BANCA D'ITALIA

Temi di discussione

del Servizio Studi

L'imputazione dei dati mancanti nelle indagini campionarie: un'applicazione delle tecniche di regressione

di Francesco Trimarchi

Numero 143 - Novembre 1990

BANCA D'ITALIA

Temi di discussione

del Servizio Studi

L'imputazione dei dati mancanti nelle indagini campionarie: un'applicazione delle tecniche di regressione

di Francesco Trimarchi

Numero 143 - Novembre 1990

La serie «Temi di discussione» intende promuovere la circolazione, in versione provvisoria, di lavori prodotti all'interno della Banca d'Italia o presentati da economisti esterni nel corso di seminari presso l'Istituto, al fine di suscitare commenti critici e suggerimenti.

I lavori pubblicati nella serie riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità dell'Istituto.

SOMMARIO

Nell'ambito delle indagini campionarie l'esistenza di mancate risposte costituisce uno dei problemi più rilevanti che si pongono all'analista, soprattutto nel caso in cui vincoli di tempo e/o di costo impediscono l'ampliamento ex-post della numerosità campionaria. In queste circostanze, molto frequenti nella pratica statistica, l'unica via per contenere la perdita di precisione degli stimatori dei parametri della popolazione che si determina di conseguenza consiste nell'utilizzare le informazioni campionarie disponibili per imputare (stimare puntualmente) i dati mancanti. Fra i metodi sviluppati a questo scopo, quelli di regressione costituiscono sotto alcune condizioni uno strumento potente e, in genere, preferibile ad altri metodi di imputazione. L'applicazione del metodo di regressione, discusso nei suoi aspetti teorici e metodologici nel cap. 2, offre risultati soddisfacenti (cap. 3) per l'imputazione dei dati mancanti e per la stima delle medie nell'indagine sugli investimenti, condotta annualmente dalla Banca d'Italia presso un campione di imprese manifatturiere.

INDICE

1- Introduzione (*)

Nell'ambito della teoria dei campioni le tecniche di imputazione sono state sviluppate principalmente per limitare la perdita di precisione degli stimatori dei parametri della popolazione, che si determina in presenza di missing data (in seguito M.D.) 1/.

Nel caso tipico della media campionaria $\bar{x}$ - stimatore di M.L. della media della popolazione qualora il carattere X sia distribuito ad esempio normalmente - tale perdita è facilmente quantificabile e dipende esclusivamente dalla frequenza relativa delle mancate risposte.

Indicando con n la numerosità campionaria ex-ante (numero dei potenziali rispondenti stabilito in base al piano di campionamento) e con $p=n-m$ $(0\leq m\leq n)$ la numerosità campionaria ex-post (numero effettivo dei rispondenti), l'efficienza (E) dello stimatore "media campionaria" in presenza di M.D. può essere misurata, infatti, assumendo come paradigma la varianza dell'analogo stimatore in assenza di M.D..

Nel campionamento casuale con ripetizione, supponendo che la varianza del carattere nella popolazione (σ^2) sia nota e che la probabilità di mancata risposta non dipenda da X, avremo:

$$E = (\sigma^2/p)/(\sigma^2/n) = n/p = 1/\tau$$

in cui $\tau=p/n$ è la frequenza relativa dei rispondenti.

Si osservi che, qualora il meccanismo generatore delle mancate risposte dipenda dal livello di X, lo stimatore "media

campionaria" in presenza di M.D., oltre ad essere meno preciso, sarà in generale anche distorto. Il rischio di distorsione si può presentare, ad esempio, nelle indagini sui redditi e la ricchezza, in cui è ragionevole ipotizzare un legame diretto fra la reticenza degli individui nel rispondere ed il livello assunto da entrambe queste variabili.

Qualora non sia possibile, per ragioni di tempo e/o di costo, rilevare unità statistiche aggiuntive estratte dalla medesima popolazione in modo tale da riportare, ex-post, il numero effettivo dei rispondenti sino al livello ex-ante prefissato, non resta altro che cercare di stimare puntualmente il valore dei M.D. utilizzando le altre informazioni campionarie disponibili.

Ciò consente di limitare la perdita di precisione dello stimatore della media nonchè degli altri parametri della distribuzione del carattere X. Ovviamente, la perdita sarà tanto più piccola quanto più, a sua volta, lo stimatore adottato per imputare i M.D. sarà preciso. L'imputazione, inoltre, fornisce una stima puntuale dei valori assunti dalle variabili nelle unità statistiche che non hanno fornito la risposta e, quindi, può essere utilizzata per migliorare l'analisi dei dati anche a livello individuale.

Nella letteratura statistica sono state proposte svariate tecniche di imputazione, basate su opportune medie dei non-M.D. 2/. Fra queste si possono distinguere le tecniche univariate da quelle multivariate; la distinzione concerne essenzialmente il tipo di media: non condizionata nel primo caso, condizionata nel

secondo.

Un esempio del primo tipo è costituito dall'imputazione mediante la media aritmetica della variabile che presenta M.D., calcolata sulla base delle risposte fornite dall'intero subcampione dei rispondenti.

Un esempio del secondo tipo è costituito dall'imputazione di un singolo valore mancante di una determinata variabile con la media aritmetica assunta da tale variabile nello strato a cui appartiene l'unità statistica che non ha fornito la risposta; in questo caso la media è condizionata alle modalità assunte in ciascuna unità statistica dalle variabili classificatorie utilizzate nel piano di campionamento.

A loro volta, le tecniche multivariate differiscono in base al modello sottostante la scelta della media condizionata a cui eguagliare i valori incogniti di alcune variabili. Fra le tecniche multivariate, quelle di regressione si configurano come uno strumento generale e potente, poichè consentono di estrarre dal campione la massima quantità di informazione e, sotto determinate ipotesi, di ricavare dai dati completati stime corrette e di minima varianza dei parametri della popolazione.

La finalità di questo saggio consiste nel verificare empiricamente le proprietà delle tecniche di regressione e di confrontarle con quelle di altre tecniche.

I dati utilizzati provengono dall'indagine campionaria sugli investimenti fissi lordi delle imprese manifatturiere, condotta

annualmente dalla Banca d'Italia presso un campione di oltre 1.000 imprese con più di 50 addetti. In tale indagine vengono richieste, mediante un questionario, informazioni relative agli investimenti, al fatturato, all'occupazione, nonchè all'utilizzo e alle variazioni della capacità produttiva negli ultimi due anni. Oltre ai dati di consuntivo, vengono richieste anche le previsioni relative, fra l'altro, al flusso degli investimenti e al numero di occupati a fine anno; le previsioni relative a queste variabili, quindi, non sono ricavate a posteriori utilizzando metodi statistici ma corrispondono ai programmi formulati dalle imprese stesse e riassunti nel budget aziendale.

Nel par. 2.1 si esplicitano brevemente gli assunti che stanno alla base delle tecniche di regressione. Nei parr. 2.2 e 2.3 si esaminano gli stimatori di regressione dei principali parametri (con particolare riferimento alla media), che scaturiscono dall'ipotesi di: i) normalità della distribuzione congiunta dei caratteri nella popolazione; ii) casualità del meccanismo generatore dei M.D.. Nel cap. 3, infine, si espongono i risultati di un esperimento compiuto allo scopo di verificare le performances relative degli stimatori di regressione della media nell'ambito della citata indagine condotta dalla Banca d'Italia.

2- Le tecniche di regressione

2.1- Definizione e fondamenti logici

In generale, date due variabili casuali X ed Y, uno stimatore di regressione dei parametri della popolazione consiste in una funzione T dei valori campionari effettivi di X (Y), se X (Y) è misurata, e stimati puntualmente sulla base della regressione di X su Y (di Y su X), se X (Y) è missing.

Una volta stimati i parametri della regressione, l'imputazione dei M.D. è possibile ove si disponga di una misura della variabile di destra nelle osservazioni cui quella di sinistra è missing.

Gran parte della letteratura sull'argomento ha preso le mosse dall'approccio della regressione multivariata, assumendo l'ipotesi di normalità della distribuzione congiunta dei caratteri nella popolazione. Questo approccio non impone restrizioni a priori (o meglio "teoriche") circa la specificazione dei modelli da utilizzare per l'imputazione, salvo la linearità del legame fra variabile dipendente e regressori, che discende necessariamente dall'ipotesi di normalità 3/. A parte questo, ciascuna variabile può assumere indifferentemente il ruolo di variabile dipendente o di regressore a seconda del contesto; le regressioni stimate posseggono, quindi, un significato meramente statistico-probabilistico che prescinde dal legame funzionale - la cui specificazione discende dalla "teoria" - fra variabili esplicative (cause) e variabili esplicate (effetti) 4/.

D'altra parte, l'ipotesi di normalità - se verificata - assicura a priori l'esistenza di una soluzione statisticamente corretta ed efficiente al problema dell'imputazione dei M.D. e della stima dei parametri della popolazione. Ne consegue che, mentre anche l'utilizzo di regressioni non suggerite dalla "teoria" è pienamente legittimo, la violazione dell'ipotesi di normalità può essere grave soprattutto nei piccoli campioni, non essendo più assicurata, in questa circostanza, la linearità delle regressioni nonchè la correttezza e l'efficienza degli stimatori dei parametri.

Altrettanto grave può essere la violazione di una seconda ipotesi che sta alla base delle tecniche di regressione, ovvero l'ipotesi di "ignorabilità" dei meccanismi generatori delle mancate risposte.

Euristicamente possiamo dire, seguendo le indicazioni di Little 5/, che "...if these mechanisms are unrelated to the values of variables measured in the survey, then the response mechanism can be ignored and the observed values treated as a random subsample of the hypotetical complex sample without nonresponse...If the response mechanism is related to the values of the variables under study, then it is nonignorable, in the sense that wich do not take this into account are subject to bias.".

Qualora entrambe le ipotesi risultano violate, i rimedi sono piuttosto limitati ed empirici.

