Appunti di Econometria

ARGOMENTO [4]: VARIABILI DIPENDENTI BINARIE

Maria Luisa Mancusi – Università Bocconi

Novembre 2009

1 Introduzione

Nei modelli econometrici studiati fino ad ora la variabile dipendente, y_i, è sempre stata una variabile continua. In economia ci sono però diverse situazioni in cui la variabile che vogliamo spiegare è di tipo qualitativo o continua solo in parte. In tutti questi casi il modello di regressione lineare risulta solitamente inappropriato. Qui ci concentreremo sui modelli in cui la variabile dipendente y_i, è una variabile dummy.

- Perché alcuni individui decidono di andare all'università ed altri no?
- Perché alcuni individui decidono di acquistare la casa, mentre altri l'affittano?
- Perché alcune donne decidono di lavorare ed altre no?

In tutti questi casi la variabile che vogliamo spiegare è di tipo binario; ad esempio:

$$y_i = \begin{cases} 1 & \text{se la casa è di proprietà} \\ 0 & \text{se la casa non è di proprietà (es. affitto)} \end{cases}$$

2 Il modello di probabilità lineare

Proviamo a considerare un modello lineare che spiega y_i in funzione del reddito e di altre caratteristiche della famiglia:

$$y_i = x_i'\beta + \varepsilon_i$$

Sotto l'usuale ipotesi che $E(\varepsilon_i|x_i) = 0$, in questo modello

$$E(y_i|x_i) = x_i'\beta$$

Poiché y_i è un variabile con distribuzione di Bernoulli:

$$\begin{aligned} E(y_i|x_i) &= 1 \cdot \Pr(y_i = 1|x_i) + 0 \cdot \Pr(y_i = 0|x_i) = \\ &= \Pr(y_i = 1|x_i) \end{aligned}$$

Ne consegue che il modello che stimiamo spiega la probabilità che un evento si realizzi (es. l'acquisto della casa):

$$y_i = \Pr(y_i = 1|x_i) + \varepsilon_i$$

ovvero la probabilità che l'evento si realizzi è una funzione lineare:

$$\Pr(y_i = 1|x_i) = x_i'\beta$$

e, per questo motivo, questo modello è detto "*modello di probabilità lineare*". Sembrerebbe dunque che si possa applicare il metodo dei minimi quadrati anche al caso in cui la variabile dipendente sia di tipo binario; l'unica differenza rispetto ai modelli fin qui studiati sembrerebbe la sua interpretazione in termini di probabilità. Sfortunatamente, lo stimatore OLS in questo modello soffre di una serie di problemi.

- **ε_i non ha una distribuzione normale**

ε_i può assumere solo due valori ed è quindi distribuito secondo una distribuzione di Bernoulli. Infatti, $\varepsilon_i = y_{i-}x_i'\beta$ e quindi:

	ε_i	Probabilità
$y_i = 1$	$1 - x_i'\beta$	$x_i'\beta$
$y_i = 0$	$-x_i'\beta$	$1 - x_i'\beta$

Sappiamo che ciò non ha conseguenze sulla non distorsione e sulla consistenza di $\widehat{\beta}_{OLS}$ (abbiamo specificato correttamente la media condizionale di y_i), ma non è più vero che $\widehat{\beta}_{OLS}$ è distribuito secondo una normale. In particolare, questo implica che i test statistici non hanno le distribuzioni note (es. *t student* per la verifica delle ipotesi su un singolo parametro), che sono derivate dall'ipotesi di normalità degli errori in piccoli campioni. Le distribuzioni dei test statistici saranno basate sulla distribuzione asintotica di $\widehat{\beta}_{OLS}$.

- **ε_i è eteroschedastico**

La distribuzione di ε_i ricavata al punto precedente, evidenzia come anche ε_i sia una variabile Bernoulliana, dunque:

$$V(\varepsilon_i|x_i) = x_i'\beta\left(1 - x_i'\beta\right)$$

ovvero la varianza dell'errore dipende dalle x. Questo non ha effetto sulla non distorsione di $\widehat{\beta}_{OLS}$, ma:

(a) $V(\widehat{\beta}_{OLS}) \neq \sigma^2 (X'X)^{-1}$

La conseguenza principale è che gli standard errors forniti dal software a cui chiediamo di stimare il modello con OLS sono basati sulla formula sbagliata:

$$s.e.(\widehat{\beta}_j) = s \cdot \sqrt{c_{jj}}$$

dove c_{jj} è l'elemento nella posizione (j, j) nella matrice $(X'X)^{-1}$ ed $s = \frac{\sqrt{\sum \widehat{\varepsilon}_i}}{(N-K)}$. Ne consegue che i test sulla significatività dei coefficienti non sono attendibili.

