Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Dep, Oslo 1. Tlf.*(02) 41 38 20

IO 77/46

22. desember 1977

OM LOG-LINEÆR ANALYSE AV FLERVEISTABELLER

av

Tor Haldorsen*

INNHOLD

* Takk til Petter Laake og Rolf Aaberge for kommentarer til manuskriptet.

1. INNLEDNING

Notatet inneholder en del momenter om log-lineær analyse av flerveistabeller. Data til et gjennomgangseksempel er tatt fra helseundersøkelsen 1975 og er om tannlegekontakter.

Flerveistabeller har ofte mange celler og det er lett å miste oversikten. Noen klarer å oppøve imponerende ferdigheter i å lese og finne fram til mønstre i slike tabeller, men analyse som er basert på direkte lesing av tabeller, vil ofte ha svakheter. Når data er beheftet med usikkerhet, vil det være vanskelig å få tatt skikkelig hensyn til dette. Videre kan det oppstå uklarheter og inkonsekvenser fordi en ikke bruker klart definerte metoder for å måle sammenheng og for å sammenligne ulike grupper. Disse problemene kan løses ved å bruke en stokastisk modell for tabellen. Siden modellen er stokastisk, vil usikkerhetsstrukturen være fastlagt og sammenheng og sammenligninger kan defineres ved hjelp av parametrene i modellen. Det er flere typer av modeller å velge blant. I dette notatet behandles en av dem.

Log-lineære modeller gir oss et system for å beskrive og finne fram til strukturen i en flerveistabell. På en forholdsvis naturlig og oversiktlig måte avbildes de kompliserte former for avhengighet/uavhengighet det kan være mellom to eller flere variable i en flerveistabell. Beregningene som kreves, kan utføres av programmet ECTA som er lagt inn på Byråets regneanlegg.

I kapittel 2 har vi en kort omtale av modellene. For tilfellet med tre variable belyser vi tolkningen av parametrene og nevner noen av de strukturer som kan beskrives ved modellene. Videre ser vi på observatoren som vi vil bruke for å teste ulike hypoteser. I kapittel 3 presenteres data og vi viser hvordan én spesiell hypotese om tabellen kan testes. Videre behandler vi situasjonen når vi vil bruke data til å lete etter en modell. Vi fortsetter letingen etter modell i kapittel 4 og drøfter da ulike trinnvise prosedyrer. I kapittel 5 tolkes den endelige modell ved hjelp av parametrene i modellen og ved alternative metoder. Hvis data består av stratifiserte utvalg og/eller en velger en spesiell betraktningsmåte for sine variable, er det aktuelt å modifisere den generelle metode. Vi behandler dette i kapittel 6. Kapittel 7 inneholder noen momenter om å sløyfe en eller flere variable i analysen.

Stoff om log-lineære modeller finnes etterhvert i flere lærebøker. Vi har brukt Bishop, Fienberg and Holland (1975) som referanse. Boken krever ikke store forkunnskaper i matematisk statistikk og inneholder mange eksempler.

2. LOG-LINEÆRE MODELLER

2a. Parametrene i modellen

Vi skal i første omgang se på parametrene i en modell med 3 variable, A, B og C. Vi antar at variablene er målt på nominalnivå og har h.h.v. I, J og K mulige verdier. Uavhengig av hverandre er n enheter krossklassifisert m.h.t. A, B og C. Vi tar utgangspunkt i en vanlig multinomisk modell for denne situasjonen. Parametre er p_{ijk} der $i = 1, \ldots, I$, $j = 1, \ldots, J$ og $k = 1, \ldots, K$, p_{ijk} står for sannsynligheten for at en vilkårlig enhet faller i celle (i, j, k). Videre setter vi X_{ijk} for antallet og $m_{ijk} = n \cdot p_{ijk}$ for forventet antall av de n som faller i celle (i, j, k). Situasjoner der ikke alle m_{ijk} større enn null og ukjente, dekkes ikke av notatet. Slike modeller behandles i kapittel 5 i Bishop et.al. (1975).

Når en vil uttrykke ulike former for sammenheng mellom A, B og C, er det nokså komplisert å bruke p_{ijk}-ene. Et alternativ er da å foreta en reformulering og bruke log-lineære modeller.

Da settes

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC} \qquad (2,1)$$

der

$$\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = 0,\ \sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = 0,\ \sum_i u_{ik}^{AC} = \sum_k u_{jk}^{AC} = 0,\ \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0$$

og

$$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0. \qquad (2,2)$$

Log x står for den naturlige logaritme til x.

Siden $\log(m_{ijk}) = \log(n \cdot p_{ijk}) = \log n + \log p_{ijk}$ er det bare en konstant som skiller denne modellen fra modellen for p_{ijk}-ene i Haldorsen (1976). Ved å ta utgangspunkt i m_{ijk} oppnår vi en mer generell formulering som også dekker andre situasjoner enn den rent multinomiske, som vi for enkelhets skyld har tatt utgangspunkt i. (Mer om dette i 2.2.3, 2.3.4, 2.4.2, og 3.2 i Bishop et.al. (1975).)

u-ene har fotskrifter. For ulike i-er vil u_i^A være elementene i en vektor u^A, for ulike (i, j)-er vil u_{ij}^{AB} være elementene i en matrise u^{AB} osv.

u-ene i (2.1) og (2.2) kan vi betrakte på flere måter. De med minst to toppskrifter vil vi av og til se på som mål for sammenhengen mellom variable. Vi kan også bruke u-ene til å "forklare" cellefrekvensene i en tabell, da vil de bli kalt effekter. u uten toppskrift angir en gjennomsnittseffekt. u-ene med en toppskrift tolkes som hovedeffekter av variable, f.eks. u_j^B angir hovedeffekt av at variabel B har verdi j. u-ene med to toppskrifter tolkes som to-faktoreffekter (samspill), f.eks. u_{ik}^{AC} angir effekten på cellefrekvensen at variabel A har verdi i <u>samtidig</u> som variabel C har verdi k. u-ene med tre toppskrifter tolkes som trefaktoreffekter (samspill av 2. orden), f.eks. u_{ijk}^{ABC} angir effekten av at variable A, B og C samtidig antar verdiene i, j og k. (I modeller for flere variable enn tre, vil vi ha u-er som gir firefaktoreffekter osv. alt etter hvor mange variable vi har.)

Vi skal begrunne tolkningene noe nærmere. u-ene kan uttrykkes som funksjoner av $\log m_{ijk}$-ene. Når vi regner ut fra (2.1) og (2.2), får vi bl.a.

$$u = \frac{1}{I \cdot J \cdot K} \sum_{i,j,k} \log m_{ijk}$$

$$u_j^B = \frac{1}{I \cdot K} \sum_{i,k} \log m_{ijk} - \frac{1}{IJK} \sum_{i,j,k} \log m_{ijk} = \frac{1}{I \cdot K} \sum_{i,k} \log m_{ijk} - u$$

$$u_{ik}^{AC} = \frac{1}{J} \sum_j \log m_{ijk} - \frac{1}{IJ} \sum_{i,j} \log m_{ijk} - \frac{1}{J \cdot K} \sum_{j,k} \log m_{ijk} + \frac{1}{IJK} \sum_{i,j,k} \log m_{ijk}$$

$$= \frac{1}{J} \sum_j \log m_{ijk} - u_i^A - u_k^C - u$$

$$u_{ijk}^{ABC} = \log m_{ijk} - \frac{1}{I} \sum_i \log m_{ijk} - \frac{1}{J} \sum_j \log m_{ijk} - \frac{1}{K} \sum_k \log m_{ijk}$$

$$+ \frac{1}{I \cdot J} \sum_{i,j} \log m_{ijk} + \frac{1}{I \cdot K} \sum_{i,k} \log m_{ijk} + \frac{1}{J \cdot K} \sum_{j,k} \log m_{ijk} - \frac{1}{I \cdot J \cdot K} \sum_{i,j,k} \log m_{ijk}$$

$$= \log m_{ijk} - u_{ij}^{AB} - u_{ik}^{AC} - u_{jk}^{BC} - u_i^A - u_j^B - u_k^C - u$$

Disse ligningene understøtter tolkningen av u-ene. Siden vi regner med logaritmen til det forventede antall i cellene (m_{ijk}), ser vi det hele i <u>logaritmisk</u> skala. Av ligningene ser vi at u, gjennomsnittseffekten, er gjennomsnittet for alle celler. u_j^B, hovedeffekten av at B = j, er gjennomsnittet i alle celler der B = j minus det som forklares av gjennomsnittseffekten u. u_{ik}^{AC}, samspillet mellom A og C når A = i og C = k, er gjennomsnittet i alle celler der A = i og C = k minus det som kan forklares ved lavere effekter u_i^A, u_k^C og u. Tilsvarende er u_{ijk}^{ABC} lik forventet antall i celle (i,j,k) minus det som forklares ved lavere effekter u_{ij}^{AB}, u_{ik}^{AC}, u_{jk}^{BC}, u_i^A, u_j^B, u_k^C og u.

Det fins også en annen måte å se på oppbyggingen av u-ene, som kan gi innsikt i hva som måles. Hvis vi splitter 3-veistabellen i K 2-veistabeller etter nivå på variabel C, kan vi sette opp en modell for hver av de K deltabellene. For C = k får vi

$$\log m_{ij}(k) = u\,(k) + u_i^A\,(k) + u_j^B\,(k) + u_{ij}^{AB}\,(k)$$

Mellom u-ene for 3-veistabellen og 2-veistabellene har vi følgende sammenhenger:

$$\begin{aligned}
u &= \frac{1}{K}\sum_k u\,(k) \\
u_i^A &= \frac{1}{K}\sum_k u_i^A\,(k) \\
u_{ij}^{AB} &= \frac{1}{K}\sum_k u_{ij}^{AB}\,(k) \qquad (2.3)\\
u_{ijk}^{ABC} &= u_{ij}^{AB}\,(k) - u_{ij}^{AB} = u_{ij}^{AB}\,(k) - \frac{1}{K}\sum_k u_{ij}^{AB}\,(k) \\
u_{ik}^{AC} &= u_i^A\,(k) - u_i^A = u_i^A\,(k) - \frac{1}{K}\sum_k u_i^A\,(k) \\
u_k^C &= u\,(k) - u
\end{aligned}$$

Av ligningene kan en bl.a. lese at u_{ijk}^{ABC} er samspillet mellom A og B i k-te deltabell minus det gjennomsnittlige samspill mellom A og B i deltabellene. En ser videre at u_{ik}^{AC} er hovedeffekt av A i k-te deltabell minus gjennomsnittlig hovedeffekt av A i deltabellene. Vi splittet her i deltabeller etter verdi av C. Det er symmetri i u-ene så vi kunne laget deltabeller m.h.t. A eller B og oppnå tilsvarende tolkninger.

De to betraktningsmåtene kan brukes når vi har flere enn tre variable. I kapittel 5 vil vi beskrive en tredje måte for å tolke parametrene.

2b. Ulike modeller

Ligningene (2.1) og (2.2) definerer hva vi vil kalle den mettede modell. De gir en en-entydig sammenheng mellom u-ene og m_{ijk}-ene (dvs. p_{ijk} når n kjent). De representerer ingen forenkling av strukturen i tabellen, de gir oss bare en hensiktsmessig omformulering p.g.a. tolkningsmulighetene til u-ene. Den videre nytten av u-ene ligger i at når vi i ligning (2.1) setter noen av u-ene lik 0, framkommer ulike interessante strukturer for tabellen. Disse strukturene kan også uttrykkes ved p_{ijk}-ene men da på en mer komplisert måte. Videre må også nevnes at selvsagt kan ikke alle strukturer defineres ved å sette u-er lik null.

