Stockholms universitet

Metodeffekter i urvalsundersökningar där deltagarna får välja mellan pappers- och webbenkät

Martina Aksberg

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:9
Matematisk statistik
September 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:9**,
http://www.math.su.se/matstat

Metodeffekter i urvalsundersökningar där deltagarna får välja mellan pappers- och webbenkät

Martina Aksberg*

September 2010

Sammanfattning

Denna uppsats behandlar urvalsundersökningar där individerna i stickprovet själva får välja mellan att besvara undersökningens frågor via pappers- eller webbenkät. I undersökningar där svaren som inkommit genom de två insamlingsmetoderna skiljer sig signifikant från varandra har vi att göra med en så kallad metodeffekt, vilket innebär att insamlingsmetoden systematiskt påverkar de svar som ges. I uppsatsen tar vi fram förslag på hur hänsyn till metodvalet kan tas i de populationsskattningar som görs. Förslagen riktar sig till undersökningar där stickprovet har erhållits genom obundet slumpmässigt urval, stratifierat urval, klusterurval eller tvåstegsurval. I undersökningar utan bortfall eller där bortfallet är helt slumpmässigt fokuserar vi på regressionsskattningen, men ger även ett kortfattat förslag på hur hänsyn till metodvalet kan tas om skattningarna istället görs med hjälp av prediktion. Skattningar i undersökningar där bortfallet inte kan anses vara helt slumpmässigt gör vi genom att använda oss av en metod som kombinerar viktning och imputering. Våra förslag går i samtliga fall ut på att betrakta metodvalet som en bakgrundsvariabel som har inverkan på de svar som ges. Om populationsskattningar ska kunna göras med detta synsätt krävs det att vi skattar det totala antalet personer i populationen som skulle ha valt respektive insamlingsmetod om de hade blivit uttagna till stickprovet. När regressionsskattning av undersökningsvariabeln görs utgår vi från tidigare gjorda undersökningar för att skatta det totala antalet pappers- och webbundersökningar i populationen. Vid prediktionsskattning betraktar vi istället metodvalet som en undersökningsvariabel och använder oss av de metoder som finns för att skatta totalsumman av antalet personer i populationen som skulle ha valt respektive insamlingsmetod, vid de valda urvalsmetoderna. För att göra skattningar i undersökningar där bortfallet inte är helt slumpmässigt betraktar vi återigen metodvalet som en undersökningsvariabel och gör utifrån stickprovet skattningar av det totala antalet webb- och pappersundersökningar i populationen med hjälp av viktning. Skattningarna av vilka metodval som skulle ha gjorts vid en totalundersökning utan bortfall (som erhålls med hjälp av de ovan beskrivna metoderna) sätter vi sedan in i de formler som finns för att skatta populationstotalen av undersökningsvariabeln. Vi erhåller därmed populationsskattningar där hänsyn till insamlingsmetoden tas.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post:ma.aks@hotmail.com . Handledare: Gudrun Brattström.

Abstract

This thesis deals with sample surveys in which the data is collected by giving the participants the opportunity to choose between paper-and-pencil and web-based questionnaires. In surveys where the answers received from the web respondents are significantly different from the answers received from the paper-and-pencil respondents there is a mode effect, which means that the survey mode systematically affects the answers given in the survey. In this thesis we present suggestions on how to take the choice of mode into account when estimating the population totals of the survey variables. The suggestions are applicable when the data is collected by simple random sampling, stratified sampling, cluster sampling or two-stage sampling. To make estimates in surveys without nonresponse or in surveys where data is missing completely at random (MCAR) we focus on using the regression estimator, but we also give a brief suggestion on how to make estimates when using the prediction approach. In surveys where data is missing at random (MAR) we use a method that combines weighting and imputation to make population estimates. The idea behind our suggestions is to consider the choice of mode to be a background variable that affects the answers given in the survey (i.e. the survey variables). To make population estimates (of the survey variables) from this point of view it is necessary to estimate the number of people in the whole population that would have chosen each of the two collection modes if they had been in the sample. When the regression estimator is used we estimate the total number of web and paper-and-pencil participants in the population by using data from former surveys. To make estimates when the prediction approach is used we analyze the data received in the survey by considering the choice of mode to be a survey variable. In surveys where data is missing at random (MAR) we again consider the choice of mode to be a survey variable, that is estimated for the whole population by weighting. The estimates of the total number of people that would have chosen each of the two collection modes in a census without nonresponse are put into the the formulas that estimate the population totals of the survey variables. Thus we receive population estimates that take the choice of mode into account.

Förord

Detta arbete utgör ett examensarbete om 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik vid Matematiska institutionen, Stockholms Universitet. Arbetet har genomförts på uppdrag av utvecklingsavdelningen på Statistiska centralbyrån. Först och främst vill jag rikta ett stort tack till mina båda handledare, Boris Lorenc, metodarkitekt på SCB, och Gudrun Brattström, docent vid Stockholms Universitet, som hjälpt mig genom denna uppsats. Tack också till Dan Hedlin, metodarkitekt på SCB, för värdefulla kommentarer och engagemang under arbetets gång.

Innehåll

1 Introduktion **5**

2 Problem **5**

3 Undersökningar utan bortfall **5**

3.1 Introduktion – Populationsskattningar i undersökningar utan bortfall 5
3.2 Prediktions- och kvotskattning 6
3.3 Inklusionssannolikhet och Horvitz-Thompsonskattning 8
3.4 Regressionsskattning 8
3.5 Förslag på regressionsskattning som tar hänsyn till metodeffekten i undersökningar utan bortfall vid OSU 10
3.6 Stratifierat urval 12
3.7 Förslag på regressionsskattning som tar hänsyn till metodeffekten i undersökningar med stratifierat urval utan bortfall 13
3.8 Klusterurval och tvåstegsurval 14
3.9 Förslag på regressionsskattning som tar hänsyn till metodeffekten vid kluster- och tvåstegsurval i undersökningar utan bortfall 15
3.10 Förslag för hantering av metodeffekter vid prediktionsskattning i undersökningar utan bortfall 17

4 Undersökningar med bortfall **18**

4.1 Bortfall 18
4.2 Viktning 18
4.3 Imputering 20
4.4 Den kombinerade metoden 21
4.5 Orsaker till bortfall 22
4.5.1 MCAR – ”Missing completely at random” 22
4.5.2 MAR - ”Missing at random” 22
4.5.3 NMAR – ”Not missing at random” 22
4.6 Hantering av bortfall vid MCAR 22
4.7 En översikt över Fannie Cobbens sätt att hantera metodeffekter i undersökningar med bortfall där individen blir tilldelad en insamlingsmetod 23
4.8 Kritik mot Cobbens metoder 23
4.9 Förslag på sätt att hantera datamaterial som erhållits genom OSU i undersökningar med bortfall där individen själv väljer pappers- eller webbenkät 24
4.10 Förslag på sätt att hantera datamaterial som erhållits genom stratifierat urval i undersökningar med bortfall där individen själv väljer pappers- eller webbenkät 26
4.11 Ett verkligt exempel på metodeffekter – sammanfattning av en rapport av Kaiser 27
4.12 Förslag på ett sätt att hantera Kaisers datamaterial för att göra skattningar av den ekonomiska utvecklingen hos alla tyska företag i servicesektorn 28

5 Diskussion **28**

5.1 Är det rimligt att skatta metodvalet i en population? 28

6 Slutsatser **30**

6.1 Slutsatser 30

7 Referenslista **30**

8 Appendix **31**

1 Introduktion

I statistiska undersökningar skattas ofta olika undersökningsvariabler för en hel population utifrån ett stickprov. När Statistiska centralbyrån genomför undersökningar får individerna som tagits ut till stickprovet i många fall själva välja mellan att fylla i enkäten på papper eller via webben. Ett problem med urvalsundersökningar i allmänhet är att datainsamlingsmetoden kan ha inverkan på hur individen svarar. Fenomenet kallas för metodeffekt och innebär att vilken typ av enkät eller intervjumetod som används för att samla in data påverkar de svar som ges. Undersökningar visar exempelvis att de mest extrema svarsalternativen ofta undviks vid webbundersökningar (Taylor, 2000) och att vissa insamlingsmetoder tenderar att påverka deltagarna att lämna svar som är socialt accepterade snarare än sanna (se till exempel Kreuter, Presser och Tourangeau, 2008).

Om insamlingsmetoden har inverkan på svaren som lämnas i en given undersökning bör man även ta hänsyn till att det i undersökningar där individen själv får välja mellan pappers- och webbenkät kan finnas ett samband mellan valet av insamlingsmetod och personliga egenskaper såsom kön och ålder. Man kan exempelvis misstänka att det generellt sett är så att andelen som föredrar postenkät framför webbenkät är större bland äldre personer än bland ungdomar. Om ett sådant samband finns kan det vara fel att bryta ner resultatet på de olika bakgrundsvariablerna och till exempel uttala sig om de äldre personerna i populationen. Detta eftersom metodvalen som gjorts har haft inverkan på de svar som lämnats. Att ta hänsyn till metodeffekten är även av vikt vid hantering av det (eventuella) bortfall som uppstått i undersökningen i och med att man har bakgrundsvariablerna som utgångspunkt när detta korrigeras.

2 Problem

Denna uppsats har som syfte att ta fram förslag på hur hänsyn till metodeffekten kan tas i urvalsundersökningar där svaren som lämnats genom de två insamlingsmetoderna skiljer sig signifikant från varandra. Förslagen ska gälla i undersökningar såväl med som utan bortfall och där urvalspersonerna själva väljer om de ska svara via pappers- eller webbenkät.

3 Undersökningar utan bortfall

3.1 Introduktion - Populationsskattningar i undersökningar utan bortfall

I statistiska undersökningar där man har som syfte att ta reda på någonting om en hel population görs i de flesta fall ett urval ur den population som man är intresserad av istället för att undersöka samtliga individer i populationen. Utifrån den information som erhålls från urvalet dras slutsatser om populationen i sin helhet. Ofta är man intresserad av att skatta medelvärdet eller totalsumman av någon variabel i populationen, exempelvis medellön i en viss yrkesgrupp eller det totala antalet arbetslösa personer i befolkningen. Det finns flera metoder för att skatta summan (och därmed även medelvärdet) av en undersökningsvariabel i en given population. I de tre följande avsnitten beskrivs översiktligt några vanliga metoder som kan användas i undersökningar utan bortfall eller i undersökningar där det är helt slumpmässigt vilka individer som låter bli att svara. Metoderna har vi senare som utgångspunkt när vi ställer upp modeller för att göra populationsskattningar där hänsyn till metodeffekten tas.

För att kunna uttala sig om en skattnings precision är det i praktiken av stor vikt att beräkna variansen. I denna uppsats lägger vi dock fokus på att ta fram punktskattningarna. Variansen för respektive skattning, eller i vissa fall en hänvisning till var variansen för respektive skattning kan hittas, i såväl fallet med som utan bortfall, återfinns därför i appendix.

3.2 Prediktions- och kvotskattning

I urvalsundersökningar innebär prediktion att göra skattningar av de individer i populationen som inte ingått i urvalet. Vi är intresserade av att skatta summan av variablerna y_i i en population av storlek N, det vill säga $T = \sum_{i=1}^{N} y_i$. För de individer som ingår i urvalet blir variablerna y_i kända genom undersökningen medan värdena för individerna i den övriga delen av populationen måste skattas. Antag att vi har tillgång till en bakgrundsvariabel, x_i, för samtliga individer i populationen och att vi, efter att ha ställt en fråga som ger svaret y_i till ett urval individer, vill uttala oss om (summan av) y_i för hela populationen. Exempelvis kan vi tänka oss att x_i betecknar ålder för individ i, y_i betecknar antalet sjukhusbesök det senaste året för samma person och att vi är intresserade av att skatta summan av antalet sjukhusbesök under det gångna året i den aktuella populationen. Om datamaterialet verkar passa, det vill säga om ett någorlunda linjärt samband mellan x_i och y_i verkar råda, kan vi tänka oss en regressionsmodell på formen

$$E(Y_i) = \beta x_i$$
$$Var(Y_i) = \sigma^2 x_i$$

Den bästa linjära väntevärdesriktiga skattningen av β kan i detta fall härledas till

$$\hat{\beta} = \frac{\sum_{i \epsilon s} y_i}{\sum_{i \epsilon s} x_i}$$

(Valliant, Dorfman och Royall, 2000, kapitel 1.2), där *s* betecknar urvalsgruppen.

