Econométrie 2

Chapitre 1 : GMM et compléments sur les variables instrumentales

ENSAE 2021-2022

Michael Visser

CREST-ENSAE

Plan

Rappels sur les MCO

Rappels sur les variables instrumentales : l'estimateur des 2MC

Un cadre unificateur : la méthode des moments généralisée (GMM)

Condition de rang et instruments faibles

Notations

- $Y \in \mathbb{R}$ sera toujours la variable dépendante (ou expliquée). Elle pourra être continue (chapitres 1 et 2 notamment) ou non (en particulier les chapitres 3 et 4).
- $X \in \mathbb{R}^K$ désignera le vecteur des variables explicatives, incluant la constante : $X = (1, X_1, ..., X_{K-1})'$. On supposera toujours par la suite que la matrice $E(XX')$ (de dimension K par K) est de rang K. Autrement dit, les composantes de X sont linéairement indépendantes. Cette condition implique en particulier que $E(XX')$ est inversible, mais *n'exclut pas* que les composantes de X sont corrélées.
- Les composantes de X pourront être continues ou discrètes. Dans le 2ème cas, on inclura toutes les indicatrices des modalités correspondantes, sauf une.
- Ex : pour la variable d'activité prenant 3 valeurs (actif en emploi, actif au chômage, inactif) on inclura par ex. $\mathbb{1}_{\text{en emploi}}$ et $\mathbb{1}_{\text{chômeur}}$.

Notations

- ε capte l'effet des facteurs inobservables, i.e., le terme d'erreur.
- L'unité statistique (individus, ménages, entreprises...) est notée i ($i \in \{1, ..., n\}$). Si nécessaire le temps est désigné par t ($t \in \{1, ..., T\}$).
- Les variables $(Y_i, X_i)_{i=1,...,n}$ seront toujours supposées i.i.d., de même loi que (Y, X). L'indice i sera omis s'il n'est pas utile.
- Le vrai paramètre sera indicé par 0 (e.g. β_0) et les estimateurs seront chapeautés (e.g. $\widehat{\beta}$).

Identification et estimation

- Le modèle principalement considéré en Econométrie 1 est le modèle linéaire suivant :
$$Y = X'\beta_0 + \varepsilon.$$
- Comme X contient une constante on peut imposer $E(\varepsilon) = 0$ sans perte de généralité.
- Sous l'hypothèse d'exogénéité de X, $E(X\varepsilon) = 0$, on a
$$E(XY) = E(XX')\beta_0.$$
Puisque $E(XX')$ est inversible, β_0 vérifie
$$\boxed{\beta_0 = E(XX')^{-1}E(XY).}$$
- Cette relation montre que β_0 est *identifié*, i.e. défini de manière unique par la distribution des observables (X, Y) et le modèle.
- En pratique on ne connaît pas la loi de (X, Y). L'estimateur des MCO de β_0 est obtenu en remplaçant les espérances théoriques par des moyennes empiriques :
$$\widehat{\beta}_{MCO} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).$$

Résultats asymptotiques

- Notons que

$$\begin{aligned}\widehat{\beta}_{MCO} &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i(X_i'\beta_0+\varepsilon_i)\right)\\ &= \beta_0 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\varepsilon_i\right).\end{aligned}$$

- Par la loi des grands nombres, cet estimateur est convergent quand $n \to \infty$:

$$\widehat{\beta}_{MCO} \xrightarrow{P} \beta_0.$$

- Cet estimateur est par ailleurs asymptotiquement normal (voir Econométrie 1) :

$$\sqrt{n}\left(\widehat{\beta}_{MCO}-\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, E(XX')^{-1}E(\varepsilon^2 XX')E(XX')^{-1}\right).$$

Inférence

- On peut estimer en particulier la variance asymptotique V par

$$\widehat{V} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 X_i X_i'\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1},$$

avec $\widehat{\varepsilon}_i = Y_i - X_i'\widehat{\beta}_{MCO}$. C'est l'estimateur *robuste* (à l'hétéroscédasticité de ε) de la variance asymptotique.

- On estime alors l'écart-type (asymptotique) de $\widehat{\beta}_{MCO,j}$ par $\widehat{V}_{j,j}^{1/2}/\sqrt{n}$. En anglais : (asymptotic) standard error, noté $se(\widehat{\beta}_{MCO,j})$.
- Il est possible de faire de l'inférence statistique (tests, intervalles de confiance) à partir de cette distribution asymptotique.
- Par exemple, pour tester $H_0 : \beta_{0,j} = 0$ contre $H_1 : \beta_{0,j} \neq 0$, on utilise le fait que, sous H_0, $t_j \equiv \widehat{\beta}_{MCO,j}/se(\widehat{\beta}_{MCO,j})$ suit approximativement une loi normale de moyenne 0 et de variance 1. On rejette H_0 si $|t_j| > q_{1-\alpha/2}$ où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$.

