Estimation of Static Discrete Choice Models Using Market Level Data
NBER Methods Lectures

Aviv Nevo
Northwestern University and NBER

July 2012
Data Structures

- Market-level data
  - cross section/time series/panel of markets

- Consumer level data
  - cross section of consumers
  - sometimes: panel (i.e., repeated choices)
  - sometimes: second choice data

- Combination
  - sample of consumers plus market-level data
  - quantity/share by demographic groups
  - average demographics of purchasers of good $j$
Market-level Data

- We see product-level quantity/market shares by "market"
- Data include:
  - aggregate (market-level) quantity
  - prices/characteristics/advertising
  - definition of market size
  - distribution of demographics
    - sample of actual consumers
    - data to estimate a parametric distribution
- Advantages:
  - easier to get
  - sample selection less of an issue
- Disadvantages
  - estimation often harder and identification less clear
Consumer-level Data

- See match between consumers and their choices
- Data include:
  - consumer choices (including choice of outside good)
  - prices/characteristics/advertising of all options
  - consumer demographics
- Advantages:
  - impact of demographics
  - identification and estimation
  - dynamics (especially if we have panel)
- Disadvantages
  - harder/more costly to get
  - sample selection and reporting error
Review of the Model and Notation

Indirect utility function for the $J$ inside goods

$$U(x_{jt}, \bar{\zeta}_{jt}, l_i - p_{jt}, D_{it}, \nu_{it}; \theta),$$

where

- $x_{jt}$ – observed product characteristics
- $\bar{\zeta}_{jt}$ – unobserved (by us) product characteristic
- $D_{it}$ – "observed" consumer demographics (e.g., income)
- $\nu_{it}$ – unobserved consumer attributes

- $\bar{\zeta}_{jt}$ will play an important role
  - realistic
  - will act as a "residual" (why don’t predicted shares fit exactly – overfitting)
  - potentially implies endogeneity
Linear RC (Mixed) Logit Model

A common model is the linear Mixed Logit model

\[ u_{ijt} = x_{jt} \beta_i + \alpha_i p_{jt} + \xi_j + \epsilon_{ijt} \]

where

\[
\begin{pmatrix}
\alpha_i \\
\beta_i
\end{pmatrix} = \begin{pmatrix}
\alpha \\
\beta
\end{pmatrix} + \Pi D_i + \Sigma v_i
\]

It will be convenient to write

\[ u_{ijt} \equiv \delta(x_{jt}, p_{jt}, \xi_j; \theta_1) + \mu(x_{jt}, p_{jt}, D_i, v_i; \theta_2) + \epsilon_{ijt} \]

where \( \delta_{jt} = x_{jt} \beta + \alpha p_{jt} + \xi_j \), and \( \mu_{ijt} = (p_{jt} x_{jt}) (\Pi D_i + \Sigma v_i) \)
Linear RC (Mixed) Logit Model

A common model is the linear Mixed Logit model

\[ u_{ijt} = x_{jt} \beta_i + \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt} \]

and

\[ u_{ijt} = \delta(x_{jt}, p_{jt}, \xi_{jt}; \theta_1) + \mu(x_{jt}, p_{jt}, D_i, \nu_i; \theta_2) + \varepsilon_{ijt} \]

Note:
(1) the mean utility will play a key role in what follows
(2) the interplay between \( \mu_{ijt} \) and \( \varepsilon_{ijt} \)
(3) the "linear" and "non-linear" parameters
(4) definition of a market
Key Challenges for Estimation

- Recovering the non-linear parameters, which govern heterogeneity, without observing consumer level data
- The unobserved characteristic, $\xi_{jt}$
  - a main difference from early DC model (earlier models often had option specific constant in consumer level models)
  - generates a potential for correlation with price (or other x’s)
  - when constructing a counterfactual we will have to deal with what happens to $\xi_{jt}$
- Consumer-level vs Market-level data
  - with consumer data, the first issue is less of a problem
  - the "endogeneity" problem can exist with both consumer and market level data: a point often missed
What would we do if we had micro data?

- Estimate in two steps.
- First step, estimate $(\delta, \theta_2)$ say by MLE

$$
\Pr(y_{it} = j | D_{it}, x_t, p_t, \xi_t, \theta) = \Pr(y_{it} = j | D_{it}, \delta(x_t, p_t, \xi_t, \theta_1), x_t, p_t, \theta)
$$

e.g., assume $\varepsilon_{ijt}$ is iid double exponential (Logit), and $\Sigma = 0$

$$
= \frac{\exp\{\delta_{jt} + (p_{jt} x_{jt}) \Pi D_i\}}{\sum_{k=0}^{J} \exp\{\delta_{kt} + (p_{kt} x_{kt}) \Pi D_i\}}
$$