Vi skal se på noen av dem som kan defineres. Når vi skriver f.eks. $u^{AC} = 0$, så betyr det at $u_{ik}^{AC} = 0$ for alle (i,k). Videre brukes $m_{ij.}$ for $\sum_k m_{ijk}$, $m_{i..}$ for $\sum_{j,k} m_{ijk}$, $m...$ for $\sum_{i,j,k} m_{ijk}$ og tilsvarende. Vi skal se på 5 ulike strukturer som framkommer når vi setter noen av u-ene lik null

1) $u^{ABC} = 0.$

Ligning (2.1) blir da

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$$

I følge våre tolkninger av u-ene betyr dette at det er samspill mellom A og B, men at dette samspillet er likt for ulike verdier av C. Det er symmetri, så vi kan også beskrive denne strukturen ved at samspillet mellom A og C er likt for ulike verdier av B eller at samspillet mellom B og C er likt for ulike verdier av A. I et konkret tilfelle kan vi velge den av de tre ekvivalente formuleringene vi finner mest opplysende.

Vi har også

$$u^{ABC} = 0 \text{ ekvivalent med } \frac{m_{ijk} \cdot m_{iJk}}{m_{Ijk} \cdot m_{IJk}} = \frac{m_{ijK} \cdot m_{iJK}}{m_{IjK} \cdot m_{IJK}} \text{ for } i < I,\ j < J \text{ og } k < K. \quad (2.4)$$

Vi skal bruke denne ligningen senere.

2) $u^{BC} = u^{ABC} = 0.$

Da er samspillet mellom B og C ikke bare likt for alle verdier av A, men det er også lik 0. Setter vi inn $u^{BC} = u^{ABC} = 0$ i (2.1), kan vi vise at det medfører at $m_{ijk} = \frac{m_{ij.} \cdot m_{i.k}}{m_{i..}}$. Vi har at B og C er betinget uavhengig gitt A.

3) $u^{BC} = u^{AC} = u^{ABC} = 0.$

Ligning (2.1) blir nå

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB}$$

Det er bare samspillet mellom A og B som er forskjellig fra null og dette er likt for ulike nivå av C. I denne situasjonen har vi $m_{ijk} = \frac{m_{ij.} \cdot m_{..k}}{m...}$ som viser at C uavhengig av (A, B).

4) $u^{AB} = u^{AC} = u^{BC} = u^{ABC} = 0$

Alle samspill er lik 0. Det gir fullstendig marginal uavhengighet mellom de tre variable, dvs. $m_{ijk} = \frac{m_{i..} \cdot m_{.j.} \cdot m_{..k}}{m_{...}^2}$ 3-veistabellen framkommer ved multiplisering av de 3 marginale fordelinger for variablene A, B og C.

5) $u^A = u^B = u^C = u^{AB} = u^{AC} = u^{BC} = u^{ABC} = 0.$

Da er ifølge (2.1) $\log m_{ijk}$ lik u, dvs. at alle celler er like sannsynlige. Varianter av modellene 2) og 3) kan en få ved å bytte om på bokstavene.

Modellene i 1)-5) har det til felles at ingen u kan settes lik 0 uten at alle u-er der topp-skriften til den første inngår også er lik 0. Vil vi ha f.eks. u^A lik 0, så må også u^{AB}, u^{AC} og u^{ABC} være lik 0. Vi skal hele tiden begrense oss til slike modeller som vi kaller hierarkiske.

2c. Multiplikativ formulering av modellene

Ligningene (2.1) og (2.2) gir en lineær modell for logaritmen til m_{ijk}. Dette er det samme som en multiplikativ modell for m_{ijk}. En ser dette ved at (2.1) gir

$$\begin{aligned} m_{ijk} &= e^{\log m_{ijk}} \\ &= e^{(u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC})} \\ &= e^u \cdot e^{u_i^A} \cdot e^{u_j^B} \cdot e^{u_k^C} \cdot e^{u_{ij}^{AB}} \cdot e^{u_{ik}^{AC}} \cdot e^{u_{jk}^{BC}} \cdot e^{u_{ijk}^{ABC}} \qquad (2.5) \\ &= \tau \cdot \tau_i^A \cdot \tau_j^B \cdot \tau_k^C \cdot \tau_{ij}^{AB} \cdot \tau_{ik}^{AC} \cdot \tau_{jk}^{BC} \cdot \tau_{ijk}^{ABC} \end{aligned}$$

Sidebetingelsene blir

$$\prod_i \tau_i^A = \prod_j \tau_j^B = \prod_k \tau_k^C = 1,$$

$$\prod_i \tau_{ij}^{AB} = \prod_j \tau_{ij}^{AB} = 1, \prod_i \tau_{ik}^{AC} = \prod_k \tau_{ik}^{AC} = 1, \prod_k \tau_{jk}^{BC} = 1 \qquad (2.6)$$

og $$\prod_i \tau_{ijk}^{ABC} = \prod_j \tau_{ijk}^{ABC} = \prod_k \tau_{ijk}^{ABC} = 1.$$

τ-ene gir en modell for forventede cellefrekvenser (evt. cellesannsynligheter). De kan være fordelaktig å bruke da det er lettere å tenke i form av cellefrekvenser enn i logaritmen til cellefrekvenser. ECTA gir anledning til å arbeide med begge modellformuleringene.

2d. Å finne modeller for data

Vi skal nå se på hvordan en kan sammenholde observerte data med ulike log-lineære modeller. I prinsippet er ikke dette så uvant som kanskje noen vil tro. I en multinomisk situasjon med enhetene krysssklassifisert m.h.t. to variable (toveistabell) har de fleste gjennomført en kji-kvadrattest for uavhengighet med observatoren

$$Z = \sum_{ij} \frac{(X_{ij} - \frac{1}{n} \cdot X_{i.} \cdot X_{.j})^2}{X_{i.} \cdot X_{.j}/n}$$

Modellen (hypotesen) som prøves er at de to variable er uavhengige. I telleren i testobservatoren sammenholdes det observerte, X_{ij}, med $\frac{1}{n} \cdot X_{i.} \cdot X_{.j}$, som er en estimator for det forventede under forutsetning av at modellen gjelder. Er avviket for stort, forkaster vi modellen (hypotesen) om uavhengighet. På tilsvarende måte vil vi vurdere ulike log-lineære modeller. La H være en hypotese om at én spesiell modell gjelder. Vi tester denne med observatoren

$$LL\ (H) = 2 \sum_\theta X_\theta\ (\log X_\theta - \log \hat{m}_\theta),$$

der $\hat{m}_\theta$ er en estimator for forventet cellefrekvens når modellen H gjelder. θ betegner de aktuelle indekser. Summasjonen foregår over alle celler i tabellen. Med to variable er altså $\theta = i,j$. Vi foretrekker observatoren LL (H) framfor generaliseringer av Z, da LL (H) har noen spesielle egenskaper som vi vil bruke når vi leter etter en passende modell. Beregningen av $\hat{m}_\theta$ og LL (H) vil ECTA gjøre for oss, nærmere om dette siden. Til hver LL (H) er knyttet et visst antall frihetsgrader som er lik antall uavhengige u-er som settes lik null under H. Til H: "$u^{ABC} = 0$" er knyttet (I-1) · (J-1) · (K-1) frihetsgrader. Telling av frihetsgrader og spesiell justering av disse når nullestimater i noen celler er behandlet i kapittel 3.8 i Bishop et al. (1975).

Under H er LL (H) asymptotisk kji-kvadratfordelt, dvs. observatoren har denne fordelingen når antall observasjoner går mot uendelig. Vi vil bruke nevnte fordeling som tilnærming i det endelige tilfellet. Vi kan ikke vise til noen bestemt regel for når tilnærmingen til kji-kvadratfordelingen er tilfredsstillende, men vil bruke tilnærmingen når langt de fleste estimerte cellefrekvensene er større enn 5.

Hvis hensikten med å samle inn data har vært å teste én helt spesifisert modell, kan det gjøres med LL (H). Vi finner da enten at mistilpasningen mellom data og modell målt ved LL (H) er så stor at modellen må forkastes, eller at mistilpasningen er såpass liten at vi ikke kan avskrive modellen. En slik test gjennomføres i kapittel 3b.

Nå er det vel sjelden at vi har så greit siktemål. Noen ganger kan vi ha en viss kunnskap om sammenhengene, men ikke mer bestemt enn at en hel familie modeller er aktuelle. Andre ganger kan vi være helt forutsetningsløse og vil bruke data til å finne fram til en modell. (Slike situasjoner behandles i kapitlene 3c og 4.) Da må vi ha klart for oss hva vi krever av en modell. To krav er

naturlige:

1) Modellen gir en god forklaring på de observerte data. Det må være små avvik mellom data og det forventede når modellen gjelder.

2) Modellen bør ha få parametre. Vi foretrekker enkle forklaringer. I vår sammenheng betyr det bl.a. at vi helst ikke vil ha samspillsparametre av orden 2 og høyere i den endelige modell.

Kravene er motstridende. I enhver situasjon vil vi oppnå perfekt tilpasning mellom modell og data med den mettede modell. Etterhvert som vi begrenser antall parametre vil mistilpasningen øke. Avhengig av omstendighetene vil vi foretrekke en forholdsvis enkel modell som dog ikke er enklere enn at den med rimelighet kan antas å ha generert de data vi observerer. I de neste kapitlene vil vi se på ulike metoder for å finne modeller.

LL (H) har en egenskap vi skal bruke gjentatte ganger. For en treveistabell har vi f.eks. hypotesene

$$H_1: \; u^{ABC} = 0.$$

$$H_2: \; u^{AB} = u^{ABC} = 0.$$

H_2 er strengere enn H_1, i den forstand at H_2 inneholdt i H_1. Noen ganger er det ønskelig å foreta testingen trinnvis. Vi tester da først H_1 med observatoren LL (H_1). Hvis H_1 ikke forkastes vil vi se om ytterligere reduksjon i parametrene mulig. Vi vil teste H_2 gitt at H_1 er sann. Dette kan vi gjøre med observatoren

$$LL\,(H_2|H_1) = LL\,(H_2) - LL\,(H_1).$$

LL ($H_2|H_1$) er under H_2 tilnærmet kji-kvadratfordelt med ($\nu_2 - \nu_1$) frihetsgrader, der ν_1, ν_2 er frihetsgradene tilknyttet h.h.v. LL (H_1) og LL (H_2). Med H_1 og H_2 som angitt ovenfor i en treveistabell, blir ν_1 lik (I-1) (J-1) (K-1) og ν_2 lik K · (I-1) · (J-1). Da blir det (I-1) (J-1) frihetsgrader knyttet til LL ($H_2|H_1$).

2e. Spesifisering av modeller i ECTA

I LL (H) inngikk $\hat{m}_e$-er, estimater for cellefrekvensene under hypotesen H. Både LL (H) og $\hat{m}_e$-er beregnes av ECTA, men det er nødvendig å forklare hvordan hypotesene spesifiseres i programmet.