I deskriptiva urvalsundersökningar ses skattningen av β enbart som ett led i att skatta populationstotalen. Analys av β-värdet i sig är därför inte av intresse, såsom att skatta dess standardfel eller att hitta ett konfidensintervall. I en population där gruppen personer som är med i urvalet betecknas med *s* och den övriga delen av populationen betecknas med *a* kan summan av variabeln y_i i populationen uttryckas som

$$T = \sum_{i \epsilon s} y_i + \sum_{i \epsilon a} y_i$$

, där den första summan är känd efter att undersökningen har genomförts och den andra summan måste skattas. Summan för den del av populationen som inte tillhör urvalet kan skattas med hjälp av väntevärdena för y_i, som fås ur regressionsmodellen, och blir

$$\sum_{i \epsilon a} \hat{y}_i = \sum_{i \epsilon a} \hat{\beta} x_i = \frac{\sum_{i \epsilon s} y_i}{\sum_{i \epsilon s} x_i} \sum_{i \epsilon a} x_i$$

Skattningen av summan i hela populationen blir därmed

$$\hat{T} = \sum_{i\epsilon s} y_i + \frac{\sum_{i\epsilon s} y_i}{\sum_{i\epsilon s} x_i} \sum_{i\epsilon a} x_i = \frac{\sum_{i\epsilon s} y_i}{\sum_{i\epsilon s} x_i} \sum_{i=1}^{N} x_i$$

Detta sätt att skatta totalsumman kallas för kvotskattning och kan efter omskrivning även uttryckas som

$$\hat{T} = N\bar{y}_s \frac{\bar{x}}{\bar{x}_s}$$

I de allra flesta undersökningar som görs har man dock tillgång till mer än en bakgrundsvariabel som är känd för hela populationen och där samband med undersökningsvariabeln kan upptäckas. Antag att vi även i dessa fall är intresserade av att skatta totalsumman av en viss variabel, y_i, i populationen. Om vi för samtliga individer i populationen har tillgång till p stycken olika bakgrundsvariabler blir regressionsmodellen

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta}$$
$$Var(\boldsymbol{Y}) = \boldsymbol{V}$$

, där $\boldsymbol{X}$ är en $N \times p$-matris med bakgrundsvariablerna, $\boldsymbol{\beta}$ är en $p \times 1$-vektor, $\boldsymbol{V}$ är kovariansmatrisen och $\boldsymbol{Y}$ är en $N \times 1$-vektor med responsvariabeln (Valliant et al., 2000, kapitel 2.1). Genom att arrangera om matriserna får vi

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_{ss}} & \boldsymbol{V_{sa}} \\ \boldsymbol{V_{as}} & \boldsymbol{V_{aa}} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_s} \\ \boldsymbol{X_a} \end{bmatrix}$$

$$\boldsymbol{Y} = (\boldsymbol{Y_s'}, \boldsymbol{Y_a'})$$

Valliant et al. (2000, kapitel 2.2) visar att en väntevärdesriktig skattning av totalsumman i populationen (med minimerad felvarians) blir

$$\hat{\theta} = (1, \dots, 1)\boldsymbol{Y_s} + (1, \dots, 1)[\boldsymbol{X_a}\widehat{\boldsymbol{\beta}} + \boldsymbol{V_{as}}\boldsymbol{V_{ss}^{-1}}(\boldsymbol{Y_s} - \boldsymbol{X_s}\widehat{\boldsymbol{\beta}})]$$

där

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X_s'}\boldsymbol{V_{ss}^{-1}}\boldsymbol{X_s})^{-1}(\boldsymbol{X_s'}\boldsymbol{V_{ss}^{-1}}\boldsymbol{Y_s})$$

, och där antaganden om kovariansmatrisen görs utifrån datamaterialets struktur.

I prediktionsteorin betraktas y_i som ett utfall av den stokastiska variabeln Y_i. Metoden för att göra skattningar av hela populationen går ut på att hitta en funktion med hjälp av vilken värden, y_i, för de element som inte ingått i stickprovet kan predikteras (Valliant et al., 2000, kapitel 1.2) och det finns inga krav på att urvalet ska ha gjorts slumpmässigt även om vissa urvalsmetoder är olämpliga (Valliant et al., 2000, kapitel 2). Ett annat synsätt presenteras i Särndal, Swensson och Wretman (1992) där y_i istället betraktas som konstanter och där sannolikheten för att ett visst element ska väljas ut till stickprovet är centralt i de skattningar som görs (Valliant et al., 2000, kapitel 2.7). I denna uppsats utgår vi främst från teorin i Särndal et al. (1992), som beskrivs i kapitel 3.3 - 3.4 nedan. Några kortfattade förslag på hur metodeffekten kan hanteras vid prediktionsskattning presenteras dock i kapitel 3.10.

3.3 Inklusionssannolikhet och Horvitz-Thompsonskattning

Antag att vi har en population med N element och att ett stickprov s dras från populationen. En viss urvalsmetod är vald så att sannolikheten att välja just stickprov s kan beräknas till ett givet värde, $p(s)$. Vi ställer upp en indikatorvariabel, I_k, som anger om element k i populationen ingår i det valda stickprovet eller inte,

$$I_k = \begin{cases} 1 \; om \; k \epsilon s \\ 0 \; annars \end{cases}$$

Sannolikheten för att element k ingår i stickprovet kallas för inklusionssannolikheten och är

$$\pi_k = \Pr(k \epsilon s) = \Pr(I_k = 1) = \sum_{k \epsilon s} p(s)$$

(Särndal et al., 1992, kapitel 2.4), eller uttryckt med ord, summan av sannolikheterna för att i urvalet erhålla respektive stickprov där element k ingår.

Ett sätt att skatta totalsumman i populationen är

$$\hat{t}_\pi = \sum\nolimits_s \frac{y_k}{\pi_k} = \sum\nolimits_s \check{y}_k$$

(Särndal et al., 1992, kapitel 2.8) där $\check{y}_k$ en vanligt förekommande notation för $\frac{y_k}{\pi_k}$. Detta sätt att skatta summan av undersökningsvariabeln i hela populationen kallas för Horvitz-Thompson-skattning efter sina upphovsmän, men är även känt under namnet π-skattning.

3.4 Regressionsskattning

Vid regressionsskattning används bakgrundsvariablerna, som redan innan undersökningens början är kända för hela populationen, för att göra skattningar av y_k. En förutsättning för att kunna använda sig av regressionsskattningen är att bakgrundsvariablerna kovarierar med responsvariabeln. Till att börja med beskrivs den så kallade differensskattningen för att med utgångspunkt från denna övergå till regressionsskattningen, som är det slutliga målet. Det teoretiska resonemanget i kapitlet kommer från Särndal et al. (1992, kapitel 6.1-6.4).

<u>Differensskattning</u>

Antag att vi i en undersökning har tillgång till J stycken bakgrundsvariabler, $x_1, \dots, x_j, \dots, x_J$, som för varje person i populationen antar några givna värden. Med andra ord har vi för person k den redan kända vektorn $\boldsymbol{x_k} = (x_{1k}, \dots x_{jk}, \dots x_{Jk})$ och variabeln y_k, som är okänd innan undersökningens början. Även i detta fall är det populationstotalen,

$$t_y = \sum\nolimits_U y_k$$

, som ska skattas. Ett stickprov av storlek s dras från hela populationen enligt någon given urvalsmetod, där samtliga element har en positiv inklusionssannolikhet. För individerna i urvalet observerar vi värden både på variabeln y_k och på bakgrundsvariablerna. Med hjälp av denna information och bakgrundsvariablerna för övriga individer i populationen, som inte

ingått i urvalet, skattas variabeln y_k i hela populationen. Detta görs genom att approximera varsitt värde på y_k för samtliga individer i populationen, $y_1^o,\ y_2^o, \ldots y_N^o$.

Skattningarna, y_k^o, erhålls genom linjärkombinationer av bakgrundsvariablerna,

$$y_k^o = \sum_{j=1}^{J} A_j x_{jk} = \boldsymbol{A}'\boldsymbol{x_k}$$

, där vi vid differensskattning gör antagandet att $\boldsymbol{A}$ är en vektor med kända koefficienter. I och med att $\boldsymbol{x_k}$ är en känd vektor för $1 \leq k \leq N$ kan vi skatta y_k^o för samtliga individer i populationen.

Populationstotalen skrivs som

$$t_y = \sum\nolimits_U y_k = \sum\nolimits_U y_k^o + \sum\nolimits_U (y_k - y_k^o) = \sum\nolimits_U y_k^o + \sum\nolimits_U D_k$$

där

$$D_k = y_k - y_k^o$$

Eftersom y_k^o har kända värden i hela populationen kan summan av dessa beräknas medan summan av differenserna, D_k, måste skattas i och med att vi inte känner till värdena på de y_k som inte ingått i urvalet. För detta ändamål används Horvitz-Thompson-skattningen, som enligt kapitel 3.3 blir

$$\sum\nolimits_s \breve{D}_k = \sum\nolimits_s \frac{D_k}{\pi_k} = \sum\nolimits_s \frac{(y_k - y_k^o)}{\pi_k}$$

En (differens-)skattning av totalsumman blir därmed

$$\hat{t}_{y,dif} = \sum\nolimits_U y_k^o + \sum\nolimits_s \breve{D}_k$$

, där den senare summan kan ses som en korrigering för det systematiska fel som uppkommit vid skattningarna av y_k^o. Differensskattningen kan även användas som ett sätt att förbättra den vanliga Horvitz-Thompson-skattningen genom att utnyttja att bakgrundsvariablerna är kända i hela populationen. Om de skattade värdena y_k^o också i detta fall antas vara linjärkombinationer av bakgrundsvariablerna, det vill säga $y_k^o = \sum_{j=1}^{J} A_j x_{jk}$, och detta uttryck insätts istället för y_k^o på de båda platserna i formeln för totalsumman $\hat{t}_{y,dif}$ erhålls efter omskrivning

$$\hat{t}_{y,dif} = \hat{t}_{\pi y} + \sum_{j=1}^{J} A_j (t_{x_j} - \hat{t}_{x_j \pi})$$

där

$$\hat{t}_{x_j \pi} = \sum\nolimits_s x_{jk}/\pi_k$$

och

$$t_{x_j} = \sum_U x_{jk}$$

och

$$\hat{t}_{\pi y} = \sum_s y_k/\pi_k$$

Regressionsskattning

Med utgångspunkt från ovanstående teori kan vi övergå till att behandla regressionsskattningen. Vi har samma förutsättningar och mål som vid differensskattningen, som kortfattat var att skatta populationstotalen under antagandet att en skattning av varje värde på y_k kan göras genom en linjärkombination av bakgrundsvariablerna. Det som skiljer regressionsskattningen från differensskattningen är att vi inte längre förutsätter att vektorn $\boldsymbol{A}$ är känd. För att kunna beräkna y_k^o måste vi därmed skatta $\boldsymbol{A}$. Detta görs, som metodens namn antyder, genom att skatta koefficienterna i $\boldsymbol{A}$ på samma sätt som koefficientskattningar görs i regressionsanalysen. För att skilja differensskattningens koefficienter från regressionsskattningens skrivs de senare för tydlighets skull från och med nu som $\widehat{\boldsymbol{B}} = (\hat{B}_1, \dots, \hat{B}_J)$, och skattas genom

$$\widehat{\boldsymbol{B}} = \left(\hat{B}_1, \dots, \hat{B}_J\right)' = \left(\sum_{\boldsymbol{s}} \boldsymbol{x_k x_k'}/\sigma_k^2\, \pi_k\right)^{-1} \left(\sum_{\boldsymbol{s}} \boldsymbol{x_k} y_k/\sigma_k^2\, \pi_k\right)$$

där

$$\sigma_k^2 = Var(Y_k)$$

och antaganden om variansen görs utifrån datamaterialets struktur (Särndal et al, 1992, kapitel 6.4).

Genom att utgå från den förbättrade π-skattningen som härleddes i avsnittet om differensskattning erhåller vi regressionsskattningen

$$\hat{t}_{yr} = \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j (t_{x_j} - \hat{t}_{x_j\pi})$$

, där komponenterna definieras på samma sätt som vid differensskattning.

