Interne notater

STATISTISK SENTRALBYRÅ

81/24

1. september 1981

STANDARDISERING SOM ANALYSETEKNIKK

AV

IB THOMSEN*

INNHOLD

* H.T. Amundsen og R. Aaberge har gitt nyttige kommentarer til et tidligere utkast.

1. INNLEDNING

La oss innledningsvis se på et eksempel på bruk av standardisering. I tabell 1 er vist gjennomsnittlig antall levendefødte etter morens seksuelle debutalder. Data er hentet fra den norske fruktbarhetsundersøkelsen. Tallene i parentes angir antall kvinner.

Tabell 1. Gjennomsnittlig antall levendefødte etter mors seksuelle debutalder

	Debutalder				
I alt	Under 16 år	16-17 år	18-19 år	20-21 år	22-35 år
1.62	1.22	1.54	1.66	1.71	1.85
(3686)	(212)	(1277)	(1302)	(527)	(368)

Av resultatene i tabell 1 er det lett å dra den slutning at fruktbarheten, målt ved antall levendefødte barn, vokser med voksende debutalder. Nå er det klart at også andre variable påvirker antall levendefødte barn, f.eks. kvinnens alder. I tabell 2 er gjennomsnittlig antall levendefødte gitt etter morens seksuelle debutalder og alder.

Tabell 2. Gjennomsnittlig antall levendefødte barn i grupper for mors alder og debutalder

		Debutalder					Tallet på
Alder	I alt	Under 16	16-17	18-19	20-21	22-35	kvinner
I alt ..	1.62	1.22	1.54	1.66	1.71	1.85	3686
18-24 ..	0.44	0.64 (137)	0.49 (504)	0.33 (277)	0.33 (54)	0 (11)	983
25-29 ..	1.39	2.00 (29)	1.58 (304)	1.38 (340)	1.10 (121)	0.78 (55)	849
30-34 ..	2.06	2.43 (23)	2.50 (211)	1.94 (299)	1.84 (148)	1.75 (122)	803
35-39 ..	2.47	3.19 (16)	2.69 (156)	2.43 (212)	2.25 (106)	2.29 (87)	577
40-44 ..	2.68	2.42 (7)	2.85 (102)	2.87 (174)	2.44 (98)	2.40 (93)	474
Tallet på kvinner		212	1277	1302	527	368	

Dersom en ser på en bestemt aldersgruppe, får en et noe annet bilde av debutalderens effekt på fruktbarheten enn tilfellet var i tabell 1. For samtlige aldersgrupper gjelder det at gjennomsnittlig antall levendefødte avtar med økende debutalder. Når det i hver aldersklasse gjelder at fruktbarheten avtar med økende debutalder, samtidig som det omvendte gjelder for hele utvalget, skyldes det at debutalderen i gjennomsnitt er lavere jo yngre kvinnen er.

Dette medfører at gjennomsnittene i tabell 1 gir et dårlig bilde av debutalderens virkning på fruktbarheten, fordi aldersfordelingen varierer meget mellom gruppene for debutalder. For å fjerne, eller redusere, effekten av at aldersfordelingen varierer, går en ofte fram på følgende måte:

Innen hver kolonne i tabell 2 utregnes et veid gjennomsnitt av tallene i kolonnen. Vektene i gjennomsnittet er tallet på kvinner ytterst til høyre i tabellen dividert med totalt antall kvinner i utvalget. For kvinner med debutalder 16-17 år utregnes det standardiserte gjennomsnitt på følgende måte:

$$\bar{X}^{st} = 0.49\frac{983}{3686} + 1.58\frac{849}{3686} + 2.50\frac{803}{3686} + 2.69\frac{577}{3686} + 2.85\frac{474}{3686} = 1.83$$

Innen de øvrige kolonner regnes ut de standardiserte gjennomsnittene på samme måten idet vektene holdes konstant. Resultatet er vist i tabell 3.

Tabell 3. Standardisert gjennomsnittlig antall levendefødte etter mors seksuelle debutalder. Standardisert med hensyn til aldersfordelingen i utvalget

Under 16 år	16-17 år	18-19 år	20-21 år	22-35 år
1.96	1.83	1.57	1.39	1.23

Tallene i tabell 3 hevdes ofte å gi en bedre beskrivelse av debutalderens "sanne" virkning på fruktbarheten, og er i dette eksemplet helt forskjellig fra det inntrykk tabell 1 gir. I dette notatet skal vi se nærmere på hvilke betingelser som må være oppfylt for at sådan tolkning skal være riktig, og se på andre måter en kan gå fram på i slike tilfeller.

Standardisering er ingen ny teknikk, når en f.eks. ønsker å sammenlikne dødelighet mellom to geografiske regioner, er det vanlig å standardisere med hensyn på alder for å fjerne eller redusere virkningen av en eventull forskjell i aldersfordeling i de to områdene. De analytiske egenskaper ved en slik framgangsmåte har vært underkastet grundige studier. En omfattende oversikt over relevant litteratur er gitt av Hoem (1979), Ahlbom (1980).

Bruken av standardisering i økonomi er kanskje enda mer kjent idet indekser ikke er noe annet enn standardisering. Også denne bruken av standardisering underkastes grundige studier, se f.eks. Gørtz (1977). I Thomsen (1973) er standardisering brukt for å redusere virkningene av frafall ved analyse av utvalgsundersøkelser.

I de senere årene har standardisering vært brukt til mer generelt analysearbeid, slik som i eksemplet overfor, spesielt når data bare er tilgjengelig i tabellform. Pullum (1978) har forsøkt å se på virkningen av forskjellige variables innflytelse på fruktbarheten. F. eks. har han estimert virkningen av utdanningen på fruktbarhet etter å ha redusert virkningen av variab-

len "varighet som gift" ved standardisering. Johansson (1978) har forsøkt å isolere virkningen av å bo i bestemt fylke ved å standardisere for alders og yrkesfordelingen. Endelig har Hellevik (1978) utført en kausul analyse ved hjelp av en teknikk som er identisk med standardisering.

Det er disse siste anvendelser av standardisering vi skal være spesielt opptatt av i dette notatet. Ved å sammenlikne standardisering med regresjonsanalyse og log-lineær analyse, skal vi gi en vurdering av standardisering som analyseteknikk, og vise at en ofte med fordel kan bruke andre teknikker enten istedenfor eller i tillegg til standardisering. I kapittel 2 skal foretas en sammenlikning mellom standardisering og regresjonsanalyse hvor alle de uavhengige variable er binære (dummy) variable. I kapittel 3 skal det vises at gjentatt bruk av standardisering etter flere variable kan sees på som en estimeringsteknikk når en estimerer parametre i et sett av lineære strukturligninger, ofte kalt kausal analyse. I kapittel 4 blir standardisering i forbindelse med analyse av hyppighetstabeller sammenliknet med log-lineære modeller.

