

Compendio su stimatore OLS del modello di regressione classico

(Ad integrazione del libro di testo, dove il tutto è trattato in maniera meno ‘concentrata’, e delle slides, dove la trattazione è necessariamente scarna, data la difficoltà di inserirvi notazione formale)

Valerio Potì

17 ottobre 2020

QUESTE DISPENSE SONO ALLO STATO DI BOZZA

L'AUTORE SAREBBE MOLTO GRATO CON IL LETTORE PER QUALUNQUE
SUGGERIMENTO O COMMENTO

(Work-in-progress, comments and feed-back very welcome)

Indice

1. Introduzione.....	3
2. Regressione (elementi essenziali).....	7
2.1 Il modello di regressione.....	8
2.2 Stimatore OLS – caso univariato.....	12
3. Distribuzione dello stimatore OLS nel modello di regressione classico.....	17
3.1 Teorema di Gauss Markov.....	18
3.2 Normalità dello stimatore OLS nel modello di regressione semplice.....	19
4. Appendici.....	24
4.1 Teorema del Limite Centrale (Central Limit Theorem, CLT).....	25
4.2 Tipi di errore ed inferenza statistica.....	26
4.3 Intervalli di confidenza.....	27
4.4 Richiami di algebra lineare.....	28

1. Introduzione

Cos'è l'econometria?

È una disciplina che si occupa della misurazione di dati economici con due fini principali:

1. fare inferenza sulla validità di modelli economici (analisi induttiva)
2. aiutare a prendere decisioni di natura economico-finanziaria.

Per far ciò, nello studio di osservazioni empiriche, utilizza metodi statistici di stima e di verifica delle ipotesi. Un tema cruciale è il cosiddetto errore di campionamento.

Cos'è l'econometria applicata?

Due definizioni si incontrano più di frequente:

1. È la branca dell'econometria che si occupa in modo particolare dello studio delle serie storiche.
2. È la branca dell'econometria che si occupa in modo particolare del problema dell'**identificazione** dei modelli econometrici, cioè dell'individuazione delle relazioni di **causa ed effetto** che legano le variabili che appaiono in tali modelli quando tali relazioni debbano essere inferite sulla base di dati raccolti **al di fuori di esperimenti controllati e ripetibili**.

Chi ha ragione? Probabilmente, entrambi. La connessione sta nel fatto che l'identificazione delle relazioni di **causa ed effetto** è particolarmente difficile quando si lavora con serie storiche, perché in questo caso siamo davvero lontani dalle condizioni tipiche di esperimenti controllati e riproducibili/ripetibili.

Popolazione e campioni

La **popolazione** è l'insieme complessivo degli oggetti o individui che vengono studiati e sui cui caratteri si vogliono inferire relazioni di causa ed effetto e/o effettuare previsioni. Ad esempio, se siamo interessati a predire l'esito di una competizione elettorale, la popolazione oggetto di studio è l'elettorato nel suo complesso e il carattere che si vuole spiegare o prevedere è il voto.

La **distribuzione** dei caratteri oggetto di studio nella popolazione viene tipicamente descritta, per ciascun carattere, come la distribuzione di tante **variabili casuali** quanti sono i membri della

popolazione stessa. A loro volta, tali variabili casuali e le relazioni che le legano costituiscono il cosiddetto “**data generating process**” (DGP).

Un **campione** è una selezione di alcuni membri della popolazione.

Un **campione casuale** è un campione di cui tutti i membri della popolazione abbiano la stessa probabilità di far parte.

Tipi di dati

Semplificando, i dati con cui si lavora in econometria sono di due/tre tipi:

1. dati cross-sezionali, cioè dati su unità differenti raccolti nello stesso momento o in un momento concettualmente equivalente
2. serie storiche, cioè dati sulla stessa unità o variabile raccolti in momenti diversi e successivi
3. dati “panel”, ovvero un ‘ibrido’ tra i due tipi di dati appena elencati

Nel caso delle variabili casuali connesse a dati di natura cross-sezionale, ci si riferisce loro collettivamente con il termine di **popolazione**.

Nel caso delle serie temporali, quando tali variabili casuali siano viste in successione temporale, ci si riferisce loro collettivamente con il termine di **processo stocastico**.

Le osservazioni di entrambi i tipi di dati possono essere interpretate come realizzazioni di altrettante variabili casuali corrispondenti agli individui facenti parte del campione estratto dalla popolazione o dal processo oggetto di studio.

Il problema dell’identificazione

Soprattutto nella prima parte del corso, la nostra preoccupazione principale sarà l’identificazione delle relazioni che legano determinate variabili, nell’ambito di un processo teso all’inferenza di un modello econometrico che descriva il DGP dei fenomeni economici oggetto d’indagine.

Se, come tipicamente è il caso, le variabili oggetto di indagine non sono deterministiche, sono cioè variabili aleatorie o casuali, il nostro interesse può essere rivolto in due direzioni:

1. identificare la distribuzione multivariata (congiunta) delle variabili suddette

2. identificare la distribuzione condizionale di una (o più) di tali variabili date le altre

È nel secondo caso che ci occupiamo propriamente di problemi connessi al concetto di **causalità/causazione**. Infatti, come vedremo meglio in seguito, il concetto di correlazione (che è un aspetto della distribuzione congiunta di due variabili) è diverso dal concetto di causazione.

Per ora, è sufficiente sottolineare che il nostro obiettivo principale sarà quello di imparare a formulare inferenze su relazioni di causa ed effetto che posseggano i requisiti della **validità interna** ed **esterna**. Si dice che un'inferenza circa la relazione causale tra le variabili di un modello, così come prevista dal modello stesso, posseda **validità interna** quando la relazione stessa sia rigorosamente dimostrabile. La relazione causale si ritiene dimostrata quando la causa prevista dalla relazione in parola **preceda** l'effetto (ovvero quando le variazioni della variabile indipendente precedano quelle della variabile dipendente), questi siano in relazione l'uno con l'altro in maniera misurabile (ovvero quando esista una **co-varianza** statisticamente significativa tra variabile esplicativa e dipendente) e il nesso causale non sia **spurio** nel senso che, per tale relazione (ovvero per la covarianza osservata), non esistano spiegazioni alternative plausibili.

Si dice invece che le inferenze riguardo una relazione di causa-effetto basate su di uno studio, campione e circostanze particolari posseggono **validità esterna** quando possono essere generalizzate alla intera popolazione oggetto di studio. La maggiore minaccia alla generalizzabilità delle inferenze risultanti da uno studio o esperimento proviene dalla possibilità che una o più variabili indipendenti dipendano a loro volta da altri fattori, la cui influenza non è presa in considerazione o non è adeguatamente descritta dal modello oggetto di studio e può dar luogo ad effetti non osservati in circostanze in cui questi fattori assumano valori diversi da quelli prevalenti durante lo studio.

