Estimating Discrete Choice Models of Demand

Data

Aggregate (market)	Consumer-level
aggregate (market-level) quantity	consumer choices
prices + characteristics (+advertising)	prices + characteristics (+advertising) of <u>all</u> options
distribution of demographics (optional) sample from distribution data to estimate a parametric distribution	consumer demographics (optional)
market size (data + assumption)	choice of the outside option
advantages: (1) easier to get; (2) sample selection less of a problem;	advantages: (1) evaluating impact of demographics; (2) estimation easier (endogeneity?); (3) dynamics/ repeated choice;
disadvantages: (1) study of demographics and dynamics; (2) estimation is harder (?);	disadvantages: (1) harder to get; (2) sample selection;

Combining the two types of data.

Estimation with Aggregate Data

Identification (informally)

(i) How can we identify a consumer-level model from aggregate data?

As we saw, the model predicts aggregate shares as a function of the distribution of consumers. The idea is to choose the value of the parameters (that govern the distribution of consumers) to minimize the distance between predicted and observed aggregate shares (quantities):

$$\min_{\theta} \left\| s\left(x, p, \delta(x, p, \xi; \theta_1); \theta_2\right) - S \right\|$$

where $s(\cdot)$ are the market shares given by the model and S are the observed market shares.

We will not do exactly this but the intuition remains.

(ii) What variation in the data identifies cross-price effects?

- Ideal experiment: (randomly) vary prices (and characteristics) across markets and see how market shares vary.
- In reality prices will not vary randomly – we will consider IV – but more importantly we have a small number of markets (sometimes just 1) so this is not really the experiment we see in the data.
- In a single cross section we try to map the relation between market share and price/ characteristics (assume for now that these are uncorrelated with the error). The cross section of products allows us to identify the tradeoff between price and characteristics. The models maps these utility parameters into price effects.
- With several markets we get closer to the ideal experiment, but rarely do we have enough markets to credibly claim that we are near the ideal. So basically the same idea as a cross-section, with more flexibility in what we can control for.

(iii) What identifies the difference between the Logit, Nested Logit, and Random Coefficients (Mixed) Logit?

- With several markets: the (co) variation in market shares as prices/characteristics/choice set change.

  Example: 3 brands A, B, and C; 2 markets;
  In market 1: B and C have the same share;
  In market 2: price of A goes up (relative to market 1);
  Logit predicts that the S_B/S_C should stay at 1 (more generally whatever it was in market 1);
  RC Logit predicts that if B is a better substitute for A then its share should go up more.

- In a cross-section the thought experiment is slightly different: which model can reduce the impact of the error (i.e., the unobserved characteristic).

  In practice, it is very hard to estimate a RC without several markets.

Estimation

As we discussed a starting point is

$$\underset{\theta}{\text{Min}} \left\| s\left(x, p, \delta(x, p, \xi; \theta_1); \theta_2\right) - S \right\|$$

2 issues:

- computation (all parameters enter non-linearly)
- more importantly:
  prices might be correlated with the ξ ("structural" error);
  standard IV methods do not work;

The estimation method proposed by Berry (1994) transfers the model to fit into the standard (linear or) non-linear simultaneous equations model. Let $Z = [z_1, ..., z_M]$ be a set of instruments such that

$$E\left[Z_m \cdot \omega(\theta^*)\right] = 0 \quad , \quad m = 1 ... M,$$

where ω, a function of the model parameters, is an error-term defined below and θ^* denotes the "true" value of the parameters. The GMM estimate is

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \; \omega(\theta)' Z \Phi^{-1} Z' \omega(\theta)$$

where Φ is a consistent estimate of $E[Z'\omega\omega'Z]$.

The error term is defined as the structural error, ξ_{jt}.

We need to express the error-term as an explicit function of the parameters of the model and the data.
The key insight:
– the error-term, ξ_{jt}, only enters the mean utility level, $\delta(\cdot)$;
– the mean utility level is a linear function of ξ_{jt};
– we need δ as a function of the variables and parameters of the model;

This is done in several steps:

(0) prepare the data including draws from the distribution of *v* and *D* (only some models);

(1) for a given value of θ_2 and δ, compute the market shares implied by equation (4) (only for some models);

(2) for a given θ_2, compute the vector δ that equates the market shares computed in Step 1 to the observed shares;

(3) for a given θ, compute the error term (as a function of the mean valuation computed in Step 2), interact it with the instruments, and compute the value of the objective function;

(4) search for the value of θ that minimizes the objective function computed in Step 3.

Prep Step:

For any models requiring simulation draw "individuals".
Note: these draws are held fixed.

Step 1: Compute the market shares predicted by the model.

Given a value of θ_2 and δ (and the data) compute

$$s_{jt}(x_{.t}, p_{.t}, \delta_{.t}; \theta_2) = \int_{A_{jt}} dP^*(D, v, \epsilon).$$

For some models this can be done analytically. For example for Logit

$$s_{jt} = \frac{\exp(\delta_{jt})}{1 + \sum_{k=1}^{J} \exp(\delta_{kt})}$$

Nested Logit and PD GEV also have closed form solutions.

