CAPITOLO 20

Il problema della causalità.
Una introduzione generale e un esempio.

Andrea Ichino

5 Marzo, 2001

20.1 Introduzione

Nel corso di questo libro abbiamo incontrato numerosi esempi di correlazioni tra variabili economiche: tra istruzione e reddito (Capitolo 2), tra salario e offerta di lavoro (Capitolo 3), tra salario e domanda di lavoro (Capitolo 4), tra domanda del prodotto e occupazione (Capitoli 5 e 6), e l'elenco potrebbe continuare con tutti gli altri capitoli e con più di un esempio per capitolo. Per ciascuna di queste correlazioni, abbiamo visto teorie che suggeriscono interpretazioni causali. Abbiamo però anche visto teorie alternative per le quali quanto osservato non implica un legame tra una causa e un effetto, ma solo una correlazione spuria generata da altri fenomeni oppure una combinazione confusa di legami causali con segno opposto. Sia a fini positivi sia a fini normativi, è evidente quanto importante sia per l'economista del lavoro stabilire la reale natura di queste correlazioni. Tuttavia, solo una corretta e sovente non semplice analisi empirica dei dati può aiutarci a verificare se queste correlazioni implichino veramente un effetto causale e a misurarne con precisione il segno e l'intensità.

Nel capitolo precedente sono stati illustrati alcuni metodi statistici che, sotto certe condizioni, consentono di identificare e stimare il nesso causale che lega due variabili economiche in un certa popolazione, dato un campione di osservazioni estratte dalla popolazione stessa. L'attenzione era tuttavia concentrata sulla metodologia statistica più che sulla nozione di causalità. In questo capitolo, invece, la nozione di causalità sarà al centro dell'attenzione al fine di comprendere meglio *quale tipo* di effetto causale sia possibile identificare con i metodi considerati nel capitolo precedente. Assumerà particolare rilevanza, come vedremo, l'idea che, a differenza di quanto assunto nel capitolo precedente, non si possa parlare di un unico effetto causale uguale per tutti gli individui della popolazione di interesse. Abbandonando questa ipotesi fortemente restrittiva, e quindi assumendo che gli effetti causali possano essere eterogenei ossia diversi da individuo a individuo, l'analisi empirica diventa necessariamente più complessa, ma anche assai più rilevante e informativa soprattutto dal punto di vista delle domande che gli economisti del lavoro si pongono.

Per essere più chiari e concreti continueremo a fare riferimento alla relazione tra istruzione e reddito esaminata sotto il profilo teorico nel Capitolo 2 e sotto quello empirico nel Capitolo 19. La teoria del capitale umano suggerisce l'esistenza di un legame causale tale per cui se potessimo aumentare di un anno l'istruzione di un individuo questi guadagnerebbe di più una volta entrato nel mercato del lavoro. Esiste a questo riguardo una regolarità quasi sorprendente nei dati di paesi anche assai diversi tra loro: in ambito Europeo, ad esempio, la Tabella 2.3 indica che ogni anno aggiuntivo di istruzione è associato a un incremento di reddito pari in media al 7% e variabile tra il

4% della Svezia e l'11'% della Gran Bretagna. Per l'Italia la stima, confermata anche nella prima colonna della Tabella 19.1, è pari al 6%.

Sulla base di questi risultati e considerando un particolare studente universitario italiano, sarebbe ragionevole scommettere che se egli continuasse gli studi per un altro anno il suo reddito aumenterebbe del 6% rispetto al caso ipotetico di interruzione degli studi? Più in generale, sotto quali condizioni sarebbe possibile affermare che un anno di istruzione avrebbe questo *stesso* effetto causale per tutti gli studenti universitari italiani? Pensando alle teorie esaminate nel capitolo 2, la risposta a queste domande è tutt'altro che scontata. Ad esempio, abbiamo visto che una variabile non osservabile come l'abilità individuale, e quindi non inclusa tra le variabili di controllo della Tabella 19.1, può influenzare sia l'istruzione sia i redditi. Quindi, è possibile che chi è più istruito guadagni di più non per un effetto causale del tempo passato a scuola, ma solo perché più abile. In questo caso è evidente che la correlazione positiva tra istruzione e reddito non implicherebbe l'esistenza di alcun nesso di causalità.

È immediato rendersi conto che la rilevanza di questo problema è assai più generale. Ci stiamo chiedendo sotto quali condizioni sia possibile affermare che tra due variabili economiche qualsiasi esista un'autentica relazione *causale* e non una semplice correlazione. L'applicazione concreta che utilizzeremo a titolo di esempio è quella della relazione tra istruzione e reddito, ma possiamo porci le stesse domande in riferimento a molte altre relazioni. Ad esempio, se la Comunità Europea organizza un corso di formazione professionale per lavoratori disoccupati come facciamo a valutare se questa attività di formazione aumenta effettivamente la loro probabilità di trovare lavoro? Coloro che frequentano il corso potrebbero trovare più facilmente un impiego semplicemente perché più motivati degli altri indipendentemente dal corso. Se nelle aziende più sindacalizzate osserviamo salari più alti a parità di altre caratteristiche, possiamo attribuire questo effetto alla forza del sindacato? Potrebbe ad esempio accadere che le aziende con margini di profitto maggiori siano quelle che possono concedere aumenti più elevati e quindi, al tempo stesso, quelle in cui i lavoratori ottengono di più indipendentemente dalla presenza del sindacato.

Anche questo elenco potrebbe continuare a lungo con ulteriori esempi per ciascuno dei capitoli. E potrebbe includere anche problemi di natura non economica, come la valutazione degli effetti di una nuova terapia sulla probabilità di guarigione dei malati di tumore, degli effetti del numero di TIR sulla frequenza di incidenti autostradali, o degli effetti di un fertilizzante sulla produzione agricola di un campo. Il problema che accomuna tutte queste domande è il problema della *causalità*: un concetto con il quale ci confrontiamo costantemente nella vita di tutti i giorni, ma la cui definizione è tutt'altro che scontata così come tutt'altro che facile è identificare e misurare relazioni

causali nei dati generalmente a nostra disposizione. Proprio dalla definizione generale di causalità, che ci porta ai confini tra economia e filosofia, inizieremo quindi la nostra analisi nella Sezione 20.2. Utilizzeremo poi, nelle sezioni successive, questa definizione per studiare l'esempio concreto della relazione causale tra istruzione e reddito, senza tuttavia perdere di vista l'importanza più generale dei problemi metodologici affrontati.

20.2 Il problema della definizione di causalità

Consideriamo una popolazione di individui e immaginiamo di misurare per ciascuno di essi due variabili D e Y. Supponiamo di stimare la regressione lineare di Y su D ottenendo un coefficiente significativo e positivo.[1] Ci chiediamo sotto quali condizioni questo risultato ci autorizzi ad affermare che D *è una causa determinante di* Y, o piú semplicemente che D causa Y.

Per rispondere a questa domanda è necessario in primo luogo dare una definizione di causalità. Una possibile definizione, che sembra adattarsi bene alle problematiche economiche, è quella che si basa sul concetto di *evidenza controfattuale*.[2] Per caratterizzare questa definizione, immaginiamo che il problema in esame sia quello di valutare l'effetto di un *trattamento* D (ad esempio un anno di istruzione, un corso di formazione professionale o una terapia medica) su una caratteristica Y, che chiameremo *risultato*, di un individuo i (ad esempio, rispettivamente, il suo reddito, il suo stato occupazionale o il suo stato di salute). Supponiamo che la variabile D possa assumere due soli valori:

$D_i = 1$ se l'individuo i è stato esposto al trattamento;

$D_i = 0$ se l'individuo i non è stato esposto al trattamento.

Si noti che entrambe queste condizioni di trattamento sono ex ante possibili, ma solo una delle due può concretamente verificarsi nella realtà, poiché l'individuo i non può essere contemporaneamente *trattato* e *non trattato*. Se l'individuo è effettivamente trattato, si dice che l'evento controfattuale è la situazione di assenza di trattamento, e viceversa.

[1] Per la definizione di regressione lineare vedi il Capitolo 19.

[2] Per altre definizioni di causalità e in generale per una rassegna del millenario dibattito filosofico e statistico sul problema della causalità, vedi ad esempio Holland (1986) e Pearl (2000). Particolarmente affascinante, per chi fosse interessato a questo dibattito, è la lezione finale del libro di Pearl, dal quale abbiamo tratto parte del materiale presentato nel Riquadro 20.1.

Riquadro 20.1: Per discutere.
David Hume, il Canto del Gallo e il Tacchino Induzionista:
relazioni causali e correlazioni spurie.

Nel "Trattato sulla Natura Umana" di David Hume, troviamo la seguente caratterizzazione della nozione di causalità: "Così come ricordiamo di aver visto quel particolare oggetto chiamato *fiamma* e di aver sentito quella particolare sensazione che chiamiamo *caldo*, allo stesso modo richiamiamo alla nostra mente la loro costante congiunzione in tutte le situazioni passate. Senza ulteriori complicazioni chiamiamo il primo *causa* e il secondo *effetto* e inferiamo l'esistenza dell'uno dall'altro."(p. 156 della versione originale inglese del 1739; nostra traduzione.)
La definizione di Hume sembra quindi indicare nella ripetuta contiguità spazio temporale il requisito fondamentale per l'identificazione di una relazione causale. Tuttavia, come commenta Judea Pearl [Pearl 2000, p. 336], "È difficile credere che Hume non fosse cosciente delle difficoltà inerenti alla sua definizione. Egli sapeva bene che il canto del gallo è in continua connessione con il sorgere del sole e, ciononostante, non *causa* questo evento. Egli sapeva che un abbassamento del barometro è in continua connessione con la caduta della pioggia, eppure non *causa* la pioggia. Oggigiorno, queste difficoltà vengono rubricate come *correlazioni spurie* ossia come *correlazioni che non implicano causazione*."
L'esempio del gallo suggerisce che la regolarità della successione temporale non è sufficiente per dimostrare la causalità e invita a chiedersi quale altro tipo di osservazione possa autorizzare a classificare una relazione come causale. Ma forse, prima ancora di affrontare questa domanda, viene naturale chiedersi che differenza faccia stabilire se una relazione tra due eventi sia causale o no.
La risposta a quest'ultima domanda è suggerita dalla storiella del "Tacchino Induzionista" (e studioso di Hume) nato nel giorno di capodanno, il quale, a cominciare dalla nascita, iniziò ad osservare che dopo l'entrata del contadino nel pollaio la mangiatoia era sempre invariabilmente piena di cibo. Forte di questa costante e ripetuta osservazione concluse in breve che l'ingresso del contadino *causava* solamente benessere. Venne però la mattina del 25 dicembre, e il Tacchino capì a sue spese, maledicendo Hume, quanto pericoloso fosse inferire l'esistenza di una relazione causale dal mero ripetersi di una coppia di eventi.
Tornando alle parole di Pearl, il motivo per cui siamo interessati a identificare relazione causali è che "conoscere *cosa causa cosa* fa una gran differenza per come ci comportiamo."

Per indicare il fatto che il risultato Y può dipendere in modo causale dal trattamento D, usiamo la notazione funzionale $Y_i(D_i)$. Con questa notazione possiamo descrivere anche i due eventi controfattuali che riguardano il risultato dell'individuo i:

$Y_i(1)$ è il risultato in caso di trattamento;

$Y_i(0)$ è il risultato in caso di assenza di trattamento.

Abbiamo a questo punto gli elementi necessari per dare una definizione di causalità.

Definizione 1 : Effetto Causale.
Il trattamento D ha un effetto causale sul risultato Y per l'individuo i se il risultato in caso di trattamento è diverso dal risultato in caso di assenza di trattamento, ossia se

$$\Delta_i \equiv Y_i(1) - Y_i(0) \neq 0. \tag{20.1}$$

In questo caso Δ_i è l'effetto causale di D su Y per i.

Questa definizione può sembrare fin troppo ovvia, ma in realtà il suo utilizzo operativo comporta problemi tutt'altro che banali. Consideriamo il caso della relazione tra istruzione e reddito, e supponiamo che l'individuo i sia laureato. Per stabilire, in base a questa definizione, se il conseguimento della laurea abbia avuto un effetto causale sul suo reddito, non basta osservare l'evento effettivamente verificatosi e le sue conseguenze. Abbiamo bisogno anche di sapere che cosa sarebbe accaduto nel caso controfattuale in cui l'individuo i non avesse conseguito la laurea. Ma ciò è evidentemente impossibile, poiché l'evidenza controfattuale non è osservabile.

La Definizione 1, nella sua ovvietà, mette in luce quanto problematica sia l'identificazione di una relazione causale. Holland (1986), con il seguente enunciato, dà una definizione precisa di questo problema.

Definizione 2 : Problema Fondamentale dell'Inferenza Causale.
È impossibile osservare per uno stesso individuo i valori $D_i = 1$ e $D_i = 0$ così come i valori $Y_i(1)$ e $Y_i(0)$; quindi, è impossibile osservare e misurare l'effetto causale di D su Y per un individuo i.

Posto in questi termini il problema della causalità sembrerebbe insolubile. Gli studiosi di scienze statistiche hanno, tuttavia, proposto alcuni metodi per aggirare il problema.[3] Elemento comune a questi metodi è l'idea che, anche se l'effetto causale per l'individuo i non è identificabile, con opportune ipotesi è invece possibile identificare l'effetto causale per l'*individuo medio* nella popolazione. Per illustrare questo metodo di aggiramento del problema, consideriamo la definizione seguente.

