

UNIVERSITÀ DEGLI STUDI DI FIRENZE

Dipartimento Statistico

DOTTORATO IN STATISTICA APPLICATA VII CICLO

**IMPUTAZIONE MULTIPLA:
UN'APPLICAZIONE
ALL'ANALISI DI DATI DI REDDITO**

Anna Giraldo

Relatore:
Prof. Luigi Biggeri

Coordinatore:
Prof. Luigi Biggeri

Indice

Introduzione	1
1 Le mancate risposte nelle indagini campionarie	5
1.1 Tipi di mancata risposta	6
1.2 Formalizzazione del problema della non risposta	10
1.2.1 Termini del problema	10
1.3 Meccanismi di non risposta non ignorabili	16
2 Trattamento delle mancate risposte parziali	19
2.1 Metodi per trattare i dati incompleti	19
2.1.1 Metodi basati sulle sole unità osservate	19
2.1.2 Metodi di imputazione e ponderazione	21
2.1.3 Metodi basati sui modelli	22
2.2 Aggiustamenti per la non risposta parziale: l'imputazione	23
2.2.1 Metodi di imputazione deterministica	28
2.2.2 Imputazione "hot-deck"	29
2.2.3 Un esempio di "hot-deck" sequenziale: la Current Population Survey	33
2.2.4 Imputazione per regressione	33
2.3 Vantaggi e svantaggi dell'imputazione singola	35

3	L'imputazione multipla	37
3.1	Introduzione	37
3.2	Esempio numerico di imputazione multipla	41
3.3	Metodi per analizzare dataset imputati più volte	46
3.4	Imputazione multipla attraverso un modello esplicito bayesiano	50
3.4.1	Esempio di modello esplicito	53
3.5	Imputazione propria ed impropria	55
3.5.1	Esempio di modello implicito	57
3.6	Imputazione multipla per meccanismi di non risposta non ignorabili	59
4	L'imputazione di dati di reddito	63
4.1	Particolarità dei dati di reddito e caratteristiche dei non rispondenti	63
4.2	Proposte in letteratura per trattare dati di reddito incompleti	68
4.2.1	Italia	68
4.2.2	Stati Uniti	71
5	L'indagine suppletiva alla Rilevazione Trimestrale sulle Forze di Lavoro: caratteristiche della non risposta	73
5.1	La Rilevazione Trimestrale sulle Forze di Lavoro (RTFL)	73
5.2	L'indagine suppletiva alla RTFL	76
5.2.1	Il questionario	80
5.2.2	Il campione	83
5.3	Caratteristiche delle mancate risposte totali	86
5.4	Procedure di correzione dei dati individuali	93
5.5	Caratteristiche delle mancate risposte parziali	94
6	Alcune proposte per l'imputazione multipla di dati di reddito	99
6.1	Generalità	99
6.2	Analisi preliminare	102
6.3	Imputazione multipla di dati di reddito	108

6.3.1	Imputazione per regressione con residui casuali	110
6.3.2	Imputazione tramite il “Mean predictive matching”	122
6.3.3	Imputazione utilizzando il modello MIMIC	122
6.4	Analisi di sensibilità	125
Conclusioni		129
Bibliografia		131
Bibliografia Annotata		143

Indice delle figure

3.1	<i>Dataset</i> con M imputazioni per ogni dato mancante	38
5.1	Struttura a salti per la rilevazione della posizione lavorativa	82
6.1	Istogramma redditi da lavoro autonomo. Casi osservati.	102
6.2	Istogramma redditi da lavoro dipendente. Casi osservati.	104
6.3	Istogramma logaritmo redditi da lavoro autonomo. Casi osservati. . .	106
6.4	Istogramma logaritmo redditi da lavoro dipendente. Casi osservati. .	106
6.5	Istogramma residui standardizzati redditi da lavoro autonomo. . . .	118
6.6	Grafico di probabilità normale dei redditi da lavoro autonomo. . . .	119
6.7	Grafico residui standardizzati contro valori predetti redditi da lavoro autonomo.	119
6.8	Istogramma residui standardizzati redditi da lavoro dipendente. . . .	120
6.9	Grafico di probabilità normale dei redditi da lavoro dipendente. . . .	120
6.10	Grafico residui standardizzati contro valori predetti redditi da lavoro dipendente.	121

