Conditional Choice Probability Estimators

Robert A. Miller

Tilburg University

November 2023

Unrestricted Estimation

Data and outside knowledge

- Recall the social surplus function is:

$$V_t(x_t) \equiv E\left\{\sum_{\tau=t}^{T}\sum_{j=1}^{J}\beta^{\tau-t-1}d_{j\tau}^{o}\left(x_\tau,\epsilon_\tau\right)\left(u_{j\tau}(x_\tau)+\epsilon_{j\tau}\right)\right\}$$

  where $d_{j\tau}^{o}\left(x_\tau,\epsilon_\tau\right)$ is the optimal choice.

- Suppose the data comes from a panel and assume we know:
  1. the discount factor β
  2. the distribution of disturbances $G_t\left(\epsilon\left|x\right.\right)$
  3. $u_{1t}\left(x\right)$ (or more generally one of the payoffs for each state and time).
  4. $u_{1t}\left(x\right)=0$ (for notational convenience).

Unrestricted Estimation

The likelihood

- Consider a panel:
  - of independently drawn individuals $n \in \{1, \dots, N\}$
  - each with history $t \in \{1, \dots, T\}$
  - and data on their state variables, x_{nt}
  - and decisions $d_{nt} = (d_{n1t}, \dots, d_{nJt})$.
- The joint probability distribution of the decisions and outcomes is:

$$\prod_{n=1}^{N} \prod_{t=1}^{T} \left(\sum_{j=1}^{J} \sum_{x'=1}^{X} d_{njt} \mathbf{1} \left\{ x_{n,t+1} = x' \right\} p_{jt}(x) f_{jt}(x'|x) \right)$$

- Taking logs yields:

$$\sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} d_{njt} \left\{ \log \left[p_{jt}(x_{nt}) \right] + \sum_{x=1}^{X} \mathbf{1} \left\{ x_{n,t+1} = x \right\} \log \left[f_{jt}(x|x_{nt}) \right] \right\}$$

Unrestricted Estimation
The reduced form

- Note the choice probabilities are additively separable from the transition probabilities in the formula for the joint distribution of decisions and outcomes.
- Hence the estimation of the joint likelihood splits:
  - One piece deals with the choice probabilities conditional on the state.
  - The other deals with the transition conditional on the choice and state.
- Maximizing each piece separately with respect to $f_j(x'|x)$ and $p_t(x_{nt})$ gives the unrestricted estimators:

$$\widehat{f}_{jt}\left(x' \left| x\right.\right) = \frac{\sum_{n=1}^{N} \mathbf{1}\left\{x_{nt} = x, d_{njt} = 1, x_{n,t+1} = x'\right\}}{\sum_{n=1}^{N} \mathbf{1}\left\{x_{nt} = x, d_{njt} = 1\right\}}$$

and:

$$\widehat{p}_{jt}\left(x\right) = \frac{\sum_{n=1}^{N} \mathbf{1}\left\{x_{nt} = x, d_{njt} = 1\right\}}{\sum_{n=1}^{N} \mathbf{1}\left\{x_{nt} = x\right\}} \tag{1}$$

- Let $\kappa_\tau(x_{\tau+1}|t, x_t, j)$ denote the probability of reaching $x_{\tau+1}$ at $\tau+1$ from x_t by following action j at t and then always choosing the first action:

$$\kappa_\tau(x_{\tau+1}|t, x_t, j) \equiv \begin{cases} f_{jt}(x_{t+1}|x_t) & \tau = t \\ \sum_{x=1}^{X} f_{1\tau}(x_{\tau+1}|x)\kappa_{\tau-1}(x|t, x_t, j) & \tau = t+1, \ldots \end{cases} \tag{2}$$

- Thus we can recursively estimate $\kappa_\tau(x_{\tau+1}|t, x_t, j)$ with:

$$\widehat{\kappa}_\tau(x_{\tau+1}|t, x_t, j) \equiv \begin{cases} \widehat{f}_{jt}(x_{t+1}|x_t) & \tau = t \\ \sum_{x=1}^{X} \widehat{f}_{1\tau}(x_{\tau+1}|x)\widehat{\kappa}_{\tau-1}(x|t, x_t, j) & \tau = t+1, \ldots \end{cases}$$

- Similarly we estimate $\psi_{jt}(x_t) \equiv V_t(x) - v_{jt}(x)$ with $\widehat{\psi}_{jt}(x_t)$ using the $\widehat{p}_{jt}(x)$ estimates of the CCPs.

