# Limited dependent variables

Paolo Zacchia

Microeconometrics

Lecture 13

# Plan of the lecture

This lecture covers a selection of econometric models that feature a **limited dependent variable** (LDV). The tools developed in this lecture have wide applicability, and are instrumental towards some particular topics treated in later lectures (14-18).

Specifically, this lecture covers three major themes.

1. Models for **multinomial responses** (multinomial logit and probit, models for ordered LDVs): the backbone of demand estimation (Lecture 14) and *entry* game models (Lecture 16).

2. LDV models for **panel data** (fixed/random effects adapted to LDVs), occasionally useful in Lectures 17 and 18.

3. The **dynamic logit** model (Rust, 1987) which is helpful for the understanding of dynamic games (Lecture 16).

Knowledge of simple logit and probit (Lecture 11) is assumed.

# Review of multinomial response models

What follows is an overview of leading econometric **multinomial** response models. The following are presented in sequence:

- the **multinomial logit model**;
- the **nested (multinomial) logit model**;
- the **mixed (multinomial) logit model**;
- the **multinomial probit model**;
- and **ordered multinomial models** (probit and logit).

Emphasis is placed on the foundational multinomial logit model; the other models, while motivated, are treated more briefly.

# The multinomial logit model (1/9)

- The **multinomial logit** is an important limited dependent variable (LDV) model for a **multinomial** outcome $Y_i$.

- That is, the support of $Y_i$ (write it $\mathbb{Y}$) is *finite* and *countable*.

- Let there be $J$ alternative realizations of $Y_i$ ($|\mathbb{Y}| = J$).

- Typically, the dependent variable is coded over a collection of integers, $Y_i = 1, 2, \dots, J$: however, numbers do **not** imply an ordered relationship of any sort.

- Thus, the outcome variable can be conveniently re-coded in terms of $J$ Bernoulli variables $Y_{ji}$ for $j = 1, \dots, J$ with:

$$Y_{ji} = \begin{cases} 1 & \text{if } Y_i = j \\ 0 & \text{otherwise.} \end{cases}$$

# The multinomial logit model (2/9)

- Interest in this model falls on the *probability* that any of the $J$ possible realizations of $Y_i$ occurs as a function of some $K$ observable characteristics $\boldsymbol{x}_{ji} = (X_{1ji}, X_{2ji}, \dots, X_{Kji})$ that are possibly **specific to alternative** $j = 1, \dots, J$.

- If for example $Y_i$ represents different product alternatives, $\boldsymbol{x}_{ji}$ may represent the subjective evaluation that a consumer makes of all these alternatives.

- Because this amounts to specifying *conditional* probabilities, the model is often called **conditional multinomial logit**.

- The (conditional) multinomial logit's defining feature is the following expression for the probability of all alternatives.

$$p_{ji} \equiv \mathbb{P}\left(Y_{ji} = 1 \middle| \boldsymbol{x}_{1i}, \dots, \boldsymbol{x}_{Ji}\right) = \frac{\exp\left(\boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_{ki}^{\mathrm{T}} \boldsymbol{\beta}\right)}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)$ is a parameter vector of interest.

# The multinomial logit model (3/9)

- Note that if $\boldsymbol{x}_{ji}$ were constant across the $J$ alternatives, that is $\boldsymbol{x}_{1i} = \boldsymbol{x}_{2i} = \cdots = \boldsymbol{x}_{Ji} = \boldsymbol{x}_i$, this model would be moot: all the $J$ choices would be equally likely.

- However, in this case one can re-formulate the model as:

$$p_{ji} \equiv \mathbb{P}\left(Y_{ji} = 1 \middle| \boldsymbol{x}_i\right) = \frac{\exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_k\right)}$$

  where $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jK})$ is one out of $J$ **alternative-specific** parameter vectors of interest.

- The inability to estimate $J$ alternative-specific parameters if $\boldsymbol{x}_{ji}$ is not constant over $j$ is an identification problem!

- Most typically, $\boldsymbol{x}_{ji}$ features both alternative-specific as well as "constant" characteristics. The elements of $\boldsymbol{\beta}_j$ associated with the former are constrained constant across alternatives.

# The multinomial logit model (4/9)

These different levels of variation for the observed characteristics $\boldsymbol{x}_{ji}$ and for the parameters $\boldsymbol{\beta}_j$ led to a use of language that may appear confusing. Many researchers call:

- a plain **multinomial logit** a model that features fixed $\boldsymbol{x}_i$ and varying $\boldsymbol{\beta}_j$;
- an actual **conditional multinomial logit** a model that on the contrary features varying $\boldsymbol{x}_{ji}$ and fixed $\boldsymbol{\beta}$;
- a **mixed multinomial logit** a model that "mixes" both.

This specific use of terminology may appear rather confusing to econometricians, who are typically accustomed to call "mixed" a multinomial logit with *random parameters* (more on this later).

For simplicity, the following treatment sticks to the "conditional multinomial logit" with varying $\boldsymbol{x}_{ji}$ and fixed $\boldsymbol{\beta}$.

# The multinomial logit model (5/9)

Make the following observations.

- One can always reformulate an alternative-invariant variable $X_i$ as a vector of length $J$: $\boldsymbol{x}^*_{ji} = (D_{1ji}X_i, \dots, D_{Jji}X_i)$; with $D_{\ell ji} = 1$ if $\ell = j$ and $D_{\ell ji} = 0$ otherwise, for $\ell = 1, \dots, J$.

- Hence, the $J$ parameters associated with $\boldsymbol{x}^*_{ji}$ correspond to alternative-specific parameters.

- If $\boldsymbol{x}_{ji}$ contains a "constant" vector that is thus dummified, its parameters are interpreted as the *realization probabilities* conditional on all other $\boldsymbol{x}_{ji}$'s being set at zero.

Although the "conditional multinomial logit" is more general, for the sake of practical implementation and estimates interpretation a researcher must always pay attention to the level of variation of the observable characteristics $\boldsymbol{x}_{ji}$'s.

# The multinomial logit model (6/9)

Like all LDV models, the multinomial logit admits a structural interpretation in terms of **latent variables**. Let:

$$V_{ji} = \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_{ji}$$

be the **utility** associated by observation $i$ to the $j$-th alternative. Here $\varepsilon_{ji}$ is a **random** component of the utility $V_{ji}$. It is assumed that alternative $j$ is "chosen" by observation $i$ if it is the one that delivers the highest utility.

$$Y_{ji} = 1 \;\Leftrightarrow\; V_{ji} = \max\left\{V_{1i}, \dots, V_{Ji}\right\}$$

Furthermore, if $\varepsilon_{ji}$ is **i.i.d.** with

$$\varepsilon_{ji} \sim \text{Gumbel}\left(0, 1\right)$$

that is, the random component follows the Gumbel distribution with standard parameters, then the realization probabilities take the multinomial logit form, as it is shown next.

# The multinomial logit model (7/9)

$$
\begin{aligned}
p_{ji} &= \mathbb{P}\left(\bigcup_{k\neq j}\{V_{ji} \geq V_{ki}\}\right) \\
&= \mathbb{P}\left(\bigcup_{k\neq j}\left\{\varepsilon_{ki} \leq \varepsilon_{ji} + (\boldsymbol{x}_{ji} - \boldsymbol{x}_{ki})^{\mathrm{T}}\,\boldsymbol{\beta}\right\}\right) \\
&= \int_{-\infty}^{\infty} \prod_{k\neq j} \exp\left(-\exp\left(-\varepsilon_{ji} - (\boldsymbol{x}_{ji} - \boldsymbol{x}_{ki})^{\mathrm{T}}\,\boldsymbol{\beta}\right)\right) \frac{\exp\left(-\varepsilon_{ji}\right)}{\exp\left(\exp\left(-\varepsilon_{ji}\right)\right)} d\varepsilon_{ji} \\
&= \int_{\infty}^{0} -\prod_{k\neq j} \exp\left(-u\exp\left((\boldsymbol{x}_{ki} - \boldsymbol{x}_{ji})^{\mathrm{T}}\,\boldsymbol{\beta}\right)\right) \frac{1}{\exp\left(u\right)} du \qquad \left[u = \exp\left(-\varepsilon_{ji}\right)\right] \\
&= \int_{0}^{\infty} \exp\left(-u\left[1 + \sum_{k\neq j}\exp\left((\boldsymbol{x}_{ki} - \boldsymbol{x}_{ji})^{\mathrm{T}}\,\boldsymbol{\beta}\right)\right]\right) du \\
&= \frac{1}{1 + \sum_{k\neq j}\exp\left((\boldsymbol{x}_{ki} - \boldsymbol{x}_{ji})^{\mathrm{T}}\,\boldsymbol{\beta}\right)} \\
&= \frac{\exp\left(\boldsymbol{x}_{ji}^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{k=1}^{J}\exp\left(\boldsymbol{x}_{ki}^{\mathrm{T}}\boldsymbol{\beta}\right)}
\end{aligned}
$$

# The multinomial logit model (8/9)

- At first the Gumbel assumption might seem rather arbitrary. Note though that for $j, k = 1, \ldots, J$:

$$(V_{ji} - V_{ki}) - (\boldsymbol{x}_{ji} - \boldsymbol{x}_{ki})^{\mathrm{T}} \boldsymbol{\beta} = \varepsilon_{ji} - \varepsilon_{ki} \sim \text{Logistic}\,(0, 1)$$

  the difference between any two random components follows the standard **logistic** distribution (Observation 14, Lecture 3) which can be thought as a more natural choice.

