

Il rendimento dell'istruzione in Italia: evidenza empirica dai quantili di regressione

Pamela Giustinelli*

Università degli Studi di Verona

Questo lavoro si propone di fornire dell'evidenza empirica aggiornata sul "rendimento dell'istruzione in Italia" attraverso i "quantili di regressione". Tale metodologia ci permette di studiare il Quantile Treatment Effect dell'istruzione sulla distribuzione condizionale del reddito (che consideriamo riflettere quella dell'abilità non osservata), e di analizzare la relazione tra abilità ed istruzione, e gli effetti sul reddito generati dalla loro interazione. Le stime da noi ottenute presentano un andamento a "U", ovvero rendimenti maggiori ai quantili estremi della distribuzione del reddito, suggerendo una relazione di sostituibilità tra istruzione ed abilità, per livelli bassi dell'abilità, e di complementarità, per livelli elevati di essa. [Cod. JEL: C21, I20, J24, J31]

1. - Introduzione

I risultati di numerosi studi empirici, sulla relazione esistente tra il grado di istruzione degli individui ed il loro reddito, mostrano che i lavoratori più istruiti percepiscono retribuzioni più elevate (Ashenfelter e Rouse, 1998; Card, 1995). Da qui l'interesse e gli sforzi che gli economisti del lavoro hanno rivolto allo stu-

* <p-giustinelli@northwestern.edu>. Sono sinceramente grata a tre *referee* anonimi per i loro commenti su questo saggio, e al Prof. Gustavo Piga per i suoi utili suggerimenti. Ringrazio inoltre di cuore il mio Relatore di tesi Diego Lubian per la motivazione e l'entusiasmo che sempre mi trasmette, e per l'incoraggiamento, gli utili commenti e i consigli su questo lavoro.

dio del rendimento dell'istruzione, nel tentativo di darne una compiuta giustificazione teorica e una precisa quantificazione.

A fronte dell'elaborazione in letteratura di modelli economici sull'acquisizione dell'istruzione (come investimento mirante ad incrementare le capacità reddituali future), che tengono conto dell'endogeneità e degli elementi di eterogeneità individuale (l'abilità, l'ambiente familiare, *etc.*) che caratterizzano tale scelta (Becker, 1967; Card, 1994), esistono diversi problemi metodologici per la stima del rendimento dell'istruzione come esso viene definito in tali modelli.

Comuni difficoltà derivano dalla possibilità che il livello di istruzione dell'individuo sia osservato con errore, e dal problema dell'omissione di variabili rilevanti non osservabili (tipicamente l'abilità), con conseguenti distorsioni nelle stime. Ma il problema principale per la stima del rendimento dell'istruzione emerge dall'osservazione che chi è istruito potrebbe guadagnare di più non tanto per un effetto causale del maggior tempo trascorso a scuola, bensì perché più abile (Ichino, 2001). Allora la stima del coefficiente dell'istruzione nella regressione del reddito non esprimerebbe una misura (distorta) della relazione causale tra le due variabili, quanto, piuttosto, della loro correlazione spuria (*Problema dell'identificazione di causalità*). La maggiore difficoltà nella stima del rendimento dell'istruzione riguarda, dunque, il fatto che l'istruzione e il reddito degli individui sono congiuntamente determinati, venendo meno la causalità tra il reddito e l'istruzione medesimi.

Nel tentativo di trattare tali aspetti è andata sviluppandosi una copiosa letteratura (v. ad esempio la rassegna in Card, 2001; Card, 1995; Ichino e Winter-Ebmer, 1999; e per l'Italia, Flabbi, 1997; Brunello e Miniaci, 1999; Brunello, Comi e Lucifora, 2001) costituita da studi empirici in cui il rendimento dell'istruzione viene stimato utilizzando il *Metodo delle variabili strumentali* (IV). La maggior parte di tali lavori stimano però il rendimento marginale medio dell'istruzione, sotto l'ipotesi che le distribuzioni condizionali del reddito risultino semplicemente traslate l'una rispetto all'altra al variare dell'istruzione. In realtà, non è detto che tale ipotesi valga, né che una stima del rendimento marginale medio

dell'istruzione permetta di sintetizzare in modo efficace il fenomeno oggetto di studio. Da qui l'idea di stimare il rendimento dell'istruzione attraverso la metodologia dei *Quantili di regressione* (Koenker e Bassett, 1978), che permette di analizzare il rendimento dell'istruzione degli individui ai diversi quantili della distribuzione dei redditi. L'effetto dell'istruzione sul reddito potrebbe infatti essere diverso per gli individui che si trovano in "punti" differenti della distribuzione reddituale. Utilizzando i quantili di regressione è possibile, dunque, controllare se la distribuzione condizionale del reddito subisca solo un effetto di traslazione al variare dell'istruzione, o intervenga anche un effetto di scala, e quindi un cambiamento nella forma della distribuzione stessa. In particolare, poiché la distribuzione condizionale dei redditi tende a riflettere la distribuzione dell'abilità non osservata, una indagine basata sui quantili di regressione dovrebbe risultare informativa ed esplicativa della relazione tra l'abilità e l'istruzione, e degli effetti sul reddito generati dall'interazione tra queste variabili, senza che venga posta alcuna restrizione *a priori* su di essa. Inoltre, lo stimatore quantile campionario ha una interpretazione in termini di *Quantile Treatment Effect* (QTE): esso misura, per ogni quantile τ , il cambiamento nel reddito richiesto per restare, dopo il trattamento (un anno aggiuntivo di istruzione), al τ -esimo quantile della distribuzione condizionale (sotto l'ipotesi di *Rank Invariance*¹).

L'obiettivo del presente lavoro è quello di evidenziare come una indagine basata sui quantili possa risultare più informativa rispetto ad una stima OLS o *IV* valida per l'individuo medio. Da una parte, questo studio fornisce dell'evidenza empirica aggiornata sui rendimenti dell'istruzione in Italia (stime OLS, sui dati tratti dalla Banca d'Italia, 1993; 1995; 1998; 2000); dall'altra, l'aspetto centrale è rappresentato dalla stima dei quantili per questi dati.

Nel secondo paragrafo viene presentato brevemente il modello endogeno di Card (1994) sulla scelta di acquisizione dell'istru-

¹ L'ipotesi di *Rank Invariance* richiede che la posizione relativa degli individui nella distribuzione dei redditi non cambi dopo il trattamento.

zione e vengono analizzate le implicazioni di tale modello sullo stimatore OLS del rendimento marginale medio dell'istruzione. Viene anche effettuata una breve rassegna dei principali lavori empirici sul rendimento dell'istruzione condotti sui dati italiani, utilizzando stimatori OLS e IV. Nel terzo paragrafo viene presentata la teoria sui quantili di regressione e la loro interpretazione nel contesto di una indagine sul rendimento dell'istruzione. Sono poi presentati alcuni studi della letteratura internazionale sulle distribuzioni condizionali dei redditi attraverso i quantili di regressione. Infine nel sesto paragrafo sono descritti i dati, e nel settimo sono riportati i risultati delle stime da noi condotte. Seguono le conclusioni.

2. - Il rendimento dell'istruzione: teoria economica e analisi econometriche

2.1 Un modello endogeno per l'acquisizione dell'istruzione

Un modello endogeno per le scelte di istruzione è stato proposto da Card (1994), a partire dal modello di Becker (1967). In questo modello gli individui compiono le loro scelte d'istruzione attraverso un processo di ottimizzazione che si basa sul confronto dei benefici e dei costi derivanti dal proseguire gli studi. Il reddito e i costi dell'individuo vengono espressi come funzioni degli anni di istruzione, nelle quali si tiene conto degli aspetti di eterogeneità tra gli individui. Ciascun individuo effettua la propria scelta d'istruzione massimizzando una funzione di utilità:

$$U(Y, S) = \log(Y) - \phi(S)$$

dove S è il numero di anni d'istruzione, Y è il reddito medio annuo dell'individuo, e $\phi(S)$ è la sua funzione di costo. In particolare, si ipotizza che l'istruzione permetta di accumulare capitale umano, e che individui maggiormente dotati di capitale umano percepiscano redditi più elevati sul mercato del lavoro; quindi, la

funzione generatrice del reddito è definita da $Y = Y(S)$, con $Y' > 0$ e $Y'' < 0$. Il "reddito da capitale umano" aumenta all'aumentare del grado di istruzione, ma in modo decrescente, come avviene in generale per tutti fattori produttivi.

La funzione di costo $\phi(S)$ è strettamente convessa; cioè, i costi associati all'istruzione aumentano all'aumentare degli anni trascorsi a scuola in modo crescente. In $\phi'(S)$ sono compresi i costi monetari diretti (tasse scolastiche, libri, trasporto, etc.), i costi monetari indiretti o costi opportunità (mancati guadagni che si sarebbero ottenuti entrando immediatamente nel mercato del lavoro) e i costi non monetari (costi psicologici).

Considerando una funzione di utilità globalmente concava in S , esiste un (unico) numero ottimo di anni di istruzione, che si ottiene dalla soluzione delle condizioni del primo ordine:

$$(1) \quad \frac{Y'(S)}{Y(S)} = \phi'(S)$$

dove $Y'(S)/Y(S)$ rappresenta il tasso marginale di rendimento associato ad un anno aggiuntivo di scuola, e $\phi'(S)$ rappresenta il costo marginale di un anno di istruzione. Il significato economico della condizione di equilibrio è che l'individuo raggiunge il proprio livello ottimo di istruzione quando i benefici marginali ad esso associati eguagliano i costi marginali.

Per rendere operativo il modello, vengono assegnate una forma funzionale al rendimento marginale e una ai costi marginali. Card (1994) ipotizza che queste due funzioni siano lineari e incorporino nelle intercette i fattori di eterogeneità individuale:

$$(2) \quad \left[\frac{Y'(S)}{Y(S)} \right]_i = \beta_i(S) = b_i - k_b S \quad (k_b \geq 0)$$

$$(3) \quad [\phi'(S)]_i = \delta_i(S) = r_i + k_r S \quad (k_r \geq 0)$$

Le scelte di istruzione variano dunque per due ragioni: 1) le differenze nell'abilità, b_i , generano eterogeneità nei rendimenti marginali dell'istruzione degli individui; 2) le differenze nei vin-

coli di liquidità e nella condizione finanziaria e culturale delle famiglie, r_p generano eterogeneità nei costi marginali fronteggiati dagli individui.

Queste ipotesi sulla caratterizzazione dei rendimenti e dei costi marginali associati all'acquisizione di istruzione implicano a loro volta delle specifiche forme funzionali per $Y(S)$ e $\phi(S)$. Integrando la (2) e prendendo il logaritmo, si ottiene la funzione generatrice dei redditi, o *earning function*:

$$(4) \quad \log(Y_i) = a + b_i S - \frac{1}{2} k_b S^2$$

In tale specificazione, l'abilità influenza la pendenza della *earning function*; per cui, con preferenze omotetiche, gli individui dotati di maggior abilità sceglieranno livelli più elevati di istruzione².

La funzione di costo individuale è invece rappresentata da:

$$(5) \quad \phi(S) = c + r_i S + \frac{1}{2} k_r S^2$$

in cui l'elemento di eterogeneità determina una maggior inclinazione della funzione stessa.

Il livello ottimo di istruzione è dato da:

$$(6) \quad S_i^* = \frac{b_i - r_i}{k}$$

con $k = k_b + k_r$.

² Al contrario, se il termine di abilità fosse inserito nell'intercetta, a_p , la conclusione sarebbe opposta: in quanto individui con maggiori opportunità di reddito per ogni livello di istruzione potrebbero ben investire meno in istruzione, dal momento che avrebbero maggiori costi opportunità se continuassero a frequentare la scuola. In tal caso, però, si dovrebbe forse attribuire un diverso significato a questa componente. CARD D. (1994), richiamando HAUSE J.C. (1972), ricorre al concetto di "abilità cognitiva" (capacità di un soggetto di rielaborare ed applicare la conoscenza acquisita; cioè una forma di abilità che il soggetto esprime nel contesto lavorativo, e che lo porta a percepire redditi più elevati, indipendentemente dal suo livello di istruzione e dalle sue abilità in ambito scolastico). Naturalmente, unendo le due specificazioni in una unica, con l'abilità incorporata sia nell'intercetta sia nella pendenza della *earning function*, l'effetto finale sulla scelta di istruzione sarebbe (*a priori*) incerto.

Dove le equazioni (6) e (4) assieme determinano la distribuzione congiunta dei redditi e dell'istruzione³.

Il rendimento marginale dell'istruzione in corrispondenza della scelta ottima è dato da:

$$(7) \quad \beta_i^* = b_i - k_b S_i^* = \left(1 - \frac{k_b}{k_b + k_r}\right) b_i + \frac{k_b}{k_b + k_r} r_i$$

e rappresenta l'effetto causale dell'istruzione sul reddito dell'individuo i^4 . Tuttavia, a causa del «problema fondamentale dell'inferenza causale» (Holland, 1986), esso non può essere né identificato né misurato. Ciò che può essere identificato, invece, e che rappresenta un utile parametro di riferimento con cui confrontare gli stimatori (OLS e IV) del rendimento dell'istruzione è il rendimento marginale medio dell'istruzione nella popolazione:

$$(8) \quad \bar{\beta} = E[\beta_i] = E[b_i - k_b S_i] = \bar{b} - k_b \bar{S} = \frac{k_r}{k_b + k_r} \bar{b} + \frac{k_b}{k_b + k_r} \bar{r}$$

³ L'evidenza empirica (CARD D. - KRUEGER A.B., 1992; PARK J.H., 1994) testimonia a favore di una relazione per dati sezionali approssimativamente lineare tra il logaritmo del reddito e gli anni d'istruzione. Tale circostanza sembrerebbe inconsistente con l'equazione (4), a meno di $k_b = 0$. In realtà, nonostante la relazione proposta dal modello sia quadratica, i dati generati dal modello stesso potrebbero presentarsi secondo una relazione lineare. Da una parte, infatti, l'andamento dovrebbe essere concavo, dato che, a parità di b_i , gli individui con tasso marginale di sconto più basso tendono a scegliere livelli più elevati di istruzione. Dall'altra, poiché, tra gli individui con diversa abilità, quelli maggiormente dotati sceglieranno livelli più elevati di istruzione, tale relazione dovrebbe risultare convessa. In definitiva, la relazione reddito-istruzione nella popolazione è determinata dalla combinazione dei due effetti, e tenderà maggiormente ad essere concava quanto minore è la varianza dell'abilità b_i rispetto a quella del tasso di sconto r_i .

⁴ Come correttamente osservato da un anonimo *referee*, che colgo l'occasione di ringraziare, sembrerebbe dunque di dover fronteggiare un problema di aggregazione, che peraltro mai viene esplicitamente rilevato nella letteratura di riferimento. In realtà, le specificazioni da noi stimate, eq. (31), (rispettivamente nella versione con gli anni di istruzione e in quella con le variabili dicotomiche per i titoli di studio) e i dati a nostra disposizione ci permettono di effettuare una indagine *cross-section* dei rendimenti marginali dell'istruzione, che sono medi, nel caso della stima OLS, e quantilici nel caso della stima con i quantili di regressione. Come infatti da noi precedentemente osservato, a fronte dei modelli economici sull'acquisizione dell'istruzione elaborati in letteratura, che incorporano ad esempio le caratteristiche di eterogeneità individuale proprie di tale scelta (BECKER G.S., 1967; CARD D., 1994), esistono diversi problemi metodologici e di stima del rendimento dell'istruzione così come esso viene definito in tali modelli.

