Kalibreringsrapport

Inledning

I denna bilaga behandlas estimationen i korta drag. I en urvalsundersökning är alltid skattningarna behäftade med urvalsfel beroende på att endast en delmängd (ett urval) av populationen studeras. Ett annat fel uppkommer om vi inte lyckas få svar från alla personer och om de som inte svarar avviker från de svarande med avseende på undersökningsvariablerna. Detta fel kallas bortfallsfel.

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

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 kalibrerings-estimatorn används, är att använda ”stark” hjälpinformation.

Vid val av hjälpvariabler är det tre kriterier som ska beaktas, se Särndal och Lundström (2005).

(i) Det första 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.

(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 en undersökning som denna, med ett stort antal frågor av skiftande karaktär, är det främst kriterierna (*i*) och (*iii*) som kan beaktas. Tänkbara hjälpvariabler, det vill säga variabler som tros uppfylla dessa kriterier, hämtades från Registret över totalbefolkningen (RTB), Inkomst- och taxeringsregistret (IoT) och Utbildnings-registret (UREG).

Det finns två typer av hjälpvariabler: den ena känner vi värdet på för alla i populationen, för den andra typen så känner vi värdet endast på individer i urvalet. Av de hjälpvariabler som har valts för denna undersökning så känner vi värdet för alla individer i populationen.

Tabell 1. Hjälpvariabler

Variabel (benämning)	Kategorier	Typ
Kön	Kvinna Man	Hela populationen
Ålder	18 – 30 år 31 – 64 år 65 – 85 år	Hela populationen
Inkomst	- 149 tkr 150 – 299 tkr 300 - tkr	Hela populationen
Bakgrund	Svensk Utländsk	Hela populationen
Utbildningsnivå	Förgymnasial samt övriga Gymnasial utbildning Kortare eftergymnasial utbildning Eftergymnasial utbildning	Hela populationen
Civilstånd	Ogift Gift eller Registrerad partnerskap Skild eller Änka/änkling	Hela populationen
Stratum	Bostadsort vid ifyllande av enkäten 25 kommuner och 14 stadsdelar	Hela populationen

Följande hjälpvektor användes:

Kön + *Ålder* + *Inkomst* + *Bakgrund* + *Utbildningsnivå* + *Civilstånd* + *Stratum*

Teknisk beskrivning av estimationen

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 variabeln *y* för person *k* och z_k värdet på en annan variabel för samma person. Vanligtvis är *y* (och även *z*) en dikotom variabel, dvs.

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

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

$$Y_d = \sum_U y_{dk} \quad (2)$$

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 erhålls när $C = 100$ och $z_k = 1$ för alla k, och *y* är definierad enligt (1). Om vi låter N_d vara antalet personer i redovisningsgrupp *d*, då kan parametern skrivas

$$P_d = 100\frac{\sum_U y_{dk}}{N_d} \tag{3}$$

Vi drar ett obundet slumpmässigt urval s_h av storleken n_h från stratum h ($h = 1,...,H$), men p.g.a. övertäckning och bortfall har vi endast svarsmängden r_h av storleken m_h att utföra beräkningarna på. Storleken på stratum *h* ger vi beteckningen N_h.
Den ”konventionella” estimatorn (för Y_d), har följande form:

$$\hat{Y}_d = \sum_{h=1}^{H}\frac{N_h}{m_h}\sum_{r_h} y_{dk} \tag{4}$$

I estimator (4) används ingen ytterligare hjälpinformation än stratifieringsinformationen.

I syfte att erhålla en estimator med mindre urvalsfel och bortfallsskevhet än estimator (4) utnyttjar vi hjälpinformation också i estimationen. Vi bildar en hjälpvektor $\mathbf{x}_k$, för person *k*.

Från populationsregistren framställer vi hjälptotalerna $\sum_U \mathbf{x}_k$. Vi utnyttjar denna hjälpinformation i kalibreringsestimatorn.
Kalibreringsestimatorn för totalen Y_d har följande utseende:

$$\hat{Y}_{wd} = \sum_r d_k v_k y_{dk} \tag{5}$$

där

$$d_k = N_h / m_h \text{ för } k \in r_h$$

och

$$v_k = 1 + (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)'\left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}\mathbf{x}_k \tag{6}$$

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

Denna kalibreringsvikt uppfyller kalibreringsvillkoret: $\sum_r w_{\mathbf{k}} \mathbf{x}_k = \sum_U \mathbf{x}_k$, vilket innebär att om vikterna läggs på variabler som ingår i hjälpvektorn summeras dessa upp till de hjälptotaler vi hämtat från registren.

Kalibreringen och skattningen har gjorts med hjälp av SAS och SAS-programmet ETOS, se Andersson (2012), som är en vidareutveckling av CLAN, se Andersson och Nordberg (1998).

Referenser

Andersson, C. (2012). *ETOS 2.0 User's guide*. Statistics Sweden.

Andersson, C. och Nordberg, L. (1998). *A User's Guide to CLAN97*- a SAS-program for computation of point- and standard error estimates in sample surveys. Statistics Sweden.

Särndal, C-E. och Lundström, S. (2009). *Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias.* Research and Development – Methodology reports from Statistics Sweden, 2009:1

Särndal, C-E. och Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Wiley & Sons.

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