

Endogénéité et estimation par variables instrumentales

Aristide E. Houndetoungan

6 Octobre 2021

Introduction

- Modèle linéaire-en-moyennes : $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- Hypothèse d'exogénéité des \mathbf{X} (hypothèse H.3, chapitre 1) : $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0 \ \forall \ i$.
- Elle implique que ($\forall \ i$) :
 - $\mathbb{E}(\varepsilon_i) = \mathbb{E}_{\mathbf{x}}(\mathbb{E}(\varepsilon_i|\mathbf{x}_i)) = 0$,
 - $\mathbb{E}(\varepsilon_i\mathbf{x}_i) = \mathbb{E}_{\mathbf{x}}(\mathbb{E}(\varepsilon_i\mathbf{x}_i|\mathbf{x}_i)) = \mathbb{E}_{\mathbf{x}}(\mathbb{E}(\varepsilon_i|\mathbf{x}_i)\mathbf{x}_i) = 0$.
- Elle est cruciale pour montrer que $\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{MCO}}) = \boldsymbol{\beta}_0$ et $\text{plim}(\hat{\boldsymbol{\beta}}_{\text{MCO}}) = \boldsymbol{\beta}_0$.

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}),$$

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon},$$

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon}.$$

Donc,

$$\mathbb{E} \left(\hat{\beta}_{\text{MCO}} | \mathbf{X} \right) = \beta + \mathbb{E} \left((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon | \mathbf{X} \right) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbb{E} (\varepsilon | \mathbf{X}) = \beta,$$

$$\mathbb{E} \left(\hat{\beta}_{\text{MCO}} \right) = \mathbb{E}_{\mathbf{X}} \left(\mathbb{E} \left(\hat{\beta}_{\text{MCO}} | \mathbf{X} \right) \right) = \mathbb{E}_{\mathbf{X}} (\beta) = \beta.$$

Aussi,

$$\text{plim} \left(\hat{\beta}_{\text{MCO}} \right) = \beta + \text{plim} \left(\left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}'\varepsilon}{n} \right),$$

$$\text{plim} \left(\hat{\beta}_{\text{MCO}} \right) = \beta + \text{plim} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \text{plim} \left(\frac{\mathbf{X}'\varepsilon}{n} \right).$$

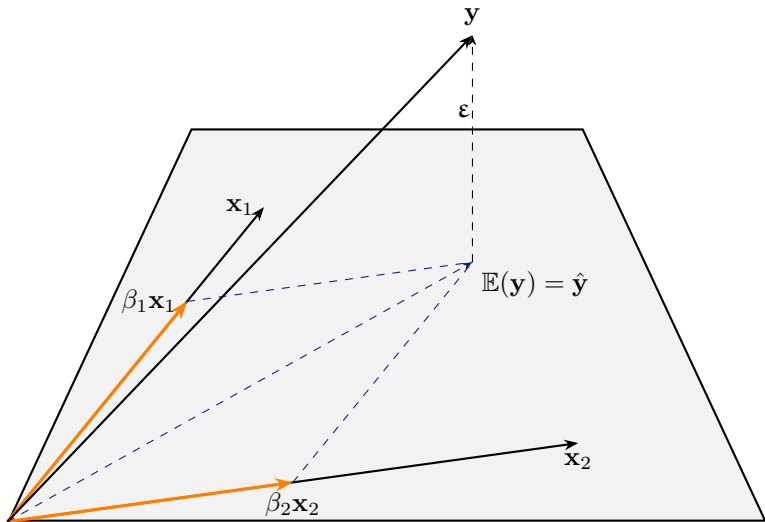
On suppose que $\text{plim} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right) = \mathbf{Q}_{\mathbf{xx}}$ et par la loi des grands nombres

(LGN), $\text{plim} \left(\frac{\mathbf{X}'\varepsilon}{n} \right) = \mathbb{E} (\varepsilon_i \mathbf{x}_i) = 0$. Donc,

$$\text{plim} \left(\hat{\beta}_{\text{MCO}} \right) = \beta + \mathbf{Q}_{\mathbf{xx}}^{-1} \times 0,$$

$$\text{plim} \left(\hat{\beta}_{\text{MCO}} \right) = \beta.$$

- Projection de y dans l'espace de x en cas **d'exogénéité**.



