ÉCONOMÉTRIE
(UGA S2)

Regression linéaire: Introduction

Michal W. Urdanivia*

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

1er mars 2023

Contenu

1. Deux exemples d'effets causaux difficiles à identifier

2. Le problème de l'identification, covariances et paramètres

Partie A, Introduction

Dans cette partie, on ne considérera que des ***modèles de variables continues*** (sans points de masse).
On n'utilisera que le cadre d'inférence de la ***Méthode des Moments*** (Généralisée), cadre qui inclut celui de la régression.

Ceci dit, l'objectif essentiel de cette partie est plus général. Il s'agit :

1. D'introduire les notions de ***modèle économétrique*** et d'***effets causaux*,**
2. de montrer que la mesure statistique d'effets causaux amène la question de leur ***identification***, cette question devant être examinée avant de se lancer dans les estimations

et :

3. de montrer que ces problèmes d'identification impliquent qu'on n'estime pas toujours les paramètres d'un modèle de forme linéaire avec les MCO.

La dernière remarque est une des conséquences de la conjonction de deux phénomènes.

1. Les économètres veulent estimer des ***effets causaux***

mais

2. ils ne disposent ***pas de données expérimentales***, seulement de données de comportement « réels ».

L'avantage des données expérimentales pour la mesure statistique d'effets causaux est qu'***elles ont justement été construites dans le but de mesurer ces effets causaux*** (on verra dans la suite ce que cela implique « concrètement »).

Les économètres n'ont en général pas cette chance, et cela leur complique singulièrement la tâche, car ils doivent :

La dernière remarque est une des conséquences de la conjonction de deux phénomènes.

1. Les économètres veulent estimer des ***effets causaux***

mais

2. ils ne disposent ***pas de données expérimentales***, seulement de données de comportement « réels ».

L'avantage des données expérimentales pour la mesure statistique d'effets causaux est qu'***elles ont justement été construites dans le but de mesurer ces effets causaux*** (on verra dans la suite ce que cela implique « concrètement »).

Les économètres n'ont en général pas cette chance, et cela leur complique singulièrement la tâche, car ils doivent :

1. Analyser le contenu des données qu'ils utilisent **(*Analyse du processus générateur des données*),**

2. Examiner ce que ces données peuvent « révéler » des ***effets causaux*** qui les intéresse (***Analyse de l'identification des effets d'intérêt***),

3a. S'ils ont de la chance, ils peuvent utiliser les ***outils usuels*** de la statistique (comparaison de moyennes, estimateurs des Moindres Carrés, …)

ou :

3b. S'ils en ont moins, ils doivent se tourner vers des ***techniques spécifiques*** permettant de contourner les problèmes d'identification mis en évidence.

et, finalement :

4. Faire les calculs d'estimation, de test, … (ouf !)

Plan de la partie A

1. Modèles économétriques, effets causaux et identification
- Variables explicatives endogènes et/ou exogènes
- Remarque sur le modèle linéaire

2. Variables explicatives exogènes : modèles de régression
- Moindres carrés, Méthode des Moments et conditions d'orthogonalité
- La projection linéaire

3. Variables explicatives endogènes : modèles à variables instrumentales et contrôle de l'endogénéité
- Méthode des Moments, estimateurs des VI et estimateurs des 2MC
- Variables et fonctions de contrôle
- Test de la régression augmentée

4. Systèmes d'équations simultanées

- Forme structurelle versus forme réduite, estimation par les MCI
- Conditions de rang versus conditions d'ordre
- Intérêt de l'estimateur des 3MC

5. Introduction à la Méthode des Moments Généralisée (?)

- Synthèse des chapitres 2 et 3 et estimation « en système »

6. Estimation dans les systèmes d'équations linéaires (?)

- Estimation en système *versus* estimation équation par équation
- Systèmes de régressions empilées et estimateur « SUR »
- Systèmes d'équations simultanées et 3MC

- Modèles de forme linéaire
- Cas d'observations issues d'échantillons aléatoires de (vecteurs de) variables indépendantes et équi-distribuées
- Souvent, termes d'erreur homoscédastiques

Chapitre 1

Modèles économétriques, effets causaux et identification

1. Deux exemples d'effets causaux difficiles à identifier

Identification des paramètres d'un modèle

Variables explicatives exogènes, variables explicatives endogènes, modèle de régression

2. Le problème de l'identification, covariances et paramètres

Les concepts présentés à partir des exemples sont présentés dans le cas général

PLAN

1. Deux exemples d'effets causaux difficiles à identifier

2. Le problème de l'identification, covariances et paramètres

1. Deux exemples d'effets causaux difficiles à identifier

1. Deux exemples d'effets causaux difficiles à identifier

1.1. Formation et salaire

On analyse ici un problème « emblématique » (qui a valu le prix Nobel à Heckman en 2000), sous une forme un peu caricaturale.