Unrestricted Estimation

Utility parameter estimates

- From the previous lecture:

$$
\begin{aligned}
u_{jt}(x_t) &= \psi_{1t}(x_t) - \psi_{jt}(x_t) \\
&\quad + \sum_{\tau=1}^{T-t} \sum_{x=1}^{X} \beta^{\tau-t} \psi_{1,t+\tau}(x) \left[\kappa_{t1,\tau-1}(x|x_t) - \kappa_{tj,\tau-1}(x|x_t) \right]
\end{aligned}
$$

- Substituting $\widehat{\kappa}_{\tau-1}(x|x_t, j)$ for $\kappa_{\tau-1}(x|x_t, j)$ and $\psi_{jt}(x_t)$ with $\widehat{\psi}_{jt}(x_t)$ then yields:

$$
\begin{aligned}
\widehat{u}_{jt}(x_t) &\equiv \widehat{\psi}_{1t}(x_t) - \widehat{\psi}_{jt}(x_t) \\
&\quad + \sum_{\tau=1}^{T-t} \sum_{x=1}^{X} \beta^{\tau-t} \widehat{\psi}_{1,t+\tau}(x) \left[\widehat{\kappa}_{t1,\tau-1}(x|x_t) - \widehat{\kappa}_{tj,\tau-1}(x|x_t) \right]
\end{aligned}
$$

- The stationary case is similar (and has the matrix representation we discussed in previous lectures).

Asymptotic efficiency

- By the Law of Large Numbers $\widehat{f}_{jt}\left(x'\left|x\right.\right)$ converges to $f_{jt}\left(x'\left|x\right.\right)$ and $\widehat{p}_{jt}\left(x\right)$ converges to $p_{jt}(x)$, both almost surely.
- By the Central Limit Theorem both estimators converge at $\sqrt{N}$ and and have asymptotic normal distributions.
- Both $\widehat{f}_{jt}\left(x'\left|x\right.\right)$ and $\widehat{p}_{jt}\left(x\right)$ are ML estimators for $f_{jt}\left(x'\left|x\right.\right)$ and $p_{jt}(x)$ and obtain the Cramer-Rao lower bound asymptotically.
- Since and $u_{jt}(x)$ is exactly identified, it follows by the *invariance principle* that $\widehat{u}_{jt}(x)$ is consistent and asymptotically efficient for $u_{jt}(x_t)$, also attaining its Cramer Rao lower bound.
- Thus the unrestricted ML and CCP estimators are identical.
- Greater efficiency can only be obtained by making functional form assumptions about $u_{jt}(x_t)$ and $f_{jt}\left(x'\left|x\right.\right)$.

Restricting the Parameter Space

Parameterizing the primitives

- In practice applications further restrict the parameter space.
- For example assume $\theta \equiv \left(\theta^{(1)}, \theta^{(2)}\right) \in \Theta$ is a closed convex subspace of Euclidean space, and:
  - $u_{jt}(x) \equiv u_j(x, \theta^{(1)})$
  - $f_{jt}(x|x_{nt}) \equiv f_{jt}(x|x_{nt}, \theta^{(2)})$
- We now define the model by (T, β, θ, g).
- Assume the DGP comes from (T, β, θ_0, g) where:

$$\theta_0 \equiv \left(\theta_0^{(1)}, \theta_0^{(2)}\right) \in \Theta^{(interior)}$$

- For example many applications assume:
  - $u_{jt}(x) \equiv x'\theta_j^{(1)}$ is linear in x and does not depend on t
  - $f_{jt}(x|x_{nt})$ is degenerate, x following a deterministic law of motion that does not depend on t.