- L'estimateur des MCO est convergent.

Endogénéité

- L'hypothèse d'exogénéité entre \mathbf{x}_i et ε_i n'est pas souvent vérifiée.
- Exemples :
 - ① Omission de variables pertinentes : (demande d'essence, E)
 - Vrai modèle : $\log(E) = \beta_1 + \beta_2 \log(\text{prix}) + \beta_3 \log(\text{revenu}) + \varepsilon$
 - Spécification : $\log(E) = \beta_1 + \beta_2 \log(\text{prix}) + \omega$
 - Le terme d'erreur $\omega = \beta_3 \log(\text{revenu}) + \varepsilon$ peut être corrélé au prix.
 - ② Erreur de mesure ou variable proxy
 - Vrai modèle : $y = \beta_1 + \beta_2 x + \varepsilon$
Mais en pratique x est mesuré par $\tilde{x} = x + u$
 - Spécification : $y = \beta_1 + \beta_2 \tilde{x} + \omega$
 - Le terme d'erreur $\omega = -\beta_2 u + \varepsilon$ peut être corrélé à \tilde{x} .

③ Simultanéité (Offre et Demande)

$$\left\{ \begin{array}{l} \text{Offre :} \quad Y_O = \beta_1 + \beta_2 \text{Prix} + \beta_3 \text{Prix_intransit} + \varepsilon_O, \\ \text{Demande :} \quad Y_D = \alpha_1 + \alpha_2 \text{Prix} + \alpha_3 \text{Revenu} + \varepsilon_D, \\ \text{Equilibre :} \quad Y_O = Y_D \end{array} \right.$$

A partir de l'équilibre,

$\text{Prix} = (\alpha_1 - \beta_1 + \alpha_3 \text{Revenu} - \beta_3 \text{Prix_intransit} + \varepsilon_D - \varepsilon_O) / (\beta_2 - \alpha_2)$. Le prix est corrélé à ε_D .

④ Effet de traitement endogène

- $\text{Revenu} = \mathbf{x}'\boldsymbol{\beta} + \gamma \text{Education} + \varepsilon$

Facteurs inobservés (par l'économètre) qui expliquent le revenu et le niveau d'éducation. L'éducation est potentiellement corrélée à ε .

- Endogénéité : $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) \neq 0$.
- Biais de l'estimateur des MCO.

$$\mathbb{E}(\hat{\beta}_{\text{MCO}}|\mathbf{X}) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \underbrace{\mathbb{E}(\varepsilon|\mathbf{X})}_{\neq 0},$$

$$\mathbb{E}(\hat{\beta}_{\text{MCO}}|\mathbf{X}) \neq \beta \quad \text{en général.}$$

- Non convergence de l'estimateur des MCO.

$$\text{plim}(\hat{\beta}_{\text{MCO}}) = \beta + \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \text{plim}\left(\frac{\mathbf{X}'\varepsilon}{n}\right).$$

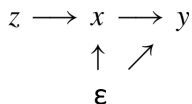
$$\text{plim}(\hat{\beta}_{\text{MCO}}) = \beta + \mathbf{Q}_{\mathbf{xx}}^{-1} \underbrace{\mathbb{E}(\varepsilon_i\mathbf{x}_i)}_{\neq 0},$$

$$\text{plim}(\hat{\beta}_{\text{MCO}}) \neq \beta \quad \text{en général.}$$

- Application avec R : script `non_convergence.mco.R`

Variables Instrumentales

- Modèle linéaire simple : $y = \beta_1 + \beta_2 x + \varepsilon$, avec endogénéité $\mathbb{E}(\varepsilon|x) \neq 0$.
- **Instrument** (ou **Variable Instrumentale**) pour x est une variable z ayant la propriété suivante : variations de z sont associées à des variations de x mais n'entraînent pas à une variation de y (hormis la voie indirecte via x).



- Autrement dit, z est un instrument de x dans le modèle $y = \beta_1 + \beta_2 x + \varepsilon$, si :
 - ① z n'est pas corrélé à ε ,
 - ② z est corrélé à x .

