STATISTICA, anno LXII, n. 1, 2002

STIMATORI COL METODO DELLA REGRESSIONE IN CASO DI NON RISPONDENTI CON UN CAMPIONAMENTO DOPPIO (*)

Angiola Pollastri

1. PREMESSA

Effettuando indagini campionarie anche molto accurate, ci si imbatte spesso nel problema dei non rispondenti. Esso dipende dal tema dell'indagine, dalla struttura del questionario, dalle caratteristiche delle unità selezionate, dal modo di condurre l'indagine e da molti altri fattori.

In questa situazione il disegno campionario viene meno e le probabilità di selezione cambiano secondo uno schema ignoto al ricercatore. È facile da dimostrare (si veda, ad esempio, Pollastri, 1997) che la distorsione dello stimatore della media è funzione della proporzione di non rispondenti e della differenza fra la media della variabile oggetto di studio fra i rispondenti e quella fra i non rispondenti. Pertanto la suddetta distorsione non diminuisce all'aumentare della numerosità del campione.

Una via per ridurre la distorsione dello stimatore è quella di estrarre un campione ridotto fra coloro che cono non rispondenti al primo tentativo e stimare uno o più parametri della popolazione tenendo conto sia dell'informazione sui rispondenti nella prima faie che sui rispondenti nella seconda fase. Spesso la procedura di raccolta dei dati è diversa nelle due fasi. Se, ad esempio, nella prima fase si è effettuata un'indagine postale, nella seconda fase si potrebbe optare per un'intervista diretta avvalendosi di intervistatori con provata esperienza.

Nel presente lavoro, dopo aver brevemente rivisto il metodo proposto nel 1946 da Hansen e Hurwitz, nel paragrafo 3 vengono analizzate le procedure per stimare la media dalle osservazioni ottenute nelle due fasi col metodo della regressione che si basa sulla conoscenza della media di una o più variabili ausiliarie. Relativamente a quest'ultima situazione si ricava la varianza dello stimatore. La dimostrazione viene riportata in appendice.

Si mostra poi, nel paragrafo 4, come la stima dei parametri (qualora non siano noti) del modello lineare che lega la variabile di cui si vuol stimare la media alle variabili ausiliarie, possa migliorare tenendo conto anche del campione effettuato

(*) Ricerca co-finanziata nell'ambito del progetto MURST ex-40% "Produzione e sperimentazione di sistemi computer assisted per rilevare la qualità della didattica universitaria e l'inseritnento professionale dei laureati". Coordinatore Nazionale Prof Luigi Fabbris.

fra i non rispondenti della prima fase. Le stime in questione vengono poi utilizzate per stimare la media col metodo della regressione.

2. STIMA DELLA MEDIA CON UN CAMPIONAMENTO DOPPIO

Data una popolazione formata da N unità statistiche, si indichi con μ_Y la media della caratteristica Y che si intende stimare.

Si supponga che la popolazione sia suddivisa in due strati di cui il primo formato da N, soggetti rispondenti e il secondo formato da N, soggetti non rispondenti dirrante la prima fase di campionamento.

Si indichi con $W_1 = N_1/N$ e con $W_2 = N_2/N$ rispettivamente l'ignota frazione dei rispondenti e dei non rispondenti nella popolazione. Prima di effettuare il campione non si sa se un individuo può essere identificato come rispondente o non rispondente e pertanto non si conosce la composizione dei due strati.

Si supponga che la media della caratteristica Y sia diversa nei due strati: si indichi con μ_1 la media nello strato dei non rispondenti e μ_2 la inedia nello strato dei non rispondenti.

Si estragga un campione casuale semplice, indicato con s, di ampiezza n. Si supponga che n, individui rispondano mentre n, non rispondano in questa fase dell'indagine. Si denoti con s_1 il gruppo di rispondenti e con s, il gruppo di non rispondenti.

Le quantità

$$\hat{w}_1 = \frac{n_1}{n} \quad \text{e} \quad \hat{w}_2 = \frac{n_2}{n}$$

sono stime corrette rispettivamente di W, e di W_2.

