Économétrie

Régresseurs non déterministes

Stéphane Adjemian

stephane.adjemian@univ-lemans.fr

Septembre 2025

Régresseurs aléatoires

- Dans les chapitres I et II, nous avons supposé que les variables explicatives, X, sont détermnistes.
- Cette hypothèse n'est généralement pas raisonnable.
- Exemple Si le modèle est : $y_t = \rho y_{t-1} + \varepsilon_t$ où ε_t est une variable aléatoire normale (AR(1)).
- Même si la condition initiale, y_0, est déterministe, la variable y est clairement aléatoire (à gauche, comme d'habitude, mais aussi à droite).
- Quelles sont les conséqences pour l'inférence ?

Erreurs de mesure

- Supposons que le DGP soit $\mathbf{y} = X^\star\beta + \varepsilon$, où les conditions idéales sont vérifiées et où la matrice des régresseurs $X^\star$ est déterministe.

- Supposons que les variables explicatives soient mesurées avec des erreurs, on observe seulement $X = X^\star + \eta$, où η est une variable aléatoire (centrée).

- Les variables explicatives sont observées à un aléa près $\Rightarrow$ Les variables explicatives, X, considérées par l'économètre sont aléatoires.

- Est-il possible d'obtenir une estimation sans biais de β en régressant $\mathbf{y}$ sur X ? Convergente ? Quelles sont les propriétés de l'estimateur des MCO ?

- Ici la variable explicative est aléatoire, on verra que cela biaise l'estimation car elle est corrélée avec l'erreur du modèle.

Double causalité

- Supposons que l'on s'intéresse à l'effet du revenu sur la santé dans une population. On considère le modèle :

$$\text{Santé}_i = \beta_0 + \beta_1 \text{Revenu}_i + u_i$$

- Les individus les plus riches mangent mieux et vivent dans de meilleurs environnements $\rightarrow \beta_1 > 0$

- Mais une meilleure santé accroît la productivité des individus, on peut donc lire la causalité dans l'autre sens, avec par exemple un modèle de la forme :

$$\text{Revenu}_i = \alpha_0 + \alpha_1 \text{Santé}_i + v_i$$

avec $\alpha_1 > 0$.

- Cette double causalité induit une corrélation entre le revenu de l'individu i et u_i. $\Rightarrow$ On verra que cela biaise l'estimation de β_1 par les MCO dans le premier modèle.

- En substituant la seconde équation dans la première (on élimine le revenu), on obtient :

$$\text{Santé}_i = \beta_0 + \beta_1 \left(\alpha_0 + \alpha_1 \text{Santé}_i + v_i\right) + u_i$$

$$\Leftrightarrow \text{Santé}_i = \frac{\beta_0 + \beta_1\alpha_0}{1 - \beta_1\alpha_1} + \frac{\beta_1 v_i}{1 - \beta_1\alpha_1} + \frac{u_i}{1 - \beta_1\alpha_1}$$

- En substituant dans la seconde equation :

$$\text{Revenu}_i = \alpha_0 + \alpha_1 \frac{\beta_0 + \beta_1\alpha_0}{1 - \beta_1\alpha_1} + \left(1 + \frac{\alpha_1\beta_1}{1 - \alpha_1\beta_1}\right) v_i + \frac{\alpha_1 u_i}{1 - \alpha_1\beta_1}$$

$$\Leftrightarrow \text{Revenu}_i = \alpha_0 + \alpha_1 \frac{\beta_0 + \beta_1\alpha_0}{1 - \beta_1\alpha_1} + \frac{v_i}{1 - \alpha_1\beta_1} + \frac{\alpha_1 u_i}{1 - \alpha_1\beta_1}$$

- La corrélation entre revenu et le terme d'erreur dans le premier modèle est donc non nulle:

$$\text{corr}\left(\text{Revenu}_i, u_i\right) = \frac{\alpha_1\sigma_u^2}{1 - \alpha_1\beta_1}$$

en supposant que u_i et v_i sont non corrélés.