Definizione 3 : Effetto Causale per l'Individuo Medio.
Il trattamento D ha un effetto causale sul risultato Y per l'individuo medio nella popolazione se il risultato medio in caso di trattamento è diverso dal risultato medio in caso di assenza di trattamento, ossia se

$$E\{\Delta_i\} \equiv E\{Y_i(1)\} - E\{Y_i(0)\} \neq 0 \tag{20.2}$$

In questo caso $E\{\Delta_i\}$ è l'effetto causale di D su Y per l'individuo medio.

Apparentemente questa definizione non ci fa fare molti passi avanti perché non possiamo osservare l'intera popolazione in entrambi gli eventi controfattuali. Ad esempio, su una popolazione di 100 individui ce ne saranno 10 con la laurea (per i quali non osserviamo la situazione in assenza di laurea) e 90 senza laurea (per i quali non osserviamo la situazione in caso di laurea). Per stimare l'effetto dell'istruzione per l'individuo medio, avremmo bisogno di calcolare *per l'intera popolazione* ciascuna delle due medie sulla destra della equazione 20.2, ma, sempre per colpa del Problema Fondamentale dell'Inferenza Causale, ciò è impossibile: non possiamo osservare il risultato medio dei trattati nel caso controfattuale di assenza di trattamento, nè il risultato medio dei non trattati nel caso controfattuale di trattamento.

Supponiamo, però, di estrarre dalla popolazione due campioni casuali, di individui che indicheremo con C e T. Poiché per costruzione i due campioni sono statisticamente identici all'intera popolazione, se in entrambi i campioni nessun individuo viene trattato il risultato medio di ciascuno campione sarà identico al risultato medio per l'intera popolazione in assenza di trattamento, ossia:

$$E\{Y_i(0)|i \in C\} = E\{Y_i(0)|i \in T\} = E\{Y_i(0)\}. \tag{20.3}$$

Viceversa se entrambi i campioni vengono trattati, il risultato medio di ciascuno sarà uguale al risultato medio per l'intera popolazione in presenza di trattamento:

$$E\{Y_i(1)|i \in C\} = E\{Y_i(1)|i \in T\} = E\{Y_i(1)\}. \tag{20.4}$$

[3] Esistono anche soluzioni non statistiche del Problema Fondamentale dell'Inferenza Causale, per le quali vedi ancora Holland (1986).

Sostituendo la 20.3 e la 20.4 nella 20.2 è immediato ricavare che

$$\begin{aligned} E\{\Delta_i\} &\equiv E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1)|i \in T\} - E\{Y_i(0)|i \in C\}. \end{aligned} \tag{20.5}$$

In questo modo possiamo aggirare il Problema Fondamentale dell'Inferenza Causale perché usiamo il campione C, chiamato campione dei *controlli*, come immagine di quello che accadrebbe al campione T, chiamato campione dei *trattati*, nella situazione controfattuale di assenza di trattamento, e viceversa.

Gli studi che si fondano su questo metodo vengono chiamati *esperimenti controllati con assegnazione casuale della popolazione ai gruppi di trattamento e di controllo* (o più semplicemente studi randomizzati) e sono frequenti in campo epidemiologico. Esistono, tuttavia, problemi etici che rendono a volte difficile la realizzazione di questo tipo di sperimentazioni.[4] Ai problemi etici, si aggiungono poi anche i problemi tecnici connessi con l'attuazione concreta di questo tipo di esperimenti.

Per gli stessi motivi, l'applicazione di questo metodo è forse ancor più complessa, e quindi assai meno frequente, in campo economico. Immaginate, ad esempio, di voler misurare l'effetto causale della laurea sul reddito dell'individuo medio in una popolazione. Non sarebbe difficile estrarre due campioni casuali di neo-diplomati della scuola superiore, ma sarebbe eticamente e anche tecnicamente molto difficile obbligare gli individui del campione T a laurearsi, eventualmente in contrasto con la loro volontà, e viceversa obbligare quelli del campione C ad interrompere gli studi anche se volessero laurearsi. Gli ostacoli etici alla sperimentazione assumono una rilevanza particolare nel campo dell'economia del lavoro, stante il coinvolgimento integrale della persona umana nel rapporto economico (il rapporto di lavoro appunto) che costituisce qui l'oggetto dello studio.

[4] Il Riquadro 20.2 descrive alcuni interessanti casi concreti in cui problemi etici, simili a quelli che un economista del lavoro dovrebbe affrontare, hanno reso difficile la sperimentazione medica.

Riquadro 20.2: Per discutere

Problemi etici e tecnici degli esperimenti statistici in medicina. Implicazioni per l'economista del lavoro.

Si suole far risalire l'inizio della ricerca epidemiologica volta a identificare relazioni causali in medicina alle osservazioni del dottor Semmelweiss intorno alla metà del secolo diciannovesimo (vedi Semmelweiss, 1941). Questo medico e docente universitario aveva notato che l'incidenza di febbre puerperale era maggiore tra le partorienti assistite da studenti della facoltà di medicina che non tra quelle assistite da ostetriche. Osservò poi che un collega era morto dopo essersi fatto un taglio nella sala delle autopsie e, con ulteriori osservazioni, ipotizzò (correttamente) che la febbre puerperale fosse causata dal fatto che gli studenti visitavano le puerpere dopo aver partecipato ad autopsie. Queste pionieristiche osservazioni "quasi-sperimentali" hanno probabilmente consentito di salvare numerosissime vite da allora. Eppure se esse fossero state originate da un esperimento controllato sarebbe difficile rimanere indifferenti rispetto ai problemi etici che tale sperimentazione, per quanto utile, avrebbe posto. Oggigiorno, proprio per questi motivi i medici che intendano effettuare esperimenti su pazienti devono sottostare non solo a dettagliati protocolli statistici senza i quali non è consentito trarre conclusioni in tema di causalità, ma anche al vaglio di una attenta valutazione etica.

Un esempio famoso in cui i problemi etici hanno giocato un ruolo importante è quello del dibattito relativo all'effetto del fumo sul tumore ai polmoni, recentemente tornato alla ribalta in Italia per via dei progetti governativi di proibizione del fumo nelle aree pubbliche. Alle origini di questo dibattito sta il rapporto, pubblicato nel 1964 dal Ministero della Sanità degli Stati Uniti, che per la prima volta attirò l'attenzione generale sulla questione. (Vedi ancora Pearl [2000, p. 353].) È interessante osservare che questo rapporto non era basato su dati sperimentali, ma su semplici correlazioni osservate nella popolazione. Queste correlazioni erano tuttavie interpretate come causali e conducevano il Ministro a concludere che se il fumo fosse stato proibito la percentuale di casi di tumore si sarebbe ridotta a quella osservata nella popolazione dei non fumatori. Il rapporto fu attaccato dalle imprese produttrici di tabacco, le quali, anche con l'appoggio di famosi statistici, affermavano che le correlazioni osservate non implicavano una connessione causale. Sostenevano, infatti, che un genotipo non osservabile poteva simultaneamente causare il bisogno di nicotina e alzare il rischio di tumore senza che tra i due fenomeni esistesse alcuna relazione diretta. In questo caso, proibire il fumo non avrebbe avuto alcun effetto sulla percentuale di casi di tumore in quanto la causa di questi non era il fumo ma il genotipo non osservabile. Allora, come oggi, esperimenti controllati avrebbero potuto dire facilmente chi avesse ragione in questo dibattito, ma non furono mai effettuati in quanto considerati inappropriati per evidenti motivi etici. Gli epidemiologi furono quindi costretti a cercare altri metodi statistici per valutare la pericolosità del fumo.

In altri casi, considerazioni di tipo etico hanno imposto l'interruzione di esperimenti ritenuti leciti inizialmente. È questo il caso del Physician's Health Study. che si proponeva di stabilire se l'aspirina protegge contro l'infarto. Dividendo in modo casuale tra trattati e controlli un campione di circa 22000 medici americani, i ricercatori iniziarono ad osservare, ben prima della conclusione pianificata dello studio, che la percentuale di infarti era notevolmente inferiore nel gruppo dei trattati a cui veniva somministrata l'aspirina. La continuazione dell'esperimento avrebbe significato impedire al gruppo dei controlli (a cui era proibito assumere aspirina) la possibilità di ricorrere ad uno strumento efficace per ridurre il rischio di infarto. Per questo motivo lo studio fu sospeso, anche se questa sospensione fu criticata da chi riteneva che in questo modo andasse persa un'occasione per stabilire se l'aspirina avesse anche effetti collaterali dannosi (vedi Cairns et al, 1991).

Particolarmente significativo è poi il caso (simmetricamente opposto) dello studio finlandese volto a dimostrare che il beta-carotene e la vitamina E proteggono i fumatori dal tumore ai polmoni, come ci si attendeva sulla base di evidenza non sperimentale. Con grande sorpresa, i ricercatori si accorsero che la mortalità era purtroppo significativamente superiore nel campione trattato con queste sostanze (vedi Palmgren et al. 1995). È facile intuire che questa scoperta sollevò numerose polemiche sulla opportunità delle sperimentazioni sull'uomo.

Recentemente in Italia, la scelta di non valutare la terapia Di Bella mediante assegnazione casuale dei malati di cancro ai gruppi di trattamento e di controllo è stata in parte motivata da simili controindicazioni etiche. Era infatti eticamente inaccettabile negare agli eventuali pazienti trattati con questa nuova terapia, la cui efficacia era ignota e potenzialmente nulla o dannosa, la possibilità di curarsi con le terapie tradizionali la cui efficacia era invece nota, per quanto parziale. (Per ulteriori informazione sulle problematiche e sui metodi di valutazione della terapia Di Bella si veda AA.VV (1999) e il relativo dibatto pubblicato dal *British Medical Journal* e reperibile sul sito http://www.bmj.com/cgi/content/full/318/7178/224).

Non solo problemi etici, ma anche difficoltà tecniche rendono spesso difficile il disegno e la implementazione di esperimenti controllati. Ad esempio è spesso molto costoso se non addirittura impossibile fare in modo che i trattati assumano effettivamente il trattamento e che viceversa i controlli ricevano solo il placebo, così come difficile è a volte la definizione stessa di trattamento, nei casi in cui questo corrisponda ad un complesso di molteplici interventi sui soggetti analizzati.

In economia del lavoro, e in generale nelle scienze sociali, siamo ben lontani dall'affrontare il problema della identificazione di relazioni causali con la stessa serietà dei protocolli adottati nelle scienze mediche. Abbiamo solo da poco iniziato a renderci conto di quanto erronee possano essere le conclusioni basate su dati osservati e non su dati generati da esperimenti controllati, come illustrato nel pionieristico studio di LaLonde [1986] che confronta stime di effetti causali basati su questi due tipi di dati. I problemi etici, tecnici e statistici che noi economisti dobbiamo affrontare per valutare, ad esempio, che effetto abbia innalzare l'obbligo scolastico o spendere denaro pubblico per la formazione professionale dei disoccupati, o diminuire la pressione fiscale, sono del tutto simili ai problemi da tempo affrontati dagli epidemiologi. Per questo abbiamo voluto attirare l'attenzione sulla esperienza di questi studiosi, dalla quale abbiamo molto da imparare.

Non sono tuttavia solo i problemi etici e tecnici a limitare il ricorso a questo metodo da parte degli economisti. Un problema aggiuntivo e particolarmente rilevante è costituito dal fatto che l'effetto causale per l'individuo medio non è necessariamente un effetto interessante per un economista. Ciò perché, in campo economico forse più che in altri campi, è ragionevole ritenere non solo che gli effetti causali di un trattamento possano differire notevolmente da individuo a individuo, ma soprattutto che gli individui scelgano liberamente di sottoporsi o meno ad un trattamento a seconda della convenienza che sperano di trarne in termini di risultato. È questo il fenomeno che viene generalmente chiamato *auto-selezione nel trattamento.*

Ad esempio, i modelli esaminati nel Capitolo 2, descrivono la scelta di raggiungere un certo livello di istruzione come una scelta razionale basata su un confronto tra i costi e i benefici derivanti al margine dalla continuazione degli studi. Poiché questi costi e questi benefici marginali sono molto probabilmente distribuiti in modo eterogeneo nella popolazione, nulla autorizza a ritenere che il rendimento dell'istruzione sia uguale per tutti. È infatti plausibile che tra i laureati vi siano individui ai quali la laurea è costata moltissimo, ma che si attendevano un incremento di reddito altrettanto elevato dal suo conseguimento. Viceversa, possono esistere individui ai quali la laurea è costata pochissimo e per questo l'hanno conseguita anche se si attendevano da essa un incremento di reddito poco elevato.

In questo contesto, nel quale l'istruzione ha effetti eterogenei sugli individui, potrebbe non avere molto senso chiedersi qual è l'incremento di reddito generato dal consegui-

mento della laurea per l'individuo medio nella popolazione, poiché questo ipotetico individuo potrebbe essere molto diverso dai gruppi di individui a cui siamo ragionevolmente più interessati: in particolare, coloro che effettivamente decidono di conseguire la laurea oppure coloro che al margine sono indifferenti tra il conseguirla o meno.

Consideriamo, ad esempio, il caso in cui lo Stato decida di offrire prestiti pubblici a tasso agevolato ai diplomati della scuola superiore con lo scopo di favorire coloro che vorrebbero conseguire una laurea senza poterne sopportare i costi. Per valutare il beneficio collettivo generato da una tale offerta non è evidentemente di alcuna utilità il rendimento generato dalla laurea per l'individuo medio. Più utile sarebbe il rendimento medio di coloro che ricevono il prestito, ma anche questo indicatore non sarebbe soddisfacente perché includerebbe il rendimento di coloro che conseguono la laurea indipendentemente dal prestito. L'effetto causale più rilevante in questo caso dovrebbe essere il rendimento medio per coloro che conseguono la laurea grazie al prestito, ma avrebbero interrotto gli studi se il prestito non fosse stato loro offerto.