A partire dalla specificazione:

$$\log(Y_i) = \alpha + \rho S_i + \varepsilon_i$$

è possibile dimostrare che il limite in probabilità dello stimatore OLS è:

$$(9) \quad \text{plim}(\hat{\rho}_{OLS}) = \left(1 - \frac{k_b}{k} + \lambda\right) \bar{b} + \left(\frac{k_b}{k} - \lambda\right) \bar{r} = \bar{\beta} + \lambda(\bar{b} - \bar{r})$$

da cui l'inconsistenza per il rendimento marginale medio dell'istruzione.

La distorsione è positiva, in quanto prodotto di due fattori positivi (λ rappresenta la frazione della varianza di S attribuibile alla variabilità di b ; mentre \bar{b} è sicuramente maggiore di \bar{r} , dato che il livello ottimo di istruzione non può essere negativo). Essa inoltre tenderà ad aumentare tanto maggiore è σ_b^2 rispetto a σ_r^2 , e tanto più ampia è la differenza $(\bar{b} - \bar{r})$. Il termine $\lambda(\bar{b} - \bar{r})$ può essere interpretato come una *endogeneity bias* dovuta al fatto che gli individui che hanno un rendimento marginale dell'istruzione maggiore (perché più abili), o un minor costo marginale (perché meno "vincolati monetariamente"), tendono a studiare più a lungo. In altri termini, lo stimatore OLS del coefficiente degli anni di istruzione è influenzato dal modo in cui gli individui, caratterizzati dal proprio livello di abilità e dai propri vincoli di liquidità, sono distribuiti nella popolazione⁵.

2.2 L'ability bias

Una ulteriore difficoltà nella stima del rendimento dell'istruzione deriva dalla possibilità che siano presenti fattori non osser-

⁵ Ammettendo l'esistenza di una correlazione negativa tra b_i ed r_i , si osserveranno presumibilmente con maggior frequenza individui con abilità elevata e bassi vincoli di liquidità e viceversa, ossia individui con un alto grado di istruzione e redditi elevati, oppure individui con un basso livello di istruzione e basso reddito. Quindi la retta di regressione avrà pendenza positiva e la sua inclinazione tenderà a crescere all'aumentare della frequenza relativa con cui si presentano tali osservazioni.

vati di eterogeneità nel livello dei redditi. L'approccio adottato in letteratura per affrontare la questione della cosiddetta *ability bias* è quello di considerare una funzione generatrice dei redditi che contenga una componente specifica individuale (Griliches, 1977).

Card (1994) introduce tale aspetto aggiungendo un'intercetta individuale a_i nella *earning function*:

$$(10) \quad \log(Y_i) = a_i + b_r S_i - \frac{1}{2} k_b S_i^2$$

Di conseguenza, l'espressione per il limite in probabilità dello stimatore OLS del coefficiente degli anni di istruzione diventa:

$$(11) \quad \begin{aligned} \text{plim}(\hat{\rho}_{OLS}) &= \frac{\text{Cov}[\log(Y_i), S_i]}{\text{Var}(S_i)} = \\ &= \frac{E\left[a_i(S_i - \bar{S}) + b_r S_i(S_i - \bar{S}) - \frac{1}{2} k_b S_i^2(S_i - \bar{S}) \right]}{\text{Var}(S_i)} \end{aligned}$$

Una correlazione tra a_i e S_i può sussistere per due motivi: perché a_i è correlato con b_i , oppure con r_i , cioè la parte di reddito che l'individuo è potenzialmente in grado di percepire grazie a sue peculiari abilità (di tipo cognitivo) è correlata ai fattori che determinano l'abilità (in ambito scolastico) e il tasso individuale di sconto, quali il benessere familiare, il substrato culturale, *etc.*, dell'individuo stesso. Analiticamente, poiché:

$$(12) \quad E[a_i(S_i - \bar{S})] = E\left[a_i \frac{(b_i - \bar{b}) - (r_i - \bar{r})}{k} \right] = \frac{1}{k} (\sigma_{ab} - \sigma_{ar})$$

l'espressione per il termine di distorsione risulta essere:

$$(13) \quad \frac{E[a_i(S_i - \bar{S})]}{\text{Var}(S_i)} = k \frac{(\sigma_{ab} - \sigma_{ar})}{(\sigma_b^2 + \sigma_r^2 - 2\sigma_{br})}$$

Si può ragionevolmente supporre che a_i e b_i siano positivamente correlati ($\sigma_{ab} > 0$), mentre a_i e r_i lo siano negativamente ($\sigma_{ar} < 0$). In particolare, una correlazione positiva tra a_i e b_i può presentarsi se a_i esprime una qualche misura di abilità cognitiva dell'individuo e se il grado di istruzione e tale genere di abilità sono complementa-

ri. Invece una correlazione negativa tra a_i e r_i può derivare dalla circostanza che i figli delle famiglie più ricche abbiano tassi di sconto più bassi, o una naturale propensione a studiare più a lungo, e che questi individui tendano in generale a raggiungere livelli di reddito più elevati in virtù di migliori accessi al mercato del lavoro.

2.3 Variabile esplicativa osservata con errore

Consideriamo, infine, la possibilità che la variabile esplicativa “anni di istruzione” sia osservata con errore, cioè gli anni di istruzione osservati (S_i^0) differiscono dal vero livello di istruzione (S_i) per un termine d'errore additivo:

$$S_i^0 = S_i + \varepsilon_i$$

dove ε_i ha media 0, varianza σ_ε^2 , ed è incorrelato con Y_i . In questo caso, il limite in probabilità dello stimatore OLS è:

$$(14) \quad \text{plim}(\hat{\rho}_{OLS}^0) = \frac{\text{Cov}[\log(Y_i), S_i^0]}{\text{Var}(S_i^0)}$$

Sfruttando la relazione:

$$\frac{\text{Cov}[\log(Y_i), S_i^0]}{\text{Cov}[S_i^0, S_i]} = \frac{\text{Cov}[\log(Y_i), S_i]}{\text{Var}(S_i)}$$

il limite in probabilità (14) può essere scritto come:

$$(15) \quad \text{plim}(\hat{\rho}_{OLS}^0) = \frac{\text{Cov}[\log(Y_i), S_i]}{\text{Var}(S_i)} \cdot \frac{\text{Cov}[S_i^0, S_i]}{\text{Var}(S_i^0)}$$

dove il primo fattore non è altro che il limite in probabilità dello stimatore OLS nel caso *standard*, mentre il secondo rappresenta il termine di distorsione. È immediato verificare che tale distorsione porta ad una sottostima del parametro ρ , infatti:

$$\text{plim}(\hat{\rho}_{OLS}^0) = \frac{\text{Cov}[\log(Y_i), S_i]}{\text{Var}(S_i)} \cdot \frac{\text{Cov}[S_i^0, S_i]}{\text{Var}(S_i^0)}$$

In conclusione, tenendo conto di tutte le problematiche ana-

lizzate inerenti alla procedura di stima, l'espressione per il limite in probabilità dello stimatore OLS sarà:

$$(16) \quad \text{plim}(\hat{p}_{OLS}^0) = \theta_0 \left[\bar{\beta} + \lambda(\bar{b} - \bar{r}) + k \frac{(\sigma_{ab} - \sigma_{ar})}{\sigma_b^2 + \sigma_r^2 - 2\sigma_{br}} \right]$$

in cui sono presenti tre fonti di distorsione nello stimatore OLS rispetto al rendimento marginale medio dell'istruzione $\bar{\beta}^6$: 1) la componente dovuta alla *endogeneity bias*, $\lambda(\bar{b} - \bar{r})$; 2) la *ability bias* attribuibile alla presenza di componenti non osservate di eterogeneità nel livello dei redditi,

$$k \frac{(\sigma_{ab} - \sigma_{ar})}{(\sigma_b^2 + \sigma_r^2 - 2\sigma_{br})}$$

3) la distorsione verso il basso generata da errori di misura della variabile anni di istruzione, θ_0 .

2.4 Il rendimento dell'istruzione in Italia nei precedenti studi

Nella seconda metà degli anni '80 e durante gli anni '90, sono stati condotti numerosi studi empirici rivolti alla stima del rendimento dell'istruzione in Italia (nella tavola 1 sono riassunti i risultati delle principali stime). Come osservato da Brunello e Miniaci (1999), i primi studi si basavano su stime OLS effettuate su dati eterogenei e non sempre rappresentativi⁷.

⁶ Occorre invece ricordare che una procedura di stima IV dovrebbe fornire stime immuni da tali distorsioni. D'altra parte, strumenti diversi tendono a generare stime differenti dei rendimenti medi per differenti sottogruppi nella popolazione (ICHINO A. - WINTER-EBMER R., 1999). Infatti, una stima IV misura il rendimento marginale degli individui che, nel contesto dell'esperimento naturale considerato, sono *compliers*, con una tendenza così a sovrastimare il rendimento marginale medio dell'istruzione nella popolazione, dato che tali individui hanno tipicamente rendimenti più elevati (*discount rate bias*) (CARD D., 1994).

⁷ Ad esempio, ANTONELLI G. (1985) stima una equazione minceriana *standard* con gli OLS, utilizzando un *data-set* a carattere regionale, e ottenendo un rendimento pari al 4,6%. Medesimo risultato viene trovato da CANNARI L. - PELLEGRINI G. - SESTITO P. (1989), i quali utilizzano un campione più ampio tratto da BANCA D'ITALIA (1986). LUCIFORA C. - REILLY B. (1990) (sulla base dei dati ENI-IRI sui redditi individuali) effettuano una stima OLS per genere, trovando che il rendimento marginale dell'istruzione è più elevato per le donne che per gli uomini.

TAV. 1

I RENDIMENTI DELL'ISTRUZIONE IN ITALIA IN PRECEDENTI STUDI

autore	dati	campione	stime		strumenti
			OLS	IV	
Antonelli (1985) Cannari, Pellegrini e Sestito (1989)	ER SHIW 1986	(uomini) (uomini)	0,046	-	
			0,046	-	
Lucifora e Reilly (1990)	ENI-IRI SHIW 1993	(donne) (uomini)	0,040	-	istruzione dei genitori
			0,036	-	
Cannari e D'Alessio (1995)	SHIW 1993	(uomini)	0,045	0,070	istruzione dei genitori
Colussi (1997)	SHIW 1993	(uomini)	0,062	0,076	istruzione dei genitori
Flabbi (1997)	SHIW 1991	(donne)	0,022	0,056	riforme 1962 e 1969, vicinanza all'università
			0,017	0,062	
Brunello e Miniaci (1999)	SHIW 1993-1995	(uomini)	0,048	0,057	istruzione dei genitori, riforma 1969
Brunello, Comi e Lucifora (2001)	SHIW 1995 (esperienza) (età)	(donne) (uomini)	0,077	-	istruzione e professione genitori, riforma 1969
			0,062	-	
Martins e Pereira (2004)	SHIW 1995	OLS	-	0,077	+ avversione al rischio
			0,048	0,059	
			0,048	0,061	
			0,062		
		Quantili	%		
			10	0,065	
			20	0,063	
			30	0,057	
			40	0,057	
			50	0,056	
			60	0,057	
			70	0,061	
			80	0,065	
			90	0,068	

Gli studi più recenti, a partire dalla seconda metà degli anni '90, utilizzano ampiamente i dati delle *Indagini sui bilanci delle famiglie*, effettuando stime IV del rendimento dell'istruzione in Italia. Cannari e D'Alessio (1995), con i dati del 1993 (Banca d'Italia, 1993), e sulla base di variabili strumentali relative al substrato familiare, ottengono una stima vicina al 7%, più elevata di quelle ottenute dalle ricerche precedenti. Colussi (1997) ottiene una stima del 6,6%, utilizzando i dati relativi al medesimo anno e un gruppo simile di strumenti.

Flabbi (1997) stima il rendimento dell'istruzione per le donne e per gli uomini separatamente (dati Banca d'Italia, 1991), utilizzando come strumenti: la variabile binaria "province", indicatore della presenza o meno di sedi universitarie nella provincia di residenza dell'individuo quando questi aveva 19 anni; e la variabile "riforme", costruita considerando come eventi esogeni le riforme del sistema scolastico italiano del '62 (media unica) e del '69 (accesso a tutte le facoltà universitarie per i diplomati presso qualunque tipo di istituto secondario superiore). Le stime IV sono 0,56 per le donne e 0,62 per gli uomini. Il risultato nuovo ottenuto dall'autore è dato dal rovesciamento della "gerarchia" nei risultati. Le stime IV del rendimento dell'istruzione risultano infatti maggiori per gli uomini, mentre le stime OLS ottenute da Flabbi (1997) sono 0,22 per le donne e 0,17 per gli uomini. Obiettivo di Flabbi (1997), (1999) è quello di chiarire se l'usuale gerarchia, ovvero un rendimento femminile superiore a quello maschile, possa essere considerata un risultato empirico acquisito nel mercato del lavoro italiano o non dipenda piuttosto dalle tecniche di stima utilizzate. La conclusione dell'autore è che tale gerarchia è confermata solo in parte. Essa tende a valere, anche se in modo non indipendente dalla specificazione, quando si applica la procedura OLS. È invece rovesciata dalle stime IV. Una prima interpretazione è rappresentata dall'osservazione che la distorsione delle stime con i minimi quadrati ordinari sarebbe tale da nascondere il vero rapporto relativo dei rendimenti in base al genere. Ma è possibile anche una interpretazione alternativa, e cioè che l'inversione non varrebbe in realtà per tutto il campione, ma solo per coloro che presentano un rendimento dell'istruzione più elevato

(ad esempio perché provenienti da famiglie meno abbienti). L'interpretazione economica di un rendimento dell'istruzione superiore per il gruppo di trattamento maschile, rispetto a quello del gruppo femminile, porterebbe allora ad ipotizzare variabilità dei rendimenti all'interno della popolazione e discriminazione di genere pre-mercato del lavoro a svantaggio delle donne.

Brunello e Miniaci (1999) e Brunello, Comi e Lucifora (2001), utilizzando i dati della Banca d'Italia '93 e '95, stimano il rendimento dell'istruzione, con variabili strumentali relative al *background* familiare (istruzione e posizione professionale dei genitori), alla riforma scolastica del '69 e (solo i secondi) al grado di avversione al rischio dell'individuo. Brunello e Miniaci (1999) ottengono una stima OLS del 4,8% e una stima IV del 5,7% per i capifamiglia maschi. Valori simili sono ottenuti da Brunello, Comi e Lucifora (2001). Le stime IV mantengono però la configurazione di quelle OLS, relativamente alla gerarchia dei rendimenti femminili e maschili. Brunello, Comi e Lucifora (2001) ottengono infatti rendimenti dell'istruzione più elevati per i campioni femminili, anche con le stime IV, a differenza di Flabbi (1997), (1999).

