Om register och imputering av binära variabler

av

Thomas Laitila[1,2], Anders Holmberg[1], Emma Snölilja[1]

[1]Statistiska Centralbyrån, SE-701 89 Örebro
[2]Handelshögskolan, Örebro universitet, SE-701 82 Örebro

Preliminär version: 2010-06-21

1. Introduktion

Ett flertal anledningar ligger bakom statistikbyråernas ökande intresse för användning av register och administrativa data vid statistikproduktion (Wallgren och Wallgren, 2007). Kostnadsbesparingar och minskad uppgiftslämnarbörda är två förväntade effekter av en ökad användning av sådana data. En tredje förväntan är minskande produktionstider och ökad aktualitet i publicerad statistik. Cerroni, Migliardo och Morganti (2010) presenterar en utvärdering av ISTATs företagsregister som bl.a. indikerar snabbare publicering av statistik.

Fastän insamling av data från register och administrativa källor skiljer sig från insamling av data vid urvalsbaserade undersökningar, så finns ett antal gemensamma problem och felkällor. Två gemensamma och väsentliga problem är bortfall och mätfel. Hantering av bortfall och mätfel kan göras enligt två strategier. I den första används estimatorer som tar hänsyn till bortfall och mätfel. Exempel på sådana estimatorer är kalibreringsestimatorn (Särndal och Lundström, 2005) och ML estimation via tillämpning av EM algoritmen (Dempster, Laird och Rubin, 1977). Ett annat exempel är Ilves och Laitila (2009) som föreslår en biaskorrigerad estimator vid mätfel. I den andra strategin anpassas data så att ordinare estimationsförfarande kan användas, d.v.s. imputering för bortfall och mätfel. Ett stort antal olika imputeringstekniker finns föreslagna i litteraturen. De kan klassificeras efter datakälla för imputering, om parametrisk eller icke parametrisk metod används, och om randomisering används eller inte. Notera att de två strategierna kan kombineras vilket bl.a.

Särndal och Lundström (2005) föreslår, där variabelbortfall hanteras med imputering och objektsbortfall hanteras med kalibrering.

Ett problem med imputering är dess effekt på uppskattningar av estimatorernas varians. Vid deterministisk imputering med medelvärdet över tillgängliga observationer underskattas variansen. Ett sätt att försöka återspegla variationen i den studerade variabeln och korrigera för underskattning är att tillämpa randomiserad imputering, d.v.s. att istället för imputering av ett förväntat värde imputeras ett slumptal draget från en skattad fördelning. En ansats för att skatta variansen hos estimatorer baserade på imputerade data är Multipel Imputation (MI) (Rubin, 1989). Vid MI genereras flera datamängder med olika randomiserade imputationer av bortfallet. Den extra variationen kan mätas via variationen hos skattningarna över datamängderna. En viktig aspekt på teorin för MI behandlas av Björnstad (2007) som utvecklar MI ansatsen för tillämpning vid officiell statistikproduktion.

Denna artikel bygger på resultat i Laitila (2010) och behandlar problemet med imputering av binära variabler för bortfall när registerdata används för skattning av populationstotaler. En viktig utgångspunkt i analysen är utgångspunkterna i teorin för design-baserad inferens, (t.ex. Särndal, Swenson and Wretman, 1992), där populationens objekt och deras egenskaper ses som fixa enheter. Resultaten visar att randomiserad imputering ger sämre precision i skattningar jämfört med deterministisk imputering och, att randomisering i sig ger ingen information om skattningarnas precision.

2. Bortfall av en binär variabel

Betrakta skattning av en populationstotal av en binär variabel, d.v.s. en variabel som antar värdet ett eller noll. Populationen betecknas med U, vilken för enkelhets skull antas motsvara registerpopulationen. Den binära variabeln betecknas med y och mängden $U_y \subseteq U$ betecknar de individer i registret för vilka det finns data på variabeln y. $\overline{U}_y$

betecknar komplementmängden till U_y avseende populationen U, d.v.s. $\overline{U}_y$ innehåller de individer för vilka data saknas för variabeln y. Antalet enheter i U respektive U_y betecknas med N and N_y.

