Simulated Maximum Likelihood

Thierry Kamionka

2021-2022

Early beginning

Euclid's algorithm is one of the oldest algorithms in mathematics (around 300 BC), it allows to calculate the greatest common divisor of two integers.

The use of simulations to approximate in mathematics is not recent. In 1733, Georges Louis Leclerc, Comte de Buffon, proposed the needle problem. The « Mémoire sur le jeu du franc carreau », presented in 1733, was published in 1777.

Let us consider a parquet on which are traced parallel grooves regularly separated by L units of length. Let's throw the needles at random, the needles measuring $L/2$ units. The probability that a needle cuts a groove is equal to $1/\pi$. The idea is then to approximate this probability by the proportion of needles which intersects the grooves to obtain an estimation of π (this is a suggestion by Pierre Simon Laplace, 1736-1813).

Monte Carlo methods are useful to deal with the multidimensional integration problems. Monte Carlo methods are numerical techniques that allow approximate integrals using random draws (Robert and Casella, 2005).

The term "Monte Carlo method" comes from researches that were conduced at Los Alamos National Laboratory at the end of the second word war (1945) for the Manhattan project related to the development of the first atomic bomb. During the second world war, was elaborated the first electronic computer -the ENIAC- at the University of Pennsylvania in Philadelphia (see Metropolis, 1987).

The principle of the use of simulation on the basis of computer ressources was proposed by John von Neumann and Stanislaw Ulam to solve the problem of "neutron diffusion in fissionable material" (Metropolis, 1987).

Interestly, the MCMC algorithms were developed by the same group of scientists who proposed the Monte Carlo method (see Robert and Casella, 2001).

In econometrics, Kloek and Van Dijk (1978) consider the estimation of the parameters of a system of equations corresponding to a macro model. They use Monte Carlo techniques in order estimate expectations corresponding to posterior moments.

Theorem

Let $F(x)$ denote a cumulative distribution function. Let $Q_X(u) = \inf\{x : F(x) \geq u\}$, for all $u \in]0,1[$, denote the quantile function. If U is a random variable uniform on [0,1], then the cdf of the random variable $X = Q_X(U)$ is $F(x)$.

Proof: $\Pr[X \leq x] = \Pr(Q_X(U) \leq x) = \Pr(F(Q_X(U)) \leq F(x)) = \Pr(U \leq F(Q_X(U)) \leq F(x)) = \Pr(U \leq F(x)) = F(x)$.

Explanation: If $U \leq F(x)$, as $U \leq F(Q_x(U))$ (by definition), then we have that $U \leq F(Q_x(U)) \leq F(x)$ for all x such that $U \leq F(x)$ (otherwise $Q_X(U)$ would not be the quantile of order U).
If $U \leq F(Q_x(U)) \leq F(x)$, then we have $U \leq F(x)$ QED.

Algorithm

1) Generate $U \sim Uniform(0, 1)$.

2) If X is a continuous random variable, then $X = F^{-1}(U)$,

If X is a discrete random variable, then
$X = \min\{x \in \mathcal{X} \mid F(x) \geq U\}$ where $\mathcal{X}$ is the set of all possible values for x.

Example: $X \sim$ exponential distribution with parameter λ ($\lambda > 0$).

$F(x) = 1 - \exp(-\lambda\, x)$ and $F^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u)$.

Let $U \sim Uniform(0, 1)$ then $X = -\frac{1}{\lambda} \ln(1 - U)$ is distributed as an exponential random variable with parameter λ.

Theorem

Let us consider f(x) a probability density function ($x \in \mathbf{R}$). Let (X, Y) be uniformaly distributed on the area A such that

$$A = \{(x, y) \in \mathbf{R}^2 \mid 0 \leq y \leq f(x)\}$$

Then $f(x)$ is the pdf of the random variable X.

Proof: $f_{(X,Y)}(x, y) = \frac{1}{\int_{-\infty}^{\infty} \int_0^{f(x)} d\,y d\,x,}$ for all $(x, y) \in A$.

As $\int_{-\infty}^{\infty} f(x) d\,x = 1$, $f_{(X,Y)}(x, y) = 1$ for all $(x, y) \in A$ and $f_{(X,Y)}(x, y) = 0$ otherwise.

The pdf of X is $\int_{-\infty}^{\infty} f_{X,Y}(x, y) d\,y = \int_0^{f(x)} d\,y = f(x)$. QED.

To generate X with pdf $f(x)$, we simply generate (X, Y) uniformly distributed on the area A.

Algorithm: Acceptation - Rejection method

1) Let $\alpha, \beta \in \boldsymbol{R}$ and $\gamma > 0$ such that for all $x \in [\alpha, \beta]$ we have $f(x) \in [0, \gamma]$.

2) Generate $U, V \sim \textit{uniform}(0, 1)$.

Let $X = \alpha + (\beta - \alpha)\ U$ and $Y = \gamma\ V$.

If $f(X) < Y$ then reject the draw and go to step 2.

If $f(X) \geq Y$, then keep X and return to step 2 if you need another draw (otherwise stop).

Example: x is a grade (between 0 and 20) and $f(x)$ is the distribution of the grades among students ($\sim$ Gaussian, or Poisson restricted to 0-20). $\alpha = 0$, $\beta = 20$ and $\gamma = 1$.

Monte Carlo Integration

One needs to evaluate the following integral

$$I = \int_A f(x)\, d\, x_1 \ldots d\, x_J$$

using a random sample of n independent draws obtained from a density function $g(x)$.

The estimate of I is

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)}.$$

where $g(x)$ is chosen such that it is easy to obtain these draws.

The expectation if $\hat{I}$ is

$$E[\hat{I}] = E[\frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)}] = \frac{1}{n} \sum_{i=1}^{n} \int_A \frac{f(x)}{g(x)} g(x)\, d\, x_1 \ldots d\, x_J = I$$

given $\frac{f(x)}{g(x)}$ is finite for every x such that $f(x) \neq 0$.

Example : A dynamic model of participation to the labour market:

The observed endogenous variable is

$$y_{it} = \mathbb{1}[x_{it}' \beta + \gamma y_{it-1} + \epsilon_{it} > 0] \tag{1}$$

where $t = 2, \ldots, T$, and $i = 1, \ldots, n$.

Let β denote a vector of parameters. Let γ denote a real scalar.

For instance, if y_{it} represents the participation to the labor market at time t ($y_{it} \in \{0, 1\}$). i represents the index of the individual.