For RC Logit simulation is needed. The most common simulator is:

$$s_{jt}(p_{.t}, x_{.t}, \delta_{.t}, P_{ns}; \theta_2) = \frac{1}{ns} \sum_{i=1}^{ns} s_{ijt} =$$

$$\frac{1}{ns} \sum_{i=1}^{ns} \frac{\exp\left(\delta_{jt} + \sum_{k=1}^{K} x_{jt}^{k} (\sigma_k v_i^{k} + \pi_{k1} D_{i1} + \ldots + \pi_{kd} D_{id})\right)}{1 + \sum_{m=1}^{J} \exp\left(\delta_{mt} + \sum_{k=1}^{K} x_{mt}^{k} (\sigma_k v_i^{k} + \pi_{k1} D_{i1} + \ldots + \pi_{kd} D_{id})\right)},$$

where:

$(v_i^1, \ldots, v_i^K)$ and $(D_{i1}, \ldots D_{id})$, $i = 1, \ldots, ns$ are draws from $P_v^*(v)$ and $P_D^*(D)$, x_{jt}^k, $k = 1, \ldots, K$, are the variables which have random slope coefficients.

Note: (i) the ε's are integrated analytically;
(ii) other simulators exist (importance sampling, Halton seq);

Step 2: Invert the shares to get mean utilities

Compute the $J \times T$-dimensional vector of mean valuations, δ_{jt}, that equates the market shares computed in Step 1 to the observed shares. For each market solve the system of equations:

$$s\left(\delta_{.t}; \theta_2\right) = S_{.t} \qquad t = 1, \ldots, T,$$

where $s(\cdot)$ are the predicted market shares and S are the observed shares.

For the Logit model this inversion can be computed analytically:

$$\delta_{jt} = \ln(S_{jt}) - \ln(S_{0t})$$

where S_{0t} is the market share of the outside good;

For the RC Logit the system is solved using the contraction mapping

$$\delta_{.t}^{h+1} = \delta_{.t}^{h} + \ln(S_{.t}) - \ln\left(s(p_{.t}, x_{.t}, \delta_{.t}^{h}, P_{ns}; \theta_2)\right), \qquad t = 1, \ldots, T, \quad h = 0, \ldots, H,$$

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

$\delta_{.t}^{H}$ is the approximation to $\delta_{.t}$.

Step 3: Compute the GMM objective

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

$$\omega_{jt} = \delta_{jt}(S_{.t}; \theta_2) - (x_{jt}\beta + \alpha p_{jt}) \equiv \xi_{jt} \ .$$

Note: θ_1 enters this term, and the GMM objective, in a linear fashion, while θ_2 enters non-linearly.

The exact error depends on what is included in the model (e.g., brand dummy variables)

This error is interacted with the IV to form

$$\omega(\theta)' Z \Phi^{-1} Z' \omega(\theta)$$

Step 4: Search for the parameters that maximizes the objective.

For Logit model (with appropriate weight matrix) this is just 2SLS (or OLS).

For other models the search is done numerically. It can be simplified in two ways.

First, "concentrate out" the linear parameters. From the FOC:

$$\hat{\theta}_1 = (X_1' Z \Phi^{-1} Z' X_1)^{-1} X_1' Z \Phi^{-1} Z' \delta(\hat{\theta}_2),$$

Now, the non-linear search can be limited to θ_2.

Second, use the Implicit Function Theorem to compute the analytic gradient and use it to aid the search,

The derivatives of the mean value with respect to the parameters are

$$D\delta_{.t} = \begin{pmatrix} \frac{\partial \delta_{1t}}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_{1t}}{\partial \theta_{2L}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \delta_{Jt}}{\partial \theta_{21}} & \cdots & \frac{\partial \delta_{Jt}}{\partial \theta_{2L}} \end{pmatrix} = - \begin{pmatrix} \frac{\partial s_{1t}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{1t}}{\partial \delta_{Jt}} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_{Jt}}{\partial \delta_{1t}} & \cdots & \frac{\partial s_{Jt}}{\partial \delta_{Jt}} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial s_{1t}}{\partial \theta_{21}} & \cdots & \frac{\partial s_{1t}}{\partial \theta_{2L}} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_{Jt}}{\partial \theta_{21}} & \cdots & \frac{\partial s_{Jt}}{\partial \theta_{2L}} \end{pmatrix},$$

where θ_{2i}, $i = 1,...,L$ denotes the i's element of the vector θ_2, which contains the non-linear parameters of the model.

