

Om analyse af kategorielle data: En kommentar til en kritisk kommentar

Ole P. Kristensen

Kjeld Møller Pedersen (KMP) har i *Økonomi og Politik* kritiseret Socialforskningsinstituttets levevilkårsundersøgelse.¹⁾ Jeg vil i det følgende kritisere de kritiske kommentarer, der vedrører forskningsteknikken. Kritikken af kritikken skal ikke opfattes som et forsvar for den forskningsteknik, der er anvendt i levevilkårsundersøgelsen, da den formentlig i en række henseender er kritisabel. Det vil jeg blot ikke gå nærmere ind på her. Når jeg føler anledning til at kommentere KMP's kritik, hænger det sammen med, at den er præget af mangel på præcision, unøjagtigheder og regulære fejl. Dette er yderst uheldigt, når man er ude i det fortrinlige pædagogiske ærinde at forbedre samfundsforskeres analyser af nominalskaladata.

Kritikken af tabelteknikken

KMP indleder sin gennemgang af analyseteknikken i levevilkårene med en generel diskussion af tabellers hensigtsmæssighed som analyseteknik til behandling af nominalskalavariabel. Til det formål opstiller KMP en række tænkelige argumenter for anvendelse af tabelteknikken (p. 101), hvoraf det vigtigste tænkelige argument imidlertid mangler: på nominalskalaniveauet kan de analytiske muligheder rent logisk fuldt udtømmes ved hjælp af tabelteknik. Det er bandt andet derfor, der i Rosenbergs bog²⁾ (som KMP refererer til) alene er anvendt tabelteknik, da bogen, som titlen siger, alene vedrører analyselogik, ikke analyseteknik. Følgelig er KMP's henvisning (p. 101) til dette værk i forbindelse med »traditionsargumentet« til støtte for tabelteknik temmeligt malplaceret. At analysemulighederne logisk kan udtømmes med tabelteknik på nominalskalaniveauet forhindrer selvsagt ikke, at andre teknikker kan være mere hensigts-

1) Kjeld Møller Pedersen, »Fordelingen af levevilkårene i Danmark – En kritisk vurdering af og kommentar til en undersøgelse«, *Økonomi og Politik*, 55 pp. 77–116.

2) Morris Rosenberg, *The Logic of Survey Analysis*, New York og London: Basic Books, 1968.

mæssige i mange tilfælde, men den potentielle gevinst ved at gå fra tabelteknik til mere »avanceret« teknik er mindre på nominalskalaniveauet end på højere måleniveauer, fordi der ikke er indeholdt information i måleresultaterne på nominalskalaniveauet, som ikke udnyttes ved tabelteknik, som det er tilfældet på højere måleniveauer. Dette er formentlig den egentlige forklaring på tabelteknikkens hyppige anvendelse i samfundsvidenskaberne til analyse af nominalskaladata. Samfundsforskerne holder sig nemlig ikke tilbage med anvendelse af »avancerede« teknikker på intervalskaladata, heller ikke i tilfælde hvor deres manglende indsigt i disse teknikker egentlig skulle mane til tilbageholdenhed.

Dernæst gennemgår KMP, hvilke krav der kan stilles til analyseteknikken, hvilket fører frem til tre typer spørgsmål (p. 104), der postuleres ikke at kunne besvares via tabelteknik. Det er imidlertid ikke korrekt; en række af disse spørgsmål vil naturligvis kunne besvares allerede ved hjælp af simpel tabelteknik (jvf. for eksempel Helleviks artikel og Rosenbergs bog.³) Dernæst argumenteres for, at analysen skal være multivariat for at sikre, at de fundne sammenhænge ikke er spuriøse (pp. 104–105). Det er naturligvis så sandt, som det er sagt; denne kritik rammer muligvis levevilkårsundersøgelsen, men bestemt ikke tabelteknik som sådan, det vil man let kunne overbevise sig om ved at læse Rosenbergs bog.⁴) KMP foretrækker imidlertid at illustrere tankegangen bag multivariat analyse ved hjælp af en lineær regressionsmodel (p. 105). Det er naturligvis acceptabelt (men tankegangen kan som nævnt fuldt ud illustreres med mange dimensionale tabeller); men det er imidlertid mindre acceptabelt, når KMP (s. 105) skriver: »men i de fleste tilfælde gælder blot, at β 's størrelse falder, når der kontrolleres for C«. For det første mener KMP formentlig, at den numeriske værdi af β falder ved kontrol, for det andet er det en alt for snæver opfattelse af spuriøse relationer alene at relatere dem til det tilfælde, hvor en ikke eksisterende relation på grund af forstyrrende tredjevariable ser ud til at eksistere. Et lige så interessant og hyppigt forekommende tilfælde er, at eksisterende relationer bliver »undertrykt« af tredjevariable.⁵) I regressionsprog kan kontrol for tredjevariable føre til, at den numeriske værdi af β både kan stige og falde, og det kan β selv i øvrigt også.