- On a $\widehat{V} \xrightarrow{P} V$ sans hypothèse supplémentaire, contrairement à l'estimateur « standard » de la variance asymptotique :

$$\widetilde{V} = \left(\frac{1}{n-K}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1},$$

qui n'est convergent que sous l'hypothèse d'homoscédasticité $E(\varepsilon^2 XX') = E(\varepsilon^2)E(XX')$ (ou celle plus forte $E(\varepsilon^2|X) = \sigma^2$).

- En effet, sous l'hypothèse d'homoscédasticité, on a

$$\sqrt{n}\left(\widehat{\beta}_{MCO} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 E(XX')^{-1}\right)$$

et l'on obtient $\widetilde{V}$ en estimant σ^2 par le premier terme, et $E(XX')^{-1}$ par le deuxième terme.

Plan

L'hypothèse d'exogénéité

- Problème récurrent : le réalisme de l'hypothèse d'exogénéité $E(X\varepsilon) = 0$.
- Notons (voir Econométrie 1) qu'il est toujours possible de définir un $(\tilde{\beta}, \tilde{\varepsilon})$ tels que
$$Y = X'\tilde{\beta} + \tilde{\varepsilon},$$
avec $E(X\tilde{\varepsilon}) = 0$. En effet, si l'on considère $\tilde{\beta} = E(XX')^{-1}E(XY)$ et $\tilde{\varepsilon} = Y - X'\tilde{\beta}$, alors :
$$E\left[X\tilde{\varepsilon}\right] = E\left[X(Y - X'\tilde{\beta})\right] = E[XY] - E(XX')\tilde{\beta} = 0.$$
- L'estimateur des MCO estime toujours de manière convergente $\tilde{\beta}$ mais si l'hypothèse d'exogénéité n'est pas vérifiée on a $\tilde{\beta} \neq \beta_0$!
- $X'\tilde{\beta}$ est le meilleur prédicteur linéaire de Y par X (ou projection linéaire de Y sur X) et $\tilde{\beta}$ est le coefficient de projection : $\tilde{\beta} = argmin_b\ E[(Y - X'b)^2]$.
- β_{0j} est l'effet causal de X_j sur Y, i.e. l'effet de X_j sur Y si tous les autres facteurs (les autres variables dans X et les variables agrégées dans ε) restaient fixées.

Corrélation vs causalité

- En dehors des expériences contrôlés, les variables explicatives peuvent parfois être liées aux facteurs inobservables et $E(X\varepsilon) \neq 0$.
- Exemple : effet de la taille des classes sur la réussite des élèves.
- Si l'on régresse la réussite à des tests sur la taille de classe (sur données françaises), on obtient $\widehat{\beta}_{taille} > 0$.
- Explication : les élèves en difficulté sont souvent placés dans des classes de plus petite taille.
- La variable *taille* est corrélée au niveau initial, qui lui même affecte le résultat final. Donc $E(X\varepsilon) \neq 0$ et $\tilde{\beta} \neq \beta_0$.
- Projet STAR : allocation aléatoire des élèves dans des classes de taille 10 ou 20. Dans ce cas, par construction *taille* est non corrélée au niveau initial ou à d'autres facteurs influant le résultat final. $\Rightarrow E(X\varepsilon) = 0$ et $\tilde{\beta} = \beta_0$ (et d'ailleurs dans ce cas $\widehat{\beta}_{taille} < 0$!).

- En l'absence d'expérience contrôlée, ou si l'expérience n'est pas parfaite, on peut utiliser des variables *instrumentales* $Z \in \mathbb{R}^L (L \geq K)$, i.e. des variables qui ont une influence sur Y uniquement à travers X (pas d'effet direct de Z sur Y).
- Formellement on suppose :

  1. $E(ZX')$ est de rang K (condition de rang) ;
  2. $E(Z\varepsilon) = 0$ (hypothèse d'exogénéité).

- On a alors $ZY = ZX'\beta_0 + Z\varepsilon$ et donc

$$E(ZY) = E(ZX')\beta_0.$$

  Comme $E(ZX')$ est de rang K, il existe une matrice Γ de taille (K, L) et de rang K telle que $E(\Gamma ZX')$ est inversible. Alors

$$\beta_0 = E(\Gamma ZX')^{-1} E(\Gamma ZY).$$

- β_0 est donc encore identifié dans ce cas.