Notatet er med vilje gjort så lite teknisk som mulig og inneholder ingen nye resultater. Hensikten med notatet er å gi en kritisk vurdering av standardisering som analyseteknikk, og peke på velkjente teknikker som ofte bør brukes istedetfor standardisering når en har adgang til hele datasettet og ikke bare data på tabellform.

2. STANDARDISERING SAMMENLIKNET MED REGRESJONSANALYSE

2.1 Standardisering i toveis tabeller

La oss først kort repetere teknikken på et enkelt konstruert eksempel:

I tabell 4 er gjennomsnittene til variabel C gitt for forskjellige verdier av variabel A og B. Tallene i parentes angir antall observasjoner.

Tabell 4.

		Variabel A			
		1	2	3	Total
Variabel B	1	0.20 (15)	0.30 (70)	0.40 (820)	0.39 (905)
	2	0.10 (180)	0.20 (1 000)	0.30 (15)	0.19 (1 195)
	Total	0.11 (195)	0.21 (1 070)	0.40 (835)	0.28 (2 100)

På grunnlag av tabell 4 ønsker en nå å si noe om virkningen av variabel B på variabel C, "kontrollert for variabel A". Hvis en først ser på marginalene til høyre i tabellen, vil en se at gjennomsnittet for variabel C ligger 0.11 over populasjonsgjennomsnittet når B har verdien 1, mens det ligger 0.09 under populasjonsgjennomsnittet når variabel B har verdien 2. Det er imidlertid klart at marginalene på samme måten som i tabell 2, er avhengig av hvordan observasjonene er fordelt langs variabel A, og marginalene er derfor ikke bare et uttrykk for effekten av variabel B. Som i innledningseksemplet kan en nå regne ut de standardiserte gjennomsnittene.

Det standardiserte gjennomsnitt når variabel B har verdien 1 er f.eks.

$$0.20 \cdot \frac{195}{2\ 100} + 0.30 \cdot \frac{1\ 070}{2\ 100} + 0.40 \cdot \frac{835}{2\ 100} = 0.33.$$

Det tilsvarende standardiserte gjennomsnitt når variabel B har verdien 2 er

$$0.10 \cdot \frac{195}{2\ 100} + 0.20 \cdot \frac{1\ 070}{2\ 100} + 0.30 \cdot \frac{835}{2\ 100} = 0.23.$$

En tolker nå resultatet på følgende måte:

Når variabel B har verdien 1 er gjennomsnittet til C 0.10 høyere enn når variabel B har verdien 2. (Hvis en ser på tabell 4 synes den konklusjonen å stemme godt med tallene. Årsaken er naturligvis at tabellen er konstruert. Vanligvis kan en ikke direkte lese resultatene på denne måten).

Vi skal nå se på de forutsetninger som må være oppfylt for at denne konklusjonen skal være riktig. Men først skal vi innføre en del notasjon.

La oss tenke oss at gjennomsnittene til variabel C er gitt for forskjellige verdier av variablene A og B i en to-veis tabell. Da definerer vi følgende:

$\bar{C}_{ij}$: Gjennomsnittet av variabel C når variabel B har verdien i og variabel A har verdien j. I det følgende skal vi kort skrive celle (i,j).

n_{ij}: Antall observasjoner i celle (i,j).

La dessuten

$$\sum_{i=1}^{I} n_{ij} = n_{\cdot j} \quad \text{og} \quad \sum_{j=1}^{J} n_{ij} = n_{i\cdot}.$$

Det følger da at de standardiserte gjennomsnittene kan skrives som

$$\bar{C}^*_{i\cdot} = \sum_{j=1}^{J} \bar{C}_{ij} \frac{n_{\cdot j}}{n}, \quad i = 1, 2, \ldots, I$$

og

$$\bar{C}^*_{\cdot j} = \sum_{i=1}^{I} \bar{C}_{ij} \frac{n_{i\cdot}}{n}, \quad j = 1, 2, \ldots, J.$$

Vi skal nå se på en velkjent modell under hvilken de standardiserte gjennomsnittene har en tolkning som er i tråd med konklusjonene som ble gjort ovenfor.

For hver person i utvalget defineres tre binære variable

$$X_{1i} = \begin{cases} 1 & \text{hvis A = 1 for person i} \\ 0 & \text{ellers} \end{cases}$$

$$X_{2i} = \begin{cases} 1 & \text{hvis A = 2 for person i} \\ 0 & \text{ellers} \end{cases}$$

$$Y_{1i} = \begin{cases} 1 & \text{hvis B=1 for person i} \\ 0 & \text{ellers} \end{cases}$$

C_i er responsen for enhet i.

Vi tenker oss da at vi har følgende sammenheng mellom variabel C på den ene siden, og variabel A og B på den andre.

$$E(C_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_0 \qquad (2.1)$$

Modellen (2.1) er en vanlig lineær regresjonsmodell, med binære, uavhengige variable.

Tolkningen av koeffisientene sees lettest ved å sette opp de forventede verdier i tabell 4 under modell (2.1). Disse er

Tabell 5. Forventede verdier til cellegjennomsnittene under modell (2.1)

		Variabel A		
		1	2	3
Variabel B	1	$\beta_1+\beta_3+\beta_0$	$\beta_2+\beta_3+\beta_0$	$\beta_3+\beta_0$
	2	$\beta_1+\beta_0$	$\beta_2+\beta_0$	β_0

Uttrykkene i tabell 5 fås ved å innsette verdiene til variabel X_{1i}, X_{2i} og Y_{1i} i modell (2.1). Verdien i øverste venstre hjørne i tabellen fås f.eks. ved å sette $X_{1i} = 1$, $X_{2i} = 0$ og $Y_{1i} = 1$.

Det er viktig å merke seg at virkningen av å gå fra linje 2 til linje 1 i tabell 5 er den samme uansett verdien på variabel A, nemlig lik β_3. Tilsvarende, kolonnevis kan β_1 og β_2 tolkes.

Vi har altså forutsatt at det i forventning gjelder at variabel B's effekt på C er uavhengig av verdien på variabel A, og at variabel A's effekt er uavhengig av verdien på variabel B. Slike modeller kalles additive uten samspill.

Et naturlig spørsmål er nå hvordan forventningen til de standardiserte gjennomsnittene ligger i forhold til parametrene i modellen (2.1).

Det er da lett å vise at

$$E(\bar{C}^*_{1.}) = \frac{n_{.1}}{n}\beta_1 + \frac{n_{.2}}{n}\beta_2 + \beta_3 + \beta_0 ,$$

og

$$E(\bar{C}^*_{2.}) = \frac{n_{.1}}{n}\beta_1 + \frac{n_{.2}}{n}\beta_2 + \beta_0 .$$

Det følger derav at

$$E(\bar{C}^*_{1.} - \bar{C}^*_{2.}) = \beta_3$$

Forskjellen mellom de to standardiserte gjennomsnittene er altså en forventningsrett estimator for β_3, som vi tidligere har tolket som et mål for variabel B's effekt på variabel C. Under modell(2.1) er det altså rimelig å påstå at differansen mellom de standardiserte gjennomsnittene i tabell 4 gir et "sannere" bilde av variablels B's effekt på variabel C, enn forskjellen mellom marginalgjennomsnittene som er forskjellig fra β_3.

