Chapitre 4 : Non respect des hypothèses du modèle de régression multiple

Partie 4-3 : Biais de l'estimateur des MCO.

A. Fadhuile (adelaide.fadhuile@univ-grenoble-alpes.fr)
Univ Grenoble Alpes

Année 2023-2024

Plan du cours

1 – Introduction

- Quelle hypothèse n'est pas respectée ?
  - H_1 : $E[u] = 0, \forall\ t$ l'espérance mathématique de l'erreur est nulle
  - H_2 : $V[u] = E(uu^T) = \sigma_u^2 I_T$ la variance de l'erreur est constante
  - H_3 : la matrice **X** est non-aléatoire **H_3 non respectée : donc la matrice X aléatoire/mesurée avec erreur/inclut de la simulatnéité**$\rightarrow$ X n'est plus exogène
  - H_4 : le modéle est correctement spécifié.
  - H_5 : la matrice **X** est de plein rang : $k + 1 < T$
- A distance finie, l'estimateur des MCO, sous les hypothèses H_1 à H_5, est:
  - **Sans biais** (i.e. $E(\hat{b}) = b$) et **à Variance minimale**

  $\Rightarrow$ **BLUE** (Best Linear Unbiased Estimator).

  $\Rightarrow$ Estimateur efficace
  - **Ni sans biais NI à Variance minimale**

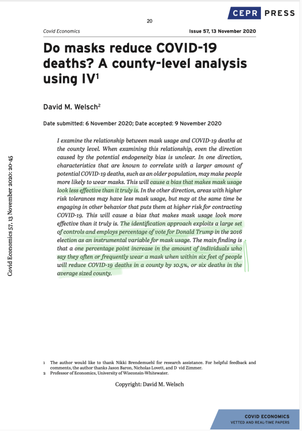
Plan du cours

2 – Introduction

2.1 – Intuitions

2.1.1 – Exemple port du masque et Covid

- Exemple et commentaires sur l'article suivant : D. Welsch (2020) Do Masks Reduce COVID-19 Deaths? A County Level Analysis using IV, Covid Economics, Issue 57, 13 November 2020, p20-45.
- vous pouvez le télécharger ici dans sa version complète Lien et sur Moodle (Lien ici) pour la version que j'ai annoté et présenté en cours.
- Forme d'endogénéité liée à une erreur de mesure sur une des variable explicatives.


CEPR PRESS
20
Covid Economics Issue 57, 13 November 2020

Do masks reduce COVID-19 deaths? A county-level analysis using IV[1]

David M. Welsch[2]

Date submitted: 6 November 2020; Date accepted: 9 November 2020

I examine the relationship between mask usage and COVID-19 deaths at the county level. When examining this relationship, even the direction caused by the potential endogeneity bias is unclear. In one direction, characteristics that are known to correlate with a larger amount of potential COVID-19 deaths, such as an older population, may make people more likely to wear masks. This will cause a bias that makes mask usage look less effective than it truly is. In the other direction, areas with higher risk tolerances may have less mask usage, but may at the same time be engaging in other behavior that puts them at higher risk for contracting COVID-19. This will cause a bias that makes mask usage look more effective than it truly is. The identification approach exploits a large set of controls and employs percentage of vote for Donald Trump in the 2016 election as an instrumental variable for mask usage. The main finding is that a one percentage point increase in the amount of individuals who say they often or frequently wear a mask when within six feet of people will reduce COVID-19 deaths in a county by 10.5%, or six deaths in the average sized county.

Covid Economics 57, 13 November 2020: 20-45

1 The author would like to thank Nikki Brendemuehl for research assistance. For helpful feedback and comments, the author thanks Jason Baron, Nicholas Lovett, and D vid Zimmer.
2 Professor of Economics, University of Wisconsin-Whitewater.

COVID ECONOMICS
VETTED AND REAL-TIME PAPERS

2 – Introduction

2.1 – Intuitions

2.1.2 – Equilibre de marché

- Exemple un modèle d'équilibre de marché : modèle est constitué de 3 équations :
  - une équation d'offre
  - une équation de demande
  - une équation d'équilibre
- Questions économiques :
  - $D(p)$: quel serait le niveau de la demande avec des prix au niveau p?
  - $S(p)$: quel serait le niveau de l'offre avec des prix au niveau p?

  $\Rightarrow$ Les **quantités** observées et les **prix** observés sont déterminés **simultanément** à l'équilibre
- **Endogénéité** car les variables sont déterminées simultanément : i.e. plusieurs variables explicatives sont le résultat d'un comportement de coordination.

