Variables instrumentales

Olivier Godechot

« Causes and effect, I know,
Our little hates and blames,
We are born and grow,
As the seeds we sow,
And right and wrong are – names
Cause and effect, I know»
Philip Green Wright, « Revulsion », The dial of heart, 1905.
(Inventor of instrumental variables)

Invention des variables instrumentales

- Wright, P.G. (1928), The Tariff on Animal and Vegetable Oils, New York: The Macmillan Company.
- Economiste, poète, mathématicien, enseignant en sociologie aussi…
- Appendice B
- Estimation des courbes d'offres et de demandes
- Rôle de son fils biologiste qui l'aurait aidé (ou écrit l'appendice)

1. Le problème: l'endogénéîté

Les limites des MCO

- Modèle linéaire notation

  $y_i=\beta_0+\beta_1.x_{1i}+\beta_2.x_{2i}+\ldots+\beta_k.x_{ki}+u_i$

  où i représente l'individu i

  ou

  $y=\beta_0+\beta_1.x_1+\beta_2.x_2+\ldots+\beta_k.x_k+u$

  ou

  $y = X.\ \beta + u$

- Méthode
  - Moindre carré des erreurs
  - On estime les paramètres $\beta_0,\beta_1, \beta_2 \ldots \beta_k$ de telle sorte que $\Sigma_i u_i^2$ soit minimale

Les 6 hypothèses des MCO

- Linéarité
- Matrice plein rang et absence d'autocorrélation des variables explicatives
- Homéoscédasticité
- Absence d'autocorrélation des résidus
- Normalité des résidus
- Absence de corrélation entre les variables explicatives et le résidu *dans le modèle théorique*.

Les problèmes d'endogénéité

- On parle d'endogénéité au sens large quand la dernière hypothèse est violée.
- Ces problèmes peuvent conduire à se tromper dans l'interprétation des paramètres.
- Les variables instrumentales peuvent constituer une technique de correction.

Trois problèmes et leurs effets sur les paramètres

- Cas 1 : on mesure très mal la variable explicative
  - On sous-estime la valeur absolue du paramètre
- Cas 2 : il nous manque une variable explicative importante corrélée positivement à la variable dépendante et cette variable explicative est corrélée fortement et positivement (resp. négativement) à l'une de nos variables explicatives
  - On surestime (resp. sous-estime) la valeur absolue du paramètre
- Cas 3 : la variable explicative dépend par ailleurs positivement de la variable expliquée
  - Effet plus complexe. Pas d'intuition évidente

Indépendance des variables explicatives et des résidus : $Cov(x_i,u)=0$

- Il se peut que dans notre *vrai* modèle $Cov(x_i,v)\neq 0$.
  - $y_i = a_{vrai} + b_{vrai}.\ x_i + v_i$ (1)
- Dans une estimation empirique par MCO, par construction, $Cov(x_i,u)=0$.
  - $y_i = a_{est} + b_{est}.\ x_i + u_i$ (2)
- Si c'est le cas, alors les paramètres estimés empiriquement ne sont pas ceux que l'on cherche à connaître.
  - $E(a_{est}) \neq a_{vrai}$
  - $E(b_{est}) \neq b_{vrai}$
- L'estimation empirique n'est pas « fausse » en soi. Mais il est faux d'interpréter les paramètres de l'estimation empirique comme ceux du vrai modèle !

Erreur de mesure sur une variable explicative

- Notre vrai modèle est le suivant

  $y = a_{vrai} + b_{vrai}.\ x_{vrai} + u$ avec $cov(x_{vrai}, u) = 0$ (1)

- Or on mesure mal x_{vrai}

  $x_{prox} = x_{vrai} + e$ avec $cov(x_{prox}, e) \neq 0$ (2)

- Notre vrai modèle modifié sera le suivant

  $y = a_{vrai} + b_{vrai}.\ x_{prox} + u'$ avec $u' = u - b_{vrai}.\ e$ (3)

- Or dans ce vrai modèle modifié, la variable approchée est corrélée au nouveau résidu u'

  $cov(x_{prox}, u') = cov(x_{prox}, u) + cov(x_{prox}, -b_{vrai}.\ e)$

  $= -b_{vrai}.\ cov(x_{prox}, e) = -b_{vrai}.\ cov(x_{vrai} + e, e) \neq 0$

- Par conséquent, notre modèle empirique ne permet pas de connaître les paramètres a_{vrai} et b_{vrai}

  $y = a_{est} + b_{est}.\ x_{prox} + v$ (4)

- Le modèle 4 n'est pas faux en soi, mais il est faux de l'interpréter comme si on estimait 3.

Nature du biais en cas d'erreur de mesure

- On peut calculer la relation entre le paramètre estimé par les MCO et le vrai paramètre

$$b_{est} \xrightarrow{P} \frac{b_{vrai}}{1 + \frac{V(e)}{V(x_{vrai})}}$$

- On sous-estime le paramètre
- Le paramètre estimé est d'autant plus mauvais que l'erreur de mesure est importante

Variable omise ou hétérogénéité inobservée

- Notre <u>vrai modèle</u> est le suivant

  $y = a_{vrai} + b_{vrai}.\ x + c_{vrai}.\ z + u$ avec $cov(x,\ u) = 0$ et $cov(z,\ u) = 0$ (1)

- Or on n'observe pas la variable z et $cov(x,z) \neq 0$
- Notre <u>vrai modèle modifié</u> sera le suivant

  $y = a_{vrai} + b_{vrai}.\ x + u'$ où $u' = u + c_{vrai}.\ z$ (3)

- Or dans ce vrai modèle modifié, la variable x est corrélée au nouveau résidu u'

$$cov(x,u') = cov(x,\ u + c_{vrai}.\ z\ ) = cov(x,\ u) + cov(x,\ c_{vrai}.\ z)$$
$$= -c.\ cov(x,\ z) \neq 0$$

- Par conséquent, notre modèle empirique ne permet pas de connaître les paramètres a_{vrai} et b_{vrai}

  $y = a_{est} + b_{est}.\ x + v$ (4)

- Le modèle 4 n'est pas faux en soi, mais il est faux de l'interpréter comme si on estimait 3.