3. - L'approccio dei "quantili di regressione" nello studio della relazione reddito-istruzione-abilità

3.1 Lo stimatore dei quantili

Il modello dei quantili di regressione, introdotto da Koenker e Bassett (1978), estende la nozione degli ordinari quantili del modello di locazione alla più generale classe dei modelli lineari, nei quali i quantili condizionali hanno appunto la forma di funzioni lineari (nei parametri).

I quantili di regressione possono dunque considerarsi una metodologia statistica, volta a stimare le funzioni quantile condizionale, e a condurre inferenza su di esse. Infatti, come il modello lineare classico costituisce un modello di media condizionale, i quantili di regressione rappresentano dei modelli di mediana con-

dizionale, o delle altre funzioni quantile condizionale. Un ben noto caso particolare di quantile di regressione è costituito dallo stimatore LAD, proposto da Koenker e Bassett (1978), che adatta la mediana ad una funzione lineare di variabili esplicative, minimizzando la somma dei valori assoluti dei residui.

La stima LAD è potenzialmente attrattiva per le medesime ragioni per cui la mediana può costituire un indice di localizzazione migliore rispetto alla media: esiste sempre all'interno della distribuzione considerata; è rappresentativa della posizione della distribuzione stessa anche in presenza di valori anomali, limitandosi a tener conto solo della modalità dell'elemento che occupa una certa posizione nella distribuzione delle osservazioni ordinate. Analogamente, gli stimatori dei quantili sono robusti alla presenza di *outliers* nelle osservazioni sulla variabile dipendente, e si dimostrano più efficienti degli stimatori OLS, nei casi in cui il termine di errore non possieda una distribuzione normale. Inoltre, dal punto di vista computazionale (ed estetico), il modello dei quantili di regressione può essere formulato come un problema di Programmazione Lineare (LP), il che agevola la stima e permette di semplificare i problemi computazionali⁸. Esso può essere anche "contestualizzato" nello schema del Metodo Generalizzato dei Momenti (GMM), il che si rivela utile per la determinazione delle proprietà asintotiche dello stimatore quantile campionario. Infine, un aspetto altamente interessante e peculiare di questa metodologia è costituito dal fatto che le differenti soluzioni per i diversi quantili possono essere interpretate come differenti risposte della variabile dipendente, nei vari punti della distribuzione condizionale della variabile dipendente stessa, alle variazioni nei regressori⁹.

⁸ La rappresentazione del problema di minimizzazione del quale lo stimatore quantile campionario è soluzione eq. (18) come programma lineare è riportata in *Appendice*. Per una trattazione sull'utilizzo della programmazione lineare nel contesto delle regressioni quantiliche v. BUCHINSKY M. (1995), e KOENKER R. - BASSETT G. (1978) per un semplice ma esplicativo esempio bivariato con cinque osservazioni. Si rimanda inoltre a HILLIER F.S. - LIEBERMAN G.J. (1990) per la necessaria teoria di LP.

⁹ Per chiarire tale affermazione, richiamiamo qualche applicazione di tale tecnica, mutuandola da KOENKER R. (2003). I quantili di regressione sono stati uti-

Il τ -esimo quantile di regressione è definito da:

$$(17) \quad Q_\tau(y_i|X=x) = x'\beta_\tau$$

dove x_i è il vettore $k \times 1$ dell' i -esima osservazione su k regressori, e y_i l' i -esima osservazione sulla variabile dipendente. Il processo $u_{\tau i} = y_i - x_i'\beta_\tau$ possiede funzione di ripartizione $F_{u_\tau}(\cdot)$ e, per definizione, $Q_\tau(u_{\tau i}|x_i) = 0$. Lo stimatore quantile campionario $\hat{\beta}_\tau$ è definito come qualsiasi soluzione al problema di minimizzazione¹⁰.

$$(18) \quad \min_{\beta_\tau \in \mathbb{R}^k} \sum_{i=1}^n (|y_i - x_i'\beta_\tau| \cdot \tau \cdot 1\{y_i > x_i'\beta_\tau\} + |y_i - x_i'\beta_\tau| \cdot (1 - \tau) \cdot 1\{y_i \leq x_i'\beta_\tau\})$$

Alla luce di tale rappresentazione, preme sottolineare la proprietà di robustezza di tale stimatore, influenzato solo dal comportamento locale della distribuzione condizionale della variabile dipendente vicino al quantile specificato. Data una soluzione $\hat{\beta}_\tau$, basata sulle osservazioni $\{y, X\}$, fintanto che non viene cambiato il segno dei residui $\hat{u}_\tau = y - X\hat{\beta}_\tau$, qualsiasi osservazione su y può essere arbitrariamente modificata senza alterare la soluzione iniziale. Solo i segni degli scarti entrano nella determinazione delle

lizzati, ad esempio, per studiare l'impatto di varie caratteristiche demografiche e del comportamento materno sul peso dei bambini alla nascita, permettendo in particolare di concentrare l'analisi sulla coda inferiore della distribuzione, ossia sui bambini sottopeso. Questa risulta infatti di particolare interesse per i ricercatori, in quanto un basso peso alla nascita viene comunemente associato ad una serie di conseguenti problemi di salute, nonché ai risultati scolastici e professionali che gli individui mediamente raggiungono. Ancora, negli studi sulle relazioni esistenti tra i risultati ottenuti dagli studenti di scuole pubbliche in esami standardizzati e alcune variabili socio-economiche, quali il reddito familiare e il livello di istruzione dei genitori, e normative, quali la numerosità delle classi e la qualifica degli insegnanti, i quantili di regressione vengono utilmente impiegati per capire se gli interventi normativi influiscono sulle *performance* degli studenti migliori nello stesso modo che su quelle dei meno brillanti. Non sembra plausibile, infatti, che gli effetti di tali variabili agiscano in modo da traslare l'intera distribuzione dei risultati dei *test* di una misura fissata.

¹⁰ In *Appendice* vengono ripercorsi i passaggi dalla definizione di quantile nel modello di locazione, e dalla sua rappresentazione come soluzione di un problema di ottimizzazione, a quella analoga nel modello lineare, rappresentata appunto dall'equazione (18).

stime dei quantili di regressione; quindi, le osservazioni estreme influenzano il risultato in quanto si trovino al di sopra o al di sotto dell'iperpiano stimato, ma quanto al di sopra, o al di sotto, è irrilevante. È utile inoltre osservare come, allo stesso modo dello stimatore OLS, anche lo stimatore dei quantili gode di una serie di "proprietà di equivarianza". E in più, a differenza di quello, lo stimatore quantile campionario soddisfa la proprietà di invarianza alle trasformazioni monotone¹¹, la quale permette di superare i problemi di stima dei modelli trasformati, che si presentano nel contesto della media condizionale¹².

Il problema (18) appena specificato può essere riscritto come:

$$(19) \quad \min_{\beta_\tau \in B_\tau} \frac{1}{n} \sum_{i=1}^n (\tau - 1/2 + 1/2 \operatorname{sgn}(y_i - x_i' \beta_\tau)) (y_i - x_i' \beta_\tau)$$

dove $\operatorname{sgn}(\alpha) = I(\alpha \geq 0) - I(\alpha < 0)$, da cui seguono le condizioni del primo ordine

$$(20) \quad \sum_{i=1}^n (\tau - 1/2 + 1/2 \operatorname{sgn}(y_i - x_i' \hat{\beta}_\tau)) x_i = 0$$

che, se considerate all'interno dello schema del *Metodo generalizzato dei momenti (GMM)*, permettono di stabilire le proprietà in grandi campioni dello stimatore quantile campionario $\hat{\beta}_\tau$. Seguendo Buchinsky (1995), si definisce la seguente funzione dei momenti:

$$(21) \quad \psi(x_i, y_i, \beta_\tau) = (\tau - 1/2 + 1/2 \operatorname{sgn}(y_i - x_i' \beta_\tau)) x_i$$

per la quale, sotto le ipotesi del modello (17), valgono le condizioni di ortogonalità:

$$E[\psi(x_i, y_i, \beta_\tau)] = 0$$

¹¹ Per ogni funzione monotona $h(\cdot)$ vale $Q_\tau(h(Y)|x) = h(Q_\tau(Y|x))$, mentre normalmente $E(h(Y)|x) \neq h(E(Y|x))$.

¹² Per una trattazione dettagliata di tali proprietà v. KOENKER R. - BASSETT G. (1978).

Nonostante tale funzione dei momenti non soddisfi la condizione di differenziabilità è possibile ottenere la distribuzione asintotica ($n \rightarrow \infty$) di $\hat{\beta}_\tau$, riassunta dalla seguente proposizione:

PROPOSIZIONE 1 (Buchinsky, 1995)

Se β_τ è nell'interno di B_τ dove B_τ è l'insieme compatto dei parametri in \mathbb{R}^k ; $F_{u_\tau}(0|x) = \tau$ con probabilità 1; $F_{u_\tau}(\cdot|x)$ è continua con densità $f_{u_\tau}(0|x) > 0$; e $n^{-1} \sum_{i=1}^n x_i x_i' \xrightarrow{p} E[x_i x_i']$, matrice definita positiva; allora:

$$(22) \quad \sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N(0, \Lambda_\tau)$$

dove:

$$(23) \quad \Lambda_\tau = \tau(1-\tau) (E[f_{u_\tau}(0|x_i) x_i x_i'])^{-1} E[x_i x_i'] (E[f_{u_\tau}(0|x_i) x_i x_i'])^{-1}$$

Nel contesto dei quantili di regressione si assume talvolta che la distribuzione di $u_{\tau i} = y_i - x_i' \beta_\tau$ non dipenda da x_i (per cui $E(u_{\tau i} | x_i) = 0$). In tal caso la (23) si semplifica in:

$$\Lambda_\tau = \frac{\tau(1-\tau)}{f_{u_\tau}^2(0)} (E[x_i x_i'])^{-1}$$

come in Koenker e Bassett (1978).

Un possibile metodo per stimare in modo consistente la matrice di varianza-covarianza Λ_τ è rappresentato dal *bootstrap*, utilizzato nella parte empirica. Brevemente, si consideri un campione casuale $(\tilde{y}_i, \tilde{x}_i)$, con $i = 1, \dots, n$, generato dalla funzione di distribuzione empirica $F_n(x, y)$. Sia $\hat{\beta}_\tau$ la stima *bootstrap* ottenuta da un quantile di regressione di \tilde{y}_i su \tilde{x}_i . Se la procedura di campionamento e di stima viene ripetuta M volte, ottenendo $\hat{\beta}_{\tau 1}, \hat{\beta}_{\tau 2}, \dots, \hat{\beta}_{\tau M}$, si ha che:

$$(24) \quad \hat{V}^M(\hat{\beta}_\tau) = \frac{n}{M} \sum_{j=1}^M \left(\hat{\beta}_{\tau j} - \hat{\beta}_\tau \right) \left(\hat{\beta}_{\tau j} - \hat{\beta}_\tau \right)'$$

costituisce lo stimatore *bootstrap* della varianza asintotica di $\hat{\beta}_\tau$.

3.2 Quantili di regressione e abilità non osservata

Come precedentemente osservato, un importante aspetto di questa metodologia è rappresentato dall'interpretazione delle stime dei diversi quantili come le differenti risposte della variabile dipendente alle variazioni nei regressori in diversi punti della distribuzione condizionale della variabile dipendente stessa. In particolare, lo stimatore quantile campionario ha una interpretazione in termini di Quantile Treatment Effect (QTE): esso misura, per ogni quantile τ , il cambiamento nella variabile dipendente richiesto per restare, dopo il trattamento, al τ -esimo quantile della distribuzione condizionale.

I quantili di regressione rappresentano pertanto un approccio più flessibile per caratterizzare l'effetto dell'istruzione su differenti percentili della distribuzione condizionale dei redditi, per analizzare la relazione tra abilità ed istruzione e gli effetti sul reddito generati dalla loro interazione.

Nel paragrafo 2 si è mostrato come l'abilità non osservata induca eterogeneità nella distribuzione condizionale dei redditi, influenzando sia l'intercetta sia il coefficiente della *earning function* (10). Consideriamo dunque, seguendo Arias, Hallock e Sosa-Escudero (2001), una specificazione più generica dell'equazione minceriana:

$$(25) \quad \ln(Y_i) = \alpha X_i + \beta_0 S_i + \varphi(S_i, v_i) + v_i$$

dove X_i raggruppa alcune variabili di controllo (quali l'età dell'individuo, l'esperienza, etc.), la funzione φ esprime l'interazione esistente tra l'istruzione e l'abilità degli individui, e v_i è un termine di errore che contiene anche l'abilità non osservata: $v_i = \gamma A_i + \varepsilon_i$. Questa specificazione corrisponde alla (10), con $\varphi = b_r S_i - 0,5 k_b S_i^2$ e $\gamma A_i = a_r$.

Si consideri ora il caso in cui il termine di interazione sia semplicemente $\varphi = \delta A_r S_r$, così che il rendimento dell'istruzione sia dato da:

$$(26) \quad \partial \ln(Y_i) / \partial S_i \equiv \beta_i = \beta_0 + \delta A_i$$

dove δ cattura l'effetto dell'abilità sul rendimento dell'istruzione. Se $\delta < 0$, il rendimento decresce all'aumentare dell'abilità, e viceversa. Questo conduce a un modello con coefficienti casuali, tale che lo stimatore OLS applicato alla (25) stima consistentemente:

$$(27) \quad \partial E(\ln(Y_i)|X_i, S_i)/\partial S_i = \beta_0 + \delta \bar{A}$$

ossia il rendimento dell'istruzione per un individuo con abilità media, o *average treatment effect*. Questo approccio però si basa su una parametrizzazione restrittiva dell'interazione tra istruzione ed abilità non osservata. Infatti, nel modello (26) si presuppone che β_i sia una funzione monotonica dell'abilità.

Invece, con S_i esogeno e la restrizione sul termine di errore $Q_\tau(v_i|S_i) = 0$, l'effetto dell'istruzione sul τ -esimo quantile condizionale di Y_i è dato da:

$$(28) \quad \partial Q_\tau(\ln(Y_i)|X_i, S_i)/\partial S_i = \partial \ln(Q_\tau(Y_i)|X_i, S_i)/\partial S_i$$

$$(29) \quad = \beta_0 + \partial Q_\tau(\varphi(S_i, v_i)|X_i, S_i)/\partial S_i$$

$$(30) \quad = \beta_0 + G_v^{-1}(\tau|X_i, S_i) \equiv \beta_\tau$$

dove G_v è una qualche trasformazione della distribuzione dell'abilità nella popolazione, e β_τ può essere considerato una misura del QTE dell'istruzione sui redditi, dato $\tau \in (0, 1)$ ¹³. I quantili di regressione per differenti valori di τ conducono alla stima di una intera famiglia di rendimenti dell'istruzione che riflettono la distribuzione dell'abilità tra gli individui. L'interazione tra istruzione ed abilità può allora essere analizzata confrontando β_τ a differenti quantili τ_k e τ_s , con $k \neq s$.

3.3 Quantili di regressione e distribuzione del reddito nella letteratura internazionale

Numerosi studi, oltre a quello di Arias, Hallock e Sosa-Escudero (2001), sono stati recentemente condotti utilizzando i quan-

¹³ Nel caso particolare della specificazione (26), l'equazione (30) diventa $\partial \ln(Q_\tau(Y_i)|X_i, S_i)/\partial S_i \equiv \beta_\tau = \beta_0 + \delta Q_\tau(A_i)$.

tili di regressione al fine di analizzare la distribuzione condizionale dei redditi di differenti paesi: tra questi, ad esempio, Buchinsky (1994); Mwabu e Schultz (1996); Martins e Pereira (2004); Fitzenberger e Kurz (2003).

