ÉCONOMÉTRIE
(UGA, M1-M2 BDA)
Inférence par le Maximum de Vraisemblance (Conditionnelle)

Michal W. Urdanivia*

*Université de Grenoble Alpes, Faculté d'Économie, GAEL,
e-mail: michal.wong-urdanivia@univ-grenoble-alpes.fr

30 septembre 2024

Contenu

Plan

1. Principe d'estimation par le MV(C)
2. Propriétés de l'estimateur du MV(C)
3. Liens avec la Méthode des Moments
4. Exemples
5. Quelques résultats utiles
6. MV Conditionnelle versus MV
7. Problème de spécification : Pseudo- ou Quasi-MV

1. Principe d'estimation par le MV(C)

1. Principe du Maximum de Vraisemblance (Conditionnel)

- On cherche ici à :

Caractériser la distribution commune des $\mathbf{y}_i \mid \mathbf{x}_i$ pour $i = 1, \ldots, N$

à partir d'un (grand) échantillon aléatoire d'observation des $(\mathbf{y}_i, \mathbf{x}_i)$ pour $i = 1, \ldots, N$ sachant que :

les variables aléatoires $(\mathbf{y}_i, \mathbf{x}_i)$ sont iid pour $i = 1, \ldots, N$ et de support $\mathcal{W}$.

- La ***fonction de densité (conditionnelle)*** commune des $\mathbf{y}_i \mid \mathbf{x}_i$ en $(\mathbf{y}, \mathbf{x})$ est notée $p_0(\mathbf{y}; \mathbf{x})$.

Modèle paramétrique pour les $\mathbf{y}_i / \mathbf{x}_i$

On suppose que $p_0(.;.)$ est membre d'une ***famille de fonctions de densité paramétrée***, par $\boldsymbol{\theta}$ ici :

$$p_0(.;.) \in \mathcal{F} \equiv \left\{ f(.;.;\boldsymbol{\theta}) : \mathcal{W} \to \mathbb{R}_{++}; \boldsymbol{\theta} \in \Theta \right\}.$$

- ***Avantages*** des modèles paramétriques :
  - ***Identification d'un paramètre*** : on cherche ici $\boldsymbol{\theta}_0$ tel que :

    $$\boldsymbol{\theta}_0 \in \Theta \text{ et } f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0) = p_0(\mathbf{y};\mathbf{x}) \text{ pour tout } (\mathbf{y},\mathbf{x}) \in \mathcal{W}$$

    versus **identification** (***difficile ; pénible***) ***d'une fonction***, $p_0(.;.)$.
  - Cadre d'inférence « évident » : ***Maximum de Vraisemblance***.
  - Certaines familles de distribution ont des propriétés très « pratiques ». *E.g.*, la loi normale (addition et conditionnement) ou les lois à valeurs extrêmes (max-stabilité et emboîtement).

- ***Inconvénient*** des modèles paramétriques :
  - ***Problème de spécification*** : $\mathcal{F}$ est mal choisie, *i.e.* $\mathcal{F}$ ne contient pas $p_0(.;.)$
  - ***En théorie*** : l'inférence statistique sur les caractéristiques statistiques de $\mathbf{y}_i \,|\, \mathbf{x}_i$ sera en général biaisée si « mal-spécification » de $\mathcal{F}$.
    - Voir tout de même, ***Quasi-MV*** (White) ou ***Pseudo-MV*** (Gouriéroux, Monfort et Trognon).
  - ***En pratique*** : modèle paramétrique certainement souvent mal spécifié !
    - Pas trop grave si choix d'une famille paramétrique « large », *i.e.* flexible mais …
    - … si c'est le cas : vraisemblance de forme compliquée, beaucoup de paramètres à estimer, …
    - Bref, gros problèmes de calculs évités par … choix de familles paramétriques simples.

- Problème pointé par les économètres venant de la statistique : Un choix de modèle paramétrique particulier peut résoudre un problème d'identification :
  - Identification dite « par les formes fonctionnelles ». Finalement assez rare mais …
  - … systématique pour les modèles de troncature.

- ***Voies de recherche actuelles*** :
  - Modélisations semi- ou non-paramétriques
  - Travaux sur l'identification non-paramétrique des caractéristiques d'intérêt des modèles, ne serait-ce que pour montrer que le choix d'une famille paramétrique ne fait pas tout
  - Modèles paramétriques à lois mélangées, *i.e*. modèles paramétriques à paramètres aléatoires dont la loi est paramétrique.
    - De loin la « généralisation » la plus utilisée en pratique.

Définition. ***Spécification correcte et identifiabilité***

Le modèle paramétrique des $\mathbf{y}_i \mid \mathbf{x}_i$ est ***correctement spécifié*** par $\mathcal{F}$ s'il existe un ***unique vecteur*** $\boldsymbol{\theta}_0 \in \Theta$ tel que :

$$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0) = p_0(\mathbf{y};\mathbf{x}) \text{ pour tout } (\mathbf{y};\mathbf{x}) \in \mathcal{W}$$

On dit alors que $\boldsymbol{\theta}_0$ ***est identifiable***.

- Le principe du MV découle de l'***inégalité de Kullback-Leibler***, cette dernière indiquant que si :

  (*i*) la distribution des $\mathbf{y}_i \mid \mathbf{x}_i$ est caractérisée par $f(.;.;\boldsymbol{\theta}_0)$

  (*ii*) $\boldsymbol{\theta}_0$ est identifiable

  alors :

$$\boxed{\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]}$$

Propriété. ***Inégalité d'information de Kullback-Leibler***

Pour tout $(\mathbf{x},\boldsymbol{\theta}) \in \mathcal{X} \times \Theta$ on a :

$$E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln p_0(\mathbf{y}_i;\mathbf{x}_i) - \ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] \geq 0.$$

- Bien entendu, si :

$$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0) = p_0(\mathbf{y};\mathbf{x}) \text{ pour tout } (\mathbf{y},\mathbf{x}) \in \mathcal{W}$$

alors on a :

$$E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0) - \ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] \geq 0$$

et, par conséquent (inégalité de Jensen) :

$$\boxed{E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] \leq E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0)\right] \text{ pour tout } (\mathbf{x},\boldsymbol{\theta}) \in \mathcal{X} \times \Theta.}$$

- Inégalité de Kullback-Leibler et entropie :

$$E_{\mathbf{y}_i/\mathbf{x}_i=\mathbf{x}}\left[\ln p_0(\mathbf{y}_i;\mathbf{x}_i) - \ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] = \int_{\mathcal{Y}_{|\mathbf{x}}} \ln\left[\frac{p_0(\mathbf{y};\mathbf{x})}{f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})}\right] p_0(\mathbf{y};\mathbf{x})\mu(d\mathbf{y})$$

sachant que :

$$\int_{\mathcal{Y}_{|\mathbf{x}}} p_0(\mathbf{y};\mathbf{x})\mu(d\mathbf{y}) = \int_{\mathcal{Y}_{|\mathbf{x}}} f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})\mu(d\mathbf{y}) = 1.$$

Propriété. ***Inégalité d'information de Kullback-Leibler et MV(C)***

Les trois propositions suivantes sont équivalentes :

(*i*) Il existe un unique vecteur $\boldsymbol{\theta}_0 \in \Theta$ tel que :

$$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0) = p_0(\mathbf{y};\mathbf{x}) \text{ pour tout } (\mathbf{y};\mathbf{x}) \in \mathcal{W}$$

(*ii*) $\boldsymbol{\theta} \in \Theta$ et $E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] = E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0)\right] \quad \Leftrightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0$

(*iii*) $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]$

- Ces équivalences utilisent l'inégalité de Kullback-Leibler et la loi des conditionnements successifs :

$$E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] = E_{(\mathbf{y}_i,\mathbf{x}_i)}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] = E_{\mathbf{x}_i}\left[E_{\mathbf{y}_i|\mathbf{x}_i}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]\right].$$

Eléments de démonstration et d’interprétation

- On montre aisément que l’inégalité de Kullback-Leibler implique que :

$$E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0)\right]=\sup_{\boldsymbol{\theta}\in\Theta} E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right],$$

  i.e. que $\boldsymbol{\theta}_0$ est un maximande de $E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]$ sur Θ. Il suffit maintenant de montrer que ce maximande est unique.

- La condition (*i*) indique que si $f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})\neq f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0)$ avec $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$ et $(\mathbf{y};\mathbf{x})\in\mathcal{W}$ où $\mathcal{W}$ est une partie mesurable de $\mathcal{W}$ alors l’inégalité de Kullback-Leibler implique alors :

$$\boxed{E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]<E_{\mathbf{y}_i|(\mathbf{x}_i=\mathbf{x})}\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0)\right],}$$

  ce qui implique que :

$$\boxed{E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right]<E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}_0)\right].}$$

$\boxed{\boldsymbol{\theta}_0 \text{ est l'unique maximande de } E\left[\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\right] \text{ sur } \Theta.}$

Principe du Maximum de Vraisemblance

- L'estimateur du MV de $\boldsymbol{\theta}_0$ résulte de l'***application du principe d'analogie*** et de la ***définition de la vraisemblance***.

Définition. Si la densité (conditionnelle) de $\mathbf{y}_i \mid (\mathbf{x}_i = \mathbf{x})$ en $\mathbf{y}$ est donnée par $f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0)$, alors, $f(.;.;\boldsymbol{\theta}_0)$ étant une fonction de densité conditionnelle membre de la famille $\mathcal{F}$, la ***vraisemblance de*** $(\mathbf{y},\mathbf{x})$ ***en*** $\boldsymbol{\theta}\in\Theta$ est donnée par :

$$\ell(\boldsymbol{\theta};\mathbf{y},\mathbf{x}) \equiv f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}).$$

- $\ell(\boldsymbol{\theta};\mathbf{y},\mathbf{x})$ est une fonction de $\boldsymbol{\theta}$ « paramétrée » en $(\mathbf{y},\mathbf{x})$.
- $\ell(\boldsymbol{\theta};\mathbf{y},\mathbf{x})$ s'interprète comme la « probabilité » d'observer $\mathbf{y}$ lorsque la distribution de $\mathbf{y}_i \mid (\mathbf{x}_i = \mathbf{x})$ est caractérisée par la fonction de densité (ou probabilité) conditionnelle $f(.;\mathbf{x};\boldsymbol{\theta})$.