- A Quasi Maximum Likelihood (QML) estimator can be obtained by estimating:
    1. $\theta_0^{(2)}$ with $\theta_{LIML}^{(2)}$ from the data on $f_{jt}\left(x|x_t,\theta^{(2)}\right)$.
    2. $\kappa_\tau(x|t,x_t,k,\theta_0^{(2)})$ with $\widehat{\kappa}_\tau(x|t,x_t,k,\theta_{LIML}^{(2)})$ using $f_{jt}\left(x|x_t,\theta_{LIML}^{(2)}\right)$.
    3. $\psi_{1t}(x)$ with $\widehat{\psi}_{1t}(x)$ by substituting cell estimators $\widehat{p}_{jt}(x)$ for $p_{jt}(x)$.
    4. $v_{jt}\left(x,\theta^{(1)},\theta_0^{(2)}\right)$ with $\widehat{v}_{jt}\left(x,\theta^{(1)},\theta_{LIML}^{(2)}\right)$ for any given $\theta^{(1)}$, given below.
    5. $p_{jt}\left(x,\theta^{(1)},\theta_0^{(2)}\right)$ with $\widehat{p}_{jt}\left(x,\theta^{(1)},\theta_{LIML}^{(2)}\right)$ by substituting $\widehat{v}_{jt}\left(x,\theta^{(1)},\theta_{LIML}^{(2)}\right)$ for $v_{jt}\left(x,\theta^{(1)},\theta_0^{(2)}\right)$ in ML estimator.

- Working through each step:
  1. This step is quite common whenever $f_{jt}\left(x|x_{nt},\theta^{(2)}\right)$ must be estimated:

$$\theta^{(2)}_{LIML} \equiv \underset{\theta^{(2)}}{\arg\max} \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{x=1}^{X} d_{njt}\mathbf{1}\left\{x_{n,t+1}=x\right\}\ln\left[f_{jt}(x|x_{nt},\theta^{(2)})\right]$$

  2. Here (2) is replaced with:

$$\widehat{\kappa}_{\tau}(x_{\tau+1}|t,x_t,j,\theta^{(2)}_{LIML}) \equiv \begin{cases} f_{jt}(x_{t+1}|x_t,\theta^{(2)}_{LIML}) & \tau=t \\ \sum_{x=1}^{X} f_{1\tau}(x_{\tau+1}|x,\theta^{(2)}_{LIML})\widehat{\kappa}_{\tau-1}(x|t,x_t,j,\theta^{(2)}_{LIML}) & \tau=t+1,\ldots \end{cases}$$

  3. For example if ϵ_t is T1EV, then $\widehat{\psi}_{1t}(x)=0.57\ldots-\ln\left[p_{1t}(x)\right]$.

Quasi Maximum Likelihood Estimation

Elaborating the last steps in QML estimation

- With respect to the last two steps:
  4. Appealing to the Representation theorem:

$$\widehat{v}_{jt}\left(x,\theta^{(1)},\theta^{(2)}_{LIML}\right) = u_{jt}(x,\theta^{(1)}) + \widehat{h}_{jt}(x)$$

  where the numeric *dynamic correction factor* $\widehat{h}_{kt}(x)$ is defined:

$$\widehat{h}_{jt}(x) \equiv \sum_{\tau=t+1}^{T}\sum_{x_\tau=1}^{X}\beta^{\tau-t}\widehat{\psi}_{1\tau}(x_\tau)\widehat{\kappa}_{\tau-1}(x_\tau|t,x,j,\theta^{(2)}_{LIML})$$

  5. In T1EV applications:

$$\widehat{p}_{jt}(x,\theta^{(1)},\theta^{(2)}_{LIML}) = \frac{\exp\left[u_{jt}(x,\theta^{(1)}) + \widehat{h}_{jt}(x)\right]}{\sum_{k=1}^{J}\exp\left[u_{kt}(x,\theta^{(1)}) + \widehat{h}_{kt}(x)\right]}$$

Minimum Distance Estimators (Altug and Miller,1998)

Minimizing the difference between unrestricted and restricted current payoffs

- Another approach is to match up the parametrization of $u_{jt}(x_t)$, denoted by $u_{jt}(x_t, \theta^{(1)})$, to its representation as closely as possible:
  1. Form the vector function where $\Psi(p, f)$ by stacking:

$$\Psi_{jt}(x_t, p, f) \equiv \psi_{1t}(x_t) - \psi_{jt}(x_t) + \sum_{\tau=1}^{T-t} \sum_{x=1}^{X} \beta^{\tau} \psi_{1,t+\tau}(x) \left[\begin{array}{c} \kappa_{kt,\tau-1}(x|x_t) \\ -\kappa_{jt,\tau-1}(x|x_t) \end{array} \right]$$