Per via del nostro interesse, in quanto econometrici, per il problema dell'identificazione della distribuzione condizionale, i modelli di **regressione** saranno i nostri principali strumenti di lavoro. Tali modelli specificano l'aspettativa condizionale di una variabile come funzione del valore assunto da altre variabili, dette regressori (condizionando appunto sul loro valore). Per semplicità, lavoreremo soprattutto con regressioni *lineari* nei coefficienti associati ai regressori.

2. Regressione (elementi essenziali)

2.1 Il modello di regressione

Preliminari: convenzioni sulla notazione. Nel seguito, si adotterà la convenzione di evidenziare ogni variabile vettoriale o matriciale con la formattazione in grassetto, ad esempio \mathbf{X} , \mathbf{x} o \mathbf{y} . In questo caso, si utilizzeranno le lettere minuscole per i vettori e le maiuscole per le matrici.

Regressione. Si consideri il modello seguente della relazione tra le variabili Y , \mathbf{X} e U che rappresentano caratteri degli individui (o elementi) di una popolazione (insieme) oggetto di studio:

$$Y = f(\mathbf{X}) + U$$

Qui sopra, Y e U sono scalari, f è una qualunque funzione a valori finiti (e reali), \mathbf{X} è in generale un vettore di dimensioni $k \times 1$. In tale modello, Y e U sono variabili casuali e \mathbf{X} può esserlo o meno. Quest'ultimo caso si ha quando le sue realizzazioni siano o possano essere considerate fisse in campionamenti ripetuti.

Assumendo di estrarre un campione dalla popolazione, si avranno T osservazioni sulle variabili Y , \mathbf{X} e U , ovvero y_t , \mathbf{x}_t e u_t , $t = 1, 2, \dots, T$. Nel seguito, interpreteremo ciascuna osservazione come la realizzazione di una variabile casuale e associeremo dunque una variabile casuale a ciascuna osservazione. Nella nostra notazione, per non complicarla oltre misura, non distingueremo però esplicitamente tra osservazione e variabile casuale di qui essa è una realizzazione, contando che la distinzione emerga dal contesto. Utilizzeremo dunque y_t , \mathbf{x}_t e u_t , $t \in \{1, 2, \dots, T\}$, per fare riferimento sia alle osservazioni che alle variabili aleatorie che le generano.

Si consideri ora il modello seguente:

$$y_t = f(\mathbf{x}_t) + u_t, \quad t = 1, 2, \dots, T$$

Qui sopra, y_t e u_t sono variabili casuali che generano le nostre osservazioni e, di nuovo, \mathbf{x}_t può esserlo o meno a seconda che \mathbf{x} lo sia o meno e, dunque, a seconda che, date le caratteristiche del metodo di campionamento, si possa ritenere che i valori assunti da \mathbf{x} possano essere considerati fissi in campionamenti ripetuti.

Quando l'aspettativa condizionale di u_t sia pari a quella non condizionale e questa sia pari a zero, cioè quando $E(u_t|\mathbf{x}_t) = E(u_t) = 0$, ci riferiamo al modello di cui sopra col termine di *regressione*. In tale modello, la variabile casuale u_t svolge il ruolo di termine di errore, che raccoglie tutte le influenze su y_t che, per la loro natura non sistematica, non si riesca o non valga la pena modellizzare esplicitamente tramite la loro inclusione come variabile esplicativa/regressore.

Questa condizione è anche nota come assunzione di *esogeneità* dei regressori e va interpretata come *condizione d'identificazione* del modello di regressione. In altre parole, se la condizione

$E(u_t|\mathbf{x}_t) = 0$ non vale, non ci troviamo in presenza di un modello di regressione ma bensì semplicemente di un generico modello che specifica y_t come la somma di una funzione di \mathbf{x}_t e di un'altra variabile u_t .

Inoltre, giacchè $E(u_t|\mathbf{x}_t) = 0$, abbiamo

$$E(y_t | \mathbf{x}_t) = f(\mathbf{x}_t)$$

e dunque, se le derivate parziali esistono,

$$dE(y_t|\mathbf{x}_t) = \frac{\partial f(\mathbf{x}_t)}{\partial x_{1,t}} dx_{1,t} + \dots + \frac{\partial f(\mathbf{x}_t)}{\partial x_{k,t}} dx_{k,t}$$

In numerose situazioni, non è particolarmente agevole calcolare le derivate nell'espressione di cui sopra, ed a volte non è facile derivarne un'espressione analitica che possa essere utilizzata in manipolazioni algebriche successive. Questo è uno dei motivi per cui, quando possibile, si ricorre ad un modello lineare.

Regressione lineare. Si consideri il modello seguente della relazione tra le variabili y e \mathbf{x} :

$$y = \alpha + \mathbf{x}'\boldsymbol{\beta} + u$$

Qui sopra, $\boldsymbol{\beta}$ è in generale un vettore di dimensioni $k \times 1$. Assumendo di estrarre un campione dalla popolazione, si avranno T osservazioni del tipo

$$y_t = \alpha + \mathbf{x}_t'\boldsymbol{\beta} + u_t$$

Quando l'aspettativa condizionale di u , e quindi di u_t , sia zero, cioè quando $E(u_t|\mathbf{x}_t) = E(u|\mathbf{x}) = 0$, ci riferiamo al modello di cui sopra col termine di regressione lineare. Data la linearità del modello, abbiamo che $\frac{\partial f(\mathbf{x}_t)}{\partial x_{i,t}} = \frac{\partial(\alpha + \boldsymbol{\beta}'\mathbf{x}_t)}{\partial x_{i,t}} = \beta_i$, dove β_i rappresenta l'elemento i -esimo del vettore $\boldsymbol{\beta}$, e dunque $dE(y_t|\mathbf{x}_t)$ diventa

$$\begin{aligned} dE(y_t|\mathbf{x}_t) &= \frac{\partial f(\mathbf{x}_t)}{\partial x_{1,t}} dx_{1,t} + \dots + \frac{\partial f(\mathbf{x}_t)}{\partial x_{k,t}} dx_{k,t} = \frac{\partial \boldsymbol{\beta}'\mathbf{x}_t}{\partial x_{1,t}} dx_{1,t} + \dots + \frac{\partial \boldsymbol{\beta}'\mathbf{x}_t}{\partial x_{k,t}} dx_{k,t} \\ &= \beta_1 dx_{1,t} + \dots + \beta_k dx_{k,t} \end{aligned}$$

Nella più utile versione discreta, quest'equazione ci dice che

$$\Delta E(y_t|\mathbf{x}_t) = \beta_1 \Delta x_{1,t} + \dots + \beta_k \Delta x_{k,t}$$