- Clairement les deux variables considérée ici (le revenu et la santé) sont aléatoires. C'est la corrélation, dans le premier modèle, entre le revenu et le terme d'erreur qui va biaiser l'estimateur de β_1.

- Si une personne est en meilleure santé pour des raisons non observées (captées par u_i), cette meilleure santé accroît aussi son revenu, le revenu observé est donc corrélé au terme d'erreur.

Le modèle de la Nature

$$\mathbf{y} = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_T)$
- X est une matrice aléatoire $T \times K$
- $\frac{X'X}{T} \xrightarrow[T\to\infty]{\text{proba}} Q$ est une finie et de plein rang.
- Les régresseurs sont linéairement indépendants avec probabilité 1.
- Les colonnes de X ne sont pas nécessairement toutes aléatoire (constante).
- **Problème:** Le modèle est incomplet puisque nous ne disons rien de la loi de X.

Cas 1: X et ε indépendants

- La distribution de ε conditionnelle à X est identique à sa distribution marginale.
- $\varepsilon | X \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_T)$
- Tous les résultats du Chapitre 1 sont valides **conditionnellement à** X.
- L'estimateur MCO conserve ses propriétés désirables.
- Les tests usuels restent valides (conditionnellement à X)

Intuition: En conditionnant sur X, on le traite les variables explicatives comme non-stochastiques. L'indépendance garantit que la valeur particulière de X soit sans conséquence.

Limites des énoncés conditionnels

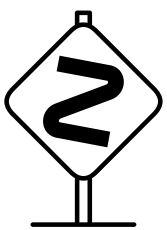
- Si X et ε sont tous deux stochastiques, l'utilité d'énoncés conditionnels sur X est limitée.

- Pour faire des inférences **non conditionnelles**, il faut connaître la distribution de X(difficile).

- Si x_{ti} sont iid et si $\mathbb{E}[x_{ti}\varepsilon_t]$ existe (nul), alors la loi des grands nombres s'applique :

$$\frac{1}{T}\sum_{t=1}^{T} x_{ti}\varepsilon_t \xrightarrow[T\to\infty]{\text{proba}} 0 \quad (i = 1, 2, \ldots, K)$$

  Cela assure la convergence de l'estimateur des MCO.



Les x_{ti} peuvent ne pas être iid, et $\mathbb{E}[x_{ti}\varepsilon_t]$ peut ne pas exister ! On ne peut alors rien dire de la convergence de l'estimateur des MCO.

Distribution asymptotique

Problème:

- On ne peut pas garantir que $\hat{\beta}$ sera asymptotiquement normal
- Les théorèmes de limite centrale ne s'appliquent pas directement car $\hat{\beta}$ est une combinaison **non-linéaire** de X et ε
- La non-linéarité vient de l'inversion de $X'X$

Sans information sur la distribution de X:

- Tests d'hypothèses valides impossibles, même asymptotiquement

Cas 2: Corrélation entre X et ε

$$\text{plim}\,\hat{\beta} = \beta + \text{plim}\frac{X'\varepsilon}{T} \neq \beta$$

$\Rightarrow \hat{\beta}$ **est inconsistant**

3.3 Variables Instrumentales: Le problème

Contexte:

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

Problème:

$$\text{plim}\frac{1}{T}X'\varepsilon \neq 0$$

$\Rightarrow$ L'estimateur MCO $\hat{\beta} = (X'X)^{-1}X'y$ est **inconsistant**

Solution: Trouver des **instruments** Z qui remplacent X dans la construction de l'estimateur

Théorème principal

Théorème 1

Supposons qu'il existe un ensemble de variables Z tel que:

1. $Q_{ZX} = plim\frac{1}{T}Z'X$ *est finie et non-singulière*
2. $\frac{Z'\varepsilon}{\sqrt{T}} \xrightarrow{d} N(0, \Psi)$