- If the scale parameter is **unrestricted**: $\varepsilon_{ji} \sim \text{Gumbel}\,(0, \sigma)$, the alternative-specific probabilities are hardly changed:

$$p_{ji} \equiv \mathbb{P}\left( Y_{ji} = 1 \middle| \boldsymbol{x}_{1i}, \ldots, \boldsymbol{x}_{Ji} \right) = \frac{\exp\left( \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} / \sigma \right)}{\sum_{k=1}^{J} \exp\left( \boldsymbol{x}_{ki}^{\mathrm{T}} \boldsymbol{\beta} / \sigma \right)}$$

  and consequently $\boldsymbol{\beta}$ and $\sigma$ are not separately identified. This motivates the normalization $\sigma = 1$.

# The multinomial logit model (9/9)

How to interpret the model's coefficients $\boldsymbol{\beta}$?

- They allow to calculate the **marginal effects** of changes in $\boldsymbol{x}_{ji}$ on the realization probability of each alternative.

$$\frac{\partial p_{ji}}{\partial \boldsymbol{x}_{ki}} = p_{ji} \left( \mathbb{1} \left[ j = k \right] - p_{ki} \right) \boldsymbol{\beta}$$

where $p_{ki}$ is understood as a function of $(\boldsymbol{x}_{1i}, \dots, \boldsymbol{x}_{Ji})$ for all $k = 1, \dots, J$. Similarly to simpler logit and probit models, such marginal effects must be computed and/or averaged at specific realizations of $(\boldsymbol{x}_{1i}, \dots, \boldsymbol{x}_{Ji})$.

- Under the structural interpretation of the model, they also bear an interpretation in terms of **marginal utilities**.

$$\frac{\partial V_{ji}}{\partial \boldsymbol{x}_{ji}} = \frac{\partial \left( \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_{ji} \right)}{\partial \boldsymbol{x}_{ji}} = \boldsymbol{\beta}$$

# Estimation of the multinomial logit model (1/4)

- The likelihood function of this model is:

$$\mathcal{L}\left(\boldsymbol{\beta}\,\middle|\,\{\mathbf{y}_i;\mathbf{x}_{1i},\ldots,\mathbf{x}_{Ji}\}_{i=1}^{N}\right)=\prod_{i=1}^{N}\prod_{j=1}^{J}p_{ji}^{y_{ji}}$$

where $p_{ji}$ is implicitly treated as a function of the *realizations* $(\mathbf{x}_{1i},\ldots,\mathbf{x}_{Ji})$ and $y_{ji}$ is the realization of $Y_{ji}$ for $j=1,\ldots,J$ stacked in an observation-specific vector $\mathbf{y}_i=(y_{1i},\ldots,y_{Ji})$. Recall that $\sum_{j=1}^{J}y_{ji}=\sum_{j=1}^{J}Y_{ji}=1$.

- Thus, the log-likelihood function is as follows.

$$\log\mathcal{L}\left(\boldsymbol{\beta}\,\middle|\,\{\mathbf{y}_i;\mathbf{x}_{1i},\ldots,\mathbf{x}_{Ji}\}_{i=1}^{N}\right)=\sum_{i=1}^{N}\sum_{j=1}^{J}y_{ji}\log\left(p_{ji}\right)$$

- Define the following quantity.

$$\bar{\mathbf{x}}_i=\sum_{j=1}^{J}p_{ji}\mathbf{x}_{ji}=\frac{\sum_{j=1}^{J}\exp\left(\mathbf{x}_{ji}^{\mathrm{T}}\boldsymbol{\beta}\right)\mathbf{x}_{ji}}{\sum_{k=1}^{J}\exp\left(\mathbf{x}_{ki}^{\mathrm{T}}\boldsymbol{\beta}\right)}$$

# Estimation of the multinomial logit model (2/4)

- The First Order Conditions are as follows:

$$\begin{aligned}\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta} \,\middle|\, \{\mathbf{y}_i; \mathbf{x}_{1i}, \dots, \mathbf{x}_{Ji}\}_{i=1}^{N}\right)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} \sum_{j=1}^{J} \frac{y_{ji}}{p_{ji}} \frac{\partial p_{ji}}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ji} \left(\mathbf{x}_{ji} - \bar{\mathbf{x}}_i\right) \\ &= \mathbf{0}\end{aligned}$$

  since $\partial p_{ji} / \partial \boldsymbol{\beta} = p_{ji} \left(\mathbf{x}_{ji} - \bar{\mathbf{x}}_i\right)$ as it is possible to verify. The parameters are "buried" within $\bar{\mathbf{x}}_i$. Since there is no closed form solution, the estimates are obtained numerically.

- Some further algebra yields the Hessian of the log-likelihood function, which may be useful for inference purposes.

$$\frac{\partial \log \mathcal{L}\left(\left.\boldsymbol{\beta}\right| \cdot\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = -\sum_{i=1}^{N} \sum_{j=1}^{J} y_{ji} \left(\mathbf{x}_{ji} - \bar{\mathbf{x}}_i\right) \left(\mathbf{x}_{ji} - \bar{\mathbf{x}}_i\right)^{\mathrm{T}}$$

# Estimation of the multinomial logit model (3/4)

- These equations differ for models that feature *only* constant characteristics $\mathbf{x}_i$ and varying parameters $\boldsymbol{\beta}_j$.

- In particular, the First Order Conditions for $\boldsymbol{\beta}_j$, $j = 1, \dots, J$ are as follows (yet again without closed form solution).

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta} \,\middle|\, \{\mathbf{y}_i; \mathbf{x}_{1i}, \dots, \mathbf{x}_{Ji}\}_{i=1}^{N}\right)}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{N} \left(y_{ji} - p_{ji}\right) \mathbf{x}_i = \mathbf{0}$$

- Recall that $p_{ji}$ is a function of *all* the parameters! Note that for $k = 1, \dots, J$ it is $\partial p_{ji} / \partial \boldsymbol{\beta}_k = p_{ji} \left(\mathbb{1}\left[j = k\right] - p_{ki}\right) \mathbf{x}_i$.

- The Hessian of the log-likelihood function instead has blocks with the following form, for $j, k = 1, \dots, J$.

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta} \middle| \cdot\right)}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k^{\mathrm{T}}} = -\sum_{i=1}^{N} \sum_{j=1}^{J} p_{ji} \left(\mathbb{1}\left[j = k\right] - p_{ki}\right) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$$

# Estimation of the multinomial logit model (4/4)

- Sometimes the alternatives **available** to single observations are not the same. Denote the choice set for observation $i$ as $\mathcal{C}_i$. In such a case, the multinomial logit is still well defined. The likelihood function changes as:

$$\mathcal{L}\left(\boldsymbol{\beta}\,|\cdot\right) = \prod_{i=1}^{N} \prod_{j \in \mathcal{C}_i} \left[ \frac{\exp\left(\mathbf{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\sum_{k \in \mathcal{C}_i} \exp\left(\mathbf{x}_{ki}^{\mathrm{T}} \boldsymbol{\beta}\right)} \right]^{y_{ji}}$$

and estimation proceeds as in the standard case.

- In other cases, the choice set is so large as to make estimation impractical. McFadden (1978) showed that one can still get consistent estimates with a likelihood function like:

$$\mathcal{L}\left(\boldsymbol{\beta}\,|\cdot\right) = \prod_{i=1}^{N} \prod_{j \in \mathcal{K}_i} \left[ \frac{\exp\left(\mathbf{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\sum_{k \in \mathcal{K}_i} \exp\left(\mathbf{x}_{ki}^{\mathrm{T}} \boldsymbol{\beta}\right)} \right]^{y_{ji}}$$

where $\mathcal{K}_i$ is a **random subset** of alternatives associated to $i$ that is selected so as to include $i$'s realized outcome $Y_i$.

# Independence of irrelevance alternatives

- The fundamental property of the multinomial logit model is the **independence of irrelevant alternatives** (IIA) that is featured by realization probabilities. In short:

$$\frac{p_{ji}}{p_{ki}} = \exp\left(\left(\boldsymbol{x}_{ji} - \boldsymbol{x}_{ki}\right)^{\mathrm{T}} \boldsymbol{\beta}\right)$$

  for any $j, k = 1, \dots, J$. Thus, for every observation pair the ratio between the realization probabilities of two alternatives is constant, and unaffected by other alternatives $\ell$ and their characteristics $\boldsymbol{x}_{\ell i}$.

