Om multicollinearitetsproblemet

Peter Toft-Nielsen

Statistisk Institut, Københavns Universitet

SUMMARY. *Many economists have in their attempts to explain the interaction amongst economic variables used the method of multiple regression analysis. If there by accident or systematically should be a very close relationship amongst the explanatory variables there is said to be a high degree of multicollinearity. As already shown by Ragnar Frisch (1934) this can lead to situations where the classical analysis indicates significant regression coefficients although a partial interpretation is nonsense. To avoid this a method containing two steps is suggested. In the first step the presence of multicollinearity is tested and the second step is an application of the bunch-map analysis developed by Frisch. Finally an example, taken from the thesis of Erling Olsen (1971), is examined.*

I de senere år har økonomer i forsøget på at forklare sammenhængen mellem forskellige økonomiske variable i vid udstrækning benyttet sig af regressionsanalysen med flere forklarende variable.

Idet den afhængige variabel benævnes X_1 kan problemet være at bestemme parametrene i følgende relation:

$$X_{1t} = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + e_t \qquad \text{for } t = 1, 2, \ldots, T \tag{1}$$

hvor T er antallet af observationer, x_{2t} og x_{3t} de forklarende variable målt som afvigelsen fra deres middeltal og e_t det stokastiske led.

For at få gode estimater af parametrene β_1, β_2 og β_3 må en række forudsætninger være opfyldt. Her skal dog alene fremhæves, at selve beregningen kræver, at ingen af de forklarende variable må kunne skrives som lineære funktioner af de øvrige. Hvis dette er tilfældet kan estimationen af de enkelte regressionskoefficienter ved mindste kvadraters metode ikke gennemføres, ligesom de forskellige EDB standardprogrammer ikke kan benyttes.

I fig. 1a er relationen (1) afbildet som et plan i 3 dimensioner. Ved mindste kvadraters metode estimeres et plan P, således at afstanden fra de T observa-

---

Artiklen er tildelt Zeuthen-prisen. Bedømmelsesudvalget har bestået af Else Zeuthen, Jon Stene, Jan Rasmussen samt tidsskriftets redaktør.

Om multicollinearitetsproblemet

Peter Toft-Nielsen

Statistisk Institut, Københavns Universitet

SUMMARY. *Many economists have in their attempts to explain the interaction amongst economic variables used the method of multiple regression analysis. If there by accident or systematically should be a very close relationship amongst the explanatory variables there is said to be a high degree of multicollinearity. As already shown by Ragnar Frisch (1934) this can lead to situations where the classical analysis indicates significant regression coefficients although a partial interpretation is nonsense. To avoid this a method containing two steps is suggested. In the first step the presence of multicollinearity is tested and the second step is an application of the bunch-map analysis developed by Frisch. Finally an example, taken from the thesis of Erling Olsen (1971), is examined.*

I de senere år har økonomer i forsøget på at forklare sammenhængen mellem forskellige økonomiske variable i vid udstrækning benyttet sig af regressionsanalysen med flere forklarende variable.

Idet den afhængige variabel benævnes X_1 kan problemet være at bestemme parametrene i følgende relation:

$$X_{1t} = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + e_t \qquad \text{for } t = 1, 2, \ldots, T \tag{1}$$

hvor T er antallet af observationer, x_{2t} og x_{3t} de forklarende variable målt som afvigelsen fra deres middeltal og e_t det stokastiske led.

For at få gode estimater af parametrene β_1, β_2 og β_3 må en række forudsætninger være opfyldt. Her skal dog alene fremhæves, at selve beregningen kræver, at ingen af de forklarende variable må kunne skrives som lineære funktioner af de øvrige. Hvis dette er tilfældet kan estimationen af de enkelte regressionskoefficienter ved mindste kvadraters metode ikke gennemføres, ligesom de forskellige EDB standardprogrammer ikke kan benyttes.

I fig. 1a er relationen (1) afbildet som et plan i 3 dimensioner. Ved mindste kvadraters metode estimeres et plan P, således at afstanden fra de T observa-

---

Artiklen er tildelt Zeuthen-prisen. Bedømmelsesudvalget har bestået af Else Zeuthen, Jon Stene, Jan Rasmussen samt tidsskriftets redaktør.