Alors l'estimateur

$$\tilde{\beta} = (Z'X)^{-1}Z'y$$

est consistant, et la distribution asymptotique de $\sqrt{T}(\tilde{\beta} - \beta)$ *est:*

$$N\left(0, Q_{ZX}^{-1}\Psi(Q'_{ZX})^{-1}\right)$$

Preuve: Consistance

Preuve de la consistance:

On a:

$$\tilde{\beta} = (Z'X)^{-1}Z'y = (Z'X)^{-1}Z'(X\beta + \varepsilon)$$
$$= \beta + (Z'X)^{-1}Z'\varepsilon$$

Donc:

$$\text{plim}\,\tilde{\beta} = \beta + Q_{ZX}^{-1} \cdot \text{plim}\frac{Z'\varepsilon}{T}$$

Puisque $\frac{Z'\varepsilon}{\sqrt{T}}$ a une distribution asymptotique bien définie, nécessairement:

$$\text{plim}\frac{Z'\varepsilon}{T} = 0$$

D'où: $\text{plim}\,\tilde{\beta} = \beta$ □

Preuve: Distribution asymptotique

Distribution asymptotique:
On a:

$$\sqrt{T}(\tilde{\beta} - \beta) = \left(\frac{Z'X}{T}\right)^{-1} \frac{Z'\varepsilon}{\sqrt{T}}$$

Puisque:

- $\frac{Z'\varepsilon}{\sqrt{T}} \xrightarrow{d} N(0, \Psi)$
- $\frac{Z'X}{T} \xrightarrow{p} Q_{ZX}$

Par le théorème de Slutsky:

$$\sqrt{T}(\tilde{\beta} - \beta) \xrightarrow{d} N\left(0, Q_{ZX}^{-1}\Psi(Q'_{ZX})^{-1}\right)$$

□

Définitions et remarques

Définition.

L'estimateur $\tilde{\beta} = (Z'X)^{-1}Z'y$ est appelé **estimateur par variables instrumentales** (IV) de β. La matrice Z est l'ensemble des **instruments** pour X.

Remarque.

Le cas typique est celui où:

$$\frac{Z'\varepsilon}{\sqrt{T}} \xrightarrow{d} N(0, \sigma^2 Q_{ZZ})$$

où $Q_{ZZ} = \text{plim}\frac{1}{T}Z'Z$

Exemple: Offre et demande (simultanéité)

Modèle de marché:
Demande (consommateurs):

$$Q_t^D = \alpha_0 + \alpha_1 P_t + \alpha_2 R_t + u_t$$

où R_t est le revenu des consommateurs, $\alpha_1 < 0$
Offre (producteurs):

$$Q_t^S = \beta_0 + \beta_1 P_t + \beta_2 W_t + v_t$$

où W_t est le coût des inputs, $\beta_1 > 0$
Équilibre:

$$Q_t^D = Q_t^S = Q_t$$

Problème: Le prix P_t est **endogène** (déterminé simultanément avec Q_t)

Exemple (suite): Prix d'équilibre

Résolution du système:
À l'équilibre, $Q_t^D = Q_t^S$:

$$\alpha_0 + \alpha_1 P_t + \alpha_2 R_t + u_t = \beta_0 + \beta_1 P_t + \beta_2 W_t + v_t$$

Le prix d'équilibre est:

$$P_t = \frac{(\beta_0 - \alpha_0) + \beta_2 W_t - \alpha_2 R_t + (v_t - u_t)}{\alpha_1 - \beta_1}$$

Observation cruciale:

$$\mathbb{E}[P_t \cdot u_t] = \mathbb{E}\left[\frac{v_t - u_t}{\alpha_1 - \beta_1} \cdot u_t\right] = \frac{-\sigma_u^2}{\alpha_1 - \beta_1} \neq 0$$

