

# Appunti di Econometria

## ARGOMENTO [3]: VARIABILI STRUMENTALI

Tommaso Nannicini – Università Bocconi

Novembre 2009

*E ho visto causa ad effetto che si scambiavano il ruolo*  
Lorenzo “Jovanotti” Cherubini, “Un buco nella tasca”, 2005

### 1 Problemi di endogeneità

Se quello “sbruffone” di Archimede chiedeva una leva per sollevare il mondo, l’econometrico si accontenta (il più delle volte) di una variabile strumentale per spiegarlo. Per una serie di problemi che in parte abbiamo anticipato nella trattazione del modello di regressione lineare, è spesso arduo dare un’interpretazione causale alle stime ottenute con i metodi discussi finora (si riveda anche l’argomento 0 di queste dispense). Problemi di **endogeneità** e **causalità inversa** sono endemici in tutti quei contesti empirici in cui è difficile sostenere che la condizione di esogeneità (anche nella sua versione più debole) regga. Gli stimatori OLS diventano quindi inconsistenti. In questo paragrafo, ricorderemo alcuni dei casi più frequenti. Nel prossimo paragrafo, analizzeremo la soluzione a questi problemi offerta dagli stimatori con variabili strumentali.

#### Variabili omesse

Nell’argomento 2 di queste dispense, abbiamo già discusso formalmente il problema delle variabili omesse. Richiamamolo brevemente per il caso bivariato, in cui il vero modello nella popolazione è  $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ , mentre, non osservando  $x_3$ , il modello stimato è  $y = \beta_1^* + \beta_2 x_2^* + \epsilon^*$ . In questo caso:

$$E[\hat{\beta}_2^*] = \beta_2 + \beta_3 \frac{\sigma_{23}}{\sigma_2^2},$$

dove  $\sigma_{23}$  è la covarianza tra il regressore  $x_2$  e la variabile omessa  $x_3$ , e  $\sigma_2^2$  la varianza di  $x_2$ . Inoltre,  $\beta_3$  può essere interpretato come l’*outcome effect* della variabile omessa e  $\sigma_{23}$  come il suo *selection effect*. Abbiamo quindi quattro possibili casi di distorsione da variabile omessa.

	$\sigma_{23} > 0$	$\sigma_{23} < 0$
$\beta_3 > 0$	distorsione positiva	distorsione negativa
$\beta_3 < 0$	distorsione negativa	distorsione positiva

In sintesi, se possiamo ragionevolmente prevedere una distorsione positiva da variabile omessa (come nel caso in cui  $y$  è il salario,  $x_2$  l'istruzione, e  $x_3$  l'istruzione del padre), l'effetto stimato può essere interpretato come un limite superiore (*upper bound*) del vero effetto nella popolazione; viceversa, in presenza di una distorsione negativa, l'effetto stimato ci fornisce un limite inferiore (*lower bound*). Se  $\hat{\beta}_2^* > 0$  e significativo, sapere che possiamo interpretarlo come un *lower bound* significa avere un test (la cui potenza è quindi aumentata) che ci dice che l'effetto di  $x_2$  su  $y$  è senz'altro positivo (anche se non riusciamo a quantificarlo). In maniera simile, se  $\hat{\beta}_2^* < 0$  e significativo, sapere che possiamo interpretarlo come un *upper bound* significa avere un test che ci dice che l'effetto di  $x_2$  su  $y$  è senz'altro negativo.

Prima di passare agli altri casi di endogeneità, si ripassi l'argomento 2 di queste dispense per la formulazione multivariata del problema delle variabili omesse.

