

## XXV CONFERENZA ITALIANA DI SCIENZE REGIONALI

### LA DOMANDA DI TRASPORTO URBANO: MODELLI LOGIT PER L'INDIVIDUAZIONE DEI FATTORI RILEVANTI

Francesca CONDINO

Istituto di Scienze Neurologiche – Cnr  
Loc. Burga, 87050; Mangone (Cs)

#### SOMMARIO

Il presente lavoro pone l'attenzione sulla possibilità di applicare modelli di tipo aleatorio, e più in particolare di regressione logistica, all'ambito della domanda di trasporto urbano. Con riguardo alla struttura del contributo, si può individuare una prima parte, relativa alla descrizione formale della metodologia proposta, ed una seconda parte riguardante invece un caso concreto di una cittadina calabrese dell'area cosentina, per la quale è stato costruito un modello di emissione per spostamenti di tipo "casa- acquisti". Scopo del lavoro è valutare, non solo la possibilità di applicazione delle modellizzazioni presentate, ma anche il potere interpretativo dei risultati ottenibili tramite questi approcci. A tal proposito, fondamentale è la fase di interpretazione dei parametri, effettuata a partire dai cosiddetti *odds ratios*, la quale consente l'immediata individuazione dei fattori maggiormente influenti sulle scelte degli individui.

## 1. INTRODUZIONE

La domanda di trasporto, è da tempo oggetto di diversi tipi di modellizzazioni, classicamente distinti in analisi di tipo descrittivo, dirette alla semplice descrizione degli spostamenti di utenti e merci, e analisi di tipo aleatorio volte invece a valutare nel complesso le dinamiche dei processi da un punto di vista inferenziale, tramite l'individuazione di legami più o meno stringenti tra due o più variabili. In generale essa è definibile come il numero di viaggiatori, relativamente alla mobilità di persone, e operatori, per il trasporto di merci, che usufruiscono di un sistema di trasporto in un fissato periodo di tempo (Domencich – McFadden, 1975); nasce dall'esigenza degli individui di spostarsi da un luogo all'altro per lo svolgimento delle proprie attività e per questo risulta dall'aggregazione degli spostamenti individuali, frutto delle singole scelte. Dette scelte riguardano diversi livelli: il luogo di residenza e di lavoro, il possesso di un autoveicolo, lo scopo dello spostamento, la fascia oraria, il modo, il percorso, la destinazione. Ognuna di queste categorie è denominata "dimensione di scelta" e nella maggior parte dei casi può considerarsi un insieme finito, in quanto finito è il numero di alternative a disposizione dell'utente in relazione al singolo ambito. L'ipotesi fondamentale su cui si basa il presente lavoro è che le scelte degli individui possano rappresentarsi come variabili aleatorie, in quanto non note con certezza a priori. Di qui la possibilità di applicare modelli di tipo comportamentale che in qualche modo diano indicazioni sulla probabilità del verificarsi di un certo evento piuttosto che un altro. Sebbene gli ambiti di scelta siano certamente interdipendenti fra di loro, si preferisce, per motivi di trattabilità analitica, scomporre la funzione di domanda globale in più sottomodelli interconnessi, ognuno relativo ad un singolo ambito di scelta. L'approccio seguito è dunque quello delle aliquote parziali (Cascetta, 1998).

## 2. L'UTILITA' ALEATORIA

L'ipotesi principale su cui si basano i modelli di utilità casuale deriva direttamente dalla teoria microeconomica del cosiddetto consumatore razionale: si assume cioè di considerare il singolo utente come un decisore razionale, ovvero massimizzatore di utilità (Manski – McFadden, 1981). Una seconda ipotesi riguarda le alternative disponibili: si ipotizza che il numero delle alternative sia finito e pari ad  $m_i$  e che l'utente, al fine di effettuare la propria scelta, consideri tutte le alternative a sua disposizione, ovvero il suo insieme di scelta  $I^i$  (si noti a questo proposito che utenti diversi possono avere differenti insiemi di scelta). Quindi si ipotizza che l'utente  $i$  associ ad ogni alternativa  $j$ , del suo insieme di scelta, un'utilità (o attrattività) percepita  $U_j^i$  e rivolga la sua scelta all'alternativa che massimizza tale utilità.

Supponendo che l'utilità sia esprimibile come funzione di fattori misurabili, quali attributi socioeconomici, attributi di livello o di servizio o attributi del sistema di attività, e che possa essere scomposta in un'utilità sistematica  $V_j^i$ , valore atteso dell'utilità percepita tra tutti gli utenti aventi lo stesso vettore di attributi  $X$ , e in un residuo aleatorio  $\epsilon_j^i$ , che rappresenta lo scostamento dalla media dell'utilità percepita dall'utente  $i$  in relazione all'attributo  $j$ , si avrà:

$$U_j^i = V_j^i + \epsilon_j^i \quad \forall j \in I^i$$

Per quanto riguarda la forma funzionale dell'utilità sistematica, sebbene questa possa naturalmente essere di vario genere, di solito si assume, per ragioni di convenienza analitica, che sia lineare nei parametri  $\alpha_k$ :

$$V_j^i = \sum_k \alpha_k X_{kj}^i$$

dove  $X_{kj}^i$  è il valore del  $k$ -esimo attributo oppure una trasformazione funzionale della variabile originaria e  $\alpha_k$  sono i parametri da stimare.

Implicitamente, vista l'ipotesi di razionalità dell'utente, la scelta dell'alternativa  $j$  comporta che il livello di utilità raggiunto, scegliendo una qualsiasi altra alternativa, risulti inferiore ad  $U_j^i$ . Formalmente ciò può essere espresso nel modo seguente:

$$p^i[j/I^i] = \Pr[V_j^i - V_{j'}^i > \epsilon_{j'}^i - \epsilon_j^i \quad \forall j' \neq j, j' \in I^i]$$

Si nota dunque come la probabilità che venga scelta l'alternativa  $j$  sia condizionata dall'insieme di scelta e dipenda tanto dalle utilità sistematiche di tutte le alternative quanto dalla legge di distribuzione congiunta dei residui aleatori.