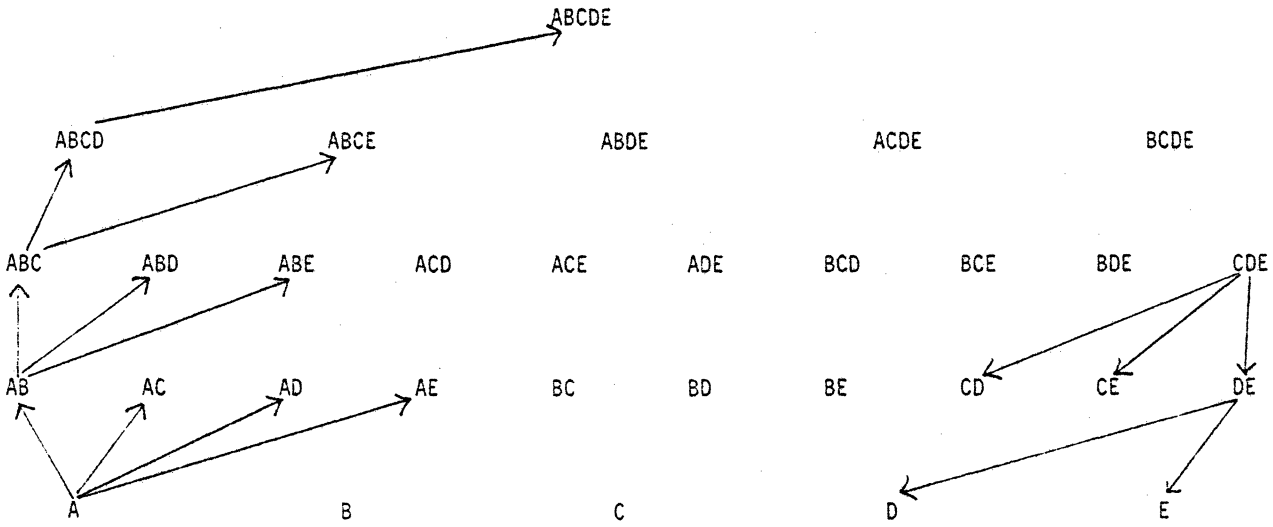
Vi må innføre notasjon for å betegne ulike tabeller. I treveistilfellet står ABD for den opprinnelige tabell, AB for toveistabellen som framkommer når en i ABC summerer over variabel C. A er tabellen en får når en i ABC summerer over variable B og C. Tabellen A kan også framkomme ved å summere over variabel B i tabellen AB. For flere enn tre variable vil vi bruke tilsvarende notasjon. Marginale tabeller benevnes altså ved de variable en ikke summerer over i den opprinnelige tabell for å lage den enkelte marginaltabell.

Modellene er bestemt av hvilke u-ledd som er tilstede. Vi ser bare på hierarkiske modeller. Det medfører at hvis et høyere ordens ledd er tilstede, så vil også visse ledd av lavere orden være tilstede. Hvis f.eks. u^{AB} er med, så må også u^A og u^B være med. Når vi skal forklare hvilke ledd som er med i modellen, er det derfor ikke nødvendig å nevne de ledd som i følge det hierarkiske prinsipp må være med. Modellen log $m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB}$. (2.7)
er bestemt når vi sier at leddene u^{AB} og u^C er med. Toppskriftene på de ledd som det således er nødvendig å ta med, betegner ulike marginaltabeller. Disse tabellene sies å være <u>de bestemmende marginaler</u>. For modellen i (2.7) blir dette tabellene AB og C. I ECTA spesifiseres modeller ved de bestemmende marginaler. En må bruke tall i stedet for bokstaver for variablene. En skriver 12, 13 og 23 for AB, AC og BC, og 12 og 3 for AB og C osv.

Dette er ikke bare en formell metode for å betegne ulike modeller. Estimeringsmetoden som brukes er slik at de estimerte frekvenser, $\hat{m}_\theta$ -ene, vil stemme overens med de observerte frekvenser for de bestemmende marginaltabellene. $\hat{m}_\theta$-ene er sannsynlighetsmaksimeringsestimater (SME). Det samme er $\hat{u}_\theta$-ene. Vil vi ha SME for andre entydige funksjoner av m_θ-ene, får vi dette ved å sette inn $\hat{m}_\theta$-ene i definisjonsligningene. I noen modeller vil $\hat{m}_\theta$-ene være eksplisitte funksjoner av de observerte cellefrekvensene, X_θ-ene, i andre modeller vil $\hat{m}_\theta$-ene være implisitt definert. I begge tilfelle vil ECTA beregne de aktuelle estimater. En fyldig behandling av estimeringsmetoden finner en i kapittel 3 i Bishop, et.al. (1975).

Når en har mange variable, kan det være vanskelig å beholde oversikten over mulige modeller og over hvilke parametre som må være med for at modellen skal være hierarkisk. Det kan da være hjelp i et skjema med toppskriftene til alle mulige parametre. I figur 2.8 finner en slikt skjema for et system med fem variable.

Figur 2.8. Et system med fem variable



På figuren har vi med piler antydet sambandet mellom ledd på ulike nivåer. Vi ser f.eks. at hvis u^{ABC} er lik null i en modell, så må også u^{ABCD}, u^{ABCE} og u^{ABCDE} være lik null hvis modellen skal være hierarkisk. Hvis u^{AB} lik null, så må alle tre-, fire- og femfaktoreffekter som kan nåes fra AB med oppadgående piler, være lik null. (Alle pilene er ikke tegnet inn på figuren). De nedadgående pilene viser ledd som må være med i modellen når et spesielt ledd er med. Hvis u^{DE} er med, så må også u^{D} og u^{E} være med. Hvis BCDE er med, så må u^{BCD}, u^{BCE}, u^{BDE}, u^{CDE} og alle ledd som kan nåes med nedadgående piler fra disse være med. (Alle disse er ikke tegnet inn).

Figuren kan også være til hjelp når en vil finne de bestemmende marginaler. For en modell vil de bestemmende marginaler være de toppskrifter som ikke kan nåes med nedadgående piler fra toppskriften til andre u-ledd i modellen. For modellen

$$\log m_{ijk\ell m} = u + u^A_i + u^B_j + u^C_k + u^D_\ell + u^E_m + u^{AB}_{ij} + u^{AC}_{ik} + u^{AD}_{i\ell} + u^{AE}_{im}$$
$$+ u^{CD}_{k\ell} + u^{CE}_{km} + u^{ACD}_{ik\ell} + u^{ACE}_{ikm}$$

vil ACD, ACE og AB være bestemmende marginaler.

Videre viser figuren at hvis BCDE er en av de bestemmende marginaler i en modell, og vi ønsker å sette u^{BCDE} lik null, så blir BCD, BCE, BDE og CDE med blant de bestemmende marginaler i den nye modellen i den grad de ikke allerede inngår i de andre bestemmende marginalene.

Når en vil utvide en gitt modell med én ny parameter, kan også skjemaet være nyttig. Anta vi har en modell bestemt av marginalene AB, AC, AD, AE, BC og BD. Da er u^{BE}, u^{CD}, u^{CE} og u^{DE} opplagte kandidater, men også u^{ABC} og u^{ABD} er mulige. Derimot er ikke u^{ACD} mulig, da vil den nye modellen ikke bli hierarkisk siden u^{CD} ikke er med. Velger vi u^{ABC} som nytt ledd, blir ABC, AD, AE og BD bestemmende marginaler for den nye modellen.

3. ET EKSEMPEL

I kapitlet vil vi gjennomføre en analyse av en femveistabell om tannlegekontakter. Data er fra Byråets helseundersøkelse i 1975, Statistisk Sentralbyrå (1977). Lignende materiale ble også samlet inn av Levekårsundersøkelsen i 1973, og er bl.a. presentert i Holst (1977). Hun arbeider stort sett med to- og treveistabeller og peker på at hun i noen tilfelle ikke kan komme med "sikre" tolkninger fordi de interessante variable ikke inngår simultant. De log-lineære modeller gir oss et redskap for simultan analyse.

3a. Om valg av data

Materialet vi skal analysere vises i tabell 3.1.

Tabell 3.1. Personer 25-54 år med egne tenner i grupper for utdanning, inntekt, reisetid og kjønn etter om hatt kontakt med tannlege siste året

Utdanning	Inntekt	Reisetid	Kjønn	Kontakt		I alt	% kontakt
				1	2		
1	1	1	1	710	208	918	77
1	1	1	2	759	109	868	87
1	1	2	1	12	4	16	75
1	1	2	2	14	7	21	67
1	2	1	1	374	145	519	72
1	2	1	2	460	113	573	80
1	2	2	1	14	12	26	54
1	2	2	2	12	3	15	80
2	1	1	1	81	78	159	51
2	1	1	2	138	63	201	69
2	1	2	1	2	1	3	67
2	1	2	2	2	3	5	40
2	2	1	1	82	85	167	49
2	2	1	2	80	59	139	58
2	2	2	1	3	10	13	23
2	2	2	2	7	11	18	39

Inndelingen av de variable er

(A) Kontakt: 1 = Under 1 år siden kontakt med tannlege
2 = 1 år og mer siden kontakt

(B) Kjønn: 1 = Mann
2 = Kvinne

(C) Reisetid: 1 = Under 1 times reisetid til nærmeste tannlegekontor
2 = 1 time og mer i reisetid

(D) Inntekt: 1 = Husholdningsinntekt 50 000 og over i 1974
2 = Under 50 000

(E) Utdanning: 1 = Framhaldskolenivå og høyere
2 = Folkeskolenivå

Når det gjelder kontakt satte vi grensen ved ett år fordi det er et vanlig råd fra tannleger, at voksne folk bør gå til kontroll minst en gang i året. Inndelingen av reisetid gir få personer i den ene gruppen, men det synes å være en erfaring fra annen statistikk om helsetjenester at reisetiden må være nokså lang før den eventuelt har noen effekt. Husholdningsinntekten er sum brutto inntekt i 1974 for personene i husholdningen. Til "folkeskolenivå" på utdanning regnes de som ikke har minst 5 måneders utdanning etter folkeskolen.

I materialet har vi begrenset oss til aldersgruppen 25-54 år. De yngre er holdt utenfor fordi mange av dem omfattes av tilbud om gratis tannbehandling. Vi antar at mønsteret er vesentlig annerledes for dem som ikke selv må betale behandlingen. De eldre er ikke tatt med fordi vi tror de har et annet behov for og en annen holdning til tannlegetjenesten. Videre er de få mellom 25 og 54 som ikke har noen egne tenner, holdt utenfor. Disse føler neppe at de har samme behov som andre for å gå til tannlege, og vi ville ikke at de eventuelt skulle tilsløre resultatene for de andre.

Vi startet opprinnelig analysen med en seksveistabell der materialet var inndelt i to aldersgrupper. En innledende analyse viste at vi kunne se bort fra alder, nærmere om dette i kapittel 7.

Alle variable er dikotomisert. Dette gjør analysen lettere og mer oversiktlig. Vi så i dette tilfellet ingen grunn til å arbeide med tre eller flere kategorier på noen variabel, men har selvsagt ingen garanti for at ikke resultatet av analysen hadde blitt et annet hvis vi hadde gjort det. På samme måte vil også valget av delepunkter for todelingen kunne influere på resultatet.

3b. Testing av ett forslag til modell

La oss anta at vårt eneste formål med dataene i tabell 3.1 var å teste en spesiell hypotese om sammenhengen mellom variablene. Hypotesen går ut på det er sammenheng mellom kjønn (B) og kontakt (A), reisetid (C) og kontakt, utdanning (E) og kontakt, men ingen sammenheng mellom inntekt (D) og kontakt. Sammenhengene påstås like for ulike nivåer av de andre tre variable. Hypotesen lar forholdet mellom variablene B, C, D og E være uspesifisert. Hypotesen kan beskrives ved at følgende modell ligger under de observerte data.

$$\log m_{ijk\ell m} = u + u_i^A + u_j^B + u_k^C + u_\ell^D + u_{im}^E$$
$$+ u_{ij}^{AB} + u_{ik}^{AC} + u_{im}^{AE} + u_{jk}^{BC} + u_{j\ell}^{BD} + u_{jm}^{BE} + u_{k\ell}^{CD} + u_{km}^{CE} + u_{\ell m}^{DE}$$
$$+ u_{jk\ell}^{BCD} + u_{jkm}^{BCE} + u_{j\ell m}^{BDE} + u_{k\ell m}^{CDE}$$
$$+ u_{jk\ell m}^{BCDE} \qquad i = 1,2 \quad j = 1,2 \quad k = 1,2 \quad \ell = 1,2 \quad m = 1,2$$

Bestemmende marginaler er BCDE, AB, AC og AE. Vi har 12 frihetsgrader og bruker 5 prosent nivå ved testen. Vi får LL (H) lik 28,02 som er større enn 21,026 (95 prosent-fraktilen i kji-kvadratfordelingen med 12 frihetsgrader). Hypotesen forkastes.