(b) $\widehat{\beta}_{OLS}$ non è BLUE

La soluzione al problema dell'eteroschedasticità consiste nell'utilizzo dei minimi quadrati ponderati. Si stima prima il modello con OLS, ottenendo così una stima consistente di β, $\widehat{\beta}_{OLS}$; quindi si stima nuovamente con OLS il modello trasformato dividendo per $\sqrt{x_i'\widehat{\beta}_{OLS}\left(1 - x_i'\widehat{\beta}_{OLS}\right)}$.

- $x_i'\widehat{\beta}_{OLS}$ **può essere esterno all'intervallo** $[\mathbf{0}, \mathbf{1}]$

Nel modello di probabilità lineare $E(y_i|x_i) = \Pr(y_i = 1|x_i)$, dunque deve essere $0 \leq E(y_i|x_i) \leq 1$. E' però possibile che $x_i'\widehat{\beta}_{OLS}$, ovvero la stima di $\Pr(y_i = 1|x_i)$, risulti maggiore di uno o minore di zero. E' quest'ultimo il vero problema del modello di probabilità lineare e il metodo di stima che presenteremo di seguito garantisce che le probabilità stimate rimangano nell'intervallo $[0, 1]$.

3 Il modello con variabile latente

Il modo migliore per introdurre questo metodo di stima consiste nel partire da un problema economico. Ad esempio, immaginiamo di avere dati su un campione di famiglie e di essere interessati a determinare le variabili rilevanti nella scelta di acquisto della casa. Basandoci sulle nostre conoscenze di microeconomia potremmo pensare ad un modello che spiega la scelta di acquisto della casa come il risultato di un processo di massimizzazione dell'utilità da parte dell'individuo/famiglia. Tale processo determina la massima disponibilità a pagare una casa (es. al mq) da parte della famiglia in funzione del proprio reddito e di una serie di altre caratteristiche della famiglia che ne descrivono le preferenze. Mentre il reddito ed alcune caratteristiche famigliari sono osservabili non è però possibile osservare direttamente la massima disponibilità a pagare, ma solo la sua naturale conseguenza: se tale disponibilità è maggiore di una certa soglia (es. il prezzo di mercato al mq) allora la famiglia acquista la casa, altrimenti sceglie l'affitto.

Dunque, la nostra analisi empirica si fonda su un modello economico che non è quello che possiamo direttamente stimare perché la variabile dipendente non è la stessa; per questo motivo possiamo riferirci al primo modello (quello teorico) come "*modello latente*" ed al secondo (quello stimabile) come "*modello con variabile dipendente limitata*". Nell'esempio la variabile dipendente osservata è di tipo qualitativo e binario (la casa è di proprietà oppure no) e può quindi essere interpretata come una variabile dummy uguale a 1 se la casa è di proprietà e 0 se non lo è.

Supponiamo che il modello teorico sia di tipo lineare:

$$y_i^* = x_i'\beta + \varepsilon_i^*$$

dove y_i^* è la variabile latente. Sono invece disponibili osservazioni sulle variabili (y_i, x_i), $i = 1, ..., N$. y_i è la variabile dipendente limitata la cui relazione con y_i^* è data da[1] :

$$y_i = \begin{cases} 1 & \text{se } y_i^* > 0 \\ 0 & \text{se } y_i^* \leq 0 \end{cases}$$

[1] In quanto segue supponiamo, per semplicità, che la soglia sia uguale a 0.

Il nostro obiettivo è di scrivere il modello da stimare nella forma che ci assicura la possibilità di avere uno stimatore consistente:

$$y_i = E\left(y_i|x_i\right) + \varepsilon_i$$

In altri termini, dobbiamo specificare correttamente $E\left(y_i|x_i\right) = \Pr(y_i = 1|x_i)$. Ma a cosa è uguale $\Pr(y_i = 1|x_i)$? Per capirlo è sufficiente fare riferimento al modello latente:

$$\begin{aligned}
\Pr(y_i &= 1|x_i) = \Pr\left(y_i^* > 0|x_i\right) = \\
&= \Pr(x_i'\beta + \varepsilon_i^* > 0|x_i) = \\
&= \Pr\left(\varepsilon_i^* > -x_i'\beta|x_i\right)
\end{aligned}$$