2 – Introduction

2.1 – Intuitions

2.1.2 – Equilibre de marché

- Considérons le système suivant:

$$\begin{cases} D(p) : Q = \alpha_1 P + \alpha_2 A + e_d \\ S(p) : Q = \beta_1 P + e_s \end{cases}$$

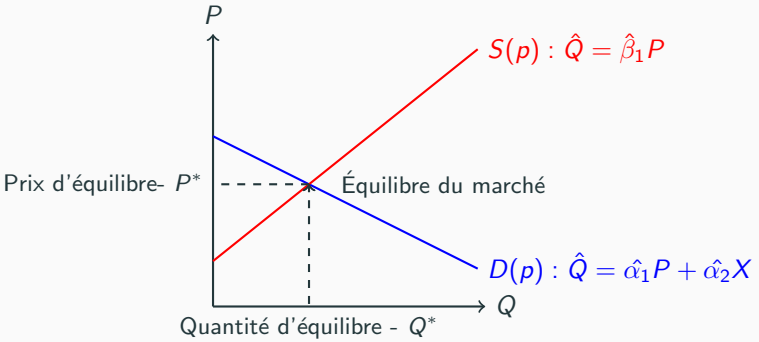
- A est le revenu, et P les prix
- Que dit la théorie économique sur le sens des paramètres estimés ?
  - α_1 ?
  - α_2 ?
  - β_1 ?
- Pour simplifier, dans un premier temps nous allons considérer un modèle simplifié

2 – Introduction

2.1 – Intuitions

2.1.2 – Equilibre de marché

$\Rightarrow$ Prix et quantités à l'équilibre du marché



2 – Introduction

2.1 – Intuitions

2.1.2 – Equilibre de marché

- Considérons le système suivant:

$$\begin{cases} D(p) : Q = \alpha_1 P + \alpha_2 A + e_d \\ S(p) : Q = \beta_1 P + e_s \end{cases}$$

- X est le revenu, et P les prix
- Que dit la théorie économique sur le sens des relations ?
  - Dans l'équation de demande : β_1?? et α_1??
  - Dans l'équation d'offre : α_2??
- Les termes d'erreur sont dans les 2 éq° !!! Supposons que

$$E(e_d) = 0,\ V(e_d) = \sigma^2_{ud},,\ E(e_s) = 0,,\ V(e_s) = \sigma^2_{us}\ cov(e_d, e_s) = 0$$

- e_d et e_s affectent simultanément D et S $\Rightarrow$ corrélation

2 – Introduction

2.1 – Intuitions

2.1.2 – Equilibre de marché

- Considérons le système suivant:

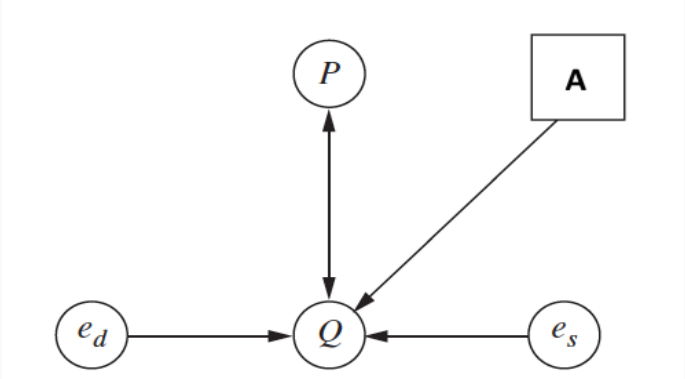
$$\begin{cases} D(p) : Q = \alpha_1 P + \alpha_2 A + e_d \\ S(p) : Q = \beta_1 P + e_s \end{cases}$$

- X est le revenu, et P les prix
- Que dit la théorie économique sur le sens des relations ?
  - Dans l'équation de demande : $\beta_1 > 0$ et $\alpha_1 < 0$
  - Dans l'équation d'offre : $\alpha_2 > 0 \Rightarrow$ Mais la seule variable exogène du système est A
- Les termes d'erreur sont dans les 2 éq° !!! Supposons que

$$E(e_d) = 0,\ V(e_d) = \sigma^2_{ud},,\ E(e_s) = 0,,\ V(e_s) = \sigma^2_{us}\ cov(e_d, e_s) = 0$$

