UNIVERSITA' DEGLI STUDI DI PISA

Dipartimento di Statistica e Matematica Applicata all'Economia

Report n. 60

Un modello logit per lo studio del fenomeno delle nuove imprese

Gilberto GHILARDI

Pisa, giugno 1992

La ricerca è stata finanziata in parte dal Ministero dell'Università e della Ricerca Scientifica e Tecnologica (fondi 60%)

INDICE

1. Introduzione

Il fenomeno delle nuove imprese è oggetto dell'attenzione degli studiosi di economia e degli operatori privati e pubblici (Lissoni, 1991) che si occupano di argomenti connessi al sistema produttivo. In particolare, in un ambito locale il settore delle imprese manifatturiere talvolta è oggetto di analisi, che sono effettuate con dati di natura amministrativa, viste le carenze delle fonti statistiche ufficiali in questo campo, oppure con i dati ricavati da indagini statistiche finalizzate per lo studio delle caratteristiche di tali imprese. In alcuni casi, si tratta di dati ottenuti mediante indagini che si avvalgono di questionari somministrati con intervista diretta ed i risultati sono resi disponibili soprattutto sotto forma di distribuzioni statistiche semplici e doppie, che sono di lettura immediata. Tuttavia, spesso si rende utile l'analisi di distribuzioni multiple, che possono essere interpretate agevolmente mettendo a punto dei modelli statistici appropriati.

In proposito, tra i problemi interessanti per lo studio del fenomeno relativo alla nascita di nuove imprese, la sopravvivenza di queste dopo un certo intervallo di tempo dalla nascita può essere interpretato come un successo e costituisce un evento che dipende da una molteplicità di elementi, quali le caratteristiche individuali dell'impresa, le difficoltà incontrate e le capacità di superarle.

Da questo punto di vista è sembrato utile cercare di mettere a punto un modello per stabilire se ed in quale misura sussiste una relazione tra il successo dell'attività di nuova impresa e le variabili che caratterizzano l'impresa stessa, quali ad esempio le condizioni che hanno favorito la nascita, l'attività svolta e la dimensione aziendale. Inoltre, si è ritenuto interessante studiare le variazioni della possibilità di successo (o di insuccesso) delle nuove imprese nate in anni diversi.

Pertanto, ci siamo posti il problema di mettere a punto un modello per l'interpretazione sistematica dei dati statistici, relativi ad un determinato insieme di nuove imprese ed alla loro sopravvivenza o meno dopo un prestabilito numero di anni dalla nascita, circostanza che viene indicata come successo o insuccesso dell'attività intrapresa. Successivamente, si è provveduto alla definizione di due indici, che possono essere considerati delle misure della probabilità di successo di un'impresa nella sua attività e che appaiono utili per valutarne l'andamento nel tempo. Infine, a scopo esemplificativo, abbiamo effettuato un'applicazione concreta, mediante le informazioni raccolte nell'ambito di un osservatorio realizzato in ambito locale sulle nuove imprese.

2. Il modello statistico

Nell'analisi del fenomeno delle nuove imprese possiamo indicare con Y* la variabile non osservabile che rappresenta la possibilità o capacità di una nuova impresa in relazione alla sopravvivenza trascorsi t anni dalla sua nascita, sopravvivenza che considereremo come un successo dell'attività d'impresa. Inoltre, indichiamo con la notazione X_1, X_2, ..., X_k le k variabili che caratterizzano l'impresa alla nascita, al fine di misurare la relazione statistica esistente tra la variabile Y* e le variabili X_j (j=1, 2, ...k), riguardando la variabile Y* come una funzione

$$Y^* = f(X_1, X_2, ..., X_k, e) \tag{1}$$

delle variabili esplicative X_1, X_2, ..., X_k e della componente stocastica e (Judge et al., 1985; Walker and Duncan, 1967). Tuttavia, dato che la variabile Y* non è osservabile, si deve fare riferimento alla variabile dicotomica Y

$$Y: \begin{array}{l} 1 \text{ (successo), se } Y^* > c \\ 0 \text{ (insuccesso), se } Y^* \leq c \end{array} \tag{2}$$

o variabile indicatrice degli eventi successo e insuccesso, che supponiamo sia legata alla variabile Y* attraverso una soglia c, in base alla quale si riscontrano gli eventi menzionati. Naturalmente, ai fini pratici dell'impiego di un modello è determinante specificare analiticamente la relazione indicata (1). Il caso più semplice (Goldberger, 1964) sarebbe quello in cui la variabile dipendente è una combinazione lineare

$$Y^* = \sum_{j=1}^{k} a_j X_j + e \tag{3}$$

delle variabili esplicative con determinati coefficienti a_j e della componente stocastica e. In queste condizioni (Maddala, 1983) per ciascuna impresa i la variabile dipendente

$$Y_i^* = \sum_{j=1}^{k} a_j x_{ij} + e_i \tag{4}$$

è una funzione delle caratteristiche individuali x_{ij} (con x_{i1}=1 per qualunque i, in modo da ottenere l'intercetta a_1) e della componente non osservabile e_i, in cui i coefficienti a_j

assumono un diverso significato a seconda che si riferiscano ad una variabile continua oppure ad una variabile dicotomica (Cox, 1970). Supposto per semplicità che il modello indicato contenga solo variabili esplicative continue o dicotomiche, quando il coefficiente a_j si riferisce ad una variabile continua, esso ha il significato usuale di un coefficiente di regressione tra le variabili X_j e Y*, mentre nel caso in cui la variabile X_j è una variabile dicotomica con modalità 0 e 1, il segno del coefficiente a_j dipende esclusivamente dalla modalità alla quale è assegnato il codice 1 e il suo valore assoluto indica la variazione subita dal valore della variabile Y*, quando è presente la modalità 1 della variabile esplicativa anzichè la modalità 0. In entrambi i casi il coefficiente a_j sarà pari a zero se in base al modello prescelto la variabile X_j non influisce sulla variabile Y*, mentre sarà diverso da zero se tra la variabile esplicativa e la variabile Y* sussiste una certa relazione.