3.5 Förslag på regressionsskattning som tar hänsyn till metodeffekten i undersökningar utan bortfall vid OSU

Antag att vi har gjort en undersökning där individerna valts ut genom obundet slumpmässigt urval, OSU, vilket innebär att alla urval av fix storlek är lika sannolika. Om vi upptäcker en signifikant skillnad mellan svaren som lämnats på pappersenkät och svaren som lämnats på webbenkät (exempelvis genom att använda oss av Pearsons χ^2-test), det vill säga att insamlingsmetoden kovarierar med responsvariabeln, är en idé att behandla insamlingsmetoden som en bakgrundsvariabel som kan anta värdena noll och ett, där noll innebär att den ena metoden valts av den aktuella individen, medan ett betecknar motsatsen. Fortsättningsvis låter vi i denna uppsats webbenkät kodas med värdet ett, medan pappersenkät kodas med värdet noll, det vill säga

$$x_{webb\,k} = \begin{cases} 1\ om\ webbenkät\ valts\ av\ individ\ k \\ 0\ om\ pappersenkät\ valts\ av\ individ\ k \end{cases}$$

Om vi väljer att betrakta metodvalet som en bakgrundsvariabel får vi enligt kapitel 3.4 regressionsmodellen

$$\hat{t}_{yr} = \hat{t}_{y\pi} + \hat{B}_1\left(t_{x_1} - \hat{t}_{x_1\pi}\right) + \ \ldots + \hat{B}_J\left(t_{x_J} - \hat{t}_{x_J\pi}\right) + \hat{B}_{webb}\left(t_{x_{webb}} - \hat{t}_{x_{webb}\pi}\right)$$

Koefficienterna skattas på vanligt vis med hjälp av formeln för $\widehat{\boldsymbol{B}}$, där ytterligare en kolumn som betecknar metodvalet har lagts till i $\boldsymbol{x_k}$. Samtliga variabler i modellen är kända utom $t_{x_{webb}}$ som betecknar det totala antalet personer i populationen som skulle ha valt webbundersökning om de hade ingått i urvalsgruppen.

En skattning får ersätta $t_{x_{webb}}$ i modellen förutsatt att vi är medvetna om risken för att ett systematiskt fel uppstår när vi tar med en skattning i en modell som egentligen är avsedd för verkliga observerade värden, se Särndal et al. (1992, kapitel 6.4, remark 6.4.3). Ett förslag på ett möjligt sätt att skatta antalet webbundersökningar i hela populationen är att gå igenom ett stort antal tidigare genomförda undersökningar där obundet slumpmässigt urval (OSU) gjorts. I första hand bör undersökningar som genomförts i samma population användas och i andra hand undersökningar i populationer som är så lika den aktuella populationen som möjligt med avseende på bakgrundsvariablerna. Genom att mäta andelen personer i respektive undersökningsmetod i de tidigare genomförda undersökningarna kan medelvärdet av dessa andelar beräknas och multipliceras med antalet personer i vår undersökning för att på så sätt få en skattning, $\hat{t}_{x_{webb}}$, som kan användas i formeln för regressionsskattningen.

Vi kan även välja att ställa upp en regressionsmodell, där vi betraktar metodvalet som undersökningsvariabel, för att göra skattningar av antalet webbsvar i de tidigare genomförda undersökningarna. Med utgångspunkt från dessa regressionsskattningar kan en skattning av andelen webbsvar i populationen göras för respektive (tidigare genomförd) undersökning. Genom att beräkna medelvärdet av de skattade andelarna och multiplicera med populationsstorleken i vår undersökning erhålls även i detta fall en skattning, $\hat{t}_{x_{webb}}$, som kan ersätta $t_{x_{webb}}$ i modellen. De båda ovan beskrivna metoderna förutsätter dock att vi antar att det ämne som frågorna i undersökningen behandlar inte har någon inverkan på andelen personer som väljer webb- respektive pappersenkät. Ytterligare en förutsättning, i fall där vi inte har tillgång till tidigare undersökningar i samma population, är som tidigare nämnts att de populationer som vi gör våra skattningar med utgångspunkt från är någorlunda lika populationen i vår undersökning med avseende på bakgrundsvariablerna. Exempelvis bör vi inte välja att använda oss av den genomsnittliga andelen webbundersökningsdeltagare i undersökningar med enbart manliga deltagare för att göra skattningar av fördelningen mellan webb- och papperssvar i en undersökning som enbart vänder sig till kvinnor eller vice versa.

När vi har skattat summan av undersökningsvariabeln, $\hat{t}_{yr}$, i hela populationen med hjälp av modellen ovan har vi ett mått på hur varje bakgrundsvariabel har inverkat på skattningen, det vill säga vi får ett siffervärde på bland annat $\hat{B}_{webb}\left(\hat{t}_{x_{webb}} - \hat{t}_{x_{webb}\pi}\right)$. Detta värde kan ses som en indikation på hur det har påverkat populationsskattningen att det antal personer som har valt webbundersökning har gjort det val som de har gjort. Vi måste även ha i åtanke att den inverkan som det har haft att ett visst antal individer har valt pappersenkät ligger inbakat i modellen. Om metoderna inte skiljer sig åt signifikant bör $\hat{B}_{webb}$ ha ett värde som inte är

signifikant skilt från noll. Ett sådant resultat skulle innebära att den metodeffekt som pappers- och webbundersökning har är lika, men inte nödvändigtvis att ingen metodeffekt finns. Detta eftersom det kan vara så att de båda insamlingsmetoderna systematiskt påverkar de svar som lämnas, men att det inte är någon signifikant skillnad mellan *hur* metoderna påverkar svaren.

Man kan ifrågasätta varför vi, om vi skulle komma fram till att en viss typ av undersökning är bäst, inte enbart använder oss av den ”bästa” metoden. Förklaringen till det är att vi måste ta hänsyn till att vi genom att erbjuda flera metoder når en betydligt större del av befolkningen. Antag att vi genom empiriska undersökningar har kommit fram till att webbundersökning ger mest sanningsenliga svar. Att enbart erbjuda webbundersökning är inte ett bra alternativ med tanke på att vi inte kan förutsätta att hela populationen har tillgång till dator eller har tillräckliga datorkunskaper för att kunna genomföra en webbundersökning. Om vi dessutom har i åtanke att det med stor sannolikhet finns ett samband mellan att av någon anledning inte ha möjlighet att genomföra en webbundersökning och exempelvis ålder inser vi att vi genom att enbart erbjuda alternativet webbundersökning tappar en betydande del av populationen, som kan ha svar att lämna som eventuellt skiljer sig avsevärt från svarsgruppens. En sammanställning av internettillgången i Nederländerna visar på tydliga samband mellan minskad tillgång till internet med ökande ålder, medan ökande utbildning respektive ökande inkomst medför ökande tillgång till internet (Cobben, 2009, kapitel 9.2.3). Om liknande samband finns i Sverige skulle det innebära att en genomsnittlig person som svarar på en undersökning där webbenkät är den enda svarsmetoden med stor sannolikhet både är yngre, har högre inkomst och högre utbildning än en genomsnittlig person i populationen (eller urvalsgruppen).

3.6 Stratifierat urval

Stratifierat urval innebär att populationen delas in i grupper, strata, efter någon eller några bakgrundsvariabler, exempelvis, kön och åldersgrupp. Ur varje stratum dras ett stickprov enligt någon given urvalsmetod. Vid behov kan man använda sig av olika urvalsmetoder inom de olika stratumen, men i denna uppsats väljer vi att begränsa oss till det mest vanligt förekommande fallet då en och samma metod används i samtliga strata.

Vi delar upp populationen, U, i H stycken strata $U_1, \dots, U_h, \dots, U_H$. Ur varje stratum dras ett stickprov s_h, så att vi ur populationen har dragit stickprovet $s = s_1 \cup s_2 \cup \dots \cup s_H$. Antalet element i de olika stratumen har de kända värdena $N_1, \dots, N_h, \dots, N_H$. Värdet på summan av en viss intressant variabel y_k i hela populationen blir

$$t_y = \sum_U y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \bar{y}_{U_h}$$

, där t_h är summan av undersökningsvariabeln i stratum h. Stratum U_h utgör $W_h = \frac{N_h}{N}$ av populationen och det genomsnittliga värdet i hela populationen blir

$$\bar{y}_U = \frac{t}{N} = \sum_{h=1}^{H} W_h \bar{y}_{U_h}$$

I verkligheten har vi inte tillgång till samtliga värden i varje stratum, vilket gör en skattning av t_y nödvändig. Särndal et al. (1992, kapitel 3.7) väljer att använda sig av Horvitz-

Thompsonskattningen för att göra skattningar av totalsumman i varje stratum, som sedan summeras över samtliga strata för att få en skattning av populationstotalen. Skattningen blir

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi} = \sum_{h=1}^{H} \left(\sum_{s_h} \frac{y_k}{\pi_k} \right)$$

Vid OSU inom varje stratum blir skattningen av totalsumman hos undersökningsvariabeln

$$\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

där $\bar{y}_{s_h}$ är genomsnittsvärdet i stickprov s_h (Särndal et al., 1992, kapitel 3.7).

3.7 Förslag på regressionsskattning som tar hänsyn till metodeffekten i undersökningar med stratifierat urval utan bortfall

Vi bygger upp modellen på motsvarande sätt som vid OSU och får även i detta fall skattningen

$$\hat{t}_{yr} = \hat{t}_{y\pi} + \hat{B}_1\left(t_{x_1} - \hat{t}_{x_1\pi}\right) + \ \ldots + \hat{B}_J\left(t_{x_J} - \hat{t}_{x_J\pi}\right) + \hat{B}_{webb}\left(t_{x_{webb}} - \hat{t}_{x_{webb}\pi}\right)$$

Komponenterna beräknas på samma sätt som tidigare (för definition se kapitel 3.4) men inte heller i det här fallet är $t_{x_{webb}}$ känt utan måste ersättas med en skattning, $\hat{t}_{x_{webb}}$.

En lösning på problemet kan vara att, precis som i kapitel 3.5, titta på ett antal tidigare gjorda undersökningar och göra skattningar av andelen webbsvar i dessa. Eftersom stratifierat urval gjorts väljer vi att göra stratumvisa skattningar av antalet webbsvar som vi sedan väger samman till en skattning av antalet webbsvar i populationen. Exempelvis kan vi tänka oss att om vi har delat in populationen i stratum efter några olika åldersgrupper så börjar vi med att beräkna andelen webbsvar i respektive åldersgrupp i de tidigare genomförda undersökningarna. Dessa skattningar görs på motsvarande sätt som i kapitel 3.6, det vill säga antingen genom att direkt beräkna medelvärdet eller genom regressionsskattning. När skattningarna är gjorda utifrån ett antal tidigare undersökningar övergår vi till att beräkna medelvärdet, $\bar{t}_{x_{webb}h}$, av dessa skattningar i respektive stratum h $(1 \leq h \leq H)$. En skattning av antalet webbsvar i stratum h i vår undersökning erhålls därefter genom att multiplicera andelen webbsvar med det totala antalet individer N_h i statumet, så att vi för stratum h får skattningen $\bar{t}_{x_{webb}h} N_h$. En skattning av antalet webbsvar i hela populationen blir därmed

$$\hat{t}_{x_{webb}} = \sum_{h=1}^{H} \bar{t}_{x_{webb}h} N_h$$

När skattningarna utifrån de tidigare undersökningarna görs är det mest optimala alternativet, precis som tidigare nämnts, att utgå från undersökningar som gjorts i samma population men om inga sådana finns bör vi åtminstone se till att inte välja populationer som skiljer sig avsevärt från vår population med avseende på bakgrundsvariablerna. Alternativt kan vi tänka oss att vi i ett fall där populationen undersöks för första gången kan välja att utgå från olika undersökningar för att göra skattningarna i de olika stratumen. Exempelvis skulle vi i sådant fall vid en stratumindelning efter ålder kunna utgå från undersökningar som riktar sig till ungdomar för att göra en skattning av metodvalet i stratumet med de yngsta individerna och

undersökningar som riktar sig till äldre för att göra skattning av metodvalet i stratumet med de äldsta individerna. Inte heller vid detta tillvägagångssätt får vi frångå kravet om att populationerna som används vid skattningarna och populationen som vi undersöker ska vara lika varandra med avseende på bakgrundsvariablerna. Med andra ord bör exempelvis gruppen med äldre individer ha en sammansättning med avseende på kön, utbildningsnivå och övriga bakgrundsvariabler (som kan tänkas ha inverkan på metodvalet) som är mycket lik sammansättningen i gruppen med äldre individer i populationen som behandlas i vår undersökning. De ovan beskrivna metoderna förutsätter även i detta fall att vi antar att de frågor som ställs i undersökningen inte har någon inverkan på metodvalet.