- This may be **unrealistic** in many settings, as illustrated by the "red bus, blue bus" famous example (McFadden, 1974). Suppose that one is studying the determinants of choosing a "red bus" ($j$) against a car ($k$) as means of transportation. A two-outcomes model would return some ratio $p_{ji}/p_{ki}$. Then a "blue bus" ($\ell$) is introduced. Realistically, $p_{ki}$ should not vary, but IIA must be violated for $p_{ji} + p_{ki} + p_{\ell i} = 1$ to hold.

# Limitations of the multinomial logit

The multinomial logit is extremely popular: it is based on simple expressions, it is easy enough to estimate, and it can be motivated in ways other than the Gumbel-distributed latent shocks $\varepsilon_{ji}$.

However, to an econometrician's eye it also features three major **limitations**.

1. The non-random "tastes" of individuals/observations, that is the $\boldsymbol{\beta}$ parameters, are unrealistically **homogeneous**.

2. As argued, the **substitution patterns** between alternatives are often unrealistic because of IIA.

3. The model is generally not well suited to data that feature autocorrelation in time or spatial correlation.

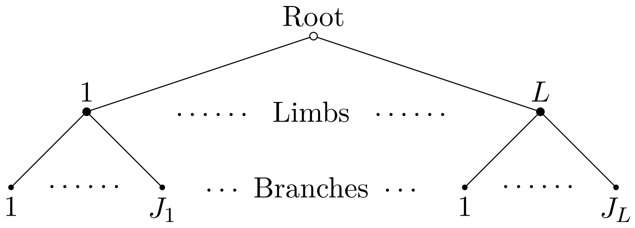
The next three multinomial models aim at addressing limitations 1 and 2, while the third one is outside the scope of this review.

# The nested logit model (1/3)

- It was McFadden himself (1978) who proposed an extension of the multinomial logit that addresses issues of IIA.

- In the **nested logit** the alternative outcomes have a "tree-like" **hierarchical structure**, with "limbs" and "branches." The $J$ alternatives are thought as "branches" grouped across $L$ "limbs;" each limb has $J_\ell$ branches with $\sum_{\ell=1}^{L} J_\ell = J$.

- Thus, alternatives are denoted by $Y_{\ell j i}$ with $j = 1, \ldots, J_\ell$ and $\ell = 1, \ldots, L$. They can be represented as follows.



# The nested logit model (2/3)

- Let there be $H$ **limb-specific** observable characteristics $\boldsymbol{z}_{\ell i}$, and $K_l$ **branch-specific** $\boldsymbol{x}_{\ell j i}$ characteristics for $\ell = 1, \dots, L$.

- In the nested logit model, the realization probabilities are:

$$p_{\ell j i} = \underbrace{\frac{\exp\left(\boldsymbol{z}_{\ell i}^{\mathrm{T}} \boldsymbol{\alpha} + \rho_\ell I_\ell\right)}{\sum_{h=1}^{L} \exp\left(\boldsymbol{z}_{hi}^{\mathrm{T}} \boldsymbol{\alpha} + \rho_h I_h\right)}}_{= p_{\ell i} \equiv \mathbb{P}(Y_{\ell i} = 1 | \cdot)} \underbrace{\frac{\exp\left(\boldsymbol{x}_{\ell j i}^{\mathrm{T}} \boldsymbol{\beta}_\ell / \rho_\ell\right)}{\sum_{k=1}^{J_\ell} \exp\left(\boldsymbol{x}_{\ell k i}^{\mathrm{T}} \boldsymbol{\beta}_\ell / \rho_\ell\right)}}_{= p_{ji|\ell} \equiv \mathbb{P}\left(Y_{\ell j i} = 1 \middle| Y_{\ell i} = 1, \cdot\right)}$$

  where $Y_{\ell i} = 1$ denotes selection of the $\ell$-th limb; whereas $I_\ell$ is defined as follows for $\ell = 1, \dots, L$.

$$I_\ell = \log\left(\sum_{k}^{J_\ell} \exp\left(\boldsymbol{x}_{\ell k i}^{\mathrm{T}} \boldsymbol{\beta}_\ell / \rho_\ell\right)\right)$$

- The model's parameters are $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_L, \rho_1, \dots, \rho_L)$. The nested structure operates through the $\boldsymbol{\rho} = (\rho_1, \dots, \rho_L)$ parameters: if $\boldsymbol{\rho} = \boldsymbol{\iota}$, this is a standard multinomial logit.

# The nested logit model (3/3)

- The latent variable representation of the nested logit is:

$$V_{\ell ji} = \boldsymbol{z}_{\ell i}^{\mathrm{T}} \boldsymbol{\alpha} + \boldsymbol{x}_{\ell ji}^{\mathrm{T}} \boldsymbol{\beta}_{\ell} + \varepsilon_{\ell ji}$$

  where $\varepsilon_{\ell ji}$ follows a joint GEV distribution that features $\boldsymbol{\rho}$ as measures of within-limb anti-correlation (McFadden, 1978).

- The likelihood function is most succinctly expressed in terms of the various realization probabilities involved:

$$\mathcal{L}\left(\boldsymbol{\theta} \middle| \cdot\right) = \prod_{i=1}^{N} \prod_{\ell=1}^{L} \left( p_{\ell i}^{y_{\ell i}} \prod_{j=1}^{J_{\ell}} p_{ji|\ell}^{y_{\ell ji}} \right)$$

  and so is the log-likelihood function to be *jointly* maximized.

$$\log \mathcal{L}\left(\boldsymbol{\theta} \middle| \cdot\right) = \sum_{i=1}^{N} \sum_{\ell=1}^{L} \left[ y_{\ell i} \log\left(p_{\ell i}\right) + \sum_{j=1}^{J_{\ell}} y_{\ell ji} \log\left(p_{ji|\ell}\right) \right]$$

- For convenience, one can **sequentially** estimate first $I_{\ell}$ and $\boldsymbol{\beta}_{\ell}/\rho_{\ell}$ in branches; and second, $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ in limbs.

# The mixed logit model (1/2)

The **random parameters logit** model, also called **mixed logit** by econometricians, is based upon a representation of the latent random utility that features **heterogeneous** "tastes."

$$V_{ji} = \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_{ji}$$

The key feature is that the parameters $\boldsymbol{\beta}_i$ are observation-specific and treated as **random**, typically jointly normal.

$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$$

For $\boldsymbol{u}_i = \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\beta}_i - \boldsymbol{\beta}\right)$, the model can be re-written as follows.

$$\begin{aligned} V_{ji} &= \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} + \upsilon_{ji} \\ \upsilon_{ji} &= \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{u}_i + \varepsilon_{ji} \end{aligned}$$

The shock $\varepsilon_{ji}$ is still assumed to be standard Gumbel distributed, and to be independent across observations and alternatives.

# The mixed logit model (2/2)

- Notice that for $j \neq k$, $\mathbb{C}\mathrm{ov}\left(v_{ji}, v_{ki} \middle| \boldsymbol{x}_{ji}, \boldsymbol{x}_{ki}\right) = \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{x}_{ki}$: this introduces correlation between alternatives, defying IIA!

- The realization probabilities are as follows:

$$p_{ji} = \int_{\mathbb{R}^K} \frac{\exp\left(\boldsymbol{x}_{ji}^{\mathrm{T}}\left(\boldsymbol{\beta} + \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u}_i\right)\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_{ki}^{\mathrm{T}}\left(\boldsymbol{\beta} + \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u}_i\right)\right)} \phi\left(\mathbf{u}_i\right) \mathrm{d}\mathbf{u}_i$$

  where $\phi\left(\cdot\right)$ is the p.d.f. of the standard multivariate normal distribution. This integral has no closed form solution.

- This model is typically estimated by MSL through a sample of $S$ simulation draws $\left\{\mathbf{u}_s\right\}_{s=1}^{S}$; $\boldsymbol{\Sigma}$ is often restricted *ex ante.*

$$\left(\widehat{\boldsymbol{\beta}}_{MSL}, \widehat{\boldsymbol{\Sigma}}_{MSL}\right) =$$

$$= \underset{(\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\arg\max} \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ji} \log \left[\frac{1}{S} \sum_{s=1}^{S} \frac{\exp\left(\boldsymbol{x}_{ji}^{\mathrm{T}}\left(\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\frac{1}{2}} \mathbf{u}_s\right)\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_{ki}^{\mathrm{T}}\left(\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\frac{1}{2}} \mathbf{u}_s\right)\right)}\right]$$

# The multinomial probit model

The **multinomial probit** model is also based on the standard representation of the latent random utility:

$$V_{ji} = \boldsymbol{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_{ji}$$

but the random component $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \ldots, \varepsilon_{Ji})$ is jointly normally distributed: a more natural choice than GEV distributions.