Per quanto riguarda la variabile non osservabile e_i, questa può essere rappresentata attraverso la differenza

$$e_i = Y_i^* - \sum_{j=1}^{k} a_j \, x_{ij} \qquad (5)$$

tra la variabile Y_i^* e la combinazione lineare delle variabili esplicative. Inoltre, se supponiamo che la componente e_i (i=1, 2,..., N) per tutte le N imprese abbia media nulla e varianza σ^2, allora il valore atteso condizionato

$$E[Y^*|x_1,x_2, \ldots,x_k] = \sum_{j=1}^{k} a_j \, x_j \qquad (6)$$

della variabile Y* dipende dalle variabili esplicative X_1, X_2, ..., X_k il cui peso è rappresentato dai coefficienti della combinazione lineare. Tuttavia, come si è detto, nei casi concreti la variabile Y* non è osservabile, mentre è osservabile la variabile dicotomica Y, il cui valore atteso condizionato

$$E[Y|x_1, x_2, \ldots, x_k] = P\,[Y=1|x_1, x_2, \ldots, x_k] \qquad (7)$$

è uguale alla probabilità di successo condizionata dalle variabili esplicative. Inoltre, notiamo che per la variabile osservabile Y non è opportuno (McFadden, 1974) usare una specificazione analoga a quella riportata (3) per la variabile Y*, perchè l'adattamento del modello potrebbe dare luogo a valori della probabilità esterni all'intervallo (0,1), valori che non rappresentano una soluzione accettabile, e perchè tale modello presuppone una distribuzione uniforme della componente stocastica, per la quale è generalmente più appropiato assumere una distribuzione diversa.

Il problema descritto brevemente può essere affrontato, considerando che un'impresa riscontri un successo, ovvero che l'impresa sia in vita dopo t anni dalla nascita, qualora essa superi un particolare livello c

$$c < E[Y^*|x_1, x_2, ..., x_k] + e_i \tag{8}$$

con una probabilità

$$P\{c<E[Y^*|x_1,x_2, ...,x_k] + e_i\} = P[Y=1|x_1,x_2, ...,x_k] \tag{9}$$

che dipende dalle caratteristiche individuali e dalla capacità individuale di superare le difficoltà incontrate. In queste condizioni, se un'impresa ha una probabilità p' di successo

$$p' = P[Y=1|x'_1, x'_2, ..., x'_k] > p \tag{10}$$

maggiore della probabilità p relativa ad un'altra impresa per il fatto di appartenere ad una categoria di imprese con caratteristiche individuali $x'_1, x'_2, ..., x'_k$, allora essa è in grado di raggiungere il livello c con una capacità individuale $e_{i'}$ inferiore a quella e_i di un'altra impresa, che ha una probabilità p di successo. Se ora scriviamo la probabilità di successo

$$\begin{aligned} &P\ [Y=1|x_1, x_2,...,x_k] = P\{E[Y^*|x_1, x_2,...,x_k]+e>c\}= \\ &= P\ \{e>c - E[Y^*|x_1, x_2,...,x_k]\} \end{aligned} \tag{11}$$

di un'impresa in funzione della componente stocastica e, si vede facilmente che tale probabilità

$$\begin{aligned} &P\ [Y=1|x_1, x_2,...,x_k] = 1 - P\{e\leq c - E[Y^*|x_1, x_2,...,x_k]\}= \\ &= 1 - F_e\ (c - E[Y^*|x_1, x_2,...,x_k]) \end{aligned} \tag{12}$$

dipende dalla funzione di ripartizione F_e (.) di tale componente. Inoltre, è immediato rilevare che se il valore atteso di Y* condizionato dai valori $x'_1, x'_2, ..., x'_k$,

$$E[Y^*|x'_1, x'_2,...,x'_k] > E[Y^*|x_1, x_2,...,x_k] \tag{13}$$

è maggiore di quello condizionato dai valori $x_1, x_2, ..., x_k$, allora il valore della funzione di ripartizione

$$F_e\ (c - E[Y^*|x'_1, x'_2,...,x'_k]) < F_e(c - E[Y^*|x_1, x_2,...,x_k]) \qquad (14)$$

della componente stocastica e nel punto c-E[Y*| x'_1, x'_2, ..., x'_k] risulta minore di quello calcolato nel punto c-E[Y*| x_1, x_2, ..., x_k] e conseguentemente la probabilità di successo

$$P[Y=1|\ x'_1, x'_2, ..., x'_k] > P[Y=1|\ x_1, x_2, ..., x_k] \qquad (15)$$

condizionata dai valori x'_1, x'_2, ..., x'_k risulta maggiore di quella condizionata dai valori x_1, x_2, ..., x_k. In generale, la probabilità di successo può essere scritta altrimenti,

$$P[e > c - \sum_j a_j x_j] = P[e > a_1^* - \sum_{j=2}^{k} a_j x_j] \qquad (16)$$