- Remarque 1 : en général, X et Z ont plusieurs composantes en commun : les régresseurs exogènes. Les régresseurs exclus de Z sont les régresseurs endogènes. Les composantes de Z exclus de X sont les instruments à proprement parler.
- Remarque 2 : si $Z = X$, on retrouve le cadre classique des MCO.
- Remarque 3 : lorsque $L > K$, β_0 est défini par plusieurs équations (correspondant à différents Γ). On dit alors qu'il est suridentifié. On a « trop » d'information, ce qui permet de tester certains aspects du modèle (voir ci-dessous).
- Si $L = K$, Γ ne joue pas de rôle et on obtient $\beta_0 = E(ZX')^{-1}E(ZY)$. Le modèle est dit juste identifié.

L'estimateur des doubles moindres carrés : définition

- Si l'on choisit $\Gamma = E(XZ')E(ZZ')^{-1}$ et en notant $X^* \equiv \Gamma Z$, on obtient
$$\beta_0 = E[X^* X']^{-1} E[X^* Y]$$
- X^* est la projection linéaire de X sur Z. Attention : il s'agit ici d'une projection d'un vecteur (X) sur un autre vecteur (Z).
- Dans le modèle $X = \Gamma Z + u$ (où u est un vecteur de dimension K) on a $E[Z(X - X^*)'] = 0$ (par construction, comme dans le cas d'une projection linéaire d'un scalaire sur un vecteur) et donc $E[X^*(X - X^*)'] = 0$. On obtient donc
$$\beta_0 = E[X^* X']^{-1} E[X^* Y] = E[X^* X^{*\prime}]^{-1} E[X^* Y].$$
- Dès lors, on peut estimer β_0 par
$$\boxed{\widehat{\beta}_{2MC} = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i \widehat{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{X}_i Y_i\right)}$$
où $\widehat{X}_i = \widehat{\Gamma} Z_i$ est le prédicteur linéaire *estimé* de X_i par Z_i ($\widehat{\Gamma}$ étant l'estimation des MCO que l'on obtient en régressant chaque élément de X sur Z).

- $\widehat{\beta}_{2MC}$ se nomme l'estimateur des doubles moindres carrés (2MC) car il s'obtient en 2 étapes :
  1. On régresse linéairement X sur Z. Soit $\widehat{X}$ le prédicteur de X ;
  2. On régresse linéairement Y sur $\widehat{X}$. Le coefficient correspondant est $\widehat{\beta}_{2MC}$.
- Remarque : dans la pratique, il suffit de régresser uniquement les variables endogènes sur Z, car la projection d'une variable exogène sur Z correspond à cette variable elle-même (car Z inclut toutes les variables exogènes). Seules les variables endogènes doivent ensuite être remplacées par leurs prédictuers lors de la 2ème étape.
- Exemple : on considère le modèle $Y = \beta_{01} + \beta_{02}X_1 + \beta_{03}X_2 + \beta_{04}X_3 + \varepsilon$, où X_1 et X_2 sont endogènes (corrélés avec ε), et X_3 exogène (non-corrélé avec ε). Il y a par ailleurs deux instruments à proprement parler, Z_1 et Z_2. Donc $Z = (1, Z_1, Z_2, X_3)'$. On est dans le cas $K = L = 4$.

- Exemple (suite) : dans la 1ère étape on régresse les variables endogènes sur toutes les variables exogènes, c'est à dire X_1 sur Z_1, Z_2, et X_3, puis X_2 sur Z_1, Z_2, et X_3. Pour tout i, on en déduit les prédicteurs

$$\widehat{X}_{1i} = \widehat{\alpha}_{01} + Z_{1i}\widehat{\alpha}_{02} + Z_{2i}\widehat{\alpha}_{03} + X_{3i}\widehat{\alpha}_{04}$$

et

$$\widehat{X}_{2i} = \widehat{\delta}_{01} + Z_{1i}\widehat{\delta}_{02} + Z_{2i}\widehat{\delta}_{03} + X_{3i}\widehat{\delta}_{04}$$

- Dans la 2ème étape on régresse Y sur $\widehat{X}_{1i}$, $\widehat{X}_{2i}$ et X_3.

