stesura: Sandra Bonfiglioli
revisione: Elena De Valerio

PARTE 1

IL METODO ECONOMETRICO:

SPECIFICAZIONE, STIMA, TEST

Il *metodo econometrico* cerca di attuare una fusione tra le tre fonti d'informazione, cioè tra: dati (organizzati in serie storiche, oppure cross-section, oppure panel) vettori di osservazioni riguardanti ad esempio le famiglie, le imprese, il debito, i finanziamenti; la teoria economica (ad esempio le teorie di Modigliani per le scelte di finanziamento delle imprese, le teorie classiche, le teorie Keynesiane in cui si dice che il consumo dipende dal reddito) e gli strumenti matematici che formalizzano i concetti delle teorie economiche e ne esprimono delle formule.

Il *metodo econometrico,* usato per la costruzione di modelli econometrici, può essere classificato in tre fasi: specificazione del modello, stima dei parametri e test.

1.1 Specificazione del modello

In questa prima fase di specificazione del modello un ruolo importante lo svolgono le ipotesi che si fanno su come è fatto il processo statistico che ha generato i nostri dati. La teoria economica suggerisce l'elenco delle variabili di interesse del problema che si intende affrontare e la direzione di causalità (ad esempio $Y = f(X)$ cioè la variabile dipendente Y è spiegata dalla variabile esplicativa X → nesso di causalità tra reddito e consumo e, tale nesso, non può essere spiegato dalla matematica o dalla statistica).

D'altro canto la teoria da sola non basta per definire compiutamente tutti gli elementi in cui si compone un modello econometrico stimabile; per questo sono necessarie ipotesi di specificazione quali la scelta della forma funzionale. La matematica trasforma la semplice relazione $Y = f(X)$, in una relazione specifica e, se imponiamo l'ipotesi di linearità, attraverso eventuali trasformazioni delle variabili (ad esempio trasformazioni logaritmiche che hanno lo scopo di smussare eventuali osservazioni anomale) evidenziamo il punto di vista quantitativo. L'ipotesi di linearità non è così restrittiva, ad esempio, se dico che Y è uguale ad "a + bX" e prendo i consumi di quest'anno, sto solo dicendo che questo dato è uguale ad "a + bX"; è invece importante capire in quale stadio ipotizzare la linearità, in modo che sia il più veritiera possibile.

La relazione è inoltre stocastica per la presenza di un termine d'errore, spesso additivo, cioè di una variabile casuale che serve a cogliere qualsiasi altro effetto non esplicitato (spesso non osservabile direttamente e non misurabile) e che dipende dalla qualità del modello (in un buon modello l'errore sarà marginale). Tale variabile casuale è

rappresentata da ε e, proprio con l'aggiunta di ε alla relazione, possiamo scrivere: $Y = a + bX + \varepsilon$, cioè ε rende esatta la relazione ipotizzata dalla teoria.

Oltre all'ipotesi (1) : $Y_i = a + bX_i + \varepsilon_i$, di linearità della funzione, si fanno altre quattro ipotesi:

ipotesi (2) : $\text{cov}(X_i, \varepsilon_i) = 0$

ipotesi (3) : $E(\varepsilon_i) = 0$

ipotesi (4) : $E(\varepsilon_i^2) = \sigma^2$

ipotesi (5) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

$\varepsilon_i \sim i.i.d(0, \sigma^2)$ le componenti ε sono tra loro indipendenti ed identicamente distribuite, con media nulla e varianza costante σ^2

Le ipotesi fatte su ciò che non conosciamo (ε) ci indicano:

la (2): che le variabili esplicative a destra dell'uguale non covariano con l'errore. Ad esempio, se uno shock sul consumo non retroagisce sul reddito, le variabili esplicative del modello di consumo (il reddito) sono indipendenti dallo shock sul consumo (termine di errore).

la (3): che in previsione, il valore atteso degli errori deve essere nullo

la (4): che la varianza di ε è ipotizzata uguale ad un numero costante σ^2

la (5): che la covarianza dei termini d'errore tra gli individui è uguale a zero, cioè vi è assenza di correlazione tra gli ε, ciò significa che i disturbi sono estremamente individuali e non influenti per il modello, se non fosse così dovremmo specificare meglio il modello e significherebbe che nel termine d'errore c'è qualcosa di sistematico.

