Några synpunkter på situationen med dubbla urval

Daniel Thorburn
Statistiska institutionen
Stockholms Universitet
106 91 Stockholm

April 19, 2004

Abstract

I promemorian behandlas situationen när man har två urval från en och samma population. Ett sannolikhetsurval med enbart bakgrundsfrågor och ett urval med okänd metod där både bakgrundsfrågor och målvariabeln mätts. Vi antar "ignorability" dvs värdet på målvariabeln beror inte av vilket urval observationen kommer från, givet bakgrundsvariablerna. Olika metoder att analysera detta tas upp och diskuteras och problem med metoderna diskuteras. Ett antal frågeställningar redovisas.

1 Inledning

Ibland är det svårt eller dyrt att dra representativa urval. I så fall kan det vara frestande att dra ett stort urval på ett enkelt men icke-representativt sätt. och sedan försöka räkna om det på något sätt till totalpopulationen. I den situation vi skall titta på här har man som hjälp ett sannolikhetsurval där man visserligen inte känner värdet på den studerade variabeln men väl på ett antal frågor som är relaterade till den och som också de i det icke-representativa urvalet får svara på. Vi kommer att ta upp två olika situationer inom denna ram. En som i någon mening svarar mot diskreta variabler och en som motsvarar vanliga kontinuerliga situationer.

De två avsnitten kan läsas oberoende av varandra och i valfri ordning. De kompletterar dock varandra genom att beskriver samma situation men sedd från två olika utgångspunkter.

2 Diskreta fallet

Här tänker vi oss data i en trevägs kontingenstabell där olika rader har olika urvalssannolikhet, olika kolumner har olika värden på den studerade variabeln och

den sista marginalen saknar samband med urvalet och den studerade variabeln. Det ena urvalet är ett riktigt slannolikhetsurval så att vi kan skatta urvalssannolikheterna bra medan vi inte observerar den studerade målvariabeln. I det andra observerar vi målvariablen men det är inget sannolikhetsurval och vi kan alltså inte skatta sannolikheterna att hamna i de olika raderna.

2.1 Beteckningar

Vi har en population U. Det finns "hjälpvariabler" I, J och K som är diskreta och en studerad variabel Y. Ingen av dessa är kända från början, men I skall tänkas vara relaterad till Y och J till urvalssannnolikheten.

Från U dras två stickprov, S som är ett slumpmässigt (OSU) urval med storlek n och T med storlek m som dras på något annat sätt. (I teorin finns dock inget som säger att T måste vara dragen från U, men det känns naturligt. Ibland låter man T vara dragen i två steg där det första steget representerar en undergrupp till U som vi kan kalla U_T (t ex resenärer med kollektivtrafik) och ur den drar man ett urval på korrekt vis. Detta tvåstegsförfarande gör bara beteckningarna krångligare utan att tillföra något)

I S observeras värdena på $I; J$ och K och våra data blir $n_{i,j,k}$ dvs antalet med kombinationen i, j och k. Sannolikheten att få en uppsättning i, j och k skattas med $\widehat{p_{i,j,k}} = n_{i,j,k}/n$.

I T observeras I, J, K och Y. Våra observationer blir $m_{i,j,k}$ dvs antalet med kombinationen i, j och k och $y_{i,j,k,l}$ där l går fråm 1 till $m_{i,j,k}$. Förväntat värde på y för uppsättningen i, j, k skattas med $\overline{y_{i,j,k}}$. "Sannolikheten" att få en uppsättning i, j och k uppskattas med $\widehat{q_{i,j,k}} = m_{i,j,k}/m$.

2.2 Skattningar

Nu går jag över till skattningar och skriver ner fyra olika tänkbara varianter.

2.2.1 Enkel skattning

Den naturliga skattningen av medelvärdet i populationen U blir då

$$\sum_i \widehat{p_{i,j,k}}\,\overline{y_{i,j,k.}}$$

För att den skall vara väntevärdesriktig krävs att $E(\overline{y_{i,j,k}})$ är densamma i både S och T och att $\widehat{p_{i,j,k}}$ och $\overline{y_{i,j,k}}$ är okorrelerade. Ofta kräver man strongly ignorable, vilket är starkare (Den andra delen påpekas sällan explicit).