L'estimateur des doubles moindres carrés : asymptotique

- Lorsqu'on remplace $\widehat{X}_i$ par $\widehat{\Gamma} Z_i$ et $\widehat{\Gamma}$ par $\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}$, on obtient

$$\begin{aligned}\widehat{\beta}_{2MC} &= \left\{\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)\right\}^{-1} \\ &\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i Y_i\right).\end{aligned}$$

- Le raisonnement est le même que pour les MCO :

$$\begin{aligned}\widehat{\beta}_{2MC} &= \beta_0 + \left\{\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)\right\}^{-1} \\ &\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i \varepsilon_i\right).\end{aligned}$$

- Par la loi des grands nombres, $\widehat{\beta}_{2MC} \xrightarrow{P} \beta_0$. Par ailleurs, l'estimateur est asymptotiquement normal (voir Econométrie 1) :

$$\sqrt{n}\left(\widehat{\beta}_{2MC} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, E(X^* X^{*\prime})^{-1} E(\varepsilon^2 X^* X^{*\prime}) E(X^* X^{*\prime})^{-1}\right).$$

- Etudions de plus près le cas unidimensionnel ($Y = \beta_{01} + X_1\beta_{02} + \varepsilon$), homoscédastique ($E(\varepsilon^2|Z) = \sigma^2$), et avec un instrument à proprement parler (Z_1). On a $X_1^* = \Gamma_0 + Z_1\Gamma_1$ avec, par défintion, $\Gamma_1 = Cov(X_1, Z_1)/V(Z_1)$. Comme Γ_0 et Γ_1 sont non-aléatoires, on a $V(X_1^*) = \Gamma_1^2 V(Z_1)$.
- La variance asymptotique de $\sqrt{n}\widehat{\beta}_{2MC,2}$ ($\widehat{\beta}_{2MC,2}$ étant l'estimateur de β_{02}), s'écrit alors :

$$V = \frac{\sigma^2}{V(X_1^*)} = \frac{\sigma^2}{\Gamma_1^2 V(Z_1)} = \frac{\sigma^2}{\sigma_{X_1}^2 \rho_{X_1,Z_1}^2},$$

où $\sigma_{X_1}^2$ est la variance de X_1 et ρ_{X_1,Z_1} le coefficient de corrélation entre X_1 et Z_1. La variance asymptotique est d'autant plus faible que cette corrélation est plus forte, i.e. que X_1 est bien expliqué par Z_1.

- Dans le cas général, on peut estimer de manière convergente la variance asymptotique par (avec $\widehat{\varepsilon}_i = Y_i - X_i'\widehat{\beta}_{2MC}$)

$$\widehat{V} = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\widehat{X}_i\widehat{X}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i'\right)^{-1}.$$

- Comme dans le cas exogène, cet estimateur est dit robuste car il ne repose pas sur des hypothèses d'homoscédasticité.
- Notons que ce n'est pas l'estimateur qu'on obtient si l'on fait une régression de Y sur $\widehat{X}$ car $\widehat{\varepsilon}_i \neq Y_i - \widehat{X}_i'\widehat{\beta}_{2MC}$.

Propriétés de l'estimateur des doubles moindres carrés

- L'estimateur est très simple à calculer, convergent et asymptotiquement normal sous des conditions faibles.
- Contrairement à l'estimateur des MCO, il n'est en général pas sans biais, même sous l'hypothèse $E(\varepsilon|Z) = 0$.
- Contrairement à l'estimateur des MCO, on ne peut pas obtenir sa loi exacte même si $\varepsilon|Z \sim N(0, \sigma^2)$. L'inférence repose sur sa loi asymptotique (ou le bootstrap).
- Est-il optimal asymptotiquement ? Cf. la suite.

Plan

Principe général

- Dans le cas où X est exogène, on a $E[XY] = E[XX']\beta_0$ soit encore $E[X(Y - X'\beta_0)] = 0$. Dans le cas instrumental (avec $K = L$) on a utilisé $E[Z(Y - X'\beta_0)] = 0$.
- Idée plus générale : soit U_i l'ensemble des données sur i (par ex. (X_i, Y_i, Z_i)), on cherche à estimer $\theta_0 \in \mathbb{R}^K$ en utilisant

$$\boxed{E\left[g(U, \theta_0)\right] = 0}$$

  où g est une fonction à valeurs dans $\mathbb{R}^L$ avec $L \geq K$.
- Si $L = K$ on résout simplement $\frac{1}{n}\sum_{i=1}^n g(U_i, \widehat{\theta}) = 0$. Il y a dans ce cas autant d'équations que d'inconnus, et a priori une solution unique existe : $\widehat{\theta}$ est l'estimateur des moments.
- Exemple (tiré de Crépon et Jacquemet) : supposons que $U = X$ avec X une variable scalaire. Pour tout i, on a $E(X_i) = \mu_0$ et $V(X_i) = \sigma_0^2$, donc $\theta_0 = (\mu_0, \sigma_0^2)'$. On définit $g_1(X_i, \theta) = X_i - \mu$ et $g_2(X_i, \theta) = (X_i - \mu)^2 - \sigma^2$.

