Frontier Topics in Empirical Economics: Week 12
Discrete Choice Model II

Zibin Huang [1]

[1] College of Business, Shanghai University of Finance and Economics

December 26, 2025

Maximum Likelihood Estimation

- We have introduced the Logit model
- Now we consider how to estimate it

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{1}$$

- There are several ways for the estimation
- A simple and naive way is to log-linearize it:

$$lnP_{ni} = V_{ni} + \underbrace{ln(\sum_j e^{V_{nj}})}_{FE_n} \tag{2}$$

- When V_{ni} is linear, we run an OLS with fixed effect at n level

Maximum Likelihood Estimation

- Widely used in trade literature, especially in estimating the gravity equation
- Very easy to implement
- However, this is NOT the main method people usually use, especially in labor:
  - In data, when $P_{ni} = 0$, log values are undefined
    Granular setting issues in spatial economics (Dingel and Tintelnot, 2020)
  - Absorbed by the FE, any parameters for n level variables in V_{ni} cannot be estimated
  - The OLS only uses part of the information, not efficient

Maximum Likelihood Estimation

- The best method is MLE

$$\text{Likelihood Function: } L(\beta) = \prod_{n}^{N} \prod_{i} (P_{ni})^{y_{ni}}$$

$$\text{Log Likelihood Function: } LL(\beta) = \sum_{n=1}^{N} \sum_{i} y_{ni} ln P_{ni}$$

$$\text{MLE Estimator: } \hat{\beta}_{MLE} = argmax_{\beta} LL(\beta)$$

- y_{ni} is whether choice i is chosen in the data by individual n
- We choose β to maximize the probability of observing such data of y_{ni}

Maximum Likelihood Estimation

- In OLS, optimization is simple
- We can have a closed-form optimal solution
- But very often, there is no closed-form solution in MLE
- We have to use numerical non-linear optimization

Maximum Likelihood Estimation

- The idea of numerical non-linear optimization is: Guess → Update → Iterate
  - Step 1: Make an initial guess of β^0
  - Step 2: Use some updating rule to update β^0 to β^1, search for the optimal point
  - Step 3: Keep update in step 2 ($\beta^2, \beta^3, ..., \beta^t$) until we find the solution
- Then the question is, how to search and update?
- Today we will give a very brief introduction to numerical methods used in Economics

Maximum Likelihood Estimation

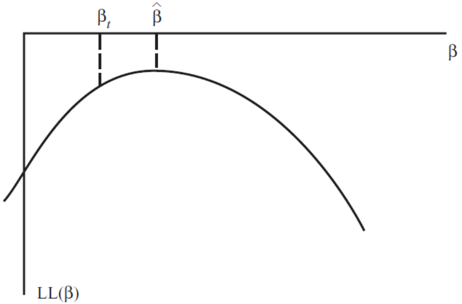

Figure 8.1. Maximum likelihood estimate.

How to search? Follow the derivative!

Maximum Likelihood Estimation

- Let's define $K \times 1$ gradient for each iteration β^t as:

$$g^t = \left(\frac{\partial LL(\beta)}{\partial \beta}\right) \tag{3}$$

- This is the vector of first order derivatives for an vector β
- Define $K \times K$ Hessian matrix for β^t as:

$$H^t = \left(\frac{\partial g^t}{\partial \beta'}\right) = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'}\right) \tag{4}$$

- This is the matrix of second order derivatives (including cross derivatives) for β

Maximum Likelihood Estimation

- For all gradient-based optimization method, we have the same form of updating rule:

$$\beta^{t+1} = \beta^{t} + \lambda M g^{t} \tag{5}$$

- g^{t} is the gradient, controlling the updating direction
- λ is a scalar called step size, M is a $K \times K$ matrix
- They control the speed of updating

Maximum Likelihood Estimation

- The most famous method is Newton-Raphson (NR):

$$\beta^{t+1} = \beta^{t} + \lambda(-H^{t})^{-1} g^{t} \tag{6}$$

- The NR method uses Hessian $(-H^{t})^{-1}$ as the speed matrix
- This is actually very intuitive
  - Gradient tells us the direction:
    Positive $\Rightarrow \beta^{t+1} \uparrow$; Negative $\Rightarrow$ need to decrease $\beta^{t+1} \downarrow$
  - Hessian is the curvature of the function:
    More curved means the slope changes quickly, need to be more conservative
- λ can help us adjust step size in case it is too large (update past maximum)

Maximum Likelihood Estimation

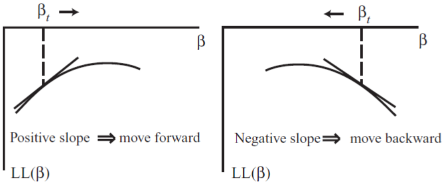

Figure 8.2. Direction of step follows the slope.

