Egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler

Sixten Lundström

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1991:19

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1991:19. Egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler / Sixten Lundström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1991-X101OP9119

Egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler

Sixten Lundström

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1991:19

Från trycket November 1991
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Åke Lönnqvist
Förfrågningar Sixten Lundström, tel. 019/17 64 96

ISSN 0283-8680
Garnisonstryckeriet, Stockholm

U/STM-Ö
Sixten Lundström

1991-06-28

Egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler.

Sixten Lundström, U/STM

Sammanfattning.

I framtiden kommer det troligen att bli mer vanligt med modeller. Ökat utnyttjande av hjälpinformation och utvecklingen av hård- och mjukvaran i databehandlingen talar för detta. Modeller kan utnyttjas i flera moment i en undersökning, men i denna rapport ska vi begränsa oss till estimationsfasen. Närmare bestämt studerar vi egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler.

En estimator kan vara direkt beroende av en modell på så sätt att om modellen är felaktig blir estimatorn biased. Detta är vanligt för estimatorer vid små redovisningsgrupper t ex då vi utnyttjar s k syntetiska estimatorer. Här studerar vi dock enbart estimatorer som är asymptotiskt unbiased d v s vi håller oss till redovisningsgrupper som inte är alltför små.

Vid utnyttjandet av modeller är det naturligt att starta med en generell ansats för att sedan välja ut en "bästa" metod. I föreliggande rapport utnyttjas en generaliserad regressionsansats ur vilken en stor mängd (i princip ett oändligt antal) specialestimatorer kan erhållas - väl kända är kvotestimatorn och den klassiska regressionsestimatorn. Förutom egenskaperna hos punktestimatorerna studeras även egenskaperna hos variansskattningarna samt konfidensintervallen.

Allmänt gäller att ju bättre hjälpinformation man har och ju bättre man förstår sambanden med undersökningsvariabeln desto bättre estimator kan man finna. Nu lyckas man inte alltid så bra och därför vill man veta vad följderna blir av detta. I rapporten visas hur modeller kan konstrueras och hur stora felspecificeringar av modellen som estimatorerna "tål". Några datorprogram är också framställda för detta.

Framställningen baserar sig i hög grad på boken "Model Assisted Survey Sampling" av Särndal, Swensson och Wretman.

INNEHÅLL

1. Inledning

Det som i hög grad skiljer stickprovsteorin från annan statistisk teori är utnyttjande av hjälpinformation. Läroböcker som Cochran [1977], Des Raj [1972], Kish[1965] visar hur man kan använda den i olika stadier av en statistisk undersökning, i urvalsdesignen för att t ex göra stratifierade urval eller i estimationen för bildande av exempelvis kvot- eller regressionsestimatorer.

Vid SCB är det vanligt att använda hjälpinformation i urvalsdesignen och då framförallt för stratifiering av populationen. I estimationen kan man se en hel del kvotestimatorer och även enstaka regressionsestimatorer. I vissa produkter "mjölkar" troligen den omfattande stratifieringen ut en stor del av hjälpinformationen, men vi menar att det finns anledning att tro att den stora mängden registervariabler vi har på SCB inte är utnyttjad till fullo. Utnyttjande av redan insamlade data är ju ett billigt alternativ för att förbättra en estimator och ökar inte heller uppgiftslämnarbördan och därför är det väsentligt att i möjligaste mån utnyttja hjälpinformation.

För de flesta urvalsundersökningar vid SCB är redan urvalet designat och därför kräver en förändring ganska stora insatser. Däremot är det lättare att för vissa viktiga variabler konstruera bättre estimatorer utifrån given urvalsdesign. Föreliggande rapport behandlar därför endast det momentet.

Nämnda läroböcker i stickprovsteori studerar huvudsakligen parametrar för hela populationen när man beskriver kvot- och klassiska regressionsestimatorer och har inte heller någon sammanhållen teori för estimatorer som utnyttjar hjälpinformation. Lyckligvis finns nu en bok, som inte har de begränsningarna, nämligen "Model Assisted Survey Sampling" av Carl-Erik Särndal (Universite´de Montréal, Kanada), Bengt Swensson (Högskolan i Örebro) och Jan Wretman (Stockholms universitet). Den generaliserade regressionsansats, som används här kommer från den boken[1]. Under mitt arbete har endast en mindre del av manuskriptet varit tillgängligt och jag har även unyttjat annat material och därför ligger ansvaret för eventuella tveksamheter och felaktigheter på mig.

En allmän beskrivning av den generaliserade regressionsansatsen finns i Lundström [1991]. Från denna ansats kan i stort sett alla kända estimatorer för totaler erhållas - den kvot- resp klassiska regressionsestimatorn utgör endast specialfall. (Fortsättningsvis sätts prefixet "klassisk" för att skilja det specialfallet från gruppen av regressionsestimatorer som erhålls från den generaliserade ansatsen.) Förutom den generella beskrivningen av punktestimatorerna finns också ett generellt uttryck för medelfelet, vilket ger möjlighet att konstruera generella datorprogram. I föreliggande arbete har sådana program konstruerats, som kan användas för metodstudier (och ev undervisning), men självklart inte är färdiga standardprogram.

De finns flera egenskaper hos dessa estimatorer, som inte är helt kända. T ex saknas en del kunskap om hur de fungerar för små stickprov eller hur robusta de är för felval av modell. För att öka kunskapen om detta genomförs simuleringar på ett antal konstruerade populationer samt även på några "verkliga" populationer.

---

1) Jag har också fått värdefulla synpunkter från i första hand Carl-Erik Särndal.

Genom att studera gamla data kan en lämplig modell väljas. Plottningen av observationerna på undersökningsvariabeln mot observationerna på hjälpvariabeln beskriver vad vi fortsättningsvis kallar för sambandsmönster. Utifrån den kan man bedöma om en linjär modell passar, om linjen ser ut att gå genom origo, hur variationen kring linjen kan modelleras, etc. Nu är det inte så lätt att endast visuellt bestämma modellen och därför behandlas detta område i rapporten. Speciellt gäller det bestämning av modellens variansstruktur och för den skull har ett datorprogram framställts för att studera skattningarnas egenskaper med hjälp av Monte Carlo simulering.

2. Estimatorer.

Vi har en ändlig population U = { 1,...,k,...N}, som är uppdelad i Q disjunkta redovisningsgrupper U_1, ...U_q,..., U_Q. Dessutom är populationen uppdelad i H disjunkta grupper, U_1. ,...,U_h. ,...U_H. bestämda av hjälpvariabler. Detta ger upphov till HxQ populationsceller, där storleken på cell hq är N_{hq}.

Undersökningsvariabeln antar värdet y_k för det k:te elementet; k=1,...,N. Vi önskar skatta totalen för redovisningsgrupp q

$$t_q = \sum_{U_q} y_k \qquad (1)$$

Vi drar ett urval s av storleken n från U. Den del av urvalet som hamnar i cell hq betecknar vi s_{hq} och har storleken n_{hq}. Till designen hör inklusionssannolikheten π_k (av första ordningen).

Den enklaste estimatorn, som inte utnyttjar hjälpinformation, är den s k Horvitz-Thompson-estimatorn och studeras här närmast för jämförelse med resultaten från regressionsestimatorerna. Den skrivs på följande sätt

$$\hat{t}_{qHT} = \sum_{s_q} \frac{y_k}{\pi_k} \qquad (2)$$

Om vi nu har tillgång till en hjälpvariabel x, som uppvisar hyfsat samband med undersökningsvariabeln y kan en regressionsestimator ge skattningar med mindre urvalsfel. I denna rapport studerar vi modeller där regressionslinjen går genom origo, d v s $\mathbf{x_k} = x_k$ i formeln nedan och fallet med ett intercept, vilket innebär att $\mathbf{x_k} = (1, x_k)$, d v s $\mathbf{x_k}$ nedan är en vektor med komponenterna 1 och x_k.