Principe général

- L'estimateur des moments est la valeur de $\widehat{\theta}$ vérifiant

$$\begin{pmatrix} \frac{1}{n}\sum_i X_i - \widehat{\mu} \\ \frac{1}{n}\sum_i (X_i - \widehat{\mu})^2 - \widehat{\sigma}^2 \end{pmatrix} = 0$$

  et donc $\widehat{\mu} = \overline{X}$ et $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2$.

- Lorsque $L > K$, les fluctuations d'échantillonnage impliquent qu'il n'existe pas de valeur unique $\widehat{\theta}$ qui annule la contrepartie empirique de toutes les L conditions de moments (sauf si l'échantillon est de taille infinie).

- Exemple (suite) : supposons maintenant que $\sigma_0^2 = 1$. On a alors toujours deux conditions de moments mais qui portent sur un paramètre unique cette-fois ci (μ_0).

Principe général

- La contrepartie empirique des deux conditions de moments est alors

$$\begin{pmatrix} \frac{1}{n}\sum_i X_i - \widehat{\mu} \\ \frac{1}{n}\sum_i (X_i - \widehat{\mu})^2 - 1 \end{pmatrix} = 0$$

  et en principe il n'existe pas de solution unique pour $\widehat{\mu}$.

- Lorsque $L > K$, l'idée est alors de rendre le vecteur $\frac{1}{n}\sum_{i=1}^n g(U_i, \widehat{\theta})$ « petit ». Plus précisement, l'estimateur des moments généralisé vérifie :

$$\boxed{\widehat{\theta}_{GMM} = \arg\min_{\theta} \left(\frac{1}{n}\sum_{i=1}^n g(U_i,\theta)\right)' W_n \left(\frac{1}{n}\sum_{i=1}^n g(U_i,\theta)\right)}$$

  où W_n est symétrique et définie positive, potentiellement aléatoire.

Pourquoi les GMM ?

- Beaucoup de modèles économiques imposent des conditions de moments.
- GMM utilisée dans d'autres chapitres (e.g. chapitre 2).
- Contrairement au maximum de vraisemblance (MV), les GMM ne s'appuient pas sur une loi paramétrique.
- L'estimateur du MV est efficace mais pas forcément convergent en cas d'erreur de spécification.
- Permet de construire des estimateurs optimaux en appliquant un principe général.
- Permet de construire des tests de suridentification (test du modèle) en appliquant un principe général.

Résultats asymptotiques

- L'estimateur des GMM est la solution de la condition du 1er ordre :

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial \theta}(U_i,\widehat{\theta}_{GMM})\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n} g(U_i,\widehat{\theta}_{GMM})\right) = 0.$$

- L'estimateur des GMM sera convergent et asymptotiquement normal, principalement dès que l'hypothèse d'identification est vérifiée :

$$E\left[g(U,\theta)\right] = 0 \Rightarrow \theta = \theta_0,$$

  i.e., θ_0 est la seule valeur du paramètre à annuler l'ensemble des moments.

- Pour déterminer la variance asymptotique, on admet que $G_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g}{\partial \theta}(U_i,\widehat{\theta}_{GMM}) \xrightarrow{P} G = E\left[\frac{\partial g}{\partial \theta}(U,\theta_0)\right]$ et $W_n \xrightarrow{P} W$, où W est symétrique et non aléatoire.

Résultats asymptotiques et optimalité

- Sous plusieurs conditions de régularité on montre alors (preuve hors cours) que

$$\sqrt{n}\left(\widehat{\theta}_{GMM} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G'WG\right)^{-1}\left(G'WHWG\right)\left(G'WG\right)^{-1}\right),$$

avec

$$\begin{aligned} H &\equiv V\left(g(U,\theta_0)\right) = E\left(\left[g(U,\theta_0) - E(g(U,\theta_0)\right]\left[g(U,\theta_0) - E(g(U,\theta_0)\right]'\right) \\ &= E\left((g(U,\theta_0))(g(U,\theta_0))'\right). \end{aligned}$$

- On peut montrer aussi que (pour tout $W \neq H^{-1}$) :

$$\left(G'WG\right)^{-1}\left(G'WHWG\right)\left(G'WG\right)^{-1} >> \left(G'H^{-1}G\right)^{-1},$$

où $A >> 0$ signifie que la matrice carrée A est positive semi définie.

- Par conséquent, le choix $W_o = H^{-1} = V\left(g(U,\theta_0)\right)^{-1}$ est optimal : il conduit à la matrice de variance asymptotique la plus faible.

- Mais W_o est en général inconnue car fonction de θ_0.