Etter dette kan en lure på om en oppnådde tilfredsstillende tilpasning hvis en føyde til leddet u^{AD}, men da er vi over i søkeprosedyrer og betinget testing som behandles i kapitlene 3c og 4.

3c. Den mettede modell og bruken av denne

For den mettede modell blir alltid LL (H) = 0,00. Av beregningene for denne modellen er det estimatene for u-ene som interesserer. For vårt eksempel er disse gjengitt i tabell 3.2. Bare u-ene med alle fotskrifter lik 1 er tatt med, de andre er bestemt av randbetingelser som i (2.2). F.eks. får vi $u^{AB}_{12} = u^{AB}_{21} = -u^{AB}_{11} = 0{,}086$ og $u^{AB}_{22} = -u^{AB}_{21} = u^{AB}_{11} = -0{,}086$. SD ($\hat{u}$) er et estimat for standardavviket til $\hat{u}$.

Tabell 3.2. Estimater for parametrene i den mettede modell

Variable	$\hat{u}$	SD ($\hat{u}$)	$\frac{\hat{u}}{SD\ (\hat{u})}$	$\hat{\tau}$
(Gjennomsnittseffekt)	3,360			28,780
A	0,268	0,064	4,155	1,307
B	-0,016	0,064	-0,253	0,984
C	1,648	0,064	25,549	5,194
D	-0,110	0,064	-1,712	0,895
E	0,525	0,064	8,136	1,690
AB	-0,086	0,064	-1,326	0,918
AC	0,145	0,064	2,252	1,156
AD	0,111	0,064	1,728	1,118
AE	0,285	0,064	4,424	1,330
BC	0,060	0,064	0,924	1,061
BD	-0,081	0,064	-1,251	0,922
BE	0,107	0,064	1,657	1,113
CD	0,234	0,064	3,630	1,264
CE	0,091	0,064	1,416	1,096
DE	0,186	0,064	2,889	1,205
ABC	-0,055	0,064	-0,861	0,946
ABD	0,088	0,064	1,369	1,092
ABE	-0,039	0,064	-0,609	0,962
ACD	-0,026	0,064	-0,400	0,975
ACE	-0,008	0,064	-0,131	0,992
ADE	-0,045	0,064	-0,692	0,956
BCD	0,070	0,064	1,085	1,072
BCE	-0,072	0,064	-1,123	0,930
BDE	-0,027	0,064	-0,413	0,974
CDE	-0,127	0,064	-1,965	0,881
ABCD	-0,130	0,064	-2,011	0,878
ABCE	0,034	0,064	0,529	1,035
ABDE	-0,002	0,064	-0,031	0,998
ACDE	0,061	0,064	0,948	1,063
BCDE	0,105	0,064	1,621	1,110
ABCDE	0,011	0,064	0,176	1,011

I programmet brukes en metode for beregning av SD ($\hat{u}$) som er best egnet for den mettede modell. I andre modeller vil oftest SD ($\hat{u}$) bli overestimert, Lee (1977).

Med tabell 3.2 kan vi foreta en foreløpig vurdering av hvilke parametre som er nødvendige for beskrivelsen av materialet. Hvis antall observasjoner er stort, vil de standardiserte verdier, $\hat{u}/SD(\hat{u})$, være tilnærmet normalfordelt. Denne tilnærmingen må vi bruke med forsiktighet, da antall observasjoner i noen av cellene er nokså lite, se tabell 3.1. Vi er ikke like interessert i alle parametrene. Gjennomsnittseffekten og de fem hovedeffektene gir uttrykk for rent marginale trekk ved variablene. Vi vil her begrense oss til modeller som i det minste inneholder disse seks parametrene. Større interesse knytter seg til de resterende 26 parametrene. Vi vil vurdere hvilke av disse som må antas å være forskjellig fra null. Testene utføres ved å sammenlikne med passende fraktiler i normalfordelingen. Med nivå 5 prosent gir separate tester av de 26 parametrene at u^{ABCD}, u^{CDE}, u^{DE}, u^{CD}, u^{AE} og u^{AC} er signifikant forskjellige fra null. Vi fant dette ved å se hvilke av de standardiserte parameterestimatene som var større enn 1,96 eller mindre enn -1,96. Ut fra dette vil den underliggende modell være bestemt av marginalene ABCD CDE AE og kan uttrykkes ved

$$\begin{aligned}\log m_{ijk\ell m} = {} & u + u_i^A + u_j^B + u_k^C + u_\ell^D + u_m^E \\ & + u_{ij}^{AB} + u_{ik}^{AC} + u_{i\ell}^{AD} + u_{im}^{AE} + u_{jk}^{BC} + u_{j\ell}^{BD} + u_{k\ell}^{CD} + u_{km}^{CE} + u_{\ell m}^{DE} \\ & + u_{ijk}^{ABC} + u_{ij\ell}^{ABD} + u_{ik\ell}^{ACD} + u_{jk\ell}^{BCD} + u_{k\ell m}^{CDE} \\ & + u_{ijk\ell}^{ABCD}\end{aligned} \qquad (3.3)$$

Målt med LL (H) er mistilpassingen 22,69 og med 11 frihetsgrader er tilpassingen nokså dårlig. (Det ser vi ved å sammenholde med fraktilene i kjikvadratfordelingen med 11 frihetsgrader.)

Vi tror at som regel vil metoden føre til modeller med vesentlig bedre tilpassing, men da kan man også ha kommet galt avsted. Vi utførte i alt 26 separate tester. Nivået på hver test er 5 prosent, men hvis alle u-ene var lik null, så vil vi med metoden ha sannsynlighet langt større enn 0,05 for feilaktig å påstå at en eller flere u-er ulik null. Med 26 parametre å teste er det svært trolig at en eller flere av "signifikansene" er resultat av tilfeldige variasjoner.

Hvis vi i de separate testene sammenlikner med $0{,}05/2\cdot26$ og $1-0{,}05/2\cdot26$ fraktilene, sikrer vi et simultant 5 prosent nivå. Med denne metode blir bare u^{AE} og u^{CD} signifikante. Modellen er bestemt av marginalene AE, CD og B og det betyr at

$$\log m_{ijk\ell m} = u + u_i^A + u_j^B + u_k^C + u_\ell^D + u_m^E + u_{im}^{AE} + u_{k\ell}^{CD} \qquad (3.4).$$

For modellen finner vi LL (H) = 147,04 og selv med 24 frihetsgrader er tilpassingen altfor dårlig. Vi kan ikke bruke (3·4) som modell, den forklarer ikke bra nok det observerte.

De to metodene kan gi utgangspunkt for trinnvis søking etter "bedre" modeller. Fra modeller med god tilpassing, men med mange parametre, kan vi utelukke en og en av u-ene så lenge reduksjonen i tilpassingen ikke er signifikant. Dette tester vi for hvert trinn med differansen mellom de aktuelle LL (H)-ene. Hvis utgangspunktet er en modell med få parametre, men med for dårlig tilpassing, kan vi på tilsvarende måte arbeide oss mot mer kompliserte modeller. Vi føyer til et og et u-ledd så lenge vi for hvert trinn oppnår signifikant øking i tilpassingen. Disse metodene krever at vi har en regel for å finne det u-ledd som skal sløyfes/tilføyes og vi må bestemme et nivå for enkelttestene. Vi kommer nærmere inn på dette i neste kapittel der vi demonstrerer en trinnvis metode som tar utgangspunkt i den mettede modell.

Vi godtok verken (3.3) eller (3.4) som endelig modell. Det har også sammenheng med at parametrene er estimert i den mettede modell. Vi holder det for opplagt at den underliggende modell er enklere. Estimeringen av u-ene foregår dermed i en overspesifisert modell, og vi er usikre på hvilken effekt dette har.

Slik som u-ene er definert, er det også litt kunstig å foreta en samtidig vurdering av signifikansen av de enkelte. Når vi i eksemplet finner u^{ABCD} ulik null, får det liten mening å spørre om f.eks. u^{ABC} ulik null. I hvertfall får det en ganske annen mening enn å spørre om det samme når u^{ABCD} er lik null. Når u^{ABCD} er ulik null, vet vi at samspillet av 2. orden mellom A, B og C er forskjellig for de to verdiene av D. Vi må da snakke om samspillene og ikke samspillet mellom A, B og C. u^{ABC} blir et veid snitt mellom to ulike samspill. Foran fant vi f.eks. u^{AC} signifikant forskjellig fra null. I det vi også har u^{ABCD} ulik null betyr det at en spesiell funksjon av samspillene mellom A og C for ulike verdier av B og D, er forskjellig fra null.

Tross dette mener vi at til enhver analyse av en flerveistabell, hører en utkjøring av parameterestimatene i den mettede modell. De gir informasjon som betraktet med en viss skepsis, er meget nyttig. De kan gi forslag til rekkefølge og/eller utgangspunkt for testingen i trinnvise prosedyrer.

4. TRINNVIS LETING ETTER MODELL

4a. En trinnvis metode

Med utgangspunkt i den mettede modell skal vi foreta en trinnvis leting etter modell for femveistabellen i eksemplet. Prosedyren vil omfatte en rekke enkelttester, og vi må bestemme nivå for hver av disse. For at vi ikke med stor sannsynlighet før eller siden skal foreta en feil forkasting, må vi velge et lavt nivå. Hvor lavt er vanskelig å angi, da vi på forhånd ikke vet hvor mange enkelttester vi kommer til å utføre. Vi vil bruke 1 prosent nivå i de enkelte testene.

Først tester vi H_1: Femfaktoreffekten lik null ($u^{ABCDE} = 0$), vi får LL (H_1) = 0,03 og forkaster ikke H_1. Vi har en frihetsgrad ved testingen. Gitt at H_1 gjelder, så tester vi: Alle firefaktoreffektene lik null. (Vi ser kun på hierarkiske modeller så H_2: $u^{ABCD} = u^{ABCE} = u^{ABDE} = u^{ACDE} = u^{BCDE} = u^{ABCDE} = 0$). Vi finner LL ($H_2|H_1$) = LL ($H_2$) - LL ($H_1$) = 6,98 - 0,03 = 6,95, med (6-1) = 5 frihetsgrader kan ikke H_2 forkastes. Da tester vi H_3: Alle trefaktoreffekter lik null, gitt at H_2 gjelder. LL ($H_3|H_2$) = LL (H_3) - LL (H_2) = 27,67 - 6,98 = 20,69, der er (16-6) = 10 frihetsgrader i testingen og med 1 prosent nivå forkastes ikke hypotesen. Da går vi videre og tester H_4: Alle tofaktoreffekter lik null, gitt at H_3 er sann. Resultatet blir LL ($H_4|H_3$) = 297,44 og vi forkaster H_4.

Vi har til nå testet grupper av u-er, f.eks. fra H_2 til H_3 ble alle 10 trefaktoreffekter testet på en gang. Nå måtte vi forkaste hypotesen (H_4) om at alle tofaktoreffekter var null. Da går vi tilbake til H_3 og tester om én tofaktoreffekt kan settes lik null. Det er da i alt 10 effekter å velge blant. Vi velger å teste den som satt lik null gir minst reduksjon i tilpassingen. Dette måles med LL (H). Resultatene for de 10 alternativene er gjengitt i tabell 4.1.