Una volta stabilite le variabili da utilizzare e dunque l'espressione dell'utilità sistematica, è necessario definire la funzione di probabilità che caratterizza le alternative di scelta.

Il fattore discriminante dei modelli maggiormente impiegati per riprodurre le scelte degli utenti è la funzione di distribuzione di probabilità congiunta ipotizzata per i residui aleatori. Nel prosieguo si supporrà che i residui seguano una distribuzione di Gumbel e si focalizzerà dunque l'attenzione sul modello Logit Multinomiale.

### 3. IL MODELLO STATISTICO

Il modello Logit Multinomiale è utilizzato in ambiti disparati (economico, medico ecc.) grazie soprattutto alla sua semplicità analitica. Occorre ricordare che esso può essere derivato dalla regressione logistica tramite una particolare trasformazione dell'aspettativa della variabile risposta, detta appunto trasformazione Logit, oppure può essere visto come caso particolare

dei modelli lineari generalizzati (Agresti, 1990). Ma oltre queste argomentazioni, si può trovare ragione di una sua applicazione in questo contesto grazie alla cosiddetta proprietà di indipendenza delle alternative irrilevanti. Si dice che un modello gode di questa proprietà quando il rapporto delle probabilità di due alternative  $j$  e  $j'$  è costante e indipendente da una qualunque terza scelta e dal numero totale  $m$  di alternative considerate. Il modello Logit gode appunto di questa proprietà e quindi consente di inserire ulteriori alternative di scelta senza la necessità di dover nuovamente effettuare la stima dei parametri (Domencich – McFadden, 1975).

Per giungere all'espressione che definisce il modello Logit Multinomiale, si può ipotizzare che i residui aleatori seguano una distribuzione di Gumbel e siano indipendenti e identicamente distribuiti (Oppenheim, 1995).

Per il  $j$ -esimo residuo casuale,  $\varepsilon_j$ , la funzione di ripartizione e quella di densità risultano, rispettivamente, essere:

$$F_{\hat{a}_j}(\hat{a}) = \exp[-\exp(-\hat{a}/\hat{e} - \hat{o})]$$

$$f(\hat{a}_j) = \frac{1}{\hat{e}} \exp\left[-\frac{\hat{a}}{\hat{e}} - \hat{o}\right] \cdot \exp\left[\exp\left(-\frac{\hat{a}}{\hat{e}} - \hat{o}\right)\right]$$

dove  $\hat{e}$  è il parametro di scala della distribuzione e  $\hat{o}$  è la costante di Eulero.

Si ha inoltre:

$$E[\hat{a}_j] = 0 \quad \forall j; \quad \text{Var}[\hat{a}_j] = \frac{\pi^2}{6} \hat{e}^2 \quad \forall j.$$

Dato che l'utilità casuale risulta dalla somma di una variabile aleatoria di Gumbel e di una costante, anch'essa seguirà una distribuzione di Gumbel di parametro  $\hat{e}$ .

Ricordando che, in generale, la distribuzione del massimo di  $n$  variabili aleatorie indipendenti è data dalla produttoria delle singole funzioni di ripartizione, indicato con

$$U_M = \max_j \{U_j\}$$

dopo alcuni passaggi si ottiene:

$$F_{U_M}(u) = \prod_j F_{U_j}(u) = \prod_j \exp\left[-\exp\left(-\frac{u - V_j}{\hat{e}} - \hat{o}\right)\right] = \exp\left[-\exp\left(-\frac{1}{\hat{e}}\left(u + \hat{e}\hat{o} + \hat{e} \ln \sum_j e^{V_j/\hat{e}}\right)\right)\right]$$

e, posto  $\theta = \ln \left( \sum_j e^{\frac{V_j}{\theta}} \right) = V_M$  si ha:

$$F_{U_M}(u) = \exp \left\{ - \exp \left[ - \frac{u - V_M}{\theta} - \theta \right] \right\}$$

che risulta essere ancora una variabile di Gumbel di parametro  $\theta$  e con media pari a:

$$E(U_M) = V_M = \theta \ln \sum_j e^{V_j/\theta}$$

Dalle ipotesi fatte si può giungere ad un'espressione in forma chiusa per la probabilità di scelta dell'alternativa  $j$  (la dimostrazione è riportata in appendice):

$$p[j] = \frac{\exp(V_j/\theta)}{\sum_{h=1}^m \exp(V_h/\theta)}$$

che definisce appunto il modello Logit Multinomiale.

Alcune osservazioni possono essere fatte analizzando la struttura dell'espressione che definisce la varianza dei residui. E' possibile notare, infatti, come la variabilità dei residui dipenda dal parametro  $\theta$  della distribuzione: maggiore è il parametro e maggiore sarà la varianza. Allo stesso tempo però, le probabilità di ogni alternativa tenderanno ad assumere lo stesso valore, pari ad  $1/m$ , in quanto al crescere di  $\theta$  gli esponenti  $V_j/\theta$  tenderanno tutti al valore zero. Di contro, minore sarà la variabilità dei residui, maggiore sarà la probabilità di scelta dell'alternativa a cui è associata l'utilità massima.

#### 4. LA STIMA E LA VERIFICA DELLE IPOTESI

Il metodo generalmente utilizzato per la stima dei modelli Logit è quello della massima verosimiglianza. Tale metodo garantisce infatti che gli stimatori ottenuti siano asintoticamente normali, asintoticamente corretti e pienamente efficienti, ovvero appartengano alla classe degli stimatori BAN (Best Asymptotically Normal).