L'estimateur en deux étapes

- On commence par obtenir un estimateur des GMM convergent $\widehat{\theta}^{(1)}_{GMM}$, en choisissant typiquement $W_n = I_L$ (matrice d'identité de taille L).
- On calcule ensuite $\widehat{W_o} = V\left(g(U, \widehat{\theta}^{(1)}_{GMM})\right)^{-1}$. On en déduit $\widehat{\theta}^{(2)}_{GMM}$.
- Comme $\widehat{\theta}^{(1)}_{GMM}$ est convergent, on a, dès que $\theta \mapsto V\left(g(U, \theta)\right)$ est continue, $\widehat{W_o} \xrightarrow{P} W_o$.
- L'estimateur $\widehat{\theta}^{(2)}_{GMM}$ est donc asymptotiquement optimal :

$$\sqrt{n}\left(\widehat{\theta}^{(2)}_{GMM} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G'H^{-1}G\right)^{-1}\right).$$

Retour aux modèles linéaires

- Dans le cas où X est exogène, $g(U,\beta) = X(Y - X'\beta)$, $L = K$ et tous les choix de W_n conduisent au même estimateur, celui des MCO.
- Dans le cas instrumental, $g(U,\beta) = Z(Y - X'\beta)$, et si $L > K$, différents choix de W_n conduisent à différents estimateurs.
- L'estimateur des GMM est alors défini comme (cas spécial de la dernière formule à la page 24) :

$$\widehat{\beta}_{GMM} = \arg\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i (Y_i - X_i'\beta) \right)' W_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i (Y_i - X_i'\beta) \right)$$

- On peut montrer que dans ce cas :

$$\widehat{\beta}_{GMM} = \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} X_i Z_i' \right) W_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right) \right\}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Z_i' \right) W_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i Y_i \right)$$

Retour aux modèles linéaires

- La matrice de pondération optimale s'écrit

$$\begin{aligned} W_o &= V(Z(Y - X'\beta_0))^{-1} = V(Z\varepsilon)^{-1} \\ &= E\left((Z\varepsilon - E(Z\varepsilon))(Z\varepsilon - E(Z\varepsilon))'\right)^{-1} \\ &= E\left((Z\varepsilon)(Z\varepsilon)'\right)^{-1} = E\left(\varepsilon^2 ZZ'\right)^{-1} \end{aligned}$$

- Si les résidus sont homoscédastiques, $W_o = E(\varepsilon^2)^{-1} E\left(ZZ'\right)^{-1}$.
- Soit $\widehat{W_o} = \widehat{E}(\varepsilon^2)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}$ un estimateur convergent de W_o.
- En remplaçant, dans l'expression de $\widehat{\beta}_{GMM}$ (page 29), W_n par $\widehat{W_o}$, on constate 2 choses : i) l'estimateur des GMM ne dépend pas de $\widehat{E}(\varepsilon^2)^{-1}$ (seulement de $\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1}$), et peut donc être obtenu en une étape ; ii) l'estimateur des GMM coincide avec l'estimateur des 2MC (première expression page 17).
- L'estimateur des 2MC est alors optimal (car l'estimateur des GMM est optimal).

Retour aux modèles linéaires

- Sinon, l'estimateur optimal se calcule en deux étapes.
- On calcule d'abord l'estimateur des 2MC standard. Cet estimateur est convergent mais pas efficace.
- On en déduit $\widehat{\varepsilon}_i = Y_i - X_i'\widehat{\beta}_{2MC}$, puis $\widehat{W_o} = \left(\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2 Z_i Z_i'\right)^{-1}$.
- L'estimateur optimal vérifie enfin (dernière formule de la page 29 avec W_n remplacé par $\widehat{W_o}$) :

$$\widehat{\beta}^{(2)}_{GMM} = \left\{\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\widehat{W_o}\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)\right\}^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i'\right)\widehat{W_o}\left(\frac{1}{n}\sum_{i=1}^n Z_i Y_i\right)$$

Tests de modèle

- Supposons que l'on ait $L > K$ conditions de moments $E[g_l(U, \theta_0)] = 0,\ l = 1, ..., L)$.

$\Rightarrow$ On dispose de plus de conditions que nécessaire pour identifier θ_0.

- Intuitivement, ces conditions (prises dans leur ensemble) sont testables.
- Idée du test : si ces conditions sont vraies $\frac{1}{n}\sum_{i=1}^{n} g(U_i, \widehat{\theta})$ doit être « petit ».

$\Rightarrow$ on rejette le modèle si la norme de ce vecteur est « grande ».

- Formellement, on montre que sous l'hypothèse $E[g(U, \theta_0)] = 0$,

$$T = n\left(\frac{1}{n}\sum_{i=1}^{n} g(U_i, \widehat{\theta}_{GMM}^{(2)})\right)' \widehat{W_o}\left(\frac{1}{n}\sum_{i=1}^{n} g(U_i, \widehat{\theta}_{GMM}^{(2)})\right) \xrightarrow{d} \chi^2_{L-K}.$$

Il suffit donc de comparer le membre de gauche avec le quantile d'un χ^2_{L-K} (par exemple le 0,95-quantile si on fait un test de niveau 5%), et de rejetter l'hypothèse nulle si la valeur de T dépasse le quantile.