Samtliga regressionsestimatorer kan skrivas

$$\hat{t}_{qr} = \sum_{s} g_{qks}\, y_k / \pi_k \qquad (3)$$

där

$$g_{qks} = \delta_k \, d_{qk} + \left(\sum_{U_q} \mathbf{x_k} - \delta_q \sum_{s_q} \mathbf{x_k} / \pi_k \right)' \hat{T}^{-1} \mathbf{x_k} / \sigma_k^2, \text{ där } d_{qk} = 1 \text{ om } k \in s_q$$

och $d_{qk} = 0$ f.ö.

och

$$\hat{T} = \sum_s \frac{\mathbf{x_k} \, \mathbf{x_k}'}{\sigma_k^2 \, \pi_k}$$

Kvantiteten δ_q bestämmmer valet mellan två estimatortyper (se Lundström(1991)).

De två värden δ_q antar är 1 resp $\frac{N_q}{\hat{N}_q}$, där $\hat{N}_q = \sum_{s_q} 1 / \pi_k$. Vi väljer den senare typen i denna studie.

För variansberäkningarna behöver vi också uttrycket för residualen, nämligen

$e_k = y_k - \mathbf{x_k}' \hat{B}$ där $\hat{B} = \hat{T}^{-1} \sum_s \frac{\mathbf{x_k} y_k}{\sigma_k^2 \, \pi_k}$.

Fortsättningsvis i denna rapport antar vi att designen är obundet slumpmässigt urval, d v s $\pi_k = n / N$.

Om vi definierar $z_k = y_k$ för $k \in U_q$ och $z_k = 0$ f.ö., så kan variansen för Horvitz-Thompson-estimatorn skrivas

$$V(\hat{t}_{qHT}) = N^2 \frac{1-n/N}{n} \frac{1}{N-1} \left\{ \sum_U z_k^2 - \frac{1}{N} \left(\sum_U z_k \right)^2 \right\} \qquad (4)$$

Den vanliga skattningen av denna varians är

$$\hat{V}(\hat{t}_{qHT}) = N^2 \frac{1-n/N}{n} \frac{1}{n-1} \left\{ \sum_s z_k^2 - \frac{1}{n} \left(\sum_s z_k \right)^2 \right\} \qquad (5)$$

En lämplig variansestimator för regressionsestimatorn (3) under obundet slumpmässigt urval erhålles om z_k i (5) ersätts av $g_{qks} \, e_k$, där e_k är residualen för objekt k (se ovan).

I denna rapport studeras två av tre modelltyper som beskrivs i Lundström [1991]. I avsnitt 5 förs en diskussion om hur man väljer modell, men tills vidare antar vi att vi har möjlighet att bilda sambandsmönster, d v s plottar av punkterna (x,y), som hämtas från annat håll (gamla data, registerinformation, e d). Analysen kan leda till olika modelltyper.

Anm. I fortsättningen avses med "regressionsestimatorer" hela gruppen som beskrivs av (3) och när vi

Modelltyp 1.

Man bör välja en skattning baserad på modelltyp 1 om den bästa beskrivningen av sambandsmönstret erhålles genom separata modeller för respektive redovisningsgrupp

$$E_\xi(Y_k) = \alpha_q + \beta_q x_k;\ V_\xi(Y_k) = \sigma_q^2 x_k^v \text{ för } k \in U_q. \qquad (6)$$

I denna rapport beräknas estimatet av totalen och variansen för estimatorn med APL-programmet MOD1I ("I" för Intercept). Om vi tror att linjen går genom origo i alla redovisningsgrupper görs beräkningen med MOD1. Variansstrukturen kan bestämmas genom val av konstanten v.

Som visas i Lundström [1991] ger denna modell en kvotestimator inom varje redovisningsgrupp då $\alpha_q = 0$ och v=1 för alla q. Om $\alpha_q \neq 0$ och v=0 så erhåller vi en klassisk regressionsestimator (beskrivs t ex i Cochran [1977]) inom varje redovisningsgrupp.

Modelltyp 2.

Man bör välja en skattning baserad på modelltyp 2 om den bästa beskrivningen av sambandsmönstret erhålles genom separata modeller för hjälpgrupperna U_1. ,..U_H.

$$E_\xi(Y_k) = \alpha_h + \beta_h x_k;\ V_\xi(Y_k) = \sigma_h^2 x_k^v \text{ för } k \in U_h. \qquad (7)$$

Datorprogrammen har getts namnen MOD2I resp MOD2.

3. Simuleringsstudier.

Genomgående dras 2000 urval med designen OSU och vanligtvis låter vi stickprovsstorleken vara 200 (i avsnitt 3.2 studerar vi även alternativet med n=400). Med hjälp av dessa upprepningar skattar vi ett antal storheter.

Anm. Programmen är gjorda så att om $n_q < 2$ för något q utgår det urvalet. Det innebär att egenskaperna vi studerar är "givet $n_q \geq 2$".

Medelfelet hos en estimator $\hat{t}_q$ skattas med

$$[V_{sim}(\hat{t}_q)]^{1/2} = \left[\frac{1}{1999} \sum_{i=1}^{2000} \left(\hat{t}_{qi} - \frac{1}{2000} \sum_{i=1}^{2000} \hat{t}_{qi} \right)^2 \right]^{1/2} \qquad (8)$$

$\hat{t}_q$ är en godtycklig punktestimator och index "i" står för samplenummer.

Vi gör även en motsvarande skattning av medelfelet för kvadratroten av $\hat{V}(\hat{t}_q)$ d v s

$\{\hat{V}(\hat{t}_{qi})\}^{1/2}$ sätts i formel (8) i stället för $\hat{t}_{qi}$.

(Benämnes "medelfel för medelfelsestimatorn" i fortsättningen),

Dessutom beräknas ett konfidensintervall med avsedd konfidens på 95%:

$$\hat{t}_{qi} \pm 1.96\,[\hat{V}(\hat{t}_{qi}]^{1/2} \quad \text{för varje urval (i=1,...,2000)} \qquad (9)$$

och därefter skattas täckningsgraden med det procentuella antalet gånger intervallen innehåller parametervärdet. Täckninggradens avvikelse från konfidensgraden bestäms av flera faktorer.

Bias i medelfelsestimatorn:
Vi beräknar genomgående storleken på eventuell under- resp överskattning av medelfelet som procentuell andel av medelfelet.

Sned population:
Om y-värdena uppvisar en kraftigt sned fördelning kan det bli problem med täckningsgraden. I vissa fall beräknar vi ett snedhetsmått, nämligen

$$G = \left(\frac{1}{N} \sum_{U} (z_k - \bar{z}_U)^3 \right) / \left(\frac{1}{N} \sum_{U} (z_k - \bar{z}_U)^2 \right)^{3/2}, \quad \text{där } \bar{z}_U = \frac{1}{N} \sum_{U} z_k$$

(z_k definieras inför uttryck (3)).

För HT-estimatorn anger Cochran [1977] en tumregel för det minsta antal observationer som krävs för att konfidensintervallen ska fungera, nämligen n > 25G.

Få observationer:
Denna punkt hör ihop med föregående eftersom en sned population kombinerat med litet antal observationer ger en estimator som inte är normalfördelad. Om däremot antalet observationer är stort är estimatorn (approximativt) normalfördelad även om populationen är sned.

Stor bias i punktestimatorn:
Om estimatorn har stor bias blir täckningsgraden mindre än konfidensgraden - Cochran[1977] säger att om absolutbeloppet av biasen dividerat med medelfelet är mindre än 0.1 så fungerar konfidensintervallet bra och även om värdet är så högt som 0.2 "duger" konfidensintervallet.

Vi börjar med att studera egenskaperna hos estimatorerna på konstruerade populationer. I avsnitt 3.2 utnyttjar vi "verkliga" populationer.

3.1 Konstruerade populationer.

Samtliga populationer som studeras består av 650 objekt, fördelade på tre redovisningsgrupper (q = 1, 2, 3) och två hjälpgrupper (h =1, 2) med följande antal i de sex cellerna:

	q			
	1	2	3	Σ
h				
1	300	125	25	450
2	100	75	25	200
Σ	400	200	50	650

Värdena på hjälpvariabeln x utgör ett OSU av 650 från heltalen 1,...,1000, d v s x är likformigt fördelad. Samma urval av x-värden används för samtliga populationer. Undersökningsvariabeln y kan bestämmas efter följande tre huvudprinciper, där variabeln ξ är en likformigt fördelad variabel på intervallet [-0.5 , 0.5] och konstanten c kan väljas fritt:

1) $y_k = \alpha_q + \beta_q x_k + c\,\xi\, x_k^{w/2}$ för $k \in U_q$

2) $y_k = \alpha_h + \beta_h x_k + c\,\xi\, x_k^{w/2}$ för $k \in U_h$.