Nel caso in cui la forma della distribuzione si discosta

dalla normale, si può ricercare una opportuna trasformazione delle variabili originali (ad esempio Box-Cox) tale per cui la distribuzione congiunta delle variabili trasformate sia normale o almeno simmetrica e unimodale (confidando nella "robustezza" e nelle proprietà asintotiche degli stimatori). La ricerca della trasformata, tuttavia, può comportare un aumento anche notevole dei costi di imputazione non necessariamente giustificato dal guadagno di precisione e dalla riduzione del bias degli stimatori 6/.

Riguardo alla violazione dell'ipotesi di "ignorabilità", poi, si deve osservare che il legame fra la probabilità di mancata risposta ed il livello della variabile da imputare non è indagabile a livello individuale.

In entrambi i casi, le proprietà degli stimatori di regressione che saranno illustrati nei parr. 2.2 e 2.3 possono essere studiate solo sperimentalmente 7/ ed essere comparate a quelle di altri stimatori. E' però evidente che i risultati delle simulazioni dipendono dai dati utilizzati e non sono generalmente estensibili a popolazioni differenti rispetto a quelle prese a riferimento.

2.2- L'imputazione nel caso bivariato

E noto che, se la distribuzione congiunta di due variabili casuali X ed Y è normale, allora: i) sia le distribuzioni condizionate che quelle marginali saranno ancora normali; ii) le regressioni di Y su X e di X su Y saranno necessariamente lineari, ovvero:

$$E(Y|X=x_i)=\mu_y + \beta\cdot(x_i - \mu_x) \qquad [2.2.1]$$

$$E(X|Y=y_i)=\mu_x + \delta\cdot(y_i - \mu_y) \qquad [2.2.2]$$

in cui $\beta=\sigma_{xy}/\sigma_x^2$, $\delta=\sigma_{xy}/\sigma_y^2$ ed i=1,2,...,n.

Nel campionamento casuale senza M.D. le medie, le varianze e la covarianza campionaria sono stimatori di M.L. dei rispettivi parametri nella popolazione (μ_x , μ_y , σ_x^2 , σ_y^2 , σ_{xy}). Il rapporto fra la covarianza campionaria e la varianza campionaria della variabile di destra è a sua volta stimatore di M.L. dei coefficienti di regressione.

Nella circostanza in cui, ferme le altre ipotesi, una o entrambe le variabili presentano M.D., la soluzione di M.L. non é ovvia nè immediata e dipenderà dal <u>pattern</u> dei M.D. nonché dalla natura dell'eventuale legame fra il livello assunto dalle variabili e la probabilità di non risposta.

Lord e Anderson <u>8</u>/ - nell'ipotesi che: i) solo una delle due variabili, poniamo Y, abbia M.D; ii) il meccanismo generatore delle mancate risposte sia ignorabile - hanno ricavato gli

stimatori di M.L. dei parametri della popolazione:

$$\hat{\mu}_x = (1/n)\cdot\left(\sum_{i=1}^{n} x_i\right) \qquad \hat{\sigma}^2_x = (1/n)\cdot\left[\sum_{i=1}^{n} (x_i - \hat{\mu}_x)^2\right]$$

$$\hat{\mu}_y = \bar{y}_c + \hat{\beta}_{yx}\cdot(\bar{\mu}_x - \bar{x}_c) \qquad \hat{\sigma}^2_y = \sigma^2_{y,c} + \hat{\beta}^2_{yx}\cdot(\sigma^2_x - \sigma^2_{x,c})$$

$$\hat{\sigma}_{yx} = \sigma_{yx,c}\cdot(\sigma^2_x/\sigma^2_{x,c}) \qquad \hat{\beta}_{yx} = \sigma_{yx,c}/\sigma^2_{x,c}$$

in cui n è la numerosità campionaria, n_c il numero di osservazioni complete ($n_c \leq n$); $\bar{x}_c$, $\bar{y}_c$, $\sigma^2_{x,c}$, $\sigma^2_{y,c}$ e $\sigma_{xy,c}$ sono le medie, le varianze e la covarianza campionarie calcolate sulla base delle n_c osservazioni complete. Indicando con $m_y = n - n_c$ il numero dei M.D. della variabile Y si dimostra che:

$$\hat{\mu}_y = (1/n)\cdot\left(\sum_{i=1}^{n_c} y_i + \sum_{i=1}^{m_y} \hat{y}_i\right) \qquad [2.2.3]$$

in cui $\hat{y}_i = \bar{y}_c + \hat{\beta}_{yx}\cdot(x_i - \bar{x}_c)$;

Si osservi che la [2.2.3] equivale a:

$$\hat{\mu}_y = \tau\cdot\left[(1/n_c)\cdot\left(\sum_{i=1}^{n_c} y_i\right)\right] + (1-\tau)\cdot\left[(1/m_y)\cdot\left(\sum_{i=1}^{m_y} \hat{y}_i\right)\right] \qquad [2.2.4]$$

in cui $\tau = n_c/n$ è il tasso di risposta.

Lo stimatore $\hat{\mu}_y$ coincide, dunque, con la media aritmetica dei

valori campionari effettivi, se Y è misurata, e imputati mediante la regressione, se Y è missing (vedi la [2.2.3]), ovvero con la media ponderata delle medie dei valori effettivi ed imputati, in cui i pesi sono costituiti rispettivamente dai tassi di risposta e di non risposta (vedi la [2.2.4]).

Lo stimatore di regressione $\hat{\mu}_y$ può essere confrontato con il seguente stimatore, che rappresenta il "prototipo" degli stimatori univariati:

$$\hat{\mu}'_y = (1/n) \cdot \left(\sum_{i=1}^{n_c} y_i + \sum_{i=1}^{m_y} \hat{y}'_i \right) \qquad [2.2.5]$$

con $\hat{y}'_i = \bar{y}_c$.

Quest'ultimo stimatore coincide con la media dei valori campionari effettivi, se Y è misurata, e imputati mediante la media dei non-M.D., se Y è missing. Evidentemente $\hat{\mu}'_y = \bar{y}_c$; pertanto:

$$\hat{\mu}_y = \hat{\mu}'_y + \hat{\beta}_{yx} \cdot (\hat{\mu}_x - \bar{x}_c) \qquad [2.2.6]$$

Lo stimatore $\hat{\mu}_y$, quindi, aggiunge all'informazione contenuta nelle n_c osservazioni complete - la sola utilizzata da $\hat{\mu}'_y$ - l'ulteriore informazione che scaturisce dalla regressione lineare di Y su X, sintetizzata dal termine $\hat{\beta}_{yx} \cdot (\hat{\mu}_x - \bar{x}_c)$.

La distribuzione asintotica non condizionata dei due stimatori è la seguente 9/:

$$\sqrt{n}(\hat{\mu}_y - \mu_y) \sim N\{0,\ \sigma^2_y/\tau - \beta^2_{yx} \cdot \sigma^2_x \cdot (1-\tau)/\tau\} \qquad [2.2.7]$$

$$\sqrt{n}(\mu'_y - \mu_y) \sim N\{0,\ \sigma^2_y/\tau\} \qquad [2.2.8]$$

in cui $\tau = \text{plim}(n_c/n) \leq 1$ é la probabilità in senso frequentista dell'evento "risposta rispetto ad Y".

Gli stimatori - entrambi corretti nell'ipotesi di "ignorabilità" - coincideranno asintoticamente, oltre che nel caso banale in cui $\tau=1$ (assenza di M.D.), anche quando $\beta_{yx}=0$ (X ed Y sono incorrelati perciò, data l'ipotesi di normalità, anche indipendenti). Se la regressione non fornisce alcuna informazione supplementare (ovvero se COV(X,Y)=0), l'approccio univariato e quello di regressione coincidono e conducono a stime della media corrette e di eguale precisione. A parte questo caso limite, l'efficienza relativa dello stimatore di regressione può essere studiata esprimendo la sua varianza in funzione del coefficiente di correlazione lineare fra X ed Y:

$$\text{VAR}(\hat{\mu}_y) = \sigma^2_y/\tau - \beta^2_{yx} \cdot \sigma^2_x \cdot (1-\tau)/\tau = \sigma^2_y \cdot [1-(1-\tau)r^2_{yx}]/\tau \qquad [2.2.9]$$

Possiamo scrivere quindi:

$$E = \text{VAR}(\hat{\mu}_y)/\text{VAR}(\mu'_y) = 1/[1-(1-\tau)r^2_{yx}] \qquad [2.2.10]$$

E' evidente che $\hat{\mu}_y$ é sempre più preciso di μ'_y ; la sua efficienza relativa dipenderà, data la probabilità di risposta,

dalla "perfezione" del legame lineare fra le due variabili e sarà massima nel caso in cui tutta la variabilità di Y risulta spiegata dalla regressione.