Dove

$$\beta_1 \cong \frac{\Delta E(y_t | \mathbf{x}_t)}{\Delta x_{1,t}} \Big|_{\Delta x_{1,t}=0, \dots, \Delta x_{k,t}=0}$$

$$\beta_2 \cong \frac{\Delta E(y_t | \mathbf{x}_t)}{\Delta x_{2,t}} \Big|_{\Delta x_{1,t}=0, \Delta x_{3,t}=0, \dots, \Delta x_{k,t}=0}$$

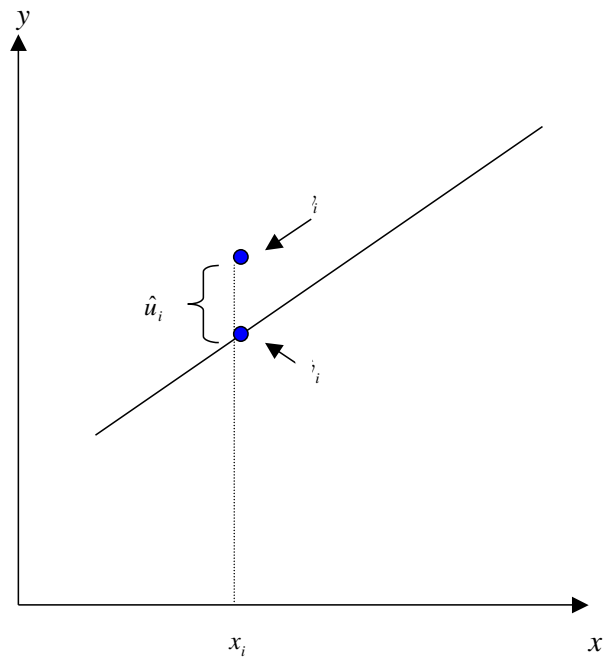
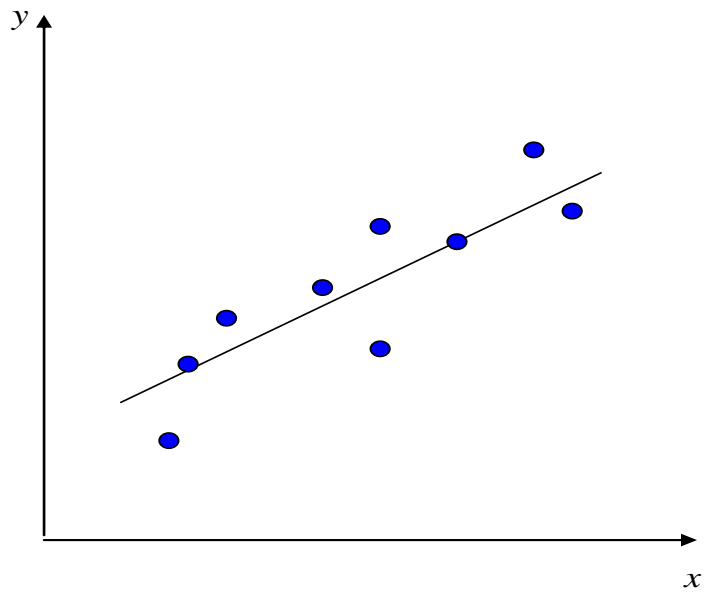
Specificazione. Il modello di regressione della popolazione (population regression function, PRF) è una descrizione del DGP. Un esempio univariato è $y = \alpha + \beta x + u$, il quale genera i campioni $y_t = \alpha + \beta x_t + u_t$.

Il modello di regressione campionario (sample regression function, SRF) è invece una stima del PRF, come per esempio $\hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$ nel qual caso abbiamo anche che $\hat{u}_t = y_t - \hat{y}_t$, oppure per esteso $y_t = \hat{\alpha} + \hat{\beta}x_t + \hat{u}_t$.

Nell'analisi di regressione, ci si riferisce col termine di *specificazione* al processo che porta a convertire una teoria in un modello di regressione identificato dall'ortogonalità tra le variabili ritenute esplicative e il termine di errore, ovvero dalla circostanza che $E(u_t | \mathbf{x}_t) = E(u | \mathbf{x}) = 0$. La finalità dell'analisi è quella di sottoporre a verifica una teoria che descrive il DGP di una o più variabili.

Tipicamente, usiamo la SRF per inferire la PRF e quindi il DGP. Per procedere nelle nostre inferenze, ci serviamo del concetto cruciale di errore di campionamento, grazie al quale possiamo sottoporre a verifica empirica ipotesi riguardo ipotesi circa la PRF e dunque specifici aspetti del DGP.

Uno stimatore particolarmente usato è quello dei *minimi quadrati* (Ordinary Least Square, *OLS*), che viene descritto di seguito. Questo metodo specifica una particolare SRF e un modo particolare di inferire la PRF.



2.2 Stimatore OLS – caso univariato

Date T osservazioni sulle variabili y e x , lo stimatore dei minimi quadrati (ovvero OLS, come si è già detto) si ottiene scegliendo i coefficienti α e β in modo da minimizzare la somma dei quadrati dei residui del modello (RSS, ovvero Residual Sum of Squares):

$$\min_{\hat{\alpha}, \hat{\beta}} \sum \hat{u}_t^2 \quad \hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$$

Minimizzare questa quantità è equivalente a minimizzare le deviazioni della variabile dipendente (*regressand*) dai valori ‘previsti’ dal modello:

$$\min_{\hat{\alpha}, \hat{\beta}} \sum (y_t - \hat{y}_t)^2 \quad \hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$$

Facendo riferimento a quest’ultima quantità come funzione obiettivo nella minimizzazione, poniamo:

$$L = \sum_t (y_t - \hat{y}_t)^2 = \sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t)^2$$

Volendo minimizzare questa funzione obiettivo rispetto ad $\hat{\alpha}$ e $\hat{\beta}$, imponiamo le FOCs:

$$\frac{\partial L}{\partial \hat{\alpha}} = -2 \sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \quad (1)$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2 \sum_t x_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \quad (2)$$

La FOC in (1) implica $\sum \hat{u}_t = 0$ ed anche

$$\sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0 \Leftrightarrow \sum_t y_t - \sum_t \hat{\alpha} - \sum_t \hat{\beta}x_t = 0 \Leftrightarrow \sum_t y_t - T\hat{\alpha} - \hat{\beta} \sum_t x_t = 0$$

Ma, denotando con \bar{y} la media aritmetica delle osservazioni sulla variabile dipendente, possiamo scrivere

$$\sum y_t = T\bar{y}$$

e inoltre, procedendo allo stesso modo:

$$\sum x_t = T\bar{x}$$

Quindi possiamo scrivere (1) come segue:

$$T\bar{y} - T\hat{\alpha} - T\hat{\beta}\bar{x} = 0$$

E dunque, dividendo entrambi i lati per T e risolvendo per \bar{y} , otteniamo

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x} \quad (3)$$

Inoltre, possiamo riscrivere (2) come segue:

$$\sum_t x_t \underbrace{(y_t - \hat{\alpha} - \hat{\beta}x_t)}_{\hat{u}_t} = 0 \quad (4)$$

Le equazioni (3) e (4) sono molto importanti. Vengono chiamate “**equazioni normali**” e rappresentano le condizioni che il modello di regressione impone al DGP dei dati.

