

# Économétrie des données de panel

## Modèles dynamiques

Thomas Chuffart

[thomas.chuffart@univ-fcomte.fr](mailto:thomas.chuffart@univ-fcomte.fr)

# Introduction

## Definition

On considère désormais un modèle de panel dynamique, i.e la variable expliquée retardée est incluse dans les variables explicatives :

$$y_{it} = \gamma y_{i,t-1} + \beta' x_{it} + \alpha_i^* + \varepsilon_{it} \quad (1)$$

pour  $i = 1, \dots, N$  et  $t = 1, \dots, T$ .  $\alpha_i^*$  modélise les effets individuels.  $E[\varepsilon_{it}] = 0$ ,  $E[\varepsilon_{it}\varepsilon_{js}] = \sigma_\varepsilon^2$  pour  $j = i$  et  $t = s$ .

Le choix entre effets fixes et aléatoires a des implications sur l'estimation du modèle encore différentes que précédemment.

# Introduction

## Remarques

- Valeur initiale, quel choix ? On verra que dans le modèle à effets aléatoires, l'interprétation dépend de l'hypothèse de la valeur de départ.
- La convergence de l'estimateur MCG est vérifiée uniquement dans certains cas.
- L'hypothèse de stricte exogénéité des variables explicative n'est plus vérifiée. L'estimateur LSDV n'est plus convergent quand  $T$  est fini.

# Outline

- Introduction
- **Biais dynamique**
- L'approche par variables instrumentales
- GMM

# Le biais dynamique

- L'estimateur LSDV est convergent dans le cadre statique pour les deux types d'effets, fixes ou random.
- L'estimateur LSDV ne converge plus quand on introduit de la dynamique dans le modèle de panel.

## Definition

Le biais de l'estimateur LSDV dans un modèle dynamique de panel est généralement connu sous le nom du biais de Nickell (1981)

# Le biais dynamique

## Definition

Le modèle de panel autoregressif d'ordre 1 est défini par :

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \alpha + \varepsilon_{it} \quad (2)$$

Avec  $|\gamma| < 1$ ,  $y_{i0}$  est observable et  $\varepsilon_{it}$  satisfait les conditions usuelles.

# Le biais dynamique

## Theorem

$$\text{plim } \hat{\gamma}_{LSDV} \neq \gamma \quad \text{et} \quad \text{plim}_{n,t \rightarrow \infty} \hat{\gamma}_{LSDV} = \gamma$$

$$\hat{\gamma}_{LSDV} = \left( \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2 \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1}) (y_{it} - \bar{y}_i) \right)$$

# Le biais dynamique

Le biais de  $\gamma_{LSDV}$  est donc :

$$\hat{\gamma}_{LSDV} = \gamma + \left( \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2 \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1}) (\varepsilon_{it} - \bar{\varepsilon}_i) \right)$$



# Le biais dynamique

Le deuxième terme peut s'écrire :

$$\frac{(1/nT) \sum_{i=1}^N \sum_{t=1}^T y_{i,t-1} \varepsilon_{it} - \sum_{i=1}^N \sum_{t=1}^T y_{i,t-1} \bar{\varepsilon}_i - \sum_{i=1}^N \sum_{t=1}^T \bar{y}_{i,-1} \varepsilon_{it} + \sum_{i=1}^N \sum_{t=1}^T \bar{y}_{i,-1} \bar{\varepsilon}_i}{(1/nT) \left( \sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2 \right)}$$

# Le biais dynamique

Le biais dépend donc du numérateur

- Si un des terme ne converge pas vers 0 en probabilité
- Il faut donc étudier chaque terme un à un.
- Dans ce cours, on ne prouvera pas la convergence de chaque terme.

# Le biais dynamique

- plim du 1er terme = 0 car  $\varepsilon_{it}$  non corrélé avec  $y_{i,-1}$
- Le 2ème terme :  $\frac{1}{nT} \sum_{i=1}^N \sum_{t=1}^T y_{i,t-1} \bar{\varepsilon}_i \Rightarrow \frac{1}{n} \sum_{i=1}^N \bar{y}_{i,-1} \bar{\varepsilon}_i$
- Le 4ème terme :  $\frac{1}{nT} \sum_{i=1}^N \sum_{t=1}^T \bar{y}_{i,-1} \bar{\varepsilon}_i \Rightarrow \frac{1}{n} \sum_{i=1}^N \bar{y}_{i,-1} \bar{\varepsilon}_i$
- Donc ces deux termes s'annulent

