Invariantie van schattingen van de omvang van een populatie en de rol van actieve en passieve variabelen

Peter G.M. van der Heijden*, Joe Whittaker†, Maarten Cruyff*,
Bart Bakker‡ en Rik van der Vliet‡

* Utrecht University
† Lancaster University
‡ Statistics Netherlands

NPSO Oktober 2011

Outline

1. Data en context van probleem
2. Optelbaarheidseigenschappen
   - Loglineaire modellen met een covariaat
   - Loglineaire modellen met twee covariaten
   - Loglineaire modellen als covariaten niet in elke registratie zitten
   - Loglineaire modellen met drie registers en een covariaat
3. Voorbeeld
4. Conclusies

1. Data en context van probleem

Data: populatie van mensen met Afghaanse, Iraanse en Iraakse nationaliteit die met of zonder geldige documenten in Nederland verblijven.

Voorbereiding van virtuele volkkstelling 2011

Nederland: uitgevoerd op registers

GBA: officiele registratie
HKS: politieregistratie van verdachten

TABLE 1
Linked registers GBA and HKS

	HKS	
GBA	Included	Not included
Included	1,085	26,254
Not included	255	-

Gebruikelijke assumpties

- Aanwezig in GBA (ja/nee) is statistisch onafhankelijk van aanwezig in HKS
- Inclusiekansen homogeen voor minimaal 1 register

Oplossingen voor schendingen

- Neem covariaten mee, gebruik loglineaire modellen
- Gebruik derde registratie
- Latente variabele model (drie registraties nodig)

Context

- Grote steekproef, dus veel covariaten kunnen meegenomen worden
- Eerder onderzoek staat toe dat covariaten kunnen worden meegenomen die niet in alle registraties zitten

- Constateringen:

(i) Verschillende loglineaire modellen van zelfde tabel kunnen identieke omvangschattingen opleveren

(ii) Onder bepaalde loglineaire modellen kun je optellen over covariaten zonder dat dit omvangschattingen beïnvloed.

2. Optelbaarheidseigenschappen

Loglineaire modellen met 1 covariaat

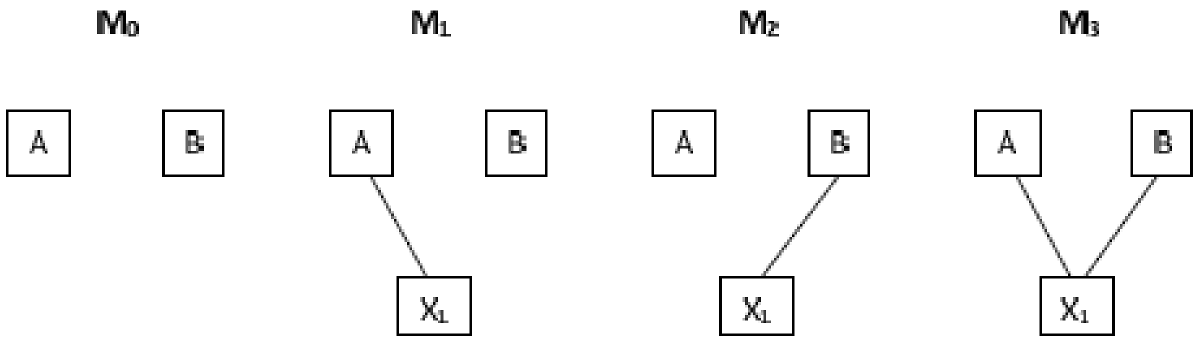


FIG 1. *Graphs for loglinear models with one covariate*

TABLE 2

Models fitted to contingency table of variables A (GBA), B (HKS) and to A, B and X_1 (gender), deviances, degrees of freedom and estimated numbers missed.