- **Exemple 1 : Simultanéité (Offre et Demande)**

$$\left\{ \begin{array}{l} \text{Offre :} \quad Y_O = \beta_1 + \beta_2 \textit{Prix} + \beta_3 \textit{Prix_intransant} + \varepsilon_O, \\ \text{Demande :} \quad Y_D = \alpha_1 + \alpha_2 \textit{Prix} + \alpha_3 \textit{Revenu} + \varepsilon_D, \\ \text{Equilibre :} \quad Y_O = Y_D \end{array} \right.$$

A partir de l'équilibre,

$\textit{Prix} = (\alpha_1 - \beta_1 + \alpha_3 \textit{Revenu} - \beta_3 \textit{Prix_intransant} + \varepsilon_D - \varepsilon_O) / (\beta_2 - \alpha_2)$. Le prix est une variable endogène dans l'Equation de demande. Un instrument possible du prix est le prix des intrants.

- **Exemple 2 : Impact de l'éducation sur le salaire**

- L'éducation est endogène.

- Instrument 1 : distance entre l'adresse de résidence et l'université ou le collège [voir Card, David. (1993). *Using geographic variation in college proximity to estimate the return to schooling.*].

- Cet instrument requiert la prise en compte de régresseurs supplémentaires tels que des indicatrices de zones non métropolitaines .

- Instrument 2 : Mois de naissance [voir Angrist, Joshua D., et Alan B. Keueger. (1991). *Does compulsory school attendance affect schooling and earnings ?*].

Cet instrument est souvent critiqué dans la littérature [voir Bound, Jaeger, et Baker. (1995). *Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak.*].

Hypothèses

- Hypothèses de la méthode des MCO

H.1. Linéarité : $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$.

H.2. Indépendance linéaire des variables explicatives : $\text{rang}(\mathbf{X}) = K$.

H.3. Exogénéité des variables explicatives : $\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0 \implies \varepsilon_i \perp \mathbf{x}_i$.

H.4. Homoscédasticité et non autocorrélation des erreurs.

H.5. Normalité des erreurs : chaque ε_i suit une distribution normale.

H.6. $(\varepsilon_1, \mathbf{x}_1), \dots, (\varepsilon_n, \mathbf{x}_n)$ sont i.i.d.

- Dans ce chapitre, l'hypothèse H.3. n'est pas vérifiée : $\mathbb{E}(\varepsilon_i | \mathbf{x}_i) \neq 0$.

- Existence d'un ensemble de variables additionnelles, \mathbf{Z}

 - exogènes par rapport à ε : $\mathbb{E}(\varepsilon | \mathbf{z}) = 0$;

 - fortement corrélées avec \mathbf{X} .

- La matrice \mathbf{Z} de dimension $n \times L$ est appelée matrice des variables instrumentales.

• Hypothèses additionnelles

H.7. $(\varepsilon_1, \mathbf{x}_1, \mathbf{z}_1), \dots, (\varepsilon_n, \mathbf{x}_n, \mathbf{z}_n)$ sont i.i.d.

H.8. $\mathbb{E}(\varepsilon_i | \mathbf{z}_i) = 0$ pour tout i (exogénéité de \mathbf{Z} par rapport à ε).

- H.8. implique que $\mathbb{E}(\mathbf{z}_i \varepsilon_i) = \mathbb{E}_{\mathbf{z}} (\mathbb{E}(\mathbf{z}_i \varepsilon_i | \mathbf{z}_i)) = \mathbb{E}_{\mathbf{z}} (\mathbf{z}_i \mathbb{E}(\varepsilon_i | \mathbf{z}_i)) = 0$.

- Par la LGN, $\text{plim} \frac{\mathbf{Z}' \boldsymbol{\varepsilon}}{n} = \mathbb{E}(\mathbf{z}_i \varepsilon_i) = \mathbf{0}$.

H.9. $\text{plim} \frac{\mathbf{Z}' \mathbf{Z}}{n} = \mathbf{Q}_{\mathbf{zz}}$ est une matrice finie et définie positive.

H.10. $\text{plim} \frac{\mathbf{Z}' \mathbf{X}}{n} = \mathbf{Q}_{\mathbf{zx}}$ est une matrice finie non nulle $L \times K$ de rang K .