- Second step, recover $\theta_1$

$$
\hat{\delta}_{jt} = x_{jt} \beta + \alpha p_{jt} + \xi_{jt}
$$

$\xi_{jt}$ is the residual. If it is correlated with price (or $x$’s) need IVs (or an assumption about the panel structure)
Intuition from estimation with consumer data

- Estimation in 2 steps: first recover $\delta$ and $\theta_2$ (parameters of heterogeneity) and then recover $\theta_1$
- Different variation identifying the different parameters
  - $\theta_2$ is identified from variation in demographics holding the level (i.e., $\delta$) constant
    - If $\Sigma \neq 0$ then it is identified from within market share variation in choice probabilities
  - $\theta_1$ is identified from cross market variation (and appropriate exclusion restrictions)
- With market-level data will in some sense try to follow a similar logic
  - however, we do not have within market variation to identify $\theta_2$
  - will rely on cross market variation (in choice sets and demographics) for both steps
  - a key issue is that $\xi_{jt}$ is not held constant
In principle, we could consider estimating $\theta$ by min the distance between observed and predicted shares:

$$\min_{\theta} \| S_t - s_j(x_t, p_t, \theta) \|$$

• Issues:
  • computation (all parameters enter non-linearly)
  • more importantly,
    - prices might be correlated with the $\xi_{jt}$ (“structural” error)
    - standard IV methods do not work
Inversion

- Instead, follow estimation method proposed by Berry (1994) and BLP (1995)
- Key insight:
  - with $\xi_{jt}$ predicted shares can equal observed shares

\[ \sigma_j(\delta_t, x_t, p_t; \theta_2) = \int 1[u_{ijt} \geq u_{ikt} \quad \forall k \neq j] \ dF(\varepsilon_{it}, D_{it}, v_{it}) = S_{jt} \]

- under weak conditions this mapping can be inverted

\[ \delta_t = \sigma^{-1}(S_t, x_t, p_t; \theta_2) \]

- the mean utility is linear in $\xi_{jt}$; thus, we can form linear moment conditions
- estimate parameters via GMM
Important (and often missed) point

- IVs play dual role (recall 2 steps with consumer level data)
  - generate moment conditions to identify $\theta_2$
  - deal with the correlation of prices and error
- Even if prices are exogenous still need IVs
- This last point is often missed
- Why different than consumer-level data?
  - with aggregate data we only know the mean choice probability, i.e., the market share
  - with consumer level data we know more moments of the distribution of choice probabilities (holding $\xi_{jt}$ constant): these moments help identify the heterogeneity parameters
I will now go over the steps of the estimation
For now I assume that we have valid IVs
  later we will discuss where these come from
I will follow the original BLP algorithm
  I will discuss recently proposed alternatives later
  I will also discuss results on the performance of the algorithm
The BLP Estimation Algorithm

1. Compute predicted shares: given $\delta_t$ and $\theta_2$ compute
   $\sigma_j (\delta_t, x_t, p_t; \theta_2)$

2. Inversion: given $\theta_2$ search for $\delta_t$ that equates $\sigma_j (\delta_t, x_t, p_t; \theta_2)$
   and the observed market shares
   - the search for $\delta_t$ will call the function computed in (1)

3. Use the computed $\delta_t$ to compute $\tilde{\xi}_{jt}$ and form the GMM
   objective function (as a function of $\theta$)

4. Search for the value of $\theta$ that minimizes the objective function
Example: Estimation of the Logit Model

- Data: aggregate quantity, price characteristics. Market share
  \( s_{jt} = q_{jt} / M_t \)
  
  - Note: need for data on market size

- Computing market share

  \[
  s_{jt} = \frac{\exp\{\delta_{jt}\}}{\sum_{k=0}^{J} \exp\{\delta_{kt}\}}
  \]

- Inversion

  \[
  \ln(s_{jt}) - \ln(s_{0t}) = \delta_{jt} - \delta_{0t} = x_{jt} \beta + \alpha p_{jt} + \xi_{jt}
  \]

- Estimate using linear methods (e.g., 2SLS) with \( \ln(s_{jt}) - \ln(s_{0t}) \) as the "dependent variable".
Step 1: Compute the market shares predicted by the model

- Given $\delta_t$ and $\theta_2$ (and the data) compute

$$
\sigma_j(\delta_t, x_t, p_t; \theta_2) = \int 1 [u_{ijt} \geq u_{ikt} \quad \forall k \neq j] \ dF(\varepsilon_{it}, D_{it}, \nu_{it})
$$

