Choice models with panel data

Serial correlation and dynamic choices

Michel Bierlaire

Mathematical Modeling of Behavior

Outline

Static model

Serial correlation

Dynamic model

Dynamic model with panel effects

Introduction

Panel data

- Type of data used so far: cross-sectional.
- Cross-sectional: observation of individuals at the same point in time.
- Time series: sequence of observations.
- **Panel data** is a combination of comparable time series.

Introduction

Panel data

Data collected over multiple time periods for the same sample of individuals.

Multidimensional

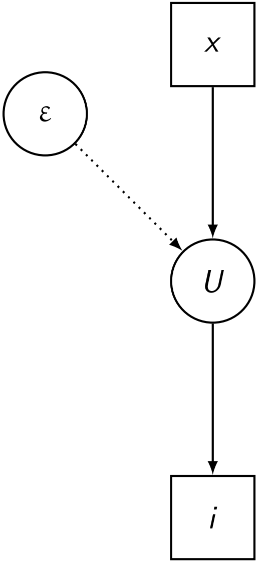
Individual	Day	Price of stock 1	Price of stock 2	Purchase
n	t	x_{1nt}	x_{2nt}	i_{int}
1	1	12.3	15.6	1
1	2	12.1	18.6	2
1	3	11.0	25.3	2
1	4	9.2	25.1	0
2	1	12.3	15.6	2
2	2	12.1	18.6	0
2	3	11.0	25.3	0
2	4	9.2	25.1	1

Introduction

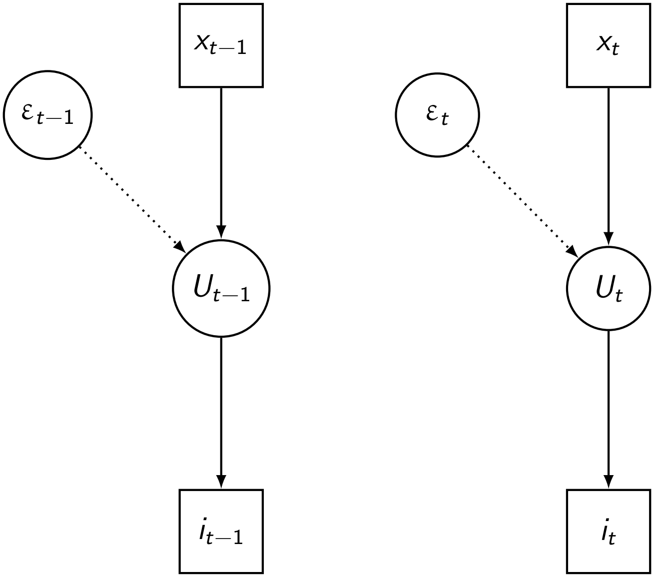
Examples of discrete panel data

- People are interviewed monthly and asked if they are working or unemployed.
- Firms are tracked yearly to determine if they have been acquired or merged.
- Consumers are interviewed yearly and asked if they have acquired a new cell phone.
- Individual's health records are reviewed annually to determine onset of new health problems.

Model: single time period



Static model



Static model

Utility

$$U_{int} = V_{int} + \varepsilon_{int}, \ \ i \in \mathcal{C}_{nt}.$$

Assumption

$$\varepsilon_{int} \text{ i.i.d. } \mathrm{EV}(0,1), \ \text{ across } i, n \text{ and } t.$$

Logit

$$P(i_{nt}) = \frac{e^{V_{int}}}{\sum_{j \in \mathcal{C}_{nt}} e^{V_{jnt}}}.$$

Static model

Estimation: contribution of individual n to the log likelihood

$$P(i_{n1}, i_{n2}, \ldots, i_{nT}) = P(i_{n1})P(i_{n2}) \cdots P(i_{nT}) = \prod_{t=1}^{T} P(i_{nt})$$

$$\ln P(i_{n1}, i_{n2}, \ldots, i_{nT}) = \ln P(i_{n1}) + \ln P(i_{n2}) + \cdots + \ln P(i_{nT}) = \sum_{t=1}^{T} \ln P(i_{nt})$$

Static model

Comments

- Views observations collected through time as supplementary cross sectional observations.
- Standard estimation procedure for cross sectional data may be used directly.
- Simple, but there are two important limitations.