Si consideri la seguente variabile:

$$Y_{ij} = \begin{cases} 1 & \text{se l'individuo sceglie l'alternativa } j \\ 0 & \text{altrimenti} \end{cases}$$

Allora la funzione di log-verosimiglianza, ipotizzando un campionamento casuale semplice è data dal prodotto delle probabilità; quindi, per la funzione di probabilità prescelta si ha:

$$\ln L(\beta) = \sum_{i=1}^n \sum_{j=1}^m Y_{ij} (\beta' X_{ij}) - \sum_{i=1}^n \ln \left[ \sum_{t \in I_i} \exp(\beta' X_{it}) \right]$$

dove è stato posto  $\beta_k = \frac{\alpha_k}{\theta}$ . I parametri di tale modello infatti, non possono essere stimati

separatamente ma sarà possibile ottenere solo le stime dei suddetti rapporti. Derivando e ponendo pari a zero l'espressione ricavata, non è possibile ottenere una soluzione in forma chiusa, in quanto l'equazione non è lineare rispetto al parametro  $\hat{\alpha}$ . E' necessario perciò ricorrere a metodi di tipo iterativo, che forniscono soluzioni vicine all'ottimo. Un metodo molto utilizzato per il calcolo dello stimatore di massima verosimiglianza è il metodo di Newton – Raphson, basato sullo sviluppo in serie di Taylor della funzione obiettivo.

Per la sua implementazione è necessaria un'ipotesi iniziale sul valore del parametro che massimizza la funzione. Inoltre è necessario che siano calcolabili le derivate prime e seconde, al fine di poter costruire il polinomio di Taylor di secondo ordine. Da esso può essere ricavato il valore di  $\hat{\alpha}$  che rende massima la funzione. In definitiva, il valore del parametro alla t-esima iterazione sarà dato da:

$$\mathbf{b}^{(t+1)} = \mathbf{b}^{(t)} - (\mathbf{H}^{(t)})^{-1} \mathbf{q}^{(t)}.$$

dove  $\mathbf{H}^{(t)}$  è la matrice hessiana associata alla funzione di log-verosimiglianza che si dimostra essere definita negativa. Ciò garantisce l'esistenza di un solo punto di massimo e dunque di un solo vettore di parametri in corrispondenza del quale sarà massima la funzione. La matrice di varianze e covarianze dei parametri coincide con l'inversa della matrice di informazione di Fisher e si dimostra essere pari a

$$\hat{\Sigma}_b = \left[ \sum_{i=1}^n X_i [E(Y_i - P_i)(Y_i - P_i)'] X_i' \right]^{-1}.$$

dove  $P_i$  indica la probabilità associata all'individuo  $i$ .

E' da notare come l'espressione ottenuta non dipenda dalla variabile aleatoria  $Y$ , ma solo dalla sua aspettativa, e ciò in quanto la funzione di probabilità utilizzata appartiene alla classe esponenziale. Va comunque osservato che, quando viene impiegato un metodo iterativo come quello appena trattato, lo stimatore della matrice di covarianza risulta essere distorto, fatta eccezione per quello ottenuto alla prima iterazione. Nelle successive iterazioni, infatti, il

calcolo degli elementi della matrice viene basato sui valori dello stimatore, detti “valori inclusivi”, ottenuti nello stadio precedente.

Una volta conclusa la fase di stima dei parametri, il modello risulta completamente specificato. A questo punto è necessario valutare la veridicità di alcune particolari ipotesi sui parametri del modello stimato, ricorrendo ad opportuni test statistici, quali ad esempio i cosiddetti test di significatività (Landenna – Marasini – Ferrari, 1998).

Per questi tipi di modelli si può pensare di considerare un’ipotesi, abbastanza generale, specificata da una forma lineare:

$$H_0 : \mathbf{Q}'\boldsymbol{\beta} = c$$

dove  $\mathbf{Q}$  è un vettore a  $K$  componenti e  $c$  è una costante.

Supponendo di voler utilizzare lo stimatore di massima verosimiglianza per testare l’ipotesi, potrà essere assunta la seguente statistica test:

$$Z = \frac{\mathbf{Q}'\boldsymbol{\beta}^{ML} - c}{\sqrt{\mathbf{Q}'\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\beta}^{ML}}\mathbf{Q}}}$$

dove  $\tilde{\boldsymbol{\Omega}}_{\boldsymbol{\beta}^{ML}}$  è una stima della matrice di varianze e covarianze dello stimatore di massima verosimiglianza.

La statistica test così costruita segue la distribuzione di una variabile casuale t- Student con  $(n-k)$  gradi di libertà. Fissato il livello di significatività, si potrà allora determinare la regione di accettazione tramite i percentili di una variabile casuale t- Student.

In alcuni casi è possibile, anziché fare riferimento alla distribuzione t-Student, considerare l’approssimazione della statistica test ad una variabile casuale avente distribuzione Normale standard. Per un numero di gradi di libertà abbastanza elevato, assumere quest’ultima approssimazione, anziché la precedente, non comporta a livello pratico, molta differenza. E’ noto, infatti, come la Normale standardizzata, al crescere dei gradi di libertà, rappresenti la distribuzione limite della variabile t-Student.

E’ interessante valutare un caso particolare dell’ipotesi nulla precedentemente formulata, considerando il vettore  $\mathbf{Q}$  come un vettore avente  $K-1$  elementi nulli ed un solo elemento, precisamente l’ $l$ -esimo, pari ad 1. Se si pone inoltre  $c=0$  si perviene alla verifica dell’ipotesi che un certo parametro del modello sia pari a zero, ovvero  $H_0 : \beta_l = 0$  contro l’ipotesi alternativa  $H_1 : \beta_l \neq 0$ .

La statistica test assumerà la seguente forma:

$$t = \frac{\beta_l^{ML}}{\text{Var}[\beta_l^{ML}]^{1/2}}$$

essendo  $\text{Var}[\beta_1^{\text{ML}}]^{1/2}$  l'elemento dell'1-esima riga e dell'1-esima colonna della matrice  $\tilde{O}_{\mathbf{b}^{\text{ML}}}$ .

Se  $H_0$  è vera, la statistica  $t$  è distribuita come una  $t$ -Student con un numero di gradi di libertà pari alla dimensione  $n$  del campione meno il numero dei parametri stimati  $K$ .

Un secondo test sul valore dei singoli parametri è quello che verifica l'ipotesi di uguaglianza tra due parametri della distribuzione. In questo caso la statistica test utilizzata è la seguente:

$$t = \frac{\beta_1^{\text{ML}} - \beta_j^{\text{ML}}}{\left[ \text{Var}(\beta_1^{\text{ML}}) + \text{Var}(\beta_j^{\text{ML}}) - 2\text{Cov}(\beta_1^{\text{ML}}, \beta_j^{\text{ML}}) \right]^{1/2}}.$$

Ancora una volta la variabile casuale  $t$ , sotto  $H_0$ , si distribuisce secondo una variabile  $t$ -Student.