La prima di tali equazioni implica che una retta di regressione passa per il ‘punto medio’ e può anche essere riscritta come segue, offrendo uno stimatore dell’intercetta α in funzione dello stimatore del coefficiente β :

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (5)$$

La seconda di tali equazioni può essere riscritta come segue:

$$\sum_t x_t \hat{u}_t = 0$$

Così riscritta, consente di dimostrare che la regressione fa sì che i residui, per costruzione, siano incorrelati con i regressori (ortogonali). Infatti, poiché abbiamo anche già dimostrato (per conseguenza della prima FOC) che $\sum \hat{u}_t = 0$, segue che

$$\begin{aligned} Cov_T(x, \hat{u}) &\equiv \frac{1}{T} \sum_t \left(x_t - \frac{1}{T} \sum_t x_t \right) \left(\hat{u}_t - \frac{1}{T} \sum_t \hat{u}_t \right) = \frac{1}{T} \sum_t \left(x_t - \frac{1}{T} \sum_t x_t \right) \hat{u}_t \\ &= \frac{1}{T} \sum_t x_t \hat{u}_t - \left(\frac{1}{T} \sum_t \hat{u}_t \right) \left(\frac{1}{T} \sum_t x_t \right) = 0 \end{aligned}$$

Per completare la derivazione dello stimatore di β , possiamo usare (5) per sostituire $\hat{\alpha}$ in (4) e così ottenere:

$$\begin{aligned} &\sum_t x_t \left(y_t - \underbrace{\hat{\alpha}}_{\bar{y} - \hat{\beta}\bar{x}} - \hat{\beta}x_t \right) = 0 \\ &\Leftrightarrow \sum_t x_t (y_t - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_t) = 0 \\ &\Leftrightarrow \sum_t x_t y_t - \bar{y} \sum_t x_t + \hat{\beta}\bar{x} \sum_t x_t - \hat{\beta} \sum_t x_t^2 = 0 \\ &\Leftrightarrow \sum_t x_t y_t - T\bar{y}\bar{x} + \hat{\beta}T\bar{x}^2 - \hat{\beta} \sum_t x_t^2 = 0 \end{aligned}$$

L’ultima di queste equazioni può essere riscritta come segue:

$$\hat{\beta}(T\bar{x}^2 - \sum x_t^2) = T\bar{y}\bar{x} - \sum x_t y_t$$

E quindi può essere agevolmente risolta per $\hat{\beta}$,

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2}$$

A sua volta, l'espressione sulla destra può essere riscritta in maniera differente, dandoci una formulazione dello stimatore più facile da ricordare. Per far ciò, per prima cosa notiamo quanto segue:

$$\begin{aligned}\sum x_t \bar{y} &= \bar{y} \left(\sum x_t \right) = \bar{y}(T\bar{x}) = T\bar{x}\bar{y} \\ \sum \bar{x} y_t &= \bar{x} \left(\sum y_t \right) = \bar{x}T\bar{y} = T\bar{x}\bar{y}\end{aligned}$$

Sottraendo ed aggiungendo $T\bar{x}\bar{y}$ al numeratore della formula per $\hat{\beta}$, ciò ci consente di riscriverlo nel modo seguente:

$$\begin{aligned}\sum x_t y_t - T\bar{x}\bar{y} &= \sum x_t y_t - T\bar{x}\bar{y} - T\bar{x}\bar{y} + T\bar{x}\bar{y} = \sum x_t y_t - \sum x_t \bar{y} - \sum \bar{x} y_t + T\bar{x}\bar{y} \\ &= \sum x_t y_t - \sum x_t \bar{y} - \sum \bar{x} y_t + \sum \bar{x}\bar{y} = \sum (x_t y_t - x_t \bar{y} - \bar{x} y_t + \bar{x}\bar{y}) \\ &= \sum [(x_t - \bar{x})(y_t - \bar{y})]\end{aligned}$$

E per quanto riguarda il denominatore della formula per $\hat{\beta}$, abbiamo quanto segue:

$$\begin{aligned}\sum x_t^2 - T\bar{x}^2 &= \sum x_t^2 - 2T\bar{x}^2 + T\bar{x}^2 = \sum x_t^2 - 2\bar{x}(T\bar{x}) + \sum \bar{x}^2 = \sum x_t^2 - 2\bar{x}\sum x_t + \sum \bar{x}^2 \\ &= \sum (x_t^2 - 2x_t \bar{x} + \bar{x}^2) = \sum (x_t - \bar{x})^2\end{aligned}$$

Dunque, possiamo riscrivere $\hat{\beta}$ come segue:

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{\sum [(x_t - \bar{x})(y_t - \bar{y})]}{\sum (x_t - \bar{x})^2}$$

Dividendo numeratore e denominatore per l'ampiezza del campione, abbiamo dunque che

$$\hat{\beta} = \frac{\frac{1}{T} \sum [(x_t - \bar{x})(y_t - \bar{y})]}{\frac{1}{T} \sum (x_t^2 - \bar{x}^2)} \equiv \frac{Cov_T(x, y)}{Var_T(x)}$$

Questa versione della formula dello stimatore evidenzia che lo stimatore stesso è un rapporto di momenti campionari secondi (covarianza e varianza per quanto riguarda, rispettivamente, il numeratore e denominatore) delle variabili x e y .

Queste formule diventano molto più semplici quando tutte le variabili hanno media campionaria nulla, come quando ne sottraiamo la media campionaria prima di usarle nella specificazione del modello di regressione e nella stima dei parametri dello stesso:

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\frac{1}{T} \sum x_t y_t}{\frac{1}{T} \sum x_t^2}$$

Una buona notizia è che, pur lavorando con variabili cui avessimo sottratto la media campionaria, potremmo poi facilmente risalire ai coefficienti del modello originario. Infatti, si consideri il modello

$$y_t - \bar{y} = a + b(x_t - \bar{x}) + e_t$$

E lo si riscriva ponendo

$$\tilde{y}_t = y_t - \bar{y}$$

ed inoltre

$$\tilde{x}_t = x_t - \bar{x}$$

Si ha dunque che

$$\tilde{y}_t = a + b\tilde{x}_t + e_t$$

Quindi, applicando la formula dello stimatore OLS del coefficiente di regressione ottenuta per il caso in cui le variabili abbiano media campionaria nulla, si ottiene

$$\hat{b} = \frac{\frac{1}{T} \sum \tilde{x}_t \tilde{y}_t}{\frac{1}{T} \sum \tilde{x}_t^2} = \frac{\frac{1}{T} \sum [(x_t - \bar{x})(y_t - \bar{y})]}{\frac{1}{T} \sum (x_t - \bar{x})^2} = \frac{Cov_T(x, y)}{Var_T(x)} = \hat{\beta}$$