Questi esempi suggeriscono che, in presenza di eterogeneità degli effetti di un trattamento[5] e di auto-selezione degli individui nel gruppo dei trattati, l'effetto statistico causale a cui l'economista dovrebbe essere interessato non è necessariamente quello medio nella popolazione, poiché a seconda della domanda a cui l'economista vuole rispondere l'attenzione potrebbe focalizzarsi su sottogruppi particolari di individui. Quindi la soluzione statistica suggerita dall'equazione 20.5 non sempre risolverebbe il problema dell'economista anche se le difficoltà di ordine etico e tecnico fossero sormontabili. Per comprendere meglio la natura dei problemi che l'economista incontra nell'identificazione di relazioni causali, è opportuno calarsi in modo più formalizzato nell'applicazione concreta che serve da esempio per questo capitolo, ossia la relazione tra istruzione e reddito.

20.3 Eterogeneità individuale ed effetto causale dell'istruzione

Consideriamo una versione semplificata del modello studiato nel capitolo 2, in cui assumiamo che un individuo scelga il numero ottimale di anni di istruzione risolvendo il seguente problema di ottimo:[6]

[5] Si noti che questa è la differenza cruciale rispetto alla situazione analizzata nel capitolo precedente, in cui si assumeva omogeneità di effetti.

[6] Il modello qui presentato, come caso particolare di quello del capitolo 2, è il modello di Becker (1975) nella versione proposta da Card (1995a).

$$\begin{aligned} \text{Max } U_i(w,S) &= w - r_i S \\ \text{sotto il vincolo: } w &= b_i S - \frac{k_b}{2} S^2 \end{aligned} \tag{20.6}$$

dove w è il logaritmo del reddito da lavoro W e S sono gli anni di istruzione. L'utilità dell'individuo aumenta con il reddito, ma diminuisce con l'istruzione perché ogni anno aggiuntivo di scuola ha un costo misurato dal parametro $r_i \geq 0$. Questo parametro misura quindi il costo marginale dell'istruzione assunto per semplicità costante rispetto a S, ma potenzialmente diverso da individuo a individuo. Per focalizzare l'analisi possiamo interpretare r_i come una misura dei vincoli di liquidità a cui l'individuo è soggetto. Le curve di indifferenza corrispondenti a questa funzione di utilità sono quindi delle rette crescenti come quelle rappresentate nella Figura 20.1. Agli individui caratterizzati da vincoli di liquidità più elevati (ossia con un maggiore r_i), corrispondono curve di indifferenza maggiormente inclinate. Al crescere dell'intercetta aumenta, ovviamente, il livello di utilità corrispondente a ciascuna curva di indifferenza.

Data questa funzione di utilità, se il reddito fosse indipendente dall'istruzione l'individuo sceglierebbe evidentemente un livello di istruzione pari a zero. Ma il vincolo del problema di ottimo implica che il (logaritmo del) reddito sia una funzione crescente e concava di S, come quella rappresentata nella Figura 20.1 che, come nel capitolo 2, chiamiamo *funzione generatrice del reddito*. Quindi, l'istruzione ha un effetto causale positivo sul reddito misurato (in termini percentuali) dal rendimento marginale dell'istruzione che ha la seguente espressione:

$$\beta_i(S) = b_i - k_b S. \tag{20.7}$$

Poiché il reddito è una funzione crescente e concava dell'istruzione, il rendimento marginale dell'istruzione è decrescente per tutti gli individui ad un tasso che dipende dal parametro costante $k_b > 0$. Esiste però eterogeneità nel rendimento marginale di individui diversi, in funzione del parametro b_i. Agli individui con b_i più elevato corrisponde una funzione generatrice del reddito più inclinata, il che implica, ad ogni livello di S, un maggior rendimento marginale dell'istruzione. Possiamo quindi interpretare il parametro b_i come un indicatore dell'abilità di un individuo intesa come capacità di ottenere un reddito più elevato da ogni anno aggiuntivo di istruzione.[7]

[7] Questa è una interpretazione particolare del concetto di abilità. Per altre interpretazioni, vedi il Capitolo 2.

La Figura 20.1 rappresenta la scelta ottimale dell'individuo i, corrispondente al punto di tangenza tra la funzione generatrice del reddito e la curva di indifferenza indicante la massima utilità raggiungibile. Seguendo sempre gli stessi passi descritti nel Capitolo 2 per il caso generale, possiamo ricavare formalmente il livello ottimale di istruzione scelto dall'individuo:[8]

$$S_i^* = \frac{(b_i - r_i)}{k_b} \tag{20.8}$$

Sostituendo la 20.8 nella 20.7 otteniamo il rendimento marginale dell'istruzione per l'individuo *i in corrispondenza della sua scelta ottima*, ossia nel punto in cui tale rendimento marginale eguaglia il costo marginale:

$$\beta_i^* = b_i - k_b S_i^* = r_i. \tag{20.9}$$

Si noti che possiamo interpretare β_i^* come l'effetto causale dell'istruzione per l'individuo i, poiché misura esattamente di quanto il reddito di i crescerebbe (al margine) se i aumentasse il suo livello di istruzione invece di fermarsi alla scelta ottima.[9] Nell'ambito di questo modello teorico quindi, β_i^* ha la stessa interpretazione causale, in termini di evidenza controfattuale, del parametro Δ_i nell'equazione 20.1. Si noti poi che r_i misura non solo l'inclinazione (costante) della curva di indifferenza, ma anche l'inclinazione della funzione generatrice del reddito nel punto di tangenza. Quindi, concettualmente, r_i non è l'effetto causale dell'istruzione sul reddito, ma ha un valore pari a tale effetto in equilibrio.

Come illustrato nel Capitolo 2, se considerassimo una popolazione costituita da individui identici, tutti sceglierebbero lo stesso numero di anni di istruzione, avrebbero lo stesso reddito e lo stesso rendimento marginale. Un campione di osservazioni su scolarità e reddito estratte da questa popolazione si presenterebbe, a meno di errori di misurazione, come una lista di osservazioni identiche. Se rappresentassimo questi dati in un grafico otterremmo un unico punto corrispondente alla tangenza tra funzione generatrice del reddito e curva di indifferenza nella Figura 20.1. Quindi, se questo fosse il vero modello che genera i dati non saremmo in grado di stimare alcunché per assenza di variabilità. Del resto questo sarebbe un cattivo modello per rappresentare la realtà poiché è evidente che nel mondo reale gli individui scelgono livelli di istruzione diversi

[8] Notate che, grazie alle ipotesi semplificatrici di questo capitolo, possiamo ricavare in modo esplicito il livello ottimale di istruzione.

[9] Come già si è detto, a differenza che nel Capitolo 19 il rendimento a cui siamo interessati dipende da i, ossia varia nella popolazione.

e, a parità di istruzione, ricevono redditi differenti. Questo caso estremo e paradossale è però utile per capire che, se gli individui effettuano le loro scelte educative come nel modello sopra descritto, solo l'esistenza di eterogeneità in termini di costo marginale e/o rendimento marginale dell'istruzione può generare una variabilità di osservazioni corrispondente a quella effettivamente osservata nella realtà.

Senza complicare troppo il modello, ma guadagnando molto ai fini di una rappresentazione più soddisfacente della realtà, consideriamo il caso più interessante di una popolazione caratterizzata da due livelli di abilità $b_H > b_L$ e due livelli di costo marginale dell'istruzione $r_H > r_L$. In questa popolazione esistono quattro gruppi di individui corrispondenti alle quattro possibili combinazioni di valori dei due parametri, che denoteremo con $g \in \{LH, HH, LL, HL\}$ (dove la prima lettera in ogni coppia indica il livello di abilità e la seconda il livello dei costi). La scelta ottimale del gruppo di individui g è data da

$$S_g^* \equiv S_{ij}^* = \frac{(b_i - r_j)}{k_b}, \tag{20.10}$$

ed essendoci quattro gruppi vi sono quattro scelte possibili che rappresentiamo graficamente nella Figura 20.2. In modo conforme a quanto ci si potrebbe attendere, a parità di abilità il livello di istruzione ottimale è maggiore per coloro che hanno vincoli di liquidità inferiori, mentre a parità di vincoli sono i più abili a frequentare la scuola per un numero maggiore di anni.

Data la linearità delle curve di indifferenza, benché esistano quattro gruppi di individui che operano scelte differenti, esistono solo due diversi rendimenti marginali dell'istruzione. Indipendentemente dalla abilità, hanno infatti lo stesso rendimento marginale tutti gli individui caratterizzati dallo stesso vincolo di liquidità. Utilizzando la 20.9 i due rendimenti sono:

$$\begin{aligned} \beta_{LH}^* &= \beta_{HH}^* = r_H \\ \beta_{LL}^* &= \beta_{HL}^* = r_L, \end{aligned} \tag{20.11}$$

dove è ancora una volta opportuno osservare che i parametri β denotano gli effetti causali dell'istruzione per i quattro gruppi mentre i parametri r sono i valori che questi effetti assumono in equilibrio.

Abbiamo così tutti gli elementi per analizzare, seppure nell'ambito di questa economia semplificata, i problemi legati alla identificazione e alla misurazione dell'effetto causale dell'istruzione sul reddito.

20.4 Identificazione dell'effetto causale dell'istruzione sul reddito

Iniziamo col notare che, a differenza del caso estremo e irrealistico, in cui gli individui erano per ipotesi identici, le osservazioni su scolarità e reddito estratte da una popolazione generata da questo modello assumono quattro possibili coppie di valori. Tuttavia, anche in presenza di questa variabilità, gli effetti causali individuali definiti teoricamente dalla 20.11 (e corrispondenti in questo contesto al parametro Δ_i della 20.1), non sono identificabili e misurabili nei dati per via del Problema Fondamentale dell'Inferenza Causale. Ciò che la variabilità ci consente di fare è, però, ricorrere alla statistica per aggirare il problema e stimare l'effetto causale dell'istruzione sull'individuo medio nella popolazione.

Tralasciamo per il momento il problema, su cui torneremo oltre, di stabilire se questo sia un effetto interessante per un economista. Ci interroghiamo, invece, sui metodi a nostra disposizione per stimarlo in modo consistente.[10] Prima, però, dobbiamo definire l'effetto causale medio nell'ambito del modello teorico della Sezione 20.3, ossia l'espressione che, nel contesto della relazione tra istruzione e reddito, corrisponde al parametro $E\{\Delta_i\}$ della 20.5.

Effetto causale medio teorico

Consideriamo una popolazione di N individui le cui scelte di istruzione e i conseguenti redditi siano generati dal modello della Sezione 20.3. Questi individui sono distribuiti nei quattro possibili gruppi $g \in \{LH, HH, LL, HL\}$ secondo frequenze che indichiamo con

$$\{P_{LH}, P_{HH}, P_{LL}, P_{HL}\} \tag{20.12}$$

Se potessimo osservare i valori assunti dai rendimenti dell'istruzione nei quattro gruppi, ossia i parametri r_L e r_H e se conoscessimo la distributione 20.12 potremmo facilmente calcolare l'effetto causale medio dell'istruzione sui redditi per questa popolazione, che risulterebbe pari a:

$$\bar{\beta} = (P_{LH} + P_{HH})r_H + (P_{LL} + P_{HL})r_L = \bar{r}, \tag{20.13}$$

e nel caso particolare in cui gli individui fossero equidistribuiti nei quatto gruppi ($P_g = P = 0,25 \ \forall \ g$), l'espressione si semplificherebbe in: $\bar{r} = \frac{r_H + r_L}{2}$.

[10] Per la definizione del concetto di consistenza, vedi il Capitolo 19.

Tuttavia, non possiamo calcolare l'effetto causale medio utilizzando la 20.13 perchè non osserviamo i valori r_L e r_H assunti dai rendimenti individuali, in quanto manca l'evidenza controfattuale. Quindi per stimare questo parametro teorico dobbiamo percorrere altre strade.

Esperimenti controllati con assegnazione casuale ai gruppi di trattamento e di controllo

Se potessimo ricorrere a un esperimento controllato, raggiungeremmo facilmente il nostro scopo adattando l'equazione 20.5 al problema in esame. Anche se, questo metodo è di fatto eticamente e tecnicamente improponibile per studiare la relazione tra istruzione e reddito, esso costituisce comunque un utile punto di riferimento per il confronto con metodi alternativi, e per questo lo esaminiamo ugualmente.

Supponiamo di estrarre due campioni casuali dalla popolazione, C e T. La distribuzione delle frequenze nella popolazione, data dalla 20.12, è quindi per costruzione identica a quella dei due campioni. Immaginiamo poi di poter indurre gli individui del campione T (i *trattati*) ad aumentare il loro livello di istruzione, mentre lasciamo gli individui del campione C (i *controlli*) liberi di scegliere come preferiscono. Per ottenere questo risultato possiamo ad esempio offrire una borsa di studio ai soli trattati, riducendo così per ciascuno di essi il costo marginale dell'istruzione r_i.