- Si le modèle paramétrique de $\mathbf{y}_i / \mathbf{x}_i$ est correctement spécifié, *i.e.* si sa fonction de densité est $f(.;.;\boldsymbol{\theta}_0)$ qui est un membre de la famille $\mathcal{F}$, et si $\boldsymbol{\theta}_0$ est identifiable alors :

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} E\left[\ln \ell_i(\boldsymbol{\theta})\right]$$

avec :

$$\ell_i(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{x}_i) = f(\mathbf{y}_i; \mathbf{x}_i; \boldsymbol{\theta}).$$

- L'application du ***principe d'analogie*** permet de définir, $\hat{\boldsymbol{\theta}}_N^{MV}$, l'***estimateur du MV de*** $\boldsymbol{\theta}_0$ par :

$$\hat{\boldsymbol{\theta}}_N^{MV} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$$

- Bien entendu on a :

$$\hat{\boldsymbol{\theta}}_N^{MV} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ln \prod_{i=1}^{N} \ell_i(\boldsymbol{\theta})$$

Définition. ***Estimateur du Maximum de Vraisemblance*** (***Conditionnel***)

Si :

(*i*) les variables aléatoires $(\mathbf{y}_i, \mathbf{x}_i)$ sont iid pour $i = 1,\ldots,N$

et :

(*ii*) la fonction de densité commune des $\mathbf{y}_i \mid \mathbf{x}_i$ est l'unique membre de $\mathcal{F}$ défini par $\boldsymbol{\theta}_0$,

alors l'estimateur du MV (C) de $\boldsymbol{\theta}_0$ est défini par :

$$\hat{\boldsymbol{\theta}}_N^{MV} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$$

avec :

$$\ell_i(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{x}_i) \equiv f(\mathbf{y}_i; \mathbf{x}_i; \boldsymbol{\theta}).$$

- ***Interprétation*** *via* l'information de Kullback-Leibler :

$N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$ est la log-vraisemblance moyenne des $(\mathbf{y}_i, \mathbf{x}_i)$ en $\boldsymbol{\theta}$.

- ***Interprétation heuristique*** usuelle de la vraisemblance de l'échantillon en $\boldsymbol{\theta}$ et de l'estimateur du MV :

$$\ell_N(\boldsymbol{\theta}) \equiv \prod_{i=1}^{N} \ell_i(\boldsymbol{\theta})$$

  - $\ell_N(\boldsymbol{\theta})$ mesure la « probabilité » d'observer $\{\mathbf{y}_i; i=1,\dots,N\}$ conditionnellement en $\{\mathbf{x}_i; i=1,\dots,N\}$ si la densité commune des $\mathbf{y}_i / \mathbf{x}_i$ est donnée par $f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}) = \ell_i(\boldsymbol{\theta})$.

  - L'estimateur $\hat{\boldsymbol{\theta}}_N^{MV}$ est la valeur de $\boldsymbol{\theta} \in \Theta$ qui maximise la valeur de $\ell_N(\boldsymbol{\theta})$, *i.e.* qui maximise la « probabilité » d'observer les données recueillies ou la valeur la plus « vraisemblable » de $\boldsymbol{\theta} \in \Theta$ étant donné les données recueillies.

  - La « probabilité » de l'échantillon est le produit des « probabilités » individuelles lorsque les observations sont indépendantes.

Plan

2. Propriétés de l’estimateur du MV(C)

2. Propriétés de l'estimateur du MV(C)

Propriété. ***Propriétés de l'estimateur du MV(C)***

Si :

(i) les variables aléatoires $(\mathbf{y}_i, \mathbf{x}_i)$ sont iid pour $i = 1,\ldots,N$

(ii) la fonction de densité des $\mathbf{y}_i / \mathbf{x}_i$ est l'unique membre de $\mathcal{F}$ défini par $\boldsymbol{\theta}_0$,

alors, sous certaines conditions de régularité, l'estimateur du MV(C) de $\boldsymbol{\theta}_0$:

$$\hat{\boldsymbol{\theta}}_N^{MV} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$$

vérifie :

(a) $$\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{MV} - \boldsymbol{\theta}_0) \xrightarrow[N\to+\infty]{L} \mathcal{N}\left(\mathbf{0}, -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}\right) \text{ avec } \mathbf{H}_i(\boldsymbol{\theta}) \equiv \frac{\partial^2 \ln \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

et :

(b) L'estimateur $\hat{\boldsymbol{\theta}}_N^{MV}$ est asymptotiquement efficace pour $\boldsymbol{\theta}_0$.

- $\hat{\boldsymbol{\theta}}_N^{MV}$ est ***as. normal***, il est donc ***convergent***.

- Avec $\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{MV} - \boldsymbol{\theta}_0) \xrightarrow[N\to+\infty]{L} \mathcal{N}\left(\mathbf{0}, -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}\right)$ lorsque « N est grand mais pas infini » :

$$\boxed{\hat{\boldsymbol{\theta}}_N^{MV} \sim_{app.} \mathcal{N}\left(\boldsymbol{\theta}_0, -N^{-1}E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}\right).}$$

- $\hat{\boldsymbol{\theta}}_N^{MV}$ est as. efficace. De fait, le calcul de $\hat{\boldsymbol{\theta}}_N^{MV}$ mobilise l'ensemble de l'information contenue dans le modèle des $\mathbf{y}_i / \mathbf{x}_i$.

- Maximum de Vraisemblance Conditionnel en $\mathbf{x}_i$: on n'utilise pas la loi de $\mathbf{x}_i$ qui n'est pas spécifiée.
  - On peut (parfois) obtenir un estimateur de plus efficace de $\boldsymbol{\theta}_0$ que $\hat{\boldsymbol{\theta}}_N^{MV}$ si on dispose de la loi jointe de $(\mathbf{y}_i, \mathbf{x}_i)$.

- Le terme $-E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$ est la ***matrice d'information de Fisher***, qui est d'autant plus grande que $\hat{\boldsymbol{\theta}}_N^{MV}$ est (as.) « précis ».
  - $E[\ln \ell_i(\boldsymbol{\theta}_0)]$, est d'autant plus « concave » en $\boldsymbol{\theta}$ au point $\boldsymbol{\theta}_0$ que $-E[\mathbf{H}_i(\boldsymbol{\theta}_0)]$ est grande (et donc que $-E[\mathbf{H}_i(\boldsymbol{\theta}_0)]^{-1}$ est petite).
  - $\boldsymbol{\theta}_0$ est d'autant plus facile à identifier (à « localiser ») que $E[\ln \ell_i(\boldsymbol{\theta}_0)]$ est concave en $\boldsymbol{\theta}$ au voisinage de $\boldsymbol{\theta}_0$

Propriété. ***Egalité des matrices d'information (de Fisher)***

Sous certaines conditions de régularité, on a :

$$-E[\mathbf{H}_i(\boldsymbol{\theta}_0)] = E[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0)]$$

ou :

$$-E\left[\frac{\partial^2 \ln \ell_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = E\left[\frac{\partial \ln \ell_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\frac{\partial \ln \ell_i(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right].$$

Propriété. ***Estimation de la variance as. de*** $\hat{\boldsymbol{\theta}}_N^{MV}$

Sous certaines conditions de régularité, on a :

$$\left[-N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\boldsymbol{\theta}}_N^{MV})\right]^{-1} \xrightarrow[N\to+\infty]{p} -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}$$

et :

$$\left[N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}}_N^{MV})\mathbf{s}_i'(\hat{\boldsymbol{\theta}}_N^{MV})\right]^{-1} \xrightarrow[N\to+\infty]{p} E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0)\right]^{-1}.$$

- ***Conditions de régularité*** importantes :
  - Continuité de $\ln \ell_i(\boldsymbol{\theta})$, et de ses dérivées, en $\boldsymbol{\theta} \in \Theta$. Problèmes avec les vraisemblances intégrées par simulation.
  - $E\left[\ln \ell_i(\boldsymbol{\theta})\right]$ peut n'être bornée que sur un Θ compact.
  - Pour l'égalité des matrices d'information : le support des $\mathbf{y}_i \mid \mathbf{x}_i$ ne doit pas dépendre de $\boldsymbol{\theta}$.

- Le terme :

$$\boxed{\mathbf{s}_i(\boldsymbol{\theta}) \equiv \frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}$$

  est le ***score*** de l'observation $(\mathbf{y}_i, \mathbf{x}_i)$ en $\boldsymbol{\theta}$.

- $\hat{\boldsymbol{\theta}}_N^{MV}$ est caractérisé par les ***équations normales*** (FOC du programme de maximisation de l'espérance de la log-vraisemblance) :

$$\boxed{\hat{\boldsymbol{\theta}}_N^{MV} \text{ est solution en } \boldsymbol{\theta} \in \Theta \text{ de } N^{-1}\sum_{i=1}^{N} \frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.}$$

- $\hat{\boldsymbol{\theta}}_N^{MV}$ est l'unique solution en $\boldsymbol{\theta} \in \Theta$ de ce système d'équation si $N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$ est strictement concave en $\boldsymbol{\theta} \in \Theta$.

- L'astuce suivante s'avère pratique pour calculer la distribution as. de certains estimateurs du MV(C) :

$$E\left[\mathbf{H}_i(\boldsymbol{\theta})\right] = E_{\mathbf{x}_i}\left[E\left[\mathbf{H}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i\right]\right]$$

et :

$$E\left[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i'(\boldsymbol{\theta})\right] = E_{\mathbf{x}_i}\left[E\left[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i'(\boldsymbol{\theta}) \mid \mathbf{x}_i\right]\right]$$

en effet les termes :

$$E\left[\mathbf{H}_i(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right] \quad \text{et} \quad E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right]$$

ont parfois des expressions beaucoup plus simples que :

$$\mathbf{H}_i(\boldsymbol{\theta}_0) \quad \text{et} \quad \mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0).$$

- En outre on a :

Propriété. ***Egalité des matrices d'information conditionnelles***

Sous certaines conditions de régularité, on a :

$$-E\left[\mathbf{H}_i(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right] = E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right].$$

de qui donne également :

$$E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right] = E_{\mathbf{x}_i}\left[E\left[\mathbf{H}_i(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right]\right] = E_{\mathbf{x}_i}\left[E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i'(\boldsymbol{\theta}_0) \mid \mathbf{x}_i\right]\right]$$

Propriété. ***Estimation de*** $E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]$

Sous certaines conditions de régularité, on a :

$$\left[-N^{-1}\sum_{i=1}^{N}\bar{\mathbf{H}}_i(\hat{\boldsymbol{\theta}}_N^{MV})\right]^{-1} \xrightarrow[N \to +\infty]{p} -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1}$$

où :

$$\bar{\mathbf{H}}_i(\boldsymbol{\theta}) \equiv E\left[\mathbf{H}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i\right].$$

Problèmes pratiques

- Les fonctions $E\left[\ln \ell_i(\boldsymbol{\theta})\right]$, et donc $N^{-1}\sum_{i=1}^{N}\ln \ell_i(\boldsymbol{\theta})$, peuvent avoir plusieurs *optima* (*maxima* et *minima*) en $\boldsymbol{\theta} \in \Theta$. Après calcul de l'estimation par le logiciel :
  - Vérifier qu'on est sur ***un maximum local***, ce qui est en général garantit par les logiciels usuels.
  - Vérifier qu'on est sur ***le maximum global***, en recommençant le calcul avec des points initiaux différents.
    - Toujours même estimation, ou à peu près : OK.
    - Estimations différentes : choisir celles qui donne la vraisemblance maximale (!)

