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 undersöknings-variablerna. 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 Parametrar

De parametrar som skattas i denna undersökning är totaler och kvoter.

3 Hjälpinformation

Viss hjälpinformation utnyttjas vanligtvis även före estimationen, t.ex. för bildande av stratifierade urvalsdesigner. I denna undersökning drogs ett tvåstegs klusterurval. I första steget drogs ett stratifierat urval av kommuner och i andra steget ett urval av rektorsområden. Elever i utvalda rektorsområ-den totalundersöktes. På grund av bortfall av elever i enkätundersökningen så används hjälpinformation på elevnivå i avseende att reducera de skevheter som detta bortfall kan leda till.

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

3.1 Tänkbara hjälpvariabler

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

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

(ii) Det andra kriteriet är att variabeln samvarierar väl med (viktiga) målvariabler. Om så är fallet minskar bortfallsbiasen för de skattningar som byggs upp av dessa målvariabler. Även variansen minskar för dessa skattningar.

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

I den här undersökningen innehåller enkäten frågor av mycket skiftande karaktär. Därför är det främst punkterna (i) och (iii) som kan beaktas, vilket innebär att kalibreringen främst tjänar till att reducera den skevhet som bortfallet troligtvis ger upphov till.

Tänkbara hjälpvariabler, det vill säga variabler som tros uppfylla de ovan uppsatta kriterierna, hämtades ifrån Registret över totalbefolkningen (RTB) Utbildningsregistret (UREG) och Elevpanelernas årskurs 8 insamling. En genomgång av variablerna resulterade i att fem variabler valdes ut. De sammanslagningar av kategorier som gjorts baseras på kunskaper från tidigare kalibreringar.

Hjälpvariablerna är definierade enligt tabell 1.

Tabell 1. **Tänkbara hjälpvariabler**

Variabel (benämning)	**Kategorier (koder)**
KÖN	1 = Man 2 = Kvinna
FÖDELSELAND	1 = Född i Sverige 2 = Född i övriga Världen
HÖGSTA UTBILDNING FÖRÄLDRAR	1 = Grundskoleutbildning eller lägre 2 = Gymnasial utbildning 3 = Eftergymnasial utbildning
REGION[1]	1 = Storstäder 2 = Förortskommuner 3 = Större städer 4 = Övriga kommuner
MERITVÄRDE ÅRSKURS 8	1 = Uppgift saknas 2 = < 160 3 = 165-190 4 = > 190

Med högsta utbildning föräldrar avses högsta avslutade utbildning till och med 2007. Utbildningen avser den förälder som har högst utbildning.

Med region avses den region där individen var folkbokförd i februari 2008.

I följande avsnitt analyserar vi variablerna i tabell 1 för att slutligen bestämma en hjälpvektor.

[1] För definition av regioner, se http://www.skl.se/artikel.asp?A=11248&C=445

4 Analys av hjälpinformation

4.1 Kriterium 1: Variabeln samvarierar med svarsbenägenheten

För att se huruvida hjälpvariablerna uppfyller det första kriteriet, studeras sambandet mellan den dikotoma variabeln svarande/bortfall och hjälp-variablerna. Det görs genom att beräkna skattad andel svarande i olika grupper, bestämda av respektive hjälpvariabel. Den totala svarsandelen har skattats till 61 procent.

Vid stora skillnader mellan svarsandelarna utgör variabeln en stark kandidat till hjälpvariabel.

Tabell 2. Andel svarande fördelat på kön

Kön	Svarsandel (%)
Kvinnor	65,7
Män	56,3
Totalt	60,9

Tabell 3. Andel svarande fördelat på födelseland

Födelseland	Svarsandel (%)
Sverige	60,9
Övriga Världen	59,7
Totalt	60,9

Tabell 4. Andel svarande fördelat på region

Region	Svarsandel (%)
Storstäder	55,4
Förortskommuner	61,9
Större städer	63,1
Övriga kommuner	61,0
Totalt	60,9

Tabell 5. Andel svarande fördelat på högsta utbildning föräldrar

Utbildningsnivå	Svarsandel (%)
Grundskoleutbildning	47,6
Gymnasial utbildning	55,4
Eftergymnasial utbildning	67,8
Totalt	60,9

Tabell 6. Andel svarande fördelat på meritvärde årskurs 8

Meritvärde årskurs 8	Svarsandel (%)
Uppgift saknas	56,2
<160	54,6
165-190	59,8
>190	71,8
Totalt	60,9

Tabellerna 2-6 visar att variablerna kön, högsta utbildning föräldrar och meritvärde årskurs 8 kan anses som starka (beträffande kriterium 1). Även för variabeln region finns en tendens att elever i storstadsområden svarar i lägre grad än elever i övriga regioner.