Nella Figura 20.3 rappresentiamo le scelte ottimali (e libere) del campione C, uguali a quelle della Figura 20.2, e le scelte indotte del campione T, spostate a destra perché le curve di indifferenza dei trattati hanno una inclinazione inferiore. Per semplificare l'analisi, ma senza perdita di generalità ai nostri fini, immaginiamo di controllare l'esperimento in modo tale che tutti gli individui del campione T vengano *trattati* con lo stesso aumento di istruzione ΔS rispetto alla loro scelta ottimale libera.[11] Quindi, in ciascun gruppo, la differenza di istruzione tra trattati e controlli è uguale a ΔS, e, ricordando che g denota i quattro gruppi $\{LH, HH, LL, HL\}$, l'uguaglianza dei trattamenti implica:

$$\Delta S_g = \Delta S \qquad \forall g. \tag{20.14}$$

Dato il loro maggiore livello di istruzione, i trattati hanno un reddito più alto dei rispettivi controlli. Tuttavia, poiché i campioni C e T sono identici, possiamo usare

[11] In altri termini, immaginiamo di differenziare opportunamente le borse di studio in modo tale che tutti i trattati aumentino dello stesso numero di anni la loro istruzione, indipendemente da quale sia la loro scelta libera in assenza di trattamento. Vedremo nel seguito di questo capitolo cosa accade quando il ricercatore non è in grado di controllare così perfettamente la somministrazione del trattamento.

l'uno come immagine di ciò che sarebbe accaduto all'altro nelle situazioni di assegnazione controfattuale non osservate. Ossia, all'interno di ciascun gruppo, il reddito dei controlli può essere considerato uguale al reddito controfattuale che i trattati avrebbero ottenuto se non avessero ricevuto il trattamento, e viceversa. Formalmente, ciò significa che valgono, in questo contesto, le equazioni 20.3, 20.4 e 20.5. Quindi la differenza tra il reddito medio di tutti i trattati e il reddito medio di tutti i controlli è pari all'incremento di reddito che l'individuo medio otterrebbe se la sua istruzione aumentasse in misura uguale a ΔS. Per verificarlo, basta scrivere questa differenza come la media ponderata delle differenze di reddito tra trattati e controlli nei quattro gruppi g, ossia:

$$E\{w_i|i \in T\} - E\{w_i|i \in C\} = (P_{LH} + P_{HH})r_H \Delta S + (P_{LL} + P_{HL})r_L \Delta S = \bar{r}\Delta S = \bar{\beta}\Delta S \tag{20.15}$$

dove le ultime due uguaglianze derivano dalla 20.13.

Poiché siamo interessati all'effetto causale medio *per unità di trattamento*, per esplicitare questo parametro di interesse dobbiamo dividere l'incremento medio di reddito dato dalla 20.15 per l'incremento medio di istruzione dato dalla 20.14. In questo modo otteniamo

$$\begin{aligned} \frac{E\{w_i|i \in T\} - E\{w_i|i \in C\}}{E\{S_i^*|i \in T\} - E\{S_i^*|i \in C\}} &= \frac{E_g\{r_g \Delta S_g\}}{E_g\{\Delta S_g\}}. \\ &= \frac{(P_{LH} + P_{HH})r_H \Delta S + (P_{LL} + P_{HL})r_L \Delta S}{\Delta S} \\ &= \bar{r} \\ &= \bar{\beta}. \end{aligned} \tag{20.16}$$

È importante osservare nella 20.16 (e analogamente nella 20.15) che l'espressione sulla sinistra della prima uguaglianza è quello che possiamo stimare sostituendo ai valori attesi per la popolazione le opportune corrispondenti stime campionarie. Sulla destra, invece, il secondo, il terzo e il quarto termine sono espressioni equivalenti del valore che il parametro teorico di interesse assume in equilibrio. Il quinto e ultimo termine, $\bar{\beta}$ denota infine il parametro teorico di interesse.

Abbiamo quindi verificato, nel contesto del modello teorico della Sezione 20.3, che mediante un esperimento controllato con assegnazione casuale ai gruppi di trattamento e di controllo siamo in grado di aggirare il Problema Fondamentale dell'Inferenza Causale e di stimare l'effetto causale per l'individuo medio. Tuttavia, come già abbiamo detto, è difficile, se non impossibile, che *dati sperimentali* di questo genere possano essere raccolti per un'analisi degli effetti dell'istruzione sul reddito. In questo tipo di

studi gli unici dati disponibili[12], sono solitamente quelli corrispondenti ad un campione casuale di scelte liberamente compiute dagli individui in esame, ossia un campione di *dati osservati*. Poiché lo strumento statistico di base con cui l'economista analizza questo tipo di dati è la regressione lineare[13], ci chiediamo ora se questo strumento consenta di stimare l'effetto causale medio dell'istruzione sul reddito, o, in alternativa, quali altre informazioni esso fornisca, ricordando sempre che, a differenza di quanto assunto nel capitolo precedente, qui consideriamo la regressione lineare in un contesto in cui l'effetto causale è eterogeneo nella popolazione.

Regressione lineare del reddito sull'istruzione

Nei quadranti della Figura 20.4 rappresentiamo graficamente quattro possibili popolazioni generate dal modello teorico della Sezione 20.3 che si differenziano l'una dall'altra per una diversa distribuzione delle frequenze nei quattro gruppi.

Consideriamo ad esempio il quadrante 20.4A. I quattro punti che appaiono nel grafico corrispondono alle quattro scelte ottimali rappresentate nella Figura 20.2. La dimensione dei cerchi intorno ai punti misura la frequenza di individui in ciascun gruppo. L'ipotesi del quadrante 20.4A è quindi che le frequenze siano uguali in ciascun gruppo, ossia: $P_g = P$ per ogni g. Iniziando con la popolazione di questo quadrante, ci chiediamo che tipo di informazione possiamo ottenere mediante la regressione lineare del logaritmo del reddito sugli anni di istruzione:[14]

$$w_i = \alpha + \rho S_i + \epsilon_i \tag{20.17}$$

dove ϵ_i è un termine di errore. Più precisamente, ci chiediamo che relazione ci sia tra il parametro ρ di questa regressione e l'effetto causale medio nella popolazione $\bar{\beta}$, ossia il parametro a cui siamo interessati e che, data la 20.13, assume in equilibrio il valore $\bar{r}$. La retta nel quadrante 20.4A rappresenta la retta di regressione stimata. Il coefficiente angolare di questa retta è il parametro ρ e la sua espressione analitica è:[15]

$$\rho \;=\; \lambda\bar{b} + (1-\lambda)\bar{r}. \tag{20.18}$$

[12] E in molti casi questa disponibilità è già un lusso

[13] Vedi il Capitolo 19.

[14] Nel Capitolo 2 abbiamo visto che, con l'aggiunta di una misura dell'esperienza lavorativa ed eventualmente di altri regressori di controllo, la stima di una simile regressione lineare è l'approccio classico per la stima dei rendimenti dell'istruzione proposto da Mincer (1974).

[15] Per la derivazione analitica di questa espressione, che richiede l'ulteriore ipotesi di simmetria delle distribuzioni di b_i e r_i nella popolazione (soddisfatta ad esempio nel quadrante 20.4A), vedi Card (1995a) e la soluzione dell'Esercizio 1 nella Sezione 20.6.

$$= \quad \bar{r} + \lambda(\bar{b} - \bar{r})$$

dove $\bar{b}$ e $\bar{r}$ sono, rispettivamente, le medie nella popolazione dell'abilità b_i e del costo marginale dell'istruzione r_i. Il parametro λ è invece una funzione delle varianze σ_b e σ_r di questi due parametri e della loro covarianza σ_{br}:

$$\lambda = \frac{\sigma_b^2 - \sigma_{br}}{(\sigma_b^2 - \sigma_{br}) + (\sigma_r^2 - \sigma_{br})} \tag{20.19}$$

Notate che, data la 20.10, r_i è necessariamente inferiore a b_i per ogni individuo i, poiché il livello ottimo di istruzione non può essere negativo. Ciò implica che $\bar{b} \geq \bar{r}$. Poiché nulla garantisce che il termine $\lambda(\bar{b} - \bar{r})$ sia nullo, il coefficiente ρ differisce dall'effetto causale medio $\bar{\beta}$ perchè quest'ultimo è uguale a $\bar{r}$ in equilibrio. Quindi, se stimiamo il coefficiente ρ, ad esempio con il metodo dei Minimi Quadrati Ordinari[16], otteniamo una stima distorta dell'effetto causale $\bar{\beta}$ a cui siamo interessati.[17] Per comprendere meglio il motivo economico di questa conclusione, ci possiamo aiutare con l'analisi grafica basata sugli altri quadranti della Figura 20.4.

Nel quadrante 20.4B la dimensione dei cerchi HH e LL è maggiore, ad indicare che in questa popolazione i corrispondenti individui sono relativamente più frequenti degli individui LH e HL. La retta di regressione corrispondente a questa situazione ha addirittura una inclinazione negativa poiché per minimizzare la somma delle distanze tra i punti e la retta dobbiamo avvicinare quest'ultima alle osservazioni HH e LL se queste sono relativamente più numerose. Nel caso estremo in cui la popolazione sia costituita solo da individui di tipo HH e LL la retta di regressione sarebbe esattamente la retta congiungente i corrispondenti punti, con un coefficiente angolare negativo. Quindi, pur in presenza di effetti causali individuali positivi dell'istruzione sul reddito, la stima del coefficiente di regressione risulterebbe negativa.

Nel quadrante 20.4C abbiamo invece la situazione in cui la distribuzione di frequenza è tale da originare una retta di regressione piatta. Ossia, in questo caso la stima del coefficiente di regressione indicherebbe un'assenza di correlazione tra w e S. Al crescere della frequenza relativa delle osservazione LH e HL l'inclinazione della retta diventa sempre più positiva come ad esempio nel quadrante 20.4D. In questo caso, il coefficiente angolare è superiore sia a r_L che a r_H, e quindi costituisce una stima distorta verso l'alto dell'effetto causale medio dell'istruzione sul reddito.

[16] Vedi il Capitolo 19.

[17] È importante comprendere che il metodo dei Minimi Quadrati stima in modo non distorto il parametro ρ così come definito per la popolazione nella 20.18: il problema è che questo parametro non è uguale al parametro della popolazione a cui siamo interessati, ossia all'effetto causale medio $\bar{\beta}$.

L'analisi dei quadranti della Figura 20.4 suggerisce che, se la popolazione è generata da questo modello, la relazione tra il coefficiente della regressione lineare e l'effetto causale dipende in modo cruciale dalla distribuzione degli individui nei quattro gruppi. Questa distribuzione dipende a sua volta da come i parametri b ed r sono correlati nella popolazione. Il caso 4A è un caso di assenza di correlazione, ossia sapendo l'abilità di un individuo non possiamo dire nulla su quale sia il valore più probabile del suo costo marginale dell'istruzione, e, viceversa, conoscendo questo costo marginale non possiamo dire nulla sulla sua abilità. I casi 4B e 4C corrispondono invece ad una correlazione positiva tra i due parametri. Ossia, le persone dotate di maggiore abilità tendono ad essere caratterizzate da costi marginali maggiori e viceversa. Il quadrante 20.4D, corrisponde, infine, a una situazione di correlazione negativa tra b ed r, per la quale, se un individuo ha un'abilità elevata, è anche probabile che il suo costo marginale sia basso e viceversa.

Quale di questi casi è il più plausibile? Poiché per quanto a nostra conoscenza non esistono stime negative o nulle del coefficiente ρ è ragionevole escludere una correlazione così positiva tra b ed r da rendere negativa o pari a zero l'inclinazione della retta di regressione. Quindi i casi 4B e 4C sono da considerarsi poco realistici. Esistono poi motivi teorici, che abbiamo esaminato nel Capitolo 2, per ritenere che la correlazione tra b ed r debba essere negativa.[18] Il caso 4D di correlazione negativa ma imperfetta appare quindi il più plausibile.

Non è, tuttavia, questa la conclusione più importante dell'analisi basata sulla Figura 20.4. Il punto fondamentale è che dalla stima della regressione lineare 20.17 possiamo attenderci, in generale, solo una stima distorta dell'effetto causale medio dell'istruzione sul reddito. La plausibilità di una correlazione negativa tra b ed r è importante solo nella misura in cui ci permette di qualificare questo punto fondamentale affermando che la distorsione è probabilmente verso l'alto. Ossia, che, in media, la stima del coefficiente ρ è maggiore dell'effetto causale medio nella popolazione.[19]

[18] Supponete, ad esempio, che esista una anche solo parziale persistenza intergenerazionale dell'abilità. In questa situazione i genitori più abili scelgono livelli di istruzione più elevati, hanno redditi maggiori e figli in media più abili. Nella generazione dei figli, quindi, i più abili tendono ad avere genitori con redditi più alti. Nella misura in cui i maggiori redditi dei genitori riducano i vincoli di liquidità dei figli, nella generazione di quest'ultimi gli individui più abili hanno costi marginali dell'istruzione inferiori. Per una discussione più approfondita di questo ed altri casi possibili vedi il Capitolo 2.

[19] Vedi su questo punto anche Griliches (1977) che per primo discusse questo genere di problemi.

Esperimenti naturali e stima con variabili strumentali

Abbiamo sin qui visto che l'analisi di dati osservati con la regressione lineare non consente, in linea generale, di ottenere l'effetto causale medio di un trattamento, se gli individui si autoselezionano nel trattamento sulla base del vantaggio che ne traggono e quindi se l'effetto del trattamento è eterogeneo nella popolazione. D'altro canto, le difficoltà legate alla realizzazione di un esperimento controllato con assegnazione casuale ai gruppi di trattamento e di controllo rendono poco frequente la disponibilità di dati sperimentali per la ricerca economica. Quindi, per un economista, sembrano esservi poche possibilità di aggirare in modo soddisfacente il Problema Fondamentale dell'Inferenza Causale. Tuttavia, la realtà offre talvolta situazioni che, dal punto di vista dell'analisi statistica della causalità, assomigliano formalmente ad esperimenti controllati anche se di fatto non lo sono. Queste situazioni vengono chiamate *esperimenti naturali*.