The expression of the error term in the equation (1) is

$$\epsilon_{it} = \alpha_i + u_{it},$$

where α_i is a non observable individual component (an individual random effect). α_i is assumed to be independent of x_{it} and $\alpha_i \sim N(0, \sigma_\alpha^2)$.

Initial conditions are treated using the method proposed by Heckman (1981).

The endogenous variable at the initial period is

$$y_{i1} = \mathbb{1}[x_{i1}' \beta_1 + \epsilon_{i1} > 0],$$

where x_{i1} is a vector explanatory variables and ϵ_{i1} is an error term. This error term can be correlated with ϵ_{it}, for all $t = 2, \ldots, T$. β_1 is a vector of parameters.

For the error term of the initial equation, we are going to assume that $\epsilon_{i1} \sim \mathrm{N}(0, \sigma_1^2)$.

We adopt an auto-regressive structure for the error terms of the other periods of time

$$u_{it} = \rho\, u_{it-1} + \nu_{it},$$

where $v_i \perp\!\!\!\perp \nu_{it}$, $t = 2, \ldots, \mathrm{T}$, and

$$\nu_{it} \sim \mathrm{N}(0, \sigma_\nu^2).$$

Moreover, the initial error term is such that the correlation between ϵ_{i1} and ϵ_{it}, for $t = 2, \ldots, T$, is, a priori, different from zero.

`Structure of the variance-covariance matrix:`

For each period, the endogenous variable is dichotomous. So, we are going to assume that

$$\sigma_\alpha^2 + \sigma_u^2 = 1,$$

and

$$\text{var}(\epsilon_{i1}) = \sigma_1^2 = 1.$$

We can show that

$$var(u_{it}) = \sigma_u^2 = \frac{\sigma_\nu^2}{(1 - \rho^2)}.$$

Let $\tilde{\epsilon}_i = (\epsilon_{i2}, \cdots, \epsilon_{iT})'$ denote the vector of error terms for the periods distinct from the initial time. We obtain that

$$\tilde{\epsilon}_i \sim \mathrm{N}(0, \Omega_\epsilon),$$

where

$$\Omega_\epsilon = var(\tilde{\epsilon}_i) = \sigma_\alpha^2 \, 1_{T-1} \, 1'_{T-1} + \frac{\sigma_\nu^2}{(1-\rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-2} \\ \rho & 1 & \rho & \dots & \rho^{T-3} \\ \vdots & & \ddots & & \vdots \\ \rho^{T-3} & \rho^{T-4} & \dots & 1 & \rho \\ \rho^{T-2} & \rho^{T-3} & \dots & \rho & 1 \end{bmatrix}.$$

It is necessary to assume that $0 < \sigma_\alpha^2 < 1$ (because $\sigma_\alpha^2 + \sigma_u^2 = 1$).

We can remark that the parameters σ_ν and ρ are identified using the correlations between the periods going from 2 to T because these correlations are identified in a multivariate Probit model. So, σ_u and, finally, σ_α are identified.

Let us consider the vector $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$. Then we have $\epsilon_i \sim N(0, \Omega)$, with

$$\Omega = \begin{bmatrix} 1 & \rho_0 & \cdots & \rho_0 \\ \rho_0 & & & \\ \vdots & & \Omega_\epsilon & \\ \rho_0 & & & \end{bmatrix},$$

where ρ_0 represents the correlation between the error term of the initial time and the error terms of the next periods. This correlation is identifiable.

`Likelihood function:`

For each period and each individual, we observe $y_{it} \in \{0; 1\}$ (where $i = 1, \ldots, n$ and $t = 1, \ldots, T$).

The contribution of individual i to the likelihood function is given by the expression:

$$L_i(\theta) = \int_{a_{i1}}^{b_{i1}} \int_{a_{i2}}^{b_{i2}} \ldots \int_{a_{iT}}^{b_{iT}} \phi(u_1, u_2, \ldots, u_T; \Omega)\, d\, u_1\, d\, u_2\, \ldots\, d\, u_T,$$

where $\phi(.; \Omega)$ is the pdf of the normal distribution with mean zero and variance-covariance matrix Ω.

Example : ECHP : 1994-2001. Annual data: T=8. Monthly data : T=96 !

The limits a_{it} and b_{it} of this integral are such that:

$$\begin{cases} a_{it} = -\infty, \text{ if } y_{it} = 0 \text{ and } 1 \leq t \leq T, \\ b_{it} = -x'_{it}\,\beta - \gamma\, y_{it-1}, \text{ if } y_{it} = 0 \text{ and } 2 \leq t \leq T, \\ b_{i1} = -x'_{i1}\,\beta_1, \text{ if } y_{i1} = 0, \end{cases}$$

and

$$\begin{cases} a_{it} = -x'_{it}\,\beta - \gamma\, y_{it-1}, \text{ if } y_{it} = 1 \text{ and } 2 \leq t \leq T, \\ a_{i1} = -x'_{i1}\,\beta_1, \text{ if } y_{i1} = 1, \\ b_{it} = +\infty, \text{ if } y_{it} = 1 \text{ and } 1 \leq t \leq T, . \end{cases}$$

The likelihood function is given by the expression

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta),$$

where $\theta = (\sigma_\alpha^2, \sigma_\nu^2, \beta', \beta_1', \rho, \rho_0)'$.

These limits a_{it} and b_{it} depend on the characteristics of the individual i at time t (i.e. x_{it}), the lagged values of the endogenous variable (i.e., $y_{i1},\ldots,y_{it-1}$), and the vector of parameters θ ($\theta \in \Theta$).

Ω is a variance-covariance matrix (Ω is a square matrix $T \times T$).

Simulation based estimation techniques are particularly adapted when the number of observations for each individual is such that $T > 3$.

Example: Probit model with random effects

Let $f(y_i \mid x_i; \alpha_i)$ denote the conditional pdf of the variable y_i given x_i and α_i. We have:

$$f(y_i \mid x_i; \alpha_i; \theta) = \prod_{t=1}^{T} \Phi\left(\left(\frac{x_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}}\right)(2y_{it}-1)\right),$$

and

$$f(y_i \mid x_i; \theta) = \underset{\alpha_i}{E}[\, f(y_i \mid x_i; \alpha_i; \theta)\,],$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\underset{\alpha_i}{E}$ represents the expectation with respect to the distribution of α_i.

