FOAD
COURS D' ECONOMETRIE 1
CHAPITRE 3 : Variable explicative endogène : La méthode des Variables Instrumentales.
23 mars 2012.

Christine Maurel
Maître de conférences en Sciences Economiques
Université de Toulouse 1 - Capitole
Toulouse School of Economics-ARQADE

Table des matières

I Introduction

Ce chapitre 3 est le seul chapitre où l'application sur un fichier de données économiques figure après la présentation de tous les outils théoriques.
Nous présenterons deux points théoriques dans cette introduction pour vous apprendre le vocabulaire particulier utilisé par les économètres qui travaillent sur les Modèles à Equations Simultanées (MES). Ensuite nous appliquerons les techniques d'estimation et de tests présentés en début de chapitre sur un fichier de données.

1. Vocabulaire et identification des Modèles à Equations Simultanées (MES).
   Source : R. Bourbonnais "Econométrie", Dunod , 6 ième édition, 2005, p 215.
   Soit le modèle à équations simultanées (MES) suivant :
   $Y1 + aY2 + b = \epsilon_1$ Equation 1 du MES
   $cY1 + Y2 + dX + e = \epsilon_2$ Equation 2 du MES
   En économétrie ce système d'équations est appelé "système d'équations simultanées" qui décrit un MES, car les deux variables endogènes Y_1et Y_2 sont déterminées simultanément ;elles sont interdépendantes.
   Ces modèles sont très utilisés en macroéconomie par exemple où les équations du modèle décrivent l'économie d'un pays et en particulier les relations entre les grandeurs macroéconomiques (les variables du MES).
   Les MES peuvent être écrits sous la forme matricielle suivante : $BY + CX = \varepsilon$.
   Sur notre exemple on a :

$$\begin{bmatrix} 1 & a \\ c & 1 \end{bmatrix} \begin{bmatrix} Y1 \\ Y2 \end{bmatrix} + \begin{bmatrix} 0 & b \\ d & e \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

où U est le vecteur unité.
A partir de ce système on note g le nombre de variables endogènes du système c'est à dire le nombre de variables dans Y, et k le nombre de variables exogènes du MES c'est à dire le nombre de variables dans X. Ici, on a $g = 2$ et $k = 2$.
Cette écriture matricielle du MES est qualifiée de "forme structurelle ", notée FS, car plusieurs variables endogènes (ici 2) figurent dnas chaque équation. La FS a une interprétation économique en terme d'élasticités par exemple ; elle décrit l'économie d'un pays. Ces équations peuvent aussi décrire le comportement des individus, des entreprises ; elles peuvent être issues de la théorie microéconomique sur le comportement des individus : quel est le calcul que fait un individu pour décider du nombre d'années d'études qu'il va poursuivre ? quel est son calcul pour choisir une formation ? la microéconomie nous enseigne que les individus font un arbitrage entre coût et bénéfices (futurs, attendus ou prévus). Nous reviendrons sur ces considérations sur la FS d'un modèle dans le paragraphe sur l'application.
Les paramètres de la FS sont qualifiés de paramètres "structurels". Si on souhaite estimer tous les paramètres de la FS , c'est à dire a, b, c, d et e alors on utilisera des techniques d'estimation que l'on qualifie de techniques en "information complète" (par exemple le méthode des Triple Moindres Carrés). En information limitée, on souhaite estimer une seule équation et on s'interesse à un nombre restreint de paramètres de la FS ou paramètres structurels.

Le problème de l'identification :
La question que l'on se pose quand on étudie l'identification est la suivante : peut - on retrouver les paramètres structurels si on dispose des paramètres de la Forme Réduite (FR) que nous présentons ci-après ?
Nous pouvons réécrire le MES sous la forme (presque) habituelle $Y = AX + u$ avec $A = -B^{-1}C$ et $u = -B^{-1}\epsilon$; dans cette écriture Y représente la liste des variables et non pas les observations de ces variables de la même manière que dans l'écriture de la FS ci-dessus. Notons que dans cette écriture toutes les variables explicatives sont exogènes ; c'est ce que l'on nomme "Forme Réduite", notée FR. Nous verrons que sur cette FR on peut appliquer les MCO.
Les matrices B et C contiennent les paramètres de la forme structurelle du modèle alors que la matrice A contient les paramètres de la forme réduite.
Avec les notations que que nous avons adoptées, on a le système $BA = -C$ ce qui donne :

$$\begin{pmatrix} 1 & a \\ c & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 \\ \hat{\alpha}_3 & \hat{\alpha}_4 \end{pmatrix} = - \begin{pmatrix} 0 & b \\ d & e \end{pmatrix}$$