Può essere interessante valutare l'ipotesi che ogni parametro assuma un certo valore. Ciò significa costruire un test che interessi simultaneamente tutti i parametri della distribuzione. L'ipotesi nulla sarà  $H_0 : \mathbf{b} = \mathbf{b}^*$  e la statistica test che si utilizza generalmente per testare questa ipotesi è la seguente:

$$\div^2(\mathbf{b}^*) = (\mathbf{b}^{\text{ML}} - \mathbf{b}^*)' \hat{O}_a^{-1} (\mathbf{b}^{\text{ML}} - \mathbf{b}^*).$$

La statistica così costruita può assumere solo valori maggiori di zero e quindi la sua distribuzione sarà certamente definita solo per valori positivi. Di più, per il teorema di Cochran (1934), si può affermare che la variabile casuale  $\div^2(\mathbf{b}^*)$  segue una distribuzione Chi-Quadrato con un numero di gradi di libertà pari alle dimensioni del vettore di parametri.

Altra categoria di test molto utilizzati sono i test del rapporto di verosimiglianza, costruiti secondo un criterio proposto da J. Neyman e E. Pearson nel 1928. Si supponga, ad esempio, di voler verificare l'ipotesi  $H_0 : \mathbf{b} = \mathbf{b}^*$ . Usando la statistica "rapporto di verosimiglianza", avremo che la variabile casuale sarà data da:

$$X^2 = -2 \ln \frac{L(\mathbf{b}^*)}{L(\mathbf{b}^{\text{ML}})} = -2 [\ln L(\mathbf{b}^*) - \ln L(\mathbf{b}^{\text{ML}})].$$

Considerando il teorema di Wilks, si può affermare che la variabile casuale  $X^2$  è asintoticamente distribuita secondo una variabile Chi-Quadrato con  $K$  gradi di libertà (dove con  $K$  si indica la dimensione del vettore  $\mathbf{b}$ ).

E' interessante valutare un caso particolare dell'ipotesi precedente: l'ipotesi  $H_0 : \hat{\mathbf{a}} = \mathbf{0}$ .

La statistica test sarà allora distribuita come una Chi-Quadrato, avente però  $K$  gradi di libertà. Inoltre supponendo un modello di tipo Logit, si verifica implicitamente l'ipotesi che le alternative di scelta siano equiprobabili. Infatti se si ipotizza che il modello abbia tutti



parametri nulli si ottiene un'utilità sistematica sempre nulla e dunque la probabilità risultante sarà, per ogni alternativa,  $p[j] = 1/J$ .

Infine, un'altra ipotesi frequentemente testata riguarda le variabili specifiche contenute nel modello. Più precisamente, si ipotizza che gli unici parametri del modello diversi da zero siano quelli corrispondenti agli attributi specifici dell'alternativa.

Sia in questo caso che nel precedente, il test porta generalmente ad un rifiuto dell'ipotesi nulla e ciò in quanto, anche intuitivamente, un modello contenente anche altre variabili oltre a quelle specifiche riesce a descrivere meglio la realtà rispetto ad un modello privo di variabili esplicative.

Nel caso in cui siano presenti variabili risposta di tipo qualitativo, la bontà di adattamento del modello può essere valutata sia in termini di capacità del modello di prevedere gli scenari futuri, sia in termini di differenza tra la probabilità calcolata utilizzando il modello stimato, e la frequenza osservata.

Ciò fa pensare ad un'analogia con la regressione lineare, in cui la misura più utilizzata per valutare la bontà di adattamento risulta essere il coefficiente di correlazione multipla, basato appunto sugli scarti tra valori osservati e valori teorici.

Si può, quindi, determinare una somma pesata dei quadrati degli scarti tra frequenza osservata e probabilità calcolata tramite il modello stimato:

$$S(\hat{\mathbf{a}}) = \frac{\sum_i \sum_j (f_{ij} - P_{ij}(\hat{\mathbf{a}}))^2}{P_{ij}^*}$$

dove i pesi,  $P_{ij}^*$ , sono dati dalle reali probabilità di scelta.

A partire da questa quantità si può definire la seguente misura (McFadden, 1975):

$$R^2 = 1 - \frac{S(\hat{\mathbf{a}})}{S(\bar{\mathbf{a}})}$$

con

$\hat{\mathbf{a}}$  = stimatore di massima verosimiglianza

$\bar{\mathbf{a}}$  = vettore di parametri tutti nulli.

Tale misura risulta essere normalizzata, ovvero assume valori nell'intervallo [0,1] e quindi consente di effettuare confronti tra la bontà di adattamento di modelli diversi. Di contro, però, l'indice non gode di proprietà statistiche desiderabili per piccoli campioni e appare molto sensibile agli errori di specificazione del modello.

Una misura più soddisfacente, suggerita da McFadden nel 1974, è ottenuta a partire dalla funzione di verosimiglianza:

$$\tilde{n}^2 = 1 - \frac{\ln L(\hat{\mathbf{a}}^{\text{ML}})}{\ln L(\mathbf{0})}$$

dove  $L(\hat{\mathbf{a}}^{\text{ML}})$  è il massimo della funzione di verosimiglianza, mentre  $L(\mathbf{0})$  è il valore della funzione di verosimiglianza calcolato per un vettore di parametri nullo.

Quest'indice assume valore pari ad uno se il modello ha una perfetta capacità di riprodurre la realtà, ed assume valore 0 nel caso opposto. Infatti, nel peggiore dei casi, le variabili introdotte non hanno alcuna capacità di spiegare la realtà e quindi, adottando un modello contenente queste variabili esplicative, si otterrebbero le stesse probabilità di scelta che si ottengono assumendo un modello in cui non sia presente nessuna di queste variabili. Algebricamente, il rapporto che compare nella relazione assumerà valore unitario, portando all'azzeramento dell'indice. Al contrario, quando il modello stimato è il migliore possibile (ovvero nel caso in cui il modello è deterministico, nel senso che descrive perfettamente la realtà) allora la probabilità di scelta dell'alternativa è pari ad uno e quindi la funzione di log-verosimiglianza sarà

$$L(\hat{\mathbf{a}}^{\text{ML}}) = \sum_i \sum_j Y_{ij} \ln P_{ij} = \sum_i \sum_j Y_{ij} \ln 1 = 0$$

e, di conseguenza l'indice assumerà valore uno (Oppenheim, 1995). Va inoltre sottolineato che si tratta di un indice normalizzato e dunque è possibile, tramite esso, effettuare dei confronti tra diversi modelli.