I slutningen af dette afsnit hævdes det endvidere: »Denne tankegang (tankegangen bag regressionsanalyse) indebærer vurdering af effekt/virkning af de uafhængige variable...«. Det er næppe rigtigt. Der er et væsentligt logisk spring fra at have påvist variabelsammenhænge, som modstår kontrol for tredjevariabel til at påstå, at der er tale om et effekt/virkingsforhold, eller et årsagsforhold om man vil. Disse overvejelser omkring multivariat analyse kan føre til en lang filosofisk diskussion, som jeg ikke vil indlade mig på her. Blot vil jeg

3) Ottar Hellevik, »Kausalanalyse ved hjælp av multivariate prosenttabeller«, *Tidsskrift for Samfunnsforskning*, 19, 1978, pp. 301–333; Morris Rosenberg, *op.cit.*

4) *Ibid.*

5) *Ibid.*, kapitel 4.

slå fast, at disse problemer lige så vel kan belyses med multivariat tabelteknik som med multivariat regressionsanalyse.

Alternative analysestrategier

Jeg er enig med KMP i, at det alligevel kan være umagen værd at se sig om efter alternative analyseteknikker. Det interessante tilfælde er det, hvor også den afhængige variabel er en nominalskalavariabel. I de tilfælde, hvor den afhængige variabel måles på intervalskalaniveau, er der en række alternativer, der er velkendte og hyppigt anvendte.

Blandt de foreliggende teknikker til analyse af rene kategorielle data har KMP valgt en teknik, som er baseret på »vægtede mindste kvadraters« metode, en metode som er velkendt fra den almindelige regressionsanalyse.⁶⁾ Denne teknik kaldes, når den anvendes på nominalskaladata, for GSK teknikken efter dens ophavsmænd: Grizzle, Starmer og Koch.⁷⁾ Denne teknik består imidlertid af en hel familie af teknikker, da den rummer mulighed for forskelligartede transformationer af de relative tabelfrekvenser, der er de egentlige afhængige variabler for analysen. Hovedsondringen går mellem lineære og logaritmiske transformationer af disse frekvenser. KMP har valgt en lineær transformation, hvilket formentlig skyldes, at man så kan give de estimerede β -koefficienter en substantiel fortolkning (mere herom nedenfor), hvilket ikke er tilfældet, hvis der foretages logaritmiske transformationer. Lineære transformationer er imidlertid problematiske, fordi de kan føre til estimater på sandsynligheder, som er negative eller større end én. På den baggrund anbefales det normalt at foretage en logaritmisk transformation.⁸⁾ KMP's motiver for at vælge en lineær transformation fremgår ikke af hans indlæg. Hvis motivet er tolkbare $\hat{\beta}$ -koefficienter, kan det undre, at KMP ikke er ganske klar i sin beskrivelse af fortolkningen af $\hat{\beta}$ -koefficienterne. Korrekt fortolket skal de opfattes som »procentdifferencer« af den type, der er velkendt fra simpel tabelanalyse.⁹⁾ Ved den af KMP anvendte udgave af GSK-teknikken tages en »reference gruppe« som udgangspunkt, hos KMP er denne gruppe kvinder i alderen 60–69. Det konstantled, der estimeres ved den kategorielle regression, skal fortolkes som sandsynligheden for, at medlemmer af referencegruppen har en bestemt værdi på den afhængige variabel, i dette tilfælde sandsynligheden for at 60–69 årige kvinder har varige helbredsproblemer. De øvrige $\hat{\beta}$ -koefficienter kan nu fortolkes

6) Se for eksempel J. Johnston, *Econometric Methods*, 2. ed., New York: McGraw-Hill, 1972, kapitel 7.