På samme måten kan en nå vise at

$$E(\bar{C}^*_{\cdot j} - \bar{C}^*_{\cdot 3}) = \beta_j \quad \text{for} \quad j = 1.2$$

Vi har altså følgende resultat:

Under modell (2.1) er $(\bar{C}^*_{\cdot j} - \bar{C}^*_{\cdot 3})$ en forventningsrett estimator for β_j, $(j = 1.2)$, og $(\bar{C}^*_{1\cdot} - \bar{C}^*_{2\cdot})$ er en forventningsrett estimator for β_3.

Ved å utføre standardisering i tabell 4 fås:

$$\hat{\beta}_1 = \bar{C}^*_{\cdot 1} - \bar{C}^*_{\cdot 3} = -0.20 \; ; \quad \hat{\beta}_2 = \bar{C}^*_{\cdot 2} - \bar{C}^*_{\cdot 3} = -0,10 \; ;$$

$$\hat{\beta}_3 = \bar{C}^*_{1\cdot} - \bar{C}^*_{2\cdot} = 0.10$$

Resultatet lar seg lett generalisere til (rxs) tabeller.
Det er to ting en må merke seg:

(i) Det finnes mange andre lineærkombinasjoner av gruppegjennomsnittene i tabell 4 som er forventningsrette estimatorer for koeffisientene i modell (2.1). Hvis en f.eks. satte $\bar{C}^*_{\cdot j}$ lik det vanlige uveide gjennomsnittet av cellegjennomsnittene i kolonne j i tabell 4, ville det fortsatt gjelde at

$$E(\bar{C}^*_{\cdot j} - \bar{C}^*_{\cdot 3}) = \beta_j, \quad (j=1,2).$$

På lignende måte kan det vises at

$$E(\bar{C}^*_{1\cdot} - \bar{C}^*_{2\cdot}) = \beta_3$$

når $\bar{C}^*_{i\cdot}$ betegner det uveide gjennomsnitt av cellegjennomsnittene på linje i, i = 1,.2

(ii) Det vanligste er å bruke minste kvadraters metode når en skal estimere parametrene i modell (2.1). Disse er optimale når variansen til C_i er konstant for gitte verdier på de uavhengige variablene.

La oss se litt på forutsetningen om at modellen (2.1) er additiv uten samspill. Dette er kanskje den viktigste begrensningen med å bruke standardisering i stedet for å bruke et vanlig regresjonsanalyse program. Dersom en f.eks. hadde følgende modell

$$E(C_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_4 X_{1i} Y_{1i} + \beta_5 X_{2i} Y_{1i} \; , \qquad (2.2)$$

hvor de siste to ledd gir uttrykk for at virkningen av variabel B på variabel C avhenger av verdien på variabel A. I denne modellen vil ikke forskjellene mellom de standardiserte gjennomsnittene være forventningsrette estimatorer for β_1, β_2 og β_3. En vil derimot finne at f.eks.

$$E(\bar{C}^*_{1\cdot} - \bar{C}^*_{2\cdot}) = \beta_3 + \frac{n_{\cdot 1}}{n}\beta_4 + \frac{n_{\cdot 2}}{n}\beta_5$$

Dersom en gjennom standardisering skal få tak i variabel B's "rene" effekt må en forutsette at modellen er additiv uten samspill.

Bruker en et regresjonsanalyse program er det ingen problemer med å estimere samtlige påraметre i (2.4), og en kan få testet om f.eks. koeffisientene β_4 og β_5 er signifikant forskjellig fra null, men dette forutsetter at en har adgang til samtlige data, og ikke bare tabeller som tabell 4.

Som konklusjon på dette avsnittet kan en altså si at en ved standardisering kan få fram forventningsrette estimatorer for parametrene i additive regresjonsmodeller uten samspill. Vi kan foreløpig ikke påstå at standardisering er den enkleste metoden, eller at den har noen optimale egenskaper. I neste kapittel skal vi vise at standardisering har spesielle fordeler når en i tillegg til modell (2.1) har en lineær sammenheng mellom de uavhengige variable, altså en sti-modell, og ønsker å estimere de direkte og indirekte effekter i en slik modell.

2.2. STANDARDISERING I FLERVEISTABELLER

Resultatene foran lar seg lett generalisere til flerveistabeller. Vi skal antyde framgangsmåten ved å innføre enda en variabel, D, i tabell 4. Vi tenker oss at D antar to verdier, og at tallene i tabell 4 spaltes opp etter variabel D.

Tabell 6.

			VARIABEL A			
			1	2	3	
VARIABEL	VARIABEL					
D = 1	B	1	0.13 (6)	0.23 (30)	0.32 (320)	(356)
		2	0.04 (80)	0.13 (400)	0.23 (6)	(486)
D = 2	B	1	0.25 (9)	0.35 (40)	0.45 (500)	(549)
		2	0.15 (100)	0.25 (600)	0.35 (9)	(709)
						(2 100)

Vi innfører nå enda en variabel, og setter opp en additiv modell uten samspill for sammenhengen mellom variabel C og variablene A, B og D.

La

$$Z_{1i} = \begin{cases} 1 \text{ hvis } D = 1 \\ 0 \text{ ellers} \end{cases}$$

Dessuten forutsetter vi følgende modell

$$E(C_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_4 Z_{1i} + \beta_0 \qquad (2.3)$$

På samme måten som for modell (2.1) er det mange lineærkombinasjoner av cellegjennomsnittene som er forventningsrette estimatorer for parametrene i modell (2.3), men av grunner som skal gis i neste kapittel, skal vi her konsentrere oppmerksomheten omkring de standardiserte gjennomsnitt og forskjellen mellom dem som estimatorer for koeffisientene i modell (2.3).

Vi finner da den enkle regel at hvis en ønsker å estimere effekten av en variabel, kan en standardisere med hensyn på de to andre, og bruke differansen mellom de standardiserte gjennomsnittene som estimatorer for koeffisientene i (2.3). For å skrive regelen ut litt mer detaljert, skal vi ta i bruk en del nye symboler.