### Errori di misura

In alcuni casi, la variabile che crea problemi potrebbe essere osservata, ma con un certo grado di approssimazione. Nel caso di errore di misura della variabile dipendente, gli stimatori OLS sono corretti e consistenti, a patto che l'errore non sia correlato con i regressori. Gli errori del modello stimato sono invece maggiori di quelli del vero modello, e le stime sono meno accurate. Nel caso di errore di misura di una variabile esplicativa, tuttavia, i problemi possono essere ben maggiori. Si consideri il caso che segue. Definiamo l'errore di misura ( $e_x$ ) come la differenza tra la variabile osservata ( $x$ ) e quella vera ( $x^*$ ). Se il vero modello è  $y = \beta_1 + \beta_2 x^* + \epsilon$ , il modello stimato diventa  $y = \beta_1 + \beta_2 x + \eta$ , dove  $\eta = \epsilon - \beta_2 e_x$ . Se l'errore di misura  $e_x$  non è correlato con  $x^*$ , è per definizione correlato con la variabile osservata  $x$ , che sarà quindi correlata con l'errore del modello stimato  $\eta$ . Gli OLS sono quindi inconsistenti, dato che:  $Cov(x, \eta) = -\beta_2 \sigma_e^2$ , dove  $\sigma_e^2$  è la varianza dell'errore di misura. Definendo  $\sigma_{x^*}^2$  come la varianza di  $x^*$ , si può far vedere che:

$$plim(\hat{\beta}_2) = \beta_2 + \frac{Cov(x, \eta)}{Var(x)} = \beta_2 - \frac{\beta_2 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2} = \beta_2 \left( \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} \right) < \beta_2.$$

Di conseguenza, la stima ottenuta sarà sempre più piccola (in valore assoluto) del vero valore di  $\beta_2$ . Si parla in questo caso di *attenuation bias*: se l'effetto è statisticamente significativo, quindi, siamo sicuri della sua direzione, ma non riusciamo a quantificarlo.

### Equazioni simultanee

Un altro esempio classico di endogeneità (forse il più classico, visto che anche la terminologia delle variabili strumentali nasce in questo contesto) è rappresentato dalle equazioni simultanee. Consideriamo il caso della domanda e dell'offerta di lavoro nel settore agricolo. La domanda di lavoro, per esempio a livello provinciale, può essere espressa come:  $L_d = \alpha_1 w + \beta_1 z_1 + \epsilon_1$ , dove  $w$  è il salario,  $z_1$  una **variabile esogena** che influenza la domanda (per esempio, le condizioni metereologiche), ed  $\epsilon_1$  l'errore stocastico. L'offerta di lavoro è invece data da:  $L_s = \alpha_2 w + \beta_2 z_2 + \epsilon_2$ , dove  $z_2$  è una variabile esogena che influenza l'offerta (per esempio, il salario nel settore manifatturiero), ed  $\epsilon_2$  l'errore stocastico. Noi vorremmo stimare separatamente queste due **equazioni strutturali**, per valutare la relazione tra le due **variabili endogene**  $L$  e  $w$ , ma il compito risulta impossibile per il fatto che osserviamo soltanto la quantità di lavoro e il salario d'equilibrio in ogni provincia. È facile far vedere che le due equazioni strutturali di cui sopra ci consegnano questa **equazione in forma ridotta** (dove una variabile endogena viene espressa in funzione delle variabili esogene) per il salario:

$$w = \pi_1 z_1 + \pi_2 z_2 + \eta,$$

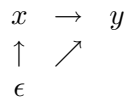
dove  $\pi_j = \beta_j / (\alpha_2 - \alpha_1)$  ed  $\eta = [\epsilon_1 / (\alpha_2 - \alpha_1)] + [\epsilon_2 / (\alpha_2 - \alpha_1)]$ . Ne consegue che:

$$Cov(w, \epsilon_j) = \frac{Var(\epsilon_j)}{(\alpha_2 - \alpha_1)^2} > 0,$$

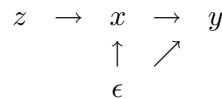
per  $j = 1, 2$ . Quindi, in entrambe le equazioni strutturali, la variabile endogena è correlata con l'errore, e le stime OLS sono inconsistenti. Come vedremo, le variabili esogene  $z_1$  e  $z_2$ , grazie al fatto che ognuna di esse compare in una equazione strutturale ma non nell'altra (**restrizione di esclusione**), possono aiutarci a identificare e stimare i parametri d'interesse  $\alpha_1$  e  $\alpha_2$ .