ricorrendo all'uguaglianza (11), in funzione dei coefficienti a_1^*, a_2, a_3,...,a_k della combinazione lineare delle variabili esplicative, dove il coefficiente a_1^* rappresenta la differenza tra le costanti c e a_1. Pertanto, non è possibile determinare separatamente questi due valori, mentre si può determinare l'effetto delle variabili esplicative X_1, X_2, ..., X_k sulle probabilità attraverso i restanti coefficienti a_2, ..., a_k. Inoltre, supposto che la componente stocastica abbia media $\mu=0$ e varianza σ^2, si può scrivere l'equazione

$$P[e > a_1^* - \sum_{j=2}^{k} a_j x_j] = P[\frac{e}{\sigma} > \frac{a_1^*}{\sigma} - \sum_{j=2}^{k} \frac{a_j}{\sigma} x_j] \qquad (17)$$

che fornisce la probabilità di successo in base alla variabile stocastica e/σ con media nulla e varianza unitaria. Infine, riparametrizzando il modello indicato, la probabilità cercata

$$P[Y=1|x_1, x_2, ...x_k] = 1 - F_{e/\sigma}\ (- \sum_j b_j x_j) \qquad (18)$$

è data dalla funzione di ripartizione della componente stocastica $z = e/\sigma$, funzione che dipende dai coefficienti $b_1= (a_1-c)/\sigma$, $b_2= a_2/\sigma$, ..., $b_k= a_k/\sigma$ delle variabili esplicative e dalla quale è evidente che se il valore della combinazione lineare $-\Sigma\, b_j x_j$ diminuisce allora la probabilità di successo aumenta e viceversa.

Per quanto concerne la stima dei coefficienti b_j (j =1, 2,..., k), supposto che le componenti e_1, e_2, ..., e_N siano indipendenti, il metodo della massima verosimiglianza generalmente è adeguato (Maddala, 1983) e si basa sulla verosimiglianza

$$L(Y_1, Y_2, \ldots Y_N; b_1, b_2, \ldots, b_k) = \prod_{1}^{N} [F_{e/\sigma}(-\sum_{j} b_j x_{ij})]^{1-y_i} [1 - F_{e/\sigma}(-\sum_{j} b_j x_{ij})]^{y_i} \quad (19)$$

dei valori Y_i (i = 1, 2,..., N) della variabile risposta Y per le N imprese. Dal punto di vista tecnico gli stimatori $\hat{b}_j$ sono forniti dalla soluzione del sistema di equazioni che si ricava ponendo uguali a zero le derivate parziali

$$\frac{\partial \log L\ (Y_1, \ldots, Y_N; b_1, \ldots, b_k)}{\partial\, b_j} \quad (20)$$

del logaritmo dell'espressione della verosimiglianza rispetto ai coefficenti b_j, sui quali l'inferenza può essere effettuata facendo riferimento alla matrice asintotica di varianze e covarianze ottenuta a partire dalle derivate seconde

$$\frac{\partial^2 \log L\ (Y_1, \ldots, Y_N; b_1, \ldots, b_k)}{\partial\, b_j\, \partial\, b_{j'}} \quad (21)$$

rispetto ai coefficienti b_j e $b_{j'}$ (j, j' = 1, ..., k). Naturalmente, il procedimento indicato presuppone di aver specificato la funzione di ripartizione $F_{e/\sigma}$ (.), oppure la funzione di densità di probabilità della componente stocastica (McCullagh and Nelder, 1989); in particolare, se consideriamo la variabille casuale,

$$z = e/\sigma \approx N(0, 1) \quad (22)$$

con distribuzione normale standardizzata allora si ha la funzione di ripartizione

$$F_z(-\sum_{j} b_j x_j) \quad (23)$$

dell'errore, che corrisponde al cosiddetto modello probit e che associa ad ogni valore della combinazione lineare delle variabili esplicative la probabilità condizionata dalle variabili X_j

$$P[Y=0|x_1, x_2, \ldots, x_k] = 1 - F_z(\sum_{j} b_j x_j) \quad (24)$$

di un insuccesso, ovvero quella

$$P[Y=1|x_1, x_2, \ldots, x_k] = F_z(\sum_{j} b_j x_j) \quad (25)$$

di un successo, dato che si possono scrivere le uguaglianze seguenti

$$P[Y=0|x_1, x_2, ..., x_k] = F_z\left(-\sum_j b_j x_j\right) = \int_{-\infty}^{-\sum_j b_j x_j} f(z)\,dz =$$

$$= 1 - P[Y=1|x_1, x_2, ..., x_k] = 1 - \int_{-\sum_j b_j x_j}^{+\infty} f(z)\,dz = 1 - \int_{-\infty}^{\sum_j b_j x_j} f(z)\,dz \qquad (26)$$

grazie alla proprietà di simmetria della distribuzione della variabile $z = e/\sigma$. Infine, vale la pena di sottolineare che se i valori della probabilità condizionata di un successo fossero noti, l'equazione seguente

$$\sum_j b_j x_j = F_z^{-1}(P\ [Y=1|x_1, x_2, ..., x_k]) \qquad (27)$$

fornirebbe il valore della combinazione lineare delle variabili esplicative X_j attraverso la funzione inversa F_z^{-1} della funzione di ripartizione della variabile casuale normale standardizzata.