1.2 Stima dei parametri

I parametri *a*, *b* sono incogniti e costituiscono l'oggetto dell'inferenza statistica che viene condotta nell'ambito del modello di regressione lineare. Poter disporre delle stime di tali parametri ci permette di quantificare la relazione di causalità fra le variabili esplicative e la variabile dipendente.

Un metodo largamente utilizzato per la stima del modello parametrico è quello dei *minimi quadrati ordinari (OLS ordinary least squares)*, che attribuisce ai parametri della relazione quei valori che minimizzano il quadrato delle distanze fra le osservazioni disponibili e la corrispondente retta di regressione; tali distanze sono anche dette residui ε_i.

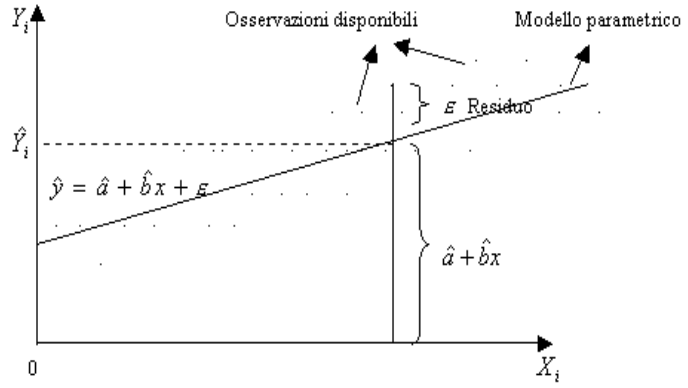

Fig. 2 - Retta di regressione stimata, osservazioni campionarie e residui

Le stime $\hat{a}$ e $\hat{b}$ vengono scelte in modo da rendere più piccoli possibile i residui ε_i quindi la miglior retta di regressione è quella

ottenuta minimizzando la somma dei quadrati dei residui, cioè minimizzando la funzione $f(.) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

Per ottenere i residui si fa la **stima del termine d'errore** :

$$\underbrace{\hat{\varepsilon}_i}_{\text{ciò che non sappiamo del modello sul comportamento di } Y_i} = \underbrace{Y_i}_{\text{ciò che voglio spiegare (dati actual)}} - \underbrace{(\hat{a} + \hat{b}X_i)}_{\text{parte spiegata del modello (dati fitted)}}$$

ponendo $\hat{Y}_i = (\hat{a} + \hat{b}X_i)$

otteniamo $\boxed{\hat{\varepsilon}_i = Y_i - \hat{Y}_i}$ che rappresenta la differenza fra il dato reale e quello spiegato (stimato) dal modello

Ora passiamo alla **stima di $\hat{a}$:**

cerchiamo quel valore che minimizza il valore dei residui e lo si ottiene a partire dall'ipotesi (3) in cui si pone la somma dei residui pari a zero:

Momento primo teorico $E(\varepsilon_i) = 0$

Momento primo empirico $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

Dato: $\hat{\varepsilon}_i = Y_i - \hat{a} - \hat{b}X_i$

Ottengo:

$$\sum_i (Y_i - \hat{a} - \hat{b}X_i) = 0$$

$$\sum_i Y_i - \underbrace{\sum_i \hat{a}}_{} - \sum_i \hat{b}X_i = 0$$

$$\sum_I Y_i - N\hat{a} - \hat{b}\sum_i X_i = 0$$

$$\hat{a} = \underbrace{\frac{\sum_i Y_i}{N}}_{\text{media campionaria di } Y_i} - \hat{b}\underbrace{\frac{\sum_i X_i}{N}}_{\text{media campionaria di } X_i}$$

$$\boxed{\hat{a} = \bar{Y} - \hat{b}\bar{X}}$$ FORMULA della STIMA di ***a***

Data l'ipotesi (2) (anche questa ipotesi sta alla base del metodo *OLS*) che descrive l'ortogonalità fra i residui e le variabili esplicative del modello, possiamo vedere come si ottiene la **stima di $\hat{b}$:**

Ipotesi (2) $cov(X_i, \varepsilon_i) = 0$

Scritta come valore atteso $E[(X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}_i) = 0$

Ovvero (in scarti) $Ex_i\varepsilon_i = 0$

Si ottiene la corrispondente restrizione stimata:

$$\boxed{\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0}$$ CONDIZIONE di ORTOGONALITA'

Nota: la scrittura del modello in scarti rispetto alla media permette calcoli più semplici. Il modello in livelli: $Y_i = a + bX_i + \varepsilon_i$, può essere riscritto in modo analogo come: $y_i = bx_i + \varepsilon_i$, dove le variabili in minuscolo sono pari agli scarti rispetto alle medie; il parametro b e il termine di errore sono gli stessi.