$\Rightarrow$ Le prix est corrélé avec l'erreur de demande!
$\Rightarrow$ MCO sur l'équation de demande est **inconsistant**

Exemple (suite): Estimation par VI

Objectif: Estimer l'équation de demande

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 R_t + u_t$$

Instruments pour P_t:

$$Z_t = (1, R_t, W_t)'$$

Justification:

- W_t (coût des inputs) affecte l'offre mais pas directement la demande
- W_t est corrélé avec P_t (via l'équation d'offre)
- W_t n'est pas corrélé avec u_t (erreur de demande)
- R_t (revenu) peut être utilisé comme instrument pour lui-même

Exemple (suite): Vérification des conditions

Conditions du théorème IV:

1. **Pertinence:** $Q_{ZX} = \text{plim}\frac{1}{T}Z'X$ finie et non-singulière
   - $X = (1, P_t, R_t)$, $Z = (1, R_t, W_t)$
   - W_t est corrélé avec P_t via l'équation d'offre
   - Si $\text{Var}(W_t) > 0$ et $\beta_2 \neq 0$, alors Q_{ZX} est non-singulière
2. **Exogénéité:** $\frac{Z'u}{\sqrt{T}} \xrightarrow{d} N(0, \Psi)$
   - Supposons que (R_t, W_t) sont stationnaires avec moments finis
   - Supposons que u_t est iid $N(0, \sigma_u^2)$ et indépendant de (R_s, W_s) pour tout t, s
   - Alors par un TCL pour processus stochastiques:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t u_t \xrightarrow{d} N(0, \sigma_u^2 Q_{ZZ})$$

où $Q_{ZZ} = \text{plim}\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$

$\Rightarrow$ L'estimateur IV $\tilde{\alpha} = (Z'X)^{-1}Z'Q$ est **consistant**

Exemple (suite): Pourquoi ces conditions sur Z?

Question: Pourquoi avons-nous besoin d'hypothèses sur (R_t, W_t) pour appliquer le TCL?

Réponse: Le TCL s'applique à la somme $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t u_t$

- Même si u_t est iid, le produit $Z_t u_t$ est **stochastique**
- Pour que le TCL s'applique, il faut que:
  1. $\mathbb{E}[Z_t u_t] = 0$ (exogénéité des instruments)
  2. $\text{Var}(Z_t u_t) < \infty$ (nécessite moments finis de Z_t)
  3. La LGN s'applique à $\frac{1}{T} \sum Z_t Z_t' u_t^2$ (stationnarité)

Exemple (suite): Ce qui peut mal tourner

Ce qui peut mal tourner sans conditions sur Z:

1. **Variance infinie:** Si Z_t a des queues trop épaisses
   - Var$(Z_t u_t)$ peut ne pas exister
   - Le TCL ne s'applique pas
2. **Dépendance temporelle:** Si Z_t est très persistant
   - Les $Z_t u_t$ ne sont pas "suffisamment indépendants"
   - La normalisation par $\sqrt{T}$ peut être incorrecte
3. **Non-stationnarité:** Si Var(Z_t) croît avec t
   - $Q_{ZZ} = \text{plim} \frac{1}{T} \sum Z_t Z_t'$ peut ne pas exister
   - La distribution asymptotique est mal définie

Conclusion: Les conditions sur Z garantissent que la somme $\sum Z_t u_t$ se comporte "régulièrement"

Mise en garde importante

Erreur courante dans la littérature:

- On affirme parfois que si Q_{ZX} est finie et non-singulière et si:

$$\text{plim}\frac{1}{T}Z'\varepsilon = 0$$

alors $\tilde{\beta}$ est consistant et $\sqrt{T}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q_{ZZ}^{-1} Q_{ZX}(Q'_{ZX})^{-1})$

FAUX!