Nature du biais en cas d'hétérogénéité inobservée

- On peut calculer la relation entre le paramètre estimé par les MCO et le vrai paramètre

$$b_{est} \xrightarrow{P} b_{vrai} + c\frac{cov(x,z)}{V(x)}$$

- Le paramètre estimé s'éloigne d'autant plus du vrai paramètre
  - que les variables x et z sont corrélées (effet de cov(x,z))
  - que l'impact de z sur y est important (effet du paramètre c)

Simultanéité

Cas classiques : les prix et les quantités sur un marché ; aspirations scolaires et niveau scolaire

- Notre vrai modèle est le suivant

$y = ax + b + u$ (1)

$x = cy + d + v$ (2)

- Or dans un tel modèle $cov(x,u) \neq 0$ et $cov(y,v) \neq 0$. En remplaçant y dans 2 par 1:

$x = cax + cb + cu + d + v$

$x = (cb + cu + d + v)/(1 - ca)$

$cov(x, u) = cov((cb + cu + d + v)/(1 - ca), u) = (c/(1 - ca)).cov(u, u)$

- Par conséquent, notre modèle empirique ne permet pas de connaître les paramètres a et b

$y = a_{est} + b_{est}.\ x + w$ (3)

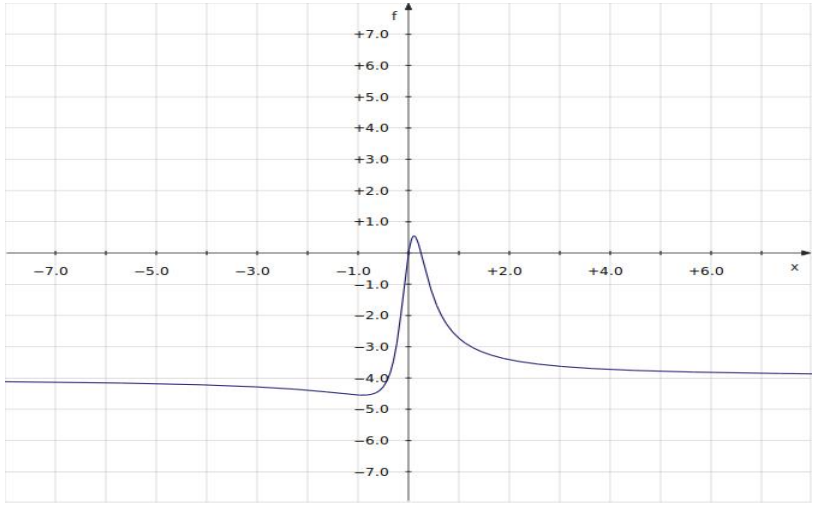
- Le modèle 3 n'est pas faux en soi, mais il est faux de l'interpréter comme si on estimait 1.

Nature du biais en cas de simultanéité

- On peut calculer la relation entre le paramètre estimé par les MCO et le vrai paramètre

$$a_{est} \xrightarrow{P} \frac{a.V(v)+c.V(u)}{V(v)+c^2.V(u)}$$

- L'effet est plus complexe. Cf la variation du biais en fonction de *v* :

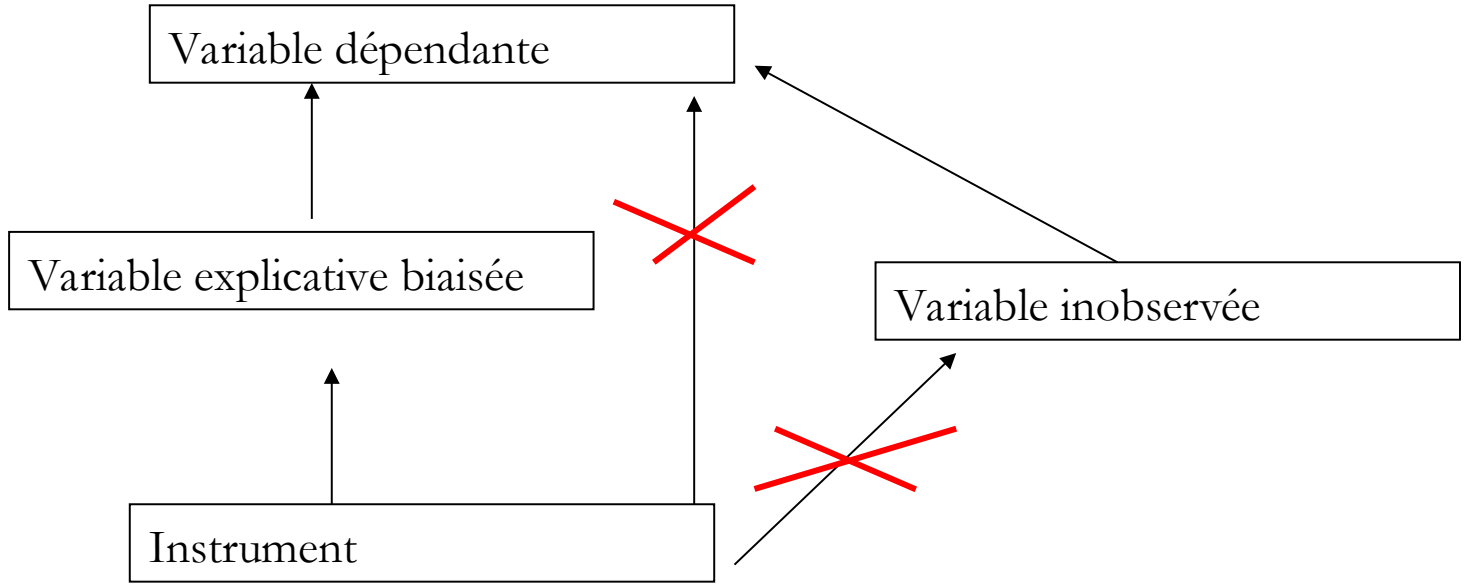


2. La solution: l'instrument

Traitement de l'endogénéité avec la technique des variables instrumentales

- Deux problèmes d'endogénéité : simultanéité et hétérogénité inobservée
- Goux et Maurin
  - $Etude_i = Etude\_voisin_k + Exogènes + Omises + u$
  - $Etude_i = Etude\_voisin_k(Etude_i) + Exogènes + Omises + + u$
  - Avec $cov(Etude\_voisin_k, Omises) \neq 0$
  - Endogénéité du modèle. Les MCO peut-être biaisés.
- Solution : variables instrumentales
  - Trouver une variable instrumentale exogène qui impacte ma *variable expliquée* uniquement par l'intermédiaire de son effet sur *la variable explicative* suspecte d'endogénéité.
  - Idée : Goux et Maurin.
    - Le mois de naissance impacte le niveau scolaire des voisins.
    - Je ne choisis pas mes voisins en fonction du mois de naissance.
    - CQ : La variation du mois de naissance des voisins impacte mon niveau scolaire uniquement du fait de son impact sur leur niveau scolaire.

Schéma des variables instrumentales



La méthode des variables instrumentales

- On a un problème d'endogénéité sur la variable x_{endo} (pour l'une des trois raisons signalées ci-dessus) dans notre vrai modèle.

$$y = a_{vrai} + b_{vrai} x_{endo} + c_{vrai} x_2 + u$$

$$cov(x_{endo}, u) \neq 0$$

- Par conséquent l'estimateur empirique des MCO ne permet pas de connaître les vrais paramètres, b_{vrai}, mais aussi a_{vrai} et c_{vrai}

$$E(b_{est.mco}) \neq b_{vrai} \quad ; \quad E(a_{est.mco}) \neq a_{vrai} \quad ; \quad E(c_{est.mco}) \neq c_{vrai}$$

- On peut corriger le problème si dispose d'un instrument z_{inst} tel que :

$$cov(z_{inst}, x_{endo}) \neq 0$$

$$cov(z_{inst}, u) = 0$$

Une procédure en deux étapes

Vrai modèle : $y = a_{vrai} + b_{vrai}x_{endo} + c_{vrai}.x_2 + u$ avec $cov(x_{endo},u) \neq 0$

- **Première étape :** on régresse la variable endogène à la fois sur l'instrument et sur les autres variables explicatives. NB : On met toutes les variables explicatives même non pertinentes en première étape.

