Dynamic Discrete Choice

Robert A. Miller

Econometrics 2

October 2018

- The course material can be found at:
  - http://comlabgames.com/econometrics/
- The lectures are in three segments:
  1. Introduction to Econometrics Modeling
  2. Asymptotic Theory for Parametric Models
  3. Nonparametric and Semiparametric Methods
- Your grade will come from:
  1. Three assignments (20 percent each)
  2. Final examination on lectures and reading material (40 percent)

- Throughout this course and its sequel we will imagine that the data is generated by a model, and in that way embrace the classical laws of probability and statistics.
- For the most part we assume the model comes from economics:
  - Individuals solve dynamic optimization problems.
  - Groups of individuals or firms play a noncooperative game using equilibrium strategies.
  - Asymmetrically informed individuals make optimal contracts with each other.
  - Individuals and firms make consumption and production choices in competitive equilibrium.
- To help understand how economic models provide the basis for estimation and testing we introduce the course by analyzing some of the first structural econometric models in:
  - dynamic discrete choice
  - competitive equilibrium models with continuous choices.

- The data typically comprise a sample of individuals for which there are records on some of their:
  - background characteristics
  - choices
  - outcomes from those choices.
- What are the challenges to making predictions and testing hypotheses when we take this approach?
  1. The choices and outcomes of economic models are typically nonlinear in the underlying parameters characterizing the model that we seek to estimate.
  2. The data variables on background, choices and outcomes might be an incomplete description about what is relevant to the model.

- Each period $t \in \{1, 2, \ldots, T\}$ for $T \leq \infty$, an individual chooses among J mutually exclusive actions.
- Let d_{jt} equal one if action $j \in \{1, \ldots, J\}$ is taken at time t and zero otherwise:

$$d_{jt} \in \{0, 1\}$$

$$\sum_{j=1}^{J} d_{jt} = 1$$

- At an abstract level assuming that choices are mutually exclusive is innocuous, because two combinations of choices sharing some features but not others can be interpreted as two different choices.
- For example in a female labor supply and fertility model, suppose:

$$j \in \{(\text{work, no birth}), (\text{work, birth}), (\text{no work, no birth}), (\text{no work, birth})\}$$

- Suppose that actions taken at time t can potentially depend on the state $z_t \in Z$.
- For Z finite denote by $f_{jt}(z_{t+1} | z_t)$, the probability of z_{t+1} occurring in period $t+1$ when action j is taken at time t.
- For example in the example above, suppose $z_t = (w_t, k_t)$ where:
  - $k_t \in \{0, 1, \ldots\}$ are the number of births before t
  - $w_t \equiv d_{1,t-1} + d_{2,t-1}$, so $w_t = 1$ if the female worked in period $t-1$, and $w_t = 0$ otherwise.
- Note that Z must be defined compatible to the transition matrix: for example setting $z_t = (w_t, k_t)$ where $k_t \in \{0, 1, \ldots\}$ are the number of births before $t-1$, is incompatible with assumption about transitions and choices.
- With up to 5 offspring, 3 levels of experience, the number of states including age (say 50 years) is 750. Add in 4 levels of education (less than high school, high school, some college and college graduate) and 3 racial categories, increases this number to 9000.

- When Z is finite there is a $Z \times Z$ transition matrix for each (j, t).
- In many applications the matrices are sparse.
- In the example above they have $9,000^2 = 81$ million cells.
- However households can only increase the number of kids one at time.
- They can only increase or decrease their work experience by one unit at most.
- Hence there are at most six cells they can move from (w_t, k_t):

$$\left\{ \begin{array}{l} (w_t, k_t), (w_t, k_t + 1), (w_t + 1, k_t), \\ (w_t + 1, k_t + 1), (w_t - 1, k_t), (w_t - 1, k_t + 1) \end{array} \right\}$$

- Therefore a transition matrix has at most $54,000$ nonzero elements, and all the nonzero elements are one.
- Given a deterministic sequence of actions sequentially taken over S periods, we can form the S period transition matrix by producting the one period transitions.

- If Z is a Euclidean space $f_{jt}(z_{t+1}|z_t)$ is the probability (density function) of z_{t+1} occurring in period $t+1$ when j is picked at time t.
- With almost identical notation we could model $z_t \in Z_t$ and in this way generalize from states of the world to histories, or information known at t, or t-measurable events.
- For example in a health application we might define $z_t \equiv \{h_s\}_{s=1}^{t-1}$ as a medical record with $h_s \in \{\text{healthy at } s, \text{ sick at } s\}$.