Si nous souhaitons estimer les pramètres de l'équation 1 du MES , nous nous intéressons à a et b et nous devons condidérer les deux premières équations de ce système $BA = -C$:
$\hat{\alpha}_1 + a\hat{\alpha}_3 = 0$
$\hat{\alpha}_2 + a\hat{\alpha}_4 = -b$
On voit que l'équation 1 est juste identifiée car il y a 2 équations à deux inconnues (a et b). On peut donc retrouver a et b à partir des paramètres estimés

de la FR.
Pour l'équation 2 du MES , c'est à dire pour les paramètres c, d et e, le système est le suivant :
$c\hat{\alpha}_1 + \hat{\alpha}_3 = -d$
$c\hat{\alpha}_2 + \hat{\alpha}_4 = -e$
Il y a cette fois 2 équations à 3 inconnues (c, d et e) et donc cette équation est sous identifiée. [1]

On peut aussi présenter des conditions nécessaires et suffisantes à l'identification de la manière suivante :
On note g' le nombre de variables endogènes de l'équation considérée et k' le nombre de variables exogènes de la même équation. Par exemple si on souhaite estimer la première équation du MES il faut vérifier l'identification de cette équation c'est à dire il faut vérifier si le nombre $g' - g + k - k'$ est supérieur , égal ou inférieur au nombre $g - 1$. Si $g' - g + k - k'$ est strictement plus grand que $g - 1$ on dit que l'équation est suridentifiée ; si ce nombre est égal à $g - 1$, alors l'équation sera juste identifiée [2].
Pour l'équation 1 du système on a $g' = 2$ et $k' = 1$ [3]. Donc pour l'équation 1 ,$g' - g + k - k' = 2 - 2 + 2 - 1 = 1$ qui est égal à $g - 1 = 1$; l'équation 1 est donc juste identifiée.
Pour l'équation 2, $g' - g + k - k' = 2 - 2 + 1 - 2 = -1$ qui est inférieur à $g - 1 = 1$. L'équation 2 est donc sous identifiée, on ne peut pas retrouver les paramètres structurels du modèle.

2. Biais des OLS en présence d'une variable explicative endogène
Revenons sur un cas particulier des MES : on souhaite estimer une seule équation qui contient une variable explicative endogène ce qui signifie que nous allons utiliser une méthode d'estimation en information incomplète. Dans ce cas on montre que les paramètres estimés par MCO sont biaisés et inconsistants.
Etudions ce biais sur un exemple très simple :
Pour simplifier, supposons que le modèle contienne une variable explicative exogène X_2 et une variable explicative endogène $Y2$ mais qu'il ne contienne pas de constante ; ce modèle s'écrit de la manière suivante :

$$Y1 = \alpha_1 Y2 + \alpha_2 X_2 + u \tag{1}$$

Si $Y2$ est endogène, cela signifie qu'elle dépend de variables explicatives. Faisons l'hypothèse que cette relation s'écrit :

$$Y2 = \beta_1 Y1 + \beta_2 X3 + v \tag{2}$$

1. Remarque : si $e = 0$ alors l'équation 1 reste identifiée et l'équation 2 le devient.
2. si ce nombre est inférieur à $g - 1$ le modèle est sous identifié et on ne peut pas retrouver les paramètres structurels.
3. La variable Unité qui est égale à 1 pour toutes les observations.

Il s'agit bien d'un MES car $Y1$ et $Y2$ sont déterminées simultanément.
Notons que les variables X_2 et X_3 sont exogènes.
Si nous substituons Y_1 par sa valeur dans la seconde équation de 1 nous obtenons :

$$Y2 = \beta_1(\alpha_1 Y1 + \alpha_2 X2 + u) + \beta_2 X3 + v$$
$$(1 - \alpha_1\beta_1)Y2 = \alpha_2\beta_1 X2 + \beta_2 X3 + v + \beta_1 u$$

Supposons que $\alpha_1\beta_1 \neq 1$, on obtient alors

$$Y2 = \frac{\alpha_2\beta_1}{1 - \alpha_1\beta_1}X2 + \frac{\beta_2}{1 - \alpha_1\beta_1}X3 + \frac{v + \beta_1 u}{1 - \alpha_1\beta_1} \quad (3)$$

Cette équation est une forme réduite car toutes les variables explicatives sont exogènes donc les OLS sont sans biais et consistants quand on les applique sur cette équation en FR. Cette propriété sera utilisée pour estimer les paramètres dans le paragraphe concernant l'application.
De cette forme réduite, on peut déduire que $Y2$ et u sont corrélées (sauf si $\beta_1 = 0$) donc les MCO dans 1 sont biaisés car l'hypothèse $E(X'u) = 0$ n'est plus vérifiée. Ce biais s'appelle "biais de simultanéité" ou "biais d'endogénéité".

Donc, si le modèle contient une variable explicative endogène alors les MCO sont biaisés et inconsistants c'est à dire que le biais ne s'atténue pas quand la taille de l'échantillon est très grande. Pour obtenir des estimateurs consistants, il faut utiliser une autre méthode d'estimation, par exemple la méthode des variables instrumentales qui est présentée dans le paragraphe qui suit.

