Variabili strumentali

2 maggio 2001

Indice

1 Introduzione

Come si è visto in precedenza, il metodo OLS fornisce un quadro di riferimento piuttosto completo per trattare il problema dell'inferenza in modelli lineari, ossia in modelli che possono essere scritti nella forma

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u};$$

sotto precise condizioni riguardanti il termine di disturbo $\mathbf{u}$, infatti, è possibile dimostrare che lo stimatore OLS è corretto, consistente e può essere

usato come base per costruire statistiche test tramite le quali è dato vagliare una vasta gamma di ipotesi sul meccanismo probabilistico che pensiamo generi i dati.

Per comprendere la necessità che motiva l'introduzione di una nuova classe di stimatori è necessario considerare con attenzione le condizioni sotto le quali abbiamo sino ad ora analizzato il modello lineare. In particolare, abbiamo analizzato le conseguenze derivanti dallo scomporre il vettore $\mathbf{y}$ in due componenti, ossia

$$\mathbf{y} = E[\mathbf{y}|\mathbf{X}] + \mathbf{u}; \tag{1}$$

in questa sede, il fatto che poi si sia fatta l'ulteriore assunzione secondo cui la media di y_t condizionale ad X_t fosse una funzione lineare è un dettaglio senza importanza. Il dato su cui mettere l'accento è, in questa sede, il fatto che l'equazione (1) definisce implicitamente una serie di proprietà del termine di disturbo u_t. In particolare, la proprietà secondo cui

$$E[\mathbf{u}|\mathbf{X}] = 0,$$

che gioca un ruolo fondamentale in vari contesti — si pensi ad esempio alla dimostrazione della correttezza di $\widehat{\boldsymbol{\beta}}$ — segue banalmente dall'equazione (1) quando si applica l'operatore valore atteso condizionale su entrambi i lati dell'equazione. In altri termini, la possibilità di stimare in modo corretto i parametri della media condizionale di $\mathbf{y}$ rispetto a $\mathbf{X}$ con i minimi quadrati ordinari è *garantita per costruzione*.

Il problema, a volte, nasce da un'altra assunzione fino ad ora implicita: noi abbiamo sempre ipotizzato che i parametri di nostro interesse fossero gli stessi che caratterizzano la media condizionale, o loro funzioni. In certi contesti, non è così. Nei paragrafi che seguono daremo alcuni esempi di situazioni in cui i parametri di nostro interesse sono differenti da quelli della media condizionale, e delineeremo le più diffuse tecniche che vengono usate in questi casi.

2 Esempi

2.1 L'abilità individuale

Facciamo finta, per rendere più semplice l'esempio, che siano definite per ogni individuo laureato tre variabili y_i, x_i e η_i, che chiamiamo rispettivamente "reddito da lavoro", "voto di laurea" e "intelligenza". Sempre per amor di semplicità, supponiamo di considerare i loro scarti dalla media, e che questi ultimi abbiano una distribuzione congiunta normale multivariata come segue:

$$\begin{pmatrix} y_i \\ x_i \\ \eta_i \end{pmatrix} \sim MN\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta\cdot\gamma & \theta \\ \theta\cdot\gamma & 1 & \gamma \\ \theta & \gamma & 1 \end{pmatrix}\right]. \tag{2}$$

Notate che, in questo esempio, $\text{Cov}(y_i, x_i) = \text{Cov}(y_i, \eta_i)\,\text{Cov}(\eta_i, x_i)$.

Consideriamo la distribuzione di y_i condizionale a x_i ed η_i; dopo un paio di passaggi semplici anche se non terribilmente eccitanti[1] , si ricava che

$$y_i|x_i, \eta_i \sim N\left[\theta\eta_i, 1-\theta^2\right]. \tag{3}$$

Dalla distribuzione condizionale ricaviamo una conclusione: nel mondo che stiamo immaginando un alto voto di laurea non fa guadagnare di più (perché x_t è assente dalla media condizionale di y_t), ma l'intelligenza sì (se θ è positivo)[2]; tuttavia, essere intelligenti porta vantaggi anche nel voto di laurea (e questo lo si vede dalla distribuzione non condizionale, sempreché γ sia positivo).

Apparentemente, una conclusione diversa la si sarebbe potuta trarre considerando la distribuzione di y_i condizionale a x_i,

$$y_i|x_i \sim N\left[\theta\gamma x_i, 1-\theta^2\gamma^2\right], \tag{4}$$

da cui risulta che effettivamente un buon voto di laurea paga in termini di reddito. È importante notare, in questo contesto, che i due risultati sono contraddittori solo in apparenza, poiché sono ottenuti condizionando la variabile "reddito" rispetto a due set informativi diversi.

Poniamo adesso la questione sotto la forma di un modello lineare: usando l'equazione (3) si potrebbe scrivere

$$y_i = \lambda \cdot x_i + \theta \cdot \eta_i + \varepsilon_i, \tag{5}$$

in cui i primi due termini a destra del segno di uguale rappresentano la media di y_i condizionale a x_i e η_i, mentre ε_i è un residuo la cui media condizionale alle stesse variabili è, *per costruzione*, 0. La stima di questa equazione con gli OLS produrrebbe dei valori con tutte le buone proprietà analizzate in precedenza, e probabilmente sarebbe anche possibile accettare l'ipotesi nulla secondo cui λ, il coefficiente associato a x_i, è zero (come in effetti è).

Cosa succederebbe, però, se l'intelligenza non fosse misurabile? In effetti, il reddito e il voto di laurea sono dati oggettivi e quantificabili, mentre l'intelligenza è una quantità più sfuggente, con buona pace dei test che

[1] Ricordo brevemente la regola di condizionamento in una normale multivariata: se

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim MN\left[\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_Y & K' \\ K & \Sigma_X \end{pmatrix}\right],$$

allora

$$Y|X \sim MN\left[\mu_Y + K'\Sigma_X^{-1}(X-\mu_X), \Sigma_Y - K'\Sigma_X^{-1}K\right].$$

[2] Forse questo paragrafo si sarebbe potuto intitolare "La rivolta degli umili".

popolano le riviste da ombrellone. Se l'intelligenza non è misurabile, niente paura: utilizziamo l'equazione (4) e scriviamo un modello del tipo

$$y_i = \beta \cdot x_i + u_i, \tag{6}$$

in cui βx_i è la media condizionale di y rispetto a x: il parametro β è semplicemente dato da $\theta \cdot \gamma$. Stimando questo modello, otteniamo un valore stimato per β che ha le solite amichevoli proprietà, e se il campione è abbastanza grande, dovremmo anche essere in grado di rifiutare, con apposito test, l'ipotesi $\beta = 0$.

Dovremmo forse concludere che l'influsso del voto di laurea dipende dall'osservabilità dell'intelligenza? Certamente, questa è una conclusione che ripugna alla logica. L'intero argomento potrebbe essere usato come esempio dei danni che si ottengono con l'"omissione di variabili rilevanti" dalla specificazione, ma questo sarebbe un modo di liquidare la questione poco interessante e, ai nostri fini, quasi fuorviante. In effetti, la statistica

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

fa né più né meno di quel che le chiediamo di fare, cioè stimare il parametro della media condizionale di y rispetto a x e, a onor del vero, lo fa egregiamente. Il punto è che a noi piacerebbe una statistica che ci informasse del fatto che, sebbene ci sia una certa correlazione fra voto e reddito, non è per merito (o colpa) del voto che si guadagna di più (o di meno). In altri termini, noi vorremmo un modello che potessimo scrivere come

$$y_i = \lambda \cdot x_i + v_i, \tag{7}$$

in cui $\lambda = 0$, per poi poterlo stimare e farci sopra test di ipotesi. Ma attenzione, un modello così lo abbiamo già, ed è dato dall'equazione (5). Basta riscriverla così:

$$y_i = \lambda \cdot x_i + (\theta \cdot \eta_i + \varepsilon_i),$$

e definire v_i come $\theta\eta_i + \varepsilon_i$. La differenza fondamentale fra questa formulazione e quelle precedenti sta nel fatto che nell'equazione (7) la variabile dipendente non viene scomposta in "media condizionale" e "residuo", ma come somma di due componenti diverse, in cui la parte che andiamo a stimare non è la media condizionale, ma è una funzione di x che è il nostro vero oggetto di interesse.