In particolare, Buchinsky (1994), pur non trattando esplicitamente l'interazione tra educazione ed abilità, analizza i cambiamenti verificatisi negli ultimi decenni nella struttura dei salari degli Stati Uniti (da una parte, le cresciute disuguaglianze retributive, anche dopo aver controllato per le caratteristiche individuali; dall'altra, l'aumento del rendimento di fattori come l'istruzione e l'esperienza), utilizzando i dati del *Current Population Survey (CPS)*. Secondo l'autore risulta essenziale esaminare tali cambiamenti ai diversi punti della distribuzione: infatti, negli anni più recenti i redditi sono sì aumentati, in generale, a tutti i quantili della distribuzione, al crescere degli anni di istruzione, ma le differenze negli incrementi sono maggiori in corrispondenza di determinati quantili (ad esempio, l'aumento del rendimento dell'istruzione viene registrato come più sensibile ai quantili più alti della distribuzione dei redditi).

Tale configurazione viene riscontrata anche da Martins e Pereira (2004), i quali conducono una stima del rendimento dell'istruzione per 16 paesi europei con l'intento di analizzare le differenze negli incrementi reddituali determinati dall'istruzione lungo l'intera distribuzione dei salari. In tal modo essi confrontano il rendimento dell'istruzione per i lavoratori "più abili" e "meno abili", con l'intento di chiarire l'effetto dell'istruzione sulla disuguaglianza dei redditi tra gli individui. Il fatto stilizzato emergente da questo studio è che il rendimento dell'istruzione è maggiore ai quantili più elevati delle distribuzioni condizionali dei redditi (coerentemente con quanto trovato da Buchinsky, 1994 per gli US). La Svezia viene citata come caso tipo; la Grecia come eccezione, poiché le stime per questo paese seguono un andamento esattamente opposto. Martins e Pereira (2004) hanno stimato il rendimento dell'istruzione anche per un campione tratto dalla Banca d'Italia (1995). Le stime ottenute (piuttosto simili alle nostre per quell'anno, data l'identica specificazione e la somiglianza nei criteri di costruzione del campione) seguono un andamento legger-

mente a U , ma non viene commentato dagli autori (v. tav. 1). Dunque, dai risultati emerge che i lavoratori più abili (che ricevono salari orari maggiori, condizionatamente alle loro caratteristiche) sono associati ad un maggiore incremento reddituale legato all'istruzione. Martins e Pereira (2004) propongono tre possibili spiegazioni. La prima fa riferimento al fenomeno della "sovra-educazione", cioè alla diffusione di situazioni in cui lavoratori con elevata istruzione sono impiegati in occupazioni che richiedono scarsa abilità o qualificazione, e che dunque sono remunerati con un basso salario. Così che, più la coda inferiore dei redditi dei lavoratori con elevata abilità è popolata da individui "sovra-istruiti" (lavoratori altamente qualificati con lavori non qualificati), più basso sarà il rendimento dell'istruzione ai quantili inferiori della distribuzione dei redditi.

Una seconda spiegazione riguarda l'abilità e la sua interazione con l'istruzione. Il ruolo delle differenze nell'abilità, dato un certo livello di istruzione, tenderebbe ad amplificarsi in modo crescente in termini di salario, via via che si considerano livelli di istruzione più elevati.

Un'ultima spiegazione si basa sulla presenza di differenze nella qualità scolastica o nei diversi indirizzi di studio (l'equazione di Mincer controlla infatti solo per la quantità dell'istruzione). Potrebbe cioè essere che gli individui che cadono nella coda inferiore della distribuzione dei redditi siano proprio coloro che hanno ricevuto una istruzione scolastica di modesta qualità o che hanno seguito indirizzi scolastici scarsamente remunerati (*ex-post*) sul mercato del lavoro. Tali differenze tendono inoltre a prevalere ai livelli di istruzione medio-alti, che presentano una maggior varietà negli indirizzi formativi, ed eventualmente nel grado qualitativo.

Mwabu e Schultz (1996) utilizzano i quantili di regressione per stimare il rendimento dell'istruzione per etnia in Sud Africa e spiegarne le differenze. Considerando i residui dei quantili di regressione come espressione dell'abilità non osservata, essi interpretano i rendimenti crescenti/decrescenti come indicatori di complementarità/sostituibilità tra questa misura di abilità e l'istruzione nell'incrementare la produttività dei lavoratori. Mwabu e Schultz (1996) trovano che tra la popolazione bianca più istruita

il rendimento dell'istruzione aumenta all'aumentare dei decili del reddito e interpretano tale evidenza in termini di complementarità tra abilità ed istruzione, mentre il rendimento derivante dall'aver frequentato la scuola secondaria presenta una correlazione negativa (istruzione e abilità sostituti, almeno tra i meno abili). Il rendimento per la popolazione nera che possiede un basso livello di istruzione (scuola primaria) diminuisce all'aumentare del decile dei redditi indicando sostituibilità; mentre segue un andamento non lineare (convesso) per quanto riguarda l'istruzione secondaria; e un andamento piuttosto irregolare per l'istruzione più elevata.

4. - Evidenza empirica dai quantili di regressione per l'Italia

4.1 I dati

I dati utilizzati per costruire i campioni sono tratti da Banca d'Italia (1993), (1995), (1998) e (2000).

Le tavole 2, 3, 4 e 5 mostrano le statistiche descrittive delle principali variabili utilizzate nelle stime e delle altre variabili di interesse. Questo per ognuno degli anni considerati, e tenendo distinti donne e uomini.

Nei campioni sono compresi individui tra i 14 ed i 60 anni, esclusivamente lavoratori dipendenti, non appartenenti al settore agricolo. Inoltre, tra gli individui di sesso maschile sono stati considerati solamente i lavoratori a tempo pieno, mentre sono incluse sia le lavoratrici a tempo parziale sia quelle a tempo pieno.

Qualche precisazione preliminare sulle variabili utilizzate:

1) Gli anni di istruzione degli individui sono ricostruiti a partire dal dato sul titolo di studio più elevato conseguito. (Agli individui che non hanno raggiunto alcun titolo di studio è stato assegnato 0). A partire dal 1995 la classificazione relativa ai titoli di studio è stata particolareggiata, prevedendo le voci: diploma professionale (3 anni) e diploma universitario o laurea breve (3 anni). Dal medesimo anno l'indagine chiede all'intervistato anche il tipo di diploma conseguito e il tipo di laurea. Attraverso que-

TAV. 2

STATISTICHE DESCRIPTIVE PER I CAMPIONI FEMMINILI*

	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
	minimo				media				massimo			
età	15	15	16	16	36,96 (10,06)	37,16 (10,13)	38,36 (10,07)	38,74 (10,07)	60	60	60	60
anni istruzione	0	0	0	0	11,22 (3,83)	11,28 (3,66)	11,94 (3,53)	11,83 (3,53)	20	21	21	21
medie inf.	-	-	-	-	0,28 (0,45)	0,26 (0,44)	0,22 (0,42)	0,23 (0,42)	-	-	-	-
medie sup.	-	-	-	-	0,45 (0,49)	0,48 (0,50)	0,52 (0,50)	0,52 (0,50)	-	-	-	-
università	-	-	-	-	0,14 (0,35)	0,14 (0,35)	0,18 (0,39)	0,18 (0,38)	-	-	-	-
esperienza pot.	0	0	0	0	19,74 (11,08)	19,88 (11,06)	20,42 (10,91)	20,91 (10,94)	54	50	53	53
operaio	-	-	-	-	0,33 (0,47)	0,37 (0,48)	0,32 (0,47)	0,33 (0,47)	-	-	-	-
part-time	-	-	-	-	0,11 (0,31)	0,13 (0,34)	0,15 (0,35)	0,16 (0,37)	-	-	-	-
ore settimanali lavorate	4	1	5	5	34,38 (9,06)	34,23 (9,48)	34,06 (9,11)	34,82 (9,43)	70	70	65	100

TAV. 2 segue

STATISTICHE DESCRITTIVE PER I CAMPIONI FEMMINILI*												
	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
	minimo				media				massimo			
	400	600	500	500	18.677,32 (7.849,34)	19.034,68 (8.247,59)	21.482,52 (9.384,24)	22.787,49 (10.100,93)	64.000	63.900	180.000	150.000
reddito annuo netto	0,80	1	1,20	1,74	12,98 (7,33)	13,39 (7,93)	15,13 (11,41)	15,44 (18,06)	102,08	150	264,29	750
salario orario netto	0	0	0	0	6,27 (4,17)	6,41 (4,10)	6,98 (4,07)	7,09 (4,43)	17	17	17	20
istruz. padre	0	0	0	0	5,39 (3,64)	5,60 (3,50)	6,09 (3,59)	6,32 (4,14)	17	17	17	20
istruz. madre	-	-	-	-	0,37 (0,48)	0,42 (0,49)	0,41 (0,49)	0,40 (0,49)	-	-	-	-
padre operaio	-	-	-	-	0,64 (0,48)	0,58 (0,49)	0,58 (0,49)	0,55 (0,50)	-	-	-	-
madre non lav.	-	-	-	-	0,24 (0,43)	0,27 (0,45)	0,21 (0,41)	0,22 (0,41)	-	-	-	-
padre autonomo	-	-	-	-	2.212	2.355	2.128	2.299	-	-	-	-
n. di osservazioni												

* Deviazioni standard in parentesi. Redditi in migliaia di lire.

TAV. 3

STATISTICHE DESCRITTIVE PER I CAMPIONI FEMMINILI*

	I quartile			mediana			III quartile					
	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
età	29	29	30	31	37	37	38	39	45	45	46	47
anni istruzione	8	8	8	8	13	13	13	13	13	13	13	13
medie inf.	-	-	-	-	-	-	-	-	-	-	-	-
medie sup.	-	-	-	-	-	-	-	-	-	-	-	-
università	-	-	-	-	-	-	-	-	-	-	-	-
esperienza pot.	11	11	12	12	18	19	20	21	28	28	28	29
operaio	-	-	-	-	-	-	-	-	-	-	-	-
part-time	-	-	-	-	-	-	-	-	-	-	-	-
ore settimanali lavorate	30	28	30	30	36	36	36	36	40	40	40	40
reddito annuo netto	14.000	14.000	16.000	17.000	19.000	19.500	22.000	23.000	24.000	24.000	26.000	28.000
salario orario netto	8,68	9,25	10,42	10,42	11,28	11,57	13,02	13,20	15,05	15,62	16,67	17,36
istruz. padre	5	5	5	5	5	5	5	5	8	8	8	8
istruz. madre	5	5	5	5	5	5	5	5	8	8	8	8
padre operaio	-	-	-	-	-	-	-	-	-	-	-	-
madre non lav.	-	-	-	-	-	-	-	-	-	-	-	-
padre autonomo	-	-	-	-	-	-	-	-	-	-	-	-

* Deviazioni standard in parentesi. Redditi in migliaia di lire.

TAV. 4

STATISTICHE DESCRIPTIVE PER I CAMPIONI MASCHILI*

	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
	minimo				media				massimo			
età	14	14	15	15	39,29 (10,79)	38,90 (10,83)	39,79 (10,42)	39,73 (10,47)	60	60	60	60
anni istruzione	0	0	0	0	9,88 (3,79)	10,31 (3,78)	10,74 (3,62)	10,73 (3,65)	20	21	21	21
medie inf.	-	-	-	-	0,41 (0,49)	0,36 (0,48)	0,35 (0,47)	0,36 (0,48)	-	-	-	-
medie sup.	-	-	-	-	0,33 (0,47)	0,39 (0,48)	0,44 (0,50)	0,43 (0,49)	-	-	-	-
università	-	-	-	-	0,09 (0,28)	0,10 (0,30)	0,11 (0,31)	0,12 (0,32)	-	-	-	-
esperienza pot.	0	0	0	0	23,41 (11,71)	22,59 (11,65)	23,05 (11,13)	22,99 (11,16)	54	54	54	54
operaio	-	-	-	-	0,50 (0,50)	0,52 (0,50)	0,49 (0,50)	0,50 (0,50)	-	-	-	-
ore settimanali lavorate	5	5	5	5	40,22 (6,86)	40,56 (7,56)	40,55 (7,40)	40,95 (7,67)	70	98	150	100

segue Tav. 4

STATISTICHE DESCRIPTIVE PER I CAMPIONI MASCHILI*

	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
	minimo				media				massimo			
reddito annuo netto	800	700	1.000	800	24.300,06 (11.621,83)	25.109,26 (12.557,43)	28.465,28 (13.201,74)	29.687,22 (15.053,04)	110.000	150.000	180.000	350.000
salario orario netto	0,56	1,67	0,65	0,69	13,42 (7,20)	13,90 (7)	15,51 (8,35)	15,74 (8,08)	166,67	93,75	156,25	191,89
istru. padre	0	0	0	0	5,27 (3,90)	5,50 (3,92)	6,01 (4,02)	6,41 (4,45)	17	17	17	20
istru. madre	0	0	0	0	4,58 (3,43)	4,81 (3,55)	5,41 (3,62)	5,53 (3,97)	17	17	17	20
padre operaio	-	-	-	-	0,43 (0,50)	0,50 (0,50)	0,46 (0,50)	0,44 (0,50)	-	-	-	-
madre non lav.	-	-	-	-	0,71 (0,46)	0,66 (0,47)	0,63 (0,48)	0,60 (0,49)	-	-	-	-
padre autonomo	-	-	-	-	0,24 (0,43)	0,24 (0,43)	0,21 (0,41)	0,19 (0,39)	-	-	-	-
n. di osservazioni					3.618	3.568	2.996	3.265				

* Deviazioni standard in parentesi. Redditi in migliaia di lire.