Tabell 4.1. Mulige hypoteser på neste trinn når H_3 ikke forkastet

Hypotese	LL (H)
$H_3 \cap (u^{AB} = 0)$	80,69
$H_3 \cap (u^{AC} = 0)$	39,30
$H_3 \cap (u^{AD} = 0)$	48,20
$H_3 \cap (u^{AE} = 0)$	173,75
$H_3 \cap (u^{BC} = 0)$	27,84
$H_3 \cap (u^{BD} = 0)$	28,50
$H_3 \cap (u^{BE} = 0)$	32,53
$H_3 \cap (u^{CD} = 0)$	44,98
$H_3 \cap (u^{CE} = 0)$	33,52
$H_3 \cap (u^{DE} = 0)$	38,09

Vi ser av tabellen at "$u^{BC} = 0$" gir best tilpassing og vi velger H_5: Alle trefaktoreffekter og u^{BC} lik null. Vi tester H_5 gitt H_3 med LL (H_5 | H_3) = LL (H_5) - LL (H_3) = 0,17. H_5 kan ikke forkastes.

I neste trinn undersøker vi om enda en to-faktoreffekt kan settes lik null. Vi finner at "$u^{BD} = 0$" gir minst reduksjon i tilpassing og velger H_6: $H_5 \cap (u^{BC} = 0) \cap (u^{BD} = 0)$. Vi tester med LL ($H_6$ | H_5) og kan ikke forkaste H_6. På denne måte fortsetter vi trinnvis til vi må forkaste en hypotese. Vi får at H_8: $H_3 \cap (u^{BC} = 0) \cap (u^{BD} = 0) \cap (u^{BE} = 0) \cap (u^{CE} = 0)$ ikke forkastes, men H_9: $H_8 \cap (u^{DE} = 0)$ forkastes fordi LL (H_9 | H_8) = 11,99 og med 1 frihetsgrad viser dette at det blir et signifikant tap i tilpassing når en ut fra H_8 setter u^{DE} lik null.

Prosedyrens forslag til modell for femveistabellen blir at

$$\log m_{ijk\ell m} = u + u^A_i + u^B_j + u^C_k + u^D_\ell + u^E_m + u^{AB}_{ij} + u^{AC}_{ik} + u^{AD}_{i\ell} + u^{AE}_{im} + u^{CD}_{k\ell} + u^{DE}_{\ell m}.$$

Bestemmende marginaler er altså AB, AC, AD, AE, CD og DE.

Når en bruker en slik mekanisk søkermetode, må en vurdere sluttresultatet nøye. For H_8 har vi LL (H_8) = 39,87. Til H_8 er knyttet 20 frihetsgrader. Det viser at tilpassingen er dårlig. Hvorvidt en skal ta det som en avgjørende innvending mot modellen, avhenger av formålet med analysen. Vårt formål er å finne fram til hovedtrekkene i materialet. På forhånd hadde vi ingen hypoteser eller viten som tilsa at spesielle parametre må være med i modellen. Videre har vi under arbeidet med dataene sett at for å oppnå en vesentlig bedre tilpassing, må vi tilføye temmelig mange flere ledd i modellen. Derfor aksepterer vi i dette tilfellet den dårlige tilpassingen.

I søkeprosedyren vil tilpassingen på sluttproduktet avhenge av nivået på enkelttestene. Men for det samme nivå på disse, kan tilpassingen på sluttresultatet variere meget alt etter strukturen i materialet en arbeider med.

Programmet kan beregne estimater for cellefrekvensene under gitte modeller. Sammenlikner en disse med de observerte cellefrekvensene kan en oppdage systematiske avvik som kan gi ideer om parametre som burde ha vært med i modellen. Vi har ikke gjennomført dette for H_8. I kapittel 4.4 i Bishop et.al. (1975) fins en del om metoder for slik residualanalyse.

Vi aksepterer H_8 som modell. Tolkning og presentasjon behandles i kapittel 5. Først har vi noen kommentarer om trinnvise prosedyrer.

4b. Andre trinnvise metoder

Som i regresjonsanalysen fins det mange ulike trinnvise metoder. Vi skal peke på noen muligheter for varianter.

I 4a) gikk vi fra det kompliserte til det enklere ved gradvis å utelukke parametre. Metoden kan kalles en baklengsmetode. Der fins også forlengsmetoder. En arbeider da fra enklere mot mer kompliserte modeller. På hvert trinn testes om parameteren som tas inn, gir en signifikant øking i tilpassingen. En stopper når en ikke kan oppnå slik øking. Madsen (1977) holder baklengsmetodene for å være best. En arbeider da hele tiden i modeller som er "riktige" i den forstand at de ihvertfall ikke er for enkle. Som regel oppnår en da størst styrke på testene.

I eksemplet startet vi med den mettede modell. Det tilsvarende i en forlengsmetode ville være å starte i den enkleste modell. Det er modellen der alle cellefrekvensene antas like, dvs. $\log m_{ijk\ell m} = u$. Uansett hvilken vei en går, kan en velge andre utgangspunkt. Vi kan f.eks. bruke resultatet av testingen av parametrene i den mettede modell (se kap. 3) som utgangspunkt.

I 4a) testet vi på en gang grupper av parametre. I annet trinn f.eks. satte vi alle fire-faktoreffekter lik null. Mer vanlig er det kanskje å teste bare ett u-ledd om gangen.

Videre kan en velge mellom å behandle ledd på ulike nivå om hverandre eller gjøre seg ferdig med alle ledd på ett nivå før en prøver med ledd på andre nivå. I en forlengsprosedyre betyr dette at med modellen bestemt av marginalene AB, AC, AD, AE og BC, er u^{BD}, u^{BE}, u^{CD}, u^{CE} og u^{DE} kandidater ved utvidelse, hvis vi holder oss til tofaktoreffekter. Tillater vi å utvide med effekter på andre nivåer, vil i tillegg u^{ABC} være en mulighet.

Ulike metoder kan også utvides ved en revurdering på hvert trinn. Slik vi gjennomførte testingen i 4a), kunne ikke et ledd som var blitt utelukket, senere bli tatt inn i modellen. For baklengsprosedyrer innebærer revurderingen at hver gang et ledd er utelukket, så undersøker en om ledd som da ikke er med i modellen, gir signifikant økning i tilpassingen. Da vurderes også ledd som tidligere er utelukket. Men en arbeider hele tiden med hierarkiske modeller og tilføyer kun et ledd av gangen. I noen tilfelle kan en også nøye seg med å foreta revurdering når prosedyren stanser fordi neste trinn gir signifikant reduksjon i tilpassingen. Som eksempel velger vi modellen H_8 i kapittel 4a. Bestemmende marginaler var AB, AC, AD, AE, CD og DE. Videre reduksjon var ikke mulig. Vi vil revurdere ledd som ikke er med i modellen. Hvis vi også tillater utvidelse med trefaktorledd, er det i alt 6 muligheter. I tabell 4.2 har vi satt opp mulighetene og den tilhørende LL (H).

Tabell 4.2. Mulige utvidelser med én parameter fra modellen H_8.

Bestemmende marginaler	LL (H)
AB, AC, AD, AE, CD, DE og <u>BC</u>	39,53
" " " " " " " <u>BD</u>	38,73
" " " " " " " <u>BE</u>	34,66
" " " " " " " <u>CE</u>	33,93
<u>ACD</u>, AB, AE, DE	39,33
<u>ADE</u>, AB, AC, CD	39,86

I tabellen har vi streket under toppskriften for det u-ledd som tilføyes. Strengt tatt er det ikke nødvendig å ta med alternativet der u^{CE} tilføyes, fordi dette er testet da vi gikk fra H_7 til H_8 i prosedyren. LL (H_8) = 39,87 og ingen av alternativene gir signifikant øking. Testene utføres med 1 prosent nivå og en har 1 frihetsgrad. For forlengsprosedyrer kan en foreta lignende revurderinger. Når et ledd er tilføyd, tester en om andre ledd i modellen kan sløyfes uten at det blir signifikant reduksjon i tilpassingen i den nye sammenhengen.

I prosedyren i kapittel 4a brukte vi LL (H) for de ulike alternativene, for å avgjøre hvilken hypotese som skulle testes på hvert trinn. Hypotesen som ga minst reduksjon i tilpassing, ble testet. Denne framgangsmåten kan være nokså arbeidskrevende. Arbeidsmengden reduseres hvis en på forhånd bestemmer rekkefølgen på testene. En hjelp til dette er de standardiserte parameterestimatene. Vi kan bruke de vi får i den mettede modell eller ta dem fra en modell som behandles underveis. Ledd med små absoluttverdier testes først. En kan få ulike rekkefølger alt etter hvilken modell en tar estimatene fra. I tabell 4.3 er dette illustrert. Der finner en rekkefølgen en får for testing av tofaktoreffekter ved estimater fra fire ulike modeller, samt av kriteriet i kapittel 4a.

Tabell 4.3. Ulike forslag til rekkefølge for testing av tofaktorledd

	Modeller				Ved kriteriet i 4a
	Mettet	Femfaktor-effekter = 0	Firefaktor-effekter = 0	Trefaktor-effekter = 0	
u^{AB}	3	3	5	7	
u^{AC}	7	7	7	8	
u^{AD}	6	6	2	5	
u^{AE}	10	10	10	10	
u^{BC}	1	1	3	2	1
u^{BD}	2	2	4	1	2
u^{BE}	5	5	6	3	3
u^{CD}	9	9	9	9	
u^{CE}	4	4	1	6	4
u^{DE}	8	8	8	4	5

For prosedyren i 4a har vi tatt med de fem første trinn, da måtte vi stoppe fordi u^{DE} ikke kunne settes lik null. Tabellen viser at vi får ulike forslag til rekkefølge, alt etter hvilke estimater/metoder vi baserer oss på. Det er nødvendig å se på forslagene som veiledende, og som oftest vil det være gunstig å foreta revurdering av leddene på ett eller flere trinn.

I kapitlene 3 og 4 har vi behandlet det som i en viss forstand er ytterpunktene i bruk av log-lineære modeller. I kapittel 3b viste vi hvordan data kan brukes til å teste en hypotese som er spesifisert på forhånd mens vi ellers har omtalt metoder der vi mer eller mindre lar data lede oss fram til en modell. Det er selvsagt fullt mulig å bruke modellene på andre måter. Ut fra interesser og a priori kunnskap kan en bestemme hvilke modeller som er aktuelle og/eller velge rekkefølge på testing osv. Vi viser til kapitlene 4 og 9 i Bishop et.al.(1975) og til drøftingene av eksempler i hele boka.

5. TOLKNING OG PRESENTASJON AV MODELLER

5a. Direkte tolkning av u-er

I kapitlet foran kom vi fram til en modell som var bestemt av marginalene AB, AC, AD, AE, CD og DE. Da har vi

$$\log m_{ijk\ell m} = u + u_i^A + u_j^B + u_k^C + u_\ell^D + u_m^E + u_{ij}^{AB} + u_{ik}^{AC} + u_{i\ell}^{AD} + u_{im}^{AE} + u_{k\ell}^{CD} + u_{\ell m}^{DE} \qquad (5.1)$$

eller skrevet som en multiplikativ modell

$$m_{ijk\ell m} = \tau \cdot \tau_i^A \cdot \tau_j^B \cdot \tau_k^C \cdot \tau_\ell^D \cdot \tau_m^E \cdot \tau_{ij}^{AB} \cdot \tau_{ik}^{AC} \cdot \tau_{i\ell}^{AD} \cdot \tau_{im}^{AE} \cdot \tau_{k\ell}^{CD} \cdot \tau_{\ell m}^{DE} \qquad (5.2)$$

for alle (i, j, k, ℓ, m). Programmet beregner estimater for u-er og τ-er. Disse er gjengitt i tabell 5.1.