Dati i residui del modello in scarti $\hat{\varepsilon}_i = y_i - \hat{b}x_i$

Ottengo:

$$\sum_{i=1}^{n} x_i(y_i - \hat{b}x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{b}\sum_{i=1}^{n} x_i^2 = 0$$

$$\boxed{\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}$$

Stima della covarianza fra X e Y

FORMULA della STIMA di ***b***

Stima della varianza di X

Le due formule di stima di $\hat{a}$ e di $\hat{b}$, come citato sopra, sono ottenute con il metodo *OLS;* a questo punto, sostituendo nella formula di $\hat{b}$ il modello scritto in "scarti" e si ottiene la **formula dello stimatore OLS**:

Dato il modello teorico in scarti $y_i = \hat{b}x_i + \hat{\varepsilon}_i$

Sostituisco y_i nella formula della stima OLS di *b*:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i (bx_i + \varepsilon_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i^2 b}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$$ da cui:

$$\hat{b} = b + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$$

FORMULA dello STIMATORE **OLS**

Nella formula dello stimatore OLS c'è la presenza della variabile ε_i che, essendo una variabile casuale, si ripercuote su $\hat{b}$ che (a priori) non è più un numero (stima) ma diventa una variabile casuale (stimatore). Si può dedurre che questa non è una formula operativa perché il parametro *b* (che appare destra dell'uguale) è sconosciuto, così come la variabile casuale ε_i. Si noti che la stima non potrà mai centrare esattamente l'obiettivo *b*. Se andassimo a ripetere le prove su campioni diversi estratti dalla stessa popolazione otterremmo stime sempre diverse. Da questo emerge che la parte più importante dell'applicazione di un metodo di stima (stimatore) sono *le proprietà* statistiche dello *stimatore (nel nostro caso gli OLS).*

Ora possiamo evidenziare che, se valgono le cinque ipotesi del modello classico di regressione lineare, si dimostra per il '*Teorema di Gauss-Markov*' che lo stimatore è **B L U E**:

Best **Linear** **Unbiased** **Estimator**

" **Miglior** **Stimatore** **Lineare** **Non distorto" (corretto)**

Tra tutti gli stimatori lineari non distorti quello *OLS* ha la varianza minima

FORMULA della VARIANZA dello STIMATORE:

$$E(\hat{b} - b)^2 = \frac{\sigma^2}{\sum_i x_i^2}$$

Lo stimatore non è distorto quando: $E(\hat{b}) = b$ ovvero quando: $E(\hat{b} - b) = 0$

1.3 Test

Le ipotesi di specificazione formano l'oggetto dei test. La prima serie di test ha a che vedere con le scelte di specificazione del modello e si concentrano sull'analisi dei residui della regressione.

Tutti questi test hanno sotto l'ipotesi nulla H_o (sottoinsieme dei valori dello spazio parametrico individuati dall'ipotesi da sottoporre a verifica) le ipotesi (1) (2) (3) (4) (5), che prese singolarmente devono essere ragionevoli e lo stimatore deve avere le proprietà desiderate/ottime.

Questa prima serie di test va sotto il nome di test di scorretta specificazione.

1.3.1 Test di scorretta specificazione sui residui

Per dimostrare che non c'è distorsione, cioè la distorsione è uguale a zero (proprietà della correttezza), si tengono in considerazione le prime tre ipotesi :

(1) +(2)+(3) → $E(\hat{b}) = b$ *quindi non c'è distorsione e lo stimatore è BLUE*

invece per dimostrare che la varianza è minima (proprietà dell'efficienza) si prendono in considerazione le ultime due ipotesi di specificazione :

(4)+(5) → $Var(\hat{b}) = \frac{\sigma^2}{\sum_i x_i^2}$ *la varianza è minima*

I test di scorretta specificazione si dividono in tre categorie a seconda dell'ipotesi che vanno a verificare. Sono tutti basati sui residui della regressione

Test di esogeneità debole

Per verificare la ragionevolezza dell'ipotesi (2) : $Cov(X_i \varepsilon_i) = 0$ alla luce dei dati campionari che sto osservando.