The contribution of individual i to the likelihood function is

$$L_i(\theta) = \int_{-\infty}^{+\infty} \prod_{t=1}^{T} \Phi\left(\left(\frac{x_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}}\right)(2y_{it}-1)\right) \frac{1}{\sigma_\alpha}\,\phi\left(\frac{\alpha_i}{\sigma_\alpha}\right) d\,\alpha_i, \quad (2)$$

where $\Phi(u)$ is the cdf of the distribution $N(0,1)$.

$\phi(u)$ is the pdf of the distribution $N(0,1)$.

For each individual i, we replace the integral (2) by an estimator.

We are going to estimate $f(y_i \mid x_i; \theta)$ with

$$\hat{f}(y_i \mid x_i; \theta) = \frac{1}{H} \sum_{h=1}^{H} f(y_i \mid x_i; \alpha_{ih}; \theta),$$

where α_{ih} are i.i.d. random draws according to a distribution with pdf $f(\alpha_i; \sigma_\alpha^2)$ (i.e., the pdf of the $N(0, \sigma_\alpha^2)$ distribution).

H is the number of random draws for each individual (in practice we take $H = 30$ or $H = 40$ or ... or $H = 100$).

We maximize the log of the simulated likelihood function.

Variable	Estimation	Stand. Err.	Estimation	Stand. Err.
	$H = 40$		$H = 100$	
Constant	-0.7588*	0.0443	-0.7890*	0.0527
Manual	0.2092*	0.0453	0.2294*	0.0483
Mar	0.0466	0.0396	0.0334	0.0416
	Variance of the individual effect			
σ_α^2	0.7560*	0.0187	0.7412*	0.0216

Data : NLS. Endogenous variable: Union. H is the number of random draws obtained using the distribution $N(0, 1)$. Random draws are specific to the individual.

▸ Go to Appendix A

Random effect Tobit model

Let us consider the random effect Tobit model

$$\begin{aligned} y_{it}^* &= x_{it}'\beta + \alpha_i + u_{it}, \\ y_{it} &= y_{it}^* \mathbb{1}[y_{it}^* > 0], \end{aligned} \tag{3}$$

where α_i, $i = 1, \ldots, n$, are individual random effects. These effects are assumed to be i.i.d. according to a $N(0, \sigma_\alpha^2)$ distribution. Let us assume that we have T observations for each individual ($t = 1, \ldots, T$).

Let $f(y_i \mid x_i; \alpha_i)$ denote the conditional pdf of the variable y_i given x_i and α_i. We have :

$$f(y_i \mid x_i; \alpha_i; \theta) = \prod_{t=1}^{T} \left[1 - \Phi\left(\frac{x_{it}'\beta + \alpha_i}{\sigma_u}\right)\right]^{\mathbb{1}[y_{it}=0]} \left[\frac{1}{\sigma_u}\phi\left(\frac{y_{it} - x_{it}'\beta - \alpha_i}{\sigma_u}\right)\right]^{\mathbb{1}[y_{it}>0]}$$

and $f(y_i \mid x_i; \theta) = E_{\alpha_i}[\, f(y_i \mid x_i; \alpha_i; \theta)\,]$, where $\alpha_i \sim N(0, \sigma_\alpha^2)$.

We can replace $f(y_i \mid x_i; \theta)$ by

$$\hat{f}(y_i \mid x_i; \theta) = \frac{1}{H} \sum_{h=1}^{H} f(y_i \mid x_i; \alpha_{ih}; \theta),$$

where α_{ih} are i.i.d. random draws according to the distribution with pdf $f(\alpha; \sigma_\alpha^2)$ (i.e., the pdf of the $N(0, \sigma_\alpha^2)$ distribution).

The expression of the log-likelihood function is

$$\ln(\hat{L}_{n,H}(\theta)) = \sum_{i=1}^{n} \ln(\hat{f}(y_i \mid x_i; \theta)) = \sum_{i=1}^{n} \ln(\frac{1}{H} \sum_{h=1}^{H} f(y_i \mid x_i; \alpha_{ih}; \theta)),$$

where $\theta \in \Theta$.

SML : General case

Let $y_i = (y_{i1}, \ldots, y_{iT})'$ denote the vector of endogenous variables observed for individual i. Let T denote the number of observation periods. Let $x_i = (x_{i1}, \ldots, x_{iT})'$ denote the vector of explanatory variables for individual i.

Let $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ denote an i.i.d. sample with size n.

The expression of the log likelihood function is

$$\ln(L(\theta)) = \sum_{i=1}^{n} \ln(f(y_i \mid x_i; \theta)),$$

and let us assume that we have no closed form for $f(y_i \mid x_i; \theta)$.

Let us assume that $\underset{\alpha_i}{E}[f(y_i \mid x_i; \alpha_i; \theta)] = f(y_i \mid x_i; \theta)$, where the distribution of α_i is indexed by an unknown parameter (respectively, a vector of unknown parameters).

SMLE : General case

We are going to generate, for each individual i, H random draws $\alpha_{i,h}$, $h = 1, \ldots, H$, for the individual effect α_i.

We are going to replace $\underset{\alpha_i}{E}[f(y_i \mid x_i; \alpha_i; \theta)]$ by $\frac{1}{H} \sum_{h=1}^{H} f(y_i \mid x_i; \alpha_{ih}; \theta)$, where the variables α_{ih} are i.i.d. random draws according to the pdf $f(\alpha_i; \theta)$.

The expression of the log of the simulated likelihood function is:

$$\ln(\hat{L}_{n,H}(\theta)) = \sum_{i=1}^{n} \ln\Big(\frac{1}{H} \sum_{h=1}^{H} f(y_i \mid x_i; \alpha_{ih}; \theta) \Big).$$

SMLE : General case

`Definition:` A Simulated Maximum Likelihood Estimator (SMLE) is a solution of the problem:

$$\max_{\theta \in \Theta} \ln(\hat{L}_{n,H}(\theta)).$$

This solution is denoted $\hat{\theta}^{sml}_{n,H}$.

