Forebyggelse og reparation - vægtning af data

Brian Larsen Thorsted

bnl@dst.dk

DANMARKS STATISTIK

Indhold

- Usikkerhed i stikprøveundersøgelser
  - Bortfald
- Estimation
  - Regressionsestimator
  - Reduktion af usikkerhed
  - Repræsentativitet genskabes
- Konkluderende bemærkninger

Den perfekte verden

'*In a perfect world a survey has no nonresponse. All selected elements cooperate and deliver all of the requested data, with no measurement error. In that perfect but non-existent world, a survey has only sampling error.*'

- Lundström og Särndal 2005, *Estimation in Surveys with Nonresponse*

Usikkerhed i stikprøver

- Kilder til usikkerhed:
- *Tilfældig fejl*
  - stokastisk struktur i udvælgelse
- *Systematisk fejl / bias*
  - målefejl
  - bortfald/missing
  - fejlsøgning
  - opregning
  - tabellering
  - mm.
- Kun den tilfældige fejl kan kvantificeres
- Ønsker at minimere den tilfældige fejl og kontrollere den systematiske fejl
- Forebyggelse og reparation

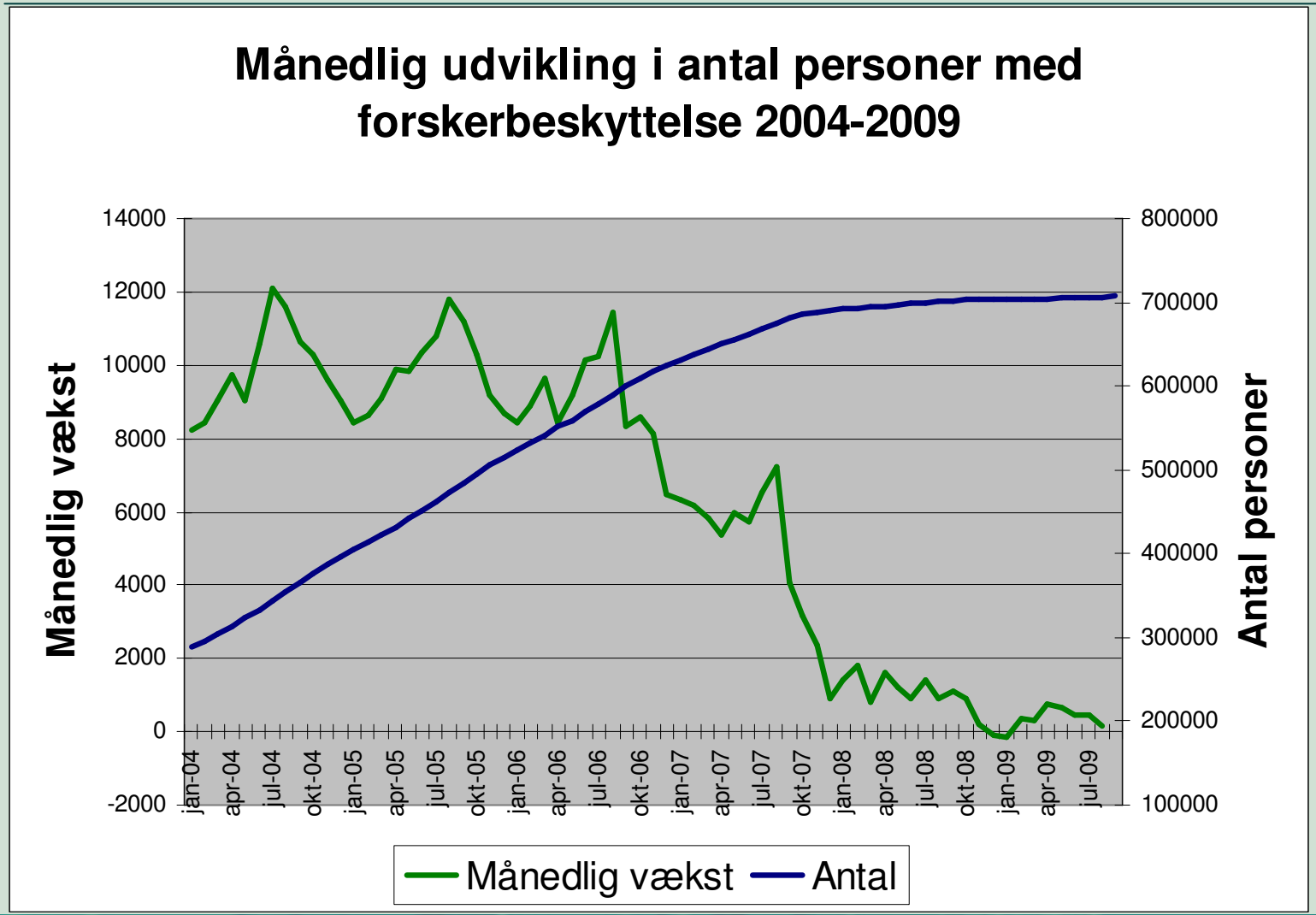
Bortfald

- Fokus på bortfaldet
- Et problem hvis bortfaldet er skævt og korreleret med spørgeskemavariable
- …resultatet er bias på estimaterne

Bortfald

- Eksempel (Särndal & Lundström 2005)
- Gennemsnitlig indkomst, M: 196 000 Kr.
- Gennemsnitlig indkomst, K: 135 000 Kr.
- Svarprocent, M: 50 pct.
- Svarprocent, K: 90 pct.
  - kvinder er overrepræsenteret
  - Gennemsnitlig indkomst er lavest for kvinder
  - Total populationsindkomst underestimeres

Bortfald - Forskerbeskyttelse



DANMARKS STATISTIK

Estimation