- e_d et e_s affectent simultanément D et S $\Rightarrow$ corrélation

2.1.2 – Equilibre de marché



2 – Introduction

2.1 – Intuitions

2.1.3 – Demande de travail des firmes

- On dispose d'une base de données contenant 4 variables l_i, q_i, c_i et s_i où
  - l_i est l'effectif total de la firme,
  - $q_i$la valeur ajoutée déflatée au coût des facteurs,
  - c_i, le cout relatif travail/capital **supposé mesuré avec erreur** et
  - s_i**, le taux de salaire réel : supposé être un instrument de** c_i
- On considère l'équation de demande de travail :

$$\log l_i = b_0 + b_1 \log q_i + b_2 \log c_i + u_i \qquad i = 1, ..., 92$$

2 – Introduction

2.2 – Hypothèses

- Quelle hypothèse n'est pas respectée ?
  - H_1 : $E[u] = 0, \forall\ t$ l'espérance mathématique de l'erreur est nulle
  - H_2 : $V[u] = E(uu^T) = \sigma_u^2 I_T$ la variance de l'erreur est constante
  - H_3 : la matrice **X** est non-aléatoire **H_3 non respectée : donc la matrice X aléatoire/mesurée avec erreur/inclut de la simulatnéité**$\rightarrow$ X n'est plus exogène
  - H_4 : le modéle est correctement spécifié
  - H_5 : la matrice **X** est de plein rang : $k + 1 < T$
- A distance finie, l'estimateur des MCO, sous les hypothèses H_1 à H_5, est:
  - **Sans biais** (i.e. $E(\hat{b}) = b$) et **à Variance minimale**

  $\Rightarrow$ **BLUE** (Best Linear Unbiased Estimator).

  $\Rightarrow$ Estimateur efficace
  - **Ni sans biais NI à Variance minimale**

2 – Introduction

2.3 – Définition

- Qu'est-ce que l'exogénétité ?
  - Une variable est considérée comme exogène si on peut la considérer comme une donnée sans perte d'information.
  - Dépend du contexte et de la question posée !!!
- Le problème de l'endogénéité est courant en sciences sociales/économie.
  - Le problème se pose lorsque des variables individuelles importantes ne peuvent pas être observées.
  - Ceux-ci sont souvent corrélés avec les informations explicatives observées.
- l'erreur de mesure peut également conduire à l'endogénéité.

2 – Introduction

2.4 – Objectif du chapitre

- Remise en cause de $H_3 \Rightarrow$ corrélation entre une variable explicative et le terme d'erreur (notamment en cas d'erreur de mesure).
- Dans ce cas : L'estimateur des MCO est **biaisé** et **non convergent**.

1 Les instruments doivent :
   - être corrélées avec la variable instrumentée,
   - Ne pas être corrélées au terme d'erreur,
   - Ne doit pas apparaître dans l'équation estimée

2 Méthodes par variables instrumentales
   - Doubles moindres carrés (DMC, 2SLS)
   - Méthodes des moments généralisés (GMM) [Pas vu ds ce cours]

3 Tests
   - Propriétés de l'estimateur
   - Validité des instruments

Plan du cours

3 – Erreur de mesure sur une variable explicative

3.1 – Définition en économétrie

- Considérons pour simplifier le modèle de regression linéaire simple :

$$y_t = b_0 + b_1 x_t + u_t \tag{1}$$

- Une variable explicative est **strictement exogène** si:

$$E[x_{jt} u_{t'}] = 0 \forall t, t'$$

- Une variable explicative est **exogène** si:

$$E[x_{jt} u_{t'}] = 0 \forall t$$

- Une variable explicative est **prédeterminée** si:

$$E[x_{jt} u_{t'}] = 0 \forall t' \geq t$$

3 – Erreur de mesure sur une variable explicative

3.1 – Définition en économétrie

- Dans le cadre standard, on a :

$$E(u_t) = 0 \forall t, V(u_t) = \sigma_u^2 \forall t, \text{et } cov(x_t, u_t) = 0 \forall t \neq t'$$

- Supposons que x_t soit mesuré avec erreur, on a:

$$\underbrace{x_t}_{\text{Valeur observée}} = \underbrace{\widetilde{x_t}}_{\text{Vraie valeur}} + \underbrace{\nu_t}_{\text{Erreur de mesure}}$$

- Cela implique que :

$$E[\nu_t] = 0 \forall t, \quad V[\nu_t] = \sigma_\nu^2 \forall t$$

$$cov[\nu_t, \nu_t'] = 0 \forall t \neq t', \quad cov[\nu_t, u_t'] = 0 \forall t, t'$$

3 – Erreur de mesure sur une variable explicative

3.1 – Définition en économétrie

- Pour estimer l'équation, il faut considérer le vrai modèle, et donc

$$y_t = b_1\widetilde{x}_t + b_0 + u_t \tag{2}$$
$$= b_1(x_t - \nu_t) + b_0 + u_t \tag{3}$$
$$= b_1 x_t + b_0 + w_t, \text{ avec } w_t = u_t - b_1\nu_t \tag{4}$$

$\Rightarrow$ L'estimateur des MCO de b_1 est biaisé et non convergent.