Consideriamo, ad esempio, la situazione generata dal fatto che gli individui in età scolare possono vivere a distanze diverse da una sede universitaria. Questa situazione genera un esperimento naturale rilevante ai nostri fini se la continuazione degli studi verso una laurea è meno costosa per coloro i quali vivono vicino ad una università.[20] Abbiamo già fatto riferimento alla distanza tra residenza e sede universitaria nel Capitolo 19 (Sezione 19.4), dove è stato introdotto il Metodo delle Variabili Strumentali nel contesto di omogeneità dell'effetto causale di interesse. Qui mostreremo che questo stesso metodo può essere utilmente interpretato come analisi statistica di un esperimento naturale e che questa interpretazione è particolarmente informativa quando l'effetto causale è eterogeneo nella popolazione.

Supponiamo di osservare un campione di lavoratori di cui sia noto il (logaritmo del) reddito w_i, gli anni di istruzione S_i e un indicatore Z_i dello stato di vicinanza ad una università durante l'età scolare. Più precisamente, l'indicatore è definito come segue:

$$Z_i = \left\{ \begin{array}{ll} 1 & \text{se } i \text{ viveva nello stesso comune di una università} \\ 0 & \text{se } i \text{ viveva in altro comune.} \end{array} \right\} \tag{20.20}$$

Confrontiamo ora questa situazione con quella ipotetica dell'esperimento controllato della Figura 20.3. Nel caso dell'esperimento controllato il ricercatore suddivide la popolazione in due campioni identici di trattati e di controlli e induce i trattati ad accrescere la loro istruzione offrendogli una borsa di studio sufficiente a ridurre il costo

[20] Questo esperimento naturale è stato studiato da Card (1995b) per gli Stati Uniti, su cui torneremo nel seguito di questa sezione, e da Flabbi (1998) per l'Italia.

marginale r_i. Nel caso dell'esperimento naturale, è la realtà, invece, a generare autonomamente i due campioni di trattati e di controlli, perché la distanza dalla sede universitaria determina valori diversi del parametro r_i.

Supponiamo che l'esperimento naturale generi i campioni C e T secondo la regola di assegnazione seguente:

$$Z_i = \left\{ \begin{array}{ll} 1 & \rightarrow i \in T \\ 0 & \rightarrow i \in C. \end{array} \right\} \tag{20.21}$$

In questo caso la soluzione proposta per l'esperimento controllato della Figura 20.3 si applica perfettamente all'esperimento naturale generato dalla distanza da una sede universitaria. Più in generale, data una variabile che possa svolgere il ruolo di Z_i, è possibile stimare l'espressione sulla sinistra della prima uguaglianza nella 20.16 e ottenere l'effetto causale medio teorico del trattamento. Gli economisti, senza fare riferimento esplicito ad una situazione sperimentale, conoscono (e utilizzano ampiamente) questo metodo di stima con il nome di Metodo delle Variabili Strumentali. Nel caso degli studi sulla relazione tra istruzione e reddito, il metodo viene applicato utilizzando l'indicatore Z_i come *strumento* per il livello di istruzione S_i nella equazione 20.17.[21] Purtroppo, però, è raro, se non impossibile, che un esperimento naturale possa sostituire perfettamente un esperimento controllato, e quindi i dati generati da esperimenti naturali vanno utilizzati con estrema cautela.

Un problema fondamentale è in realtà comune a entrambi i tipi di sperimentazione, ma si presenta in forma più grave, benché per certi versi più interessante, in quelli naturali. Si tratta del problema creato dalla impossibilità di controllare perfettamente il trattamento. Nel caso degli esperimenti naturali, questa impossibilità è evidente, perché il ricercatore non controlla nè l'assegnazione al trattamento nè la sua effettiva somministrazione agli individui. Ad esempio, nel nostro caso, la distanza da una sede universitaria (ossia l'assegnazione al trattamento) non è stabilita dal ricercatore. Inoltre è plausibile che questa variabile abbia un effetto diverso sul costo marginale dell'istruzione a seconda della ricchezza di un individuo. A livelli elevati di ricchezza, la vicinanza ad una università è probabilmente un fattore del tutto irrilevante mentre può essere cruciale per persone al limite della povertà e con forti vincoli di liquidità. Per i *ricchi*, quindi, questa vicinanza potrebbe non implicare un aumento di istruzione, ossia l'assegnazione al trattamento potrebbe non implicare la sua somministrazione. Per i *poveri*, invece, la corrispondenza tra assegnazione e somministrazione sarebbe più

[21] Vedi il Capitolo 19. La soluzione dell' Esercizio 2, nella Sezione 20.6, illustra invece l'equivalenza tra la formulazione statistico-sperimentale qui proposta e la formulazione econometrica classica del Capitolo 19.

probabile. Questo aspetto del problema prende il nome di *non-compliance*, ossia di mancata ubbidienza all'assegnazione da parte dei soggetti studiati.[22]

Nel caso degli esperimenti con assegnazione casuale, il problema del controllo imperfetto della sperimentazione è meno grave perché si presuppone che il ricercatore determini almeno l'assegnazione al trattamento. Tuttavia il controllo dell'assegnazione non garantisce il controllo della somministrazione. Ad esempio, nel caso dell'esperimento ipotizzato nella Figura 20.3, abbiamo supposto che l'offerta di una borsa di studio induca tutti i destinatari dell'offerta ad aumentare la loro istruzione di un pari ammontare, ma nulla garantisce che ciò effettivamente accada. Questo esperimento potrebbe definirsi perfettamente controllato solo se i potenziali trattati fossero costretti (e non solo indotti) ad aumentare la loro istruzione esattamente di ΔS. Nemmeno gli esperimenti controllati in campo medico ed epidemiologico sfuggono al problema. Ad esempio, se i potenziali trattati devono prendere una pillola al giorno, è possibile che se ne dimentichino, che ne prendano più di una o che decidano di abbandonare la sperimentazione.

Tornando al nostro esperimento naturale, possiamo analizzare graficamente il problema mediante la Figura 20.5 nella quale assumiamo che solo gli individui con un costo marginale dell'istruzione pari a r_H, ossia quelli con maggiori vincoli di liquidità, siano sensibili alla distanza dall'università. Quindi, solo nei gruppi $\{LH\}$ e $\{HH\}$ gli individui assegnati al trattatamento aumentano effettivamente il loro livello di istruzione. Poiché, solo per questi individui esiste una differenza di reddito tra trattati e controlli, è ragionevole attendersi che la stima generata dall'esperimento si avvicini all'effetto causale per questi individui. Si noti che questo effetto è maggiore di quello medio nella popolazione, poiché l'inclinazione della funzione generatrice del reddito è maggiore nei punti di tangenza corrispondenti ai gruppi $\{LH\}$ e $\{HH\}$.

In generale, se l'esperimento non è perfettamente controllato, non vale l'ipotesi di uguaglianza dei trattamenti 20.14 poiché $\Delta S_g \neq \Delta S_k$ quando $g \neq k$. Se la differenza media di istruzione tra i trattati e i controlli non è uguale a ΔS in tutti i gruppi, non vale più l'equazione 20.16. Vale invece l'equazione seguente:

$$\frac{E\{w_i|Z_i=1\} - E\{w_i|Z_i=0\}}{E\{S_i^*|Z_i=1\} - E\{S_i^*|Z_i=0\}} = \frac{E_g\{r_g \Delta S_g\}}{E_g\{\Delta S_g\}}. \tag{20.22}$$

[22] Si noti che la *non-compliance* può prendere varie forme. Ad esempio, in contrasto con le intenzioni dei ricercatori, gli assegnati al trattamento possono rifiutarlo, e, viceversa, gli assegnati al campione di controllo possono cercare di ricevere il trattamento a tutti i costi. Per una analisi illuminante della relazione tra *compliance* e Metodo delle Variabili Strumentali, sulla quale torneremo spesso nel corso di questo capitolo, vedi Angrist, Imbens and Rubin (1996).

$$= \frac{P_{LH} r_H \Delta S_{LH} + P_{HH} r_H \Delta S_{HH} + P_{LL} r_L \Delta S_{LL} + P_{HL} r_L \Delta S_{HL}}{P_{LH} \Delta S_{LH} + P_{HH} \Delta S_{HH} + P_{LL} \Delta S_{LL} + P_{HL} \Delta S_{HL}}$$

Si noti che il lato sinistro (ciò che possiamo stimare) è come nella 20.16, ma le espressioni sulla destra, invece, non sono necessariamente uguali a $\bar{\beta}$ (che sappiamo dalla 20.13 essere uguale a $\bar{r}$ in equilibrio). Ossia, in queste condizioni l'esperimento naturale non consente di stimare l'effetto causale medio nella popolazione. Consente solo di stimare un parametro che si avvicinerà a r_H o a r_L a seconda della dimensione relativa degli incrementi di istruzione ΔS_g e di come gli individui sono distribuiti nei quattro gruppi. Sostituento sulla sinistra della 20.22 le opportune medie campionarie otterremmo ancora una volta una stima potenzialmente distorta dell'effetto causale medio teorico dell'istruzione che ci siamo posti come obiettivo.

Ad esempio, nel caso specifico della Figura 20.5 in cui abbiamo assunto $\Delta S_{LL} = \Delta S_{HL} = 0$, il parametro che il nostro esperimento naturale può stimare in modo consistente è:

$$\frac{E\{w_i|Z_i=1\} - E\{w_i|Z_i=0\}}{E\{S_i^*|Z_i=1\} - E\{S_i^*|Z_i=0\}} = \frac{P_{LH} r_H \Delta S_{LH} + P_{HH} r_H \Delta S_{HH}}{P_{LH} \Delta S_{LH} + P_{HH} \Delta S_{HH}} = r_H = \beta_{LH}^* = \beta_{HH}^* \tag{20.23}$$

In questa equazione due sono le cose importanti da notare. In primo luogo, l'esperimento naturale permette di stimare in modo consistente solo l'effetto causale per gli individui il cui trattamento cambia con l'assegnazione, ossia nel caso specifico quelli dei gruppi LH e HH caratterizzati da un costo marginale più elevato e quindi più sensibili alla distanza dalla sede universitaria.[23] In secondo luogo, gli effetti causali per questi individui, β_{LH} e β_{HH}, assumono in equilibrio, data la 20.11, lo stesso valore r_H che risulta maggiore dell'effetto medio nella popolazione.

Più in generale, dovremmo attenderci stime elevate dell'effetto causale dell'istruzione tutte le volte che queste siano ottenute sulla base di esperimenti naturali in cui gli individui sensibili alla assegnazione siano caratterizzati da un alto rendimento dell'istruzione in equilibrio. Troviamo conferma di questa aspettativa nella Tabella 20.1 che riporta alcuni esempi basati su esperimenti naturali. Il primo è proprio quello utilizzato da Card (1995b) a cui è ispirata l'analisi sopra presentata. Utilizzando la distanza dalla sede universitaria come regola di assegnazione al trattamento in modo analogo a quanto descritto nella 20.21, le stime ottenute da Card per gli USA con il Metodo delle Variabili strumentali (per brevità IV, da *Instrumental Variables*) risultano più elevate di quelle ottenute con il Metodo dei Minimi Quadrati (per brevità OLS, da *Ordinary*

[23] Nella terminologia di Angrist Imbens e Rubin (1996) gli individui che *ubbidiscono* all'assegnazione, sono detti *compliers*.

Least Squares). Poichè sotto le ipotesi plausibili del quadrante 20.4D la stima OLS è distorta verso l'alto rispetto all'effetto causale medio nella popolazione, a maggior ragione il risultato di Card (1995b) indica che le sue stime IV misurano l'effetto causale di individui caratterizzati da un rendimento particolarmente elevato dell'istruzione in equilibrio.

Simile risultato e simile interpretazione si applicano all'articolo di Kane e Rouse (1993) che costruisce la regola di assegnazione combinando la distanza dalla sede universitaria con il livello delle tasse di iscrizione nelle università pubbliche, variabile da stato a stato negli USA. È plausibile infatti che una maggiore distanza possa essere compensata da minori tasse di iscrizione.

Radicalmente diverso è invece l'esperimento naturale su cui è basato l'articolo di Angrist e Krueger (1991), i quali osservano che negli USA (come del resto in Italia) bambini nati nello stesso anno solare iniziano la scuola nello stesso momento. Poichè l'obbligo scolastico è definito in termini di una età minima da raggiungere prima di poter abbandonare la scuola, i bambini nati nell'ultimo trimestre di ogni anno solare dovranno rimanere obbligatoriamente a scuola per un periodo di tempo più lungo rispetto ai bambini nati nei precedenti trimestri. Se il trimestre di nascita è casuale[24] ci troviamo di fronte ad un esperimento naturale tale per cui, se gli individui obbedissero perfettamente all'assegnazione, i nati nel quarto trimestre dovrebbero ricevere più istruzione. In realtà, gran parte della popolazione continua oltre l'obbligo scolastico indipendentemente dal trimestre di nascita, risultando quindi indifferente all'assegnazione. Similmente indifferente è il gruppo, assai meno numeroso, di coloro che abbandonano comunque la scuola prima dell'obbligo. Per entrambe queste componenenti della popolazione, gli assegnati al trattamento e gli assegnati al controllo si comportano nello stesso modo, analogamente a quanto accade per i gruppi HL e LL della Figura 20.5. Diversa è invece la situazione di coloro che continuano la scuola solo se obbligati. Alla luce del modello presentato in questo capitolo è ragionevole pensare che questi individui, i quali ubbidiscono all'assegnazione, abbiano un costo marginale dell'istruzione particolarmente elevato, tanto che solo se obbligati vanno a scuola. Per questo possiamo attenderci per essi un rendimento dell'istruzione relativamente alto, che giustificherebbe la differenza positiva tra la stima IV e la stima OLS ottenuta da Angrist e Krueger.