## 2 Stimatori con variabili strumentali

I problemi appena visti possono essere sintetizzati nel fatto che la variabile esplicativa d'interesse ( $x$ ) è correlata con il termine d'errore ( $\epsilon$ ), ed è per questo definita endogena.



In questi casi, una soluzione all'inconsistenza degli OLS può arrivare dall'esistenza di una o più variabili strumentali ( $z$ ) correlate con il regressore endogeno, ma non con l'errore.



Questo schema contiene tre assunzioni chiave sulle variabili  $z$ : a) che sono correlate con il regressore endogeno  $x$  (**rilevanza**); b) che non sono correlate con l'errore  $\epsilon$  (**esogeneità**); c) che non influenzano direttamente la variabile dipendente  $y$  (**esclusione**). Se la b) e la c) sono verificate (assunzioni che non possiamo testare ma soltanto giustificare o, come vedremo, valutare in forma indiretta), gli strumenti sono **validi**. L'assunzione a) può essere testata direttamente analizzando l'equazione in forma ridotta che esprime la variabile endogena in funzione degli strumenti esogeni: se la correlazione tra  $z$  e  $x$  è bassa, lo strumento è ancora valido ma **debole**. Si veda la Tabella 1 in Angrist e Krueger (2001) per alcuni esempi di variabili strumentali usate (con maggiore o minore successo) in noti studi empirici.

### Caso bivariato

Nel modello di regressione bivariato  $y = \alpha + \beta x + \epsilon$ , se la  $x$  è endogena e una variabile  $z$  soddisfa le tre assunzioni appena viste, abbiamo che:

$$Cov(z, y) = Cov(z, \alpha + \beta x + \epsilon) = \beta Cov(z, x) + Cov(z, \epsilon) \Rightarrow \beta = \frac{Cov(z, y)}{Cov(z, x)} - \frac{Cov(z, \epsilon)}{Cov(z, x)} = \frac{Cov(z, y)}{Cov(z, x)},$$

dove l'ultima eguaglianza deriva dall'esogeneità della  $z$ . Questa relazione ci suggerisce un semplice stimatore per  $\beta$ :

$$\hat{\beta}_{IV} = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (x_i - \bar{x})(z_i - \bar{z})}.$$

Infatti, per la legge dei grandi numeri:

$$plim[\hat{\beta}_{IV}] = plim \left[ \frac{\sum (z_i - \bar{z})(y_i - \bar{y})/N}{\sum (x_i - \bar{x})(z_i - \bar{z})/N} \right] = \frac{Cov(z, y)}{Cov(z, x)} = \beta.$$

Lo stimatore delle variabili strumentali di  $\beta$  è quindi consistente. Lo stesso vale per:  $\hat{\alpha}_{IV} = \bar{y} - \hat{\beta}_{IV}\bar{x}$ . Lo stimatore ha anche un'interpretazione molto intuitiva: si assuma che un aumento di una unità di  $z$  provochi, nello stesso tempo, un aumento di 0.2 anni di studio e di 500 euro di stipendio. Poiché  $z$  non ha effetti diretti sullo stipendio (restrizione di esclusione) possiamo attribuire la variazione di 500 euro all'aumento nell'istruzione indotto da  $z$ . Quindi, l'effetto di un anno in più di istruzione sul salario (secondo questo stimatore delle variabili strumentali fatto a mano) è dato da:  $500/0.2=2,500$  euro.<sup>1</sup>

Nel caso particolare in cui la variabile  $z$  è binaria, lo stimatore delle variabili diventa:

$$\hat{\beta}_{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0},$$

dove  $\bar{y}_j$  è la media di  $y$  per  $z = j$  e  $\bar{x}_j$  è la media di  $x$  per  $z = j$ , con  $j = 0, 1$ . Questa formulazione è chiamata **stimatore di Wald**.