- Faire attention avec les ***fonctions exponentielles*** :
  - « Grosses » exponentielles
    - gourmandes en temps de calcul (et mémoire)
    - peuvent « exploser » (*overflow*).
  - Travailler avec des données (**y** et **x**) avec les bonnes échelles
  - Eviter, sinon stocker, au maximum les exponentielles calculées dans les programmes de maximisation de log-vraisemblance.

- Comme d'habitude : un ***problème d'identification de*** $\boldsymbol{\theta}_0$ est révélé par :
  - Des matrices $-N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\boldsymbol{\theta}}_N^{MV})$ ou $N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}}_N^{MV})\mathbf{s}_i'(\hat{\boldsymbol{\theta}}_N^{MV})$ non inversibles ou « énormes »
  - Des algorithmes d'optimisation qui ne convergent pas, ou mal

Plan

3. Liens avec la Méthode des Moments

3. Liens avec la Méthode des Moments

$$\hat{\boldsymbol{\theta}}_N^{MV} \text{ est solution en } \boldsymbol{\theta} \in \Theta \text{ de } N^{-1}\sum_{i=1}^{N} \frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

- En utilisant à rebours le principe d'analogie :

$$\boldsymbol{\theta}_0 \text{ est solution en } \boldsymbol{\theta} \in \Theta \text{ de } E\left[\frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = E\left[\mathbf{s}_i(\boldsymbol{\theta})\right] = \mathbf{0}.$$

$$E\left[\frac{\partial \ln \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \text{ est la CO1 du programme } \max_{\boldsymbol{\theta} \in \Theta} E\left[\ln \ell_i(\boldsymbol{\theta})\right]$$

- La méthode consistant à définir l'estimateur de $\boldsymbol{\theta}_0$ comme la solution de l'équation de score s'appelle la ***Méthode des Scores*** (***MS***) :

$$\boxed{\hat{\boldsymbol{\theta}}_N^{MS} \text{ est solution en } \boldsymbol{\theta} \in \Theta \text{ de } N^{-1}\sum_{i=1}^{N} \frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.}$$

L'estimateur correspondant est l'***Estimateur de la Méthode des Scores***.

- Un ***estimateur de la Méthode des Scores*** est :
  - un ***estimateur de la MM***, ce qui est pratique pour déterminer la distribution as. des estimateurs par étapes
  - un ***estimateur du MV*** s'il maximise la log-vraisemblance du modèle, ce qui suppose qu'il est solution de l'équation des scores.

- On peut déterminer la distribution as. de l'estimateur du MV à partir des résultats disponibles pour les estimateurs de la MM(G) et de la propriété d'égalité des matrices d'information.

- En tant qu'estimateur de la MS l'estimateur du MV est l'estimateur de la MM (cas juste-identifié) de fondé sur :

$$E\left[\frac{\partial \ln \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right] = E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\right] = \mathbf{0},$$

le principe du MV sélectionnant, si besoin, le maximum global de la log-vraisemblance parmi l'ensemble des maxima locaux.

- L'application des résultats sur la MM(G) donne alors que :

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N^{MV} - \boldsymbol{\theta}_0) \xrightarrow[N\to+\infty]{L} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_0\right)$$

avec :

$$\boldsymbol{\Sigma}_0 = \left(E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right] E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right]^{-1} E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]\right)^{-1}$$

et :

$$\boldsymbol{\Sigma}_0 = E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1} E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right] E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1},$$

grâce à la juste-identification (MM).

- L’égalité des matrices d’information :

$$E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right] = -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]$$

donne alors que :

$$\boxed{\begin{aligned}\boldsymbol{\Sigma}_0 &= E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1} E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right] E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1} \\ &= -E\left[\mathbf{H}_i(\boldsymbol{\theta}_0)\right]^{-1} \\ &= -E\left[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\right]^{-1}\end{aligned}}$$

Plan

4. Exemples

4. Exemples

- Modèle de Régression Linéaire Gaussien
- Modèle Probit et Logit : choix dichotomique
- Modèle Tobit Généralisé : modèle de variable censurée ou de solution en coin
- Chapitre suivant : Système de Modèles de Régression Linéaires Gaussien
- Chapitre suivant : Système d'Equations Simultanées Linéaires Gaussiennes

Modèle de Régression Linéaire Gaussien

$$y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i \quad \text{avec} \quad u_i / \mathbf{x}_i \sim \mathcal{N}(0;\sigma_0^2)$$

ou

$$y_i / \mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i'\mathbf{a}_0;\sigma_0^2)$$

- Densité de $y_i / \mathbf{x}_i$ en $(y,\mathbf{x})$:

$$f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = \sigma_0^{-1} \times \varphi(\sigma_0^{-1}\mathbf{a}_0'\mathbf{x})$$

- Log-vraisemblance de $(\mathbf{a},\sigma)$ en $(y_i,\mathbf{x}_i)$:

$$\ln \ell_i(\mathbf{a},\sigma) = -\ln\sigma + \ln\varphi(\sigma^{-1}\mathbf{a}'\mathbf{x})$$

Propriété. ***Fonctions de densité et changement de variable***

Si (i) $f_{\tilde{e}|\tilde{\mathbf{x}}}(e;\mathbf{x})$ est la fonction de densité de $\tilde{e}\,|\,\tilde{\mathbf{x}}=\mathbf{x}$ en e, (ii) $\tilde{y}=g(\tilde{e},\tilde{\mathbf{x}})$ et (iii) $g(e,\mathbf{x})$ est bijective en e pour $(e,\mathbf{x})\in \mathcal{S}_{(\tilde{e},\tilde{\mathbf{x}})}$ avec pour fonction réciproque $g^{(-y)}(y,\mathbf{x})$, alors la fonction de densité de $\tilde{y}\,|\,\tilde{\mathbf{x}}=\mathbf{x}$ en $\mathbf{y}$, $f_{\tilde{y}|\tilde{\mathbf{x}}}(y;\mathbf{x})$, est définie par :

$$f_{\tilde{y}|\tilde{\mathbf{x}}}(y;\mathbf{x})=\left|\frac{\partial g\left(g^{(-y)}(y,\mathbf{x})\right)}{\partial e}\right|^{-1} f_{\tilde{e}|\tilde{\mathbf{x}}}\left(g^{(-y)}(y,\mathbf{x});\mathbf{x}\right)$$

$$=\left|\frac{\partial g^{(-y)}(y,\mathbf{x})}{\partial y}\right| f_{\tilde{e}|\tilde{\mathbf{x}}}\left(g^{(-y)}(y,\mathbf{x});\mathbf{x}\right).$$

Ici : $y=g(e;\mathbf{x})=\mathbf{a}'\mathbf{x}+\sigma\times e$ et $g^{(-y)}(y;\mathbf{x})=\sigma^{-1}\times(y-\mathbf{a}'\mathbf{x})$

Propriété. ***Fonctions de densité et changement de variable***

Si :

(*i*) $f_{\tilde{\mathbf{u}}|\tilde{\mathbf{x}}}(\mathbf{u};\mathbf{x})$ est la fonction de densité de $\tilde{\mathbf{u}}\,|\,\tilde{\mathbf{x}}=\mathbf{x}$ en $\mathbf{u}$,

(*ii*) $\tilde{\mathbf{y}}=\mathbf{g}(\tilde{\mathbf{u}},\tilde{\mathbf{x}})$

et :

(*iii*) $\mathbf{g}(\mathbf{u},\mathbf{x})$ est bijective en $\mathbf{u}$ pour $(\mathbf{u},\mathbf{x})\in\mathcal{S}_{(\tilde{\mathbf{u}},\tilde{\mathbf{x}})}$ avec pour fonction réciproque $\mathbf{g}^{(-\mathbf{y})}(\mathbf{y},\mathbf{x})$,

alors la fonction de densité de $\tilde{\mathbf{y}}\,|\,\tilde{\mathbf{x}}=\mathbf{x}$ en $\mathbf{y}$, $f_{\tilde{\mathbf{y}}|\tilde{\mathbf{x}}}(\mathbf{y};\mathbf{x})$, est définie par :

$$\begin{aligned}
f_{\tilde{\mathbf{y}}|\tilde{\mathbf{x}}}(\mathbf{y};\mathbf{x}) &= \left|\det\left[\frac{\partial\mathbf{g}\left(\mathbf{g}^{(-\mathbf{y})}(\mathbf{y},\mathbf{x})\right)}{\partial\mathbf{u}'}\right]\right|^{-1} f_{\tilde{\mathbf{u}}|\tilde{\mathbf{x}}}\left(\mathbf{g}^{(-\mathbf{y})}(\mathbf{y},\mathbf{x});\mathbf{x}\right) \\
&= \left|\det\left[\frac{\partial\mathbf{g}^{(-\mathbf{y})}(\mathbf{y},\mathbf{x})}{\partial\mathbf{y}}\right]\right| f_{\tilde{\mathbf{u}}|\tilde{\mathbf{x}}}\left(\mathbf{g}^{(-\mathbf{y})}(\mathbf{y},\mathbf{x});\mathbf{x}\right)
\end{aligned}.$$

Modèles Probit et Logit

Modèle Probit

$$\begin{vmatrix} y_i = 1 \text{ si } y_i^* > 0 \\ y_i = 0 \text{ sinon} \end{vmatrix} \quad \text{avec} \quad y_i^* = \mathbf{x}_i'\mathbf{a}_0 + u_i \ \text{ et } \ u_i / \mathbf{x}_i \sim \mathcal{N}(0;\sigma_0^2)$$

Modèle Logit

$$\begin{vmatrix} y_i = 1 \text{ si } y_i^* > 0 \\ y_i = 0 \text{ sinon} \end{vmatrix} \quad \text{avec} \quad y_i^* = \mathbf{x}_i'\mathbf{a}_0 + u_i \ \text{ et } \ u_i / \mathbf{x}_i \sim \mathcal{L}$$

- Log-vraisemblance de $(\mathbf{a},\sigma)$ en $(y_i, \mathbf{x}_i)$:

Probit

$$\ln \ell_i(\mathbf{a}\sigma^{-1}) = y_i \times \ln \Phi(\sigma^{-1}\mathbf{a}'\mathbf{x}) + (1-y_i)\times \ln\left(1-\Phi(\sigma^{-1}\mathbf{a}'\mathbf{x})\right)$$

Remarque. $\mathbf{a}_0\sigma_0^{-1}$ identifiable, mais pas $\mathbf{a}_0$ et σ_0 séparément. Souvent normalisation par $\sigma_0 \equiv 1$.
Mais on peut également choisir $a_{k,0} \equiv c$ pour un *k* et un *c* choisis.