3.8 Klusterurval och tvåstegsurval

Klusterurval är en användbar metod när det inte finns någon sammanställning av varje individ i den population som man vill undersöka eller när individerna i populationen är spridda över ett geografiskt område som är så stort att kostnaderna för att genomföra personliga intervjuer blir orimligt höga. Det senare alternativet bör dock inte vara aktuellt i vårt fall eftersom datamaterialen i de undersökningar som behandlas i denna uppsats samlas in via pappers- och webbenkät, vilket bör hålla kostnaderna på mer eller mindre samma nivå oavsett geografisk spridning. Som exempel på när klusterurval kan vara användbart kan vi tänka oss att vi vill undersöka åsikterna i någon fråga bland personer som arbetar i sjukvården. Om inget register över samtliga sjukvårdsanställda i Sverige finns tillgängligt är det i praktiken omöjligt att dra ett stickprov genom OSU eller stratifierat urval. Istället kan vi välja ett urval av alla svenska sjukhus och antingen ställa frågorna till samtliga eller några av de anställda på de utvalda sjukhusen. Vi kallar de olika sjukhusen för kluster och om vi väljer att ge enkäten till samtliga anställda på de utvalda sjukhusen säger vi att vi gör ett klusterurval medan vi om vi endast låter vissa av de anställda delta i undersökningen gör ett tvåstegurval. Urval i tre eller fler steg kan också göras, men vi väljer att begränsa oss till tvåstegsurval i denna uppsats.

Särndal et al. (1992, kapitel 4.1-4.3) delar upp populationen i N_I stycken kluster, $U_1, \dots, U_i, \dots, U_{N_I}$, av storlekarna $N_1, \dots, N_i, \dots, N_{N_I}$. De definierar $U_I = \{1, \dots, i, \dots, N_I\}$ där varje element representerar ett kluster. Ett stickprov, s_I, bestående av n_I kluster dras enligt någon given urvalsmetod. I urvalet ingår $s = \bigcup_{i \epsilon s_I} U_i$ och det totala antalet individer i stickprovet är $n_s = \sum_{s_I} N_i$. Inklusionssannolikheten för ett visst kluster är

$$\pi_{Ii} = \sum_{i \epsilon s_I} p_I(s_I)$$

där p_I bestäms utifrån vilken urvalsmetod som används.

Sannolikheten för att ett visst element ska ingå i urvalet blir

$$\pi_k = \Pr(k \epsilon s) = \Pr(i \epsilon s_I) = \pi_{Ii}$$

Summan av värdena på undersökningsvariabeln i ett visst kluster blir

$$t_i = \sum_{U_i} y_k$$

och i hela populationen

$$t = \sum_U y_k = \sum_{U_I} t_i$$

.

Särndal et al. (1992, kapitel 4.2.1) skattar populationstotalen genom

$$\hat{t}_\pi = \sum_{s_I} \frac{t_i}{\pi_{Ii}}$$

Om klustren väljs genom OSU, och samtliga element i de aktuella klustren observeras, erhålls populationsskattningen genom att multiplicera antalet kluster i populationen med det genomsnittliga värdet bland elementen i de dragna klustren

$$\hat{t}_\pi = N_I \bar{t}_{s_I} = N_I \frac{\sum_{s_I} t_i}{n_I}$$

En utveckling av det vanliga klusterurvalet är tvåstegsurval, där man efter att ha valt kluster även gör ett urval av vilka individer i de olika klustren som ska delta i undersökningen, istället för att som vid klusterurval välja samtliga individer i stickprovet. En orsak till att välja tvåstegsurval framför klusterurval är att variansen normalt sett blir högre vid klusterurval än vid exempelvis OSU eftersom individerna inom ett givet kluster ofta liknar varandra. För att minska variansen skulle en lösning naturligtvis kunna vara att helt enkelt välja fler kluster och observera samtliga element i dessa, men eftersom det ofta är dyrt är en mer kostnadseffektiv lösning att göra tvåstegsurval. På så sätt får man med fler kluster i undersökningen utan att välja fler individer (Särndal et al., 1992, kapitel 4.3.1).

Särndal et al. (2000, kapitel 4.3.1) delar även i detta fall in populationen i N_I delar som betecknas $U_1, \dots, U_i, \dots, U_{N_I}$. Mängden kluster i populationen skrivs $U_I = \{1, \dots, i, \dots, N_I\}$. Ett stickprov, s_I, dras från U_I och för varje $i \epsilon s_I$ dras ett stickprov s_i med element från U_i. Vid tvåstegsurval blir populationsskattningen enligt Särndal et al. (2000, kapitel 4.3.2)

$$\hat{t}_\pi = \sum_{s_I} \frac{\hat{t}_{i\pi}}{\pi_{Ii}}$$

där

$$\hat{t}_{i\pi} = \sum_{s_i} \frac{y_k}{\pi_{k|i}}$$

och π_{Ii} betecknar inklusionssannolikheten för kluster i och

$$\pi_{k|i} = \frac{\pi_k}{\pi_{Ii}}$$

3.9 Förslag på regressionsskattning som tar hänsyn till metodeffekten vid kluster- och tvåstegsurval i undersökningar utan bortfall

Vi gör antagandet att vi har ett fall där det inte finns någon sammanställning av varje enskild individ som ingår i populationen som vi vill undersöka (det vill säga vi utesluter att klusterurval görs på grund av att populationen är geografiskt utspridd, eftersom klusterurval i sådant fall inte bör vara nödvändigt i och med att svaren samlas in via post och internet). När man vill använda sig av regressionsskattningen efter att ha gjort klusterurval skiljer man på två olika fall beroende på vilken information om bakgrundsvariablerna som man har tillgång till. Vi kan antingen ha ett fall där vi har tillgång till summan av varje bakgrundsvariabel i respektive kluster i populationen, eller ett fall där vi känner till bakgrundsvektorn för varje enskild individ i de utvalda klustren men saknar information om bakgrundsvariablerna för den övriga delen av populationen.

Särndal et al. (1992, kapitel 8.2) låter $\boldsymbol{u_i}$ beteckna en vektor med summan av bakgrundsvektorerna i kluster *i*, det vill säga $\boldsymbol{u_i} = (u_{1i}, \dots, u_{\boldsymbol{vi}}, \dots, u_{Ji})'$, där $u_{\boldsymbol{vi}}$ är summan av bakgrundsvariabel *v* i kluster *i*, vilket exempelvis skulle kunna representera antalet män i det aktuella klustret. I det första fallet blir regressionsskattningen vid klusterurval enligt Särndal et al. (1992, kapitel 8.4),

$$\hat{t}_{yAr} = \sum\nolimits_{s_I} \left(\frac{\sum_{U_i} y_k}{\pi_{Ii}} \right) + \left(\sum\nolimits_{U_I} \boldsymbol{u_i} - \sum\nolimits_{\boldsymbol{s_I}} \boldsymbol{u_i}/\pi_{\boldsymbol{Ii}} \right)' \widehat{\boldsymbol{B}}_{\boldsymbol{I}}$$

och regressionsskattningen vid tvåstegsurval

$$\hat{t}_{yAr} = \sum\nolimits_{s} \frac{y_k}{\pi_k} + \left(\sum\nolimits_{U_I} \boldsymbol{u_i} - \sum\nolimits_{\boldsymbol{s_I}} \boldsymbol{u_i}/\pi_{\boldsymbol{Ii}} \right)' \widehat{\boldsymbol{B}}_{\boldsymbol{I}}$$

där $\boldsymbol{u_i}$ betecknar en vektor med summan av bakgrundsvektorerna i kluster *i* och

$$\widehat{\boldsymbol{B}}_{\boldsymbol{I}} = \left(\sum\nolimits_{s_I} \boldsymbol{u_i} \boldsymbol{u_i'}/\sigma_{Ii}^2 \pi_{\boldsymbol{Ii}} \right)^{-1} \sum\nolimits_{s_I} \boldsymbol{u_i} t_{yi}^*/\sigma_{Ii}^2 \, \pi_{\boldsymbol{Ii}}$$

där

$$t_{yi}^* = \sum\nolimits_{U_i} y_k$$

vid klusterurval och

$$t_{yi}^* = \sum\nolimits_{s_i} \frac{y_k}{\pi_{k|i}}$$

vid tvåstegsurval.

Om vi även i detta fall väljer att betrakta metodvalet som en bakgrundsvariabel som påverkar populationsskattningen är problemet återigen att göra en rimlig skattning av det totala antalet personer i populationen, $\hat{t}_{x_{webb}}$, som skulle ha valt webbundersökning om de hade deltagit i undersökningen. Vi väljer att även i detta fall förlita oss på att rimliga skattningar kan göras med utgångspunkt från tidigare undersökningar i liknande populationer. Om klusterurval har gjorts har vi tillgång till det sanna antalet webbundersökningar i de utvalda klustren. Vi behöver därför enbart en skattning av det totala antalet webbundersökningar i populationen. Om klustren verkar någorlunda lika med avseende på de ursprungliga bakgrundsvariablerna

(som vi ju har tillgång till summan av i samtliga kluster i populationen) kan vi välja att göra en skattning av det totala antalet webbundersökningar i hela populationen direkt utifrån tidigare undersökningar i liknande eller samma population på motsvarande sätt som nämnts tidigare i uppsatsen, det vill säga med hjälp av medelvärden eller regressionsskattningar där antalet webbundersökningar betraktas som en undersökningsvariabel. Om stora skillnader i bakgrundsvariablerna finns kan vi istället välja att göra en skattning per kluster (som görs med hjälp av någon av de beskrivna metoderna) utifrån undersökningar gjorda i populationer med liknande sammansättning med avseende på bakgrundsvariablerna i det aktuella klustret, och beräkna summan av dessa skattningar. I undersökningar där tvåstegsurval gjorts måste det senast nämnda tillvägagångssättet användas för att skatta antalet webbsvar i respektive kluster. Skattningen av det totala antalet webbsvar kan däremot göras enligt vilken som helst av de ovan nämnda metoderna även vid tvåstegsurval.

Vi övergår till en situation där vi har tillgång till bakgrundsvektorerna för de enskilda individerna i de utvalda klustren, men inte i hela populationen. Regressionsskattningen blir i detta fall enligt Särndal et al. (1992, kapitel 8.9)

$$\hat{t}_{yCr} = \sum_{\boldsymbol{s_I}} \left(\frac{\sum_{\boldsymbol{U_i}} \boldsymbol{x_k}' \widehat{\boldsymbol{B}}}{\pi_{Ii}}\right) + \sum_{\boldsymbol{s}} \frac{y_k - \boldsymbol{x_k'}\widehat{\boldsymbol{B}}}{\pi_k}$$

där

$$\widehat{\boldsymbol{B}} = \left(\sum_{\boldsymbol{s}} \boldsymbol{x_k x_k'}/\sigma_k^2\,\pi_k\right)^{-1}\left(\sum_{\boldsymbol{s}} \boldsymbol{x_k} y_k/\sigma_k^2\,\pi_k\right)$$

Vi ser att vi i den första summan behöver en skattning av det totala antalet webbundersökningar i de utvalda klustren. För detta ändamål används den ovan beskrivna metoden för att göra skattningar i respektive kluster.

3.10 Förslag för hantering av metodeffekter vid prediktionsskattning i undersökningar utan bortfall

Denna uppsats lägger fokus på regressionsskattningen. I detta kapitel ger vi dock ett kortfattat förslag på hur prediktionsskattningar där hänsyn till insamlingsmetoden tas kan göras, vid samtliga beskrivna urvalsmetoder utom tvåstegsurval. Vid urval i ett steg (det vill säga vid samtliga urvalsmetoder som beskrivs i denna uppsats utom tvåstegsurval) är det rimligt att göra antagandet att $\boldsymbol{V_{as}} = \boldsymbol{V_{sa}} = \boldsymbol{0}$ (Valliant, Dorfman och Royall, 2000, kapitel 2.2) vilket ger

$$\hat{\theta} = (1, \dots, 1)\boldsymbol{Y_s} + (1, \dots, 1)\boldsymbol{X_a}\widehat{\boldsymbol{\beta}}$$

där $\widehat{\boldsymbol{\beta}}$ beräknas enligt kapitel 3.2.