Objectif. On veut estimer l'effet moyen, en termes de salaire, d'une formation BAC+5 par rapport à un BAC seul. On veut l'***effet causal*** moyen de la formation BAC+5 :

$\tilde{x}_i \rightarrow y_i$: effet d'acquis de connaissances sur le salaire

C'est une mesure de l'efficacité « économique » du BAC+5.

Les données. Un grand échantillon ($i = 1,..., N$) de jeunes salariés avec le « BAC » ($\tilde{x}_i = 0$) et avec « BAC+5 » ($\tilde{x}_i = 1$). On dispose de leurs salaires, y_i, 10 ans après leur BAC. Ces salariés ont tous le même âge, sexe, …

Examen théorique de la question posée

Une manière simple de poser le problème consiste à écrire le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0.$$

On a alors :

Salaire d'un « BAC » ($\tilde{x}_i = 0$):	$\alpha_0 + u_i$
Salaire moyen d'un « BAC » ($\tilde{x}_i = 0$):	α_0
Salaire d'un « BAC+5 » ($\tilde{x}_i = 1$):	$\alpha_0 + b_0 + u_i$
Salaire moyen d'un « BAC+5 » ($\tilde{x}_i = 1$):	$\alpha_0 + b_0$
Effet BAC+5 / BAC :	b_0

Reste maintenant à estimer b_0, mais avant ça il faut savoir si b_0 est :

« Estimable » avec les données disponibles $\Leftrightarrow$ identifiable

Examen de l'identification des paramètres du modèle $\mathbf{a}_0 \equiv (\alpha_0, b_0)$

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

On va en fait montrer ici que :

$\mathbf{a}_0 \equiv (\alpha_0, b_0)$ est identifiable avec les données disponibles, *i.e.* les $(y_i, \tilde{x}_i)$ avec $i = 1, ..., N$, ***si et seulement si*** $Cov[\tilde{x}_i; u_i] = 0$**.**

L'intuition sous-jacente est simple :

(*i*) Le seul estimateur de $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ qu'on sache estimer à partir des données disponibles est l'estimateur des MCO.

(*ii*) L'estimateur des MCO de $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ est biaisé si $Cov[\tilde{x}_i; u_i] \neq 0$.

(*iii*) On ne peut estimer $Cov[\tilde{x}_i; u_i]$ puisque les u_i ne sont pas observés.

On montrera ensuite que, dans le cas de l'effet de « BAC+5 » sur le salaire, on a vraisemblablement $Cov[\tilde{x}_i; u_i] > 0$.

L'estimateur des MCO de $\mathbf{a}_0$***,*** $\hat{\mathbf{a}}_N^{MCO}$***, ne converge pas vers*** $\mathbf{a}_0$ ***si*** $Cov\left[\tilde{x}_i;u_i\right]\neq 0$

Avec $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ et $\mathbf{x}_i \equiv (1, \tilde{x}_i)$, on écrit ici le modèle sous la forme générale « compacte » :

$$y_i = \mathbf{a}_0'\mathbf{x}_i + u_i \quad \text{avec} \quad E\left[u_i\right] \equiv 0.$$

L'estimateur des MCO de $\mathbf{a}_0$ est donné par (voir le cours de régression) :

$$\hat{\mathbf{a}}_N^{MCO} \equiv \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$$

avec :

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} \quad \text{et} \quad \mathbf{X} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}_{N\times K} = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{K,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,N} & x_{2,N} & \cdots & x_{K,N} \end{bmatrix}$$

Nous écrirons ici l'estimateur des MCO sous la forme :

$$\hat{\mathbf{a}}_N^{MCO} = \left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$$

en utilisant :

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \text{ et } \mathbf{X}'\mathbf{y} = \sum_{i=1}^{N} \mathbf{x}_i y_i$$

et en multipliant par N^{-1} les deux $\sum_{i=1}^{N}$... Cette écriture est « lourde » mais facilite l'analyse de la convergence de $\hat{\mathbf{a}}_N^{MCO}$ puisqu'on utilise la LGN.