Si noti inoltre che, per conseguenza della prima FOC, dovrà necessariamente darsi che la media dei residui debba essere nulla, ovvero $\frac{1}{T} \sum \hat{e}_t = 0$. Quindi, poiché la media di \tilde{x}_t e di \tilde{y}_t è anche essa nulla, ne consegue che l'intercetta debba anche essa essere nulla, ovvero $\hat{a} = 0$. Dunque, la retta di regressione stimata di variabili con media nulla passa necessariamente per l'origine. Per recuperare l'intercetta del modello di regressione delle variabili originarie (ovvero quelle date prima che ne sottraessimo la media), lo potremmo ottenere usando la equazione (5), ovvero

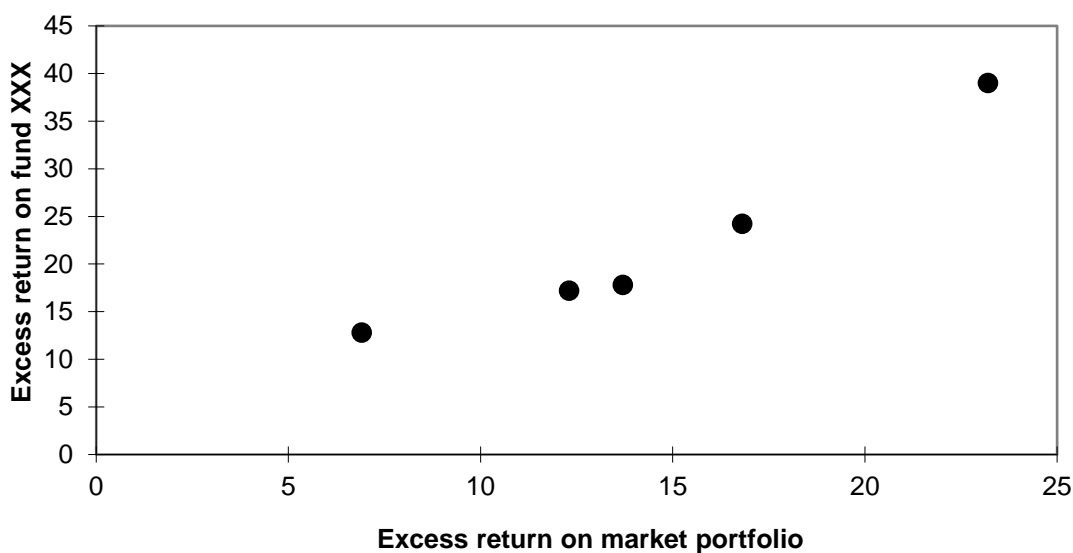
$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Semplice esempio su OLS

Si supponga di avere i dati seguenti sui rendimenti del fondo XXX e dell'indice di mercato (dati in percentuale) in eccesso al rendimento dell'attività priva di rischio:

Year, t	Excess return $= r_{XXX,t} - r_{f_t}$	Excess return on market index $= r_{m_t} - r_{f_t}$
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

Graficamente,



Qual è il beta di questo fondo? E il manager, se la cava?

Usando le formule derivate più sopra, abbiamo:

$$\hat{\alpha} = -1.74 \quad \hat{\beta} = 1.64$$

L'intercetta negativa suggerisce una performance del manager non proprio brillante, ma prima di rivolgerci alla società di gestione in malo modo dovremmo verificare se la stima puntuale di tale parametro è negativa in misura statisticamente significativa.

In ogni caso, possiamo usare le nostre stime per porci domande del tipo: “se ci aspettassimo un rendimento in eccesso a quello dell'attività senza rischio del 20%, cosa potremmo aspettarci in termini di rendimento (sempre in eccesso a quello dell'attività priva di rischio) di questo fondo?”.

La risposta, in termini percentuali, sarebbe:

$$\hat{y}_i = -1.74 + 1.64 \times 20 = 31.06$$

3. Distribuzione dello stimatore OLS nel modello di regressione classico

3.1 Teorema di Gauss Markov.

Perchè OLS si usa tanto? Perchè, a certe condizioni, è BLUE (Best, Linear, Unbiased Estimator). Il teorema di Gauss-Markov, così chiamato in onore dei matematici Carl Friedrich Gauss e Andrej Markov, afferma infatti che, in un modello lineare in cui i disturbi abbiano valore atteso nullo e siano incorrelati e omoschedastici, gli stimatori lineari corretti efficienti sono quelli ottenuti con il metodo dei minimi quadrati.

Il teorema di Gauss-Markov si focalizza dunque sulla proprietà dell'efficienza. Le altre proprietà, per cui lo stimatore OLS è BLUE, si dimostrano facilmente.

Si ricorda che per stimatore efficiente si intende quello con errore quadratico medio (Mean Square Error o MSE) più ridotto. Il MSE, a sua volta, è definito come segue:

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] = Var(\hat{\beta}) + [E(\hat{\beta}) - \beta]^2$$

L'errore quadratico medio è dunque uguale alla somma della varianza e del quadrato della distorsione, o bias, di uno stimatore.

Nel contesto dell'enunciato del teorema di Gauss-Markov, che esplicitamente richiama il fatto che lo stimatore OLS sia corretto e quindi non distorto, dire che lo stimatore è efficiente equivale ad affermare che sia quello con varianza minima.

Le condizioni seguenti sono sufficienti, anche se non necessarie, perchè OLS sia BLUE:

1. $E(u_t) = 0$
2. $Var(u_t) = \sigma^2 < \infty$
3. $Cov(u_i, u_j) = 0$
4. La matrice X è non-stocastica, cioè è fissa in campionamenti ripetuti
5. $u_t \sim N(0, \sigma^2)$

Queste condizioni sono anche note come assunzioni del modello di regressione classico (CLRM assumptions).

3.2 Normalità dello stimatore OLS nel modello di regressione semplice

In presenza delle condizioni del modello di regressione classico e assumendo per comodità che le variabili abbiano media campionaria nulla, lo stimatore OLS del coefficiente della variabile dipendente in un modello di regressione semplice ha la distribuzione seguente:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{t=1}^T x_t^2}\right)$$

Dimostrazione.

Quanto sopra è chiaramente un caso speciale del risultato che si è già dimostrato nel caso multivariato, ovvero della regressione multipla. Qui di seguito, lo dimostreremo però in maniera alternativa, senza fare ricorso all'algebra matriciale.

Allora, riordiamo che

$$\hat{\beta} = \frac{Cov_T(x, y)}{Var_T(x)}$$

Nell'usare la formula di cui sopra, teniamo presente però che la x è, per assunzione, non stocastica e dunque utilizziamo i momenti campionari che appaiono al numeratore e denominatore senza interpretarli come stimatori dei momenti della distribuzione di queste variabili (dunque, di fatto, solo per convenienza notazionale).