II Estimation par la méthode des Variables Instrumentales : éléments théoriques

Soit le modèle $Y = X\beta + u$ où u est d'espérance nulle et de variance conditionnelle égale à σ^2. Ce modèle reprend les notations habituelles du chapitre 1.
Le problème étudié dans ce chapitre est le suivant : nous soupçonnons une variable explicative d'être endogène [4]. Supposons qu'il s'agisse de la dernière variable du modèle, notée Xk. Dans ce cas, comme nous l'avons vu dans l'introduction de ce chapitre, la variable Xk est corrélée avec le terme d'erreur et l'hypothèse $E(Xk_i, u_i) = 0$ n'est plus vérifiée.
Ainsi si on applique les MCO sur le modèle :$Y = X\beta + u$ qui contient une variable explicative endogène on obtient des paramètres estimés biaisés et inconsistants. On montre que dans ce cas, on peut obtenir des estimateurs consistants grâce à la méthode des variables instrumentales qui est une méthode d'estimation en information limitée c'est à dire que nous nous intéressons à une seule équation.
On appelle variable instrumentale, une variable exogène (non corrélée avec le terme

4. dans ce cours, il y aura une seule variable explicative soupçonnée d'être endogène.

d'erreur) et corrélée avec la variable explicative qui est soupçonnée d'être endogène ici Xk.
Soit Z la matrice des variables instrumentales de dimension (N,p) [5] avec $p \geq k$ c'est à dire que le nombre d'instruments doit être supérieur ou égal au nombre de variables explicatives du modèle. Si $p = k$ le modèle est juste identifié ; si $p > k$ le modèle est suridentifié. Attention si $p < k$ le modèle est sous identifié et on ne peut pas retrouver les paramètres structurels.

La formule générale pour l'estimateur des variables instrumentales , noté $\hat{\beta}_{IV}$ est la suivante :

$$\hat{\beta}_{IV} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y \tag{4}$$

Si le modèle est juste identifié c'est à dire si $p = k$, alors l'équation 4 se réduit à :

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y \tag{5}$$

En pratique la matrice Z contient toutes les variables exogènes du modèle. Supposons par exemple que l'équation structurelle soit la suivante :
$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \cdots + \beta_{k-1} Xk - 1_i + \beta_k Xk_i + u_i$
dans laquelle on soupçonne Xk d'être endogène. Notons $Z_1, \ldots, Z_m$ les variables instrumentales. Dans ce cas la matrice Z aura pour colonne : $(1, X1, X2, \ldots, Xk - 1, Z1, \ldots, Zm)$ avec $p = k + m$ où p est le nombre de colonnes de Z. Attention : les notations peuvent prêter à confusion ; ici la matrice Z contient toutes les variables exogènes c'est à dire d'abord toutes les variables exogènes de la FS ou encore $1, X1, X2, \ldots, Xk - 1)$ ainsi que toutes les variables instrumentales $(Z1, \ldots, Zm)$. Remarquons de plus que les variables instrumentales Z_i ne figurent pas dans l'équation structurelle II ; ces restrictions d'exclusion permettent d'identifier le modèle (voir introduction de ce chapitre).
On montre dans ce cas que, pour obtenir l'estimateur des Variables Instrumentales il faut procéder selon les deux étapes suivantes :
Etape 1 : on regresse Xk sur $(1, X1, X2, \ldots, Xk - 1, Z1, \ldots, Zm)$ et on calcule $\hat{X}k$ grâce à cette regression. Notons $\hat{X}$ la matrice qui contient les k colonnes suivantes :$1, X1, X2, \ldots, Xk - 1, \hat{X}k$ où Xk a été remplacée par $\hat{X}k$. Remarquons que la regression de cette première étape n'a pas d'interprétation économique. A nouveau elle pourrait être nommée "regression auxiliaire" car elle aide à la correction du biais.
Etape 2 : on regresse Y sur $\hat{X}$ c'est à dire sur les variables $(1, X1, X2, \ldots, Xk - 1, \hat{X}k)$. Les paramètres estimés à la fin de cette seconde étape sont les estimateurs par la méthode des variables instrumentales. Cette estimation en deux étapes correspond aussi à la méthode des Doubles Moindres Carrés, DMC (ou 2SLS). On obtient dans cette seconde étape :

$$\hat{\beta}_{2SLS} = (\hat{X}'X)^{-1}\hat{X}'Y = \hat{\beta}_{IV} \tag{6}$$

5. rappel X est de dimension (N,k)

Remarquons que la variable $\hat{X}$ est considérée comme la matrice d'instruments Z [6].
Démontrons que $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$
On a vu que $\hat{X} = Z'(Z'Z)^{-1}Z'X = P_Z X$ où P_Z est la matrice de projection sur Z. Cette matrice est idempotente et symétrique. Donc $\hat{X}'X = X'P_Z X = X'P_Z P_Z X = (P_Z X)'P_Z X = \hat{X}'\hat{X}$. Ainsi

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'X)^{-1}\hat{X}'Y \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \\ &= [(P_Z X)'P_Z X]^{-1}(P_Z X)'Y \\ &= [X'P_Z P_Z X]^{-1}X'P_Z Y \\ &= \hat{\beta}_{IV}\end{aligned}$$