On veut ici montrer que :

$$\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N \to +\infty]{p} \mathbf{a}_0 \text{ si } Cov\left[\tilde{x}_i ; u_i\right] = 0$$

et :

$$\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N \to +\infty]{p} \mathbf{a}_0 + \boldsymbol{\beta} \text{ avec } \boldsymbol{\beta} \neq \mathbf{0} \text{ si } Cov\left[\tilde{x}_i ; u_i\right] \neq 0.$$

Rmq. ***Pour tous les estimateurs de la*** **Partie A** : $\hat{\mathbf{a}}_N = \left[...\right]^{-1} \left[....\right] N^{-1} \sum_{i=1}^{N} ... y_i$

Le modèle de y_i nous donne que $y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i$, on a donc :

$$\hat{\mathbf{a}}_N^{MCO} = \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} N^{-1}\sum_{i=1}^{N} \mathbf{x}_i(\mathbf{x}_i'\mathbf{a}_0 + u_i).$$

Après développement, on obtient :

$$\hat{\mathbf{a}}_N^{MCO} = \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\mathbf{a}_0 + \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} N^{-1}\sum_{i=1}^{N} \mathbf{x}_i u_i$$

et après simplifications :

$$\hat{\mathbf{a}}_N^{MCO} = \mathbf{a}_0 + \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} N^{-1}\sum_{i=1}^{N} \mathbf{x}_i u_i$$

car :

$$\left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\mathbf{a}_0 = \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right]^{-1} \left[N^{-1}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i' \right] \mathbf{a}_0 = \mathbf{a}_0.$$

Rmq. ***Pour tous les estimateurs de*** **Partie A** : $\hat{\mathbf{a}}_N = \mathbf{a}_0 + [...]^{-1}[....]N^{-1}\sum_{i=1}^{N}...u_i$

$$\hat{\mathbf{a}}_N^{MCO} = \mathbf{a}_0 + \left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i$$

L'estimateur des MCO est convergent pour $\mathbf{a}_0$, *i.e.* $\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N \to +\infty]{p} \mathbf{a}_0$, si :

$$\left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i \xrightarrow[N \to +\infty]{p} \mathbf{0}.$$

On sait, par la loi LGN, que :

$$N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \xrightarrow[N \to +\infty]{p} E\left[\mathbf{x}_i \mathbf{x}_i' \right] \text{ et } N^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i \xrightarrow[N \to +\infty]{p} E\left[\mathbf{x}_i u_i \right]$$

En combinant ces résultats on obtient :

$$\left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i \xrightarrow[N \to +\infty]{p} E\left[\mathbf{x}_i \mathbf{x}_i' \right]^{-1} E\left[\mathbf{x}_i u_i \right] = E\left[\mathbf{x}_i \mathbf{x}_i' \right]^{-1} \begin{bmatrix} E\left[1 \times u_i \right] \\ E\left[\tilde{x}_i \times u_i \right] \end{bmatrix}$$

Finalement on obtient :

$$\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N\to+\infty]{p} \mathbf{a}_0 + E\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\begin{bmatrix} 0 \\ Cov\left[\tilde{x}_i;u_i\right] \end{bmatrix}.$$

et donc :

$$\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N\to+\infty]{p} \mathbf{a}_0 \text{ si } Cov\left[\tilde{x}_i;u_i\right]=0$$

et :

$$\hat{\mathbf{a}}_N^{MCO} \xrightarrow[N\to+\infty]{p} \mathbf{a}_0 + \boldsymbol{\beta} \text{ avec } \boldsymbol{\beta}\neq\mathbf{0} \text{ si } Cov\left[\tilde{x}_i;u_i\right]\neq 0.$$

Puisque $\hat{\mathbf{a}}_N^{MCO}$ est le seul estimateur de $\mathbf{a}_0$ qu'on sache calculer à partir des données disponibles on a :

$\mathbf{a}_0$ n'est identifiable à partir des $(y_i,\tilde{x}_i)$ ***que si*** $Cov\left[\tilde{x}_i;u_i\right]=0.$

On peut retrouver le problème lié à $Cov\left[\tilde{x}_i;u_i\right]\neq 0$ à partir de calculs simples.