Mentre lo stimatore di regressione della media presenta una formulazione intuitiva, l'analogo stimatore della varianza ($\hat{\sigma}^2_y$) presenta alcune caratteristiche peculiari che meritano un commento. Si può dimostrare che:

$$\hat{\sigma}^2_y = \hat{\sigma}^2_{y,b} + \hat{c}_y \qquad [2.2.11]$$

in cui:

$$\hat{\sigma}^2_{y,b} = (n_c/n)\cdot(\hat{\sigma}^2_{y,c} - \hat{\beta}^2_{yx}\cdot\hat{\sigma}^2_{x,c}) + \hat{\beta}^2_{yx}\cdot\hat{\sigma}^2_x = (1/n)\cdot\left[\sum_{i=1}^{n_c}(y_i - \hat{\mu}_y)^2 + \sum_{i=1}^{m_y}(y_i - \hat{\mu}_y)^2\right]$$

$$\hat{c}_y = (m_y/n)\cdot(\hat{\sigma}^2_{y,c} - \hat{\beta}^2_{yx}\cdot\hat{\sigma}^2_{x,c}) = (m_y/n)\cdot\hat{\sigma}^2_{y,c}\cdot(1 - r^2_{yx,c})$$

Il termine $\hat{\sigma}^2_{y,b}$ altro non è che la varianza campionaria calcolata sulla base dei valori effettivi, se Y é misurata, e imputati mediante la regressione, se Y é missing. Evidentemente $\hat{\sigma}^2_{y,b}$ é uno stimatore distorto di σ^2_y poichè aggiunge alla stima della varianza spiegata dalla regressione ($\hat{\beta}^2_{yx}\cdot\hat{\sigma}^2_x$) solamente la componente misurabile della varianza residua, ovvero il termine:

$$(n_c/n)\cdot(\hat{\sigma}^2_{y,c} - \hat{\beta}^2_{yx}\cdot\hat{\sigma}^2_{x,c}) = (n_c/n)\cdot\hat{\sigma}^2_{y,c}\cdot(1 - r^2_{yx,c})$$

L'aggiunta del fattore di correzione $\hat{c}_y$ (stima della componente residua non osservata) fa sì che che lo stimatore $\hat{\sigma}^2_y$

risulti asintoticamente corretto. Si osservi, peraltro, che la distorsione di $\hat{\sigma}^2_y$ tende ad annullarsi all'aumentare del tasso di risposta (τ) e del coefficiente di correlazione ($r^2_{yx,c}$). Nel caso bivariato sinora illustrato, peraltro, la correzione è implicita nella definizione stessa dello stimatore $\hat{\sigma}^2_y$, che scaturisce direttamente, come nel caso dello stimatore della media, dalla massimizzazione della funzione di verosimiglianza opportunamente fattorizzata. Si osservi, invece, che nessuna correzione è implicita nella definizione dello stimatore della covarianza; infatti:

$$\hat{\sigma}_{yx} = \hat{\sigma}_{yx,b} \qquad [2.2.12]$$

in cui $\hat{\sigma}_{yx,b}$ è la covarianza campionaria calcolata sulla base dei valori effettivi, se Y é misurata, e imputati mediante la regressione, se Y é missing.

Quest'ultimo risultato vale esclusivamente nel caso bivariato. Come vedremo in seguito, se le variabili sono più di due occorrerà correggere anche la covarianza ogni qualvolta una coppia di variabili è congiuntamente non misurata nell'ambito di una determinata unità statistica.

2.3- L'estensione al caso multivariato

Considerazioni analoghe a quelle sviluppate nel paragrafo precedente sono valide anche nel caso multivariato, a condizione che il pattern dei M.D. sia monotono. Questa circostanza si realizza allorchè le variabili e le osservazioni possono essere ordinate gerarchicamente in modo che, per ogni osservazione, se la j-esima variabile è misurata siano misurate anche le j-1 variabili antecedenti. Ad esempio:

Fig 1

OBS.	Y_1	Y_2	Y_3
1	p	p	p
2	p	p	p
3	p	p	p
4	p	p	m
5	p	p	m
6	p	p	m
7	p	m	m
8	p	m	m

in cui p ed m indicano, rispettivamente, la presenza o l'assenza di risposta.

Si osservi che il caso bivariato illustrato nel par. 2.2 è un caso speciale di pattern monotono.

Un'altro caso speciale di pattern monotono ricorre allorchè, date k variabili, le prime p variabili non hanno M.D., mentre le rimanenti k-p variabili hanno M.D. in corrispondenza delle medesime osservazioni. Ad esempio:

Fig. 2

OBS.	Y_1	Y_2	Y_3	Y_4
1	p	p	p	p
2	p	p	p	p
3	p	p	p	p
5	p	p	m	m
6	p	p	m	m

La caratteristica comune degli stimatori di M.L. dei parametri di una popolazione normale nel caso di M.D. con pattern monotono è che la loro definizione scaturisce direttamente dalla massimizzazione della funzione di verosimiglianza opportunamente fattorizzata. In altri termini il sistema di equazioni normali è risolvibile algebricamente e fornisce l'espressione esplicita degli stimatori stessi.

In generale, date k variabili, le stime delle k medie nel caso di un pattern monotono in cui $m_{y_j} < m_{y_h}$ se $j<h$ e $m_{y_1} = 0$ si otterranno: a) stimando le k-1 regressioni di ciascuna generica variabile Y_j sulle j-1 variabili "antecedenti", sulla base delle n osservazioni in cui Y_j è misurata; b) imputando sequenzialmente i M.D. di ciascuna variabile y_j utilizzando la corrispondente regressione stimata; c) calcolando le medie di ciascuna variabile sulla base dei dati così completati. Nella Fig. 3 è sintetizzato il procedimento appena descritto con riferimento al caso trivariato (k=3) rappresentato a scopo esemplificativo nella Fig. 1 in cui:

$$n_{y_1} = 8,\quad m_{y_1} = 0,\quad n_{y_2} = 6,\quad m_{y_2} = 2,\quad n_{y_3} = 3,\quad m_{y_3} = 5,\quad n_c = 3$$

Fig. 3

1- STIMA O.L.S. DELLE REGRESSIONI

$$\text{di } Y_2 \text{ su } Y_1 : \quad \hat{y}_{i,2} = \hat{\beta}_{0,2} + \hat{\beta}_{1,2} \cdot y_{i,1} \qquad (i=1,2,\ldots,n_{y_2})$$

$$\text{di } Y_3 \text{ su } Y_1 \text{ e } Y_2 : \quad \hat{y}_{i,3} = \hat{\beta}_{0,3} + \hat{\beta}_{1,3} \cdot y_{i,1} + \hat{\beta}_{2,3} \cdot y_{i,2} \qquad (i=1,2,\ldots,n_{y_3})$$

2- IMPUTAZIONE DEI M.D. E COMPLETAMENTO DEL CAMPIONE

$$Y_2 : \quad y^*_{i,2} = \begin{cases} y_{i,2} & (i=1,2,\ldots,n_{y_2}) \\ \hat{\beta}_{0,2} + \hat{\beta}_{1,2} \cdot y_{i,1} & (i=1,2,\ldots,m_{y_2}) \end{cases}$$

$$Y_3 : \quad y^*_{i,3} = \begin{cases} y_{i,3} & (i=1,2,\ldots,n_{y_3}) \\ \hat{\beta}_{0,3} + \hat{\beta}_{1,3} \cdot y_{i,1} + \hat{\beta}_{2,3} \cdot y^*_{i,2} & (i=1,2,\ldots,m_{y_3}) \end{cases}$$

3- STIMA DELLE MEDIE

$$\hat{\mu}_{y_1} = (1/n) \cdot \left(\sum_{i=1}^{n} y_{i,1} \right), \quad \hat{\mu}_{y_2} = (1/n) \cdot \left(\sum_{i=1}^{n} y^*_{i,2} \right), \quad \hat{\mu}_{y_3} = (1/n) \cdot \left(\sum_{i=1}^{n} y^*_{i,3} \right).$$

Qualora il pattern non sia monotono la soluzione di M.L. non è esplicita. Dempster et alii hanno proposto un algoritmo numerico (E.M.-algorithm) che, attraverso iterazioni successive, conduce all'insieme dei valori che massimizzano la funzione di verosimiglianza, ovviamente a meno di uno scarto non controllabile 10/.

Una alternativa di second best ad un algoritmo complesso e approssimato come l'E.M. può essere individuata nelle seguenti due strategie:

a) individuare preliminarmente tutti i possibili subcampioni (non necessariamente disgiunti) a ciascuno dei quali è associato un ben determinato pattern monotono ed applicare il metodo di imputazione riassunto nella Fig. 3 al subcampione più numeroso, imputando i rimanenti M.D. con altri metodi;

b) scomporre il campione in due o più subcampioni disgiunti al cui interno i M.D. siano ordinabili secondo un pattern monotono ed applicare separatamente a ciascun subcampione il metodo di imputazione appena richiamato.

In particolare, sotto l'ipotesi di casualità del meccanismo generatore delle mancate risposte e di indipendenza delle estrazioni delle singole unità statistiche, i vari subcampioni "monotoni" che si otterrano seguendo la strategia (b) possono essere considerati come un insieme di subcampioni indipendenti di differente numerosità estratti da una medesima popolazione; ne consegue che questa strategia è ammissibile e corretta anche se meno efficiente dell'E.M.-algorithm.

3- Un'applicazione

3.1- La struttura dell'esperimento

Come sottolineato nei capitoli precedenti, gli stimatori di regressione posseggono le note proprietà di ottimo solo se la distribuzione del carattere nella popolazione è normale e se il meccanismo di generazione dei M.D. è ignorabile.

Nella realtà è difficile che queste condizioni siano rispettate congiuntamente. Nel caso dell'indagine Banca d'Italia, la distribuzione nel campione delle imprese per classi di investimento è asimmetrica ed heavy-tailed 11/; inoltre non si può escludere a priori l'esistenza di un legame sistematico fra le variabili e la probabilità di mancata risposta.

In queste circostanze si può verificare solo sperimentalmente se, pur in presenza di una violazione di tali ipotesi, gli stimatori di regressione conservano una superiorità rispetto ad altri stimatori, che d'altra parte soffrono essi stessi di tale violazione.

A questo scopo, sulla base del campione del 1987, è stato effettuato un esperimento che prevede i seguenti quattro passi.

1) Dal campione di 1.077 unità statistiche sono state inizialmente estratte quelle che non presentavano né M.D., né valori nulli in corrispondenza delle seguenti variabili:

E1=occupazione alla fine del 1986
E2=occupazione alla fine del 1987
E3=occupazione prevista alla fine del 1988
I1=investimenti effettuati nel 1986
I2=investimenti previsti per il 1987
I3=investimenti effettuati nel 1987
I4=investimenti previsti per il 1988

Questo subcampione (814 unità) rappresenta l'universo virtuale di riferimento. Nella realtà, le variabili E1, E2, I1, I3 presentano una frequenza relativa di M.D. molto contenuta, a differenza delle variabili I2, I4 ed E3 rispetto alle quali la frequenza di M.D. appare rilevante e crescente (fra il 5 ed il 15 per cento). Inoltre, il pattern dei M.D. non è monotono il che imporrebbe l'utilizzo dell'E.M.-algorithm per ottenere stime efficienti. Tuttavia, circa il 94% delle 1077 unità statistiche è ordinabile secondo un pattern monotono del tipo:

E1	E2	I1	I3	I2	I4	E3
p	p	p	p	p	p	p
p	p	p	p	p	p	m
p	p	p	p	p	m	m
p	p	p	p	m	m	m

Considerando missing i valori nulli, la frazione delle unità statistiche ordinabili secondo tale pattern scende all'89%.

