\delta_j}} \\
\delta_j &= \ln s_j - \ln s_0, \quad j = 1, \ldots, J \\
\varepsilon_j &= \ln s_j - \ln s_0 - x_j\beta + \alpha p_j
\end{aligned}
$$

- See paper for more details.

Generalized Method of Moments (GMM)

References:

- Hansen (1982), *Econometrica*
- Hansen and Singleton (1982,1988)
- Also known as "minimum distance estimators"
- Suppose that we have some data $\{x_t\}$ $t = 1...T$ and we want to test hypotheses about $E(x_t = \mu)$.
- How do we proceed? By a CLT

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (x_t - \mu) \sim N(O, V_0)$$

$$\begin{aligned} V_0 &= E\left((x_t - \mu)(x_t - \mu)'\right) \quad \text{if } x_t \text{ iid} \\ &= \lim_{t \to \infty} E\left(\frac{1}{\sqrt{T}} v \sum_t (x_t - \mu) \frac{1}{\sqrt{T}} (x_t - \mu)'\right) \quad \text{(general case)} \end{aligned}$$

- We can decompose $V_0 = QDQ'$ where
  $QQ' = I, \quad Q^{-1} = Q^1, \quad D =$ matrix of eigenvalues

$$\begin{aligned} V_0 &= QD^{1/2}D^{1/2}Q' \\ Q'V_0Q &= D^{1/2}D^{1/2} \\ D^{-1/2}Q'V_0QD^{-1/2} &= I \end{aligned}$$

- under rule H_0,

$$\left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right]' V_0^{-1} \left[\frac{1}{\sqrt{T}}\sum (x_t - u_0)\right]$$

$$= \left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right]' QD^{-1/2}D^{-1/2}Q'$$

$$\left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right] \sim X^2(n)$$

- where n is the number of moment conditions.
- How does test statistic behave under alternative? $(\mu \neq \mu_0)$
- should get large

- Write as

$$\left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right]' V_0^{-1} \left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right] \tag{4}$$

$$= \left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right]' V_0^{-1} \left[\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right]$$

$$+ \frac{2}{\sqrt{T}}\sum_t (\mu - \mu_0)' V_0^{-1} \frac{1}{\sqrt{T}}\sum_t (x_t - \mu) + \ldots$$

$$+ \frac{1}{\sqrt{T}}\sum_t (\mu - \mu_0)' V_0^{-1} \frac{1}{\sqrt{T}}\sum_t (\mu - \mu_0) \tag{5}$$

- last term * is $0(T)$.

$\lambda = T(\mu - \mu_0)' V_0^{-1} (\mu - \mu_0)$ is the noncentrality parameter

- Problems:
  - V_o is not known *a priori*.
- Estimate $V_T \longrightarrow V_0$
- In the setting, use sample covariance matrix.
- In general setting, approximate limit by finite T
  - μ not known
- Suppose we want to test $\mu = \varphi(\beta)$
  φ specified
  β unknown

- Can estimate by *min-x^2 estimation*.

$$\min_{\beta \in B} \left[\frac{1}{\sqrt{T}} \sum \left(x_t - \varphi \left(\beta \right) \right) \right]' V_T^{-1} \left[\frac{1}{\sqrt{T}} \sum \left(x_t - \varphi \left(\beta \right) \right) \right]$$

$$\sim x^2 \left(n - k \right)$$

k =dimension of β

n =number of moments

- Note: Searching over k dimensions, you lose one degree of freedom (will show next).