$\boldsymbol{x_{endo}} = a_0 + a_1.z_{inst} + a_2.x_2 + u_{prem}$

- On récupère de cette première régression x'_{endo}, la prédiction de la variable endogène x_{endo} :

$x'_{endo} = a_0 + a_1.z_{inst} + a_2.x_2 = x_{endo} - u_{prem}$

- **Deuxième étape :** on introduit cette prédiction dans la régression à la place de x_{endo}

$y = a_{est} + \boldsymbol{b_{est}x'_{endo}} + c_{est}.x_2 + u_{deux}$

- comme z_{inst} et x_2 ne sont pas corrélés avec le résidu u, alors x'_{endo} n'ést plus corrélé avec u, l'estimateur des variables instrumentales permet d'estimer sans biais b_{vrai} (et aussi a_{vrai} et c_{vrai}).

$E(b_{est}) = b_{vrai}$; $E(a_{est}) = a_{vrai}$; $E(c_{est}) = c_{vrai}$

Exemple : La correction des erreurs de mesure

- Notre <u>vrai modèle modifié</u> était le suivant

  $y = a_{vrai} + b_{vrai}.\ x_{prox} + u'$ avec $u' = u - b_{vrai}.\ e$ (1)
- On estime en première étape avec notre instrument z_{inst}

  $x_{prox} = dz_{inst} + v$ avec $cov(z_{inst}, v) = 0$ (2)
- Deuxième étape : on introduit en alors dans la régression empirique x'_{prox} la prédiction de x_{prox}

  $y = a_{vrai} + b_{vrai}.\ x'_{prox} + w$ (3)
- Or si dans le vrai modèle modifié (1), on remplace la variable x_{prox} par sa prédiction instrumentée x'_{prox}, cette dernière n'est plus corrélée au nouveau résidu $(u - b_{vrai}.\ e + b_{vrai}.\ v)$

  $cov(x'_{prox}, v - b_{vrai}.\ e + b_{vrai}.\ u) = cov(dz_{inst}, u) + cov(dz_{inst}, -b_{vrai}.\ e) + cov(dz_{inst}, b_{vrai}.\ v)$

  $= 0$
- Par conséquent, le modèle empirique des MCO permet d'estimer sans biais les paramètres a_{vrai} et b_{vrai} du modèle 3 (et donc ceux du modèle 1)

Goux et Maurin (Revue économique, 2005)

Tableau 6. *Une analyse de l'effet de contexte endogène par la technique des variables instrumentales*

Modèle à probabilité linéaire	Variables dépendantes		
	% voisins en retard	Être en retard à 15 ans	
	Première étape	MCO	VI
% voisins en retard		0.22 (0.01)	0.20 (0.11)
[Garçon = 1]	.001 (.004)	0.11 (0.01)	0.11 (0.01)
Distribution des mois de naissance .			
% Janvier-juin	− .14 (.01)		
% Juillet-novembre	− .07 (.01)		
% Décembre	Ref.		
R^2	0.001	0.04	0.02
Nombre d'observations	24 367	24 367	24 367

Champ : pour l'enquête *t*, enfants nés en *t* − 15 résidant dans le voisinage depuis plus d'un an.
Source : INSEE, Enquêtes « Emploi », 1991-2002.

Goux et Maurin (tentative de réplication). Modèle des MCO

```
Call:
lm(formula = RET15 ~ VRET15 + S1, data = gm2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5928 -0.4043 -0.2771  0.5355  0.7453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.254656   0.005871   43.38   <2e-16 ***
VRET15      0.224473   0.009309   24.11   <2e-16 ***
S1          0.113680   0.006228   18.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4818 on 23948 degrees of freedom
  (405 observations deleted due to missingness)
Multiple R-squared: 0.03663,    Adjusted R-squared: 0.03655
F-statistic: 455.2 on 2 and 23948 DF,  p-value: < 2.2e-16
```

La forme réduite. Impact des instruments sur la variable dépendante

(non présentée dans l'article de 2005)

```
Call:
lm(formula = RET15 ~ S1 + VJANJUN + VJULNOV, data = gm2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4947 -0.4553 -0.3435  0.5433  0.6576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.381750   0.017307  22.057   <2e-16 ***
S1           0.112928   0.006304  17.914   <2e-16 ***
VJANJUN     -0.039376   0.018725  -2.103   0.0355 *
VJULNOV     -0.036094   0.019145  -1.885   0.0594 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4876 on 23940 degrees of freedom
Multiple R-squared:  0.01339,        Adjusted R-squared:  0.01326
F-statistic: 108.3 on 3 and 23940 DF,  p-value: < 2.2e-16
```

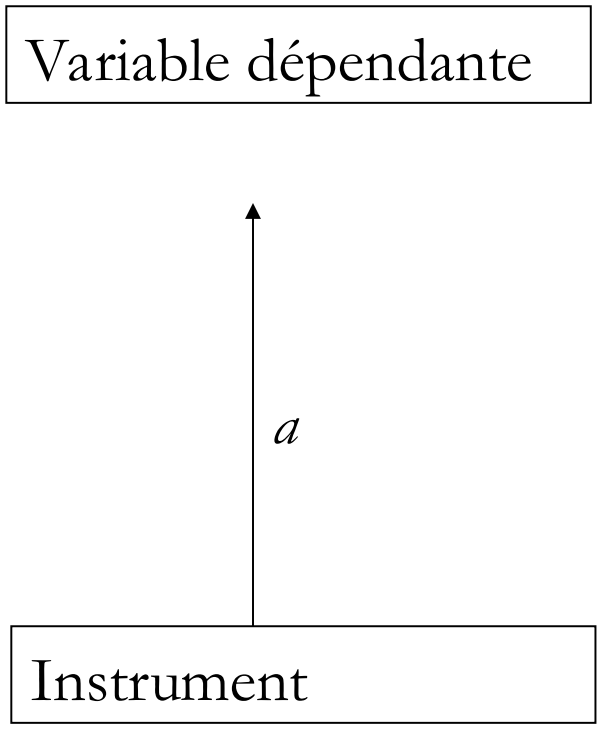
Goux et Maurin. Régression de première étape

```
2SLS estimates for 'eq1' (equation 1)
Model Formula: VRET15 ~ VJANJUN + VJULNOV + S1
Instruments: ~VJANJUN + VJULNOV + S1

               Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  0.50992739  0.01182131  43.13629 < 2.22e-16 ***
VJANJUN     -0.14299532  0.01278992 -11.18032 < 2.22e-16 ***
VJULNOV     -0.06400824  0.01307643  -4.89493  9.898e-07 ***
S1          -0.00293271  0.00430569  -0.68112     0.4958
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.333023 on 23940 degrees of freedom
Number of observations: 23944 Degrees of Freedom: 23940
SSR: 2655.052533 MSE: 0.110904 Root MSE: 0.333023
Multiple R-Squared: 0.008394 Adjusted R-Squared: 0.008269
```