- La consistance est vraie
- Mais l'énoncé sur la distribution asymptotique est **faux**
- Raison: $\text{plim}\frac{1}{T}Z'\varepsilon = 0$ n'implique pas que $\frac{Z'\varepsilon}{\sqrt{T}} \xrightarrow{d} N(0, \sigma^2 Q_{ZZ})$

Difficulté de vérification

Problème pratique:
Il est clairement plus facile de vérifier:

$$\text{plim}\frac{1}{T}Z'\varepsilon = 0$$

que de vérifier:

$$\frac{Z'\varepsilon}{\sqrt{T}} \text{ a une distribution asymptotique bien définie}$$

Question naturelle: Existe-t-il des conditions suffisantes facilement vérifiables?
Réponse: Si Z est non-stochastique et $Q_{ZZ} = \text{plim}\frac{1}{T}Z'Z$ est finie, c'est suffisant.
Mais: Si Z est stochastique, c'est compliqué. Les conditions suffisantes qui existent sont trop exigeantes pour être vraiment utiles.

Conditions fortes (mais insuffisantes)

Tentative de condition: Les observations Z_t sont iid et indépendantes de toutes les observations sur ε

Mais même cela ne suffit pas!
Contre-exemple: Reprenons l'exemple de la Section 3.1 où:

$$Z_t = \frac{1}{X_t}$$

avec X_t iid $N(0, \sigma_X^2)$ et indépendant de ε.

- Les Z_t sont aussi iid et indépendants de ε
- **Mais:** $Z_t \varepsilon_t$ a une distribution de Cauchy (dont la moyenne n'existe pas)
- Donc $\frac{1}{T} \sum_{t=1}^{T} Z_t \varepsilon_t$ aussi
- On n'a même pas plim $\frac{1}{T} Z' \varepsilon = 0$
- Encore moins une distribution asymptotique bien définie pour $\frac{Z' \varepsilon}{\sqrt{T}}$

3.4 Erreurs de Mesure: Le modèle

Modèle de régression linéaire simple:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Problème: y et x sont observés avec erreur
Variables observées:

$$y_t^* = y_t + v_t = \alpha + \beta x_t + (\varepsilon_t + v_t)$$
$$x_t^* = x_t + u_t$$

Forme alternative:

$$y_t^* = \alpha + \beta x_t^* + (\varepsilon_t + v_t - \beta u_t)$$

Observation: La perturbation composée dépend de x_t^* via u_t!

Théorème d'inconsistance

Théorème 2

Considérons le modèle ci-dessus, où ε_t, v_t, u_t et x_t sont tous indépendants entre eux, et sont de plus iid comme $N(0, \sigma_\varepsilon^2)$, $N(0, \sigma_v^2)$, $N(0, \sigma_u^2)$ et $N(\mu, \sigma_x^2)$ respectivement.
Alors l'estimateur MCO $\hat{\beta}$ de β est ***inconsistant*** *tant que $\sigma_u^2 \neq 0$, et:*

$$\textit{plim}\,\hat{\beta} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

Biais: plim $\hat{\beta} < \beta$ (en valeur absolue)
C'est un **biais d'atténuation** vers zéro.

Preuve du théorème

Preuve:
L'estimateur MCO de β est:

$$\hat{\beta} = \frac{\sum_t (x_t^* - \bar{x}^*)(y_t^* - \bar{y}^*)}{\sum_t (x_t^* - \bar{x}^*)^2} = \frac{(1/T)\sum_t (x_t^* - \bar{x}^*)(\varepsilon_t + v_t - \beta u_t)}{(1/T)\sum_t (x_t^* - \bar{x}^*)^2}$$

Puisque $x_t^* = x_t + u_t$, sous les hypothèses:

$$\text{plim} \frac{1}{T} \sum_t (x_t^* - \bar{x}^*)(\varepsilon_t + v_t - \beta u_t) = -\beta \sigma_u^2$$

$$\text{plim} \frac{1}{T} \sum_t (x_t^* - \bar{x}^*)^2 = \sigma_x^2 + \sigma_u^2$$

D'où:

$$\text{plim}\, \hat{\beta} = \beta \frac{-\sigma_u^2}{\sigma_x^2 + \sigma_u^2} + \beta = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

□

La mesure de y ne pose pas de problème

Remarque.