Logit

$$\begin{aligned}\ln \ell_i(\mathbf{a}) &= y_i \times \ln\left[\exp(\mathbf{x}_i'\mathbf{a}_0)\left(1+\exp(\mathbf{x}_i'\mathbf{a}_0)\right)^{-1}\right] \\ &\quad + (1-y_i)\times \ln\left[\left(1+\exp(\mathbf{x}_i'\mathbf{a}_0)\right)^{-1}\right] \\ &= y_i \times \mathbf{x}_i'\mathbf{a}_0 - \ln\left(1+\exp(\mathbf{x}_i'\mathbf{a}_0)\right)\end{aligned}$$

Remarque

L'« échelle » de mesure de y_i^* est fixée par le choix de $u_i/\mathbf{x}_i \sim \mathcal{N}(0;1)$ ou $u_i/\mathbf{x}_i \sim \mathcal{L}$.

Les valeurs des éléments de $\mathbf{a}_0$ ne sont donc pas interprétables dans l'absolu.

Cette remarque ne s'applique pas au Probit si on choisit une normalisation telle que $a_{k,0} \equiv c$ pour un *k* et un *c* choisis,

i.e. si on estime σ_0 dans $u_i/\mathbf{x}_i \sim \mathcal{N}(0;\sigma_0)$.

Modèle Tobit Généralisé (Heckit)
Modèle de variable censurée, de solution en coin et …

$$\begin{cases} y_{1i} = y_{1i}^* & \text{si } y_{2i}^* > 0 \\ y_{1i} = m & \text{si } y_{2i}^* \leq 0 \end{cases} \quad \text{et} \quad \begin{bmatrix} y_{1i}^* \\ y_{2i}^* \end{bmatrix} \Big/ \mathbf{x}_i \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_{1,i}' \mathbf{a}_{1,0} \\ \mathbf{x}_{2,i}' \mathbf{a}_{2,0} \end{bmatrix} ; \begin{bmatrix} \sigma_{1,0}^2 & \rho_0 \sigma_{1,0} \sigma_{2,0} \\ \rho_0 \sigma_{1,0} \sigma_{2,0} & \sigma_{2,0}^2 \end{bmatrix} \right)$$

$\mathbf{x}_i$ toujours observée, $\mathbf{x}_{1,i}$ et $\mathbf{x}_{2,i}$ sont des sous-vecteurs de $\mathbf{x}_i$.

- On pose souvent $\sigma_{2,0} \equiv 1$ et il est recommandé que $\mathbf{x}_i = \mathbf{x}_{2,i} = (\mathbf{x}_{1,i}, \mathbf{z}_i)$.
- L'égalité $y_{1i} = m$ est essentiellement un choix de codage si censure.
- Le modèle Tobit (simple) est un cas particulier du modèle Tobit généralisé caractérisé par :

$$y_{1,i}^* = y_{2,i}^* \text{ et donc par } \rho_0 = 1 \text{ et } \sigma_{1,0} = \sigma_{2,0}.$$

Modèles associés au modèle Tobit généralisé

Modèle de régression Gaussienne : ***variable latente d'intérêt***
$y_{1,i}^{*} \big/ \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}'_{1,i}\mathbf{a}_{1,0};\sigma_{1,0}^{2}\right) \quad \Leftrightarrow \quad y_{1,i}^{*} = \mathbf{x}'_{1,i}\mathbf{a}_{1,0} + u_{1,i}$ avec $u_{1,i} \big/ \mathbf{x}_i \sim \mathcal{N}\left(0;\sigma_{1,0}^{2}\right)$

Modèle Probit associé au modèle Tobit : ***mécanisme de sélection***
$\left\| \begin{array}{l} s_i = 1 \quad \text{si } y_{2,i}^{*} > 0 \\ s_i = 0 \quad \text{si } y_{2,i}^{*} \leq 0 \end{array} \right.$ et $y_{2,i}^{*} \big/ \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}'_{2,i}\mathbf{a}_{2,0};\sigma_{2,0}^{2}\right)$

Modèle Tobit généralisé $(\sigma_{2,0} \equiv 1)$

$$\left| \begin{array}{l} y_{1i} = y_{1i}^* \quad \text{si} \quad s_i = 1 \\ y_{1i} = m \quad \text{si} \quad s_i = 0 \end{array} \right. \quad \text{et} \quad \left| \begin{array}{l} s_i = 1 \quad \text{si} \quad y_{2i}^* > 0 \\ s_i = 0 \quad \text{si} \quad y_{2i}^* \leq 0 \end{array} \right.$$

avec :

$$\begin{bmatrix} y_{1,i}^* \\ y_{2,i}^* \end{bmatrix} \Big/ \mathbf{x}_i \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{x}'_{1,i}\mathbf{a}_{1,0} \\ \mathbf{x}'_{2,i}\mathbf{a}_{2,0} \end{bmatrix} ; \begin{bmatrix} \sigma_{1,0}^2 & \rho_0\sigma_{1,0} \\ \rho_0\sigma_{1,0} & 1 \end{bmatrix} \right), \quad \mathbf{x}_i \text{ toujours observée}$$

On modélise ici le couple de variables observées (y_{1i}, s_i) conditionnellement en $\mathbf{x}_i$

- Log-vraisemblance du modèle Tobit généralisé :

$$\ln \ell_i(\boldsymbol{\theta}) = (1-s_i)\times\ln\left(1-\Phi(\mathbf{x}'_{2,i}\mathbf{a}_2)\right) + s_i\times\left\{\begin{array}{c} -\ln\sigma_1 + \ln\varphi\left(\sigma_1^{-1}(y_{1,i}-\mathbf{x}'_{1,i}\mathbf{a}_1)\right) \\ +\ln\Phi\left[(1-\rho^2)^{-1/2}\left(\mathbf{x}'_{2,i}\mathbf{a}_2+\rho\sigma_1^{-1}(y_{1,i}-\mathbf{x}'_{1,i}\mathbf{a}_1)\right)\right] \end{array}\right\}$$

avec

$$\boldsymbol{\theta} \equiv (\mathbf{a}_1, \mathbf{a}_2, \rho, \sigma_1)$$

Plan

5. Quelques résultats utiles

5. Quelques résultats utiles

- **Rappels : *Loi normale*** :

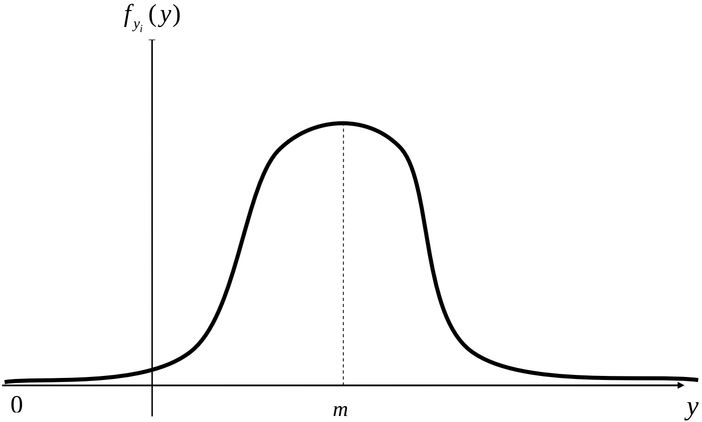
  - Si $y_i \sim \mathcal{N}(m;\omega)$ alors :

$$\boxed{f_{y_i}(y) = \omega^{-1/2}\varphi\left[\frac{y-m}{\omega^{1/2}}\right] = \frac{1}{\sqrt{\omega 2\pi}}\exp\left[-\frac{1}{2}\frac{(y-m)^2}{\omega}\right].}$$

  - Si $y_i \sim \mathcal{N}(m;\omega)$ alors :

$$y_i = m + u_i \text{ avec } u_i \sim \mathcal{N}(0;\omega).$$

Loi normale



- ***Loi normale censurée*** :
  - On a :

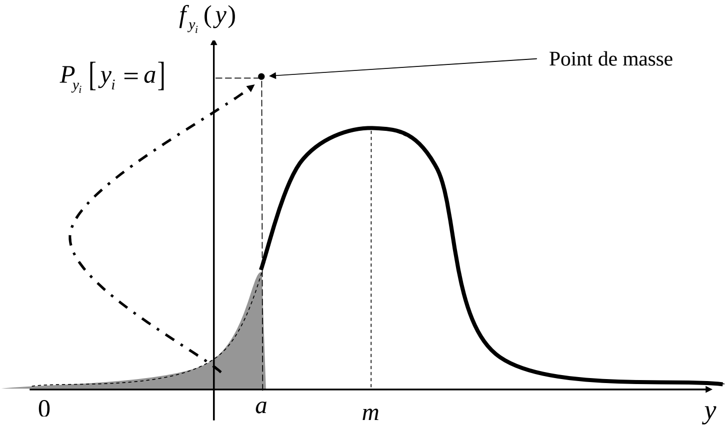
$$y_i = \begin{cases} y_i^* & \text{si } y_i^* > a \\ a & \text{si } y_i^* \le a \end{cases} \quad \text{ou } y_i = y_i^* \times \mathbf{1}(y_i^* > a) + a \times \mathbf{1}(y_i^* \le a)$$

avec $y_i^* \sim \mathcal{N}\left(m;\omega\right)$ alors :

$$f_{y_i}(y) = \begin{cases} f_{y_i^*}(y) = \varphi\left[\dfrac{y-m}{\omega^{1/2}}\right] & \text{si } y > a \\ F_{y_i^*}(a) = \Phi\left[\dfrac{a-m}{\omega^{1/2}}\right] & \text{si } y = a \\ 0 & \text{si } y < a) \end{cases}$$

$$P_{y_i}\left[y_i = a\right] = P_{y_i^*}\left[y_i^* \le a\right] = F_{y_i^*}(a) = \Phi\left[\dfrac{a-m}{\omega^{1/2}}\right].$$

Loi normale censurée sous a à a



- ***Loi normale tronquée*** :