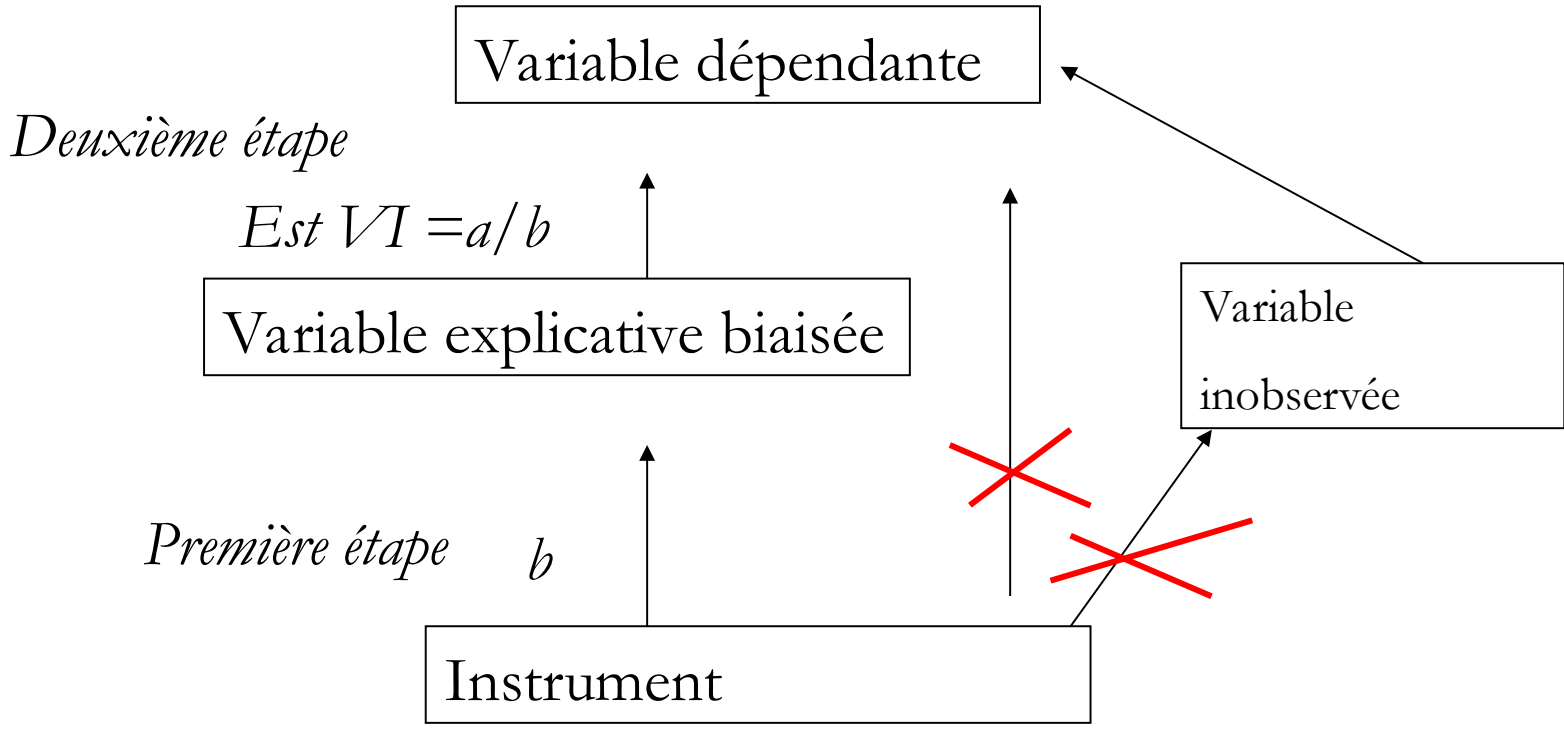
Goux et Maurin. Régression de deuxième étape

```
2SLS estimates for 'eq2' (equation 2)
Model Formula: RET15 ~ VRET15 + S1
Instruments: ~VJANJUN + VJULNOV + S1

              Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 0.28178700 0.04207210  6.69772 2.1639e-11 ***
VRET15      0.15861092 0.10186035  1.55714    0.11945
S1          0.11330863 0.00624303 18.14963 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.48229 on 23941 degrees of freedom
Number of observations: 23944 Degrees of Freedom: 23941
SSR: 5568.769685 MSE: 0.232604 Root MSE: 0.48229
Multiple R-Squared: 0.034594 Adjusted R-Squared: 0.034514
```

Estimation en forme réduite et estimation en deux étapes



Forme réduite



Estimateur des variables instrumentales
(Moindre carré en deux étapes)

Les variables instrumentales réduites à leur expression la plus simple

(inspiré de Goux et Maurin, 2007, Tableau 3-3)

Différence de moyennes	(Forme réduite)	(Première étape)	(Estimateur des VI)
Variable dépendante	Probabilité pour ego d'être en retard à 16 ans	Proportion de voisins en retard à 15 ans	Proba. ego retard Diff col. 1 / Diff. col. 2
Variable explicative			
Voisins nés au 1er sem. >50%	0.552	0.383	
Voisins nés au 1er sem. <=50%	0.575	0.426	
Différence en points de %	-2.3**	-4.4 ***	+51.2**

- >50% des voisins nés au 1er semestre
  - =>-4.4 points de % de voisins en retard à l'âge de 15 ans
  - => -2.3 points de % de retard d'ego à l'âge de 16 ans
- Si tout l'effet de la proportion des voisins nés au premier trimestre sur le redoublement d'égo passe par le fait qu'ils soient moins souvent en retard
  - => Quand 100% des voisins en retard à 15 ans : -0.023/-0.044 = 51.2 points de % de chance supplémentaire d'être en retard à 16 ans.