La $\bar{C}^*_{i(BD)}$ være det standardiserte gjennomsnitt en får ved å ta for seg alle gjennomsnitt i tabell 6 for hvilke variabel A antar verdien i, og standardisere disse for variablene B og D (i = 1, 2, 3)

Fra tabell 6 finner en at

$$\bar{C}^*_{1(BD)} = \frac{356}{2100}\,0.13 + \frac{486}{2100}\,0.04 + \frac{549}{2100}\,0.25 + \frac{709}{2100}\,0.15 = 0.151$$

$$\bar{C}^*_{2(BD)} = \frac{356}{2100}\,0.23 + \frac{486}{2100}\,0.13 + \frac{549}{2100}\,0.35 + \frac{709}{2100}\,0.25 = 0.251$$

$$\bar{C}^*_{3(BD)} = 0.35$$

La dessuten $\bar{C}^*_{j(AD)}$ betegne det standardiserte gjennomsnittet en får når variabel B er lik j og det standardiseres med hensyn til variablene A og D. (j = 1, 2)

Fra tabell 6 finner en at

$$\bar{C}^{*}_{1(AD)} = \frac{86}{2100}\,0.13 + \frac{430}{2100}\,0.23 + \frac{326}{2100}\,0.32 + \frac{109}{2100}\,0.25 + \frac{640}{2100}\,0.35 +$$

$$\frac{509}{2100}\,0.45 = 0.32 \;;$$

$$\bar{C}^{*}_{2(AD)} = 0.22$$

La til slutt $\bar{C}^{*}_{K(AB)}$ betegne det standardiserte gjennomsnittet en får når variabel D settes lik K og det standardiseres med hensyn til variablene A og B (K = 1, 2).

$$\bar{C}^{*}_{1(AB)} = \frac{15}{2100}\,0.13 + \frac{70}{2100}\,0.23 + \frac{820}{2100}\,0.32 + \frac{180}{2100}\,0.04 + \frac{1000}{2100}\,0.13$$

$$= 0.201 \,,$$

$$\bar{C}^{*}_{2(AB)} = 0.32$$

Det kan da vises at

$$E(\bar{C}^{*}_{1(BD)} - \bar{C}^{*}_{3(BD)}) = \beta_1 \,,$$

$$E(\bar{C}^{*}_{2(BD)} - \bar{C}^{*}_{3(BD)}) = \beta_2 \,,$$

$$E(\bar{C}^{*}_{1(AD)} - \bar{C}^{*}_{2(AD)}) = \beta_3$$

og

$$E(\bar{C}^{*}_{1(AB)} - \bar{C}^{*}_{2(AB)}) = \beta_4 .$$

I tabell 6 finner en følgende estimater for parametrene i modell (2.3).

$\hat{\beta}_1 = -0.20$, $\hat{\beta}_2 = -0.10$, $\hat{\beta}_3 = 0{,}10$ og $\hat{\beta}_4 = -0.12$

2.3. Eksempler

2.3.1 Standardisering brukt på data fra Fruktbarhetsundersøkelsen

I det første eksempel skal vi se på data fra den norske fruktbarhetsundersøkelsen. I likhet med analysen som er gjort i Pullum (1978) skal vi se på et forsøk på å estimere effekten av utdanning på fruktbarhet på grunnlag av data gitt i tabell 7 nedenfor, hvor gjennomsnittlig antall levende fødsler er gitt for tre aldersklasser og syv utdanningskategorier.

Tabell 7. Gjennomsnittlig antall levende fødte etter mors utdanning og alder

Alder	Utdanningsnivå							Total	Antall observasjoner
	1	2	3	4	5	6	7		
18-24	1,5	1,2	0,8	0,6	0,2	0,2	0,1	0,5	339
25-34	2,6	2,1	1,8	2,0	1,3	1,7	1,5	1,9	667
35-44	3,2	2,9	2,6	2,5	2,5	2,3	2,5	2,7	472
Total	3,0	2,3	1,2	1,8	0,9	1,4	1,7	1,8	1 478
Antall observasjoner ...	121	189	98	678	97	94	201		

I tabell 7 ser det ut som om en additiv modell er rimelig. Vi skal derfor definere følgende variable, og sette opp en lineær, additiv modell for sammenhengen mellom alder, utdanning og fruktbarhet.

Vi har en (3 x 7) tabell, og trenger derfor å definere (2 + 6) "dummy-variable".

La

$$X_{1i} = \begin{cases} 1 & \text{hvis alder er mellom 18 og 24 år} \\ 0 & \text{ellers} \end{cases}$$

$$X_{2i} = \begin{cases} 1 & \text{hvis alder er mellom 25 og 34 år} \\ 0 & \text{ellers} \end{cases}$$

$$Y_{1i} = \begin{cases} 1 & \text{hvis utdanningsnivå er 1} \\ 0 & \text{ellers} \end{cases}$$

$$Y_{2i} = \begin{cases} 1 & \text{hvis utdanningsnivå er 2} \\ 0 & \text{ellers} \end{cases}$$

$$\vdots$$

$$Y_{6i} = \begin{cases} 1 & \text{hvis utdanningsnivå er 6} \\ 0 & \text{ellers.} \end{cases}$$

F_i er antall levende fødsler for kvinne i. Vi setter opp følgende modell for fruktbarhetens avhengighet av alder og utdanning:

$$E(F_i) = \beta_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \beta_1 Y_{1i} + \beta_2 Y_{2i} + \beta_3 Y_{3i} + \beta_4 Y_{4i} + \beta_5 Y_{5i} + \beta_6 Y_{6i}.$$

På samme måten som overfor kan f.eks. β_3 estimeres ved å regne ut de standardiserte gjennomsnittene i kolonne 3 og 7 i tabell 7, og subtrahere.

Dette gir følgende resultat

$$\beta_3 = \frac{339}{1478} 0.8 + \frac{667}{1478} 1.8 + \frac{472}{1478} 2.6 - (\frac{339}{1478} 0.1 + \frac{667}{1478} 1.5 + \frac{472}{1478} 2.5)$$

$$= 0.3276$$

De øvrige βene kan estimeres på samme måten ved å subtrahere det standardiserte gjennomsnittet i kolonne 7 i tabell 7 fra de standardiserte gjennomsnittene i de andre kolonner. En får følgende resultat

$$\hat{\beta}_1 = 1.0407; \hat{\beta}_2 = 0.6506; \hat{\beta}_3 = 0.3276; \hat{\beta}_4 = 0.3352; \hat{\beta}_5 = -0.0673;$$

$$\hat{\beta}_6 = 0.0491.$$

For å estimere α_1 regnes ut det standardiserte gjennomsnittet på første linje i tabell 7 og fra dette tallet subtraheres det standardiserte gjennomsnittet i linje tre i tabell 7.

$$\alpha_1 = \frac{121}{1478} 1.5 + \frac{189}{1478} 1.2 + \frac{98}{1478} 0.8 + \frac{678}{1478} 0.6 + \frac{97}{1478} 0.2 + \frac{94}{1478} 0.2$$

$$+ \frac{201}{1478} 0.1 - (\frac{121}{1478}3.2+\frac{189}{1478}2.9+\frac{98}{1478}2.6+\frac{678}{1478}2.5+\frac{97}{1478}2.5+\frac{94}{1478}2.3+\frac{201}{1478}0.7) =$$

$$-1.96.$$

På tilsvarende måte estimeres α_2 ved $\alpha_2 = -0,6849$.