Sotto l'assunzione di omoschedasticità ( $E(\epsilon^2|z) = \sigma^2$ ), la **varianza asintotica** dello stimatore delle variabili strumentali è data da:

$$Var.As.(\hat{\beta}_{IV}) = \frac{\sigma^2}{N\sigma_x^2\rho_{xz}^2},$$

dove  $\sigma_x^2$  è la varianza di  $x$  e  $\rho_{xz}^2$  il coefficiente di correlazione tra  $x$  e  $z$  (nella popolazione) al quadrato. Questa quantità può essere stimata inserendovi gli analoghi campionari: lo stimatore  $s^2$  al posto di  $\sigma^2$ , la varianza campionaria di  $x$  al posto di  $\sigma_x^2$ , e  $R_{xz}^2$  al posto di  $\rho_{xz}^2$ . Questo ci permette di calcolare l'errore standard dello stimatore e applicare le procedure inferenziali che già conosciamo. Un altro aspetto di questa formula merita di essere sottolineato: se  $R_{xz}^2$  ha un valore piccolo (cioè, se lo strumento è debole) la varianza dello stimatore è grande e le stime perdono di accuratezza.

La perdita di accuratezza non è il solo problema degli strumenti deboli. Si consideri la possibile distorsione asintotica di  $\hat{\beta}_{IV}$  nel caso in cui le condizioni di validità non sono rispettate in pieno:

$$plim[\hat{\beta}_{IV}] = \frac{Cov(z, y)}{Cov(z, x)} = \beta + \frac{Cov(z, \epsilon)}{Cov(z, x)} = \beta + \frac{Corr(z, \epsilon)\sigma_\epsilon}{Corr(z, x)\sigma_x},$$

dato che, in generale,  $Cov(z, w) = Corr(z, w)\sigma_z\sigma_w$ . In maniera simile, data la presenza di endogeneità, si può riscrivere la distorsione asintotica di  $\hat{\beta}_{OLS}$  come:

$$plim[\hat{\beta}_{OLS}] = \beta + \frac{Cov(x, \epsilon)}{Var(x)} = \beta + \frac{Corr(x, \epsilon)\sigma_\epsilon}{\sigma_x}.$$

Di conseguenza, abbiamo che:

$$\frac{plim[\hat{\beta}_{IV}] - \beta}{plim[\hat{\beta}_{OLS}] - \beta} = \frac{Corr(z, \epsilon)/Corr(x, \epsilon)}{Corr(z, x)}.$$

Questa formula ci dice che, in presenza di uno **strumento debole**, anche se il problema di endogeneità della  $z$  è molto più piccolo di quello della  $x$ , la distorsione asintotica dello stimatore con variabili strumentali può

<sup>1</sup>È facile far vedere che  $\hat{\beta}_{IV} = (dy/dz)/(dx/dz)$ , sostituendo le stime OLS dei due effetti marginali della  $z$ .

essere maggiore della distorsione OLS. La debolezza dello strumento, quindi, esaspera l'eventuale problema di non validità dello stesso.

### Caso multivariato “appena” identificato

Passiamo adesso al caso multivariato:  $\mathbf{y} = X\beta + \epsilon$ . In particolare, assumiamo che  $k_1$  regressori siano endogeni ( $X_1$ ) e  $k_2$  regressori siano esogeni ( $X_2$ ), con  $k = k_1 + k_2$ :  $X = [X_1 X_2]$ . Assumiamo anche di disporre di  $r$  variabili esogene ( $Z$ ), tra cui includiamo le  $X_2$  più  $r - k_2$  variabili strumentali che soddisfano le condizioni di rilevanza e validità già discusse ( $Z_1$ ):  $Z = [Z_1 X_2]$ . Una prima condizione di identificazione è quella di poter disporre di abbastanza variabili strumentali per quante sono le variabili endogene:  $r \geq k$ , o  $r - k_2 \geq k_1$  (**condizione d'ordine**). Se questa condizione è soddisfatta come eguaglianza abbiamo un numero “appena” sufficiente di variabili strumentali e il modello è detto *just identified*. In questo caso, la semplice estensione dello stimatore delle variabili strumentali visto nel modello bivariato è data da:

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

dove a quella d'ordine dobbiamo aggiungere un'altra condizione per l'identificazione:  $\text{rang}(Z'X) = k$  (**condizione di rango**). Lo stimatore può essere riscritto come:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'(X\beta + \epsilon) = \beta + (Z'X)^{-1}Z'\epsilon.$$

Da cui, visto che stiamo assumendo che  $\text{plim}[(1/N)Z'\epsilon] = \mathbf{0}$  (esogeneità) e  $\text{plim}[(1/N)Z'X] \neq 0$  (rilevanza), si ricava che:  $\text{plim}[\hat{\beta}_{IV}] = \beta$  (consistenza).<sup>2</sup>

Sotto l'assunzione di omoschedasticità, la varianza asintotica dello stimatore delle variabili strumentali è data da:

$$\text{Var.As.}(\hat{\beta}_{IV}) = \sigma^2(Z'X)^{-1}(Z'Z)(Z'X)^{-1}.$$

E possiamo stimarla sostituendo  $\sigma^2$  con  $s^2$ , per poi applicare le procedure inferenziali che già conosciamo. Nel caso di errori non sferici, la formula della varianza asintotica (e quindi il suo stimatore) possono essere modificati tenendo conto che:

$$\text{Var.As.}(\hat{\beta}_{IV}) = \sigma^2(Z'X)^{-1}(Z'\Omega Z)(Z'X)^{-1}.$$

Come nel caso bivariato, strumenti deboli provocano una perdita di accuratezza.

Infine, si può dimostrare (anche se non lo faremo qui) che quanto abbiamo visto per il caso bivariato riguardo al problema d'inconsistenza si estende al caso multivariato: strumenti deboli aumentano l'eventuale distorsione asintotica dello stimatore delle variabili strumentali.

---

<sup>2</sup>Si noti che lo stimatore delle variabili strumentali è consistente, ma distorto in campioni finiti anche se vale  $E(\epsilon|Z) = \mathbf{0}$ . Infatti:

$$E[\hat{\beta}_{IV}] = \beta + E_{Z,X,\epsilon}[(Z'X)^{-1}Z'\epsilon] = \beta + E_{Z,X}[(Z'X)^{-1}Z'E(\epsilon|Z,X)],$$

dove l'ultimo passaggio segue dalla legge delle aspettative iterate. Di conseguenza,  $E[\hat{\beta}_{IV}] = \beta$  soltanto se  $E(\epsilon|Z, X) = \mathbf{0}$ , ma questa assunzione è troppo forte visto che implicherebbe anche che  $E(\epsilon|X) = \mathbf{0}$ , nel qual caso non esisterebbe un problema di endogeneità neanche per gli stimatori OLS.

### Caso multivariato sovraidentificato

Se abbiamo più strumenti che variabili endogene ( $r > k$  o  $r - k_2 > k_1$ ), lo stimatore discusso sopra non ha senso per un problema di conformità. Una soluzione potrebbe essere quella di buttare nel cestino un po' di strumenti per tornare al caso appena identificato. Ma non è mai efficiente rinunciare alle informazioni contenute in alcune variabili. Molto meglio usare una combinazione degli strumenti  $Z$  che ci restituisca una matrice  $\tilde{Z}$  di dimensioni  $[N \times k]$ . Un metodo per farlo è quello dei cosiddetti minimi quadrati a due stati o *two-stage least squares* (**2SLS**). Nel primo stadio, si regredisce ogni variabile  $x_j$  sulle  $Z$  e si ottengono i valori predetti che possono essere organizzati nella matrice  $[N \times k]$ :

$$\hat{X} = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X = P_zX,$$

dove  $P_z$  è la matrice di proiezione, simmetrica ( $P_z' = P_z$ ) e idempotente ( $P_z'P_z = P_z$ ). Nel secondo stadio, si regredisce la variabile  $y$  sui valori predetti  $\hat{X}$  invece che sui regressori osservati  $X$ , ottenendo lo stimatore 2SLS:<sup>3</sup>