Hansen e Hurwitz (1946) proposero di prelevare un sottocampione, indicato con s_2', fra i non rispondenti del primo campione di ampiezza $n_2' = n, . k$ con $0 < k < 1$.

Assumendo che tutte le unità del sottocampione rispondano, una stima corretta per μ_Y è data da

$$\bar{y}^* = \bar{y}_1\hat{w}_1 + \bar{y}_2\hat{w}_2$$

dove $\bar{y}_1$ è la stima della media μ_1 dello strato dei rispondenti basata su n, e $\bar{y}_2$ è la stima della media μ_2 dello strato dei non rispondenti al primo tentativo basata sulle n_2' osservazioni.

Si può dimostrare che (si veda Hedayat e Sinha, 1991) lo stimatore corrispondente $\bar{Y}$ ha varianza data da

$$\mathrm{Var}(\bar{Y}') = \frac{\sigma_y^2}{n} + \frac{W_2\sigma_{2y}^2}{n}\frac{(1-k)}{k}$$

dove σ_y^2 è la varianza della caratteristica nella popolazione mentre σ_{2y}^2 è la varianza della caratteristica fra i non rispondenti. Il primo termine coinciderebbe con la varianza della stima se tutti rispondessero. Il secondo termine è l'incremento di

varianza dovuto al fatto che si sono campionate fra i non rispondenti n_2' unità invece di n_2.

Nel caso in cui il campionamento venga effettuato in blocco (Thompson, 1992) la varianza dello stirnatore $\bar{Y}_B$ è

$$\mathrm{Var}(\bar{Y}_B^*) = \frac{(N-n)\sigma_y^2}{(N-1)n} + \frac{W_2\sigma_{2y}^2}{n}\frac{(1-k)}{k}$$

3. STIMA DELLA MEDIA CON INFORMAZIONI AUSILIARIE SULLA POPOLAZIONE

In molte situazioni può accadere di essere in possesso di informazioni relative ad alcune variabili nella popolazione. Tenere conto di queste informazioni può migliorare notevolmente l'efficienza degli stimatori ed ottenere l'effetto di compensare in qualche modo le perdite di informazioni subite a causa dei non rispondenti.

3.1. *Nota la media della variabile ausiliaria ed il coefficiente angolare*

Si supponga di conoscere la inedia μ_x di una variabile ausiliaria X e si supponga che sia ragionevole assumere un legame approssimativamente lineare tra la variabile oggetto di studio Y e la variabile ausiliaria. Si ipotizzi che il coefficierite di regressione b di Y su X sia noto da passate esperienze.

La stima di μ_1 col metodo della regressionc nello strato dei rispondenti diventa

$$\bar{y}_{lr1} = \bar{y}_1 + b(\mu_x - \bar{x}_1)$$

dove $\bar{x}_1$ è la stirna di μ_x basata sulle n, osservazioni effettuate sui rispondenti.

La stima di μ_2 col metodo della regressione è

$$\bar{y}_{lr2} = \bar{y}_2 + b(\mu_x - \bar{x}_2)$$

dove $\bar{x}_2$ è la stima di p, basata sulle n_2' osservazioni effettuate fra i non rispondenti.

La stima della inedia di Y è data da

$$\bar{\bar{y}}_{lr} = \hat{w}_1\bar{y}_{lr1} + \hat{w}_2\bar{y}_{lr2} = \bar{y}^* + b(\mu_x - \bar{x}^*)$$

dove

$$x = \bar{x}_1\hat{w}_1 + \bar{x}_2\hat{w}_2$$

Si può facilmente dimostrare che

$$\mathrm{E}(\bar{Y}_{lr}) = \boldsymbol{p},$$

La varianza del suddetto stimatore è data da

$$\mathrm{Var}(\bar{Y}_{lr}) = \left[\frac{\sigma_y^2}{n} + W_2\frac{\sigma_{2y}^2}{n}\frac{(1-k)}{k}\right](1-\rho^2)$$

Se il campionamento è in blocco

$$\mathrm{Var}(\bar{Y}_{lr}^{B}) = \left[\frac{\sigma_y^2}{n} \frac{N-n}{N-1} + W_2 \frac{\sigma_{2y}^2}{n} \frac{(1-k)}{k} \right] (1-\rho^2)$$

ove σ_{2y}^2 è la varianza della variabile Y fra i non rispondenti.