2.2.2 Radvis eller imputering

Nu tänker jag göra lite starkare antaganden. Antag att I är vald så att $E(\overline{y_{i,j,k}})$ bara beror av i dvs $E(\overline{y_{i,j,k}}) = f(i)$. (I praktiken får man skatta vad som är i genom t ex regression av y på i, j, k). I så fall kan $\overline{y_{i,j,k}}$ ersättas av $\overline{y_{i,.,.}}$ (eller ett vägt medelvärde där hänsyn också tas till precisionen i skattningarna). Nu blir en naturlig skattning av medelvärdet $\sum_{ijk} \widehat{p_{i,j,k}}\,\overline{y_{i,.,.}}$. Detta kan förenklas till

$$\sum_i \widehat{p}_i\,\overline{y_{i,.,.}}$$

där $p_i = n_i/n$ och n_i är antalet observationer med värdet i.

2.2.3 Kolumnvis eller vägning

Nu släpper vi det antagandet och koncentrerar oss i stället på J. Vi kommer att titta på två olika alternativa antaganden. Det första säger att urvalssannolikheten $p_{i,j,k}$ bara beror av j dvs $p_{i,j,k} = g(j)$. Då kan man precis som ovan ersätta $\widehat{p_{i,j,k}} = n_{i,j,k}/n$ med $\widehat{p}_j/N_{I,K} = n_j/\left(nN_{I,K}\right)$, där $N_{I,K}$ betyder antal klasser för $i \times k$ och n_j antalet S-observationer i klass j. Skattningen blir nu $\sum_{i,j,k} \left(\widehat{p}_j/N_{I,K}\right)\ \overline{y_{i,j,k}}$. Detta kan också förenklas till

$$\sum_j \widehat{p}_j\ \overline{\overline{y_{.,j,.}}}$$

där $\overline{\overline{y_{.,j,.}}}$ är det raka medelvärdet $\sum_{i,k} \overline{y_{i,j,k}}/N_{I,K}$.

2.2.4 Kolumnvis eller propensity score

Det andra tänkbara antagandet för J är det som görs vid propensity score, att kvoten $p_{i,j,k}/q_{i,j,k}$ bara beror av j dvs $p_{i,j,k} = g(j)\ q_{i,j,k}$. Då kan man ersätta $\widehat{p_{i,j,k}} = n_{i,j,k}/n$ med $\left(\sum_{i,k} n_{i,j,k}/\sum_{i,k} m_{i,j,k}\right) m_{i,j,k}/n = (mn_j/nm_j)\,\widehat{q_{i,j,k}}$. och få skattningen $\sum_{i,j,k} (mn_j/nm_j)\,\widehat{q_{i,j,k}}\,\overline{y_{i,j,k}}$. Detta är den vanliga propensity-score skattningen. Det ser inte direkt ut att vara någon förenkling men eftersom $\widehat{q_{i,j,k}}$ och $\overline{y_{i,j,k}}$. kommer från samma stickprov kan summan förenklas till

$$\sum_{,j} (mn_j/nm_j)\,\widehat{q}_j\,\overline{y_{.,j,.}} = \sum_{i,j,k} \widehat{\widehat{p}}_j\ \overline{y_{.,j,.}}$$

där $\overline{y_{.,j,.}}$ är det raka medelvärdet $\sum_{i,k} y_{i,j,k}/m_{i,k}$. och $\widehat{\widehat{p}}_j\ = (mn_j/nm_j)\,\widehat{q}_j$. Detta är den vanliga skattningen vid propensity score (bortsett från att någon ny klassindelning av J inte görs).

2.3 Frågeställningar

Nu kan man ställa ett antal forskningsbara frågor. Flera av dem kan i sin tur delas upp i mindre frågor. Jag bara skissar på några av dem.

Hur bestämma I resp J om man inte vet dem på förhand? Med linjär resp. logistisk regression? Hur påverkas skattningens bias av att dessa inte är kända?