Test di eteroschedasticità

Per vedere se ha senso dire che la varianza del termine d'errore è costante, ipotesi (4), cioè:

- Se $E(\varepsilon_i)^2 = \sigma^2$ varianza costante → situazione di *omoschedasticità*
- Se $E(\varepsilon_i)^2 \neq \sigma^2$ le varianze sono diverse → situazione di *eteroschedasticità*

Test di autocorrelazione

Per verificare se c'è correlazione fra ε_i e gli ε_j cioè verifica dell'ipotesi (5)

1.3.2 Test di significatività dei parametri

Questo secondo gruppo di test sono invece basati sulle stime dei parametri. Sarebbe necessario fare un test sul parametro *b* ma, dato che non lo si conosce, si fa dell'inferenza e si usa la stima di b e la stima di ε.

Se alle precedenti 5 ipotesi si aggiunge anche l'ipotesi (6) → di Normalità della distribuzione degli errori, allora è possibile calcolare specifici *intervalli di stima* per i parametri del modello ed effettuare dei *test di verifica dell'ipotesi* sui parametri, utilizzando i valori critici della distribuzione *t di Student.*

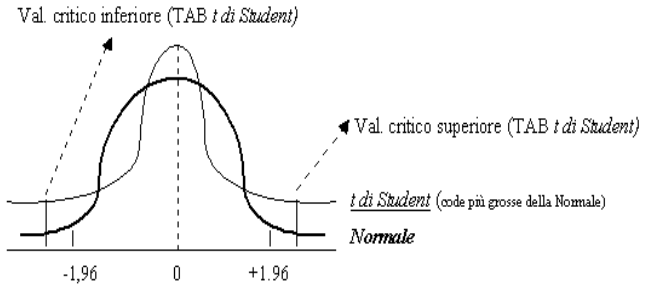

Fig. 3 - Rappresentazione distrib. Normale a confronto con distrib. *t di Student*

Esempio:

- se la distribuzione di questa statistica è una ***t di Student***
- data l'ipotesi nulla H_o: $b=1$
- data l'ipotesi alternativa H_1: $b \neq 1$
- supponiamo di disporre di una stima del modello di consumo:

  $C_t = 7{,}6 + 0{,}872 R_{t-1} + \hat{\varepsilon}_t$

  dove le stime dei parametri sono : $\hat{a} = 7{,}6$

  $\hat{b} = 0{,}872$

  i corrispondenti errori standard sono: $s_a = 2{,}1$

  $s_b = 0{,}012$

- abbiamo 42 osservazioni → N=42
- vi sono 2 parametri, a e b → K=2
  - Si calcola la Statistica test:

$$\frac{\hat{b} - b}{s / \sqrt{\sum_i x_i^2}} = \frac{0{,}872 - 1}{0{,}012} = \frac{-0{,}128}{0.012} = -10{,}666$$

  - Si estraggo dalla tabella della distribuzione *t di Student* i valori critici corrispondenti ai seguenti gradi di libertà:
    (N-K) = (42 – 2) = 40 gradi di libertà
    che corrispondono ai valori critici **±2,021** (tabella *t di Student)*

  - Si verifica se il risultato ottenuto con la statistica test sta nella zona di accettazione dell'ipotesi nulla o di rifiuto:

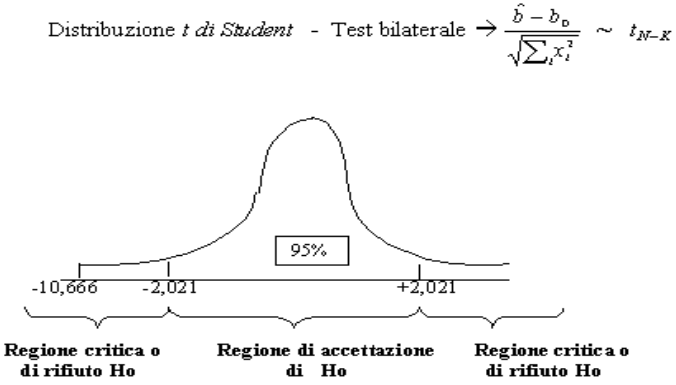

Fig. 4 - Test di significatività bilaterale (*test t*) con distribuzione t di Student

Dato un livello di significatività del 5%, il valore –10,666 è nella zona in cui si rifiuta H_o: $b=1$ quindi equivale accettare l'ipotesi alternativa H_1: b ≠ 1.

Quando la statistica *t* cade nella regione critica, l'evidenza empirica del fenomeno studiato porta a ritenere che l'ipotesi H_o non debba considerarsi valida e perciò non può essere considerata come vera.