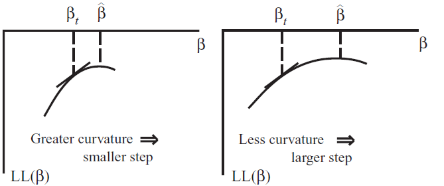

Figure 8.3. Step size is inversely related to curvature.

Maximum Likelihood Estimation

- There are two drawbacks of the NR method
  - Calculation of Hessian is computationally-intensive
  - No guarantee to follow the direction of gradient if LL is not globally concave
    Hessian is not necessarily positive definite!
- There are other methods using different speed matrix to overcome these issues
- For more details, please refer to Chapter 8 in Train (2009)
- When doing MLE, usually we can use packages like *Optim* in Julia, or *fminsearch* in Matlab
- No need to do it by yourself

Endogeneity in DCM: Issues

- So far we assume the unobserved errors ϵ is independent of the explanatory variables
- But this cannot always be the case
- Assume a case that consumers want to buy cars (BLP, 1995)
- The effect of the car price on the purchase probability is hard to be identified
- Since there can be unobserved car traits correlated with prices (smoothness of ride, brand reputation...)
- These are endogeneity issues in DCM
- IV estimation in linear regression cannot be directly used, even if we have an IV

Endogeneity in DCM: 1. BLP

- How to use IV in non-linear models such as DCM?
- Today we are going to introduce three methods
- (1) BLP; (2) Probit-IV; (3) Control function

Endogeneity in DCM: 1. BLP

- The first method is called BLP, introduced in Berry, Levinsohn, and Pakes (1995)
- The idea is to transform the endogeneity issue in a nonlinear model to a linear one
- Then we can use well-developed IV method to solve it

Endogeneity in DCM: 1. BLP

- Assume we have the car buying problem
- There are M markets with J_m options (brands) in each market
- Utility for consumer n in market j to choose brand m is:

$$U_{njm} = V(p_{jm}, x_{jm}, s_n, \beta_n) + \xi_{jm} + \epsilon_{njm}$$

- p_{jm} price; s_n personal attributes; x_{jm} product attributes; ξ_{jm} unobserved product attributes; ϵ_{njm} i.i.d. T1EV shock
- In simple case, V is linear
- Price is correlated with unobserved brand attributes $\xi_{jm} \not\perp\!\!\!\perp p_{mj}$
- Important feature of BLP: endogeneity comes from market-product level ξ_{jm}

Endogeneity in DCM: 1. BLP

- The idea of BLP employs a two-step approach
- First, add in a product-market level FE, absorb ξ_{jm}
- Estimate the equation with fixed effect
- Second, open the box of product-market level FE, estimate the remaining parameters
- The key point here is that endogeneity happens only at product-market level!

Endogeneity in DCM: 1. BLP

- We can decompose the observed utility value into

$$V_{njm} = \underbrace{\bar{V}(p_{jm}, x_{jm}, \bar{\beta})}_{\text{varies only over product-market}} + \underbrace{\tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)}_{\text{varies also over consumer}}$$

- Then we have the utility to be

$$U_{njm} = \underbrace{\left[\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}\right]}_{\text{a product-market level fixed effect}} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \epsilon_{njm}$$

- We just combine all terms varying only at product-market level together

Endogeneity in DCM: 1. BLP

- We define product-market level FE as:

$$\delta_{jm} = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm} \tag{7}$$

$$U_{njm} = \delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \epsilon_{njm} \tag{8}$$

- Equation (8) does not entail any endogeneity
- Step 1: We run a Logit model with *jm* level FE to estimate parameters $\tilde{\beta}$
- Step 2: We get estimates of δ_{jm} in step 1, and run IV regression for equation and get $\bar{\beta}$ (7)

Endogeneity in DCM: 1. BLP

The essence of BLP

- We cannot run IV regression directly in DCM
- We first pack all terms at the level where endogeneity happens into FE
- Then we estimate a DCM with these FEs
- We have estimated FEs, then unpack it and run linear IV regression
- Transform non-linear IV to be linear IV
- BLP tells you how to use an IV in a DCM, but only in a specific model structure.