För varje huvudprincip kan parametrarna α, β och w väljas fritt. Variansen för $c\xi$ fixeras till $0.083c^2$ där c^2 bestäms så att (produktmoment)-korrelationskoefficienten är ungefär lika (omkring 0,85) i varje population. Det följer att variansstrukturen är densamma för alla redovisningsgrupper (det vore enkelt att också framställa populationer med varierande variansstrukturer i olika redovisningsgrupper, men analysen av resultaten blir mer komplicerad), nämligen

$$V(y_k) = 0.083c^2x_k^w$$

De APL-program som framställer önskad population har namnen P1 resp P2 - ett program för respektive huvudprincip.

POPULATION 1.

Populationen framställs med pgm P1 där $\alpha_q = 0$ och $\beta_q = 1$ för q = 1, 2, 3 och dessutom sätter vi w=1 och c=30.

Totalerna för redovisningsgrupperna är 206162, 96473 och 20992. För hela populationen är (produktmoment)-korrelationskoefficienten r = 0.83 (ung. lika inom varje redovisningsgrupp).

Stickprovsstorleken är 200 vilket leder till att det förväntade antalet observationer i resp redovisningsgrupp är 123, 62 resp 15.

I tabell 1 studerar vi estimatoralternativen utan intercept för olika v-värden.

Tabell 1

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	14556	11242	5434
	94	94	93
MOD1, v=0	5939	4173	2048
	94	94	92
MOD1, v=0.5	5939	4170	2043
	94	94	92
MOD1, v=1	5938	4168	2039
	94	94	92
MOD1, v=1.5	5940	4166	2038
	94	94	92

Medelfelen för regressionsestimatorerna är betydligt mindre än medelfelen för HT-estimatorn för samtliga redovisningsgrupper. Regressionsestimatorna visar dock inbördes upp nästan identiska medelfel, d v s valet av v-värde är betydelselöst för denna population.

Samtliga v-värden ger också samma täckningsgrad.

Medelfelet för medelfelsestimatorn är för HT-estimatorn 582, 803 och 884 för respektive redovisningsgrupp och motsvarande värde för kvotestimatorn är 298, 318 och 327. Övriga regressionsestimatorer uppvisar i stort sett samma bild som kvotestimatorn.

Huvudskälet till att täckningsgraden är för låg för den minsta redovisningsgruppen är att medelfelen underskattas särskilt mycket för den (se tabell 2).

Tabell 2

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	-2.6	-0.4	-0.4
MOD1, v=0	-2.6	-1.1	-7.4
MOD1, v=0.5	-2.5	-1.0	-7.0
MOD1, v=1	-2.5	-0.9	-6.3
MOD1, v=1.5	-2.5	-0.8	-6.3

Syftet med nästa simulering är att studera hur en estimator baserad på en modell med intercept fungerar för dessa redovisningsgrupper. Vi tittar därför på den klassiska regressionsestimatorn (MOD1I, v=0) samt MOD1I, v=1. Dessutom tar vi återigen med fallen MOD1, v=0 samt den klassiska kvotestimatorn (MOD1, v=1). Vi har två skäl till detta. För det första vill vi för samma urval jämföra estimatorer med resp utan intercept och för det andra studera hur stabil simuleringen är.

Tabell 3

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	14211	11187	5517
	95	94	93
MOD1, v=0	5802	4208	2082
	95	94	90
MOD1, v=1	5799	4206	2077
	95	94	91
MOD1I, v=0	5819	4231	2131
	95	93	89
MOD1I, v=1	5813	4208	2106
	94	94	90

Vi ser att alternativen med intercept ger något större medelfel, men skillnaden är ytterst liten.

Även täckningsgraden är i stort sett densamma för de båda typerna av estimatorer - kanske kan man ana en något sämre täckningsgrad för fallen med intercept för den minsta redovisningsgruppen. Bilden förklaras delvis av följande skattningar:

Tabell 4

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	0.0	-0.2	-2.1
MOD1, v=0	-0.6	-2.1	-9.6
MOD1, v=1	-0.4	-1.9	-8.7
MOD1I, v=0	-0.8	-2.6	-12.0
MOD1I, v=1	-0.7	-2.0	-10.0

Vi ser alltså en något större underskattnning av medelfelet för estimatorerna med intercept.

Om vi jämför tabell 1 och tabell 3 får vi en uppfattning om stabiliteten i simuleringsskattningarna av medelfelen för estimatorer utan intercept. Skillnaderna är mindre än 2.5 % och får sägas vara små. Stabiliteten i skattningen av täckningsgraden är sämre, eftersom man kan observera en skillnad på två procentenheter (MOD1, v=0, minsta redovisningsgruppen).

Jämför vi tabell 2 och 4 får vi en uppfattning om stabiliteten av skattningen av medelfelet för medelfelsestimatorn. Även där kan man se relativt stora skillnader mellan de två simuleringsstudierna.

En viss instabilitet i simuleringsskattningarna kan alltså konstateras (ändå är det sällan man i litteraturen ser så många upprepningar som 2000).

Låt oss även skatta medelfelet för medelfelsestimatorn.

Tabell 5

Medelfel för olika medelfelsestimatorer.

Medelfels-estimator för följande estimatorer	Redovisningsgrupp 1	2	3
HT-est	556	778	887
MOD1, v=0	298	329	319
MOD1, v=1	297	330	325
MOD1I, v=0	303	332	332
MOD1I, v=1	297	329	329

Medelfelsestimatorn för HT-estimatorn uppvisar genomgående större medelfel än regressionsestimatorerna. För HT-estimatorn ökar medelfelet för medelfelsestimatorn

med minskande redovisningsgrupper, men för regressionsestimatorerna är motsvarande storhet relativt lika för olika storlekar av redovisningsgrupper.

Populationen är konstruerad så att varje redovisningsgrupp uppvisar samma struktur, vilket borde innebära att vi får bättre resultat med alternativet där regressionskoefficienter skattas för hela populationen. Vi prövar därför MOD2 -alternativet, där vi slår samman de två hjälpgrupperna.

Tabell 6

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

	Redovisningsgrupp		
Estimatorer	1	2	3
HT-est	13670	11184	5333
	96	95	94
MOD2, v=0	5945	4222	2015
	94	94	95
MOD2, v=0.5	5944	4222	2015
	94	94	95
MOD2, v=1	5943	4222	2015
	94	94	95
MOD2, v=1.5	5944	4223	2017
	94	94	95

Skillnaderna mellan tabell 1 och 3 vad avser medelfelen är ytterst små och kan förklaras av slumpen. Däremot är täckningsgraden bättre med MOD2-alternativet än med MOD1-alternativet för den minsta redovisningsgruppen. Här uppvisar medelfelsestimatorn en svag överskattning (1.7-1.9) jämfört med en underskattning -6.3 - -7.4 i det förra fallet.

Vi har även genomfört simulering med intercept-estimatorerna (i övrigt på samma sätt som i föregående simulering) och funnit att dessa fungerar i stort sett lika bra som estimatorerna utan intercept.

POPULATION 2.

Populationen framställs med pgm P1 där $\alpha_q = 200$ samt $\beta_q = 1$ för $q = 1, 2, 3$ och dessutom sätter vi w=1 och c=30. (Interceptet är ca 30% av medelvärdet av y-variabeln inom resp redovisningsgrupp.)

Totalerna som skattas är 276076, 140274 och 31415 i resp redovisningsgrupp. För hela populationen är (produktmoment)-korrelationskoefficienten r = 0.82.

Vi genomför först en simulering med de estimatorer vars resultat vi ser i tabell 1, d v s vi vill studera det fall då vi felaktigt trodde att "populationen" inte hade något intercept. (Vid studium av spridningsdiagrammet är det lätt att negligera interceptet).