Kan man kombinera de två ansatserna? Dvs vad göra om vi både vet att väntevärdet för Y bara beror av I och att propensity score (urvalssannolikheten) bara beror av J? Blir kombinationen ännu bättre?

Skatta standardavvikelserna i de fyra fallen 1,2,3 och 4 ovan. dvs när man verkligen känner I resp J och inte behöver beräkna den. Går det att säga om/när 2 eller 4 (3) är bäst? Vilka antaganden behövs för att komma åt osäkerheten t ex strongly ignorable plus något om osäkerheten i T.

Om T utgör en substantiell del av U, kan man förbättra skattningen (slumpfelsskattningarna) genom att bara försöka skatta den resterande delen av den ändliga populationen U?. Det borde vara möjligt om vi för observationerna i S kan avgöra om de också finns i T. Går det att utnyttja liknande om S är en substantiell del av U?

Vad händer när man inte har kvalitativa utan kvantitativa bakgrundsvariabler? Är klassindelning som föreslås vid propensity scores verkligen den bästa ickeparametriska metoden eller är det ännu bättre att använda någon annan skattningsform t ex kärnskattningar? (Rent parametriska metoder är naturligtvis också tänkbara).

Hur skall detta formuleras om urvalet S dras på annat sätt än SRS men fortfarande sannolikhetsurval med kända inklusionssannolikheter? Om T också är ett sannolikhetsurval, har man någon glädje av S?

3 Det kontinuerliga fallet

Nu går vi över till fallet då alla studerade variabler kan tänkas vara kontinuerliga och då tekniker som linjära modeller eller generaliserade linjära modeller är naturliga. Avsnittet är i stort sett fristående från dett diskreta fallet.

3.1 Bakgrund och definitioner

Populationen P består av N element som alla är förknippade med två variabler Y och X (en kolumnvektor $k \times 1$). Från populationen dras dels ett OSU-urval, S, med n_S individer, dels ett annat urval, T, med n_T individer men med okänd urvalsmetod. I S observeras bara X medan i T observeras både Y och X.

Urvalet T antas vara draget så att urvalet inte beror av Y på annat sätt än genom motsvarande X-värden.

3.2 Imputering

Vi antar nu att fördelningen för Y_j uppfyller

$$Y_j|X_j \in D(AX_j, \sigma^2_{Y|X})$$

där D betyder en fördelning med argumenten som väntevärde och varians och A är en radvektor. Vi inför också en beteckning $\sigma^2_Y = \mathrm{Var}(Y_j)$ för variansen av Y_j obetingat av X_j. Man inser lätt att

$$\sigma^2_Y = \sigma^2_{Y|X} + A\Sigma_{X|S}A'$$

där $\Sigma_{X|S}$ betyder kovariansmatrisen för X-värdena i populationen (och stickprovet).

3.2.1 Generell linjär regressionsestimator (GREG)

Vi börjar med att skatta A med vanliga regressionsmetoder med hjälp av T-urvalet, dvs

$$A^* = Y_T X'_T (X_T X'_T)^{-1}$$

där Y_T är en radvektor med alla Y-värden i T och X_T är en $k \times n_T$-matris med alla X-värden i T.

Om man antar att alla Y_i är oberoende (eg behövs bara okorrelerade) för olika i (givet T och X_T), blir variansen

$$\mathrm{Var}(A^*) = \sigma^2_{Y|X}(X_T X'_T)^{-1}$$

.