1.4 Estensione del modello a più variabili

La presente sezione è volta ad estendere l'analisi delle caratteristiche e delle proprietà dello stimatore OLS al caso di più variabili esplicative. In particolare, si affronterà il caso emblematico di un modello lineare con due variabili esplicative:

$Y = a + b X_1 + c X_2 + \varepsilon$ (1a)

che può essere riscritto come:

$y = b x_1 + c x_2 + \varepsilon$ (1b)

dove Y è la variabile dipendente; X_1 e X_2 sono le variabili esplicative; *a*, *b* e *c* sono parametri (in particolare, *a* è la costante); ε è il termine di errore stocastico. Quando il modello viene specificato con lettere minuscole (1b), indicheremo scostamenti dalla media; ad esempio $x_1 = X_1 - \bar{X}_1$.

Le problematiche che affronteremo con riferimento ad un modello con due esplicative sono le stesse di un modello con un numero di variabili esplicative superiore a due.

Per la derivazione degli stimatori OLS nel modello classico di regressione lineare con due variabili esplicative, anche se il calcolo risulta più complicato di quello visto per una sola esplicativa, il problema della minimizzazione dei quadrati dei residui del modello è sempre lo stesso: si tratta di trovare quelle stime dei parametri che soddisfano le condizioni di somma dei residui uguale a zero e di ortogonalità dei residui con le variabili esplicative del modello; in particolare, visto che ora si affronta il caso di un modello con due esplicative X_1 e X_2 (oltre alla costante) avremo due (e non più una) condizioni di ortogonalità. Il risultato sarà un sistema di tre equazioni normali dalla cui soluzione si ricava la formula della stima OLS dei

parametri *a* (la costante), *b* (relativo a X_1) e *c* (relativo a X_2) del modello lineare. Se si dispone di N osservazioni, i gradi di libertà del modello di regressione saranno in generale N-K, con K pari al numero di parametri stimati (nel presente caso K=3 e, quindi, i gradi di libertà sono N-3).

Ora supponiamo di specificare il seguente modello a più variabili:

fase1: specificazione del modello a più variabili (in questo caso, due: X e Z)

- Ipotesi (1):

  $Y_i = a + bX_i + cZ_i + \varepsilon$ Modello sui livelli

  $y_i = bx_i + cz_i + \varepsilon_i$ Scarti dei livelli rispetto alla propria media

come si può notare abbiamo una relazione lineare con due variabili esplicative *X* e *Z*, entrambe sono esogene (non covariano con il termine d'errore):

- Ipotesi (2):

  $\text{cov}(X_i, \varepsilon_i) = 0$
  $\text{cov}(Z_i, \varepsilon_i) = 0$

- Ipotesi (3)-(5), le stesse che per il modello con solo una esplicativa:

  $\varepsilon_i \sim i.i.d(0, \sigma^2)$

fase2: stima del modello a più variabili

si ricavano le condizioni di ortogonalità dall'uguaglianza dei momenti teorici (ipotesi di specificazione) con i corrispondenti momenti empirici:

1. per $\hat{a}$ $\quad \sum_i \hat{\varepsilon}_i = 0$
2. per $\hat{b}$ $\quad \sum_i x_i \hat{\varepsilon}_i = 0$
3. per $\hat{c}$ $\quad \sum_i z_i \hat{\varepsilon}_i = 0$

Dopodichè si ricavano le formule per la stima di: $\hat{a}$, $\hat{b}$, $\hat{c}$.

La stima di *a* si ottiene facilmente a partire dalla prima condizione di ortogonalità:

$$\hat{a} = \overline{Y} - \hat{b}\overline{X} - \hat{c}\overline{Z}$$

FORMULA della STIMA di $\hat{a}$

Le stime di $\hat{b}$ e di $\hat{c}$ sono simmetriche, si ottengono sostituendo la definizione di residuo dentro la seconda e terza condizione di ortogonalità. Di seguito possiamo vedere come, aggiungendo una variabile, le formule si complicano:

$$\hat{b} = \frac{\sum_i x_i y_i \sum_i z^2{}_i - \sum_i z_i y_i \sum_i x_i z_i}{\sum_i x^2{}_i \sum_i z^2{}_i - (\sum_i x_i z_i)^2}$$

FORMULA della STIMA di $\hat{b}$

$$\hat{c} = \frac{\sum_i z_i y_i \sum_i x^2{}_i - \sum_i x_i y_i \sum_i x_i z_i}{\sum_i x^2{}_i \sum_i z^2{}_i - (\sum_i x_i z_i)^2}$$

