

Économétrie des données de panel

M2 Économie de la Santé et Économie Appliquée

Université Paris Cité

Carla Morvan

2026-2027

Part V

Les variables instrumentales

Les variables instrumentales

1. Introduction

- 1.1 Retour sur l'endogénéité
- 1.2 Intuition de la variable instrumentale

2. Le modèle IV

- 2.1 IV en panel
- 2.2 Diagnostics et limites

3. Application

- 3.1 pollution de l'air et COVID-19

Plan

1. Introduction

1.1 Retour sur l'endogénéité

1.2 Intuition de la variable instrumentale

2. Le modèle IV

3. Application

Objectifs du chapitre

- Saisir l'intuition d'une variable instrumentale
- Connaître les deux conditions de validité d'un bon instrument
- Estimer un modèle IV en panel (within-2SLS)
- Diagnostiquer un instrument faible et interpréter un résultat IV

Ce que le DiD ne résout pas

Le DiD est puissant, mais il a des conditions strictes :

- Il faut un **traitement daté** (une réforme, un choc, à un moment précis)
- Il faut un **groupe contrôle** crédible (tendances parallèles)

Mais si l'endogénéité vient d'une variable qui **varie en continu dans le temps** (pas un choc ponctuel daté), ou s'il n'existe **aucun bon groupe contrôle**, alors le DiD ne s'applique plus.

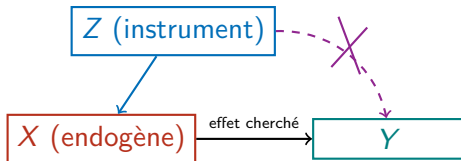
Exemple : quel est l'effet des dépenses de santé par habitant sur l'espérance de vie ? Pas de "réforme" à une date précise, pas de groupe "non traité" par les dépenses : le DiD est hors jeu ici.

Le problème, encore une fois

On veut l'effet causal de X (dépenses de santé) sur Y (espérance de vie), mais X est endogène : $Cov(X, \varepsilon) \neq 0$.

On ne peut pas faire varier X "à la main", toutes choses égales par ailleurs. On n'observe que des variations de X mélangées avec d'autres facteurs cachés dans ε .

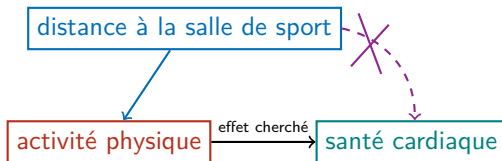
L'idée de l'IV : trouver une variable Z qui fait bouger X , mais qui n'a **aucun lien direct** avec Y , seulement via X .



Une analogie

Imaginez vouloir mesurer l'effet de l'exercice physique sur la santé cardiaque, mais les gens qui font du sport sont aussi ceux qui mangent mieux, dorment mieux, fument moins (facteurs de confusion).

Une variable instrumentale possible : la **distance à la salle de sport la plus proche**. Elle influence la probabilité de faire du sport (pertinence), mais n'a **aucune raison** d'affecter directement la santé cardiaque autrement qu'en changeant la pratique sportive (exclusion).



On utilise cette variation induite par Z , plutôt que toute la variation de X , polluée par la confusion.

Plan

1. Introduction

2. Le modèle IV

2.1 IV en panel

2.2 Diagnostics et limites

3. Application

Les deux conditions d'un bon instrument

Un instrument Z doit vérifier **deux conditions**, correspondant aux deux flèches du schéma précédent :

1. Pertinence (*relevance*) : Z doit être corrélé avec X : $Cov(Z, X) \neq 0$. Vérifiable directement dans les données.

2. Exclusion : Z ne doit affecter Y **que** via X : $Cov(Z, \varepsilon) = 0$. **Jamais vérifiable**, seulement défendable par argumentation économique.

L'estimation : les 2SLS

Two-Stage Least Squares = deux régressions successives.

Étape 1 : régresser X sur Z

$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

On récupère $\hat{X}_i$, la partie de X expliquée par Z , exogène.

Étape 2 : régresser Y sur $\hat{X}$

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \varepsilon_i$$

$\hat{\beta}_1$ est l'estimateur IV recherché.

IV et panel : pourquoi combiner les deux ?

On a vu que les effets fixes neutralisent l'hétérogénéité **constante dans le temps**. Mais certaines sources d'endogénéité **varient** dans le temps.

Exemple : les dépenses de santé et l'espérance de vie peuvent être liées par simultanéité **chaque année**, indépendamment de l'hétérogénéité entre pays. Un effet fixe pays ne règle pas ce problème.

Solution : combiner les deux outils, effets fixes pour l'hétérogénéité invariante, IV pour l'endogénéité qui varie dans le temps.