Invece, se la funzione di ripartizione dell'errore è rappresentata dalla funzione logistica, allora si ha l'equazione

$$F_{e/\sigma}\left(-\sum_j b_j x_j\right) = p[Y=0|x_1, x_2, ..., x_k] =$$
$$= 1/(1 + e^{\sum_j b_j x_j}) \qquad (28)$$

del modello logit, la cui denominazione deriva dal fatto che la combinazione lineare delle variabili esplicative

$$\sum_j b_j x_j = \log \frac{P[Y=1|x_1, x_2, ..., x_k]}{1 - P[Y=1|x_1, x_2, ..., x_k]} \qquad (29)$$

è uguale alla trasformazione logaritmica, detta logit, del rapporto tra probabilità condizionate di successo e di insuccesso.

3. L'interpretazione dei parametri del modello logit

La soluzione del sistema delle equazioni (21) fornisce gli stimatori $\hat{b}_j$ dei coefficienti b_j (j=1, 2,.., k) della combinazione lineare

$$\sum_j b_j x_j = F^{-1}_{e/\sigma}(P[Y=1|x_1, x_2, ...,x_k]) \qquad (30)$$

a partire dalla quale è possibile valutare la probabilità di successo condizionata da particolari valori delle variabili esplicative. Se ora supponiamo che la distribuzione dell'errore sia normale e che quindi la funzione $F_{e/\sigma}(.)$ rappresenti la funzione di ripartizione di una variabile normale standardizzata, allora attraverso l'espressione

$$P[Y=1|x_1, x_2, ...,x_k] = F(\sum_j b_j x_j) = \int_{-\infty}^{\sum_j b_j x_j} f(e/\sigma)\, de/\sigma \qquad (31)$$

che dipende dalla combinazione lineare delle variabili esplicative, è immediato vedere come queste variabili influiscono sulla probabilità di successo. In particolare, per valori $b_j>0$, al crescere di X_j aumenta la probabilità $P\,[Y=1|x_1, ..., x_k]$, e viceversa per valori $b_j<0$. Tuttavia, oltre le considerazioni che si potrebbero fare dal punto di vista teorico e pratico (Cox, 1970) sul modello più adeguato, notiamo che i valori di tali coefficienti non possono essere posti facilmente in relazione diretta con i valori dei rapporti tra probabilità di successo e di insuccesso, a differenza di quanto accade per tali parametri quando la distribuzone della componente stocastica è specificata diversamente attraverso la funzione logistica.

Quando la funzione di ripartizione $F_{e/\sigma}(.)$ della componente stocastica è rappresentata dalla funzione logistica, allora la probabilità di successo è data dalla seguente equazione,

$$P[Y=1|x_1, x_2, ...,x_k]= 1 - F_{e/\sigma}(-\sum_j b_j x_j) = \frac{1}{1+ e^{-\sum_j b_j x_j}} \qquad (32)$$

nella quale compare la combinazione lineare delle variabili esplicative, che può essere scritta

$$\sum_j b_j x_j = \log \frac{P[Y=1|x_1, x_2, ...,x_k]}{1 - P[Y=1|x_1, x_2, ...,x_k]} \qquad (33)$$

come logaritmo del rapporto tra probabilità condizionate di successo e di insuccesso nella forma già indicata (29). Da questa equazione è evidente che se il valore della combinazione lineare è positivo, allora il rapporto tra le due probabilità considerate risulta

maggiore di uno e la probabilità di successo è più grande della probabilità di insuccesso. Inoltre, si vede altrettanto facilmente che a valori più elevati dei coefficienti b_j della combinazione lineare corrispondono valori relativamente più elevati della probabilità di successo, e viceversa. Invece, per individuare le variabili esplicative che contribuiscono a modificare maggiormente la probabilità di successo si può far riferimento alla significatività dei coefficienti (Zelen, 1971), oppure alle variabili che permettono di migliorare l'adattamento del modello, una proprietà che si traduce nella capacità del modello stesso (Amemiya, 1980) di prevedere il successo o l'insuccesso di un'impresa sulla base delle sue caratteristiche individuali.

Proseguendo nell'interpretazione dei parametri del modello logit, consideriamo ora il rapporto

$$R = \frac{P[Y=1|x_1, x_2, ...,x_k]}{1 - P[Y=1|x_1, x_2, ...,x_k]} \tag{34}$$

tra le probabilità di successo e di insuccesso e notiamo che esso varia tra zero ed infinito, cosicchè il logaritmo di tale rapporto

$$-\infty \leq \log R \leq +\infty \tag{35}$$

assume i valori negativi e positivi, compreso il valore zero, che si ha quando la probabilità di successo risulta pari a 0.5 e il rapporto R vale uno. Pertanto, dal valore della combinazione lineare (33) delle variabili esplicative si ha immediatamente un'indicazione sui valori del rapporto in questione e delle probabilità interessanti per l'analisi del fenomeno oggetto di studio.

Passando a considerare in dettaglio il caso in cui le variabili esplicative sono dicotomiche (Cox, 1970), è utile sottolineare il significato dei coefficienti che compaiono nel modello logit. Infatti, se la variabile X_j è dicotomica, il relativo coefficiente b_j è un indice del modo in cui la probabilità di successo varia, quando la variabile X_j passa dalla modalità 0 alla modalità 1. Inoltre, se tutte le variabili esplicative sono dicotomiche, allora quando per tutte queste variabili si considera il valore 0 escluso la prima variabile, alla quale si fa corrispondere per tutte le unità il valore 1, si ha l'equazione

$$\log R = b_1 \tag{36}$$

che consente di ricavare il termine noto della combinazione lineare e che può essere usato per ottenere la probabilità di successo,

$$P[Y=1|X_1=1,\ X_2=0,...,\ X_k=0] = \frac{1}{1+e^{-b_1}} \tag{37}$$

in funzione del coefficiente b_1. Come è facile vedere, da questa equazione si rileva anche che tale coefficiente

$$b_1 = \log \frac{P[Y=1|X_1=1,\ X_2=0,...,\ X_k=0]}{1 - P[Y=1|X_1=1,\ X_2=0,\ ...,\ X_k=0]} \tag{38}$$