Tabell 7

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	16187	14523	7295
	95	94	93
MOD1, v=0	6531	4564	2204
	95	94	92
MOD1, v=0.5	6677	4658	2236
	95	94	92
MOD1, v=1	6991	4878	2371
	95	94	93
MOD1, v=1.5	7903	5974	3621
	95	96	97

Vi ser att medelfelen för regressionsestimatorerna ökar med v-värdena. T ex ger den klassiska kvotestimatorn ca 7 % större medelfel än estimatorn MOD1, v=0 och detta trots att w=1. Alternativet MOD1, v=1.5 ger 21, 31 resp 64 % större medelfel än alternativet MOD1, v=0.

Täckningsgraden är bra för samtliga estimatorer för de två största redovisningsgrupperna. För den minsta redovisningsgruppen är täckningsgraden för liten då v<1.5 men för stor då v=1.5. Detta beror i första hand på medelfelsestimatorn (se tabellen nedan).

Tabell 8

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	0.4	-0.9	-0.4
MOD1, v=0	0.3	-1.2	-4.7
MOD1, v=0.5	-0.4	-2.2	-5.4
MOD1, v=1	-0.9	-2.7	-5.9
MOD1, v=1.5	1.0	7.3	10.1

Även medelfelet för medelfelsestimatorn ökar med v, men med en dramatisk ökning från v=1 till v=1.5 (429, 459, 545 för resp redovisningsgrupp till 719, 1295, 2212).

Låt oss därefter studera hur estimatorer med intercept fungerar då vi i "populationen" har intercept, d v s vi väljer "rätt" estimatortyp.

Tabell 9

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	16311	14471	7258
	95	94	92
MOD1I, v=0	5941	4121	2155
	94	94	89
MOD1I, v=0.5	5929	4101	2125
	95	94	90
MOD1I, v=1	5920	4092	2116
	95	94	90
MOD1I, v=1.5	5918	4091	2123
	95	94	91

Medelfelen är mindre för estimatorerna med intercept än de utan intercept. Simuleringarna visar att kvotestimatorn (MOD1, v=1) ger för respektive redovisningsgruppp 15, 15 och 9 % större medelfel än den klassiska regressionsestimatorn (MOD1I, v=0).

Val av v-värde spelar uppenbart liten roll.

Täckningsgraden är bra för samtliga estimatorer för de två största redovisningsgrupperna. För den minsta redovisningsgruppen är täckningsgraden genomgående för liten. Detta beror i första hand på medelfelsestimatorn (se tabellen nedan).

Tabell 10

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	-0.4	-0.6	0.1
MOD1I, v=0	-1.8	-3.2	-12.2
MOD1I, v=0.5	-1.6	-2.8	-10.9
MOD1I, v=1	-1.3	-2.4	-9.6
MOD1I, v=1.5	-1.0	-2.1	-7.7

Tabellen visar att underskattningen av medelfelet ökar med minskande redovisningsgrupp. Underskattningen minskar också med ökande v-värde.

Populationen är sådan att varje redovisningsgrupp har samma "linje" och därför kan man tro att MOD2I-alternativet med hopslagna h-grupper är bättre än MOD1I-alternativet. Vi genomför en simulering där MOD2I, v=0 resp v=1 samt MOD2, v=0 resp v=1.

Tabell 11

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	15898	14122	7384
	95	95	93
MOD2, v=0	6737	4168	2266
	94	97	94
MOD2, v=1	7231	4413	2422
	94	96	94
MOD2I, v=0	5936	3973	2071
	95	96	94
MOD2I, v=1	5944	3967	2069
	95	96	94

Ingen genomgående minskning av medelfelet kan upptäckas vid övergång från MOD1-alternativen till MOD2-alternativen. Däremot erhåller man en bättre täckningsgrad för den minsta redovisningsgruppen. För den mellanstora redovisningsgruppen kan vi observera ett något oväntat resultat - täckningsgraden överskattas något liksom medelfelet.

Tabell 12
Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	2.2	2.2	-1.0
MOD2, v=0	-1.6	8.1	-4.5
MOD2, v=1	-2.4	3.4	-4.1
MOD2I, v=0	-1.5	1.7	-4.1
MOD2I, v=1	-1.4	1.8	-4.0

POPULATION 3.

Populationen framställs med pgm P1 där $\alpha_1 = \alpha_2 = 200$, $\alpha_3 = 400$ samt $\beta_1 = 2$, $\beta_2 = \beta_3 = 1$ och dessutom sätter vi w=1 och c=30. För redovisningsgrupp 2 har vi samma struktur som för population 2, men för redovisningsgrupp 1 har vi $\beta_1 = 2$ och för redovisningsgrupp 3 har vi $\alpha_3 = 400$.

Totalerna som skattas är 485101, 134808 och 43811 i resp redovisningsgrupp. För hela populationen är (produktmoment)-korrelationskoefficienten r = 0.82. För respektive redovisningsgrupp har vi värdena 0.95, 0.83 och 0.82.

Vi studerar nu endast två v-värden för respektive estimatortyp.

Tabell 13

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	28977	13889	9614
	95	95	94
MOD1, v=0	6246	4554	3232
	95	95	92
MOD1, v=1	6597	4765	3864
	95	95	92
MOD1I, v=0	5776	4308	2388
	95	94	89
MOD1I, v=1	5769	4306	2323
	95	94	89

Regressionestimatorna med intercept ger som väntat de minsta medelfelen och bästa resultatet erhålles i redovisningsgrupp 3 där vi har det största "populations"-interceptet. Däremot är täckningsgraden för interceptalternativen inte mer än 89% för den minsta redovisningsgruppen. Om vi däremot valt "fel" modell, t ex kvotestimatorn, erhåller vi bättre täckningsgrad. Detta förklaras delvis av tabell 14.

Om vi jämför de procentuella relativa medelfelen (100 x medelfel / t_q vilken vi kallar rel-M) för estimatorerna som återfinns i tabell 13 med motsvarande estimatorer i tabell 7 ser vi bl a följande:

- För redovisningsgrupp 1 minskar rel-M med ca 1 procentenhet för samtliga studerade regressionsestimatorer. I den gruppen har "populationslinjens" lutning ökats från 1 till 2 medan interceptet är detsamma.

- För redovisningsgrupp 2 har vi i stort sett oförändrat värde på rel-M. För den gruppen har inga förändringar av "poulationslinjen" gjorts.

- För redovisningsgrupp 3 ökar rel-M med 0.4 resp 1.3 procentenheter för MOD1, v=0 resp v=1 och minskar med ca 1.4 procentenheter för interceptalternativen. "Populationslinjen" har samma lutning men interceptet har ökats från 200 till 400 för den redovisningsgruppen.

Tabell 14

Procentuell under- resp överskattning av medelfel.

	Redovisningsgrupp		
Estimatorer	1	2	3
HT-est	0.7	1.2	0.6
MOD1, v=0	2.3	-0.3	-6.6
MOD1, v=1	1.2	-1.0	-11.6
MOD1I, v=0	1.4	-3.4	-14.0
MOD1I, v=1	1.5	-3.2	-12.3

För samma population tittar vi även på estimatorerna för modelltyp 2, vilket får ses som exempel på ett fall där man misslyckats med val av modell.

Tabell 15

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	29087	14260	9789
	94	95	94
MOD2, v=0	6104	7192	3766
	98	98	93
MOD2, v=1	6073	7659	4004
	97	98	94
MOD2I, v=0	6719	5888	3125
	97	99	93
MOD2I, v=1	7040	5558	2974
	97	99	93

Fortfarande ger regressionestimatorerna betydligt mindre medelfel än HT-estimatorn, men när vi jämför de olika regressionsestimatorerna möts vi av en splittrad bild. Ingen av estimatorerna är bäst. Om man i modellarbetet hade förstått att "poulationen" har intercept skulle man fått betydligt mindre medelfel för redovisningsgrupp 2 och 3 men större medelfel för redovisningsgrupp 1. Täckningsgraden är dock för hög i den första redovisningsgruppen och särskilt hög i den andra. I dessa två redovisningsgrupper har vi kraftiga överskattningar av medelfelen (se nedanstående tabell).