- Find distribution theory for $\hat{\beta}$:
  Q: This is a M-estimator, so how do you proceed?
- FOC

$$\sqrt{T}\left.\frac{\partial \varphi}{\partial \beta}\right|_{\hat{\beta}_T}^{\prime} V_T^{-1} \frac{1}{\sqrt{T}}\left[x_t-\varphi\left(\hat{\beta}_T\right)\right]=0$$

- Taylor expand $\varphi\left(\hat{\beta}_T\right)$ around $\varphi\left(\hat{\beta}_0\right)$

$$\varphi\left(\hat{\beta}_T\right)=\varphi\left(\hat{\beta}_0\right)+\left.\frac{\partial \varphi}{\partial \beta}\right|_{\hat{\beta}^*}\left(\hat{\beta}_T-\beta_0\right) \qquad \beta^* \text{ between } \beta_0, \hat{\beta}_T$$

- get

$$\sqrt{T}\left.\frac{\partial \varphi}{\partial \beta}\right|_{\hat{\beta}_T}^{\prime} V_T^{-1} \frac{1}{\sqrt{T}} \sum_t\left[x_t-\varphi\left(\beta_0\right)-\left.\frac{\partial \varphi}{\partial \beta}\right|_{\hat{\beta}^*}\left(\hat{\beta}_T-\beta_0\right)\right]=0$$

- Rearrange to solve for $\left(\hat{\beta}_T - \beta_0\right)$

$$+\left[\sqrt{T}\left.\frac{\partial\varphi}{\partial\beta}\right|'_{\hat{\beta}_T} V_T^{-1}\left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}^*}\right]\left(\hat{\beta}_T - \beta_0\right)\frac{T}{\sqrt{T}}$$

$$= \sqrt{T}\left.\frac{\partial\varphi}{\partial\beta}\right|'_{\hat{\beta}_T} V_T^{-1}\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right)$$

if $\left.\frac{\partial\varphi}{\partial\beta}\right|'_{\hat{\beta}_T} \rightarrow D_0$ (Convergence of random function)

$V_T \rightarrow V_0$

$\left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}^*} \rightarrow D_0$

- Apply CLT to $\frac{1}{\sqrt{T}} \sum_t x_t - \varphi\left(\hat{\beta}_T\right) \sim N\left(0, V_0\right)$
- Then $\sqrt{T}\left(\hat{\beta}_T - \beta_0\right) \sim N\left(0, \left(D_0' V_0^{-1} D_0\right)^{-1} \left(D_0' V_0^{-1} D_0\right)^{-1'}\right)$

- Why is the limiting distribution $\chi^2(n-k)$?
- Write

$$\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right) = \frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0) + \frac{1}{\sqrt{T}}\sum_t \left(\mu_0 - \varphi\left(\hat{\beta}_T\right)\right)$$

- Recall, we had

$$\varphi\left(\hat{\beta}_T\right) = \varphi(\beta_0) + \left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}^*} \tag{6}$$

$$\left\{\left(\left.\frac{\partial\varphi}{\partial\beta}\right|'_{\hat{\beta}_T} V_T^{-1} \left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}^*}\right)^{-1} \cdot \left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}_T} \cdot V_T^{-1} \frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)\right\}$$

definition of $\sqrt{T}\left(\hat{\beta}_T - \beta_0\right)$ derived easier. (7)

- Note that the second term (6) is a linear combination of the first.

$$\Rightarrow \frac{1}{\sqrt{T}} \sum_t x_t - \varphi\left(\hat{\beta}_T\right) = B \frac{1}{\sqrt{T}} \sum_t (x_t - \mu_0)$$

- where $B = I - \left.\frac{\partial \varphi}{\partial \beta}\right|_{\beta^0} \left[\left.\frac{\partial \varphi}{\partial \beta}\right|'_{\beta_0} V_0^{-1} \left.\frac{\partial \varphi}{\partial \beta}\right|_{\beta_0} \right]^{-1} \left.\frac{\partial \varphi}{\partial \beta}\right|'_{\beta^0} V_0^{-1} = B_0$

$$\text{Note that } \left.\frac{\partial \varphi}{\partial \beta}\right|'_{\beta_0} V_0^{-1} B_0 = 0$$

- This tells us that certain linear combinations of B will give a degerate distribution (along k dimensions)
- This needs to be taken into account when testing.
- Recall that we had $V_0 = QDQ'$
  $QQ' = I \qquad V_0^{-1} = QD^{-1/2}D^{-1/2}Q'$

$$D^{-1/2}Q'\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right) = D^{-1/2}Q'B\frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)$$