Un'altra misura spesso utilizzata, anche se non troppo accurata, è la cosiddetta “% right”, detta anche *sample reconstitution test*. Quest'indice è dato dalla percentuale delle osservazioni campionarie per cui l'alternativa effettivamente scelta coincide con quella che, dal modello, risulta l'alternativa di massima utilità. Questa misura fornisce quindi la percentuale di casi esattamente previsti dal modello (Akiva B., Lerman S. R., 1985).

## 5. L'INTERPRETAZIONE DEI PARAMETRI

Nella regressione logistica l'interpretazione dei parametri costituisce una fase fondamentale dell'analisi. E' possibile infatti, a partire dalle stime ottenute, trarre alcune interessanti conclusioni riguardo la probabilità che la variabile risposta assuma un valore piuttosto che un altro. Per meglio chiarire questo concetto, si consideri una certa variabile esplicativa X, di

tipo dicotomico, che assume valore 1 se la caratteristica X è presente e valore 0 se, invece è assente. Sia  $\beta$  il parametro associato a tale variabile.

Si può dimostrare che  $\beta$  corrisponde alla variazione del *logit* di  $P(Y)$ , che si ha quando X passa dal valore 0 al valore 1:

$$\text{logit}[\Pr(Y = 1|X = 1)] - \text{logit}[\Pr(Y = 1|X = 0)] = \hat{\alpha}.$$

Vista la definizione di *logit* quale logaritmo naturale del rapporto tra la probabilità di successo di Y (dove per successo s'intende l'evento  $Y=1$ ) e la probabilità di insuccesso, dato un certo valore di X, la precedente relazione può essere scritta come segue:

$$\hat{\alpha} = \ln \left[ \frac{\Pr(Y = 1|X = 1)}{\Pr(Y = 0|X = 1)} \right] - \ln \left[ \frac{\Pr(Y = 1|X = 0)}{\Pr(Y = 0|X = 0)} \right]$$

Gli argomenti dei logaritmi su scritti vengono denominati *odds* e rappresentano il numero di successi per ogni insuccesso, dato il valore di X. Si dice che l'*odds* è pari al rischio che Y assuma valore 1, fissato il valore di X.

Dalla precedente relazione si ottiene

$$\hat{\alpha} = \ln \frac{\Pr[Y = 1|X = 1] \cdot \Pr[Y = 0|X = 0]}{\Pr[Y = 0|X = 1] \cdot \Pr[Y = 1|X = 0]}.$$

In questo caso, l'argomento del logaritmo è pari al rapporto degli *odds*, calcolati rispettivamente per  $X=1$  e  $X=0$ . Tale rapporto viene detto *odds ratio* o *rapporto crociato* e non è altro che la variazione dell'*odds* dovuta alla variazione della variabile X.

In altre parole l'*odds ratio* permette di valutare di quanto cresce o decresce il “rischio” che la variabile risposta assuma valore 1 anziché valore 0, a seguito della variazione del valore assunto dal regressore (Fabbris, 1997).

L'estensione dell'*odds ratio* al caso in cui il regressore sia quantitativo discreto è immediata: in questo caso l'*odds ratio* rappresenta la variazione del rischio dovuta all'incremento unitario della variabile X:

$$e^{\hat{\alpha}} = \frac{\Pr[Y = 1|X = x + 1] \cdot \Pr[Y = 0|X = x]}{\Pr[Y = 0|X = x + 1] \cdot \Pr[Y = 1|X = x]}.$$

Si può concludere dunque che un valore negativo del parametro, implicando un valore dell'*odds ratio* compreso tra zero ed uno, porta alla seguente disuguaglianza:

$$\frac{\Pr[Y = 1|X = x + 1]}{\Pr[Y = 0|X = x + 1]} < \frac{\Pr[Y = 1|X = x]}{\Pr[Y = 0|X = x]}$$

ovvero il “rischio” che Y sia presente decresce al crescere di X. Il contrario avviene se  $\beta$  è positivo. In questo caso, infatti, il rischio è crescente rispetto ad X.

## 6. IL MODELLO DI EMISSIONE: UN CASO REALE

Verrà ora illustrata l'applicazione della metodologia fin qui descritta ad un caso reale. I dati utilizzati sono riferiti ad un'indagine domiciliare realizzata nei primi mesi del 2001 presso una cittadina calabrese di medie dimensioni della provincia di Cosenza. Il piano di campionamento impiegato per la raccolta dei dati è di tipo casuale semplice. Agli intervistati sono stati somministrati due diversi questionari: il primo relativo all'intero nucleo familiare, il secondo diretto ad ogni singolo individuo, purché di età superiore ai quattordici anni. La popolazione campionata risulta pari a 1207 abitanti, su un totale di 12.681 residenti, per un tasso di campionamento quasi pari al 10%. L'età media degli individui intervistati è di circa 37 anni, mentre il numero medio di componenti per famiglia risulta essere di 3,61. La presenza maschile di cui si compone il campione è pari al 51,1%, di cui il 57,2% in possesso di patente. Quest'ultima percentuale si abbassa invece ad un livello pari al 42,8% quando si fa riferimento alle donne.

Il modello proposto è un modello di emissione di tipo Logit avente come motivo in origine “casa” e come motivo in destinazione “acquisti”.<sup>1</sup> E' da sottolineare il fatto che tali spostamenti, a differenza di altri, quali ad esempio gli spostamenti per lavoro, non sono di natura sistematica e dunque ben si prestano a modellizzazioni di tipo aleatorio.