Minste kvadraters estimater er

$$\tilde{\beta}_1 = 1.08; \tilde{\beta}_2 = 0.61; \tilde{\beta}_3 = 0.44; \tilde{\beta}_4 = 0.35; \tilde{\beta}_5 = -0.11; \tilde{\beta}_6 = -0.05;$$

$$\tilde{\alpha}_1 = -1.96; \tilde{\alpha}_2 = -0.68$$

I dette eksemplet gir de to estimeringsmetoder ganske like resultater. Fordelene ved å bruke et regresjonsprogram er likevel mange. Blant annet får en regnet ut usikkerhetene på koeffisientene, samt et mål for modellens tilpasning til data.

2.3.2 Eksempel på bruk av variansanalyse i stedet for standardisering

I eksemplet overfor var forutsetningen om additivitet åpenbart rimelig, og det viste seg da også at standardisering ga nesten samme estimater som minste kvadraters metode. I mange praktiske tilfeller er det vanskelig å avgjøre om en med rimelighet kan forutsette en additiv modell. I tabell 8 er gjennomsnittlig antall sykedager gitt for fire aldersgrupper og to geografiske regioner. Data er hentet fra Levekårsundersøkelsen 1973, og eksemplet er inspirert av Johansson (1978), som i utstrakt grad bruker standardisering for å estimere og teste effekten av å bo i et bestemt len i Sverige.

Tabell 8. Gjennomsnittlig antall sykedager etter alder og bostedstype

	17 - 24 år	25 - 49 år	50 - 66 år	67 år og over
Kommuner med 50 000 innbyggere eller flere	10.27	6.20	16.45	27.98
Landet ellers	7.76	9.56	21.04	28.11
	(417)	(1190)	(893)	(466)

Det er klart at gjennomsnittlig antall sykedager avhenger av alder. Derimot er det mindre klart om bostedstype har samme effekt for alle aldersgrupper. I dette tilfeldet synes det derfor mer naturlig å utføre en variansanalyse for å teste om samspillet kan settes lik null, og deretter teste om geografisk region har noen signifikant effekt på gjennomsnittlig antall sykedager.

En slik analyse gir som resultat at det bare er alder som har signifikant effekt på sykeligheten.

Dersom en likevel velger å utføre standardisering for å estimere effekten av å bo i en større kommune, får en at gjennomsnittlig antall sykedager for de store kommuner ligger 2.4 dager under gjennomsnittet for landet ellers. I forhold til tallene i tabell 8 er 2.4 dager en ikke ubetydelig tidsperiode, men variansanalysen opplyser oss altså om at dette tallet ikke er signifikant forskjellig fra null.

2.4. Transformasjon for å redusere samspillet

Ofte vil forutsetningen om at samspillet er null ikke være oppfylt, men en kan da i noen tilfeller transformere data, for derved å fjerne, eller redusere, samspillet. Slike teknikker er særlig viktig i forbindelse med analyse av hyppighetstabeller, noe vi skal vende tilbake til i kap. 4.

Vi skal se på enda et konstruert eksempel:

I tabell 9 er gjennomsnittet av en ny variabel C gitt for forskjellige verdier av variablene A og B.

Tabell 9.

		Variabel A			
		1	2	3	
Variabel B	1	0.20 (15)	0.30 (70)	0.40 (15)	0.3 (100)
	2	0.10 (180)	0.15 (1 000)	0.20 (820)	 (2 000)
		0.11 (195)	0.16 (1 070)	0.20 (835)	0.17 (2 100)

I tabell 9 synes det ikke som om effekten av variabel B er uavhengig av verdien på variabel A, og en additiv modell synes derfor av liten verdi.

Derimot vil en se i tabell 9 at det gjelder at

$$\frac{\bar{C}_{11}}{\bar{C}_{21}} = \frac{\bar{C}_{12}}{\bar{C}_{22}} = \frac{\bar{C}_{13}}{\bar{C}_{23}} = 2$$

Det vil si at

$$\ln\bar{C}_{11} - \ln\bar{C}_{21} = \ln\bar{C}_{21} - \ln\bar{C}_{22} = \ln\bar{C}_{13} - \ln\bar{C}_{23} = \ln 2$$

Hvis vi altså definerer

$$\bar{C}^{*}_{ij} = \ln\bar{C}_{ij},$$

gjelder det at modellen for $\bar{C}^{*}_{ij}$ er additiv. Hvis vi derfor definerer X_{1i}, X_{2i} og Y_{1i} som i kapittel 2, er følgende modell rimelig:

$$E(\bar{C}^{*}_{ij}) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_0,$$

og parametrene kan estimeres ved standardisering etter at variabel C er transformert.

I tabell 9b er gitt logaritmen til gjennomsnittene i tabell 9.

Tabell 9b

		Variabel A		
Variabel B	1	-1.609 (15)	-1.204 (70)	-0.916 (15)
	2	-2.303 (180)	-1.897 (1 000)	-1.609 (820)

Ved standardisering estimeres nå parametrene til

$\hat{\beta}_1 = -0.694$

$\hat{\beta}_2 = -0.288$

$\hat{\beta}_3 = +0.693$

$\hat{\beta}_0 = -1609$

Nå innvendes det ofte at det er vanskelig å tolke resultatene når data er transformert. Dette er riktig i mange tilfeller, men i dette tilfellet har vi funnet at β_3 estimeres til 0.694. Dvs. at logaritmen til gjennomsnittene øker med 0.693 fra linje 2 til linje 1 i tabell 9b. Dette vil igjen si at $\bar{C}_{1j}/\bar{C}_{2j} = e^{0.693} = 2$. Konklusjonen er altså at gjennomsnittet til variabel C når variabel B er 2 ligger 50% under gjennomsnittet når variabel B er lik 1.

3. KAUSALANALYSE VED HJELP AV STANDARDISERING

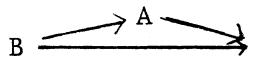
3.1. Kausalanalyse i toveis tabeller

La oss nå gå tilbake til tabell 4. Her ses det at innen hver kolonne øker gjennomsnittet med 0.10 når en går fra linje 2 til linje 1. Marginalt ses det likevel at gjennomsnittet øker med 0.20 når en går fra linje 2 til linje 1. Hva er forklaringen til dette? Det skyldes at når en har verdien 1 på variabel B er det en tendens til å ha en høyere verdi på variabel A enn når variabel B har verdien 2. Det er altså en sammenheng mellom variablene A og B. Hvis denne sammenhengen er en årsaks-virkning sammenheng, er det rimelig å modellere denne på samme måten som vi modellerte sammenhengen mellom A, B og C.