Abbiamo dunque

$$\begin{aligned}\hat{\beta} &= \frac{Cov_T(y, x)}{Var_T(x)} = \frac{Cov_T(\alpha + \beta x + u, x)}{Var_T(x)} = \frac{Cov_T(\beta x, x) + Cov_T(u, x)}{Var_T(x)} \\ &= \frac{\beta Cov_T(x, x) + Cov_T(u, x)}{Var_T(x)} \\ &= \frac{\beta Var_T(x) + Cov_T(u, x)}{Var_T(x)} = \frac{\beta Var_T(x)}{Var_T(x)} + \frac{Cov_T(u, x)}{Var_T(x)}\end{aligned}$$

$$\begin{aligned}
&= \beta + \frac{Cov_T(u, x)}{Var_T(x)} \\
&= \beta + \frac{\sum_{t=1}^T u_t(x_t - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2}
\end{aligned}$$

Ricordando che abbiamo assunto (per semplicità) di lavorare con variabili con media campionaria nulla, si ha dunque

$$\begin{aligned}
\hat{\beta} &= \beta + \frac{\sum_{t=1}^T u_t x_t}{\sum_{t=1}^T x_t^2} = \beta + \sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) u_t \\
E(\hat{\beta}) &= E \left(\beta + \sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) u_t \right) = \beta + E \left(\sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) u_t \right) \\
&= \beta + \sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) E(u_t) = \beta + E(u_t) \sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) \\
&= \beta + 0 \times \sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) = \beta
\end{aligned}$$

Quindi, poiché abbiamo assunto che la variabile esplicativa (ovvero x) sia non stocastica, abbiamo che $E \left(\sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) u_t \right) = \sum_{t=1}^T \frac{x_t}{\sum_{t=1}^T x_t^2} E(u_t) = E(u_t) \sum_{t=1}^T \frac{x_t}{\sum_{t=1}^T x_t^2}$. Poiché abbiamo anche assunto che $E(u) = 0$, segue dunque che $E \left(\sum_{t=1}^T \left(\frac{x_t}{\sum_{t=1}^T x_t^2} \right) u_t \right) = E(u_t) \sum_{t=1}^T \frac{x_t}{\sum_{t=1}^T x_t^2} = 0$. Quindi, abbiamo che

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

Per quanto riguarda la normalità, segue dal fatto che $\sum_{t=1}^T u_t x_t$ non è altro che una combinazione lineare degli errori (si ricordi di nuovo che abbiamo assunto la non stocasticità di x).

L'errore standard, infine, è dato da

$$\begin{aligned}
Var(\hat{\beta}) &= Var \left(\frac{\sum_{t=1}^T u_t x_t}{\sum_{t=1}^T x_t^2} \right) = \frac{\sum_{t=1}^T x_t^2 Var(u_t)}{(\sum_{t=1}^T x_t^2)^2} = \frac{\sum_{t=1}^T x_t^2 \sigma^2}{(\sum_{t=1}^T x_t^2)^2} = \frac{\sum_{t=1}^T x_t^2}{(\sum_{t=1}^T x_t^2)^2} \sigma^2 \\
&= \frac{1}{\sum_{t=1}^T x_t^2} \sigma^2 = \frac{\sigma^2}{\sum_{t=1}^T x_t^2}
\end{aligned}$$

La seconda di queste eguaglianze è, di nuovo, giustificata dal fatto che la variabile esplicativa (ovvero x) sia non stocastica e la terza dal fatto che la varianza degli errori sia assunta uguale per ogni osservazione e pari ad un dato σ^2 (ovvero, $Var(u_t) = \sigma^2 \forall t \in \{1, 2, \dots, T\}$).

Q.E.D.

Quanto sopra implica che la seguente statistica può essere utilizzata per sottoporre a verifica, col cosiddetto test z , l'ipotesi nulla che il coefficiente β sia pari a un valore dato, poniamo β_0 , contro l'ipotesi alternativa che sia diverso (in un test cosiddetto a “due code” o bilaterale) oppure maggiore o minore (in un test cosiddetto a “una coda” o unilaterale):

$$z(\hat{\beta}_i) = \frac{\hat{\beta}_i - \beta_0}{\sqrt{\sigma^2 / \sum_{t=1}^T x_t^2}}$$

Abbiamo infatti che $z(\hat{\beta}_i)$ rappresenta una variabile normale standardizzata, a motivo della normalità di $\hat{\beta}_i$ e del fatto che $\hat{\beta}_i$ abbia valore atteso pari a $\beta_{i,0}$, secondo l'ipotesi nulla, e varianza pari a $\frac{\sigma^2}{\sum_{t=1}^T x_t^2}$. Si ha dunque che

$$z(\hat{\beta}_i) = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\sigma^2 / \sum_{t=1}^T x_t^2}} \sim N(0,1)$$

Più nel dettaglio, se l'ipotesi sottoposta a verifica è data dall'ipotesi nulla

$$H_0 : \beta_i = \beta_{i,0}$$

contro l'ipotesi alternativa

$$H_1 : \beta_i \neq \beta_{i,0},$$

il test z richiede che si confronti la statistica $z(\hat{\beta}_i)$ oppure il suo valore assoluto, a seconda che il test sia a una o due code, con il valore critico z_α che, se l'ipotesi nulla fosse vera, verrebbe eguagliato o ecceduto dalla statistica stessa con probabilità pari ad α , e si rifiuta H_0 quando ciò avvenga. Dunque, in un test a due code, si rifiuta H_0 quando

$$|z(\hat{\beta}_i)| \geq z_\alpha$$

La probabilità α è nota come il livello di significatività del test (si veda l'Appendice per una definizione più dettagliata).

In pratica, per costruire la statistica di cui sopra, ci serve uno stimatore della varianza dei termini di errore nella 'popolazione', ovvero di σ^2 , poichè tale quantità non è in generale nota. Lo stimatore OLS è

$$\hat{\sigma}^2 = \frac{1}{T-k} \sum_{t=1}^T \hat{u}_t^2 \equiv s^2$$

In quest'espressione, cui spesso ci si riferisce con il simbolo s^2 (oppure, in alcuni testi, semplicemente s), k è il numero di regressori (a tal riguardo si ricordi che si assume di lavorare con variabili con media campionaria nulla e dunque non vi è l'intercetta). Quando si adotti un tale stimatore della varianza delle stime OLS, si può usare la seguente statistica per sottoporre a verifica l'ipotesi nulla che il coefficiente dell' i -esimo regressore sia $\beta_{i,0}$:

$$t(\hat{\beta}_i) = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\hat{\sigma}^2 / \sum_{t=1}^T x_t^2}} \sim t_{T-k}$$

Si noti che la statistica di cui sopra ha una distribuzione t di Student con $T - k$ gradi di libertà (non si fornisce una dimostrazione per ragioni di spazio e di snellezza della discussione, ma l'idea centrale è che questa è la distribuzione di un rapporto tra una normale standardizzata e la radice quadrata di una variabile con una distribuzione Chi-quadrato).

