



THE WORLD BANK



Session IV

Variables instrumentales

Christel M. J. Vermeersch

janvier 2008

Un exemple pour commencer...

- Disons que nous voulons évaluer un programme volontaire de formation professionnelle
 - Toute personne sans emploi est éligible
 - Certaines personnes choisissent de s'inscrire (« Traitements »)
 - D'autres choisissent de ne pas le faire (« Comparaisons »)

- Quelques moyens simples (mais pas très bons) d'évaluer le programme :
 - Comparez la situation avant et après dans le groupe traitement
 - Comparez la situation des traitements et des comparaisons après l'intervention
 - Comparez la situation des traitements et des comparaisons avant et après

Programme volontaire de formation professionnelle

Assumons que nous décidons de comparer les résultats des participants aux non-participants:
au moyen du modèle simple suivant:

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

Ou $T = 1$ si la personne participe à la formation

$T = 0$ si la personne ne participe pas à la formation

x = variables de contrôle (exogènes & observables)

Pourquoi est-ce que ce modèle ne fonctionne pas bien?

Programme volontaire de formation professionnelle



Assumons que nous décidons de comparer les résultats des participants aux non-participants:
au moyen du modèle simple suivant:

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

Ou $T = 1$ si la personne participe à la formation

$T = 0$ si la personne ne participe pas à la formation

x = variables de contrôle (exogènes & observables)

Pourquoi est-ce que ce modèle ne fonctionne pas bien? 2 raisons:

- ▷ Variables omises mais importantes
- ▷ La décision des personnes à participer est endogène/dépend d'elles-mêmes.

Problème #1 : Variables omises

Même dans un modèle bien étudié, il manquera:

- ▷ des caractéristiques "oubliées": nous ne savions pas qu'elles étaient importantes
- ▷ des caractéristiques trop compliquées à mesurer

Exemples :

- ▷ Le talent et la motivation des personnes
- ▷ Le niveau d'information des personnes
- ▷ Les coûts d'opportunité à participer des personnes
- ▷ Le niveau d'accès aux services de formation et autres

Le modèle complet "correct" est: $y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$

Le modèle que nous utilisons est: $y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$

Problème #2 : La décision à participer est endogène

Comme la formation est volontaire, la participation est une variable de décision. Donc: une variable endogène.

(C'est à dire: la participation T elle-même dépend des personnes.)

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

$$T = \pi + \pi_2 D + \xi$$

Problème #2 : La décision à participer est endogène

La participation T est une variable endogène:

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

$$T = \pi + \pi_2 D + \xi$$

$$\Rightarrow y = \alpha + \beta_1 (\pi + \pi_2 D + \xi) + \beta_2 x + \varepsilon$$

$$\Rightarrow y = \alpha + \beta_1 \pi + \beta_2 x + \beta_1 \pi_2 D + \beta_1 \xi + \varepsilon$$

Ici aussi, nous voyons que le modèle complet "correct" peut contenir une variable D que nous ne connaissons pas.

- Le modèle « correct » est :
- Modèle simplifié :

- Disons que nous estimons l'effet de traitement γ_1 avec $\beta_{1,OLS}$
- Si D est corrélé avec T, et que nous n'incluons pas D dans le modèle simplifié, alors l'estimateur du paramètre sur T prendra une partie de l'effet de D. Ceci se passera dans la mesure où D et T sont corrélés.
- Ainsi : notre estimateur OLS $\beta_{1,OLS}$ de l'effet de traitement γ_1 saisit l'effet d'autres caractéristiques (D) en plus de l'effet traitement.
- Ceci signifie qu'il y a une différence entre $E(\beta_{1,OLS})$ et γ_1
 - la valeur attendue de l'estimateur OLS β_1 n'est pas γ_1 , le véritable effet traitement
 - $\beta_{1,OLS}$ est un estimateur biaisé de l'effet traitement γ_1 .

CV4

CV4

$$\text{corr}(T, \varepsilon) = \text{corr}(T, \gamma^3 D + \eta) = \gamma^3 \text{corr}(T, D)$$

wb226893, 05/23/2006

- Le modèle « correct » est :
- Modèle simplifié :

- Ceci signifie qu'il y a une différence entre $E(\beta_{1,OLS})$ et γ_1
 - la valeur espérée de l'estimateur OLS β_1 n'est pas γ_1 , l'effet traitement véritable
 - $\beta_{1,OLS}$ est un estimateur biaisé de l'effet traitement γ_1 .

- Pourquoi cela s'est-il passé ?
 - Une des conditions de base pour que l'OLS soit BLUE a été violée :

- En d'autres mots $E(\beta_{1,OLS}) \neq \gamma_1$ (estimateur biaisé)
- Pire encore..... $plim(\beta_{1,OLS}) \neq \gamma_1$ (estimateur inconsistant)

CV6

$$\text{corr}(T, \varepsilon) = \text{corr}(T, \gamma^3 D + \eta) = \gamma^3 \text{corr}(T, D)$$

wb226893, 05/23/2006

Que pouvons-nous faire pour résoudre ce problème ?

$$y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$$

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

- ❑ Essayer de nettoyer la corrélation entre T et ε :
- ❑ Isoler la variation dans T qui n'est pas corrélée avec ε à travers la variable D omise
- ❑ Nous pouvons faire ceci en utilisant une variable instrumentale (VI)

Idée de base derrière l'estimation par

$$y = \alpha + \gamma_1 T + \gamma_2 x + \gamma_3 D + \eta$$

□ Le problème de base est que $\text{corr}(T, D) \neq 0$

□ Trouver une variable Z qui satisfasse deux conditions :

1. Z est corrélée avec T : $\text{corr}(Z, T) \neq 0$

---- Z et T sont corrélées, ou Z prédit une part de T

2. Z n'est pas corrélée avec ε : $\text{corr}(Z, \varepsilon) = 0$

---- En soi, Z n'a pas d'influence sur y . La seule façon qu'elle peut influencer y est parce qu'elle influence T . Tout l'effet de Z sur y passe à travers T .

