# Overview

Robert A. Miller

Structural Econometrics

October 2021

# Lectures on Structural Econometrics

## Website, topics and themes

- The lecture material, some assignments and background reading for these 28 sessions can be found at:
  - http://comlabgames.com/structuraleconometrics/
- There are two sets of lectures with four segments in the first group:
  1. Introduction to Structural Econometrics
  2. Summarizing the Data
  3. Probability
  4. Asymptotic Theory for Nonlinear Models
- There are three segments in the second set of lectures.
  1. Dynamic Discrete Choice
  2. Market Microstructure
  3. Optimal Contracting
- Throughout these lectures we will imagine the data is generated by a model, and 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 optimally contract 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
  - market microstructure
  - optimal contracting with moral hazard.

- 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 of the model we wish 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 $t \in \{1, 2, \ldots, T\}$ and each state $z_t \in Z$ the individual's expected utility is:

$$E\left\{\sum_{t=1}^{T}\sum_{j=1}^{J}\beta^{t-1}d_{jt}u_{jt}(z_t)\,|z_1\right\}$$

  where $\beta \in (0, 1)$ is the subjective discount factor, and at each period $t$ the expectation is taken over $z_2, \ldots, z_T$.
- Formally $\beta$ 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 $d_t^o\left(z_t\right)$ 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:

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

- In terms of period $t+1$:

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

- Appealing to Bellman's (1958) principle we obtain, when $Z$ is finite:

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

- A similar expression holds when $Z$ is Euclidean using an integral.

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

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

- 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 $d_t^o(z) \rightarrow d^o(z)$ for large $T$.

# Inference

## Estimating a model when all heterogeneity is observed

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

$$v_{jt}(z_t) \equiv u_{jt}(z_t) + \beta \sum_{z \in Z} V_{t+1}(z) f_{jt}(z|z_t)$$

- By definition:

$$d^o_{jt}(z_t) \equiv I\{v_{jt}(z_t) \geq v_{kt}(z_t) \forall\ k\}$$

- Suppose we observe the states $z_{nt}$ and decisions $d_{nt} \equiv (d_{n1t}, \ldots, d_{nJt})$ 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 $f_{jt}(z_{t+1}|z_t)$ can be obtained from the proportion of observations in the $(t, j, z_t)$ cell transitioning to $z_{t+1}$.
  2. There are $(J-1)\sum_{n=1}^{N} I\{z_{nt} = z_t\}$ 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. Can we 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 $d_{njt} = 1$. 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.

- 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, $\epsilon_t$.
- 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

## Transition density given optimal behavior

- 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:

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

- Note that both $d_{jt}^{(n)}$, an indicator that the data shows $n$ chooses $j$ at $t$, and also $d_{jt}^{o}\left(x_t^{(n)}, \epsilon_t\right)$, what $n$ would optimally choose $j$ at $t$, appear in this formula.
- Thus $H_{nt}\left(x_{t+1}, \epsilon_{t+1} \left| x_t^{(n)}, \epsilon_t\right.\right)$ embeds the assumption that the density for $(x_{t+1}, \epsilon_{t+1})$ is generated by the joint transition $d_{jt}^{o}\left(x_t^{(n)}, \epsilon_t\right) f_{jt}\left(x_{t+1}, \epsilon_{t+1} \left| x_t^{(n)}, \epsilon_t\right.\right)$ for the observed choice.

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

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

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

where $g\left(\epsilon_1 \left| x_1^{(n)} \right.\right)$ is the density of $\epsilon_1$ conditional on $x_1^{(n)}$.

- Let $\theta \in \Theta$ uniquely index a specification of $u_{jt}(z_t)$, $f_{jt}\left(z_{t+1} \middle| z_t\right)$ and $\beta$ 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 (ML) estimator, $\theta_{ML}$, selects $\theta \in \Theta$ to maximize the joint probability of the observed occurrences conditional on the initial conditions:

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

- This model is point identified if and only if (iff) $\theta_0$ is the unique solution when $\theta \in \Theta$ is chosen to maximize:

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

- If the model is point identified, $\theta_{ML}$ is $\sqrt{N}$ consistent, asymptotically normal, and asymptotically efficient:
  1. a model is *point identified* if no other model in the $\Theta$ set of models has the same *data generating process*.
  2. an estimator of an identified model is *consistent* if it converges to $\theta_0$ in some probabilistic sense as $N$ increases without bound.
  3. the *rate of convergence*, $1/2$ in this case, is the greatest $\alpha$ leaving the limit of $N^{\alpha}(\theta_{ML} - \theta_0)$ bounded in some probabilistic sense.
  4. asymptotic normality means the *limiting distribution* (again as $N$ increases without bound), of $\sqrt{N}(\theta_{ML} - \theta_0)$ is normal.
  5. *asymptotic efficiency* refers to the lowest asymptotic variance of all consistent estimators with the same rate of convergence.