7) Teknikken er først beskrevet i James E. Grizzle, C. Frank Starmer and Gary G. Koch, »Analysis of Categorical Data by Linear Models«, *Biometrics*, 25, 1969, pp. 489–504.

8) Se for eksempel Hans Theil, »On the Estimation of Relationships Involving Qualitative Variables«, *American Journal of Sociology*, 76, 1970, pp. 103–154.

9) Om procentdifferencer, se for eksempel Hellevik, *op.cit.*

som ændringen i denne sandsynlighed, når man bevæger sig fra referencegruppen til en anden gruppe svarende til den eller de pågældende $\hat{\beta}$ -koefficienter. Eksempelvis er ifølge analyseresultatet sandsynligheden for, at 20–29 årige kvinder har varige helbredsproblemer: $(\text{konstantleddet} + \hat{\beta}_1) = 0.52 + (-0.37) = 0.15$, en sandsynlighed som også fremgår direkte af tabel 6.A.1. Det fremgår på ingen måde klart af KMP's artikel, at den ovenstående fortolkning er den rette fortolkning. Adskillige gange anvendes en formulering, der med lidt god vilje kan læses som værende i overensstemmelse med den rette fortolkning. Men på side 109 går det galt; her anvendes vendingen »... nok er 0.49 gange mindre end for kvinder...« i forbindelse med omtalen af $\hat{\beta}_5$. Denne formulering kan kun forstås i retning af, at en sandsynlighed skal multipliceres med 0.49. Det er midlertid forkert; 0.49 skal subtraheres eller, for at være helt præcis, (-0.49) adderes til en sandsynlighed. Det er dog forståeligt, at KMP er blevet upræcis netop på dette sted, for estimatet på β_5 er indlysende forkert. Af tabel 6.A.1. fremgår det, at den gennemsnitlige procentdifference mellem mænd og kvinder er -1% , og den højeste negative afvigelse i nogen undergruppe er -3% (for 60–69 årige), og i en enkelt underkategori er procentdifference endda positiv. På den baggrund er et estimat, der angiver en difference på -49% åbenbart forkert. Hvis man herefter er i tvivl om, at der er noget galt, kan man med udgangspunkt i de estimerede β -koefficienter prøve at beregne den estimerede sandsynlighed for, at 20–29 årige mænd har varige helbredsproblemer; den er: $(\text{konstantleddet} + \hat{\beta}_1 + \hat{\beta}_5) = 0.52 + (-0.37) + (-0.49) = -0.34$. Ovenfor omtalte jeg, at den af KMP anvendte metode kunne føre til negative sandsynligheder, men dog slet ikke så grelt som det er tilfældet her; det er åbenbart, at noget er galt. Denne mistanke styrkes yderligere af det forhold, at $\hat{\beta}_5$ ikke er signifikant (dvs. signifikant forskellig fra nul). Det forekommer overraskende, at det numeriske største af β -estimerne ikke skulle være signifikant, mens de andre og numerisk mindre koefficienter er det. Der er således tre grunde til, at KMP burde have sagt sig selv, at estimatet på β_5 er forkert. Disse grunde fik forfatteren af disse linier til at gøre KMP's analyse efter.¹⁰⁾

Denne analyse bekræfter, at det af KMP anførte estimat på β_5 er forkert med en faktor 10^{-2} . Årsagen til denne fejl er åbenbart, at KMP fejllæser koefficienterne i edb-outputtet. Disse koefficienter er skrevet i såkaldt exponentiel form, dvs. for eksempel som $0.49E-02$, hvilket betyder 0.49×10^{-2} . Dvs. det rette estimat af β_5 er af størrelsesordenen 0.005 og ikke som af KMP anført 0.5. Hermed falder brikkerne på plads. Den samme fejl går igen i KMP's tabel 2; her er en række koefficienter ikke blevet multipliceret med 10^{-1} . Også i dette

10) Jeg vil gerne takke min kollega Peter Nannestad, der beredvilligt stillede sin udgave af det såkaldte NONMET-program til min rådighed.

tilfælde er der forhold, som skulle vække mistanke. Sådant som interaktionen er mellem køn og alder, skulle man forvente, at den numeriske værdi af hovedeffekten fra køn skulle stige, når interaktionseffekterne introduceres, og ikke falde som hos KMP. Dertil kommer, at ikke mindst KMP's estimat på interaktionseffekten mellem køn og aldersgruppen 40–49 er absurd høj på baggrund af tabel 6.A.1. Disse og andre mærkværdigheder forsvinder, når man læser edb-outputtet korrekt.