Model	Deviance	df	Missed
M_0: $[A][B]$	0.0	0	6,170.3
M_1: $[AX_1][B]$	548.5	1	6,170.3
M_2: $[A][BX_1]$	1.1	1	6,170.3
M_3: $[AX_1][BX_1]$	0.0	0	5,696.1

… totale populatieomvang identiek maar verschillende onderliggende schattingen van strata

TABLE 3

Observed and fitted counts for the three-way table of A (GBA), B (HKS) and X_1 (gender); for A and B level 1 is present and for X_1 level 1 is male.

A	B	X_1	obs	M_1	M_2	M_3
1	1	1	972	629.2	976.5	972.0
2	1	1	234	234.0	229.5	234.0
1	2	1	14,883	15,225.8	14,883.0	14,883.0
2	2	1	0	5,662.2	3,497.9	3,582.9
1	1	2	113	455.8	108.5	113.0
2	1	2	21	21.0	25.5	21.0
1	2	2	11,371	11,028.2	11,371.0	11,371.0
2	2	2	0	508.1	2,672.5	2,113.2

Loglineaire modeller met 2 covariaten

Tabel *niet* optelbaar over variabelen die op *short* path van A naar B liggen

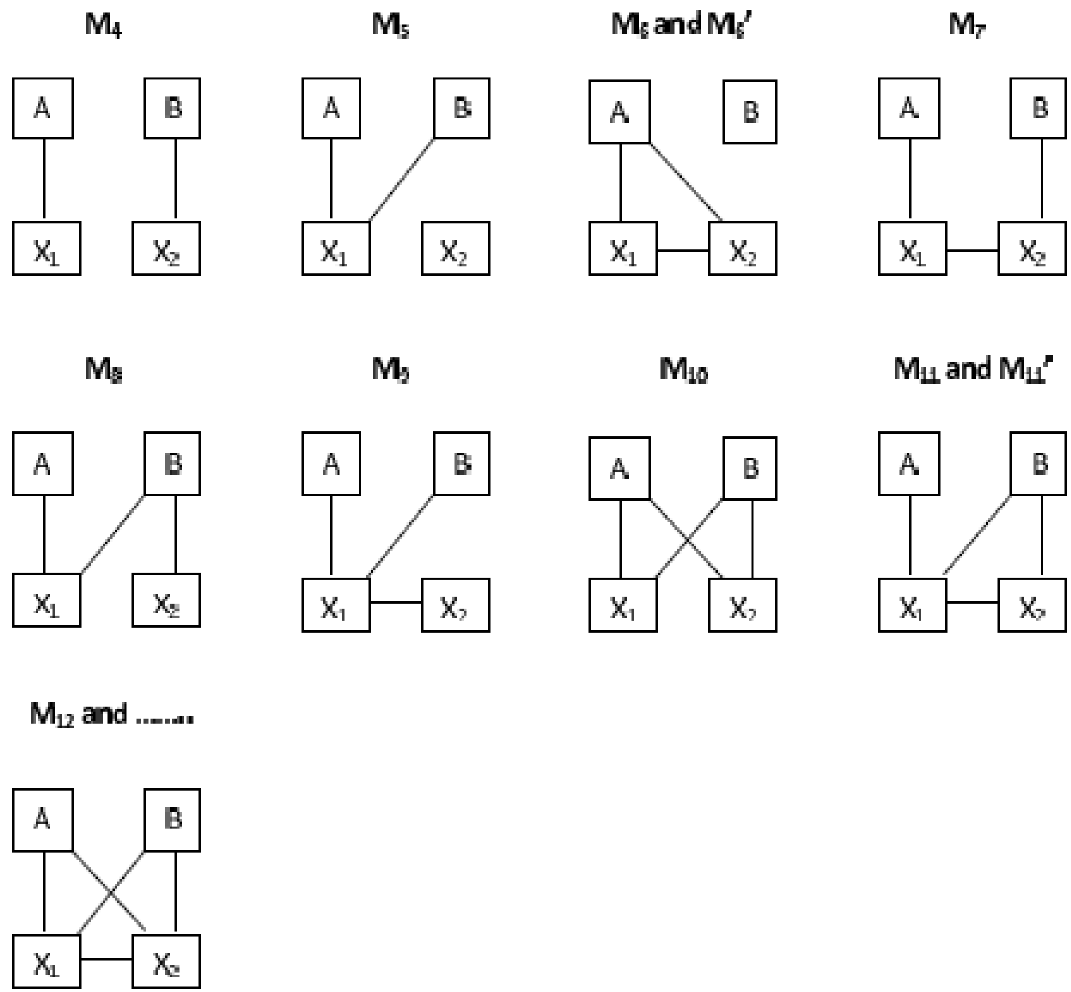


FIG 2. *Graphs of loglinear models with two covariates*

TABLE 4

Models fitted in four-way array of variables A, B, X_1 and X_2; registers A (GBA), B (HKS), covariates X_1 (gender), X_2 (age coded in four levels); deviances, degrees of freedom and estimated numbers missed.

	Model	Deviance	df	Missed
