'MATH+ECON+CODE' MASTERCLASS ON MATCHING MODELS, OPTIMAL TRANSPORT AND APPLICATIONS

Alfred Galichon (New York University)

Thursday: "Multinomial choice"
Block 11. Parametric multinomial choice

Learning objectives: block 11

- Parametric estimation of multinomial choice models
- Minimax regret estimation
- Simulation methods

References for block 11

- Savage, L. (1951). The theory of statistical decision. *JASA*.
- Bonnet, Fougère, Galichon, Poulhès (2019). Minimax estimation of hedonic models.

- Assume the utilities are parameterized as follows: $U = X\beta$ where $(\beta_k) \in \mathbb{R}^p$ is a parameter, and (X_{yk}) is a $|\mathcal{Y}| \times p$ matrix. Assume $\hat{s}_y$ is the observed market share of outcome y, and N is the total number of observations in the sample.
- The log-likelihood function is given by

$$I(\beta) = N \sum_y \hat{s}_y \log \sigma_y(X\beta)$$

- A common estimation method of β is by maximum likelihood

$$\max_\beta I(\beta).$$

MLE is statistically efficient; the problem is that the problem is not guaranteed to be convex, so there may be computational difficulties (e.g. local optima).

- In the logit case,

$$I(\beta) = N\left\{\hat{s}^{\mathsf{T}} X\beta - \log \sum_{y} \exp(X\beta)_y\right\}$$

so that the max-likehood amounts to

$$\max_{\beta}\left\{\hat{s}^{\mathsf{T}} X\beta - G(X\beta)_y\right\}$$

whose value is the Legendre-Fenchel transform of $\beta \to G(X\beta)$ evaluated at $X^{\mathsf{T}}\hat{s}$.

- Note that the vector $X^{\mathsf{T}}\hat{s}$ is the vector of empirical moments, which is a sufficient statistics in the logit model.
- As a result, in the logit case, the MLE is a convex optimization problem, and it is therefore both statistically efficient and computationally efficient.

- The previous remark will inspire an alternative procedure based on the moments statistics $X^{\mathsf{T}}\hat{s}$.
- The social welfare is given in general by $W(\beta) = G(X\beta)$. One has $\partial_{\beta_k} W(\beta) = \sum_y \sigma_y(X\beta) X_{yk}$, that is

$$\nabla W(\beta) = X^{\mathsf{T}} \sigma(X\beta),$$

  which is the vector of predicted moments.
- Therefore the program

$$\max_{\beta} \left\{ \hat{s}^{\mathsf{T}} X\beta - G(X\beta)_y \right\}$$

  picks up the parameter β which matches the empirical moments $X^{\mathsf{T}}\hat{s}$ with the predicted ones $\nabla W(\beta)$. This procedure is not statistically efficient, but is computationally efficient becauses it arises from a convex optimization problem.

- Until now, every decision-maker in a given observation $i \in \mathcal{I}$ was facing the same average utility associated with choosing outcome y, namely $U_y = \sum_k X_{yk}\beta_k$, independent on i. This can be overly restrictive. In the travel mode example, the time travelled to the airport, as well as the size of the family or the income, etc. all vary with each observation.
- Extend the analysis to model the utility associated with outcome y in observation i as

$$u_{iy} = \sum_k \Phi_{iyk}\beta_k,$$

  which we denote in a matrix way by $u = \Phi\beta$, where Φ is an $|\mathcal{I}|\,|\mathcal{Y}| \times K$ matrix. The rows of this matrix will be indexed by iy, and the columns will be indexed by k.
- The choice of the decision maker in observation $i \in \mathcal{I}$ is captured by $\mu_{iy} \in \{0, 1\}$, a dummy for choosing y.

- The analysis is left unchanged for the most part. In the logit case, the log-likelihood associated with observation i is

$$l_i(\beta) = \sum_{y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\beta)_{iy} - \log \sum_{y \in \mathcal{Y}} \exp(\Phi\beta)_{iy}$$

  and the max-likelihood rewrites as

$$\max_{\beta} \left\{ \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\beta)_{iy} - \sum_{i \in \mathcal{I}} \log \sum_{y \in \mathcal{Y}} \exp(\Phi\beta)_{iy} \right\}$$

- With other random utility structures, this yields a moment matching procedure to estimate β, namely

$$\max_{\beta} \left\{ \hat{\mu}\Phi\beta - \sum_{i \in \mathcal{I}} G((\Phi\beta)_{i\cdot}) \right\},$$

  where G is the Emax operator associated with the distribution of the random utility.