- where

$$D^{-1/2}Q'B = D^{-1/2}Q' - D^{-1/2}Q' \cdot \left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}_T} \left[\left.\frac{\partial\varphi}{\partial\beta}\right|'_{\hat{\beta}_T} QD^{-1/2}D^{-1/2}Q' \left.\frac{\partial\varphi}{\partial\beta}\right|_{\hat{\beta}_T}\right]^{-1} \left.\frac{\partial\varphi}{\partial\beta}\right|'_{\beta_0}$$

$$QD^{-1/2}D^{-1/2}Q'$$

$$= \left(I - A\left(A'A\right)^{-1}\right)D^{-1/2}Q' \text{ (idempotent matrix } M_A\text{)}$$

- Thus

$$D^{-1/2}Q'\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right) = M_A \cdot D^{-1/2} \cdot Q' \cdot \frac{1}{\sqrt{T}}\sum_t (x_t - \mu_0)$$

- This matrix M_A accounts for the fact that we performed the minimization over β

How is distribution theory affected?

- Have a quadratic form in normal r.v's with idempotent matrix

$$\text{ex. } \hat{\varepsilon} \cdot \hat{\varepsilon} = \varepsilon' M_x \varepsilon \qquad M_x = I - x\left(x'x\right)^{-1} x'$$

- Me facts
  - (i) Theorem
    - Let $Y \sim N\left(\theta, \sigma^2 In\right)$ and let P be a symmetric matrix of rank γ
    - Then $Q = \frac{(Y-B)'P(Y-B)}{\sigma^2} \sim x_r^2$ iff $p^2 = p$ (i.e. P idempotent)
    - See Seber, p.37
  - (ii) if $Q_i \sim X^2_{r_i} \qquad i = 1, 2 \qquad r_1 > r_2$ and $Q = Q_1 - Q_2$ is independent of Q_2, then $Q \sim X^2_{r_1 - r_2}$
- Apply these results to

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} = \frac{\varepsilon' M_x \varepsilon}{\sigma^2} \sim x^2\left(\text{rank } M_x\right)$$

- Rank of an idempotent matrix is equal to its trace and

$$tr\left(A\right) = \sum_{i=1}^{n} \lambda_i \qquad \lambda_i \text{ eigenvalues} \tag{8}$$

  for idempotent matrix, eigenvalues are all 0 or 1.
- (note rank = #non-zero eigenvalues for idempotent eigenvalues all 0 or 1)

$$rank\left(I - x\left(x'x\right)^{-1}x'\right) = rank\left(I\right) - rank\left(x\left(x'x\right)^{-1}x'\right)$$
$$\text{where } rank\left(I\right) = n$$

$$rank\left(x\left(x'x\right)^{-1}x'\right) = trace\left(x'x\left(x'x\right)^{-1}\right) \text{ since} tr\left(AB\right) = tr\left(BA\right)$$
$$= trace\left(I_k\right)$$

$$rank\left(I - x\left(x'x\right)^{-1}x'\right) = nk$$

- Thus, by same reasoning limiting distribution of

$$\left[\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right)\right] V_T^{-1} \left[\frac{1}{\sqrt{T}}\sum_t x_t - \varphi\left(\hat{\beta}_T\right)\right]$$
$$\sim x^2\left(n-k\right) = rank\left(A\right)$$

- We preserve x^2 but lose degrees of freedom in estimating β.
- In case where $n = k$ (just identified case)
- we can estimate β but have no degrees of freedom left to perform the test.
- *Would GMM provide a method fo estimating β if we used a weighting matrix other than V_T^{-1}?*
- Why not replace $V_0^{-1}$by w_0?

$$\min_{\beta \in B} \frac{1}{\sqrt{T}} \sum_t \left(x_t - \varphi\left(\hat{\beta}_T\right) \right)' w_0 \frac{1}{\sqrt{T}} \sum_t x_t - \varphi(\beta)$$

- Could choose $w_0 = I$ (avoid need to estimate weighting matrix)
- Result: Asymptotic covariance is altered and asymptotic distribution of criterion is different, but $\frac{1}{\sqrt{T}} \sum_t \left(x_t - \varphi\left(\hat{\beta}_T\right) \right)$ will still be normal.