TAV. 5

STATISTICHE DESCRITTIVE PER I CAMPIONI FEMMINILI*

	1993	1995	1998	2000	1993	1995	1998	2000	1993	1995	1998	2000
	I quartile				mediana				III quartile			
età	31	30	31	31	40	39	40	40	48	48	48	49
anni istruzione	8	8	8	8	8	8	11	11	13	13	13	13
medie inf.	-	-	-	-	-	-	-	-	-	-	-	-
medie sup.	-	-	-	-	-	-	-	-	-	-	-	-
università	-	-	-	-	-	-	-	-	-	-	-	-
esperienza pot.	14	13	14	14	23	22	23	23	32.75	32	32	32
operaio	-	-	-	-	-	-	-	-	-	-	-	-
ore settimanali lavorate	38	38	38	38	40	40	40	40	40	40	42	42
reddito annuo netto	18.000	18.200	21.000	22.000	22.500	23.800	26.000	27.000	28.000	30.000	32.000	33.000
salario orario netto	9,40	9,90	10,94	11,51	11,98	12,50	13,89	14,17	15,36	16,03	17,36	18,06
istru. padre	5	5	5	5	5	5	5	5	8	8	8	8
istru. madre	0	5	5	5	5	5	5	5	5	5	8	8
padre operaio	-	-	-	-	-	-	-	-	-	-	-	-
madre non lav.	-	-	-	-	-	-	-	-	-	-	-	-
padre autonomo	-	-	-	-	-	-	-	-	-	-	-	-

* Deviazioni standard in parentesi. Redditi in migliaia di lire.

st'ultima informazione è stato possibile diversificare il numero di anni di istruzione assegnati ai laureati delle diverse facoltà, tenendo conto della differente durata dei vari corsi di laurea; 2) le variabili relative ai titoli di studio (medie inferiori, medie superiori, studi universitari), allo *status* di operaio dell'individuo e del padre di questi, al tempo parziale, allo *status* di lavoratore autonomo del padre e alla condizione non lavorativa della madre sono variabili dicotomiche. Per gli anni 1995, 1998 e 2000 nella variabile medie superiori sono compresi anche coloro che hanno conseguito un diploma professionale, e nella variabile relativa agli studi universitari sono uniti tutti i soggetti che hanno conseguito un diploma universitario o laurea breve, una laurea regolare ed anche una specializzazione *post-laurea*; 3) l'esperienza potenziale è così calcolata: età meno anni di istruzione meno 6 (considerato l'inizio dell'obbligo di frequenza previsto dal sistema scolastico italiano); 4) le ore lavorate in media a settimana sono comprensive dello straordinario; 5) il reddito annuo da lavoro dipendente è al netto delle imposte e dei contributi; 6) il salario orario netto è definito come: $(\text{reddito netto annuo}) / (\text{mesi lavorati} * \text{ore settimanali lavorate} * 4)$.

Dalle tavole 2, 3, 4 e 5 emerge che per le donne l'età media è tra i 37 e i 39 anni, e vicina ai 40 per gli uomini; più bassa quindi per le prime, ma con una tendenza all'aumento nel corso degli anni. Gli anni di istruzione sono più elevati in media per le donne (11-12 anni) che per gli uomini (10-11 anni), ed aumentano nel tempo per entrambi. Il comportamento delle variabili relative al titolo di studio conseguito rispecchia naturalmente quello della variabile anni di istruzione. In particolare, negli anni tendono ad aumentare le percentuali di individui (donne e uomini) che raggiungono un diploma di scuola superiore o una laurea, a scapito di coloro che raggiungono solamente la licenza media.

Per quanto concerne le caratteristiche lavorative, si può notare come il lavoro a tempo parziale costituisca ancora un fenomeno piuttosto marginale, ma in crescita (dall'11% del 1993 al 16% del 2000). Gli uomini tendono a lavorare (in media) un maggior numero di ore delle donne, pur escludendo le lavoratrici a tempo parziale. Gli uomini hanno redditi annui medi significati-

vamente maggiori rispetto alle donne, ma il divario viene notevolmente a ridursi quando si considerino i salari orari; tale differenza tende inoltre a diminuire leggermente negli anni.

Le informazioni sul substrato familiare sono disponibili a partire dall'indagine del 1993. Guardando ai valori medi, si nota la presenza di una persistenza generazionale sia per il livello di istruzione sia per le scelte occupazionali, soprattutto per le donne: queste infatti, condizionatamente al fatto di essere mediamente più istruite degli uomini nei campioni, tendono ad avere genitori più istruiti, mentre gli uomini hanno in maggior percentuale padri dipendenti come operai e madri non occupate.

4.2 Stime OLS del rendimento dell'istruzione (in anni)

La prima stima effettuata sui campioni descritti è una stima OLS della seguente equazione minceriana (Mincer, 1974):

$$(31) \quad \ln(w_i) = \alpha + \beta S_i + \gamma_1 X_i + \gamma_2 X_i^2 + \varepsilon_i$$

dove $\ln(w)$ è logaritmo del salario orario netto, S sono gli anni di istruzione, X è l'esperienza potenziale e ε è il termine di errore.

Le stime OLS del coefficiente degli anni di istruzione per tutti gli anni considerati (1993, 1995, 1998 e 2000) sono presentate nell'ultima colonna della tavola 6, tenendo distinti donne e uomini. La regressione per i campioni femminili include anche la variabile *dummy partime*, che assume il valore 1 se la lavoratrice è occupata a tempo parziale e 0 altrimenti. Nella tavola 6, in parentesi sotto le stime, sono riportati gli *standard error*, robusti alla presenza di eteroschedasticità (HCSE) (a fronte dei risultati ottenuti effettuando il *test* di Breusch-Pagan, che portano a rifiutare l'ipotesi nulla di omoschedasticità; v. in proposito la tavola 7).

Le stime mostrano i seguenti valori del rendimento dell'istruzione: 8,7% (1993), 7,5% (1995), 6,2% (1998) e 6,1% (2000) per quanto riguarda i campioni femminili; mentre per gli uomini si ha: 6,7% (1993), 6,6% (1995), e circa 6% (1998 e 2000). Si può

TAV. 6

STIME DEL RENDIMENTO DELL'ISTRUZIONE
(anni)*

quantile	0,05	0,10	0,25	0,50	0,75	0,90	0,95	OLS
				donne				
1993	0,0919 (0,0072)	0,0829 (0,0053)	0,0784 (0,0030)	0,0826 (0,0030)	0,0893 (0,0025)	0,0955 (0,0032)	0,0973 (0,0041)	0,0867 (0,0024)
1995	0,0792 (0,0057)	0,0698 (0,0043)	0,0692 (0,0031)	0,0735 (0,0025)	0,0828 (0,0031)	0,0879 (0,0040)	0,0876 (0,0046)	0,0749 (0,0026)
1998	0,0615 (0,0053)	0,0527 (0,0040)	0,0526 (0,0030)	0,0571 (0,0027)	0,0729 (0,0032)	0,0751 (0,0037)	0,0741 (0,0046)	0,0622 (0,0028)
2000	0,0549 (0,0072)	0,0575 (0,0062)	0,0539 (0,0025)	0,0596 (0,0030)	0,0685 (0,0028)	0,0778 (0,0033)	0,0766 (0,0052)	0,0612 (0,0028)
				uomini				
1993	0,0612 (0,0052)	0,0567 (0,0027)	0,0539 (0,0017)	0,0629 (0,0018)	0,0715 (0,0024)	0,0790 (0,0023)	0,0829 (0,0040)	0,0676 (0,00177)
1995	0,0647 (0,0039)	0,0651 (0,0024)	0,0609 (0,0022)	0,0598 (0,0016)	0,0671 (0,0018)	0,0732 (0,0028)	0,0769 (0,0039)	0,0660 (0,00174)
1998	0,0523 (0,0060)	0,0522 (0,0045)	0,0514 (0,0021)	0,0534 (0,0019)	0,0624 (0,0026)	0,0741 (0,0030)	0,0830 (0,0038)	0,0599 (0,00215)
2000	0,0640 (0,0042)	0,0562 (0,0037)	0,0491 (0,0025)	0,0532 (0,0016)	0,0632 (0,0022)	0,0721 (0,0027)	0,0719 (0,0042)	0,0603 (0,00194)

* In parentesi gli standard error (bootstrap). Stime effettuate con R.

TAV. 7

TEST DIAGNOSTICI

anni	1993		1995		1998		2000	
	donne	uomini	donne	uomini	donne	uomini	donne	uomini
R ² (OLS)	0,44695	0,42595	0,35140	0,43385	0,26911	0,35439	0,27804	0,36538
test di B-P*	0,0003	0,00021	0,0000	0,00257	0,00000	0,00004	0,00002	0,07050
(p-value)								
eterosch.**	119.315	163.041	200.300	71.160	185.342	108.299	295.589	121.678
p-value	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
simmetria**	1.789.723	3.828.609	1.981.360	5.021.722	1.739.939	3.404.891	1.689.244	4.510.693
p-value	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
R ² (OLS)	0,45033	0,43232	0,35056	0,43183	0,27216	0,34902	0,27619	0,37004
test di B-P*	0,00227	0,00001	0,00000	0,00261	0,00000	0,00017	0,00004	0,06823
(p-value)								
eterosch.**	160.972	176.297	212.283	88.517	210.779	157.659	338.073	195.671
p-value	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
simmetria**	1.691.700	3.933.010	2.061.016	4.608.752	1.241.258	3.462.240	1.425.513	4.136.447
p-value	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000
n. di osserv.	2.212	3.618	2.365	3.568	2.128	2.996	2.299	3.265

* Test di Breusch-Pagan.

** BUCHINSKY M. (1995).

notare dunque come le stime del rendimento dell'istruzione siano maggiori per le femmine che per i maschi (medesimo risultato è stato ottenuto da Brunello, Comi e Lucifora (2001) su un campione simile di individui). In realtà le intercette, che rappresentano il logaritmo del reddito di un lavoratore a tempo pieno senza istruzione e con esperienza potenziale pari a 0, sono maggiori per gli uomini; inoltre il *gap* tra lavoratrici e lavoratori si riduce via via negli anni considerati. Flabbi (1997), (1999) affrontando in modo specifico (e più esaustivo di quanto non ci si proponga qui) l'aspetto delle differenze di rendimento in base al genere, propone alcune interpretazioni dell'evidenza empirica ottenuta, e nota come i risultati siano legati alle tecniche di stima utilizzate (OLS, OLS con correzione à la Heckman, IV). Una interpretazione dei maggiori rendimenti dell'istruzione delle lavoratrici rispetto ai lavoratori potrebbe derivare comunque, alla luce del modello di Card, dalla presenza di una maggior "distorsione da abilità" dello stimatore OLS per i campioni femminili. Ipotizzando che l'incidenza della distorsione derivante da errori nell'osservazione della variabile esplicativa "anni di istruzione" non sia diversa per le donne e per gli uomini, e che \bar{b} e \bar{r} non differiscano per femmine e maschi, si tratta di capire se è possibile che per le donne esista una maggiore correlazione negativa tra le forme di abilità (a_i e b_i) e i vincoli di liquidità, oltre ad una maggiore correlazione positiva tra a_i e b_i stesse. La correlazione negativa tra abilità e vincoli di liquidità può essere facilmente giustificata ammettendo l'esistenza di 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 riducono i vincoli di liquidità dei figli, nella generazione di questi ultimi, gli individui più abili hanno costi marginali dell'istruzione inferiori. Da quanto ottenuto nelle statistiche descrittive relative ai campioni qui utilizzati sembra plausibile ritenere che per le femmine esista una maggior persistenza generazionale, o comunque un maggior legame con il proprio ambiente familiare; le donne hanno in media una maggiore

istruzione, oltre che genitori anch'essi in media più istruiti. Inoltre minore, rispetto a quella degli uomini, è la percentuale di donne nei campioni che hanno un padre occupato come operaio e una madre non lavoratrice.

5. - Stime del rendimento dell'istruzione (in anni) attraverso i quantili di regressione

Sempre nella tavola 6 sono presentate le stime dei quantili di regressione secondo la medesima specificazione utilizzata per le stime OLS, e sulla base degli stessi campioni. In particolare, la tavola 6 contiene per ogni anno (1993, 1995, 1998 e 2000), e distinguendo tra donne e uomini, le stime del coefficiente degli anni di istruzione per diversi quantili di regressione, $\tau = \{0,05, 0,10, 0,25, 0,50, 0,75, 0,90, 0,95\}$. Sono riportati anche i valori degli *standard error*, calcolati attraverso il metodo Bootstrap non parametrico. I tratti del grafico 2 raffigurano, per il coefficiente dell'istruzione, le stime ai vari quantili, attraverso una linea continua, mentre le linee tratteggiate rappresentano l'intervallo di confidenza al 90%. È poi sovrapposta una linea punteggiata che rappresenta la stima OLS per quel coefficiente, anch'essa con una banda di confidenza al 90%. Nella tavola 7 sono riportati i risultati dei *test* di eteroschedasticità e di simmetria effettuati per le stime dei quantili di regressione (costruiti seguendo Buchinsky, 1995; 1998). I campioni sono i medesimi utilizzati per le stime OLS. I *p-value* ottenuti portano a rifiutare le ipotesi nulle di omoschedasticità e di simmetria per tutti i campioni.

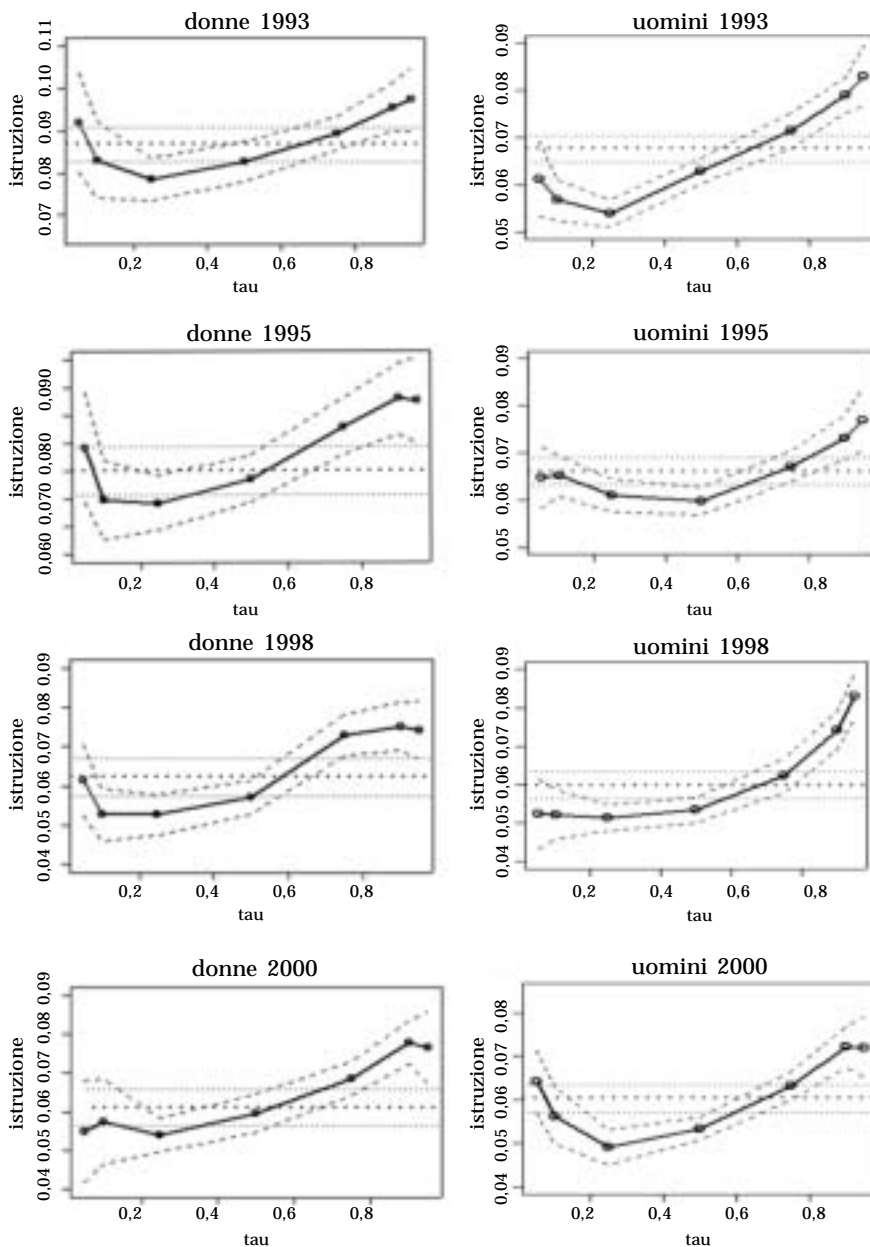
Le stime dei quantili, inizialmente elevate (per $\tau = 0,05$), tendono a calare per gli individui che si collocano al di sotto del primo quartile della distribuzione dei redditi ($\tau = 0,25$), poi invece cominciano a crescere, tendendo più o meno a stabilizzarsi ai quantili più elevati del reddito ($\tau = 0,90$; $\tau = 0,95$). Inoltre, il valore della stima OLS, che rappresenta l'effetto medio, tende ad uguagliare la stima corrispondente a $0,50 < \tau < 0,75$, cioè il valore medio del rendimento marginale dell'istruzione è generalmente maggiore del rendimento di coloro che si trovano al quantile media-