	MCO « naïf »	*Forme réduite*	*Première Étape*	*Variable instrumentale*
	P(retard à 16 ans)	P(retard à 16 ans)	Prop voisins en retard à 15 ans	P(retard à 16 ans)
Prop. voisins en retard à 15 ans	0.224***			0.382*
Prop. voisins nés entre jan. et mai		-0.0283*	-0.074***	
Contrôles	Oui	Oui	Oui	Oui

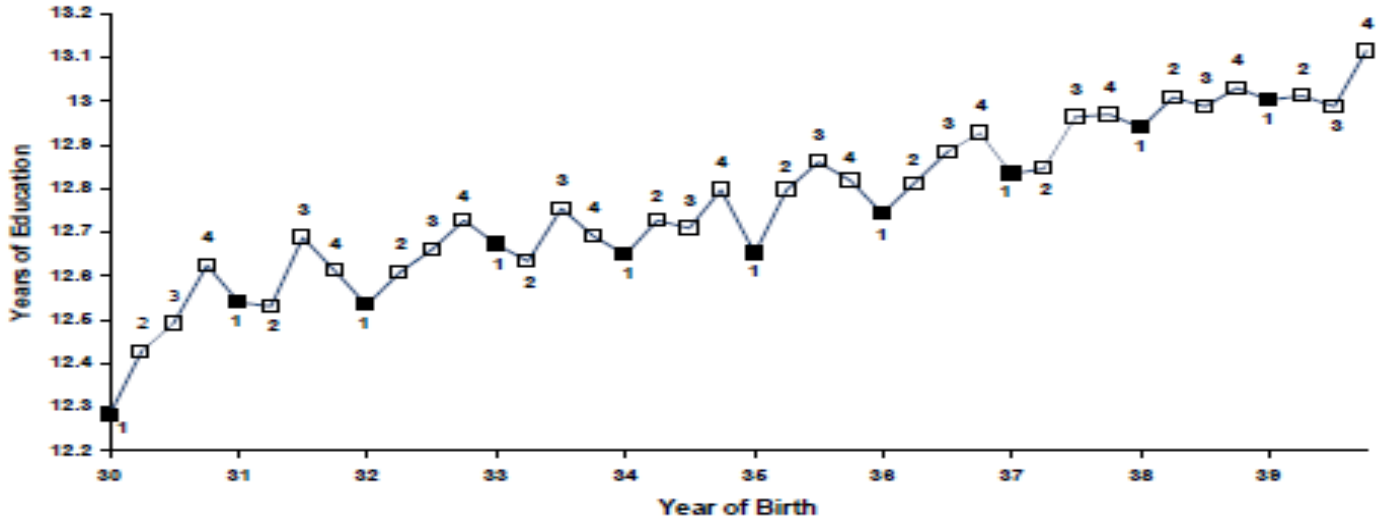
- Dans le cas d'une seule variable endogène et d'un seul instrument, l'estimateur des variables instrumentales (0.382) = l'effet forme réduite/effet de première étape (-0.0283/-0.074)

Un exemple canonique (Angrist, Krueger, 1991)

- Mesure de l'effet de l'école sur le revenu.
  - Le nombre d'années d'études mesure mal l'effet propre de l'école car il dépend aussi de l'intelligence des élèves.
  - Variable instrumentale : utiliser les effets des règles concernant les règles de scolarisation obligatoire
  - Aux États-Unis
    - Entrée obligatoire en élémentaire: avoir 6 ans avant le 1er janvier
    - Scolarité obligatoire jusqu'à 16 ans pile (ou 17, ou 18 ans selon les états). Interruption possible en cours d'année.
    - Conséquences : les enfants nés en début d'année entrent à l'école plus âgés que ceux nés en fin d'année, mais ils peuvent en sortir au même âge.
    - Ils vont donc moins longtemps à l'école.
    - Variation exogène de la durée de scolarité ne dépendant pas de « l'intelligence ».

Un effet petit mais certain

A. Average Education by Quarter of Birth (first stage)



B. Average Weekly Wage by Quarter of Birth (reduced form)

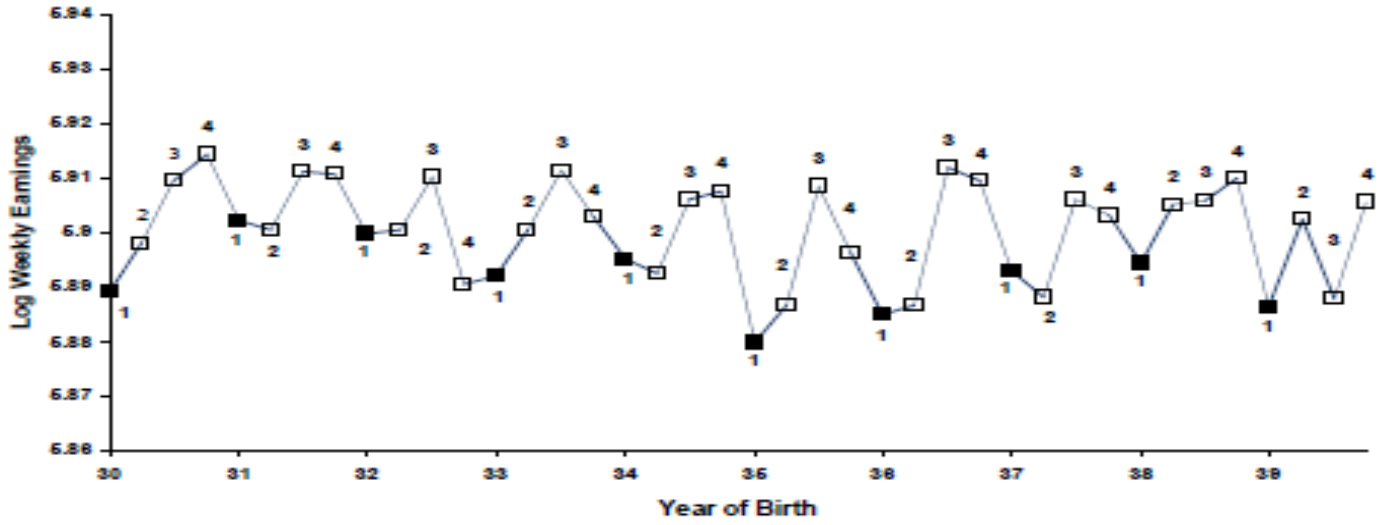


Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

Les variables instrumentales réduites à leur expression la plus simple (Angrist & Pischke 2008)

Différence de moyennes	(Forme réduite)	(Première étape)	(Estimateur des VI)
Variable dépendante			Log salaire hedo.
Variable explicative	Log salaire hebdomadaire	Nombre d'années d'éducation	Diff col. 1 / Diff. col. 2
Né au 1er semestre	5.892	12.69	
Né au 2ème semestre	5.905	12.84	
Différence en points de %	-0.013***	-0.15 ***	+0.087***

- Être né le 1er semestre de naissance
  - =>-0.15 année d'école en moins (1.8 mois d'école)
  - => -1.3% de salaire en moins
- Si tout l'effet du semestre de naissance sur le salaire passe par la durée d'éducation (et pas par d'autres canaux)
  - => 1 an d'école en plus : -0.013/-0.15 = +8.7% de salaire

Estimation de l'effet par la technique des variables instrumentales

- Variable expliquée : log du salaire hebdo
- Variable endogène : nombre d'années d'étude
- Instruments : trimestres de naissance * année de naissance.