Omskrivning ger

$$\hat{\theta} = \sum_{s} y_k + \sum_{a} \hat{\beta}_1 x_{1i} + \ \dots \ + \sum_{a} \hat{\beta}_p x_{pi}$$

Om vi även i detta fall väljer att betrakta insamlingsmetoden som en bakgrundsvariabel återstår enbart att skatta det totala antalet webbundersökningar i den del av populationen som inte ingått i urvalet eftersom $\sum_a \hat{\beta}_{webb} x_{webb\,i} = \hat{\beta}_{webb} \sum_a x_{webb\,i}$ och $\hat{\beta}_{webb}$ kan beräknas med hjälp av stickprovet. Genom att betrakta metodvalet som svaret på frågan ”Vilken av följande insamlingsmetoder föredrar du?” (vilket vi indirekt frågat eftersom vi gett urvalspersonerna möjlighet att själva välja metod) kan vi använda metoderna i kapitel 3.6 och 3.8 för att göra en skattning av det totala antalet webbundersökningar vid stratifierat urval och klusterurval. Vid OSU kan vi istället anta att urvalet är representativt för hela populationen och helt enkelt multiplicera andelen webbundersökningar med antalet personer i populationen. Skattningen av antalet webbundersökningar i gruppen som inte ingått i urvalet erhålls slutligen genom att subtrahera det skattade antalet webbundersökningar från det observerade antalet i stickprovet.

4 Undersökningar med bortfall

4.1 Bortfall

I teorin som hittills tagits upp har vi utgått från den orealistiska förutsättningen att vi vid undersökningens slut har erhållit de efterfrågade svaren från hela urvalet. I verkligheten existerar i princip aldrig sådana undersökningar. På Statistiska centralbyrån hade man år 2000 bortfall på mellan 20% och drygt 30% i många undersökningar (Särndal och Lundström, 2005, kapitel 2.2), vilket är en såpass hög andel att det är absolut nödvändigt att ta hänsyn till bortfallet för att de skattningar som görs ska kunna sägas representera hela den aktuella populationen. Bortfall definieras som alla de personer som slumpats fram till att delta i undersökningen, men som man av någon anledning inte fått in ett användbart svar från. Det kan med andra ord vara allt från personer som är omöjliga att kontakta, personer som man får kontakt med men som inte kan språket, personer med fysiska eller psykiska hinder för att svara, personer som lämnat svar som man direkt inser är orimliga, till personer som helt enkelt vägrar att delta i undersökningen.

Det finns framförallt två vanliga sätt att hantera bortfall – viktning och imputering. Viktning innebär att man ger olika vikter till de svar som inkommit från personerna som svarat på undersökningen, för att korrigera för bortfallet. Vid imputering fyller man istället i värden på de frågor där svar saknas. Man skiljer mellan två olika sorters bortfall - objektsbortfall och partiellt bortfall. Den första typen av bortfall innebär att urvalspersonen överhuvudtaget inte har deltagit i undersökningen, medan partiellt bortfall betyder att personen har svarat på enkäten, men utelämnat vissa frågor. Vid hantering av bortfall är ett vanligt tillvägagångssätt att kombinera de två metoderna så att imputering används för att korrigera för objektsbortfallet medan viktning används för att korrigera för det partiella bortfallet. Det är dock även möjligt att välja att enbart använda sig av en av de två metoderna.

4.2 Viktning

För att beräkna vikterna använder vi oss av en metod som kallas för kalibrering. För att vi ska kunna använda oss av metoden krävs det att vi har tillgång till bakgrundsvektorn **x** för varje individ i svarsgruppen, men att även de övriga kriterierna för antingen InfoU, InfoS eller InfoUS nedan är uppfyllda. Särndal och Lundström (2005, kapitel 6.2) inför följande uppdelning för att skilja på de tre fallen:

InfoU: Vi känner till totalsumman av alla bakgrundsvektorer i populationen, $\sum_U \boldsymbol{x}_k^*$ (vi har information om antalet personer i de olika ålderskategorierna, antal män, antal kvinnor och så vidare i populationen), och värdena i varje enskild vektor $\boldsymbol{x}_k^*$ för de personer som svarat på undersökningen.

InfoS: Vi känner till bakgrundsvektorerna $\boldsymbol{x}_k^o$ för varje individ i stickprovet (och därmed i svarsgruppen), men inte för hela populationen.

InfoUS: Både InfoU och InfoS är kända och används. Vi känner med andra ord till totalsumman av alla vektorer i populationen, $\sum_U \boldsymbol{x}_k^*$, men vet inte nödvändigtvis hur varje enskild vektor ser ut, utom för personerna som svarat på undersökningen och övriga personer i stickprovet. Bakgrundsvektorn blir i detta fall $\boldsymbol{x}_k = \begin{pmatrix} x_k^* \\ x_k^o \end{pmatrix}$

Vid kalibrering (och viktning i allmänhet) är syftet återigen att skatta populationstotalen. Med kända vikter w_k blir skattningen

$$\hat{Y}_W = \sum_r w_k y_k$$

, där r är den grupp individer i stickprovet som svarat på undersökningen.

Om vi vill kunna använda oss av metoden som går ut på att vi lägger in metodval i bakgrundsvektorn är InfoU den information som vi är närmast att ha tillgång till. Endast information om det totala antalet personer som skulle ha valt respektive insamlingsmetod i hela populationen saknas. InfoS och InfoUS utesluts direkt (utom vid OSU, se förklaring nedan). Visserligen känner vi till samtliga ursprungliga bakgrundsvariabler för individerna i både stickprovet och populationen, men metodvalet är endast känt för de svarande personerna. Skattningar av metodvalet för varje enskild individ i bortfallet bör knappast bli särskilt tillförlitliga. Däremot bör vi kunna göra en rimlig skattning av det totala antalet personer i populationen som skulle ha valt respektive metod. Därför väljer vi att enbart visa metoden för att skatta populationstotalen i fallet då vi har tillgång till InfoU. Vikterna ska i detta fall uppfylla nedanstående ekvation

$$\sum_r w_k \boldsymbol{x}_k^* = \sum_{\boldsymbol{U}} \boldsymbol{x}_k^*$$

, som kallas kalibreringsekvationen.

Vikterna skattas genom

$$w_k = \frac{1}{\pi_k} v_k$$

där

$$v_k = 1 + \boldsymbol{\lambda}' \boldsymbol{x}_{\boldsymbol{k}}^*$$

och

$$\boldsymbol{\lambda}' = \left(\sum_{\boldsymbol{U}} \boldsymbol{x}_{\boldsymbol{k}}^* - \sum_{\boldsymbol{r}} \frac{1}{\pi_k} \boldsymbol{x}_{\boldsymbol{k}}^* \right)' \left(\sum_r \frac{1}{\pi_k} \boldsymbol{x}_{\boldsymbol{k}}^* \boldsymbol{x}_{\boldsymbol{k}}^{*\prime} \right)^{-1}$$

Kalibreringsskattningen i hela populationen blir alltså

$$\hat{Y}_W = \sum_r w_k y_k =$$
$$= \sum_r \frac{1}{\pi_k}\left(1 + \left(\sum_{\boldsymbol{U}} \boldsymbol{x_k^*} - \sum_{\boldsymbol{r}} \frac{1}{\pi_k}\boldsymbol{x_k^*}\right)' \left(\sum_r \frac{1}{\pi_k}\boldsymbol{x_k^* x_k^{*\prime}}\right)\boldsymbol{x_k^*}\right) y_k$$

En exakt härledning av ovanstående uttryck återfinns i Särndal och Lundström (2005, kapitel 6.4) men sammanfattningsvis kan vi säga att resonemanget bygger på att i och med att det finns bortfall i undersökningen kommer den vanliga Horvitz-Thompsonskattningen att underskatta undersökningsvariabeln vid skattning av hela populationen. Därför krävs det att man ger varje lämnat svar, y_k, en större vikt, v_k, än den som erhålls genom att dividera med inklusionssannolikheten. Insättning av v_k i w_k och sedan insättning av w_k i kalibreringsekvationen ger efter omskrivningar uttrycket för $\boldsymbol{\lambda'}$.

Ett specialfall är dock då alla individer i populationen har samma sannolikhet att ingå i urvalsgruppen (exempelvis vid OSU). I detta fall är vi lika nära att ha tillgång till InfoS som InfoU eftersom vi endast behöver skatta det totala antalet webbundersökningar i stickprovet för att kunna ställa upp kalibreringsekvationen som har utseendet

$$\sum_r w_k \boldsymbol{x_k^o} = \sum_{\boldsymbol{s}} \frac{1}{\pi_k}\boldsymbol{x_k^o}$$

och som på grund av att π_k är konstant blir

$$\sum_r w_k \boldsymbol{x_k^o} = \frac{1}{\pi_k}\sum_{\boldsymbol{s}} \boldsymbol{x_k^o}$$

För en beskrivning av hur vikterna beräknas i vid InfoS hänvisas till Särndal och Lundström (2005, kapitel 6.4).

4.3 Imputering

Imputering innebär att man skattar värden på de svar som saknas i undersökningen, som sedan används i beräkningarna. Det finns flera metoder för att komma fram till vilka värden som ska användas. Särndal och Lundström (2005, kapitel 12.1) nämner tre stycken generella sätt att välja imputeringsvärden:

1. Att använda statistiska metoder för att prediktera, till exempel med hjälp av regression.
2. Att titta på individer med liknande egenskaper som bortfallsindividen, och använda samma svar.
3. Att låta en expert göra en bedömning.

I en och samma undersökning kan olika sätt att imputera värden användas eftersom de olika metoderna är lämpliga i olika situationer. Särndal och Lundström (2005, kapitel 12.1)

framhåller som exempel att det vid företagsundersökningar kan vara lämpligare att använda sig av expertutlåtanden än statistiska metoder om de företag som inte svarat skiljer sig avsevärt från de övriga företagen, exempelvis med avseende på storlek eller omsättning, eftersom statistiska metoder bygger på att det finns vissa likheter mellan elementen i undersökningen. Imputering är en metod som ibland kritiseras på grund av att den bygger på att lägga in värden i modellen som man på förhand med stor säkerhet vet är fel, men Särndal och Lundström (2005, kapitel 12.1) påpekar att det inte finns några bevis för att imputering skulle göra skattningarna mindre tillförlitliga än andra statistiska metoder. Nedan följer en översiktlig beskrivning av ett fåtal av de möjliga metoder som finns att tillgå när man vill använda sig av imputering i praktiken. Fler och mer ingående beskrivningar av imputeringsmetoder presenteras exempelvis i Särndal och Lundström (2005, kapitel 12) och Little och Rubin (2002, kapitel 4-5 samt kapitel 10.2).

Regressionsimputering

$$y_k = \boldsymbol{x}'_k \widehat{\boldsymbol{\beta}}_i$$

där

$$\widehat{\boldsymbol{\beta}}_i = \left(\sum\nolimits_{r_i} a_k \boldsymbol{x}_k \boldsymbol{x}'_k\right)^{-1} \sum\nolimits_{r_i} a_k \boldsymbol{x}_k y_k$$

och där r_i betecknar det antal personer som svarat på fråga nummer i och a_k är vikter som valts på lämpligt sätt. Det finns flera kända metoder för att hitta vikter som lämpar sig för olika typer av undersökningar. För en närmare presentation av möjliga tillvägagångssätt hänvisas till Särndal och Lundström (2005, kapitel 12.7).

Medelvärdesimputering

Medelvärdesimputering innebär att allt partiellt bortfall i en viss fråga ersätts med medelvärdet bland de svar som inkommit. Metoden ger vanligtvis en bra punktskattning medan variansen underskattas. En utveckling av den vanliga medelvärdesimputeringen är att först dela in populationen i mindre grupper där urvalsobjekten liknar varandra med avseende på bakgrundsvariablerna och svar på andra frågor undersökningen, för att sedan beräkna medelvärden gruppvis och använda dessa värden för att korrigera för det partiella bortfallet i respektive grupp (Särndal et al., 1992, kapitel 15.7).

”Hot deck”-imputering

Metoden innebär att man för varje bortfallsindivid slumpar fram ett av de svar som faktiskt inkommit på frågan, och ger det även till individen som låtit bli att svara. Eftersom man vid ”hot deck”-imputering inte använder sig av någon bakgrundsinformation alls finns det en överhängande risk att värden som ligger mycket långt ifrån det sanna värdet för individen väljs. Särndal och Lundström (2005, kapitel 12.7) framhåller därför att metoden inte är att rekommendera om andra bättre imputeringsalternativ är möjliga, utan att den snarare ska ses som ”en sista utväg”.

4.4 Den kombinerade metoden

Som tidigare nämnts är det mest vanligt förekommande tillvägagångssättet för korrigering av bortfall att kombinera imputering och viktning. Man börjar med att imputera värden för det partiella bortfallet, det vill säga värden läggs till i de enkäter där svarspersonen har hoppat över enstaka frågor. När man på detta sätt har erhållit fullständiga enkäter för samtliga personer som deltagit i undersökningen övergår man till att använda sig av viktning för att göra skattningar av undersökningsvariablerna i hela populationen.