Static model: limitations

Serial correlation

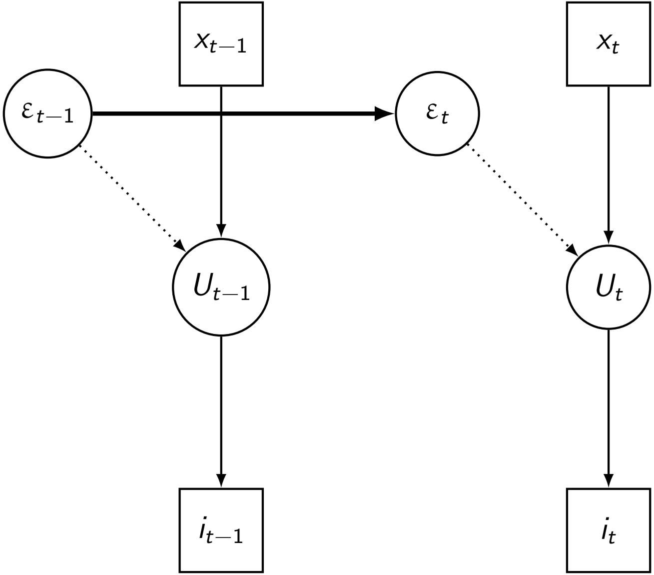
- unobserved factors persist over time,
- in particular, all factors related to individual n,
- $\varepsilon_{in(t-1)}$ cannot be assumed independent from ε_{int}.

Dynamics

- Choice in one period may depend on choices made in the past,
- e.g. learning effect, habits.

Outline

Dealing with serial correlation



Panel effects

Relax the assumption that ε_{int} are independent across t.

Assumption about the source of the correlation

- individual related unobserved factors,
- persistent over time.

The model

$$\varepsilon_{int} = \alpha_{in} + \varepsilon'_{int}$$

It is also known as

- agent effects,
- unobserved heterogeneity.

Panel effects

- Assuming that ε'_{int} are independent across t,
- we can apply the static model.
- Two versions of the model:
  - with fixed effects: α_{in} are unknown parameters to be estimated,
  - with random effects: α_{in} are distributed.

Static model with fixed effects

Utility

$$U_{int} = V_{int} + \alpha_{in} + \varepsilon'_{int}, \; i \in \mathcal{C}_{nt}.$$

Assumptions

- ε'_{int} i.i.d. EV(0, 1), across i, n and t.
- α_{in} unknown parameters to be estimated.
- α_{in} independent from ε'_{int}.

Logit

$$P(i_{nt}) = \frac{e^{V_{int}+\alpha_{in}}}{\sum_{j\in\mathcal{C}_{nt}} e^{V_{jnt}+\alpha_{jn}}}$$

Static model with fixed effects

Estimation: contribution of individual n to the log likelihood

$$P(i_{n1}, i_{n2}, \ldots, i_{nT}) = P(i_{n1}) P(i_{n2}) \cdots P(i_{nT}) = \prod_{t=1}^{T} P(i_{nt})$$

$$\ln P(i_{n1}, i_{n2}, \ldots, i_{nT}) = \ln P(i_{n1}) + \ln P(i_{n2}) + \cdots + \ln P(i_{nT}) = \sum_{t=1}^{T} \ln P(i_{nt})$$

Static model with fixed effects

Comments

- α_{in} capture permanent taste heterogeneity.
- For each n, one α_{in} must be normalized to 0.
- The α's are estimated consistently only if $T \to \infty$.
- This has an effect on the other parameters that will be inconsistently estimated.
- In practice,
  - T is usually too short,
  - the number of α parameters is usually too high,

  for the model to be consistently estimated and practical.

Static model with random effects

- Denote α_n the vector gathering all parameters α_{in}.
- Assumption: α_n is distributed with density $f(\alpha_n)$.
- For instance:

$$\alpha_n \sim N(0, \Sigma).$$

- We have a mixture of static models.
- Given α_n, the model is static, as ε'_{int} are assumed independent across t.

Static model with random effects

Utility

$$U_{int} = V_{int} + \alpha_{in} + \varepsilon'_{int}, \;\; i \in \mathcal{C}_{nt}.$$

Assumptions

- ε'_{int} i.i.d. EV(0, 1), across i, n and t.
- $\alpha_n \sim N(0, \Sigma)$, with pdf f.
- α_n independent from ε'_{int}.