- H.10. requiert que \mathbf{Z} soit corrélé à \mathbf{X} et que $L \geq K$. Modèle non identifié si $L < K$, juste identifié si $L = K$ et sur identifié si $L > K$
- Seuls des régresseurs du modèle initial ne peuvent pas constituer \mathbf{Z} .
- Naturellement, toutes les variables explicatives non corrélées à ε sont incluses dans \mathbf{Z} . Ces variables sont par la suite complétées par des variables additionnelles (autres que des régresseurs) comme instruments des variables explicatives endogènes (exclues de \mathbf{Z}).

Méthode des variables instrumentales (IV)

- S'utilise seulement lorsque $L = K$.

$$\text{plim} \left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right) = \text{plim} \left(\frac{\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n} \right) = \mathbf{0}$$

$$\text{plim} \left(\frac{\mathbf{Z}'\boldsymbol{\varepsilon}}{n} \right) = \text{plim} \left(\frac{\mathbf{Z}'\mathbf{y}}{n} \right) - \text{plim} \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right) \boldsymbol{\beta} = \mathbf{0}$$

$$\text{plim} \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right) \boldsymbol{\beta} = \text{plim} \left(\frac{\mathbf{Z}'\mathbf{y}}{n} \right)$$

Si $L = K$, alors $\mathbf{Z}'\mathbf{X}$ est une matrice carrée et $\text{plim} \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right)$ est inversible.

Donc,

$$\boldsymbol{\beta} = \left[\text{plim} \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right) \right]^{-1} \text{plim} \left(\frac{\mathbf{Z}'\mathbf{y}}{n} \right)$$

- L'estimateur de variable instrumentale est alors (si $L = K$).

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y} \quad (1)$$

- Par construction, $\hat{\beta}_{IV}$ est convergent : $\text{plim}(\hat{\beta}_{IV}) = \beta$.
- En remplaçant $y = \mathbf{X}\beta + \varepsilon$ dans l'Eq. (1), on a,

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\varepsilon = \beta + (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\varepsilon,$$

$$\hat{\beta}_{IV} = \beta + \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{Z}'\varepsilon}{n}.$$

- En grand échantillon, $\frac{\mathbf{Z}'\varepsilon}{n} \overset{a}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n} \mathbf{Q}_{zz}\right)$. Donc,

$$\hat{\beta}_{IV} \overset{a}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{n} \mathbf{Q}_{zx}^{-1} \mathbf{Q}_{zz} \mathbf{Q}_{zx}^{-1}\right) \quad (2)$$

- Un estimateur convergent de σ^2 est,

$$\hat{\sigma}_{IV}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \mathbf{x}'_i \hat{\beta}_{IV} \right)^2. \quad (3)$$

- En pratique, la variance asymptotique de $\hat{\beta}_{IV}$ est estimée par,

$$\text{Est.Asy. Var}(\hat{\beta}_{IV}) = \hat{\sigma}_{IV}^2 (\mathbf{Z}'\mathbf{X})^{-1} (\mathbf{Z}'\mathbf{Z}) (\mathbf{X}'\mathbf{Z})^{-1}. \quad (4)$$

Méthode des doubles moindres carrés

- Modèle : $y = X\beta + \varepsilon$.
- Méthode plus générale ($L \geq K$).
- Si $L > K$, alors $X'Z$ n'est plus une matrice carrée et est donc non inversible.
- Doubles moindres carrés : méthode en deux étapes.
 - **Etape 1** : Projeter chaque terme du modèle dans l'espace formé par Z et on obtient,

$$P_z y = P_z X\beta + P_z \varepsilon. \quad (5)$$

où $P_z = Z(Z'Z)^{-1}Z'$.

Autrement dit, on régresse y et chaque X sur Z et on calcule la variable dépendante prédite.