La conseguenza ovvia è che, se la parte sistematica non è più la media condizionale, la media condizionale del residuo non è più zero; infatti, si può far vedere facilmente che

$$E\left[v_i | x_i\right] = \theta\gamma.$$

Nei libri di econometria un po' *passé*, questo veniva chiamato il problema della "correlazione fra regressori e disturbi". Nei libri più moderni, si dice che "la variabile x non è debolmente esogena per il parametro di interesse λ". Di fatto, la situazione è sempre la stessa: non è possibile fare inferenza sul parametro di nostro interesse partendo da una stima, buona quanto si vuole, dei parametri della media condizionale.

2.2 L'errore di misura

Questo esempio è un classico della didattica econometrica, e può essere ben illustrato con un esempio tratto da una controversia macroeconomica risalente agli anni '50 che riguardava la funzione del consumo. La corrente di pensiero allora dominante si rifaceva ad un passo della "Teoria Generale" di Keynes, in cui veniva sostenuta l'ipotesi di una "legge psicologica fondamentale" in base alla quale non tutto il reddito disponibile viene speso, ma solo una sua frazione. Poiché però anche chi non ha reddito deve pur consumare qualche cosa per la sopravvivenza, se ne deduce che la relazione fra reddito e consumo deve avere la forma

$$C = C_0 + cY.$$

Come è noto, il parametro c prende il nome di "propensione marginale al consumo", e gioca un ruolo fondamentale in tutta la costruzione teorica keynesiana. Milton Friedman dissentiva da questa ipotesi argomentando che il reddito, di per sé, non è di nessun uso se non speso per consumi, e di conseguenza la propensione marginale al consumo non poteva che essere 1. Si pensò, nell'epoca della adolescenza dell'econometria[3], di poter dirimere la controversia misurando la propensione marginale al consumo con un modello lineare. In una serie di stime compiute in modo indipendente su basi di dati relativi ad economie diverse, emerse con una certa regolarità un valore stimato di c inferiore ad 1.

La contro-argomentazione di Friedman fu basata sull'idea di "reddito permanente": nella teoria omonima, i consumi non sono funzione del reddito corrente, ma di quello permanente, che è una sorta di media ponderata di tutti i redditi presenti e futuri; in quanto tale, esso può essere benissimo diverso dal reddito corrente. Analizziamo la situazione in modo più formalizzato: secondo il pensiero di Friedman[4], il legame fra reddito e consumo può essere illustrato dalle due seguenti equazioni:

$$c_t = \beta y_t^* + \varepsilon_t \tag{8}$$

$$y_t = y_t^* + \eta_t, \tag{9}$$

[3]Oggi la controversia non si pone più in questi termini, e comunque se anche i termini fossero questi, si dovrebbe far ricorso a strumenti empirici ben più sofisticati.

[4]Non è vero: quella che segue è una esposizione del pensiero di Friedman ignobilmente stilizzata. D'altro canto, non è peggiore di quanto viene comunemente attribuito a Friedman nei testi di macroeconomia.

dove c_t è il consumo al tempo t, y_t è il reddito corrente e y_t^* è il reddito permanente. La variabile ε_t è il consueto termine di disturbo, mentre il termine η_t riflette la differenza fra reddito corrente e permanente, che possiamo tranquillamente immaginare come una variabile casuale a media 0, varianza costante σ_η^2, indipendente sia da y_t^* che da ε_t. Il punto centrale della teoria del reddito permanente è che $\beta = 1$.

Poiché il reddito permanente è inosservabile, le stime effettuate usano come variabile esplicativa il reddito corrente. Tuttavia, questo conduce ad utilizzare un modello lineare che, combinando le equazioni (8) e (9), si può scrivere come

$$c_t = \beta y_t + (\varepsilon_t - \beta\eta_t) = \beta y_t + u_t. \tag{10}$$

Va notato che, nella formulazione dell'equazione (10), c'è per costruzione una correlazione negativa fra y_t e il termine di disturbo. Infatti, una delle componenti di u_t è η_t, che è correlata a y_t attraverso l'equazione (9). Moltiplicando tale equazione da ambo i lati per η_t, ed applicando l'operatore valore atteso al risultato, si vede facilmente che

$$\mathrm{Cov}\,(y_t, \eta_t) = \sigma_\eta^2.$$

Questo risultato ci consente di calcolare il limite in probabilità dallo stimatore OLS di β facendo ricorso ad una semplice versione della legge dei grandi numeri: poiché

$$\hat{\beta} = \frac{\sum y_t c_t}{\sum y_t^2},$$

sostituendo a c_t la sua espressione data dalla equazione (10) si ha

$$\hat{\beta} = \beta + \frac{1/T \sum y_t \varepsilon_t}{1/T \sum y_t^2} - \frac{1/T \sum y_t \eta_t}{1/T \sum y_t^2} \tag{11}$$

Assumendo che $1/T \sum y_t^2 \xrightarrow{\mathrm{p}} Q$, dove Q è un qualche numero positivo (non altro che il momento secondo di y_t), si ha che

$$\frac{1}{T}\sum y_t \varepsilon_t \xrightarrow{\mathrm{p}} 0 \quad \text{e} \quad \frac{1}{T}\sum y_t \eta_t \xrightarrow{\mathrm{p}} \sigma_\eta^2,$$

in quanto le variabili casuali che si ottengono moltiplicando y_t per ε_t ed η_t soddisfano alle condizioni per l'applicazione della legge dei grandi numeri, e di conseguenza le loro medie aritmetiche convergono in probabilità ai loro valori attesi. Per le regole di composizione dei limiti in probabilità si ha che

$$\hat{\beta} \xrightarrow{\mathrm{p}} \beta\left(1 - \frac{\sigma_\eta^2}{Q}\right), \tag{12}$$

e quindi, anche se il vero valore di β fosse 1, la statistica $\hat{\beta}$ convergerebbe in probabilità ad un valore inferiore ad 1 per costruzione[5], in quanto vige

[5]Questo fenomeno viene detto in alcuni testi "*attenuation*".

la relazione $0 < \sigma_\eta^2 < Q$. Pertanto, Friedman non ritenne che l'evidenza empirica presentata fosse sufficiente a chiudere il dibattito.

Dovrebbe essere evidente che l'esempio qui riportato è generalizzabile con facilità irrisoria a tutte le situazioni nelle quali utilizziamo come variabile esplicativa un misuratore imperfetto di una variabile teorica. Volendo riproporre il problema nei termini utilizzati nell'esempio precedente, si può dire che la stima OLS misura con la massima precisione possibile il parametro della media condizionale rispetto alla variabile "sbagliata", o per meglio dire misura un parametro che non è riconducibile al nostro parametro di interesse teorico.

2.3 I sistemi di equazioni simultanee

Un altro esempio che si può fare riguarda i sistemi di equazioni simultanee. Le procedure inferenziali che riguardano sistemi di equazioni, anziché equazioni singole, hanno una storia antica e venerabile, poiché nei primi anni della pratica econometrica questo era il terreno su cui si orientava la punta di diamante della ricerca. Qui non ne parleremo. Ci basterà accennare al fatto che una delle caratteristiche che rendono i modelli a più equazioni insidiosi da stimare possono essere descritte con gli strumenti adoperati fin qui.

Consideriamo uno dei più semplici modelli di equazioni simultanee che si possano concepire: un modello microeconomico di domanda e offerta di un bene. Un modello del genere si compone naturalmente di due equazioni (domanda e offerta, appunto):

$$q_t = \alpha_0 - \alpha_1 p_t + u_t \tag{13}$$
$$p_t = \beta_0 + \beta_1 q_t + v_t, \tag{14}$$

dove l'equazione (13) è la funzione di domanda, l'equazione (14) è la funzione di offerta, e le variabili hanno i seguenti significati:

q_t	quantità scambiata del bene al tempo t
p_t	prezzo del bene al tempo t
u_t	termine di disturbo della funzione di domanda
v_t	termine di disturbo della funzione di offerta

Se le due equazioni fossero considerate singolarmente, si potrebbe pensare di stimare i vettori di parametri che le caratterizzano col metodo OLS. In realtà, non è così, in quanto quella che noi vogliamo poter considerare la parte sistematica, e cioè tutto quel che sta a destra del segno di uguale salvo il termine di disturbo, risulta essere diversa dalla media condizionale.