# Le biais dynamique

## Remarque

*L'expression du biais asymptotique de l'estimateur LSDV s'écrit :*

$$- \operatorname{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^N \bar{y}_{i,-1} \bar{\varepsilon}_i \quad (3)$$

# Le biais dynamique

## Theorem

*Si les  $\varepsilon_{it}$  sont IID, alors le biais est égale à :*

$$\text{plim}_{n \rightarrow \infty} (\gamma - \gamma_{LSDV}) = - \frac{(1 - \gamma) (T - T\gamma - 1 + \gamma^T)}{(1 - \gamma) \left( T - T^2 - \frac{2\gamma}{(1-\gamma)^2} (T - T\gamma - 1 - \gamma^T) \right)} \quad (4)$$

# Le biais dynamique

Mais, qu'est-ce que cela signifie concrètement ?

- Si  $T \rightarrow \infty$ , le biais tend vers une constante non nulle.  
L'estimateur est alors convergent
- Sinon, l'estimateur LSDV est biaisé et ne converge pas !!!
- Le biais est causé par l'élimination des effets individuels  $\alpha$  pour chaque observation ce qui crée une corrélation d'ordre  $\frac{1}{T}$  entre les variables explicatives et les résidus.

# Biais dynamique

```
# Fonction permettant de créer l'index temporel
get_year <- function(t,n){
  return(rep(1:t,n))
}

# Fonction permettant de créer l'index individuel
get_id <- function(t,n){
  id <- rep(0,(t*n))
  for (i in 1:n){
    id[(1+(t*(i-1))):(t*i)] <- rep(i,t)
  }
  return(id)
}

# Fonction simulant les données
coeff_lsdv_arsim <- function(t,n,g){
  alpha <- runif(n,-1,1) # Simulation des paramètres non-observés
  y <- array(rep(0, (t+1)*n), dim=c(t+1, n)) # Initialisation de la variable dépendante
  e <- array(rnorm((t+1)*n), dim=c(t+1, n)) # Simulation des erreurs
  for (t in 2:(t+1)){ # On simule la variable expliquée
    y[t,] <- alpha + g*y[t-1,] + e[t,]
  }
  y0 <- y[2:t,] # y0 est la variable dépendante
  y1 <- y[1:(t-1),] # y1 est le lag de la variable dépendante
  y0 <- c(y0)
  y1 <- c(y1)
  df <- data.frame(id,year,y0,y1) # Construction du dataframe
  # Estimateur LSDV
  lsdv <- plm(y0 ~ y1, index = c("id","year"), data = df, model = "within")
  gam_hat <- lsdv$coefficients
  return(gam_hat)
}
```

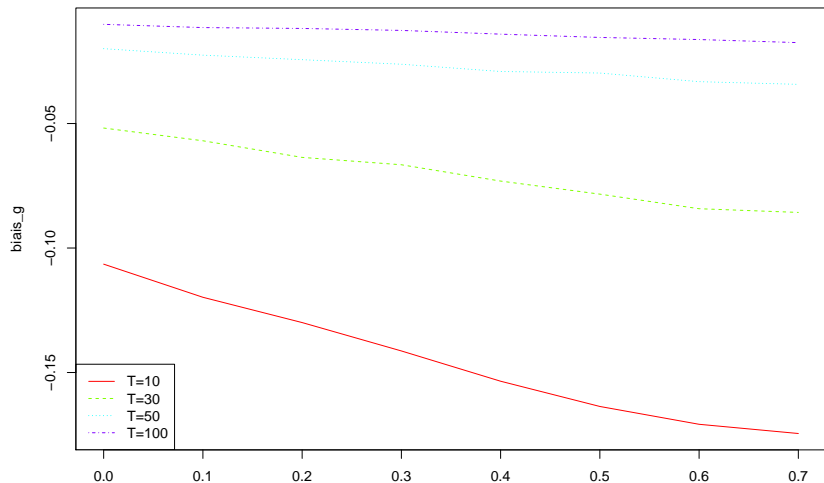
# Biais dynamique

```
g = (0:7)/10
t = c(10,20,50,100)
R = 500 ## Nombre de réplifications
biais_g <- matrix(0, nrow = length(g), ncol = length(t))
biais_gam_hat <- rep(NA,R)
for (l in 1:length(t)){
  for (k in 1:length(g)){
    G <- g[k] ## Paramètre autorégressif
    T <- t[l] ## Nombre de périodes
    N <- 100 ## Nombre d'individus
    year <- get_year(T,N) ## Construction de l'index temporel
    id <- get_id(T,N)
    for (r in 1:R){ ## Boucle sur les réplifications
      biais_gam_hat[r] <- coeff_lsdv_arsim(T,N,G) - G
    }
    biais_g[k,l] <- mean(biais_gam_hat)
  }
}
```