De plus on montre que :

$$\hat{\sigma^2} = \frac{1}{N-k}SCR$$

où SCR est la Somme des Carrés des Résidus des 2SLS et

$$\hat{Var}(\hat{\beta}_{2SLS}) = \hat{\sigma^2}(\hat{X}'X)^{-1}$$

.
Attention ne pas faire “manuellement" les deux étapes en utilisant deux fois la procédure “reg" car les écart-types ne seraient pas corrects. [7]

Nous avons vu qu'une variable instrumentale doit vérifier deux conditions : la première est qu'elle doit être exogène c'est à dire non corrélée avec le terme d'erreur. Cette condition ne peut pas être vérifiée. On peut tester une version de cette condition si le modèle est suridentifié ; c'est ce que nous étudierons dans une des sections suivantes. Nous avons vu aussi qu'une variable instrumentale doit vérifier une seconde condition : elle doit être corrélée, du moins suffisament, avec la variable endogène explicative. Que se passe -t - il si cela n'est pas le cas ? Pour répondre à cette question nous allons utiliser un exemple de Wooldridge 2006 p 520.
La liste des variables du fichier est la suivante :
bwght : birth weight, in ounces
lbwght : log de bwght
packs : packs smoked per day while pregnant
cigprice : cig. price in home state, 1988
L'équation structurelle est : $lbwght_i = \alpha + \beta packs_i + u_i$ que l'on estime par MCO. Les résultats sont présentés dans la figure 1.

Le paramètre estimé de la variable *packs* a le signe attendu : il est négatif ; au plus la femme enceinte fume pendant sa grossesse au plus le poids du bébé sera faible.

6. $\hat{X}k$ vérifie les deux conditions pour être considérée comme un instrument pour Xk : cette variable n'est pas corrélée avec le terme d'erreur mais elle est corrélée avec la variable Xk.

7. Wooldridge, 2006 p 526 ou Wooldridge, 2002 p 91.

Dependent Variable: lbwght

Number of Observations Read	1388
Number of Observations Used	1388

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	1	0.99778	0.99778	27.98	<.0001
Error	1386	49.42256	0.03566		
Corrected Total	1387	50.42034			

Root MSE	0.18883	R-Square	0.0198
Dependent Mean	4.76003	Adj R-Sq	0.0191
Coeff Var	3.96708		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	4.76940	0.00537	888.26	<.0001
packs	1	-0.08981	0.01698	-5.29	<.0001

FIGURE 1 – MCO sur le modèle pour lbwght

Dans ce modèle on soupçonne la variable *packs* d'être corrélée avec le terme d'erreur [8]. Les MCO sont biaisés dans ce cas et on choisit d'estimer le modèle par DMC. On choisit donc comme instrument la variable *cigprice* qui peut être supposée exogène [9]. Quand nous appliquons les DMC (nous présenterons la synthaxe de SAS dans l'application), nous estimons les deux équations ci dessous :
$lbwght = a0 + a1packs + u$
et la forme réduite pour la variable *packs* : $packs = b0 + b1cigprice + v$.
Remarquons que ce modèle est juste identifié pusique nous disposons d'un instrument pour une variable explicative endogène.
Par 2SLS, nous obtenons les résultats présentés dans la figure 2.

Le paramètre structurel $a1$ n'est pas significatif ; ce qui n'est peut être pas surprenant pour un produit avec addiction : le nombre de paquets de cigarette consommé est peut être peu lié au prix. Par contre le signe de ce paramètre est devenu positif ce qui est en contradiction avec le résultat attendu. Le problème est que la variable instrumentale *cigprice* n'est pas corrélée avec *packs* comme on peut le voir dans les résultats : le paramètre $b1$ n'est pas significativement différent de 0 ; dans ce cas l'instrument *cigprice* est qualifié d'instrument "pauvre" (poor instrument). Donc il faut toujours vérifier que les paramètres des variables instrumentales soient significativement différents de 0 dans la regression auxiliaire avant d'appliquer les 2SLS. Cette condition sur les instruments est facile à vérifier.
Nous présentons dans le paragraphe suivant d'autres tests à effectuer pour la méthode des Variables Instrumentales.

8. la valeur de la variable packs résulte d'un calcul pour la femme enceinte
9. pour les individus ici les femmes enceintes

Non linéaire 2SLS Récapitulatif des erreurs résiduelles

Équation	Modèle DF	Erreur DF	SSE	MSE	Racine MSE	R-carré	R carré aj.
lbwght	2	1386	1221.7	0.8815	0.9389	-23.230	-23.248
packs	2	1386	123.7	0.0892	0.2987	0.0001	-0.0006

Estimations 2SLS Parameter non linéaires

Parameter	Estimation	Erreur standard appr.	Valeur du test t	Approx Pr > |t|
a0	4.448136	0.9082	4.90	<.0001
a1	2.988676	8.6989	0.34	0.7312
b0	0.067426	0.1025	0.66	0.5109
b1	0.000283	0.000783	0.36	0.7179

Nombre d'observations		Statistiques pour Système	
Used	1388	Objective	1.426E-25
Missing	0	Objective*N	1.979E-22

FIGURE 2 – DMC ou 2SLS sur le modèle pour lbwght

III Test d'endogénéité : Eléments théoriques

Source : Wooldridge 2006, p 532.