Endogeneity in DCM: 1. BLP

- But then it comes to another problem in estimation step 1
- Sometimes we may have so many product-market combinations
- The dimension for the *jm*-level FE will be very high
- Traditional non-linear optimization algorithm in the MLE process can be slow or even impossible

Endogeneity in DCM: 1. BLP

- Here we have to use "Barry Contraction"
- It is a contraction mapping method to numerically estimate the FEs
- This is also a standard method for solving FE with high dimensions
- This procedure rests on that FEs determine predicted market shares for each product in each market
- Therefore, in each iteration, we can set the predicted shares from model to equal actual shares in data

Endogeneity in DCM: 1. BLP

- Let S_{jm} be the real market share of product j in market m in data (Share of BYD in Shanghai)
- Similarly, define $\hat{S}_{jm} = \sum_n \hat{P}_{njm}/N_m$ as the predicted share from your model
  - $\sum_n P_{njm}$ is the predicted total sales of product j in market m
  - N_m is the total sales in market m
- Denote δ as the vector of δ_{jm} for all j, m

Endogeneity in DCM: 1. BLP

- Let's see the Berry Contraction algorithm for Step 1
  - (1) Take an initial guess for parameters $\tilde{\beta}_n$ in $\tilde{V}(p_{jm}, x_{jm}, s_m, \tilde{\beta}_n)$
  - (2) Take an initial guess of δ
  - (3) For each guess of $\tilde{V}^t$ and δ^t, we calculate the choice value $\hat{U}^t$ for each consumer of each product in each market:
    $\hat{U}^t_{njm} = \delta^t_{jm} + \tilde{V}^t(p_{jm}, x_{jm}, s_m, \tilde{\beta}_n)$
  - (4) Given the calculated choice values, we calculate the predicted choice probability:
    $\hat{P}^t_{njm} = \frac{exp(\hat{U}^t_{njm})}{\sum_{j'} exp(\hat{U}^t_{nj'm})}, \quad \hat{S}^t_{jm} = \sum_n \hat{P}^t_{njm} / N_m$
  - (5) Given the choice probability, update FEs as follows: $\delta^{t+1}_{jm} = \delta^t_{jm} + ln(\frac{S_{jm}}{\hat{S}^t_{jm}(\delta^t)})$
  - (6) Iterate (3)-(5) until convergence of FEs for each guess of $\tilde{\beta}_n$
  - (7) Using traditional non-linear optimization method to iterate (1)-(6) until optimal point is found for $\tilde{\beta}_n$

Endogeneity in DCM: 1. BLP

- The idea is to separate the estimation of FEs
- In the outer loop (1)-(6), we estimate $\tilde{\beta}_n$ in $\tilde{V}(p_{jm}, x_{jm}, s_m, \tilde{\beta}_n)$ using traditional MLE algorithm
- In the inner loop (3)-(5), for each value of $\tilde{\beta}_n$, we iterate FEs δ_{jm}
- This procedure rests on that FEs determine predicted market shares for each product in each market
- Therefore, in each iteration, we can set the predicted shares from model to equal actual shares in data

Endogeneity in DCM: 1. BLP

- This algorithm in BLP is a contraction mapping

Definition: Contraction Mapping

Let (X, d) be a metric space. Then a map $T : X \to X$ is called a contraction mapping on X if there exists $q \in [0, 1)$ such that:

$$d(T(x), T(y)) \leq qd(x, y), \forall x, y \in X$$

- The contraction mapping means a function that squeezes points closer together in a space

Endogeneity in DCM: 1. BLP

- Then we have the famous Banach fixed-point theorem

Banach Fixed-point Theorem

Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed-point x^* in X, that is, $T(x^*) = x^*$. Furthermore, x^* can be found as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $(x_n)_{n \in N}$ by $x_n = T(x_{n-1})$ for $n \geq 1$. Then, $lim_{n \to \infty} x_n = x^*$

- The existence of a contraction mapping $T \Rightarrow$ Unique fixed point $T(x^*) = x^*$
- We can find x^* by iterate some arbitrary initial x_0 with T
- $x_0, T(x_0), T(T(x_0)), T(T(T(x_0)))...)$

Endogeneity in DCM: 1. BLP

- What does it mean in this BLP algorithm?
- It means that, as long as the updating iteration is a contraction mapping
- We will converge to the same point of δ, no matter what is our initial guess δ^0
- This terminal point is the fixed point, $\delta^* = F(\delta^*)$
- With δ^*, the predicted share equals the data share

Endogeneity in DCM: 1. BLP

- This idea is widely used in the computation problems
- For instance, calculating the equilibrium in a complicated model
- With a contraction mapping, we find the fixed-point which pins down the equilibrium
- More details will be discussed in my course next semester

Endogeneity in DCM: 2. Control Function

- BLP is not always feasible (error structure...)
- The algorithm of estimating BLP is complicated (involving some contraction)
- Highly recommend you to read BLP part in Train's book (or better, BLP 1993)
- The second important non-linear IV approach is Control Function (CF)