SMLE : General case

Theorem

Under regularity assumptions (cf. Gouriéroux, and Monfort, 1991), if $n \longrightarrow +\infty$ *and* $H \longrightarrow +\infty$, *so that* $\sqrt{n}/H \longrightarrow 0$, *the SMLE* $\hat{\theta}^{sml}_{n,H}$ *is consistent, asymptotically distributed as a normal distribution and asymptotically efficient.*

$$\sqrt{n}\,(\hat{\theta}^{sml}_{n,H} - \theta_0) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\theta_0)),$$

where θ_0 *is the true value of the parameter.*

SMLE : General case

Moreover, we have

$$I(\theta) = E\left[\frac{\partial \ln(\hat{L}_i(\theta))}{\partial \theta}\frac{\partial \ln(\hat{L}_i(\theta))}{\partial \theta'}\right]$$

$$= E\left[\frac{\sum_{h=1}^{H}\frac{\partial f}{\partial \theta}(y_i \mid x_i;\alpha_{ih};\theta)}{\sum_{h=1}^{H} f(y_i \mid x_i;\alpha_{ih};\theta)}\frac{\sum_{h=1}^{H}\frac{\partial f}{\partial \theta'}(y_i \mid x_i;\alpha_{ih};\theta)}{\sum_{h=1}^{H} f(y_i \mid x_i;\alpha_{ih};\theta)}\right],$$

GHK Algorithm

For each individual, we can estimate the integral

$$L_i(\theta) = \int_{a_{i1}}^{b_{i1}} \int_{a_{i2}}^{b_{i2}} \dots \int_{a_{iT}}^{b_{iT}} f(s_1, s_2, \dots, s_T \mid \Omega)\, d\, s_1\, d\, s_2 \dots d\, s_T,$$

using random draws obtained on some area defined by the values of the limit a_{it} and b_{it} for $t = 1, \dots, T$ (some "pavé").

Let us assume that f is the pdf of the normal distribution with mean 0 and variance-covariance matrix Ω.

All these random draws are useful because they allow to enhance the approximation of this integral on the corresponding area.

This algorithm was first proposed by John Geweke (1989).

GHK algorithm : G for John Geweke, H for Vassilis Hajivassiliou, K for Mikael Keane.

GHK Algorithm: dimension $T = 2$

We want to calculate :
$\text{Prob}[\epsilon \in P] = \text{Prob}[\epsilon \in [a_1; b_1] \times [a_2; b_2]]$.

The variable ϵ is assumed to be distributed as a $N(0, \Omega)$. We are going to use a standard normal distribution. So, let us use the Cholesky decomposition of the matrix Ω. We have then:

$$\epsilon = L\, u,$$

$$L = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix},$$

$$u \sim N(0, I_2).$$

Let $\epsilon = (\epsilon_1, \epsilon_2)'$ and $u = (u_1, u_2)'$.

GHK Algorithm: dimension $T = 2$

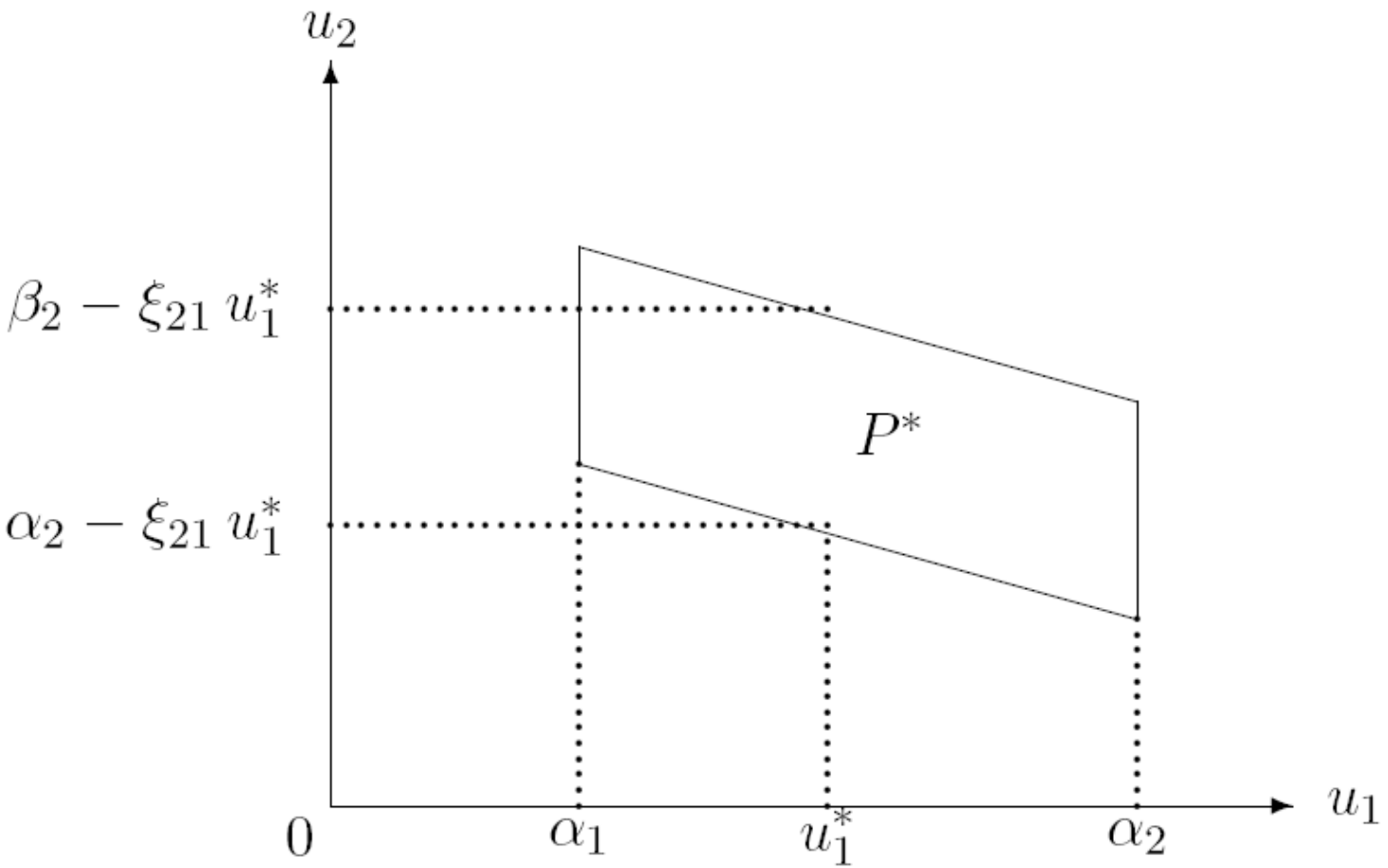
$$
\begin{aligned}
\text{Prob}[\epsilon \in P] &= \text{Prob}[a_1 < \epsilon_1 < b_1; a_2 < \epsilon_2 < b_2], \\
&= \text{Prob}[a_1 < l_{11}\, u_1 < b_1; a_2 < l_{21}\, u_1 + l_{22}\, u_2 < b_2], \\
&= \text{Prob}[\tfrac{a_1}{l_{11}} < u_1 < \tfrac{b_1}{l_{11}}; \tfrac{a_2}{l_{22}} < \tfrac{l_{21}}{l_{22}}\, u_1 + u_2 < \tfrac{b_2}{l_{22}}], \\
&= \text{Prob}[\alpha_1 < u_1 < \beta_1; \alpha_2 < \xi_{21}\, u_1 + u_2 < \beta_2].
\end{aligned}
$$

where $\alpha_t = \frac{a_t}{l_{tt}}$, $\beta_t = \frac{b_t}{l_{tt}}$, $t = 1, 2$ and $\xi_{21} = \frac{l_{21}}{l_{22}}$.