Den populationstotal som skattas är $t_y = \sum_U y_k = \sum_{U_y} y_k + \sum_{\overline{U}_y} y_k$. Imputerade värden betecknas med $\hat{y}_k$ och den imputeringsbaserade estimatorn av populationstotalen t_y är

$$\hat{t}_y = \sum_{U_y} y_k + \sum_{\overline{U}_y} \hat{y}_k \tag{1}$$

Vid randomiserad imputering, antag att imputerade värden genereras från oberoende bernoullifördelninger enligt $\hat{y}_k \sim Bern(\pi_k)$, $k \in \overline{U}_y$. Här kan π_k vara en konstant eller en funktion definierad på tillgänglig hjälpinformation. Den randomiserade imputeringsestimatorn betecknas med $\hat{t}_{yR}$ och har väntevärdet

$$E(\hat{t}_{yR}) = \sum_{U_y} y_k + \sum_{\overline{U}_y} \pi_k \tag{2}$$

och variansen

$$V(\hat{t}_{yR}) = \sum_{\overline{U}_y} \pi_k (1 - \pi_k) \tag{3}$$

Definiera den deterministiska imputeringsestimatorn enligt

$$\hat{t}_{yD} = \sum_{U_y} y_k + \sum_{\overline{U}_y} \pi_k \tag{4}$$

Via definition av enpunktsfördelningar för imputerade värden erhålls väntevärdet

$$E(\hat{t}_{yD}) = \sum_{U_y} y_k + \sum_{\overline{U}_y} \pi_k \tag{5}$$

och variansen noll, d.v.s. $V(\hat{t}_{yD}) = 0$.

Estimatorerna $\hat{t}_{yR}$ och $\hat{t}_{yD}$ har samma väntevärde och bias

$$B(\hat{t}_{yR}) = B(\hat{t}_{yD}) = \sum_{\overline{U}_y} (\pi_k - y_k) \tag{6}$$

Notera att bias begränsas till intervallet

$$\sum_{\overline{U}_y}(\pi_k - 1) \le B(\hat{t}_y) \le \sum_{\overline{U}_y} \pi_k$$

där $\overline{N}_y = N - N_y$. Intervallets längd är $\overline{N}_y$ och med $\pi_k = 0.5$ centreras intervallet kring 0. Eftersom estimatorerna har samma bias följer att $\hat{t}_{yD}$ har mindre MSE (Mean Squared Error) än $\hat{t}_{yR}$, d.v.s.

$$MSE(\hat{t}_{yD}) < MSE(\hat{t}_{yR})$$

Vid skattningar av populationsparametrar är det brukligt att illustrera skattningarnas osäkerhet m.h.a. konfidensintervall. Variansen hos den randomiserade estimatorn $\hat{t}_{yR}$ ges av ekvation (3) och ett konfidensintervall kan bildas enligt

$$\hat{t}_{yR} \pm 1.96 \cdot \sqrt{\sum_{\overline{U}_y} \pi_k (1 - \pi_k)} \qquad (7)$$

Källan till variation i $\hat{t}_{yR}$ är det ”slumpmässiga urvalet” av värden från fördelningarna $\hat{y}_k \sim Bern(\pi_k)$. Intervallet (7) illustrerar därför osäkerheten hos $\hat{t}_{yR}$ som estimator av det kända värdet $E(\hat{t}_{yR}) = \sum_{U_y} y_k + \sum_{\overline{U}_y} \pi_k = \hat{t}_{yD}$, inte som estimator av t_y.

3. Uppskattning av antal svenska arbetspendlare till Norge

Snölilja (2010) studerar egenskaper hos personer som arbetspendlar till Norge från svenska gränskommuner i västra Svealand och nord-västra Götaland. Redovisning av inkomststatistiken problematiseras av att uppgifter om inkomster från Norge blir tillgängliga efter publicering av den svenska inkomststatistiken, vilket inför en underskattning av de totala inkomsterna i gränskommunerna. I Snölilja (2010) prövas en ansats där en modell för arbetspendling utvecklas baserat på data för tidigare inkomstår, varefter modellen används för uppskattning av arbetspendling innevarande år. I hennes arbete används inkomststatistik från 2006 för utveckling av modell, varefter modell och skattning utvärderas med inkomststatistik för 2007. Analysen avgränsas till kommunerna Strömstad, Årjäng och Eda.