è pari al logaritmo del rapporto tra la probabilità di successo e la probabilità di insuccesso per le unità che possiedono la modalità 0 per tutte le variabili esplicative X_j (j=2, 3,...,k). Per quanto riguarda l'intepretazione dei restanti coefficienti b_j (j=2,3,..,k), ciascuno di questi

$$b_j = \log \frac{R_1}{R_0} =$$

$$= \log \frac{P[Y=1|x_1,...,x_{j-1},X_j=1,x_{j+1},...,x_k]/(1-P[Y=1|x_1,...,x_{j-1},X_j=1,x_{j+1},...,x_k])}{P[Y=1|x_1,...,x_{j-1},X_j=0,x_{j+1},...,x_k]/(1-P[Y=1|x_1,...,x_{j-1},X_j=0,x_{j+1},...,x_k])} \tag{39}$$

mostra come, a parità del valore assunto dalle restanti variabili esplicative, varia il logaritmo del quoziente, che ha al numeratore il rapporto tra probabilità di successo e di insuccesso, dato $X_j = 1$, e al denominatore il rapporto tra probabilità di successo e di insuccesso, dato $X_j = 0$. Pertanto, la potenza

$$e^{b_j} = \frac{R_1}{R_0} \tag{40}$$

indica se il rapporto tra probabilità di successo e di insuccesso, dato $X_j = 1$, è favorevole rispetto al caso in cui $X_j = 0$. In particolare, quando il coefficiente b_j assume il valore 0, al variare di X_j, il rapporto tra probabilità di successo e di insuccesso non varia, mentre se tale coefficiente assume valori maggiori (minori) di zero indica che si ha una situazione più (meno) favorevole per il caso $X_j = 1$ rispetto al caso $X_j = 0$.

Oltre che ai fini dell'interpretazione dei risultati dello studio sul fenomeno delle nuove imprese con riferimento all'individuazione delle variabili che appaiono influenti per il successo dell'attività intrapresa, il modello logit può essere utile anche per affrontare un

problema di cui ci occuperemo nelle pagine seguenti e che riguarda l'andamento nel tempo della possibilità di successo che caratterizza le nuove imprese.

4. Un'analisi dinamica del successo delle nuove imprese

L'equazione (32) di regressione logistica della probabilità condizionata di successo delle nuove imprese rappresenta uno strumento utile per ottenere delle informazioni sulle caratteristiche individuali che sembrano influenti nel determinare la sopravvivenza di un'impresa a t anni dalla sua nascita. Un altro aspetto interessante del fenomeno delle nuove imprese è costituito dalla dinamica dei successi e degli insuccessi delle nuove iniziative che si registrano nel corso dei vari anni. Da questo punto di vista, è utile valutare la possibilità di sopravvivenza per le nuove imprese che si sono registrate nei vari anni o fare una previsione dei successi per anni diversi da quelli che hanno consentito di determinare i coefficienti delle variabili esplicative più significative in rapporto al calcolo delle probabilità di successo, purchè si supponga che per questi anni la relazione tra probabilità di successo e variabili esplicative non subisca variazioni. In particolare, il calcolo effettuato retrospettivamente delle probabilità di successo per diversi anni consente di rilevare l'andamento di tali probabilità nel tempo e può essere utilizzato come un indicatore delle difficoltà che le nuove imprese incontrano nel corso dei vari anni e della capacità di superarle.

Per quanto riguarda le previsioni sulla sopravvivenza delle nuove imprese di un certo anno attraverso le probabilità di successo determinate sulla base dei dati rilevati in anni precedenti, esse possono essere utili per diversi scopi, quale ad esempio la messa a punto delle misure da prendere per sostenere in maniera appropriata le nuove iniziative d'impresa. A posteriori, ci possiamo domandare se le probabilità di successo si sono modificate. In particolare, se le variabili esplicative sono di tipo dicotomico, per ogni gruppo di imprese aggregate in base alle loro caratteristiche, si potrebbero confrontare le frequenze attese dei successi (o insuccessi) e le frequenze effettive delle imprese che trascorsi t anni dalla nascita non hanno cessato la loro attività.

Un altro modo, per capire se la situazione in cui le imprese operano cambia nel tempo in rapporto alle difficoltà incontrate ed alla capacità di superarle, consiste nell'effettuare delle stime annuali delle probabilità di successo e nel confrontarle. In questo caso, si può essere interessati ad osservare come tali probabiltà variano per un gruppo di imprese con determinate caratteristiche, oppure per le imprese nel loro complesso. Al fine di rilevare l'andamento delle probabilità di successo per un gruppo di imprese con determinate caratteristiche, si può pensare di confrontare direttamente i valori di probabilità riferiti ad anni diversi. Invece, volendo effettuare dei confronti tra i

quozienti di sopravvivenza per tutte le nuove imprese nate in due anni diversi, occorre considerare che le differenze tra i quozienti dipendono dalle differenze esistenti tra le caratteristiche individuali delle imprese stesse e dalla variazione subita dalla relazione tra caratteristiche individuali e probabilità di successo. Un modo di valutare statisticamente le variazioni della probabilità di successo nel complesso per tutte le imprese è di adattare ai dati di due anni r e s il modello