Sedan imputerar vi alla de saknade Y-värdena i S med den bästa linjära prediktorn $\widehat{Y_i} = A^* X_i$ och skattar populationsmedelvärdet med S-stickprovsmedelvärdet av de imputerade värdena dvs

$$\widehat{\overline{Y_P}} = \widehat{\overline{Y_S}} = A^* \overline{X_S}$$

(Denna formel håller bara om det inte finns några gemensamma värden i S och T. Om det finns det vet vi ju motsvarande Y-värde och behöver inte imputera det. Vi imputerar bara saknade Y-värden)

Man inser lätt att det förväntade medelkvadratfelet hos skattningen är

$$\begin{aligned}
\mathrm{E}(\widehat{\overline{Y_P}} - \overline{Y_P})^2 &= \mathrm{E}(\widehat{\overline{Y_S}} - \overline{Y_S})^2 + \mathrm{E}(\overline{Y_S} - \overline{Y_P})^2 \\
&= \sigma^2_{Y|X}\overline{X_S}'(X_T X'_T)^{-1}\overline{X_S} + \frac{N - n_S}{N n_S}\sigma^2_Y \\
&= \sigma^2_{Y|X}\overline{X_S}'(X_T X'_T)^{-1}\overline{X_S} + \frac{N - n_S}{N n_S}(\sigma^2_{Y|X} + A\Sigma_{X|S}A')
\end{aligned}$$

(om Y_i är oberoende för olika i givet T och X_T). Första delen här beror på att regressionskoefficienterna A är skattade med ett skattningsfel. Andra delen är det sedvanliga urvalsfelet, som beror på att S är ett stickprov.

Remark 1 *Om $\sigma^2_{Y|X}$ får bero av x på ett känt sätt går ett liknande reonemag igenom, men med skattningen $A^* = Y_T D X'_T (X_T D X'_T)^{-1}$ där $D = \mathrm{diag}(1/\sigma^2_{Y|X})$ och motsvarande ändringar i $\mathrm{Var}(A^*)$ och $\mathrm{E}(\widehat{\overline{Y_P}} - \overline{Y_P})^2$.*

Remark 2 *Det måste vara fullt möjligt och kanske till och med lämpligt att i praktiken skatta parametrar och osäkerhet med multipel imputering. Men om man vill få fram formler för att jämföra precisioner rent teoretiskt är nog denna formelbaserade ansats bättre.*

3.2.2 "Imputerad score-skattning"

Antag nu att man är osöker på om modellen verkligen är linjär. En något mer icke-parametrisk variant skulle vara att först skatta funktionen A^*X som ovan, men bara använda den som en stratifieringsgrund. Dvs dela in det slumpmässiga urvalet i M delar .

Vi bestämmer alltså $M-1$ gränser, $-\infty = a_0, a_1, ..., a_{M-1}, a_M = \infty$, (t ex så att $\#\{\ i;\ \ i \in S$ och $a_j < A^*X_i < a_{j+1}\}$ är lika stort för alla j $(= n_S/M)$). Därefter skulle man kunna skatta totalen med

$$\sum_j \frac{\#\{i;\ i \in S\ och\ a_j < A^*X_i < a_{j+1}\}}{\#\{i;\ i \in T\ och\ a_j < A^*X_i < a_{j+1}\}} \sum_{i;\ i\in T\ och\ a_j<A^*X_i<a_{j+1}} Y_i$$

Det skulle betyda att man ansätter att nivån, $\mathrm{E}(Y_i)$, är en växande funktion av AX men att den inte behöver vara linjär. Men observera att vi fortfarande antar att alla med samma värde på AX har samma förväntade värde på målvariabeln.

Detta är en parallell med Terhanians propensity score skattning som beskrivs längre fram. Eftersom det imputerade värdet är helt linjär i väntevärdet, är det förmodligen enklare att beräkna variansen och slumpfelet hos denna skattning än för Terhanians metod där olinjariteten hos propensityfunktionen också stör. Men att det kan uppstå problem visar vi med tre exempel som illustrerar olika aspekter på problemen.