Let us consider u_1^*, a random draw in the $N(0,1)$ distribution truncated on the interval $[\alpha_1; \beta_1]$ and, u_2^*, is a random draw in the $N(0,1)$ distribution restricted on the interval $[\alpha_2 - \xi_{21}\, u_1^*; \beta_2 - \xi_{21}\, u_1^*]$.

GHK Algorithm: dimension $T = 2$

Illustration:



GHK Algorithm: dimension $T = 2$

The random draws we obtain are in the good area defined by the limits of the integral.

We need, now, to construct an estimate of the probability Prob$[\epsilon \in P]$. Let us assume that

$$g(u_1^*) = [\Phi(\beta_1) - \Phi(\alpha_1)]\,[\Phi(\beta_2 - \xi_{21}\,u_1^*) - \Phi(\alpha_2 - \xi_{21}\,u_1^*)].$$

We can verify that

$$\begin{aligned} E[g(u_1^*)] &= \int_{\alpha_1}^{\beta_1} g(u_1^*)\,\frac{\phi(u_1^*)}{\Phi(\beta_1) - \Phi(\alpha_1)}\,d\,u_1^*, \\ &= \int_{\alpha_1}^{\beta_1} [\Phi(\beta_2 - \xi_{21}\,u_1^*) - \Phi(\alpha_2 - \xi_{21}\,u_1^*)]\,\phi(u_1^*)\,d\,u_1^*, \\ &= \text{Prob}[u \in P^*] = \text{Prob}[\epsilon \in P]. \end{aligned}$$

GHK Algorithm: dimension $T = 2$

In practice, in order to generate u_1^* distributed as a truncated normal on the interval $[\alpha_1; \beta_1]$, we use the inversion formula:

$$u_1^* = \Phi^{-1}([\Phi(\beta_1) - \Phi(\alpha_1)]\, v_1 + \Phi(\alpha_1)),$$

where v_1 is a random draw in a uniform distribution on the interval $[0; 1]$ ($v_1 \sim U(0, 1)$).

These random draws u_{1h}^*, $h = 1, \ldots, H$, are assumed to be independent.

We can remark that, in order to estimate the probability $\text{Prob}[\epsilon \in P]$, we need only to generate u_1^*.

Example : Bivariate Probit (GSOEP Data)

Two endogenous variables :

- Active practice of sport (Sport)
- Full time work (Ft)

To illustrate the method we use a cross section (but, as we have two variables and one date only it is equivalent to the situation where $T = 2$ and a single variable).

Explanatory variables: Age, schooling (Educ), gender (Male).

Two competing estimation methods: MLE and SMLE (H=20, H=40).

Example : Bivariate Probit (GSOEP Data)

Variable	Estim.	Stand.	Estim.	Stand.	Estim.	Stand.
	GHK H=20		*GHK H=40*		*Numerical (QLIM)*	
			Full time employment			
Const.	-1.3298*	0.0819	-1.3298*	0.0819	-1.3299*	0.0819
Male	1.0338*	0.0277	1.0339*	0.0277	1.0339*	0.0277
Age	0.4432*	0.1148	0.4431*	0.1148	0.4431*	0.1148
Educ	0.0663*	0.0059	0.0663*	0.0059	0.0663*	0.0059
			Sport			
Const.	-1.0562*	0.0792	-1.0563*	0.0792	-1.0562*	0.0792
Male	0.1410*	0.0279	0.1410*	0.0279	0.1409*	0.0279
Age	-1.6670*	0.1198	-1.6671*	0.1198	-1.6664*	0.1198
Educ	0.0964*	0.0057	0.0965*	0.0057	0.0964*	0.0057
			Correlation			
ρ=tanh(h)	-0.0407*	0.0181	-0.0394*	0.0183		
ρ					-0.0431*	0.0184

Data : GSOEP. Endogenous variables: Ft and Sport. N=9197.

GHK Algorithm: General case

We need that $\epsilon = (\epsilon_1, \ldots, \epsilon_T)' \in P$, where $\epsilon \sim N(0, \Omega)$. We are going to use the Cholesky decomposition of the matrix Ω in order to obtain a vector $u = (u_1, \ldots, u_T)'$ such that $\epsilon = L\,u$, where L is a lower triangular matrix with elements l_{mn}, for all $m, n = 1, \ldots, T$.

If $\epsilon \in P$ then u belongs to P^*. Then, we need that u to be such that:

$$\alpha_1 \leq u_1 \leq \beta_1,$$

$$\alpha_2 \leq u_2 + \xi_{21}\, u_1 \leq \beta_2,$$

$$\vdots$$

$$\alpha_T \leq u_T + \xi_{T1}\, u_1 + \ldots + \xi_{T,T-1} u_{T-1} \leq \beta_T,$$

where $\xi_{m,n} = l_{m,n}/l_{m,m}$, $\alpha_m = a_m/l_{m,m}$ and $\beta_m = b_m/l_{m,m}$.