In questo contesto,  $j$  può rappresentare due sole alternative (indicate brevemente con Trip e NoTrip) e più precisamente può essere scritta come segue:

$$j = \begin{cases} 1 & \text{se viene effettuato almeno uno spostamento "Casa - Acquisti" (Trip)} \\ 0 & \text{altrimenti (NoTrip)} \end{cases}$$

Si avranno dunque le espressioni delle due utilità sistematiche, ciascuna combinazione delle variabili inserite nel modello tramite i parametri. Una volta determinate le utilità sistematiche si potrà allora calcolare la probabilità che l'individuo effettui uno spostamento “casa – acquisti” come segue:

---

<sup>1</sup> La destinazione degli spostamenti a cui ci si riferisce è ipotizzata essere quella primaria, dove per destinazione primaria s'intende quella dove si svolge l'attività considerata più importante.

$$P(\text{Trip}) = \frac{\exp(V(\text{Trip})/\hat{\epsilon})}{\exp(V(\text{Trip})/\hat{\epsilon}) + \exp(V(\text{NoTrip})/\hat{\epsilon})}$$

Nella tabella seguente vengono riportate le variabili inserite nel modello con le relative stime ottenute tramite la procedura descritta in precedenza. Contestualmente si riportano i risultati dei test t-Student sui singoli parametri associati alle suddette variabili.

*Tabella 1* Stime e valore del test t-Student per i parametri del modello grezzo

Variabili	Stime	Deviazione standard	Test t	p-value
Utilità sistematica dell'alternativa "Trip"				
Età	-0,004036	0,00618	-0,653	0.257
Sesso	-1,132	0,2433	<b>4,65</b>	<b>&lt;0,001</b>
Veicoli	0,2683	0,1178	<b>2,28</b>	<b>0,012</b>
Occupato	-0,709	0,2814	<b>-2,52</b>	<b>0,006</b>
Responsabile	0,1865	0,1056	<b>6,17</b>	<b>0,039</b>
Bambini	0,02807	0,09372	0,266	0,382
N° di Componenti	-0,07363	0,3022	-0,786	0,404
Possesso di patente	0,1405	0,2110	0,666	0,253
Utilità sistematica dell'alternativa "NoTrip"				
Altri spostamenti	0,4910	0,1209	<b>4,06</b>	<b>&lt;0,001</b>
Intercetta	2,170	0,4466	<b>4,86</b>	<b>&lt;0,001</b>

Sia il test del rapporto di verosimiglianza con ipotesi nulla  $H_0 = \hat{\mathbf{a}} = \mathbf{0}$  che il test con ipotesi nulla  $H_0 : \hat{\mathbf{a}} = \hat{\mathbf{a}}_c$ , (ovvero si suppone che gli unici parametri diversi da zero siano quelli associati alle costanti specifiche) conducono a valori della statistica test ( $X^2=826,052$ ;  $X^2=176,504$  rispettivamente) corrispondenti ad un p-value inferiore a 0,001. Dunque in entrambi i casi l'ipotesi è nettamente rifiutata. Inoltre, come si può riscontrare dai dati riportati in tabella, ad alcune variabili corrisponde un p-value superiore alla soglia prefissata di 0,05 e per esse non si può rifiutare l'ipotesi nulla. Queste variabili verranno dunque escluse dal modello in quanto risultano essere poco informative.

Utilizzando i valori della funzione di log-verosimiglianza in corrispondenza del vettore di stimatori calcolato e in corrispondenza di un vettore nullo, è possibile calcolare l'indice  $\tilde{n}^2$ , che nello specifico risulta pari a 0,494. Tale valore risulta essere abbastanza alto per modelli di questo tipo (per un raffronto si veda lo studio effettuato da Biggiero (1991) per la città di Genova, oppure il modello di emissione proposto da Festa et Al. (2000)).

A questo punto, secondo una procedura di tipo backward si può procedere nell'analisi stimando un secondo modello, questa volta privo delle variabili poco significative.

I test chi-quadrato sul vettore di parametri portano ancora a valori dei p-value inferiori a 0,001 ( $X^2=824,444$ ;  $X^2=174,896$ ) e dunque al rifiuto delle ipotesi nulle. L'indice  $\tilde{n}^2$  risulta pari a 0,493; ovviamente la variazione di quest'indice rispetto al valore precedente è dovuta alla riduzione del numero di repressori inseriti nel modello, ma va comunque sottolineato come tale calo sia di dimensioni contenute, a riprova di una buona stabilità del modello in questione. Nella prossima tabella sono riportate le stime dei parametri e le relative deviazioni standard ottenute. Inoltre viene mostrato il valore del test t- Student, effettuato per ogni parametro, con ipotesi  $H_0 : \hat{\alpha}_l = 0$ ;  $l = 1, \dots, 6$ .

*Tabella 2* Stime e valore del test t-Student per i parametri del modello finale

Variabili	Stime	Deviazione standard	I. C. (95%)	Test t	p-value
Utilità sistematica della prima alternativa					
Sesso	-1,115	0,2313	-1,57 ; -0,66	-4,82	<0,001
Veicoli	0,2415	0,1053	0,04 ; 0,45	2,29	0,011
Occupato	-0,6744	0,2761	-1,22 ; -0,13	-2,44	0,007
Responsabile	1,819	0,2311	1,37 ; 2,27	7,87	<0,001
Utilità sistematica della seconda alternativa					
Altri spostamenti	0,4556	0,1165	0,23 ; 0,68	3,91	<0,001
Intercetta	2,468	0,2846	1,91 ; 3,03	8,67	<0,001

In questo caso, a differenza del precedente, il test t-Student con  $\alpha=0,05$ , porta a rifiutare l'ipotesi nulla per tutti i parametri considerati; ciò significa che si è giunti alla specificazione finale del modello in cui ciascuna variabile considerata apporta un valido contributo alla descrizione del fenomeno in esame.

Gli intervalli di confidenza ottenuti tramite il metodo della Quantità Pivot e riportati in tabella, mostrano in altri termini la validità delle variabili considerate. Infatti, se il parametro assume valore nullo, questo indica che una variazione del regressore associato non coincide con la variazione della variabile risposta. In altre parole il regressore non è in grado di cogliere le variazioni della variabile dipendente. Come si può osservare gli intervalli sono abbastanza piccoli e comunque nessuno di essi contiene lo zero.