  2. Estimate the reduced form $\widehat{p}$ and $\widehat{f}$.
  3. Minimize the quadratic form to obtain:

$$\theta_{MD}^{(1)} = \underset{\theta^{(1)} \in \Theta^{(1)}}{\arg\min} \left[u(x, \theta^{(1)}) - \Psi\left(\widehat{p}, \widehat{f}\right) \right]' \widetilde{W} \left[u(x, \theta^{(1)}) - \Psi\left(\widehat{p}, \widehat{f}\right) \right]$$

  where $\widetilde{W}$ is a square $(J-1)\,TX$ weighting matrix.
- Note $\theta_{MD}^{(1)}$ has a closed form if $u(x; \theta_0^{(1)})$ is linear in $\theta_0^{(1)}$.

- We could form a Methods of Simulated Moments (MSM) estimator from:
  1. Simulate a lifetime path from x_{nt_n} onwards for each j, using $\widehat{f}$ and $\widehat{p}$.
  2. Obtain estimates of $\widehat{E}\left[\epsilon_{jt}\middle| d^o_{jt}=1, x_t\right]$ from $\widehat{p}$.
  3. Stitch together a simulated lifetime utility outcome from the j^{th} choice at t_n onwards for n, to form $\widehat{v}_{jt}\left(x_{nt_n}, \theta^{(1)}, \widehat{f}, \widehat{p}\right)$.
  4. Form the $J-1$ dimensional vector $I_n\left(x_{nt_n}; \theta^{(1)}, \widehat{f}, \widehat{p}\right)$ from:

$$\begin{aligned} I_{nj}\left(x_{nt_n}; \theta^{(1)}, \widehat{f}, \widehat{p}\right) \equiv\; & \widehat{v}_{jt_n}\left(x_{nt_n}, \theta^{(1)}, \widehat{f}, \widehat{p}\right) - \widehat{v}_{Jt_n}\left(x_{nt_n}, \theta^{(1)}, \widehat{f}, \widehat{p}\right) \\ & + \widehat{\psi}_{jt}(x_{nt_n}) - \widehat{\psi}_{Jt}(x_{nt_n}) \end{aligned}$$

  5. Given a weighting matrix W_S and an instrument vector z_n minimize:

$$N^{-1}\left[\sum_{n=1}^N z_n I_n\left(x_{nt_n}; \theta^{(1)}, \widehat{f}, \widehat{p}\right)\right]' W_S \left[\sum_{n=1}^N z_n I_n\left(x_{nt_n}; \theta^{(1)}, \widehat{f}, \widehat{p}\right)\right]$$

- In the first step, given the state simulate a choice using $\widehat{p}$, and simulate the next state using $\widehat{f}$. In this way generate $\widehat{x}_{ns}$ and $\widehat{d}_{ns} \equiv \left(\widehat{d}_{n1s}, \ldots, \widehat{d}_{nJs}\right)$ for all $s \in \{t_n + 1, \ldots, T\}$.
- Generating this path does not exploit knowledge of G, only the CCPs.
- In the second step $\widehat{E}\left[\epsilon_{jt} \middle| d^o_{jt} = 1, x_t\right] \equiv$

$$p_{jt}^{-1}(x_t) \int_{\epsilon_t} \prod_{k=1}^{J} I\left\{\widehat{\psi}_{jt}(x_t) - \widehat{\psi}_{kt}(x_t) \leq \epsilon_{jt} - \epsilon_{kt}\right\} \epsilon_{jt} g(\epsilon_t)\, d\epsilon_t$$

- In Step 4 $\widehat{v}_{jt}\left(x_{nt_n}, \theta^{(1)}, \widehat{f}, \widehat{p}\right)$ is stitched together as:

$$u_{jt}(x_{nt_n}, \theta^{(1)}) + \sum_{s=t+1}^{T} \sum_{k=1}^{J} \beta^{t-1} 1\left\{\widehat{d}_{nks} = 1\right\} \left\{ \begin{array}{l} u_{ks}(\widehat{x}_{ns}, \theta^{(1)}) \\ +\widehat{E}\left[\epsilon_{js} \middle| \widehat{x}_{ns}, \widehat{d}_{njs} = 1\right] \end{array} \right\}$$

- The solution has a closed form if $u_{jt}(x, \theta^{(1)})$ is linear in $\theta^{(1)}$.