Le stime rimanenti nella Tabella 20.1, ottenute da Ichino e Winter-Ebmer (1999) su dati tedeschi, confermano l'ipotesi che l'effetto causale stimato sulla base di una

[24] Ipotesi questa difesa da Angrist e Krueger (1991) ma contestata con argomenti rilevanti da Bound et al. (1995).

data variabile strumentale deve dipendere dalle caratteristiche degli individui che ubbidiscono alla regola di assegnazione di quello specifico strumento. Nella penultima riga della tabella lo strumento utilizzato è un indicatore che assume valore 1 per gli individui il cui padre era un militare durante la seconda guerra mondiale, e zero negli altri casi. L'idea sottostante è che il coinvolgimento in guerra del padre e più in generale l'aver sperimentato una situazione di guerra in età scolare, siano fattori che rendono più difficile, a parità di altre condizioni, la frequentazione scolastica. Ci si attende inoltre che questo condizionamento sia tanto più forte quanto più precario è il benessere della famiglia d'origine. Se ciò è vero, gli individui la cui istruzione è maggiormente diminuita a causa della guerra sono probabilmente individui con maggiori vincoli di liquidità e quindi con maggiori costi marginali dell'istruzione. Analogamente a quanto osservato per gli esperimenti naturali sopra descritti, anche in questo caso ci attenderemmo stime IV particolarmente elevate, e questo è esattamente il risultato ottenuto da Ichino e Winter-Ebmer. Usando, però un diverso strumento *con gli stessi* dati, questi autori ottengono una stima notevolmente diversa. Nell'ultima riga della Tabella 20.1 lo strumento utilizzato è costruito in base al livello di istruzione del padre. L'idea sottostante è che avere un padre laureato faciliti il raggiungimento di un livello di istruzione più elevato nei figli.[25] Ichino e Winter-Ebmer, tuttavia, ipotizzano che coloro i quali prolungano gli studi solo perchè il padre è laureato siano tipicamente individui di abilità inferiore ma dotati di redditi familiari alti. Nei termini del modello considerato in questo capitolo si tratterebbe quindi di individui appartenenti al gruppo LL per i quali è lecito attendersi un basso rendimento dell'istruzione. Questa aspettativa è confermata dalle stime riportate nella Tabella 20.1. Mentre l'effetto causale dell'istruzione stimato usando la guerra mondiale come esperimento naturale è pari al 14% e superiore alla stima OLS, l'effetto ottenuto con lo strumento generato dalla istruzione paterna è pari al 4,8% ed inferiore alla stima OLS.[26]

L'analisi che ha portato alle equazioni 20.22 e 20.23 insieme all'evidenza empirica della Tabella 20.1 mostrano quindi che, se siamo interessati all'effetto causale medio

[25] Questo strumento ancor più di quello di Angrist e Krueger (1991) è stato variamente criticato in letteratura. Vedi a questo proposito Ichino e Winter-Ebmer (1999) e la letteratura ivi citata. Per una discussione dettagliata della validità di strumenti basati sulla seconda guerra mondiale vedi invece Ichino e Winter-Ebmer (2000).

[26] Questo risultato può essere analizzato formalmente considerando la popolazione della Figura 20.5. Immaginiamo che lo strumento Z_i il cui effetto è illustrato nella figura sia quello basato sulla seconda guerra mondiale. In una figura analoga possiamo invece descrivere l'effetto dello strumento basato sull'istruzione paterna che influenza le scelte scolastiche dei soli individui con costo marginale $r_i = r_L$ e abilità $b_i = b_L$. È facile vedere, applicando opportunamente a questo caso l'equazione 20.22, che il Metodo delle Variabili Strumentali fornirebbe una stima consistente di $\beta_{LL} = r_L$.

Table 20.1: **Stime OLS e IV basate su alcuni esperimenti naturali**

Articolo	Campione e strumento	Effetto stimato	
		OLS	IV
Card (1995b)	USA, NLS, maschi della coorte 1966. Strumento: distanza dalla sede universitaria.	0.073 (0.004)	0.132 (0.049)
Kane e Rouse (1993)	USA, NLS, femmine della coorte 1972. Strumento: tasse di iscrizione e distanza dalla sede universitaria.	0.080 (0.005)	0.091 (0.033)
Angrist e Krueger (1991)	USA, Censimento, coorte 1920-29 osservata nel 1970. Strumento: trimestre di nascita in interazione con l'anno di nascita.	0.070 (0.000)	0.101 (0.033)
Ichino e Winter-Ebmer (1999)	Germania, SOEP, inchiesta del 1986. Strumento: dummy per padre in guerra	0.055 (0.005)	0.140 (0.078)
Ichino e Winter-Ebmer (1999)	Germania, SOEP, inchiesta del 1986. Strumento: dummy per padre laureato	0.055 (0.005)	0.048 (0.013)

Note: La tabella è basata in parte sulla rassegna contenuta in Card (1995a). Per ulteriori informazioni si vedano gli articoli citati. Altre stime possono essere trovate nella rassegna più recente contenuta in Card (1999). La colonna OLS riporta stime ottenute con il metodo dei Minimi Quadrati Ordinari (Ordinary Least Squares) mentre la colonna IV riporta stime ottenute con il Metodo delle Variabili Strumentali (Instrumental Variables). La sigla NLS per gli USA indica la National Longitudinal Survey, mentre la sigla SOEP per la Germania indica il Socio-Economic Panel. In parentesi sono riportati gli errori standard.

nella popolazione, un esperimento naturale non può generalmente esserci di aiuto per via del problema della *non-compliance*. Tuttavia, non tutto il male vien per nuocere, perché, come abbiamo detto nella Sezione 20.2, l'effetto causale medio nella popolazione non è necessariamente un parametro di particolare interesse per un economista. Al contrario, nella misura in cui il fenomeno della *non-compliance* abbia una interpretazione economica, come nei casi della Tabella 20.1, la stima dell'effetto causale medio per coloro che ubbidiscono all'assegnazione può risultare di maggiore interesse.

Ad esempio, abbiamo visto che l'esperimento naturale generato dalla distanza da una sede universitaria permette di stimare l'effetto causale dell'istruzione solo per gli individui che sono maggiormente sensibili a riduzioni del suo costo in quanto caratterizzati da vincoli di liquidità più stringenti. Da un punto di vista normativo, questo effetto causale è probabilmente più interessante di quello medio nella popolazione se ci proponiamo di individuare strumenti di intervento per incrementare il livello di istruzione. Pensiamo ad esempio alla situazione di un paese in via di sviluppo, in cui la vicinanza a una scuola è un fattore determinante per le decisioni educative delle famiglie più povere. In questo contesto è di particolare interesse la stima dell'effetto causale medio dell'istruzione per gli individui provenienti da queste famiglie, poiché essi sono il *target* primario di una possibile campagna di alfabetizzazione. Se la campagna prescelta fosse l'incremento del numero di scuole nel territorio, l'esperimento naturale basato sulla distanza offrirebbe evidentemente informazioni assai rilevanti.[27]

La stima con variabili strumentali basate su esperimenti naturali può essere interessante non solo per fini normativi ma anche per un'analisi positiva della realtà: ad esempio, per descrivere l'intervallo di variazione degli effetti causali di un trattamento nella popolazione se questi sono eterogenei.[28] Poiché ogni strumento ha il suo corrispondente gruppo di individui che ubbidiscono all'assegnazione, mediante una scelta accorta degli strumenti, questo metodo di stima potrebbe consentire, almeno in linea teorica, di ottenere informazioni sull'intero intervallo di variazione degli effetti.

Il condizionale è tuttavia d'obbligo perché gli esperimenti naturali ... non si trovano a ogni angolo di strada e l'individuazione di strumenti validi per l'applicazione del metodo di stima sopra descritto è tutt'altro che facile. Il capitolo 19 descrive le ipotesi assai stringenti che uno strumento deve soddisfare per poter fornire una stima consistente dell'effetto causale di un trattamento. Quella solitamente più difficile da soddisfare nelle applicazioni economiche è l'ipotesi che prende il nome di *restrizione di*

[27] Vedi su questo l'Esercizio 3 nella Sezione 20.6.

[28] Sono utilizzate a questo fine, ad esempio, le stime ottenute da Ichino e Winter-Ebmer (1999) e riportate nella Tabella 20.1. Per ulteriori approfondimenti rimandiamo a questo articolo e alla letteratura ivi citata.

esclusione. Questa ipotesi richiede che, a parità di trattamento, l'assegnazione non abbia alcun effetto sul risultato. Ossia, nel caso dell'esperimento naturale sopra considerato, se un individuo è laureato il suo reddito non deve dipendere da quanto vicina era una università durante la sua infanzia. Analogamente la distanza da una sede universitaria non deve provocare differenze di reddito tra coloro che non sono laureati. In altri termini, è necessario che la distanza influenzi i redditi *solo* attraverso l'effetto intermedio sulle scelte scolastiche. É evidente come questa ipotesi possa non essere soddisfatta nell'esempio in questione, se il vivere in una grande città implica al tempo stesso la vicinanza ad una sede universitaria e l'accesso ad un mercato del lavoro caratterizzato da salari più elevati.[29] Ugualmente difficile da soddisfare, essa risulta in numerose altre appicazioni del Metodo delle Variabili Strumentali.

20.5 Riepilogo

In questo capitolo, partendo da una precisa definizione di causalità, basata sul concetto di evidenza controfattuale, abbiamo visto che l'identificazione di relazione economiche causali è particolarmente difficile con gli strumenti statistici disponibili, principalmente a causa della impossibilità di effettuare esperimenti controllati. Per questo gli economisti del lavoro devono spesso accontentarsi di esperimenti naturali, i quali, per via delle numerose ipotesi che la loro applicazione richiede, non sempre consento conclusioni soddisfacenti. Tuttavia, per quanto imperfetti e difficili da trovare, gli esperimenti naturali, in combinazione con il Metodo delle Variabili Strumentali, hanno fino a oggi offerto la strategia più frequentemente usata dagli economisti del lavoro per risolvere il Problema Fondamentale dell'Inferenza Causale.[30] Se le opposizioni etiche e le diffi-

[29] Vedi tuttavia la difesa convincente di questo esperimento proposta da Card (1995b).

[30] Vi sono anche altre strategie per risolvere il problema, come ad esempio il metodo dei *Propensity Scores*, originariamente proposto da Rosembaum and Rubin (1983) e successivamente usato da Dehejia and Wahba (1999) come possibile soluzione ai problemi posti da LaLonde(1986) a cui abbiamo accennato nel riquadro 20.2. Per una introduzione a questo metodo e per ulteriori riferimenti bibliografici vedi la Sezione 5 delle dispense di Ichino (2001).

Nel caso specifico della stima dei rendimenti dell'istruzione, particolarmente interessante è anche la strategia basata sul confronto tra gemelli monozigoti, utilizzata da Ashenfelter e Krueger (1994). Per una rassegna di questi e altri metodi che per ragioni di spazio non possiamo qui considerare, vedi Card (1999) e, più in generale, Angrist and Krueger (1999).

Infine si vedano anche i saggi di Heckman (1978, 1979, 1997) e Heckman e Robb (1985) nei quali viene introdotto e discusso il metodo di stima noto agli economisti del lavoro come "Metodo di Heckman", originariamente disegnato per la studio della offerta di lavoro femminile (vedi Capitolo 3) e poi utilizzato in molteplici altre applicazioni. Si noti però che anche per questo metodo sono

coltà tecniche che ostacolano la diffusione degli "esperimenti controllati" non verranno almeno parzialmente superate, questa strategia continuerà a essere in molti casi l'unica strategia percorribile, non solo per stimare l'effetto causale dell'istruzione sui redditi ma anche per stimare i numerosi altri effetti causali discussi in questo libro. È quindi bene che gli economisti del lavoro ne conoscano approfonditamente le potenzialità e i problemi.

---

necessarie restrizioni di esclusione equivalenti a quelle richieste dal Metodo delle Variabili Strumentali. Per approfondimenti su questo punto vedi ancora Ichino (2001) e e il dibattito finale in Angrist, Imbens and Rubin (1996).

20.6 Esercizi

Esercizio 1: *Coefficiente di regressione, abilità media e costo marginale medio.*
Utilizzando il modello teorico presentato in questo capitolo e la definizione di coefficiente di regressione lineare (vedi Capitolo 19), derivare l'equazione 20.18.

Soluzione
Per definizione il coefficiente della regressione lineare 20.17 è dato da

$$\rho \equiv \frac{Cov\{w_i, S_i\}}{Var\{S_i\}} \equiv \frac{E\{(w_i - \bar{w})(S_i - \bar{S}\}}{Var\{S_i)\}} \tag{20.24}$$

dove $\bar{w}$ e $\bar{S}$ sono rispettivamente la media del logaritmo del reddito e la media degli anni di istruzione nella popolazione.

Utilizzando il vincolo della 20.6 e la 20.8 possiamo riscrivere il numeratore della 20.24 come

$$\begin{aligned} E\{(w_i - \bar{w})(S_i - \bar{S})\} &= E\{b_i S_i (S_i - \bar{S}) - \frac{1}{2} k_b S_i^2 (S_i - \bar{S})\} \qquad (20.25) \\ &= \frac{1}{k_b^2} E\{b_i (b_i - r_i)[(b_i - \bar{b}) + (r_i - \bar{r})]\} - E\{\frac{1}{2} k_b S_i^2 (S_i - \bar{S})\} \end{aligned}$$

e il denominatore come

$$Var\{S_i\} = \frac{1}{k_b^2}(\sigma_b^2 + \sigma_r^2 - 2\sigma_{br}). \tag{20.26}$$

dove σ_b^2 e σ_r^2 sono rispettivamente le varianze di b_i e r_i nella popolazione, e σ_{br} è la loro covarianza.