Tabell 5.1. Estimater for parametrene i den endelige modell

Variable	$\hat{u}$	SD ($\hat{u}$)	$\frac{\hat{u}}{SD(\hat{u})}$	$\hat{\tau}$
(Gjennomsnittseffekt)	3,375			29,237
A	0,217	0,061	3,538	1,242
B	0,062	0,061	1,011	1,064
C	1,641	0,061	26,799	5,163
D	-0,082	0,061	-1,342	0,921
E	0,604	0,061	9,858	1,829
AB	-0,133	0,061	-2,165	0,876
AC	0,199	0,061	3,255	1,221
AD	0,088	0,061	1,431	1,092
AE	0,274	0,061	4,472	1,315
CD	0,207	0,061	3,376	1,230
DE	0,075	0,061	1,227	1,078

Først noterer vi oss at det ikke er trefaktoreffekter i modellen. Tofaktoreffektene gir da uttrykk for en sammenheng mellom to variable som er ens for ulike kombinasjoner av de tre andre variablene.

I tabellen har vi $\hat{u}^{AB}_{11} = -0,133$. Det er negativ sammenheng mellom kontakt og kjønn. Slik som variablene er inndelt betyr det at menn sjeldnere enn kvinner har hatt kontakt med tannlege i løpet av det siste året. Siden vi ikke har noen trefaktoreffekter som inkluderer AB, gjelder dette i samme grad for alle 8 delgrupper m.h.t. reisetid, inntekt og utdanning. Det betyr ikke at andelene med kontakt for h.h.v. menn og kvinner, er konstante i de 8 delgruppene. Det er "forholdet" mellom andelene for menn og kvinner som er konstante for ulike kombinasjoner av de tre andre variable. Vi skriver "forholdet" fordi i de log-lineære modeller sammenlignes andeler på en spesiell måte. Mer om dette i 5b. Tilsvarende gjelder også for tolkningen av andre tofaktoreffekter. I tabellen er $\hat{u}^{AC}_{11} = 0,199$. Det viser at de med under 1 times reisetid har oftere vært hos tannlege siste året enn de med over 1 times reisetid. Videre står $\hat{u}^{AD}_{11} = 0,088$. Det er positiv sammenheng mellom kontakt og inntekt. Med våre inndelinger betyr det at de i gruppa med høyest husholdningsinntekt, oftere har vært hos tannlege i løpet av siste året. Til slutt er $\hat{u}^{AE}_{11} = 0,274$. Det er de med høyest utdanning som oftest har vært hos tannlege i løpet av siste året. u^{CD} og u^{DE} gir oss informasjon om forholdet mellom bakgrunnsvariablene, men skal ikke kommenteres her fordi det primære for oss er samspillene mellom kontakt og andre variable.

Så langt har vi bare brukt fortegnene på u-ene i tolkningen, vi har snakket om positiv eller negativ sammenheng. Vi kan rangere u-ene, og (eventuelt etter en test) påstå at sammenhengen mellom kontakt og utdanning er sterkere enn sammenhengen mellom kontakt og inntekt siden $\hat{u}^{AE}_{11}$ er større enn $\hat{u}^{AD}_{11}$. Men på grunnlag av definisjonen av u-ene er det vanskelig å si, hva det innebærer at $\hat{u}^{AD}_{11}$ er lik 0,088 i stedet for 0,274. I 5b vil vi presentere en synsmåte som gjør det noe lettere.

5b. Sammenhengen mellom odds og parametrene i den log-lineære modell

En summering i tabell 3.1 viser at av 3 661 personer hadde 2 750 kontakt og 911 ikke kontakt med tannlege i løpet av siste året. Oddsen for kontakt 2 750/911 = 3,02. Vi vil skrive dette Odds (A) = 3,02. Videre får vi av tabell 3.1 toveistabellen.

Tabell 5.3. Tannlegekontakt etter kjønn

		Kontakt	
		1	2
Kjønn	1	1 278	543
	2	1 472	368

For menn finner vi at oddsen for kontakt er 1 278/543 = 2,35. Vi skriver Odds (A|B = 1) = 2,35. For kvinner er tilsvarende odds 1 472/368, dvs. Odds (A|B = 2) = 4,00. Sammenligner vi disse betingete oddsene ved å beregne kvotienten mellom dem, finner vi at denne er lik kryssproduktet i tabell 5.3. Dette betegner vi KP (A, B). Vi har KP (A, B) = Odds (A|B = 1)/Odds (A|B = 2).

Dette er odds og kryssprodukt i marginale tabeller, i flerveistabeller vil vi bruke tilsvarende betegnelser. Fra tabell 3.1 har vi bl.a. at Odds (A|B = 1, C = 1, D = 1, E = 1) = 710/208 = 3,41 og Odds (A|B = 2, C = 1, D = 1, E = 1,) = 759/109 = 6,96. Videre er KP (A, B|C = 1, D = 1, E = 1) = Odds (A|B = 1, C = 1, D = 1, E = 1)/Odds (A|B = 2, C = 1, D = 1, E = 1) = (710 · 109)/(759 · 208) = 0,49. KP (A, B|C = 1, D = 1, E = 1) er et betinget kryssprodukt, det er kryssproduktet i deltabellen over kjønn og kontakt når vi ser på dem med lengst utdanning, høyest inntekt og kortest reisetid.

Kryssproduktene ovenfor ble beregnet på grunnlag av observasjonsmaterialet. De har sine teoretiske motstykker i kryssprodukt uttrykt ved de forventede cellefrekvenser. KP (A, B|C = 1, D = 1, E = 1) er et estimat for $(m_{11111} \cdot m_{22111})/(m_{21111} \cdot m_{12111})$ i den mettede modell.

Kryssproduktet har en del egenskaper som gjør det naturlig å bruke som sammenhengsmål i toveistabeller, se Haldorsen (1976).

I kapittel 4 fant vi fram til en modell for dataene, denne er gitt av ligning (5.2). Når modellen gjelder, finner vi for et vilkårlig av de åtte betingete kryssproduktene mellom A og B.

$$(m_{11k\ell m} \cdot m_{22k\ell m})/(m_{21k\ell m} \cdot m_{12k\ell m}) =$$

$$(\tau \tau_1^A \tau_1^B \tau_k^C \tau_\ell^D \tau_m^E \tau_{11}^{AB} \tau_{1k}^{AC} \tau_{1\ell}^{AD} \tau_{1m}^{AE} \tau_{k\ell}^{CD} \tau_{\ell m}^{DE} \cdot \tau \tau_2^A \tau_2^B \tau_k^C \tau_\ell^D \tau_m^E \tau_{22}^{AB} \tau_{2k}^{AC} \tau_{2m}^{AE} \tau_{k\ell}^{CD} \tau_{\ell m}^{DE})/$$

$$(\tau \tau_2^A \tau_1^B \tau_k^C \tau_\ell^D \tau_m^E \tau_{21}^{AB} \tau_{2k}^{AC} \tau_{2\ell}^{AD} \tau_{2m}^{AE} \tau_{k\ell}^{CD} \tau_{\ell m}^{DE} \cdot \tau \tau_1^A \tau_2^B \tau_k^C \tau_\ell^D \tau_m^E \tau_{12}^{AB} \tau_{1k}^{AC} \tau_{1m}^{AE} \tau_{k\ell}^{CD} \tau_{\ell m}^{DE})$$

$$= (\tau_{11}^{AB} \cdot \tau_{22}^{AB})/(\tau_{21}^{AB} \tau_{12}^{AB}) = (\tau_{11}^{AB})^4 \qquad (5.4)$$

Ligningen viser to vesentlige forhold. Når modellen gjelder, så er de åtte kryssproduktene like, og det er en direkte sammenheng mellom kryssproduktet og samspillsparameteren i modellen.

For tilfellet med tre variable har vi vært inne på det samme. Ligning (2.4) viser at hvis trefaktoreffekten er lik null, så vil kryssproduktene mellom to variable være like for ulike nivåer av den tredje variabel. Generelt i flerveistabeller vil kryssproduktet mellom to variable være uavhengig av nivået på de andre variable hvis alle trefaktoreffekter som inkluderer de to, er lik null.

Når vi presenterer resultatene av analysen, kan vi bruke (5.4). I stedet for å gi estimater $\hat{\tau}^{AB} = 0{,}876$ eller $\hat{u}^{AB} = -0{,}133$, gir vi at under modellen er kryssproduktet $(\hat{m}_{11k\ell m} \hat{m}_{22k\ell m})/(\hat{m}_{21k\ell m} \hat{m}_{12k\ell m}) = (0{,}876)^4 = 0{,}589$ uavhengig av (k, ℓ, m). Med ord blir dette at forholdet mellom oddsene for kontakt hos menn og kvinner er det samme for ulike nivåer av utdanning, inntekt og reisetid og er lik 0,589.

For de andre tofaktoreffektene som inkluderer A, får vi for de aktuelle kryssproduktene når modellen gjelder:

$$(\hat{m}_{1j1\ell m} \cdot \hat{m}_{2j2\ell m})/(\hat{m}_{2j1\ell m} \cdot \hat{m}_{1j2\ell m}) = (\hat{\tau}^{AC})^4 = 2{,}223$$

$$(\hat{m}_{1jk1m} \cdot \hat{m}_{2jk2m})/(\hat{m}_{2jk1m} \cdot \hat{m}_{1jk2m}) = (\hat{\tau}^{AD})^4 = 1{,}422$$

$$(\hat{m}_{1jk\ell 1} \cdot \hat{m}_{2jk\ell 2})/(\hat{m}_{2jk\ell 1} \cdot \hat{m}_{1jk\ell 2}) = (\hat{\tau}^{AE})^4 = 2{,}990$$

Forholdet mellom oddsene for de med h.h.v. kort og lang reisetid er 2,223 uansett kjønn, inntekt og utdanning. Eventuelt kan vi uttrykke dette ved at målt med kryssproduktet er sammenhengen mellom reisetid og kontakt lik 2,223 for alle kombinasjoner av kjønn, inntekt og utdanning.

Videre er forholdet mellom oddsene for de med h.h.v. høy og lav inntekt lik 1,422 uansett kjønn, reisetid og utdanning. Til slutt har vi at forholdet mellom oddsene for kontakt for de med h.h.v. lang og kort utdanning er 2,990. Dette gjelder for alle nivå av kjønn, reisetid og inntekt.

Denne presentasjonen av resultatene bygger på Page (1977). Det vesentlige er at den gir en substansiell tolkning av resultatene. Kryssproduktene kan beregnes på to måter. Enten ved hjelp av sammenhengen med samspillsparameteren eller direkte ved hjelp av de estimerte cellefrekvensene for en vilkårlig kombinasjon av de tre andre variable. Ved begge metoder brukes estimater som beregnes under forutsetning av at modellen (5.2) gjelder.

I tabell 5.5 har vi gjengitt estimater for frekvenser, odds og andel med kontakt i vår endelige modell.

Tabell 5.5. Estimater i den endelige modell

Utdanning	Inntekt	Reisetid	Kjønn	Kontakt		Odds for kontakt	% kontakt
				1	2		
1	1	1	1	683,67	188,18	3,63	78
1	1	1	2	787,45	127,53	6,17	86
1	1	2	1	11,38	6,96	1,64	62
1	1	2	2	13,11	4,71	2,78	74
1	2	1	1	384,73	150,37	2,56	72
1	2	1	2	443,13	101,91	4,35	81
1	2	2	1	14,65	12,71	1,15	54
1	2	2	2	16,88	8,61	1,96	66
2	1	1	1	101,66	83,70	1,21	55
2	1	1	2	117,09	56,72	2,06	67
2	1	2	1	1,69	3,09	0,55	35
2	1	2	2	1,95	2,10	0,93	48
2	2	1	1	77,27	90,35	0,86	46
2	2	1	2	89,01	61,23	1,45	59
2	2	2	1	2,94	7,64	0,38	28
2	2	2	2	3,39	5,18	0,65	40

Vi vil sammenligne tabell 5.5 med tabell 3.1 som inneholdt de observerte frekvenser. I tabell 5.5 har vi oppnådd en "glatting" av det observerte materialet. Vi har fått fjernet en del av de uregelmessigheter i tabell 3.1 som en kan anta skyldes tilfeldige variasjoner. Sammenligner en kolonnene med "prosent kontakt", ser en at det er store endringer i rad 3,8 og 11. I disse radene er det få observasjoner og det er stor usikkerhet knyttet til de observerte andelene. Vi har grunnlag for å tro at de estimerte andelene i tabell 5.5 gir oss de beste anslag for de underliggende sannsynligheter.