# Biais Dynamique

```
col_set <- rainbow(4)
matplot(g,biais_g, type = 'l', col = col_set)
legend("bottomleft", c("T=10", "T=30","T=50", "T=100"), col
```



# Le biais dynamique

Expérience de Monte-Carlo	T	n	$\gamma$	$\hat{\gamma}_{LSDV}$	$\bar{biais}$
	10	10	0.4	0.238	-0.162
	10	100	0.4	0.246	-0.154
	100	10	0.4	0.386	-0.014
	100	100	0.4	0.386	-0.014
	10	10	0.1	-0.0251	-0.125
	10	100	0.1	-0.017	-0.120
	100	10	0.1	0.0895	-0.011
	100	100	0.1	0.089	-0.011

# Le biais dynamique

Que faire ?

- ML et MCG mais nécessite une hypothèse supplémentaire (on aime pas ça) sur la valeur initiale
- LSDV corrigé du biais
- Variables instrumentales (Anderson and Hsiao, 1982)
- GMM (Arenallo and Bond, 1985)

# Outline

- Introduction
- Biais dynamique
- **L'approche par variables instrumentales**
- GMM

## L'approche par IV

Vous ne l'appliquerez pas dans votre projet mais vous devez la connaître :

- Permet de faire un rappel sur les variables instrumentales
- Retour sur l'estimation 2SLS (two stage least square)

# L'approche par IV

## Rappels

Soit le modèle suivant :

$$y = X\beta + \varepsilon$$

- $y$  est de taille  $N$
- $X$  est une matrice de variables explicatives  $N \times K$
- $\beta$  est un vecteur  $K \times 1$
- $\varepsilon$  est un vecteur  $N \times 1$  avec  $E[\varepsilon] = \mathbf{0}$  et  $V[\varepsilon|X] = \sigma_\varepsilon^2 I_N$

# L'approche par IV

## Rappels

Que se passe-t'il l'hypothèse d'exogénéité n'est pas respectée ?

$$E[\varepsilon|X] \neq \mathbf{0}, \quad \text{plim}_{n \rightarrow \infty} \frac{1}{N} X' \varepsilon = \gamma \neq \mathbf{0}$$

- $E[\hat{\beta}] \neq \beta$
- $\text{plim } \hat{\beta} = \beta + Q^{-1}\gamma$

# L'approche par IV

## Rappels

### Definition

Soit  $z_h \in \mathcal{R}^N$ , un ensemble de  $H$  variables. Ces variables sont des instruments si elles sont **exogènes** aux erreurs  $E[\varepsilon|Z] = 0$  et corrélées aux variables explicatives  $E[x_{ik}z_{ih}] \neq 0$ .



# L'approche par IV

## Rappels

### Exemple

On veut estimer l'impact de l'éducation et de l'expérience sur le salaire.

- On suspecte l'éducation d'être endogène. ie. l'éducation peut-être expliquée par diverses choses qui pourraient aussi avoir un impact sur le salaire.
- On introduit l'éducation du père : généralement corrélée avec l'éducation du fils mais non corrélée avec la variable expliquée.

# L'approche par IV

## Rappels

Hypothèses :

- $\text{plim } \frac{1}{N} Z' Z = Q_{ZZ}$
- $\text{plim } \frac{1}{N} Z' X = Q_{ZX}$
- $\text{plim } \frac{1}{N} Z' \varepsilon = \mathbf{0}$  ou  $E[z_i (y_i - x_i' \beta)] = 0$

# L'approche par IV

## Rappels

La troisième hypothèse signifie donc que l'on a  $H$  équations et  $K$  paramètres inconnus :

- Si  $H > K$  le modèle est sur-identifié, on utilise la méthode 2SLS
- Si  $H < K$  le modèle est sous-identifié, on ne peut rien faire
- $H = K$  le modèle est identifié, on utilise la méthode IV classique  $\beta_{IV} = (Z'X)^{-1}Z'y$

# L'approche par IV

## Rappels

Si  $H > K$ , la matrice  $Z'X$  ne peut pas s'inverser. Il faut utiliser le 2SLS.