□ Exemples de Z dans le cas du programme volontaire de formation professionnelle ?

Doubles moindres carrés (DMC - 2SLS)

Rappelez-vous le modèle original avec T endogène:

$$y = \alpha + \beta_1 T + \beta_2 x + \varepsilon$$

Première étape: Faire une régression de T sur la variable instrumentale Z et les autres variables exogènes

$$T = \delta_0 + \delta_1 x + \theta_1 Z + \tau$$

- ▷ Calculer la valeur prédite de T pour chaque observation: \hat{T}
- ▷ Vu que Z et x ne sont pas corrélés avec ε , \hat{T} ne sera pas non plus corrélé avec ε .
- ▷ Attention: vous aurez besoin d'au moins une variable instrumentale par régresseur endogène.

Doubles moindres carrés (DMC - 2SLS)

Deuxième étape: Faire une régression de y sur la valeur prédite \hat{T} et les autres variables exogènes:

$$y = \alpha + \beta_1 \hat{T} + \beta_2 x + \varepsilon$$

- ▷ *Attention:* les erreurs-type de la seconde étape doivent être corrigées, parce que \hat{T} n'est pas un régresseur fixe
- ▷ En pratique: nous utilisons la commande STATA `ivreg`, laquelle calcule les deux étapes et les erreurs-type correctes.
- ▷ Intuition: en utilisant \hat{T} au lieu de T , nous éliminons la corrélation de T avec ε .
- ▷ Il peut être démontré que, sous certaines conditions, l'estimation par variables instrumentales donne un estimateur consistant de γ_1 (théorie des grands échantillons).

Usages des variables instrumentales

- Simultanéité : X et Y s'occasionnent l'une l'autre
 - instrument X

- Variables omises : X prend l'effet des autres variables qui sont omises
 - instrument X avec une variable qui n'est pas corrélée avec la/les variable(s) omises(s)

- Erreur de mesure : X n'est pas mesurés avec précision
 - instrument X

Exemples

- Hoxby (2000) et Angrist (1999)
 - L'effet de la taille de la classe sur les résultats scolaires

- Chaudhury, Gertler, Vermeersch (travail en cours)
 - Estimer l'effet de l'autonomie de l'école sur les apprentissages au Népal

Autonomie de l'école au Népal

□ Le but est d'évaluer

- A. La gestion scolaire autonome par les communautés
- B. Bulletins scolaires
- C. Réseaux d'information scolaire

□ Données

- Nous pourrions intégrer environ 1500 écoles dans l'évaluation
- Chaque communauté choisit librement de participer ou non.
- Bulletins scolaires et réseaux scolaires faits par les ONG
 - Le réseau scolaire ne peut être fait que dans des écoles autonomes
 - Les bulletins scolaires peuvent être faits dans n'importe quelle école.

Autonomie de l'école au Népal

- Le but est d'évaluer
 - A. La gestion scolaire autonome par les communautés
 - B. Bulletins scolaires
 - C. Réseaux d'information scolaire
- Données
 - Nous pourrions intégrer environ 1500 écoles dans l'évaluation
 - Chaque communauté choisit librement de participer ou non.
 - Bulletins scolaires et réseaux scolaires faits par les ONG
 - Le réseau scolaire ne peut être fait que dans des écoles autonomes
 - Les bulletins scolaires peuvent être faits dans n'importe quelle école.
 - Assumer que chaque communauté a exactement une école
- Tâche : concevoir l'exécution du programme pour qu'il puisse être évalué – proposer une méthode d'évaluation.

Autonomie de l'école au Népal

Interventions: A. La gestion scolaire autonome par les communautés
 B. Bulletins scolaires
 C. Réseaux d'information scolaire

| | Creation of demand for devolution to the community (A) | Feedback of performance indicators (B) | Network support after devolution to the community (C) | Number of schools in the group |
|-------------------|---|---|--|---------------------------------------|
| Group O (Control) | No | No | No | 200 |
| Group A | Yes | No | No | 300 |
| Group B | No | Yes | No | 100 |
| Group AB | Yes | Yes | No | 300 |
| Group AC | Yes | No | Yes | 300 |
| Group ABC | Yes | Yes | Yes | 300 |
| Total | | | | 1500 |

Rappel et avertissement....

□ $\text{corr}(Z, \varepsilon) = 0$

- si $\text{corr}(Z, \varepsilon) \neq 0$ « mauvais instrument » Problème !
- Il est difficile de trouver un bon instrument !
- Utiliser la théorie et le bon sens pour en trouver un !
- Nous pouvons penser à des conceptions qui produisent de bons instruments.

□ $\text{corr}(Z, T) \neq 0$

- « Instruments faibles » : la corrélation entre Z et T doit être suffisamment solide.
- Sinon, le biais reste grand, même pour de grandes tailles d'échantillons.