- par extension, si x_t est aléatoire et corrélée avec ν_t, l'estimateur de b_1 sera forcément biaisé :

$$E[x_t w_t] = E[(\widetilde{x}_t + \nu_t)w_t] = E[(\widetilde{x}_t + \nu_t)(u_t - b_1\nu_t)] = -b_1\sigma_\nu^2 \neq 0$$

Plan du cours

4 – Estimations

4.1 – Meth 1 : Variables instrumentales (IV/IV)

- L'estimateur par Variables instrumentales (noté VI) est l'application des MCO au modèle transformé suivant:

$$P_Z y = P_Z X b + P_Z u, \text{ avec } P_Z = Z\left(Z^\intercal Z\right)^{-1} Z^\intercal \tag{5}$$

  - $P_Z X$ revient à ne conserver que l'information apportée par les variables explicatives qui est asymptotiquement orthogonale aux perturbations.
  - Définir un nombre d'instruments $p > k + 1$ supérieur ou égal à $k + 1$.
  - L'instrument doit être corrélé avec la variable endogène mais non corrélé avec le terme d'erreur.
  - **Plus les instruments sont corrélés avec les variables explicatives, plus l'estimateur par VI est précis.**
  - A distance finie, il est généralement biaisée, et il n'existe pas de formulation générale pour sa matrice des variances covariances.

4 – Estimations

4.1 – Meth 1 : Variables instrumentales (IV/IV)

- L'estimateur par variables instrumentales est donné par

$$\hat{b}_{VI} = (X^\intercal P_Z X)^{-1} X^\intercal P_Z y \qquad P_Z = Z\,(Z^\intercal Z)^{-1} Z^\intercal$$

avec Z la matrice des p instruments tels que :

$$\operatorname*{plim}_{T\to\infty} \frac{X^\intercal P_Z X}{T} \Rightarrow \text{Z et X sont corrélés asymptotiquement}$$

existe et est définie positive.

$$\operatorname*{plim}_{T\to\infty} \frac{Z^\intercal u}{T} = 0 \Rightarrow \text{Z et u sont } \textbf{non} \text{ corrélés asymptotiquement}$$

4 – Estimations

4.1 – Meth 1 : Variables instrumentales (IV/IV)

- Propriétés asymptotiques
  - Convergent $\underset{T\to\infty}{\text{plim}}\, \hat{b}_{VI} = b$
  - Asymptotiquement normal

$$\sqrt{T}\left(\hat{b}_{VI} - b\right) \underset{T\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma_u^2 Q_{XP_ZX}^{-1}\right)$$

  - de variance

$$V\left[\sqrt{T}\left(\hat{b}_{VI} - b\right)\right] = \sigma_u^2 Q_{XP_ZX}^{-1}$$

$$\hat{V}\left[\hat{b}_{VI}\right] = \hat{\sigma}_{u,VI}^2 \left(\frac{X^\intercal P_Z X}{T}\right)^{-1}$$

  - avec

$$\hat{\sigma}_{u,VI}^2 = \frac{SCR_{VI}}{T} \qquad SCR_{VI} = \overline{u}^\intercal \overline{u} \text{ où } \overline{u} = y - X\hat{b}_{VI}$$

4 – Estimations

4.1 – Meth 1 : Variables instrumentales (IV/IV)

- Si $T \to \infty$
  - $\underset{T\to\infty}{\text{plim}}\widehat{\sigma}^2_{u,VI} = \sigma^2_u$, i.e. $\sigma^2_{u,VI}$ est un estimateur convergent de la variance des perturbations σ^2_u
  - $\underset{T\to\infty}{\text{plim}}\widehat{V}[\hat{b}_{VI}] = V[\sqrt{T}(\hat{b}_{VI} - b)]$, i.e. $\widehat{V}[\hat{b}_{VI}]$ est un estimateur convergent de la matrice des variances covariances asymptotique de $\sqrt{T}(\widehat{b}_{VI} - b)$
- Si l'estimateur par VI est convergent, il n'est pas asymptotiquement efficace, i.e. il n'est pas l'estimateur convergent utilisant les instruments Z, dont la matrice des var-cov est minimale.
- Comme l'estimateur par VI n'a pas de bonnes propriétés à distance finie, les tests de restriction sont des test de *Wald* basés sur les propriétés asymptotiques de l'estimateur par VI.