### Definition

L'estimateur  $\hat{\beta}_{2SLS}$  est donné par :

$$\hat{\beta}_{2SLS} = (\hat{X}'X)^{-1}\hat{X}'y \quad (5)$$

avec  $\hat{X} = Z(Z'Z)^{-1}Z'X$

# L'approche par IV

## Rappels

- Stage 1 : On régresse chaque variable explicative sur chaque instrument :

$$x_{ki} = \alpha_1 z_{1i} + \cdots + \alpha_h z_{hi} + u_i$$

On construit ensuite  $\hat{x}_{ki}$

- Stage 2 : On régresse  $y_j$  sur  $\hat{x}_{ki}$  :

$$y_i = \beta_1 \hat{x}_{1i} + \cdots + \beta_k \hat{x}_{ki} + \varepsilon_i$$

# L'approche par IV

## IV dans le panel

Soit le modèle dynamique suivant :

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \beta' x_{it} + \rho' \omega_i + \varepsilon_{it}$$

Hypothèses :

- $E[\varepsilon_{it}] = 0, E[\alpha_i] = 0, E[\varepsilon_{it}\varepsilon_{js}] = \sigma_\varepsilon^2, j = i, t = s$
- $E[\alpha_i\alpha_j] = \sigma_\alpha^2, E[\alpha_i x_{it}] = [\alpha_i \omega_i] = 0$
- $E[\varepsilon_{it} x_{it}] = [\varepsilon_{it} \omega_i] = 0$

# L'approche par IV

IV dans le panel

$$y_i = y_{i,-1}\gamma + \alpha_i e + x_i\beta + \omega_i'\rho e + \varepsilon_i$$

Anderson and Hsiao, 1982 :

- Étape 1 : Transformation différence première
- Étape 2 : Sélection des instruments et estimation de  $\gamma$  et  $\beta$
- Étape 3 : Estimation de  $\rho$
- Étape 4 : Estimation de  $\sigma_\varepsilon^2$  et  $\sigma_\alpha^2$

# L'approche par IV

## IV dans le panel

Étape 1 : Transformation différence première

$$\Delta y_{it} = y_{it} - y_{i,t-1} = \gamma \Delta y_{i,t-1} + \beta' \Delta x_{it} + \Delta \varepsilon_{it}$$

- Comme la transformation Within, cette transformation permet de retirer les effets individuels.



# L'approche par IV

## IV dans le panel

Étape 2 : Sélection des instruments et estimation de  $\gamma$  et  $\beta$

- $E[z_{it}(\varepsilon_{it} - \varepsilon_{i,t-1})] = 0$
- $E[z_{it}(y_{i,t-1} - y_{i,t-2})] \neq 0$

Deux choix possible :  $z_{it} = y_{i,t-2}$  et  $z_{it} = y_{i,t-2} - y_{i,t-3}$ .

# L'approche par IV

## IV dans le panel

Étape 3 : Estimation de  $\rho$

$$\blacksquare \bar{y}_i - \hat{\gamma}_{IV} \bar{y}_{i,-1} - \hat{\beta}_{IV} \bar{x}_i = \rho'_i \bar{\omega}_i + u_i$$

Étape 4 : Estimation de  $\sigma_\varepsilon^2$  et  $\sigma_\alpha^2$

# Outline

- Introduction
- Biais dynamique
- L'approche par variables instrumentales
- **GMM**

# GMM

## Le concept

Lars Peter Hansen (1982) : Generalised Method of Moments

- Theory-driven : croyance dans la modélisation paramétrique
- mais des hypothèses peuvent ne pas être vérifiées
- Adrian Pagan (2003) : la modélisation doit être un mix entre la cohérence théorique et empirique

# GMM

## Le concept

Les estimateurs GMM utilisent des hypothèses sur les moments des variables aléatoires pour dériver une fonction objective :

- Les moments présumés des variables aléatoires fournissent des conditions sur les moments de la population
- Les données sont utilisées pour calculer les moments empiriques
- L'estimations des paramètres se fait en respectant au mieux les conditions sur les moments
- Minimisation d'une fonction objective.