Dunque, la corretta specificazione della media condizionale di y_i dipende dalla distribuzione dell'errore nel modello latente. Ipotizziamo dunque che:

$$\varepsilon_i^*\ i.i.d. \begin{cases} \text{F(.)} & \text{funzione di ripartizione} \\ \text{f(.)} & \text{funzione di densità} \end{cases}$$

e che $f(.)$sia simmetrica intorno allo 0. Dunque:

$$\Pr\left(\varepsilon_i^* > -x_i'\beta|x_i\right) = \Pr\left(\varepsilon_i^* < x_i'\beta|x_i\right) = F(x_i'\beta)$$

Ovvero il modello da stimare è:

$$y_i = F(x_i'\beta) + \varepsilon_i$$

Il modello così ottenuto è non lineare ed assicura che la probabilità stimata sia compresa tra zero e uno, poiché $0 \leq F(x_i'\beta) \leq 1$ per definizione della funzione di ripartizione. Una prima ed importante conseguenza è che gli effetti marginali non sono più costanti, ma variano al variare di x_i. Si noti, infatti, che i parametri β rappresentano gli effetti marginali del modello teorico latente:

$$\beta_j = \frac{\partial E\left(y_i^*|x_i\right)}{\partial x_{ij}}$$

Gli effetti marginali nel modello stimato ($y_i = F(x_i'\beta) + \varepsilon_i$) sono diversi ed hanno anche una diversa interpretazione, indicano cioè la variazione della probabilità che un evento si realizzi ($\leftrightarrow$ che $y_i = 1$) conseguente ad una variazione unitaria della variabile j:

$$\begin{aligned}
ME_j &= \frac{\partial E\left(y_i|x_i\right)}{\partial x_{ij}} = \\
&= \frac{\partial F(x_i'\beta)}{\partial x_{ij}} = \\
&= f\left(x_i'\beta\right)\beta_j
\end{aligned}$$

dove f indica la funzione di densità di ε_i^*. Questi effetti marginali non sono costanti, ma sono funzione delle variabili esplicative (l'effetto marginale assume un valore diverso per ogni individuo i in funzione del valore assunto dalle variabili esplicative).

A questo punto possiamo stimare il modello in modo efficiente usando il metodo di massima verosimiglianza. La distribuzione di ogni y_i è:

y_i	Probabilità
1	$F(x_i'\beta)$
0	$1 - F(x_i'\beta)$

ovvero:

$$f\left(y_i|x_i,\beta,\sigma\right) = \left[F(x_i'\beta)\right]^{y_i}\left[1 - F(x_i'\beta)\right]^{(1-y_i)}$$

Quindi il logaritmo della funzione di verosimiglianza è:

$$\ln L = \sum_{i=1}^{N}\left\{y_i \ln F(x_i'\beta) + (1-y_i)\ln\left[1 - F(x_i'\beta)\right]\right\}$$

Dunque:

$$\widehat{\beta} = \arg\max_{\beta} \ln L$$

ovvero $\widehat{\beta}$ risolve le seguenti K condizioni del primo ordine:

$$\frac{\partial \ln L}{\partial \beta} = 0 \Leftrightarrow \sum_{i=1}^{N}\left\{y_i \frac{f\left(x_i'\widehat{\beta}\right)}{F(x_i'\widehat{\beta})}x_i - (1-y_i)\frac{f\left(x_i'\widehat{\beta}\right)}{1 - F(x_i'\widehat{\beta})}x_i\right\} = 0$$

Una volta ottenuto $\widehat{\beta}$, per avere un indicatore sintetico degli effetti marginali si possono seguire due strade:

a) si calcola l'effetto marginale in corrispondenza del valore medio delle variabili esplicative

$$\widehat{ME} = f\left(\overline{x}'\widehat{\beta}\right)\widehat{\beta}_j$$

b) si calcola la media degli effetti marginali:

$$\widehat{ME} = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i'\widehat{\beta}\right)\widehat{\beta}_j$$

4 Significatività e bontà della stima nel modello con variabile latente

4.1 Significatività dei coefficienti

Per valutare la significatività del singolo coefficiente si usa sempre il test t, ma poiché lo stimatore MLE è solo asintoticamente normale, la statistica t è distribuita secondo una $N(0,1) \Rightarrow$ occorre fare riferimento ai valori critici della distribuzione normale standard e non più a quelli della t di student.