Les 2SLS sont moins efficaces que les OLS quand les variables explicatives sont toutes exogènes. Donc si les variables explicatives sont toutes exogènes, il faut utiliser les OLS. On dispose d'un test d'endogénéité qui permet de savoir si les 2SLS sont nécessaires. L'idée du test est que la différence entre les estimateurs des OLS et des 2SLS doit être faible si la variable explicative est exogène. Si cette différence est "grande" on conclut que la variable explicative suspectée est endogène. Pour savoir si cette différence est faible, on peut utiliser une regression avec le raisonnement suivant : supposons que nous ayions une seule variable explicative suspectée d'être endogène, notée Y2 dans la suite. Le modèle structurel est le suivant :
$Y1 = \beta_0 + \beta_1 Y2 + \beta_2 X1 + \beta_3 X2 + u_1$.
On dispose de deux variables exogènes supplémentaires, les deux instruments, notés $Z1$ et $Z2$.
La forme réduite s'écrit $Y2 = \pi_0 + \pi_1 X1 + \pi_2 X2 + \pi_3 Z1 + \pi_4 Z2 + u_2$.
Si $Y2$ n'est pas corrélée avec u_1 nous devons utiliser les OLS. Dans quel cas la corrélation entre $Y2$ et U_1 est nulle ? on a $corr(Y2, u_1) = corr(\pi_0 + \pi_1 X1 + \pi_2 X2 + \pi_3 Z1 + \pi_4 Z2 + u_2, u_1) = corr(u_2, u_1)$. Donc la variable Y2 n'est pas corrélée avec le terme d'erreur u_1 si ce terme d'erreur n'est pas corrélée avec u_2. Comment tester la nullité de cette corrélation ? Ecrivons $u_1 = \delta u_2 + \varepsilon$ où ε vérifie toutes les hypothèses des OLS. Si u_1 et u_2 ne sont pas corrélées alors $\delta = 0$. Il suffit donc de remplacer u_1 par l'expression ci dessus dans l'équation structurelle pour obtenir $Y1 = \beta_0 + \beta_1 Y2 + \beta_2 X1 + \beta_3 X2 + \delta u_2 + \varepsilon$. On remplace u_2 par le résidu de la FR et on teste $\delta = 0$.
En pratique on procède donc de la manière suivante :
Etape 1 : on estime la forme réduite par MCO en regressant Y2 sur Z c'est à dire

sur $X1, X2, \ldots Z1, Z2 \ldots Zm$ et on calcule le résidu de cette équation, noté $\hat{u}_2$
Etape 2 : on estime l'équation structurelle par MCO en ajoutant ce résidu comme variable explicative c'est à dire que l'on estime
$Y1 = \beta_0 + \beta_1 Y2 + \beta_2 X1 + \beta_3 X2 + \delta \hat{u}_2 + u_1$
et on teste $\delta = 0$ dans cette équation. Tester cette hypothèse revient à tester l'existence d'un biais d'endogénéité ou de simultanéité. Si on rejette H_0 on conclut que $Y2$ est endogène car u_1 et u_2 sont corrélés et qu'il faut donc utiliser les 2SLS. Dans le cas contraire on utilisera les MCO.
Remarque : les paramètres estimés par MCO dans l'étape 2 sont identiques à ceux estimés par DMC. Nous le verrons dans l'application.

IV Test de restrictions suridentifiantes : Eléments théoriques

Source : Wooldridge 2006, p 533.
On applique ce test si on dispose de plus d'un instrument pour la variable suspectée d'être endogène. On définit le nombre de restrictions suridendifiantes, noté q, comme le nombre total d'instruments moins le nombre de variable explicative suspectée d'être endogène.
Nous avons vu qu'une variable instrumentale doit respectée deux conditions :
- Elle doit être corrélée avec la variable explicative endogene : nous avons déjà testé cette condition
- Elle doit être exogène et donc non corrélée avec le terme d'erreur. Pour tester cette seconde condition nous allons utliser le test de Sargan (1958) :
Etape 1 : On enregistre les résidus des DMC.
Etape 2 : On regresse ce résidu sur toutes les variables exogènes du modèle (les variables X et les instruments). On effectue un test du multiplicateur de Lagrange sur cette regression auxiliaire en calculant $N \times R^2$ que l'on compare une loi du χ^2 à q degré de liberté , où q est le nombre de restrictions suridentifiantes. Si la valeur observée dépasse la valeur théorique de la table , on rejette $H0$: les instruments sont exogènes et donc au moins une variable instrumentale n'est pas exogène.
Nous reviendrons sur ce test dans l'application.