- Dans le cas du modèle linéaire instrumental avec $L > K$, le test revient à vérifier si les résidus sont bien orthogonaux aux instruments. La statistique du test devient alors :

$$T = n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \widehat{\varepsilon}_i \right)' \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 Z_i Z_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i \widehat{\varepsilon}_i \right)$$

- Dans le cas homoscédastique, on peut obtenir un test très simplement en régressant $\widehat{\varepsilon}_i$ sur Z_i (test de Sargan). Le R^2 de la régression vérifie alors, sous l'hypothèse nulle,

$$nR^2 \xrightarrow{d} \chi^2_{L-K}.$$

Plan

La condition de rang

- β_0 est identifié ssi $E(ZX')$ est de rang K.
- Cette hypothèse (ou plutôt son contraire) est toujours testable, contrairement à $E(Z\varepsilon) = 0$ qui ne l'est que si $L > K$.
- Dans le cas d'un seul régresseur endogène X_1 et d'un seul instrument Z_1 ($X = (1, X_1, ..., X_{K-1})$, $Z = (1, Z_1, X_2, ..., X_{K-1})$), considérons la régression
$$X_1 = \alpha_{01} + Z_1\alpha_{02} + X_2\alpha_{03} + ... + X_{K-1}\alpha_{0K} + \eta.$$
On peut montrer que dans ce cas la condition de rang est rejetée ssi $\alpha_{02} = 0 \Rightarrow$ test de Student.
- Toujours avec un seul régresseur endogène mais plusieurs instruments $Z_1, ..., Z_{L-K+1}$:
$$X_1 = \alpha_{01} + Z_1\alpha_{02} + ... + Z_{L-K+1}\alpha_{0L-K+2} + X_2\alpha_{0L-K+3} + ... + X_{K-1}\alpha_{0L} + \eta.$$
La condition de rang est rejetée dans ce cas ssi $\alpha_{02} = ... = \alpha_{0L-K+2} = 0 \Rightarrow$ test de Fisher.
- Dans le cas général (plusieurs variables endogènes et plusieurs instruments), tester la condition de rang est plus compliqué (hors cours).

Violation de la condition de rang

- Que se passe-t-il si $\text{rg}(E(ZX')) < K$? Supposons pour simplifier $K = L = 2$. On a alors, pour le coefficient β_{02} de X_1 :

$$\widehat{\beta}_{2MC,2} = \beta_{02} + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)X_{1i}}.$$

- On a, par le théorème central limite,

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)\varepsilon_i - Cov(Z_1, \varepsilon) \\ \frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)X_{1i} - \text{Cov}(Z_1, X_1) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Comme $Cov(Z_1, \varepsilon) = 0$ (Z_1 étant exogène), et sous l'hypothèse $\text{Cov}(Z_1, X_1) =$, on a alors

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \overline{Z}_1)X_{1i} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Violation de la condition de rang

- D'après le théorème de mappage continu (voir Wooldridge, chapitre 3, lemme 3.6), on a

$$\widehat{\beta}_{2MC,2} - \beta_{02} \xrightarrow{d} \frac{U_1}{U_2},$$

où $(U_1, U_2) \sim \mathcal{N}(0, \Sigma)$.

- On sait alors que $E(U_1|U_2) = (\Sigma_{12}/\Sigma_{22})U_2$, et donc $U_1 = (\Sigma_{12}/\Sigma_{22})U_2 + U_3$, avec $U_3 \perp\!\!\!\perp U_2$.
- Donc on a

$$\widehat{\beta}_{2MC,2} - \beta_{02} \xrightarrow{d} \frac{U_1}{U_2} = \frac{\Sigma_{12}}{\Sigma_{22}} + \frac{U_3}{U_2},$$

- $\widehat{\beta}_{2MC} - \beta_{02}$ converge donc en loi vers $\frac{U_1}{U_2}$, une variable aléatoire qui suit une distribution de Cauchy (dont la moyenne n'existe pas). Par conséquence, $\widehat{\beta}_{2MC,2}$ n'est pas un estimateur convergent de β_{02}.