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

Observe that if $\boldsymbol{\Sigma}$ is non-diagonal, the alternatives are correlated, like in the mixed logit. Moreover, IIA does not hold in this model.

For all its advantages, this model features quite a major problem: its realization probabilities can be very difficult to compute.

$$p_{ji} = \int_{\mathbb{R}^K} \prod_{k \neq j} \mathbb{1}\left(\mathbf{x}_{ji}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_{ji} \geq \mathbf{x}_{ki}^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_{ki}\right) \frac{1}{|\boldsymbol{\Sigma}|} \phi\left(\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\varepsilon}_i\right) \mathrm{d}\boldsymbol{\varepsilon}_i$$

Even simulation methods struggle to estimate this model quickly. Furthermore, identification requires careful restrictions on $\boldsymbol{\Sigma}$.

# Ordered multinomial models (1/2)

- What if the alternatives are naturally **ordered** (for example, $Y_i$ represents a ladder of a product's qualities)? The models reviewed thus far are unsuited to address the problem.

- The solution are the **ordered multinomial models** that posit a latent variable representation

$$Y_i^* = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$$

  that implies selection of the $j$-th alternative if it "passes" a certain associated **threshold** $\alpha_{j-1}$, for $j = 1, \ldots, J$.

$$Y_i = j \Leftrightarrow \alpha_{j-1} < Y_i^* \leq \alpha_j$$

- There are $J$ thresholds $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ that are treated as **parameters** to be estimated, alongside $\boldsymbol{\beta}$.

- Note that the observable characteristics $\boldsymbol{x}_i$ only vary at the level of units of observation here.

# Ordered multinomial models (2/2)

- Let $F_{\varepsilon|\boldsymbol{x}}\left(\varepsilon_i|\, \boldsymbol{x}_i\right)$ be the c.d.f. for $\varepsilon_i$ given $\boldsymbol{x}_i$: for example, the standard normal $\Phi\left(\cdot\right)$ for the **ordered probit**, the standard logistic $\Lambda\left(\cdot\right)$ for the **ordered logit**, or others. Then:

$$\begin{aligned} p_{ji} &\equiv \mathbb{P}\left(Y_i = j|\, \boldsymbol{x}_i\right) \\ &= \mathbb{P}\left(\alpha_{j-1} < Y_i^* \leq \alpha_j|\, \boldsymbol{x}_i\right) \\ &= \mathbb{P}\left(\alpha_{j-1} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} < \varepsilon_i \leq \alpha_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\middle|\, \boldsymbol{x}_i\right) \\ &= F_{\varepsilon|\boldsymbol{x}}\left(\alpha_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\middle|\, \boldsymbol{x}_i\right) - F_{\varepsilon|\boldsymbol{x}}\left(\alpha_{j-1} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\middle|\, \boldsymbol{x}_i\right) \end{aligned}$$

- ...which enables MLE via a familiar log-likelihood function.

$$\log \mathcal{L}\left(\boldsymbol{\beta}\,\middle|\left\{y_i; \mathbf{x}_i\right\}_{i=1}^N\right) = \sum_{i=1}^N \sum_{j=1}^J y_i \log\left(p_{ji}\right)$$

- The **marginal effects** obtain from the p.d.f.s $f_{\varepsilon|\boldsymbol{x}}\left(\varepsilon_i|\, \boldsymbol{x}_i\right)$.

$$\frac{p_{ji}}{\partial \boldsymbol{x}_i} = \left[f_{\varepsilon|\boldsymbol{x}}\left(\alpha_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\middle|\, \boldsymbol{x}_i\right) - f_{\varepsilon|\boldsymbol{x}}\left(\alpha_{j-1} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\middle|\, \boldsymbol{x}_i\right)\right]\boldsymbol{\beta}$$

# Instrumental variables for multinomial models

An alternative estimation approach for **unordered** multinomial models is based on **moment conditions** of the following form:

$$\mathbb{E}\left[Y_{ji} - p_{ji}\left(\boldsymbol{x}_{ji};\boldsymbol{\theta}\right)\middle|\,\boldsymbol{z}_{ji}\right] = \mathbb{E}\left[\boldsymbol{z}_{ji}\left(Y_{ji} - p_{ji}\left(\boldsymbol{x}_{ji};\boldsymbol{\theta}\right)\right)\right] = \mathbf{0}$$

for $j = 1, \ldots, J$. These moment conditions feature:

- $p_{ji}\left(\boldsymbol{x}_{ji};\boldsymbol{\theta}\right)$: the realization probability for the $j$-th choice, as a function of the characteristics $\boldsymbol{x}_{ji}$ and some parameters $\boldsymbol{\theta}$; for example, this can be a multinomial probit *simulated* $p_{ji}$;

- $\boldsymbol{z}_{ji}$: a vector of **instruments**; possibly it is $\boldsymbol{z}_{ji} = \boldsymbol{x}_{ji}$, more generally it includes a different/larger set of shifters.

If one suspects that the latent variable error $\varepsilon_{ji}$ correlates with $\boldsymbol{x}_{ji}$ and $p_{ji}\left(\boldsymbol{x}_{ji};\boldsymbol{\theta}\right)$ is correctly specified, estimating $\boldsymbol{\theta}$ using these moments in a (G)MM/MSM framework can be the sound choice. However, this is generally less efficient than MLE.

# Review of panel models for discrete outcomes

What follows is an overview of selected approaches to unobserved heterogeneity in LDV models, when **panel data** are available to the econometrician. The models outlined next are:

- the **conditional logit model** for binary outcomes;
- the **dynamic logit model** with fixed effects;
- the **fixed effects multinomial logit model**;
- the **random effects model probit model**;
- the **correlated random effects models**.

Emphasis is placed on the statistical interpretation of each model.

# The incidental parameter problem

Practitioners of econometrics are accustomed to a fairly seamless implementation of fixed or random effects in *linear* models. With a hindsight this should be a surprise, because in general, a model written as:

$$Y_{it} = h\left(\boldsymbol{\alpha}_i + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right) + \varepsilon_{it}$$

where $h(\cdot)$ is some arbitrary **non-linear** function, should pose econometric challenges if the longitudinal dimension of the panel $T$ is small (as it is extremely common in practice).

In fact, estimation of the individual effects $\boldsymbol{\alpha}_i$ is **inconsistent** with small $T$, and this also makes the estimates of $\boldsymbol{\beta}$ inconsistent via the M-Estimation First Order Conditions. This is known as the **incidental parameter problem**.

This does not occur in linear models thanks to the Frisch-Waugh-Lovell Theorem (Lecture 7). This is all but a **coincidence**.

# Logit and probit with fixed effects

Adding fixed effects $\alpha_i$ to the logit or the probit model in presence of panel data gives, respectively:

$$\mathbb{P}\left(Y_{it} = 1 \mid \boldsymbol{x}_{it}\right) = \Lambda\left(\alpha_i + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)$$
$$\mathbb{P}\left(Y_{it} = 1 \mid \boldsymbol{x}_{it}\right) = \Phi\left(\alpha_i + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)$$

where $\Lambda(\cdot)$ and $\Phi(\cdot)$ are the c.d.f.s of the standard **logistic** and standard **normal** distributions, respectively.

There is no obvious solution to the incidental parameter problem in the probit's case. However, the logit can be **transformed** so as to remove the fixed effects $\alpha_i$. This is yet another coincidence, this time due to the logistic distribution's functional form.

The transformation obtains by **conditioning** on $\sum_{t=1}^{T} Y_{it}$, which is a **sufficient statistic** for $\alpha_i$.

# The conditional fixed effects logit (1/4)

In the panel data logit model, write the conditional density of all the outcomes $\boldsymbol{y}_i = (Y_{i1}, \ldots, Y_{iT})$ of observation $i$.

$$f_{\boldsymbol{y}_i}\left(\mathbf{y}_i \middle| \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\right) =$$

$$= \prod_{t=1}^{T} \left( \frac{\exp\left(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)}{1 + \exp\left(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)} \right)^{y_{it}} \left( \frac{1}{1 + \exp\left(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)} \right)^{1-y_{it}}$$

$$= \frac{\exp\left(\alpha_i \sum_{t=1}^{T} y_{it}\right) \exp\left(\sum_{t=1}^{T} y_{it} \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\prod_{t=1}^{T} \left[1 + \exp\left(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)\right]}$$

Note that this result can be generalized for any arbitrary vector of "hypothetical" individual-level outcomes $\boldsymbol{v}_i = (V_{i1}, \ldots, V_{iT})$.

$$f_{\boldsymbol{v}_i}\left(\mathbf{v}_i \middle| \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\right) = \frac{\exp\left(\alpha_i \sum_{t=1}^{T} v_{it}\right) \exp\left(\sum_{t=1}^{T} v_{it} \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\prod_{t=1}^{T} \left[1 + \exp\left(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)\right]}$$

# The conditional fixed effects logit (2/4)

The **conditional fixed effects logit** model (not to be confused with the multinomial "conditional" logit) is constructed by noting (Chamberlain, 1980) that:

$$f_{\boldsymbol{y}_i}\left(\mathbf{y}_i \middle| \sum_{t=1}^{T} y_{it}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\right) = \frac{f_{\boldsymbol{y}_i}\left(\mathbf{y}_i \middle| \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\right)}{f_{\boldsymbol{y}_i}\left(\sum_{t=1}^{T} y_{it} \middle| \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\right)}$$
$$= \frac{\exp\left(\sum_{t=1}^{T} y_{it}\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{\mathbf{v}_i \in \mathbb{V}_i} \exp\left(\sum_{t=1}^{T} v_{it}\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)}$$

where $\mathbb{V}_i \equiv \left\{\mathbf{v}_i : \sum_{i=1}^{T} (v_{it} - y_{it}) = 0\right\}$ is the set of all the possible configurations of the individual binary outcomes that yield the same count of "successes" for $i$ as the one actually observed.