TABLE IV

OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

a. Standard errors are in parentheses. Sample size is 247,199. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the State, County, and Neighborhoods 1 percent samples of the 1970 Census (15 percent form). The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

3. Valider la solution

Un dispositif empirique difficile à valider

- La validité de cette technique repose d'abord sur la qualité de l'argumentation
  - Convaincre le lecteur que l'instrument influence la variable explicative biaisée et n'influence que celle-ci
- Il existe des tests statistiques mais ces tests supposent que les instruments soient bons.
- Validités des tests, condition (généralement) nécessaire pour montrer la qualité des régressions avec variable instrumentale mais non suffisante.

La stratégie argumentative

- 1. Le raisonnement
- 2. Montrer que l'instrument est aussi bon qu'une ventilation aléatoire
  - Non corrélation avec les autres variables explicatives
- 3. La variable suspecte est elle endogène ?
  - Test d'endogénéité dit de Wu-Hausman
- 4. Les instruments sont ils bien exogènes ?
  - Test d'exogénéité des instruments (ou de validité jointe des deux instruments) dit de Sargan
- 5. Les instruments sont ils suffisamment puissants pour corriger le biais ?
  - Faiblesse des instruments

As good as random

- On étudie les corrélations (avec coefficients de corrélations, tests simples de Student ou régression) entre l'instrument et les autres variables explicatives observables non biaisés pour montrer qu'il n'y a pas de corrélation
  - => montrer que l'instrument est aussi bon qu'une ventilation aléatoire
  - Ex: la distribution des mois de naissance des voisins est-elle liée aux caractéristiques d'ego ?
- Si l'instrument est corrélé aux variables observables (ce n'est pas un problème en soi car le contrôle par les observables va corriger ça), l'instrument a beaucoup de chance d'être aussi corrélé aux inobservables.
- Mais si l'instrument non corrélé aux variables observables ne prouve pas pour autant l'absence de corrélations avec les inobservables…
- Et parfois il peut y avoir des corrélations avec certaines observables pour de bonnes raisons (ventilation aléatoire à l'intérieur de clusters)

Le mois de naissance aussi bon qu'une randomisation aléatoire?

Table A4

Adolescents' Characteristics and Neighbours' Dates of Birth

	Distribution of dates of birth of 15-*year-old* neighbours	
Individual characteristics	Prop. born January–May	Prop. born June–November
Father college grad.	42.4 (0.7)	49.7 (0.7)
Father not college grad.	42.3 (0.3)	49.6 (0.3)
Born January–May	42.7 (0.4)	49.2 (0.4)
Born June–November	42.0 (0.4)	49.9 (0.4)
Boy	42.3 (0.6)	49.0 (0.6)
Girl	42.2 (0.6)	50.6 (0.6)
French	42.3 (0.2)	49.6 (0.2)
non-French	41.7 (1.1)	50.9 (1.1)

Source: LFS $t = 1991, \ldots 2002$, Insee. *Sample*: 15-year-old respondents, observed at t and $t + 1$, who have been living in their neighbourhood for more than one year.
Note: The average proportion of peers born in January–May is 42.3% for boys and 42.2% for girls.

Table A3

Relationships Between an Adolescent's Characteristics and the Distribution of Dates of Birth of Other Adolescents in the Neighbourhood

Independent variables	Dependent variables: Proportion neighbours born January–May	Proportion neighbours born June–December	Neighbours' average month of Birth
Date of birth (continuous specification	–	–	0.005 (0.005)
Date of birth (dummies)			
Born January–May	0.008 (0.010)	−0.006 (0.010)	–
Born June–November	0.001 (0.010)	0.001 (0.010)	–
December	Ref.	Ref.	–
Boy	−0.008 (0.005)	0.013 (0.05)	0.001 (0.036)
Non-French	0.002 (0.012)	0.001 (0.012)	−0.007 (0.008)
Father's education			
College grad.	−0.004 (0.009)	0.005 (0.009)	0.009 (0.063)
High-school grad.	0.010 (0.010)	−0.011 (0.011)	−0.051 (0.072)
Vocational	Ref.	Ref.	Ref.
No Dip.	−0.010 (0.009)	0.008 (0.009)	0.022 (0.063)
Missing	−0.003 (0.010)	0.007 (0.010)	0.028 (0.070)
Mother's education			
College grad.	−0.000 (0.009)	0.004 (0.009)	0.009 (0.063)
High-school grad.	0.006 (0.009)	−0.003 (0.010)	−0.025 (0.066)
Vocational	Ref.	Ref.	Ref.
No Dip.	−0.002 (0.009)	−0.006 (0.009)	0.050 (0.059)
Missing	−0.002 (0.009)	0.016 (0.009)	0.077 (0.063)
No. Obs.	13,116	13,116	13,116
R^2	0.002	0.003	0.001
Fisher (4 dummies father Educ. = 0)	0.93 (0.42)	0.83 (0.47)	0.46 (0.76)
Fisher (4 dummies mother Educ. = 0)	0.17 (0.91)	0.28 (0.83)	0.42 (0.74)

Source: LFS $t = 1991, \ldots 2002$, Insee. *Sample*: 15-*year-old* respondents, observed at t and $t + 1$, who have been living in their neighbourhood for more than one year. All regressions include a set of eleven years dummies as additional control variables. Standard deviation in parenthesis.

Test l'endogénéité de la variable: le test de Wu-Hausman

- Le test est le suivant, on compare l'estimation des MCO et de la régression instrumentée. Si le paramètre estimé n'est pas différent alors il n'y a pas d'endogénéité.
- Mise en œuvre simple par la technique de la régression augmentée
  - Au lieu d'introduire la prédiction à la place de la variable biaisée, on introduit le résidu de la régression de première étape en plus de la variable biaisée
  - première étape : $x_{endo} = a_0 + a_1.z_{inst} + a_2.x_2 + u_{prem}$
  - deuxième étape : $y = a_{est} + b_{est}x_{endo} + c_{est}.x_2 + \boldsymbol{d_{est}.u_{prem}} + u_{deux}$

Test de Wu-Hausman

- Si le résidu de première étape u_{prem} est significatif, la variable était bien endogène et on a eu raison d'instrumenter
- Si le résidu n'est pas significatif, la variable ne semble pas endogène.
- Il vaut alors mieux utiliser les MCO que les variables instrumentales car les MCO sont plus précis

- Limites : Pour pouvoir mener ce test il faut disposer d'un bon instrument

Goux et Maurin : Test de Wu Hausman

```
Call:
lm(formula = RET15 ~ VRET15 + S1 + res, data = gm3)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5960 -0.4040 -0.2779  0.5356  0.7521

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.281787   0.042029   6.705 2.06e-11 ***
VRET15      0.158611   0.101755   1.559    0.119
S1          0.113309   0.006237  18.168  < 2e-16 ***
res         0.066532   0.102184   0.651    0.515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4818 on 23940 degrees of freedom
Multiple R-squared: 0.03663,    Adjusted R-squared: 0.03651
F-statistic: 303.4 on 3 and 23940 DF,  p-value: < 2.2e-16
```

Test de validité des instruments
Ou le test de Sargan (dit de « suridentification »)

- Les instruments sont-ils corrects ? Sont-ils bien exogènes, ie non corrélés avec u ?
- Condition : il faut disposer d'au moins une variable instrumentale de plus que de variables endogènes.
- Quand une variable est endogène, il faut alors avoir deux instruments.