Example 3 ***Olinjariteten.*** *För att se ett problem med olinjariteten antar vi att ett T-populationen är samlad i punkterna $(0,0)$, $(1,0)$ och $(1,1)$ och att rätt modell är $E(Y) = (X_1 + X_2)^2$. I så fall skattas det linjära sambandet med $E(Y) = X_1 + 3X_2$. Om den verkliga populationen skiljer sig från fördelningen i T blir det fel på grund av icke-linjariteten. Om populationen t ex är lika fördelad i punkterna $(0,0)$, $(1,0)$, $(0,1)$ och $(1,1)$ blir medelvärdesskattningen 2 istället för rätt värde $1,5$.*

Example 4 ***Stratumindelningen*** *innebär också att om T och S har olika skev fördelning inom klasserna blir väntevärdena inom stratum olika. T ex Låt*

modellen vara $E(Y) = X_1 + X_2$ *och antag att ett stratum är* $0 < x_1 + x_2 < 1$. *Om* T *är likformig på linjen* $x_1 = x_2$ *och* $U = S$ *ligger likformigt i övre kvadranten* $(x_1 > 0;\ x_2 > 0)$*, blir väntevärdet för* T *i stratat lika med* $1/2$ *och för värdena i* U *borde de vara* $2/3$. *Liknande problem kan uppstå beroende på att gränsen mellan strata inte bestäms rätt.*

Example 5 ***Slumpmässigheten*** *i stratumindelning kan också ge problem som inte skulle uppstå om kurvan kunde uppskattas perfekt. Låt* T *vara likformig på linjen* $x_1 + x_2$ *mellan* $(0,0)$ *och* $(1,1)$ *och* S *likformig på linjen mellan* $(0,0)$ *och* $(2,0)$. *Antag vidare att regressionslinjen blir* $(5x_1 + 3x_2)/4$ *eller* $(3x_1 + 5x_2)/4$ *med samma sannolikhet (Rätt värde som ovan* $x_1 + x_2$*). Om nu* S *delas i fem lika delar svarar de i det första fallet mot intervallen mellan punkterna* $(0,0), (0.25, 0.25), (0.5, 0.5), (0.75, 0.75), (1,1)$ *och* $(1.25, 1.25)$ *i* T *respektive* $(0,0), (0.15, 0.15), (0.3, 0.3), (0.45, 0.45), (0.6, 0.6)$ *och* $(0.75, 0.75)$ *också i* T. *Den översta punkten måste i båda fallen ersättas avv* $(1,1)$ *eftersom* T *inte går längre.Medelvärdet i de fem intervallen blir* $0.25, 0.75, 1.25, 1.75$ *resp.* 2. *Den sista skattningen blir dock tveksam eftersom det bara finns en gemensam punkt i stratat och* T *vilket ger problem när man bara drar ändliga stickprov och den punkten kanske inte kommer med. Motsvarigheten för den andra lutningen blir* $0.15, 0.45, 0.75, 1, 05$ *resp* 1.6. *I det här fallet blev ändå det förväntade värdet rätt men det är inte nödvändigt att så blir fallet..*

3.3 Propensity

3.3.1 HT-skattning med propensity

Låt $\pi(x)$ vara sannolikheten att en enhet med X-värdet, $X = x$, kommer att ingå i stickprovet T. Motsvarande sannolikhet att enheten ingår i S är självklart lika med n_S/N. Observera att denna sannolikhet kan bero både på fördelningen för X och på urvalsmekanismen för T. Det är en modellbaserad sannolikhet och inte en designbaserad sannolikhet. (Designbaserade sannolikheter definieras givet hela populationens alla X-värden. Då kommer denna sannolikhet nämligen att kunna bli 0 eller 1 (beroende på om x-värdet ingår i X_T eller ej).

Vi inför nu en sannolikhet som vi kallar propensity,

$$\pi_{ps}(x) = \pi(x)/(\pi(x) + n_S/N)$$

Den kan sägas betyda sannolikheten att en enhet tillhör T givet att den har X-värdet $X = x$ och dessutom tillhör $S \cup T$ (Denna tolkning måste modifieras något eftersom enheten kan finnas både i S och T. Men det är som sagt bara tolkningen som behöver modifieras. Definitionen är det inget fel på). Omvänt kan man lösa ut

$$\pi(x) = n_S \pi_{ps}(x)/(N(1 - \pi_{ps}(x)))$$

Vid propensity-teorin antar man ofta att propensity har en känd funktionell form. Här ansätter vi ett logistiskt uttryck (men andra former som t ex probit-

modeller är tänkbara)