Infine è stato calcolato il cosiddetto *sample reconstitution test*. Questa misura permette di affermare che il modello, così come è stato costruito, ricostruisce l'84,26% del campione, ovvero le osservazioni campionarie per cui l'alternativa effettivamente scelta coincide con quella prevista dal modello sono 1017 su un totale di 1207.

Si è visto come il procedimento di stima abbia portato ad ottenere due valori negativi per i parametri considerati. In particolare si tratta di due parametri associati entrambi a variabili di tipo dicotomico: “sesso” e “occupato”.

Per quanto riguarda la variabile “sesso” si fa notare che questa è stata definita nel modo seguente:

$$\text{"sesso"} = \begin{cases} 1 & \text{se l'individuo è di sesso maschile} \\ 0 & \text{se l'individuo è di sesso femminile} \end{cases}.$$

La stima ottenuta in corrispondenza di questa variabile porta ad un valore dell’odds ratio pari a 0,327. Il “rischio” che si effettui uno spostamento dato che l’individuo è di sesso maschile è quindi minore del “rischio” che si effettui uno spostamento dato che l’individuo è di sesso femminile. Si può dunque affermare che si ha una tendenza a che siano le donne ad effettuare spostamenti per il motivo “casa – acquisti”.

Per la variabile “occupato” si effettua un analogo tipo di analisi. L’odds ratio è pari a 0,509 e quindi, anche in questo caso il “rischio” di effettuare lo spostamento si abbassa se il soggetto considerato è un occupato anziché un disoccupato. Ciò conferma l’ipotesi che il fattore occupazione sia un fattore deterrente per lo spostamento. In questo caso infatti si può presumere che, se in famiglia vi è un individuo libero da vincoli di lavoro, si tende a demandare a quest’ultimo il compito di effettuare spostamenti per il soddisfacimento dei bisogni della famiglia stessa.

Un risultato del genere è stato ottenuto anche da altri autori. In particolare, nel lavoro già citato di Luigi Biggiero (1991), viene proposta una stima dei parametri per i diversi motivi che generano gli spostamenti: per il motivo acquisti il parametro relativo alla variabile “occupazione” assume segno negativo. Solo nel caso in cui lo spostamento considerato sia effettuato per il motivo “affari professionali” il fattore occupazione risulta incentivante per lo spostamento. In questo caso infatti si può presumere che sia l’individuo occupato quello più propenso ad effettuare spostamenti per motivi di affari.

Un’altra variabile dicotomica utilizzata nel modello è la variabile “responsabile”. In questo caso il valore del parametro è positivo e conduce ad un odds ratio pari 6,156. Allora, non solo si può affermare che la probabilità relativa che venga effettuato lo spostamento è maggiore nel caso in cui “responsabile=1” anziché nel caso in cui “responsabile=0”, ma si può anche dire che il “rischio” di effettuare lo spostamento quando l’individuo è il capofamiglia o il coniuge è circa sei volte maggiore rispetto al “rischio” di effettuare lo spostamento se l’individuo considerato è un figlio o un altro parente.

Per quanto riguarda invece le variabili “Veicoli” e “Altri Spostamenti”, di tipo quantitativo, l’odds ratio è in entrambi i casi maggiore di uno. Va comunque sottolineato che la variabile

“Altri Spostamenti” compare nell’utilità sistematica dell’alternativa “NoTrip”, e quindi deve essere interpretata in senso opposto a quanto fatto per le altre variabili.

Per la variabile “Altri Spostamenti”, quindi, il “rischio di non fare lo spostamento” è crescente rispetto al numero di spostamenti già effettuati (odds ratio=1,577). Ciò vuol dire che si ha la propensione a non effettuare altri spostamenti al crescere del numero di spostamenti già effettuati.

Infine, per quanto riguarda il numero di veicoli, dall’odds ratio si deduce che la probabilità relativa di fare uno spostamento cresce al crescere del numero di veicoli. Di più si può affermare che il “rischio” è quasi il doppio se il numero dei veicoli aumenta di 3 unità; infatti se si pone una variazione pari a 2 e una variazione del numero di veicoli pari a c si ottiene:

$$2 = \exp(\hat{c}_3) \Rightarrow c \cong 3$$

dove  $\hat{c}_3$  è appunto il parametro associato alla variabile “Veicoli”.

## 7. CONCLUSIONI

Nel presente lavoro si è analizzato un aspetto specifico della mobilità: la domanda di trasporto. In particolare si è cercato di valutare la possibilità di applicare modelli di tipo Logit a tale fenomeno. Nella prima parte è stata presentata la metodologia basata su un approccio di tipo comportamentale, descrivendo i modelli di utilità aleatoria e fornendo la derivazione analitica dei modelli di tipo Logit per l’ambito specifico. Sono stati successivamente considerati i diversi aspetti della costruzione, della stima e della verifica delle ipotesi statistiche per i modelli proposti. Nell’ultima parte si è poi analizzato un caso di studio, valutando nello specifico il fenomeno della generazione degli spostamenti tramite un modello di emissione di tipo Logit Binomiale.

Ai fini della scelta delle variabili si è tenuto conto di modelli già esistenti, nonché di tecniche statistiche dirette a fornire indicazioni sul tipo di relazione intercorrente tra la variabile oggetto di studio e la variabile indipendente di volta in volta considerata.

Tramite test statistici si è potuto valutare la bontà complessiva del modello proposto nonché il contributo di ciascuna variabile alla modellizzazione del fenomeno in esame. A seguito di queste considerazioni si è giunti alla specificazione di un modello contenente un numero di regressori minore, notando comunque una buona stabilità del modello rispetto alla riduzione delle variabili esplicative.

Si è visto poi come sia possibile valutare quali siano i fattori che contribuiscano ad indirizzare la scelta degli individui in un senso piuttosto che in un altro.