3.2. *Modello unico su tutta la popolazione note le medie di p variabili ausiliare e il vettore b*

Si supponga di conoscere le medie μ_h di p variabili ausiliarie X_h $(h = 1, \ldots, p)$. Si ipotizzi che il legame fra la variabile oggetto di indagine Y e le variabili X_h sia del tipo:

$$Y = b_0 + \sum_{h=1}^{p} b_h X_h$$

Si supponga inoltre che i coefficienti b_h $(h = 1, \ldots, p)$ siano noti da passate esperienze.

La stima col metodo della regressione della media dello strato dei rispondenti è

$$\bar{y}_{lr1} = \bar{y}_1 + \sum_{h=1}^{p} b_h (\mu_h - \bar{x}_{h1})$$

La stima col metodo della regressione della media dello strato dei non rispondenti è

$$\bar{y}_{lr2} = \bar{y}_2 + \sum_{h=1}^{p} b_h (\mu_h - \bar{x}_{h2})$$

La stima della media complessiva μ_Y è

$$\bar{y}_{lr}^{*} = \hat{w}_1 \bar{y}_{lr1} + \hat{w}_2 \bar{y}_{lr2} = \hat{w}_1 \bar{y}_1 + \hat{w}_2 \bar{y}_2 + \sum_{h=1}^{p} b_h (\mu_h - \bar{x}_k^{*}) = \bar{y}^{*} + \sum_{h=1}^{p} b_h (\mu_h - \bar{x}_k^{*}) \qquad (1)$$

dove

$$\bar{x}_k^{*} = \hat{w}_1 \bar{x}_{k1} + \hat{w}_2 \bar{x}_{k2}$$

Dopo alcuni passaggi (riportati in appendice 1) si può dimostrare che:

$$\mathrm{Var}(\bar{Y}_{lr}^{*}) = \frac{\sigma_y^2}{n} + \sum_{h=1}^{p} \sum_{l=1}^{p} b_h b_l \frac{\sigma_{hl}}{n} - 2 \sum_{h=1}^{p} b_h \frac{\sigma_{hy}}{n} + W_2 \frac{(1-k)}{k}$$

$$\left[\frac{\sigma_{2y}^2}{n} + \sum_{h=1}^{p} \sum_{l=1}^{p} b_h b_l \frac{\sigma_{hl}}{n} - 2 \sum_{h=1}^{p} b_h \frac{\sigma_{hy}}{n} \right]$$

dove

$\sigma_{hl} = \mathrm{Cov}(X_h, X_l)$
$\sigma_{hy} = \mathrm{Cov}(X_h, Y)$

4. STIMA DI b NEI. CASO DI MODELLO UGUALE SUI DUE STRATI

Pur essendo valido il modello (*), in molte situazioni i parametri b_i non sono noti e devono essere stimati dal campione. Se si hanno buone ragioni (ad es. verifica su un campione doppio effettuato in una occasione precedente) per ritenere che il vettore dei parametri del modello nello strato dei rispondenti iniziali (indicato con $\underline{b}_1$) sia uguale a quello dello strato dei non rispondenti al primo tentativo (indicato con b,) allora lo stimatore BLUE usa le informazioni del primo e del secondo campione soggette alla condizione che $\underline{b}_1$ e $\underline{b}_2$ siano uguali.

