ORGANIZACIÓN INDUSTRIAL EMPÍRICA IN7E0

Carlos Noton

Clase 6 - Jueves 16 de Agosto
Primavera - 2018

Outline

Introduction

- We introduce the Random Coefficient Model (RCM) that addressed most of the concerns about IIA.
- RCM (also called Mixed Logit) allows for correlated shocks to similar products. Hence, we obtain more realistic patterns of substitution.
- BLP are able to estimate the parameters using market level data and assuming a known distribution for the consumer heterogeneity.
- Today, we cover a non-parametric alternative if you have individual level data.

Standard Logit (or Random Coefficient Model)

The standard Logit considers the following linear utility of consumer i for the model j in market t :

$$U_{ijt} = \beta_1(y_i - p_{jt}) + x_{jt}\beta_2 + \xi_{jt} + \varepsilon_{ijt}$$

where y_i is consumer's income, p_{jt} is the price and x_{jt} is the row vector of K observable characteristics, ξ_{jt} is an unobserved scalar product characteristic, and ε_{ijt} is a homoscedastic mean-zero stochastic term.

Standard Multinomial Logit

Rewrite the utility as follows

$$U_{ijt} = \beta z_{ijt} + \varepsilon_{ijt}$$

where z_{ijt} summarize all characteristics for products and individuals. Also $\beta = (\beta_1, \beta_2, \gamma_j)$, summarizes all utility coefficients where γ_j is the product dummy or product fixed effect that identifies ξ_{jt} when using individual level data.

Recall that in standard multinomial logit models we assume **homogeneous preferences**: same $\beta, \forall i$

Individual Probabilities in Multinomial Logit

Suppose you data on individual i making decisions or choosing between mutually exclusive products $y_{it} = \{0, 1, .., J\}$ over time $t = \{1, .., T\}$, where option 0 is the outside good.

One can compute the individual probabilities $s_{ijt}(\beta)$:

$$s_{ijt}(\beta) = Pr(y_{it} = j \mid \beta) = \frac{\exp(\beta z_{ijt})}{1 + \sum_k \exp(\beta z_{ikt})}$$

Notice that β is not consumer-specific (ie. homogenous consumers)

MLE

Define the dummy y_{ijt} is one if individual i chose product j at time t and zero otherwise.

$$L(\beta) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=0}^{J} s_{ijt}(\beta)^{y_{ijt}} = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=0}^{J} \left(\frac{\exp(\beta z_{ijt})}{1 + \sum_k \exp(\beta z_{ikt})} \right)^{y_{ijt}}$$

Then the estimation will maximize $\mathfrak{L} = \ln L(\beta)$

$$\mathfrak{L}(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} y_{ijt} \ln \left(\frac{\exp(\beta z_{ijt})}{1 + \sum_k \exp(\beta z_{ikt})} \right)$$

Standard optimization applies to find vector β that maximizes the objective function. Notice that β is not consumer-specific (ie. homogenous consumers)

Mixed Logit or Random Coefficient Model

The standard Mixed Logit considers the following linear utility of consumer i for the model j in market t :

$$\begin{aligned} U_{ijt} &= \alpha_i(y_i - p_{jt}) + x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \\ &= \beta_i z_{ijt} + \varepsilon_{ijt} \end{aligned}$$

where y_i is consumer's income, p_{jt} is the price and x_{jt} is the row vector of K observable characteristics, ξ_{jt} is an unobserved scalar product characteristic, and ε_{ijt} is a homoscedastic mean-zero stochastic term.

Recall we assume **heterogenous preferences**: β_i is different for each consumer and z_{ijt} contains all the observable characteristics and fixed effects.

Predicted Individual Probabilities

Same as before, if ε_{ijt} is type I extreme value, the individual probability is given by:

$$s_{ijt}(\beta_i) = Pr(y_{it} = j \mid \beta_i) = \frac{\exp(\beta_i z_{ijt})}{1 + \sum_{h=1}^{J} \exp(\beta_i z_{iht})}$$

Can we write the Log-Likelihood without the distribution of the vector β_i?