4.2 Significatività della regressione

Testare la significatività della regressione significa, come al solito, testare la seguente ipotesi:

$$
\begin{aligned}
H_0 &: \beta_j = 0 \,\forall j \text{ (ad esclusione della costante)} \\
H_1 &: \text{ esiste almeno un } \beta_j \neq 0
\end{aligned}
$$

Questo test viene costruito sulla base dei seguenti valori della funzione di verosimiglianza:

- L_0: valore della verosimiglianza associato a max lnL ottenuta dal modello "*ristretto*" in cui tutti i parametri, tranne la costante, sono posti uguali a zero ($\leftrightarrow$ si stima il modello con la sola costante come variabile esplicativa e si valuta lnL in corrispondenza del valore trovato)
- L_1: valore della verosimiglianza associato a max lnL ottenuta dal modello completo ($\leftrightarrow$ si stima il modello con tutte le variabili esplicative e si valuta lnL in corrispondenza dei valori dei parametri trovati)

La statistica test è:

$$2\,[\ln L_1 - \ln L_0] \overset{a}{\sim} \chi^2_{K-1}$$

dove $(K-1)$ è uguale al numero di variabili esplicative nel modello completo, meno la costante (ovvero il numero di restrizioni). Se la statistica test calcolata è maggiore del valore critico corrispondente al livello di significatività prescelto allora si rigetta l'ipotesi nulla in favore della significatività della regressione.

4.3 Bontà della stima

Per valutare la bontà della stima nei modelli con variabile latente non è possibile utilizzare l'R^2. Sono dunque state costruite delle misure alternative basate su L_1 ed L_0 definite come sopra e, in particolare, sulla distanza tra $\ln L_1$ ed $\ln L_0$: maggiore è questa distanza, maggiore è il contributo alla spiegazione della variabile dipendente fornito dal modello completo rispetto a quello ristretto. Due sono gli indicatori di bontà di stima maggiormente utilizzati:

$$
\begin{aligned}
\text{pseudo } R^2 &= 1 - \frac{1}{1 + 2(\ln L_1 - \ln L_0)/N} \\
\text{McFadden } R^2 &= 1 - \frac{\ln L_1}{\ln L_0}
\end{aligned}
$$

Poiché la verosimiglianza è una probabilità congiunta: $0 \leq L \leq 1 \Rightarrow \ln L \leq 0$. Inoltre, essendo il massimo non vincolato sempre maggiore del massimo vincolato: $lnL_1 \geq \ln L_0$. Ne consegue che:

$$0 \leq \text{pseudo } R^2 \leq 1 \text{ e } 0 \leq \text{McFadden } R^2 \leq 1$$

Un valore pari a zero si ha quando tutti i coefficienti delle variabili esplicative sono uguali a zero, ovvero quando $lnL_1 = \ln L_0$.

5 Due applicazioni importanti del modello con variabile latente

5.1 Modello PROBIT

Supponiamo che $\varepsilon_i^* \overset{iid}{\sim} N(0,1)$, ovvero:

$$\begin{aligned} F(x_i'\beta) &= \Phi(x_i'\beta) = \int\limits_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz = \\ &= \int\limits_{-\infty}^{x_i'\beta} \phi\left(z\right) dz \end{aligned}$$

Riprendendo dunque quanto ricavato nella sezione 2, possiamo scrivere il modello che stimiamo come:

$$y_i = \Phi(x_i'\beta) + \varepsilon_i$$

Gli effetti marginali nel modello stimato sono:

$$ME_j = \frac{\partial E\left(y_i|x_i\right)}{\partial x_{ij}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{ij}} = \phi\left(x_i'\beta\right)\beta_j$$

dove ϕ indica la funzione di densità della $N(0,1)$.

Anche in questo caso il modello il modello è stimato usando il metodo di massima verosimiglianza, ovvero:

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \left\{ y_i \ln \Phi(x_i'\beta) + (1-y_i) \ln\left[1 - \Phi(x_i'\beta)\right] \right\} \\ \widehat{\beta} &= \arg\max_{\beta} \ln L \end{aligned}$$

Nel modello probit le condizioni del secondo ordine sono verificate (in assenza di multicollinearità) $\leftrightarrow$ lnL è concava $\Rightarrow$ esiste un massimo ed è dunque sempre possibile trovare $\widehat{\beta}$ anche se la funzione che si massimizza non è lineare ed è quindi richiesto l'utilizzo di algoritmi iterativi.