- **Etape 2** : Appliquer la méthode des MCO à l'Eq. (5).

$$\begin{aligned}\hat{\beta}_{2SLS} &= [X'Z (Z'Z)^{-1} Z'X]^{-1} [X'Z (Z'Z)^{-1} Z'y] \\ \hat{\beta}_{2SLS} &= [X'P_z X]^{-1} [X'P_z y]\end{aligned} \quad (6)$$

- Modèle projeté : $\mathbf{P}_z \mathbf{y} = \mathbf{P}_z \mathbf{X} \beta + \mathbf{P}_z \varepsilon$.
- Intuition : Nouvelle variable explicative $\tilde{\mathbf{X}} = \mathbf{P}_z \mathbf{X}$ exogène par rapport au nouveau terme d'erreur $\tilde{\varepsilon} = \mathbf{P}_z \varepsilon$. En effet,

$$\text{plim} \left(\frac{\tilde{\mathbf{X}}' \tilde{\varepsilon}}{n} \right) = \text{plim} \left(\frac{(\mathbf{P}_z \mathbf{X})' \mathbf{P}_z \varepsilon}{n} \right) = \text{plim} \left(\frac{\mathbf{X}' \mathbf{P}_z' \mathbf{P}_z \varepsilon}{n} \right) = \text{plim} \left(\frac{\mathbf{X}' \mathbf{P}_z \varepsilon}{n} \right).$$

$$\text{plim} \left(\frac{\tilde{\mathbf{X}}' \tilde{\varepsilon}}{n} \right) = \text{plim} \left(\frac{\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \varepsilon}{n} \right),$$

$$\text{plim} \left(\frac{\tilde{\mathbf{X}}' \tilde{\varepsilon}}{n} \right) = \text{plim} \left[\left(\frac{\mathbf{X}' \mathbf{Z}}{n} \right) \left(\frac{\mathbf{Z}' \mathbf{Z}}{n} \right)^{-1} \left(\frac{\mathbf{Z}' \varepsilon}{n} \right) \right]$$

$$\text{plim} \left(\frac{\tilde{\mathbf{X}}' \tilde{\varepsilon}}{n} \right) = \text{plim} \left(\frac{\mathbf{X}' \mathbf{Z}}{n} \right) \text{plim} \left[\left(\frac{\mathbf{Z}' \mathbf{Z}}{n} \right)^{-1} \right] \text{plim} \left(\frac{\mathbf{Z}' \varepsilon}{n} \right),$$

$$\text{plim} \left(\frac{\tilde{\mathbf{X}}' \tilde{\varepsilon}}{n} \right) = \mathbf{0}.$$

- Donc $\hat{\beta}_{2\text{SLS}}$ est convergent.

- $\hat{\beta}_{2SLS} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}]$

- En remplaçant \mathbf{y} par $\mathbf{X}\beta + \varepsilon$, on a,

$$\hat{\beta}_{2SLS} = \beta + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\varepsilon],$$

$$\hat{\beta}_{2SLS} = \beta + \left[\left(\frac{\mathbf{X}'\mathbf{Z}}{n} \right) \left(\frac{\mathbf{Z}'\mathbf{Z}}{n} \right)^{-1} \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right) \right]^{-1} \left(\frac{\mathbf{X}'\mathbf{Z}}{n} \right) \left(\frac{\mathbf{Z}'\mathbf{Z}}{n} \right)^{-1} \left(\frac{\mathbf{Z}'\varepsilon}{n} \right).$$

- En grand échantillon, $\frac{\mathbf{Z}'\varepsilon}{n} \stackrel{a}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n}\mathbf{Q}_{zz}\right)$. Donc,

$$\hat{\beta}_{2SLS} \stackrel{a}{\sim} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}(\mathbf{Q}'_{zx}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1}\right) \quad (7)$$

- Un estimateur convergent de σ^2 est,

$$\hat{\sigma}_{2SLS}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \mathbf{x}'_i \hat{\beta}_{2SLS} \right)^2. \quad (8)$$

- En pratique, la variance asymptotique de $\hat{\beta}_{2SLS}$ est estimée par,

$$\text{Est.Asy. Var} \left(\hat{\beta}_{2SLS} \right) = \hat{\sigma}_{2SLS}^2 (\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1}. \quad (9)$$