4.5 Orsaker till bortfall

För att kunna dra korrekta slutsatser utifrån ett datamaterial är det av stor vikt att skilja på nedanstående typer av bortfall.

4.5.1 MCAR - Missing completely at random

”Missing completely at random” (MCAR) innebär att huruvida en person väljer att svara eller att inte svara på en viss fråga är helt oberoende av vad svaret på frågan är. Inte heller andra kända bakgrundsvariabler har någon inverkan. I en undersökning där man vill undersöka om individerna röker eller inte har det med andra ord ingen betydelse för chansen att svara om en viss person är rökare eller icke-rökare, man eller kvinna, tillhör en viss åldersgrupp och så vidare.

Little och Rubin (2002, kapitel 1.3) definierar det kompletta datamaterialet som matrisen $Y = (y_{ij})$. Att vi har bortfall för en viss person och fråga indikeras i $I \times J$- matrisen $M = (m_{ij})$ där $m_{ij} = 0$ om observationen saknas och $m_{ij} = 1$ om svar på frågan har erhållits. Okända parametrar betecknas med φ. I fall där vilka observationer som saknas är helt oberoende av observationernas värden säger vi att datamaterialet är MCAR. Vi kan skriva

$$f(M|Y, \varphi) = \mathrm{f}(\mathrm{M}|\varphi) \text{ för alla } Y, \varphi$$

4.5.2 MAR - Missing at random

”Missing at random” innebär, tvärtemot vad namnet antyder, att bortfallet inte är helt slumpmässigt. Little och Rubin (2002, kapitel 1.3) nämner som exempel en undersökning där man har med de två variablerna ålder och inkomst. Vi kan säga att variabeln inkomst är MAR i undersökningar där det är olika stor sannolikhet för att svara på frågor om inkomsten för olika åldersgrupper om det samtidigt är helt slumpmässigt om personer inom en och samma ålderskategori svarar eller inte. Huruvida en individ ger information om sin inkomst är med andra ord slumpmässigt betingat på ålder.

4.5.3 NMAR - Not missing at random

”Not missing at random”, NMAR, innebär att om individerna svarar eller inte svarar beror på det sanna svaret på frågan, alltså värdena i det kompletta datamaterialet Y (Little och Rubin, 2002, kapitel 1.3). Om anhängare till ett visst politiskt parti deltar i en partisympatiundersökning i lägre utsträckning än övriga partiers anhängare är det ett exempel på NMAR. En annan vanlig beteckning för NMAR är NN – ”Nonignorable nonresponse”. I denna uppsats förutsätter vi att bortfallet inte är NMAR.

4.6 Hantering av bortfall vid MCAR

Om bortfallet är helt slumpmässigt (MCAR) kan modellerna där hänsyn till metodeffekten tas i kapitel 3.5, 3.7, 3.9 och 3.10 användas direkt på det datamaterial som erhållits i undersökningen. Bortfallet behandlas i dessa fall på samma sätt som om vi aldrig hade bett personerna i bortfallet att delta i undersökningen. Vi ser det med andra ord som om vi helt enkelt hade valt ett mindre stickprov redan vid undersökningens början i vilket samtliga individer svarat.

4.7 En översikt över Fannie Cobbens sätt att hantera metodeffekter i undersökningar med bortfall där individen blir tilldelad en insamlingsmetod

I avhandlingen ”Nonresponse in Sample Surveys – Methods for Analysis and Adjustment” (2009) presenterar Cobben några sätt att hantera bortfall i undersökningar med flera insamlingsmetoder. Samtliga metoder gäller i undersökningar där individen *inte* själv får välja insamlingsmetod, utan där metodvalet görs av statistikbyrån som genomför undersökningen. Detta skiljer sig således från det mest vanligt förekommande tillvägagångssättet på SCB som går ut på att individen själv får välja metod.

Inledningsvis ställer Cobben upp en modell där hon helt bortser från att olika insamlingsmetoder har använts. Istället analyseras datamaterialet i sin helhet, som om ingen metodeffekt fanns. För att skatta populationstotalen används viktning (kalibrering) enligt metoden som beskrivs i kapitel 4.2. Det är även det vanliga tillvägagångssättet på SCB i dagsläget, men i kombination med att pilotstudier genomförs för att bland annat undersöka eventuella metodeffekter och i största möjliga mån undanröja orsaken till att dessa uppkommer, exempelvis genom att ändra enkäternas utformning innan den slutliga versionen av undersökningen skickas ut.

Nästa metod som beskrivs går även den ut på att använda sig av viktning (där InfoS är den information som är känd eftersom enbart personerna i stickprovet tilldelas en metod), men med skillnaden att hänsyn tas till insamlingsmetoden. Cobben inför variabeln metod, M_i, som kan anta värdena noll och ett, det vill säga

$$M_i = \begin{cases} 1 \; om \; person \; i \; har \; tilldelats \; insamlingsmetod \; 1 \\ 0 \; om \; person \; i \; har \; tilldelats \; insamlingsmetod \; 2 \end{cases}$$

, och lägger in den bland bakgrundsvariablerna. Det nämns att det kan finnas ett samband mellan de ursprungliga bakgrundsvariablerna och svarsbeteende i de olika insamlingsmetoderna och därför föreslås två metoder för att handskas med detta. Den ena är att införa samspelstermer mellan metodvariabeln och de bakgrundsvariabler som man väntar sig ska ha inverkan på benägenheten att svara. Ett exempel skulle kunna vara att unga i lägre utsträckning har tillgång till fast telefon och därför inte tilldelas metoden telefonundersökning. Om det dessutom finns ett samband mellan att vara ung och att inte delta i undersökningar kan det finnas anledning att införa en samspelsterm mellan ålder och metod.

Eftersom modellen ofta blir mycket stor vid införandet av samspelstermer föreslås ytterligare ett tillvägagångssätt som innebär att man delar upp urvalsgruppen i två delar, så att man får en grupp per insamlingsmetod. Detta innebär att olika koefficienter fås framför bakgrundsvariablerna, förutsatt att grupperna inte är lika. Cobben inför sannolikheter för att en viss individ tilldelas metod 1 respektive 2 och datamaterialet kan därmed analyseras på samma sätt som vid stratifierat urval, där de två metoderna ses som varsitt strata.

4.8 Kritik mot Cobbens metoder

De modeller som beskrivs i Cobben (2009, kapitel 9.5.1) bygger på att varje individ i stickprovet blir tilldelad en insamlingsmetod. Genom att exempelvis utgå från listor över personer med telefonabonnemang eller bredband tilldelas individerna telefon- respektive

webbundersökning. Problemet med ett sådant tillvägagångssätt är att man måste vara beredd på att antingen acceptera ett stort bortfall (som med stor sannolikhet hade kunnat minskas om man hade erbjudit fler insamlingsmetoder) eller att man som statistikbyrå trots allt erbjuder de individer som inte svarar ytterligare en metod i syfte att minska bortfallet. Det senare fallet innebär i praktiken att individen delvis själv är med och väljer metod eftersom inga säkra skattningar av det totala antalet webb- och pappersundersökningar i stickprovet kan göras om denna möjlighet finns. Om vi låter några individer i stickprovet byta metod och om de trots det inte svarar – vilken metod ska vi då säga att de har tilldelats? Dessutom blir det omöjligt att beräkna samspelstermerna som Cobben föreslår i en av sina modeller eftersom dessa enligt Särndal och Lundström (2005, kapitel 7.6) bygger på att vi känner till egenskaperna hos de individer som skulle ha svarat via respektive metod (antalet kvinnor i stickprovet som skulle ha valt pappersundersökning et cetera) vilket vi omöjligt kan göra när vi inte känner till metodvalet för individerna som inte deltagit i undersökningen. En genomgång av hur skattningar kan göras med hjälp av multivariata probit-modeller då nya insamlingsmetoder erbjuds till de individer som inte svarar på undersökningen beskrivs i Cobben (2009, kapitel 9.5.2), men eftersom endast fallet då individen själv väljer insamlingsmetod redan vid första utskicket är aktuellt i denna uppsats utelämnar vi dessa beskrivningar.

4.9 Förslag på sätt att hantera datamaterial som erhållits genom OSU i undersökningar med bortfall där individen själv väljer pappers- eller webbenkät

Antag att vi har en undersökning där ett stickprov dragits ur en populationen genom OSU. I stickprovet har ett bortfall uppstått, det vill säga ett visst antal personer har av någon anledning låtit bli att lämna in svar, antingen på hela enkäten eller på enstaka frågor. Vi förutsätter att bortfallet är MAR och att deltagarna i undersökningen själva har valt om de ska svara via webb- eller pappersenkät.

För att analysera datamaterialet väljer vi att använda oss av den kombinerade metoden som introducerades i kapitel 2.6. Vi imputerar alltså värden för de personer som lämnat svar på minst en fråga, för att sedan övergå till viktning för att göra populationsskattningar för respektive fråga. Vilken typ av imputering som är lämpligast att använda måste, som nämndes i kapitel 4.3, bedömas från fall till fall och värdena kan komma från alltifrån expertutlåtanden till regressionsmodeller. Vi övergår därför direkt till steget där imputeringen är gjord och det är dags att göra populationsskattningar genom viktning.

Enligt kapitel 4.2 är vi lika nära att ha tillgång till informationen InfoU som InfoS. För enkelhets skull väljer vi dock att enbart presentera tillvägagångssättet när vi har tillgång till InfoU, som innebär att vi känner till totalsumman av alla bakgrundsvektorer i populationen, $\sum_U \boldsymbol{x}_k^*$, och värdena på varje enskild vektor $\boldsymbol{x}_k^*$ för de personer som svarat på (minst en fråga i) undersökningen. Det som saknas för att vi ska kunna påstå oss känna till InfoU är tillgång till information om det totala antalet personer som skulle ha valt respektive insamlingsmetod i hela populationen.

Enligt SCB:s noteringar är bortfall särskilt vanligt hos befolkningen i storstäderna, bland unga personer (i synnerhet unga män) och bland låginkomsttagare. En av teorierna är att många som hamnar i den senare gruppen är nyanlända invandrare som inte kan svenska, vilket gör det omöjligt att svara på enkäten. Cobben (2009, kapitel 9.2.3) refererar i sin avhandling till en undersökning bland studenter gjord av Kwak och Radler (2002) som visar att män och unga personer svarar via webbenkät i större utsträckning än övriga delar av befolkningen. Om vi dessutom lägger till det faktum att det kan finnas ekonomiska orsaker till att inte ha tillgång

till dator och internet, som medför att pappersenkät blir det naturliga alternativet, inser vi att det lätt kan uppstå en under- eller överrepresentation av någon av metoderna i svarsgruppen jämfört med hur resultatet skulle ha sett ut vid en totalundersökning utan bortfall. Detta faktum medför att vi vid skattning av antalet individer i populationen som skulle ha valt webbenkät behöver en metod som tar hänsyn till bakgrundsvariablerna.

Eftersom vi har tillgång till samtliga ursprungliga bakgrundsvariabler för individerna i populationen kan vi använda oss av viktning för att skatta antalet webbsvar i populationen. I detta fall har vi med andra ord tillgång till InfoUS i och med att vi känner till totalsumman av samtliga vektorer i populationen, $\sum_U \boldsymbol{x}_k^*$, och utseendet på varje enskild vektor för individerna i stickprovet (och därmed naturligtvis även i svarsgruppen). Vi återgår till resonemanget i kapitel 4.2 och genomför motsvarande härledning av vikterna som där genomfördes för InfoU. Målet är att även i detta fall göra en skattning av det totala antalet personer i populationen som skulle ha valt webbenkät genom

$$\hat{t}_{x_{webb}} = \sum_r w_k x_{webb_k}$$

Särndal och Lundström (2005, kapitel 6.4) visar att vid tillgång till informationen InfoUS får kalibreringsekvationen utseendet

$$\sum_r w_k \begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix} = \begin{pmatrix} \sum_U \boldsymbol{x}_k^* \\ \sum_s \frac{1}{\pi_k} \boldsymbol{x}_k^o \end{pmatrix}$$

där $\begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix}$ och $\begin{pmatrix} \sum_U \boldsymbol{x}_k^* \\ \sum_s \frac{1}{\pi_k} \boldsymbol{x}_k^o \end{pmatrix}$ tolkas som matriser med två rader.