Avec $E\left[u_i\right]\equiv 0$ et $y_i=\alpha_0+\tilde{x}_i b_0+u_i$, on a :

$$E\left[y_i\right]=\alpha_0+b_0E\left[\tilde{x}_i\right]$$

et donc :

$$\alpha_0=E\left[y_i\right]-b_0E\left[\tilde{x}_i\right].$$

Les termes $E\left[y_i\right]$ et $E\left[\tilde{x}_i\right]$ sont estimables, *i.e.* identifiables, puisque par la LGN on a :

$$\overline{y}_N\equiv N^{-1}\sum_{i=1}^{N}y_i\xrightarrow[N\to+\infty]{p}E\left[y_i\right]\quad\text{et}\quad\overline{\tilde{x}}_N\equiv N^{-1}\sum_{i=1}^{N}\tilde{x}_i\xrightarrow[N\to+\infty]{p}E\left[\tilde{x}_i\right].$$

Donc α_0 est identifiable si b_0 est identifiable.

Reste donc à savoir si on peut estimer b_0 à partir des données disponibles.

Dans le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0,$$

b_0 ***représente l'effet causal de*** $\tilde{x}_i$ ***vers*** y_i**.**

On sait ***estimer simplement la covariance entre*** $\tilde{x}_i$ ***et*** y_i :

$$N^{-1} \sum_{i=1}^{N} (\tilde{x}_i - \overline{\tilde{x}}_N)(y_i - \overline{y}_N) \xrightarrow[N \to +\infty]{p} Cov[\tilde{x}_i ; y_i].$$

Par application des propriétés des covariances on a :

$$Cov[\tilde{x}_i ; y_i] = Cov[\tilde{x}_i ; 1]\alpha_0 + Cov[\tilde{x}_i ; \tilde{x}_i] b_0 + Cov[\tilde{x}_i ; u_i],$$

c'est-à-dire :

$$Cov[\tilde{x}_i ; y_i] = V[\tilde{x}_i] b_0 + Cov[\tilde{x}_i ; u_i].$$

Une covariance mesure une corrélation, i.e. c'est un concept « symétrique », c'est une mesure (très) imparfaite d'une relation causale

L'équation $Cov[\tilde{x}_i; y_i] = V[\tilde{x}_i] b_0 + Cov[\tilde{x}_i; u_i]$ donne :

$$V[\tilde{x}_i]^{-1} Cov[\tilde{x}_i; y_i] = b_0 + V[\tilde{x}_i]^{-1} Cov[\tilde{x}_i; u_i].$$

Or on sait que :

$$\hat{b}_N^{MCO} = \frac{N^{-1}\sum_{i=1}^{N}(\tilde{x}_i - \bar{\tilde{x}}_N)(y_i - \bar{y}_N)}{N^{-1}\sum_{i=1}^{N}(\tilde{x}_i - \bar{\tilde{x}}_N)^2} \xrightarrow[N\to+\infty]{p} V[\tilde{x}_i]^{-1} Cov[\tilde{x}_i; y_i],$$

ce qui donne ici :

$$\hat{b}_N^{MCO} \xrightarrow[N\to+\infty]{p} b_0 + V[\tilde{x}_i]^{-1} Cov[\tilde{x}_i; u_i].$$

On pourrait éventuellement corriger le biais $V[\tilde{x}_i]^{-1} Cov[\tilde{x}_i; u_i]$ si on pouvait calculer un estimateur de $Cov[\tilde{x}_i; u_i]$ (on sait estimer $V[\tilde{x}_i]^{-1}$) mais :

On ne peut estimer $Cov[\tilde{x}_i; u_i]$ ***puisque*** u_i ***n'est pas observé***

Pour résumer. On vient de montrer que :

$\mathbf{a}_0$ n'est identifiable à partir des $(y_i, \tilde{x}_i)$ ***que si*** $Cov\left[\tilde{x}_i; u_i\right] = 0.$

Ceci nous amène à introduire des définitions importantes :

Définition. Si $Cov\left[\tilde{x}_i; u_i\right] = 0$ alors $\tilde{x}_i$ est ***exogène*** dans le modèle (linéaire) considéré.

Définition. Le modèle : $y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad$ avec $\quad E\left[u_i\right] \equiv 0$ est un ***modèle de régression*** (linéaire simple) si $Cov\left[\tilde{x}_i; u_i\right] = 0.$

Interprétation. C'est un modèle dont « on a le droit d'estimer les paramètres par les méthodes de régression » (car ces méthodes reposent sur des estimateurs, MC, convergents dans ce cas).

Définition. Si $Cov\left[\tilde{x}_i ; u_i\right] \neq 0$ alors $\tilde{x}_i$ est ***endogène*** dans le modèle considéré.