Si $y_i^* \sim \mathcal{N}\left(m;\omega\right)$ et $y_i = y_i^* \mid (y_i^* > a)$ alors :

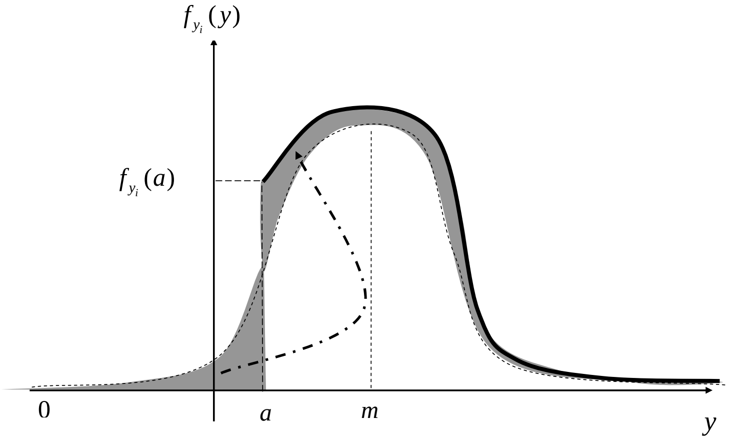
$$f_{y_i}(y) = \begin{cases} \left(1 - F_{y_i^*}(a)\right)^{-1} f_{y_i^*}(y) & \text{si } y > a \\ 0 & \text{si } y \le a \end{cases}$$

ou :

$$f_{y_i}(y) = \begin{cases} \left\{1 - \Phi\left[\dfrac{a-m}{\omega^{1/2}}\right]\right\}^{-1} \varphi\left[\dfrac{y-m}{\omega^{1/2}}\right] & \text{si } y > a \\ 0 & \text{si } y \le a \end{cases}$$

$$P_{y_i^*}\left[y_i^* > a\right] = 1 - F_{y_i^*}(a) = 1 - \Phi\left[\frac{a-m}{\omega^{1/2}}\right].$$

Loi normale tronquée sous *a*



Calcul de la densité de $y_i = y_i^* \mid (y_i^* > a)$: *passer par* $F_{y_i}(y)$ *puis dériver*

- On a évidemment $f_{y_i}(y) = f_{y_i^* \mid (y_i^* > a)}(y) = 0$ si $y \leq a$.
- Si $y > a$, on a :

$$F_{y_i}(y) = P_{y_i}\left[y_i \leq y\right] = P_{y_i^* \mid (y_i^* > a)}\left[y_i^* \leq y\right] \text{ (définition de } y_i\text{)}$$

$$= \frac{P_{y_i^*}\left[a < y_i^* \leq y\right]}{P_{y_i^*}\left[a < y_i^*\right]} \text{ (loi de Bayes)}$$

$$= \frac{F_{y_i^*}(y) - F_{y_i^*}(a)}{1 - F_{y_i^*}(a)}$$

et :

$$\boxed{f_{y_i}(y) = \frac{\partial F_{y_i}(y)}{\partial y} = \frac{f_{y_i^*}(y)}{1 - F_{y_i^*}(a)}.}$$

- **Rappels :** ***Loi normale conditionnelle en*** **x** :

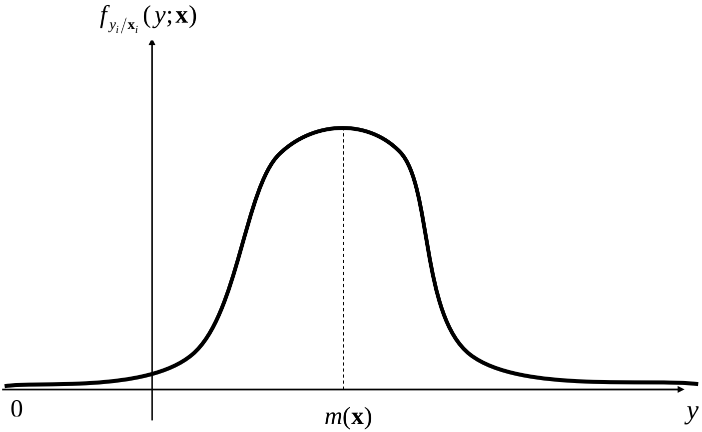
  - Si $y_i / \mathbf{x}_i = \mathbf{x} \sim \mathcal{N}\left(m(\mathbf{x}); \omega(\mathbf{x})\right)$ alors :

$$\boxed{\begin{aligned} f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) &= \omega(\mathbf{x})^{-1/2}\varphi\left[\frac{y-m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right] \\ &= \frac{1}{\sqrt{\omega(\mathbf{x})2\pi}}\exp\left[-\frac{1}{2}\frac{(y-m(\mathbf{x}))^2}{\omega(\mathbf{x})}\right]. \end{aligned}}$$

  - Si $y_i / \mathbf{x}_i = \mathbf{x} \sim \mathcal{N}\left(m(\mathbf{x}); \omega(\mathbf{x})\right)$ alors :

$$y_i = m(\mathbf{x}_i) + u_i \text{ avec } u_i / \mathbf{x}_i = \mathbf{x} \sim \mathcal{N}\left(0; \omega(\mathbf{x})\right).$$

Loi normale



- ***Loi normale censurée*** :
  - Si $y_i^* \big/ \mathbf{x}_i = \mathbf{x} \sim \mathcal{N}\left(m(\mathbf{x});\omega(\mathbf{x})\right)$ et $y_i = \begin{cases} y_i^* & \text{si } y_i^* > a(\mathbf{x}_i) \\ a(\mathbf{x}_i) & \text{si } y_i^* \leq a(\mathbf{x}_i) \end{cases}$

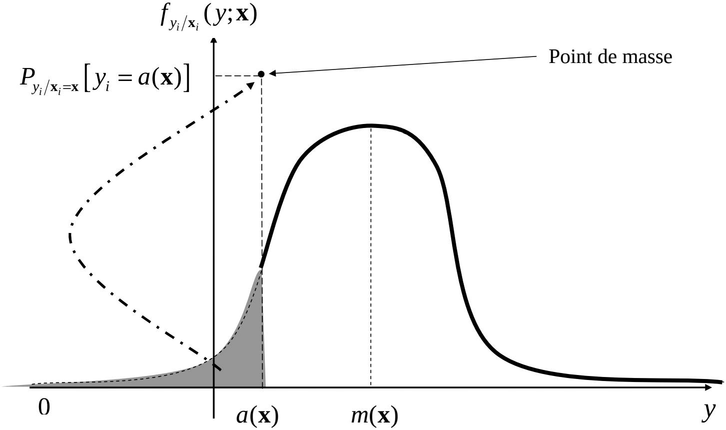
    alors :

$$
f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = \begin{cases} f_{y_i^*/\mathbf{x}_i}(y;\mathbf{x}) = \varphi\left[\dfrac{y - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right] & \text{si } y > a(\mathbf{x}) \\ F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x}\right] = \Phi\left[\dfrac{a(\mathbf{x}) - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right] & \text{si } y = a(\mathbf{x}) \\ 0 & \text{si } y < a(\mathbf{x}) \end{cases}
$$

car :

$$
P_{y_i/\mathbf{x}_i=\mathbf{x}}\left[y_i = a(\mathbf{x}_i)\right] = P_{y_i^*/\mathbf{x}_i=\mathbf{x}}\left[y_i^* \leq a(\mathbf{x}_i)\right] = F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x}\right] = \Phi\left[\frac{a(\mathbf{x}) - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right].
$$

Loi normale censurée sous $a(\mathbf{x})$ à $a(\mathbf{x})$



- ***Loi normale tronquée*** :

Si $y_i^* \big/ \mathbf{x}_i = \mathbf{x} \sim \mathcal{N}\left(m(\mathbf{x});\omega(\mathbf{x})\right)$ et $y_i = y_i^* \Big/ \left(\mathbf{x}_i, y_i^* > a(\mathbf{x}_i)\right)$ alors :

$$f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = \begin{cases} \left\{1 - F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x}\right]\right\}^{-1} f_{y_i^*/\mathbf{x}_i}(y;\mathbf{x}) & \text{si } y > a(\mathbf{x}) \\ 0 & \text{si } y \le a(\mathbf{x}) \end{cases}$$

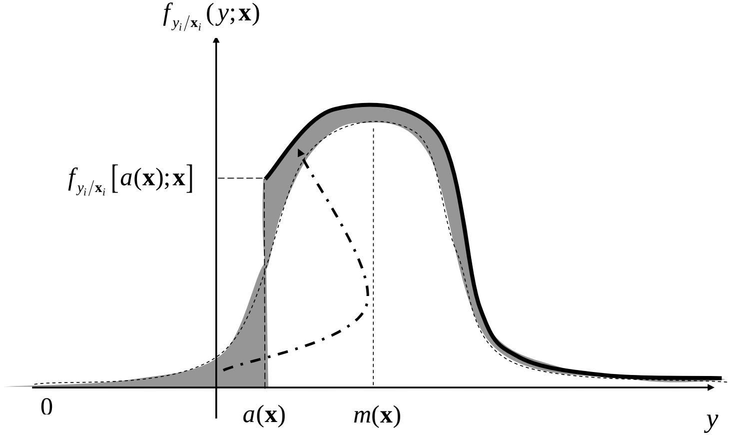
ou :

$$f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = \begin{cases} \left\{1 - \Phi\left[\dfrac{a(\mathbf{x}) - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right]\right\}^{-1} \varphi\left[\dfrac{y - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right] & \text{si } y > a(\mathbf{x}) \\ 0 & \text{si } y \le a(\mathbf{x}) \end{cases}$$

car :

$$P_{y_i^*/\mathbf{x}_i=\mathbf{x}}\left[y_i^* > a(\mathbf{x}_i)\right] = 1 - F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x}\right] = 1 - \Phi\left[\frac{a(\mathbf{x}) - m(\mathbf{x})}{\omega(\mathbf{x})^{1/2}}\right].$$

Loi normale tronquée sous $a(\mathbf{x})$



Calcul de la densité de $y_i = y_i^* \big/ \left(\mathbf{x}_i = \mathbf{x}, y_i^* > a(\mathbf{x}_i) \right)$

- On a évidemment $f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = f_{y_i^*\big/\left(\mathbf{x}_i, y_i^* > a(\mathbf{x}_i)\right)}(y;\mathbf{x}) = 0$ si $y \le a(\mathbf{x})$.