- Back to the logit case. Recall we have

$$I(\beta) = \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\beta)_{iy} - \sum_{i \in \mathcal{I}} \log \sum_{y \in \mathcal{Y}} \exp (\Phi\beta)_{iy}$$

- Assume that we restrict ourselves to $\beta^\top z > 0$. Then we can write $\beta = \theta / T$ where $T = 1/\beta^\top z$ and $\theta = \beta T$. Call $\Theta = \left\{\theta \in \mathbb{R}^p, \theta^\top z = 1\right\}$, so that $\beta = \theta / T$ where $\theta \in \Theta$ and $T > 0$. We have

$$I(\theta, T) = \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} / T - \sum_{i \in \mathcal{I}} \log \sum_{y \in \mathcal{Y}} \exp \left((\Phi\theta)_{iy} / T \right)$$

and we define the fixed temperature maximum likelihood estimator by

$$\theta(T) = \arg\max_{\theta} I(\theta, T)$$

- Note that $\theta(T) = \arg\max_{\theta\in\Theta} TI(\theta, T)$ where

$$TI(\theta, T) = \sum_{i\in\mathcal{I}, y\in\mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} - T \sum_{i\in\mathcal{I}} \log \sum_{y\in\mathcal{Y}} \exp\left((\Phi\theta)_{iy} / T \right)$$

and we note that as $T \to 0$,

$$TI(\theta, T) \to \sum_{i\in\mathcal{I}, y\in\mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} - \sum_{i\in\mathcal{I}} \max_{y\in\mathcal{Y}} \left\{ (\Phi\theta)_{iy} \right\}.$$

- Let $\theta(0) = \lim_{T\to 0} \theta(T)$. We have

$$\theta(0) \in \arg\max_{\theta\in\Theta} \left\{ \sum_{i\in\mathcal{I}, y\in\mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} - \sum_{i\in\mathcal{I}} \max_{y\in\mathcal{Y}} \left\{ (\Phi\theta)_{iy} \right\} \right\},$$

or

$$\theta(0) \in \arg\min_{\theta\in\Theta} \left\{ \sum_{i\in\mathcal{I}} \max_{y\in\mathcal{Y}} \left\{ (\Phi\theta)_{iy} \right\} - \sum_{i\in\mathcal{I}, y\in\mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} \right\}.$$

5. Minimax-regret estimation, principle

- We shall now describe a third class of estimation method called **minimax regret**.
- For each observation i, the actual choice y_i is observed. This choice yields utility $u_{y_i}(\theta)$, where θ is the parameter to be estimated. The regret associated with another possible choice y if the parameter value is θ is

$$R_{iy}(\theta) = u_y(\theta) - u_{y_i}(\theta).$$

- The *maximal regret* associated with parameter, which is utility associated with optimal choice minus utility associated with observed choice (if parameter value is θ), is therefore

$$\max_{y \in \mathcal{Y}} R_{iy}(\theta)$$

- The minimax regret procedude consists of picking the value of θ that *minimizes* the maximal regret:

$$\min_{\theta \in \Theta} \max_{y \in \mathcal{Y}} R_{iy}(\theta).$$

Dates back to Savage (1951). Very popular in online learning today.

- In our setting, the regret associated with choosing y in observation i is

$$R_{iy}(\theta) = (\Phi\theta)_{iy} - \sum_{y' \in \mathcal{Y}} \hat{\mu}_{iy'} (\Phi\theta)_{iy'} .$$

- The total max-regret is therefore

$$\sum_{i \in \mathcal{I}} \max_{y \in \mathcal{Y}} \{R_{iy}(\theta)\} = \sum_{i \in \mathcal{I}} \max_{y \in \mathcal{Y}} \left\{ (\Phi\theta)_{iy} \right\} - \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy}$$

- The minmax regret procedure thus consists of

$$\min_{\theta \in \Theta} \left\{ \sum_{i \in \mathcal{I}} \max_{y \in \mathcal{Y}} \left\{ (\Phi\theta)_{iy} \right\} - \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} (\Phi\theta)_{iy} \right\} .$$

hence

zero-temperature MLE=minimax regret!