$$P\,[Y=1|x_1, ..., x_k] = 1/(1 + e^{-\sum_j b_j x_j + \sum_j \delta_j (x_{k+1} \cdot x_j)}) \qquad (41)$$

nel quale figura la variabile X_{k+1}, che rappresenta l'anno di riferimento ($X_{k+1} = 1$ per l'anno r e $X_{k+1} = 0$ per l'anno s) e che permette di introdurre le interazioni ($X_{k+1} * X_j$; j = 1, ..., k) tra l'anno X_{k+1} e le variabili X_j. Infatti, per tale modello i coefficienti b_j risultano uguali a quelli b_{sj} che si otterrebbero considerando il modello logit

$$P_{ss}[Y=1|x_1, ..., x_k] = 1/(1 + e^{-\sum_j b_{sj} x_{sj}}) \qquad (42)$$

per l'anno s; invece i parametri del modello logit

$$P_{rr}[Y=1|x_1, ..., x_k] = 1/(1 + e^{-\sum_j b_{rj} x_{rj}}) \qquad (43)$$

relativo all'anno r possono essere calcolati

$$b_{rj} = b_j + \delta_j \quad (j = 1, ..., k) \qquad (44)$$

come somma tra i coefficienti b_j delle variabili esplicative e quelli δ_j che si riferiscono all'interazione ($X_{k+1} * X_j$). Tenuto conto di quanto è stato detto, attraverso la verifica d'ipotesi sui coefficienti δ_j si può stabilire se i coefficenti b_{rj} sono significativamente diversi dai coefficienti b_{sj} e quindi anche se il modello logit adattato per l'anno r risulta diverso da quello adattato per l'anno s.

Volendo approfondire l'esame delle variazioni della probabilità di successo per le nuove imprese di due anni r e s, ricordiamo che tale probabilità è suscettibile di variazioni nel tempo per due motivi eventuali: i mutamenti che hanno interessato la distribuzione delle imprese secondo le caratteristiche individuali, oppure le variazioni dei parametri che costituiscono un indice della relazione esistente tra le caratteristiche individuali e la probabilità di successo. Per affrontare questo problema, consideriamo che la differenza tra i valori medi

$$\bar{P}_{rr}[Y = 1|x_1,..., x_k] - \bar{P}_{ss}[Y = 1|x_1, ..., x_k] \tag{45}$$

delle probabilità di successo relative ai due anni r e s può essere scritta come una somma di due termini,

$$\begin{aligned}&(\bar{P}_{rr}[Y = 1|x_1,..., x_k] - \bar{P}_{sr}[Y = 1|x_1,..., x_k])+\\ &+ (\bar{P}_{sr}[Y = 1|x_1,..., x_k] - \bar{P}_{ss}[Y = 1|x_1,..., x_k])\end{aligned} \tag{46}$$

dopo aver aggiunto e tolto la quantità $\bar{P}_{sr}[Y=1|x_1, ..., x_k]$ che rappresenta la media delle probabilità di successo

$$P_{sr}[Y=1|x_1, ..., x_k] = 1/(1 + e^{-\sum_j b_{sj}x_{rj}}) \tag{47}$$

calcolate per le imprese dell'anno r attraverso i coefficienti determinati per l'anno s. Allora, il primo termine dipende dalla variazione subita dalla probabilità di successo a causa delle differenze riscontrate nei parametri b_{rj}, b_{sj} relativi ai due anni r e s. Invece, il secondo termine rappresenta la variazione subita dalla probabilità di successo a causa dei mutamenti intervenuti nella distribuzione delle nuove imprese secondo le loro caratteristiche individuali.

Pertanto, la variazione complessiva (45) dei valori medi delle probabilità di successo risulta ripartita in due componenti, che sono interpretabili come effetto parametri ed effetto variabili. Di norma, tali componenti saranno presenti contemporaneamente e contribuiranno a determinare la variazione complessiva della probabilità di successo.

Sulla base della descrizione fornita per i modelli logit considerati è stato possibile effettuare un'applicazione sui dati empirici relativi ad un osservatorio sulle nuove imprese, i cui risultati sono riportati sinteticamente nel paragrafo seguente.

5. Un'applicazione del modello logit

Il modello logit brevemente descritto è stato applicato ai dati di un osservatorio sulle nuove imprese. L'osservatorio in questione è stato realizzato a partire dal 1987 nella provincia di Lucca per l'industria manifatturiera dalla locale Camera di Commercio, che ringraziamo per aver messo a disposizione le informazioni necessarie. I dati utilizzati riguardano le nuove imprese individuali che sono state rilevate nel 1987 e nel 1988 e che a distanza di due anni dalla nascita sono state sottoposte ad un'indagine ulteriore, cosicchè

per tali imprese si dispone di una serie di informazioni individuali, compresa quella che si riferisce alla loro situazione dopo due anni dalla nascita. Per quanto riguarda l'applicazione del modello logit, si è fatto riferimento ad alcune variabili osservate nel corso della prima rilevazione statistica (alla nascita dell'impresa) ed al fatto che fossero risultate ancora in attività al momento della seconda rilevazione, un evento che è stato contraddistinto come un successo. In particolare, per ogni anno ci siamo posti il problema di misurare la relazione statistica tra successo o insuccesso (sopravvivenza o meno dopo due anni dalla nascita) e caratteristiche individuali delle nuove imprese.

