PCA/MIH
Peter Werner

Bilaga 2

Kalibreringsrapport

1 Inledning

I en urvalsundersökning är alltid skattningarna behäftade med *urvalsfel* beroende på att endast en delmängd (urval) av populationen studeras. Ett annat fel uppkommer om vi inte lyckas få svar från alla personer (bortfall) och om de avviker från de svarande med avseende på undersökningsvariablerna. Detta fel kallas för *bortfallsfel.*

För att underlätta användningen av statistiken är det värdefullt om storleken på felen kan uppskattas. Av nämnda feltyper är det endast storleken på urvalsfelet som kan skattas med hjälp av urvalsinformation. Kunskap om bortfallsfelet kan i regel bara fås på ett indirekt och approximativt sätt genom att utnyttja registervariabler.

Både urvalsfel och bortfallsfel kan reduceras genom att använda ett effektivt uppräkningsförfarande. I följande avsnitt redovisas hur det görs i denna undersökning.

2 Hjälpinformation

Viss hjälpinformation utnyttjas vanligtvis även före estimationen, t.ex. för bildande av stratifierade urvalsdesigner. Det kan dock finnas ytterligare hjälpinformation som är effektiv i estimationen.

Det centrala arbetet för att få god kvalitet på skattningarna, då kalibreringsestimatorn används, är att använda ”stark” hjälpinformation. I följande avsnitt beskrivs detta arbete för denna undersökning.

2.1 Tänkbara hjälpvariabler

Vid val av hjälpvariabler är det tre kriterier som ska beaktas (se Lundström och Särndal 2001):

- Det <u>första</u> kriteriet är att variabeln samvarierar väl med svarsbenägenheten (-sannolikheten). Det är det viktigaste kriteriet eftersom det leder till en minskning av bortfallsskevheten för alla skattningar.

- Det <u>andra</u> kriteriet är att variabeln samvarierar väl med (viktiga) målvariabler. Om så är fallet minskar bortfallsskevheten för de skattningar som byggs upp av dessa målvariabler. Även variansen minskar för dessa skattningar.

Bilaga 2

- Det tredje kriteriet är att variabeln avgränsar (viktiga) redovisningsgrupper. Det leder framförallt till minskad varians i skattningar för dessa redovisningsgrupper.

I en undersökning med ett stort antal frågor av skiftande karaktär är det främst kriterierna (i) och (iii) som kan beaktas. Eftersom det i denna undersökning rör sig om två olika typer av populationer man vill skatta diverse storheter i, har två olika uppsättningar av hjälpvariabler utnyttjats: en för barnpopulationerna och en för föräldrapopulationerna.

Tänkbara hjälpvariabler, det vill säga variabler som tros uppfylla ovanstående kriterier, hämtades ifrån RTB (Registret över totalbefolkningen) och Utbildningsregistret. Hjälpvariablerna är definierade enligt tabell 1 (barnpopulationerna) och tabell 2 (föräldrapopulationerna).

Tabell 1. Tänkbara hjälpvariabler, barnpopulationerna

Variabel (benämning)	**Kategorier (koder)**
ANT (=antal vårdnadshavare)	1 = ingen vårdnadshavare 2 = en vårdnadshavare 3 = två vårdnadshavare
FOD (=vårdnadshavarnas födelseländer)	1 = ingen v.h. född utanför Sverige 2 = minst en v.h. född utanför Sverige
UTB (=högsta utbildningsnivå bland vårdnadshavarna)	1 = Förgymnasial (inkl. okänd/saknas) 2 = Gymnasial 3 = Eftergymnasial
INK (=hushållets årsinkomst, kr)	1 = 0-399 999 2 = 400 000-599 999 3 = 600 000-
REG (=barnets bostadsregion)	1 = Stockholm, Göteborg, Malmö 2 = Övriga kommuner >= 50 000 invånare 3 = Övriga kommuner < 50 000 invånare

Bilaga 2

Tabell 2. Tänkbara hjälpvariabler, föräldrapopulationerna

Variabel (benämning)	**Kategorier (koder)**
KON (=förälderns kön)	1 = man 2 = kvinna
CIV (=förälderns civilstånd)	1 = Gift/registrerat partnerskap 2 = Övriga
FOD (=förälderns födelseland)	1 = Sverige 2 = Övriga världen
UTB (=förälderns utbildnings-nivå)	1 = Förgymnasial (inkl. okänd/saknas) 2 = Gymnasial 3 = Eftergymnasial
INK (=förälderns årsinkomst, kr)	1 = 0-199 999 2 = 200 000-299 999 3 = 300 000-
REG (=förälderns bostads-region)	1 = Stockholm, Göteborg, Malmö 2 = Övriga kommuner >= 50 000 invånare 3 = Övriga kommuner < 50 000 invånare

I följande avsnitt analyserar vi variablerna i tabell 1-2 för att slutligen bestämma hjälpvektorer.