Tests de spécification

Tests de Hausman et de Wu

- Une variable peut être considérée à tort comme endogène.
- Avec un instrument valide pour cette variable, $\hat{\beta}_{IV}$ ou $\hat{\beta}_{2SLS}$ est pourtant convergent.
- Puisque la variable n'est pas endogène, $\hat{\beta}_{MCO}$ est aussi convergent. Mieux encore, $\hat{\beta}_{MCO}$ est BLUE (variance plus petite que celles de $\hat{\beta}_{IV}$ et $\hat{\beta}_{2SLS}$).
- Important de tester si la variable est effectivement endogène (hypothèse alternative) ou non (hypothèse nulle).
- Si le test ne rejette pas l'hypothèse nulle, il est alors préférable de garder $\hat{\beta}_{MCO}$ qui est plus précis que $\hat{\beta}_{IV}$ et $\hat{\beta}_{2SLS}$.
- Deux tests couramment utilisés : test de **Hausman** et test de **Wu**.

- **Hypothèse nulle** : X est exogène ; **Hypothèse alternative** : X est endogène ;

- **Test de Hausman**

- **Intuition** : Sous l'hypothèse nulle, $\hat{\beta}_{2SLS}$ (ou $\hat{\beta}_{IV}$), ainsi que $\hat{\beta}_{MCO}$ sont convergents et devraient être très proches. Le test compare donc $d = \hat{\beta}_{IV} - \hat{\beta}_{MCO}$ à 0 . Si d est trop éloigné de 0 , on rejette l'hypothèse nulle. Dans le cas contraire, on ne la rejette pas.
- Statistique du test :

$$H = d' \{ \text{Est.Asy. Var} (d) \}^{-1} d \quad (10)$$

- Hausman montre que,

$$\text{Est.Asy. Var} (d) = \hat{\sigma}_{MCO}^2 \left((X'P_Z X)^{-1} - (X'X)^{-1} \right). \quad (11)$$

- La statistique d'Hausman $H \sim \chi^2(K)$ si \mathbf{X} et \mathbf{Z} n'ont pas de variables communes, ce qui est rarement le cas (généralement le vecteur de "uns" associé à l'intercept est inclus dans \mathbf{X} et \mathbf{Z}).
- En présence de variables communes dans \mathbf{X} et \mathbf{Z} , le rang de la matrice à inverser, $\text{Est.Asy. Var}(\mathbf{d})$, est $K^* < K$, où K^* est le nombre de variables explicatives endogènes, et on a besoin de recourir à un inverse généralisé.
- La statistique H suit alors une $\chi^2(K^*)$ sous l'hypothèse nulle.

- **Test de Wu**

- Soit \mathbf{X}^* les K^* variables explicatives dans \mathbf{X} exclues de \mathbf{Z} .
- On estime par MCO le modèle,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}^*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \quad (12)$$

où $\hat{\mathbf{X}}^*$ est la variation dépendante prédite lorsque \mathbf{X}^* est régressé sur \mathbf{Z} ; i.e., $\hat{\mathbf{X}}^* = \mathbf{P}_z \mathbf{X}^*$.

- Tester l'endogénéité revient à tester si $\boldsymbol{\gamma} = 0$ (test classique de nullité de coefficients). Si on rejette l'hypothèse $\boldsymbol{\gamma} = 0$, alors \mathbf{X}^* est endogène.

Tests de spécification

Test de suridentification

- Méthode de IV est développée autour des conditions d'orthogonalité :
$$\mathbb{E}(\mathbf{z}_i \varepsilon_i) = 0.$$
- Contrepartie empirique est l'équation de moments $\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i = 0.$
- Si $L = K$, alors l'équation de moments est un système de K équations à K inconnues et la solution est $\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}.$
- Lorsque $L > K$, le système n'a pas de solution. Mais le modèle est estimé par projection dans l'espace des Z . Rien ne garantit que la solution $\hat{\beta}_{2SLS}$ vérifie toujours l'équation de moments.
- Le test de suridentification consiste donc à vérifier si $\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i = 0$ dans le cas où $L > H$. L'hypothèse nulle est $\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \varepsilon_i = 0.$

- Soit $\bar{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \hat{\varepsilon}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \left(y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{2SLS} \right)$.
- La statistique du test est,

$$\bar{S} = \bar{\mathbf{m}}' \mathbb{V}\text{ar}(\bar{\mathbf{m}})^{-1} \bar{\mathbf{m}}, \quad (13)$$