Test de Sargan d'exogénéité des instruments

- On estime la régression augmentée suivante de deuxième étape

$$y = a_{est} + b_{est} x_{endo} + c_{est} . x_2 + \boldsymbol{d_{est} . u_{prem}} + u_{deux}$$

- On récupère le résidu u_{deux} et on le régresse sur les instruments z_{inst1} et z_{inst2}.

$$u_{deux} = f . z_{inst1} + g . z_{inst2} + w \qquad (2)$$

- Si l'une des variables est significative, alors ça veut dire qu'elle n'est pas exogène : elle est corrélée au résidu, ce qui est contraire aux hypothèses.
- Un résumé du test de Sargan pour l'ensemble des variables exogènes est aussi obtenu par la statistique suivante à partir de la régression 2 :
- N*R2 que l'on compare à une loi de Chi2 au degré de liberté suivant (Nb instruments-Nb variables endogènes)
- Si c'est significatif, les instruments ne sont pas (tous) exogènes

Limites du test de Sargan

- Il est déjà difficile de trouver un instrument… en trouver deux est deux fois plus dur.
  - Il existe des trucs pour obtenir plusieurs instruments à partir d'une seule variable comme élever son instrument au carré, découper son instrument en classes, etc.
- Le test de Sargan vérifie le fait que les deux instruments corrigent de la même façon le biais
  - Au moins un des instruments doit être bon
    - Deux mauvais instruments peuvent être validés par le test de Sargan ! (Cf. simulation)
  - Les populations sensibles à l'instrument doivent être les mêmes
    - Deux bons instruments mais impactant des populations différentes peuvent ne pas être validés par le test de Sargan
    - Car l'estimateur des VI estime les effets moyens locaux (LATE) et non les effets moyens
    - Ex: taille de la famille expliquée par les naissances de jumeaux ou par le sexe ratio des deux premiers nés.

Goux et Maurin : Validité des instruments

```
Call:
lm(formula = et2b$residuals ~ gm2$VJANJUN + gm2$VJULNOV)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6056 -0.4048 -0.2784  0.5354  0.7547

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01912    0.01683   1.136    0.256
gm2$VJANJUN -0.01669    0.01850  -0.902    0.367
gm2$VJULNOV -0.02594    0.01892  -1.371    0.170

Residual standard error: 0.4818 on 23941 degrees of freedom
Multiple R-squared: 8.91e-05,    Adjusted R-squared: 5.572e-06
F-statistic: 1.067 on 2 and 23941 DF,  p-value: 0.3442
```

Goux et Maurin : Validité des instruments

- On connait le R2=8.91e-05 et l'effectif N= 23944

```
stat<-8.91e-05*23944
stat
```

=> 2.133

```
1-pchisq(2.133,df=1)
```

=> 0.144

Le test n'est pas significatif. Les instruments ne sont pas endogènes. Ils sont donc corrects

Faiblesse des instruments

- L'estimateur permet d'estimer les vrais paramètres quand on tend vers l'infini.
- Sur un petit échantillon, on peut tomber sur un biais substantiel.
- La ou les variables instrumentales doivent avoir un pouvoir explicatif important sur la variable endogène, sinon sur un échantillon fini l'estimateur des VI se rapproche de l'estimateur des MCO en pire (en moins précis). On parle alors ***d'instrument faible***.

Détection de la faiblesse des instruments

- A ma connaissance il n'existe pas de « tests » au sens strict, mais des règles de détection.
- Règle du pouce 1 (F<10): si la statistique de Fisher de nullité jointe des instruments dans la régression de première étape est inférieur à 10, alors les instruments sont dits faibles.

Faiblesse des instruments

F Test

```
Analysis of Variance Table
Model 1: VRET15 ~ S1
Model 2: VRET15 ~ VJANJUN + VJULNOV + S1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  23942 2677.5
2  23940 2655.1  2    22.419 101.07 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Une critique célèbre (Bound, Jaeger, Baker, 1995)

- Colonnes 3 à 6 Réplication d'Angrist Krueger. Instruments faible
- Colonne 2 : Modèle plus simple avec comme instruments uniquement le trimestre de naissance

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, ×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 × 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).

Effet moyen ou effet local

- MCO mesure des effets moyens (qui peuvent être biaisés) : ATE : Average Treatment Effect
- L'estimateur des variables instrumentales estime l'effet du traitement sur ceux qui réagissent à l'instrument.
- Ces effets non biaisés ne sont pas forcément moyens mais des effets locaux : LATE. Local Average Treatment Effect
- Ex. Bound & Alii ➔ effet d'une année d'école : +14% de salaire.
- Effet d'une année d'école plus important aux environs de 16 ans qu'après.

4. Bilan des variables instrumentales

Trouver les variables instrumentales : that is the question ?

- La recherche sauvage d'instruments
  - Recherche (inavouable) d'instruments qui répondent aux tests
  - Ensuite on raconte une histoire plus ou moins convaincante sur les bonnes raisons de l'instrumentation
  - Ex : Rendement de l'éducation => instrumenter par l'éducation des parents
  - Hypothèse : tout l'effet de l'éducation des parents n'impacte le salaire que par son effet sur l'éducation des enfants.
  - Pratique en voie de disparition
- La recherche « d'expériences naturelles »

Exemples de variables instrumentales fondées sur des expériences naturelles

- Sex ratio.
  - Aléatoire (sauf si avortement différentiel)
  - A une incidence sur différents comportements potentiellement endogènes : Taille des ménages
  - Permet d'estimer les effets de la taille du ménage sur l'activité de la femme, le divorce, etc.
- Météorologie
  - Aléatoire
  - Incidence sur la production agricole
  - Modèle d'offre et de demande

Variables instrumentales (exemples)

- Dates de naissance
  - Répartition aléatoire dans l'année (pas tout à fait le cas : saisonnalité des naissances).
  - Effet sur l'éducation et d'autres aspects
- Mesures de politiques publiques
  - Seuil de déclenchement de politique publique
  - Individus se trouvent plus ou moins aléatoirement autour de ce seuil de déclenchement
  - La distance au seuil de déclenchement peut être un instrument
  - Ex : Aurélie Ouss (Maurin, Ouss, 2009), Effet des remises de peine sur la récidive. 14 juillet, date traditionnelle de remise automatique de peine.
    - Si date de sortie officielle avant le 14 juillet, pas de remise de peine.
    - Après remise de peine
  - (date de sortie officielle - 14 juillet) => instrument de la remise de peine.

Variables instrumentales (exemples)

- Localisation dans l'espace
  - Position dans l'espace sous certain rapport exogène.
  - Influence de l'esclavage en Afrique au 17ème/18ème sur la confiance interpersonnelle en 2007(Nunn and Wantchekon 2011)
    - Mécanisme explicatif: le commerce d'esclave a détruit la confiance interpersonnelle dans les sociétés qui l'ont subi
    - Causalité inverse: les groupes africains qui s'adonnaient au commerce d'esclave avaient des degrés de confiance bas.
    - Instrument: Proximité à la côte comme instrument du commerce d'esclaves.

Usages en sociologie – En progression

- Effet des magazines sur le développement des sociétés anti-esclavagistes aux EU au 19è (King, Haveman, 2008)
  - Problème: causalité inverse. Les sociétés anti-esclavagistes peuvent conduire au développement de la presse
  - Instrument: Le nombre de bureau de postes
- Effet de la position dans le réseau sur la probabilité d'avoir un poste. (Godechot, Mariot, 2004)
  - Problème : la position dans le réseau capture la qualité de la thèse qu'on mesure mal
  - Instrument: position dans le réseau des autres docteurs du même directeur de thèse (capital social exogène du directeur).
  - Limite : appariement docteur-directeur se fait par niveau de qualité

Usages en sociologie

- Acquisition de la nationalité sur l'emploi. (Fougère Safi, 2006)
  - Problème causalité inverse : emploi -> cause de l'acquisition de la nationalité
  - Instruments : le nombre d'étrangers résidant dans le même département au moment du recensement, et le nombre d'étrangers de même origine résidant dans le même département au moment du recensement
  - Ces deux variables affectent la longueur de la file d'attente des candidats à la naturalisation, et donc la probabilité individuelle d'acquérir la nationalité française entre deux recensements successifs
- Effet du sentiment de discrimination sur la satisfaction (Safi, 2010)
  - Pb de causalité inverse possible : les satisfaits se sentent moins discriminés
  - Instrument : Appartenance religieuse au judaisme ou à l'islam
  - Hypothèse : cela affecte le sentiment de discrimination. Pas d'effet direct sur la satisfaction
  - Limite: on peut discuter de cette dernière hypothèse

Les variables instrumentales : De l'apologie au doute

- Limites de la technique
  - Econométrie plus complexe
  - Difficiles à trouver
  - Aspect « âge du capitaine »
  - Pas toujours vraiment exogènes
  - Exogénéité difficile à prouver
  - Potentiellement faibles
  - Et n'estimant que des effets locaux
  - Amélioration par rapport aux MCO discutables
- Evolution en économie
  - De la recherche systématique d'instruments
  - Aux expériences aléatoires contrôlées.
    - Un échantillon auquel on applique la mesure
    - Un échantillon témoin auquel on n'applique pas la mesure
  - Si l'expérience est sans biais (tirage, attrition, etc.) on mesure alors la différence de résultat par une différence de moyenne et on contrôle la significativité par un simple test de student.

5. Programmes

R : fonction ivreg du package AER