- For some models this can be done analytically (e.g., Logit, Nested Logit and a few others)
- Generally the integral is computed numerically
- A common way to do this is via simulation

$$
\tilde{\sigma}_j(\delta_t, x_t, p_t, F_{ns}; \theta_2) = \frac{1}{ns} \sum_{i=1}^{ns} \frac{\exp\{\delta_{jt} + (p_{jt} \times x_{jt})(\prod D_i + \Sigma \nu_i)\}}{1 + \sum_{k=1}^{J} \exp\{\delta_{kt} + (p_{kt} \times x_{kt})(\prod D_i + \Sigma \nu_i)\}}
$$

where $\nu_i$ and $D_i, \ i = 1, \ldots, ns$ are draws from $F_v^*(\nu)$ and $F_D^*(D)$,

- Note:
  - the $\varepsilon$'s are integrated analytically
  - other simulators (importance sampling, Halton seq)
  - integral can be approximated in other ways (e.g., quadrature)
Step 2: Invert the shares to get mean utilities

- Given $\theta_2$, for each market compute mean utility, $\delta_t$, that equates the market shares computed in Step 1 to observed shares by solving

$$\tilde{\sigma}(\delta_t, x_t, p_t, F_{ns}; \theta_2) = S_t$$

- For some model (e.g., Logit and Nested Logit) this inversion can be computed analytically.
- Generally solved using a contraction mapping for each market

$$\delta_{t}^{h+1} = \delta_{t}^{h} + \ln(S_t) - \ln(\tilde{\sigma}(\delta_{t}^{h}, x_t, p_t, F_{ns}; \theta_2)) \quad h = 0, \ldots, H,$$

where $H$ is the smallest integer such that $\|\delta_{t}^{H} - \delta_{t}^{H-1}\| < \rho$

- $\delta_{t}^{H}$ is the approximation to $\delta_t$
- Choosing a high tolerance level, $\rho$, is crucial (at least $10^{-12}$)
Step 3: Compute the GMM objective

- Once the inversion has been computed the error term is defined as

\[
\xi_{jt}(\theta) = \tilde{\sigma}^{-1}(S_t, x_t, p_t; \theta_2) - x_{jt}\beta - \alpha p_{jt}
\]

- Note: \(\theta_1\) enters this term, and the GMM objective, in a linear fashion, while \(\theta_2\) enters non-linearly.

- This error is interacted with the IV to form

\[
\xi(\theta)'ZWZ'\xi(\theta)
\]

where \(W\) is the GMM weight matrix.
Step 4: Search for the parameters that maximize the objective

- In general, the search is non-linear
- It can be simplified in two ways.
  - “concentrate out” the linear parameters and limit search to $\theta_2$
  - use the Implicit Function Theorem to compute the analytic gradient and use it to aid the search
- Still highly non-linear so much care should be taken:
  - start search from different starting points
  - use different optimizers
Identification

- Ideal experiment: randomly vary prices, characteristics and availability of products, and see where consumers switch (i.e., shares of which products respond)
- In practice we will use IVs that try to mimic this ideal experiment
- Next lecture we will see examples
- Is there "enough" variation to identify substitution?
- Solutions:
  - supply information (BLP)
  - many markets (Nevo)
  - add micro information (Petrin, MicroBLP)
- For further discussion and proofs (in NP case) see Haile and Berry
The Limit Distribution for the Parameter Estimates

- Can be obtained in a similar way to any GMM estimator
- With one cross section of observations is

$$J^{-1}(\Gamma'\Gamma)^{-1}\Gamma'V_0\Gamma(\Gamma'\Gamma)^{-1}$$

- where
  - $\Gamma$ – derivative of the expectation of the moments wrt parameters
  - $V_0$ – variance-covariance of those moments evaluated

- $V_0$ has (at least) two orthogonal sources of randomness
  - randomness generated by random draws on $\xi$
  - variance generated by simulation draws.
  - in samples based on a small set of consumers: randomness in sample

- Berry Linton and Pakes, (2004 RESTUD): last 2 components likely to be very large if market shares are small.

- A separate issue: limit in $J$ or in $T$
  - in large part depends on the data
Challenges

• Knittel and Metaxoglou found that different optimizers give different results and are sensitive to starting values
• Some have used these results to argue against the use of the model
• Note, that its unlikely that a researcher will mimic the KM exercise
  • start from one starting point and not check others
  • some of the algorithms they use are not very good and rarely used
• It ends up that much (but not all!) of their results go away if
  • use tight tolerance in inversion \(10^{-12}\)
  • proper code
  • reasonable optimizers
• This is an important warning about the challenges of NL estimation
Judd and Su, and Dube, Fox and Su, advocate the use of MPEC algorithm instead of the Nested fixed point.

The basic idea: maximize the GMM objective function subject to the "equilibrium" constraints (i.e., that the predicted shares equal the observed shares).

Avoids the need to perform the inversion at each and every iteration of the search.

- performing the inversion for values of the parameters far from the truth can be quite costly.

The problem to solve can be quite large, but efficient optimizers (e.g., Knitro) can solve it effectively.