En type modell som ofte brukes i slike situasjoner, er da

$$\begin{aligned} E(C_i) &= \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 y_{1i} + \beta_0 \\ E(X_{1i}) &= \gamma_1 y_{1i} + \gamma_0 \\ E(X_{2i}) &= \delta_1 y_{1i} + \delta_0, \end{aligned} \tag{3.1}$$

hvor samtlige variable er definert på side 5. Det er viktig å merke seg at X_{1i} og X_{2i} tilsammen definerer verdien på variabel A, slik at sti-diagrammet for modell (3.1) ser ut som følger:



Modellen (3.1) er en rekursiv, additiv modell med en eksogen variabel, nemlig Y_1. Dvs. når Y_1 er gitt, og parametrene i (3.1) er kjente, kan forventningene til de øvrige variablene bestemmes ved å innsette i (3.1). Vanligvis brukes også i slike modeller minste kvadraters metode, men her skal vi vise at alle parametre kan estimeres ved standardisering. (I kapittel 4 skal vi rette en del kritikk mot modeller av typen $E(X_{1i}) = \gamma_1 Y_{1i} + \gamma_0$, når X_{1i} er en dummy-variabel, men i denne sammenhengen konstaterer vi bare at modeller av typen (3.1) er mye brukt, og av den grunn kan de følgende resultatene være nyttige.)

Ved i (3.1) å "innsette" de to siste likninger i den første fås

$$\begin{aligned} E(C_i) &= \beta_1\gamma_1 y_{1i} + \beta_2\delta_1 y_{1i} + \beta_3 y_{1i} + \beta_0 + \beta_1\gamma_0 + \beta_2\delta_0 \\ &= (\beta_3+\beta_1\gamma_1+\beta_2\delta_1) y_{1i} + \beta_0^*. \end{aligned} \tag{3.2}$$

I (3.2) kalles ofte β_3 for den <u>direkte effekt</u> av Y_1 på C, mens $(\beta_1\gamma_1+\beta_2\delta_1)$ kalles den <u>indirekte effekt</u> av Y_1 på C.

Vanligvis estimeres de indirekte og direkte effekter ved hjelp av minste kvadraters estimatorer for β_1, γ_1 og δ_1. Vi skal bruke følgende resultat:

La $\bar{C}^*_{i.}$ betegne det standardiserte gjennomsnitt på linje i i tabell 4, og la $\bar{C}_{i.}$ betegne det vanlige gjennomsnitt på linje i. Da gjelder det under modell (3.1) at estimatoren $\{(\bar{C}_{1.} - \bar{C}^*_{1.}) - (\bar{C}_{2.} - \bar{C}^*_{2.})\}$ har forventningen lik $\beta_1\gamma_1 + \beta_2\delta_2$.

Vi har altså funnet en forventningsrett estimator for den indirekte effekt i modellen. (Denne måten er ikke identisk med den måten som er brukt i Pullum (1978).) Den direkte effekt, β_3, estimeres som i kapitlet foran ved

$$\hat{\beta}_3 = (\bar{C}^{st}_{1.} - \bar{C}^{st}_{2.}).$$

I tabell 1 finner vi at

$$(\bar{C}_{1.} - \bar{C}^{st}_{1.}) - (\bar{C}_{2.} - \bar{C}^{st}_{2.}) = 0.1,$$

og tidligere har vi funnet at

$$\hat{\beta}_3 = 0.1.$$

Den observerte forskjell i marginalene ytterst i høyre i tabell 4 har vi gjennom modell (3.1) fått spaltet opp i en direkte effekt, som er forutsatt konstant i alle kolonner, samt en indirekte effekt som skyldes avhengigheten mellom variabel A og B.

I kapittel 2 understreket vi at å estimere koeffisientene i en regresjonsmodell ved hjelp av standardisering kun er en av mange mulige måter, og at den hverken er den enkleste, eller har noen optimalitetsegenskaper. I kausalmodeller har vi nå vist at en ved hjelp av de standardiserte gjennomsnittene og de vanlige marginale gjennomsnitt på en enkel måte kan estimere de direkte og indirekte effekter. Det er altså spesielt når en lager kausalanalyse på tabellerte data at standardisering peker seg ut som en i visse tilfeller fornuftig teknikk.

3.2. Kausalanalyse i flerveistabeller

I avsnitt 2.2 ble det gitt et eksempel på estimering av regresjonskoeffisientene i en lineær, additiv modell på grunnlag av data gitt i en (2x2x3) tabell. I dette avsnittet skal vi antyde hvordan en ved standardisering på en enkel måte kan oppdele effektene i indirekte og direkte effekter,

når modellen som beskriver sammenhengene mellom de fire variablene er en additiv, rekursiv modell.

La modellen for variablene i tabell 6 være

$$\text{(i)} \quad E(C_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 Y_{1i} + \beta_4 Z_{1i} + \beta_0$$

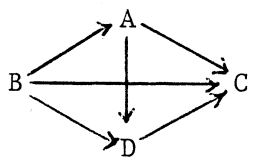
$$\text{(ii)} \quad E(Z_{1i}) = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 Y_{1i} + \gamma_0 \qquad (3.3)$$

$$\text{(iii)} \quad E(X_{1i}) = \delta_1 Y_{1i} + \delta_0$$

$$\text{(iv)} \quad E(X_{2i}) = \varepsilon_1 Y_{1i} + \varepsilon_0,$$

hvor alle variable er definert i avsnitt 2.1 og 2.2.

Ofte ser en modellen (3.3) beskrevet ved følgende stidiagram:



Som i avsnitt 3.1 "innsettes" nå (iv), (iii) og (ii) i (i), og en får:

$$E(C_i) = \beta_1\delta_1 Y_{1i} + \beta_2\varepsilon_1 Y_{1i} + \beta_3 Y_{1i} + (\beta_4\gamma_1\delta_1 + \beta_4\gamma_2\varepsilon_1 + \beta_4\gamma_3)Y_{1i} + \beta_0.$$

Koeffisientene foran den eksogene variablelen Y_{1i} kan nå oppdeles i direkte og indirekte effekter på følgende måte:

(i) Den direkte effekt: β_3

(ii) De indirekte effekter: I: via variabel D: $\beta_4\gamma_3$

II: via variabel A: $(\beta_1\delta_1+\beta_2\varepsilon_1)$ +

$(\beta_4\gamma_1\delta_1+\beta_4\gamma_2\varepsilon_1)$.

Den indirekte effekt via A kan oppdeles ytterligere i to:

Bare via variabel A: $\beta_1\delta_1 + \beta_2\varepsilon_1$

Via A og D: $\beta_4\gamma_1\delta_1 + \beta_4\gamma_2\varepsilon_1$.

Vanligvis estimeres samtlige koeffisienter i modellen (3.3) ved å utføre vanlig regresjonsanalyse på hver av ligningene.