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'\mathbf{y} = (X'P_zX)^{-1}X'P_z\mathbf{y}.$$

Un altro modo per interpretare questo stimatore è quello di accorgersi che non è altro che uno stimatore delle variabili strumentali che usa i valori fittati del primo stadio come strumenti:  $\tilde{Z} = P_zX = \hat{X}$ .

Se torniamo al caso in cui  $r = k$ , le matrici  $X'Z$ ,  $Z'Z$  e  $X'X$  sono tutte quadrate di ordine  $[k \times k]$  e invertibili (per le condizioni di rango e di assenza di perfetta multicollinearità). Di conseguenza, è facile far vedere che lo stimatore 2SLS non è altro che lo stimatore IV:

$$\begin{aligned}\hat{\beta}_{2SLS} &= (X'P_zX)^{-1}X'P_z\mathbf{y} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\mathbf{y} \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}(X'Z)(Z'Z)^{-1}Z'\mathbf{y} = (Z'X)^{-1}\mathbf{y} = \hat{\beta}_{IV}.\end{aligned}$$

Sotto l'assunzione di omoschedasticità, la varianza asintotica dello stimatore 2SLS è data da:<sup>4</sup>

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

Sostituendo  $\sigma^2$  con  $s^2$  possiamo stimare la varianza asintotica e l'errore standard dello stimatore, per poi applicare le procedure inferenziali che conosciamo.

### **3 Le variabili strumentali all'opera**

Lo stimatore delle variabili strumentali, come detto, si fonda sull'esogeneità delle variabili strumentali e sulla loro esclusione dall'equazione strutturale. Queste assunzioni non sono testabili, neanche la restrizione di esclusione, dato che anche includendo le  $Z$  nell'equazione strutturale i parametri stimati sarebbero inconsistenti per via dell'endogeneità delle  $X_1$ . La scelta di una variabile strumentale, quindi, deve innanzitutto essere sorretta da solide argomentazioni di carattere teorico o istituzionale. Esistono alcune procedure di test, tuttavia, che possono fornire qualche pezza d'appoggio alle argomentazioni sulla validità degli strumenti usati. La condizione di rilevanza, inoltre, può essere testata direttamente attraverso la stima dell'equazione in forma ridotta, che esprime la variabile endogena in funzione di quelle esogene.

<sup>3</sup>Si noti che, nel modello trasformato  $\mathbf{y} = \hat{X}\beta + (X - \hat{X})\beta + \epsilon = \hat{X}\beta + v$ , i nuovi regressori  $\hat{X}$  sono incorrelati con l'errore  $v$  per via dell'esogeneità delle  $Z$  e della condizione di ortogonalità tra valori predetti  $\hat{X}$  e residui  $(X - \hat{X})$ .

<sup>4</sup>Di nuovo, l'estensione della formula al caso con eteroschedasticità è possibile.