GHK Algorithm: General case

We are going to proceed as follow

1) we generate u_1^* using a N(0,1) distribution truncated over the interval $[\alpha_1; \beta_1]$,
2) we generate u_2^* using a N(0,1) distribution truncated over the interval $[\alpha_2 - \xi_{21}\, u_1^*; \beta_2 - \xi_{21}\, u_1^*]$,

   $\vdots$

T-1) we generate u_{T-1}^* using a distribution N(0,1) truncated over the interval $[\alpha_{T-1} - \xi_{T-1,1}\, u_1^* - \ldots - \xi_{T-1,T-2}\, u_{T-2}^*; \beta_{T-1} - \xi_{T-1,1}\, u_1^* - \ldots - \xi_{T-1,T-2}\, u_{T-2}^*]$,

GHK Algorithm: General case

In order to estimate Prob$[\epsilon \in P]$ = Prob$[u \in P^*]$ we can use

$$g(u_1^*, \ldots, u_T^*) = [\Phi(\beta_1) - \Phi(\alpha_1)] \prod_{t=2}^{T} [\Phi(\beta_t - \xi_{t,1}\, u_1^* - \ldots - \xi_{t,t-1}\, u_{t-1}^*) - \Phi(\alpha_t - \xi_{t,1}\, u_1^* - \ldots - \xi_{t,t-1}\, u_{t-1}^*)]$$

For each individual, we can generate H random draws independently $u_{ih}^* = (u_{ih1}^*, \ldots, u_{ihT}^*)'$, $h = 1, \ldots, H$, acording to this algorithm.

The expression of the contribution to the simulated likelihood function is

$$\hat{L}_i(\theta) = \frac{1}{H} \sum_{h=1}^{H} g(u_{ih}^*). \tag{4}$$

GHK Algorithm: General case

The expression of the likelihood function is :

$$\hat{L}(\theta) = \prod_{i=1}^{n} \hat{L}_i(\theta),$$

where the expression of $\hat{L}_i(\theta)$ is given by the equation (4).

In order to implement this algorithm we need to draw in a truncated standard normal distribution over the interval $[\alpha; \beta]$.

$\longrightarrow$ see the inversion theorem.

Let us remark that there exist now many software packages that allow to estimate commonly used models using SML estimation.

For instance, Alexander Plum (2016) proposes a **Stata** command to estimate a bivariate random effects probit model (command name bireprov).

Haan and Uhlendorff (2006) propose a Stata routine to estimate a logit model with unobserved heterogeneity by maximum simulated likelihood. McFadden and Train (2000) propose a program for mixed logit models in **Gauss**.

Mark Stewart (2006) proposes a Stata command to estimate by SMLE a random effects dynamic probit model with autocorrelated errors (command name redpace).

For other models, one may have to program (**Stata** (MATA), **Gauss**, **Matlab**, **R**).

Vehicle Choice Behavior

Let us consider an illustration consisting in a model of vehicle choice. One an be interested by the impact of vehicle attributes, brand loyalty, dealership on market shares (Train and Winston, 2007).

The car attributes are price, size (wheelbase, length minus wheelbase), power (horsepower divided by weight), transmission type (automatic), operating costs (fuel consumption), reliability and body type (SUV, minivan, Pickup truck, Luxury or sport car).

Dealership is the number of dealership within 50 miles of the center of a respondent's zip code.

Brand loyalty is the number of previous consecutive brand purchases.

In 2000, U.S. manufacturers represented 64% of the market share of cars and light trucks (44% in 2019, Statistica 2020).

Let us consider a model based on **random utility** function that allow to characterize the choices of consumers - the agents - by make (e.g. GM, Toyota) and model (e.g. Corvette, Corolla).

Let us consider n consumers ($i = 1, \ldots, n$) and J models of new vehicles ($j = 1, \ldots, J$).

The utility consumer i obtains from vehicle j is

$$U_{ij} = \delta_j + \beta' x_{ij} + \mu_i' w_{ij} + \epsilon_{ij}$$

where δ_j is the part of the utility that is the same for all consumers.

x_{ij} is a vector of consumers characteristics related with vehicle j and dealership (components of observed heterogeneity). The vector β is the mean coefficient of these variables in the population.

w_{ij} is a vector of consumers characteristics related with vehicle j (associated to unobserved heterogeneity). μ_i is a vector of random terms with mean zero associated to these variables.

ϵ_{ij} is an error term corresponding to all other - unobserved - elements that can have an impact on the utility of consumer i for vehicle j.

For elements of w_{ij} that do not correspond to elements of x_{ij}, the corresponding elements of μ_i can be considered as **random coefficients with zero mean**.

For elements of w_{ij} that correspond to elements of x_{ij}, the corresponding elements of β can be considered **the average coefficients** and the corresponding elements of μ_i can be considered as **random variations around the average**.

Let us remark that the unobserved component $\mu_i' w_{ij}$ induces correlations between vehicles. The presence of this term in the utility allows to surmount the IIA restriction.

Let $f(\mu \mid \Sigma)$ denote the pdf of μ_i where Σ is a parameter corresponding, for instance, to the variance-covariance matrix of μ_i (however, f may depend of some explanatory variables).

Let us assume that ϵ_{ij} are i.i.d. extreme value distributed. The conditional probability that consumer i chooses the vehicle j when he buys a new vehicle is

$$P_{ij} = \int \frac{\exp(\delta_j + \beta' x_{ij} + \mu' w_{ij})}{\sum_{k=1}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik})} f(\mu \mid \Sigma) \, d\mu$$

It is a **mixed logit model**.

Let S_j denote the market share of consumers who buy vehicle j.

The predicted share $S_j(\beta, \delta, \Sigma)$ is obtained by averaging P_{ij} over the n consumers. For a given value of (β, Σ) there exists a unique value of δ such that these predicted market share are equal to the actual ones (Berry, 1994).

Let $\delta(\beta, \Sigma, S)$, where $S = (S_1, \ldots, S_J)$, the value of δ such that

$$S_j = S_j(\beta, \delta(\beta, \Sigma, S), \Sigma) = \frac{1}{n} \sum_{i=1}^{n} P_{ij}(\beta, \Sigma, \delta(\beta, \Sigma, S))$$

where (β, Σ) are estimated by MLE and δ is such that the predicted market shares are equal to the observed market shares for the estimated value of (β, Σ) .

$\longrightarrow$ Here we combine information from the sample and external - **aggregated** - information (S) to estimate the model.

Let assume that the average utility $\delta_j(\beta, \Sigma, S)$ depends of vehicle's j observed attributes (z_j) and unobserved characteristics (ξ_j)

$$\delta_j(\beta, \Sigma, S) = \alpha' z_j + \xi_j, \tag{5}$$

where some elements of z_j can be common with elements of w_{ij}.

The price of the vehicle - an element of z_j - can have an impact on the unobserved component ξ_j. Let y_j denote a vector of instruments. y_j includes the non price elements of z_j and other exogenous variables.

Let us assume that $E[\xi_j \mid y_j] = 0$, for all $j = 1, \ldots, J$. The IV estimator of α is then consistent and asymtoticaly normal given (β, Σ).