Tabell 16
Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	0.2	-1.5	-1.8
MOD2, v=0	19.5	16.6	-3.1
MOD2, v=1	12.0	20.8	-1.1
MOD2I, v=0	11.1	35.4	-1.2
MOD2I, v=1	10.1	40.6	-0.7

För att demonstrera hur MOD2 resp MOD2I fungerar, när populationen framställs under huvudprincip 2 bildar vi population 4.

POPULATION 4.

Populationen framställs med pgm P2 där $\alpha_1 = 200$, $\alpha_2 = 400$, $\beta_1 = 2$ samt $\beta_2 = 1$ och dessutom sätter vi w=1 och c=30.

Totalerna som skattas är 455170, 214608 och 50290 i resp redovisningsgrupp. För hela populationen är (produktmoment)-korrelationskoefficienten r = 0.87.

Låt oss först titta på estimatorer som är baserade på modelltyp 2, d v s vi väljer de "rätta" estimatorerna.

Tabell 17

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

	Redovisningsgrupp		
Estimatorer	1	2	3
HT-est	26524	22478	11656
	95	95	93
MOD2, v=0	7161	4824	2626
	95	96	96
MOD2, v=1	7738	5401	2971
	95	94	94
MOD2I, v=0	6276	4082	2081
	95	95	94
MOD2I, v=1	6275	4065	2077
	95	95	94

Som väntat uppvisar interceptalternativen de minsta medelfelen. Av alternativen utan intercept har estimatorn med v=0 minsta medelfelen. Täckningsgraden är genomgående bra, vilket förklaras av att biasen för medelfelsestimatorn är liten.

Låt oss därefter studera estimatorer som härleds från modelltyp 1, d v s "felaktig" modell.

Tabell 18

Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	26136	22619	11403
	95	94	93
MOD1, v=0	8574	6547	3404
	96	94	93
MOD1, v=1	8937	6890	3802
	96	94	93
MOD1I, v=0	8110	6149	3053
	95	93	90
MOD1I, v=1	8128	6154	3004
	94	94	91

Felvalet av modelltyp leder till relativt kraftiga ökningar av medelfelet. Kvotestimatorn inom varje redovisningsgrupp ger 42, 69 resp 83 % större medelfel än den optimala estimatorn (MOD2I, v=1). Om vi jämför den klassiska regressionsestimatorn med den optimala ser vi för respektive redovisningsgrupp 29, 51 resp 47% större medelfel. Dessutom ger estimatorerna i tabell 18 för låga täckningsgrader för den minsta redovisningsgruppen. Detta förklaras av nedanstående tabell.

Tabell 19

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	3.4	-0.7	-2.5
MOD1, v=0	2.3	-0.3	-6.6
MOD1, v=1	1.2	-1.0	-11.6
MOD1I, v=0	1.4	-3.4	-14.0
MOD1I, v=1	1.5	-3.2	-12.3

3.2 Kyrkokommuner.

F/SE har genomfört en undersökning med benämningen "Kyrkokommunernas bokslut" (KYBOK). Den består av en ekonomisk redogörele för år 1989, där variablerna mäter utgifter och inkomster för olika delar av den kyrkokommunala verksamheten. Vi har fått utnyttja den del av undersökningen som utförts som totalundersökning och från den

plockat ut ett antal variabler som antingen får tjäna som undersökningsvariabel eller som hjälpvariabel. I en verklig undersökningssituation skulle det (antagligen) vara möjligt att från annat håll erhålla uppgift om den totala församlingsskatten och även den totala lönen för präster och kyrkomusiker och därför låter vi dessa vara hjälpvariabler. Den undersökningsvariabel vi studerar mest är "Kostnad för central förvaltning".

Som redovisningsgrupper har vi valt "Samfälligheter", "Pastorat" och "Församlingar" (motsvarar posttyper i undersökningen). Antalet objekt i respektive redovisningsgrupp är 60, 830 och 200.

Vi har ingen H-gruppsindelning och därför är H=1 när MOD2-alternativen används.

POPULATION A.

Den första studien analyserar hur olika estimatorer fungerar när vi ska skatta totala kostnaden för central förvaltning (y) och utnyttjar uppgiften om församlingsskatt (x). I den första redovisningsgruppen har vi fyra objekt som har mycket höga x-värden och därmed också höga y-värden. I en rimlig design skulle dessa väljas ut med sannolikheten 1 och för att få en realistisk bedömning av medelfelen plockar vi bort dessa från populationen. Efter denna reduktion har vi 56, 830, resp 200 kyrkokommuner i respektive redovisningsgrupp.

Stickprovsstorleken är 200 och därmed är förväntade antalet observationer 10, 153, och 37 i resp redovisningsgrupp.

Parametervärdena är 230694, 554597 och 6795 för respektive redovisningsgrupp. Standardavvikelsen för undersökningsvariabeln är 3510, 821 och 58.

För hela populationen är (produktmoment)-korrelationskoefficienten r = 0.95 och för redovisningsgrupperna 0.93, 0.92 och 0.75. De skattade (med programmet VSKATT- se avsnitt 5) v-värdena i respektive redovisningsgrupp är 1.4, 1.3 och 1.6. Spridningsdiagram för redovisningsgrupperna finns i bil. 1.

Tabell 20

Undersökningsvariabel: Kostnad för central förvaltning
Hjälpvariabel: Församlingsskatt.
n=200.
Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	83637	51443	1919
	90	94	88
MOD1, v=0	23752	19845	1127
	82	93	88
MOD1, v=1	23304	19845	1090
	83	94	90
MOD1I, v=0	25367	20098	1183
	80	93	86
MOD1I, v=1	24124	20263	1117
	82	93	89

Samtliga regressionsestimatorer har betydligt mindre medelfel än HT-estimatorn för de två första redovisningsgrupperna. För den tredje redovisningsgruppen, som har liten standardavvikelse och även liten korrelationskoefficient, är skillnaden liten.

Diagrammen i bil. 2 antyder att regressionslinjerna går genom origo och därför passar estimatorn beräknad med pgmet MOD1. För redovisningsgrupp 1 kan vi se att det ligger en vinst i val av detta alternativ. Speciellt om vi jämför den klassiska regressionsestimatorn med kvotestimatorn ser vi att den senare ger 8.1% mindre medelfel för denna redovisningsgrupp. Det innebär att "large-sample"- egenskaperna som säger att regressionsestimatorn aldrig har större medelfel än kvotestimatorn inte gäller här ($E(n_1)$ = 10). Kvotestimatorn har också bättre täckningsgrad än den klassiska regressionsestimatorn. Om vi däremot väljer interceptvarianten med v=1 är skillnaden mot kvotestimatorn liten både avseende medelfel och täckningsgrad.

Täckningsgraden är låg för flera estimatorer och redovisningsgrupper. För HT-estimatorn är förklaringen snedheten i kombination med litet stickprov. Det viktigaste skälet till den låga täckningsgraden för regressionsestimatorerna är underskattningen av medelfelen, vilken framgår av tabell 21.

Vi har även genomfört simuleringar på ursprungspopulationen, d v s populationen före exkluderingen av objekten med de fyra största x-värdena, och därmed funnit att för redovisningsgrupp 1 har HT-estmatorn ett medelfel på ca 142000 och regressionsestimatorerna ett medelfel på ca 34000, d v s att de senare klarar av outliers mycket bättre än den förra.

Dessutom har vi prövat MOD1-alternativet med v=0.8, 1.3, 1.8 resp 2.3 och funnit att skillnaden mellan de tre första estimatorerna är små vad gäller medelfelen. För v=2.3 får vi ökade medelfel (mest för den första redovisningsgruppen med 11.4%).

Täckningsgraden förbättras däremot med ökande v-värde och speciellt bra värden erhålles för v=2.3. Som vi har sett tidigare ger en överskattning av det verkliga v-värdet en justering för benägenheten att underskatta medelfelen i små redovisningsgrupper och därmed kan täckningsgraden förbättras.

Tabell 21

Procentuell under- resp överskattning av medelfel. (n=200).

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	-2.5	2.1	-3.3
MOD1, v=0	-25.0	-4.8	-13.5
MOD1, v=1	-21.5	-2.9	-8.1
MOD1I, v=0	-30.5	-6.3	-17.9
MOD1I, v=1	-24.7	-3.9	-11.5

Underskattningen av medefelen är mycket stora för regressionsestimatorerna för de små redovisningsgrupperna och speciellt stora tal uppvisar den klassiska regressionsestimatorn.

