

# Appunti di Econometria

## ARGOMENTO [0]: COMMENTI INTRODUTTIVI

Tommaso Nannicini – Università Bocconi

Settembre 2009

### 1 Istruzioni per l'uso

Questi appunti hanno un carattere **sintetico** e **riepilogativo**: non pretendono di sostituire le argomentazioni affrontate a lezione, ma se ne pongono come un utile (speriamo) complemento. In particolare, gli aspetti formali (definizioni, teoremi, dimostrazioni) trattati a lezione saranno ripercorsi schematicamente in questi appunti, mentre intuizioni ed esempi saranno lasciati alla discussione in classe.

Per farla corta: gli **studenti frequentanti** possono prepararsi su una combinazione lineare di queste note, dei loro appunti presi a lezione, degli esercizi assegnati durante il corso e delle esercitazioni empiriche in aula informatica; gli **studenti non-frequentanti** possono prepararsi su una combinazione lineare di queste note, degli esercizi resi disponibili sul Learning Space del corso e di uno dei libri di testo indicati nel programma (limitatamente alle parti specificate).

### 2 Cos'è l'econometria

L'econometria ruota intorno all'interazione continua tra teoria economica, metodi statistici e dati. Il rapporto tra econometria e **teoria economica** è bidirezionale: la maggior parte delle volte, la teoria economica ci fornirà gli spunti necessari per specificare modelli o individuare ipotesi da testare rispetto alle variabili di interesse; ma esistono casi in cui si realizza un progetto econometrico per pura curiosità su relazioni poco studiate e i risultati finiscono poi per suggerire spunti alla teoria economica.

L'applicazione di **metodi statistici** a domande di natura economica è il cuore dell'econometria. Anche qui il rapporto non è univoco: l'econometria mutuata dalla statistica gli strumenti e i risultati necessari per identificare e stimare relazioni economiche<sup>1</sup>, ma nel corso del tempo la particolare natura dei dati e delle domande poste in economia ha portato gli econometrici a sviluppare contributi metodologici originali, che sono stati fatti propri anche da altre discipline. A questo riguardo, faremo a volte riferimento alla differenza tra modelli *macro-econometrici* (per i quali è particolarmente utile l'econometria delle serie storiche) e modelli *micro-econometrici* (per i quali sono particolarmente utili le tecniche di *program evaluation*).

Infine, i **dati** sono—ovviamente—una condizione ineludibile (necessaria ma non sufficiente) per ogni progetto di analisi empirica. Non solo: la natura dei dati oggetto di indagine pone vincoli importanti sui

---

<sup>1</sup>Il connotato preciso dei termini *identificare* e *stimare* risulterà chiaro in seguito. Si tratta di due concetti molto importanti!

metodi statistici che si possono (o debbono) utilizzare, e quindi anche sulle domande economiche cui si può dare credibilmente risposta. Per esempio, come vedremo, metodi statistici diversi sono più adatti per dati *cross-section* (che variano nello spazio tra unità di osservazione diverse), piuttosto che per serie storiche (che variano nel tempo per la stessa unità di osservazione) o per dati longitudinali (detti anche *panel*, che variano nello spazio e nel tempo). Ma c'è un'altra importante differenza da tenere presente: quella tra **dati sperimentali** e **dati osservazionali**. Con dati sperimentali, il ricercatore controlla direttamente certe variabili di interesse: per esempio, l'assegnazione di certe unità di osservazione a un trattamento (un nuovo medicinale) piuttosto che a un altro (la cura tradizionale). I dati osservazionali, invece, sono quelli che raccogliamo direttamente nel mondo reale, frutto delle migliaia di decisioni e interazioni degli agenti economici. In economia, difficilmente (anzi, quasi mai) si dispone di dati sperimentali, ma in molti casi è utile considerare una situazione sperimentale come termine di paragone, per capire quello che si può (o non si può) concludere con semplici dati osservazionali. I metodi statistici oggetto di studio in questo corso si riferiscono a dati osservazionali, sulla cui natura dovremo imporre assunzioni piuttosto stringenti.