Questo stimatore (Vernizzi, 1987) è dato da

$$\hat{\underline{b}}_1^{RLS} = \hat{\underline{b}}_2^{RLS} = \left[\frac{1}{\sigma_1^2}\underline{S}_1 + \frac{1}{\sigma_2^2}\underline{S}_2\right]^{-1}\left[\frac{1}{\sigma_1^2}\underline{s}_1 + \frac{1}{\sigma_2^2}\underline{s}_2\right] = \hat{\underline{b}}_1^{OLS} - \sigma_1^2\underline{S}_1^{-1}\underline{A}^{-1}\underline{h}$$

dove

$$\underline{S}_i = (\underline{X}_i'\underline{X}_i) \qquad \underline{s}_i = \underline{X}_i'\underline{y}_i \quad \text{e} \quad \hat{\underline{b}}_1^{OLS} = S_i^{-1}s_i \qquad i = 1,2$$

$$\hat{h} = \hat{\underline{b}}_1^{OLS} - \hat{\underline{b}}_2^{OLS} \quad \text{e} \quad \underline{A} = [\sigma_1^2\underline{S}_1^{-1} + \sigma_2^2\underline{S}_2^{-1}]$$

La matrice di varianze e covarianze di $\hat{b}_1^{RLS}$ è

$$\text{Var}(\hat{\underline{b}}_1^{RLS}) = \sigma_1^2\underline{S}_1^{-1} - \sigma_1^2\underline{S}_1^{-1}\underline{A}^{-1}\sigma_1^2\underline{S}_1^{-1}$$

che è inferiore a quello di $\hat{b}_1^{OLS}$ essendo

$$\text{Var}(\hat{\underline{b}}_1^{OLS}) = \sigma_1^2\underline{S}_1^{-1}$$

Utilizzando ora le stime contenute nel vettore $\hat{b}_1^{RLS}$ nella (1) si può stimare μ_Y.

5. CONCLUSIONI

Il presente lavoro ha messo in luce la possibilità di migliorare lo stimatore della media in caso di non rispondenti utilizzando sia informazioni assunte fra i non rispondenti iniziali che la conoscenza di medie di variabili correlate con la variabile oggetto di studio.

Si riesce inoltre a costruire uno stimatore più efficiente anche dei parametri del modello lineare se si tiene conto delle informazioni assunte nella prima fase e nella fase successiva in cui si va ad estrarre un campione fra chi non ha risposto al primo tentativo. Le stime dei suddetti parametri vengono poi utilizzate al fine di stimare la media col metodo della regressione.

Dipartimento Metodi Quantitatiui per *L'Economia* ANGIOLA POLLASTRI
Università degli Studi Milano - Bicocca

APPENDICE

Dimostrazione della varianza dello stimatore $\bar{Y}^{*}_{lr}$

La stima di p, coi dati ottenuti nelle dtie fasi di estrazione può essere scritta nel seguente modo:

$$\bar{y}_{lh} = \hat{w}_1\bar{y}_{lr1} + \hat{w}_2\bar{y}_{lr2} = \hat{w}_1\bar{y}_{lr1} + \hat{w}_2\bar{y}_{lr2} + \hat{w}_2(\bar{y}_{lr2} - \bar{y}^{*}_{lr2}) = y_{lh} + \hat{w}_2(\bar{y}_{lr2} - \bar{y}^{*}_{lr2})$$

dove $\bar{y}^{*}_{lr2}$ sarebbe la stima della media col metodo della regressione sul campione s_2 se tutte le unità rispondessero al primo tentativo.

Partendo dalla relazione

$$\bar{Y}_{lr2} - \mu_2 = (\bar{Y}_{lr2} - \bar{Y}^{*}_{lr2}) + (\bar{Y}^{*}_{lr2} - \mu_2)$$

elevando al quadrato entrambi i membri e calcolandone il valore atteso, si ottiene

$$\mathrm{E}(\bar{Y}_{lr2} - \mu_2)^2 = \mathrm{E}(\bar{Y}_{lr2} - \bar{Y}^{*}_{lr2})^2 + \mathrm{E}(\bar{Y}_{lr} - \mu_2)^2 \qquad (2)$$

esiendo il doppio prodotto pari a 0 in quanto, dato il campione s_2, la v.c $\bar{Y}^{*}_{lr2}$ si riduce ad una costante e $\mathrm{E}(\bar{Y}_{lr2}/s_2) = \bar{Y}^{*}_{lr2}$ e pertanto

$$\mathrm{E}[\bar{Y}^{*}_{lr2}(\bar{Y}_{lr2} - \bar{Y}^{*}_{lr2})] = \mathrm{E}_{s_2}\{\mathrm{E}[(Y^{*}_{lr2}(Y_{lr2} - Y^{*}_{lr2})/s_2]\} = E,(0) = 0$$