Når data er gitt på tabellform kan en bruke standardisering, og må da gå fram på følgende måte:

Den direkte effekt av variabel B på variabel C_1,β_3, fås ved å standardisere med hensyn til de to andre variableme A og D. Vi fant i avsnitt 2.2 at

$$\hat{\beta}_3 = (\bar{C}^*_{1(AD)} - \bar{C}^*_{2(AD)})$$

er en forventningsrett estimator for β_3.

I tabell 6 får en at $\hat{\beta}_3 = \bar{C}^*_{1(AD)} - \bar{C}^*_{2(AD)} = 0.10$

Hvis en ønsker å estimere de indirekte effekter kan en i (3.3) "innsette" (ii) i (i).

$$\text{(i)} \quad E(C_i) = (\beta_1 + \beta_4\gamma_1)X_{1i} + (\beta_2 + \beta_4\gamma_2)X_{2i} + (\beta_3 + \beta_4\gamma_3)Y_{1i} + \beta_0 \qquad (3.4)$$

$$\text{(ii)} \quad E(X_{1i}) = \delta_1 Y_{1i} + \delta_0$$

$$\text{(iii)} \quad E(X_{2i}) = \varepsilon_1 Y_{1i} + \varepsilon_0$$

Koeffisientene i ligningssystemet (3.4) kan nå estimeres på samme måten som i avsnitt 3.1, ved at det på grunnlag av tabell 6 lages en ny tabell ved å slå sammen etter variabel D, og deretter standardisere for variabel A. I vårt eksempel fås tabell 4 når en i tabell 6 slår sammen etter variabel D. Vi har tidligere estimert koeffisienten foran variabel Y_{1i} i (i) til å være 0,1. Av (3.4) ser en at denne koeffisienten nå er en sum av to effekter, nemlig den direkte effekt av variabel B på C og den indirekte effekt via variabel D. Ovenfor estimerte vi den direkte effekt av varibel B på C til å være 0.1. Den estimerte indirekte effekt via variabel D blir altså 0,0. Variabel D's direkte effekt på variabel C er tidligere estimert til å være 0.12.

For å estimere den indirekte effekt via variabel A, trenger en å finne den indirekte effekt i ligningssystemet (3.4). Ved å "innsette" (ii) og (iii) i (i) i (3.4) viser det seg at den indirekte effekt er lik $(\beta_1\delta_1 + \beta_2\varepsilon_1) + (\beta_4\gamma_1\delta_1 + \beta_4\gamma_2\varepsilon_1)$, som vi tidligere har definert som den indirekte effekt via variabel A. I eksemplet estimeres den indirekte effekt via variabel A til å være 0.1.

I tabell 6 har vi dermed delt opp variabel B's effekt på variabel C i direkte og indirekte effekter ved hjelp av standardisering.

4. STANDARDISERING I FORBINDELSE MED ANALYSE AV HYPPIGHETSTABELLER

4.1. Innledning

I kapitlene foran har vi vurdert bruken av standardisering ved analyse av tabeller som viser gjennomsnitt av en variabel innen grupper etter en eller flere variable. I dette kapitlet skal vi se på standardisering brukt i forbindelse med analyse av hyppighetstabeller, og vi skal sammenligne denne framgangsmåten med bruken av log-lineære modeller, som i de senere årene er blitt blant de vanligst brukte analyseteknikker.

Hele diskusjonen i dette avsnittet knyttes til en analyse utført av Hellevik (1978, 1979). Det viser seg at kausalanalysen utført i Hellevik (1978) er identisk med standardisering, og derfor tolkbar innen et sett med lineære modeller uten samspill for hyppighetene. Deretter sammenlignes denne analyseteknikk med log-lineær analyse, og det viser seg at sistnevnte teknikk er nyttig når en har samspill.

Alt i alt må konklusjonen på dette avsnittet bli at standardisering ikke løser noen problemer som ikke kan løses ved hjelp av log-lineær analyse av hyppighetstabeller, snarere tvert imot.

4.2. Tolking av standardisering under en lineær modell for hyppighetene

Følgende tabell som viser samvariasjonen mellom variablene "egen stemmegiving", "fars yrke" og "fars stemmegiving", er hentet fra Hellevik (1978).

Tabell 10. Andelen som stemte sosialistisk i 1957 etter fars yrke og stemmegiving. Prosent.

	Fars yrke		Antall
	Arbeider	Ikke arbeider	spurt
Fars stemmegiving:			
Sosialistisk	85	69	224
Ikke sosialistisk	46	28	397
I alt	70	34	
Antall spurte	278	343	621

I tabell 10 ser det ut som om andelen som stemmer sosialistisk varierer både med "fars yrke" og "fars stemmegiving". Dessuten ser det ut som om de to variablene har en additiv effekt på stemmegivingen, og at det ikke er noe samspillseffekter.

La oss derfor definere følgende variable:

La

$$FY_i = \begin{cases} 1 & \text{hvis fars yrke er arbeider} \\ 0 & \text{ellers} \end{cases}$$

$$FS_i = \begin{cases} 1 & \text{hvis far stemmer sosialistisk} \\ 0 & \text{ellers} \end{cases}$$

$$S_i = \begin{cases} 1 & \text{hvis personen selv stemmer sosialistisk} \\ 0 & \text{ellers} \end{cases}$$

En mulig modell er da

$$E(S_i) = \beta_0 + \beta_1 FY_i + \beta_2 FS_i. \qquad (4.1)$$

Modellen (4.1) er av samme type som (2.1) med den forskjell at den avhengige variabelen, S_i, er dikotom.

I Hellevik (1978) er en også opptatt av sammenhengen mellom "fars yrke" og "fars stemmegiving". En mulig modell som er konsistent med Helleviks analyse, er følgende

$$E(FS_i) = \gamma_1 FY_i + \gamma_0. \qquad (4.2)$$

Sammen utgjør (4.1) og (4.2) en rekursiv, lineær modell, men en eksogen variabel, "fars yrke".

Bortsett fra at alle variable i (4.1) og (4.2) er dikotome, er denne modellen identisk med modellen (3.1) i kapittel 3 foran. Vi viste da at en ved standardisering kunne estimere alle parametre som inngår i modellen. Disse resultater gjelder naturligvis også for dikotome variable.

Ved å gå fram som i kapittel 3, kan den direkte effekt av "fars yrke", β_1, estimeres ved å regne ut den standardiserte andel sosialistiske stemmer gitt at fars yrke er "arbeider", og subtrahere fra dette tallet den standardiserte andel sosialistiske stemmer gitt at fars yrke er "ikke-arbeider".

En finner av tabell 10 at

$$\hat{\beta}_1 = 0.6 - 0.43 = 0.17.$$

For å estimere den indirekte effekt går vi også fram som i modell (3.1). Den indirekte effekt fås som en differens mellom to tall, som er forskjellen mellom de standardiserte andel sosialistiske stemmer og de ustandardiserte, når det er gitt at fars yrke er "arbeider" og "ikke-arbeider" henholdsvis.