Tali caratteristiche o variabili individuali sono state usate, per semplicità e per la numerosità ridotta dei casi, in forma dicotomica, come risulta dall'elenco seguente, in cui compare tra l'altro, anche una denominazione abbreviata delle caratteristiche stesse:

SUCC : Successo o sopravvivenza dell'impresa (1: Si; 0: No);
ANNO : Anno di nascita (1: anno 1987; 0: anno 1988);
SEDE : Comune capoluogo (1: Si; 0: No);
ISAR : Iscrizione all'albo degli artigiani (1: Si; 0: No);
OCCU : Numero di addetti (1: un solo addetto; 0: più addetti);
CARP : Caratteristiche del prodotto (1: prodotto noto; 0: prodotto nuovo);
DESP : Destinazione della produzione (1: il mercato; 0: altro);
DIFF : Difficoltà iniziali ad operare (1: Si; 0: No).

Dai risultati ottenuti (tab.1) in seguito all'adattamento del modello logit per ciascun anno, si nota che per quanto concerne il successo delle nuove imprese nate nel 1987, dal punto di vista statistico esso apparre significativamente dipendente dall'iscrizione all'albo degli artigiani (ISAR), dalla lavorazione di prodotti già noti sul mercato (CARP) e dallo svolgimento dell'attività produttiva direttamente per il mercato (DESP).Verosimilmente, questi risultati risentono del fatto che spesso le nuove imprese individuali iscritte all'albo degli artigiani sono più motivate, determinate ed assistite nello svolgimento della loro attività e che l'immissione sul mercato di prodotti conosciuti facilita le operazioni nella fase del commercio dei prodotti stessi.

Circa i risultati ottenuti dall'analisi dei dati relativi alle nuove imprese nate nell'anno 1988, oltre la conferma della rilevanza statistica dell'iscrizione delle imprese all'albo degli artigiani, appare anche una relazione significativa tra il successo dell'impresa e la piccola dimensione in termini di addetti (OCCU). Pertanto, il fatto di avere la sede nel capoluogo di provincia o di avere incontrato difficoltà nella fase di avvio dell'attività di impresa non sembrano aver avuto un ruolo importante nel determinare il successo delle nuove attività di impresa.

Naturalmente, le considerazioni fatte devono essere prese con una certa cautela, perchè per i due anni prescelti i risultati si basano su un numero limitato di casi (182 per

l'anno 1987 e 189 per l'anno 1988) e dipendono, almeno in parte, dal modo in cui sono stati dicotomizzate le modalità inerenti ad alcune caratteristiche individuali delle imprese. Tuttavia ci sembra che il modello logit preso in esame offra un'interessante chiave di lettura dai dati sull'osservatorio, considerando simultaneamente più variabili.

Passando ad esaminare la variazione subita dalla probabilità di successo dall'anno 1987 all'anno 1988, ci siamo posti il problema di valutare se i due modelli adottati separatamente per i due anni risultavano significativamente diversi dal punto di vista statistico, ovvero se potevano considerarsi differenti per motivi di natura casuale. In proposito, la valutazione è stata effettuata mediante l'adattamento del modello logit (41) ai dati dei due anni, includendo le variabili prescelte, l'anno di riferimento e le interazioni tra questa variabile e le altre variabili.

Tab. 1 - Parametri b_{rj}, b_{sj}, b_j e δ_j dei modelli logit adattati separatamente per gli anni 1987, 1988 e contemporaneamente per entrambi gli anni.

Variabili o intercetta	1987 b_{rj}	1988 b_{sj}	1987-88 b_j	1987-88 δ_j
Intercetta	-0.1685	-0.7066	-0.7066	---
ANNO	---	---	---	+0.5381
SEDE	-0.0307	+0.5260	+0.5260	-0.5567
ISAR	+1.2007*	+1.4709*	+1.4709*	-0.2702
OCCU	+0.4140	+0.9747*	+0.9747*	-0.5608
CARP	+0.8717*	+0.0892	+0.0892	+0.7826
DESP	+1.3430*	-0.1308	-0.1308	+1.4738*
DIFF	-0.5442	+0.1781	+0.1781	-0.7223

* Valori significativi al livello 0.05.

Come si può vedere dai risultati contenuti nelle ultime due colonne della tabella 1, i coefficienti calcolati per le singole variabili coincidono con quelli determinati separatamente per l'anno 1988 ($X_{k+1} = 0$ corrisponde all'anno s). Invece, i coefficienti calcolati per l'anno 1987 ($X_{k+1} = 1$ corrisponde all'anno r) sono pressochè uguali alla somma dei valori dei coefficienti riferiti alle singole variabili e dei corrispondenti coefficienti calcolati per l'interazione δ tra l'anno e le variabili esplicative. In particolare, notiamo che l'intercetta (-0.1685) calcolata per l'anno 1987 risulta dalla somma tra

l'intercetta (-0.7066) ed il coefficiente (0.5381) riferito all'anno, che sono stati calcolati attraverso il modello adattato congiuntamente ai dati dei due anni.

Dai valori dei coefficienti significativamente diversi da zero e contrassegnati con un asterisco, sembra di poter dire che l'iscrizione all'albo degli artigiani (ISAR) e la piccola dimensione in termini del numero di addetti (OCCU) rappresentano due variabili influenti sul successo dell'attività di nuova impresa. Inoltre, sembra di poter dire che i due modelli logit adattati per ciascun anno sono diversi in rapporto alla relazione esistente tra il successo e la destinazione della produzione sul mercato. Infatti la variazione subita dal coefficiente interessato (DESP) e rappresentata dal coefficiente relativo all'interazione con l'anno (1.4738), risulta significativamente diversa da zero. Pertanto, le probabilità di successo riferite ai due anni appaiono significativamente diverse, anche se, come si è già detto, i risultati devono essere interpretati con una certa cautela per diversi motivi.