Les instruments faibles

- Que se passe-t-il si $\text{Cov}(X_1, Z_1) \neq 0$ mais petit ? Question d'importance en pratique (cf. exemple ci-dessous).
- L'estimateur est convergent et asymptotiquement normal. Mais, en pratique, l'approximation asymptotique peut être très mauvaise, même pour n très grand.
- Si les corrélations entre le régresseur X_1 et les instruments sont faibles (dans un certain sens) on peut alors montrer que (i) cet estimateur est biaisé en direction des MCO et (ii) le biais empire lorsqu'on augmente le nombre d'instruments faibles.
- Dans le cas d'une variable endogène (et L grand par rapport à K), on a par ailleurs l'approximation suivante :

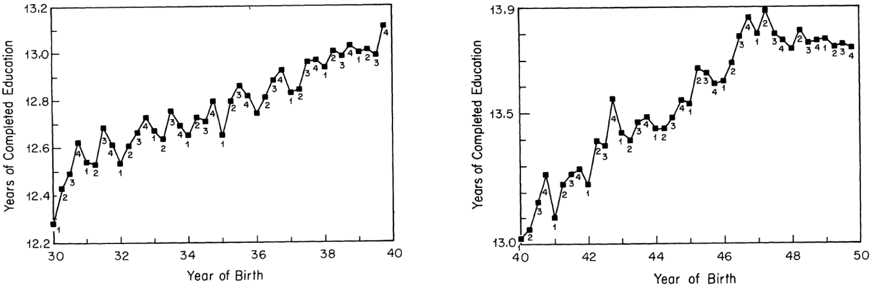
$$E(\widehat{\beta}_{2MC} - \beta_0) \simeq \frac{E(\widehat{\beta}_{MCO} - \beta_0)}{F+1},$$

  où F est la statistique de Fisher de la régression de 1ère étape (de l'hypothèse nulle $H_0 : \alpha_{02} = ... = \alpha_{0L-K+2} = 0$).
- Règle du pouce : l'estimateur des variables instrumentales est crédible seulement si $F \geq 10$! Dans la pratique, il est indispensable de reporter F.

Rendements de l'éducation et trimestre de naissance

- Exemple : mesure des rendements de l'éducation aux US en utilisant le trimestre de naissance comme instrument (Angrist et Krueger, 1991).
- Idée : l'école est obligatoire jusqu'à 16 ans et dans la plupart des états, les enfants doivent commencer l'école l'année de leur 6 ans.
- Donc les enfants nés à la fin de l'année sont susceptibles de suivre un peu plus longtemps l'école.
- Si l'on suppose que le trimestre de naissance n'affecte pas directement les salaires, il peut donc être utilisé comme instrument du niveau d'éducation.
- Angrist et Krueger (1991) utilisent pour ce faire un échantillon à 5% du recensement de 1980, soit $n = 329\,509$.

Rendements de l'éducation et trimestre de naissance



Nombre d'années d'éducation en fonction de l'année et du trimestre de naissance (tiré d'Angrist et Krueger, 1991)

- Résultats d'Angrist et Krueger :

Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings (standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.029)
F (excluded instruments)		13.486		4.747		1.613
Partial R^2 (excluded instruments, $\times$100)		.012		.043		.014
F (overidentification)		.932		.775		.725
Age Control Variables						
Age, Age^2	x	x			x	x
9 Year of birth dummies			x	x	x	x
Excluded Instruments						
Quarter of birth		x		x		x
Quarter of birth $\times$ year of birth				x		x
Number of excluded instruments		3		30		28

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. F (first stage) and partial R^2 are for the instruments in the first stage of IV estimation. F (overidentification) is that suggested by Basmann (1960).

Figure 1 – Résultats d'Angrist et Krueger (tiré de Bound, Jaeger et Baker, 1995)

- Qu'obtiendrait-on avec des instruments simulés, indépendants de l'éducation ?

Table 3. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Using Simulated Quarter of Birth (500 replications)

Table (column)	*1 (4)*	*1 (6)*	*2 (2)*	*2 (4)*
		Estimated Coefficient		
Mean	.062	.061	.060	.060
Standard deviation of mean	.038	.039	.015	.015
5th percentile	−.001	−.002	.034	.035
Median	.061	.061	.060	.060
95th percentile	.119	.127	.083	.082
		Estimated Standard Error		
Mean	.037	.039	.015	.015

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509.

Figure 2 – Estimateur des 2MC avec instruments simulés (tiré de Bound, Jaeger et Baker, 1995)

- Conclusion : une grande taille d'échantillon ne suffit pas ! Une bonne régression de 1ère étape est indispensable.

L'essentiel

- L'estimateur des MCO : définition et propriétés asymptotiques sous l'hypothèse d'exogénéité.
- Le problème d'endogénéité.
- L'estimateur des 2MC : définition et propriétés asymptotique sous les deux hypothèses des variables instrumentales.
- Estimateur des GMM : définition, normalité asymptotique, estimateur optimal en deux étapes.
- Instruments faibles : définition, comment les détecter.