Modèle linéaire à variables instrumentales

Exposé économétrie variables qualitatives

GOUBA Leyla, KABORE Adeline, YAMEOGO Saïdou

Enseignant : Dr. Israël SAWADOGO

Institut Supérieur des Sciences de la Population (ISSP)

12 Juin 2025

Plan du travail

Introduction

L'économétrie vise à estimer les effets causaux entre variables. Toutefois, la présence d'endogénéité c'est-à-dire une corrélation entre une variable explicative et l'erreur du modèle compromet cette estimation.
Pour y remédier, les économètres utilisent les modèles à variables instrumentales (VI), qui permettent d'identifier des effets causaux fiables à condition de disposer d'instruments valides.
Comme l'a souligné Joshua Angrist, « un bon instrument est comme une expérience aléatoire que la nature nous offre », illustrant la puissance de cette méthode dans les contextes où les expérimentations contrôlées sont impossibles.

Contexte et problématique

Modèle Linéaire Simple : Rappels

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \quad (1)$$

- **Linéarité** : La relation entre X et Y est linéaire.
- **Matrice plein rang** : $Rang(X) = k => X'X$ est inversible.
- **Homoscédasticité** : $\text{Var}(\varepsilon|X) = \sigma^2$.
- **Normalité des erreurs** : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
- **Absence d'autocorrélation des résidus** : $Cov(\varepsilon_i, \varepsilon_j|X) = 0$, $i \neq j$.
- **Exogénéité stricte : absence de corrélation entre X et** ε : $\mathbb{E}(\varepsilon_i|X) = 0 <=> \text{Cov}(X, \varepsilon) = 0$

La violation de la dernière condition conduit à des estimateurs MCO **Biaisés** : Il s'agit d'un problème **d'endogénéité**

Contexte et problématique

Définition clé

Endogénéité : Corrélation entre variables explicatives et terme d'erreur

$$Cov(X_k, \varepsilon) \neq 0$$

Conséquences :

- Biais des estimateurs MCO
- Inconsistance des paramètres
- Interprétation causale compromise

Solution : Modèles à Variables Instrumentales (VI)

"Un bon instrument est comme une expérience aléatoire que la nature nous offre" (Joshua Angrist)

D'ou vient cette endogénéité ?

Sources d'endogénéité

1. Erreurs de mesure

- Biais d'atténuation
- Ex : Temps d'étude mal mesuré

$$X = X^* + v$$

2. Variables omises

- Variable Z corrélée à X et Y
- Ex : Motivation dans éducation-salaire

3. Simultanéité

- Causalité bilatérale
- Ex : Prix et quantité

$$\begin{cases} Q_d = \alpha - \beta P + u \\ P = \gamma - \delta Q_d + v \end{cases}$$

Applications des modèles VI

Table 1 – Applications pratiques des variables instrumentales.

Domaine	Problème	Instruments typiques
Économie de l'éducation	Effet de l'éducation sur les salaires	Distance aux écoles, réformes éducatives
Santé publique	Impact des traitements médicaux	Assignation aléatoire, réformes d'assurance
Politiques publiques	Évaluation de programmes sociaux	Critères d'éligibilité, seuils d'attribution
Économie du développement	Accès au microcrédit	Distance aux agences, programmes pilotes
Économie du travail	Heures travaillées et productivité	Lois sur le temps de travail

Conditions de validité des instruments

Triade des conditions

1. **Inclusion** : $Cov(Z, X) \neq 0$
   - L'instrument affecte la variable endogène
   - Test : Statistique F > 10
2. **Exclusion** : $Cov(Z, \varepsilon) = 0$
   - Pas de lien direct avec l'erreur
   - Condition théorique non testable directement
3. **Monotonicité**
   - Aucun comportement "anticonformiste"
   - Condition technique pour l'interprétation causale

Identification du modèle

Identification du modèle

- Sous-identifié : $p < k$
- Juste identifié : $p = k$
- Suridentifié : $p > k$

Identification des instruments

- Repose sur connaissance théorique
- Nécessite justification économique
- Tests d'endogénéité)

Estimation des paramètres

Soit le modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \quad (2)$$

Nous soupçonnons la variable X_k d'être endogène (correlée avec le terme d'erreur ε)

On posons $Z = (1, X_1, X_2, \ldots, X_{k-1}, Z_1, \ldots, Z_m)$ avec $p = k + m$ où p est le nombre de colonnes de Z et $Z_1, \ldots, Z_m$ les variables instrumentales.