Vi erhåller återigen

$$w_k = \frac{1}{\pi_k} v_k$$

men där

$$v_k = 1 + \boldsymbol{\lambda}' \begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix}$$

och

$$\boldsymbol{\lambda}' = \left(\begin{pmatrix} \sum_U \boldsymbol{x}_k^* \\ \sum_s \frac{1}{\pi_k} \boldsymbol{x}_k^o \end{pmatrix} - \sum_r \frac{1}{\pi_k} \begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix} \right)' \left(\sum_r \frac{1}{\pi_k} \begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_k^* \\ \boldsymbol{x}_k^o \end{pmatrix}' \right)^{-1}$$

För ett resonemang om ovanstående teori, se kapitel 4.2. För fullständig härledning, se Särndal och Lundström (2005, kapitel 6).

När vi med hjälp av kalibreringen ovan har erhållit en skattning av antalet personer som skulle ha valt respektive typ av enkät kan vi återgå till att använda oss av den kombinerade metoden för att göra en populationsskattning av undersökningsvariabeln. Efter att imputeringen är gjord motsvarar det den första modellen som Cobben presenterar men i ett fall där individen själv väljer metod. Cobbens metod går som bekant ut på att lägga in metodvariabeln,

$$M_i = \begin{cases} 1 \; om \; person \; i \; har \; tilldelats \; insamlingsmetod \; 1 \\ 0 \; om \; person \; i \; har \; tilldelats \; insamlingsmetod \; 2 \end{cases}$$

, bland bakgrundsvariablerna och betrakta denna som en ursprunglig bakgrundsvariabel.

I vårt fall har vi efter beräkning av $\hat{t}_{x_{webb}}$ tillgång till InfoU (till skillnad från Cobben som känner till InfoUS) eftersom vi känner till både $\sum_U \boldsymbol{x}_k^*$ och värdet på varje enskild vektor $\boldsymbol{x}_k^*$ i svarsgruppen. Vi väljer alltså att betrakta det skattade värdet, $\hat{t}_{x_{webb}}$, på samma sätt som om det vore en siffra som var känd redan vid undersökningens början och använder oss av metoden i kapitel 4.2 för att göra populationsskattningen.

Den andra metoden som Cobben presenterar går ut på att dela upp datamaterialet i två delar och göra två olika skattningar genom viktning – en för webbenkät och en för pappersenkät. Något motsvarande tillvägagångssätt kan inte tillämpas i undersökningar där individerna själva väljer metod. Detta eftersom en förutsättning är att vi känner till totalsumman av samtliga bakgrundsvariabler i de två grupperna, $\sum_{\substack{webb \\ population}} \boldsymbol{x}_k^*$ och $\sum_{\substack{papper \\ population}} \boldsymbol{x}_k^*$, vilket är information som vi inte har tillgång till. I och med att vi inte vet vilka enskilda individer som skulle ha valt respektive metod i populationen kan vi omöjligt veta hur många kvinnor som skulle ha valt webbundersökning, inkomsten för individerna i populationen som skulle ha valt pappersundersökning et cetera. Att använda skattningar av alla dessa summor bör göra modellen alltför osäker för att ge tillförlitliga populationsskattningar av undersökningsvariabeln och vi väljer därför att inte använda oss av denna metod.

4.10 Förslag på sätt att hantera datamaterial som erhållits genom stratifierat urval i undersökningar med bortfall där individen själv väljer pappers- eller webbenkät

Antag att vi har ett stickprov som samlats in genom stratifierat urval och att bortfallet är MAR. Samtliga ursprungliga bakgrundsvariabler är kända för samtliga individer i populationen, men metodvalet är endast känt för de individer som lämnat in svar på minst en fråga i undersökningen. Återigen är vi ute efter ett sätt att skatta det totala antalet individer som skulle ha valt respektive metod i hela populationen. Om vi antar metodvalet inte är helt slumpmässigt utan att bakgrundsvariablerna har inverkan behöver vi en metod som tar hänsyn till detta val för att skatta det totala antalet webbsvar. Inom de olika stratumen väljer vi därför att även i detta fall göra skattningar med hjälp av viktning. Skattningarna inom stratumen går till på samma sätt som i kapitel 4.9 och vi har även i detta fall tillgång till InfoUS. För stratum h erhåller vi

$$\hat{t}_{x_{webb}h} = \sum_{k=1}^{n_h} w_k x_{webb_k}$$

där vikterna beräknas enligt formlerna i kapitel 4.2 och n_h betecknar antalet svarande personer i stratum h.

I en undersökning med H strata blir en skattning av det totala antalet personer i populationen som skulle ha valt webbundersökning därmed

$$\hat{t}_{webb} = \sum_{i=1}^{H} \hat{t}_{x_{webb}h}$$

När detta värde erhållits skattar vi populationstotalen av undersökningsvariabeln med hjälp av den kombinerade metoden som beskrivits i kapitel 4.4.

4.11 Ett verkligt exempel på metodeffekter – sammanfattning av en rapport av Kaiser

I en rapport från Center for European Economic Research analyserar Ulrich Kaiser (2001) ett datamaterial från en kvartalsvis undersökning, SSBS (Service Sector Business Survey), bland tyska företag i servicesektorn. Syftet med analysen är att komma fram till om skillnader mellan de svar som inkommit via webb- respektive pappersenkät kan upptäckas, alltså att hitta en eventuell metodeffekt. I undersökningen ombes företagen bedöma sin ekonomiska utveckling både tidigare och kommande kvartal, genom att svara på frågor om nuvarande och förväntade priser, vinster, anställningar, försäljning och efterfrågan. Från att tidigare enbart ha skickat ut en pappersenkät till företagen i stickprovet infördes 2001 möjligheten att välja webbenkät, som vid införandet valdes av 8,5 % av deltagarna.

Urvalet är stratifierat och består av 4000 företag, där man vid stratifieringen tagit hänsyn till vilken bransch inom servicesektorn företaget tillhör (tio indelningar), antalet anställda (fem storleksklasser) och om det är beläget i östra eller västra Tyskland.

Kaiser (2001) ställer upp två separata (binära probit-) modeller – en som beskriver deltagandet i undersökningen och en annan som beskriver metodvalet. I modellerna ingår ett flertal dummyvariabler som beskriver variablerna som legat till grund för stratifieringen.

Utifrån modellerna drar Kaiser slutsatsen att den enda av bakgrundsvariablerna som har inverkan på deltagandet i undersökningen är huruvida företaget har deltagit i samma undersökning tidigare. Inget systematiskt fel bör därför uppstå i bortfallet på grund av skillnader i bakgrundsvariablerna, förutom den som anger tidigare deltagande i undersökningen.

I modellen som behandlar metodvalet hos de deltagande företagen upptäcks en signifikant skillnad mellan större och mindre företag, där företagen med mer än 59 anställda i högre grad väljer webbundersökning framför pappersundersökning jämfört med företagen med 1-19 anställda. En signifikant skillnad mellan mjukvaruföretag och företag i renhållningsbranschen upptäckts också, där de förra väljer webbundersökning i större utsträckning än de senare. När en variabel som anger antal datorer per anställd läggs till i modellen försvinner dock signifikansen för mjukvaruföretagen, medan det fortfarande är en signifikant skillnad mellan svaren med avseende på företagens storlek, trots att den tillagda variabeln inte är signifikant i sig själv.

För att testa om det är någon (signifikant) skillnad mellan svaren som inkommit via de två insamlingsmetoderna genomförs Pearsons χ^2-test på respektive fråga i undersökningen. Resultatet blir att det enbart är signifikant skillnad mellan svaren på frågorna som behandlar bedömning av priser och försäljning, där det visar sig att företagen som svarat via pappersenkät är mer optimistiska i sin bedömning än de som lämnat in webbsvar. För att testa om skillnaden beror på bakgrundsvariablerna ställs en modell upp som har företagens benägenhet att ange försämrade, lika eller förbättrade resultat som responsvariabel och där bakgrundsvariablerna är desamma som tidigare (bortsett från variabeln som behandlar tidigare deltagande i undersökningen, vilken utesluts ur modellen). Resultatet blir att de som svarar på webben bedömer försäljningen som sämre än de som svarar på papper. Däremot upptäcks ingen signifikant skillnad när det gäller bedömningen av priserna mellan de två

insamlingsmetoderna. Kaiser finner även att det partiella bortfallet är högre bland de som svarat via webben än bland de som valt pappersenkäten.

4.12 Förslag på ett sätt att hantera Kaisers datamaterial för att göra skattningar av den ekonomiska utvecklingen hos alla tyska företag i servicesektorn

Kaiser kommer fram till att inga kända bakgrundsvariabler har inverkan på huruvida ett visst företag väljer att delta i undersökningen, utom om företaget tidigare deltagit i undersökningen. Bortfallet kan därmed sägas vara MAR, alltså slumpmässigt betingat på tidigare deltagande i undersökningen och vi måste därmed ta hänsyn till bortfallet i våra skattningar. Vi använder oss av metoden i kapitel 4.10 ovan för att göra skattningar av de olika undersökningsvariablerna hos samtliga tyska serviceföretag i fallen där en signifikant skillnad mellan insamlingsmetoderna finns, exempelvis bedömningen av det förväntade priserna kommande kvartal.

5 Diskussion

5.1 Är det rimligt att skatta metodvalet i en population?

I den typ av undersökningar som metoderna i denna uppsats är avsedda för utgår vi från att personerna i urvalsgruppen själva väljer insamlingsmetod. Man kan ifrågasätta om det överhuvudtaget är rimligt att göra skattningar av hur många individer i populationen som skulle ha valt pappers- respektive webbundersökning om de hade ingått i urvalet. En invändning skulle exempelvis kunna vara att en stor del, eller kanske till och med majoriteten, av den svenska befolkningen förmodligen är kapabel till att besvara enkäter både via internet och på papper och att man därför skulle kunna misstänka att det varierar kraftigt vilken av metoderna en viss individ väljer (vilket i sin tur bör kunna medföra stor variation i det sammanlagda antalet webb- och papperssvar i populationen).

Fram till mitten av 2007 lämnades 10-15% webbsvar i de SCB-undersökningar som gjorts där upplägget varit att individen själv väljer insamlingsmetod (Holmberg, Lorenc och Werner, 2007). Holmberg et al. (2007) visar genom en empirisk undersökning att man genom förändringar vid kontakten med urvalspersonerna med avseende på hur och när de två insamlingsmetoderna presenteras kan åstadkomma stora skillnader i andelen webbsvar. Undersökningen genomfördes i samband med en hälsoenkät som gick ut till drygt 22500 personer. Syftet med undersökningen var (förutom att få svar på frågorna i enkäten) att testa fem olika kontaktstrategier för att se om dessa skulle ha någon inverkan på antalet webbsvar. Urvalsgruppen randomiserades in i fem mindre grupper med likadan sammansättning med avseende på kön, ålder och tidigare deltagande i samma undersökning, för vilka varsin kontaktstrategi användes. Kontaktstrategierna gick ut på att de två insamlingsmetoderna presenterades på olika sätt och vid olika tillfällen i de fem grupperna. Resultatet blev att de olika strategierna gav stora skillnader i andelen webbsvar utan att ge någon större skillnad i bortfallsandelen, som vid samtliga strategier var mellan 24% och 29%. Genom att vid den första kontakten i en av grupperna enbart erbjuda webbundersökning utan att informera om kommande möjlighet till pappersenkät, för att vid den första påminnelsen återigen enbart skicka inloggningsuppgifter till webbenkäten men samtidigt informera om att både pappers- och webbenkät skulle dyka upp vid de två efterföljande påminnelserna erhölls så mycket som 64.7% webbsvar. Siffran kan jämföras med att det vid undersökningens slut hade inkommit 14.5% webbsvar från gruppen där individerna själva fick välja insamlingsmetod redan vid första utskicket.

Att drygt 35% av svarspersonerna valde pappersenkät trots att det var ett alternativ som blev tillgängligt först vid det tredje utskicket tyder på att en relativt stor andel av populationen har starka preferenser för att svara via pappersenkät. Motsvarande preferens för webbenkät verkar inte finnas, utan i gruppen där både det första och andra utskicket enbart innehöll en pappersenkät och där information om den kommande möjligheten att svara via webbenkät gavs först vid det andra utskicket erhölls så lite som 2.6% webbsvar trots att det sista utskicket enbart innehöll inloggningsuppgifter till webbenkäten. Att andelen webbsvar i den senare nämnda undersökningen var så pass mycket lägre än andelen papperssvar i den tidigare nämnda undersökningen (trots att man i båda fallen försökt ”påtvinga” individerna respektive metod) tyder på att det finns bakomliggande orsaker till att inte välja webbenkät medan så gott som alla svarspersoner kunde tänka sig att svara via pappersenkät. Det är rimligt att tro att några av bakgrundsvariablerna kan förklara preferensen för pappersenkät hos de 35% som valt denna insamlingsmetod trots att möjligheten inte fanns förrän vid det tredje utskicket. Ålder, avsaknad av dator och internet i hemmet och huruvida dator används i yrket kan exempelvis tänkas ha inverkan, vilket bör kunna visa sig genom att bakgrundsvariablerna åldersgrupp, inkomst och yrkesgrupp har inverkan på metodvalet.