This derivation shows that $\sum_{t=1}^{T} Y_{it}$ is a sufficient statistic for $\alpha_i$ (see Lecture 4). Intuitively, this is because $\alpha_i$ is a measure of the average propensity to obtain a Bernoulli "success" $Y_{it} = 1$.

# The conditional fixed effects logit (3/4)

The likelihood function associated with this model is as follows.

$$\mathcal{L}\left(\boldsymbol{\beta}\,\middle|\,\left\{\sum_{t=1}^{T} y_{it}; \mathbf{y}_i; \mathbf{X}_i\right\}_{i=1}^{N}\right) = \prod_{i=1}^{N} \frac{\exp\left(\sum_{t=1}^{T} y_{it}\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{\mathbf{v}_i \in \mathbb{V}_i} \exp\left(\sum_{t=1}^{T} v_{it}\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)}$$

In this expression, $\mathbf{y}_i$ and $\mathbf{X}_i$ represent individual observations $(y_{it}, \mathbf{x}_{it}^{\mathrm{T}})$ stacked over the panel. Some observations are due.

- The effective unit of observation is the panel unit $i$.
- The observations $i$ for which the set $\mathbb{V}_i$ has dimension 1 do not contribute to the likelihood function.
- This occurs for example if $\sum_{t=1}^{T} Y_{it} = 0$ or $\sum_{t=1}^{T} Y_{it} = T$.
- Estimation requires specifying the set $\mathbb{V}_i$ for $t = 1, \dots, T-1$. This can be cumbersome for moderate values of $T$.
- Estimation of this model is otherwise standard.

# The conditional fixed effects logit (4/4)

There are two more important observations to make.

- Similarly as in linear models with fixed effects, identification follows from the **time variation** in the regressors $\boldsymbol{x}_{it}$. This is best exemplified by the simple case with $T = 2$, where:

$$\mathbb{P}\left(Y_{i1} = 0 \cup Y_{i2} = 1 \middle| Y_{i1} + Y_{i2} = 1\right) = \frac{\exp\left(\mathbf{x}_{i2}^{\mathrm{T}} \boldsymbol{\beta}\right)}{\exp\left(\mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta}\right) + \exp\left(\mathbf{x}_{i2}^{\mathrm{T}} \boldsymbol{\beta}\right)}$$
$$= \frac{\exp\left(\left(\mathbf{x}_{i2} - \mathbf{x}_{i1}\right)^{\mathrm{T}} \boldsymbol{\beta}\right)}{1 + \exp\left(\left(\mathbf{x}_{i2} - \mathbf{x}_{i1}\right)^{\mathrm{T}} \boldsymbol{\beta}\right)}$$

  and symmetrically if $Y_{i1} = 1$ and $Y_{i2} = 0$.

- The elimination of the fixed effects prevents the calculation of standard **marginal effects** of $\boldsymbol{\beta}$ on $\Lambda\left(\alpha_i + \boldsymbol{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)$. Still, it is possible to evaluate the marginal effect of changes in the *time variation* of the regressors, e.g. in $\left(\mathbf{x}_{i2} - \mathbf{x}_{i1}\right)$ for $T = 2$.

# Adding a lagged dependent variable

Suppose interest falls on the following model:

$$\mathbb{P}\left(Y_{it}=1\,\middle|\,\boldsymbol{x}_{it};Y_{i(t-1)}\right)=\Lambda\left(\alpha_i+\boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}+\gamma Y_{i(t-1)}\right)$$

where, similarly to dynamic linear models, it is empirically salient to disentangle the fixed effect $\alpha_i$ (*unobserved heterogeneity*) from the effect of past outcomes $Y_{i(t-1)}$ (*state dependence*).

When $\boldsymbol{\beta}=\mathbf{0}$, a derivation similar to the previous one applies.

$$f_{\boldsymbol{y}_i}\left(\mathbf{y}_i\,\middle|\,y_{i1},\sum_{t=1}^{T}y_{it},y_{iT}\right)=\frac{\exp\left(\gamma\sum_{t=2}^{T-1}y_{it}y_{i(t-1)}\right)}{\sum_{\mathbf{w}_i\in\mathbb{W}_i}\exp\left(\gamma\sum_{t=2}^{T-1}w_{it}w_{i(t-1)}\right)}$$

Here, $\mathbb{W}_i\equiv\left\{\mathbf{w}_i:\sum_{i=1}^{T}\left(w_{it}-y_{it}\right)=0,w_{i1}=y_{i1},w_{iT}=y_{iT}\right\}$ also restricts the first and last "pseudo-outcomes" to match the real ones. For this reason, this dynamic logit requires $T\geq 4$. When $\boldsymbol{\beta}\neq\mathbf{0}$, more complications arise (Honoré and Kyriziadou, 2000).

# The multinomial logit with fixed effects (1/2)

This logic also extends to the multinomial logit:

$$p_{jit} \equiv \mathbb{P}\left(Y_{jit} = 1 \middle| \boldsymbol{x}_{1it}, \ldots, \boldsymbol{x}_{Jit}\right) = \frac{\exp\left(\alpha_{ij} + \boldsymbol{x}_{jit}^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{k=1}^{J}\exp\left(\alpha_{ik} + \boldsymbol{x}_{kit}^{\mathrm{T}}\boldsymbol{\beta}\right)}$$

where $\alpha_{ik}$ for $k = 1, \ldots, J$ can be interpreted as the tendency of individual $i$ to make the $k$-th choice over the $T$ periods.

In this case, the sufficient statistic approach gives:

$$f_{\mathbf{Y}_i}\left(\mathbf{Y}_i \middle| \mathbf{Y}_i \boldsymbol{\iota}; \mathbf{X}_{1i}, \ldots, \mathbf{X}_{Ji}\right) = \frac{\exp\left(\sum_{t=1}^{T}\sum_{j=1}^{J} y_{jit}\mathbf{x}_{jit}^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sum_{\mathbf{u}_i \in \mathbb{U}_i}\exp\left(\sum_{t=1}^{T}\sum_{j=1}^{J} u_{jit}\mathbf{x}_{jit}^{\mathrm{T}}\boldsymbol{\beta}\right)}$$

where here $\mathbf{u}_{it} = (u_{1it}, \ldots, u_{Jit})$ is a vector of pseudo-outcomes for observation $i$ at times $t$, matrices $\mathbf{Y}_i$, $\mathbf{U}_i$ and $\mathbf{X}_{ji}$ obtain by stacking $\mathbf{y}_{it}$, $\mathbf{u}_{it}$ and $\mathbf{x}_{jit}$ horizontally over $t$ (for $j = 1, \ldots, J$), and $\mathbb{U}_i \equiv \{\mathbf{U}_i : (\mathbf{Y}_i - \mathbf{U}_i)\boldsymbol{\iota} = \mathbf{0}\}$ is the set of all configurations of $\mathbf{U}_i$ that yield, across *all* the $J$ options, the real total count.

# The multinomial logit with fixed effects (2/2)

It is worth making some additional considerations.

- This model is most appropriately called "multinomial logit with fixed effects" as the adjective *conditional* is most often associated with the model's plain cross-sectional version.

- The baseline structure of the multinomial choice problem (in every period an observation makes at least one choice, be it even an outside option) ensures that $\mathbb{U}_i$ is never a singleton.

- However, $\mathbb{U}_i$ may be very difficult to completely characterize for large $J$ and $T$. In this case, one should adopt a strategy to *uniformly sample* from $\mathbb{U}_i$ and construct the denominator of the conditional density of $\mathbf{Y}_i$ accordingly. This is analogous to McFadden's (1978) analysis of the many-alternatives case.

- The model extends to unbalanced panels and heterogeneous choice sets; for *dynamics* see Honoré and Kyriziadou (2000).

# The random effects probit model (1/2)

In the probit case, there is no special "trick" to easily remove $\alpha_i$. The standard approach is thus to treat $\alpha_i$ as a random variable, and to account for its distribution while estimating the model.