M_4	$[AX_1][BX_2]$	617.6	13	6,170.3
M_5	$[AX_1][BX_1][X_2]$	228.6	15	5,696.1
M_6	$[AX_1X_2][B]$	718.2	7	6,170.3
M_6'	$[AX_1][AX_2][X_1X_2][B]$	725.6	10	6,170.3
M_7	$[AX_1][BX_2][X_1X_2]$	588.6	10	6,179.4
M_8	$[AX_1][BX_1][BX_2]$	69.1	12	5,696.1
M_9	$[AX_1][BX_1][X_1X_2]$	200.2	12	5,696.1
M_{10}	$[AX_1][BX_2][AX_2][BX_1]$	65.9	9	5,837.1
M_{11}	$[AX_1][BX_1X_2]$	4.9	6	5,696.1
M_{11}'	$[AX_1][BX_1][BX_2][[X_1X_2]$	34.4	9	5,696.1
M_{12}	$[AX_1X_2][BX_1X_2]$	0.0	0	5,910.1
M_{12}'	$[AX_1X_2][BX_1][BX_2]$	23.3	3	6,257.1
M_{12}''	$[AX_1][AX_2][BX_1][BX_2][X_1X_2]$	31.2	6	5,831.4

Covariaten niet in elke registratie

TABLE 5
Covariate X_1 is only observed in register A and X_2 is only observed in B

Panel 1: Observed counts

		$A=1$		$A=2$
		$X_1=1$	$X_1=2$	X_1 missing
$B=1$	$X_2=1$	259	539	13,898
	$X_2=2$	110	177	12,356
$B=2$	X_2 missing	91	164	-

Panel 2: Fitted values under $[AX_2][BX_1][X_1X_2]$

		$A=1$		$A=2$	
		$X_1=1$	$X_1=2$	$X_1=1$	$X_1=2$
$B=1$	$X_2=1$	259.0	539.0	4,510.8	9,387.2
	$X_2=2$	110.0	177.0	4,735.8	7,620.3
$B=2$	$X_2=1$	63.9	123.5	1,112.4	2,150.2
	$X_2=2$	27.1	40.5	1,167.9	1,745.4

Maximale model is [AX2][BX1][X1X2]