- Si $y > a(\mathbf{x})$, on a :

$$F_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = P_{y_i/\mathbf{x}_i=\mathbf{x}}\left[y_i \le y \right] = P_{y_i^*\big/\left(\mathbf{x}_i=\mathbf{x}, y_i^* > a(\mathbf{x})\right)}\left[y_i^* \le y \right] \text{ (définition de } y_i \big/ \mathbf{x}_i \text{)}$$

$$= \frac{P_{y_i^*/\mathbf{x}_i=\mathbf{x}}\left[a(\mathbf{x}) < y_i^* \le y \right]}{P_{y_i^*/\mathbf{x}_i=\mathbf{x}}\left[a(\mathbf{x}) < y_i^* \right]} \text{ (loi de Bayes)}$$

$$= \frac{F_{y_i^*/\mathbf{x}_i}(y;\mathbf{x}) - F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x} \right]}{1 - F_{y_i^*/\mathbf{x}_i}\left[a(\mathbf{x});\mathbf{x} \right]}$$

et :

$$\boxed{f_{y_i/\mathbf{x}_i}(y;\mathbf{x}) = \frac{\partial F_{y_i/\mathbf{x}_i}(y;\mathbf{x})}{\partial y} = \frac{f_{y_i^*/\mathbf{x}_i}(y;\mathbf{x})}{1 - F_{y_i^*/\mathbf{x}_i}(a(\mathbf{x});\mathbf{x})}.}$$

Plan

6. MV Conditionnelle versus MV

6. MV Conditionnelle *versus* MV

- Dans l'immense majorité des cas : utilisation du MVC
  - En effet : très souvent maximiser la log-vraisemblance des observations (MV) revient à maximiser la log-vraisemblance conditionnelle des observations (MVC)

- Dans certains cas, notamment en cas de stratification des échantillons, l'estimation par le MV est plus efficace que celle par le MVC.
  - Dans ces cas : la distribution des variables de conditionnement contient de l'information sur le paramètre à estimer $\boldsymbol{\theta}_0$.
    - Voir, *e.g.*, Imbens (1992) et, Imbens et Lancaster (1996)
    - Liens avec l'exogénéité au sens de Engle, Hendry et Richard (1983)

- ***Logique ici*** : on s'est concentré sur $\mathcal{D}(\mathbf{y}_i \mid \mathbf{x}_i)$ qu'on a caractérisé par $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0)$.
    - $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0)$: densité de $\mathbf{y}_i$ en $\mathbf{y}$ conditionnelle en $\mathbf{y}_i \mid \mathbf{x}_i = \mathbf{x}$
    - Donne :

$$\ell_i^c(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta};\mathbf{y}_i \mid \mathbf{x}_i) = f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$$

vraisemblance en $\boldsymbol{\theta}$ ***de*** $\mathbf{y}_i$ ***sachant*** $\mathbf{x}_i$.

- On peut considérer $\mathcal{D}\big((\mathbf{y}_i,\mathbf{x}_i)\big)$ alors caractérisée par $f_{(\mathbf{y},\mathbf{x})}(\mathbf{y},\mathbf{x};\boldsymbol{\theta}_0,\boldsymbol{\delta}_0)$.
    - $f_{(\mathbf{y},\mathbf{x})}(\mathbf{y},\mathbf{x};\boldsymbol{\theta}_0,\boldsymbol{\delta}_0)$: densité de $(\mathbf{y}_i,\mathbf{x}_i)$ en $(\mathbf{y},\mathbf{x})$.
    - Donne :

$$\ell_i(\boldsymbol{\theta},\boldsymbol{\delta}) = \ell(\boldsymbol{\theta},\boldsymbol{\delta};\mathbf{y}_i,\mathbf{x}_i) = f_{(\mathbf{y},\mathbf{x})}(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta},\boldsymbol{\delta})$$

vraisemblance en $(\boldsymbol{\theta},\boldsymbol{\delta})$ ***de*** $(\mathbf{y}_i,\mathbf{x}_i)$.

- La loi de Bayes donne :

$$f_{(\mathbf{y},\mathbf{x})}(\mathbf{y},\mathbf{x};\boldsymbol{\theta},\boldsymbol{\delta}) = f_{\mathbf{y}|\mathbf{x}}(\mathbf{y};\mathbf{x};\boldsymbol{\theta},\boldsymbol{\delta})\times f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\delta})$$

et nous, comme (presque) tout le monde, nous avons considéré le cas où :

$$f_{(\mathbf{y},\mathbf{x})}(\mathbf{y},\mathbf{x};\boldsymbol{\theta},\boldsymbol{\delta}) = f_{\mathbf{y}|\mathbf{x}}(\mathbf{y};\mathbf{x};\boldsymbol{\theta})\times f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\delta}),$$

i.e. dans le cas où $\mathcal{D}(\mathbf{x}_i)$ ne contient pas d'information sur $\boldsymbol{\theta}_0$.

  - Engle, Hendry et Richard (1983) discutent cette hypothèse de séparabilité des densités de $\mathbf{y}_i\,|\,\mathbf{x}_i$ et $\mathbf{x}_i$ en leurs paramètres.

- Pratique : pas besoin de spécifier la loi $\mathcal{D}(\mathbf{x}_i)$ pour estimer efficacement $\boldsymbol{\theta}_0$, le MVC suffit car dans :

$$\ln \ell_i(\boldsymbol{\theta},\boldsymbol{\delta}) = \ln f_{(\mathbf{y},\mathbf{x})}(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta},\boldsymbol{\delta}) = \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}) + \ln f_{\mathbf{x}}(\mathbf{x}_i;\boldsymbol{\delta})$$

il n'y a que $\ln \ell_i^c(\boldsymbol{\theta}) = \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$ qui dépend de $\boldsymbol{\theta}$.

- $\mathcal{D}(\mathbf{x}_i)$ peut donc être quelconque.

- Imbens (1992) et Imbens et Lancaster (1996) discutent des cas où l'échantillon utilisé est stratifié sur les valeurs de $(\mathbf{y}_i, \mathbf{x}_i)$, avec $s_i = s(\mathbf{y}_i, \mathbf{x}_i)$:
  - Imbens (1992) : modèles paramétriques de choix discret
  - Imbens et Lancaster (1996) : modèles paramétriques en général
- La stratification de l'échantillon implique que $\mathcal{D}_{\mathcal{P}}\left((\mathbf{y}_i, \mathbf{x}_i)\right)$ et $\mathcal{D}_{\mathcal{E}}\left((\mathbf{y}_i, \mathbf{x}_i)\right)$ diffèrent.
  - Certaines strates étant sur-représentées, les autres étant sous-représentées.
- Lorsque la stratification porte sur les valeurs des endogènes, $\mathbf{y}_i$ ici, alors le conditionnement par les appartenances aux strates implique que les valeurs des exogènes, $\mathbf{x}_i$ ici, contiennent de l'information sur $\boldsymbol{\theta}_0$.

- *I.e.* $\mathcal{D}_{\mathcal{E}}(\mathbf{x}_i \mid s_i)$ dépend de $\boldsymbol{\theta}_0$, et l’estimation par le MV est plus efficace que celle par le MVC dans un échantillon stratifié sur $\mathbf{y}_i$.

Plan

7. Problème de spécification : Pseudo- ou Quasi-MV

7. Problème de spécification : Pseudo- ou Quasi-MV

- On a utilisé l'hypothèse de ***spécification correcte*** du modèle :

Le modèle paramétrique des $\mathbf{y}_i \mid \mathbf{x}_i$ est ***correctement spécifié*** par la famille paramétrique $\mathcal{F} \equiv \left\{ f(.;.;\boldsymbol{\theta}) : \mathcal{W} \to \mathbb{R}_{++}; \boldsymbol{\theta} \in \Theta \right\}$ s'il existe un ***unique vecteur*** $\boldsymbol{\theta}_0 \in \Theta$ tel que : $$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_0) = p_0(\mathbf{y};\mathbf{x}) \text{ pour tout } (\mathbf{y};\mathbf{x}) \in \mathcal{W}.$$

- Deux idées :
  - $p_0(.) \in \mathcal{F}$, *i.e.* spécification correcte de la famille paramétrique
  - $\boldsymbol{\theta}_0$ est identifiable (essentiellement technique) :

$$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}) = p_0(\mathbf{y};\mathbf{x}),\ \forall (\mathbf{y};\mathbf{x}) \in \mathcal{W} \quad \Leftrightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0.$$

- Hypothèse la plus discutable, et la plus discutée : ***choix correct de la famille paramétrique*** $p_0(.)$.

- Problème de spécification dans les modèles à variables latentes, *i.e.* tels que :

  $\mathbf{y}_i = \mathbf{f}(\tilde{\mathbf{y}}_i)$ où $\mathbf{f}(.)$ est connue mais $\tilde{\mathbf{y}}_i$ n'est observée qu'en partie.

  si problème sur le choix de $\mathcal{D}(\tilde{\mathbf{y}}_i \,|\, \mathbf{x}_i)$, *e.g.* :

  - Endogénéité de variables explicatives négligée
  - Héteroscédasticité du modèle de $\tilde{\mathbf{y}}_i \,|\, \mathbf{x}_i$ négligée
  - Mauvais choix de la famille paramétrique de $p_0(.)$, la fonction de densité de $\tilde{\mathbf{y}}_i \,|\, \mathbf{x}_i$
    - ***Remarque***. Les deux premiers problèmes sont des cas particuliers du dernier.

Hétéroscédasticité : problème de spécification « sérieux » en paramétrique

Remarque générale quant aux cadres d'inférence utilisés

(*i*) Les résultats sur les estimateurs présentés ici (comme dans 99%) des cas reposent sur l'hypothèse de bonne spécification du modèle,

(*ii*) à l'exception de ceux ne reposant que sur les propriétés des projections linéaires.

Idée de White (et de beaucoup d'autres, dont ceux qui travaillent en « non-paramétrique ») :

Hypothèse de « bonne spécification » difficile à défendre *stricto sensu* …
… même si hypothèse de « bonne approximation » peut être acceptable.

Raison pour laquelle il a développé des résultats et notions (*i*) sur le quasi-MV et pseudo-MV (avec Gouriéroux, Monfort et Trognon) et (*ii*) sur les M-estimateurs en cas de problèmes de spécification

Remarque générale quant aux problèmes « gérés » avec le Pseudo-MV

On listé trois problèmes de spécification ;

(i) Endogénéité de variables explicatives de $\mathbf{y}_i$,

(ii) Choix de famille de distribution erroné pour $p_0(.)$

(iii) Dont hétéroscédasticité des variables latentes

Ceux considérés par la suite sont des cas (ii) ***ou*** (iii) car : (a) Ces problèmes impliquent que $p_0(.) \notin \mathcal{F}$. ($b$) Ils peuvent, dans certains cas au moins, n'être pas trop « grave », *i.e.* $\hat{\boldsymbol{\theta}}_N^{MV}$ peut alors estimer un paramètre « proche » de $\boldsymbol{\theta}_0$. (b) Ce qui suit procure une solution, partielle, à ces problèmes.