Mine resultater er angivet i tabel 1 og 2 svarende til KMP's tabel 1 og 2.

Tabel 1. Estimerede hovedeffekter for alder og køn

	Konstantled	$\hat{\beta}_1$ 20–29 år	$\hat{\beta}_2$ 30–39 år	$\hat{\beta}_3$ 40–49 år	$\hat{\beta}_4$ 50–59 år	$\hat{\beta}_5$ Mænd
Estimat	0.518	–0.371	–0.326	–0.238	–0.115	–0.005
X ² værdi	820.2	345.5	251.7	108.9	23.6	0.2
p-værdi	0.0000	0.0000	0.0000	0.0000	0.0000	0.6392

Note: Det ses, at der er små afvigelser fra KMP's X² værdier. Dette skyldes formentlig, at jeg ikke har haft adgang til den absolute fordeling i tabel 6.A.1., men kun den relative fordeling. Jeg har derfor måttet regne baglæns fra den relative fordeling til den absolute. Denne fremgangsmåde rummer mulighed for små afvigelser fra den sande absolute fordeling.

I øvrigt kan det undre, at KMP kun har opgivet X² værdier. Hvis man ikke kan huske X² fordelingen uden ad, er det ikke særligt oplysende. P-værdierne er mere oplysende.

Tabel 2. Hoved- og samspilseffekter for alder og køn

	Konstant	Hovedeffekter				Mænd M
		20–29 år	30–39 år	40–49 år	50–59 år	
Estimat	0.530	–0.380	–0.331	–0.280	–0.120	–0.031
X ² værdi	508.4	192.7	133.0	81.2	13.5	0.8
p-værdi	0.0000	0.0000	0.0000	0.0000	0.0002	0.3682

	Samspilseffekter			
	20–29×M	30–39×M	40–49×M	50–59×M
	0.021	0.013	0.090	0.013
	0.3	0.1	4.0	0.1
	0.6053	0.7591	0.0465	0.7864

Note: Samme bemærkninger som til tabel 1.

Tilbage står spørgsmålet, om det generelt er umagen værd at anvende GSK-teknikken. Svaret må være bekræftende i en række tilfælde. Blot skal man gøre

sig klart, at hovedgevinsten ikke er de estimerede β -koefficienter; de siger ikke noget nyt i forhold til, hvad der direkte fremgår af tabellen. Hovedgevinsten er, at man får testet, om de hoved- og interaktionseffekter, man kan se af tabellen, er signifikante. Det er derfor at skyde gråspurve med kanoner, når KMP på side 109, efter at have gennemført den første kategorielle regression, konkluderer: »Der er således noget, der peger på, at der forekommer samspil mellem alder og køn«. Det fremgår allerede klart af tabel 6.A.1., at der er et sådant samspil; det eneste spørgsmål, der er tilbage efter at have kastet et blik på tabellen, er, om dette samspil er signifikant. Her kan GSK-analysen give et bekræftende svar.

Hvis nogen skulle have fået lyst til at sætte sig yderligere ind i teknikker til analyse af nominalskala-data, kan jeg anbefale en lille bog af H. T. Reynolds,¹¹⁾ som er yderst let læselig. Den behandler imidlertid alene de såkaldte log-lineære modeller. For en nogenlunde let læselig introduktion til teknikker baseret på »vægtede mindste kvadrater«, som for eksempel GSK-teknikken, vil jeg anbefale to artikler af Lehen og Koch samt Kritzer.¹²⁾

Men især for »mindstekvadrater«-teknikkerne må man gøre sig klart, at det kræver en vis indsats at sætte sig grundigt ind i dem. Indtil man er virkelig fortrolig med teknikkerne, er det nok klogest at holde sig til tabellerne.

11) H. T. Reynolds, *Analysis of Nominal Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-007, Beverly Hills and London: Sage Publications, 1977.

12) Robert G. Lehen and Gary G. Koch, »A General Linear Approach to the Analysis of Non-metric Data: Applications for Political Science«, *American Journal of Political Science*, 18, 1974, pp. 283-313; Hervert M. Kritzer, »An Introduction to Multivariate Contingency Table Analysis«, *American Journal of Political Science*, 22, 1978, pp. 187-226.