Per esaminare con un dettaglio ulteriore la variazione delle probabilità di successo che si è registrata passando dall'anno 1987 all'anno 1988, notiamo che la differenza

$$\overline{P}_{87,87}[\,Y = 1|x_1, ..., x_k] - \overline{P}_{88,88}[\,Y = 1|x_1, ..., x_k] = 0.813 - 0.725 = 0.088 \qquad (48)$$

tra le medie delle probabilità di successo calcolate per i due anni coincide con la differenza esistente tra le frequenze relative (0.813 = 148/182) dei successi per il primo anno e la frequenza relativa (0.725 = 137/189) dei successi riferita all'anno 1988. Tale differenza denota una diminuzione dei successi che può essere scomposta nella somma di due termini

$$\begin{aligned}&(\overline{P}_{87,87}[\,Y = 1|x_1, ..., x_k] - \overline{P}_{88,87}[\,Y = 1|x_1, ..., x_k]) + \\ &+(\overline{P}_{88,87}[\,Y = 1|x_1, ..., x_k] - \overline{P}_{88,88}[\,Y = 1|x_1, ..., x_k]) = \\ &= (0.813 - 0.713) + (0.731 - 0.725) = 0.082 + 0.006 \qquad (49)\end{aligned}$$

calcolati in base ai valori medi delle probabilità di successo, che si hanno combinando nella maniera opportuna i dati riferiti a ciascun anno con i coefficienti dei due modelli logit adattati.

Il primo termine rappresenta la variazione registrata nella frequenza relativa dei successi ed è attribuibile alla variazione subita dai parametri del modello logit, passando dal 1987 al 1988; esso denota che una parte consistente (0.082) della variazione complessiva (0.088) della frequenza relativa dei successi è dovuta al cosiddetto effetto parametri. Invece, il secondo termine (0.006) rappresenta l'effetto delle caratteristiche individuali delle imprese sulla variazione della probabilità di successo ed il suo valore è al di sotto del 10% della variazione complessiva; perciò sembra di poter dire che la

componente riferibile alle differenze esistenti in rapporto alle caratteristiche individuali tra le nuove imprese dei due anni considerati non ha avuto un peso determinante sulla variazione della probabilità di successo. Pertanto, la situazione in cui le nuove imprese dell'anno 1987 hanno operato, in rapporto alle difficoltà incontrate ed alla capacità di superarle, appare sostanzialmente più favorevole di quella che ha caratterizzato l'attività delle nuove imprese dell'anno 1988. Inoltre, le differenze sono consistenti e statisticamente significative per quanto riguarda la relazione tra probabilità di successo e caratteristiche individuali.

6. Considerazioni conclusive

L'analisi del fenomeno delle nuove imprese presenta diversi aspetti interessanti, tra i quali lo studio della relazione esistente tra le caratteristiche individuali delle imprese e la probabilità di sopravvivenza dopo un certo numero t di anni dalla nascita. A questo riguardo, si è ritenuto utile mettere a punto e sperimentare un modello logit, considerando il successo delle nuove attività di impresa come una variabile dipendente da alcune variabili esplicative, che sono rappresentate dalle caratteristiche individuali delle imprese. In particolare, ci siamo posti il problema di individuare le variabili statisticamente significative nel determinare il successo delle nuove imprese. Inoltre, abbiamo proposto un modo per valutare la variazione subita nel tempo dai quozienti di sopravvivenza, mettendo in evidenza che tale variazione è attribuibile a due componenti. Una componente è riferibile al fatto che nel tempo la relazione tra le caratteristiche individuali delle nuove imprese ed il loro successo subisce delle variazioni (effetto dei parametri del modello logit sulla probabilità di successo). Invece, l'altra componente è individuabile considerando che anche la distribuzione delle nuove imprese secondo le caratteristiche individuali cambia nel tempo (effetto delle variabili sulla probabilità di successo). L'applicazione del modello logit ai dati di un osservatorio locale sulle nuove imprese manifatturiere ha mostrato l'utilità del metodo proprosto e costituisce esempio stimolante per ulteriori sperimentazioni e verifiche dal punto di vista teorico e pratico.

Riferimenti bibliografici

AMEMIYA T. (1980). Selection of Regressors. *Int. Econ. Review*, 21, pp.331-354.

COX D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

GOLDBERGER A.S.(1964). *Econometric Theory*.J.Wiley, New York.

JUDGE G.G., GRIFFITHS W.E., HILL R.C., LÛTKEPOHL H., LEE T. (1985). *The Theory and Practice of Econometrics*. J.Wiley, New York.

LISSONI F.(1991). La formazione di nuove imprese: due studi sul caso italiano. *Economia e Politica Industriale*, n.71, pp.185-200.

MADDALA G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

McFADDEN D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*. In: ZAREMBKA P. (ed.). *Frontier in Econometrics*. Academic Press, New York, pp.105-142.

McCULLAGH P., NELDER J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

WALKER S.H., DUNCAN D.B. (1967). Estimation of the Probability of an Event as a Function of Several Indipendent Variables. *Biometrika*, 54, pp.167-179.

ZELEN M. (1971). The Analysis of Several 2x2 Contingency Tables. *Biometrika*, 58, pp.129-137.