FORMULA della STIMA di $\hat{c}$

Ora si può dimostrare, se le cinque ipotesi sono vere, che le varianze di $\hat{b}$ e di $\hat{c}$ sono:

$$var(\hat{b}) = \frac{\sigma^2}{\sum_i x_i^2 (1-\rho)}$$

FORMULA della VARIANZA OLS per $\hat{b}$

$$var(\hat{c}) = \frac{\sigma^2}{\sum_i z_i^2 (1-\rho)}$$

FORMULA della VARIANZA OLS per $\hat{c}$

E' importante tenere conto di tutte le correlazioni incrociate tra tutte le variabili esplicative. Notiamo che ρ è il coefficiente di correlazione lineare tra le variabili *X* e *Z* :

$$\rho = \frac{\sum_i x_i z_i}{\sqrt{\sum_i x_i^2 \sum_i z_i^2}}$$

COEFFICIENTE di CORRELAZIONE lineare

Compreso –1 < ρ < 1

fase3: test (le stesse del paragrafo 1.4.4)

La visione delle formule che abbiamo ricavato per gli stimatori OLS con due esplicative ci spingono ad approfondire alcuni temi di interesse. Ad esempio l'effetto sulle proprietà degli stimatori OLS di due particolari errori di specificazione del modello: l'omissione di variabili esplicative rilevanti e l'inclusione di variabili esplicative irrilevanti.

1.4.1 Omissione di variabili rilevanti

In generale, l'omissione dal modello di una variabile rilevante implica la distorsione (rispetto al parametro corrispondente nel modello "vero") dello stimatore OLS dell'effetto della esplicativa inclusa nell'equazione. Perciò per valutare un'economia è importante includere tutte le variabili necessarie che spiegano correttamente la realtà ed il modello di regressione lineare ci aiuta ad isolare l'effetto

singolo delle variabili esplicative rispetto la variabile dipendente, per cercare di capire il contributo parziale di ogni variabile d'interesse.

Ci si accorge se sono state correttamente incluse nel modello tutte le variabile necessarie guardando lo stimatore OLS:

- Se lo stimatore è BLUE, il modello è corretto, sono valide le ipotesi fatte e nell'esempio sopra riportato avremo:

  $E(\hat{b}) = b$

  $E(\hat{c}) = c$

- Se lo stimatore è distorto avremo invece:

  $E(\hat{b}) \neq b$

  questo è il caso di un modello in cui si specifica ad esempio solo una variabile esplicativa quando Y_t, nella realtà, è spiegata da due variabili esplicative

1.4.2 Inclusione di variabili irrilevanti

L'inclusione di una variabile irrilevante, al contrario, non implica alcuna distorsione degli stimatori OLS ma la perdita della proprietà dell'efficienza (varianza minima) quindi, ha una ricaduta sulla procedura di test di significatività dei parametri.
Se prendiamo l'esempio del modello sopra indicato, supponendo che la variabile *Z* sia irrilevante, otteniamo:

$E(\hat{b}) = b$ lo stimatore è BLUE e il modello è ancora corretto

$E(\hat{c}) = 0$ si esclude la variabile esplicativa che non serve

si dimostra che, anche se si includono troppe variabili esplicative, lo stimatore rimane corretto.

Tuttavia il prezzo da pagare, come abbiamo detto, è in termini di inefficienza in quanto si sprecano delle informazioni e si stimano dei parametri per niente e questo spreco ha poi effetti su $\text{var}(\hat{b})$ (varianza dello stimatore) che non è più quella minima. Se però *X e Z* non covariano tra loro non si hanno effetti su $\text{var}(\hat{b})$.
Per quanto detto a proposito della correlazione fra le variabili esplicative e vista la definizione degli errori standard delle stime OLS in un modello con più esplicative, le due precedenti conclusioni non valgono solo se i regressori esclusi (o inclusi e irrilevanti) sono ortogonali con quelli presenti nel modello.

1.4.3 Correlazione fra regressori - multicollinearità

In questo paragrafo prendiamo in considerazione l'effetto che la correlazione fra le variabili esplicative del modello esercita sul calcolo delle stime OLS. Dalla visione delle formule OLS per *b* e *c* si nota che la stima OLS di *b* è influenzata dalla presenza nel modello di x_i, analogamente per la stima OLS di *c*. Al limite, quando $x_i = z_i$ non posso stimare né *b* né *c*; inoltre, si dimostra che se x_i e z_i sono ortogonali, la precedente stima di *b* è ottenibile da una regressione semplice in cui la y_i è spiegata dalla sola x_i (in seguito definiremo questa come regressione parziale), analogamente per la stima di z_i.