Si consideri la varianza dello stimatore dclla media col metodo della regressione quando ci si avvale dclla conoscenza delle medie di p variabili ausiliarie

$$\mathrm{Var}(\bar{Y}_{lr}) = \mathrm{E}(\bar{Y}_{lr} - \mu_y)^2 = \mathrm{E}[(\bar{Y} - \mu_y) + \sum_{k=1}^{p} b_k(\mu_k - \bar{X}_k)]^2$$

$$= \mathrm{E}(\bar{Y} - \mu_Y)^2 + \sum_{k=1}^{p}\sum_{l=1}^{p} b_k b_l\, \mathrm{E}[(\mu_k - \bar{X}_k)(\mu_l - \bar{X}_l)]$$

$$+ 2\sum_{k=1}^{p} b_k\, \mathrm{E}[(\bar{Y} - \mu_Y)(\mu_k - \bar{X}_k)]$$

$$= \frac{\sigma_y^2}{n} + \sum_{k=1}^{p}\sum_{l=1}^{p} b_k b_l \frac{\sigma_{kl}}{n} - 2\sum_{k=1}^{p} b_k \frac{\sigma_{yk}}{n}$$

Quindi la relazione *(2)*diventa:

$$\frac{\sigma_{2y}^2}{n_2'} + \sum_{k=1}^{p}\sum_{l=1}^{p} b_k b_l \frac{\sigma_{kl}}{n_2'} - 2\sum_{k=1}^{p} b_k \frac{\sigma_{yk}}{n_2'}$$

$$= \mathrm{E}(\bar{Y}_{lr2} - \bar{Y}^{*}_{lr2})^2 + \frac{\sigma_{2y}^2}{n_2'} + \sum_{k=1}^{p}\sum_{l=1}^{p} b_k b_l \frac{\sigma_{kl}}{n_2'} - 2\sum_{k=1}^{p} b_k \frac{\sigma_{yk}}{n_2}$$

Da cui segue che

$$\mathrm{E}(\bar{Y}_{lr2} - \bar{Y}^{*}_{lr2})^2 = \frac{\sigma_{2y}^2}{n_2'}\frac{(1-k)}{k} + \sum_{k=1}^{p}\sum_{l=1}^{p} b_k b_l \frac{\sigma_{kl}}{n_2'}\frac{(1-k)}{k} - 2\sum_{k=1}^{p} b_k \frac{\sigma_{yk}}{n_2}\frac{(1-k)}{k}$$

Tenendo conto del fatto che prima di effettuare la prima fase di campionamento anche la proporzione di non rispondenti è una variabile casuale in quanto non si sa quante unità ira le n estratte non risponderanno, si può scrivere:

$$\mathrm{Var}[\hat{W}_2(\overline{Y}_{lr2} - \overline{Y}^*_{lr2})] = \mathrm{V}_{\hat{W}_2}\{\mathrm{E}_{\hat{Y}_{lr}}[\hat{W}_2(\overline{Y}_{lr2} - \overline{Y}^*_{lr2})/\hat{W}_2 = \hat{w}_2]\}$$

$$+ \mathrm{E}_{\hat{W}_2}\{\mathrm{V}_{\hat{Y}_{lr}}[\hat{W}_2(\overline{Y}_{lr2} - \overline{Y}^*_{lr2})/\hat{W}_2 = \hat{w}_2]\}$$