MAR assumptie

Verzadigd model

X1 in A, X2 in B, X3 in A and B

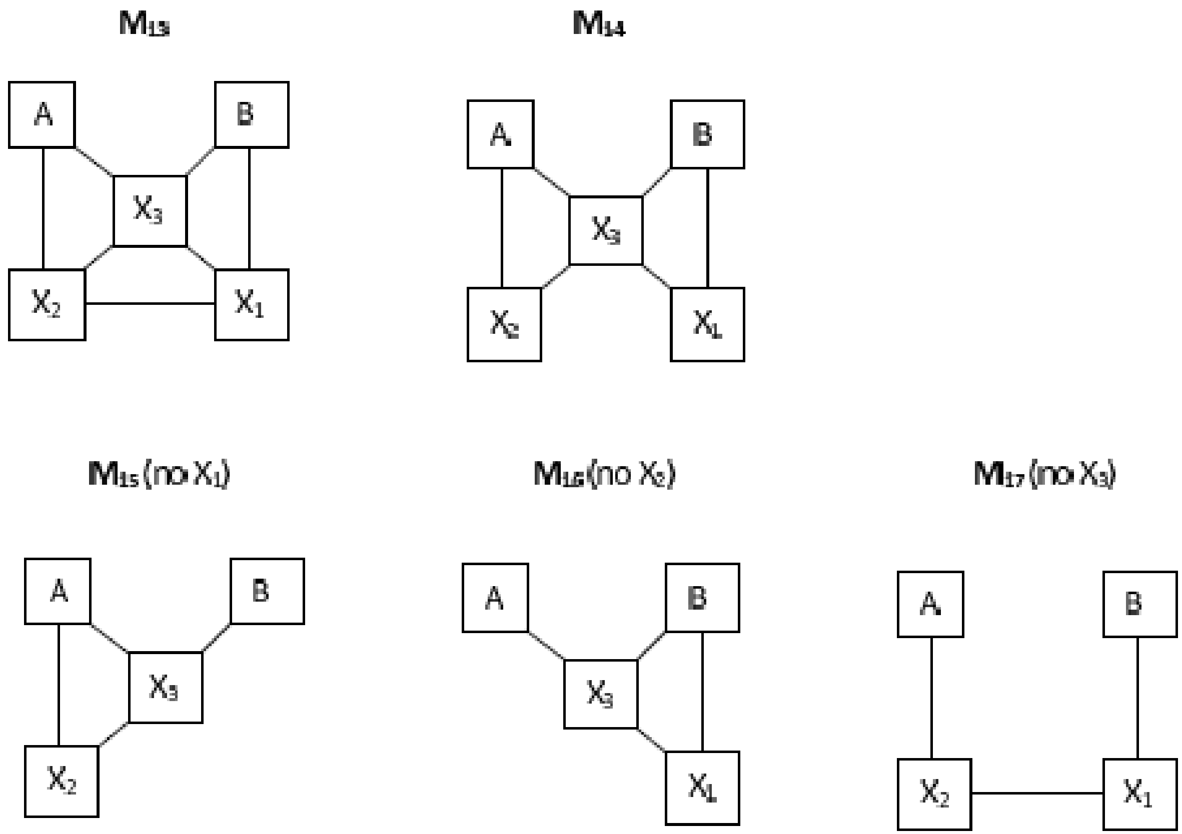