4 – Estimations

4.1 – Meth 1 : Variables instrumentales (IV/IV)

- Remarque : si les perturbations sont autocorrélées et/ou hétéroscédastiques, l'estimateur par VI reste convergent. Mais sa distribution asymptotique est donnée par

$$V[\sqrt{T}(\hat{b}_{VI} - b)] \underset{T\to\infty}{\overset{\mathcal{L}}{\to}}$$

$$N\left\{0, \sigma_u^2 \operatorname*{plim}_{T\to\infty} \left(\frac{X^T P_Z X}{T}\right)^{-1} \frac{X^T P_Z \Omega P_Z X}{T} \left(\frac{X^T P_Z X}{T}\right)^{-1}\right\}$$

- Rappel: l'estimateur par VI revient à appliquer les MCO au modèle 6. Mais, la mat des var-cov est désormais égale à :

$$V[P_Z u] = \sigma_u^2 P_Z \Omega P_Z$$

- Cela conduit à estimer le modèle:

$$\Omega^{-1/2} y = \Omega^{-1/2} X b + \Omega^{-1/2} u \qquad (6)$$

Plan du cours

5 – Application 1

- Analyse de la réussite universitaire

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel}_i + u_i$$

- $E[u|X] \neq 0$ à cause d'un biais d'omission
- Instrument proposé : Z = Colocataire (désigné au hasard) joue a des jeux vidéo
  - Pertinence ?
  - Exogénéité ?

5 – Application 1

- Analyse de la réussite universitaire

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel}_i + u_i$$

- Pertinence ?
  - Nombre d'heures de travail personnel est-il corrélé aux nombres d'heures de jeux vidéo (Z_i), ie

$$Corr(Z, X) \neq 0$$

  - "First Step"

$$(\text{Heures de travail personnel})_i = \gamma_0 + \gamma_1 Z_i + \nu_i$$

  - Avec $\widehat{\gamma_1} = -0.668^{**}$ (significatif au seuil de 5%)

5 – Application 1

- Analyse de la réussite universitaire

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel})_i + u_i$$

- Exogène ?
  - la corrélation entre Z et Y "passe" par X, ie

$$Corr(Z, u) = 0$$

  - Cette condition est valable si les colocataires jouant des jeux vidéo (Z) n'affecte la moyenne générale de GPA que par le biais du nombre d'heures passées à étudier (X).
  - Existe-t-il d'autres canaux par lesquels Z est lié à GPA ?
    - Et si le fait d'être assigné à un colocataire disposant de jeux vidéo faisait dormir les étudiants moins longtemps ?
    - Et si les étudiants masculins étaient plus susceptibles d'apporter des jeux vidéo et qu'ils avaient en moyenne de moins bonnes notes ? (à titre d'exemple)

5 – Application 1

- Analyse de la réussite universitaire

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel})_i + u_i$$

- Existe-t-il d'autres canaux par lesquels Z est lié à GPA ?
  - Et si le fait d'être assigné à un colocataire disposant de jeux vidéo faisait dormir les étudiants moins longtemps ?
  - Et si les étudiants masculins étaient plus susceptibles d'apporter des jeux vidéo et qu'ils avaient en moyenne de moins bonnes notes ? (à titre d'exemple)
- Il va être nécessaire de "contrôler" pour les caractéristiques observables pour "réduire le biais d'omission"

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel})_i + \beta_2(\text{heure de sommeil})_i + \beta_3(\text{genre})_i + u_i$$

5 – Application 1

- Analyse de la réussite universitaire

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel})_i + u_i$$

- Existe-t-il d'autres canaux par lesquels Z est lié à GPA ?
  - Et si le fait d'être assigné à un colocataire disposant de jeux vidéo faisait dormir les étudiants moins longtemps ?
  - Et si les étudiants masculins étaient plus susceptibles d'apporter des jeux vidéo et qu'ils avaient en moyenne de moins bonnes notes ? (à titre d'exemple)
- Il va être nécessaire de "contrôler" pour les caractéristiques observables pour "réduire le biais d'omission"

$$GPA_i = \beta_0 + \beta_1(\text{Heures de travail personnel})_i + \beta_2(\text{heure de sommeil})_i + \beta_3(\text{genre})_i + u_i$$

5 – Application 1

- Combien d'instruments ?
- Un seul instrument suffira, mais une meilleure prédiction de la variable endogène avec plus d'un instrument : on dit que le coefficient est sur-identifié
- Dans le cas d'un seul instrument et d'une seule variable endogène : **identification exacte**.
- mais avec une deuxième variable endogène, alors un instrument n'est pas suffisant car le coefficient à estimer est **sous-identifié**.