On exclut le cas (i), car : (a) Ce problème doit être examiné *ex ante*. En effet, si a un problème d'endogénéité c'est que ce qu'on souhaite mesurer n'est pas $p_0(.)$. (b) Ce qui suit n'offre aucune solution à ce problème.

Questions

(i) L'estimateur du MV(C) de $\boldsymbol{\theta}_0$ estime quoi si $p_0(.) \notin \mathcal{F}$?

(ii) S'il estime quelque chose d'intéressant,
quelles sont ses propriétés dans ce cas ?

- En fait, l'estimateur défini par :

$$\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta}) \quad \text{où} \quad \ell_i(\boldsymbol{\theta}) \equiv f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}).$$

ne peut être intéressant que s'il converge quelque part (au moins en proba).

Hypothèse. ***Existence d'une pseudo-vraie valeur de*** $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})].$$

Interprétation. $f(;\boldsymbol{\theta}_*)$ est la ***meilleure approximation de*** $p_0(.)$ ***dans*** $\mathcal{F}$ ***au sens de l'information de Kullback-Leibler***.

- L'inégalité d'information de Kullback-Leibler implique que :

$$E[\ln f(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta}_*)] \leq E[\ln p_0(\mathbf{y}_i,\mathbf{x}_i)] = 0$$

et :

$$E[\ln f(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta}_*)] < E[\ln p_0(\mathbf{y}_i,\mathbf{x}_i)] = 0 \text{ si } p_0(.) \notin \mathcal{F}.$$

- En supposant continuité et dérivabilité de $\ln f(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta}) = \ln \ell_i(\boldsymbol{\theta})$ en $\boldsymbol{\theta}$ lorsque nécessaire on a :

$$E\left[\frac{\partial \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = E[\mathbf{s}_i(\boldsymbol{\theta})] = \mathbf{0} \quad \Leftrightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}_*,$$

par définition de $\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})]$ (en supposant que le support de $f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})$ ne dépend pas de $\boldsymbol{\theta}$).

- En revanche, on n'a plus, en général :

$$E[\mathbf{s}_i(\boldsymbol{\theta}_*) \mid \mathbf{x}_i] = \mathbf{0},$$

et c'est le début des problèmes …

Propriété. ***Espérances des dérivées de*** $\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$ en $\boldsymbol{\theta}$

On note :

$$E_{\boldsymbol{\theta}}[g(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})\,|\,\mathbf{x}_i]\equiv\int_{\Upsilon(\mathbf{x}_i)} g(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\,f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\mu(d\mathbf{y}),$$

où $f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})$ est une fonction de densité « candidate » de $\mathbf{y}_i\,|\,\mathbf{x}_i$ dont le support, $\Upsilon(\mathbf{x}_i)$, ne dépend pas de $\boldsymbol{\theta}$. On note :

$$\ell_i(\boldsymbol{\theta})\equiv f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}) \text{ et } \mathbf{s}_i(\boldsymbol{\theta})\equiv\frac{\partial\ln\ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$

On a :

$$(i)\quad \frac{\partial E_{\boldsymbol{\theta}}[\ln\ell_i(\boldsymbol{\theta})\,|\,\mathbf{x}_i]}{\partial\boldsymbol{\theta}}=E_{\boldsymbol{\theta}}\left[\mathbf{s}_i(\boldsymbol{\theta})\,|\,\mathbf{x}_i\right]=\mathbf{0}$$

$$(ii)\quad E_{\boldsymbol{\theta}}\left[\frac{\partial\mathbf{s}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\,|\,\mathbf{x}_i\right]=\frac{\partial}{\partial\boldsymbol{\theta}'}E_{\boldsymbol{\theta}}\left[\mathbf{s}_i(\boldsymbol{\theta})\,|\,\mathbf{x}_i\right]=-E_{\boldsymbol{\theta}}\left[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})'\,|\,\mathbf{x}_i\right]=-V_{\boldsymbol{\theta}}\left[\mathbf{s}_i(\boldsymbol{\theta})\,|\,\mathbf{x}_i\right]$$

- Propriété à l’origine de l’égalité des matrices d’information.

Propriété. ***Equation de score conditionnelle et égalité des matrices d'information conditionnelles, cas de la spécification correcte***

Si $f(.;\boldsymbol{\theta}) = p_0(.) \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0$ alors :

$$(i) \quad \boldsymbol{\theta}_0 \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln f(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}) \,|\, \mathbf{x}_i]$$

$$(ii) \quad E[\mathbf{s}_i(\boldsymbol{\theta}_0) \,|\, \mathbf{x}_i] = \mathbf{0}$$

$$(iii) \quad E[\mathbf{H}_i(\boldsymbol{\theta}_0) \,|\, \mathbf{x}_i] = -E[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)' \,|\, \mathbf{x}_i] \text{ où } \mathbf{H}_i(\boldsymbol{\theta}) \equiv \frac{\partial \mathbf{s}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial^2 \ln \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

Démonstration

Il suffit d'appliquer la propriété précédente en remarquant que :

$$(i) \quad f(.;\boldsymbol{\theta}_0) = p_0(.) \text{ ou } E[.\,|\,\mathbf{x}_i] = E_{\boldsymbol{\theta}_0}[.\,|\,\mathbf{x}_i]$$

$$(ii) \quad E[g(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta}) \,|\, \mathbf{x}_i] \equiv \int_{\Upsilon_0(\mathbf{x}_i)} g(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\, p_0(\mathbf{y};\mathbf{x}_i)\mu(d\mathbf{y}).$$

Principales conséquences de $p_0(.) \notin \mathcal{F}$, i.e. d'une mauvaise spécification du modèle de $\mathbf{y}_i \,|\, \mathbf{x}_i$

Si $\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln f(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})]$ et $f(.; \boldsymbol{\theta}_*) \neq p_0(.)$ alors :

(i) $E[\mathbf{s}_i(\boldsymbol{\theta}_*)] = \mathbf{0}$ par la définition de $\boldsymbol{\theta}_*$

(ii) L'équation $E[\mathbf{s}_i(\boldsymbol{\theta}_*) \,|\, \mathbf{x}_i] = \mathbf{0}$ n'est pas garantie

(iii) L'équation $E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)' \,|\, \mathbf{x}_i] = -E[\mathbf{H}_i(\boldsymbol{\theta}_*) \,|\, \mathbf{x}_i]$ n'est pas garantie

(iv) L'équation $E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)'] = -E[\mathbf{H}_i(\boldsymbol{\theta}_*)]$ n'est pas garantie

- Le résultat (*i*) vient de ce que :

$$E[\mathbf{s}_i(\boldsymbol{\theta}_*)\,|\,\mathbf{x}_i] \equiv \int_{\mathcal{Y}_0(\mathbf{x}_i)} \frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}}\, p_0(\mathbf{y};\mathbf{x}_i)\mu(d\mathbf{y})$$

et :

$$f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}_*) \neq p_0(\mathbf{y};\mathbf{x}) \text{ sur une partie mesurable de } \mathcal{Y}_0(\mathbf{x}_i).$$

- Les résultats (*iii*) et (*iv*) sont des conséquences, entre autres, du résultat (*ii*).

Intuition *par l'information de Kullback-Leibler*

Les termes $\boldsymbol{\theta}_0$ et $\boldsymbol{\theta}_*$ sont définis par $\arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})]$, et donc par :

$$\arg\max_{\boldsymbol{\theta}\in\Theta} E_{\mathbf{x}_i}\left[\int_{\mathcal{Y}_0(\mathbf{x}_i)} \left(\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) - \ln p_0(\mathbf{y};\mathbf{x}_i)\right) p_0(\mathbf{y};\mathbf{x}_i)\mu(d\mathbf{y})\right]$$

Les égalités $E[\mathbf{s}_i(\boldsymbol{\theta}_0)\,|\,\mathbf{x}_i] = \mathbf{0}$ et $E[\mathbf{H}_i(\boldsymbol{\theta}_0)\,|\,\mathbf{x}_i] = -E[\mathbf{s}_i(\boldsymbol{\theta}_0)\mathbf{s}_i(\boldsymbol{\theta}_0)'\,|\,\mathbf{x}_i]$ viennent de ce que $f(.;\boldsymbol{\theta}_0) = p_0(.)$, ce qu'on n'a pas avec $\boldsymbol{\theta}_*$.

Définition. ***Estimateur du Pseudo-MV ou du Quasi-MV***

Si $p_0(.) \notin \mathcal{F}$ (ou en tous cas s'il y a des chances que ce soit le cas), alors l'estimateur défini par :

$$\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta}) \quad \text{où} \quad \ell_i(\boldsymbol{\theta}) \equiv f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$$

est désigné sous le nom d'estimateur du Pseudo-MV ou du Quasi-MV.

- L'interprétation de l'estimation obtenue est importante. Elle part du résultat, rappelé dans la suite, qui dit que :

$$p\lim_{N\to+\infty} \hat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})],$$

i.e. sous certaines conditions de régularité.

On en déduit que :

Interprétation de l'estimation par le pseudo- ou quasi-MV

Si (*i*) la fonction de densité de $\mathbf{y}_i \mid \mathbf{x}_i$ est donnée par $p_0(\mathbf{y};\mathbf{x})$ et (*ii*) $\hat{\boldsymbol{\theta}}_N$ est défini par :

$$\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$$

alors $f(.;\hat{\boldsymbol{\theta}}_N)$ est un estimateur convergent de :

$$f(.;\boldsymbol{\theta}_*) \equiv \arg\max_{f(.;\boldsymbol{\theta})\in\mathcal{F}} E_{\mathbf{x}_i}\left[\int_{\mathcal{Y}_0(\mathbf{x}_i)} \big(\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) - \ln p_0(\mathbf{y};\mathbf{x}_i)\big)\, p_0(\mathbf{y};\mathbf{x}_i)\mu(d\mathbf{y})\right],$$

i.e. $f(.;\hat{\boldsymbol{\theta}}_N)$ est le meilleur, au sens de l'information de Kullback-Leibler, estimateur de $p_0(.)$ dans la famille $\mathcal{F}$.

Si $p_0(.)\in\mathcal{F}$, *i.e.* s'il existe $\boldsymbol{\theta}_0$ tel que $f(.;\boldsymbol{\theta}_0) = p_0(.)$, le modèle de $p_0(.)$ est correctement spécifié. Dans ce cas $f(.;\hat{\boldsymbol{\theta}}_N)$ est un estimateur convergent de $p_0(.)$ puisque $\hat{\boldsymbol{\theta}}_N$ est l'estimateur du MV(C) de $\boldsymbol{\theta}_0$.