no della distribuzione dei redditi. È confermato un rendimento dell'istruzione maggiore per i campioni femminili che per quelli maschili. Inoltre si può notare una differenza tra i campioni femminili e quelli maschili nel comportamento delle stime ai vari quantili. Le stime relative agli uomini appartenenti alla coda inferiore della distribuzione dei redditi si trovano infatti al di sotto della stima OLS, e fuori dalla sua banda di confidenza, in modo più marcato rispetto a quanto avvenga per le corrispondenti stime effettuate sui campioni femminili; d'altro canto per i primi le stime tendono poi a crescere maggiormente, soprattutto ai quantili più elevati, di quanto facciano le seconde. Per comprendere tali risultati, occorre ricordare che il coefficiente degli anni di istruzione nella regressione di un dato quantile viene interpretato come la variazione (del logaritmo) del reddito dell'individuo necessaria affinché tale individuo rimanga al medesimo quantile della distribuzione dei redditi, dopo aver studiato un anno in più (avendo cioè ricevuto il trattamento). Ad esempio, se si considera il campione delle donne per il 1995, nella seconda riga della tavola 6 leggiamo che il logaritmo del reddito di una lavoratrice che si trova al quantile mediano della distribuzione dei redditi ($\tau = 0,50$) dovrebbe aumentare del 7,35% perché essa si possa collocare nuovamente al quantile mediano della distribuzione dei redditi delle lavoratrici con un anno di istruzione in più. Dalle stime ottenute, si evince dunque che maggiore deve essere l'incremento dei redditi degli individui che si trovano ai quantili bassi e a quelli alti della distribuzione dei redditi, perché essi, con un anno aggiuntivo di istruzione, mantengano il medesimo quantile. Questo significa che la forma delle distribuzioni condizionali cambia; non si ha cioè in questo caso una semplice traslazione della posizione della distribuzione della variabile dipendente, ma anche la sua scala e la sua forma risultano modificate.

È inoltre possibile interpretare i risultati ottenuti in termini di sostituibilità/complementarietà tra i "fattori" istruzione ed abilità. Dal grafico 2, i rendimenti decrescenti per i primi quantili (cioè per gli individui meno abili) testimoniano a favore della presenza di un certo grado di sostituibilità tra istruzione ed abilità. Il rapporto tra i due fattori tende a diventare però di comple-

GRAF. 2

STIME DEL RENDIMENTO DELL'ISTRUZIONE PER DIVERSI QUANTILI DELLA DISTRIBUZIONE CONDIZIONALE DEI REDDITI (anni di istruzione)



mentarietà per i quantili medio-alti di reddito (cioè per gli individui con maggiore abilità), infatti le stime crescono all'aumentare del quantile a partire dai quantili intermedi di reddito.

5.1 *Stime OLS del rendimento dell'istruzione per titolo di studio*

Analizziamo ora delle regressioni *standard* e quantili che contengono come variabili esplicative, in luogo degli anni di istruzione, delle variabili dicotomiche relative al titolo di studio raggiunto dall'individuo. La motivazione che spinge a stimare questo modello si basa sull'osservazione che in alcuni sistemi scolastici — incluso quello italiano — un maggior investimento in istruzione che non conduca all'ottenimento di un titolo di studio potrebbe non garantire rendimenti maggiori sul mercato del lavoro. Per trattare tale aspetto, stimiamo dunque per ogni anno considerato la regressione del logaritmo del salario orario netto sulle variabili dicotomiche “medie inferiori”, “medie superiori”, “università”, controllando per l'esperienza potenziale e il suo quadrato, e mantenendo la variabile *partime* per i campioni femminili. I risultati delle stime OLS sono presentati nell'ultima colonna della tavola 8. Le stime dei coefficienti delle variabili binarie sui titoli di studio vanno interpretate come differenziali rispetto al rendimento base, ottenuto dagli individui senza alcun titolo di studio o che hanno raggiunto solamente la licenza elementare. Ad esempio, per il campione del 1995, un lavoratore maschio in possesso di un diploma di scuola superiore percepisce, in media, il 48% in più di reddito orario di un lavoratore con la medesima esperienza potenziale, ma in possesso della licenza elementare. La configurazione delle stime ottenute utilizzando i livelli di istruzione, invece che gli anni, conferma la presenza di una relazione positiva monotonica tra il rendimento dell'istruzione e il maggior titolo di studio acquisito.

Inoltre, si trova che per il 1993 il rendimento dell'istruzione delle lavoratrici è maggiore di quello dei lavoratori, ma nel 1995 la differenza tende a ridursi, e negli anni successivi le stime dei rendimenti per gli uomini sono superiori rispetto a quelle otte-

TAV. 8

STIME DEL RENDIMENTO DELL'ISTRUZIONE
(titoli di studio)*

quantile	0,05	0,10	0,25	0,50	0,75	0,90	0,95	OLS
	donne							
	medie inf.							
1993	0,2947 (0,0909)	0,3075 (0,0691)	0,1997 (0,0446)	0,18656 (0,0324)	0,2236 (0,0286)	0,2347 (0,0320)	0,2595 (0,0422)	0,2352 (0,0302)
1995	0,3034 (0,1212)	0,1821 (0,1004)	0,1692 (0,0437)	0,14204 (0,0345)	0,1982 (0,0318)	0,1818 (0,0659)	0,1993 (0,0736)	0,1802 (0,0361)
1998	0,1074 (0,0928)	0,1302 (0,0764)	0,1053 (0,0390)	0,16207 (0,0453)	0,1493 (0,0369)	0,2039 (0,0823)	0,2138 (0,1948)	0,1140 (0,0444)
2000	0,3883 (0,11046)	0,3262 (0,0865)	0,1311 (0,0583)	0,10079 (0,0409)	0,0737 (0,0456)	0,0117 (0,0684)	0,0480 (0,1111)	0,1198 (0,0455)
	medie sup.							
1993	0,7535 (0,0749)	0,6807 (0,0734)	0,5577 (0,0495)	0,5352 (0,0328)	0,6229 (0,0329)	0,6987 (0,0349)	0,7730 (0,0518)	0,6265 (0,0316)
1995	0,7319 (0,1238)	0,5414 (0,0971)	0,4855 (0,0422)	0,4669 (0,0358)	0,5587 (0,0301)	0,5905 (0,0602)	0,6316 (0,0734)	0,5193 (0,0367)
1998	0,4373 (0,0785)	0,3930 (0,0755)	0,3621 (0,0393)	0,3861 (0,0474)	0,4284 (0,0400)	0,5072 (0,0812)	0,5502 (0,1950)	0,3799 (0,0439)
2000	0,6927 (0,0995)	0,5866 (0,0892)	0,3663 (0,0623)	0,3377 (0,0452)	0,3674 (0,0475)	0,3721 (0,0683)	0,4273 (0,1078)	0,3919 (0,0456)
	università							
1993	1,1565 (0,0825)	1,0444 (0,0833)	0,9746 (0,0601)	1,00354 (0,0392)	1,1148 (0,0366)	1,1693 (0,0404)	1,1827 (0,0508)	1,0816 (0,0362)
1995	1,0370 (0,1272)	0,8113 (0,1041)	0,8055 (0,0533)	0,87338 (0,0488)	1,0294 (0,0367)	1,0499 (0,0662)	1,0866 (0,0776)	0,9032 (0,0422)
1998	0,5880 (0,0846)	0,5965 (0,0792)	0,6260 (0,0445)	0,70640 (0,0536)	0,8786 (0,0418)	0,9239 (0,0831)	0,9422 (0,2098)	0,7159 (0,0503)
2000	0,8360 (0,1172)	0,8331 (0,0997)	0,6146 (0,0622)	0,66452 (0,0487)	0,7639 (0,0520)	0,7783 (0,0699)	0,8193 (0,1244)	0,7028 (0,0504)

* In parentesi gli standard error bootstrap. Stime condotte con R.

segue Tav. 8

STIME DEL RENDIMENTO DELL'ISTRUZIONE
(titoli di studio)*

quantile	0,05	0,10	0,25	0,50	0,75	0,90	0,95	OLS
	uomini							
	medie inf.							
1993	0,2344 (0,0622)	0,1923 (0,0271)	0,1424 (0,0171)	0,1782 (0,0157)	0,1762 (0,0233)	0,2321 (0,0255)	0,2423 (0,0282)	0,1825 (0,0162)
1995	0,1754 (0,0608)	0,2040 (0,0354)	0,2006 (0,0246)	0,2054 (0,0197)	0,2059 (0,0251)	0,2258 (0,0273)	0,2075 (0,0507)	0,2096 (0,0189)
1998	0,1047 (0,0587)	0,1421 (0,0418)	0,1689 (0,0256)	0,1513 (0,0187)	0,1976 (0,0187)	0,1702 (0,0274)	0,1973 (0,0492)	0,1649 (0,0244)
2000	0,3485 (0,0939)	0,3047 (0,0546)	0,1803 (0,0355)	0,1417 (0,0236)	0,1461 (0,0242)	0,1409 (0,0661)	0,1424 (0,0492)	0,1720 (0,0267)
	medie sup.							
1993	0,5188 (0,0573)	0,4658 (0,0291)	0,4025 (0,0174)	0,4562 (0,0174)	0,5083 (0,0274)	0,5801 (0,0295)	0,6065 (0,0340)	0,4950 (0,0179)
1995	0,5087 (0,0589)	0,4947 (0,0351)	0,4550 (0,0246)	0,4346 (0,0197)	0,4699 (0,0274)	0,5227 (0,0287)	0,5756 (0,0612)	0,4810 (0,0201)
1998	0,3670 (0,0578)	0,3496 (0,0444)	0,3498 (0,0276)	0,3472 (0,0225)	0,4203 (0,0225)	0,5164 (0,0367)	0,6055 (0,0512)	0,4023 (0,0259)
2000	0,6178 (0,0870)	0,4922 (0,0561)	0,3473 (0,0376)	0,3379 (0,0256)	0,4123 (0,0285)	0,4812 (0,0678)	0,4859 (0,0514)	0,4106 (0,0278)
	università							
1993	0,8136 (0,0757)	0,7654 (0,0470)	0,7460 (0,0362)	0,8541 (0,0299)	0,9309 (0,0324)	1,0440 (0,0501)	1,1173 (0,0518)	0,8845 (0,0267)
1995	0,8016 (0,0720)	0,8187 (0,0538)	0,8125 (0,0369)	0,8325 (0,0330)	0,9242 (0,0347)	0,9902 (0,0396)	1,0201 (0,0682)	0,8784 (0,0264)
1998	0,6293 (0,0825)	0,6718 (0,0570)	0,6898 (0,0389)	0,7207 (0,0296)	0,8686 (0,0296)	0,9218 (0,0534)	1,0447 (0,1164)	0,7763 (0,0357)
2000	0,8686 (0,1028)	0,8082 (0,0715)	0,6656 (0,0424)	0,6980 (0,0320)	0,8296 (0,0372)	0,8940 (0,0732)	0,9292 (0,0673)	0,7774 (0,0334)

* In parentesi gli standard error bootstrap. Stime condotte con R.

nute per i campioni femminili. Si nota anche che le stime relative alle lavoratrici mostrano un andamento decrescente nell'arco degli anni considerati.

5.2 Stime del rendimento dell'istruzione per titolo di studio attraverso i quantili di regressione

Le stime dei quantili di regressione ottenute per la specificazione con i titoli di studio sono presentate nella tavola 8 e con l'ausilio dei grafici 3 e 4.

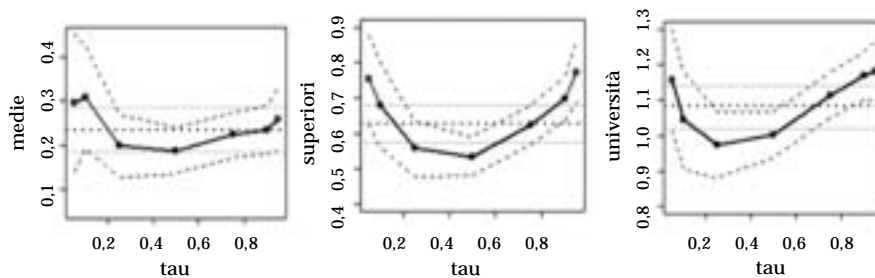
In questa specificazione le stime vanno interpretate (dato un certo quantile) come l'aumento percentuale di reddito che un individuo con al più la licenza elementare dovrebbe avere per mantenersi al medesimo quantile della distribuzione condizionale del reddito, se conseguisse, rispettivamente, la licenza media, il diploma, e la laurea. Dalle tabelle si evince che il rendimento dell'istruzione è generalmente maggiore per i campioni femminili che per i maschili a tutti i livelli di istruzione, anche se la differenza a favore delle femmine scompare nel 1998 e nel 2000, quando in corrispondenza dei quantili più alti della distribuzione dei redditi si rileva un rendimento superiore per gli uomini. Distinguendo per titolo di studio, diversamente da quanto non sia possibile fare per le stime basate sugli anni di istruzione, si può notare un differente comportamento dei rendimenti al variare dei quantili per i diversi livelli di istruzione. In generale, dai grafici risulta evidente come il rendimento derivante dall'aver conseguito la licenza di scuola media inferiore cada, per quasi tutti i quantili e in quasi tutti i campioni, all'interno dell'intervallo di confidenza della corrispondente stima OLS. Cioè, non sembra esserci una particolare variabilità delle stime tra quantili. Nei termini del rapporto (di sostituibilità/complementarietà) tra istruzione ed abilità, potremmo dire che le stime non presentano un andamento specifico al variare del quantile. Forse solo per il 2000 si nota un andamento decrescente al crescere del quantile che porterebbe ad una lettura in chiave di sostituibilità tra i due fattori.