Plan du cours

6 – Application 2

- On dispose d'une base de données contenant 4 variables l_i, q_i, c_i et s_i où
  - l_i **est l'effectif total de la firme,**
  - q_i**la valeur ajoutée déflatée au coût des facteurs,**
  - c_i**, le cout relatif travail/capital (supposé mesuré avec erreur) et**
  - s_i**, le taux de salaire réel (Supposé être l'instrument de** c_i**)**
- On considère l'équation de demande de travail :

$$\log l_i = b_0 + b_1 \log q_i + b_2 \log c_i + u_i \qquad i = 1, ..., 92$$

6 – Application 2

1. On régresse par les MCO la variable endogène sur les instruments (régression auxiliaire)
   - $\log c_i = \eta_0 + \eta_1 \log q_i + \eta_2 \log s_i + \upsilon_i \quad (1)$
     Rq: Il faut η_2 significativement différent de zéro pour que l'instrument ne soit pas faible.
     $\Rightarrow$ On calcule les valeurs prédites de la variable endogène ($\widehat{\log c_i}$) $\Rightarrow$ Le R^2 est un indicateur de la qualité des instruments
2. On estime par les MCO le modèle initial en remplaçant la variable endogène par les valeurs prédites avec (1):
   - $\log l_i = b_0 + b_1 \log q_i + b_2 \widehat{\log c_i} + u_i$

- Attention: pour réaliser une inférence statistique correcte, les "bons". résidus doivent être utilisés:
  - $\widehat{u_i} = \log l_i - \widehat{b_0} - \widehat{b_1} \log q_i - \widehat{b_2} \log c_i$
  - L'utilisation des "commandes automatiques" des logiciels réalise la correction.

6 – Application 2

Figure 1: Estimations du modèle par les MCO

Variable dépendante: ll

	coefficient	éc. type	t de Student	p. critique	
const	0,227112	0,195764	1,160	0,2491	
lq	0,916531	0,0315131	29,08	3,14e−47	***
lc	−0,127663	0,0269566	−4,736	8,19e−06	***

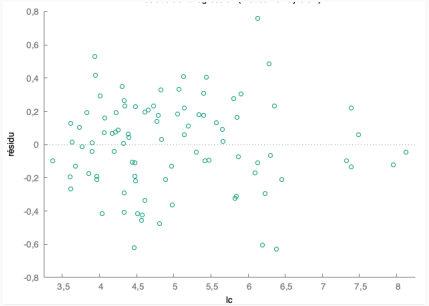
Moyenne var. dép.	3,636557	Éc. type var. dép.	0,883837
Somme carrés résidus	6,605093	Éc. type régression	0,272423
R2	0,907083	R2 ajusté	0,904995
F(2, 89)	434,4243	P. critique (F)	1,20e−46
Log de vraisemblance	−9,380758	Critère d'Akaike	24,76152
Critère de Schwarz	32,32688	Hannan−Quinn	27,81496

6 – Application 2

- S'il y a une erreur de mesure, on observe une corrélation entre le terme d'erreur et la variable explicative concernée (ici le coût relatif du travail).
- Ici, il semble qu'il y ait une relation croissante entre c_i et les perturbations. Il est donc possible que c_i soit endogène.
- La relation est parfois perceptible sur le graphique des résidus.
- Test d'exogénéité des variables explicatives

Coefficients de corrélation, utilisant les observations 1 – 92
valeur critique à 5% (bilatérale) = 0,2050 pour n = 92

ll	lq	lc	ls	uhat_ols	
1,0000	0,9400	−0,1548	−0,1763	0,3048	ll
	1,0000	−0,0019	0,1150	−0,0000	lq
		1,0000	0,4063	−0,0000	lc
			1,0000	−0,7291	ls
				1,0000	uhat_ols



6 – Application 2

Figure 2: Estimations du modèle par les MCO et les 2SLS

Variable dépendante: ll

	coefficient	éc. type	t de Student	p. critique	
const	0,227112	0,195764	1,160	0,2491	
lq	0,916531	0,0315131	29,08	3,14e−47	***
lc	−0,127663	0,0269566	−4,736	8,19e−06	***

Moyenne var. dép.	3,636557	Éc. type var. dép.	0,883837
Somme carrés résidus	6,605093	Éc. type régression	0,272423
R2	0,907083	R2 ajusté	0,904995
F(2, 89)	434,4243	P. critique (F)	1,20e−46
Log de vraisemblance	−9,380758	Critère d'Akaike	24,76152
Critère de Schwarz	32,32688	Hannan−Quinn	27,81496

Modèle 3: 2MC, utilisant les observations 1−92
Variable dépendante: ll
Instrumenté: lc
Instruments: const lq ls

	coefficient	éc. type	t de Student	p. critique	
const	2,50736	0,736401	3,405	0,0010	***
lq	0,915501	0,0647134	14,15	1,71e−24	***
lc	−0,583785	0,135263	−4,316	4,11e−05	***