Assumendo che le distribuzioni di b_i e r_i siano simmetriche, e quindi abbiamo momenti centrali terzi uguali a zero, sostituendo la 20.25 e la 20.26 nella 20.24 otteniamo la 20.18 definendo λ come nella 20.19. In assenza di simmetria l'espressione che definisce ρ contiene anche termini in cui compare il momento centrale terzo.

Esercizio 2: *Metodo Statistico-Sperimentale e Metodo delle Variabili Strumentali*
Considerate l'effetto di un trattamento binario $D = \{0, 1\}$ su un risultato Y. Sia anche dato uno strumento binario $Z = \{0, 1\}$. Adattando la 20.16 a questo caso di trattamento binario, il Metodo Statistico Sperimentale propone di stimare l'effetto

causale del trattamento sostituendo le opportune medie campionarie nella espressione:

$$\Delta_{SS} = \frac{E\{Y \mid Z = 1\} - E\{Y \mid Z = 0\}}{E\{D = 1 \mid Z = 1\} - E\{D = 1 \mid Z = 0\}} \tag{20.27}$$

Il Metodo delle Variabili Strumentali ottiene invece lo stesso obiettivo sostituendo le opportune statistiche campionarie nella espressione:

$$\Delta_{IV} = \frac{Cov\ \{Y, Z\}}{Cov\ \{D, Z\}} \tag{20.28}$$

Dimostrate che $\Delta_{SS} = \Delta_{IV}$ e quindi i due metodi di stima sono equivalenti.

Soluzione

$$\begin{aligned}
\Delta_{IV} &= \frac{E\{Y, Z\} - E\{Y\}E\{Z\}}{E\{D, Z\} - E\{D\}E\{Z\}} \\
&= \frac{E\{Y \mid Z = 1\} Pr\{Z = 1\} - E\{Y\}Pr\{Z = 1\}}{Pr\{D = 1, Z = 1\} - Pr\{D = 1\}Pr\{Z = 1\}} \\
&= Pr\{Z = 1\} \frac{E\{Y \mid Z = 1\} - E\{Y \mid Z = 1\}Pr\{Z = 1\} - E\{Y \mid Z = 0\}Pr\{Z = 0\}}{Pr\{D = 1, Z = 1\} - [Pr\{D = 1, Z = 1\} + Pr\{D = 1, Z = 0\}] Pr\{Z = 1\}} \\
&= Pr\{Z = 1\} \frac{E\{Y \mid Z = 1\} [1 - Pr\{Z = 1\}] - E\{Y \mid Z = 0\}Pr\{Z = 0\}}{Pr\{D = 1, Z = 1\} [1 - Pr\{Z = 1\}] - Pr\{D = 1, Z = 0\} Pr\{Z = 1\}} \\
&= Pr\{Z = 1\} \frac{Pr\{Z = 0\} [E\{Y \mid Z = 1\} - E\{Y \mid Z = 0\}]}{[Pr\{D = 1 \mid Z = 1\} - Pr\{D = 1 \mid Z = 0\}] Pr\{Z = 1\} Pr\{Z = 0\}} \\
&= \frac{E\{Y \mid Z = 1\} - E\{Y \mid Z = 0\}}{Pr\{D = 1 \mid Z = 1\} - Pr\{D = 1 \mid Z = 0\}} = \Delta_{SS}
\end{aligned}$$

Q.E.D.

Esercizio 3: *Effetto di una campagna di alfabetizzazione in un paese in via di sviluppo.* Considerate un paese in via di sviluppo in cui il livello medio di istruzione sia molto basso nella popolazione.

Il governo ha iniziato una campagna di alfabetizzazione che consiste nell'aprire dei centri ricreativi in alcuni villaggi. Ciascuno di questi centri dispone di un grande schermo televisivo sul quale oltre ai programmi normali vengono trasmessi anche corsi di lingua e matematica. Tenete presente che nessuno nella popolazione di questi villagi ha un proprio televisore.

Il governo ha già aperto un certo numero di centri. Dopo un anno di sperimentazione, un campione casuale della popolazione è stato esaminato con un test per valutare le capacità di lettura, scrittura e matematica. Sulla base di questo test, per ogni persona esaminata è stato calcolato un indice di alfabetizzazione Y. Ad ogni persona è anche stato chiesto se avesse seguito i corsi e la risposta può essere considerata totalmente veritiera.

Il governo vi chiede di valutare se la campagna di alfabetizzazione è stata efficace e in particolare se valga la pena di aumentare il numero di centri ricreativi.

Con l'aiuto delle seguenti domande, discutete come potreste procedere per identificare e stimare l'effetto della campagna.

1. Avete a vostra disposizione dati *osservati* o dati *sperimentali*? Cos'è il trattamento e cosa'è il risultato in questo caso. Vi aspettate che possa esserci autoselezione della popolazione nel trattamento? Scrivete l'equazione che mette in relazione il risultato e il trattamento usando la notazione introdotta nel capitolo.

2. Sarebbe utile per il governo effettuare un esperimento controllato "obbligando" un campione casuale di trattati a guardare la televisione durante le lezioni e un campione casuale di controlli a non frequentare il centro ricreativo durante le lezioni?

3. Sarebbe utile invece stimare una regressione lineare dell'indice di alfabetizzazione sull'indicatore di trattamento?

4. Potete disporre anche della seguente informazione aggiuntiva: per ciascun individuo sapete se un grande fiume separa la sua abitazione dal centro ricreativo. Nel paese considerato i fiumi possone essere superati ma con notevole difficoltà per via della loro dimensione e della mancanza di ponti. Questa informazione può essere utile per valutare l'effetto causale della campagna sul risultato Y? Come?

Soluzione

1. Il governo mette a vostra disposizione dei dati *osservati*. Non ha effettuato nessun esperimento controllato con assegnazione casuale ai gruppi di trattamento e di controllo. Si è limitato ad offrire ai cittadini la possibilità di seguire i corsi. I cittadini possono scegliere liberamente se seguire il corso o meno. È quindi ragionevole attendersi che ci sia autoselezione nel trattamento: solo coloro che pensano di trarre vantaggio dal corso lo frequenteranno. Il campione di chi frequenta il corso non è un campione casuale della popolazione. I dati a vostra disposizione sono essenzialmente costituiti da due variabili osservate in un campione estratto dalla popolazione: l'indice di alfabetizzazione Y e l'indicatore D che assume valore 1 se la persona ha seguito i corsi e 0 in caso contrario. La relazione tra queste due variabili per l'individuo i è esprimibile formalmente come:

$$Y_i(D_i) = D_i Y_i(1) + (1 - D_i)Y_i(0) \tag{20.29}$$

2. Un esperimento controllato, oltre ad essere tecnicamente ed eticamente difficile da realizzare, non darebbe risultati utili al governo perchè l'effetto sull'individuo medio non è in questo caso un parametro interessante. Ciò che interessa maggiormente è l'effetto per coloro che scelgono liberamente di seguire il corso.

3. Il coefficiente di una regressione lineare di Y su D non sarebbe un parametro interessante. Non sarebbe uguale all'effetto causale per l'individuo medio per via della autoselezione nel trattamento. Ipotizzando che i più abili siano anche coloro ai quali la frequenza costa marginalmente di meno, il coefficiente della regressione lineare sovrastimerebbe l'effetto casuale per l'individuo medio. Ne vi è alcuna garanzia che questo coefficiente sia uguale all'effetto per coloro che scelgono liberamente di frequentare.

4. Questa informazione aggiuntiva vi offre un "esperimento naturale" in base al quale potete costruire un indicatore binario Z che assume valore 1 se la persona vive sulla sponda del fiume in cui è collocato il centro ricreativo e 0 in caso contrario. Con questo strumento potete quindi applicare il metodo delle variabili strumentali sostituendo le opportune statistiche campionarie nella 20.27 o equivalentemente nella 20.28.

   Notate, però, che l'effetto causale stimato sulla base di questo esperimento naturale sarebbe quello per coloro che frequentano il corso solo perchè non devono attraversare il fiume e non lo frequenterebbero in caso contrario. Questo non è

esattamanete l'effetto causale per tutti coloro che scelgono liberamente di frequentare. Tuttavia l'effetto stimabile grazie all'esperimento è interessante per il governo perchè approssima proprio l'incremento di alfabetizzazione per coloro che al margine deciderebbero di istruirsi solo se un centro ricreativo fosse aperto più vicino alla loro abitazione. Questo è il parametro rilevante per decidere i costi e i benefici di un aumento del numero di centri ricreativi.