Indice delle tabelle

3.1	Dati osservati.	42
3.2	Imputazioni multiple con meccanismo ignorabile e non ignorabile. . .	43
3.3	<i>Dataset</i> imputati.	44
3.4	Stime rapporto e associate varianze delle stime per i <i>dataset</i> completi.	44
3.5	Stime combinate e varianze per il <i>dataset</i> imputato più volte.	45
5.1	Campione della RTFL di aprile 1989 in Lombardia e campione dell'indagine suppletiva.	84
5.2	Risultati accoppiamento delle informazioni raccolte nelle due indagini sullo stesso campione.	85
5.3	Collaborazione delle famiglie. Famiglie contattate ed intervistate. . .	86
5.4	Distribuzioni delle famiglie rispondenti e non rispondenti per numero di componenti.	88
5.5	Distribuzioni delle famiglie rispondenti e non rispondenti per età del capofamiglia.	88
5.6	Distribuzione delle famiglie rispondenti e non rispondenti per titolo di studio del capofamiglia.	89
5.7	Distribuzioni delle famiglie rispondenti e non rispondenti per provin- cia e partecipazione all'indagine.	90
5.8	Collaborazione individui. Individui contattati e intervistati.	90
5.9	Distribuzioni degli individui nati prima del 31.5.1975 per sesso e partecipazione all'indagine.	91

5.10	Distribuzioni degli individui nati prima del 31.5.1975 per età e partecipazione all'indagine.	91
5.11	Distribuzioni degli individui nati prima del 31.5.1975 per titolo di studio e partecipazione all'indagine.	92
5.12	Dati mancanti per i quesiti relativi ai redditi dei lavoratori autonomi.	96
5.13	Dati mancanti per i quesiti relativi ai redditi dei lavoratori dipendenti.	97
6.1	Dati mancanti redditi mensili e annui.	100
6.2	Lavoratori autonomi e alle dipendenze.	101
6.3	Indici descrittivi dei redditi medi mensili per autonomi e dipendenti.	103
6.4	Indici descrittivi dei logaritmi dei redditi medi mensili per autonomi e dipendenti.	105
6.5	Variabili esplicative: tipo e numero di dati mancanti.	109
6.6	Pattern dati mancanti per lavoratori autonomi.	112
6.7	Pattern dati mancanti per lavoratori dipendenti.	112
6.8	Passi seguiti per stimare la funzione di regressione. Lavoratori autonomi.	114
6.9	Passi seguiti per stimare la funzione di regressione. Lavoratori dipendenti.	114
6.10	Stime coefficienti di regressione lavoratori autonomi.	116
6.11	Stime coefficienti di regressione lavoratori dipendenti.	117
6.12	Stime parametri modello MIMIC lavoratori autonomi.	125
6.13	Stime parametri modello MIMIC lavoratori dipendenti.	126
6.14	Risultati procedure di imputazione: autonomi I.	127
6.15	Risultati procedure di imputazione: autonomi II.	127
6.16	Risultati procedure di imputazione: dipendenti I.	128
6.17	Risultati procedure di imputazione: dipendenti II.	128

Introduzione

La conoscenza della distribuzione quantitativa del reddito personale è importante dal punto di vista economico, sociale e fiscale. Le fonti utilizzate di solito da politici ed economisti per ottenere informazioni sui redditi sono di natura fiscale o amministrativa, ma non sono facilmente disponibili e spesso sono incomplete e inattendibili.

Uno strumento alternativo per rilevare il reddito è costituito dalle indagini di tipo campionario, che consentono un maggior controllo della qualità delle informazioni e una maggiore ricchezza delle stesse. Tramite indagini campionarie è possibile rilevare assieme ai redditi, siano essi da lavoro, da trasferimenti o da investimenti, anche informazioni di carattere socio-demografico sui percettori.

Tra i problemi che si possono riscontrare quando si dispone di dati da indagine campionarie vi è quello delle mancate risposte. Nel caso di variabili come il reddito tale problema diviene ancor più rilevante in quanto vi è una naturale e ben nota reticenza a fornire informazioni delicate quali sono quelle sui redditi ed inoltre è ragionevole assumere che la probabilità di ottenere l'informazione sia legata all'ammontare del reddito stesso. Ad esempio, redditi elevati e bassi sono rilevati con minore probabilità di redditi medi. Gli individui che non forniscono l'informazione richiesta (non rispondenti) sono dunque diversi per molti aspetti dagli individui che invece la forniscono (rispondenti) e ciò porta, se si utilizzano solo i dati disponibili, a distorsioni nelle stime delle medie o dei totali delle variabili di interesse.

Una delle strategie per trattare il problema delle mancate risposte in indagini

campionarie è quello di ricorrere a tecniche di imputazione che consistono nel sostituire al dato mancante un valore scelto in modo opportuno.