La dimostrazione è piuttosto semplice. Prendiamo la funzione di domanda (13): se effettivamente l'espressione $(\alpha_0 - \alpha_1 p_t)$ fosse la media di q_t

condizionale a p_t, ne deriverebbe che il valore atteso condizionale di u_t dovrebbe essere 0. Ma se prendiamo la funzione di offerta (14) e sostituiamo a q_t la sua definizione data dalla (13) otteniamo, con alcuni semplici passaggi,

$$\begin{aligned} p_t &= \beta_0 + \beta_1(\alpha_0 - \alpha_1 p_t + u_t) + v_t \\ &= (\beta_0 + \beta_1\alpha_0) - (\beta_1\alpha_1)p_t + (v_t + \beta_1 u_t) \Rightarrow \\ (1 + \beta_1\alpha_1)p_t &= (\beta_0 + \beta_1\alpha_0) + (v_t + \beta_1 u_t) \Rightarrow \\ p_t &= \pi_1 + \eta_t, \end{aligned}$$

dove il parametro π_1 è definito come $\frac{\beta_0+\beta_1\alpha_0}{1+\beta_1\alpha_1}$. È importante notare che il nuovo termine di disturbo η_t è definito come $v_t+\beta_1 u_t$, e quindi la covarianza fra p_t e u_t non è nulla[6]. Di conseguenza:

- Se $\mathrm{Cov}\,(p_t, u_t) \neq 0$, allora
- $E\,[u_t|p_t]$ non può essere 0, perciò
- $(\alpha_0 - \alpha_1 p_t)$ non può essere $E\,[q_t|p_t]$, e quindi
- non c'è speranza che lo stimatore OLS applicato alla (13) ci fornisca i risultati desiderati.

L'argomento può essere generalizzato con facilità, ma questo ci costringerebbe ad una disamina meno superficiale del trattamento statistico dei sistemi di equazioni simultanee, ciò che si farà in un'altra sede.

3 Lo stimatore GIVE

Prendiamo in esame un modello lineare standard: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. Come abbiamo visto, l'assunzione che $\mathbf{X}\boldsymbol{\beta} = E\,[\mathbf{y}|\mathbf{X}]$ è cruciale per stimare in modo consistente il vettore di parametri $\boldsymbol{\beta}$; anzi, ci si può spingere più in là e dire che questa assunzione *definisce* il vettore $\boldsymbol{\beta}$, nel senso che il vettore $\boldsymbol{\beta}$ è il vettore che rende vera la relazione

$$E\left[\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = 0. \tag{15}$$

Lo stimatore OLS $\hat{\boldsymbol{\beta}}$ è invece implicitamente definito dalla relazione

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0. \tag{16}$$

L'equazione (16) corrisponde alle condizioni di primo ordine per la minimizzazione della funzione obiettivo che dà il nome alla statistica OLS, ma può anche essere vista come il corrispondente campionario della (15). Non

[6]Dimostrazione lampo: $u_t p_t = \pi_1 u_t + u_t\eta_t$. Se u_t ha media 0, il valore atteso dell'espressione precedente è la covarianza fra u_t e p_t. Ma $E\,[u_t\eta_t] = \mathrm{Cov}\,(v_t, u_t) + \beta_1 V(u_t)$ è evidentemente diverso da 0, e da questo segue il risultato.

sorprende, pertanto, che la statistica OLS funzioni bene come stimatore del suo corrispondente teorico.

Se però il vettore $\boldsymbol{\beta}$, che contiene i parametri di nostro interesse, non è definito dall'equazione (15) ma da qualche altra proprietà, allora il problema della stima può essere affrontato per analogia, definendo uno stimatore $\widetilde{\boldsymbol{\beta}}$ come quel vettore che soddisfa l'equivalente campionario di una proprietà che vale per $\boldsymbol{\beta}$.

Il metodo di cui ci occupiamo in questo capitolo viene usato nelle situazioni in cui si ipotizza che esista un certo numero di variabili osservabili, che raggruppiamo nella matrice $\mathbf{W}$, per le quali valga la relazione

$$E\left[\mathbf{W}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = 0. \tag{17}$$

Lo stimatore corrispondente sarà allora quel vettore $\widetilde{\boldsymbol{\beta}}$ per cui vale

$$\mathbf{W}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) = 0. \tag{18}$$

Le variabili $\mathbf{W}$ prendono il nome di *strumenti*, o più precisamente *variabili strumentali*.

In molte situazioni, fra cui quelle che abbiamo citato a titolo d'esempio nel paragrafo precedente, è verosimile che variabili strumentali siano non solo disponibili, ma in qualche misura suggerite dalla natura stessa del problema. Per rendere l'esposizione più scorrevole, tuttavia, non approfondiremo subito questo punto: dapprima analizzeremo le proprietà dello stimatore GIVE, indicando solo in seguito quali siano i requisiti che una variabile deve soddisfare perché possa essere usata come strumento.

3.1 Definizione dello stimatore

Lo stimatore GIVE (*Generalized Instrumental Variables Estimator*) è una statistica così definita:

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y}. \tag{19}$$

La sua derivazione può essere illustrata come segue: come abbiamo detto, ci si trova a volte in situazioni nelle quali

$$E\left[\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \neq 0,$$

poiché il vettore di parametri di interesse $\boldsymbol{\beta}$ è diverso da quello che definisce la media condizionale. Se tuttavia esiste una matrice $\mathbf{W}$, che per il momento supponiamo abbia la stessa dimensione di $\mathbf{X}$, che soddisfa la relazione

$$E\left[\mathbf{W}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = 0,$$

allora si potrebbe pensare di costruire una statistica $\widetilde{\boldsymbol{\beta}}$ che rispetti l'uguaglianza

$$\mathbf{W}'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) = 0 \qquad \Longrightarrow \qquad \mathbf{W}'\mathbf{X}\widetilde{\boldsymbol{\beta}} = \mathbf{W}'\mathbf{y}.$$

Poiché il numero di colonne di $\mathbf{W}$ è uguale a quello di $\mathbf{X}$, si vede immediatamente che, se la matrice $\mathbf{W}'\mathbf{X}$ è invertibile, la statistica $\widetilde{\boldsymbol{\beta}}$ risulta definita da

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y}, \tag{20}$$

e prende il nome di stimatore IV (*Instrumental Variables*). Il requisito dell'invertibilità esclude naturalmente che lo stimatore sia definito quando il numero di strumenti (le colonne di $\mathbf{W}$) è inferiore al numero di regressori (le colonne di $\mathbf{X}$). Ma cosa accade nel caso opposto, in cui ci sono più strumenti che regressori? La questione può essere impostata in modo non dissimile da un semplice problema di stima di un modello lineare.

Supponiamo quindi che $\mathbf{X}$ sia una matrice $(T \times k)$ e $\mathbf{W}$ sia una matrice $(T \times m)$, con $m > k$. Supponiamo inoltre, per semplicità, che

$$E\left[\mathbf{u}\mathbf{u}'|\mathbf{W}\right] = \sigma^2\mathbf{I},$$

dove $\mathbf{u}$ è definito come $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Ricordo che $E\left[\mathbf{u}|\mathbf{W}\right] = 0$ per ipotesi. Per le proprietà del valore atteso potremo dunque scrivere

$$E\left[\mathbf{W}'\mathbf{u}\mathbf{u}'\mathbf{W}|\mathbf{W}\right] = \sigma^2\mathbf{W}'\mathbf{W} = \sigma^2\Omega.$$

Poiché Ω è una matrice di varianze-covarianze, dev'essere positiva definita, così come la sua inversa, e deve quindi esistere una matrice quadrata K tale per cui $K'\Omega K = \mathbf{I}$, per cui vale anche la relazione $KK' = \Omega^{-1}$. Come sia fatta questa matrice K non importa: ci basta sapere che esiste. Potremo quindi scrivere

$$E\left[K'\mathbf{W}'\mathbf{u}\mathbf{u}'\mathbf{W}K|\mathbf{W}\right] = E\left[\mathbf{e}\mathbf{e}'|\mathbf{W}\right] = \sigma^2\mathbf{I},$$

dove $\mathbf{e}$ è un vettore di m elementi definito come $K'\mathbf{W}'\mathbf{u}$. La definizione di $\mathbf{e}$, tuttavia, fa sì che si possa anche scrivere l'equazione

$$v = C\boldsymbol{\beta} + \mathbf{e}, \tag{21}$$

dove le grandezze v e C sono definite come

$$\underset{m\times 1}{v} = K'\mathbf{W}'\mathbf{y}$$

e

$$\underset{m\times k}{C} = K'\mathbf{W}'\mathbf{X}.$$

L'equazione (21) può essere letta come un modello lineare in cui il termine di disturbo è a media nulla, omoschedastico e serialmente incorrelato. In questo particolare modello lineare, il numero di "regressori" è k, ma il numero di "osservazioni" è m. Se tuttavia, come abbiamo ipotizzato, $m > k$, allora calcolando la statistica OLS per il modello dell'equazione (21) si ha

$$\widetilde{\boldsymbol{\beta}} = (C'C)^{-1}C'v = (\mathbf{X}'\mathbf{W}'KK'\mathbf{W}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}'KK'\mathbf{W}'\mathbf{y};$$

poiché $KK' = \mathbf{W}'\mathbf{W}^{-1}$, ecco che si riottiene la (19), che riportiamo qui per completezza:

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y}.$$

È un semplice esercizio di algebra lineare dimostrare che la (20) si può ottenere come caso particolare della (19) quando $m = k$. In questo ultimo caso, il modello si dice *esattamente identificato*, in quanto la stima dei parametri si basa su un numero di statistiche che è uguale al numero dei parametri stessi, nel senso che il sistema di equazioni

$$\mathbf{W}'\mathbf{X}\widetilde{\boldsymbol{\beta}} = \mathbf{W}'\mathbf{y},$$

che definisce implicitamente lo stimatore, è un sistema di k equazioni in k incognite. Nel caso invece in cui $m > k$, non è detto che esista un vettore per cui la precedente relazione sia vera, poiché abbiamo un numero di equazioni m che è maggiore del numero di incognite. Come abbiamo visto, il problema può essere risolto ri-esprimendolo come un problema di minimi quadrati, ma ciò non toglie che, volendo, potremmo buttar via un numero $(m - k)$ di equazioni e uno stimatore lo troveremmo lo stesso. In questo caso, si dice che il sistema è *sovraidentificato* e il numero $(m - k)$ si chiama *rango di sovraidentificazione*.

3.2 Interpretazione dello stimatore

Il vantaggio di avere derivato lo stimatore GIVE come stimatore OLS di un modello trasformato (nel caso di sovraidentificazione) ci è anche utile perché ci consente di considerare lo stimatore GIVE come soluzione di un problema di ottimo. In effetti, si vede facilmente che lo stimatore GIVE può essere definito come

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\text{Argmin}}\ \mathbf{u}(\boldsymbol{\beta})'\mathbf{P}_{\mathbf{W}}\mathbf{u}(\boldsymbol{\beta}), \tag{22}$$

ossia come quel $\boldsymbol{\beta}$ che minimizza la somma dei quadrati dei residui in (21). Nel caso esattamente identificato, il minimo della funzione obiettivo vale esattamente 0, poiché $\widetilde{\boldsymbol{\beta}}$ è appunto quel vettore che provoca $\mathbf{W}'\widetilde{\mathbf{u}} = 0$.

Naturalmente, come sottoprodotto della stima di $\boldsymbol{\beta}$ si ottiene anche un vettore di residui $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}$, che può essere usato a sua volta per costruire uno stimatore della varianza

$$\tilde{\sigma}^2 = \frac{\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}}}{T}.$$

Il problema della stima puntuale è così risolto. Come vedremo nel sottoparagrafo successivo, si può dimostrare che sotto le ipotesi di partenza gli stimatori $\widetilde{\boldsymbol{\beta}}$ e $\tilde{\sigma}^2$ sono consistenti. In più, $\widetilde{\boldsymbol{\beta}}$ è anche asintoticamente normale, cosa che ci consente di procedere nel modo consueto ed utilizzare l'approssimazione asintotica per interpretare le opportune statistiche test.

Lo stimatore GIVE viene a volte anche chiamato stimatore "a due stadi", o stimatore 2SLS (*2 Stages Least Squares*). Il motivo per cui questo accade ha perlopiù un interesse storico, ma non solo. Se si considera l'equazione (19), si ha che lo stimatore GIVE può anche essere scritto

$$\widetilde{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \tag{23}$$

dove $\hat{\mathbf{X}} = \mathbf{P}_{\mathbf{W}}\mathbf{X}$. In altri termini, la matrice $\hat{\mathbf{X}}$ contiene, per ogni colonna, il valore fittato che si otterrebbe regredendo la corrispondente colonna di $\mathbf{X}$ sulla matrice $\mathbf{W}$ (primo stadio). Lo stimatore $\widetilde{\boldsymbol{\beta}}$ può poi essere calcolato semplicemente regredendo $\mathbf{y}$ su $\hat{\mathbf{X}}$ (secondo stadio)[7]. Questa procedura, che è effettivamente piuttosto macchinosa, era però la più semplice dal punto di vista computazionale quando gli elaboratori erano rari e poco potenti. In effetti, utilizzando questa procedura non c'è bisogno di una routine apposita per il calcolo dello stimatore GIVE, ma questo viene ricondotto a $k + 1$ regressioni OLS.

Questo modo di vedere lo stimatore GIVE dà anche adito ad una sua interpretazione intuitiva: poiché $\hat{\mathbf{X}}'\hat{\mathbf{X}} = \hat{\mathbf{X}}'\mathbf{X}$, allora $\widetilde{\boldsymbol{\beta}}$ è anche uguale a

$$(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y},$$

e quindi lo stimatore GIVE è, in realtà, uno stimatore IV in cui gli strumenti ($\hat{\mathbf{X}}$) sono le combinazioni lineari degli strumenti $\mathbf{W}$ che riproducono al meglio il contenuto informativo delle variabili $\mathbf{X}$. In questo senso, si può dire che $\mathbf{P}_{\mathbf{W}}\mathbf{X}$ è la migliore approssimazione possibile a quella parte di $\mathbf{X}$ che non è contaminata dalla correlazione col termine di disturbo, in quanto costruita sulla base dei soli strumenti, che sono incorrelati coi disturbi per ipotesi; in questo senso $\hat{\mathbf{X}}$ è la matrice degli strumenti "ottimali", cioè quella combinazione lineare delle $\mathbf{W}$ che riproduce al meglio $\mathbf{X}$.

3.3 Proprietà asintotiche dello stimatore GIVE

La consistenza dello stimatore GIVE è una conseguenza della convergenza in probabilità di determinate statistiche. Quella che segue è una prova valida quando le variabili osservate possano essere pensate come realizzazioni di variabili casuali indipendenti ed identiche, in possesso dei momenti fino al secondo ordine. Tuttavia, un tipo di prova non dissimile può essere applicata anche a contesti più generali, specificando le opportune condizioni.

Dato il modello lineare $y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t$, se:

1. $(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{w}_t' = \frac{\mathbf{X}'\mathbf{W}}{T} \xrightarrow{\text{p}} A$, dove A ha rango k;

[7] Nota bene: anche se questa procedura fornisce un metodo per calcolare $\widetilde{\boldsymbol{\beta}}$ come statistica OLS, *non* fornisce una stima accettabile di σ^2. Infatti, la somma dei quadrati dei residui ottenuta nel secondo stadio è $(\mathbf{y} - \hat{\mathbf{X}}\widetilde{\boldsymbol{\beta}})'(\mathbf{y} - \hat{\mathbf{X}}\widetilde{\boldsymbol{\beta}})$, che è evidentemente diversa da $\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}}$.

2. $(1/T)\sum_{t=1}^{T}\mathbf{w}_t\mathbf{w}_t' = \frac{\mathbf{W}'\mathbf{W}}{T} \xrightarrow{\text{p}} B$, dove B è non singolare;

3. $(1/T)\sum_{t=1}^{T}\mathbf{w}_t u_t = \frac{\mathbf{W}'\mathbf{u}}{T} \xrightarrow{\text{p}} 0$;

allora $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y} \xrightarrow{\text{p}} \boldsymbol{\beta}$.