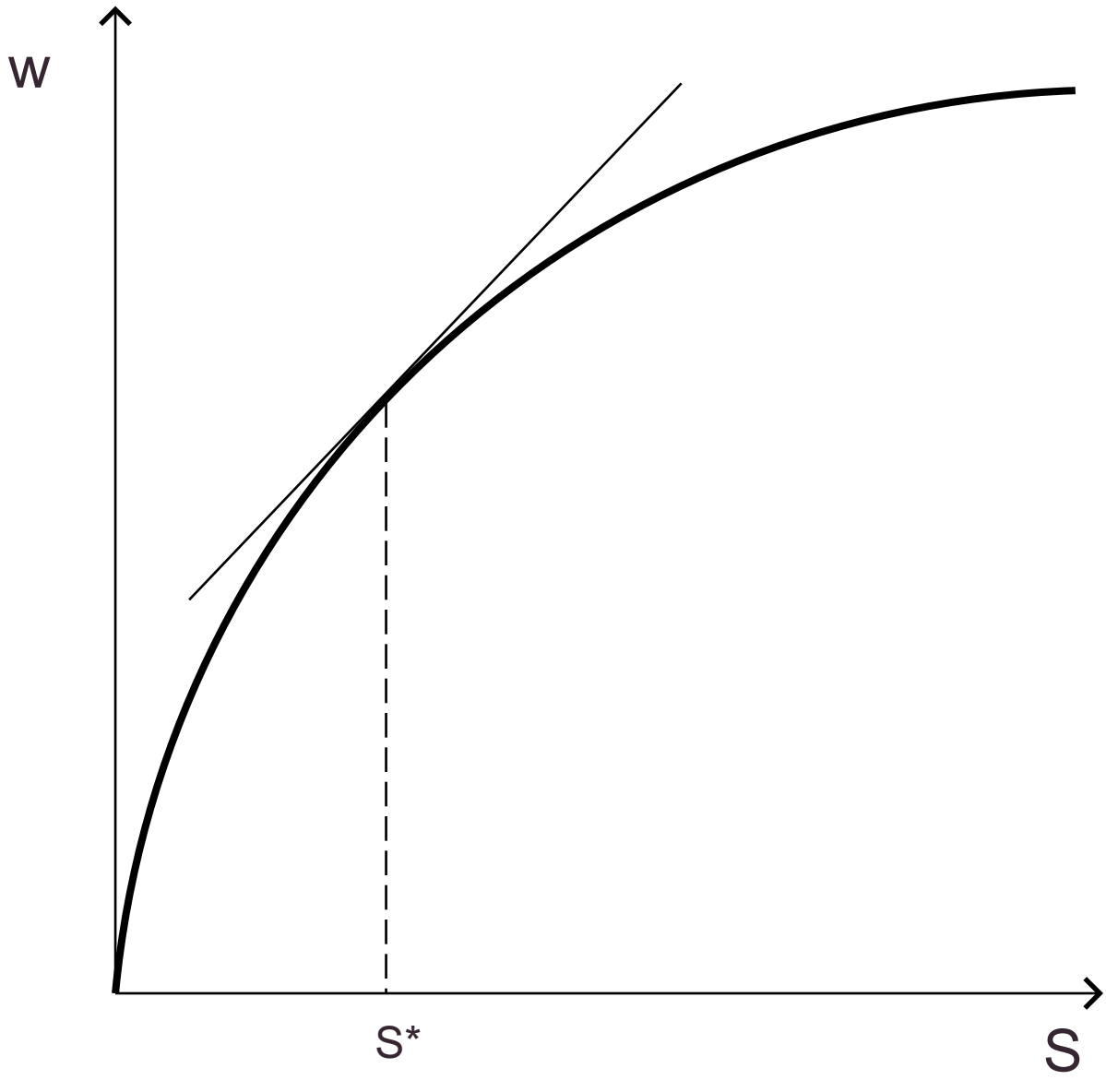

Fig. 20.1: Scelta ottimale di un individuo

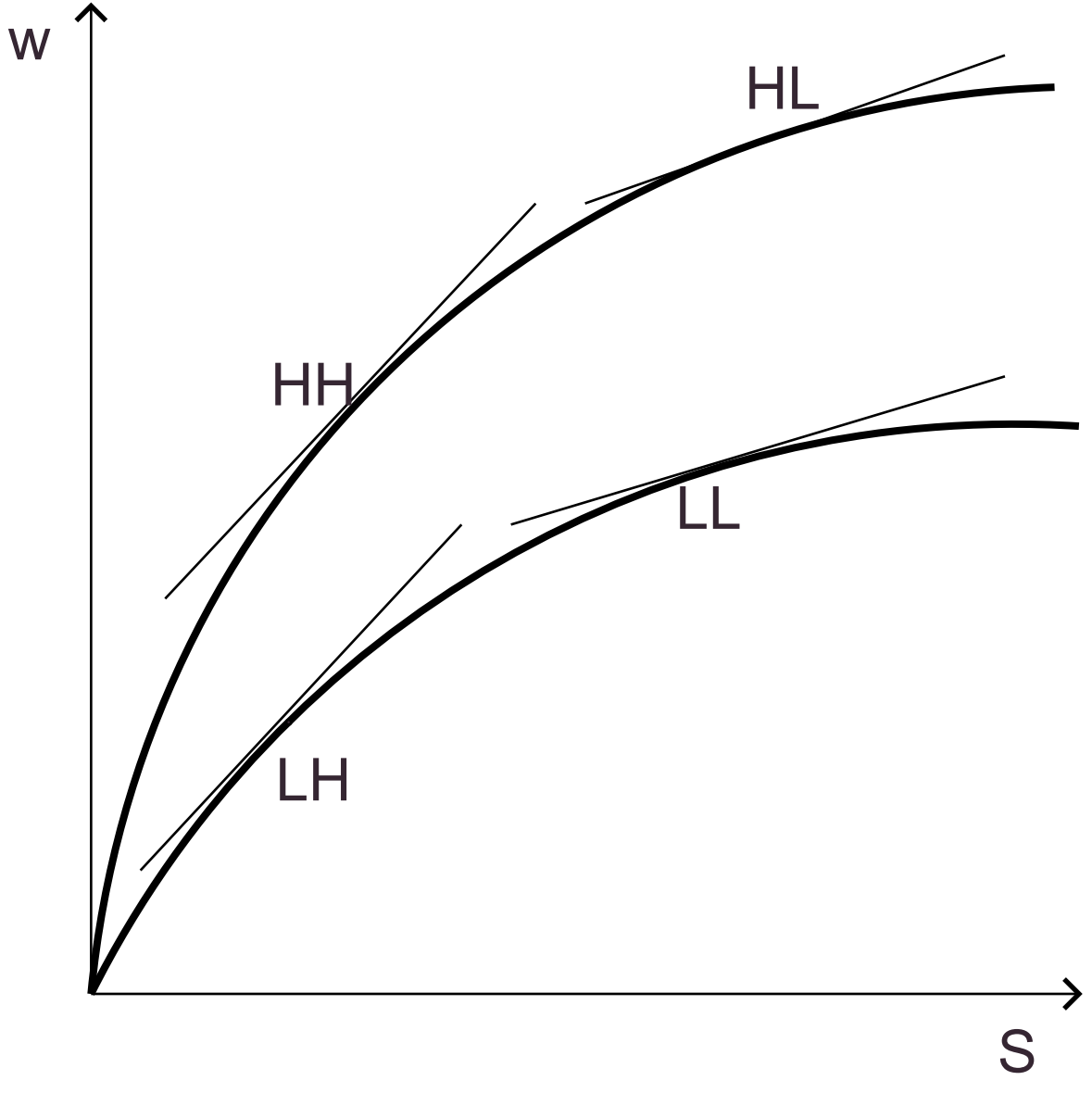

Fig. 20.2: Scelte con abilita` e costi eterogenei

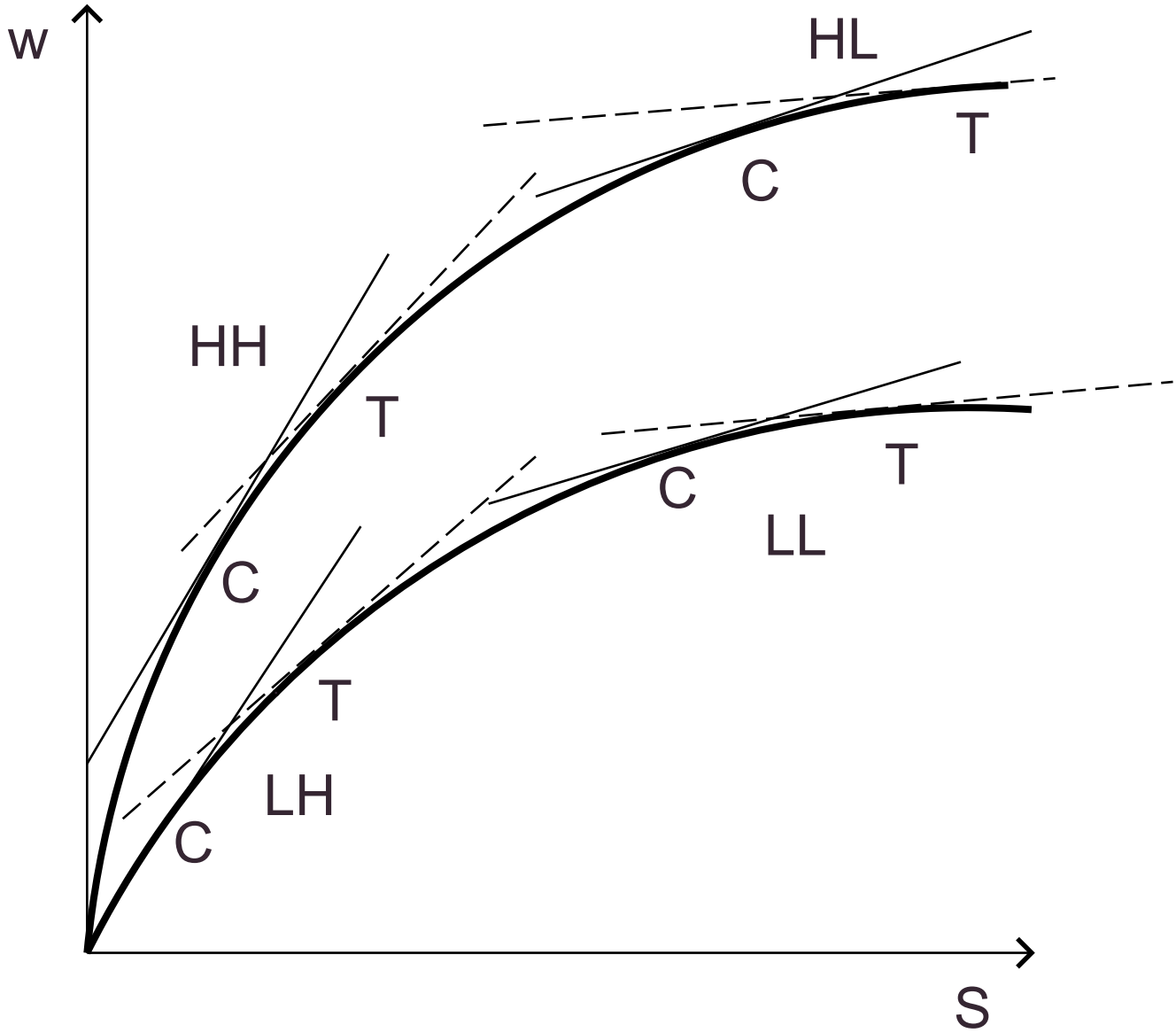

Fig. 20.3: Scelte dei controlli e dei trattati in un esperimento controllato

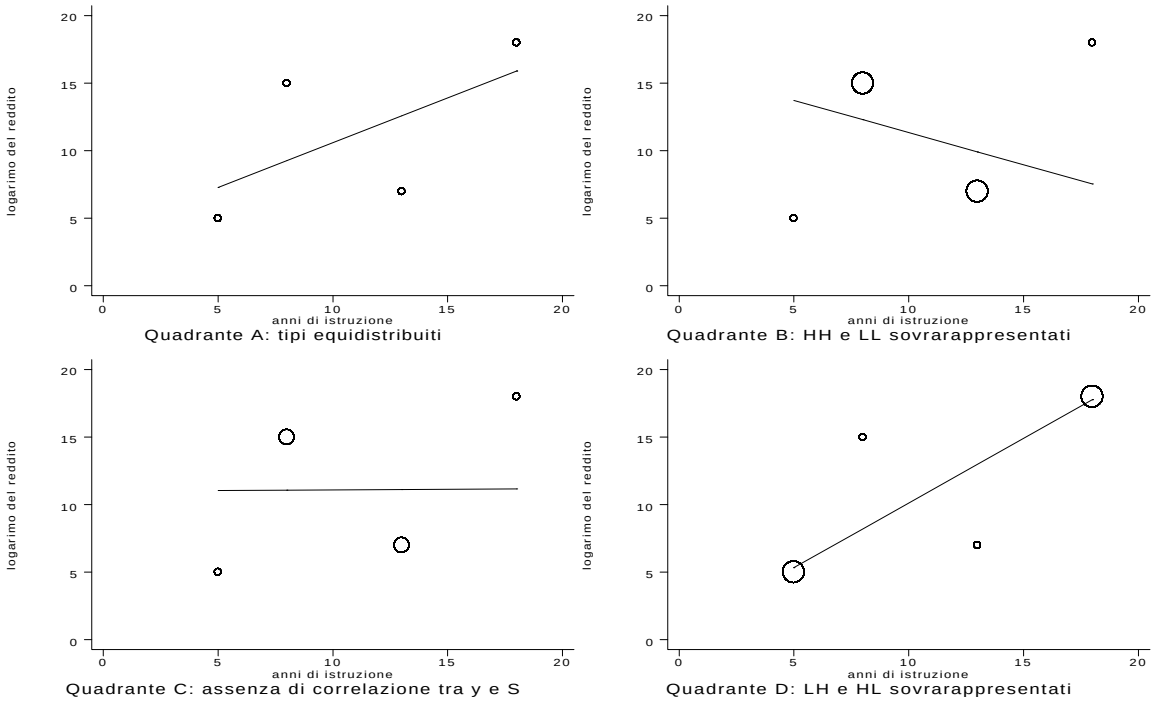

Fig. 20.4: regressioni di w su S con diverse frequenze dei tipi

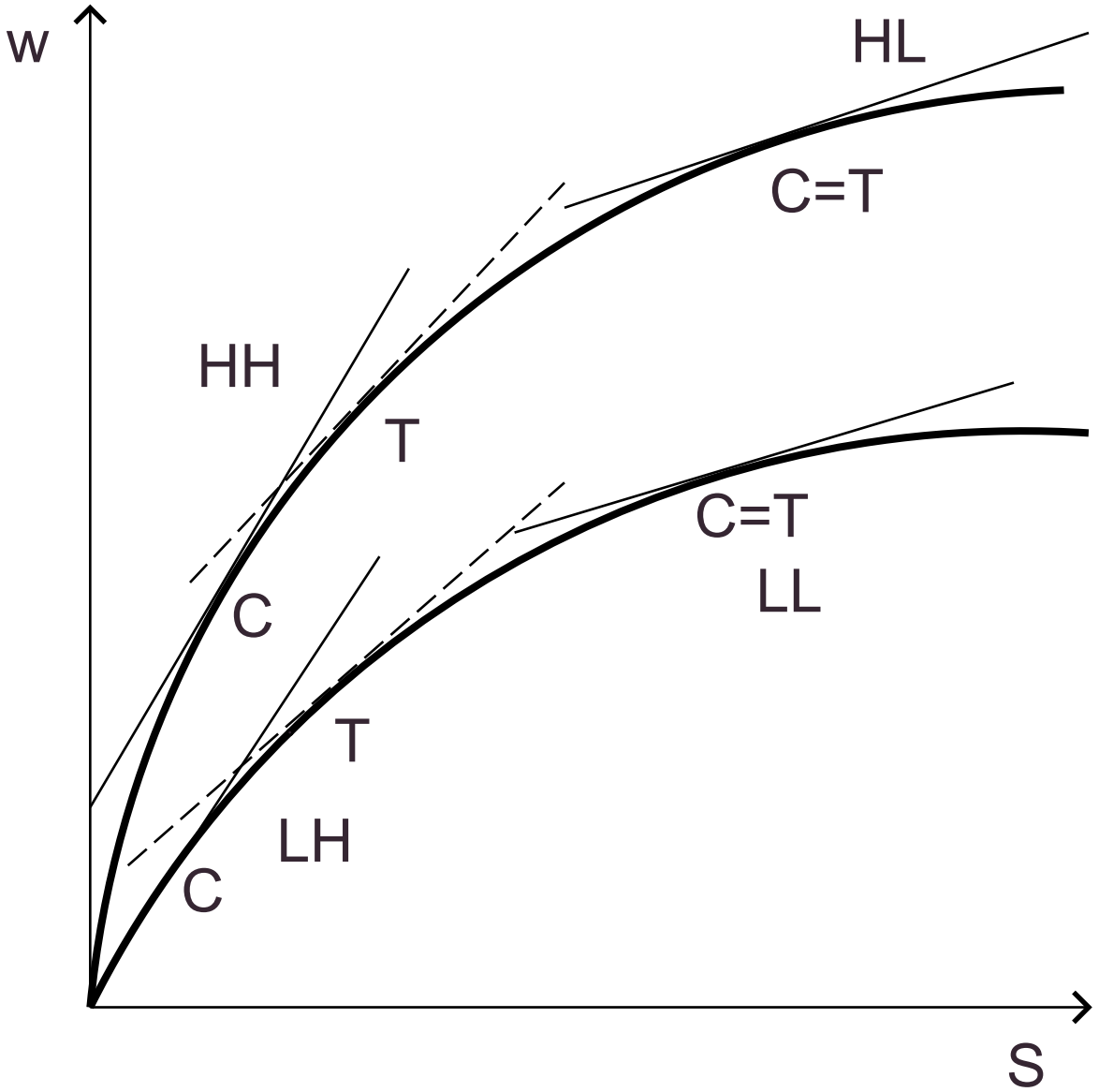

Fig. 20.5: Scelte dei controlli e dei trattati in un esperimento naturale