In conclusione si può affermare che l’analisi di regressione logistica porta a risultati abbastanza soddisfacenti. Non solo l’applicazione di un tale tipo di modello risulta piuttosto



semplice, ma i risultati ottenuti sono facilmente interpretabili e leggibili in chiave di “propensione al rischio”.

## 8. APPENDICE

L'espressione fornita nel paragrafo 3 circa la probabilità di scelta dell'alternativa j, può essere ottenuta come segue:

$$\begin{aligned} p(j) &= \Pr(U_h < U_j) = \Pr[V_h + \hat{a}_h < V_j + \hat{a}_j] = \\ &= \Pr[\hat{a}_h < \hat{a}_j + V_j - V_h] \quad \forall h \neq j. \end{aligned}$$

Dato che le variabili sono indipendenti, questa probabilità si può ottenere come la convoluzione delle funzioni di densità:

$$\Rightarrow \int_{-\infty}^{+\infty} \prod_{h \neq j} F(\hat{a}_j + V_j - V_h) \cdot f(\hat{a}_j) d\hat{a}_j.$$

Il termine all'interno dell'integrale può essere espresso nel modo seguente:

$$\begin{aligned} \prod_{h \neq j} e^{-e^{\frac{-\hat{a}_j - V_j + V_h}{\hat{e}} - \hat{o}}} \cdot \frac{1}{\hat{e}} e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}} \cdot e^{-e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}}} = \\ = e^{\sum_{h \neq j} \frac{-\hat{a}_j - V_j + V_h}{\hat{e}} - \hat{o}} \cdot \frac{1}{\hat{e}} e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}} \cdot e^{-e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}}} = \frac{1}{\hat{e}} e^{-e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}} \left( 1 + \sum_{h \neq j} \frac{e^{V_h/\hat{e}}}{e^{V_j/\hat{e}}} \right)} \cdot e^{\frac{-\hat{a}_j}{\hat{e}} - \hat{o}}. \end{aligned}$$

$$\text{Se si pone} \quad \hat{a}_j^* = \frac{\hat{a}_j}{\hat{e}} + \hat{o} \quad \Rightarrow \quad d\hat{a}_j = \hat{e} d\hat{a}_j^*$$

l'integrale diventa

$$\int_{-\infty}^{+\infty} \frac{1}{\hat{e}} \exp \left[ -\hat{a}_j^* - e^{-\hat{a}_j^*} \cdot \left( 1 + \sum_{j \neq h} \frac{e^{V_h/\hat{e}}}{e^{V_j/\hat{e}}} \right) \right] \cdot \hat{e} d\hat{a}_j^*.$$

Se si pone inoltre

$$\ddot{e}_j = \log \left( 1 + \sum_{h \neq j} \frac{e^{V_h/\ddot{e}}}{e^{V_j/\ddot{e}}} \right) = \log \left( \sum_{h=1}^m \frac{e^{V_h/\ddot{e}}}{e^{V_j/\ddot{e}}} \right)$$

dopo alcuni semplici passaggi, si ottiene

$$\begin{aligned} & \int_{-\infty}^{+\infty} \exp \left( -\hat{a}_j^* - e^{-(\hat{a}_j^* - \ddot{e}_j)} \right) d\hat{a}_j^* \\ &= \exp(-\ddot{e}_j) \cdot \int_{-\infty}^{+\infty} \exp \left( -\hat{a}_j - e^{-\hat{a}_j} \right) d\hat{a}_j \quad \text{dove} \quad \hat{a}_j = \hat{a}_j^* - \ddot{e}_j. \end{aligned}$$

A questo punto si può affermare che

$$\int_{-\infty}^{+\infty} \exp \left( -\hat{a}_j - e^{-\hat{a}_j} \right) d\hat{a}_j = 1$$

in quanto tale quantità risulta essere l'area sottesa alla funzione di densità di una variabile di Gumbel, avente  $\ddot{o} = 0$  e  $\ddot{e} = 1$ .

Si ottiene perciò:

$$p(j) = \exp(-\ddot{e}_j) = \frac{1}{\exp(\ddot{e}_j)} = \frac{e^{V_j/\ddot{e}}}{\sum_{h=1}^m e^{V_h/\ddot{e}}} \quad \text{c.v.d.}$$

## 9. BIBLIOGRAFIA

- Agresti A. (1990) *Categorical data analysis*, A Wiley – Interscience Publication, New York.
- Akiva B., Lerman S. R. (1985) *Discrete choice analysis: theory and application to travel demand*, MIT Press, Cambridge USA.
- Biggiero L. (1991) Un modello comportamentale per la generazione degli spostamenti non sistematici in area urbana, *Trasporti e Trazione*, 4.
- Cascetta E. (1998) *Teoria e metodi dell'ingegneria dei sistemi di trasporto*, Utet, Torino.
- Domenich T. A. , McFadden D. (1975) *Urban travel demand*, North Holland Publishing Company.
- Fabbris L. (1997) *Statistica multivariata – Analisi esplorativa dei dati*, McGraw Hill, Milano.
- Festa D., Mazzulla G., Condino D. (2001) Nuova città, Nuova mobilità, *Convegno Input2001, Isole Tremiti (FG)*.
- Landenna G., Marasini D., Ferrari P. (1998) *La verifica delle ipotesi statistiche*, Il mulino, Bologna.
- Manski C. F. , McFadden D. (1981) *Structural analysis of discrete data with econometric application*, The MIT Press, Cambridge.
- Oppenheim N. (1994) *Urban travel demand modeling*, Jhon Wiley & Sons, Inc.



## **ABSTRACT**

This work attempts to evaluate the possibility of application of some statistical models to urban travel demand. In particular the logistic regression is considered as possible tool in order to analyse the impact of some variables on the different choices. The structure of this paper presents a first part on the methodology and a second part on a real case of one city in Cosenza's area for which it has been estimated an emission model. The goal of this article is not only to determine the possibility of application of these kinds of models to urban travel demand, but also to show the possibility to investigate some interesting aspects of mobility through this approach. To this end, very important is estimate of model parameters, that leads the construction of odds ratios. In the logistic regression the odds ratios permit to calculate the relative risk to have a certain modality of response variable given a particular predictor value and so, in this contest, to individuate the most important factors that influence the transport system customers' choices.