On a un ***problème d'identification*** si $\tilde{x}_i$ est ***endogène*** dans $y_i = \alpha_0 + b_0 \tilde{x}_i + u_i$.

L'endogénéité des variables explicatives un problème très fréquent en économétrie.

Retour à l'exemple « BAC+5 »

La question de l'identification de l'effet « BAC+5 » se résume ici de la manière suivante :

L'hypothèse $Cov\left[\tilde{x}_i ; u_i\right] = 0$ ***est-elle valide ?***

On a de bonnes raisons de penser que ce n'est pas le cas ici. **Scéance 1**

L'hypothèse $Cov[\tilde{x}_i ; u_i] = 0$ ***est-elle valide ?***

Le terme d'erreur u_i est une variable inobservée, l'examen de cette question doit d'abord être « théorique ». Il s'agit d'examiner le contenu de u_i et ses liens avec $\tilde{x}_i$, ce qui repose sur l'analyse du PGD de $(y_i, \tilde{x}_i)$

Dans le modèle $y_i = \alpha_0 + b_0 \tilde{x}_i + u_i$ ***le terme d'erreur*** u_i contient les effets de tout ce qui détermine y_i et qui n'est pas représenté par $\alpha_0 + b_0 \tilde{x}_i$.

Ici $y_i = salaire_i$ et $\tilde{x}_i = bac5_i$:

- u_i contient les effets de la chance (santé, opportunités, …), des aptitudes et des autres « caractéristiques » non mesurées … de *i* (y_i ***résulte du choix de*** *i* ***et de son employeur, s'il en a***)

et :

- $\tilde{x}_i$ dépend, en partie, des aptitudes (et des autres caractéristiques non mesurées) de *i* car $\tilde{x}_i$ ***résulte d'un choix*** $\pm$ ***contraint de*** *i*.

Si *j* est un individu avec « des aptitudes » au-delà de la moyenne alors :

- u_j est relativement élevé

et :

- l'individu *j* est certainement un BAC+5, *i.e.* $P\left[\tilde{x}_j = 1\right]$ est élévée.

Selon toute vraisemblance, on a donc $Cov\left[\tilde{x}_i ; u_i\right] > 0$ et dans :

$$Cov\left[\tilde{x}_i ; y_i\right] = V\left[\tilde{x}_i\right] b_0 + Cov\left[\tilde{x}_i ; u_i\right] > V\left[\tilde{x}_i\right] b_0,$$

la quantité $Cov\left[\tilde{x}_i ; u_i\right]$ ne peut être estimée, et donc b_0 ne peut être estimé à partir des données disponibles, b_0 ***n'est pas identifiable ici***.

Solutions. Soit on abandonne, soit on améliore notre modèle de y_i pour tenir compte de ce que $Cov\left[\tilde{x}_i ; u_i\right] > 0$, *i.e.* ***on apporte de l'information supplémentaire à notre modèle***, relative à $Cov\left[\tilde{x}_i ; u_i\right]$.

En résumé, dans un modèle de la forme :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

tel que $Cov[\tilde{x}_i ; u_i] \neq 0$ le paramètre $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ n'est pas identifiable.

Résoudre ce problème suppose un apport d'information permettant de gérer le fait que $Cov[\tilde{x}_i ; u_i] \neq 0$:

Problème d'identification = Déficit d'information.

C'est ce qu'on va apprendre à faire dans cette partie du cours.

Dans l'exemple « BAC+5 » le problème d'identification = ***endogénéité*** de $\tilde{x}_i$
C'est ***un problème très fréquent en économétrie***.

Dans l'exemple considéré, l'endogénéité de $\tilde{x}_i$ est essentiellement liée à une ***variable explicative « omise »***, l'aptitude (pour les études et la « carrière »).

Rmq. ***Données expérimentales versus données « réelles »***
Pour évaluer l'effet d'un traitement en médecine, on fait un groupe de patients « placebo » ($\tilde{x}_i = 0$) et un groupe de patients « traités » ($\tilde{x}_i = 1$), puis on mesure et on compare leur état de santé y_i.

La répartition aléatoire des patients dans les groupes « placebo » et « traités » garantit que dans le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

on a :

$$Cov[\tilde{x}_i ; u_i] = 0.$$

Alternative : On considère que les $\tilde{x}_i$ sont ***fixes*** et que ***seul*** u_i ***est aléatoire***.

Ce modèle est un modèle de régression dont on peut estimer les paramètres par les MCO, voire par simple comparaison de moyennes.