$$\pi_{ps}(x) = 1/(1 + \exp(Bx))$$

där B är en radvektor med okända parametrar, vilket omvänt ger att

$$\pi(x) = \frac{n_S}{N \exp(Bx)}.$$

Nu studerar vi det sammanslagna stickprovet $S \oplus T$, som innehåller alla element i $S \cup T$, men de element som finns i både S och T finns med två gånger i $S \oplus T$. Med sedvanliga metoder från logistisk regression kan man där lätt skatta B med B^* och kovariansmatrisen $\mathrm{Var}(B^*) = \mathrm{Var}(B^*|X_T)$. (För att denna variansskattning skall vara riktig måste man göra sedvanliga oberoendeantaganden om inklusionen av olika enheter i T givet X). Variansskattningen vid logistisk regression är inget slutet uttryck men de flesta program kan redovisa den skattade kovariansmatrisen numeriskt. En vanlig metod är att använda den inverterade observerade Fisher-informationen (likelihoodfunktionens andraderivata).

Man kan också sätta in B^* i uttrycken för sannolikheterna och får

$$\begin{aligned} \widehat{\pi(x)} &= n_S \widehat{\pi_{ps}(x)}/(N(1 - \widehat{\pi_{ps}(x)})) \\ &= \frac{n_S}{N(\exp(B^* X)} \end{aligned}$$

Eftersom detta är en skattning av sannolikheten att en individ med X-värdet x hamnar i stickprovet T, kan vi formellt skriva upp HT-estimatorn av summan av alla Y-värden i populationen.

$$Y^*_{Pop} = \sum_T Y_i / \widehat{\pi(X_i)} = \frac{N}{n_S} \sum_T Y_i \exp(B^* X_i)$$

En variant på HT-estimatorn (som brukar anses något bättre) är

$$Y^{**}_{Pop} = N \frac{\sum_T Y_i \exp(B^* X_i)}{\sum_T \exp(B^* X_i)} = N\widetilde{Y}$$

där tilde, $\widetilde{\bullet}$, betecknar HT-skattningen av $\bullet$.

Detta svarar i viss mening mot den vanliga skattningen vid propensity score under förutsättningen att klassindelningen görs infinitesimalt liten och logistisk regression är rätt modell. Men den väger med den estimerade modellurvalssannolikheten, inte med den observerade andelen i varje stratum.

Nå hur kan man skatta variansen för denna estimator om modellen är rätt? Det finns två osäkerhetskällor, dels är B^* en skattning, dels finns ett urvalsfel. Dessa kan separeras med följande formel

$$\begin{aligned} \mathrm{Var}(Y^{**}_{Pop}) &= E(\mathrm{Var}(Y^{**}_{Pop}|T)) + \mathrm{Var}(E(Y^{**}_{Pop}|T)) \\ &\approx \mathrm{Var}(Y^{**}_{Pop}|Y_T, X_T) + \mathrm{Var}(Y^{**}_{Pop}|\pi = \widehat{\pi}) \end{aligned}$$

Den första termen beror på osäkerheten i B^* och kan uppskattas med sedvanlig Gaussapproximation och den ovan angivna metoden för att beräkna $\mathrm{Var}(B^*|X_T)$. Den andra termen är urvalsvariansen och kan t ex skattas med Horvitz-Thompsons förslag eller den vanliga Sen-Yates-Grundy estimatorn. För att det skall vara möjligt att göra dessa variansskattningar måste man dock göra ytterligare några antaganden. Vi behöver nämligen andra ordningens inklusionssannolikheter för att kunna beräkna urvalsvariansen. Här kommer vi att anta Poisson-urval eftersom det är enkelt.

Gaussapproximationen ger efter vissa räkningar att

$$\begin{aligned} & \mathrm{Var}(Y^{**}_{Pop}|X_T) \\ = \; & \mathrm{Var}(\frac{\sum_T Y_i(\exp(B^*X_i))}{\sum_T \exp(B^*X_i)}) \\ \approx \; & (\widetilde{YX} - \widetilde{Y}\widetilde{X})' \,\mathrm{Var}(B^*|X)(\widetilde{YX} - \widetilde{Y}\widetilde{X}) \end{aligned}$$

där $\widetilde{YX}$ och $\widetilde{X}$ är kolumnvektorer av vägda skttningar och ' betecknar transponat.