Medelfelet för medelfelsestimatorn är för HT-estimatorn 20169, 9602 och 549 för respektive redovisningsgrupp och motsvarande värden för kvotestimatorn (MOD1, v=1) är 6081, 2933 och 279. Övriga regressionsestimatorer uppvisar ungefär samma resultat - de skillnader som finns följer i stort sett skillnaden i medelfel för punktestimatorerna. Det innebär att även medelfelsskattningar är mycket stabilare för regressionsestimatorerna än för HT-estimatorn.

Det är intressant att se hur mycket en ökad stickprovsstorlek reducerar bristerna i estimatorerna ovan och därför genomför vi en simulering med n=400. Det innebär att vi förväntar oss följande antal observationer i resp redovisningsgrupp: 20, 306 och 74.

Tabell 22

Undersökningsvariabel: Kostnad för central förvaltning
Hjälpvariabel: Församlingsskatt.
n=400.
Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	50903	33027	1224
	93	95	91
MOD1, v=0	13301	12064	735
	90	94	91
MOD1, v=1	13206	12109	715
	91	94	92
MOD1I, v=0	13675	12129	754
	90	94	90
MOD1I, v=1	13323	12387	726
	91	94	91

Jämför man tabell 20 och tabell 22 ser man att alla värden har förbättrats. Fortfarande har vi låg täckningsgrad för vissa estimatorer och redovisningsgrupper, t ex för den klassiska regressionsestimatorn och redovisningsgrupp 3 har vi endast 90% av de 95%-iga konfidensintervallen som innesluter parametervärdet detta trots att vi förväntar oss 74 observationer. Underskattningen av medelfelet visar sig fortfarande vara så stort som 11.6% (se nedan).

Tabell 23

Procentuell under- resp överskattning av medelfel. (n=400).

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	1.1	0.2	-3.0
MOD1, v=0	-8.1	-0.9	-9.4
MOD1, v=1	-6.5	-0.1	-5.8
MOD1I, v=0	-11.0	-1.6	-11.6
MOD1I, v=1	-7.1	-0.8	-7.5

Antag att vår tillgängliga information visar att "populationslinjerna" inte skiljer sig åt inom varje redovisningsgrupp (detta visar kanske också diagrammen i bil 2?). Eftersom vi här inte har någon H-uppdelning är det rimligt att välja MOD2-alternativen med H=1. Simuleringen ger följande resultat.

Tabell 24

Undersökningsvariabel: Kostnad för central förvaltning
Hjälpvariabel: Församlingsskatt.
n=200.
Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	83902	53102	1905
	91	93	89
MOD2, v=0	20999	19358	2003
	88	95	94
MOD2, v=1	20662	19355	1962
	92	95	93
MOD2I, v=0	21397	19416	2025
	87	95	95
MOD2I, v=1	20736	19317	1983
	92	95	92

Vid övergången från modelltyp 1 till modelltyp 2 förändras medelfelet endast obetydligt för den andra redovisningsgruppen, men för den första får vi en påtaglig minskning av medelfelen och för tredje redovisningsgruppen en ökning. Beträffande täckningsgraden får vi en betydande förbättring för de mindre redovisningsgrupperna. Sammanfattningsvis är den senare estimatortypen att föredra.

Medelfelsestimatorn för denna estimatortyp fungerar också bättre än för de estimatorer som kommer från modelltyp 1 (jmf tabellen nedan med tabell 21).

Tabell 25

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	-3.1	-1.1	-3.6
MOD2, v=0	-12.7	-0.1	9.9
MOD2, v=1	-6.8	-0.6	9.6
MOD2I, v=0	-14.6	-0.1	16.5
MOD2I, v=1	-7.3	-0.7	5.6

POPULATION B.

I denna population är undersökningsvariabeln "Kostnad för församlingsfastigheter" (y) och hjälpvariabel är "Församlingsskatt" (x). Liksom för population A tar vi bort objekten med de fyra största x-värdena. Parametervärdena är 537209, 1529744 och 56093 för resp redovisningsgrupp. Korrelationen är 0.91 för hela populationen och 0.89, 0.83 och 0.85 för respektive redovisningsgrupp.Skattningen av det "verkliga" värdet av v i respektive redovisningsgrupp (se ansnitt 5) antar för respektive redovisningsgrupp värdena 1.5, 1.6 och 1.2.

I denna simulering jämför vi MOD1- resp MOD2-alternativen med v=0 och v=1.

Tabell 26

Undersökningsvariabel: Kostnad för församlingsfastigheter.
Hjälpvariabel: Församlingsskatt.
n=200.
Rad 1 : Medelfel
" 2 : Täckningsgrad vid 95%-igt konfidensintervall.

	Redovisningsgrupp		
Estimatorer	1	2	3
HT-est	192840	144710	13338
	90	93	92
MOD1, v=0	62146	74896	5930
	84	93	92
MOD1, v=1	62479	74906	5991
	84	93	93
MOD2, v=0	60693	74505	6552
	87	93	94
MOD2, v=1	67411	74290	6228
	89	93	94

De två estimatortyperna ger beträffande medelfelen en komplex bild, där ingen bästa estimator kan urskiljas. Om man i jämförelsen även tar med täckningsgraden verkar någon av MOD2-alternativen vara att fördra.

I nedanstående tabell visas hur medelfelsestimatorn fungerar för respektive estimator.

Tabell 27

Procentuell under- resp överskattning av medelfel.

Estimatorer	Redovisningsgrupp 1	2	3
HT-est	-3.4	-2.5	0.0
MOD1, v=0	-22.0	-5.4	-10.0
MOD1, v=1	-19.1	-3.5	-6.4
MOD2, v=0	-16.0	-2.3	6.0
MOD2, v=1	-6.3	-2.9	3.4

Medelfelsestimatorn för den sista punktestimatorn fungerar bäst av studerade regressionestimatorer.

3.3 Slutsatser.

I föregående avsnitt har vi testat en mängd varianter av populationer och regressionsestimatorer. Resultaten baserar sig på 2000 upprepningar, vilket ger goda skattningar. Vid analysen bör man dock inte glömma att smärre avvikelser kan förklaras av slumpen. Vissa tabeller innehåller estimatorer, som också finns med i andra simuleringar och därmed har man möjlighet att bedöma den ungefärliga slumpvariationen.

Följande slutsatser kan dras av genomförda simuleringar:

Skattningar baserade på modelltyp 1

1) Regressionsestimatorerna ger betydligt mindre medelfel än HT-estimatorn då korrelationen mellan y- och x-variabeln är hög. Detta är ju inget nytt - t ex visar Cochran[1977] att vid skattning av medelvärdet i en population är variansen för den klassiska regressionsestimatorn $100(1-r^2)\%$ mindre än variansen för HT-estimatorn (detta är ett "large-sample"-resultat, som inte alltid stämmer i redovisade simuleringar där vi har små urval), i fallet obundet slumpmässigt urval.

2) Interceptalternativen ger mindre medelfel än estimatorerna utan intercept om "populationen" uppvisar ett intercept och endast obetydligt större medelfel (för små redovisningsgrupper) om "populationen" saknar intercept.

3) Regressionestimatorerna underskattar medelfelet för små redovisningsgrupper (om inte v-värdet har "överskattats" - se nästa punkt), vilket också leder till dålig täckningsgrad. Interceptalternativen uppvisar de största problemen. T ex visar tabell 20 och 21 att medelfelet för den klassiska regressionestimatorn i den sista redovisningsgruppen ($E[n_3] = 37$) underskattas med 17.9 % och motsvarande värde för kvotestimatorn är 8.1%.

4) De redovisade konstruerade populationerna är alla framställda med w=1 (vi har även prövat andra värden). Därför säger vi fortsättningsvis att v-värdet "överskattas" då v>1 . Om man valt rätt modelltyp spelar valet bland underskattade v-värden ingen nämnvärd roll för medelfelet och täckningsgraden. Även en viss överskattning av v ger ingen effekt på medelfelet, men däremot kan medelfelet för medelfelsestimatorn öka kraftigt. Om man

däremot väljer ett estimatoralternativ utan intercept där "populationen" har intercept ökar medelfelen med v och underskattningen av medelfelet minskar och övergår till överskattningar med ökande v. Denna effekt är långsammare för det fall "populationen" saknar intercept och estimator med intercept utnyttjas.