Moyenne var. dép.	3,636557	Éc. type var. dép.	0,883837
Somme carrés résidus	27,85324	Éc. type régression	0,559426
R2	0,714749	R2 ajusté	0,708339
F(2, 89)	109,6732	P. critique (F)	9,67e−25
Log de vraisemblance	−327,8385	Critère d'Akaike	661,6770

6 – Application 2

Figure 3: First stage

Variable dépendante: lc

	coefficient	éc. type	t de Student	p. critique	
const	−4,16318	2,22447	−1,872	0,0646	*
lq	−0,0576565	0,113820	−0,5066	0,6137	
ls	1,83793	0,434357	4,231	5,63e−05	***

Moyenne var. dép.	4,989245	Éc. type var. dép.	1,059399
Somme carrés résidus	85,02628	Éc. type régression	0,977421
R2	0,167485	R2 ajusté	0,148776
F(2, 89)	8,952463	P. critique (F)	0,000287
Log de vraisemblance	−126,9162	Critère d'Akaike	259,8325
Critère de Schwarz	267,3979	Hannan−Quinn	262,8859

- R^2 de cette première étape : indicateur de la qualité de l'ajustement
- Test à réaliser pour la qualité de l'instrument

Plan du cours

7 – Tests

- Vérifier l'exogenéité d'une variable explicative $\Rightarrow$ **Test d'Hausman**
- Vérifier la validité du choix des instruments $\Rightarrow$ **Test Sargan**, s'il y a plus d'un instrument, ou F-Test (test de contraintes comme précédemment)

7 – Tests

7.1 – Test Hausman

- Vérifier l'exogenéité des variables explicatives : $\Rightarrow$ **Test d'Hausman**
  - un estimateur convergent et asymptotiquement efficace sous H_0 (non corrélation), mais non convergent sous H_1
  - à un estimateur convergent mais non efficace sous H_0 et H_1
- **Hypothèses :**
  - $H_0 : \underset{T\to\infty}{plim} \frac{X^T u}{T} = 0$ et $H_1 : \underset{T\to\infty}{plim} \frac{X^T u}{T} \neq 0$
- **Statistique de Test:**

$$Q_H = (\widehat{b_{IV}} - \widehat{b_{MCO}})^T [\widehat{V}(\hat{b}_{IV}) - \widehat{V}(\hat{b}_{MCO})]^{-1} (\widehat{b_{IV}} - \widehat{b_{MCO}}) \sim \chi^2(k+1)$$

- **Règle de décision:**
  - Si $Q_H > Q_{th}$, on rejette $H_0 \Rightarrow$ rejet corrélation entre les var expl et le terme d'erreur
- Application
  - $Q_H = 168.74 \cdots \cdots Q_{th}(\cdots) = \cdots \Rightarrow$

Figure 4: Estimations du modèle par les MCO et les 2SLS

```
Modèle 3: 2MC, utilisant les observations 1-92
Variable dépendante: ll
Instrumenté: lc
Instruments: const lq ls

             coefficient   éc. type    t de Student   p. critique
  ------------------------------------------------------------------
  const        2,50736     0,736401        3,405       0,0010     ***
  lq           0,915501    0,0647134      14,15        1,71e-24   ***
  lc          -0,583785    0,135263       -4,316       4,11e-05   ***

Moyenne var. dép.     3,636557   Éc. type var. dép.    0,883837
Somme carrés résidus  27,85324   Éc. type régression   0,559426
R2                    0,714749   R2 ajusté             0,708339
F(2, 89)              109,6732   P. critique (F)       9,67e-25
Log de vraisemblance -327,8385   Critère d'Akaike      661,6770
Critère de Schwarz    669,2423   Hannan-Quinn          664,7304

Test de Hausman -
  Hypothèse nulle: Les estimateurs MCO sont non biaisés
  Statistique du test asymptotique: Khi-deux(1) = 168,745
  avec p. critique = 1,39074e-38

Test d'instruments faibles -
  Statistique F, première étape (1, 89) = 17,9045
  Valeurs critiques de la taille maximale désirée des 2MC,
pour des tests au seuil de confiance de 5%:

   taille      10%      15%      20%      25%
   valeur    16,38     8,96     6,66     5,53

  La taille maximale est probablement de moins de 10%
```

7 – Tests

- "First stage"

$$\log c_i = \eta_0 + \eta_1 \log q_i + \eta_2 \log s_i + \upsilon_i$$

  - $H_0 : \eta_1 = \eta_2 = 0$
  - Ici : $F = 17.90 > 16.38$: le biais de l'estimateur des DMC/2SLS est inférieur à 10% par rapport aux MCO.
  - Cela signifie que nous avons moins de 10% de chances d'accepter à tort une variable comme significative en utilisant les règles standard.
- Concrétement
  - Si $\hat{F} > 10$: Z est un instrument "fort"
  - Si $\hat{F} \leqslant 10$: Z est un instrument "faible"