- Mr Zurcher maximizes the expected discounted sum of payoffs:

$$E\left\{\sum_{t=1}^{\infty}\beta^{t-1}\left[d_{t2}(\theta_1 x_t+\theta_2 s+\epsilon_{t2})+d_{t1}\epsilon_{t1}\right]\right\}$$

  where:

  - $d_{t1}=1$ and $x_{t+1}=1$ if Zurcher replaces the engine
  - $d_{t2}=1$ and bus mileage advances to $x_{t+1}=x_t+1$ if he keeps the engine
  - buses are also differentiated by a fixed characteristic $s\in\{0,1\}$.
  - the choice-specific shocks ϵ_{tj} are *iid* Type 1 extreme value (T1EV).

- Define the conditional value function for each choice as:

$$v_j(x,s)=\begin{cases}\beta V(1,s) & \text{if } j=1\\ \theta_1 x+\theta_2 s+\beta V(x+1,s) & \text{if } j=2\end{cases}$$

  where $V(x,s)$ denotes the social surplus function.

- We suppose the data comprises a cross section of N observations of buses $n \in \{1, \ldots, N\}$ reporting their:
  - fixed characteristics s_n,
  - engine miles x_n,
  - and maintenance decision (d_{n1}, d_{n2}).
- Let $p_1(x, s)$ denote the conditional choice probability (CCP) of replacing the engine given x and s.
- Stationarity and T1EV imply that for all t :

$$\begin{aligned} p_1(x, s) &\equiv \int_{\epsilon_t} d_1^o(x, s, \epsilon_t)\, g(\epsilon_t)\, d\epsilon_t \\ &= \int_{\epsilon_t} \mathbf{1}\{\epsilon_{t2} - \epsilon_{t1} \leq v_1(x, s) - v_2(x, s)\}\, g(\epsilon_t | x_t)\, d\epsilon_t \\ &= \{1 + \exp[v_2(x, s) - v_1(x, s)]\}^{-1} \end{aligned}$$

- An ML estimator could be formed off this equation following the steps described above.

Bus Engines

Exploiting the renewal property

- The previous lecture implies that if ϵ_{jt} is T1EV, then for all (x, s, j):

$$V(x, s) = v_j(x, s) - \beta \log [p_j(x, s)] + 0.57\ldots$$

- Therefore the conditional value function of not replacing is:

$$\begin{aligned} v_2(x, s) &= \theta_1 x + \theta_2 s + \beta V(x, s+1) \\ &= \theta_1 x + \theta_2 s + \beta \{ v_1(x+1, s) - p_1(x+1, s) + 0.57\ldots \} \end{aligned}$$

- Similarly:

$$v_1(x, s) = \beta V(1, s) = \beta \{ v_1(1, s) - \ln [p_1(1, s)] + 0.57 \} \ldots$$

- Because bus engine miles is the only factor affecting bus value given s:

$$v_1(x+1, s) = v_1(1, s)$$

Bus Engines

Using CCPs to represent differences in continuation values

- Hence:

$$v_2(x,s) - v_1(x,s) = \theta_1 x + \theta_2 s + \beta \ln\left[p_1(1,s)\right] - \beta \ln\left[p_1(x+1,s)\right]$$

- Therefore:

$$\begin{aligned} p_1(x,s) &= \frac{1}{1+\exp\left[v_2(x,s) - v_1(x,s)\right]} \\ &= \frac{1}{1+\exp\left\{\theta_1 x + \theta_2 s + \beta \ln\left[\frac{p_1(1,s)}{p_1(x+1,s)}\right]\right\}} \end{aligned}$$

- Intuitively the CCP for current replacement is the CCP for a static model with an offset term.
- The offset term accounts for differences in continuation values using future CCPs that characterize optimal future replacements.