L'expérience a été construite pour ça : pour simplifier l'analyse des effets causaux.

Remarque. ***Modèle linéaire et effets hétérogènes***

Pour l'analyse statistique des effets de « BAC+5 », on a posé le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E\left[u_i\right] \equiv 0$$

Ce modèle suppose en fait que l'effet de « BAC+5 » est homogène pour tous individus, il est mesuré par le paramètre « fixe » b_0.

En pratique, on pose plutôt le modèle suivant :

$$y_i = \alpha_0 + b_i \tilde{x}_i + u_i \quad \text{avec} \quad E\left[u_i\right] \equiv 0,$$

i.e., un modèles à paramètres aléatoires, b_i et $\alpha_i \equiv \alpha_0 + u_i$.

L'analyse et l'estimation de tels modèles est l'objectif de la **Partie C** du cours.

1.2. Equilibre proie-prédateur

On analyse ici un problème non économique, celui de la mesure du nombre de proies nécessaire à chaque prédateur dans un écosystème en l'équilibre.

Objectif. On veut estimer le nombre de proies nécessaire à la vie d'un prédateur dans un écosystème en équilibre. L'effet causal est que les proies permettent la survie des prédateurs par un effet « nourriture ».

Les données. Un échantillon ($i = 1,..., N$) d'écosystèmes, et on a mesuré pour chacun d'entre eux le nombre de proies ($\tilde{x}_i$) et de prédateurs (y_i).

Une manière simple de poser le problème consiste à poser le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E\left[u_i\right] \equiv 0$$

et $\alpha_0 \simeq 0$ si les écosystèmes sont à l'équilibre « en moyenne ». Le nombre de proies nécessaires à chaque prédateur est b_0^{-1} à l'équilibre.

Avec l'exemple précédent on a vu que $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ est identifiable à partir des $(y_i, \tilde{x}_i)$ uniquement, si $Cov[\tilde{x}_i; u_i] = 0$ (c'est une condition nécessaire).

La condition $Cov[\tilde{x}_i; u_i] = 0$ ***est-elle une hypothèse valide ?***

Il y a de fortes chances que non. ***Les nombres de proies et de prédateurs se déterminent « simultanément »***, l'écosystème cherchant toujours à retourner à l'équilibre par ajustement simultané des nombres de proies et prédateurs :

- le nombre de proies détermine le nombre de prédateur selon un effet « nourriture disponible » : $\tilde{x}_i \rightarrow y_i$

mais on a également :

- le nombre de prédateurs détermine le nombre de proies, selon un effet « élimination par la chasse » : $y_i \rightarrow \tilde{x}_i$.

Cette analyse du PDG des $(y_i, \tilde{x}_i)$ montre que si y_i est fonction de $\tilde{x}_i$, $\tilde{x}_i$ est également fonction de y_i, *i.e* $\tilde{x}_i$ ***et*** y_i ***se déterminent « simultanément ».***

Dans le modèle :

$$y_i = \alpha_0 + b_0 \tilde{x}_i + u_i \quad \text{avec} \quad E[u_i] \equiv 0$$

$\tilde{x}_i$ étant fonction de y_i, elle est également fonction de u_i avec ici :

$$Cov[\tilde{x}_i ; u_i] < 0,$$

un excès de prédateurs (u_i) diminuant le nombre de proies $(\tilde{x}_i)$.

Conclusion. $\tilde{x}_i$ est ***endogène*** dans le modèle considéré, et $\mathbf{a}_0 \equiv (\alpha_0, b_0)$ n'est pas identifiable à partir des seules données considérées.

On retrouve ce ***problème d'endogénéité***, dit ***problème de simultanéité***, en économétrie, par exemple pour l'analyse de ***fonctionnement de marchés*** dans lesquels les prix et les quantités échangées se déterminent conjointement, dans le cadre de l'***équilibre de marché***.

PLAN

1. Deux exemples d'effets causaux difficiles à identifier

2. Le problème de l'identification, covariances et paramètres

2. Le problème de l'identification, covariances et paramètres

2. Le problème de l'identification, covariances et paramètres

On considère ici le modèle linéaire sous sa forme générale :

$$y_i = \mathbf{x}_i'\mathbf{a}_0 + u_i = \alpha_0 + \tilde{\mathbf{x}}_i'\mathbf{b}_0 + u_i \quad \text{avec} \quad E[u_i] \equiv 0.$$

Si $\mathbf{b}_0$ est identifiable, *i.e.* peut être estimé à partir des données, alors la constante α_0 est identifiable par :

$$\alpha_0 = E[y_i] - E[\tilde{\mathbf{x}}_i']\mathbf{b}_0.$$

Pour identifier $\mathbf{b}_0$ on ne sait estimer que des covariances.