Urvalsvariansen (för den första HT-skattningen) skattas med (något i stil med)

$$\begin{aligned} \mathrm{Var}(Y^*_{Pop}|B^*) \; & \approx \; \sum_{T\times T} \frac{\widehat{\pi(X_i,X_j)} - \widehat{\pi(X_i)}\widehat{\pi(X_j)}}{\widehat{\pi(X_i,X_j)}\widehat{\pi(X_i)}\widehat{\pi(X_j)}} Y_iY_j \\ & \approx \; \sum_T \frac{1-\widehat{\pi(X_i)}}{\widehat{\pi(X_i)}^3} Y_i^2 \end{aligned}$$

där $\widehat{\pi(X_i,X_j)}$ är den skattade andra ordningens inklusionssannolikhet att både i och j ingår i T givet X_i och X_j. Vi ansatte Poissonurval där $\widehat{\pi(X_i,X_j)} = \widehat{\pi(X_i)}\widehat{\pi(X_j)}$ om $i \neq j$, vilket gav den andra raden. Poissonurval innebär att vi antagit att n_T är stokastiskt.

Det är fullt möjligt att approximativt skatta även variansen för den andra skattningen, Y^{**}_{Pop}. Ett sätt är att använda Taylorlinjarisering och gaussapproximation på hela kvoten. Ett annat sätt är att anta urvalssannolikheterna $\exp(B^*X_i)/\sum_T \exp(B^*X_j)$ och fix urvalsstorlek. Det finns då inte något lika naturligt val av andra ordningens inklusionssannolikheter, men man borde kunna välja något i stil med

$$\frac{\exp(B^*(X_i+X_j))}{\sum_T \exp(B^*X_k)((\sum_T \exp(B^*X_k)) - (\exp(B^*X_j) - \exp(B^*X_i))/2)}$$

Eftersom vi antagit fix urvalsstorlek går Yates-Grundy-Sen-estimatorn att använda.

3.3.2 Terhanians propensity score

Propensity score skattningen enligt Terhanian är mindre modellberoende än skattningen ovan. Han använder bara propensity skattningen till att bestämma

stratumgränser, inte för att bestämma uppräkningsfaktorer. Dessa bestäms istället empiriskt.

Han bestämmer alltså $M-1$ gränser, $-\infty = b_0, b_1, ..., b_{M-1}, b_M = \infty$, (t ex så att $\#\{\ i;\ \ i \in S$ och $b_j < B^* X_i < b_{j+1}\}$ är lika stort för alla j $(= n_S/M)$). Därefter skattar han totalen med

$$\sum_j \frac{\#\{i;\ i \in S\ och\ b_j < B^* X_i < b_{j+1}\}}{\#\{i;\ i \in T\ och\ b_j < B^* X_i < b_{j+1}\}} \sum_{i;\ i \in T\ \ och\ \ b_j < B^* X_i < b_{j+1}} Y_i$$

Även här torde det gå att approximativt uppskatta variansen numeriskt genom att dela upp den i två komponenter. Urvalsdelen givet B^* och b består här av två komponenter, dels urvalet som ger S, dels urvalet som ger T. Även här behövs ett antagande om hur T väljs. Ett rimligt och enkelt antagande skulle kunna vara att, givet vilket startum vi är i, väljs enheterna oberoende av varandra från en superpopulation (som beror på urvalsmekanismen till T) och att antalet T-observationer i varje stratum är känt.

I exemplen under "imputerad score" pekade jag på några problem med stratumindelningen även om modellen är korrekt. T ex så såg vi att om fördelningen i T och S skiljer sig åt så kan icke-linjariteten och stratumindelningen innebära att en skattning inte längre blir väntevärdesriktig Samma problem uppträder här och dessutom ytterligare några som beror på att inklusionssannolikheten inte är linjär i BX..