V Application : Le rendement de l'éducation

Nous disposons d'une extrait de l'enquête Formation Qualification Professionnelle (FQP) de l'INSEE pour 1993 dans le fichier de données fqp au format R. Nous disposons pour 500 salariés des variables suivantes :
salaire : salaire mensuel en francs
etudes : nombre d'années d'étude (à partir de 6 ans où la scolarité est obligatoire en France)
sexe : codé 1 pour les hommes et 2 pour les femmes

CSP : CSP du père du salarié, codée 1 pour cadre supérieur, 2 pour cadre moyen et 3 pour employés ou ouvriers
Frater : nombre de frères et soeurs du salarié

1. Spécification du modèle : il faut tout d'abord recoder la variable sexe en une variable indicatrice que nous appellerons s. Puis nous créons la variable log(salaire) ;en effet dans la spécification en log pour le salaire, le paramètre de la variable “etudes" mesure directement le rendement de l'éducation . Le modèle est le suivant :

$$log(salaire) = \beta_0 + \beta_1 Etudes + \beta_2 S + u_i$$

β_1 est la variation du salaire en pourcentage quand on augmente le nombre d'années d'études de 1 an. Le paramètre β_2 représente la différence approximative en pourcentage entre le salaire des hommes et celui des femmes si on code “s=1" pour les hommes et “s=0" pour les femmes. Le programme SAS est le suivant :

```
data tpfoad.fqp ;
set tpfoad.efqp ;
if sexe=1 then s=1 ;*H ;
if sexe=2 then s=0 ;*F ;
lnsal=log(salaire) ;
run ;
proc reg data=tpfoad.fqp ;
model lnsal=etudes s ;
run ;
```

Avec R :

```
ols <- lm(LNSAL~ETUDES+S, data=fqp)
summary(ols)
```

Quand on estime le modèle par MCO on obtient un rendement de l'éducation de 7,46%. La différence entre le salaire des hommes et celui des femmes est d'environ 20%.
Dans cette regression, on soupçonne la variable “etudes" d'être corrélée avec le terme d'erreur ; en effet la valeur de la variable étude résulte d'un calcul économique de la part de l'individu. Dans ce cas les MCO présentés ci-dessus sont biaisés et inconsistants. Nous cherchons donc des variables instrumentales c'est à dire des variables qui peuvent être supposées exogènes (pour l'individu). Nous devons trouver des variables instrumentales qui ne figurent pas dans l'équation du salaire c'est à dire qu'elles ne doivent pas avoir un effet direct sur le salaire. Ces contraintes d'exclusion nous permettent d'identifier les paramètres structurels comme nous l'avons vu dans l'introduction de ce chapitre.
Nous avons déjà indiqué qu'il ne faut pas appliquer deux fois la “proc reg” car les écart-types ne sont pas corrects (voir la formule de la Variance de β_{2SLS}).

Pour vérifier qu'en appliquant deux fois la proc reg , on obtient des écart-types estimés différents de ceux des 2SLS, nous donnons ci-dessous les résultats avec la proc reg utilisée deux fois ; nous estimons la FR de etudes avec CSP et FRATER ; nous enregistrons la variable estimée pour Etudes, puis nous estimons l'équation du log du salaire avec cette variable etude estimée et la variable indicatrice s avec le programme suivant :

```
proc reg data=tpfoad.fqp ;
model etudes =s csp frater ;
output out=t p=ethat ;
quit ;run ;
proc reg data=t ;
model lnsal=ethat s ;
quit ;run ;
```

Avec R :

```
ehat1 <- lm(ETUDES~S+FRATER+CSP, data =fqp)
summary(ehat1)
ehat2 <- lm(LNSAL~fitted(ehat1)+S, data=fqp)
summary(ehat2)
```

Nous obtenons les résultats présentés dans la figure 3.
Le rendement estimé en appliquant deux fois la proc reg est égal à 7.415% et

Number of Observations Read 500
Number of Observations Used 500

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	57.77270	28.88635	958.59	<.0001
Error	497	14.97675	0.03013		
Corrected Total	499	72.74945			

Root MSE	0.17359	R-Square	0.7941
Dependent Mean	9.16418	Adj R-Sq	0.7933
Coeff Var	1.89425		

Résultats estimés des paramètres

Variable	Libellé	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	Intercept	1	8.13637	0.03478	233.95	<.0001
ethat	Predicted Value of ETUDES	1	0.07415	0.00350	21.17	<.0001
s		1	0.20807	0.02420	8.60	<.0001

FIGURE 3 – Regression : appliquer 2 fois la proc reg

son écart-type estimé est de 0.0035.
Nous allons maintenant présenter les 2SLS et nous pourrons ainsi constater que l'écart-type que nous venons d'obtenir n'est pas correct. Nous verrons que l'écart-type de rendement avec la méthode des 2SLS est égal à 0.00256.

2. Estimation par DMC
   Nous choisissons la CSP du père du salarié et son nombre de frères et soeurs comme variables instrumentales. Elles peuvent être supposées exogènes pour

le salarié.
Nous allons vérifier qu'elles sont corrélées avec la variable “etudes" en faisant une regression sur la FR :

```
proc reg data=tpfoad.fqp;
model etudes =s frater csp;
test frater=0,csp=0;
run;
```

Avec R :

```
etudes1 <- lm(ETUDES~S+FRATER+CSP, data =fqp)
summary(etudes1)
etudes2 <- lm(ETUDES~S, data =fqp)
anova(etudes2,etudes1)
```

Les résultats sont présentés dans la figure 4.