3 Analys av hjälpinformation

3.1.1 Kriterium 1: Variabeln samvarierar med svarsbenägenheten

För att se huruvida hjälpvariablerna uppfyller det första kriteriet, studeras sambandet mellan den dikotoma variabeln svarande/bortfall och hjälpvariablerna. Det görs genom att beräkna andel svarande i olika grupper, bestämda av respektive hjälpvariabel. Vid stora skillnader mellan svarsandelarna utgör variabeln en stark kandidat till hjälpvariabel.

Först analyseras hjälpvariablerna för barnpopulationerna. Tabell 3-7 visar som exempel svarsandelar inom gruppen barn 9-12 år.

Tabell 3 Andel svarande barn fördelat på antal vårdnadshavare

	1	2	3
Svarsandel (%)	–	28,4	43,2

Tabell 4 Andel svarande barn fördelat på vårdnadshavarnas födelseländer

	1	2
Svarsandel (%)	44,6	36,0

Bilaga 2

Tabell 5 Andel svarande barn fördelat på högsta utbildningsnivå

	1	2	3
Svarsandel (%)	33,8	38,5	49,8

Tabell 6 Andel svarande barn fördelat på hushållsinkomst

	1	2	3
Svarsandel (%)	27,5	45,3	49,4

Tabell 7 Andel svarande barn fördelat på bostadsregion

	1	2	3
Svarsandel (%)	39,8	46,3	41,3

Tabellerna 3-7 visar att samtliga hjälpvariabler, möjligen med undantag för region, är starka beträffande kriterium 1. Exempelvis är svarsbenägenheten betydligt högre hos barn med två vårdnadshavare (43,2 %) än hos barn med en vårdnadshavare (28,4 %).

Nedan analyseras hjälpvariablerna för föräldrapopulationerna. Tabell 8-13 visar som exempel svarsandelar inom gruppen föräldrar till barn 5-8 år.

Tabell 8 Andel svarande föräldrar fördelat på kön

	1	2
Svarsandel (%)	44,6	43,7

Tabell 9 Andel svarande föräldrar fördelat på civilstånd

	1	2
Svarsandel (%)	46,0	40,9

Tabell 10 Andel svarande föräldrar fördelat på födelseland

	1	2
Svarsandel (%)	46,9	34,2

Tabell 11 Andel svarande föräldrar fördelat på utbildningsnivå

	1	2	3
Svarsandel (%)	31,9	43,8	54,9

Tabell 12 Andel svarande föräldrar fördelat på inkomst

	1	2	3
Svarsandel (%)	33,5	42,0	52,2

Bilaga 2

Tabell 13 Andel svarande föräldrar fördelat på bostadsregion

	1	2	3
Svarsandel (%)	45,3	46,1	41,1

Tabellerna 8-13 visar att hjälpvariablerna, med undantag för kön och möjligen också region, är starka beträffande kriterium 1. Exempelvis är svarsbenägenheten betydligt högre hos föräldrar med eftergymnasial utbildningsnivå (54,9 %) än hos föräldrar med förgymnasial utbildningsnivå (31,9 %).

3.1.2 Kriterium 3: Variabeln avgränsar (viktiga) redovisningsgrupper

Om hjälpvariabeln avgränsar viktiga redovisningsgrupper kan kvaliteten bli bättre i dessa grupper med avseende på skattningarnas urvalsfel.

Hjälpvariabeln förälderns kön avgränsar viktiga redovisningsgrupper i föreliggande undersökning.

3.2 Slutligt val av hjälpvektor

Efter en sammanvägning av analysen kring ovanstående kriterier samt efter kontroll av vikternas fördelning används följande hjälpvektor för respektive barnpopulation:

ANT + FOD + UTB + INK

För respektive föräldrapopulation används följande hjälpvektor:

KON + CIV + FOD + UTB + INK

4 Teknisk beskrivning av urval och estimation

Vi har en population *U* bestående av *N* personer. De parametrar vi är intresserade av är vanligtvis funktioner av två totaler $Y = \sum_U y_k$ och $Z = \sum_U z_k$, där y_k är värdet på variabel *y* för person *k* och z_k värdet på en annan variabel för samma person. Vanligtvis är *y* (och även *z*) en dikotom variabel, d.v.s.