Ici la covariance pertinente est :

$$Cov[\tilde{\mathbf{x}}_i; y_i] = Cov[\tilde{\mathbf{x}}_i; 1] + Cov\left[\tilde{\mathbf{x}}_i; \tilde{\mathbf{x}}_i'\right]\mathbf{b}_0 + Cov[\tilde{\mathbf{x}}_i; u_i] = V[\tilde{\mathbf{x}}_i]\mathbf{b}_0 + Cov[\tilde{\mathbf{x}}_i; u_i]$$

$$Cov\left[\mathbf{x}_i ; y_i\right] = V\left[\mathbf{x}_i\right]\mathbf{a}_0 + Cov\left[\mathbf{x}_i ; u_i\right]$$

Deux cas sont à considérer :

(*i*) Toutes les variables explicatives du modèle sont ***exogènes***, *i.e.* on a :

$$Cov\left[\tilde{\mathbf{x}}_i ; u_i\right] = \mathbf{0}.$$

Alors $\mathbf{b}_0$, et donc $\mathbf{a}_0 \equiv (\alpha_0, \mathbf{b}_0)$, est identifiable (sous certaines conditions). L'estimateur des MCO de $\mathbf{b}_0$ est convergent.

(*ii*) Certaines variables explicatives du modèle sont ***endogènes***, *i.e.* on a :

$$Cov\left[\tilde{\mathbf{x}}_i ; u_i\right] \neq \mathbf{0} \quad \Leftrightarrow \quad \text{Il existe } k \in \{2,...,K\} \text{ tel que } Cov\left[x_{k,i} ; u_i\right] \neq 0.$$

Alors $\mathbf{b}_0$, et donc $\mathbf{a}_0 \equiv (\alpha_0, \mathbf{b}_0)$, n'est pas identifiable. L'estimateur des MCO de $\mathbf{a}_0$ n'est pas convergent en général.

A partir du chapitre 3 nous apprendrons à gérer le cas (*ii*), ce qui suppose une bonne compréhension du cas (*i*) examiné en détail dans le chapitre suivant.

Remarques importantes

On a utilisé des résultats/techniques fréquemment employés par la suite.

Procédure 1. ***Estimation d'une espérance mathématique***

Si on veut estimer l'espérance mathématique commune des matrices $\mathbf{W}_i$, $\mathbf{M}_0 \equiv E[\mathbf{W}_i]$, à partir d'un (grand) échantillon d'observations de ces variables, il suffit d'utiliser la ***contre-partie empirique*** de $E[\mathbf{W}_i]$, *i.e.* la moyenne des $\mathbf{W}_i$, $N^{-1}\sum_{i=1}^{N}\mathbf{W}_i$.
En effet, la LGN (sous certaines conditions de régularité) donne que :

$$\bar{\mathbf{W}}_N \equiv N^{-1}\sum_{i=1}^{N}\mathbf{W}_i \xrightarrow[N\to+\infty]{p} E[\mathbf{W}_i] = \mathbf{M}_0$$

i.e. que $\bar{\mathbf{W}}_N \equiv N^{-1}\sum_{i=1}^{N}\mathbf{W}_i$ est un estimateur convergent de $\mathbf{M}_0 \equiv E[\mathbf{W}_i]$.

Procédure 2. ***Estimation de l'espérance mathématique d'une fonction paramétrée***

Si on veut estimer l'espérance mathématique commune des $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta}_0)$, $E\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta}_0)\right]$, à partir d'un (grand) échantillon d'observations des $\mathbf{w}_i$ et d'un estimateur convergent de $\boldsymbol{\beta}_0$, $\hat{\boldsymbol{\beta}}_N \xrightarrow[N\to+\infty]{p.} \boldsymbol{\beta}_0$, il suffit d'utiliser la contre-partie empirique de $E\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta}_0)\right]$, *i.e.* la moyenne des $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta}_0)$ en remplaçant $\boldsymbol{\beta}_0$ par son estimateur convergent, $\hat{\boldsymbol{\beta}}_N$.