- The individual's current period payoff from choosing j at time t is determined by z_t, which is revealed to the individual at the beginning of the period t.
- The current period payoff at time t from taking action j is $u_{jt}(z_t)$.
- Given choices $(d_{1t}, \ldots, d_{Jt})$ in each period $t \in \{1, 2, \ldots, T\}$ the individual's expected utility is:

$$E\left\{\sum_{t=1}^{T}\sum_{j=1}^{J}\beta^{t-1}d_{jt}u_{jt}(z_t)\right\}$$

  where $\beta \in (0, 1)$ is the subjective discount factor, and at each period t the expectation is taken over $z_{t+1}, \ldots, z_T$.
- Formally β is redundant if u is subscripted by t; we typically include a geometric discount factor so that infinite sums of utility are bounded, and the optimization problem is well posed.

- Write the optimal decision at period t as a decision rule denoted by $d_t^o(z)$ formed from its elements $d_{jt}^o(z_t)$.
- Let $V_t(z_t)$ denote the value function in period t, conditional on behaving according to the optimal decision rule:

$$V_t(z_t) \equiv E\left[\sum_{\tau=t}^{T}\sum_{j=1}^{J}\beta^{\tau-t}d_{j\tau}^o(z_\tau)\,u_{j\tau}(z_\tau)\right]$$

- In terms of period $t+1$:

$$\beta V_{t+1}(z_{t+1}) \equiv \beta E\left\{\sum_{\tau=t+1}^{T}\sum_{j=1}^{J}\beta^{\tau-t-1}d_{j\tau}^o(z_\tau)\,u_{j\tau}(z_\tau)\right\}$$

- Appealing to Bellman's (1958) principle we obtain:

$$
\begin{aligned}
V_t(z_t) &= \sum_{j=1}^{J} d_{jt}^{o} u_{jt}(z_t) \\
&\quad + \sum_{j=1}^{J} d_{jt}^{o} \sum_{z=1}^{Z} E\left\{ \sum_{\tau=t+1}^{T} \sum_{j=1}^{J} \beta^{\tau-t} d_{j\tau}^{o}\left(z_\tau\right) u_{j\tau}(z_\tau) \,|z \right\} f_{jt}\left(z|z_t\right) \\
&= \sum_{j=1}^{J} d_{jt}^{o} \left[u_{jt}(z_t) + \beta \sum_{z=1}^{Z} V_{t+1}(z) f_{jt}\left(z|z_t\right) \right]
\end{aligned}
$$

when Z is finite with a similar expression holding (using an integral) when Z is Euclidean.

- To compute the optimum for T finite, we first solve a static problem in the last period to obtain $d_T^o(z)$.
- Applying backwards induction $i \in \{1, \ldots, J\}$ is chosen to maximize:

$$u_{it}(z_t) + E\left\{ \sum_{\tau=t+1}^{T} \sum_{j=1}^{J} \beta^{\tau-t-1} d_{j\tau}^o(z_\tau) u_{j\tau}(z_\tau) \,|z_t, d_{it} = 1 \right\}$$

- In the stationary infinite horizon case we assume $u_{jt}(z) \equiv u_j(z)$ and that $u_j(z) < \infty$ for all (j, z).
- Consequently expected utility each period is bounded and the contraction mapping theorem applies, proving $d_t^o(z) \to d^o(z)$ for large T.

- Let $v_{jt}(z_t)$ denote the flow payoff of action j plus the expected future utility of behaving optimally from period $t+1$ on:

$$v_{jt}(z_t) \equiv u_{jt}(z_t) + \beta \sum_{z_{t+1}=1}^{Z} V_{t+1}(z_{t+1}) f_{jt}\left(z_{t+1} | z_t\right)$$

- By definition:

$$d_{jt}^{o}\left(z_t\right) \equiv I\left\{v_{jt}(z_t) \geq v_{kt}(z_t) \forall\ k\right\}$$

- Suppose we observe the states z_{nt} and decisions d_{nt} of individuals $n \in \{1, \ldots, N\}$ over time periods $t \in \{1, \ldots, T\}$.
- Could we use such data to infer the primitives of the model:
  1. A consistent estimator of $f_{jt}\left(z_{t+1} | z_t\right)$ can be obtained from the proportion of observations in the (t, j, z_t) cell transitioning to z_{t+1}.
  2. There are $(J-1)N$ inequalities relating the pairs of mappings $v_{jt}(z_t)$ and $v_{kt}(z_t)$ for each observation on d_{nt} at (t, z_t).
  3. We can recursively derive the values of $u_{jt}(z_t)$ from the $v_{jt}(z_t)$ values.

Inference

Why unobserved heterogeneity is introduced into data analysis

- Note that if two people in the data set with the same (t, z_t) made different decisions, say j and k, then $v_{jt}(z_t) = v_{kt}(z_t)$. This raises two potential problems for modeling data this way:
  1. In a large data set it is easy to imagine that for every choice $j \in \{1, \ldots, J\}$ and every (t, z_t) at least one sampled person n sets $d_{nt} = 1$. If so, we would conclude that the population was indifferent between all the choices, and hence the model would have no empirical content because no behavior could be ruled out.
  2. This approach does not make use of the information that some choices are more likely than others; that is the proportions of the sample taking different choices at (t, z_t) might vary, some choices being observed often, others perhaps very infrequently.
- For these two reasons, treating all heterogeneity as observed, and trying to predict the decisions of individuals, is not a very promising approach to analyzing data.