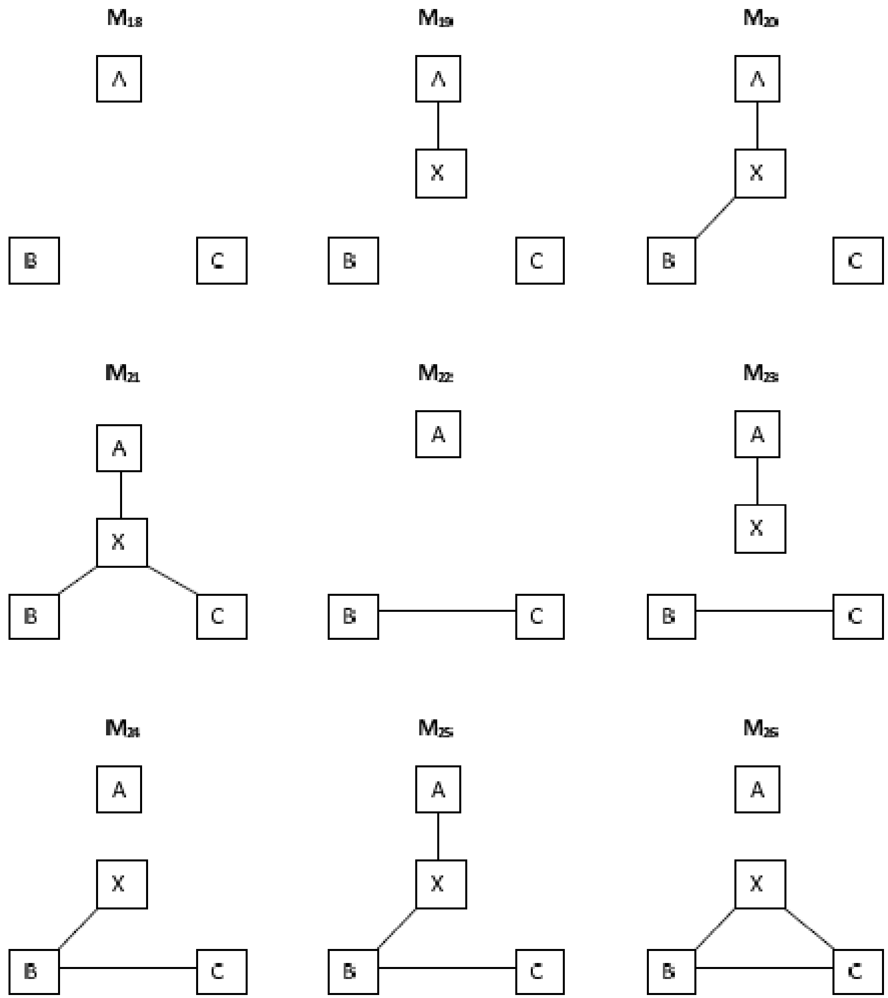
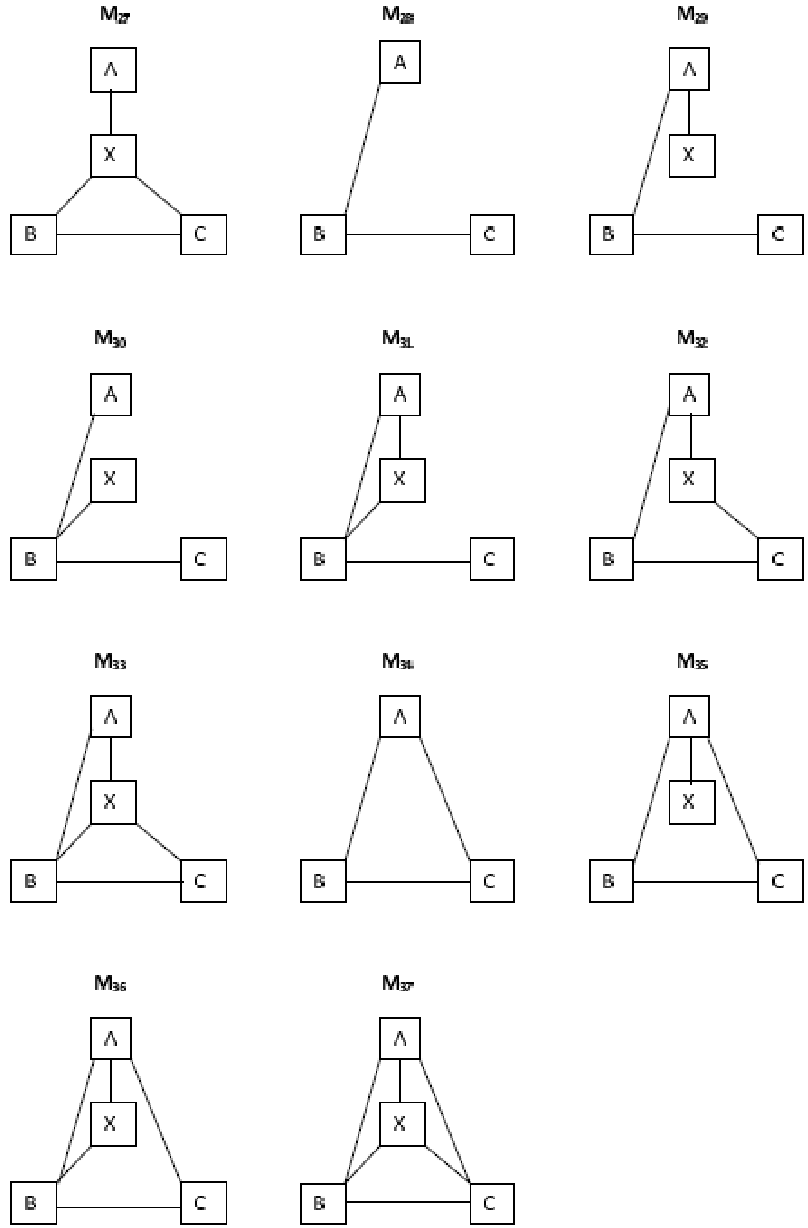