The set of vehicle that an individual considered seriously before making its choice provides additional information on preferences. Let us assume that the sample contains information on the list of vehicles the consumer considered with the corresponding ranking.

A consumer who considered seriously only two vehicles j and h and preferred vehicle j is such that $U_{i,j} > U_{i,h} > U_{ik}$ for all $k \neq j, h$.

Luce and Suppes (1965) has shown that, under the assumption that ϵ_{ij} are i.i.d. extreme value, the probability of the ranking is a product of logit functions. For a consumer who list only two vehicles j and h (j is prefered to h and h to others vehicles), the conditional probability of the list is

$$q_i(\mu) = \frac{\exp(\delta_j + \beta' x_{ij} + \mu' w_{ij})}{\sum_{k=1}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik})} \frac{\exp(\delta_h + \beta' x_{ih} + \mu' w_{ih})}{\sum_{k=1, h \neq j}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik})}.$$

The conditional probability of the ranking for consumer who listed more than two vehicles is defined similarly (more elements in the product).

The unconditional probability of the ranking made by i is

$$q_i = \int q_i(\mu)\, f(\mu \mid \Sigma)\, d\,\mu$$

Berry (1994) underlined that it was not possible to estimate unobserved taste variation without considering consumer's rankings (by only considering consumers' purshases).

δ is calculated for each value of (β, Σ) that is considered during the estimation of the parameters.

δ can be calculated iteratively using the formula

$$\delta_j^k(\beta, \Sigma, S) = \delta_j^{k-1}(\beta, \Sigma, S) + \ln(S_j) - \ln(S_j(\beta, \Sigma, \delta_j^{k-1}(\beta, \Sigma, S))) \tag{6}$$

where $j = 1, \ldots, J$.

Introduction of brand loyalty: Past experience of the consumer with one manufacturer determine its loyalty with this manufacturer. This phenomenon can be interpreted as a kind of state dependence.

Let η_{im} denote the **preference of the consumer** i **for manufacturer** m, where $m = 1, \ldots, K$ (Ford, GM, Chrysler, Japanese, European, Korean).

Let us assume that the unconditional distribution of $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})$ is denoted $g(\eta_i)$.

The actual distribution of η_i given consumer has chosen previously manufacturer m is

$$h(\eta_i \mid \eta_{im} > \eta_{ik}, k \neq m) = \frac{\mathbb{1}[\eta_{im} > \eta_{ik}, \forall k \neq m]\ g(\eta_i)}{\int \mathbb{1}[\eta_{im} > \eta_{il}, \forall l \neq m]\ g(\eta_i)\ d\,\eta_i}$$

where $g(\eta)$ is a pdf.

Let us modify the utility function in order to take into account brand loyalty. The utility consumer i obtains from vehicle j produced by manufacturer $s(j)$ is

$$U_{ij} = \delta_j + \beta' x_{ij} + \mu_i' w_{ij} + \lambda \eta_{is(j)} + \epsilon_{ij}$$

where λ is a coefficient ($\lambda \in \mathbb{R}$).

The conditional probability that consumer i chooses the vehicle j produced by manufacturer $s(j)$ given that he has chosen to buy in the past a vehicle produced by manufacturer m is

$$P_{ij} = \int \int \frac{\exp(\delta_j + \beta' x_{ij} + \mu' w_{ij} + \lambda \eta_{is(j)})}{\sum_{k=1}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik} + \lambda \eta_{is(k)})} \tag{7}$$
$$f(\mu \mid \Sigma)\ h(\eta_i \mid \eta_{im} > \eta_{ik}, k \neq m)\ d\,\mu\ d\,\eta_i$$

The conditional probability that consumer i choses to buy vehicle j and ranks a second vehicle ℓ is

$$q_i = \int \int \frac{\exp(\delta_j + \beta' x_{ij} + \mu' w_{ij} + \lambda \eta_{is(j)})}{\sum_{k=1}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik} + \lambda \eta_{is(k)})}$$
$$\times \frac{\exp(\delta_\ell + \beta' x_{i\ell} + \mu' w_{i\ell} + \lambda \eta_{is(\ell)})}{\sum_{k=1, k\neq j}^{J} \exp(\delta_k + \beta' x_{ik} + \mu' w_{ik} + \lambda \eta_{is(k)})} \tag{8}$$
$$f(\mu \mid \Sigma)\ h(\eta_i \mid \eta_{im} > \eta_{ik}, k \neq m)\ d\,\mu\ d\,\eta_i$$

It can be taken into account in a similar way the fact that the individual has ranked more than two vehicles.

Estimation: The conditional choice probability (7) and (8) are integrals that cannot be calculated analytically. Simulations can be used in order to approximated these quantities. For instance, the simulated conditional probability that consumer i choses vehicle j is

$$\hat{P}_{ij} = \frac{1}{H} \sum_{r=1}^{H} \frac{\exp(\delta_j(\beta, \Sigma, S) + \beta' x_{ij} + \mu_r' w_{ij} + \lambda \eta_{ris(j)})}{\sum_{k=1}^{J} \exp(\delta_k(\beta, \Sigma, S) + \beta' x_{ik} + \mu_r' w_{ik} + \lambda \eta_{ris(k)})}$$

where μ_r are i.i.d. draws obtained from density $f(\mu \mid \Sigma)$ and η_{ris} are i.i.d. draws obtained from the conditional density h.

The conditional probability that the individual buys vehicle j from manufacture $s(j)$ and ranked second the vehicle ℓ given he has already purchased vehicle from manufacturer m in the past can be obtained in a similar way.

For instance, under the assumption that all the consumers have given only an information on the vehicle purchased and the first vehicle ranked, the simulated log-likelihood is

$$\ln(L(\theta)) = \sum_{i=1}^{n} \ln(\hat{q}_i)$$

where $\theta = (\beta, \lambda, \Sigma)$ is the vector of parameters of the models (in practice consumers can rank more than one vehicle). The vector $\delta = (\delta_1, \ldots, \delta_J)$ of parameters is obtained using the expression (6). Train and Winston (2007) used $H = 200$ to estimate the model (it is better to use specific draw for each consumer).

The parameters of the model for average utility δ_j are then estimated using the regression (5) where z_j are attributes of vehicle j. As already noted, price of the vehicle should be instrumented. The price can be correlated with omitted characteristics of the vehicle.