I rischi di correlazioni spurie, tipici nelle scienze sociali in cui il piano sperimentale non è sotto controllo (ci si limita ad osservare le realizzazioni delle variabili di interesse), vengono affrontati mediante la specificazione di modelli con più variabili esplicative. In tal caso, il parametro della esplicativa x_i rappresenta il coefficiente di

correlazione parziale che misura l'effetto di x_i su y_i di cui non si dà conto con le altre variabili esplicative del modello (nel caso discusso in questa sede, z_i). Il punto è stato evidenziato con un esempio in cui la variabile dipendente "abilità di lettura" y_i di un gruppo di alunni elementari, viene spiegata prima da un modello in cui compare la sola variabile "taglia delle scarpe" x_i e, successivamente, tale modello è esteso per dare conto di un ulteriore effetto "età dell'allievo" z_i. Nel primo caso si assiste ad una tipica correlazione spuria: dato che l'età z_i è in parte la concausa sia di x_i sia di y_i, l'apparente relazione significativa fra x_i e y_i è spiegata dall'omissione dal modello di un esplicito effetto di età z_i.

La formula per la stima OLS del modello in cui compaiono sia x_i sia z_i, ricavata precedentemente, è replicata ora seguendo un percorso alternativo, basato su una successione di regressioni parziali: prima regredisco y_i su z_i e salvo i residui y*, poi regredisco x_i su z_i e salvo i residui x_i*, infine regredisco y* (la parte di y_i dopo che ho rimosso l'effetto di z_i) su x_i* (la parte di x_i dopo che ho rimosso l'effetto di z_i); come già detto, la stima OLS del parametro relativo a x_i* è esattamente la stessa di quella del parametro di x_i nel modello completo ma in questo modo si illustra, passo a passo, cosa succede quando si stimano i parametri di un modello con più esplicative: dato che x_i* e y* sono entrambe incorrelate per costruzione con z_i (si vedano le condizioni di minimizzazione dei residui del metodo OLS), la regressione di y* su x_i* collega la parte di y_i che non è correlata con z_i con la parte di x_i che non è correlata con z_i e, quindi, questo è il significato da attribuire alla stima del parametro *b* nella regressione

OLS completa. Ovviamente, la stima del legame fra x_i* e y_i* non sarà significativa.

Le argomentazioni relative al modello con due variabili esplicative possono essere generalizzate al caso di un modello lineare con k parametri. In tale situazione l'algebra delle matrici permette una compattazione della specificazione del modello lineare:

$Y_i = bX_i + \varepsilon$

in cui Y_i è il vettore (N x_i) della variabile dipendente, X_i è la matrice (N x_i k) delle k variabili esplicative, *b* è un vettore (k x_i) di parametri e ε è il vettore (N x_i) degli errori. Si dimostra che, se valgono le ipotesi del modello classico di regressione lineare, vale il teorema di Gauss-Markov: lo stimatore OLS dei parametri *b* e *c* è BLUE. Nel caso di k parametri (k>2), la seconda ipotesi di specificazione del modello lineare classico (la variabile esplicativa X_i è fissa) si scompone in due parti:

(2a) la matrice X_i dei regressori è non stocastica (fissa, cioè "esogena")

(2b) il rango di X_i è k (con k<N): ipotesi di rango di X_i pieno, cioè si suppone che non ci sia nessuna relazione lineare *esatta* fra le variabili (colonne) che compaiono in X_i (una colonna non può essere una combinazione lineare di un'altra colonna).

Quest'ultimo punto ci porta ad individuare un potenziale problema del modello lineare con più variabili esplicative: la multicollinearità. Al limite, quando rango(X_i)<k c'è collinearità perfetta fra le variabili del modello e quindi il problema di minimo del quadrato dei residui non ha soluzione. Normalmente, esiste sempre un po' di correlazione fra

alcune variabili esplicative del modello e, al crescere di questa correlazione, aumenta la varianza dello stimatore OLS di *b*, tutto ciò, a sua volta, evidenzia non significatività della stima dei parametri del modello; quando questa correlazione è molto elevata, si parla di multicollinearità. Intuitivamente, se fra le esplicative del modello ce ne sono alcune che "si assomigliano molto" (multicollinearità), allora sarà difficile stimare con precisione i singoli effetti (parametri) che le esplicative esercitano sulla dipendente e, quindi, le varianze degli stimatori saranno alquanto elevate (cioè gli intervalli di stima saranno ampi). La presenza di multicollinearità è diagnosticata quando:

(a) l'*R-quadro* è elevato e allo stesso tempo le *statistiche t*, di significatività di singoli parametri rispetto a zero, sono basse;

(b) se si elimina un solo regressore dal modello, l'R-quadro non scende di molto ma, se elimino congiuntamente tutti i regressori non significativi, il *test F* rifiuta l'ipotesi nulla.