Le within-2SLS

On applique d'abord la transformation *within*, puis l'IV sur les données transformées :

Étape 1 : régresser $\tilde{x}$ sur $\tilde{z}$

$$\tilde{x}_{it} = \pi_1 \tilde{z}_{it} + u_{it}$$

On récupère $\hat{\tilde{x}}_{it}$, la variation *within* de x expliquée par la variation *within* de z .

Étape 2 : régresser $\tilde{y}$ sur $\hat{\tilde{x}}$

$$\tilde{y}_{it} = \beta_1 \hat{\tilde{x}}_{it} + \varepsilon_{it}$$

Concrètement : α_i disparaît comme d'habitude par la transformation *within* ; l'IV s'applique ensuite sur la variation **intra-individu**.

Conséquence importante : l'instrument Z doit être pertinent pour expliquer la variation **dans le temps** de X , pas seulement les différences entre individus.

Instruments internes vs externes au panel

Instrument externe

Une variable en dehors du panel, choisie pour sa pertinence économique (ex. une variable géographique, climatique, historique).

Instrument interne

Construit à partir du panel lui-même — typiquement, les **valeurs retardées** de X ($x_{i,t-1}$, $x_{i,t-2}$...), comme dans les modèles dynamiques (Arellano-Bond).

Risque des instruments internes : si ε_{it} est autocorrélé, les valeurs retardées de X peuvent aussi être corrélées à ε_{it} → violation de l'exclusion.

Le problème de l'instrument faible

Si Z est **faiblement** corrélé à X (pertinence limite), l'estimateur IV devient très instable, même si Z est théoriquement valide.

Un instrument faible peut donner des estimations **très biaisées** en petit échantillon, avec des erreurs standards énormes.

→ Comment le détecter : regarder **le F-stat de première étape**

On teste la significativité conjointe de Z dans la régression $x_{it} = \pi_1 z_{it} + u_{it}$.

Règle empirique:

- $F > 10$: instrument considéré comme suffisamment fort
- $F < 10$: instrument **faible**, les résultats IV sont peu fiables

Ce seuil de 10 est une convention, pas une loi absolue. Un F proche de 10 mérite déjà de la prudence, pas seulement un F en dessous.

Quand on a plusieurs instruments

Si on dispose de **plusieurs** instruments pour une seule variable endogène, on peut tester leur cohérence mutuelle.

Test de Sargan / Hansen

H_0 : tous les instruments sont valides (exclusion respectée pour chacun)

H_1 : Certains instruments sont corrélés avec l'erreur, donc non valides

Limite importante : Ce test permet de détecter des instruments non valides, mais pas de prouver qu'un instrument est valide

L'endogénéité était-elle vraiment un problème ?

Test de Wu-Hausman : On peut tester si $\hat{\beta}_{OLS}$ et $\hat{\beta}_{IV}$ diffèrent significativement.

Hypothèses

$H_0 : X$ est exogène $\Rightarrow$ OLS consistant et plus efficient

$H_1 : X$ est endogène $\Rightarrow$ seul IV reste consistant

Si H_0 n'est pas rejetée, l'OLS classique (plus efficient) reste préférable (l'IV n'était peut-être pas nécessaire). Encore une fois : le raisonnement économique doit primer sur le test seul.

Plan

1. Introduction

2. Le modèle IV

3. Application

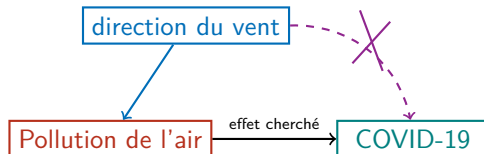
3.1 pollution de l'air et COVID-19

Contexte : pollution de l'air et COVID-19

Morvan, Lepeule & Lavaine : quel est l'effet causal de la pollution de l'air ($PM_{2.5}$) sur la sévérité du COVID-19 ?

Le problème d'endogénéité : la pollution est corrélée à des facteurs socio-économiques et démographiques qui influencent aussi la sévérité du COVID-19 (densité urbaine, âge de la population, accès aux soins...). Une simple régression OLS serait biaisée.

Solution : instrumenter la pollution locale par la **direction du vent**



Les données

Panel **commune** × **semaine**, France entière, 2020-2021.