## Test di Hausman

Gli stimatori IV dovrebbero essere usati soltanto quando ce n'è bisogno (cioè, in presenza di endogeneità), altrimenti gli stimatori OLS sono più efficienti. E, come sappiamo, la perdita di efficienza può essere rilevante, specialmente di fronte a strumenti deboli. Sotto l'assunzione che gli strumenti siano validi, il **test di Hausman** può essere usato per testare l'ipotesi nulla di non endogeneità dei regressori,  $H_0 : Cov(x, \epsilon) = 0$ . Infatti, se  $H_0$  fosse vera, gli OLS sarebbero consistenti ed efficienti, gli stimatori IV consistenti ma inefficienti. Se  $H_0$  fosse falsa, gli OLS sarebbero inconsistenti e gli stimatori IV consistenti. Quindi, intuitivamente, se le due stime sono simili, preferiamo gli OLS, e viceversa. Formalmente:

$$H = (\hat{\beta}_{IV} - \hat{\beta}_{OLS})' [Var(\hat{\beta}_{IV}) - Var(\hat{\beta}_{OLS})]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \sim \chi_k^2.$$

Se il valore osservato della statistica è maggiore del valore critico prescelto, abbiamo un problema di endogeneità ed è meglio usare gli stimatori IV. Ma tutto si basa, ricordiamocelo, sull'assunzione che gli strumenti che stiamo usando siano validi.

## Test di sovraidentificazione

Anche se non è testabile nel caso *just identified*, l'assunzione di esogeneità può essere valutata, almeno in parte, nel caso sovraidentificato. Per esempio, se abbiamo due strumenti per un unico regressore endogeno, assumendo che almeno uno dei due sia valido, il coefficiente dell'altro nell'equazione strutturale può essere consistentemente stimato: se il coefficiente è significativo, la restrizione di esclusione per questo secondo strumento non regge. Sviluppando questa intuizione, se  $r > k$ , il **test di sovraidentificazione** (o **test di Sargan**) sottopone a verifica l'ipotesi nulla  $H_0 : Cov(z, \epsilon) = 0$ , sotto l'assunzione (cruciale) che almeno  $k$  strumenti siano validi. La statistica test è pari a:

$$S = \frac{\mathbf{e}' P_z \mathbf{e}}{\mathbf{e}' \mathbf{e} / (N - k)} \sim \chi_{r-k}^2.$$

In pratica, una versione particolare di questo test può essere implementata in tre passaggi: a) si stima l'equazione strutturale con la procedura 2SLS e si ottengono i residui; b) si regrediscono i residui su tutte le variabili esogene e si calcola l' $R^2$ ; c) si sottopone a test  $H_0$  usando la statistica  $N \cdot R^2 \sim \chi_{r-k}^2$ . Se rifiutiamo  $H_0$ , c'è evidenza empirica che alcuni strumenti non sono esogeni. Ma, ricordiamocelo, tutto si basa sull'assunzione che almeno  $k < r$  strumenti siano validi. Meglio, quindi, se gli strumenti che usiamo provengono da "storie" diverse.

## Alla caccia di strumenti deboli

Visti i problemi provocati da strumenti deboli, è importante valutare la dimensione della correlazione tra gli strumenti e i regressori endogeni. L' $R^2$  e il test  $F$  dell'equazione in forma ridotta ( $x_1 = \mathbf{z}_1' \pi_1 + \mathbf{x}_2' \pi_2 + v$ , assumendo che sia presente una singola variabile endogena) non sono del tutto indicativi a riguardo, visto che i risultati potrebbero essere determinati dalle  $X_2$  piuttosto che dalle  $Z_1$ . Meglio usare, allora, il test  $F_p$  parziale ( $H_0 : \pi_1 = \mathbf{0}$ ) con soltanto  $r - k_2$  restrizioni. E l' $R_p^2$  parziale della regressione:  $x_1 - \tilde{x}_1 = (\mathbf{z}_1 - \tilde{\mathbf{z}}_1)' \gamma + \eta$ , dove  $\tilde{x}_1$  sono i valori predetti della regressione di  $x_1$  sulle  $X_2$  e  $\tilde{\mathbf{z}}_1$  i valori predetti delle regressioni di tutte le variabili in  $Z_1$  su  $X_2$  (separatamente). Se l' $R_p^2$  e la statistica  $F_p$  non sono abbastanza grandi, ne concludiamo che gli strumenti contenuti in  $Z_1$  sono deboli.