En effet, une variante de LGN, donne que (sous certaines conditions de régularité):

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

i.e. que $N^{-1}\sum_{i=1}^{N}\mathbf{g}(\mathbf{w}_i, \hat{\boldsymbol{\beta}}_N)$ est un estimateur convergent de $E\left[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\beta}_0)\right]$.

Procédure 3. ***Estimation d'une fonction de paramètres estimables***

On veut estimer $\mathbf{H}_0 \equiv \mathbf{H}\left[\mathbf{B}_0, (\mathbf{\Gamma}_0)^{-1}\right]$ dont on sait que $\mathbf{H}[\mathbf{B}, \mathbf{G}]$ est une fonction continue en les éléments de $\mathbf{B}$ et $\mathbf{G}$ sur le domaine de définition des éléments de $\mathbf{B}_0$ et $(\mathbf{\Gamma}_0)^{-1}$.
Si on dispose d'un estimateur convergent de chacun des termes $\mathbf{B}_0$ et $\mathbf{\Gamma}_0$, $\hat{\mathbf{B}}_N \xrightarrow[N\to+\infty]{p} \mathbf{B}_0$ et $\hat{\mathbf{\Gamma}}_N \xrightarrow[N\to+\infty]{p} \mathbf{\Gamma}_0$, alors les propriétés de la convergence en probabilité donne que :

$$\hat{\mathbf{H}}_N \equiv \mathbf{H}\left[\hat{\mathbf{B}}_N, (\hat{\mathbf{\Gamma}}_N)^{-1}\right] \xrightarrow[N\to+\infty]{p} \mathbf{H}_0 \equiv \mathbf{B}_0\mathbf{H}\left[\mathbf{B}_0, (\mathbf{\Gamma}_0)^{-1}\right],$$

sachant que l'estimateur $\hat{\mathbf{H}}_N$ existe avec une probabilité approchant 1.

Ce dernier résultat indique que $\hat{\mathbf{\Gamma}}_N$ peut ne pas être inversible, et donc $\hat{\mathbf{H}}_N$ peut ne pas exister, mais la probabilité que cela arrive devient nulle si $N \to +\infty$.

Ces techniques proviennent des résultats suivants, qui sont utilisés :

(*i*) pour analyser la convergence d'estimateurs

et :

(*ii*) pour construire des estimateurs (avec les techniques données ci-avant).

Propriété 8. ***Loi (faible) des Grands Nombres d'une fonction paramétrée***

Soient (*i*) $\{\mathbf{w}_i; i = 1,2,...\}$ une suite de vecteurs aléatoires de $\mathbb{R}^W$ tels que les $\mathbf{w}_i$ sont iid pour $i = 1,2,...$ et (*ii*) $\boldsymbol{\beta}_0$ un vecteur de réels. On a (sous certaines conditions de régularité) :

$$N^{-1}\sum_{i=1}^{N}\mathbf{g}(\mathbf{w}_i,\hat{\boldsymbol{\beta}}_N)\xrightarrow[N\to+\infty]{p}E\left[\mathbf{g}(\mathbf{w}_i,\boldsymbol{\beta}_0)\right] \text{ si } \hat{\boldsymbol{\beta}}_N\xrightarrow[N\to+\infty]{p}\boldsymbol{\beta}_0.$$

.

Propriété 9. ***Inversion d'une suite de matrices convergeant en probabilité***

Soit $\{\mathbf{M}_N; N = 1,2,...\}$ une suite de matrices aléatoires telle que $\mathbf{M}_N \xrightarrow[N\to+\infty]{p} \mathbf{M}_0$ et $\mathbf{M}_0$ est inversible. On alors :

(i) la matrice $(\mathbf{M}_N)^{-1}$ existe avec une probabilité approchant 1 et et :

(ii) $(\mathbf{M}_N)^{-1} \xrightarrow[N\to+\infty]{p} (\mathbf{M}_0)^{-1}$.

Propriété 10. ***Transformation continue d'une suite convergeant en probabilité***

Soient $\{\mathbf{w}_N; N = 1,2,...\}$ une suite de vecteurs aléatoires et $\mathbf{H}(\mathbf{w})$ une fonction continue en $\mathbf{w}$ sur le domaine des $\mathbf{w}_N$. On alors :

$$\mathbf{w}_N \xrightarrow[N\to+\infty]{p} \mathbf{w}_0 \quad \Rightarrow \quad \mathbf{H}(\mathbf{w}_N) \xrightarrow[N\to+\infty]{p} \mathbf{H}(\mathbf{w}_0).$$