DFS report significant speed improvements.
MPEC (cont)

- Formally

\[
\min_{\theta, \xi} \xi' Z W Z' \xi
\]

subject to \( \tilde{\sigma}(\xi; x, p, F_{ns}, \theta) = S \)

- Note
  - the min is over both \( \theta \) and \( \xi \): a much higher dimension search
  - \( \xi \) is a vector of parameters, and unlike before it is not a function of \( \theta \)
  - avoid the need for an inversion: equilibrium constraint only holds at the optimum
  - in principle, should yield the same solution as the nested fixed point

- Many bells and whistles that I will skip
Lee (2011) builds on some of the ideas proposed in dynamic choice to propose what he calls Approximate BLP

The basic idea

Start with a guess to $\xi$, denoted $\xi^0$, and use it to compute a first order Taylor approximation to $\sigma(\xi_t, x_t, p_t; \theta)$ given by

$$\ln s(\xi_t; \theta) \approx \ln s^A(\xi_t; \theta) \equiv \ln s(\xi^0_t; \theta) + \frac{\partial \ln s(\xi^0_t; \theta)}{\ln \xi^0'}(\xi_t - \xi^0_t)$$

From $\ln S_t = \ln s^A(\xi_t; \theta)$ we get

$$\xi_t = \Phi_t(\xi^0_t, \theta) \equiv \xi^0_t + \left[\frac{\partial \ln s(\xi^0_t; \theta)}{\ln \xi^0'}\right]^{-1}(\ln S_t - \ln s(\xi^0_t; \theta))$$

Use this approximation for estimation

$$\min_\theta \Phi(\xi^0, \theta)'ZWZ'\Phi(\xi^0, \theta)$$
Nest this idea into K-step procedure

Step 1: Obtain new GMM estimate

\[ \theta^K = \arg \min_{\theta} \Phi(\xi^{K-1}, \theta)' \mathbf{ZWZ}' \Phi(\xi^{K-1}, \theta) \]

Step 2: Update \( \xi \)

\[ \xi^K = \Phi(\xi_{t-1}^{K}, \theta^K) \]

Repeat until convergence

Like MPEC avoids inversion at each stage, but has low dimensional search

Lee reports significant improvements over MPEC

Disclaimer: still a WP and has not been significantly tested
Comparing the methods

Patel (2012, chapter 2 NWU thesis) compared the 3 methods

Table 12. Conditional on convergence, the average time in seconds until convergence from $\theta_C$

<table>
<thead>
<tr>
<th>Markets</th>
<th>(a) MPEC</th>
<th></th>
<th>(b) NFP</th>
<th></th>
<th>(c) ABLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alternatives</td>
<td>Alternatives</td>
<td>Alternatives</td>
<td>Alternatives</td>
<td>Alternatives</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>25</td>
<td>50</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>50</td>
<td>2.9</td>
<td>4.7</td>
<td>8.5</td>
<td>2.91</td>
<td>4.71</td>
</tr>
<tr>
<td>100</td>
<td>4.4</td>
<td>6.0</td>
<td>40.3</td>
<td>4.54</td>
<td>8.38</td>
</tr>
<tr>
<td>250</td>
<td>7.0</td>
<td>11.1</td>
<td>38.7</td>
<td>9.49</td>
<td>14.93</td>
</tr>
<tr>
<td>1000</td>
<td>46.9</td>
<td>127.0</td>
<td>128.3</td>
<td>30.67</td>
<td>59.09</td>
</tr>
<tr>
<td>2500</td>
<td>64.6</td>
<td>209.7</td>
<td>954.4</td>
<td>70.65</td>
<td>133.12</td>
</tr>
</tbody>
</table>
Comparing the methods

Table 17. Conditional on convergence, the average time in seconds until convergence from $\theta_F$

<table>
<thead>
<tr>
<th>Markets</th>
<th>Alternatives</th>
<th>(a) MPEC</th>
<th>(b) NFP</th>
<th>(c) ABLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>25</td>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td>50</td>
<td>12.8</td>
<td>11.9</td>
<td>17.4</td>
<td>3.90</td>
</tr>
<tr>
<td>100</td>
<td>34.8</td>
<td>11.2</td>
<td>39.7</td>
<td>5.77</td>
</tr>
<tr>
<td>250</td>
<td>13.9</td>
<td>24.2</td>
<td>320.2</td>
<td>12.26</td>
</tr>
<tr>
<td>1000</td>
<td>-</td>
<td>128.8</td>
<td>-</td>
<td>38.84</td>
</tr>
<tr>
<td>2500</td>
<td>764.3</td>
<td>831.5</td>
<td>-</td>
<td>95.51</td>
</tr>
</tbody>
</table>

Word of caution: MPEC results can probably be significantly improved with better implementation. This is just an example of what one might expect if asking a (good) RA to program these methods.