- *What is the advantage of focusing on minimum x^2 estimation?*
- Choosing $w_0 = V_0^{-1}$ gives smallest covariance matrix. Get most efficient estimator for β and most powerful test of restrictions.
- *Show this:* Show $(D_0' w_0 D_0)^{-1} \left(D_0' w_0 V_0^{-1} w_0 D_0\right) \left(D_0' w_0 D_0^{-1}\right)^{-1} - (D_0' V_0 D_0)^{-1}$ is P.S.D
- where $(D_0' w_0 D_0)^{-1} \left(D_0' w_0 V_0^{-1} w_0 D_0\right) \left(D_0' w_0 D_0^{-1}\right)^{-1}$ is the covariance matrix for $\sqrt{T}\left(\hat{\beta}_T - \beta_0\right)$ when general weighting matrix is used.
- Equivalent to showing

$$\left(D_0 V_0^{-1} D_0\right) - \left(D_0' w_0 D_0\right) \left(D_0' w_0 V_0^{-1} w_0 D_0\right)^{-1} \left(D_0' w_0 D_0\right)$$

is P.S.D

Show that it can be written as a quadratic form

- Take any vector α

$$\alpha' \left[D' V_0^{-1} D_0 - (D_0' W_0 D_0) (D_0' W_0 V_0 W_0 D_0)^{-1} (D_0' W_0 D_0) \right] \alpha$$

$$= \alpha' D_0' V_0^{-1/2} \left[I - V_0'^{1/2} W_0 D_0 \left(D_0' W_0 V_0^{1/2} \cdot Y_0^{1/2} W_0 D_0 \right)^{-1} D_0' W_0 V_0^{1/2} \right] Y_0^{-1/2} D_0 \alpha$$

$$= \alpha' D_0' V_0^{-1/2} \left[I - \tilde{V} \left(\tilde{V}' \tilde{V} \right)^{-1} \tilde{V}' \right] V^{-1/2} D_0 \alpha \geq 0 \quad (= 0 \text{ if } W_0 = V_0^{-1})$$

- Therefore $W_0 = V_0^{-1}$ is the optimal choice for the weighting matrix.

Many standard estimators can be interpreted as GMM estimators

- Some examples:
  1. OLS

$$y_t = x_t\beta + u_t \quad E(u_t x_t) = 0 \quad \Rightarrow E((y_t - x_t\beta)x_t) = 0$$

$$\min_{\beta \in B} \left(\frac{1}{\sqrt{T}} \sum_t (y_t - x_t\beta) x_t \right)' V_T^{-1} \left(\frac{1}{\sqrt{T}} \sum_t (y_t - x_t\beta) x_t \right)$$

- where y_t is the estimator for $E(x_t u_t u_t' x_t')$ for idd case $= \sigma^2 E(x_t x_t')$ if homoskedastic.

(2) Instrumental Variables

$$y_t = x_t\beta + u_t$$

$$\begin{aligned} E\left(u_t x_t\right) &\neq 0 \\ E\left(u_t z_t\right) &= 0 \\ E\left(x_t z_t\right) &\neq 0 \end{aligned}$$

$$\hat{\beta}_T = \arg\min_{\beta \in B} \left(\frac{1}{\sqrt{T}} \sum \left(y_t - x_t\beta\right) Z_t \right)' V_T^{-1} \frac{1}{\sqrt{T}} \sum \left(y_t - x_t\beta\right)$$

where $V_T = \hat{E}\left(z_t u_t u_t' z_t'\right)$ in idd case

- Estimator for

$$\lim_{T\to\infty} E\left(\frac{1}{\sqrt{T}}\sum_t z_t u_t \left(\frac{1}{\sqrt{T}}\sum z_s u_s\right)'\right) \quad \text{(time series case)}$$