Inoltre, nell'ambito delle sole osservazioni che presentano M.D. in corrispondenza di almeno una delle 7 variabili, quelle ordinabili secondo tale pattern sono il 69% (54% se si considerano missing i valori nulli).

Rispetto alle unità che presentano questo pattern monotono la

stima puntuale efficiente dei M.D. può essere realizzata senza ricorrere all'E.M.-algorithm, sulla base delle seguenti regressioni lineari stimate con il metodo dei minimi quadrati ordinari:

$$I2 = f_1(E1, E2, I1, I3; \Phi_1) \qquad [3.1.1]$$

$$I4 = f_2(E1, E2, I1, I3, I2; \Phi_2) \qquad [3.1.2]$$

$$E3 = f_3(E1, E2, I1, I3, I2, I4; \Phi_3) \qquad [3.1.3]$$

in cui Φ indica il vettore dei parametri della regressione.

2) Nel subcampione costituito dalle 814 unità statistiche che non presentavano nè M.D. nè valori nulli in corrispondeza di tutte e 7 le variabili sono stati generati 50 patterns monotoni indipendenti del tipo sopra illustrato, con p(I2=.)=0.05, p(I4=.)=0.10 e p(E3=.)=0.15 12/.

3) Con riferimento ad ognuno dei 50 patterns monotoni indipendenti sono state stimate le regressioni [3.1.1-3] nei logaritmi (di qui la necessità di eliminare preliminarmente le osservazioni con valori nulli). Sulla base delle regressioni sono stati imputati i M.D.; dai dati così completati si è calcolata la media delle variabili I2, I4 ed E3. Per l'opportuno confronto, sono state effettuate anche imputazioni alternative: "tasso di variazione medio per addetto di cella", "media per addetto di cella" e "tasso di variazione medio di cella",quest'ultimo solo per la variabile E3.

4) Al termine delle 50 simulazioni, sono state calcolate le medie

e le varianze delle 50 stime della media, con riferimento a ciascuno degli stimatori utilizzati.

3.2- Gli stimatori alternativi utilizzati

La definizione degli stimatori alternativi utilizzati per l'imputazione dei M.D. nell'esperimento è la seguente (gli indici "j" ed "i" indicano rispettivamente l'unità statistica e la cella, mentre la variabile sopra segnata indica la media):

a) Tasso di Variazione Medio per Addetto di Cella (TVMAC)

$$\hat{I2}_{j,i} = I1_{j,i} \cdot \overline{I2}_i / \overline{E2}_i \cdot \left(1 + \frac{\overline{I2}_i / \overline{E2}_i - \overline{I1}_i / \overline{E1}_i}{\overline{I1}_i / \overline{E1}_i}\right) \cdot E2_{j,i} \qquad [3.2.1]$$

$$\hat{I4}_{j,i} = I3_{j,i} \cdot \overline{I4}_i / \overline{E3}_i \cdot \left(1 + \frac{\overline{I4}_i / \overline{E3}_i - \overline{I3}_i / \overline{E2}_i}{\overline{I3}_i / \overline{E2}_i}\right) \cdot E3_{j,i} \qquad [3.2.2]$$

b) Media per Addetto di Cella (MAC)

$$\hat{I2}_{j,i} = E2_{j,i} \cdot \overline{I2}_i / \overline{E2}_i \qquad [3.2.3]$$

$$\hat{I4}_{j,i} = E3_{j,i} \cdot \overline{I4}_i / \overline{E3}_i \qquad [3.2.4]$$

c) Tasso di Variazione Medio di Cella, solo per l'occupazione (TVMC)

$$\hat{E3}_{j,i} = E2_{j,i} \cdot \left(1 + \frac{\overline{E3}_i - \overline{E2}_i}{\overline{E2}_i}\right) \qquad [3.2.5]$$

Nell'esperimento le celle sono 12 (3 classi dimensionali per

4 settori di attività economica). Si osservi dalle [3.2.2] e [3.2.4] che l'imputazione dei M.D. nella variabile I4 presuppone la conoscenza della variabile E3; se anche tale variabile è missing occorre procedere in due tappe: a) imputare i M.D. di E3 utilizzando il metodo del "tasso di variazione medio di cella"; b) imputare i M.D. di I4 mediante i due stimatori alternativi proposti ("tasso di variazione medio per addetto di cella" o "media per addetto di cella"), utilizzando i dati relativi alla variabile E3 "completati" nella prima tappa.

Gli stimatori della media di I4, I2, E3, come nel caso della regressione, sono definiti come media aritmetica dei valori effettivi ed imputati.

Si osservi che gli stimatori [3.2.1-5] possono configurarsi come stimatori di regressione "incompleti" nel senso che si basano su un modello lineare che descrive la relazione intercorrente fra alcune variabili e non sulla relazione lineare intercorrente fra tutte e 7 le variabili.

Infatti, quando la regressione passa per l'origine ed i residui sono eteroschedastici con varianze proporzionali al regressore, ovvero quando 13/:

$$y_i = \beta \cdot x_i + \epsilon_i \quad \text{con } Var(\epsilon_i) = \sigma_\epsilon^2 \cdot x_i$$

allora lo stimatore W.L.S. del coefficiente della regressione ($\hat{\beta}=\bar{x}_c/\bar{y}_c$) darà luogo alla seguente imputazione dei M.D.:

$$\hat{y}^{*}_i = \bar{y}_c/\bar{x}_c \cdot x_i$$

Conseguentemente lo stimatore della media sarà:

$$\hat{\mu}_{y}^{"} = \bar{y}_{c} / \bar{x}_{c} \cdot \hat{\mu}_{x} = (1/n) \cdot \left(\sum_{i=1}^{n_c} y_i + \sum_{i=1}^{m_y} \hat{y}_i^{"} \right)$$

In pratica, quindi, se X è una variabile di scala (ad esempio gli addetti) oppure rappresenta il medesimo fenomeno di Y in un periodo precedente, allora gli stimatori "media per addetto" o "tasso di variazione medio" si configurano come un caso particolare di stimatori di regressione. Si osservi, d'altronde, che pure l'imputazione mediante la media aritmetica dello strato a cui appartiene l'unità statistica che non ha fornito la risposta coincide con la stima puntuale della regressione della variabile missing sulle dummies variables rappresentative delle caratteristiche in base alle quali è stata effettuata la stratificazione (settore di attività, dimensione, ecc.).

Sulla base di queste considerazioni lo stimatore "tasso di variazione medio per addetto" presuppone un modello di regressione del tipo:

$$Y_{i,t} / X_{i,t} = \beta \cdot (Y_{i,t-1} / X_{i,t-1}) + \epsilon_i \qquad [3.2.6]$$

con $VAR(\epsilon_i) = \sigma_\epsilon^2 \cdot (Y_{i,t-1} / X_{i,t-1})$ ed $i=1,2,\ldots,n$.

Poniamo che Y ed X rappresentino, rispettivamente, gli investimenti e l'occupazione, mentre gli indici "i" e "t" indichino l'unità statistica e il tempo. La [3.2.6] esprime dunque gli investimenti per addetto correnti in funzione degli investimenti per addetto del periodo precedente. Moltiplicando

entrambi i membri della [3.2.6] per $X_{i,t}$ otteniamo:

$$Y_{i,t} = \beta \cdot Z_{i,t} + u_i \qquad [3.2.7]$$

in cui $Z_{i,t} = (Y_{i,t-1}/X_{i,t-1}) \cdot X_{i,t}$, $u_i = \epsilon \cdot X_{i,t}$ e $VAR(u_i) = \sigma^2_{\epsilon} \cdot Z_{i,t}$

Lo stimatore W.L.S. di β sarà dunque:

$$\hat{\beta} = \bar{Y}_t / \bar{Z}_t \tilde{} (\bar{Y}_t / \bar{X}_t)/(\bar{Y}_{t-1} / \bar{X}_{t-1}) \qquad [3.2.8]$$

Si osservi che l'approssimazione di cui alla [3.2.8] dipende dalla circostanza che, in genere, la media di un rapporto (prodotto) non coincide con il rapporto (prodotto) delle medie.

Si osservi, inoltre, che la considerazione di eventuali effetti di cella implica, rispetto alla [3.2.8], un modello del tipo:

$$Y_{ij,t}/X_{ij,t} = \sum_{j=1}^{h} \beta_j \cdot (Y_{ij,t-1}/X_{ij,t-1}) \cdot D_{ij} + \epsilon_i \qquad [3.2.9]$$

in cui l'indice "j" indica la cella (j=1,2,...,h) mentre D è una <u>dummy</u> (0,1) che consente di modellare opportunamente gli effetti di cella.

3.3- I dati

Nell'esperimento le regressioni lineari [3.1.1-3] sono state stimate sia nei logaritmi che, a scopo di confronto, nei valori originali.

La scelta della trasformata logaritmica rappresenta un compromesso fra l'esigenza di eliminare la non-normalità della distribuzione di tutte e 7 le variabili in esame e l'esigenza di non dilatare a dismisura la fase di search della trasformata migliore (vedi anche la nota 6).

In effetti, la trasformata logaritmica consente di ridurre almeno in parte la forte asimmetria che contraddistingue le distribuzioni marginali dei valori originali, come si può vedere esaminando la Tav. 1 in cui sono riportati la media e la moda di ogni variabile nonchè i due indici di forma (asimmetria e curtosi) ed i quartili.