I possibili rimedi sono:

- aumentare, quando possibile, l'informazione statistica N;
- vincolare alcuni parametri (seguendo indicazioni teoriche), cioè ridurre il numero dei parametri stimati.

In ogni caso, la multicollinearità è un problema di natura numerica: lo stimatore OLS è ancora BLUE anche in presenza di multicollinearità, ovviamente a patto che valgano ancora le ipotesi di specificazione del modello classico di regressione lineare.

1.4.4 Test F e R^2

In questo paragrafo consideriamo l'inferenza sui parametri del modello (test di significatività).

Le procedure di significatività di specifici valori dei parametri del modello si fondano sul *test t (di Student)* e sul *test F (di Fisher)*. L'ipotesi nulla del test t si riferisce ad un singolo parametro del modello, mentre col test F si verificano congiuntamente più ipotesi relative a più parametri del modello. Nel caso particolare di impiego di t e F per la verifica dello stesso (singolo) vincolo, il quadrato della statistica t è identica alla F della stessa regressione, e i loro valori di probabilità sono gli stessi. Sotto l'ipotesi nulla, il modello è vincolato perché i suoi parametri assumono gli specifici valori ipotizzati dall'ipotesi nulla, mentre sotto l'ipotesi alternativa il modello è non vincolato, in quanto nessuno dei suoi parametri è "costretto" ad assumere particolari valori; evidentemente la somma del quadrato dei residui del modello vincolato è maggiore (o, al limite, uguale) a quella del modello non vincolato. Anche se le ipotesi nulle dei due precedenti test possono essere le più svariate (solitamente basate su ipotesi parametriche suggerite dalla teoria economica), l'output della regressione di qualsiasi software econometrico (ad esempio, EViews) presenta automaticamente l'esito di specifici test:

(a) il test t di significatività rispetto a zero di tutti i parametri (presi uno per volta) del modello stimato;

(b) il test F della regressione, la cui ipotesi nulla è che tutti i parametri del modello (tranne la costante) siano congiuntamente nulli.

In quest'ultimo caso si tratta di un test di "qualità" del modello non vincolato appena stimato perchè lo si confronta con il modello (vincolato) più semplice in assoluto, in cui y è spiegata dalla sola costante pari alla sua media; per queste ragioni il test F della regressione corrisponde, secondo una particolare formula, al coefficiente di determinazione (R-quadro).

Test F

Per ottenere il valore critico della distribuzione di Fisher è necessario scegliere il livello di significatività, di solito del 5%, solo in una coda come si vede da figura 5, prendendolo dalle tavole dei limiti di significatività della distribuzione di Fisher; una volta trovato il valore critico (ad esempio in figura 5 è stato calcolato $F(1,4)_{5\%} = 7{,}7$) si confronta con la statistica e se il valore della t-statistic > del valore critico, si ricade nella zona di rifiuto dell'ipotesi nulla e viceversa .

$$\text{Distribuzione } F \text{ - Test unilaterale} \rightarrow \frac{\frac{RSS_v - RSS_{nv}}{v}}{\frac{RSS_{nv}}{N-K}} \sim F(V, N-K)$$

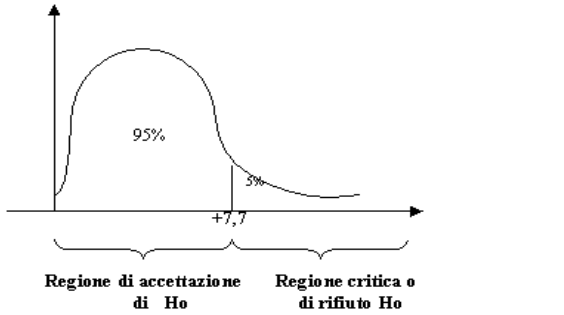

Fig. 5 - Test di significatività unilaterale (*test F*) con distribuzione F