Dette illustrerer en annen mulig bruk av de loglineære modeller. En kan glatte observasjonsmaterialer der en p.g.a. at en opererer med mange variable og/eller kategorier, har en del celler med relativt få observasjoner. I stedet for å arbeide med observerte cellefrekvenser, kan en bruke de estimerte cellefrekvenser i en redusert modell. I de fleste tilfelle ville en ikke våge å bruke en så redusert modell som vi har her. En ville kanskje nøye seg med en modell der fire- og femfaktoreffektene ble satt lik null.

Siden det er en en-entydig forbindelse mellom odds for kontakt og andel med kontakt kan en presentere resultatene av analysen ved de glattede prosentene i tabell 5.5. Men en bør vise en viss forsiktighet når en tolker disse. De log-lineære modeller medfører at vi måler sammenhengen mellom forspaltevariablene og kontakt ved forholdet mellom odds. Hvis en ønsker å måle sammenheng på annen måte, bør en ta det i betraktning fra starten av. Sammenligner vi f.eks. ved hjelp av prosentdifferanser, ser vi at tabell 5.5 gir (86-78) = 8, (74-62) = 12, (81-72) = 9, ..., (40-28) = 12 for forskjellen mellom kjønnene gitt ulike verdier av de andre variable. Målt på denne måten er det i tabellen forskjell på sammenhengen mellom kjønn og kontakt i de 8 deltabellene. Vi skal ikke her ta stilling til om disse forskjellene er signifikante, vi vil bare peke på at ulike metoder for måling av sammenheng kan føre til ulike konklusjoner.

5c. Tolkning i mer kompliserte modeller

I den endelige modell er det ikke samspillsparametre av orden to eller høyere (trefaktorledd og ledd på høyere nivå). Videre kan hver variabel bare anta to ulike verdier. Begge forhold forenkler presentasjonen av modellen. Vi skal kort se på mulighetene for presentasjon i de mer kompliserte tilfelle.

Da vi vurderte vår endelige modell, testet vi bl.a. om modellen bestemt av marginalene ACD AB AE DE ga signifikant økning i tilpassingen. La oss anta dette hadde blitt vår endelige modell. I denne har vi trefaktoreffekten u^{ACD}. Under modellen får vi følgende estimater, $\hat{u}^{ACD} = -0{,}037$, $SD(\hat{u}^{ACD}) = 0{,}062$, $\hat{u}^{ACD}/SD(\hat{u}^{ACD}) = -0{,}598$ og $\tau^{ACD} = 0{,}964$. Vi minner om at trefaktoreffekter kunne tolkes som avvik mellom tofaktoreffekter i deltabeller av den opprinnelige tabell. Vi har u^{ACD} forskjellig fra null. Tenker vi oss materialet delt i to m.h.t. inntekt (D) betyr det at sammenhengen mellom kontakt (A) og reisetid (C) er forskjellig for de to deltabellene. Siden $\hat{u}^{ACD}$ mindre enn null, vil sammenhengen mellom A og C være mindre blant dem med høy inntekt (D=1) enn blant dem med lav inntekt (D=2). Når her skrives "mindre", tenker en på sammenligning mellom to reelle tall. En vil kanskje foretrekke ulike formuleringer alt etter som sammenhengene mellom kontakt og reisetid for de to inntektsnivå var begge positive, begge negative eller hadde hvert sitt fortegn. Her er begge positive og trefaktoreffekten kan forklares som at for begge inntektsnivå vil lang reisetid være en hindring for tannlegekontakt, men dette gjelder i sterkest grad for dem med lav inntekt.

I tilfellet med trefaktoreffekter kan en også finne en sammenheng mellom ulike kryssprodukt og τ-parametrene. Modellen vi nå forutsetter gjelder, er gitt ved

$$m_{ijk\ell m} = \tau \; \tau_i^A \tau_j^B \tau_k^C \tau_\ell^D \tau_m^E \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{i\ell}^{AD} \tau_{im}^{AE} \tau_{k\ell}^{CD} \tau_{\ell m}^{DE} \tau_{ik\ell}^{ACD} \qquad (5.6).$$

For kryssproduktene mellom A og C får vi da

$$\frac{m_{1j1\ell m} \cdot m_{2j2\ell m}}{m_{2j1\ell m} \cdot m_{1j2\ell m}} = (\tau_{11}^{AC})^4 \cdot (\tau_{11\ell}^{ACD})^4$$

$$= \begin{cases} (\tau_{11}^{AC} \cdot \tau_{111}^{ACD})^4 & \text{når } D = 1 \\ (\tau_{11}^{AC} / \tau_{111}^{ACD})^4 & \text{når } D = 2 \end{cases} \qquad (5.7)$$

De avhenger av nivået til D men er uavhengig av B og E. Når vi skal anslå kryssproduktene under modellen, kan vi sette inn estimater for τ-ene i (5.7) eller vi kan bruke cellefrekvensene som de er estimert under modellen. Vi velger da to vilkårlige deltabeller der D er h.h.v. lik 1 og 2 og beregner kryssproduktene mellom A og C i disse. Vi finner da 1,845 for D=1 og 2,474 for D=2. For de med h.h.v. kort og lang reisetid er forholdet mellom odds for kontakt 1,845 i høyeste inntektsgruppe og 2,474 i gruppen med lavere inntekter. Dette utsagnet er noe komplisert og noen vil foretrekke: "For begge inntektsgrupper er det en positiv sammenheng mellom tannlegekontakt og kort reisetid, dette gjelder i sterkest grad for dem med lav inntekt." Men da får en bare fortalt hvilken vei sammenhengen går. Det blir ikke klart hvor "stor" den er.

Foran har vi snakket om trefaktoreffekten mellom A, C og D som sammenhengen mellom A og C for ulike nivåer av D. Det er symmetri, så vi kunne også se på den som sammenhengen mellom A og D for ulike nivåer av C eller som sammenhengen mellom C og D for ulike A-nivåer. Det siste er dog noe unaturlig i vårt spesielle eksempel.

Skal en tolke effekter av orden fire eller høyere, kan en se på disse som sammenheng mellom to variable gitt ulike nivåer av to eller flere andre variable. Vi mener dette er lite aktuelt. En modell med firefaktoreffekter er så komplisert at den sjelden tilfredsstiller kravet om at en modell skal gi en relativt enkel forklaring på våre data. Får en slike effekter i de endelige modellene, kan det tyde på at data er en blanding fra ulike populasjoner. En bør da søke en naturlig deling av data og lage separate modeller på grunnlag av delmaterialene.

Hvis en eller flere av variablene kan anta mer enn to verdier, blir framstillingen noe mer komplisert, men ikke prinsipielt annerledes. Har vi f.eks. 4 kategorier på én variabel, vil de enkelte u-parametrene der denne inngår være bestemt av tre verdier og kravet om at summen skal være null. Sammenhengen mellom denne variablen og en av de andre vil være bestemt ved tre kryssprodukter. Hvis A har I kategorier og B har J kategorier vil sammenhengen måtte beskrives med $(I-1) \cdot (J-1)$ kryssprodukter. Vi kan velge kryssproduktene $(m_{ij} \cdot m_{IJ})/(m_{Ij} \cdot m_{iJ})$ for $i \leq I - 1$ og $j \leq J - 1$. Hvis A og B inngår i en tabell med flere variable, vil disse kryssproduktene på samme måte som før avhenge/ikke avhenge av nivået på de andre variable alt etter om modellen har/ikke har trefaktoreffekter med A og B i topposkriften.

6. METODE VED STRATIFISERTE UTVALG

I kapitlene foran har rammen vært at n enheter er krysssklassifisert m.h.t. tre eller flere variable. Men har vi data fra stratifiserte utvalg, kan vi som regel også bruke metodene. Vi sørger bare for at de parametre der kun stratifiseringsvariablene inngår (i toppskriftene), er med i de aktuelle modellene. Disse parametrene lar vi forbli uanalysert, vi forsøker f.eks. ikke å teste om noen av dem kan settes lik null. Hvis tallene i tabell 3.1 (gjennomgangseksemplet), var resultat av at vi hadde trukket gitte antall personer fra fire kjønns(B)-utdannings(E)-grupper og hadde klassifisert disse etter kontakt (A), reisetid (C) og inntekt (D), ville vi begrenset oss til modeller der u^{BE}, u^{B} og u^{E} var med. Estimater for disse parametrene ville vi ikke kommentere, de er uttrykk for det antall personer vi på forhånd bestemte å trekke fra kjønns-utdanningsgruppene. Hvis tallet på personer i hver av de seksten forspaltegruppene var bestemt på forhånd, ville vi begrense oss til å se på hierarkiske modeller der u^{BCDE} inngikk. Hvis stratifiseringsplanen er mer komplisert, f.eks. krav på antallet i kjønns-reisetidsgrupper og kjønns-inntektgrupper, må en vurdere spesielt hvilke modeller som er aktuelle og hvordan estimeringen bør foretas, se kapittel 3.2 i Bishop, et.al. (1975).

Som eksempel på analyse av stratifiserte data, skal vi gå gjennom beregningene for eksemplet i tabell 3.1 når vi forutsetter at tallet på personer i alle seksten forspaltegrupper er gitt på forhånd. Vi bruker en baklengsprosedyre med 1 prosent nivå på enkelttestene. Først testes om u^{ABCDE} lik null, det forkastes ikke. Gitt dette testes om alle firefaktoreffekter unntatt u^{BCDE} er lik null, det forkastes ikke. Gitt dette testes om trefaktoreffektene u^{ABC}, u^{ABD}, u^{ABE}, u^{ACD}, u^{ACE}, u^{ADE} alle lik null, det forkastes ikke. Men gitt dette, forkaster vi at tofaktoreffektene u^{AB}, u^{AC}, u^{AD} og u^{AE} alle lik null. Da tester vi om en av dem er lik null, dette forkastes. Vi har foreløpig modellen bestemt av marginalene AB, AC, AD, AE og BCDE. Ingen av trefaktoreffektene som inkluderer A, gir signifikant økning i tilpassingen og dette blir den endelige modell. For effekter som inkluderer A er dette samme resultat som i kapittel 4-5. Av modellen får vi følgende estimater for tofaktoreffektene som inkluderer A (kontakt).

Tabell 6.1. Noen estimater i den endelige modell når marginalen BCDE holdes fast

Variable	$\hat{u}$	SD ($\hat{u}$)	$\frac{\hat{u}}{SD(\hat{u})}$	$\hat{\tau}$
AB	-0,144	0,062	-2,310	0,866
AC	0,174	0,062	2,788	1,189
AD	0,089	0,062	1,432	1,093
AE	0,277	0,062	4,446	1,319

Vi ser at estimatene avviker lite fra de tilsvarende tall i tabell 5.1. I tabell 6.2 finner vi estimater for cellefrekvenser, odds og andeler under modellen.