Se inoltre

4. $(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{w}_t u_t = \frac{\mathbf{W}'\mathbf{u}}{\sqrt{T}} \xrightarrow{\text{d}} N(0, Q)$;

allora $\sqrt{T}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} N(0, \Sigma)$, dove

$$\Sigma = \left[AB^{-1}A'\right]^{-1} AB^{-1}QB^{-1}A' \left[AB^{-1}A'\right]^{-1}.$$

Le prove impiegano le normali regole di composizione dei limiti in probabilità ed in distribuzione. Per quanto riguarda la consistenza

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{u} \\ &= \boldsymbol{\beta} + \left[\left(\frac{\mathbf{X}'\mathbf{W}}{T}\right)\left(\frac{\mathbf{W}'\mathbf{W}}{T}\right)^{-1}\left(\frac{\mathbf{W}'\mathbf{X}}{T}\right)\right]^{-1}\left(\frac{\mathbf{X}'\mathbf{W}}{T}\right)\left(\frac{\mathbf{W}'\mathbf{W}}{T}\right)^{-1}\left(\frac{\mathbf{W}'\mathbf{u}}{T}\right)\end{aligned}$$

e quindi

$$\widetilde{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta} + \left[AB^{-1}A'\right]^{-1} AB^{-1}\cdot 0 = \boldsymbol{\beta}. \tag{24}$$

Va notato che i requisiti 1 e 2 sul rango delle matrici A e B giocano un ruolo determinante: per quanto riguarda B, la cosa è ovvia (l'inversa deve esistere); per quanto riguarda A, il requisito che essa abbia rango k è fondamentale per dire che la matrice $\left[AB^{-1}A'\right]$ è invertibile. Da un punto di vista pratico, i due requisiti possono essere traslati in requisiti sulle variabili: dire che B dev'essere invertibile significa dire che non devono esserci strumenti (asintoticamente) collineari; dire che A dev'essere di rango pieno k significa invece dire che ogni combinazione strumento-regressore deve contenere un suo messaggio informativo intrinseco non replicato nelle altre combinazioni. Questo esclude dagli strumenti, ad esempio, quelle variabili che sono incorrelate asintoticamente con i regressori, altrimenti A avrebbe una colonna di zeri. Perché quindi una variabile sia uno strumento valido, pertanto, non basta che essa sia incorrelata col disturbo (requisito 3), ma anche che sia correlata con le variabili esplicative (requisito 1).

Con un ragionamento appena un po' più complesso, ma sostanzialmente analogo, si perviene anche a

$$\sqrt{T}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} N\left(0, \left[AB^{-1}A'\right]^{-1} AB^{-1}QB^{-1}A' \left[AB^{-1}A'\right]^{-1}\right)$$

se poi si avesse che $Q = \sigma^2 B$ (ciò che accade se $E\left[\mathbf{u}\mathbf{u}'|\mathbf{W}\right] = \sigma^2\mathbf{I}$) allora si ritrova l'espressione standard

$$\sqrt{T}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} N\left(0, \sigma^2\left[AB^{-1}A'\right]^{-1}\right). \tag{25}$$

A proposito di tale matrice di varianze-covarianze, un punto importante, su cui avremo modo di tornare, riguarda il fatto che la consistenza dello stimatore GIVE è assicurata anche nelle condizioni in cui lo stimatore OLS sarebbe esso stesso consistente per i parametri di interesse. Sotto queste condizioni, tuttavia, lo stimatore GIVE è meno efficiente dell'OLS. Infatti, la differenza fra le varianze asintotiche dei due stimatori è, come minimo, semidefinita positiva. Si può dimostrare in modo rigoroso, ma lunghetto. In modo non rigoroso, si può far notare che la differenza fra i due stimatori delle varianze è essa stessa positiva definita. Infatti, la matrice $(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}$ è p. d. se e solo se lo è anche $\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X}$; ma poiché quest'ultima matrice è uguale a $\mathbf{X}'\mathbf{M}_\mathbf{W}\mathbf{X}$, dev'essere per forza almeno semidefinita positiva.

Per quanto riguarda la consistenza di $\tilde{\sigma}^2$, poiché

$$\widetilde{\mathbf{u}} = \mathbf{X}\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right) + \mathbf{u},$$

si può scrivere

$$\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}} = \left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right) + 2\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{u} + \mathbf{u}'\mathbf{u}$$

e quindi

$$\frac{1}{T}\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}} = \left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)' \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right) + 2\left(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\right)' \left(\frac{\mathbf{X}'\mathbf{u}}{T}\right) + \left(\frac{\mathbf{u}'\mathbf{u}}{T}\right);$$

prendendo il limite in probabilità si ha

$$\tilde{\sigma}^2 = \frac{1}{T}\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}} \xrightarrow{\text{p}} 0'\,(H)\,0 + 2 \cdot 0'(k) + \sigma^2 = \sigma^2,$$

dove H è il limite in probabilità di $\frac{\mathbf{X}'\mathbf{X}}{T}$ e k è il limite in probabilità di $\frac{\mathbf{X}'\mathbf{u}}{T}$. Si noti che per la consistenza di $\tilde{\sigma}^2$ non c'è bisogno che k sia 0.

Infine, la consistenza di $\tilde{\sigma}^2$ comporta che possiamo stimare la matrice di varianze e covarianze asintotica di $\widetilde{\mathbf{u}}$ in modo consistente con $\tilde{\sigma}^2(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1}$, e ciò permette di effettuare tutte le procedure inferenziali asintotiche in modo del tutto standard.

3.4 Le variabili strumentali

Per il momento, abbiamo sempre supposto che la matrice degli strumenti fosse già bell'e pronta: nella realtà, ovviamente, non è così. Quale che sia il problema in esame, bisogna trovare un numero (almeno pari a quello dei regressori) di variabili che abbiano la proprietà di essere correlate con le variabili esplicative e incorrelate con i disturbi, così da avere $\frac{\mathbf{X}'\mathbf{W}}{T} \xrightarrow{\text{p}} A$ di rango pieno e $\frac{\mathbf{W}'\mathbf{u}}{T} \xrightarrow{\text{p}} 0$; in una certa misura, va detto, queste due proprietà

sono contraddittorie: se infatti i regressori sono correlati ai residui, una variabile che fosse perfettamente correlata al regressore non potrebbe essere ortogonale al residuo. Tuttavia, se la correlazione non è completa la cosa è possibile: in una logica tipo due stadi, si può pensare ad uno strumento come ad una variabile correlata con quella "parte" del regressore che non è contaminata dalla correlazione col disturbo. Vedremo negli esempi che variabili di questo tipo sono spesso suggerite dalla natura stessa del problema in esame.

Inoltre, c'è un'altra considerazione che spesso contribuisce ad alleviare la natura del problema: non è detto che *tutti* i regressori siano correlati coi disturbi. In effetti, se alcune delle variabili esplicative sono libere da questo problema (o, come si dice, sono *esogene*), non ci sono problemi ad includere anche queste variabili fra gli strumenti. Queste variabili, si può dire, diventano strumenti di se stesse, cosicché il problema può essere circoscritto a trovare un numero di strumenti (almeno) pari al numero di variabili effettivamente non esogene (o *endogene*). Esempi ovvi di variabili che normalmente sono considerate senza esitazione esogene sono tutte le variabili deterministiche, come ad esempio la costante.

Un'altra questione non banale riguarda il numero degli strumenti da utilizzare: naturalmente questi devono essere abbastanza perché lo stimatore sia calcolabile, e cioè almeno quanti i regressori. Se però fossimo nella fortunata situazione di disporre di un numero molto grande di strumenti, è auspicabile utilizzarli tutti? La logica direbbe di sì, poiché non si vede perché non si dovrebbe utilizzare tutta l'informazione disponibile; in effetti, la teoria asintotica conferma l'intuizione, in quanto si può mostrare in modo non troppo complesso che più strumenti si usano, tanto più lo stimatore risulta asintoticamente efficiente. In realtà, le cose non sono così ovvie in campioni finiti, in quanto è stato dimostrato in diversi contesti che le proprietà in piccoli campioni dello stimatore GIVE tendono a *peggiorare* quando il numero degli strumenti sia molto più grande del numero dei regressori. Ma qui bisogna rinviare alla letteratura specialistica per una trattazione più rigorosa: una esposizione molto chiara del problema, completa di tutti i richiami bibliografici pertinenti, sta in Davidson & MacKinnon (1993).

4 Gli esempi rivisti

4.1 L'abilità individuale

Riprendiamo l'esempio dell'abilità individuale e riscriviamo il modello che ci interessa stimare

$$y_i = \lambda \cdot x_i + v_i, \tag{26}$$

dove λ è il nostro parametro di interesse e v_i è il termine di disturbo che, come sappiamo, include l'intelligenza (che è inosservabile). Immaginiamo

che nel nostro campione gli individui provengano da due atenei diversi: l'ateneo A e l'ateneo B. Una volta tanto, i nomi non sono dati a caso. Infatti, l'ateneo A ha come tradizione quello di dare voti di laurea che, a parità di preparazione, sono più alti di quelli che dà l'ateneo B ('A' e 'B' stanno quindi per 'alto' e 'basso'). Questo accade per complesse ragioni storiche che potremmo anche divertirci ad immaginare, ma sono inessenziali al punto. Supponiamo anche che l'intelligenza media degli studenti dei due atenei sia la stessa[8].