Suppose for example that $\alpha_i | \boldsymbol{x}_{it} \sim \mathcal{N}(0, \sigma_\alpha^2)$. Then:

$$\mathbb{P}(Y_{it} = 1 | \boldsymbol{x}_{it}) = \int_{\mathbb{R}} \mathbb{P}(Y_{it} = 1 | \boldsymbol{x}_{it}; \alpha_i) \frac{1}{\sigma_\alpha} \phi\left(\frac{\alpha_i}{\sigma_\alpha}\right) d\alpha_i$$

where $\phi(\cdot)$ is the standard normal density. If $\mathbb{P}(Y_{it} = 1 | \boldsymbol{x}_{it}; \alpha_i)$ proceeds according to the familiar probit form, the full likelihood function is as follows, and it can be optimized numerically.

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma_\alpha^2 \middle| \{y_{i1}, \dots, y_{iT}; \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}\}_{i=1}^N\right) =$$

$$= \prod_{i=1}^N \prod_{t=1}^T \int_{\mathbb{R}} \left[\Phi\left(\alpha_i + \boldsymbol{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)\right]^{y_{it}} \left[1 - \Phi\left(\alpha_i + \boldsymbol{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)\right]^{1-y_{it}} \times$$

$$\times \frac{1}{\sigma_\alpha} \phi\left(\frac{\alpha_i}{\sigma_\alpha}\right) d\alpha_i$$

# The random effects probit model (2/2)

Some observations apply to this **random effects probit** model.

- As in linear models, this approach relies on the random effect $\alpha_i$ being independent of the regressors $\boldsymbol{x}_{it}$. In many practical applications, this can be inappropriate.

- This approach can be extended to the logit, as well as to any parametric non-linear model with fixed effects (even beyond binary outcomes). In some cases, the integral expressing the likelihood function has a closed form.

- Similarly, the approach can be extended to dynamic models with lagged outcomes among the regressors.

- One can specify a *discrete* support for $\alpha_i$ with an *unrestricted* mass function $p_\alpha(\alpha_j) = \pi_j$. This renders the approach akin to a mixture model (see Lecture 17 for a succinct summary of linear mixture models).

# Correlated random effects models

To overcome the assumption about independence between $\alpha_i$ and $\boldsymbol{x}_{it}$, one can specify a full-fledged parametric correlation structure between them. For example, Chamberlain's (1980) version of the **correlated random effects model** posits:

$$\alpha_i | \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT} \sim \mathcal{N}\left(\boldsymbol{x}_{i1}^{\mathrm{T}} \boldsymbol{\pi}_1 + \cdots + \boldsymbol{x}_{iT}^{\mathrm{T}} \boldsymbol{\pi}_T ; \sigma_\alpha^2\right)$$

leading to a more general likelihood function where $(\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_T)$ are parameters to estimate, alongside $\sigma_\alpha^2$.

In applications, the more restricted, easier-to-estimate version by Mundlak (1978) is often preferred: it assumes the following.

$$\alpha_i | \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT} \sim \mathcal{N}\left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_{it}^{\mathrm{T}} \boldsymbol{\pi} ; \sigma_\alpha^2\right)$$

These models enable the computation of **marginal effects** that also embody the *indirect* effect of the regressors $\boldsymbol{x}_{it}$ through $\alpha_i$.

# The dynamic logit model (1/10)

This lecture is concluded by reviewing the **dynamic logit** model as in the original formulation by Rust (1987).

- This is **not** a logit model with a lagged dependent variable.
- This is a model for longitudinal data where individuals take **forward-looking** choices.
- More specifically, **state variables** depend on **past choices**.
- Rust frames it via a famous example: Harold Zurcher (**HZ**), a superintendent for bus maintenance from Madison, WI.
- HZ is faced with a peculiar **optimal stopping** problem of econometric interest: when to replace the bus engines?
- The original model about HZ is reviewed next.

# The dynamic logit model (2/10)

Think of a bus in HZ's depot observed over time $t = 1, 2, \ldots$.

- Let $X_t$ represent **mileage** of the bus: the state variable.
- Let $I_t \in \{0, 1\}$ represent **engine replacement** for this bus: this is an endogenous decision by HZ.

Let $\boldsymbol{\varepsilon}_t = (\varepsilon_{0t}, \varepsilon_{1t})$ and $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_1', \chi)$. HZ's **per-period payoff** is:

$$\pi\left(X_t, I_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}_1\right) = \begin{cases} -c\left(X_t; \boldsymbol{\theta}_1'\right) + \varepsilon_{0t} & \text{if } I_t = 0 \\ \chi - c\left(0; \boldsymbol{\theta}_1'\right) + \varepsilon_{1t} & \text{if } I_t = 1 \end{cases}$$

where here: *i.* $c\left(X_t; \boldsymbol{\theta}_1'\right)$ are *regular* maintenance costs, dependent upon some parameters $\boldsymbol{\theta}_1'$; *ii.* $\chi$ is the *replacement cost* of engines, with $\chi < 0$; *iii.* $\varepsilon_{0t}$ and $\varepsilon_{1t}$ are two payoff shocks that are known to HZ, *but not to the econometrician.*

# The dynamic logit model (3/10)

This would be a simple logit/probit if HZ took "myopic" decisions in every period $t$. However, HZ is forward-looking and maximizes the present value of future payoffs. His **value function** is:

$$\mathcal{V}\left(X_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}\right) = \max_{\{I_\tau\}_{\tau=t}^{\infty}} \mathbb{E}\left[\sum_{\tau=t}^{\infty} \beta^{\tau-t} \pi\left(X_\tau, I_\tau, \boldsymbol{\varepsilon}_\tau; \boldsymbol{\theta}_1\right) \middle| X_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}_2\right]$$

where $\beta \in [0,1]$ is the **discount factor**; $\boldsymbol{\theta}_2$ is the parameter set that governs how **future** $X_\tau$, $\varepsilon_{0\tau}$ and $\varepsilon_{1\tau}$ are determined, whose knowledge is implicit in the expectation; and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

The value function can be represented via a **Bellman equation**:

$$\mathcal{V}\left(X_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}\right) = \max_{I_t \in \{0,1\}} \left[\pi\left(X_t, I_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}_1\right) + \beta \mathcal{E}\mathcal{V}\left(X_t, I_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}\right)\right]$$

where $\mathcal{E}\mathcal{V}\left(\cdot\,; \boldsymbol{\theta}\right)$ is the *continuation value*: that is, a **function** for the expected utility from periods later than $t$, *given* a choice $I_t$.

# The dynamic logit model (4/10)

Specifically, the expected future value is as follows.

$$\mathcal{EV}\left(X_t, I_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}\right) =$$
$$= \int_{\mathbb{R}^3} \mathcal{V}\left(Y, \eta_0, \eta_1; \boldsymbol{\theta}\right) p\left(Y, \eta_0, \eta_1 \middle| X_t, I_t, \varepsilon_{0t}, \varepsilon_{1t}; \boldsymbol{\theta}_2\right) dY \, d\eta_0 d\eta_1$$

Rust introduces a **conditional independence assumption**:

$$p\left(X_{t+1}, \varepsilon_{0(t+1)}, \varepsilon_{1(t+1)} \middle| X_t, I_t, \varepsilon_{0t}, \varepsilon_{1t}; \boldsymbol{\theta}_2\right) =$$
$$= f\left(\varepsilon_{0(t+1)}, \varepsilon_{1(t+1)} \middle| X_{t+1}, X_t, I_t, \varepsilon_{0t}, \varepsilon_{1t}\right) \cdot$$
$$\cdot \, q\left(X_{t+1} \middle| X_t, I_t, \varepsilon_{0t}, \varepsilon_{1t}; \boldsymbol{\theta}_2\right)$$
$$= f\left(\varepsilon_{0(t+1)}, \varepsilon_{1(t+1)} \middle| X_{t+1}\right) q\left(X_{t+1} \middle| X_t, I_t; \boldsymbol{\theta}_2\right)$$

where the second line follows from additional simplifications. All parameters in $f\left(\cdot\middle|\cdot\right)$ are assumed away (say, normalized) in the analysis. Note how $X_t$ follows a **first-order Markov process**.

# The dynamic logit model (5/10)

The model's likelihood function helps appreciate the usefulness of the assumption. Suppose that a **sample** of $N$ **buses** is available, and write $\boldsymbol{i}_{it} = \{I_{i\tau}\}_{\tau=0}^{t}$ and $\boldsymbol{x}_{it} = \{X_{i\tau}\}_{\tau=0}^{t}$ for $i = 1, \dots, N$ and $t = 1, \dots, T$, where $T$ is *finite*. Then:

$$\begin{aligned}
\mathcal{L}\left(\theta \middle| \{\boldsymbol{i}_{iT}, \boldsymbol{x}_{iT}\}_{i=1}^{N}\right) &= \prod_{i=1}^{N} \prod_{t=1}^{T} \mathbb{P}\left(I_{it}, X_{it} \middle| \boldsymbol{i}_{i(t-1)}, \boldsymbol{x}_{i(t-1)}; \theta\right) \\
&= \prod_{i=1}^{N} \prod_{t=1}^{T} \mathbb{P}\left(I_{it} \middle| X_{it}; \theta\right) q\left(X_{it} \middle| X_{i(t-1)}, I_{it}; \theta_2\right)
\end{aligned}$$

where the second line follows by Rust's assumption. This suggests a **two-step** approach to estimation.