Conditional choice probability

$$P(i_{nt}|\alpha_n) = \frac{e^{V_{int}+\alpha_{in}}}{\sum_{j \in \mathcal{C}_{nt}} e^{V_{jnt}+\alpha_{jn}}}$$

Static model with random effects

Contribution of individual n to the log likelihood, given α_n

$$P(i_{n1}, i_{n2}, \ldots, i_{nT} | \alpha_n) = \prod_{t=1}^{T} P(i_{nt} | \alpha_n).$$

Unconditional choice probability

$$P(i_{n1}, i_{n2}, \ldots, i_{nT}) = \int_{\alpha} \prod_{t=1}^{T} P(i_{nt} | \alpha) f(\alpha) d\alpha.$$

Static model with random effects

Estimation

- Mixture model.
- Usually requires simulation.
- Generate draws $\alpha^1, \ldots, \alpha^R$ from $f(\alpha)$.
- Approximate

$$P(i_{n1}, i_{n2}, \ldots, i_{nT}) = \int_\alpha \prod_{t=1}^{T} P(i_{nt}|\alpha) f(\alpha) d\alpha \approx \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T} P(i_{nt}|\alpha^r).$$

- The product of probabilities can generate very small numbers.

$$\sum_{r=1}^{R} \prod_{t=1}^{T} P(i_{nt}|\alpha^r) = \sum_{r=1}^{R} \exp\left(\sum_{t=1}^{T} \ln P(i_{nt}|\alpha^r) \right).$$

Static model with random effects

Comments

- Parameters to be estimated: β's and Σ's
- Maximum likelihood estimation leads to consistent and efficient estimators.
- Ignoring the correlation (i.e. assuming that α_n is not present) leads to consistent but not efficient estimators (not the true likelihood function).
- Accounting for serial correlation generates the true likelihood function and, therefore, the estimates are consistent and efficient.

Relax the i.i.d. assumption

i.i.d. assumption

- ✓ Same η for all alternatives i: relaxed.
- ✓ Same η for all observations n: relaxed.
- ✓ Same μ for all alternatives i: relaxed.
- ✓ Same μ for all observations n: relaxed.
- ✓ Independence across alternatives i: relaxed.
- ▶ Independence across observations n: relaxed in this lecture.

Outline

Static model

Serial correlation

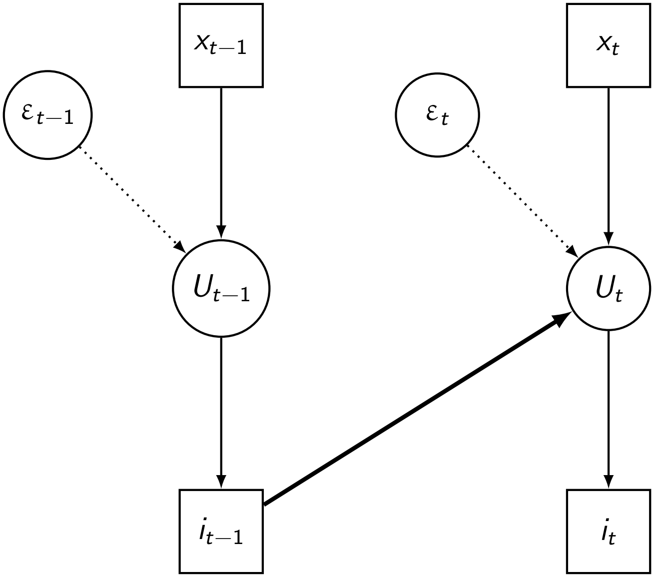
Dynamic model

Dynamic model with panel effects

Dynamics

- Choice in one period may depend on choices made in the past
- e.g. learning effects, habits.
- Simplifying assumption:
  - the utility of an alternative at time t
  - is influenced by the choice made at time $t-1$ only.
- It leads to a dynamic Markov model.