La cosa interessante è che possiamo definire una variabile a_i in questo modo:

$$a_i = \begin{cases} 1 & \text{se l'individuo } i \text{ viene dall'ateneo A} \\ 0 & \text{altrimenti} \end{cases} \tag{27}$$

Questa variabile è una perfetta candidata a fare da strumento nel nostro problema: infatti, è correlata col voto di laurea (perché tale è la politica dell'ateneo A), ma è incorrelata con l'intelligenza.

Facendo un po' di conti (che il lettore è invitato a fare effettivamente come esercizio), si perviene ad uno stimatore IV dato da

$$\lambda_A = \frac{\bar{Y}_A}{\bar{X}_A}, \tag{28}$$

dove $\bar{Y}_A$ è la media aritmetica dei redditi dei laureati di A e $\bar{X}_A$ è la media aritmetica dei loro voti di laurea. Se, come abbiamo supposto,

$$\frac{1}{T}\sum_{i=1}^{T} a_t y_t \xrightarrow{\text{p}} 0$$

e

$$\frac{1}{T}\sum_{i=1}^{T} a_t x_t \xrightarrow{\text{p}} B \neq 0,$$

allora

$$\lambda_A \xrightarrow{\text{p}} 0,$$

come richiesto.

Naturalmente, un ragionamento del tutto analogo avrebbe potuto essere fatto utilizzando una variabile che prendesse come base l'ateneo B anziché l'ateneo A. In effetti, definendo la variabile $b_i = 1 - a_i$, il ragionamento fila immutato, e ci porta a definire un secondo stimatore IV che, per analogia, risulterà essere

$$\lambda_B = \frac{\bar{Y}_B}{\bar{X}_B}. \tag{29}$$

[8]Si potrebbe argomentare che, se uno è intelligente, si iscrive nell'ateneo A piuttosto che nell'ateneo B. Ma se poi in realtà il voto di laurea non conta (come sappiamo) ai fini del reddito, non si vede perché dovrebbe farlo, a parità di altre condizioni. E poi, suvvia, un po' di complicità.

Ma a questo punto, visto che abbiamo non uno, ma due strumenti validi non collineari, non c'è motivo per non utilizzarli entrambi e stimare λ con uno stimatore GIVE. Se raggruppiamo all'inizio del campione tutti i laureati di A per semplicità notazionale, avremo che le matrici rilevanti possono essere scritte così:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} \boldsymbol{\iota} & 0 \\ 0 & \boldsymbol{\iota} \end{bmatrix},$$

con notazione ovvia; le matrici prodotto sono:

$$\mathbf{W}'\mathbf{W} = \begin{bmatrix} T_A & 0 \\ 0 & T_B \end{bmatrix} \qquad \mathbf{W}'\mathbf{X} = \begin{bmatrix} T_A\bar{X}_A \\ T_B\bar{X}_B \end{bmatrix} \qquad \mathbf{W}'\mathbf{y} = \begin{bmatrix} T_A\bar{Y}_A \\ T_B\bar{Y}_B \end{bmatrix},$$

dove T_A e T_B sono, rispettivamente, il numero di laureati di A e B presenti nel campione. Lo stimatore GIVE può a questo punto essere facilmente calcolato:

$$\begin{aligned} \tilde{\lambda} &= \left\{ \begin{bmatrix} T_A\bar{X}_A & T_B\bar{X}_B \end{bmatrix}' \begin{bmatrix} \frac{1}{T_A} & 0 \\ 0 & \frac{1}{T_B} \end{bmatrix} \begin{bmatrix} T_A\bar{X}_A \\ T_B\bar{X}_B \end{bmatrix} \right\}^{-1} \times \\ &\times \begin{bmatrix} T_A\bar{X}_A & T_B\bar{X}_B \end{bmatrix}' \begin{bmatrix} \frac{1}{T_A} & 0 \\ 0 & \frac{1}{T_B} \end{bmatrix} \begin{bmatrix} T_A\bar{Y}_A \\ T_B\bar{Y}_B \end{bmatrix}, \end{aligned}$$

ossia

$$\tilde{\lambda} = \frac{T_A\bar{X}_A\bar{Y}_A + T_B\bar{X}_B\bar{Y}_B}{T_A\bar{X}_A^2 + T_B\bar{X}_B^2}. \tag{30}$$

Se, come abbiamo supposto, $\bar{Y}_A$ e $\bar{Y}_B$ convergono ambedue a zero, e $\bar{X}_A$ e $\bar{X}_B$ convergono ambedue a valori diversi da zero, allora consegue che il limite in probabilità di $\tilde{\lambda}$ è zero come richiesto. Inoltre, è possibile stimare il suo errore standard con la statistica

$$\begin{aligned} \widehat{\mathrm{se}(\tilde{\lambda})} &= \sqrt{\tilde{\sigma}^2 \cdot \left\{ \begin{bmatrix} T_A\bar{X}_A & T_B\bar{X}_B \end{bmatrix}' \begin{bmatrix} \frac{1}{T_A} & 0 \\ 0 & \frac{1}{T_B} \end{bmatrix} \begin{bmatrix} T_A\bar{X}_A \\ T_B\bar{X}_B \end{bmatrix} \right\}^{-1}} = \\ &= \frac{\tilde{\sigma}}{\sqrt{T_A\bar{X}_A^2 + T_B\bar{X}_B^2}}; \end{aligned}$$

quest'ultimo risultato ci permette quindi di costruire la statistica test

$$Z = \frac{\tilde{\lambda}}{\widehat{\mathrm{se}(\tilde{\lambda})}},$$

sulla quale possiamo basare un test di azzeramento, poiché si distribuisce asintoticamente come una normale standardizzata, e quindi considerare "statisticamente significativo" il parametro λ se $|Z| > 1.96$, come al solito.

4.2 L'errore di misura

Per quanto riguarda la verifiche empiriche della teoria del reddito permanente, Friedman usò, nel suo libro del 1957, la tecnica di analizzare dei sottogruppi in modo simile all'esempio precedente. In pratica, vennero confrontate le propensioni medie al consumo di vari gruppi umani variamente definiti: afroamericani contro bianchi, popolazione rurale contro popolazione urbana, e così via.

In realtà, trovare degli strumenti per il reddito permanente è piuttosto problematico. Nella letteratura applicata, il problema si è spostato piuttosto su altre implicazioni pratiche della teoria del reddito permanente, quali ad esempio il fatto che la teoria, così come esposta sopra, non funziona se una quota dei consumatori non è in grado di farsi prestare soldi.

In un contesto più generale di errore di misura, il problema può essere affrontato usando un seconda misurazione il cui errore sia indipendente da quello della prima. Supponiamo infatti di avere un modello con errore di misura simile a quello delle equazioni (8)–(9):

$$\begin{aligned} y_t &= \beta x_t^* + \varepsilon_t \\ x_t &= x_t^* + \eta_t, \end{aligned}$$

in cui chiamiamo y_t la variabile dipendente e x_t^* la variabile esplicativa, osservabile solo nella sua versione contaminata x_t. Come abbiamo visto, la statistica $\left(\sum x_t^2\right)^{-1} \sum x_t y_t$ non converge in probabilità a β. Supponiamo però di avere una seconda misurazione — anch'essa affetta da errore — della variabile x_t^*, che chiamiamo, non a caso, w_t.

$$w_t = x_t^* + \omega_t$$

Si può dimostrare (e al lettore farà bene farlo) che, se η_t e ω_t sono incorrelate, allora w_t si può usare come strumento, e lo stimatore

$$\tilde{\beta} = \frac{\sum w_t y_t}{\sum w_t x_t}$$

è effettivamente consistente per β.

4.3 I sistemi di equazioni simultanee

Consideriamo di nuovo il modello dato dalle equazioni (13) e (14), riportate qui di seguito:

$$\begin{aligned} q_t &= \alpha_0 - \alpha_1 p_t + u_t \\ p_t &= \beta_0 + \beta_1 q_t + v_t. \end{aligned}$$

Come abbiamo visto, con gli OLS possiamo stimare in modo consistente il parametro $\pi_1 = \frac{\beta_0 + \beta_1 \alpha_0}{1 + \beta_1 \alpha_1}$, ossia il valore atteso di p_t: basta regredire p_t

su una costante (o, che è lo stesso, calcolare la sua media aritmetica). Con procedura analoga, potremmo stimare $\pi_0 = E[q_t]$, che è uguale a $\frac{\beta_1 - \beta_0\alpha_1}{1+\beta_1\alpha_1}$. Il problema è che questa procedura non è sufficientemente informativa sui parametri di nostro interesse, che sono le α e le β. Per stimare queste ultime, è necessario trovare degli strumenti adeguati. In un sistema di equazioni simultanee, questo obiettivo viene spesso raggiunto considerando le variabili esogene del sistema stesso.