- Sous l'hypothèse nulle $\bar{S} \sim \chi^2(L - K)$.
- Sous l'hypothèse nulle,

$$\mathbb{V}\text{ar}(\bar{\mathbf{m}}) = \frac{\hat{\sigma}^2}{n} \mathbf{Z}'\mathbf{Z}.$$

En pratique cette variance peut être remplacée par un estimateur convergent,

$$\text{Est. } \mathbb{V}\text{ar}(\bar{\mathbf{m}}) = \frac{\hat{\sigma}_{2SLS}^2}{n} \mathbf{Z}'\mathbf{Z}.$$

Instruments faibles

- Deux conditions d'un instrument :
 - ① **exogènes par rapport** à ε : $\mathbb{E}(\varepsilon|\mathbf{Z}) = 0$;
 - ② **fortement corrélées** avec \mathbf{X} .
- Les chercheurs se sont plus focalisés sur l'exogénéité (condition 1). Une littérature de plus en plus abondante soutient qu'une plus grande attention doit être également accordée à la corrélation entre l'instrument et la variable endogène.
- L'exogénéité garantit la convergence de l'estimateur. Toutefois, lorsque la corrélation entre \mathbf{Z} et \mathbf{X} est faible, i.e., quand $(1/n)\mathbf{Z}'\mathbf{X} \approx 0$, un certain nombre de problèmes ont été mis en lumière :
 - estimateur imprécis car $\text{Var}\left(\hat{\beta}_{2\text{SLS}}\right)$ est plus grande ;
 - estimateur fortement biaisé vers celui des MCO.

- Tester la faiblesse/force des instruments.
- Dans le cas d'une seule variable x^* endogène instrumentée par \mathbf{z}^* , on estime le modèle,

$$x_i^* = \mathbf{z}_i^{*'} \boldsymbol{\pi} + v_i. \quad (14)$$

Pour que l'instrument soit fort, la statistique de Fisher du modèle doit être supérieure à 10.

- Cette méthode est seulement valide dans le cas d'une seule variable explicative endogène.

- En présence de plusieurs variables explicatives endogènes, Godfrey (1999) propose une méthode alternative.
- Pour chaque variable explicative endogène, on calcule,

$$R_k^2 = \frac{\left[(\mathbf{X}'\mathbf{X})^{-1} \right]_{kk}}{\left[(\mathbf{X}'\mathbf{P}_z\mathbf{X})^{-1} \right]_{kk}}, \quad (15)$$

où l'indice kk signifie le $k^{\text{ème}}$ élément de la diagonale de la matrice.

- R_k^2 et il s'interprète comme un R^2 usuel. Lorsqu'il est proche de 1 la $k^{\text{ème}}$ variable explicative endogène est bien instrumentée (instruments forts).

Méthode des moments généralisée (GMM)

- Conditions de moments de la population conduisent à des contreparties empiriques qui peuvent être utilisées pour estimer les paramètres.
- Exemple :
 - Conditions de moments : $\mathbb{E}(\mathbf{z}_i \varepsilon_i) = \mathbb{E}(\underbrace{\mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})}_{\text{Fonct. de Moment}}) = 0$.
 - Contrepartie Empirique : $\frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta}) = 0$ qui implique $\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}$, si $L = K$.
- L'estimateur GMM généralise celui des MCO, IV, 2SLS et bien d'autres.
- Basé sur la contrepartie empirique de conditions de moments.
- Les conditions de moments nécessitent une fonction de moments.

- **Fonction de moments** : généralement notée $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$, une fonction de dimension $L \geq K$, où $K = \dim(\boldsymbol{\theta})$ et $\boldsymbol{\theta}$ est le paramètre à estimer, qui satisfait la condition,

$$\mathbb{E}(\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})) = \mathbf{0}. \quad (16)$$

La contrepartie empirique implique,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{0}. \quad (17)$$

- Lorsque $K = L$, l'Eq. (17) admet généralement une unique solution et on parle simplement de méthode des moments (MM).