1. In the **first step**, estimate $\theta_2$ using solely data about $\boldsymbol{x}_T$, *conditional* on non-replacement of the engine.

2. In the **second step**, and for a fixed value of $\beta$ (more on this later), estimate $\theta_1$ using a "dynamic logit."

# The dynamic logit model (6/10)

The first step is fairly simple: it is a simple maximum likelihood problem. One could for example maintain a continuous support for $X_{it}$, formulate a functional form assumption about $q(\cdot\,;\boldsymbol{\theta}_2)$, and estimate $\boldsymbol{\theta}_2$ accordingly.

Alternatively, one could non-parametrically estimate the **matrix** of **transition probabilities** after discretizing $X_{it}$. For example, if $X_{it}$ is measured in kilometers; $\Delta X_{it} = X_{it} - X_{i(t-1)}$, and:

$$\mathbb{P}(\Delta X_{it}) = \begin{cases} \theta_{2\,low} & \text{if } 0 \leq \Delta X_{it} < 5000 \\ \theta_{2\,medium} & \text{if } 5000 \leq \Delta X_{it} < 10000 \\ \theta_{2\,high} & \text{if } 10000 \leq \Delta X_{it} < \infty \end{cases}$$

this is an exercise about estimating a categorical distribution's parameters with $\theta_{2\,low} + \theta_{2\,medium} + \theta_{2\,high} = 1$. In Rust's original paper, mileage is discretized over 90 intervals.

# The dynamic logit model (7/10)

To build the dynamic logit for the second step it is necessary to make assumptions about $f(\cdot)$. If both $\varepsilon_{0it}$ and $\varepsilon_{1it}$ are standard Gumbel shocks, independent of one another and of $X_{it}$, one gets:

$$\mathbb{P}\left(I_{it} \middle| X_{it}; \theta\right) = \\ = \frac{\exp\left(\widetilde{\pi}\left(X_{it}, I_{it}; \theta_1\right) + \beta \mathcal{E}\mathcal{V}\left(X_{it}, I_{it}, \boldsymbol{\varepsilon}_t; \theta\right)\right)}{\sum_{J_{it} \in \{0,1\}} \exp\left(\widetilde{\pi}\left(X_{it}, J_{it}; \theta_1\right) + \beta \mathcal{E}\mathcal{V}\left(X_{it}, J_{it}, \boldsymbol{\varepsilon}_t; \theta\right)\right)}$$

where $\widetilde{\pi}\left(X_{it}, I_{it}; \theta_1\right) \equiv \chi I_{it} - c\left(X_{it}\left(1 - I_{it}\right); \theta_1'\right)$ for $I_{it} \in \{0, 1\}$.

The main challenge here is computational: $\mathcal{E}\mathcal{V}\left(\cdot\,; \theta\right)$ depends on the parameters in a non-trivial way, as the solution of a dynamic optimization problem. More elaborate assumptions on $f(\cdot)$ bring about additional complications.

Naturally, assumptions about $c\left(X_{it}; \theta_1'\right)$ are also necessary; since Rust, a linear specification is usually preferred.

# The dynamic logit model (8/10)

To estimate $\boldsymbol{\theta}_1$ Rust suggests an iterative "outer loop, inner loop" **nested fixed point** algorithm. Given $\widehat{\boldsymbol{\theta}}_2$ as obtained in the first step, at every iteration of $\boldsymbol{\theta}_1$ proceed as follows.

- In the **inner loop**, use numerical methods to evaluate the **expected** value function and thus $\mathbb{P}\left(I_{it} \middle| X_{it}; \boldsymbol{\theta}\right)$; here:

$$\mathcal{EV}\left(X_{it}, I_{it}; \widetilde{\boldsymbol{\theta}}\right) =$$

$$= \int_{\mathbb{R}} \log\left[\sum_{J\in\{0,1\}} \exp\left(\widetilde{\pi}\left(Y, J; \boldsymbol{\theta}_1\right) - \beta\mathcal{EV}\left(Y, J; \widetilde{\boldsymbol{\theta}}\right)\right)\right] \cdot$$

$$\cdot\, q\left(Y \middle| X_{it}, I_{it}; \widehat{\boldsymbol{\theta}}_2\right) dY$$

  where $\widetilde{\boldsymbol{\theta}} = \left(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2\right)$, and similarly if $X_{it}$ is discretized.

- In the **outer loop**, search for the value of $\boldsymbol{\theta}_1$ that, given $\widehat{\boldsymbol{\theta}}_2$, maximizes the joint likelihood function of the data.

# The dynamic logit model (9/10)

The **expected** value function in the inner loop is given in closed form: how convenient! To appreciate it, a digression is useful.

If $\{V_i\}_{i=1}^{N}$ is a sequence of $N$ i.i.d. random variables such that

$$V_i \sim \text{Gumbel}\left(\delta_i, 1\right)$$

then the *maximum* $V_{(N)}$ is also Gumbel-distributed. In fact:

$$\begin{aligned} F_{V_{(N)}}\left(v\right) = \prod_{i=1}^{N} \mathbb{P}\left(V_i \leq v\right) &= \prod_{i=1}^{N} \exp\left(-\exp\left(-\left(v-\delta_i\right)\right)\right) \\ &= \exp\left(-\exp\left(-\left(v-\log\sum_{i=1}^{N}\exp\left(\delta_i\right)\right)\right)\right) \end{aligned}$$

hence:

$$\mathbb{E}\left[V_{(N)}\right] = \gamma + \log\sum_{i=1}^{N}\exp\left(\delta_i\right)$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant.

# The dynamic logit model (10/10)

In Rust's model, the discount factor $\beta$ is typically held fixed (e.g. calibrated) because it is **non-parametrically unidentified**.

In short, two models are *observationally equivalent* at explaining any given $(\boldsymbol{i}_{iT}, \boldsymbol{x}_{iT})$ sequence:

- a **myopic model**, where $\chi$ is *low* and, for $t = 1, \dots, T$:

$$I_{it} = \arg\max_{J_{it} \in \{0,1\}} \pi\left(X_{it}, J_{it}, \boldsymbol{\varepsilon}_{it}; \theta_1\right)$$

- a **farsighted model**, where $\chi$ is *high* and, for $t = 1, \dots, T$:

$$I_{it} = \arg\max_{J_{it} \in \{0,1\}} \pi\left(X_{it}, J_{it}, \boldsymbol{\varepsilon}_{it}; \theta_1\right) + \mathcal{EV}\left(X_{it}, J_{it}, \boldsymbol{\varepsilon}_{it}; \theta\right)$$

and $\boldsymbol{x}_{iT}$ is determined accordingly. For more details, see Magnac and Thesmar (2002).

# Conditional choice probability estimation (1/6)

- Rust's model was path-breaking, but the nested fixed point estimation algorithm has proven to be too computationally expensive beyond relatively simple cases.

- Researchers have thus attempted alternative approaches.

- The **conditional choice probability** estimation approach by Hotz and Miller (1993) is a successful one such attempt.

- The key idea is that $\mathbb{P}\left( I_{it} \middle| X_{it} \right)$ *can be estimated in the data.*

- The parameters $\boldsymbol{\theta}$ are backed up by matching such empirical estimates to the *model-implied* probabilities $\mathbb{P}\left( I_{it} \middle| X_{it}; \boldsymbol{\theta} \right)$.

- This leads to both simpler and **faster** estimation, and it can be more easily generalized (multinomial choice, non-Gumbel shocks, etc.). This presentation is based on the HZ setting.

# Conditional choice probability estimation (2/6)

Conditional choice probability estimation also entails **two steps**: the first one is about estimating $\boldsymbol{\theta}_2$ *as well as* $\mathbb{P}\left(\left. I_{it} \right| X_{it}\right)$.