Dependent Variable: ETUDES

Number of Observations Read	500
Number of Observations Used	500

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	3	5886.87540	1962.29180	795.62	<.0001
Error	496	1223.31660	2.46636		
Corrected Total	499	7110.19200			

Root MSE	1.57047	R-Square	0.8279
Dependent Mean	12.29600	Adj R-Sq	0.8269
Coeff Var	12.77217		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	22.52862	0.46756	48.18	<.0001
CSP	1	-3.35424	0.21421	-15.66	<.0001
FRATER	1	-0.43268	0.05401	-8.01	<.0001
s	1	-1.37047	0.25568	-5.36	<.0001

FIGURE 4 – Regression de Etudes

La valeur de la statistique du test de Fisher pour l'hypothèse “CSP=0 et Frater=0" est égale à 497.98 avec une probabilité très faible ; SAS affiche ” p<.0001". Donc les deux variables instrumentales sont corrélées avec la variable Etudes.
Pour estimer le modèle par 2SLS on utilise la proc model ; le programme est le suivant :

```
proc model data=tpfoad.fqp ;
parms b0 b1 b2 a0 a1 a2 a3 ;
endogenous lnsal etudes ;
lnsal =b0 + b1* etudes +b2*s ;
etudes = a0+a1* csp + a2*s +a3*frater ;
fit lnsal etudes/ 2sls ;
instruments s csp frater ;
quit ;run ;
```

On peut aussi mettre seulement “fit lnsal / 2sls ;" si on ne veut pas les résultats de la regression de la FR de la variable Etudes.
Avec R :

```
iv1 <- ivreg(LNSAL~ETUDES+S | CSP+FRATER+S,data=fqp)
summary(iv1)
```

Les résultats sont présentés dans la figure 5.
Par 2SLS, le rendement de l'éducation est de 7.4147% donc à peu près 7.41%

Non linéaire 2SLS Récapitulatif des erreurs résiduelles

Équation	Modèle DF	Erreur DF	SSE	MSE	Racine MSE	R-carré	R carré aj.
lnsal	3	497	8.0037	0.0161	0.1269	0.8900	0.8895

Estimations 2SLS Parameter non linéaires

Parameter	Estimation	Erreur standard appr.	Valeur du test t	Approx Pr > |t|
b0	8.136366	0.0254	320.03	<.0001
b1	0.074147	0.00256	28.96	<.0001
b2	0.208071	0.0177	11.76	<.0001

Nombre d'observations		Statistiques pour Système	
Used	500	Objective	0.000665
Missing	0	Objective*N	0.3327

FIGURE 5 – Regression du Lnsal par DMC

avec un écart-type estimé de 0.00256. Le rendement n'est pas très différent de celui estimé par MCO qui était de 7.46%.
La théorie nous enseigne que la variance de l'estimateur diminue [10] avec le nombre de variables instrumentales. Vérifions cet enseignement théorique sur les données. Supposons que nous disposions seulement de CSP comme variable instrumentale ; dans ce cas nous obtenons un rendement égal à 7.1093 % avec un écart-type estimé de 0.00265. Supposons maintenant que le seul instrument soit Frater ;dans ce cas nous obtenons un rendement égal à 8.0798 % avec un écart-type de 0.00297. Nous constatons donc que plus on a d'instruments plus la précision du rendement est grande (plus son écart-type est faible). Ainsi en

10. jusqu'à un certain point mais en pratique ce point n'est pas atteint car il concerne le cas où le nombre d'instruments est beaucoup plus élevé que le nombre qu'il est possible de trouver.

pratique il est préférable d'estimer un modèle suridentifié plutôt qu'un modèle juste identifié.
Nous résumons les différentes estimations du rendement de l'éducation dans le tableau ci dessous :

Méthode	Rendement	Ecart-type
OLS	7.46	0.00209
2 fois OLS	7.415	0.00350
DMC avec csp,frater	7.4147	0.00256
DMC avec csp	7.1093	0.00265
DMC avec frater	8.0798	0.00297

3. Test d'endogénéité de la variable "etudes"
La manière de procéder pour ce test présenté dans la partie théorique, est assez simple. Il suffit d'exécuter le programme suivant :

```
proc reg data=tpm1.fqp;
model etudes=csp frater s;
output out=sortie r=res;run;
proc reg data=sortie;
model lnsal=etudes s res;
run;
```

Avec R :

```
test <- lm(LNSAL~ETUDES+S+residuals(etudes1),data=fqp)
summary(test)
```

Les résultats sont présentés dans la figure 6.
La variable "résidu " est non significative donc on ne rejette pas H_0 : la

Dependent Variable: lnsal

Number of Observations Read 500
Number of Observations Used 500

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	3	64.74805	21.58268	1337.89	<.0001
Error	496	8.00140	0.01613		
Corrected Total	499	72.74945			