The derivatives of the share function are

$$\frac{\partial s_{jt}}{\partial \delta_{jt}} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{jti}}{\partial \delta_{jt}} = \frac{1}{ns}\sum_{i=1}^{ns} s_{jti}(1 - s_{jti})$$

$$\frac{\partial s_{jt}}{\partial \delta_{mt}} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{jti}}{\partial \delta_{mt}} = -\frac{1}{ns}\sum_{i=1}^{ns} s_{jti} s_{mti}$$

$$\frac{\partial s_{jt}}{\partial \sigma_k} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{jti}}{\partial \sigma_k} = \frac{1}{ns}\sum_{i=1}^{ns} s_{ji}\left(x_{jt}^k v_i^k - \sum_{m=1}^{J} x_{mt}^k v_i^k s_{mti} \right) = \frac{1}{ns}\sum_{i=1}^{ns} v_i^k s_{jti}\left(x_{jt}^k - \sum_{m=1}^{J} x_{mt}^k s_{mti} \right)$$

$$\frac{\partial s_{jt}}{\partial \pi_{kd}} = \frac{1}{ns}\sum_{i=1}^{ns} \frac{\partial s_{jti}}{\partial \pi_{kd}} = \frac{1}{ns}\sum_{i=1}^{ns} s_{jti}\left(x_{jt}^k D_{id} - \sum_{m=1}^{J} x_{mt}^k D_{id} s_{mti} \right) = \frac{1}{ns}\sum_{i=1}^{ns} D_{id} s_{jti}\left(x_{jt}^k - \sum_{m=1}^{J} x_{mt}^k s_{mti} \right)$$

The gradient of the objective function is

$$2 * D\delta' * Z * \Phi^{-1} * Z' * \omega .$$

Estimation with Consumer-Level Data

Identification is more straight-forward (subject to the caveat below);

Estimation is usually done using ML or Simulated ML. See Train (2003) for details.

Endogeneity is usually ignored.

> "Standard" argument: prices are not set separately for each consumer and therefore this is not an issue;
>
> This argument is wrong;
>
> Correlation with the error can arise:
> – unobserved characteristics (if not controlled for);
> – unobserved promotional activity;
> – through the consumer maximization problem;

How do we deal with the problem?

(1) Add more controls. For example, brand fixed effects to control for unobserved characteristics;

(2) Use the consumer data to generate the market shares and proceed as in the aggregate case. (Consumer data can generate both shares and serve as "simulation" draws)

(3) Use Control Function approach (Blundell and Powell, 2004; or for discrete choice models Petrin and Train, 2004).

Control Function

Consider the linear (demand) model: $y_t = x_t'\beta + \epsilon_t$

where x_t' is a K -dimension row vector;

Assume that $E(x_t u_t) \neq 0$ but that $E(z_t \epsilon_t) = 0$.

A standard way to proceed is by 2SLS. It is well-known that the resulting estimates can be expressed in different ways.

- fitted value approach: by using the "fitted value" from the first stage <u>instead</u> of the endogenous variable:

$$\beta_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

where $\hat{X}' = Z\hat{\Pi}$ and $\hat{\Pi} = (Z'Z)^{-1}Z'X$, and X, Z, Y, are $T \times K$, $T \times M$ and $T \times 1$ data arrays.

- control function approach: by including the first stage residual in the regression together with the endogenous variable:

$$\begin{pmatrix} \hat{\beta}_{2SLS} \\ \hat{\rho}_{2SLS} \end{pmatrix} = (\hat{W}'\hat{W})^{-1}\hat{W}'Y$$

where

$$\hat{W} = [X \ \hat{V}] \text{ and } \hat{V} = X - \hat{X} = X - Z\hat{\Pi}$$

This approach treats the endogeneity problem as an omitted variable problem. The solution is similar to a Heckman sample selection correction

- In non-linear models the fitted value approach does not generally lead to consistent estimates (aka "the forbidden regression");

- The control function approach will yield consistent estimates in non-linear models. It depends, however, on getting the correct functional form. Blundell and Powell (*RES*, 2004).

- Petrin and Train (2004) apply the control function approach and

the inversion approach (that we discussed for aggregate data) to several data sets. For the control function they simply include the residual from the "first stage" linearly and additive in the equation.

They find that the two approaches yield almost identical results in several examples;

One can construct examples where they two approach yield very different results;

P-T do not characterize when the control function approach will work. They also do not provide guidance of how to choose the functional form for the control function.

They advocate the use of control function even for cases where the inversion could work.

- Generally, their conclusions are not well received. It is not clear why one would one to use the control function approach when the inversion is available. If the inversion approach is unavailable the control function seems more reasonable.

Estimation Using Both Aggregate and Consumer-Level Data

Up to now we assumed that you have either aggregate or consumer level data. However, there are cases where you might have both. How can this be used?

Basic idea: use the methods of Imbens-Lancaster (94) to combine the moments from the aggregate and consumer data.

Usual motivation is to use the aggregate data to improve the efficiency of the estimates from the consumer data. But can also be used to deal with issues of sample selection (see Nevo *JBES,*2003).

In our context:
– Das, Olley and Pakes (94) use "shares" by income groups. Petrin (2003) uses purchase probabilities by family size and income to improve efficiency. The idea is the same as using market shares in different cities, where demographics are different.

– Berry, Levinsohn and Pakes (*JPE*, 2004) use aggregate and consumer level data. The consumer level data also has second choice data, which is what they focus on.

– Blattberg and (2004) use store level and consumer level data scanner date to address issues of sample selection in consumer level data.