- Consider the following CCP estimator:
  1. Form a first stage estimator for $p_1(x, s)$ from the relative frequencies:
  $$\hat{p}_1(x,s) \equiv \frac{\sum_{n=1}^{N} d_{n1} I(x_n = x) I(s_n = s)}{\sum_{n=1}^{N} I(x_n = x) I(s_n = s)}$$
  2. Substitute $\hat{p}_1(x, s)$ into the likelihood as incidental parameters to estimate $(\theta_1, \theta_2, \beta)$ with a logit:
  $$\frac{d_{n1} + d_{n2} \exp(\theta_1 x_n + \theta_2 s_n + \beta \ln \left[\frac{\hat{p}_1(1,s_n)}{\hat{p}_1(x_n+1,s_n)}\right]}{1 + \exp(\theta_1 x_n + \theta_2 s_n + \beta \ln \left[\frac{\hat{p}_1(1,s_n)}{\hat{p}_1(x_n+1,s_n)}\right]}$$
  3. Correct the standard errors for $(\theta_1, \theta_2, \beta)$ induced by the first stage estimates of $p_1(x, s)$.
- Note that in the second stage $\ln \left[\frac{\hat{p}_1(1,s_n)}{\hat{p}_1(x_n+1,s_n)}\right]$ enters the logit as an individual specific component of the data, the β coefficient entering in the same way as θ_1 and θ_2.

Monte Carlo Study (Arcidiacono and Miller, 2011)

Modifying the bus engine problem

- Suppose bus type $s \in \{0, 1\}$ is equally weighted.
- Two state variables affect wear and tear on the engine:
  1. total accumulated mileage:

$$x_{1,t+1} = \begin{cases} \Delta_t \text{ if } d_{1t} = 1 \\ x_{1t} + \Delta_t \text{ if } d_{2t} = 1 \end{cases}$$

  2. a permanent route characteristic for the bus, x_2, that systematically affects miles added each period.
- More specifically we assume:
  - $\Delta_t \in \{0, 0.125, \ldots, 24.875, 25\}$ is drawn from a discretized truncated exponential distribution, with:

$$f(\Delta_t | x_2) = \exp\left[-x_2(\Delta_t - 25)\right] - \exp\left[-x_2(\Delta_t - 24.875)\right]$$

  - x_2 is a multiple 0.01 drawn from a discrete equi-probability distribution between 0.25 and 1.25.

Monte Carlo Study

Including the age of the bus in panel estimation

- Let θ_{0t} denote other bus maintenance costs tied to its vintage.
- This modification renders the optimization problem nonstationary.
- The payoff difference from retaining versus replacing the engine is:

$$u_{t2}(x_{t1}, s) - u_{t1}(x_{t1}, s) \equiv \theta_{0t} + \theta_1 \min\{x_{t1}, 25\} + \theta_2 s$$

- Denoting $x_t \equiv (x_{1t}, x_2)$, this implies:

$$\begin{aligned} v_{t2}(x_t, s) - v_{t1}(x_t, s) &= \theta_{0t} + \theta_1 \min\{x_{t1}, 25\} + \theta_2 s \\ &\quad + \beta \sum_{\Delta_t \in \Lambda} \left\{ \ln \left[\frac{p_{1t}(\Delta_t, s)}{p_{1t}(x_{1t} + \Delta_t, s)} \right] \right\} f(\Delta_t | x_2) \end{aligned}$$

Monte Carlo Study

Extract from Table 1 of Arcidiacono and Miller (2011)

Monte Carlo for Optimal Stopping Problem+

	DGP	FIML	CCP	Time effects CCP
	(1)	(2)	(3)	(4)
θ_0 (Intercept)	2	2.0100	1.9911	
		(0.0405)	(0.0399)	
θ_1 (Mileage)	-0.15	-0.1488	-0.1441	-0.1440
		(0.0074)	(0.0098)	(0.0121)
θ_2 (Type)	1	0.9945	0.9726	0.9683
		(0.0611)	(0.0668)	(0.0636)
β (Discount Factor)	0.9	0.9102	0.9099	0.9172
		(0.0411)	(0.0554)	(0.0639)
Time (Minutes)		130.29	0.078	0.079
		(19.73)	(0.0041)	(0.0047)

+ Mean and standard deviations for fifty simulations. For columns (2) and (3), the observed data consist of 1000 buses for 20 periods. For column (4), the intercept (θ_0) is allowed to vary over time and the data consist of 2000 buses for 10 periods.