- The previous problem can be computed in linear programming by

$$\min_{\theta_k, u_i} \sum_{i \in \mathcal{I}} u_i - \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} \left(\Phi\theta\right)_{iy}$$
$$u_i - \left(\Phi\theta\right)_{iy} \geq 0 \ \forall i \in \mathcal{I}, y \in \mathcal{Y}$$
$$z^\top \theta = 1$$

- Recall that $u_{iy} = u_i 1_{\{y \in \mathcal{Y}\}}$ is represented as $1_{\mathcal{Y}} \otimes u$, and

$$1_{\mathcal{Y}} \otimes u = (1_{\mathcal{Y}}.1) \otimes (I_{\mathcal{I}}.u) = (1_{\mathcal{Y}} \otimes I_{\mathcal{I}}) \,.\, (1 \otimes u) = (1_{\mathcal{Y}} \otimes I_{\mathcal{I}}) \,.u,$$

thus the previous program can be written using a Kronecker product as

$$\min_{\theta_k, u_i} 1_{\mathcal{I}}^\top u - \hat{\mu}^\top \Phi\theta$$
$$s.t. \begin{pmatrix} 1_{\mathcal{Y}} \otimes I_{\mathcal{I}} & -\Phi \\ 0 & z^\top \end{pmatrix} \begin{pmatrix} u \\ \theta \end{pmatrix} \geq \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- Note that the set of θ that enter the solution to the problem above is not unique, but is a convex set. Denoting r the value of program, we can look for bounds of $\theta^\intercal d$ for a chosen direction d by

$$\begin{aligned}
&\min_{\theta_k, u_i} \theta^\intercal d \\
&\quad \sum_{i \in \mathcal{I}} u_i - \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} \left(\Phi\theta\right)_{iy} = r \\
&\quad u_i - \left(\Phi\theta\right)_{iy} \geq 0 \ \forall i \in \mathcal{I}, y \in \mathcal{Y} \\
&\quad z^\top \theta = 1
\end{aligned}$$

- Consider now reintroducing random utility by simulation. Clone each observation i a number B of times and for $b \in \mathcal{B} := \{1, ..., B\}$, simulate $\varepsilon_{biy} \sim P$, where P is a distribution over $\mathbb{R}^{\mathcal{Y}}$. The utility associated with outcome y in observation i as $u_{biy} = \sum_k Phi_{iyk}\beta_k + \varepsilon_{biy}$.
- The minmax regret estimator will now be approximated by

$$\min_{\theta_k, u_{ib}} \frac{1}{B} \sum_{i \in \mathcal{I}, b \in \mathcal{B}} u_{ib} - \sum_{i \in \mathcal{I}, y \in \mathcal{Y}} \hat{\mu}_{iy} \left(\Phi\theta\right)_{iy}$$
$$u_{ib} - \left(\Phi\theta\right)_{iy} \geq \varepsilon_{biy}$$

which can be written using a Kronecker product as

$$\min_{\theta_k, u_{ib}} \frac{1}{B} 1_{\mathcal{IB}}^{\top} u - \hat{\mu}^{\top} \Phi\theta$$
$$s.t. \ \begin{pmatrix} 1_{\mathcal{Y}} \otimes I_{\mathcal{IB}} & -\Phi \otimes 1_{\mathcal{B}} \end{pmatrix} \begin{pmatrix} u \\ \theta \end{pmatrix} \geq \varepsilon$$

- We can simulate the logit model, by taking ε distributed as a Gumbel variable, and will approximate solutions to the logistic regression.

- For $n = |\mathcal{I}|$, denote $\widehat{\mathbb{P}}_n$ the empirical distribution of the sample (x_i, y_i), $\widehat{\mathbb{E}}_n$ the corresponding expectation, and rewrite the minimax regret estimator as

$$\max_{\beta} \left\{ \widehat{\mathbb{E}}_n \left[\Phi(X,Y)^\top \beta - G\left(\Phi(X,.)^\top \beta \right) \right] \right\},$$

where $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$. Let $\widehat{\beta}$ be the solution to that problem.