Inference

Unobserved heterogeneity

- A more modest objective is to predict the probability distribution of choices margined over factors that individuals observe, but data analysts do not.
- In this respect we seek to predict the behavior of a population, not each individual, essentially obliterating that difference between macroeconomics and microeconomics.
- We now assume the states can be partitioned into those which are observed, x_t, and those that are not, ϵ_t.
- Thus $z_t \equiv (x_t, \epsilon_t)$.
- Suppose the data consist of N independent and identically distributed draws from the string of random variables $(X_1, D_1, \ldots, X_T, D_T)$.
- The n^{th} observation is given by $\left\{x_1^{(n)}, d_1^{(n)}, \ldots, x_T^{(n)}, d_T^{(n)}\right\}$ for $n \in \{1, \ldots, N\}$.

Inference

Data generating process

- Denote the probability (density) of the pair $(x_{t+1}, \epsilon_{t+1})$, conditional on $\left(x_t^{(n)}, \epsilon_t\right)$ and the optimal action taken by n at t, as:

$$H_{nt}\left(x_{t+1}, \epsilon_{t+1}\left|x_t^{(n)}, \epsilon_t\right.\right) \equiv$$

$$\sum_{j=1}^{J} I\left\{d_{jt}^{(n)} = 1\right\} d_{jt}^{o}\left(x_t^{(n)}, \epsilon_t\right) f_{jt}\left(x_{t+1}, \epsilon_{t+1}\left|x_t^{(n)}, \epsilon_t\right.\right)$$

- Conditional on $x_1^{(n)}$ the joint probability of $\left\{d_1^{(n)}, x_2^{(n)}, \ldots, x_T^{(n)}, d_T^{(n)}\right\}$ is:

$$\Pr\left\{d_1^{(n)}, x_2^{(n)}, \ldots, x_T^{(n)}, d_T^{(n)}\left|x_1^{(n)}\right.\right\} =$$

$$\int_{\epsilon_T} \ldots \int_{\epsilon_1} \left[\begin{array}{l} \sum_{j=1}^{J} I\left\{d_{jT}^{(n)} = 1\right\} d_{jT}^{o}\left(x_T^{(n)}, \epsilon_T\right) \times \\ \prod_{t=1}^{T-1} H_{nt}\left(x_{t+1}^{(n)}, \epsilon_{t+1}\left|x_t^{(n)}, \epsilon_t\right.\right) g\left(\epsilon_1\left|x_1^{(n)}\right.\right) \end{array}\right] d\epsilon_1 \ldots d\epsilon_T$$

- Let $\theta \in \Theta$ uniquely index a specification of $u_{jt}(z_t)$, $f_{jt}(z_{t+1}|z_t)$ and β under consideration.
- Conditional on $x_1^{(n)}$ suppose $\left\{d_1^{(n)}, x_2^{(n)}, \ldots, d_T^{(n)}\right\}_{n=1}^{N}$ was generated by $\theta_0 \in \Theta$.
- Define $\epsilon \equiv (\epsilon_1, \ldots, \epsilon_T)$. The maximum likelihood estimator, θ_{ML}, selects $\theta \in \Theta$ to maximize the joint probability of the observed occurrences conditional on the initial conditions:

$$\theta_{ML} \equiv \underset{\theta \in \Theta}{\arg\max} \left\{ N^{-1} \sum_{n=1}^{N} \log \left(\Pr \left\{ d_1^{(n)}, x_2^{(n)}, \ldots, x_T^{(n)}, d_T^{(n)} \middle| x_1^{(n)}; \theta \right\} \right) \right\}$$

- If there is a unique maximum in $\theta \in \Theta$ to:

$$\int_{x_1^{(n)}} \log\left(\Pr\left\{d_1^{(n)}, x_2^{(n)}, \ldots, x_T^{(n)}, d_T^{(n)} \left| x_1^{(n)}; \theta\right.\right\}\right) dF\left(x_1^{(n)}\right)$$

  then the model is identified, and under standard conditions θ_{ML} is $\sqrt{N}$ consistent, asymptotically normal, and asymptotically efficient.

- Intuitively:

  1. a model is identified if no other model in the Θ set of models has the same data generating process.
  2. an estimator of an identified model is consistent if it converges to θ_0 in some probabilistic sense as N increases without bound.
  3. the rate of convergence, $\sqrt{N}$ in this case, gives the order of the convergence.
  4. asymptotically normality refers to the limiting distribution, in N, of $\sqrt{N}\left(\theta_{ML} - \theta_0\right)$.
  5. asymptotic efficiency refers to the lowest asymptotic variance of all consistent estimators with the same rate of convergence.