Skattningar baserade på modelltyp 2.

5) Om populationens sambandsmönster uppvisar små eller inga skillnader mellan redovisningsgrupper men däremot skillnader mellan H-grupper då är estimatorer från modelltyp 2 att föredra. Som väntat ger de senare mindre medelfel, men även medelfelsestimatorn fungerar mycket bra (se tabell 17 och 18) och därmed blir täckningsgraden bra. Även i de fall inga skillnader mellan H-grupper råder ger denna modelltyp lika små medelfel, men ger bättre medelfelsestimat än estimatorer baserade på den första modelltypen. För POPULATION A är sambandssmönstret för respektive redovisningsgrupp relativt lika (diffusa skillnader som det är i "verkligheten") och MOD2-alternativen uppvisar också i det stora hela de bästa egenskaperna.

6) Om man däremot har en markant skillnad i sambandsmönster mellan redovisningsgrupper (se POPULATION 3) och man väljer något MOD2-alternativ med hopslagna H-grupper då kan medelfelet bli betydligt större än för den "rätta" estimatorn (jmf tabell 13 och 15). MOD2-alternativen överskattar medelfelen för de större redovisningsgrupperna, vilket leder till för hög täckningsgrad (MOD2I, v=1 i tabell 16 visar en överskattning av medelfelet för redovisningsgrupp 2 med 40.6% !).

Vi har också skattat biasen för regressionsestimatorerna (inte redovisat tidigare) och det har genomgående visat sig att den är liten och påverkar täckningsgraden endast obetydligt. Endast i något fall har Cochran's gräns på 0.2 (se sid. 5) passerats.

4. Orsak till att regressionsestimatorn ibland urartar.

I vissa beräkningar (ej redovisade) har det visat sig att regressionsestimatorerna ibland urartar, d v s såväl punktestimat som medelfelsestimat antingen inte går att beräkna eller uppvisar fullständigt oacceptabla värden. Det visar sig inträffa då något x-värde är mycket litet och eventuellt kombinerat med stort v-värde. I detta avsnitt ska vi utveckla formlerna för att teoretiskt visa var problemet uppkommer.

Om vi väljer fallet utan intercept kan g_{qks} (se (3)) skrivas (fortfarande antar vi att designen är OSU och därmed är $\pi_k = n/N$).

$$g_{qks} = \delta_q d_{qk} + \left(\sum_{U_q} x_k - \delta_q \frac{N}{n} \sum_{s_q} x_k \right) C_{qk} \text{ , där } C_{qk} = \frac{n\, x_k^{1-v}}{N \sum_{s} x_k^{2-v}} k$$

(k är en konstant som beror på modelltyp, t ex = σ_q)

Det innebär att för $v \leq 1$ så kan även x anta värdet 0 utan att g-värdet blir stort. Om däremot $v > 1$ och x_k är litet för något k blir C_{qk} stort och därmed också g-värdet. För $x_k = 0$ kan inte C_{qk} beräknas.

Om vi väljer estimatorerna med intercept kan man visa att g_{qks} kan skrivas på följande sätt:

$$g_{qks} = \delta_k d_{qk} + \frac{n}{N(s_0 s_2 - s_1^2)} \left(\sum_{U_q} x_k - \delta_q \frac{N}{n} \sum_{s_q} x_k \right) S_k, \text{ där}$$

$$S_k = (x_k^{1-v} s_0 - x_k^{-v} s_1) \text{ och } s_i = \sum_s x_k^{i-v} .$$

Om v=0 erhåller vi

$$s_i = \sum_s x_k^i \text{ och } S_k = x_k s_0 - s_1.$$

Vi ser att s_i inte får något extremt värde för små x-värden och inte heller S_k och därför har vi inga problem för v=0. Det innebär att den konventionella regressionsestimatorn urartar inte p g a små x-värden.

Om vi sätter v=1 får vi

$$s_i = \sum_s x_k^{i-1} \text{ och } S_k = s_0 - \frac{s_1}{x_k} .$$

Det innebär att om $x_k = 0$ för något k kan inte s_0 beräknas. Om x_k är litet urartar såväl s_0 som S_k.

Även för v > 1 urartar regressionsestimatorn med intercept om något x-värde är litet.

5. Hur väljer man modell ?

De estimatorer som behandlas i denna rapport är alla (asymptotiskt) unbiased. Ett felaktigt val av modell förändrar inte den egenskapen, men som vi sett i simuleringarna, ökar medelfelet för estimatorn. Det är därför viktigt att få en så bra modell som möjligt och det kräver bra information. Till stöd för val av modell har man bl a de resultat som kommit fram i redovisade simuleringar.

Det man i första hand bör söka är gamla data som beskriver sambanden mellan y- och x-variabeln. Om undersökningen genomförs löpande finns goda möjligheter att succesivt bygga upp en bra modell och därmed en effektiv estimator.

SCBs undersökningar har ofta stratifierade urval och varierande urvalssannolikheter, vilket leder till att ett sambandsdiagram för hela populationen (utan vägning av observationerna) kan bli missvisande. Om man tror att sambandsmönstren varierar kraftigt mellan strata, så bör man hantera varje stratum för sig och t ex välja MOD2 / MOD2I -varianten, där hjälpgrupperna motsvarar strata. Om man har sett att sambandsmönster i vissa strata liknar varandra kan dessa strata slås samman till en gemensam hjälpgrupp och sedan ger det gemensamma sambandsmönstret information om vilken modell som är lämplig.

Väl man funnit modelltyp (t ex valt bland de två typerna (6) - (7)) är det fråga om att bestämma alternativ med eller utan intercept. Nu visar dock resultaten i tidigare avsnitt att interceptalternativet är bra i de flesta fallen.

Nästa moment består i att välja variansstruktur. I de modeller som studeras i föreliggande arbete innebär det att finna ett bästa v-värde i respektive "modellgrupp". Nu är det inte så enkelt att välja modell utifrån ett sambandssdiagram endast genom att titta på det. Det kan bero på att man har olika skalor på y- resp x-axeln vid olika tillfällen, det är svårt att avgöra effekten av enstaka avvikande värden etc. Inom teorin för regressionsanalys finns dock några förslag till hur man skattar strukturen. I bilaga 2 redovisas Harvey's metod vid skattning av v -värdet i den typ av variansstruktur vi använder i modellerna. Dessutom har vi konstruerat två datorprogram VSKATT för fallet utan intercept och VSKATTI för fallet med intercept.

REFERENSER

Cochran, W.G. (1977). Sampling techniques. John Wiley & Sons, Inc., New York.

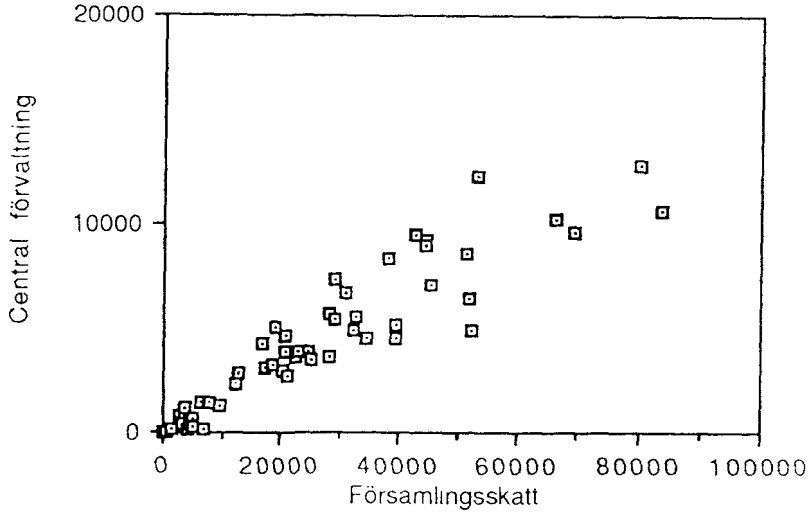
Des Raj (1972). The design of sample surveys. McGraw-Hill.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, Inc., New York.