Den indirekte effekt blir da estimert til

$$(0.70-0.60) - (0.34-0.43) = 0.19.$$

Disse resultater er identiske med resultatene i Hellevik (1978), og ved å se litt nøyere på metoden i Hellevik (1978) viser det seg at den er identisk med standardisering, og har altså en klar tolkning i modellen gitt ved (4.1) og (4.2).

Metoden er i dette eksemplet også identisk med minste kvadraters metode.

Eksemplet ovenfor viser, etter min mening, at en ofte kan få verdifull innsikt i samvariasjonen mellom dikotome variable ved hjelp av lineære modeller. Mer grundig behandling av dette synet er gitt i Amundsen (1974). Likevel er lineære modeller ofte kritisert i forbindelse med analyse av dikotome, avhengige variable, og det er blitt stadig mer vanlig å bruke andre typer modeller. Vi skal ikke her ta stilling til denne diskusjonen, men finner det naturlig å sammenligne standardisering med log-lineære modeller, og demonstrere teknikken på de data som er brukt i tabell 10.

4.3. Standardisering sammenlignet med log-lineære modeller

Sammenligningene mellom standardisering og regresjonsanalyse som er gjort ovenfor viser at det er en relativ enkel sammenheng mellom de to teknikkene når det ikke er samspill. Helt så enkel er ikke dette tilfellet når vi nå skal sammenligne standardisering med log-lineære modeller. Formelt er sammenhengen helt uoversiktlig, men i praksis vil det likevel vise seg at de to angrepsmetoder fører til lignende resultater.

La p_i være sannsynligheten for at individ i skal stemme sosialistisk. Da kan logit modellen uten samspill skrives som

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 FS_i + \beta_2 FY_i, \qquad (4.3)$$

hvor variablene FY_i og FS_i er definert overfor.

Modellen (4.3) er svært lik modellen (2.1). Ved å transformere sannsynlighetene til $\ln (p_i/1-p_i)$, ofte kalt log-odds, har vi fått modeller med kategoriske, avhengige variable, som er nesten identiske med modellene for variable på intervallnivå. Hvis vi altså transformerer tallene i tabell 10 til log-odds og deretter standardiserer, kan vi estimere parametrene i (4.3). I tabell 10a er de relative hyppigheter i tabell 10 transformert til log-odds.

Tabell 10a. Log-odds til andelen som stemte sosialistisk i 1957 etter fars yrke og stemmegiving

	Fars yrke		Antall
	Arbeider	Ikke arbeider	spurte
Fars stemmegiving:			
Sosialistisk	1.7345	0.7998	224
Ikke sosialistisk	-0.1613	-0.9467	397
Antall spurte	278	343	621

Tallene i tabell 10a tyder på at en kan sette samspillet lik null. Vi kan nå estimere parametrene i modell (4.3) på samme måten som i kapitlene 2 og 3.

Parametrene i modell (4.3) estimeres ved standardisering til

$$\hat{\beta}_1 = 1.811$$

$$\hat{\beta}_2 = 0.844.$$

Vi kan altså estimere parametrene i modellene (4.3) ved hjelp av standardisering. Den vanligste metoden for estimering i modeller av typen (4.3) er likevel å bruke maximum likelihood prinsippet, Fienberg (1978). I eksemplet ovenfor kan en bruke programmet ECTA til å tilpasse en log-lineær modell til data i tabell 10. Det viser seg da at følgende log-lineære modell gir god tilpassing til data:

$$\ln(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(i,j)} + u_{13(i,k)} + u_{23(j,k)}, \qquad (4.4)$$

hvor m_{ijk} er det forventede antall observasjoner i celle (i, j, k) og vi har de vanlige betingelsene på parametrene.

Den log-lineære modellen (4.4) svarer til modellen (4.3), og vi har altså fått testet om modellen er "god". Dessuten ble parametrene i (4.4) estimert til:

$$u = 4.025;\ u_{1(i)} = 0.191;\ u_{2(j)} = 0.019;\ u_{3(k)} = -0.391;$$

$$u_{12(i,j)} = 0.198;\ u_{13(i,k)} = 0.467;\ u_{23(j,k)} = 0.478.$$

Nå er det lett å vise følgende relasjoner mellom parametrene i modellene (4.3) og (4.4):

$$\beta_1 = {}^{4}u_{13(i,k)}$$

$$\beta_2 = {}^{4}u_{12(i,j)}$$

Maximum likelihood estimatene for β_1 og β_2 blir da

$$\tilde{\beta}_1 = 1.868 \text{ og } \tilde{\beta}_2 = 0.792 ,$$

som stemmer godt med de andre resultatene vi fant ovenfor. Fordelene ved å bruke ECTA er at vi i tillegg til å estimere parametrene i modellen også får testet om det er rimelig å sette samspillet lik null.

Konklusjonen må igjen bli at de problemer en kan studere ved hjelp av standardisering, eller lignende teknikker, lar seg bedre studere ved andre metoder, f.eks. log-lineær analyse. Log-lineær analyse utføres på tabellerte data slik at en av hensiktene med å bruke standardisering faller bort.

6. REFERANSER

Ahlbom, A (1980). Standardisering - en metod att öka jämnførbarheten. Statistisk tidsskrift. Stockholm.

Amundsen, H.T. (1974). Binary Variable Mulitple Regressions. Scand.J.Statist. 1 : 59-70.

Fienberg, S.E. (1978). Tha Analysis of Cross-Classified Categorical Data. MIT Press.

Gørtz, E. og Hansen, J.D. (1977). Indeks teori. Odense Universitetsforlag.

Hellevik, O. (1978). Kausal analyse ved hjelp av multivariate prosenttabeller. Tidsskrift for samfunnsforskning vol. 19.

Hellevik, O. (1979). Causal analysis of non-metric data. Politico 33, Institutt for statsvitenskap, Universitetet i Oslo.

Hoem, J.M. (1979). Statistical Analysis of a Multiplicative Model and its Application to the Standardization of Vital Rates. Working Paper No. 21, Laboratory of Actuarial Mathematics, University of Copenhagen.

Johansson, S. (1978). Medborgerrapporter. Institutet for Social Forskning, Stockholm

Little, R and Pullum, T.W. (1979). The General Linear Model and Direct Standardization: A comparison. Occasional Papers No. 20. World Fertility Survey, London.

Pullum, T.W. (1978). Standardization. Technical Bulletins, World Fertility Survey, London.

Selén, J. (1979). Artikkel om tvillingsansatsen i "Tre bidrag til välfärdsmätningarnas metodik". Institutet for Social Forskning. Stockholm.

Thomsen, I. (1973). A note on the efficienly of weighting subclass means to reduce the effects of non-response when analyzing survey data. Statistisk tidsskrift. Stockholm.