```
#installation du package AER
install.packages("AER")
library("AER")

#Syntaxe
myreg<-ivreg(y ~ x_endo+x2+x3|instr+x2+x3,data=db)
summary(myreg)

#Pratique : pour tous les tests
summary(myreg,diagnostics=TRUE)

#Limite : n'imprime pas les régressions de première étape… -> à
  estimer à la main
```

ivreg avec diagnostics

```
Call:
ivreg(formula = RET15 ~ VRET15 + S1 | VJANJUN + VJULNOV + S1,
    data = gm2)
Residuals:
    Min      1Q  Median      3Q     Max
-0.5537 -0.3951 -0.2976  0.5520  0.7182
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.281787   0.042072   6.698 2.16e-11 ***
VRET15      0.158611   0.101860   1.557    0.119
S1          0.113309   0.006243  18.150  < 2e-16 ***
Diagnostic tests:
                  df1   df2 statistic p-value
Weak instruments    2 23940   101.071  <2e-16 ***
Wu-Hausman          1 23940     0.424   0.515
Sargan              1    NA     2.129   0.145
```

R : systemfit

```
#installation du package systemfit
install.packages("systemfit")
library("systemfit")

#Syntaxe
premetap<- x_endo ~ instr+x2+x3
deuxetap <- y ~ x_endo+x2+x3
system <- list( premetap, deuxetap)
inst <- ~ instr+x2+x3
fit2sls <- systemfit( system, "2SLS",inst, data=db)
summary(fit2sls)
```

Sous Stata

- Ivregress & ivprobit

  ivregress 2sls y x2 x3 **(**x_endo=instr1 instr2**)**

  ivprobit y x2 x3 **(**x_endo=instr1 instr2**)**

  ivregress 2sls y x2 x3 **(**x_endo1 x_endo2=instr1 instr2**)**
- Régression de première étape

  ivregress 2sls y x2 x3 **(**x_endo=instr1 instr2**), first**
- Tests d'endogénéité

  estat endogenous
- Test de suridentification

  estat overid
- Détection des instruments faibles

  estat firststage

Sous SAS

```
proc syslin 2SlS data=mabase FIRST;
model y = x_endo x2 x3 /overid ;
endogenous y x_endo;
instruments z1 z2 x2 x3 ;
run;
```

SPSS

- Pour les utilisateurs de SPSS : 2SLS
  - En ligne de commande :

    2sls y with x w

    / instruments z w

    / constant.

  - Avec les menus déroulants
    - Analyze → Regression → Two-Stage Least Squares
    - DEPENDENT, EXPLANATORY, and INSTRUMENTAL

Références

Angrist, Krueger, 1991."Does Compulsory School Attendance Affect Schooling and Earnings?" , *Quarterly Journal of Economics*, 106(4), 979-1014

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), 443-450.

Fougère D. et Safi M., « L'acquisition de la nationalité française : quels effets sur l'accès à l'emploi des immigrés ? », *France Portrait Social*, Edition 2005-2006, Insee, p. 163-184

Godechot, O., & Mariot, N. (2004). « Les deux formes du capital social », *Revue française de sociologie,* vol. 45, n°2, pp. 243-282. Goux, D., & Maurin, É. (2005). « Composition sociale du voisinage et échec scolaire », *Revue économique*, Vol. 56, No. 2, pp. 349-361.

Goux, D., & Maurin, E. (2007). Close neighbours matter: Neighbourhood effects on early performance at school*. *The Economic Journal*, *117*(523), 1193-1215.

King, M. et Haveman H. "Antislavery in America: The press, the pulpit, and the rise of antislavery societies". *Administrative Science Quarterly*, 2008, vol. 53, no 3, p. 492-528.

Maurin, Eric and Ouss, Aurelie, « Sentence Reductions and Recidivism: Lessons from the Bastille Day Quasi Experiment ». IZA Discussion Paper No. 3990, 2009.

Nunn, N., & Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American economic review*, *101*, 3221-3252.

Safi, M. (2010). Immigrants' life satisfaction in Europe: Between assimilation and discrimination. *European Sociological Review*, *26*(2), 159-176.