$$= \mathrm{V}_{\hat{W}_2}\{\hat{W}_2(\mu_2 - \mu_2)\} + \mathrm{E}_{\hat{W}_2}\left\{\hat{W}_2^2 \frac{(1-k)}{k}\left[\frac{\sigma^2_{2y}}{n} + \sum_{h=1}^{p}\sum_{l=1}^{p} b_h b_l \frac{\sigma_{hl}}{n} - 2\sum_{h=1}^{p} b_h \frac{\sigma_{hy}}{n}\right]\right\}$$

pertanto, essendo

$$V(\overline{Y}^*_{lr}) = V(\overline{Y}_{lr}) + V[\hat{W}_2(\overline{Y}_{lr2} - \overline{Y}^*_{lr2})]$$

si può concludere che:

$$\mathrm{Var}(\overline{Y}^*_{lr}) = \frac{\sigma^2_y}{n} + \sum_{h=1}^{p}\sum_{l=1}^{p} b_h b_l \frac{\sigma_{hl}}{n} - 2\sum_{h=1}^{p} b_h \frac{\sigma_{hy}}{n} + W_2 \frac{(1-k)}{k}$$

$$\left[\frac{\sigma^2_{2y}}{n} + \sum_{h=1}^{p}\sum_{l=1}^{p} b_h b_l \frac{\sigma_{hl}}{n} - 2\sum_{h=1}^{p} b_h \frac{\sigma_{hy}}{n}\right]$$

RIFERIMENTI BIBLIOGRAFICI

G. CICCHITELLI, A. HERZEL, G.E. MONTANARI (1997), *Il campionamento statistico,* Il Mulino, Bologna.

C. CECCON, G. DIANA, A. SALVAN (1991), *Approccio classico al campionamento da popolazioni finite: alcuni risultati recenti,* CLEUP, Padova.

W.G. COCHRAN (1977), *Sampling techniques,* J. Wiley & Sons, New York.

B.V. FROSINI, M. MONTINARO, G. NICOLINI (1994), *Il campionamento da popolazioni finite,* UTET, Torino.

M.H. HANSEN, W.N. HURWITZ (1946), *The problem of non-response in sample survey,* "Journal of American Statistical Association", 41, pp. 517-529.

A.S. HEDAYAT, B.K. SINHA (1991), *Desing and inference in finite population sampling,* J. Wiley & Sons, New York.

J.T. LESLES, W.D. KALSBEEK (1992), *Non-sampling error in surveys,* J. Wiley & Sons, New York.

A. POLLASTRI (1997), *Elementi di teoria dei campioni,* CUSL, Milano.

S.R.S. RAO (1986), *Estimation par le quozient dans le cas d'un sous-échantillonage des non-répondants,* "Techniques d'enquete", *12,* 2, pp. 225-238.

C.E. SARNDAL (1986), *Estimation par la méthode de régression en situation de non-réponse,* "Techniques d'enquete", 12, 2, pp. 215-224.

S.K. THOMPSON (1992), *Sampling,* J. Wiley & Sons, New York.

A. VERNIZZI (1987), *On the problem of estimation with approximative linear restrictions,* "Metron", XLV, 3-4, pp. 195-211.

RIASSUNTO

Stimatori col metodo della regressione in caso di non rispondenti con un campionamento doppio

Nel presente lavoro vengono studiate le procedure per stimare la media di una caratteristica in presenza di non rispondenti in una prima fase sia avvalendosi delle informazioni desunte andando ad indagare su un piccolo campione estratto ira i non rispondenti sia tenendo conto delle informazioni che si possono avere a disposizione relativamente a medie di variabili ausiliarie. Viene ricavata la varianza dello stimatore col metodo della regressione.

Si propone poi di stimare i parametri del modello lineare che lega la variabile oggetto di studio e le variabili ausiliarie tenendo conto delle informazioni desunte nelle due fasi di campionamento.

SUMMARY

Regression estimators in presence of non-respondents with a double sample

The aim of this paper is to study the procedures to estimate the mean of a variablc in presence of non-respondents in a first sanipling using the informations of a seconde sample among the non-respondcnts and the informations icgarding the means of one or more auxiliaries variables. It is shown the variance of the regression estimator.

Then it is proposcd the estimation of the parameters of the linear model using the information of the first and the second sample.