# Criteria for Evaluating Estimators

## Three criteria for assessing an estimator

- Three criteria for evaluating an estimator of a point-identified model are:

  1. Large sample properties:
     - Does the estimator converge to the identified set?
     - If so, what is the rate of convergence?
     - What is the asymptotic distribution of the estimator?
  2. Finite sample properties:
     - At what sample size do the finite sample properties accurately reflect the asymptotic distribution?
     - For a given sample size, what is the standard deviation and mean squared error of the estimator ?
  3. Implementation:
     - Is the estimator defined by an algorithm or only a set of conditions to be satisfied?
     - Are numerical approximations involved?
     - Does the estimator require tuning parameters or instruments?

# Large Sample or Asymptotic Properties

In what sense does an estimator converge, and what does it converge to?

- There are several types of convergence, such as: almost sure, in mean square, and in probability.
- Given a type of convergence, we ask:
  1. Does the estimator converge to a set that includes the identified set? In other words is the estimator tight?
  2. Is the set of parameters to which the estimator converges included in the identified set? In other words is the estimator sharp?
- If both conditions are satisfied, then we say the estimator is consistent.
- For example if the identified set is a singleton, that is the model is pointwise identified, then an estimator is consistent if it converges to that singleton.
- Note that if the model is not point identified, we would not expect an extremum estimator (such as a conventionally defined ML) to converge.

- The other two criteria are extensively analyzed in econometric theory, and can typically be applied to dynamic discrete choice models in a straightforward way.
- For example, suppose the parameter space is $\Theta$, the data is generated by $\theta_0 \in \Theta$, the model in point identified, and the estimator, denoted by $\theta_N$ is consistent with:

$$\theta_N \underset{p}{\longrightarrow} \theta_0$$

- The rate of convergence is defined by $N^{\alpha}$ where:

$$\alpha = \arg\sup_{a} \left[ N^{a} \left( \theta_N - \theta_0 \right) \right] \underset{p}{\longrightarrow} 0$$

- Structural estimates of dynamic discrete choice models are typically $\sqrt{N}$ consistent.

- Suppose $\theta_N$ converges in probability to $\theta_0$ at rate $\alpha$.
- Let $\xi$ be drawn from the limiting distribution of $N^{\alpha}(\theta_N - \theta_0)$:

$$N^{\alpha}(\theta_N - \theta_0) \underset{d}{\longrightarrow} \xi$$

- Structural estimates of dynamic discrete choice models are typically asymptotically normal.
- An estimator is asymptotically efficient if $\xi$ is $\mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right)$ where:

$$\mathcal{I}(\theta) \equiv E\left[\frac{\partial l(d, x | x_1 ; \theta)}{\partial \theta} \frac{\partial l(d, x | x_1 ; \theta)}{\partial \theta}'\right] = -E\left[\frac{\partial^2 l(d, x | x_1 ; \theta)}{\partial \theta \partial \theta'}\right]$$

  and the likelihood is based on the sequence $(d, x)$ conditional on the state at date one, $x_1$.
- The ML estimator for dynamic discrete choice models typically attain $\mathcal{I}(\theta_0)^{-1}$ the Cramer-Rao lower bound.

# Implementation

## Does an algorithm define the estimator?

- Ideally an estimator is defined by an algorithm that depends on the data for each sample size $N$. In that case the estimator:
    1. can be implemented mechanically, so is easy to explain;
    2. is easy to replicate on the same and on different data sets, a virtue in scientific enquiry.
- Cell estimators and hence unrestricted ML estimators satisfy this definition.
- An OLS estimator also satisfies the first definition because algorithms exist to invert matrices exactly, within a finite number of steps.
- Similarly Gaussian methods, successively substituting out parameters, solve linear systems quickly within a finite number of steps.

# Implementation

## Is the estimator defined by a set of conditions it must satisfy?

- A weaker, more inclusive definition is that an estimator solves a set of conditions jointly satisfied by the parameter values and the data.
- Since the algorithm used to implement the estimator is not defined, such estimators are almost invariably, less transparent, and therefore harder to replicate with data.
- Extremum estimators for nonlinear models defined this way include:
  - nonlinear least squares;
  - full solution estimators to dynamic discrete choice models;
  - CCP estimators in which $G$ or $\beta$ is estimated.
- It is useful to know whether a unique solution exists. For example:
  - Is the minimization (maximization) problem strictly convex (concave)?
- If not, can all the parameters, bar one or two, be solved in terms of the one or two remaining parameters?
  - In the first case, the concentrated objective function can be plotted.
  - In the second equi-value contours can be plotted.

# Implementation

## Are numerical approximations involved?

- Because ML estimation of dynamic discrete choice models is relatively imposing in terms of programming demands and computational time, researchers economize on both by using numerical approximations:
  1. approximating distant horizons with zero;
  2. approximating smoothed integrals with rectangles and quadrilaterals;
  3. linearizing the value function;
  4. interpolating the state space to obtain estimates of continuation values;
  5. approximating $E[\max\{x, y\}]$ with $\max\{E[x], E[y]\}$;
  6. reducing the impact of the state space by treating the continuation value as a sufficient statistic for the state space;
  7. more generally only allowing the individuals to condition on a smaller set of values than there are state variables.
- These approximation errors open a gap between the defined estimator and its numerical counterpart.