$$y_k = \begin{cases} 1 \text{ om person } k \text{ har studerade egenskap} \\ 0 \text{ för övrigt} \end{cases} \qquad (4.1)$$

Bilaga 2

Vanligtvis är vi också intresserade av parametrar för redovisningsgrupper. Låt oss benämna dessa $U_1,...,U_d,...,U_D$, där $U = \bigcup_{d=1}^{D} U_d$. Totalen för redovisningsgrupp *d* kan skrivas

$$Y_d = \sum_U y_{dk} \tag{4.2}$$

där $y_{dk} = \begin{cases} y_k & \text{för } k \in U_d \\ 0 & \text{för övrigt.} \end{cases}$.

Z_d bildas på likartat sätt.

En generell parameter för redovisningsgrupp *d* (*d* kan också avse hela populationen) kan skrivas $\theta_d = C\frac{Y_d}{Z_d}$, där *C* är en konstant.

Den vanligaste parametern är en procentuell andel, som erhålles när $C = 100$ och $z_k = 1$ för alla k, och *y* är definierad enligt (4.1). Om vi låter N_d vara antalet personer i redovisningsgrupp *d*, då kan parametern skrivas

$$P_d = 100\frac{\sum_U y_{dk}}{N_d} \tag{4.3}$$

Vi drar ett obundet slumpmässigt urval (OSU) s av storleken n från populationen *U*, men p.g.a. övertäckning och bortfall har vi endast svarsmängden r av storleken m att utföra beräkningarna på.

Den ”konventionella” estimatorn (för Y_d), har då följande form:

$$\hat{Y}_d = \frac{N}{m}\sum_r y_{dk} \tag{4.4}$$

I estimator (4.4) används ingen hjälpinformation.

I syfte att erhålla en estimator med mindre urvalsfel och bortfallsskevhet än estimator (4.4) utnyttjar vi hjälpinformation i estimationen. Vi bildar en hjälpvektor $\mathbf{x}_k$, som anger till vilka kategorier av hjälpvariablerna som person *k* hör. Från RTB och Utbildningsregistret framställer vi hjälptotalerna $\sum_{U_d} \mathbf{x}_k$. Vi utnyttjar denna hjälpinformation i en kalibreringsestimator.

Kalibreringsestimatorn för totalen Y_d har följande utseende:

Bilaga 2

$$\hat{Y}_{wd} = \sum_r d_k^* g_k y_{dk} \qquad (4.5)$$

där

$d_k^* = d_k \cdot f_k = 1/\left(\pi_k \hat{\theta}_k\right)$ för $k \in r$,

så att w_k = total vikt för objekt *k*
π_k = inklusionssannolikhet för objekt *k*
$\hat{\theta}_k$ = skattad svarssannolikhet där det antas att personer svarar med samma sannolikhet och oberoende av varandra
$d_k = (1/\pi_k) =$ designvikt
$f_k = (1/\hat{\theta}_k) =$ bortfallsvikt
$g_k =$ justeringsfaktor som baseras på hjälpinformationen

och

$$g_k = 1 + (\sum_U \mathbf{x}_k - \sum_r d_k^* \mathbf{x}_k)' \left(\sum_r d_k^* \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \qquad (4.6)$$

Vid skattning av en parameter av typen $\theta_d = C\frac{Y_d}{Z_d}$ skattas respektive total med hjälp av kalibreringsvikterna $d_k^* g_k$.

Anmärkning: Den tekniska beskrivningen ovan gäller estimation av storheter i barnpopulationerna. Urvalet av föräldrar har dock i denna undersökning dragits i två steg. I det första steget drogs ett OSU av barn och i det andra steget drogs för varje utvalt barn ett OSU av en förälder. Designvikterna är i detta fall konstruerade så att hänsyn tas till att föräldrar med många barn har större sannolikhet att komma med i urvalet än föräldrar med få barn, samt att föräldrar som är ensamma vårdnadshavare har större sannolikhet att komma med än föräldrar med gemensam vårdnad (givet första steget). I övrigt genomförs kalibreringen på i princip samma sätt som för barnpopulationerna.

Referenser

Lundström S. och Särndal C.-E. (2001). *Estimation in the Presence of Nonresponce and Frame Imperfection*. Stockholm: Statistics Sweden

Andersson C. och Nordberg L. (1998). *A User's Guide to CLAN 97 – a SAS-program for computation of point- and standard error estimates in sample surveys.* Statistics Sweden

Särndal, Swensson och Wretman (1992): *Model Assisted Survey Sampling.* New York: Springer Verlag.