- Beregn en vægt for hver interviewperson
- Angiver hvor mange andre hver interviewperson repræsenterer
- Udgangspunkt er designvægten

$$v = \frac{N}{n}$$

Traditionel estimation

- Traditionel opregning deler baggrunds- og svarpopulation op i flerdimensional tabel
- Tabel dannes typisk på baggrund af køn, alder og geografi
- Der tages højde for skævt bortfald
- Fælles vægt til alle i samme celle
- Beregningsmæssigt ikke tungt
- Men…
  - demografiske faktorer forklarer typisk kun lille del af skævt bortfald
  - stikprøvestørrelse er begrænsning

Regressionsestimat

- Modellerer **Y** vha. j hjælpevariable, $X_1, \cdots, X_j$, i regressionsmodel

$$y_k = B_1 x_{1k} + \cdots + B_j x_{jk} + e_k$$

- Koefficienterne $B_1, \cdots, B_j$ bestemmes vha. mindste kvadraters metode
- Jo større R^2 jo større effektiv stikprøve

Regressionsestimat

- Populationstotalen estimeres da som

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

$$= \sum_{i \in S} \frac{N}{n} y_i g_i$$

- Bemærk: Model assisted estimation!

Regressionsestimat, varians

- Variansen er

$$\hat{V}(\hat{t}_y^{greg}) = N^2\left(1-\frac{N}{n}\right)\frac{S_y^2}{n}$$

- Hvor

$$\hat{S}_y^2 = \frac{\sum_{i\in S} g_i^2\left(y_i - \sum_{j=1}^{J}\hat{B}_j x_{ji}\right)^2}{n-1} = \frac{\sum_{i\in S} g_i^2 e_i^2}{n-1}$$

DANMARKS STATISTIK

Regressionsestimator

- Regressionsestimat kan håndtere flere registervariable end traditionelt
- Forklaringskraft i registervariable udnyttes i opregning
- Motiveret af stigende bortfald og stigende krav til effektivitet