C'est l'**erreur de mesure sur** x qui cause le problème.
L'erreur v_t sur y est indistinguable de la perturbation habituelle ε_t.
Pour le reste de cette section, on incorporera donc ε_t dans v_t.

Simplification:
On considère désormais:

$$y_t^* = y_t + v_t = \alpha + \beta x_t + v_t$$
$$x_t^* = x_t + u_t$$

Lien avec les variables instrumentales

Question: Peut-on utiliser un estimateur IV pour corriger l'erreur de mesure?

Réponse: Oui! L'erreur de mesure est un cas particulier d'endogénéité.

Rappel du problème:

$$\begin{aligned} y_t &= \alpha + \beta x_t + v_t \\ x_t^* &= x_t + u_t \\ \Rightarrow y_t &= \alpha + \beta x_t^* + \underbrace{(v_t - \beta u_t)}_{\text{erreur composée}} \end{aligned}$$

Corrélation: $\text{Cov}(x_t^*, v_t - \beta u_t) = -\beta\sigma_u^2 \neq 0$

$\Rightarrow$ Problème d'endogénéité classique, où x_t^* est corrélé avec l'erreur!

Variables instrumentales pour erreurs de mesure

Solution IV: Trouver un instrument Z_t tel que:

1. **Pertinence:** Z_t corrélé avec x_t (ou x_t^*)

$$\text{Cov}(Z_t, x_t) \neq 0$$

2. **Exogénéité:** Z_t non corrélé avec les erreurs u_t et v_t

$$\text{Cov}(Z_t, u_t) = 0 \quad \text{et} \quad \text{Cov}(Z_t, v_t) = 0$$

Exemples d'instruments possibles:

- **Mesures multiples:** Si on a deux mesures indépendantes de x_t
  - $x_t^{(1)} = x_t + u_t^{(1)}$ et $x_t^{(2)} = x_t + u_t^{(2)}$
  - On peut utiliser $x_t^{(2)}$ comme instrument pour $x_t^{(1)}$
  - Si $u_t^{(1)} \perp u_t^{(2)}$, alors $\text{Cov}(x_t^{(2)}, u_t^{(1)}) = 0$
- **Variables liées:** Une variable Z_t qui affecte x_t mais pas directement y_t

Exemple: Mesures multiples

Exemple concret: Mesure du revenu avec erreur

- Modèle: $\text{Consommation}_t = \alpha + \beta \text{Revenu}_t + v_t$
- Problème: Le revenu est mesuré avec erreur
  - Déclaration fiscale: $\text{Revenu}_t^{(1)} = \text{Revenu}_t + u_t^{(1)}$
  - Enquête auprès des ménages: $\text{Revenu}_t^{(2)} = \text{Revenu}_t + u_t^{(2)}$
- Si $u_t^{(1)}$ et $u_t^{(2)}$ sont des erreurs de mesure indépendantes:
  - Utiliser $\text{Revenu}_t^{(2)}$ comme instrument pour $\text{Revenu}_t^{(1)}$
  - Ou vice-versa

Estimateur IV:

$$\tilde{\beta} = \frac{\sum_t (x_t^{(2)} - \bar{x}^{(2)})(y_t - \bar{y})}{\sum_t (x_t^{(2)} - \bar{x}^{(2)})(x_t^{(1)} - \bar{x}^{(1)})}$$

Cet estimateur est **consistant** pour β!

Références

- Greene, William H. (2017). *Econometric analysis*. Pearson.
- Schmidt, Peter (1976). *Econometrics*. Taylor & Francis.