- Estimation of $\boldsymbol{\theta}_2$ proceeds as in Rust. When $X_{it}$ has discrete or discretized support $\mathbb{X} = \{\Xi_1, \dots, \Xi_Q\}$ of dimension $Q$, this step returns $Q$ matrices of size $2 \times Q$ expressed as follows.

$$\widehat{\mathbf{Q}}\left(X_{it}\right) \equiv \begin{pmatrix} q\left(\Xi_1 \middle| X_{it}, 0; \widehat{\boldsymbol{\theta}}_2\right) & \dots & q\left(\Xi_Q \middle| X_{it}, 0; \widehat{\boldsymbol{\theta}}_2\right) \\ q\left(\Xi_1 \middle| X_{it}, 1; \widehat{\boldsymbol{\theta}}_2\right) & \dots & q\left(\Xi_Q \middle| X_{it}, 1; \widehat{\boldsymbol{\theta}}_2\right) \end{pmatrix}$$

- In addition, $\mathbb{P}\left(\left. I_{it} \right| X_{it}\right)$ is also estimated, non-parametrically or parametrically (e.g. via a logit). This returns vectors like:

$$\widehat{\mathbf{p}}\left(X_{it}\right) \equiv \begin{pmatrix} \widehat{\mathbb{P}}\left(\left. 0 \right| X_{it}\right) \\ \widehat{\mathbb{P}}\left(\left. 1 \right| X_{it}\right) \end{pmatrix}$$

a “reduced form” of the model, one that is silent about $\boldsymbol{\theta}_1$.

# Conditional choice probability estimation (3/6)

The second step is formulated as an intuitive minimum distance problem over $\boldsymbol{\Theta}_1$, the parameter space of $\boldsymbol{\theta}_1$, given $\widehat{\boldsymbol{\theta}}_2$:

$$\widehat{\boldsymbol{\theta}}_1 = \underset{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1}{\arg\min} \left\| \mathbf{p}_{I=1} - \mathbf{p}_{I=1}\left(\boldsymbol{\theta}_1\right) \right\|$$

where, for different values of $X_{it} \in \mathbb{X}$ (e.g. those from the data):

- $\mathbf{p}_{I=1}$ is a vector of *empirical* conditional choice probabilities from the first step: $\widehat{\mathbb{P}}\left(1 | X_{it}\right)$; and,

- $\mathbf{p}_{I=1}\left(\boldsymbol{\theta}_1\right)$ is a vector of *structural*, *model-implied* conditional choice probabilities for given $\boldsymbol{\theta}_1$ and $\widehat{\boldsymbol{\theta}}_2$: $\mathbb{P}\left(1 \middle| X_{it}; \boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2\right)$.

Given $\boldsymbol{\theta}_1$, vector $\mathbf{p}_{I=1}\left(\boldsymbol{\theta}_1\right)$ may be constructed via **simulation**. In addition, if $\mathbb{X}$ is discrete a **faster**, simpler approach based on linear algebra is *also* possible.

For exposition's sake it is maintained next that $\mathbb{X} = \left\{\Xi_1, \dots, \Xi_Q\right\}$ is discrete.

# Conditional choice probability estimation (4/6)

To illustrate, express the **ex-ante value function** as follows.

$$V\left(X_{it};\boldsymbol{\theta}\right)\equiv\sum_{I_{it}\in\{0,1\}}\mathbb{P}\left(I_{it}\middle|X_{it}\right)\Bigg[\widetilde{\pi}\left(X_{it},I_{it};\boldsymbol{\theta}_1\right)+$$

$$+\mathbb{E}\left[\varepsilon_{I_{it}t}|I_{it},X_{it};\boldsymbol{\theta}\right]+\beta\sum_{\varXi\in\mathbb{X}}q\left(\varXi\middle|X_{it},I_{it};\boldsymbol{\theta}_2\right)V\left(\varXi;\boldsymbol{\theta}\right)\Bigg]$$

Further write the **choice-specific mean value function** as:

$$\mathcal{U}\left(X_{it},I_{it};\boldsymbol{\theta}\right)\equiv\widetilde{\pi}\left(X_{it},I_{it};\boldsymbol{\theta}_1\right)+\beta\sum_{\varXi\in\mathbb{X}}q\left(\varXi\middle|X_{it},I_{it};\boldsymbol{\theta}_2\right)V\left(\varXi;\boldsymbol{\theta}\right)$$

which can be computed if for all $\varXi\in\mathbb{X}$, $V\left(\varXi;\boldsymbol{\theta}\right)$ is known. With Gumbel shocks, the entries of $\mathbf{p}_{I=1}\left(\boldsymbol{\theta}_1\right)$ are calculated as follows.

$$\mathbb{P}\left(I_{it}=1\middle|X_{it};\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_2\right)=\frac{\exp\left(\mathcal{U}\left(X_{it},1;\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_2\right)\right)}{\sum_{J_{it}\in\{0,1\}}\exp\left(\mathcal{U}\left(X_{it},J_{it};\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_2\right)\right)}$$

# Conditional choice probability estimation (5/6)

The choice-specific mean value function can be **simulated** using the first step estimates. Construct $S$ simulated *sequences*:

$$\left\{\left(\mathbf{i}'_{s1}, \mathbf{x}'_{s1}\right), \dots, \left(\mathbf{i}'_{sT'}, \mathbf{x}'_{sT'}\right)\right\}_{s=1}^{S}$$

obtained via $\widehat{\boldsymbol{\theta}}_2$ and $\widehat{\mathbf{p}}\left(X_{it}\right)$ from an initial value $\left(I_0, X_0\right)$. Then:

$$\widetilde{\mathcal{U}}\left(X_0, I_0; \boldsymbol{\theta}\right) = \frac{1}{S}\sum_{s=1}^{S}\Bigg\{\chi I_0 + c\left(X_0\left(1-I_0\right); \boldsymbol{\theta}'_1\right) + \sum_{\tau=1}^{T'}\beta^{\tau}\Bigg[\chi I'_{s\tau} - \\ - c\left(X'_{s\tau}\left(1-I'_{it}\right); \boldsymbol{\theta}'_1\right) + \mathbb{E}\left[\varepsilon_{I_\tau}\middle| I'_{s(\tau-1)}, X'_{s(\tau-1)}; \boldsymbol{\theta}\right]\Bigg]\Bigg\}$$

is an appropriate simulator for $\mathcal{U}\left(X_0, I_0; \boldsymbol{\theta}\right)$ as $T' \to \infty$, though in practice this is truncated at some finite $T'$. When the $\boldsymbol{\varepsilon}_t$ shocks are standard Gumbel, one can show that for $\tau \in \mathbb{N}_0$:

$$\mathbb{E}\left[\varepsilon_{I_{\tau+1}}\middle| I'_{\tau}, X'_{\tau}; \boldsymbol{\theta}\right] = \gamma - \log \widehat{\mathbb{P}}\left(I'_{\tau}\middle| X'_{\tau}\right)$$

else this conditional expectation must be obtained numerically.

# Conditional choice probability estimation (6/6)

The faster method is summarized here for Gumbel shocks. Let:

$$\widehat{\boldsymbol{\pi}}\left(X_{it};\boldsymbol{\theta}_1\right)=\begin{pmatrix}\widetilde{\pi}\left(X_{it},0;\boldsymbol{\theta}_1\right)+\gamma-\log\widehat{\mathbb{P}}\left(0\middle|X_{it}\right)\\ \widetilde{\pi}\left(X_{it},1;\boldsymbol{\theta}_1\right)+\gamma-\log\widehat{\mathbb{P}}\left(1\middle|X_{it}\right)\end{pmatrix}$$

and:

$$\mathbf{v}\left(\boldsymbol{\theta}\right)=\begin{pmatrix}V\left(\varXi_1;\boldsymbol{\theta}\right)\\ \vdots\\ V\left(\varXi_Q;\boldsymbol{\theta}\right)\end{pmatrix}\qquad\widehat{\boldsymbol{\pi}}\left(\boldsymbol{\theta}_1\right)=\begin{pmatrix}\widehat{\boldsymbol{\pi}}\left(\varXi_1;\boldsymbol{\theta}_1\right)\\ \vdots\\ \widehat{\boldsymbol{\pi}}\left(\varXi_Q;\boldsymbol{\theta}_1\right)\end{pmatrix}$$

and:

$$\widehat{\boldsymbol{\Psi}}=\begin{pmatrix}\widehat{\mathbf{p}}^{\mathrm{T}}\left(\varXi_1\right)&\dots&\mathbf{0}^{\mathrm{T}}\\ \vdots&\ddots&\vdots\\ \mathbf{0}^{\mathrm{T}}&\dots&\widehat{\mathbf{p}}^{\mathrm{T}}\left(\varXi_Q\right)\end{pmatrix}\qquad\widehat{\mathbf{Q}}=\begin{pmatrix}\widehat{\mathbf{Q}}\left(\varXi_1\right)\\ \vdots\\ \widehat{\mathbf{Q}}\left(\varXi_Q\right)\end{pmatrix}$$

then:

$$\mathbf{v}\left(\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_2\right)=\left[\mathbf{I}-\beta\widehat{\boldsymbol{\Psi}}\widehat{\mathbf{Q}}\right]^{-1}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\pi}}\left(\boldsymbol{\theta}_1\right)$$

from which $\mathcal{U}\left(X_{it},I_{it};\boldsymbol{\theta}\right)$, and so $\mathbf{p}_{I=1}\left(\boldsymbol{\theta}_1\right)$, are obtained easily.