The price of a vehicle depends on the attributes of all other vehicles (Nash equilibrium in price). The attributes of all other vehicles are correlated with the price of this vehicle but not with the omitted characteristics of this vehicle.

Let $\bar{z}_{jk}$ denote the difference in one attribute - such as fuel economy - between the two vehicles j and k. Train and Winston (2007) suggest four instruments for each vehicle k and each attribube:

$\sum_{j \in M_{s(k)}} \bar{z}_{jk}$ where $s(k)$ is a manufacturer and $M_{s(k)}$ is the set of all the vehicles made by this manufacturer;

$\sum_{j \notin M_{s(k)}} \bar{z}_{jk}$, the sum of the difference for all the vehicles made by other manufacturers;

$\sum_{j \in M_{s(k)}} \bar{z}_{jk}^2$, the sum of the square of the difference for all vehicles made by this manufacturer;

$\sum_{j \notin M_{s(k)}} \bar{z}_{jk}^2$, the sum of the square of the difference for all the vehicles made by other manufacturers.

The two last quantities allows to take into account the importance of the dispersion in the non price attributes of other vehicles compared to non price attributes of vehicle k.

Validity assumption: we have to assume that the other attributes of the vehicle (distinct of the price) are not endogeneous.

Berry, S., 1994, Estimating discrete-choice models of product differentiation, RAND Journal of Economics, 25, 242-262.

Gourieroux, Ch., Monfort, A., 1991. Simulation based inference in models with heterogeneity, Annales d'Economie et de Statistique, 20/21, 69-107.

Haan, P., Uhlendorff, A., 2006. Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood, The Stata Journal, 6(2), 229–245.

Heckman, J.J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete stochastic process and some Monte Carlo evidence on their practical importance. in C. Manski and D. McFadden, eds. Structural analysis of discrete data, Cambridge, MA: MIT Press.

Kloek, T., van Dijk, H.K., 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo, Econometrica, 46(1), 1-19.

Leclerc G.L., 1777, Histoire naturelle, générale et particulière : supplément. Tome Quatrième, Essais d'arithmétique morale, 95–105.

Luce, R., Suppes, P, 1965. Preference, utility and subjective probability. in R. Luce, R. Bush and E. Galanter eds., Handbook of Mathematical Psychology, III (New-York: Wiley).

McFadden, D., Train, K., 2000. Mixed MNL Models for Discrete Response, Journal of Applied Econometrics 15(5), pp. 447-470.

Metropolis, N., 1987. The Beginning of the Monte Carlo Method, Los Alamos Science Special Issue, 15, 125–130.

Plum, A., (2016), bireprob: An estimator for bivariate random-effects probit models. The Stata Journal, 16(1), 96–111.

Robert, Ch., Casella, G., 2001. A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data, Statistical Science 2011, 26(1), 102–115.

Robert, Ch., Casella, G., 2005. Monte Carlo Statistical Methods, Springer New York.

Stewart, M., 2006. Maximum simulated likelihood estimation of random-effects dynamic probit models with autocorrelated errors, The Stata Journal, 6(2), 256–272.

Train, K. E., Winston, C., 2007. Vehicle choice behavior and the declining market share of U.S. automakers. International Economic Review, 48(4), 1469–1496.

Appendix A : Probit model with random effect

Data Set description

Data extracted from NLS (National Longitudinal Survey, USA).

N=545 men for the period going from 1980 to 1987.

A record in the sample : an individual for a given year.

WAGE : Log of the hourly wage (endogenous variable).

UNION (wage set by a collective bargaining), RUR (lives in a rural area), MAR (married), HLTH (problem with health), EXPER (Experience).

OCC1 (technical), OCC2 (managers), OCC3 (sales workers), OCC4 (clerical), OCC5 (craftsmen), OCC6 (operatives), OCC7 (farmers), OCC8 (foreman), OCC9 (service workers).

◂ Go Back

```
LIBNAME OUT 'w:'; run;
Data males;
 set out.males;
run;
Proc sort data=males;
 by Indivi Year;
run;
Data Simu(drop=h);
 Array sim20 s1-s20;
 Seed = 123456;
 Do i=1 to 545; /* We have 545 individuals */
    Do h=1 to 20; /* 20 random draws specific to the individual */
      call rannor(Seed,sim{h});
    End;
    Output;
 End;
run;
```

```
Data Males2(Keep=I U1-U8 M1-M8 MAn1-Man8);
 Set Males;
 retain I 0 M1-M8 0 U1-U8 0 MAN1-MAN8 0;
 If (Year eq 1980) then Do;
   I = I+1;
   M1 = Mar;Man1 = Man;U1 = Union;
 End;
 If (Year eq 1981) then Do;
   M2 = Mar;Man2 = Man;U2 = Union;
 End;
 If (Year eq 1982) then Do;
   M3 = Mar;U3 = Union;Man3 = Man;
 End;
 If (Year eq 1983) then Do;
   M4 = Mar;U4 = Union;Man4 = Man;
 End;
```

```
If (Year eq 1984) then Do;
  M5 = Mar;U5 = Union;Man5 = Man;
End;
If (Year eq 1985) then Do;
  M6 = Mar;U6 = Union;Man6 = Man;
End;
If (Year eq 1986) then Do;
  M7 = Mar;U7 = Union;Man7 = Man;
End;
If (Year eq 1987) then Do;
  M8 = Mar;U8 = Union;Man8 = Man;
  Output;
End;
Run;
```

```
/* We regroup the information coming from the data and from
simulations */

/* A line for each individual */

Data Males_Simu;
 merge Males2 Simu;
 by I;
run;
```

```
Proc NLMixed data=Males_Simu;
 parms acste=-1 aMar=0 aManual=0 sigma2=0.2;
 array pesti{20} p1-p20; array Ma{8} M1-M8; array Un{8} U1-U8;
 array Man{8} MAN1-MAN8; array Sim{20} s1-s20;
 Do h = 1 to 20;
    prod = 1;
    Do t = 1 to 8;
      xb = acste + aMar*Ma{t} + aManual*Man{t};
      If (Unt eq 1) then prob =
        probnorm((xb+sqrt(sigma2)*sim{h})/sqrt(1-sigma2));
      If (Unt eq 0) then prob =
        1-probnorm((xb+sqrt(sigma2)*sim{h})/sqrt(1-sigma2));
      Prod = prod*prob;
    End;
    pesti{h} = Prod;
 End;
 ll = log(sum(of p1-p20)/20);model I ~ general(ll); run;
```

Go Back