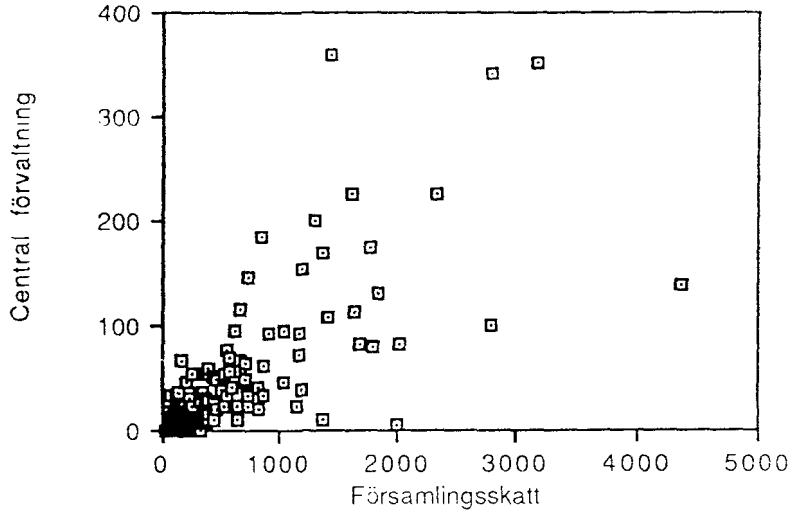
Lundström, S. (1990). Översikt av estimatorer för stora och små redovisningsgrupper.

Sambandssdiagram för redovisningsgrupper i POPULATION A.

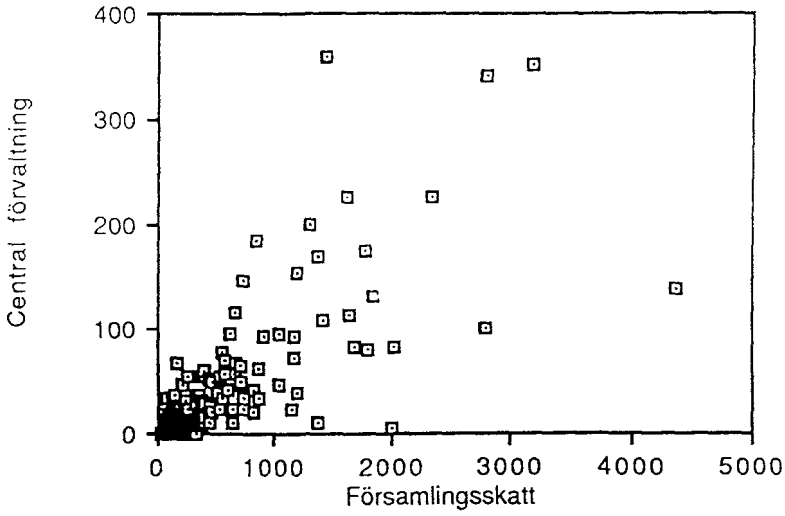
1. Samfälligheter.



2. Pastorat.



3. Församlingar.



Skattning av graden av heteroskedasticitet.

Harvey [1976] visar ett sätt att skatta v i en regressionsmodell av följande utseende (han har t o m en ändå generellare ansats):

$Y_k = \alpha + \beta x_k + u_k$, där $u_k \approx N(0, \sigma^2 x_k^v)$ samt u_k och u_l ober (k≠l).

(k=1,..., n_0)

Genom att fördelningen på residualen antas känd kan en ML-skattning av de okända parametrarna. Lösningen görs med "the method of scoring", vilken består i ett iterativt förfarande.

Sätt

$\beta = (\alpha, \beta)'$; $\delta = (\log \sigma^2, v)'$; $\mathbf{x_k} = (1, x_k)'$ och $\mathbf{z_k} = (1, \log x_k)'$

Harvey's metod går ut på att initialt anta homoscedasticitet (v=0). De erhållna skattningarna av parametrarna används som startvärden i iterationen. (σ^2skattas med hjälp av de observerade residualerna.).

Steg nr t+1 har följande utseende:

$$a)\ \hat{\delta}(t+1) = \hat{\delta}(t) + \left(\sum_{s_0} \mathbf{z_k}\, \mathbf{z_k}' \right)^{-1} \sum_{s_0} \mathbf{z_k} \left[\hat{\sigma}^{-2}(t)\, x_k^{-v(t)} (y_k - \mathbf{x_k}' \hat{\beta}(t))^2 - 1 \right]$$

och

$$b)\ \hat{\beta}(t+1) = \hat{\beta}(t) + \left(\sum_{s_0} \hat{\sigma}^{-2}(t)\, x_k^{-v(t)}\, \mathbf{x_k}\, \mathbf{x_k}' \right)^{-1} \sum_{s_0} \mathbf{x_k}\, \hat{\sigma}^{-2}(t)\, x_k^{-v(t)} (y_k - \mathbf{x_k}'\, \beta(t))$$

där $\hat{\delta}(t)$, $\hat{\beta}(t)$, $\hat{\sigma}^2(t)$ och $\hat{v}(t)$ är skattningar av δ, β, σ^2 resp v erhållna vid t:te iterationen.

Nedanstående program, VSKATT, beräknar v (vi kallar värdet v_{opt} i rapporten) i fallet utan intercept. Även för fallet med intercept har vi konstruerat ett program (VSKATTI). Harvey's iterationsmetod konvergerar inte alltid om vi har små x-värden. I programmen kan man via parametern GRANS utesluta de x-värden som är mindre än GRANS.

REFERENSER

Harvey, A.C. (1976), Estimating Regression Models with Multiplicative Heteroscedasticity. Econometrica, Vol. 44, No. 3.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown /beige covers).

Reports published during 1991:

1991:1 (grön)	Computing elementary aggregates in the Swedish consumer price index **(Jörgen Dalén)**
1991:2 (grön)	Översikt av estimatorer för stora och små redovisningsgrupper **(Sixten Lundström)**
1991:3 (grön)	Interacting nonresponse and response errors **(Håkan L Lindström)**
1991:4 (grön)	Effects of nonresponse on survey estimates in the analysis of competing exponential risks **(Ingrid Lyberg)**
1991:5 (grön)	Study of the estimation precision in the Chinese MIS surveys **(Bengt Rosén)**
1991:6 (beige)	Abstracts I - Sammanfattning av metodrapporter från SCB
1991:7 (grön)	Kvalitetsfonden 1990: Projekt - handläggning - utvärdering **(Roland Friberg och Håkan L Lindström)**
1991:8 (grön)	Tre återintervjustudier: Inkomstfördelningsundersökningen **(Jan Eriksson)**, Högskoleenkätuppföljningen **(Anders Karlsson)**, Årsarbetskraften **(Majvor Karlsson och Solveig Thudin)**
1991:9 (gul)	What metainformation should accompany statistical macrodata **(Bo Sundgren)**
1991:10 (grön)	The Family Expenditure Survey: An Experiment with Incentives **(Håkan L. Lindström)**, Seasonal Variation and Response Behaviour in Swedish Households Expenditure **(Peter Lundquist)**, Reducing Nonresponse Rates in Family Expenditure Surveys by Forming Ad Hoc Task Forces **(Lars Lyberg)**, A Study of Errors in Swedish Consumption Data **(Martin G. Ribe)**
1991:11 (gul)	Statistical Metainformation and Metainformation Systems **(Bo Sundgren)**

1991:12 (beige)	Abstracts II - Sammanfattning av metodrapporter från SCB
1991:13 (grön)	Bortfallsbarometern nr 6 **(Mats Bergdahl, Sonia Ekman, Anders Lindberg, Peter Lundquist, Monica Rennermalm)**
1991:14 (grön)	Kvalitetsrapport 1991 - Utveckling av kvaliteten för SCBs statistikproduktion **(Jan Eklöf, Per Nilsson)**
1991:15 (grön)	Variance estimation for systematic pps-sampling **(Bengt Rosén)**
1991:16 (beige)	Abstracts III - Sammanfattning av metodrapporter från SCB
1991:17 (grön)	Samband mellan SCBs statistikprodukter m m
1991:18 (grön)	Utredning beträffande "Random digit dialing" **(Anders Holmberg)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.