4.3 Kriterium 3: Variabeln avgränsar (viktiga) redovisningsgrupper

Om hjälpvariabeln avgränsar viktiga redovisningsgrupper kan kvaliteten bli bättre i dessa grupper. Framförallt blir skattningarna säkrare om hjälpvariabeln väl avgränsar redovisningsgruppen.

Variabeln kön avgränsar redovisningsgrupper och bör därför vara med i hjälpvektorn om det är möjligt.

4.4 Slutligt val av hjälpvektor

I den slutliga hjälpvektorn användes alla fem hjälpvariabler. Födelseland var inte en stark variabel enligt kriterium 1. Att variabeln födelseland ändå inkluderats i hjälpvektorn motiveras med att födelseland kan vara en redovisningsgrupp av intresse.

Efter en sammanvägning av analysen kring de tre kriterierna samt efter kontroll av vikternas fördelning används följande hjälpvektor:

Kön+ Födelseland + Region + Högsta utbildning föräldrar + Meritvärde årskurs 8

5 Teknisk beskrivning av urval och estimation

Vi har en population U bestående av N personer. De parametrar vi är intresserade av är främst 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. Vi kan definiera y (och även z) som en dikotom variabel, d.v.s.

$$y_k = \begin{cases} 1 \text{ om person } k \text{ har studerade egenskap} \\ 0 \text{ för övrigt} \end{cases} \qquad (5.1)$$

Det finns givetvis också intresse av parametrar för olika 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

$$Y_d = \sum_U y_{dk} \qquad (5.2)$$

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 (5.1). Om vi låter N_d vara antalet personer i redovisningsgrupp *d*, då kan parametern skrivas

$$P_d = 100\frac{\sum_U y_{dk}}{N_d} \qquad (5.3)$$

Vi drar ett tvåstegs klusterurval s av storleken n, 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 följande form:

$$\hat{Y}_d = \sum_r w_{1k} w_{2k} w_{3k} y_{dk} \qquad (5.4)$$

där

w_{1k} = totalt antal kommuner i respektive stratum/
antal utvalda kommuner i respektive stratum

w_{2k} = totalt antal RO-områden i utvald kommun/
antal utvalda RO-områden i utvald kommun

w_{3k} = totalt antal elever i respektive RO-område/
antal elever som svarat på enkäten i respektive RO-område

I estimator (5.4) används ingen ytterligare hjälpinformation än stratifierings-informationen. Denna estimationsmetod brukar kallas ”rak uppräkning inom strata”.

I syfte att erhålla en estimator med mindre urvalsfel och bortfallsskevhet än estimator (5.4) utnyttjar vi hjälpinformation också i estimationen. Vi bildar en hjälpvektor $\mathbf{x}_k$, som anger till vilka kategorier av

Kön+ Födelseland + Region + Högsta utbildning föräldrar + Meritvärde årskurs 8

som elev *k* tillhör. Från Utbildningsregistret och Registret över totalbefolkningen framställer vi sedan hjälptotalerna $\sum_s \mathbf{x}_k$. Vi utnyttjar denna hjälpinformation i en kalibreringsestimator.

Kalibreringsestimatorn för totalen Y_d har följande utseende:

$$\hat{Y}_{wd} = \sum_r w_{1k} w_{2k} v_{3k} y_{dk} = \sum_r w_k y_{dk} \qquad (5.5)$$

där

$$v_{3k} = 1 + (\sum_s \mathbf{x}_k - \sum_r w_{3k} \mathbf{x}_k)' \left(\sum_r w_{3k} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \qquad (5.6)$$

Vid skattning av en parameter av typen $\theta_d = C\frac{Y_d}{Z_d}$ skattas respektive total med hjälp av kalibreringsvikterna $w_k = w_{1k} w_{2k} v_{3k}$.

Referenser:

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