# GMM

## Le concept

La méthode des moments, Pearson (1895) :

- On cherche à estimer la moyenne d'une distribution par la moyenne empirique,
- La variance par sa variance empirique, ect...
- $\mu = \mathbb{E}[y]$ 
  - La condition sur ce moment est :  $\mathbb{E}[y] - \mu = 0$
  - Soit  $\frac{1}{N} \sum y_i - \mu = 0$

# GMM

## Le concept

Soit  $y_i = x_i\beta + \varepsilon_i$  :

- $\mathbb{E}[\varepsilon|x] = 0 \Rightarrow \mathbb{E}[x\varepsilon] = 0$
- Population condition  $\mathbb{E}[x(y - x\beta)] = 0$
- Échantillon condition :  $\frac{1}{N} \sum_{i=1}^N (x_i(y_i - x_i\beta)) = 0$

# GMM

## Le concept

- La MM fonctionne uniquement quand le nombre de conditions est égale au nombre de paramètre à estimer.
- Si il y en a plus, le système est sur-identifié et ne peut se résoudre
- Les GMM minimisent une fonction sur les conditions des moments. Si la condition d'exogénéité n'est pas respectée, on peut écrire les conditions comme :

$$\mathbb{E}[z(y - x\beta)] = 0 \quad (6)$$



# GMM

## Le panel dynamique

Soit le modèle de panel dynamique :

$$y_{it} = \gamma y_{i,t-1} + \beta' x_{it} + \rho' \omega_i + \alpha_i + v_{it} \quad (7)$$

- $\alpha_i$  sont les effets individuels non-observés
- $x_{it}$  est un vecteur de  $k_1$  variables explicatives
- $\omega_i$  est un vecteur de  $k_2$  variables explicatives invariantes

# GMM

## Le panel dynamique

### Hypothèses :

- $v_{it} = \varepsilon_{it} + \alpha_i$ ,  $\mathbb{E}(\alpha_i) = 0$  et  $\mathbb{E}(\varepsilon_{it}) = 0$
- $\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \sigma_\varepsilon^2$  et  $\mathbb{E}(\alpha_i\alpha_j) = \sigma_\alpha^2$
- $\mathbb{E}(\alpha_i x_{it}) = 0$  et  $\mathbb{E}(\alpha_i \omega_i) = 0$

### Definition

L'estimation GMM est basé sur un modèle en différence première pour éliminer les  $\alpha_i$  et les  $\omega_i$  :

$$(y_{it} - y_{i,t-1}) = \beta' (x_{it} - x_{i,t-1}) + \varepsilon_{it} - \varepsilon_{i,t-1} + \gamma (y_{i,t-1} - y_{i,t-2}) \quad (8)$$

# GMM

## Le panel dynamique

Intuition sur les moment conditions :

- $y_{i,t-2}$  et  $y_{i,t-2} - y_{i,t-3}$  ne sont pas les seuls instruments valides.
- Toutes les variables retardées  $y_{i,t-2-j}$  valident :

$$\mathbb{E}(y_{i,t-2-j}(\varepsilon_{i,t} - \varepsilon_{i,t-1})) = 0$$

$$\mathbb{E}(y_{i,t-2-j}(y_{i,t-1} - y_{i,t-2})) \neq 0$$

# GMM

## Le panel dynamique

### Remarque

*Intuition* : les  $m + 1$  conditions

$$\mathbb{E}(y_{i,t-2-j}(\varepsilon_{it} - \varepsilon_{i,t-1})) = 0 \quad (9)$$

*peuvent être utilisées pour estimer le vecteur de paramètres*  
 $\theta = \{\beta, \gamma, \rho, \sigma_{\alpha}^2, \sigma_{\varepsilon}^2\}$

M. ARELLANO et S. BOND (1991). « Some Tests of Specification for Panel Data : Monte Carlo Evidence and an Application to Employment Equations ». In : Review of Economic Studies 58.3, p. 277-297. DOI : 10.3982/ECTA11319

# GMM

## Le panel dynamique

### Definition

A chaque période, on a ces conditions orthogonales :

$$\mathbb{E}(q_{it}\Delta\varepsilon_{it}) = 0, \quad q_{it} = \{y_{i0}, y_{i1}, \dots, y_{it-2}, x'_i\} \quad (10)$$

Période	Nombre de conditions
$t = 2$	$1 + Tk_1$
$t = 3$	$2 + Tk_1$
$\vdots$	$\vdots$
$t = T$	$(T-1)Tk_1$
Total	$T(T-1)(\frac{k_1+1}{2})$

# GMM

## Le panel dynamique

### Definition

L'estimateur GMM minimise le critère suivant :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \quad q(y, \theta) = \underset{\theta}{\operatorname{argmin}} \quad \hat{m}(y, \theta)' S^{-1} \hat{m}(y, \theta) \quad (11)$$

avec  $S^{-1}$  est une matrice de poids.