Endogeneity in DCM: 2. Control Function

- The utility of consumer n buying product j is:

$$U_{nj} = V(x_{nj}, w_{nj}, \beta_n) + \epsilon_{nj}$$

- x_{nj} is endogenous, $x_{nj} \not\perp\!\!\!\perp \epsilon_{nj}$
- We assume that there is an instrument z_{nj}, related with x_{nj} by first stage:

$$x_{nj} = W(z_{nj}, \gamma) + \mu_{nj} \tag{9}$$

- Assume that $\epsilon_{nj}, \mu_{nj} \perp\!\!\!\perp z_{nj}$, $\epsilon_{nj} \not\perp\!\!\!\perp \mu_{nj}$
- $\epsilon_{nj} \not\perp\!\!\!\perp \mu_{nj}$ implies that x_{nj} and ϵ_{nj} are correlated

Endogeneity in DCM: 2. Control Function

- We can do a CEF decomposition (given μ_{nj}) for ϵ_{nj}:

$$\epsilon_{nj} = \underbrace{E(\epsilon_{nj}|\mu_{nj})}_{CF(\mu_{nj},\lambda)} + \tilde{\epsilon}_{nj}$$

- By construction: $\tilde{\epsilon}_{nj} \perp\!\!\!\perp \mu_{nj}$
- Thus, we have $\tilde{\epsilon}_{nj} \perp\!\!\!\perp x_{nj}$ (x is correlated with ϵ only through μ)
- We call $CF(\mu_{nj}, \lambda)$ a control function, where λ is some parameter
- After controlling for $CF(\mu_{nj}, \lambda)$, x is no longer correlated with the error

Endogeneity in DCM: 2. Control Function

- Then we have the utility function as

$$U_{nj} = V(x_{nj}, w_{nj}, \beta_n) + CF(\mu_{nj}, \lambda) + \tilde{\epsilon}_{nj} \tag{10}$$

- Step 1: Estimate first stage equation (9), get residual of the first stage $\hat{\mu}$
- Step 2: Plug $\hat{\mu}$ in the CF (10), estimate equation (10) using simple Logit
- In step 2, we need to assume a functional form for CF
- Usually we can choose flexible non-parametric form (e.g. high-order polynomials)

Endogeneity in DCM: 2. Control Function

- The logic of CF approach is as follows:
  - We know that instrument z is not correlated with the error ϵ
  - Thus, endogenous variable x correlates with ϵ only through first stage error μ
  - Then by controlling the correlated parts of μ and ϵ, we can eliminate the correlation of x and ϵ
- CF is a pretty general method
- But it requires you to set a function form for CF

Endogeneity in DCM: 3. IV-Probit

- The last method we illustrate is IV-Probit
- It has very strong model structure assumptions
- Consider the following model:

$$
\begin{aligned}
y_1^* &= \delta_1 z_1 + \alpha_1 y_2 + u_1 \\
y_2 &= \delta_{21} z_1 + \delta_{22} z_2 + v_2 \\
y_1 &= \mathbf{1}(y_1^* > 0)
\end{aligned}
$$

- y_1^* is the latent utility; y_2 is the endogenous variable; z_1 is exogenous control
- (u_1, v_2) is bivariate normal; z_2 is the instrument with $(u_1, v_2) \perp\!\!\!\perp z_2$

Endogeneity in DCM: 3. IV-Probit

- In IV probit, we employ the assumption of the joint distribution of error (u_1, v_2)
- (u_1, v_2) is bivariate normal, then we can explicitly write down the likelihood function

$$f(y_1, y_2|z) = f(y_1|y_2, z)f(y_2|z) \tag{11}$$

$$= \Phi[\frac{\delta_1 z_1 + \alpha_1 y_2 + (\rho_1/\tau_2)(y_2 - z\delta_2)}{(1-\rho_1^2)^{\frac{1}{2}}}] \tag{12}$$

- We can use MLE to jointly estimate all these parameters
- One good thing is that we have stata package *ivprobit*

Endogeneity in DCM: Main Takeaways

Main Takeaways about IV in DCM

- Don't naively use IV method in linear model to solve endogeneity issue in non-linear model! (e.g., 2SLS)
- You can use BLP, CF, or IV-Probit
- BLP fits Logit model, but needs the endogeneity happens at higher level Product-market level in consumers' problem
- CF is pretty general, but needs you to non-parametrically estimate it
- IV-probit is simple in using Stata package, but has strong functional form assumption

References

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4):841–890.

Dingel, Jonathan I and Felix Tintelnot. 2020. "Spatial Economics for Granular Settings." Tech. rep., National Bureau of Economic Research.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge university press.