Figure 5: Estimation par les 2SLS/DMC

```
Modèle 3: 2MC, utilisant les observations 1-92
Variable dépendante: ll
Instrumenté: lc
Instruments: const lq ls

             coefficient   éc. type    t de Student   p. critique
  ------------------------------------------------------------------
  const        2,50736     0,736401        3,405       0,0010     ***
  lq           0,915501    0,0647134       14,15       1,71e-24   ***
  lc          -0,583785    0,135263        -4,316      4,11e-05   ***

Moyenne var. dép.      3,636557   Éc. type var. dép.     0,883837
Somme carrés résidus   27,85324   Éc. type régression    0,559426
R2                     0,714749   R2 ajusté              0,708339
F(2, 89)               109,6732   P. critique (F)        9,67e-25
Log de vraisemblance  -327,8385   Critère d'Akaike       661,6770
Critère de Schwarz     669,2423   Hannan-Quinn           664,7304

Test de Hausman -
  Hypothèse nulle: Les estimateurs MCO sont non biaisés
  Statistique du test asymptotique: Khi-deux(1) = 168,745
  avec p. critique = 1,39074e-38

Test d'instruments faibles -
  Statistique F, première étape (1, 89) = 17,9045
  Valeurs critiques de la taille maximale désirée des 2MC,
pour des tests au seuil de confiance de 5%:

   taille      10%      15%      20%      25%
   valeur    16,38     8,96     6,66     5,53

  La taille maximale est probablement de moins de 10%
```

7 – Tests

7.2 – Test de Sargan : Validité du choix des instruments

- **Hypothèses :**
  - $H_0 : \underset{T\to\infty}{plim}\ \frac{Z^T u}{T} = 0$ contre H_1: $\underset{T\to\infty}{plim}\ \frac{Z^T u}{T} \neq 0$
- **Statistique de Test:**
  - $Q_S = \frac{\bar{u}^T Z (Z^T Z)^{-1} Z^T \bar{u}}{\hat{\sigma}^2_{u,VI}} = \mathbf{TR^2} \sim \chi^2(\mathbf{p} - (\mathbf{k}+\mathbf{1}))$
    - où le R^2 est obtenu par une régression auxiliaire des résidus des MCO, sur toutes les variables explicatives supposées exogènes.
    - T est le nombre d'observations, p est le nombre d'instruments et k le nombre de variables explicatives, i.e. $p-(k+1)$ est le nombre de restrictions sur-identifiantes.
- **Règle de décision**
  - Si $Q_S > \chi^2(p-(k+1))$ on rejette H_0 contre $H_1 \rightarrow$ les variables instrumentales ne sont pas de bons instruments

Plan du cours

8 – Conclusion

- Remise en cause de $H_3 \Rightarrow$ corrélation entre une variable explicative et le terme d'erreur (notamment en cas d'erreur de mesure).
- Dans ce cas : L'estimateur des MCO est **biaisé** et **non convergent**.

1 Les instruments doivent :
   - être corrélées avec la variable instrumentée,
   - Ne pas être corrélées au terme d'erreur,
   - Ne doit pas apparaître dans l'équation estimée

2 Méthodes par variables instrumentales
   - Doubles moindres carrés (DMC, 2SLS)

3 Tests
   - Propriétés de l'estimateur : IV vs OLS $\rightarrow$ Test Hausman
   - Validité des instruments

8 – Conclusion

- Quelle hypothèse n'est pas respectée ?
  - H_1 : $E[u] = 0, \forall\ t$ l'espérance mathématique de l'erreur est nulle
  - H_2 : $V[u] = E(uu^T) = \sigma_u^2 I_T$ la variance de l'erreur est constante
  - H_3 : la matrice **X** est non-aléatoire **H_3 non respectée : donc la matrice X aléatoire/mesurée avec erreur/inclut de la simulatnéité**$\rightarrow$ X n'est plus exogène
  - H_4 : le modéle est correctement spécifié
  - H_5 : la matrice **X** est de plein rang : $k + 1 < T$
- A distance finie, l'estimateur des MCO, sous les hypothèses H_1 à H_5, est:
  - **Sans biais** (i.e. $E(\hat{b}) = b$) et **à Variance minimale**

  $\Rightarrow$ **BLUE** (Best Linear Unbiased Estimator).

  $\Rightarrow$ Estimateur efficace
  - **Ni sans biais NI à Variance minimale**