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

Indice delle figure

Indice delle tabelle

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 riposta 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.