The mixing distribution of the Random Coefficients

The mixing distribution of the parameters is given by:

$$\beta_i \overset{iid}{\sim} f$$

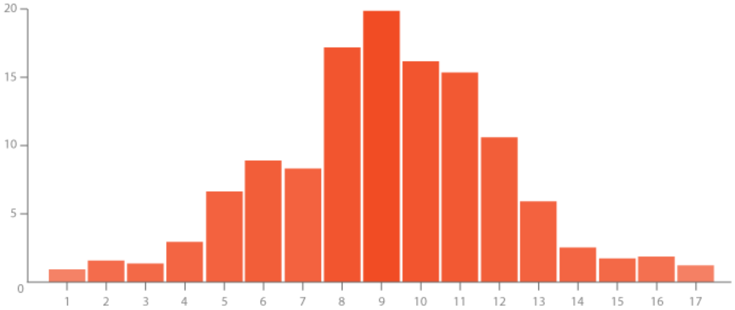
hence, the probability of $\beta_i = \overline{\beta}$ depends on the unknown density function f.

WLOG we assume that this distribution has a finite support S for random vector β_i.

Approximating density functions

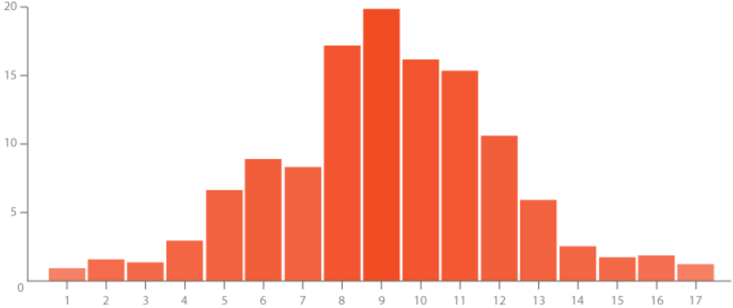
What do we know about a density function of a random variable with support S?

Can you see that any density function as just a collection of weights in the support S?



Weights

What weights do you need to characterize the following data?



What if we have a random vector with a multivariate distribution?

Non-Parametric Approach

This section heavely relies on Bajari, Fox and Ryan (2007); Train (2016), Nevo, Turner, Williams (2016) and Fox, Kim and Yang (2016)

The general approach is based on two steps:

1. Simulate many draws, say R, from the random vector β_r in a given support S.
2. Estimate the weights θ_r for each draw β_r that matches the data.

Fox, Kim and Yang (2016)

Suppose we have K random coefficients, $\beta \in \mathbb{R}^K$.

We take a large number of draws, R, of the vector β, denoting each draw by β^r. Draws are taken from a grid already chosen.

The probability, s_{ij}^r, that individual i chooses product j given a particular draw β^r is given by:

$$s_{ij}^r = s_{ij}(\beta^r) \quad = \quad Pr(y_i = j \mid \beta^r) = \frac{\exp(\beta^r z_{ij})}{1 + \sum_{h=1}^{J} \exp(\beta^r z_{ih})}$$

For simplicity, we suppress the subscript t as this approach does not exploit the time variation.

Expected Value

What is the expected probability that individual i chooses product j, , $E(s_{ij})$?

$$E(s_{ij}) = \int s_{ij}(\beta) dF(\beta)$$

Let us denote by θ^r the weight for draw β^r. Obviously, $\sum_{r=1}^{R} \theta^r = 1$ and $0 \leq \theta^r \leq 1, \forall r$. Therefore:

$$\widehat{E}(s_{ij}) = \sum_{r=1}^{R} \theta^r s_{ij}(\beta^r)$$

First Stage Simulation

Take a large number, R of draws of vector β^r

Compute a large set of conditional probabilities

$$s_{ij}^r = \frac{\exp(\beta^r z_{ij})}{1 + \sum_{h=1}^{J} \exp(\beta^r z_{ih})}$$

You will have a large $R \times 1$ vector of simulated probabilities s_{ij}^r.

Second Stage Optimization

Estimate of the unknown vector of weights, denoted by $\widehat{\theta}$, given by:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^R} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{ij} - \sum_{r=1}^{R} \theta^r s_{ij}^r \right)^2 \tag{1}$$

subject to:

$$\sum_{r=1}^{R} \theta^r = 1 \tag{2}$$

$$0 \leq \theta^r \leq 1, \forall r \tag{3}$$

where the dummy y_{ij} is one if individual i chose product j and zero otherwise.

Comments on Second Stage

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^R} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{ij} - \sum_{r=1}^{R} \theta^r s_{ij}^r \right)^2 \tag{4}$$

subject to:

$$\sum_{r=1}^{R} \theta^r = 1 \tag{5}$$

$$0 \leq \theta^r \leq 1, \forall r \tag{6}$$

Notice that in this stage the values s_{ij}^r are given and never recalculated.