Al fine di superare i limiti propri di alcune tra le più comuni tecniche di imputazione (*hot-deck*, *cold-deck*, imputazione di medie) è stata di recente proposta l'imputazione multipla che sostituisce al dato mancante un vettore di $M \geq 2$ valori accettabili. Si possono così ottenere M *dataset* completi sostituendo di volta in volta ad ogni valore mancante le componenti del vettore di imputazioni. Questi M *dataset* vengono poi analizzati con le usuali tecniche per dati completi e le inferenze relative ad ognuno di essi vengono combinate in modo da riflettere propriamente la variabilità campionaria dovuta alla mancata risposta.

L'imputazione multipla trova giustificazione in ambito bayesiano. Le imputazioni sono infatti estrazioni casuali dalla distribuzione a posteriori dei valori mancanti, condizionatamente ai valori osservati e al valore dei parametri.

Obiettivo di questo lavoro è la verifica delle potenzialità dell'imputazione multipla nell'analisi di un insieme di dati di reddito raccolti in un'indagine suppletiva alla Rilevazione Trimestrale sulle Forze di Lavoro (RTFL). Tale rilevazione, svoltasi in Lombardia nella primavera del 1989, aveva lo scopo di ricostruire le storie lavorative dei componenti delle famiglie e di integrare i dati sulla partecipazione al lavoro con informazioni sui redditi da lavoro e sull'insieme degli altri redditi personali e familiari. I quesiti relativi ai redditi individuali hanno una percentuale abbastanza elevata di mancate risposte, non tanto per i redditi da lavoro dipendente, attorno al 2%, quanto per i redditi da lavoro autonomo, circa il 16%. Per compensare le mancate risposte ai quesiti sul reddito vengono applicati diversi metodi di imputazione multipla e i risultati vengono confrontati tra loro.

Si vuole sottolineare la novità dell'approccio seguito. In Italia l'imputazione multipla non è mai stata applicata in indagini campionarie di una certa dimensione. In questo lavoro si vuole dimostrare che tale tecnica, ampiamente utilizzata negli Stati Uniti, dà buoni risultati soprattutto per quanto concerne la stima della variabilità delle stime e non comporta grossi problemi di implementazione.

Una traccia del contenuto del lavoro è la seguente. Nel primo capitolo vi è una breve introduzione al problema della mancata risposta nelle indagini campionarie, ai tipi di mancata risposta e ai modi per limitarla. Sono poi forniti alcuni risultati teorici riguardo al processo generatore dei dati mancanti; vengono date le condizioni che permettono di ignorare tale meccanismo quando si analizzano i dati, sia ponendosi nell'ambito dell'inferenza basata sulla randomizzazione che nell'ambito dell'inferenza per popolazioni finite basata sui modelli.

Nel secondo capitolo vengono esaminati i metodi proposti in letteratura per compensare le mancate risposte; particolare enfasi viene data all'imputazione e ai suoi vantaggi e svantaggi.

Nel terzo capitolo viene proposta l'imputazione multipla, che viene indicata come la tecnica in grado di risolvere i problemi dell'imputazione singola. Dopo un esempio numerico viene fornita la giustificazione teorica di tale procedura e altri risultati utili per la sua applicazione.

Nel quarto capitolo si discutono brevemente i problemi connessi alla rilevazione campionaria dei dati di reddito, le caratteristiche dei non rispondenti e le possibili soluzioni per ottenere informazioni migliori. Vengono poi brevemente esaminate le principali indagini che rilevano i redditi sia in Italia che all'estero e le soluzioni proposte in letteratura per trattare le mancate risposte.

Nel quinto capitolo, dopo aver accennato alla Rilevazione Trimestrale sulle Forze di Lavoro, si introduce l'indagine scelta per condurre l'analisi empirica, l'indagine suppletiva alla RTFL. Oltre ad informazioni di tipo tecnico sull'indagine, vengono fornite indicazioni sulle dimensioni e sulle caratteristiche delle mancate risposte.

Nell'ultimo capitolo si esaminano alcuni metodi di imputazione multipla per compensare le mancate risposte. Nelle Conclusioni viene fornita un'analisi critica dei risultati di tali tecniche e i possibili sviluppi del lavoro.

Nella Bibliografia vengono riportati solamente i riferimenti bibliografici citati nel testo.

Nella Bibliografia Annotata che segue vi sono invece i più importanti contributi della letteratura al problema dei dati mancanti. L'enfasi è su dati mancanti in indagini campionarie e sui metodi per compensarle: imputazione e ponderazione. Riguardo agli altri argomenti, dati mancanti nelle serie storiche, dati mancanti nell'analisi degli esperimenti solo per citarne alcuni, vi sono i principali riferimenti discretamente aggiornati.

La Bibliografia Annotata vuole essere una guida per il ricercatore che si accosta per la prima volta al problema, ma anche per coloro che hanno un problema specifico di mancata risposta e cercano gli strumenti per risolverlo.