Root MSE	0.12701	R-Square	0.8900
Dependent Mean	9.16418	Adj R-Sq	0.8893
Coeff Var	1.38595		

Résultats estimés des paramètres

Variable	Libellé	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	Intercept	1	8.13637	0.02545	319.75	<.0001
ETUDES		1	0.07415	0.00256	28.93	<.0001
s		1	0.20807	0.01771	11.75	<.0001
res	Residual	1	0.00136	0.00444	0.31	0.7590

FIGURE 6 – Test de l'exogénéité de Etudes

variable Etudes est exogène. Ainsi sur cet échantillon il faut utiliser les OLS pour estimer le rendement de l'éducation.
Les résultats de ce test dépendent du choix des variables instrumentales. Nous allons donc tester la "validité" des instruments dans le paragraphe ci-dessous.

4. Test de restrictions suridentifiantes
   Pour tester l'exogénéité des instruments on utilise le test de Sargan ; on sauve le résidu des 2SLS :

```
proc model data=tpfoad.fqp ;
parms b0 b1 b2 a0 a1 a2 a3 ;
endogenous lnsal etudes ;
lnsal =b0 + b1* etudes +b2*s ;
etudes = a0+a1* csp + a2*s +a3*frater ;
fit lnsal / 2sls out=dd outresid ;
instruments s csp frater ;
quit ;run ;
```

   Aller ouvrir le fichier *dd* et vous constaterez que SAS appelle le résidu du nom de la variable endogène : dans ce fichier *dd* le residu s'appelle "lnsal" (faites la moyenne de cette variable si vous n'êtes pas convaincu). Pour ne pas me tromper je renomme cette variable "res" :

```
data res ;set dd ;
if _type_ ='RESIDUAL' ;rename lnsal=res ;run ;
```

   Nous effectuons le test de Sargan dans le programme ci-dessous :

```
proc reg data=res ;
model res=s csp frater ;run ;
data calcul ;
r2= 0.0416 ;n=500 ;
obs=n*r2 ;
p=1-probchi(obs,1) ;
run ;proc print ;run ;
```

   Avec R :

```
sargan <- lm(residuals(iv1)~S+CSP+FRATER,data=fqp)
summary(sargan)
jour_sargan <- summary(sargan)
class(jour_sargan)
names(jour_sargan)
r2iv <- jour_sargan$r.square
r2iv
obs <- jour_sargan$r.square*500
print(obs)
pvalue <- 1-pchisq(obs,1)
print(pvalue)
```

   Le nombre de degré de liberté de la statistique de χ^2 est le nombre de restriction

suridentifiante : ici nous avons un seul instrument “en plus".
La probabilité associée est égale à 0.000005098 ;elle est très petite on rejette donc H_0 : les variables instrumentales sont exogènes .

Conclusion sur l'application : nous avons conclu dans le test sur l'exogénéité de Etudes que l'on ne rejettait pas l'hypothèse selon laquelle Etudes est exogène. Nous avons aussi signalé que le résultat de ce test dépend du choix des instruments. Nous venons de constater que les instruments ne peuvent pas être supposés exogènes . Il est donc difficile de conclure sur ce fichier de données. Il nous semble que Etudes est plutôt une variable endogène et que peut être le choix des instruments devrait être reconsidéré. Sur le fichier dont nous disposons il n'y a pas d'autres variables instrumentales ; nous nous contenterions de signaler nos doutes sur les instruments et de reporter aussi les résultats des OLS car ils sont sensiblement identiques.

VI Conclusion du Chapitre 3

Dans l'application nous avons constaté que la méthode des variables instrumentales est difficile à appliquer pour au moins deux raisons : la première est que l'économètre doit choisir quelle est la variable potentiellement endogène parmi la liste des variables explicatives. Pour identifier cette variable on doit utiliser la théorie économique. La seconde raison est que cette méthode est difficile à mettre en oeuvre en pratique car il est souvent difficile de trouver des variables instrumentales. De plus leur choix doit être justifié par des tests sur les restrictions suridentifiantes. En pratique pour ce test comme pour beaucoup d'autres, l'économètre doit résouble un problème si le test rejette H_0. Si le test ne rejette pas l'hypothèse nulle, cela ne veut pas dire que nous sommes sûrs que ce problème n'existe pas dans nos données mais que nous ne l'avons pas détecté. Dans ce cas le choix que nous avons fait sur les variables instrumentales a “passé" le test des restrictions suridentifiantes. C'est un “minimum requis".

Pour conclure sur le choix des instruments, la première chose que l'économètre doit faire est, comme toujours, un tour d'horizon de la littérature appliquée au domaine étudié. Ainsi de nombreuses études appliquées existent dans le domaine de l'économie du travail pour estimer le rendement de l'éducation. Certains devoirs associés à ce chapitre utiliseront une base de données très connue sur ce thème.

VII Références bibliographiques

J.D. Sargan , “The estimation of economic relationships using instrumental variables.” Econometrica 26 (1958), pp. 393-415.