L'estimation se fait en deux étapes connue sous le nom de **Double MCO (2SLS)**

Estimation des paramètres

Étape 1 : Régression auxiliaire

$$X_k = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{k-1} X_{k-1} + \delta_1 Z_1 + \cdots + \delta_m Z_m + v$$

$\rightarrow$ Obtention de $\widehat{X}_k$

Étape 2 : Régression principale

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k \widehat{X}_k + \varepsilon$$

Estimateur VI :

$$\widehat{\beta}_{2SLS} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y$$

Avec $\widehat{X} = (1, X_1, ..., X_{k-1}, \widehat{X}_k)$

La régression auxiliaire n'a pas d'interprétation économique

Test de spécification du modèle

Test d'endogénéité

- Durbin-Wu-Hausman
- Compare MCO et VI

$$H = (\hat{\beta}_{2SLS} - \hat{\beta}_{MCO})'[\text{Var}(\hat{\beta}_{2SLS}) - \text{Var}(\hat{\beta}_{MCO})]^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{MCO})$$

- H_0 : Exogénéité

Test de sur-identification

- Sargan-Hansen
- $S = n \times R^2 \sim \chi^2(m)$
- H_0 : Instruments valides

Test de spécification du modèle

Test de sous-identification

- Kleibergen-Paap LM
- Anderson canonique
- H_0 : Instruments pertinents

Test de pertinence

- $StatistiqueF > 10$
- R^2 première étape
- Règle de Stock-Yogo

Interprétation des résultats des tests

Table 2 – Guide d'interprétation des tests VI.

Test	Valeur critique	Interprétation
Endogénéité	$p - value < 0.05$	Rejet de H_0 : présence d'endogénéité
Sargan	$p - value > 0.05$	Instruments valides
Pertinence (F-stat)	$F > 10$	Instruments forts
Sous-identification	$p - value < 0.05$	Modèle identifié

Interprétation des paramètres

Effet causal local (LATE)

$\widehat{\beta}_{IV} \rightarrow$ Effet moyen pour les individus sensibles à l'instrument

Comparaison MCO vs VI :

- MCO : Effet corrélationnel (biaisé en cas d'endogénéité)
- VI : Effet causal local (sous conditions de validité)

Exemple éducation-salaire :

- $\widehat{\beta} = 0.09$: +9% de salaire par année d'éducation
- Interprétation : Effet pour ceux dont la scolarité est influencée par l'instrument (distance à l'école)

Problématique du R^2 dans les modèles VI

R^2 première étape

$$R_1^2 = \frac{\sum(\hat{X}_k - \bar{X}_k)^2}{\sum(X_k - \bar{X}_k)^2}$$

- Mesure la pertinence des instruments
- Valeur élevée souhaitable

R^2 deuxième étape

$$R_2^2 = 1 - \frac{\sum(y_i - \hat{y}_i^{IV})^2}{\sum(y_i - \bar{y})^2}$$

- Peut être négatif
- Ne mesure pas la qualité d'ajustement
- Utilité limitée

- Validité conditionnelle des instruments
- Sensibilité aux choix d'instruments
- Interprétation locale (LATE vs ATE)

Pratique

Notre étude porte sur la modélisation de l'effet de l'éducation sur le salaire annuel. La base de données, extraite du package WOOLDRIDGE, est composée de 7 variables :

- **lwage** : Salaire annuel
- **educ** : Années d'éducation
- **nearc4** : Vivre à proximité de l'université (1 dollar) ou loin de l'université (0 dollar)
- **exper** : Années d'expérience expersq : Années d'expérience (terme de marché)
- **black** : Noir (no 1), pas noir (-0)
- **south** : Vivant dans le sud (no 1) ou non (-0)

Nous avons utilisé le logiciel R.

Conclusion et perspectives

Apports principaux :

- Méthode robuste pour l'inférence causale
- Solution élégante au problème d'endogénéité
- Large applicabilité empirique

Limites et défis :

- Difficulté à trouver des instruments valides
- Problème des instruments faibles
- Interprétation locale (LATE)

Remerciements

Merci pour votre attention !