Si osservi preliminarmente che il segno positivo di entrambi gli indici di forma implica una distribuzione con la moda "spostata a sinistra" rispetto alla media e "più appuntita" (leptocurtica) rispetto ad una distribuzione normale con stessa media e varianza.

Si osservi, tuttavia, che il passaggio ai logaritmi consente di ridurre sensibilmente entrambi gli indici di forma nonchè la discrepanza fra media, moda e mediana (parametri che in una distribuzione normale devono coincidere). In particolare, nelle distribuzioni dei valori originali la media` è sistematicamente superiore al terzo quartile, mentre nelle distribuzioni dei

Tav. 1

DISTRIBUZIONI DELLE VARIABILI NELL'UNIVERSO DI RIFERIMENTO (1)
PRINCIPALI INDICATORI

VALORI ORIGINALI

INDICATORI	I1	I2	I3	I4	E1	E2	E3
MEDIA	8846,15	10522,40	10332,50	11581,40	946,09	924,96	911,36
MODA	200,00	500,00	300,00	2000,00	50,00	75,00	68,00
ASIMMETRIA	16,59	18,42	21,15	20,97	19,59	19,59	19,92
CURTOSI	313,97	410,37	516,70	513,96	465,77	465,72	478,37
1° QUARTILE	502,25	600,00	631,50	600,00	130,00	131,00	133,00
2° QUART. (MEDIANA)	1604,00	1990,00	1800,00	2000,00	296,00	291,50	290,00
3° QUARTILE	4829,25	5620,00	5151,50	6884,25	693,00	673,75	674,25

LOGARITMI

INDICATORI	I1	I2	I3	I4	E1	E2	E3
MEDIA	7,398	7,567	7,515	7,665	5,803	5,794	5,791
MODA	5,298	6,215	5,704	7,601	3,912	4,317	4,220
ASIMMETRIA	0,199	0,214	0,063	0,065	0,783	0,801	0,793
CURTOSI	0,254	0,230	0,708	0,284	0,855	0,906	0,918
1° QUARTILE	6,219	6,397	6,448	6,397	4,868	4,875	4,890
2° QUART. (MEDIANA)	7,380	7,596	7,496	7,601	5,690	5,675	5,670
3° QUARTILE	8,482	8,634	8,547	8,837	6,541	6,513	6,514

(1) Subcampione di 814 osservazioni.

logaritmi viene praticamente a coincidere con la mediana (secondo quartile), pur risultando ancora superiore alla moda per effetto della residua asimmetria che permane anche dopo la trasformazione.

Nella Tav. 2 sono riportati i valori del test di accostamento alla distribuzione normale (D di Kolmogorov-Smirnov) nonchè le probabilità di un valore della D superiore a quello effettivo, sotto l'ipotesi nulla di normalità.

L'esame della Tav. 2 mostra che il passaggio dai valori originali ai logaritmi comporta sempre una sensibile riduzione della D, il che indica in ogni caso un migliore accostamento della funzione di ripartizione effettiva a quella teorica. Inoltre, per le variabili rappresentative degli investimenti (I1, I2, I3 e I4) il test non consente di rifiutare l'ipotesi di normalità della distribuzione; il medesimo test applicato ai valori originali, invece, porta al rifiuto di questa ipotesi con riferimento alle distribuzioni di tutte e 7 le variabili.

Si osservi, tuttavia, che la normalità delle distribuzioni marginali, rispetto alle quali sono stati calcolate le D, non implica necessariamente la multinormalità della distribuzione congiunta, mentre è vero il viceversa; questa circostanza si configura, d'altra parte, come un caso limite che può verificarsi solo in presenza di particolari distribuzioni multivariate 14/.

In generale, questi risultati mostrano che la trasformata logaritmica, pur non configurandosi necessariamente come la trasformazione migliore, produce un'approssimazione tutto sommato

Tav. 2

DISTRIBUZIONE DELLE VARIABILI NELL'UNIVERSO DI RIFERIMENTO (1)
TEST DI ACCOSTAMENTO ALLA DISTRIBUZIONE NORMALE (2)

VARIABILI	VALORI ORIGINALI		LOGARITMI	
	* D	* Pr(D>D)	* D	* Pr(D>D)
I1	0,431	<0,01	0,023	>0,15
I2	0,428	<0,01	0,028	0,135
I3	0,437	<0,01	0,028	0,126
I4	0,432	<0,01	0,031	0,053
E1	0,416	<0,01	0,058	<0,01
E2	0,414	<0,01	0,052	<0,01
E3	0,414	<0,01	0,051	<0,01

(1) Subcampione di 814 osservazioni. (2) Kolmogorov-Smirnov

accettabile delle condizioni sotto le quali gli stimatori di regressione esaminati nel cap. 2 sono stimatori di M.L..

Nel seguito si vedrà che questa trasformazione dei dati consente effettivamente di aumentare la precisione dello stimatore di regressione della media rispetto all'analogo stimatore calcolato sulla base dei valori originali. Risultati migliori non sono stati ottenuti con la trasformata di Box-Cox, attribuendo al parametro valori compresi fra 0 e 1. Questo risultato costituisce, dunque, una verifica ex-post dell'accettabilità della trasformata logaritmica.

3.4- I risultati

Nell'universo virtuale di riferimento le stime dei tre modelli lineari nei logaritmi [3.1.1-3] appaiono nel complesso soddisfacenti, nonostante la presenza di un'elevata collinearità, la cui riduzione - d'altra parte - comporterebbe l'eliminazione di qualche regressore e, di conseguenza, l'abbandono dell'approccio "completo" qui utilizzato.

Nella Tav. 3 sono riportati i risultati delle stime O.L.S. (fra parentesi i valori del test t). Si osservi che i valori dell'R^2 e della F sono piuttosto elevati; i coefficienti sono in buona parte significativi.

La regressione [3.1.3] relativa alla variabile E3, in particolare, presenta un R^2 pari a 0,995 ed un errore quadratico medio, rapportato alla media, pari a poco più dell'1,0%, contro il 6,4% e l'8,5% delle regressioni [3.1.2-3].

La soddisfacente performance del modello nell'universo è un presupposto ovviamente importante per la performance dello stimatore di regressione della media in presenza di M.D..

Come si è visto nel par. 2.2, in caso di indipendenza fra la variabile dipendente ed i regressori la regressione non fornisce informazioni supplementari; di conseguenza, lo stimatore di regressione della media non è più preciso dello stimatore univariato ("media semplice dei non-M.D."). Come vedremo in seguito, lo strettissimo accostamento della regressione [3.1.3] ai valori effettivi nell'universo di riferimento si traduce in una dispersione del corrispondente stimatore della media in presenza

Tav. 3

REGRESSIONI NELL'UNIVERSO DI RIFERIMENTO (1)

INDICATORI	VARIABILI DIPENDENTI		
	I2 (2)	I4 (3)	E3 (4)
R^2	0,92	0,86	0,99
F	2341,02	1010,35	28914,07
CV% (5)	6,35	8,47	1,40
INTERCETTA	0,083	0,118	0,026
	(0,978)	(1,034)	(1,808)
I1	0,222	0,128	-0,001
	(9,387)	(3,802)	(-0,023)
I2		0,597	-0,014
		(12,590)	(-2,212)
I3	0,665	0,110	0,011
	(29,272)	(2,504)	(1,908)
I4			0,011
			(2,613)
E1	0,269	-0,209	-0,315
	(1,311)	(-0,753)	(-9,086)
E2	-0,124	0,426	1,301
	(-0,583)	(1,484)	(36,252)

(1) Subcampione di 814 osservazioni. (2) Equazione [3.1.1]. (3) Equazione [3.1.2]. (4) Equazione [3.1.3] (5) Errore quadratico medio in percent. della media.

di M.D. relativamente più bassa rispetto a quella degli analoghi stimatori di regressione delle medie di I2 e I4.

Occorre tenere presente, naturalmente, che sia la distribuzione nell'universo sia, di conseguenza, i parametri della regressione non sono noti - almeno in genere - in situazioni diverse da quelle sperimentali qui riprodotte.

Nella Tav. 4 sono riportati le medie e le deviazioni standard dei diversi stimatori della media di E3, I2 e I4, calcolate sulla base dei dati - effettivi e imputati mediante i diversi metodi - relativi a 50 campioni con pattern dei M.D. monotono, generati indipendentemente l'uno dall'altro (cfr. par. 3.1) 15/.

Fra i possibili stimatori della media della variabile E3 sono stati presi in considerazione gli stimatori di regressione (**REGL** sui logaritmi e **REGO** sui valori originali) e "tasso di variazione medio di cella" (**TVMC**). Per la stima della media di I2 e I4 sono stati presi in considerazione, oltre ai citati stimatori di regressione, anche gli stimatori "tasso di variazione medio per addetto di cella" (**TVMAC**) e "media per addetto di cella" (**MAC**).

Nella Tav. 4 sono riportati (fra parentesi, rispettivamente sotto la media e la deviazione standard di ciascun stimatore) sia la distorsione calcolata sulla base della media dell'universo di riferimento, sia l'efficienza di ciascun stimatore calcolata rispetto a quello di minima varianza per ogni data variabile.