Tabell 6.2. Estimater i den endelige modell når marginalen BCDE holdes fast

Utdanning	Inntekt	Reisetid	Kjønn	Kontakt		Odds for kontakt	% kontakt
				1	2		
1	1	1	1	716,68	201,32	3,56	78
1	1	1	2	749,53	118,47	6,33	86
1	1	2	1	10,24	5,76	1,78	64
1	1	2	2	15,95	5,05	3,16	76
1	2	1	1	370,38	148,62	2,49	71
1	2	1	2	476,46	105,54	4,43	82
1	2	2	1	14,42	11,58	1,25	55
1	2	2	2	10,33	4,67	2,21	69
2	1	1	1	85,95	73,05	1,18	54
2	1	1	2	135,98	65,02	2,09	68
2	1	2	1	1,11	1,89	0,59	37
2	1	2	2	2,55	2,45	1,04	51
2	2	1	1	75,43	91,57	0,82	45
2	2	1	2	82,59	56,41	1,46	59
2	2	2	1	3,79	9,21	0,41	29
2	2	2	2	7,60	10,40	0,73	42

Tallene for odds og andeler er svært like de vi fant i tabell 5.5. Estimatene for cellefrekvensene avviker en del. I tabell 6.2 summerer de seg til de observerte antall i forspaltegruppene.

For modellen er LL (H) = 8,32. Med 11 frihetsgrader er dette en bra tilpassing. Modellen er ikke mer komplisert enn modellen gitt ved ligning (3.3) og tilpassingen er langt bedre. Modellen i (3.3) fikk vi fra en vurdering av signifikansen til parameterestimatene i den mettede modell. Vi har fått demonstrert at disse estimatene må brukes med forsiktighet.

Når antallet i forspaltegruppene er gitt på forhånd, vil noen formulere modellen annerledes enn vi gjorde. De hadde satt kontakt på venstresiden og de andre variable på høyresiden i en slags regresjonsligning. En måte å gjøre dette på, er å lage modell for oddsen for kontakt (evt. logaritmen til oddsen for kontakt, (logitmodeller)). Vi setter $\omega_{jk\ell m} = m_{1jk\ell m}/m_{2jk\ell m}$ og ser på modeller:

$$\log \omega_{jk\ell m} = v + v^B_j + v^C_k + v^D_\ell + v^E_m + v^{BC}_{jk} + v^{BC}_{j\ell} + v^{BE}_{jm} + v^{CD}_{k\ell} + v^{CE}_{km} + v^{DE}_{\ell m} + v^{BCD}_{jk\ell} + v^{BCE}_{jkm} + v^{BDE}_{j\ell m} + v^{CDE}_{k\ell m} + v^{BCDE}_{jk\ell m} \tag{6.3}$$

der v-er med indeks(er) summerer seg til null m.h.t. hver av indeksene, f.eks. $\sum_j v^{BDE}_{j\ell m} = 0$. For oddsen blir modellen

$$\omega_{jk\ell m} = \upsilon\, \upsilon^B_j\, \upsilon^C_k\, \upsilon^D_\ell\, \upsilon^E_m\, \upsilon^{BC}_{jk}\, \upsilon^{BD}_{j\ell}\, \upsilon^{BE}_{jm}\, \upsilon^{CD}_{k\ell}\, \upsilon^{CE}_{km}\, \upsilon^{DE}_{\ell m}\, \upsilon^{BCD}_{jk\ell}\, \upsilon^{BCE}_{jkm}\, \upsilon^{BDE}_{j\ell m}\, \upsilon^{CDE}_{k\ell m}\, \upsilon^{BCDE}_{jk\ell m} \tag{6.4}$$

der produktet av ledd med indekser over en av indeksene er lik 1, f.eks.

$$\prod_k \upsilon^{BCE}_{jk\ell} = 1.$$

Så lenge en holder seg til sannsynlighetsmaksimeringsestimering vil ikke (6.3) og (6.4) gi andre resultater enn den framgangsmåte vi har beskrevet før i dette kapitlet. Estimatene for v-ene (υ-ene) blir de samme enten vi starter med (6.3) ((6.4)) som modell eller tar utgangspunkt i modellen for log $m_{ijk\ell m}$ ($m_{ijk\ell m}$). Ut fra den siste har vi nemlig:

$\log \omega_{jk\ell m} = \log m_{1jk\ell m} - \log m_{2jk\ell m}$ og

$\omega_{ijk\,m} = m_{1jk\ell m}/m_{2jk\ell m}$ får vi

$$\log \omega_{jk\ell m} = (u^A_1 - u^A_2) + (u^{AB}_{1j} - u^{AB}_{2j}) + \ldots + (u^{ABCDE}_{1jk\ell m} - u^{ABCDE}_{2jk\ell m})$$

$$= 2u^A_1 + 2u^{AB}_{1j} + \ldots + 2u^{ABCDE}_{1jk\ell m} \tag{6.5}$$

og

$$\omega_{jk\ell m} = (\tau^A_1 \cdot \tau^{AB}_{1j} \cdot \ldots \tau^{ABCDE}_{1jk\ell m})/(\tau^A_2 \cdot \tau^{AB}_{2j} \cdot \ldots \cdot \tau^{ABCDE}_{2jk\ell m})$$

$$= (\tau^A_1)^2 \cdot (\tau^{AB}_{ij})^2 \cdot \ldots \cdot (\tau^{ABCDE}_{1jk\ell m})^2 \tag{6.6}$$

u-ledd og τ-ledd uten A som toppskrift faller vekk. v-ene (υ-ene) blir entydige funksjoner av u-ene (τ-ene).

I kapitlene 2-5 tok vi utgangspunkt i modeller for log $m_{ijk\ell m}$ (evt. $m_{ijk\ell m}$). I kommentarene til eksemplet fant vi det delvis naturlig å skille mellom variablene. Vi beskrev hvordan variablene kjønn, reisetid, inntekt og utdanning virket inn på kontakten med tannlege. Kontakt var responsvariabel og de andre faktorvariable.

Vi tok ikke hensyn til deiingen i respons- og faktorvariable i estimeringen. Goodman (1971) synes å mene at en skal ta hensyn til dette skillet uansett hvordan data er samlet inn. Selv om data var en krossklassifisering m.h.t. fem variable i et tilfeldig utvalg, ville han holde marginalen BCDE fast under letingen etter modell. En gjennomfører da en betinget analyse gitt de observerte antall i faktorgruppene. Vi skjønner ikke helt dette. Intuitivt vil en oppnå bedre celleestimater hvis en med rette setter noen av samspillene mellom faktorvariablene lik null under estimeringen, som videre burde føre til bedre estimater for de andre samspillsparametrene. Vi har sett at i eksemplet blir det liten forskjell i interessente estimater for de to metodene. Bishop (1969) inneholder momenter om forholdet mellom logitmodeller og generelle log-lineære modeller.

Noen vil hevde at siden kontakt er respons og de andre faktorvariable så bør en se på v-ene i (6.3) uansett innsamlingsmetode. Vi ser ikke noe galt i det. Det gjelder bare å avgjøre hvordan de best estimeres når vi tar hensyn til innsamlingsmetoden. Ut fra den lager vi en modell og v-ene vil være funksjoner av parametrene i denne. Når vi først har definert en passende modell er vi fri til å se på funksjoner av de opprinnelige parametre, og til å gi disse og de opprinnelige parametre de tolkninger vi finner interessante.

7. Å SE PÅ AGGREGERTE TABELLER

Når en skal undersøke et forhold og har tatt med de fleste variable en vet eller tror har betydning, finner en ofte at det blir få eller ingen observasjoner i mange av cellene i den flerveistabell en vil arbeide med. Det er da aktuelt å undersøke om en kan oppnå de samme resultater med aggregerte tabeller, der en ser bort fra én eller flere variable. På side 47 i Bishop et.al. (1975) finner en eksakte regler for dette når en arbeider innenfor rammen av log-lineære modeller.

Som nevnt i kapittel 3 startet vi analysen av tannlegekontakter med seks variable. Materialet var delt i to aldersgrupper. Vi anså at det viktigste var å undersøke forholdet mellom kontakt og de fem forklaringsvariable. Hvis vi skulle se på en aggregert tabell, måtte denne gi et korrekt bilde av disse forholdene. Theorem 2.5-1 i Bishop et.al. (1975) krever da at vi undersøker om noen av parametrene u^{AB}, u^{AC}, u^{AD}, u^{AE}, u^{AF} kan settes lik null (F står for alder). Materialet tillot at vi satte u^{AF} lik null. Teoremet sikrer da at alle parametre der A inngår som en av toppskriftene, vil bli korrekt avbildet i femveistabellen der en har sett bort fra alder (F). Vi kontrollerte ikke hvilke av de andre parametrene som ville bli korrekt avbildet, da dette er av mindre betydning.

Det er mange mulige tester av hypotesen, $u^{AF} = 0$. Madsen (1976) påpeker at med fire variable vil en tilsvarende hypotese kunne testes i 41 ulike sammenhenger. Med seks variable vil det være langt flere. Madsen (1976) viser i et eksempel at så lenge en holder seg til modeller som beskriver data godt får en relativt store verdier på testobservatoren og det blir liten forskjell mellom alternative metoder. En konservativ prosedyre vil da være å holde seg til tester i modeller med mange parametre. Oppnår vi ikke forkasting av hypotesen "$u^{AF} = 0$" i disse, har vi god grunn til å anta at u^{AF} er lik null i den underliggende modell.

Forslagene til endelige modeller i kapittel 4 og 6 inneholder bare to-faktorledd m.h.t. variabel A (kontakt.) Dette betyr ikke at sammenhengen mellom A og de andre fire variable kan avbildes i fire separate toveistabeller. Det ser en av theorem 2.5-1 i Bishop et.al. (1975). Vi kan også demonstrere det med estimater for de aktuelle kryssprodukter.

Tabell 7.1. Estimater i tre ulike situasjoner

Kryssprodukter	Estimert i endelig modell kapittel 4	Estimert i endelig modell kapittel 6	Estimert i i fire separate toveis-tabeller
$\hat{\delta}^{AB}$	0,59	0,56	0,87
$\hat{\delta}^{AC}$	2,22	2,00	2,41
$\hat{\delta}^{AD}$	1,42	1,43	1,54
$\hat{\delta}^{AE}$	2,99	3,03	3,08

Vi ser at selv i dette enkle materialet der responsvariablen inngår bare i to-faktoreffekter og det er relativt enkle sammenhenger mellom faktorvariablene, så fører det galt avsted å studere toveis-tabeller framfor å foreta en simultan analyse. Med toveistabeller vil en undervurdere samspillet mellom kontakt (A) og kjønn (B), og en vil overvurdere samspillet mellom kontakt og inntekt (C).

REFERANSER

[1] Bishop, Yvonne M. M. (1969): "Full contingency tables, logits, and split contingency tables" Biometrics 25: 383-399.

[2] Bishop, Yvonne M. M., Stephen E. Fienberg and Paul W. Holland (1975): "Discrete Multivariate Analysis" The MIT Press, Cambridge, Mass.

[3] Goodman, Leo A. (1971): "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Metods for Building Models for Multiple Classifications." Tecnnometrics Vol. 13, No. 1: 33-61

[4] Haldorsen, Tor (1976): "Forelesninger om avhengighetsmål i kontingenstabeller av universitetslektor Harald Goldstein" Statistisk Sentralbyrå, Arbeidsnotat IO 76/27.

[5] Holst, Dorthe (1977): "Levekårsundersøkelsen 1973. Noen odontologiske resultater" Den norske tannlægeforenings tidende. 87, feb.: 68-74.

[6] Lee, S. Keith (1977): "On the Asymptotic Variances of $\hat{u}$ Terms in Log-linear Models of Multidimensional Contingency Tables" J. Amer. Statist. Assoc. Vol. 72, No. 358: 412-419.

[7] Madsen, Mette (1976): "Statistical Analysis of Multiple Contingency Tables. Two Examples." "Scand. J. Statist., Vol. 3, No. 3: 97-106.

[8] Page, William F. (1977): "Intepretation of Goodman's Log-linear Model Effects "Sociological Methods & Research, Vol. 5, No. 4: 419-435.

[9] Statistisk Sentralbyrå (1977): "Helseundersøkelse 1975" NOS A894