Baserat på data från 2006 skattas en logistisk regressionsmodell för variabeln

$$y_k = \begin{cases} 1 & \text{om individ k har inkomst från Norge 2006} \\ 0 & \text{i annat fall} \end{cases}$$

Den skattade modellen appliceras på data från 2007 och sannolikheter $\pi_k = \Pr(y_k = 1)$beräknas enligt den skattade modellen. För 2007 beräknas två deterministiska imputationsskattningar:

$$\hat{t}_{yD1} = \sum_{U_y} y_k + \sum_{\overline{U}_y} 1(\pi_k \geq 0.5)$$

$$\hat{t}_{yD2} = \sum_{U_y} y_k + \sum_{\overline{U}_y} \pi_k$$

Beräknade skattningar presenteras i Tabell 1.

Den första imputeringsestimatorn underskattar antalet pendlare kraftigt. Om en mindre andel arbetspendlar kan fördelningen av $\pi_k = \Pr(y_k = 1)$ över populationen förväntas vara skev mot små värden. En tröskelgräns på 0.5 ger därmed en underskattning av antalet pendlare. I extrema fall kan en sådan tröskelgräns ge uppskattningar på noll arbetspendlare. Ett alternativ är att sänka tröskelgränsen från 0.5 till ett mindre tal. Ett annat alternativ är att använda den estimator som föreslås i ekvation (4). I exemplet har den estimatorn en liten bias, -10%, fastän den modell som används för beräkning av imputeringsvärden $\hat{y}_k = \pi_k$ är baserad på en ”cold deck” ansats med data från ett föregående år.

I tabell 1 inkluderas även en skattning baserad på randomiserad imputering. För denna realisering är bias på -11%. Estimatorns bias är dock densamma som för $\hat{t}_{yD2}$ enligt (6). Variansen för estimatorn med randomiserad imputering är $V(\hat{t}_{yR}) = \sum_{\overline{U}_y} \pi_k(1-\pi_k) = 596.9$, vilket ger ett litet högre MSE vid randomiserad imputering jämfört med deterministisk imputering. I detta exempel domineras MSE av bias. Ett 95% KI enligt (7) ger intervallet 2075 ± 47.9, vilket inkluderar $\hat{t}_{yD2}$. Däremot inkluderas inte populationstotalen $t_y = 2342$ i konfidensintervallet.

Tabell 1: Registrerat och skattat antal personer med inkomst från Norge 2007.

Estimator/Register	Skattning/Värde	Relativt Bias	MSE
$\hat{t}_{yD1}$	1580	-33%	580644
$\hat{t}_{yD2}$	2110	-10%	53824
$\hat{t}_{yR}$ [a)]	(2075) [a)]	-10% [b)]	54421 [b)]
Register (t_y)	2342 [c)]	---	---

[a)] En realisering av estimatorn med randomiserad imputering. [b)] Bias och MSE för estimatorn $\hat{t}_{yR}$. [c)] Värde enligt SCBs inkomst och taxeringsregister.

Referenser

Björnstad, J.F. (2007). Non-Bayesian multiple imputation, *Journal of Official Statistics*, **23:4**, 433-452.

Cerroni, F, Migliardo, S. and E. Morganti (2010). Quality evaluation analysis of the Italian business register on enterprise groups. Paper presented at Q2010, Helsinki, 3-6 May, 2010.

Dempster, A.P, Laird, N.M. and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.

Ilves, M. and T. Laitila (2009). Probability-Sampling Approach to Editing, *Austrian Journal of Statistics*, **38(3)**, 171-182.

Laitila, T. (2010). On imputation of binary variables in registers, Mimeo, Statistics Sweden.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Snölilja, E. (2010). *Inkomststatistik och pendling – Prediktion av arbetspendlare till Norge.* Kandidatuppsats i statistik, Örebro universitet.

Särndal, C.-E., Swensson, B. och J. Wretman (1992). Model Assisted Survey Sampling, Springer, New York.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*, Wiley, Chichester, England.

Wallgren, A. och B. Wallgren (2007). *Register-based Statistics*, Wiley, Chichester.