In linea generale, un sistema di equazioni lineari può essere scritto come

$$\Gamma \mathbf{y}_t = B\mathbf{x}_t + \mathbf{u}_t; \tag{31}$$

il vettore $\mathbf{y}_t$ è un vettore a n dimensioni che contiene tutte le endogene, mentre $\mathbf{x}_t$ contiene le m esogene. Ovviamente, Γ è una matrice non singolare $n \times n$ e B è una matrice $n \times m$. Nell'esempio precedente,

$$\Gamma = \begin{bmatrix} 1 & \alpha_1 \\ -\beta_1 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}.$$

L'equazione (31) è detta *forma strutturale*, perché i parametri contenuti nella matrici Γ e B sono i nostri parametri di interesse. Premoltiplicando la forma strutturale per Γ^{-1}, si ottiene la cosiddetta *forma ridotta*:

$$\mathbf{y}_t = \Pi\mathbf{x}_t + \mathbf{w}_t, \tag{32}$$

nella quale, naturalmente, $\Pi = \Gamma^{-1}B$ e $\mathbf{w}_t = \Gamma^{-1}\mathbf{u}_t$. Nel nostro esempio, la matrice Π è un vettore colonna, che contiene π_0 e π_1, così come li abbiamo definiti poc'anzi.

Poiché si suppone che $\mathbf{x}_t$ sia incorrelato con $\mathbf{u}_t$, ne consegue che $\mathbf{x}_t$ è anche incorrelato con $\mathbf{w}_t$, e quindi, come abbiamo già visto, i parametri della forma ridotta possono essere stimati in modo consistente con gli OLS, possibilità che non è data per i parametri della forma strutturale. Ma un'altra conseguenza interessante si ottiene postmoltiplicando la forma ridotta per $\mathbf{x}_t'$:

$$\mathbf{y}_t\mathbf{x}_t' = \Pi\mathbf{x}_t\mathbf{x}_t' + \mathbf{w}_t\mathbf{x}_t';$$

prendendo il valore atteso dell'espressione precedente, si noterà che esso è uguale a $\Pi E[\mathbf{x}_t\mathbf{x}_t']$; a meno di casi specialissimi, questa è una matrice piena (cioè senza zeri). Se le variabili fossero espresse in scarti dalla media, sarebbe anche la matrice delle covarianze fra l'intero vettore $\mathbf{y}_t$ e l'intero vettore $\mathbf{x}_t$. Se ne deduce che ognuna delle esogene è, in generale, correlata con ognuna delle endogene senza essere correlata con i disturbi. Ma questo è precisamente il requisito necessario a far sì che ognuna delle esogene possa essere usata come strumento.

Nell'esempio precedente, il tutto serve a ben poco, visto che il numero dei regressori in ogni equazione della forma strutturale (due) è maggiore del

numero delle esogene (la sola costante). Siamo in presenza di sottoidentificazione. Consideriamo però una modifica del modello espresso dalle equazioni (13)-(14) data da:

$$q_t = \alpha_0 - \alpha_1 p_t + \alpha_2 y_t + u_t \quad (33)$$
$$p_t = \beta_0 + \beta_1 q_t + \beta_2 m_t + v_t, \quad (34)$$

dove le nuove variabili (assunte ambedue esogene) hanno i seguenti significati:

y_t	reddito pro capite al tempo t
m_t	costo delle materie prime al tempo t

In questo caso, ambedue le equazioni del modello sono stimabili con il metodo GIVE, in quanto abbiamo tre regressori e tre strumenti, ossia le tre esogene (costante, y_t e m_t). L'argomento può essere generalizzato dando luogo alla cosiddetta *condizione di ordine* per la stima di un'equazione che fa parte di un sistema. Condizione necessaria per l'identificazione, e quindi la stimabilità, di un'equazione è che il numero di endogene *incluse* nell'equazione sia minore o uguale al numero di esogene *escluse* dall'equazione, appunto perché le ultime devono servire come strumenti per le prime.

È possibile che il lettore si senta a questo punto leggermente preso in giro: per uscire d'impaccio e applicare lo stimatore GIVE ho aggiunto due esogene al modello, a mo' di *deus ex machina*. In parte è così, ma questo deriva dallo stratagemma didattico di partire con un modello ridotto il più possibile all'osso. Nella realtà, ogni equazione di un modello simultaneo appena appena realistico contiene un numero di endogene incluse molto inferiore al numero di esogene escluse, per cui il problema della sottoidentificazione è in genere più teorico che reale.

5 I test di Sargan e di Hausman

La normalità asintotica dello stimatore GIVE sia normale apre naturalmente la strada a tutte le procedure standard di test che si usano nei modelli stimati con la tecnica OLS, ivi comprese le procedure diagnostiche. In questa sezione illustreremo due test che però, per la loro stessa natura, hanno senso soltanto nell'ambito della stima di un modello con variabili strumentali.

5.1 Il test di Sargan

Naturalmente, il fatto che noi ipotizziamo che certe variabili abbiano i requisiti necessari per essere usate come strumenti non implica che tali requisiti li abbiano per davvero. In particolare, non è detto che la correlazione fra le variabili $\mathbf{w}_t$, che usiamo come strumenti, e i disturbi u_t sia effettivamente

0, o almeno non è detto che tale condizione valga per tutti gli elementi del vettore $\mathbf{w}_t$.

Se fossero osservabili i disturbi $\mathbf{u}$, non sarebbe difficile costruire un test al proposito: sotto l'ipotesi nulla $\mathbf{W}'\mathbf{u} = 0$, infatti, si avrebbe

$$\frac{1}{\sqrt{T}}\mathbf{W}'\mathbf{u} \xrightarrow{\text{d}} MN\left(0, \sigma^2\mathbf{W}'\mathbf{W}\right)$$

e quindi, dato uno stimatore consistente di σ^2, si dimostra che sotto l'ipotesi nulla la quantità

$$\frac{\mathbf{u}'\mathbf{P}_{\mathbf{W}}\mathbf{u}}{\hat{\sigma}^2} \tag{35}$$

converge in distribuzione ad una χ^2 con m gradi di libertà, ossia con tanti gradi di libertà quanti sono gli strumenti. Purtroppo, questa quantità non è utilizzabile come statistica test perché non è una statistica, visto che il vettore $\mathbf{u}$ non è osservabile. Potremmo chiederci però se si può ottenere un risultato analogo adoperando i residui $\widetilde{\mathbf{u}}$ in luogo dei disturbi $\mathbf{u}$, così come abbiamo fatto per costruire uno stimatore consistente di σ^2.

Quyesta strategia conduce al *test di Sargan*. La cosa importante da notare, a proposito di questo test, è che questo test ha però una distribuzione asintotica diversa da quella dell'espressione in (35), in quanto il numero dei gradi di libertà della χ^2 a cui converge in distribuzione la statistica non è pari a m, bensì a $m - k$. In formule,

$$S = \frac{\widetilde{\mathbf{u}}'\mathbf{P}_{\mathbf{W}}\widetilde{\mathbf{u}}}{\hat{\sigma}^2} \xrightarrow{\text{d}} \chi^2_{m-k}. \tag{36}$$

Questo risultato, a prima vista bizzarro, si spiega facilmente considerando che il numeratore dell'espressione in (36) è il valore minimo della funzione obiettivo già vista nell'espressione (22). Come sappiamo, nel caso di esatta identificazione, tale valore è zero. Più in generale, si può dimostrare che il numeratore della (36) si può scrivere come una forma quadratica con una matrice di rango $(m - k)$, da cui il risultato. Questo test, pertanto, non ha come ipotesi nulla la validità degli strumenti, ma solo degli eventuali vincoli di sovraidentificazione.