I risultati confermano la superiorità di entrambi gli stimatori di regressione, che presentano in ogni caso una

Tav. 4

PERFORMANCES DEGLI STIMATORI DELLA MEDIA (1)

STIMATORI	$\overline{I2}$		$\overline{I4}$		$\overline{E3}$	
	MEDIA	SIGMA	MEDIA	SIGMA	MEDIA	SIGMA
REGL	10519,51 (-0,03)	10,69 (100,0)	11527,74 (-0,51)	43,48 (100,0)	911,51 (0,00)	0,41 (100,0)
REGO	10580,11 (+0,55)	11,22 (105,0)	11579,24 (-0,02)	46,30 (106,5)	911,48 (0,00)	0,47 (114,6)
TVMAC	10534,93 (+0,12)	22,68 (212,2)	11577,65 (-0,03)	51,39 (118,2)		
MAC	10532,86 (+0,10)	25,52 (238,7)	11597,63 (+0,14)	56,59 (130,2)		
TVMC					915,32 (+0,43)	4,92 (1200,0)
UNIVERSO	10522,42		11581,41		911,36	

(1) 50 simulazioni

dispersione minore rispetto a quella degli stimatori alternativi. Fra gli altri due stimatori (**TVMAC** e **MAC**) il primo risulta leggermente più preciso del secondo; il guadagno di precisione è compreso fra il 9,2% (variabile I4) e l'11,1% (variabile I2). Anche lo stimatore **TVMC**, utilizzato solo per l'occupazione, è meno preciso rispetto a quelli di regressione.

Fra i due stimatori di regressione, quello che offre i risultati nettamente migliori, in accordo con le considerazioni svolte nel par. 3.3, è quello calcolato sulla base dei logaritmi.

Si osservi, ancora in accordo con le considerazioni del par. 3.3, che la precisione relativa dello stimatore **REGL** è legata alla proporzione della varianza spiegata nell'universo dalla regressione ed è massima nel caso dello stimatore **REGL** relativo alla variabile E3, la cui corrispondente regressione [3.1.3] presenta un R^2 elevatissimo:

INDICATORI	E3	I2	I4
SIGMA/MEAN% (1)	0,045	0,102	0,377
R-SQUARE (2)	0,995	0,920	0,861

(1) Coefficiente di variazione dello stimatore **REGL**. (2) Relativo alla corrispondente regressione nell'universo.

Si osservi che gli stimatori di regressione non prevedono effetti di cella né sull'intercetta né sui coefficienti; in effetti la simulazione con un modello che prevedeva tali effetti (incorporati in opportune dummies di cella) non ha dato risultati migliori. D'altra parte uno dei due parametri in base ai quali

sono costruite le celle - l'occupazione - è incorporato nei modelli di regressione in maniera fra l'altro più precisa poichè - come regressore - compare l'occupazione della singola impresa (effetto individuale) invece che la dummy associata alla rispettiva classe dimensionale (effetto di cella).

Lo stimatore **REGL** mostra una soddisfacente performance anche nella stima puntuale dei M.D. e, dunque, è utile anche per analizzare con sufficiente affidabilità i dati individuali completati.

Anche in questo caso la precisione dipende dal grado di accostamento della regressione ai valori effettivi nell'universo. Le stime puntuali dei M.D. della variabile E3 presentano, infatti, uno scostamento quadratico medio pari solo all'1,6% della media; il medesimo indicatore relativo alle stime dei M.D. della variabile I4 sale al 4,3%.

Naturalmente l'imputazione dei M.D. elimina una parte della variabilità del campione così "completato"; in termini informali possiamo dire che elimina, in ogni osservazione imputata, la componente della variabilità non spiegata dalla regressione. Come abbiamo visto nel par. 2.2 quanto più questa componente residua è piccola per effetto di un elevato accostamento della regressione ai valori effettivi, tanto più la distorsione della stima della varianza basata sui dati campionari effettivi ed imputati è piccola.

Nel caso dello stimatore di regressione questa distorsione è praticamente nulla con riferimento alle variabili I2 ed E3 e sale anche se a un livello trascurabile con riferimento alla variabile I4 a cui corrisponde una regressione nell'universo con un R^2 relativamente più basso:

	Standard deviation	
VARIABILI	EFFETTIVA	STIMATA
I2	57687,49	57687,56
E3	4072,09	4072,72
I4	67271,01	67256,52

Si osservi, infine, che la distribuzione dello stimatore REGL relativa alle 50 simulazioni è asintoticamente normale, con riferimento alle variabili I2 ed E3; l'ipotesi di normalità asintotica, invece, dovrebbe essere rifiutata con riferimento alla variabile I4.

Nella Tav. 5 sono riportati la media, la moda, i due indici di forma ed i quartili dello stimatore della media delle 3 variabili in esame; sono riportati, inoltre, il test di accostamento alla distribuzione normale (W di Shapiro-Wilk) e la probabilità di un valore di W inferiore a quello effettivo sotto l'ipotesi nulla di normalità. E' facile osservare che solo per le variabili I2 e, soprattutto, E3 la discrepanza fra media moda e mediana è contenuta, mentre il valore assunto dalla W porta all'accettazione dell'ipotesi di normalità; nel caso della variabile I4 la distribuzione di REGL presenta, invece, una certa asimmetria positiva mentre il valore assunto dalla W porta al rifiuto dell'ipotesi nulla.

Tav. 5

DISTRIBUZIONE DEGLI STIMATORI DI REGRESSIONE (1)
PRINCIPALI INDICATORI

INDICATORI	$\overline{I2}$	$\overline{I4}$	$\overline{E3}$
MEDIA	10519,5	11527,7	911,5
MODA	10493,3	11441,8	910,6
ASIMMETRIA	-0,431	-0,334	0,298
CURTOSI	0,526	-1,153	-0,288
1° QUARTILE	10514,1	11484,3	911,2
2° QUART. (MEDIANA)	10519,9	11540,4	911,4
3° QUARTILE	10526,8	11561,7	911,8
W^* (2)	0,974	0,923	0,978
$Pr(W<W^*)$	0,516	<0,01	0,654

(1) 50 simulazioni. (2) Shapiro-Wilk.

4- Alcune osservazioni conclusive

I risultati dell'esperimento, date le finalità essenzialmente operative, sono piuttosto soddisfacenti.

Gli stimatori di regressione della media, infatti, posseggono le proprietà desiderabili di correttezza e di efficienza (relativamente ad altri stimatori comunemente usati) anche in presenza di una violazione delle ipotesi fondamentali di normalità della distribuzione congiunta dei caratteri nella popolazione, attenuata però dalla trasformazione logaritmica, e di ignorabilità del meccanismo generatore delle mancate risposte.

Essi si prestano, dunque, ad un impiego operativo per l'imputazione dei dati mancanti e per la stima della media nell'indagine sugli investimenti della Banca d'Italia.

I risultati dell'esperimento, tuttavia, dipendono dalla distribuzione empirica dei dati utilizzati e dalle particolari ipotesi introdotte circa il meccanismo di generazione dei M.D.. Non è quindi lecito generalizzare queste conclusioni ad altre popolazioni.

D'altra parte si è visto che la non-normalità dei dati può essere eliminata mediante opportune trasformazioni. Nulla si può fare, operando solo sui dati, per eliminare l'influenza di un meccanismo generatore delle mancate risposte di tipo non casuale. In questa ultima circostanza occorre procedere su due vie complementari:

a) ridurre il più possibile il numero dei M.D. nella fase di

rilevazione, ovvero contenere alla radice il problema;

b) formulare ipotesi al riguardo ed incorporarle nel modello di regressione da utilizzare per l'imputazione 16/.

N O T E

(*) Ringrazio Luigi Cannari del Servizio Studi della Banca d'Italia, che mi ha stimolato ad approfondire la problematica offrendomi utili spunti ed indicazioni. Ringrazio pure Ettore Romagnano per l'editing. Ovviamente rimango unico responsabile di eventuali errori ed omissioni.

1/ Nel testo il termine osservazione è usato come sinonimo di unità statistica. L'incompletezza di un campione consiste nella mancata misurazione di una o più variabili in una o più osservazioni. Il termine M.D., se riferito ad una variabile, indica la mancata misurazione di quella variabile in una o più osservazioni; lo stesso termine, se riferito ad una osservazione, indica la mancata misurazione di una o più variabili in quella osservazione. Poichè nelle indagini campionarie la mancata misurazione è spesso dovuta alla mancata risposta, le due espressioni assumono convenzionalmente il medesimo significato.

2/ Per una rassegna estesa della letteratura si veda AA.VV. (1983). Accanto alle tecniche di imputazione basate su medie dei non-M.D., si collocano altre tecniche fra le quali quelle di hot-deck in cui il valore incognito di una determinata variabile viene sostituito dal valore assunto da tale variabile in una unità statistica estratta casualmente dall'intero campione dei rispondenti o dallo strato a cui appartiene l'unità statistica che non ha fornito la risposta.

3/ Si veda D. PICCOLO-C. VITALE (1981), pagg. 247-249.

4/ Una trattazione approfondita di questa problematica si trova, ad esempio, in H. WOLD (1966), cap. 2: "...noi dobbiamo in primo luogo stabilire lo scopo per il quale la regressione in esame sarà impiegata. Limiteremo la nostra indagine ai due principali fini dell'analisi della regressione: a) stimare o prevedere una variabile, data una o più altre variabili; b) ottenere una spiegazione causale di una variabile, in funzione di una o più variabili. Da un punto di vista formale la procedura dell'analisi di regressione è la stessa nelle due situazioni. La differenza consiste nel tipo di quesito posto al materiale statistico e nell'interpretazione data alla relazione di regressione. Nella situazione a) la regressione viene impiegata per stimare una variabile non conosciuta in termini di una o più variabili note, e la questione più importante è di fare le stime più accuratamente possibile.".

5/ Per una trattazione più rigorosa del problema e, in particolare, delle condizioni di ignorabilità del meccanismo generatore dei M.D. si veda R. J. A. LITTLE (1983), da cui è stato tratto il passo citato nel testo.