Di conseguenza, il seguente è un intervallo di confidenza dell' $(1 - \alpha) \times 100$ per cento 'a due code' intorno a β_i , ovvero l'intervallo di valori alla destra e alla sinistra di $\hat{\beta}_{i,T}$ che contiene β_i con probabilità $1 - \alpha$:

$$\hat{\beta}_{i,T} - t_\alpha s \leq \beta_i \leq \hat{\beta}_{i,T} + t_\alpha s$$

Qui t_α denota il valore critico della distribuzione t di Student della statistica $t(\hat{\beta}_{i,T})$ e, come spiegato appena sopra, $\sqrt{s^2} = \sqrt{\hat{\sigma}^2 / \sum_{t=1}^T x_t^2}$. Poichè l'intervallo è 'a due code', questo è il valore per cui $\Pr\{|t| \leq t_\alpha\} = \alpha$ ovvero $\Pr\{z \leq t_\alpha\} = \alpha/2$ e $\Pr\{t \geq t_\alpha\} = \alpha/2$. Stabilire se $\beta_{i,0}$, cioè il valore del parametro nell'ipotesi nulla, sia all'interno di questo intervallo consente di sottoporre a verifica l'ipotesi nulla stessa. Diciamo infatti che, se $\beta_{i,0}$ risulta al di fuori dell'intervallo di confidenza, rigettiamo la ipotesi nulla $H_0: \beta_i = \beta_{i,0}$ al livello di significatività α .

In maniera del tutto equivalente, ma forse più pratica, possiamo calcolare la probabilità del valore campionario della statistica, ovvero di $\hat{\beta}_{i,T}$, utilizzando la distribuzione di probabilità valida nell'ipotesi nulla, e rifiutare quest'ultima se la probabilità risulti minore del livello di significatività α prefissato, poniamo il 5%.

4. Appendici

4.1 Teorema del Limite Centrale (Central Limit Theorem, CLT)

Teorema. La media campionaria di T variabili aleatorie indipendentemente ed identicamente distribuite, ovvero i.i.d., con valore atteso μ e varianza σ^2 finita, converge in distribuzione (quindi, quando T sia ‘grande’) ad una normale con valore atteso μ e varianza pari a $\frac{\sigma^2}{T}$. Tra l’altro, ciò implica anche che la media campionaria converge in probabilità al valore atteso.

Cioè, un pò più formalmente, la successione di medie aritmetiche $\bar{X}_T \equiv \frac{1}{T} \sum_{t=1}^T X_t$ di T variabili casuali X_i , ciascuna con valore atteso μ e varianza pari a σ^2 , ovvero $X_i \sim \phi(\mu, \sigma^2)$, converge in distribuzione, al crescere di T , ad una variabile aleatoria normalmente distribuita con lo stesso valore atteso e varianza $\frac{1}{T} \sigma^2$, ovvero $\bar{X} \sim N(\mu, \frac{1}{T} \sigma^2)$, dove $\bar{X} \equiv \frac{1}{T} \sum_{t=1}^T X_t$.

Quanto sopra è il **Teorema del Limite Centrale** o Central Limit Theorem (CLT). La **legge dei grandi numeri** è una specie di generalizzazione al caso in cui non sia possibile fare l’assunzione che la varianza sia finita. In questo caso è solo possibile dire che la media campionaria converge in probabilità al valore atteso, ovvero che $\bar{X}_T \xrightarrow{p} \mu$.

Nota: l’assunto che la varianza σ^2 sia finita implica che anche il momento primo, cioè il valore atteso μ , lo sia.

4.2 Tipi di errore ed inferenza statistica

Errori del tipo I.

L'errore di tipo I nella verifica (test) di una ipotesi consiste nel rigettare l'ipotesi nulla quando questa sia vera, ovvero quando l'alternativa sia falsa.

Errori del tipo II.

L'errore di tipo II nella verifica (test) di una ipotesi consiste nel non rifiutare l'ipotesi nulla quando questa sia falsa, ovvero quando l'alternativa sia vera.

La probabilità prestabilita (o **nominale**) di incorrere in un errore del tipo I è nota come il **livello di significatività** del test e viene specificata prima di effettuare il test stesso.

La ipotesi nulla, infatti, viene rifiutata in favore dell'ipotesi alternativa quando la statistica utilizzata nel test, che è funzione esclusivamente dei dati osservati, assuma un valore compreso in un intervallo di valori cui corrisponda una probabilità cumulata pari al livello prestabilito di significatività del test secondo la distribuzione teorica dei valori che la statistica stessa, data l'ipotesi nulla, possa assumere.

Il **valore critico** della statistica utilizzata nel test è il valore della stessa per cui il test rifiuta la ipotesi nulla ad un determinato livello di significatività. L'insieme dei valori di tale statistica per cui il test rigetta l'ipotesi nulla è noto come **regione di rifiuto**, mentre l'insieme dei valori per cui il test non rigetta l'ipotesi nulla è noto come **regione di accettazione**. La probabilità **vera** di incorrere in un errore del tipo I è nota come il **livello minimo** (o *size*) del test. La probabilità **vera** di non incorrere in un errore del tipo II è nota come la **potenza** (o *power*) del test.

4.3 Intervalli di confidenza

Definizione.

L'intervallo di confidenza di livello $1 - \alpha$ di un parametro θ che descrive la distribuzione di un determinato carattere in una popolazione è dato dall'insieme di valori ipotizzabili per tale parametro che non siano rifiutabili da un test a due code al livello di significatività α .

Esempio.

Abbiamo visto che la media campionaria di T variabili i.i.d., con valore atteso μ e varianza σ^2 finita, converge in distribuzione ad una normale con valore atteso μ e varianza pari a $\frac{\sigma^2}{T}$. Dunque, supponendo che i risultati di un certo esperimento costituiscano un campione di variabili aleatorie i.i.d. $x_i \sim \phi(\mu, \sigma^2)$, dove $\phi(\cdot)$ denota una distribuzione arbitraria con momenti secondi finiti e $i \in [1, 2, \dots, T]$, e che la media campionaria dei risultati stessi sia \bar{x} , l'intervallo di confidenza al 95% del valore atteso μ è dato da $|\mu - \bar{x}| \leq 1.96 \frac{\sigma}{\sqrt{T}}$ ovvero

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{T}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{T}}$$

Probabilità di copertura.

Data un certo intervallo di confidenza di livello $1 - \alpha$ di un parametro θ che descrive la distribuzione di un determinato carattere in una popolazione, la probabilità di copertura è la probabilità che l'intervallo in parola contenga il vero valore di θ . Dunque, un 'buon' test deve avere una probabilità di copertura uguale al livello di confidenza nominale $1 - \alpha$ (o da questo non troppo dissimile).