Da un punto di vista pratico, vale la pena di far notare che il test si può calcolare molto facilmente come il prodotto fra l'ampiezza campionaria T e l'indice R^2 non centrato della regressione ausiliaria

$$\widetilde{\mathbf{u}} = \mathbf{W}\gamma + \text{residui}.$$

5.2 Il test di Hausman

Sino ad ora abbiamo esaminato casi in cui supponevamo di sapere dal principio che nel nostro modello di regressione comparivano variabili esplicative

non esogene, e quindi la stima OLS non ci forniva informazioni sui parametri di interesse. Nella realtà, questo non sempre è il caso. Supponiamo ad esempio di considerare una funzione di costo del tipo:

$$\log C_i = \beta \log Y_i + \sum_{j=1}^{N} \gamma_j \log p_{ij} + \varepsilon_i; \tag{37}$$

se l'impresa è *price-taker*, evidentemente i prezzi per lei sono dati, e si possono tranquillamente considerare esogeni. Ma se l'impresa fosse monopsonista o oligopsonista sul mercato del j-esimo fattore, allora la domanda di quel fattore sarebbe influenzata significativamente dal comportamento dell'impresa, e dovremmo considerare l'equazione (37) come parte di un sistema simultaneo. Nei casi intermedi, non è chiaro come procedere.

Si potrebbe argomentare che, per sicurezza, nei casi incerti si potrebbe usare comunque lo stimatore GIVE; questa strada, tuttavia, può non essere ottimale. Infatti, se il problema della endogeneità non ci fosse, lo stimatore OLS assicurerebbe un vantaggio in termini di efficienza che in certi contesti può essere considerevole. Il test di Hausman[9] si basa sull'idea di calcolare entrambi gli stimatori, e decidere *ex post*, sulla base del loro confronto, quale dei due è più adatto ai nostri scopi.

In realtà, il principio su cui si basa il test è molto più generale, e può essere illustrato con una metafora. Una Ferrari corre più di una Land Rover, e quindi arriva prima, ma solo se la strada è liscia. Nel caso di strada accidentata, la Land Rover non fa una piega, mentre la Ferrari non è neanche detto che arrivi. Per sapere com'è la strada, pertanto, tutto ciò che dobbiamo fare è far correre la Ferrari e la Land Rover. Se arrivano tutt'e due, vuol dire che la strada era liscia. Se no, vuol dire che era accidentata.

Fuor di metafora: supponiamo di avere due stimatori — chiamiamoli $\hat{\theta}$ e $\tilde{\theta}$ — per lo stesso parametro incognito θ. Immaginiamo che $\hat{\theta}$ sia più efficiente di $\tilde{\theta}$, ma consistente sotto un insieme più ristretto di condizioni. L'ipotesi nulla di un test di Hausman è appunto che valgano tali condizioni. Se è così, i due stimatori sono entrambi consistenti, e la loro differenza non dovrebbe risultare statisticamente significativa. Il test è pertanto basato sulla statistica $\hat{\theta} - \tilde{\theta}$. Le caratteristiche che i due stimatori devono avere perché il test funzioni sono sintetizzate nella tabella 1.

Quindi, se definiamo la statistica $\delta = \hat{\theta} - \tilde{\theta}$, il test ha la forma

$$H = \delta' \left[\widehat{V(\delta)}\right]^{-1} \delta; \tag{38}$$

se $\widehat{V(\delta)}$ è uno stimatore consistente della varianza asintotica di δ, allora si può dimostrare con metodi asintotici standard che il test, sotto l'ipotesi

[9]Per onestà, bisogna dire che lo stesso test era stato proposto, prima di Hausman, da Durbin e da Wu indipendentemente, tant'è che su alcuni testi viene chiamato test di Wu-Hausman, o test di Durbin-Wu-Hausman. Ma, per ingiusto che sia, ci rimettiamo all'uso corrente.

Tabella 1: Caratteristiche degli stimatori nel test di Hausman

	Sotto H_0	Sotto H_1
$\hat{\theta}$	Consistente – Asintoticamente Normale – Efficiente	Non consistente
$\tilde{\theta}$	Consistente – Asintoticamente Normale	Consistente

nulla, ha una distribuzione asintotica χ^2 con un numero di gradi di libertà pari alla dimensione di δ.

Nel nostro caso, gli stimatori da considerare sono lo stimatore OLS $\widehat{\boldsymbol{\beta}}$ e lo stimatore GIVE $\widetilde{\boldsymbol{\beta}}$, per cui $\delta = \widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$. Per implementare il test di Hausman, a questo punto, ci manca solo uno stimatore consistente di $V(\delta)$. Poiché

$$V(\delta) = V(\widehat{\boldsymbol{\beta}}) + V(\widetilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}),$$

sembrerebbe di dover disporre di uno stimatore consistente di $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}})$ per poter calcolare il test. In realtà questo non è necessario, perché risulta

$$\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}) = V(\widehat{\boldsymbol{\beta}}),$$

per cui

$$V(\delta) = V(\widetilde{\boldsymbol{\beta}}) - V(\widehat{\boldsymbol{\beta}}).$$

Questo miracoloso risultato deriva da una considerazione abbastanza generale: se $\hat{\theta}$ e $\tilde{\theta}$ sono due stimatori consistenti di un parametro incognito θ, e $\hat{\theta}$ è asintoticamente efficiente, allora la covarianza fra i due è pari alla varianza di quello più efficiente. Abbozzo di prova nel caso scalare:

$$AVar \begin{pmatrix} \hat{\theta} \\ \tilde{\theta} \end{pmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Consideriamo ora la statistica $\grave{\theta} = \lambda\hat{\theta} + (1-\lambda)\tilde{\theta}$, dove λ è un numero reale qualunque. Ovviamente $\grave{\theta} \xrightarrow{\mathrm{p}} \theta$ per costruzione. La sua varianza asintotica è data da

$$\begin{pmatrix} \lambda & 1-\lambda \end{pmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} . \begin{pmatrix} \lambda \\ 1-\lambda \end{pmatrix} = \lambda^2 a + 2\lambda(1-\lambda)b + (1-\lambda)^2 c.$$

Si dimostra facilmente che la scelta di λ che rende minima $V(\grave{\theta})$ è

$$\frac{b-c}{a-2b+c};$$

ma poiché λ dev'essere 1 (altrimenti $\hat{\theta}$ non sarebbe efficiente), si deduce che $a = b$.

Poiché sotto l'ipotesi nulla lo stimatore OLS è consistente, allora anche $\hat{\sigma}^2$ è uno stimatore consistente della varianza, e pertanto la matrice

$$\hat{\sigma}^2 \left[(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \right]$$

è uno stimatore consistente di $V(\delta)$. Il test di Hausman, pertanto, si può calcolare con la formula

$$H = \frac{\left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)' \left[(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \right]^{-1} \left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2}. \tag{39}$$

In pratica, il calcolo del test è ancora più semplice, visto che il test di Hausman si può calcolare molto facilmente con una regressione ausiliaria: consideriamo infatti il modello

$$\mathbf{y} = \mathbf{X}\beta + \hat{\mathbf{X}}\gamma + \text{residui}. \tag{40}$$

Mostreremo ora che il test di Hausman è numericamente uguale ad un test di azzeramento del parametro γ. Usando il teorema di Frisch e Waugh si ha infatti che

$$\hat{\gamma} = \left[\hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}} \right]^{-1} \hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\mathbf{y};$$

dalla definizione di $\mathbf{M}_\mathbf{X}$ si può scrivere

$$\begin{aligned} \hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}} &= \hat{\mathbf{X}}'\hat{\mathbf{X}} - \hat{\mathbf{X}}'\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\hat{\mathbf{X}} = \\ &= \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right) \left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1} \right] \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right), \end{aligned}$$

dove la seconda uguaglianza è giustificata dal fatto che $\hat{\mathbf{X}}'\hat{\mathbf{X}} = \hat{\mathbf{X}}'\mathbf{X}$. Con ragionamento assolutamente analogo si perviene a

$$\hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\mathbf{y} = \hat{\mathbf{X}}'\mathbf{y} - \hat{\mathbf{X}}'\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right) \left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)$$

e quindi

$$\hat{\gamma} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} \left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1} \right]^{-1} \left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right).$$

Poiché un test di tipo Wald di azzeramento di γ è dato dalla statistica

$$W = \frac{\hat{\gamma}' \left[\hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}} \right] \hat{\gamma}}{\hat{\sigma}^2},$$

si vede chiaramente, con un po' di sostituzioni, che la stessa statistica può essere scritta come

$$W = \frac{\left[\mathbf{y}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}} \right]' \hat{\gamma}}{\hat{\sigma}^2} = \frac{\left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)' \left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1} \right]^{-1} \left(\widetilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2},$$

che è appunto la definizione del test di Hausman.