To estimate just need the observed dummy y_{ij} and the $R \times 1$ vector of simulated probabilities s_{ij}^r.

Advantages

The advantage of using a logit model for the mixing distribution is that it allows for easy and flexible specification of relative probabilities.

The researcher just need to specify the support for the draws and the positive weights will describe the shape of the distribution (the constrained optimization guarantees positive probability at each point and the sum in the denominator sum to one over points.)

The specification is entirely general in the sense that any choice model with any mixing distribution can be approximated to any degree of accuracy by a model presented above.

Optimization Tip

Can you see that the optimization resembles an OLS estimation?

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^R} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{ij} - \sum_{r=1}^{R} \theta^r s_{ij}^r \right)^2$$

You could run a constrained OLS regression just ensuring the

non-negativity constraints $0 \leq \theta^r \forall r$.

Afterwards you can divide each estimated coefficient by the sum of all the positive coefficients, meeting all the constraints by construction.

Non Standard MLE as in Train (2016)

Suppose the person i made a **sequence of chosen alternatives** in T periods denoted by subscripts $j1, .., jt, .., jT$. The probability that person i made this sequence of choices, conditional on β^r, is:

$$L_i(\beta^r) = \prod_{t=1}^{T} \prod_{j=1}^{J} s_{ijt}(\beta^r)^{y_{ijt}} = \prod_{t=1}^{T} \prod_{j=1}^{J} \left(\frac{\exp(\beta^r z_{ijt})}{1 + \sum_{h=1}^{J} \exp(\beta^r z_{iht})} \right)^{y_{ijt}}$$

Notice that $L_i(\beta_r)$ does not depend on the weight θ^r.

Unconditional Choice Probabilities

The Unconditional Choice Probabilities (or expected value) that individual i chooses that particular sequence is P_i:

$$P_i(\theta) \;=\; \sum_{r \in S} \theta^r L_i(\beta^r)$$

Notice that we have integrated out over distribution of β. The probability is only a function of θ, the weights of the mixing distribution.

The simulated log-likelihood function for θ given a sample of individuals indexed by $i = \{1, .., N\}$ is:

$$SLL(\theta) = \sum_{i=1}^{N} \ln \{P_i(\theta)\} = \sum_{i=1}^{N} \ln \left\{ \sum_{r \in S} L_i(\beta_r)\theta^r \right\} \text{subject to:}$$

$$\sum_{r=1}^{R} \theta^r = 1$$

$$0 \leq \theta^r \leq 1, \forall r$$

The log-likelihood function can be simulated in the usual way by using random draws of β_r for each person.
However, the objective function is very non-linear.

Method of Moments as in NTW (2016)

A straightforward generalization can be:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^R} m_k(\theta)' \widehat{V}^{-1} m_k(\theta)$$

$$\text{subject to:}$$

$$\sum_{r=1}^{R} \theta^r = 1$$

$$0 \leq \theta^r \leq 1, \forall r$$

where $m_j(\theta)$ is a discrepancy between a moment observed in the data and the moment predicted by the model.

Moments as in NTW (2016)

Suppose we observed expenditure for each individual, E_i, and we have a structural model to predict expenditures given coefficients β^r, denoted by $\mathfrak{E}_i(\beta^r)$. Hence, formally, $m_j(\theta) = \widehat{m}_k^{dat} - m_k^{mod}\theta$, where:

- $\widehat{m}^{dat}$ is the vector of moments recovered from the data, for example Average Expenditure: $\overline{E} = \frac{1}{N}\sum_i E_i$.
- $m_k^{mod}\theta = \sum_r \theta^r \mathfrak{E}_i(\beta^r)$ is a weighted average of the moments predicted by the model, predicted expenditure in this example
- $\widehat{V}^{-1}$ is a weighting matrix, that we know there is an optimal weighting matrix derived by Hansen (1982) for GMM.

On the Optimality of Halton Sequences

- A natural starting point for a grid in R^K is a uniform grid.
- Also, you can have a draw from uniform distributions in R^K
- However, there are more efficient draws to ensure a better coverage, i.e., better accuracy of the integral given a fixed number of points
- The Halton sequences based on prime numbers is easy and very powerful to use. See Train's book.

Conclusions

- We have covered a cutting edge estimation technique to estimate Mixed Logit when individual level data are available.
- This approach allows us to explore non-parametric estimation of the mixing distribution of parameters.
- We have reached the frontier on non-parametric approaches to demand estimation of differentiated products.