Regressionsestimator

- Traditionelle kategoriske variable (alder, køn, geografi) kan bruges
- I kombination med en eller flere andre variable
- Hovedvirkning – dvs. ingen kombination!!
- Kategoriske og kontinuerte variable
- Repræsentativitet genskabes på en lang række registervariable
- Forskellige vægte

Reduktion af usikkerhed

- Arbejdskraftundersøgelsen – bortfald 43 pct.
- Antal arbejdsløse (AKU) 2. kvt. 2009

Metode	Arbejdsløse	Usikkerhed	Relativ usikkerhed (CV)
Direkte estimat	143.000	+/-11.000	3,9 pct.
Alder*køn*region	160.000	+/-13.000	4,2 pct.
Anvendt estimat	177.000	+/-14.000	4,0 pct.

- Effektiv stikprøve forøges med 10 pct.

Repræsentativitet genskabes

	Population	Simpel tilfældig stikprøve	Stikprøve efter forskerbeskyttelse	Svarstikprøven efter alt bortfald	Vægte de svar
	Pct.				
Køn					
Mand	50	49	49	47	50
Kvinde	50	51	51	53	50
Aldersgruppe					
16-19 år	7	6	6	5	7
20-29 år	15	15	13	10	15
30-39 år	19	17	15	15	19
40-49 år	20	21	22	23	20
50-59 år	18	18	19	20	18
60-74 år	21	23	25	27	21
Uddannelse					
Grundskole	37	36	35	29	37
Gymnasial og erhvervsfagliguddannelse	44	45	45	47	44
Mellemlang videregående uddannelse	13	15	15	18	13
Lang videregående uddannelse	6	5	5	6	6

Repræsentativitet genskabes

	Population	Simpel tilfældig stikprøve	Stikprøve efter forskerbeskyttelse	Svarstikprøven efter alt bortfald	Vægtede svar
	Pct.				
Socioøkonomisk status					
Selvstændig	4	5	5	5	4
Lønmodtager	60	60	60	65	60
Arbejdsløs	2	2	2	2	2
Uddannelsessøgende	8	7	7	5	8
Pensionist/Efterlønsmodtager	16	18	19	18	16
Uden for arbejdsstyrken, børn, øvrige	9	9	7	5	9
Gennemsnitlige familieindkomst					
0-49.999 kr.	29	27	27	25	29
50.000-99.999 kr.	14	14	14	14	14
100.000-199.999 kr.	30	32	33	33	30
200.000-299.999 kr.	15	15	15	15	15
over 300.000 kr.	12	12	12	13	12
Flyttet					
Flyttet efter flytteblanket	50	47	40	36	50
Ikke flyttet	50	53	60	64	50

Kontrol af estimation

- Estimation af kendt registerindkomst

Metode	Gennemsnit (Før skat i kr)	Spredning
Direkte estimat	180.000	118.000
Køn*alder*region	181.000	118.000
Anvendt estimat (uden indkomst)	177.000	116.000
Register	175.000	116.000

Repræsentativitet genskabes

- Estimater er baseret på fordeling fra registre – ikke fra svarene
- Kvalitetsløft
  - Reduktion af bias

Reduktion af usikkerhed

- Ved korrelation med registervariable sænkes usikkerhed
- Kvalitetsløft
  - Smallere konfidensintervaller
  - Større effektiv stikprøve

Analyse

- Regressionsanalyser bør laves på vægtede svar
- Sikrer repræsentativitet og middelrette estimater
- Konservativ signifikansniveauer
  - skyldes empirisk varians fra før…

$$\hat{S}_y^2 = \frac{\sum_{i \in S} g_i^2 e_i^2}{n-1}$$

DANMARKS STATISTIK

Konkluderende bemærkninger

- Forebyggelse/reparation nødvendig
- Muligt at genskabe repræsentativitet på en række parametre
- Samtidig med gevinst på effektiv stikprøve
- …men (survey)-verden bliver ikke perfekt af den grund