Ciò si verifica quando tutte le assunzioni alla base della definizione dell'intervallo di confidenza sono valide. Se alcuna di queste assunzioni non lo è, la probabilità di copertura può essere maggiore o minore del livello di confidenza nominale (o, secondo un'espressione più colloquiale, della probabilità di copertura 'nominale') del test, ovvero $1 - \alpha$. Nel primo caso, si dice che l'intervallo e il relativo test sono 'conservativi', in quanto il test rigetta l'ipotesi nulla corretta più frequentemente del livello nominale α , mentre nel caso contrario si dice che sono 'permissivi'.

Ovvero, in un test conservativo, cosiddetto **livello minimo** o *size* del test (la probabilità vera di **errori del tipo I**) è più elevata del livello nominale α , e viceversa nel caso di un test permissivo.

4.4 Richiami di algebra lineare

Prodotto matriciale. Si consideri il prodotto matriciale:

$$AB = C$$

Qui, A , B e C sono tre matrici di dimensioni, rispettivamente, $m \times n$, $n \times l$ e $m \times l$. Tali matrici hanno elementi a_{ij} , b_{ij} , c_{ij} , dove i e j rappresentano, come di consueto, gli indici di riga e colonna. Cioè, si ha $A \equiv [a_{ij}]$, $B \equiv [b_{ij}]$ e $C \equiv [c_{ij}]$.

La matrice C , in virtù della definizione di prodotto matriciale, ha elementi:

$$c_{ij} = \sum_{s=1}^n a_{is}b_{sj}$$

Prodotti interni di matrici di dati

Si supponga di avere T osservazioni su k variabili. Si consideri il vettore riga \mathbf{x}_t' $1 \times k$ corrispondente alla t -esima tra tali osservazioni. Si organizzino le osservazioni allineando uno sull'altro i vettori riga in una matrice $T \times k$ di dati \mathbf{X} :

$$\mathbf{X} \equiv \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{T1} & \cdots & x_{Tk} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \dots \\ \mathbf{x}_T' \end{bmatrix}$$

Abbiamo quindi:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{T1} & \cdots & x_{Tk} \end{bmatrix}' \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{T1} & \cdots & x_{Tk} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \dots \\ \mathbf{x}_T' \end{bmatrix}' \begin{bmatrix} \mathbf{x}_1' \\ \dots \\ \mathbf{x}_T' \end{bmatrix} = \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t'$$

Si noti inoltre che il termine t -esimo di quest'ultima sommatoria può essere espresso come

$$\begin{aligned} \mathbf{x}_t \mathbf{x}'_t &\equiv [(\mathbf{x}_t \mathbf{x}'_t)_{ij}] = \left[\sum_{s=1}^1 x_{is} x'_{sj} \right] = [x_{i1} x'_{1j}] = [x_{i1} x_{j1}] = [x_i x_j] \\ &= \begin{bmatrix} x_1 x_1 & \cdots & x_1 x_k \\ \vdots & \ddots & \vdots \\ x_k x_1 & \cdots & x_k x_k \end{bmatrix} = \begin{bmatrix} x_1^2 & \cdots & x_1 x_k \\ \vdots & \ddots & \vdots \\ x_k x_1 & \cdots & x_k^2 \end{bmatrix} \end{aligned}$$

E per la sommatoria dei T termini abbiamo:

$$\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t = [(\mathbf{x}_1 \mathbf{x}'_1)_{ij}] + [(\mathbf{x}_2 \mathbf{x}'_2)_{ij}] + \cdots + [(\mathbf{x}_T \mathbf{x}'_T)_{ij}]$$

Diseguaglianza triangolare e diseguaglianza di Cauchy-Schwartz (intuizione)

Empiricamente, possiamo notare che la somma delle lunghezze di due lati qualsiasi di un triangolo non può essere minore della lunghezza del lato rimanente. Utilizzando la norma come misura della lunghezza di un vettore, abbiamo quindi:

$$\| \mathbf{v} - \mathbf{s} \| \leq \| \mathbf{v} \| + \| \mathbf{s} \| \quad \forall \mathbf{v} \text{ ed } \mathbf{s} \text{ appartenenti ad uno spazio vettoriale normato } R^n$$

Ma ciò deve valere anche per $\| -\mathbf{s} \| = \| \mathbf{s} \|$ e quindi

$$\| \mathbf{v} + \mathbf{s} \| = \| \mathbf{v} - (-\mathbf{s}) \| \leq \| \mathbf{v} \| + \| \mathbf{s} \| \quad \forall \mathbf{v} \text{ ed } \mathbf{s} \text{ appartenenti ad uno spazio vettoriale normato } R^n$$

Ovvero

$$\| \mathbf{v} + \mathbf{s} \|^2 \leq (\| \mathbf{v} \| + \| \mathbf{s} \|)^2$$

Quindi:

$$\| \mathbf{v} + \mathbf{s} \|^2 \leq \| \mathbf{v} \|^2 + \| \mathbf{s} \|^2 + 2 \| \mathbf{v} \| \| \mathbf{s} \|$$

E risolvendo per $\| \mathbf{v} \| \| \mathbf{s} \|$,

$$\| \mathbf{v} \| \| \mathbf{s} \| \geq (\| \mathbf{v} + \mathbf{s} \|^2 - \| \mathbf{v} \|^2 - \| \mathbf{s} \|^2) / 2$$

Per il termine in parentesi abbiamo:

$$\| \mathbf{v} + \mathbf{s} \|^2 - \| \mathbf{v} \|^2 - \| \mathbf{s} \|^2 = (v_1 + s_1)^2 + (v_2 + s_2)^2 + \dots + (v_k + s_k)^2 - \| \mathbf{v} \|^2 - \| \mathbf{s} \|^2$$

$$\begin{aligned}
&= (v_1^2 + v_2^2 + \dots + v_k^2) + (s_1^2 + s_2^2 + \dots + s_k^2) + 2 \mathbf{v}'\mathbf{s} - \|\mathbf{v}\|^2 - \|\mathbf{s}\|^2 \\
&= \|\mathbf{v}\|^2 + \|\mathbf{s}\|^2 + 2 \mathbf{v}'\mathbf{s} - \|\mathbf{v}\|^2 - \|\mathbf{s}\|^2 \\
&= 2 \mathbf{v}'\mathbf{s}
\end{aligned}$$

Quindi, sostituendo:

$$\|\mathbf{v}\| \|\mathbf{s}\| \geq (2 \mathbf{v}'\mathbf{s})/2 = \mathbf{v}'\mathbf{s}$$

Ovvero

$$\mathbf{v}'\mathbf{s} \leq \|\mathbf{v}\| \|\mathbf{s}\| \quad \text{disuguaglianza di Cauchy-Schwartz}$$