I och med att såpass få personer valde webbundersökning när pappersenkät var det enda alternativet vid de två första utskicken kan vi misstänka att grovt sett är strax över 60% av svarspersonerna flexibla när det gäller metodval (eftersom andelen som valde webbenkät pendlar mellan 2.6% och 64.7%) och att dessa personer i hög grad påverkas av vilken kontaktstrategi som används vid utskicket. Att enbart 15% valde webbenkät då valmöjlighet fanns redan vid första utskicket kan förmodligen förklaras av att pappersenkät är det alternativ som för många ligger närmast till hands och att det därför blir det naturliga förstahandsvalet. Att det går att påverka andelen webbsvar tyder på att det också går att prediktera andelen webbsvar. Vi har anledning att tro att en viss del av populationen alltid kommer att välja pappersenkät, oavsett kontaktstrategi, medan en stor del av populationen kan tänka sig båda alternativen men har vissa preferenser som visar sig om vi ger valmöjlighet redan vid första utskicket. Att även preferenserna kan predikteras för de personer som kan tänka sig att använda sig av båda svarsmetoderna verkar troligt. Exempelvis skulle vi kunna tänka oss att äldre personer med dator i hemmet generellt sett använder den i mindre utsträckning och för andra ändamål än yngre personer. Att personer som vanligtvis använder datorn i låg utsträckning är mindre benägna att välja webbenkät (trots att de har alla förutsättningar som krävs för att göra det) bör vara ett rimligt antagande. Därmed har vi anledning att tro att det i fallet som denna uppsats behandlar, där individerna väljer metod redan vid första utskicket, går att hitta ett samband mellan bakgrundsvariablerna och metodvalet även för de personer som kan tänka sig att byta metod om de måste.

En helt annan möjlighet, som inte tidigare nämnts i denna uppsats, när man vill beräkna det totala antalet personer i populationen som skulle ha valt webbenkät är att gå igenom de deklarationer som lämnats in till Skatteverket eftersom det är så nära en totalundersökning som erbjuder både pappers- och webbalternativ man kan komma i dagsläget. I Sverige är det obligatoriskt att deklarera för individer med en förvärvsinkomst på minst 18104 kronor eller som uppfyller vissa andra kriterier (se Skatteverket, Vem ska lämna deklaration 2010?) och möjlighet att välja mellan att deklarera på såväl internet som pappersblankett finns. Vi kan därför tänka oss att ett alternativ för att komma åt det totala antalet personer i populationen som föredrar webbundersökning skulle kunna vara att använda sig av den information som finns om vilket val respektive person har gjort vid sin senast inlämnade deklaration och förutsätta att samma val skulle ha gjorts i de SCB-undersökningar som görs. Invändningarna

mot ett sådant tillvägagångssätt är dock många. Förutom att man inte kan komma åt varje enskild person i befolkningen eftersom en betydande andel personer inte behöver deklarera måste man ta hänsyn till att ett tredje alternativ, sms-deklaration, finns. Dessutom är det ytterst tveksamt om obligatoriska uppgifter som är straffbara att inte lämna kan jämföras med en urvalsundersökning. Att webbdeklarationen är utformad för att förenkla för individerna och dessutom ger fördelar i form av tidigare skatteåterbäring bör ge individerna incitament att välja webbalternativet som inte finns vid SCB-undersökningar. Innan man överväger att använda sig av möjligheten att titta på antalet webbdeklarationer behöver därför omfattande undersökningar göras som jämför metodval i urvalsundersökningar med metodval vid deklarationen för att bekräfta eventuella samband.

6 Slutsatser

6.1 Slutsatser

I denna uppsats har vi tagit fram förslag på hur hänsyn till insamlingsmetoden kan tas i några vanliga modeller som används för att göra populationsskattningar av undersökningsvariabler. Beroende på varför vi tror att metodeffekten har uppkommit i en given undersökning kan vi använda modellerna för olika ändamål. Under förutsättning att vi tror att individerna gör de metodval som de gör på grund av att de besitter vissa egenskaper som både påverkar metodvalet och svaren på de frågor som ställs bör de presenterade modellerna ge mer korrekta populationsskattningar rakt av. Om vi exempelvis har en undersökning där frågor angående datoranvändning ställs till ett antal personer är det troligt att de som väljer att svara via webbenkät är personer som i det dagliga livet använder datorer i större utsträckning än de som väljer att svara via pappersenkät. I en sådan undersökning vore det inte särskilt förvånande om svaren som inkommer via webbenkät skiljer sig signifikant från övriga svar med tanke på att den förra gruppen generellt sett har större datorvana än den senare. Det är därför troligt att skattningarna av undersökningsvariablerna blir bättre om vi tar hänsyn till insamlingsmetoden än om vi inte gör det.

En annan, vanligare, tolkning av metodeffekten är att det själva verket inte är någon signifikant skillnad mellan individerna som valt pappers- och webbundersökning, men att den valda insamlingsmetoden påverkar undersökningsdeltagarna att lämna svar som skiljer sig från de sanna svaren (som förutsätts finnas även i attitydundersökningar). Även under sådana antaganden finns det anledning att ta med metodvalet i modellen. Vi kan aldrig veta om ett enskilt svar som lämnas i en undersökning är sant eller inte. Däremot kan vi genom empiriska undersökningar där vi ställer frågor om uppgifter som vi redan har tillgång till (exempelvis taxerad inkomst, kön, ålder, födelseland, utbildning och liknande) komma fram till om det är så att en av metoderna generellt sett ger svar som är mer sanningsenliga än den andra. Om så är fallet bör modellerna som presenteras i denna uppsats kunna ses som ett första led i att korrigera de populationsskattningar som görs. Ett lämpligt nästa steg vore därför att testa hur väl modellerna fungerar på verkliga data och att fundera över hur modellerna kan användas för att korrigera skattningarna i undersökningar som görs i syfte att få värden som ligger så nära sanningen som möjligt.

7 Referenslista

Cobben F.(2009), *Nonresponse in Sample Surveys – Methods for Analysis and Adjustment*, Statistics Netherlands

Holmberg A., Lorenc B., Werner P., (2007), *Lotta P2 – Optimal kontaktstrategi vid blandad insamling papper och webb*, SCB

Kaiser U. (2001), *Differences in Response Patterns in a Mixed Mode – Online/Paper & Pencil Business Survey*, Center for European Economic Research http://econstor.eu/bitstream/10419/24471/1/dp0150.pdf

Kreuter F, Presser S., Tourangeau (2008), *Social Desirability Bias in CATI, IVR and Web Surveys – The Effects of Mode and Question Sensitivity*, Public Opinion Quarterly, Vol 72, No. 5

Kwak N., Radler B. (2002): *A Comparison between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality*, Journal of Official Statistics (18)

Little R.J.A., Rubin D.B. (2002), *Statistical Analysis with Missing Data*, Wiley Interscience

Skatteverket (2010), *Vem ska lämna deklaration 2010?* http://www.skatteverket.se/privat/skatter/deklarera2010/omdeklarationen/vemskalamnadeklaration.4.233f91f71260075abe8800014666.html?posid=6&sv.search.query.allwords=deklaration

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

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

Taylor, H, (2000), *Does Internet Research Work?*, International Journal of Market Research (42)

Valliant R., Dorfman A.H., Royall R.M. (2000), *Finite Population Sampling and Inference – A Prediction Approach*, Wiley Interscience

8 Appendix

Horvitz-Thompson-skattningens varians

$$Var(\hat{t}_\pi) = \sum\sum_U \Delta_{kl}\check{y}_k\check{y}_l$$

där

$$\Delta_{kl} = Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$$

Om $\pi_{kl} > 0 \; för \; alla \; k, l \; \epsilon \; U$ är en väntevärdesriktig skattning

$$\widehat{Var}(\hat{t}_\pi) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\check{y}_k\check{y}_l$$

För bevis hänvisas till Särndal et al.(1992, sid 44-48)

Differensskattningens varians

$$Var(\hat{t}_{y,dif}) = \sum\sum_U \Delta_{kl}\breve{D}_k\breve{D}_l$$

En väntevärdesriktig skattning är

$$\widehat{Var}(\hat{t}_{y,dif}) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\breve{D}_k\breve{D}_l$$

För bevis hänvisas till Särndal et al.(1992, sid 223)

Regressionsskattningens varians

Variansen beräknas approximativt med

$$ApprVar(\hat{t}_{yr}) = \sum\sum_U \Delta_{kl}\frac{(y_k - y_k^o)}{\pi_k}\frac{(y_l - y_l^o)}{\pi_l}$$

En skattning är

$$\widehat{Var}(\hat{t}_{yr}) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(g_{ks}\frac{(y_k - \breve{y}_k)}{\pi_k}\right)(g_{ls}\frac{(y_l - \breve{y}_l)}{\pi_l})$$

där

$$g_{ks} = 1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})'\widehat{\boldsymbol{T}}^{-1}\boldsymbol{x}_k/\sigma_k^2$$

och

$$\widehat{\boldsymbol{T}} = \sum_{\boldsymbol{s}}\frac{\boldsymbol{x}_k\boldsymbol{x}_k{}'}{\sigma_k^2\pi_k}$$

och

$$\boldsymbol{t}_x = (t_{x_1}, \dots, t_{x_J})'$$

och

$$\hat{\boldsymbol{t}}_{x\pi} = (\hat{t}_{x_1\pi}, \dots, \hat{t}_{x_J\pi})'$$

För bevis hänvisas till Särndal et al.(1992, sid 236-238)

Regressionsskattningens varians vid tvåstegsurval då vi har tillång till summan av bakgrundsvariablerna i respektive kluster

En skattning av variansen är

$$ApprVar(\hat{t}_{yAr}) = \sum\sum_{s_I}\Delta_{Iij}\left(\frac{g_{is_I}d_i}{\pi_{Ii}}\right)\left(\frac{g_{js_I}d_j}{\pi_{Ij}}\right) - \sum_{s_I}\frac{1}{\pi_{Ii}}\left(1 - \frac{1}{\pi_{Ii}}\right)g_{is_I}^2\,\hat{V}_i + \sum_{s_I}g_{is_I}^2\frac{\hat{V}_i}{\pi_{Ii}^2}$$

där

$$d_i = (t_{yi}^* - \boldsymbol{u}_i{}'\widehat{\boldsymbol{B}}_I)$$

och

$$\hat{V}_i = \sum\sum_{s_i} \Delta_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}}$$

Vikterna g_{is_I}är komplicerade att beräkna och approximeras ofta med $g_{is_I} = 1$ (Särndal et al., 1992, kapitel 8.4). För det exakta uttrycket för g_{is_I} och för en härledning av variansen hänvisas till Särndal et al. (1992, kapitel 8.4).

Regressionsskattningens varians vid klusterurval

Formlerna för tvåstegsurval gäller om vi sätter

$$t_{yi}^* = \sum_{U_i} y_k$$

och

$$\hat{V}_i = 0$$

för alla *i*.

Prediktionsskattningens varians

$$Var(\hat{\theta}) = \boldsymbol{g}_s{}'\boldsymbol{V}_{ss}\boldsymbol{g}_s$$

där

$$\boldsymbol{g}_s = -\boldsymbol{V}_{ss}^{-1}\boldsymbol{X}_s(\boldsymbol{X}_s{}'\boldsymbol{V}_{ss}^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}'(1, \dots, 1)$$

För bevis hänvisas till Valliant et al. (2000, kapitel 2.2)

Regressionsskattningens varians vid kluster- och tvåstegsurval då vi har tillgång till bakgrundsvektorn för de enskilda individerna i de utvalda klustren

Läsaren hänvisas till Särndal et al. (1992, kapitel 8.9, result 8.9.2)

Varians vid användning av den kombinerade metoden för hantering av bortfall

I praktiken används ofta fler än en imputeringsmetod i en och samma undersökning, vilket gör variansskattningen komplicerad. Inga universella metoder existerar och läsaren hänvisas därför till kapitel 13.4 i Särndal och Lundström (2005) för en genomgång av några möjliga tillvägagångssätt för att beräkna variansen approximativt.