- **Y** : hospitalisations et décès COVID-19 (pour 100 000 habitants), par âge, sexe, comorbidité
- **PM_{2.5}** : agrégée en moyenne hebdomadaire communale
- **Direction du vent** : Météo France, 875 stations, regroupée en 4 catégories cardinales
- **Contrôles** : météo, couverture vaccinale

Le design IV : première étape

La direction du vent, **interagit avec le département**, i.e pour chaque géographies, prédit la pollution locale :

Première étape

$$PM_{m,w} = \sum_{r=1}^R \sum_{d=1}^D \varphi_w^{r,d} \mathbb{1}(\text{Vent}_{m,w} = d) \times \mathbb{1}(\text{Dept}_m = r) + \\ \alpha_m + \delta_w + \theta_{s,d} + \gamma X_{m,w} + \varepsilon_{m,w}$$

Seconde étape

$$COVID_{m,w} = \beta \widehat{PM}_{m,w} + \alpha_m + \delta_w + \theta_{s,c} + \gamma X_{m,w} + \varepsilon_{m,w}$$

Quels sont tous ces effets fixes ?

$\alpha_m \rightarrow$ effets fixes individuel (municipal ici)

$\delta_w \rightarrow$ effets fixes temporel (week ici)

$\theta_{s,d} \rightarrow$ effets fixes double temporel $\times$ individuel (saison $\times$ département)

Estimer IV panel avec R

```
library(fixest)
```

```
model_iv <- feols(  
  covid_rate ~ meteo_controls + vaccination |  
    commune + semaine + dept_saison |  
  pm25 ~ vent * departement,  
  data = panel_covid,  
  cluster = ~commune  
)
```

```
summary(model_iv, stage = 1) # première étape (force de l'instrument)  
summary(model_iv, stage = 2) # seconde étape (effet causal)
```

Syntaxe fixest pour l'IV : $Y \sim \text{contrôles} \mid \text{effets_fixes} \mid$
 $X_{\text{endogène}} \sim \text{instruments}$
On écrit la première étape après le deuxième " $\mid$ "

Discussion de l'instrument

Pertinence : Z doit être corrélé avec X .

→ F-stat de première étape **élevé dans toutes les spécifications** la direction du vent prédit fortement la concentration locale de $PM_{2.5}$.

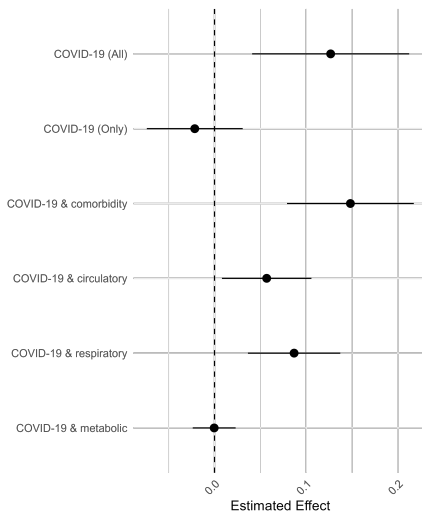
Exclusion : Z ne doit affecter Y que via X .

Pas de test direct possible pour l'exclusion — l'hypothèse est que la direction du vent est **exogène** aux facteurs socio-économiques (ne dépend ni de la richesse, ni des comportements, ni des politiques locales).

Menace potentielle à l'exclusion : et si le vent affectait aussi **directement** le comportement des gens (ex. rester chez soi par mauvais temps, donc moins de contacts sociaux et moins de transmission du COVID) ? Le papier contrôle pour la météo (température, pluie, vitesse du vent) pour limiter ce risque.

Résultats principaux

Effet d'une hausse de $PM_{2.5}$ sur les décès COVID-19



L'effet est concentré chez les 80 ans et plus

Chez les **80 ans et plus avec comorbidité** : $+1 \mu\text{g}/\text{m}^3$ de $\text{PM}_{2.5} \rightarrow$ **+2,39 décès** et **+4,17 hospitalisations** pour 100 000 habitants.

Aucun effet significatif détecté chez les moins de 60 ans, quel que soit le statut de comorbidité.

La pollution agit comme un **facteur aggravant** chez les individus déjà vulnérables (âge, comorbidités)
Pas d'effet détectable dans la population générale en bonne santé.

Ce qu'on retient de cette application

- L'instrument (vent $\times$ département) est **fort** (F-stat élevé) et économiquement défendable (le vent n'affecte la santé que via la pollution qu'il transporte)
- Un résultat **hétérogène** : effet nul en population générale, mais net chez les plus vulnérables.

Conclusion du chapitre

- Les variables instrumentales permettent de surmonter les problèmes d'endogénéité en identifiant des sources exogènes de variation
 - Un bon instrument doit vérifier deux conditions : **pertinence** (Z corrélé à X) et **exclusion** (Z n'influence Y que par X)
- En pratique, il est souvent difficile de trouver de bonnes variables instrumentales
- En panel, l'IV se combine avec les effets fixes (**within-2SLS**), l'instrument doit expliquer la variation **dans le temps** de X , pas seulement les différences entre individus

Lorsque les instruments sont bien choisis, l'IV reste une méthode puissante et utile pour répondre à des questions d'évaluation d'impact cruciales !