5.2 Modello LOGIT

Questo secondo modello è costruito sull'ipotesi che gli errori del modello teorico siano distribuiti secondo una logistica standard. In questo caso la funzione di ripartizione e quella di densità sono:

$$\begin{aligned} F(x_i'\beta) &= \Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \\ f(x_i'\beta) &= \frac{e^{x_i'\beta}}{\left(1+e^{x_i'\beta}\right)^2} = \\ &= \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \frac{1}{1+e^{x_i'\beta}} = \\ &= \Lambda(x_i'\beta)\left[1 - \Lambda(x_i'\beta)\right] \end{aligned}$$

La funzione di densità della logistica ha una forma a campana simile a quella della normale e simmetrica intorno alla propria media. Anche in questo caso possiamo dunque scrivere il modello come:

$$y_i = \Lambda(x_i'\beta) + \varepsilon_i$$

ovvero

$$y_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} + \varepsilon_i$$

da cui si ricavano gli effetti marginali:

$$ME_j = \frac{\partial E\left(y_i|x_i\right)}{\partial x_{ij}} = f(x_i'\beta)\beta_j = \frac{e^{x_i'\beta}}{\left(1 + e^{x_i'\beta}\right)^2}\beta_j$$

Veniamo ora alla funzione di verosimiglianza.y_i è sempre una variabile Bernoulliana la cui distribuzione è data da:

y_i	Probabilità
1	$\Lambda(x_i'\beta)$
0	$1 - \Lambda(x_i'\beta)$

Quindi:

$$\ln L = \sum_{i=1}^{N} \left\{ y_i \ln \Lambda(x_i'\beta) + (1 - y_i) \ln \left[1 - \Lambda(x_i'\beta)\right] \right\}$$

E' particolarmente interessante, in questo caso, derivare le condizioni del primo ordine per la massimizzazione di lnL rispetto a β. Il vettore delle derivate di lnL rispetto a β è

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} \left\{ y_i x_i - \Lambda(x_i'\beta) x_i \right\}$$

e le condizioni del primo ordine:

$$\frac{\partial \ln L}{\partial \beta} = 0 \Leftrightarrow \sum_{i=1}^{N} \left(y_{i-}\Lambda(x_i'\widehat{\beta}) x_i \right) = 0$$

Confrontando l'espressione delle condizioni del primo ordine con il modello stimato si deduce che, nel modello logit, lo stimatore di massima verosimiglianza si ricava dalla condizione di ortogonalità tra gli errori e le variabili esplicative:

$$\sum_{i=1}^{N} \widehat{\varepsilon}_i x_i = 0$$

Inoltre, indicando con il valore stimato della $\Pr(y_i = 1|x_i) = \frac{e^{x_i'\widehat{\beta}}}{1+e^{x_i'\widehat{\beta}}}$, sempre dalle condizioni del primo ordine si ottiene:

$$\sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N} \widehat{p}_i x_i$$

Quindi, se tra le variabili esplicative è inclusa la costante:

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \widehat{p}_i$$

Dividendo per N entrambi i lati, l'uguaglianza sopra riportata ci dice che nel modello logit lo stimatore di massima verosimiglianza verifica l'uguaglianza tra frequenza effettiva (numero relativo di osservazioni per cui $y_i = 1$) e frequenza stimata.

Anche nel modello logit le condizioni del secondo ordine sono verificate (in assenza di multicollinearità) e quindi è sempre possibile trovare $\widehat{\beta}$ che massimizza lnL.

Infine, nel modello logit è disponibile un modo alternativo per descrivere l'effetto delle variabili esplicative sulla probabilità, in termini di *odds-ratio*, ovvero:

$$\Omega\left(y_i = 1|x_i\right) = \frac{\Pr(y_i = 1|x_i)}{\Pr(y_i = 0|x_i)} = \frac{\Lambda(x_i'\widehat{\beta})}{1 - \Lambda(x_i'\widehat{\beta})} = \exp\left(x_i'\beta\right)$$

In particolare, consideriamo una variazione della variabile x_j

$$x_i^0 = \begin{pmatrix} 1 \\ x_{1i} \\ \cdots \\ x_{ji} \\ \cdots \\ x_{ki} \end{pmatrix} \to x_i^1 = \begin{pmatrix} 1 \\ x_{1i} \\ \cdots \\ x_{ji} + \Delta x_{ji} \\ \cdots \\ x_{ki} \end{pmatrix}$$

l'effetto conseguente sulla probabilità relativa è:

$$\frac{\Omega\left(y_i = 1|x_i^1\right)}{\Omega\left(y_i = 1|x_i^0\right)} = \frac{\exp(x_i^{0\prime}\beta + \Delta x_{ji}\beta_j)}{\exp(x_i^{0\prime}\beta)} = \exp(\Delta x_{ji}\beta_j)$$

Se consideriamo variazioni unitarie $(\Delta x_{ji} = 1)$, l'effetto sulla probabilità relativa è semplicemente e^{β_j}. Quindi sarà nullo se $\beta_j = 0$ $(e^{\beta_j} = 1)$, positivo se $\beta_j > 0$ $(e^{\beta_j} > 1)$ e negativo se $\beta_j < 0$ $(e^{\beta_j} < 1)$.