FIG 3. *Graphs of loglinear models with partially observed covariates*

Three registers, een covariaat




M27
M28
M29
M30
M31
M32
M33
M34
M35
M36
M37
A
X
B
C

3. Voorbeeld

Voorbeeld: X1 is geslacht, X2 is leeftijd, X3 is nationaliteit, X4 is burgerlijke staat (alleen in GBA), X5 is politie regio (alleen in HKS)

TABLE 6

Models fitted to example of variables A, B, X_1 to X_5, deviances, degrees of freedom, AIC's, estimated population size and 95 percent confidence intervals.

	Model	Deviance	df	AIC	Pop.size	CI
N_1	$[AX_1X_2X_3][BX_1X_2X_3]$	0	0	144.0	33,098.6	32,209 - ∞
N_2	$[AX_1X_2][BX_1X_2X_3]$	24.9	16	136.8	33,504.1	32,480 - 35,468
N_3	$[AX_1X_2X_3][BX_1X_2]$	28.8	16	140.7	33,504.1	32,480 - 35,468
N_4	$[AX_1X_2X_5][BX_1X_2X_3X_4]$	75.7	72	315.7	33,504.1	32,480 - 35,468
N_5	$[AX_1X_2X_5][BX_1X_2X_3X_4][X_4X_5]$	75.7	71	317.7	33,503.8	32,395 - 35,543
N_6	$[AX_1X_2X_3X_5][BX_1X_2X_4]$	523.8	72	763.7	33,504.1	32,480 - 35,468
N_7	$[AX_1X_2X_3X_5][BX_1X_2X_4][X_4X_5]$	289.1	71	531.4	33,510.9	32,363 - 35,432

Table 7

Estimates for GBA with each of the other variables under model N_4.

	in HKS	not in HKS	male	female
in GBA	1,085.0	26,254.0	15,855.0	11,484.0
not in GBA	255.0	5,910.0	3,874.7	2,290.3

	15-25	25-35	35-50	50-64
in GBA	7,234.0	8,361.0	9,185.0	2,559.0
not in GBA	1,292.2	2,167.3	1,925.9	779.7

	Afghan	Iraqi	Iranian
in GBA	12,818.8	8,743.3	5,776.8
not in GBA	2,950.9	1,914.5	1,299.7

	unmarried	married	4 large cities	elsewhere
in GBA	14,698.2	12.640,8	9,720.0	17,619.0
not in GBA	3,302.3	2,862.7	2,182.6	3,982.5

4. Discussie en conclusie

- Mogelijk om veel covaraten op te nemen
- Beschrijving van populatie in termen van die covariaten
- Terminologie: actieve en passieve covariaten
- Met covariaten die slechts in 1 register zitten heb je weinig power om interacties vast te stellen ->

  Deze worden daardoor vaak passieve covariaten