- By first order conditions,

$$\widehat{\mathbb{E}}_n \left[\Phi(X,Y) \right] = \widehat{\mathbb{E}}_n \left[\sum_{y \in \mathcal{Y}} \Phi(X,y) \frac{\partial G}{\partial U_y} \left(\Phi(X,.)^\top \beta \right) \right]$$

which becomes in the logit case a true mle estimate, namely

$$\widehat{\mathbb{E}}_n \left[\Phi(X,Y) \right] = \widehat{\mathbb{E}}_n \left[\sum_{y \in \mathcal{Y}} \Phi(X,y) \frac{\exp\left(\Phi(X,y)^\top \beta \right)}{\sum_z \exp\left(\Phi(X,z)^\top \beta \right)} \right]$$

- Now, let β be the "true parameter", which is solution to the problem in the population

$$\max_{\beta} \left\{ \mathbb{E} \left[\Phi(X,Y)^\top \beta - G\left(\Phi(X,.)^\top \beta \right) \right] \right\}.$$

- Introduce
$$I(\beta, Z) = \Phi(X, Y)^\top \beta - G\left(\Phi(X, .)^\top \beta\right)$$
so that the problem becomes $\max_\beta \widehat{\mathbb{E}}_n [I(\beta, Z)]$.
- **Theorem**. One has
$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \to N\left(0, F^{-1} S F^{-1}\right)$$
where $S = \mathbb{E}\left[\partial_\beta I(\beta, Z) \left(\partial_\beta I(\beta, Z)\right)^\top\right]$, and $F = \mathbb{E}\left[\partial^2_{\beta\beta} I(\beta, Z)\right]$, and
$$\frac{\partial I(\beta, y)}{\partial \beta^k} = \Phi^k(X, Y) - \sum_{y \in \mathcal{Y}} \Phi^k(X, y) \frac{\partial G}{\partial U_y}\left(\Phi(X, .)^\top \beta\right)$$
$$\frac{\partial^2 I(\beta, y)}{\partial \beta^k \partial \beta^\kappa} = -\sum_{y \in \mathcal{Y}, z \in \mathcal{Y}} \Phi^k(X, y) \Phi^\kappa(X, z) \frac{\partial^2 G}{\partial U_y \partial U_z}\left(\Phi(X, .)^\top \beta\right)$$

- **Proof**. We take a first order expansion of $\widehat{\mathbb{E}}_n \left[\partial_\beta l \left(\widehat{\beta}, Z \right) \right] = 0$, which yields

$$\left(\widehat{\mathbb{E}}_n - \mathbb{E} \right) \left[\partial_\beta l \left(\widehat{\beta}, Z \right) \right] + \mathbb{E} \left(\partial^2_{\beta\beta} l \left(\beta, Z \right) \right) \left(\widehat{\beta} - \beta \right) + o_P \left(1/\sqrt{n} \right) = 0$$

which obtains

$$\widehat{\beta} - \beta = - \left(\mathbb{E} \left[\partial^2_{\beta\beta} l \left(\beta, Z \right) \right] \right)^{-1} \left(\left(\widehat{\mathbb{E}}_n - \mathbb{E} \right) \left[\partial_\beta l \left(\beta, Z \right) \right] \right) + o_P \left(1/\sqrt{n} \right)$$

and by the central limit theorem,

$$\sqrt{n} \left(\widehat{\mathbb{E}}_n - \mathbb{E} \right) \left[\partial_\beta l \left(\beta, Z \right) \right] \to N \left(0, \mathbb{E} \left[\partial_\beta l \left(\beta, Z \right) \left(\partial_\beta l \left(\beta, Z \right) \right)^\top \right] \right)$$

in distribution.

- Back to the travel mode example. For each individual i, we have access to: y=travel mode (bus car etc); T_{iy}=time taken (observed); C_{iy}=generalized cost for passenger (observed); I_i=income; y_j=travel mode actually chosen.
- Our model is

$$u_{iy} = U_y - T_{iy}\left(a + bI_i\right) - C_{iy}$$
$$u_{iy} = \sum_y U_y 1_{\{y_j = y\}} - T_{ij}\left(a + bI_i\right) - C_{ij}c$$

- **Let's code it!**