Ma questa configurazione cambia, quando si analizzino le sti-

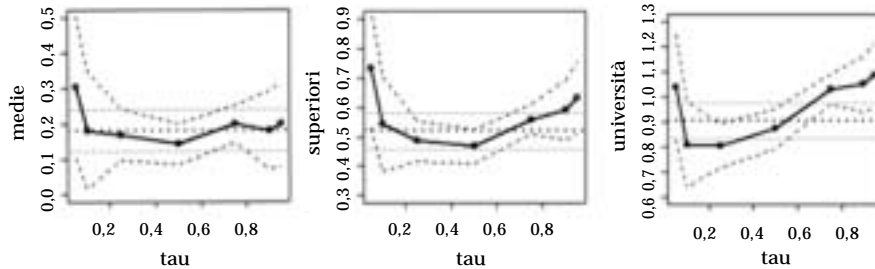
GRAF. 3

STIME DEL RENDIMENTO DELL'ISTRUZIONE PER I CAMPIONI FEMMINILI A DIVERSI QUANTILI DELLA DISTRIBUZIONE CONDIZIONALE DEI REDDITI (titoli di studio)

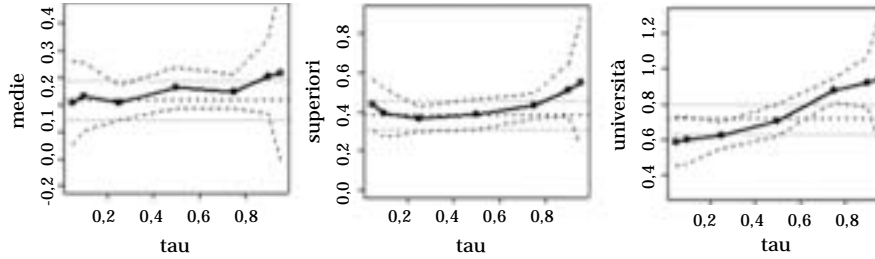
donne 1993



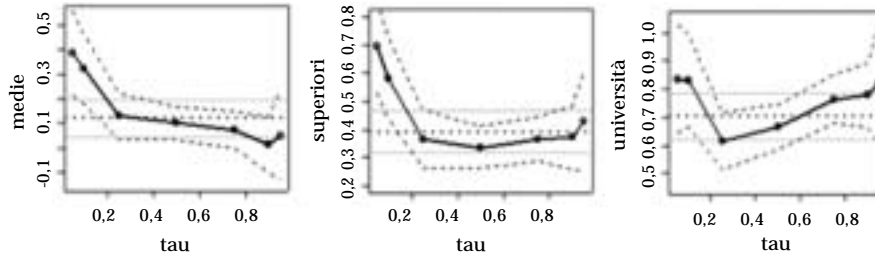
donne 1995



donne 1998

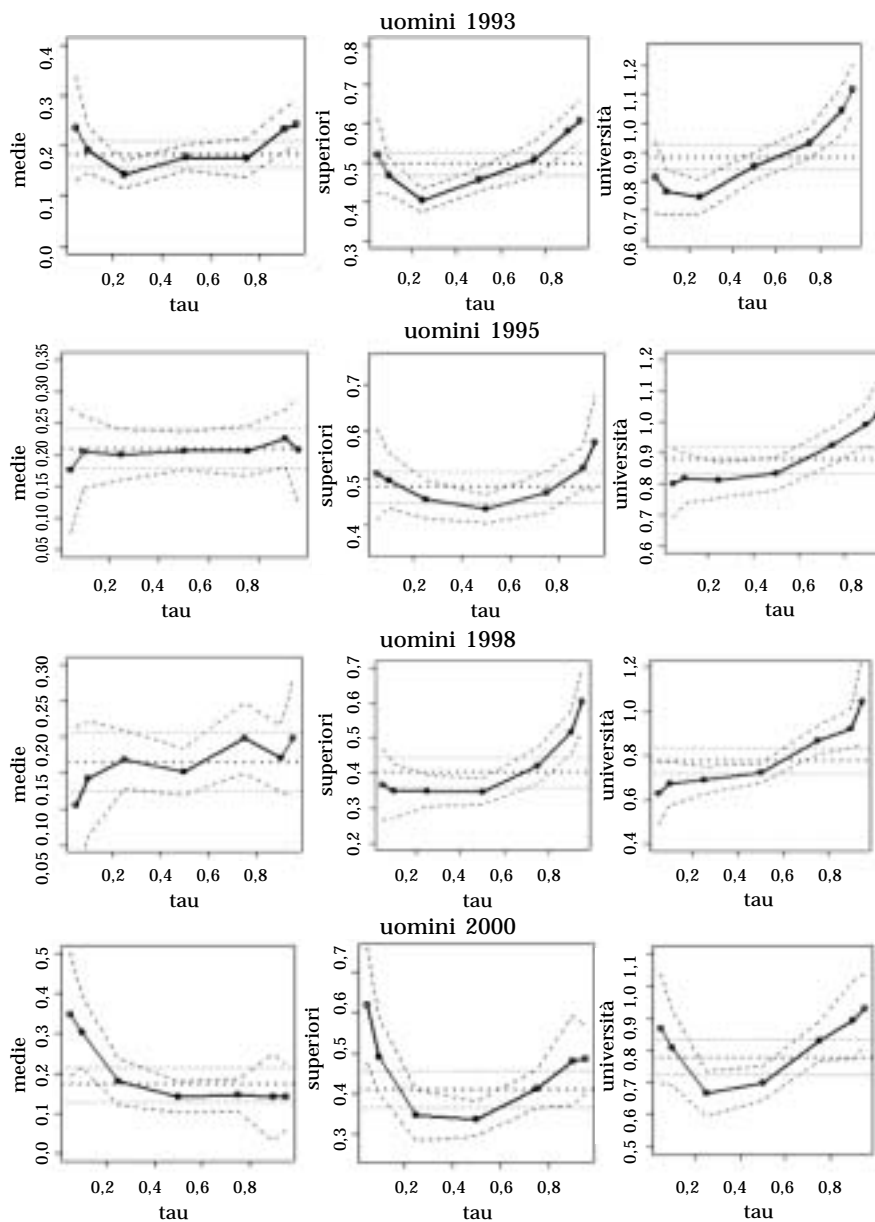


donne 2000



GRAF. 4

STIME DEL RENDIMENTO DELL'ISTRUZIONE PER I CAMPIONI MASCHILI A DIVERSI QUANTILI DELLA DISTRIBUZIONE CONDIZIONALE DEI REDDITI (titoli di studio)



me dei coefficienti delle variabili *medie superiori* e *università*. Per quanto riguarda la prima, infatti, si osserva un andamento convesso ai vari quantili, in modo analogo a quanto visto in precedenza con la variabile anni di istruzione. Cioè, ai quantili bassi e a quelli elevati della distribuzione dei redditi, gli individui necessitano (a parità di condizioni) di un maggior incremento di reddito per mantenere la stessa posizione nella distribuzione dei redditi dei lavoratori diplomati, rispetto agli individui che si trovano ai quantili centrali della medesima distribuzione. Quindi anche qui, cambia la forma della distribuzione condizionale del reddito dei diplomati rispetto agli individui con istruzione elementare. Per quanto riguarda l'istruzione secondaria superiore possiamo dunque interpretare i risultati in termini di sostituibilità tra istruzione ed abilità ai quantili bassi, e di complementarità ai quantili alti.

Anche le stime dei coefficienti della variabile *università* mostrano la presenza di differenze tra quantili, ma in questo caso l'andamento tende ad essere meno convesso, o meglio, meno simmetrico, cioè tende a prevalere un *trend* crescente all'aumentare del quantile a favore dell'interpretazione in termini di complementarità tra i "fattori". Questo si riscontra in particolar modo per quanto riguarda gli anni intermedi (1995 e 1998) e per i campioni maschili. Mentre per le donne l'andamento delle stime presenta una forma maggiormente convessa, con valori molto elevati ai quantili estremi inferiori ($\tau = 0,05$ e $\tau = 0,10$), pari o spesso anche superiori alle stime ottenute ai quantili più alti ($\tau = 0,90$ e $\tau = 0,95$); per gli uomini prevale l'aspetto della crescita del rendimento all'aumentare del quantile di reddito considerato. Questo ad eccezione dell'anno 2000, anno in cui si riscontra un andamento piuttosto convesso anche per il campione maschile, mentre il campione femminile vede addirittura una tendenza decrescente all'aumentare del quantile di reddito considerato (sostituibilità).

Un altro aspetto interessante è rappresentato dalla circostanza che l'andamento piatto delle stime dei coefficienti delle variabili medie inferiori e medie superiori caratterizza particolarmente il campione femminile 1998, l'anno cioè in cui si registra il

maggior calo del rendimento marginale medio dell'istruzione per le donne. Per tale anno una certa differenziazione tra i rendimenti ai diversi quantili di reddito è presentata solo dalle lavoratrici laureate. In particolare, si notano stime più elevate ai quantili alti della distribuzione dei redditi (complementarietà).

6. - Conclusioni

Questo lavoro, si propone di fornire dell'evidenza empirica aggiornata sui rendimenti dell'istruzione in Italia attraverso la sua stima condotta sia con i minimi quadrati ordinari, sia sulla base di un certo numero di quantili di regressione (Koenker e Bassett, 1978), evidenziando come una indagine basata sui quantili possa risultare più informativa rispetto ad una stima OLS *standard* valida per l'individuo medio.

I risultati ottenuti dalle stime OLS (effettuate sui dati tratti dalla Banca d'Italia 1993, 1995, 1998 e 2000) sono molto vicini a quelli precedentemente trovati nella letteratura italiana e confermano la presenza di un rendimento dell'istruzione maggiore per le lavoratrici. L'interpretazione da noi proposta è che tale maggior rendimento potrebbe essere dovuto ad una più forte incidenza dell'*endogeneity bias* e dell'*ability bias* sulle stime relative ai campioni femminili, ammettendo una maggiore persistenza generazionale e una maggiore correlazione dell'abilità e dei vincoli di liquidità delle donne con quelli della loro famiglia di provenienza.

Il risultato più originale è ottenuto però dalle stime dei quantili. In generale, per quanto riguarda il coefficiente degli anni di istruzione, i quantili stimati hanno un andamento a *U*. Maggiore, cioè, deve essere l'incremento dei redditi degli individui che si trovano ai quantili bassi e a quelli alti della distribuzione dei redditi, perché essi, dopo aver studiato un anno in più, si mantengano sul medesimo quantile. In altri termini, non si ha una semplice traslazione della distribuzione del reddito, ma anche la sua scala e la sua forma risultano modificate al variare di τ . Nei termini del rapporto di sostituibilità/complementarietà tra istruzione e abilità, è possibile considerare dunque i due fattori come sostituiti ai

quantili inferiori della distribuzione dei redditi, e come complementi per i quantili medio-alti (cioè per gli individui più abili).

L'aspetto più significativo delle stime relative alla specificazione con i titoli di studio è la presenza di un diverso comportamento del rendimento ai vari quantili per i tre livelli di istruzione considerati. Infatti, mentre i rendimenti derivanti dall'aver conseguito la licenza di scuola media inferiore cadono, per quasi tutti i quantili e in quasi tutti i campioni, all'interno dell'intervallo di confidenza della corrispondente stima OLS, la configurazione cambia, quando si analizzano le stime dei coefficienti delle variabili *medie superiori e università*. Le prime presentano un andamento convesso al variare dei quantili (sostituibilità tra istruzione ed abilità per i meno abili e complementarietà per i più abili), analogamente a quelle relative agli anni di istruzione. Cambia dunque la forma della distribuzione condizionale della variabile dipendente dei diplomati rispetto agli individui con istruzione elementare. Anche le stime dei coefficienti della variabile *università* mostrano la presenza di differenze tra quantili, anche se meno marcate, infatti tende a prevalere un *trend* crescente all'aumentare del quantile (complementarietà).

Si comprende quindi la capacità dei modelli basati sui quantili di regressione di meglio caratterizzare l'impatto dell'istruzione sull'intera distribuzione del reddito, e la loro attrattività per diverse applicazioni economiche. Purtroppo, spesso il trattamento si presenta come endogeno o soggetto a *self-selection*; rendendo i quantili di regressione convenzionali (così come lo stimatore OLS) inadeguati a stimare consistentemente l'effetto del trattamento stesso. Da qui il principale limite e possibilità di sviluppo futuro di questo lavoro, dato che un caso tipico di endogeneità è rappresentato dalla scelta dell'istruzione. Sono state infatti proposte in letteratura delle metodologie di stima che utilizzano le variabili strumentali assieme ai quantili di regressione per trattare tale problema (Honoré e Hu, 2004; Arias, Hallock e Sosa-Escudero, 2001; Chernozhukov e Hansen, 2005). Uno sviluppo futuro di questa indagine sui dati italiani potrebbe essere dunque quello di stimare il rendimento dell'istruzione utilizzando le dette metodologie, nel tentativo di coniugare i vantaggi informativi derivanti da

una indagine basata sui quantili con i vantaggi derivanti dall'uso delle variabili strumentali, che permettono di trattare i problemi legati all'endogenità e all'eterogenità nelle scelte di istruzione.

Un'altra direzione di sviluppo del presente lavoro¹⁴ potrebbe essere quella di utilizzare specificazioni più articolate per l'equazione stimata, che includano ad esempio, da una parte il voto di maturità e di laurea come *proxy* per l'abilità e dall'altra alcune variabili relative al settore o alla qualifica del lavoro dell'individuo. Queste ultime permetterebbero quindi di effettuare un'indagine più mirata, focalizzata su particolari settori o realtà, distinguendo ad esempio tra settore pubblico e privato. Per quanto riguarda l'Italia, tale aspetto è trattato ad esempio da Brunello, Comi e Lucifora, 2001, che utilizzano però stime OLS e *IV*. Per quanto concerne invece la prima proposta, essa si inserisce in quel filone della letteratura che, seguendo Griliches, 1977, utilizza i voti riportati in esami standardizzati (ad esempio il *test IQ*), come variabili in grado di catturare l'abilità non osservata degli individui. Un "contro" di carattere generale di tale approccio consiste nell'implicita definizione che verrebbe così data del concetto di abilità, e nel fatto che in tal modo verrebbe catturato solo un certo "genere di abilità" individuale.

¹⁴ Come osservato da un anonimo *referee*, che ringrazio.

APPENDICE**1. - Dalla localizzazione alla regressione: quantili in ottimizzazione**

Ogni variabile casuale a valori reali Y può essere caratterizzata attraverso la sua funzione di ripartizione:

$$F(y) = \text{Prob}(Y \leq y)$$

Per ogni $0 < \tau < 1$, si definisce il τ -esimo quantile di Y nel modo seguente:

$$(32) \quad \text{Quant}(\tau) = \inf \{y | F(y) \geq \tau\}$$

La mediana, $\text{Quant}(1/2)$, rappresenta il quantile centrale. Come la funzione di ripartizione, anche la funzione quantile fornisce una caratterizzazione completa della variabile casuale Y .

I quantili possono essere formulati come soluzione di un semplice problema di ottimizzazione. Per ogni $0 < \tau < 1$, si definisce la seguente funzione di perdita assoluta simmetrica, lineare a tratti:

$$(33) \quad \rho_\tau(u) = u(\tau - 1) \{u \leq 0\}$$

illustrata nel grafico 1.