6/ In pratica si ricorre spesso, almeno come prima

approssimazione, alla trasformata logaritmica che equivale alla trasformata di Box-Cox nel caso in cui $\lambda=0$. Si osservi peraltro che in questo modo si generano dei M.D. in corrispondenza dei valori nulli delle variabili originarie. Questa circostanza assorbe in parte i guadagni di efficienza legati alla trasformazione. Inoltre il meccanismo generatore dei M.D. viene alterato non casualmente poichè la probabilità di non risposta viene a dipendere dal livello della variabile da imputare e dei regressori. D'altra parte, nel caso dell'indagine sugli investimenti condotta dalla Banca d'Italia è possibile che una parte dei valori nulli siano effettivamente mancate risposte, erroneamente interpretate in fase di elaborazione come valori nulli. Questa circostanza si realizza prevalentemente quando le imprese indicano la mancata risposta in maniera diversa da quanto espressamente previsto dalle istruzioni, ovvero quando appongono nella relativa casella del questionario, anzichè la prevista sigla ND, simboli del tipo "=" oppure "/" che possono essere assimilati allo 0.

7/ Una verifica sperimentale delle performances di altri metodi di imputazione dei M.D. (hot deck, tasso di variazione medio di cella, impresa tipo) è stata effettuata da L. CANNARI (1988) con riferimento all'indagine sugli investimenti delle imprese manifatturiere condotta dalla Banca d'Italia, a cui fa riferimento anche questo saggio. Una analoga verifica degli stimatori di regressione è stata sperimentata dal medesimo Autore con riferimento ai dati individuali dell'indagine sui bilanci delle famiglie italiane, anch'essa elaborata dalla Banca d'Italia.

8/ Si vedano, in proposito, i saggi di F.M. LORD (1955) e T.W. ANDERSON (1957). A questo risultato i due autori giungono seguendo vie differenti. Lord, infatti, considera la verosimiglianza di

$$\Theta = (\mu_x, \mu_y, \sigma^2_x, \sigma^2_y, \sigma_{xy})$$

corrispondente alla fattorizzazione della distribuzione normale bivariata di Y ed X nella distribuzione congiunta delle n osservazioni complete e nella distribuzione marginale delle m osservazioni in cui solo X è misurata. Anderson, invece, considera la verosimiglianza di

$$\Phi = (\mu_x, \sigma^2_x, \delta_{yx}, \beta_{yx}, \sigma^2_{y|x})$$

corrispondente alla fattorizzazione della distribuzione normale bivariata di Y ed X nella distribuzione marginale di X e nella distribuzione condizionata di Y dato X. I parametri

$$\delta_{yx}, \beta_{yx}, \sigma^2_{y|x}$$

rappresentano rispettivamente l'intercetta e il coefficiente angolare della regressione di Y su X nonchè la varianza della distribuzione condizionata di Y dato X. Dalla massimizzazione di

$$L(\Phi)=L(\mu_x, \sigma^2_x)\cdot L(\delta_{yx}, \beta_{yx}, \sigma^2_{y|x})$$

discendono le seguenti stime di M.L. dei parametri della popolazione:

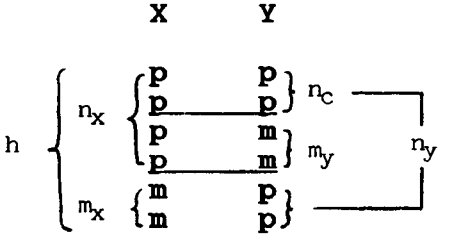
$$\hat{\mu}_x = (1/n)\cdot\left(\sum_{i=1}^{n} x_i\right) \qquad \hat{\sigma}^2_x = (1/n)\cdot\left\{\sum_{i=1}^{n} (x_i - \hat{\mu}_x)^2\right\}$$

$$\hat{\delta}_{yx} = \bar{y}_c - \hat{\beta}_{yx}\cdot\bar{x}_c \qquad \hat{\beta}_{yx} = \sigma_{yx,c}/\sigma^2_{x,c}$$

$$\hat{\sigma}^2_{y|x} = \sigma^2_{y,c} - (\sigma^2_{yx,c}/\sigma^2_{x,c})$$

Per ottenere, infine, le stime di M.L. della media e varianza di Y nonchè della covarianza fra X ed Y è sufficiente applicare le note proprietà che legano, nell'ipotesi di normalità, i parametri delle distribuzioni congiunte, condizionate e marginali (vedi la Nota 3).

9/ In una serie di articoli pubblicati fra il 1967 ed il 1969 (una sintesi si trova in MADDALA (1977)) Afifi ed Elashoff hanno esaminato diverse tecniche di imputazione ed i relativi stimatori dei parametri della popolazione, limitando l'analisi al caso bivariato con pattern dei M.D. non monotono, ovvero ad esempio:



in cui:

n	= numero delle	osservazioni	totali
n_c	= "	"	complete
n_y	= "	"	in cui Y é misurata
n_x	= "	"	in cui X é misurata
m_y	= "	"	in cui Y é missing
m_x	= "	"	in cui X é missing

da cui deriva: $m_y=n-n_y$, $m_x=n-n_x$.

Fra gli stimatori della media esaminati dagli Autori meritano un cenno particolare quelli che derivano dallo "zero order method" (in seguito: ZOM) e dal "first order regression method"

(in seguito FORM).

La differenza fra i due metodi consiste nel metodo di imputazione dei M.D. di ciascuna variabile: media delle osservazioni non missing (ZOM); stima puntuale di regressione di X su Y e di Y su X (FORM).

METODO	VARIABILE	IMPUTAZIONE	STIMATORE DELLA MEDIA
FORM	Y	$\hat{y}_i = \bar{y}_c + \hat{\beta}(x_i - \bar{x}_c)$	$\hat{\mu}_y = \frac{n_x}{n}\bar{y}_c + \frac{m_x}{n}\bar{y}_m + \frac{m_y}{n}\hat{\beta}(\bar{x}_m - \bar{x}_c)$
	X	$\hat{x}_i = \bar{x}_c + \hat{\delta}(y_i - \bar{y}_c)$	$\hat{\mu}_x = \frac{n_y}{n}\bar{x}_c + \frac{m_y}{n}\bar{x}_m + \frac{m_x}{n}\hat{\delta}(\bar{y}_m - \bar{y}_c)$
ZOM	Y	$\hat{y}'_i = \bar{y}_c$	$\hat{\mu}'_y = \bar{y}_c$
	X	$\hat{x}'_i = \bar{x}_c$	$\hat{\mu}'_x = \bar{x}_c$

in cui $\hat{\beta}=\sigma_{xy,c}/\sigma^2_{x,c}$, $\hat{\delta}=\sigma_{xy,c}/\sigma^2_{y,c}$; $\bar{x}_m$ e $\bar{y}_m$ sono le medie di X ed Y relative, rispettivamente, alle m_y ed m_x osservazioni in cui Y ed X sono missing. La distribuzione asintotica non condizionata dei due stimatori è la seguente:

$$\text{ZOM: } \sqrt{n}\,(\hat{\mu}'_y - \mu_y) \sim N\{0,\ \sigma^2_y/\tau_y\}$$

$$\text{FORM: } \sqrt{n}(\hat{\mu}_y - \mu_y) \sim N\{0, (\sigma^2_y - \beta^2\sigma^2_x)[\tau_x^2 + (1-\tau_x)] + \beta^2\sigma^2_x[(\tau_x + 1 - \tau_y)^2 + (2 - \tau_x - \tau_y)]\}$$

in cui $\tau_x = \text{plim}(n_x/n)$, $\tau_y = \text{plim}(n_y/n)$ e $\tau_c = \text{plim}(n_c/n)$.

E' facile dimostrare che, nel caso in cui $\tau_x = 1$ i due stimatori coincidono con quelli esaminati nel cap. 3.

10/ Si veda A.P. DEMPSTER-N.M. LAIRD-D.B. RUBIN (1977). La tecnica - denominata dagli Autori EM algorithm - prevede i seguenti passi:

1) Costruzione, sulla base delle n_c osservazioni complete delle stime iniziali del vettore delle medie (μ) e della matrice di

varianze-covarianze (Σ) delle k variabili oggetto dell'indagine.

2) Imputazione dei M.D. sulla base delle m_i regressioni fra le m_i variabili che nella i-esima osservazione (i=1,2,...,n) sono missing e le $p_i = k - m_i$ variabili che nella medesima osservazione sono misurate, stimando i parametri delle regressioni sulla base di μ e Σ.

3) Ricalcolo di μ e Σ sulla base dei dati completati. Per eliminare la distorsione delle varianze-covarianze dovuta al fatto che i valori veri, ma incogniti, di alcune variabili sono imputati (vedi cap. 3) bisogna correggere il contributo di ciascuna osservazione alle varianze-covarianze, ovvero:

$$(y_{ij} - \hat{\mu}_j) \cdot (y_{ih} - \hat{\mu}_h) \qquad i=1,2,\ldots,n \quad j,h=1,2,\ldots,k$$

aumentandolo di un termine $c_{i,jh}$ che rappresenta la covarianza residua (varianza se j=h) fra le variabili Y_j ed Y_h calcolata sulla base delle regressioni di Y_j e di Y_h sulle p_i variabili che nella i-esima osservazione sono misurate. Il termine $c_{i,jh}$ ovviamente sarà nullo se y_{ij} ed y_{ih} non sono missing.

4) Iterazione della procedura sino a che la variazione delle stime da un'iterazione all'altra scende al di sotto di un valore prefissato.

11/ Si veda L. CANNARI, op. cit.

12/ Allo scopo di verificare la performance degli stimatori di regressione della media nell'ipotesi di non ignorabilità del meccanismo generatore delle mancate risposte e di assicurare nel contempo l'indipendenza, ogni singola generazione è stata realizzata:
a) assegnando ad ogni unità statistica del subcampione un numero casuale K_i compreso fra 0 e l'unità, così determinato:

$$K_i = (K_{1,i} + K_{2,i} + e_i)/3$$

in cui K_1 e K_2 sono due scalari deterministici, compresi per definizione fra 0 e l'unità, mentre e è un numero casuale, anch'esso compreso fra 0 e l'unità, estratto da una distribuzione uniforme. Gli scalari K_1 e K_2 sono definiti sulla base delle seguenti relazioni lineari:

$$K_{1,i} = a + b \cdot E2_i , \qquad K_{2,i} = c + d \cdot S_i$$

in cui S è il codice del settore di attività (S=1, 2, 3, 4), E2 è il logaritmo dell'occupazione alla fine del 1987 mentre i

parametri assumono i seguenti valori:

$$a = Max(E2)/[Max(E2)-Min(E2)]$$
$$b = -1/[Max(E2)-Min(E2)]$$
$$c = -Min(S)/[Max(S)-Min(S)]$$
$$d = 1/[Max(S)-Min(S)]$$

Dalla definizione scaturisce che K_1 aumenta al diminuire dell'occupazione ed assume i valori estremi, 0 e 1, rispettivamente in corrispondenza della più grande e della più piccola impresa. Analogamente, K_2 assume i predetti valori estremi in corrispondenza delle "Industrie di Base" (S=1) e delle "Industrie Manifatturiere Varie" (S=4). Per ogni unità statistica K_1 e K_2 rimangono costanti di generazione in generazione, mentre e varia casualmente.