Exemples de fonctions de moments

- **Modèle linéaire-en-moyennes avec exogénéité**

- Si \mathbf{x} est exogène, alors

$$\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \quad (18)$$

satisfait les conditions de moments.

- En effet, $\mathbb{E}(\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})) = \mathbb{E}(\mathbf{x}_i\varepsilon_i) = \mathbf{0}$.
- Avec la contrepartie empirique $\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$, on peut montrer que l'estimateur de MM est $\hat{\boldsymbol{\beta}}_{\text{MM}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$.
- L'estimateur des MCO est donc un estimateur de MM.

● Modèle linéaire-en-moyennes avec endogénéité

- La condition d'orthogonalité des instruments peut être utilisée pour définir la fonction de moments :

$$\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}). \quad (19)$$

- En effet, $\mathbb{E}(\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})) = \mathbb{E}(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$.
- Avec la contrepartie empirique $\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$, on peut montrer que si $L = K$, l'estimateur de MM est $\hat{\boldsymbol{\beta}}_{\text{MM}} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}$.
- L'estimateur de variables instrumentales est donc un estimateur de MM.

- **Modèle non linéaire**

- Modèle : $y_i = h(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i$, où h est une fonction non linéaire.
- Le terme d'erreur est $\varepsilon_i = y_i - h(\mathbf{x}_i, \boldsymbol{\theta})$. La condition d'orthogonalité des instruments peut être utilisée pour définir la fonction de moments.

$$\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{m}_i (y_i - h(\mathbf{x}_i, \boldsymbol{\theta})) . \quad (20)$$

où \mathbf{m}_i peut contenir des instruments ou des variables explicatives exogènes.
 \mathbf{m}_i peut être aussi une fonction d'instrument et de variables explicatives exogènes.

- Il est important d'inclure suffisamment d'instruments et de variables explicatives exogènes dans \mathbf{m}_i quitte à ce que $\dim(\mathbf{m}_i) = \dim(\boldsymbol{\theta})$

Estimateur GMM

- Même si $L = K$, ce n'est pas toujours simple de trouver un θ qui satisfait la contrepartie empirique des conditions de moments (surtout dans le cas d'un modèle non linéaire). L'estimateur de méthode de moments de θ , notée θ_{MM} , est définie comme étant la valeur de θ qui minimise,

$$\left[\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \theta) \right]' \left[\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \theta) \right]. \quad (21)$$

- Lorsque $L > K$, comme dans le cas de la méthode IV, la contrepartie empirique des conditions de moments n'admet pas de solution. On a alors recourt à l'estimateur GMM, noté $\hat{\theta}_{GMM}$, en minimisant

$$\left[\sum_{i=1}^n \frac{1}{n} \mathbf{g}(\mathbf{z}_i, \theta) \right]' \mathbf{W} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{z}_i, \theta) \right], \quad (22)$$

où \mathbf{W} est une matrice de poids de dimension $L \times L$.

- Dans le cas d'un modèle linéaire avec endogénéité, $g(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$,
Donc,

$$\hat{\boldsymbol{\theta}}_{\text{GMM}} = [\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y}], \quad (23)$$

- Si $\mathbf{W} = \left(\frac{1}{N}\mathbf{Z}'\mathbf{Z}\right)^{-1}$, alors $\hat{\boldsymbol{\theta}}_{\text{GMM}} = \hat{\boldsymbol{\theta}}_{\text{2SLS}}$.
- L'estimateur de doubles moindres carrés est aussi un estimateur GMM.
- Il est également possible de définir \mathbf{W} pour que l'estimateur GMM soit optimale (faible variance), noté $\hat{\boldsymbol{\theta}}_{\text{OGMM}}$. Dans ce cas, l'estimation se fait en deux étapes.
 - Etape 1 : On calcule $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ avec $\mathbf{W} = \mathbf{I}$.
 - Etape 2 : On calcule ensuite $\hat{\boldsymbol{\theta}}_{\text{OGMM}}$ avec $\mathbf{W} = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i'\hat{\boldsymbol{\theta}}_{\text{GMM}})\mathbf{z}_i\mathbf{z}_i'$.
- La variance de l'estimateur OGMM est robuste à l'hétéroscédasticité.
- Application avec R : script `iv-gmm.R`