- Suppose

$$E\left(u_t u_t' \mid z_t\right) = \sigma^2 I$$

- then

$$W_0 = \sigma^2 E\left(z_t z_t'\right)$$

- Can verify that 2SLS and GMM give same estimator

$$\hat{\beta}_{2SLS} = \left(x'z\left(z'z\right)^{-1}z'x\right)^{-1}\left(x'z\left(z'z\right)^{-1}z'y\right)$$

- Note: In first stage regress x on Z

$$\begin{aligned}\hat{x} &= z\left(z'z\right)^{-1}z'x \\ \hat{y} &= z\left(z'z\right)^{-1}z'y\end{aligned}$$

$$var\left(\hat{\beta}_{2SLS}\right) = \left[E\left(x_i z_i\right)E\left(z_i z_i'\right)^{-1}E\left(z_i x_i\right)\right]^{-1}E\left(x_i z_i\right)E\left(z_i z_i'\right)^{-1}E\left(z_i u_i u_i' z_i'\right)$$
$$\cdot E\left(z_i z_i'\right)^{-1}E\left(x_i z_i\right)'\left[E\left(x_i z_i\right)E\left(z_i z_i'\right)^{-1}E\left(z_i x_i\right)\right]^{-1}$$

- Under GMM

$$
\begin{aligned}
var\left(\hat{\beta}_{\text{GMM}}\right) &= \left(D_0 V_0^{-1} D_0\right)^{-1} \quad (\text{when } W_0 = V_0^{-1}) \\
D_0 &= \frac{\partial \varphi}{\partial \beta} \mid_{\beta_0} = \text{p}\lim_{\frac{1}{n}} \sum x_i z_i^{'} = E\left(x_i z_i'\right) \\
W_0 &= \left(\sigma^2\right)^{-1} E\left(z_i z_i'\right)^{-1}
\end{aligned}
$$

In the presense of heterskedastocity, weighting matrix would be different (and 2SLS and GMM not the same)

$$W_0 = E\left(z'uu'z\right)^{-1} = E\left(z'E\left(uu' \mid z\right)z\right)^{-1} = E\left(z'vz\right)^{-1}$$

- with panel data could have

$$E\left(uu' \mid z\right) = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ & v_2 & \cdots & \\ & & \ddots & 0 \\ 0 & & & v_T \end{pmatrix} = y$$

- allow for correlation over time for given individual, but iid across individuals.

Nonlinear least squares

$$y_t = \varphi(x_t, \beta) + u_t \quad E(u_t \varphi(x_t; \beta)) = 0$$

$$\min_{\beta \in B} \left[\frac{1}{\sqrt{T}} \sum (y_t - \varphi(x_t; \beta)) \varphi(x_t; \beta) \right]$$
$$V_T^{-1} \left[\frac{1}{\sqrt{T}} \sum (y_t - \varphi(x_t; \beta)) \varphi(x_t; \beta) \right]$$

- General Method of Moments

$$\min_{\beta \in B} \sum_t f_t(\beta)' V_0^{-1} \left[\sum_t f_t(\beta) \right]$$

Nonlinear least squares

- where

$$\frac{1}{\sqrt{T}} \sum f_t(\beta_0) \quad \rightarrow \quad N(O, V_0)$$

$$\frac{1}{\sqrt{T}} \sum f_t \quad \rightarrow \quad Ef_t$$

- In general, f_t is a random function.

- Suppose we want to estimate β 5×1 and we have 6 potential instruments. Can we test validity of the instruments? What if we have 5 instruments? If we assume $E(\varepsilon_i \mid x_i) = 0$ instead of $E(\varepsilon_i x_i) = 0$ (i. e. conditional instead of unconditional) then have infinite number of moment conditions.

$$E(\varepsilon_i f(x_i)) = E(E(\varepsilon_i \mid x_i) f(x_i)) = 0 \quad \text{any } f(x_i)$$

- How to optimally choose which moment conditions to use is current area of research. How might you use GMM to check if a variable is normally distributed?

Return to main text

THE UNIVERSITY OF CHICAGO