Dynamic Markov model



Notation

$$y_{jnt} = \begin{cases} 1 & \text{if } i_{nt} = j \\ 0 & \text{otherwise.} \end{cases}$$

Example

$$i_{nt} = 2 \Leftrightarrow y_{nt} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

Dynamic Markov model

The model

$$U_{int} = V_{int} + \gamma y_{in(t-1)} + \varepsilon_{int}, \;\; i \in \mathcal{C}_{nt}.$$

$$y_{in(t-1)} = \begin{cases} 1 & \text{if alternative } i \text{ was chosen by } n \text{ at time } t-1 \\ 0 & \text{otherwise.} \end{cases}$$

Estimation: same as for the static model

except that observation $t = 0$ is lost

Outline

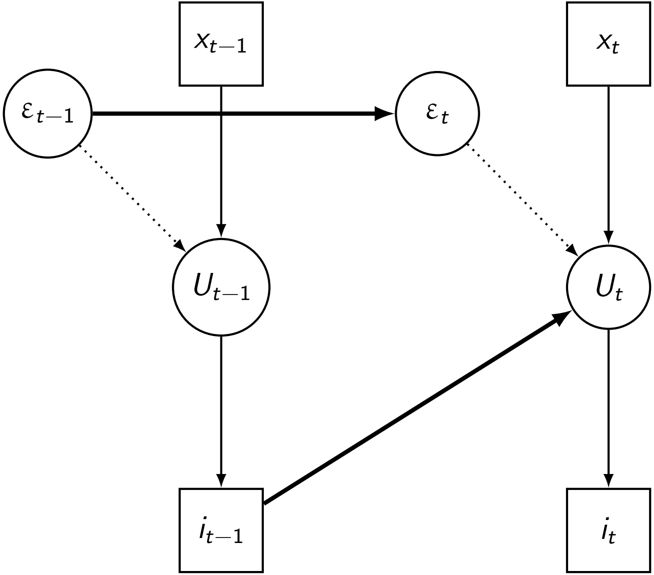
Static model

Serial correlation

Dynamic model

Dynamic model with panel effects

Dynamic Markov model with serial correlation



Dynamic Markov model

Extension: combine Markov with panel effects

$$U_{int} = V_{int} + \alpha_{in} + \gamma y_{in(t-1)} + \varepsilon'_{int}, \; i \in \mathcal{C}_{nt}.$$

Dynamic Markov model with fixed effects

- Similar to the static model with fixed effects.
- Similar limitations.

Dynamic Markov model with random effects

The initial condition problem.

The initial condition problem

History

- The dynamic choice process usually has an history before the sampling period.
- The only information about history is y_{n0}.
- Because of persistence of unobserved factor, y_{n0} is correlated with these unobserved factors.
- Examples: heavy smokers, car lovers, etc.

Endogeneity

- Cause: y_{n0} is correlated with α_n.
- Problem: inconsistent parameter estimates.

Dynamic Markov model with panel effects

Utility function

$$U_{int} = V_{int} + \alpha_{in} + \gamma y_{in(t-1)} + \varepsilon'_{int}, \; i \in \mathcal{C}_{nt}.$$

Contribution of individual n to the log likelihood, given i_{n0} and α_n

$$P(i_{n1}, i_{n2}, \ldots, i_{nT} | i_{n0}, \alpha_n) = \prod_{t=1}^{T} P(i_{nt} | i_{n0}, \alpha_n).$$

Dynamic Markov model with panel effects

Wooldridge's model

Assume a distributions of α, depending on the first choice.

$$f(\alpha_n | i_{n0})$$

We integrate out α_n

$$P(i_{n1}, i_{n2}, \ldots, i_{nT} | i_{n0}) = \int_\alpha \prod_{t=1}^{T} P(i_{nt} | i_{n0}, \alpha) f(\alpha | i_{n0}) d\alpha.$$

[Wooldridge, 2005]

Dynamic Markov model with random effects

- The main difference between static model with RE and dynamic model with RE is the term

$$f(\alpha | i_{n0})$$

- It captures the distribution of the panel effects, knowing the first choice.

Modeling

$$\alpha_n = a + b^T y_{n0} + c^T x_n + \xi_n, \quad \xi_n \sim N(0, \Sigma_\alpha).$$

- a, b and c are vectors and Σ_α a matrix of parameters to be estimated.
- x_n capture the entire observed history $(t = 1, \ldots, T)$ for agent n.
- This addresses the endogeneity issue.

Summary

- Panel data consist in observations of the same individuals over time.
- Static model suffers from two limitations.
- Serial correlation is addressed with the agent effect.
- Dynamic choices are captured by the Markov model.
- Initial condition problem: endogeneity.

Bibliography I

Wooldridge, J. M. (2005).
Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity.
Journal of applied econometrics, 20:39–54.