Si pensi a questo semplice esempio di problema economico/econometrico:

$$y = f(x_2, x_3, x_4, x_5, x_6),$$

dove  $y$  è il salario,  $x_2$  l'età,  $x_3$  il sesso,  $x_4$  il livello d'istruzione,  $x_5$  le caratteristiche della famiglia di origine, e  $x_6$  la bellezza individuale. Ci proponremo di applicare metodi statistici a un campione di dati (raccolti sotto determinate condizioni) per testare la plausibilità di certe ipotesi su come le variabili esplicative  $x_j$  influiscono sulla variabile spiegata  $y$ . Una maggiore istruzione significa anche un maggiore salario nel mercato del lavoro? Lavoratori più belli ricevono salari più alti? Esiste una discriminazione salariale tra uomini e donne? A questo proposito, è già utile chiederci: tutte le possibili determinanti del salario sono incluse tra le nostre  $x_j$ ? Che dire dell'abilità innata (variabile impossibile da misurare)? Non averla inclusa tra le variabili esplicative ci creerà qualche problema? È probabile: per esempio, se l'abilità innata influisse sia sul livello d'istruzione sia sul livello del salario, la correlazione tra salario e istruzione rilevata nei dati avrebbe natura **spuria** piuttosto che **causale** (cioè, sarebbe interamente provocata dalla variabile abilità, anche in assenza di un vero impatto dell'istruzione sul salario). Ovviamente, la lista degli esempi sulla differenza tra correlazione spuria e rapporto di causalità è infinita. Nel corso delle lezioni, torneremo spesso su questi concetti. Di seguito, ci limitiamo a una formalizzazione puramente introduttiva, facendo uso della definizione di risultati potenziali.

### 3 Il metro di paragone “sperimentale”

L'esempio precedente sulla relazione (causale o spuria) tra istruzione e reddito può essere interpretato in termini di risultati potenziali. Il **risultato potenziale**, per ogni singola unità di osservazione, è il valore che una certa variabile (detta appunto risultato) acquista qualora si realizzi un valore specifico del **trattamento** (per esempio,  $T_i = 1$  se il lavoratore  $i$  è laureato,  $T_i = 0$  se il lavoratore non lo è):  $Y_{1i}$  è il reddito in caso di laurea, mentre  $Y_{0i}$  è il reddito in assenza di laurea. È come se ognuno di noi, in una serie di mondi paralleli in cui vivono nostri cloni con livelli di istruzione diversi, avesse un certo livello di reddito potenziale a seconda dell'istruzione ricevuta. Di conseguenza, il reddito osservato per il lavoratore  $i$  (nel nostro mondo) è:  $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})T_i$ . Se volessimo tentare (nel nostro mondo, non potendo osservare tutti gli altri) di stimare l'effetto causale dell'istruzione sul reddito, comparando (magari in un grande campione) la media condizionata del reddito a seconda del livello di istruzione, otterremmo:

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0] = \underbrace{E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1]}_{\text{effetto causale medio sui trattati}} + \underbrace{E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 0]}_{\text{distorsione da autoselezione}}.$$

In altri termini, il confronto tra risultati osservati di lavoratori con livelli diversi di istruzione ci fornisce una stima distorta dell'effetto causale medio, perché persone con risultati potenziali diversi (in assenza di trattamento) si auto-selezionano per ricevere il trattamento stesso. Per esempio, persone con abilità diversa, e quindi con risultati potenziali diversi anche senza laurea, decidono se andare all'università oppure no.

Se, per pura ipotesi, potessimo realizzare un **esperimento controllato**, obbligando persone diverse (es-tratte casualmente da una popolazione di riferimento) a prendere una laurea oppure no: allora, in dati del genere, potremmo stimare facilmente l'effetto causale dell'istruzione sul reddito. Infatti, in quel caso:

$$\begin{aligned}
 E[Y_i|T_i = 1] - E[Y_i|T_i = 0] &= E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 0] = \\
 &= \underbrace{E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1]}_{\text{effetto causale medio sui trattati}} = \underbrace{E[Y_{1i} - Y_{0i}]}_{\text{effetto causale medio}},
 \end{aligned}$$

dove la seconda eguaglianza è dovuta all'ipotesi di esperimento randomizzato, che fa sì che  $E[Y_{0i}|T_i = 0] = E[Y_{0i}|T_i = 1]$ . Ovviamente, nelle scienze sociali, esperimenti di questo tipo sono quasi impossibili da realizzare (per problemi etici, pratici o economici). Ma tenere a mente questo "metro di paragone" è utile per capire con quali tecniche econometriche sia possibile avvicinarsi maggiormente, sulla base dei dati disponibili e del problema economico oggetto di indagine, all'identificazione di un vero effetto causale.