Propriété. ***Convergence et normalité as. des estimateurs Pseudo-MV***

Si $\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})]$ et :

$$\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$$

alors sous certaines conditions de régularité, on a :

$$(i)\ \ \hat{\boldsymbol{\theta}}_N \xrightarrow[N\to+\infty]{p} \boldsymbol{\theta}_*$$

et :

$$(ii)\ \ \sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_*) \xrightarrow[N\to+\infty]{L} \mathcal{N}\left(\mathbf{0}, E[\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1} E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)'] E[\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1}\right).$$

- L'estimateur $\hat{\boldsymbol{\theta}}_N \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} N^{-1}\sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta})$ est défini pour estimer $\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})]$ selon la logique du principe d'analogie.
- Se montre en remarquant que $\hat{\boldsymbol{\theta}}_N$ peut être interprété comme un estimateur de la MM fondé sur la condition de moment juste-identifiante $E[\mathbf{s}_i(\boldsymbol{\theta}_*)] = \mathbf{0}$.

- Propriétés de $\hat{\boldsymbol{\theta}}_N$ pour l'estimation de $\boldsymbol{\theta}_*$ similaires à celles de $\hat{\boldsymbol{\theta}}_N^{MV}$ pour l'estimation de $\boldsymbol{\theta}_0$ mais …
    - … malheureusement, pas aussi simple que ça.
- Puisque $p_0(.) \notin \mathcal{F}$, certains résultats du MV(C) ne s'appliquent plus.
- L'égalité des matrices d'information ne s'applique plus :

$$E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)'] \neq -E[\mathbf{H}_i(\boldsymbol{\theta}_*)], \text{ en général.}$$

- Dommage car l'astuce consistant à utiliser la décomposition :

$$E[\mathbf{H}_i(\boldsymbol{\theta}_*)] = E_{\mathbf{x}_i}\left[E[\mathbf{H}_i(\boldsymbol{\theta}_*) \mid \mathbf{x}_i]\right]$$

  est également de peu d'utilité. Le modèle étant mal spécifié, le calcul du terme $E[\mathbf{H}_i(\boldsymbol{\theta}_*) \mid \mathbf{x}_i]$ n'offre en général aucune simplification.
    - Il est alors nécessaire de faire le pénible calcul de $\mathbf{H}_i(\boldsymbol{\theta})$ « en entier ».

Propriété. ***Estimation de*** $\boldsymbol{\Sigma}_* \equiv [\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1} E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)'] E[\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1}$

Sous les conditions de régularité assurant la normalité as. de $\hat{\boldsymbol{\theta}}_N$, on a :

$$\hat{\boldsymbol{\Sigma}}_N \xrightarrow[N \to +\infty]{p} \boldsymbol{\Sigma}_* = [\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1} E[\mathbf{s}_i(\boldsymbol{\theta}_*)\mathbf{s}_i(\boldsymbol{\theta}_*)'] E[\mathbf{H}_i(\boldsymbol{\theta}_*)]^{-1}$$

où :

$$\hat{\boldsymbol{\Sigma}}_N \equiv \left(N^{-1}\sum_{i=1}^N \mathbf{H}_i(\hat{\boldsymbol{\theta}}_N)\right)^{-1} \left(N^{-1}\sum_{i=1}^N \mathbf{s}_i(\hat{\boldsymbol{\theta}}_N)\mathbf{s}_i(\hat{\boldsymbol{\theta}}_N)'\right)\left(N^{-1}\sum_{i=1}^N \mathbf{H}_i(\hat{\boldsymbol{\theta}}_N)\right)^{-1}.$$

Remarque importante. $\boldsymbol{\theta}_0$ ***versus*** $\boldsymbol{\theta}_*$

Cas de spécification correcte de $p_0(.)$, **MV**

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})] \text{ car } f(.;\boldsymbol{\theta}_0) = p_0(.) \text{ (conséquence)}$$

Cas de problème de spécification sur $p_0(.)$, **Pseudo-MV**

$$\boldsymbol{\theta}_* \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} E[\ln \ell_i(\boldsymbol{\theta})], \text{ } i.e. \text{ définition de } \boldsymbol{\theta}_*$$

Conséquences :

(*i*) Le vecteur $\boldsymbol{\theta}_0$ contient généralement des paramètres « économiques ».
Descriptions économiques et statistiques du PGD coïncident.

(*ii*) $\boldsymbol{\theta}_*$ est un vecteur de paramètres statistiques.
Descriptions économiques et statistiques du PGD ne coïncident pas.

Problème. Impossible de définir ce que mesure réellement $\boldsymbol{\theta}_*$.

On retrouve ici des questions analogues à celle rencontrée lors de la comparaison entre (*i*) un ***modèle de régression linéaire*** d'une part, et (*ii*) une ***projection linéaire*** d'autre part.

(*i*) ***Modèle de régression*** :

$$y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i \quad \text{et} \quad E[u_i \mid \mathbf{x}_i] = E[u_i] \equiv 0 \quad \Rightarrow \quad E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\mathbf{a}_0$$

$\mathbf{a}_0$ est un paramètre défini par le modèle

(*i*) ***Projection linéaire*** :

$$E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\mathbf{a}_* \quad \Rightarrow \quad y_i = \mathbf{x}_i'\mathbf{a}_* + e_i \quad \text{et} \quad E[\mathbf{x}_i e_i] = 0$$

$\mathbf{a}_*$ est un paramètre statistique, $\mathbf{a}_* = E[\mathbf{x}_i\mathbf{x}_i']^{-1}E[\mathbf{x}_i y_i]$, défini par la technique de projection linéaire, *i.e*. la régression linéaire.

Dans les deux cas on utilise les MCO, mais pour des raisons différentes. Ici logiques d'identification (de $\mathbf{a}_0$) et d'ajustement (de y_i par $\mathbf{x}_i'\mathbf{a}_*$) coïncident.

Démonstration. ***Espérances des dérivées de*** $\ln f(\mathbf{y}_i;\mathbf{x}_i;\boldsymbol{\theta})$ en $E_{\boldsymbol{\theta}}[.]$

Résultat (i). Avec :

$$\frac{\partial}{\partial\boldsymbol{\theta}}\int_{\mathcal{Y}(\mathbf{x}_i)} f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\mu(d\mathbf{y}) = \int_{\mathcal{Y}(\mathbf{x}_i)} \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mu(d\mathbf{y})$$

et :

$$\int_{\mathcal{Y}(\mathbf{x}_i)} f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\mu(d\mathbf{y}) = 1$$

on a :

$$\int_{\mathcal{Y}(\mathbf{x}_i)} \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mu(d\mathbf{y}) = \mathbf{0}.$$

Avec :

$$\frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})^{-1}\frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$

sur le support de $f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})$, on a :

$$\int_{\mathcal{Y}(\mathbf{x}_i)} \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mu(d\mathbf{y}) = \mathbf{0} = \int_{\mathcal{Y}(\mathbf{x}_i)} \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})} \mu(d\mathbf{y})$$

et donc :

$$E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \,|\, \mathbf{x}_i] = \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}[\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \,|\, \mathbf{x}_i] = \mathbf{0}.$$

Ce résultat est en lien direct avec l'inégalité de Kullback-Leibler qui dit que :

$$\sup_{\boldsymbol{\delta}\in\Theta} E_{\boldsymbol{\theta}}[\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\delta}) \,|\, \mathbf{x}_i] = E_{\boldsymbol{\theta}}[\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \,|\, \mathbf{x}_i],$$

ou, de manière équivalente si $f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})$ est continûment différentiable en $\boldsymbol{\theta}\in\Theta$:

$$\frac{\partial}{\partial \boldsymbol{\delta}} E_{\boldsymbol{\theta}}[\ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\delta}) \,|\, \mathbf{x}_i] = \mathbf{0} \text{ si } \boldsymbol{\delta} = \boldsymbol{\theta}.$$

Avec l'interversion possible des opérations d'intégration et de différentiation, ceci garantit que :

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid \mathbf{x}_i\right] = E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i] = \mathbf{0}.$$

Maintenant, on sait que :

$$\frac{\partial E_{\boldsymbol{\theta}}[\ln \ell_i(\boldsymbol{\theta}) \mid \mathbf{x}_i]}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\Upsilon(\mathbf{x}_i)} \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \mu(d\mathbf{y})$$

$$= \int_{\Upsilon(\mathbf{x}_i)} \left[\frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) + \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \mu(d\mathbf{y})$$

$$= \int_{\Upsilon(\mathbf{x}_i)} \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mu(d\mathbf{y})$$

Le résultat (*ii*) implique donc que :

$$\int_{\Upsilon(\mathbf{x}_i)} \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mu(d\mathbf{y}) = \mathbf{0}.$$

Résultat (*ii*). Avec :

$$\frac{\partial}{\partial \boldsymbol{\theta}'} E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i] = \int_{\Upsilon(\mathbf{x}_i)} \frac{\partial\big(\mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})\big)}{\partial \boldsymbol{\theta}'} \mu(d\mathbf{y}) \quad \text{et} \quad E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i] = \mathbf{0}$$

on a :

$$\int_{\Upsilon(\mathbf{x}_i)} \frac{\partial \mathbf{s}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta}) \mu(d\mathbf{y}) = -\int_{\Upsilon(\mathbf{x}_i)} \mathbf{s}_i(\boldsymbol{\theta}) \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \mu(d\mathbf{y}).$$

Avec :

$$\frac{\partial f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})^{-1} \frac{\partial \ln f(\mathbf{y};\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$$

on a :

$$\int_{\Upsilon(\mathbf{x}_i)} \mathbf{s}_i(\boldsymbol{\theta}) \frac{\partial f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \mu(d\mathbf{y}) = \int_{\Upsilon(\mathbf{x}_i)} \mathbf{s}_i(\boldsymbol{\theta}) \mathbf{s}_i(\boldsymbol{\theta})' f(\mathbf{y};\mathbf{x};\boldsymbol{\theta}) \mu(d\mathbf{y}) = E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \mathbf{s}_i(\boldsymbol{\theta})'].$$

Avec :

$$\frac{\partial \mathbf{s}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

on a :

$$E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})'] = -E_{\boldsymbol{\theta}}\left[\frac{\partial \mathbf{s}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial \ln f(\mathbf{y};\mathbf{x}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$$

et avec :

$$E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta}) \mid \mathbf{x}_i] = \mathbf{0}$$

on a:

$$E_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})'] = V_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta})].$$