Quindi, nell'ambito di ogni singola generazione, K assumerà valori diversi da una unità statistica all'altra per l'effetto di una componente deterministica (costante di generazione in generazione) e di una componente stocastica, il cui meccanismo generatore - ma non ovviamente la determinanzione empirica - rimane invariato di generazione in generazione.

b) ordinando le unità statistiche in base ai valori crescenti di K; in questo modo, ad esempio, una impresa "piccola" ed appartenente al settore delle "Industrie Manifatturiere Varie" ha più scarse probabilità di trovarsi ai primi posti della graduatoria.

c) ponendo: E3=. in corrispondenza delle unità statistiche per le quali $g_i >= (1-0.15)\cdot N$, I4=. per $g_i >= (1-0.10)\cdot N$ e, infine, I2=. per $g_i >= (1-0.05)\cdot N$, in cui g_i ed N rappresentano rispettivamente il posto occupato in graduatoria dalla i-esima unità statistica ed N la numerosità del subcampione (814 unità).

13/ Si veda G. S. MADDALA (1979).

14/ Si veda D. PICCOLO-C. VITALE (1981), pag. 257.

15/ Si osservi che l'esperimento è stato disegnato considerando il campione dell'indagine della Banca d'Italia come la popolazione virtuale di riferimento ed estraendo da tale popolazione campioni di numerosità pari alla popolazione stessa. In questo modo si elimina la varianza degli stimatori della media dovuta al campionamento mentre rimane la varianza dovuta all'imputazione dei M.D. generati di volta in volta.

16/ Si veda, in proposito, l'approccio al "non-ignorable case" di R. J. A. LITTLE (1983).

BIBLIOGRAFIA

AA.VV (1983), Incomplete data in sample surveys, Academic Press, New York.

AFIFI A.A. - ELASHOFF R.M. (1966), Missing Observations in Multivariate Statistics I: Review of the literature, in "Journal of American Statistical Association", n. 61.

AFIFI A.A. - ELASHOFF R.M. (1967), Missing Observations in Multivariate Statistics II: Point Estimation in Simple Linear Regression, in "Journal of American Statistical Association", n. 62.

AFIFI A.A. - ELASHOFF R.M. (1969a), Missing Observations in Multivariate Statistics III: Large Sample Analysis of Simple Linear Regression, in "Journal of American Statistical Association", n. 64.

AFIFI A.A. - ELASHOFF R.M. (1969b), Missing Observations in Multivariate Statistics III: A Note on Simple Linear Regression, in "Journal of American Statistical Association", n. 64.

ANDERSON T.W. (1957), Maximun likelihood estimates for a multivariate normal distribution when some observations are missing, in "Journal of American Statistical Association", n. 52.

BOX G.E.P. - COX D.R. (1964), An Analysis of Trasformations (with discussion), in "Journal of the Royal Statistical Society", n. 2.

CANNARI L. (1988), L'imputazione di informazioni mancanti: una sperimentazione, Temi di Discussione del Servizio studi della Banca d'Italia, Roma, n. 100.

CESARI R. - SIGNORINI L.F. (1989), Stime regionali con "pochi dati": analisi e simulazione di stimatori alternativi per investimenti, occupazione e fatturato nelle imprese manifatturiere, Roma, dattiloscritto.

DEMPSTER A.P. - LAIRD N.M. - RUBIN D.B. (1977), Maximum Likelihood from Incomplete Data via E.M. Algorithm (with discussion), in "Journal of the Royal Statistical Society", series B, n. 39.

HUNSEN M.H. - HURWITZ W.N. - MADOW W.G.(1953), Sample Survey Methods and Theory, Wiley, New York.

KENDALL M.G. - STUART A. (1967), Advanced Theory of Statistics, Griffin, London.

LITTLE R.J.A. (1983), The Nonignorable Case, in "Incomplete data in sample surveys", Academic Press, New York, cap. 22.

LORD F.M. (1955), Estimation of parameters from incomplete data, in "Journal of American Statistical Association", n. 50.

MADDALA M. (1977), Econometrics, McGraw Hill, New York.

PICCOLO D. - VITALE C. (1981), Metodi statistici per l'analisi economica, Il Mulino, Bologna.

WILKS S.S. (1932), Moments and distributions of estimates of population parameters from fragmentary samples, in "Annals of Matematical Statistics", n. 3.

WOLD H. (1966), Analisi della domanda, Feltrinelli, Milano.

ELENCO DEI PIÙ RECENTI TEMI DI DISCUSSIONE (*)

n. 116 — *LDCs' repayment problems: a probit analysis*, di F. DI MAURO - F. MAZZOLA (maggio 1989).

n. 117 — *Mercato interbancario e gestione degli attivi bancari: tendenze recenti e linee di sviluppo*, di G. FERRI - P. MARULLO REEDTZ (giugno 1989).

n. 118 — *La valutazione dei titoli con opzione di rimborso anticipato: un'applicazione del modello di Cox, Ingersoll e Ross ai CTO*, di E. BARONE - D. CUOCO (giugno 1989).

n. 119 — *Cooperation in managing the dollar (1985-87): interventions in foreign exchange markets and interest rates*, di E. GAIOTTI - P. GIUCCA - S. MICOSSI (giugno 1989).

n. 120 — *The US current account imbalance and the dollar: the issue of the exchange rate pass-through*, di C. MASTROPASQUA - S. VONA (giugno 1989).

n. 121 — *On incentive-compatible sharing contracts*, di D. TERLIZZESE (giugno 1989).

n. 122 — *The adjustment of the US current account imbalance: the role of international policy coordination*, di G. GOMEL - G. MARCHESE - J. C. MARTINEZ OLIVA ((luglio 1989).

n. 123 — *Disoccupazione e dualismo territoriale*, di G. BODO - P. SESTITO (agosto 1989).

n. 124 — *Redditi da lavoro dipendente: un'analisi in termini di capitale umano*, di L. CANNARI G. PELLEGRINI - P. SESTITO (settembre 1989).

n. 125 — *On the estimation of stochastic differential equations: the continuous-time maximum-likelihood approach*, di R. CESARI (settembre 1989).

n. 126 — *La misurazione dell'efficienza nei modelli di "frontiera"*, di M. GRESTI (settembre 1989).

n. 127 — *Do intergenerational transfers offset capital market imperfections? Evidence from a cross-section of Italian households*, di L. GUISO - T. JAPPELLI (settembre 1989).

n. 128 — *La struttura dei rendimenti per scadenza secondo il modello di Cox, Ingersoll e Ross: una verifica empirica*, di E. BARONE - D. CUOCO - E. ZAUTZIK (ottobre 1989).

n. 129 — *Il controllo delle variabili monetarie e creditizie: un'analisi con il modello monetario della Banca d'Italia*, di I. ANGELONI - A. CIVIDINI (novembre 1989).

n. 130 — *L'attività in titoli delle aziende di credito: un'analisi di portafoglio*, di G. FERRI - C. MONTICELLI (dicembre 1989).

n. 131 — *Are asymmetric exchange controls effective?* di F. PAPADIA - S. ROSSI (gennaio 1990).

n. 132 — *Misurazione dell'offerta di lavoro e tasso di disoccupazione*, di P. SESTITO (marzo 1990).

n. 133 — *Progressing towards European Monetary Unification: Selected Issues and Proposals*, di L. BINI SMAGHI (aprile 1990).

n. 134 — *Il valore informativo delle variabili finanziarie: un'analisi con il modello econometrico trimestrale della Banca d'Italia*, di I. ANGELONI e A. CIVIDINI (aprile 1990).

n. 135 — *A Model for Contingent Claims Pricing on EMS Exchange Rates*, di A. ROMA (maggio 1990).

n. 136 — *Le attività finanziarie delle famiglie italiane*, di L. CANNARI - G. D'ALESSIO - G. RAIMONDI - A. I. RINALDI (luglio 1990).

n. 137 — *Sistema pensionistico e distribuzione dei redditi*, di L. CANNARI - D. FRANCO (luglio 1990).

n. 138 — *Time Consistency and Subgame Perfection: the Difference between Promises and Threats*, di L. GUISO - D. TERLIZZESE (luglio 1990).

n. 139 — *Test di integrazione e analisi di cointegrazione: una rassegna della letteratura e un'applicazione*, di G. BODO - G. PARIGI - G. URGA (luglio 1990).

n. 140 — *The Experience with Economic Policy Coordination: the Tripolar and the European Dimensions*, di G. GOMEL - F. SACCOMANNI - S. VONA (luglio 1990).

n. 141 — *The Short-Term Behavior of Interest Rates: Did the Founding of the Fed Really Matter?*, di P. ANGELINI (ottobre 1990).

n. 142 — *Evoluzione e performance dei fondi comuni mobiliari italiani*, di F. PANETTA - E. ZAUTZIK (ottobre 1990).

(*) I «Temi» possono essere richiesti a:
Banca d'Italia - Servizio Studi - Divisione Biblioteca e Pubblicazioni - Via Nazionale, 91 - 00184 Roma.

BANCA D'ITALIA - CENTRO STAMPA