Minimizzando il valore atteso di $\rho_\tau(Y - \xi)$ rispetto a ξ , si ottengono le soluzioni $\xi(\tau)$, la più piccola delle quali è $\text{Quant}(\tau)$, definito nell'equazione (2). Formalmente, si può scrivere:

$$\min_{\xi \in \mathbb{R}} E[\rho_\tau(Y - \xi)]$$

o, equivalentemente:

$$\xi^*(\tau) = \text{argmin}_{\xi \in \mathbb{R}} E[\rho_\tau(Y - \xi)]$$

con:

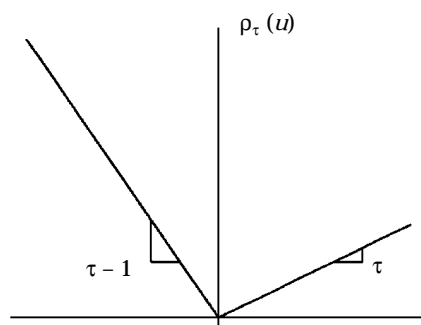
$$(34) \quad Quant(\tau) = \inf \xi^*(\tau)$$

L'equivalente campionario di $Quant(\tau)$, a partire da un campione casuale $\{y_1, \dots, y_n\}$, è detto τ -esimo quantile campionario, e può essere determinato risolvendo:

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^n \rho_{\tau}(y_i - \xi)$$

GRAF. 1

FUNZIONE DI PERDITA ASSOLUTA ASIMMETRICA



ovvero:

$$\hat{\xi}^*(\tau) = \operatorname{argmin}_{\xi \in \mathbb{R}} \sum_{i=1}^n \rho_{\tau}(y_i - \xi)$$

con:

$$(35) \quad \widehat{Quant}(\tau) = \inf \hat{\xi}^*(\tau)$$

in modo analogo a quanto visto sopra.

Utilizzando poi il fatto che:

$$u = -|u| \cdot 1\{u \leq 0\} + |u| \cdot (1 - 1\{u \leq 0\}) = |u| \cdot (1 - 2 \cdot 1\{u \leq 0\})$$

la *check function* può essere utilmente riscritta nel modo seguente:

$$\begin{aligned}
 \rho_{\tau}(u) &= |u| (1 - 2 \cdot 1\{u \leq 0\}) (\tau - 1\{u \leq 0\}) \\
 &= |u| (\tau - 2\tau \cdot 1\{u \leq 0\} + 1\{u \leq 0\}) \\
 &= |u| (\tau \cdot 1\{u \leq 0\} + (1 - \tau) \cdot 1\{u \leq 0\})
 \end{aligned}$$

da cui deriva, ad esempio, che la mediana corrisponde a:

$$(36) \quad \rho_{\frac{1}{2}}(u) = |u| \left(\frac{1}{2} (1 - 1\{u \leq 0\}) + \frac{1}{2} \cdot 1\{u \leq 0\} \right) = \frac{1}{2} |u|$$

mentre per gli altri quantili si pesano diversamente, in modo asimmetrico, gli scarti positivi e quelli negativi.

Per comprendere l'utilità e la validità di tale formulazione, si consideri la funzione obiettivo:

$$(37) \quad Q_{\tau}(\xi) = \sum_{i=1}^n (|y_i - \xi| \cdot \tau \cdot 1\{y_i > \xi\} + |y_i - \xi| \cdot (1 - \tau) \cdot 1\{y_i \leq \xi\})$$

La derivata di questa funzione rispetto a ξ (tranne che per $y_i = \xi$, in corrispondenza delle quali la derivata non esiste) è:

$$(38) \quad \frac{\partial Q_{\tau}(\xi)}{\partial \xi} = \sum_{i=1}^n (-\tau \cdot 1\{y_i > \xi\} + (1 - \tau) \cdot 1\{y_i \leq \xi\})$$

Se la frazione di osservazioni (rispetto al totale) con $y_i \leq \xi$ è pari a τ , questa derivata è nulla:

$$\frac{\partial Q_{\tau}(\xi)}{\partial \xi} = \sum_{i=1}^n [-\tau(1 - 1\{y_i > \xi\}) + (1 - \tau) \cdot 1\{y_i \leq \xi\}] = 0$$

$$\sum_{i=1}^n (-\tau + 1\{y_i \leq \xi\}) = 0$$

$$\frac{1}{n} \sum_{i=1}^n 1\{y_i \leq \xi\} = \tau$$

Quest'ultima è proprio la condizione sugli scarti; essendo infatti ξ_τ il τ -esimo quantile, il numero degli scarti negativi e nulli sarà pari a $m\tau$.

Sarebbe più comune definire i quantili campionari attraverso l'ordinamento delle osservazioni, ma la formulazione in termini di problema di minimizzazione ha il vantaggio di condurre ad una naturale generalizzazione per i quantili condizionali.

Allo stesso modo con cui l'idea di stimare la media non condizionale come:

$$\hat{\mu} = \operatorname{argmin}_{\mu \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2$$

può essere estesa alla stima della funzione lineare media condizionale $E(Y|X=x) = x'\beta$ risolvendo:

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^k} \sum_{i=1}^n (y_i - x'_i\beta)^2$$

la funzione lineare quantile condizionale:

$$(39) \quad Q_\tau(Y|X=x) = x'_i\beta_\tau$$

può essere stimata risolvendo:

$$(40) \quad \hat{\beta}_\tau = \operatorname{argmin}_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_\tau(y_i - x'_i\beta)$$

2. - Quantili di regressione e programmazione lineare

È possibile mostrare che il problema (18) può essere rappresentato anche come un problema di programmazione lineare. Una conseguenza attrattiva di questo fatto è che, sia dal punto di vista teorico sia da quello pratico, la programmazione lineare può essere utilizzata per meglio caratterizzare la soluzione $\hat{\beta}_\tau$ specificata nel problema (40). In particolare, il *Teorema della Dualità* e il *Teorema dell'Equilibrio* della programmazione lineare fanno luce sulle proprietà della soluzione e sulla sua sensibilità agli *outliers* (Buchinsky, 1995).

Si consideri il modello di base:

$$(41) \quad \begin{aligned} y_i &= x_i' \beta_\tau + u_i \\ Q_\tau(y_i|x_i) &= x_i' \beta_\tau \end{aligned}$$

con $Q_\tau(u_{\tau|x_i}) = 0$ (per costruzione), e il problema di minimizzazione (18)

$$(42) \quad \min_{\beta_\tau \in \mathbb{R}^k} \sum_{i=1}^n (|y_i - x_i' \beta_\tau| \cdot \tau \cdot 1\{y_i > x_i' \beta_\tau\} + |y_i - x_i' \beta_\tau| \cdot (1 - \tau) \cdot 1\{y_i \leq x_i' \beta_\tau\})$$

Si noti poi che y_i può essere riscritto come funzione di soli elementi positivi:

$$(43) \quad y_i = \sum_{j=1}^k x_{ij} \beta_{\tau j} + u_{\tau i} = \sum_{j=1}^k x_{ij} (\beta_{\tau j}^1 - \beta_{\tau j}^2) + (u_{\tau i}^+ - u_{\tau i}^-)$$

dove: $\beta_{\tau j}^1 \geq 0$, $\beta_{\tau j}^2 \geq 0$, (con $j = 1, \dots, k$), e $u_{\tau i}^+ \geq 0$, $u_{\tau i}^- \geq 0$, (con $i = 1, \dots, n$).

Segue quindi che il problema (18), scritto nei termini delle componenti della (43), assume la forma seguente:

$$\begin{cases} \min[\tau u^+ + (1 - \tau) u^-] \\ y = X \beta_\tau + u_\tau^+ - u_\tau^- \\ (\beta_\tau, u_\tau^+, u_\tau^-) \in \mathbb{R}^k \times \mathbb{R}_+^{2n} \end{cases}$$

o, equivalentemente, in notazione matriciale, più compatta:

$$\begin{cases} \min_z c'z \\ Az = y \\ z \geq 0 \end{cases}$$

dove $A = (X, -X, I_n, -I_n)$ è una matrice $n \times (2n + 2k)$; $y = (y_1, \dots, y_n)'$ è un vettore $n \times 1$ che raccoglie le osservazioni sulla variabile di-

pendente; $z = (\beta_\tau^1, \beta_\tau^2, u_\tau^+, u_\tau^-)'$ è un vettore $(2k + 2n) \times 1$; $c = (0', 0', \tau\iota', (1 - \tau)\iota)'$; $X = (x_1, \dots, x_n)'$ è una matrice $n \times k$ con la i -esima riga data da x_i' , con $i = 1, \dots, n$; I_n è una matrice identità di dimensione n ; $0'$ è un vettore $k \times 1$ di zeri e ι è un vettore $n \times 1$ di uno. Nel sistema di LP, u_τ^+ , u_τ^- vengono utilizzati in modo da rendere vincoli di uguaglianza i vincoli del sistema, che risulta così espresso in forma *standard*, secondo le comuni regole di trasformazione valide per i sistemi di programmazione lineare. I vincoli, in generale, sarebbero costituiti da disuguaglianze, nella forma canonica del programma: per le osservazioni della variabile dipendente al di sotto dell'iperpiano, infatti, $y_i - x_i'\beta_\tau < 0$; per le osservazioni al di sopra di esso, $y_i - x_i'\beta_\tau > 0$. In altri termini, la variabile $u_{\tau i}$ non è vincolata in segno. Tale problema rappresenta un problema primale di programmazione lineare. Di conseguenza, il problema duale è definito da:

$$\begin{cases} \max_w & w'y \\ & w'A \leq c' \end{cases}$$

Esso equivale approssimativamente alle FOC (20) del problema di minimizzazione Buchinsky (1995).

BIBLIOGRAFIA

- ANTONELLI G., *Risorse umane e redditi da lavoro*, Milano, F. Angeli, 1985.
- ARIAS O. - HALLOCK K.F. - SOSA-ESCUADERO W., «Individual Heterogeneity in the Returns to Schooling: Instrumental Variables Quantile Regression Using Twins Data», *Empirical Economics*, vol. 26, 2001, pp. 7-40.
- ASHENFELTER O. - ROUSE C., «Income, Schooling and Ability: Evidence from a New Sample of Identical Twins», *Quarterly Journal of Economics*, vol. 113, 1998, pp. 253-84.
- BANCA D'ITALIA, *Indagine sui bilanci delle famiglie italiane*, distribuzione elettronica dei microdati, Roma, Banca d'Italia, 1986, 1991, 1993, 1995, 1998, 2000.
- BECKER G.S., *Human Capital and the Personal Distribution of Income*, Ann Arbor (Michigan), University of Michigan Press, 1967.
- BRUNELLO G. - MINIACI R., «The Economic Returns to Schooling for Italian Men. An Evaluation Based on Instrumental Variables», *Labour Economics*, vol. 6, n. 4, 1999, pp. 509-19.
- BRUNELLO G. - COMI S. - LUCIFORA C., «The Returns to Education in Italy: A New Look at the Evidence», in HARMON C. - WALKER I. - WESTERGARD-NIELSEN N. (a cura di), *The Returns to Education in Europe*, Cheltenham (UK), Edward Elgar, 2001.
- BUCHINSKY M., «Changes in the US Wage Structure 1963-1987: Application of Quantile Regression», *Econometrica*, vol. 62, 1994, pp. 405-58.
- —, «Estimating the Asymptotic Covariance Matrix for Quantile Regression Models: A Monte Carlo Study», *Journal of Econometrics*, vol. 68, 1995, pp. 303-38.
- —, «Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research», *Journal of Human Resources*, vol. 33, 1998, pp. 88-126.
- CANNARI L. - D'ALESSIO G., «Il rendimento dell'istruzione: alcuni problemi di stima», Roma, Banca d'Italia, *Temi di Discussione*, n. 253, 1998.
- CANNARI L. - PELLEGRINI G. - SESTITO P., «Redditi da lavoro dipendente: un'analisi in termini di capitale umano», Roma, Banca d'Italia, *Temi di Discussione*, n. 124, 1989.
- CARD D., «Earnings, Schooling, and Ability Revisited», Cambridge (MA), *NBER Working Paper*, n. 4832, 1994.
- —, «Using Geographic Variation in College Proximity to Estimate the Return to Schooling», *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, Toronto, Università di Toronto, 1995, pp. 201-22.
- —, «Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems», *Econometrica*, vol. 69, 2001, pp. 1127-60.
- CARD D. - KRUEGER A.B., «Does School Quality Matter? Return to Education and the Characteristics of Public Schools in the United States», *Journal of Political Economy*, vol. 100, 1992, pp. 1-40.
- CHERNOZHUKOV V. - HANSEN C., «An IV Model of Quantile Treatment Effect», *Econometrica*, vol. 73, n. 1, 2005.
- COLUSSI A., «Una Analisi Cross Section del Tasso di Rendimento dell'Istruzione in Italia», *Politica Economica*, vol. 13, n. 2, 1997.
- FITZENBERGER B. - KURZ C., «New Insights on Earnings Trends Across Skill Groups and Industries in West Germany», *Empirical Economics*, vol. 28, 2003, pp. 479-514.

- FLABBI L., «Investire in istruzione: è meglio per lui o per lei? Stima per genere dei rendimenti dell'istruzione in Italia», Milano, Università degli studi di Milano - Bicocca, *Working Paper*, n. 08, 1997.
- —, «Returns to Schooling in Italy: OLS, IV and Gender Differences», Milano, Università Bocconi, *Working Paper*, n. 1, 1999.
- GRILICHES Z., «Estimating the Returns to Schooling: Some Econometric Problems», *Econometrica*, vol. 45, 1977, pp. 1-22.
- HAUSE J.C., «Earnings Profile: Ability and Schooling», *Journal of Political Economy*, vol. 80, 1972, pp. 108-38.
- HILLIER F.S. - LIEBERMAN G.J., *Introduction to Operations Research*, New York, McGraw-Hill, 1990.
- HOLLAND P.W., «Statistics and Causal Inference», *Journal of the American Statistical Association*, vol. 81, 1986, pp. 945-60.
- HONORÉ B.E. - HU L., «On Performance of Some Robust Instrumental Variables Estimators», *Journal of Business and Economic Statistics*, vol. 22, n. 1, 2004, pp. 30-9.
- ICHINO A., «Il problema della causalità. Una introduzione generale e un esempio», *Manuale di economia del lavoro*, Bologna, Il Mulino, 2001, pp. 457-83.
- ICHINO A. - WINTER-EBMER R., «Lower and Upper Bounds of Returns to Schooling: An Exercise in IV Estimation with Different Instruments», *European Economic Review*, vol. 43, 1999, pp. 889-901.
- KOENKER R., «Short Course on Quantile Regression», *Technical Report*, Urbana-Campaign, University of Illinois, 2003.
- KOENKER R. - BASSETT G., «Regression Quantiles», *Econometrica*, vol. 46, 1978, pp. 33-50.
- LUCIFORA C. - REILLY B., «Wage Discrimination and Female Occupational Intensity», *Labour*, vol. 4, n. 2, 1990, pp. 147-68.
- MARTINS P.S. - PEREIRA P.T., «Does Education Reduce Wage Inequality? Quantile Regression and Evidence From 16 Countries», *Labour Economics*, vol. 11, 2004, pp. 355-71.
- MINCER J., *Schooling, Experience and Earnings*, New York, Columbia University Press, 1974.
- MWABU G. - SCHULTZ T.P., «Education Returns Across Quantiles of the Wage Function: Alternative Explanations for Returns to Education by Race in South Africa», *American Economic Review*, vol. 86, n. 2, 1996.
- PARK J.H., «Returns to Schooling: A Peculiar Deviation from Linearity», Princeton (NJ), Princeton University, *Working Paper*, n. 335, 1994.