Om man antar att skattningen ovan alltid är väntevärdesriktig tror jag att man approximativt kan bortse från osäkerheten i B^*, (eftersom även variansen betingad av B^* är rimlig att presentera. Detta enligt betingningsprincipen, eftersom B^* är approximativt ancillär). Men om skattningen är biased ger detta en underskattning. Det kanske går att beräkna variansen med resampling t ex bootstrap. Möjligen går det till och med att skatta biasen med resamplingmetoder.

Jag skrev formeln ovan så att det inte krävs att intervallen är lika stora vilket Terhanian kräver och som jag antog i exemplet i parentesen. Hur stora intervall bör man välja? (Kanske Hodges-Dalenius kumulativa rot-regel tillämpad på $B * X$ eller $(B * X)^2$).

3.4 Kalibrering

Kalibrering är en annan form av omvägning som skulle gå att använda här. Jag går inte igenom den här. Men bl a Sixten Lundström har arbetat 1med detta och visat hur man kan göra formella variansskattningar genom att utnyttja s.k g-vikter. Det är inte en modell-baserad ansats utan snarare algoritm-baserad. Resultatet beror mycket på vilken avståndsfunktion man väljer att minimera och det är inte nu självklart för mig vilken som är bäst i denna situation.

3.5 Diverse lösa tankar och frågeställningar

Vi har sett lite olika varianter av ignorability antagandet

1) Y beror av X (och π) endast genom en linjär funktion av X: AX.
2) Y och π är oberoende givet X
3) Y och π är oberoende givet AX
4) Y och π är oberoende givet BX
5) π beror av X (och Y) endast genom en linjär funktion av X; BX
hänger dessa ihop? Man inser lätt att 5) medför 2) och 4) och att 1) medför 2) och 3). Observera att det första och sista formuleringen (1) och 5)) är Glim-antagande.

Jag tycker att man borde sätta sig ner och jämföra hur stora de olika varianserna blir och jämföra de två (tre) skattningsmetoderna. Man borde också fundera på om de skulle gå att kombinera.

Det viktiga för att hantera variansskattningen i första fallet är att Y är oberoende givet X och T. Då blir variansskattningen för A^* rätt och därefter variansen för hela skattningen. Dessutom behövs att väntevärdet av Y är linjärt i X. (men det kan naturligtvis ersättas av annat t ex Bernoulli-logististisk om Y är en $0-1$-variabel.

I det andra fallet med omvägning med skattad propensity score behövs mer. Dels att inklusionen av olika element i T sker oberoende av varandra (givet X). Annars blir variansen för B^* inte rätt och inte heller skattningen av urvalsvariansen. Dessutom behövs att inklusionssannolikheten i T är ett exponentiellt uttryck i X. (Om man ansätter probit-regression blir det naturligtvis något annat här).

Vad som behövs för Terhanians version är inte helt utrett ovan. Jag gissar att ungefär samma villkor som för parametrisk propensity score behövs. Om inte inklusionen av olika element i T sker oberoende gissar jag att B^* inte ens behöver bli väntevärdesriktig. I så fall skulle inte heller Terhanians skattning fungera. Eftersom Terhanians metod är mer icke-parametrisk för inklusionssannolikheten, skulle kanske antagandet om att den är exponentiell kunna försvagas.

Jag har i score avsnitten inte krävt att intervallen är lika stora vilket Terhanian gjorde. Hur stora intervall bör man ha för att få bra skattningar.

Som någon kanske råkat höra någon gång är bayesianska metoder att föredra rent generellt. Inget här är dock gjort bayesianskt utan alla subjektiva antaganden är enpunktsfördelade (som brukligt i klassisk teori), t ex antaganden som Strong Ignorability eller konstant varians. Mycket här skulle kunna förbättras med bayesiansk teori, men detta är nog tillräckligt besvärligt att tränga igenom.

Skattningsmässigt skulle det i några fall ligga nära till hands att använda MCMC-metoder för att hitta ML-skattningar. Det skulle nog vara möjligt att analysera problemet med dem.