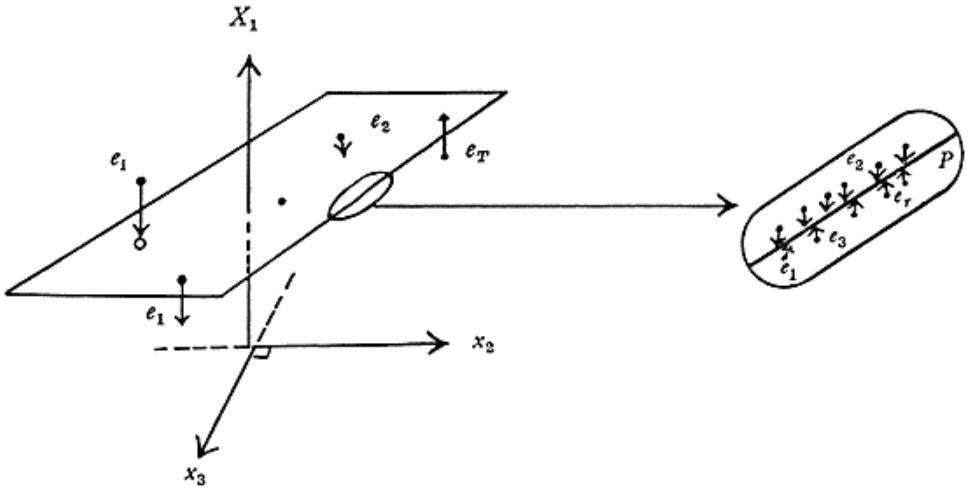

FIG. 1A. *Grafisk fremstilling af* $X_{1t} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e_t$ *samt snit gennem planet* P.

tioner til planet er mindst mulig. Af tværsnittet ses det, at afstanden er minimeret i X_1's retning, dvs. lodret. Minimaliseringsretningen determineres af hvilken variabel, man betragter som afhængig.

Som et kvalitetsmål for den fundne relation benyttes traditionelt determinationsgraden R^2 defineret som forholdet mellem den del af variationen i den afhængige variabel X_1 som forklares af x_2 og x_3 og den samlede variation i X_1.

Økonomer er som regel interesserede i at tolke de estimerede koefficienter b_2 og b_3 hver for sig d.v.s. som de partielle bidrag fra henholdsvis x_2 og x_3 til forklaring af X_1. Det må hertil bemærkes, at en sådan partiel fortolkning ofte er misvisende, fordi størrelsesforholdet mellem b_2 og b_3 afhænger af samvariationen mellem x_2 og x_3. Det er kun i de tilfælde, hvor x_2 og x_3 er ukorrelerede, at b_2's størrelse er uafhængig af størrelsen af b_3. Hvis derimod x_2 og x_3 for eksempel er positivt korrelerede, vil b_2 og b_3 være negativt korrelerede.

Man beregner sædvanligvis standardafvigelsen til hver enkelt regressionskoefficient. Derefter udføres et test for om den beregnede parameter er signifikant forskellig fra nul i de tilfælde, hvor man er interesseret i at undersøge, om den pågældende variabel har nogen betydning for forklaringen af variationen i den afhængige variabel. Hvis dette test falder positivt ud, antages det ofte, at en partiel fortolkning er tilladelig. Dette er imidlertid farligt. Hvis nemlig to eller flere af de forklarende variable tilnærmelsesvist er lineært afhængige, men hvor afrundings-, index- eller regnefejl bevirker, at selve udregningen af parametrene kan gennemføres, således at man ikke opdager denne sammenhæng mellem de forklarende variable, kan den klassiske metode lede til upålidelige eller i værste fald til nonsensestimationer.

Herom skrev Ragnar Frisch allerede i 1934: »In practice, particularly in the social sciences, these cases are apt to arrive much more frequently than is usually recognized. As a matter of fact I believe that a substantial part of the regression and correlation analysis which have been made on economic data in recent years is nonsens for this very reason«.

Med EDB teknikkens udvikling kan det ofte virke fristende at udføre regressionsanalyser med flere og flere variable, men herved øges også risikoen for multicollinearitet.

Om multicollinearitet

Hvis to eller flere forklarende variable er korrelerede er det vanskeligt at separere deres indflydelse på den afhængige variabel X_1. Hvis korrelationen er høj tales der om multicollinearitetsproblemet. Lad os igen illustrere dette grafisk.

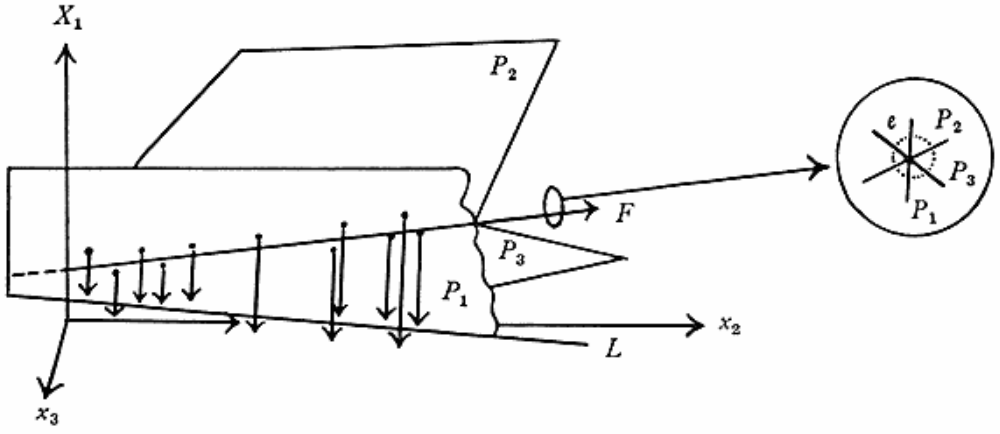

FIG 1B. *Grafisk fremstilling af »fuldstændig« multicollinearitet i regressionsmodellen* $k = 3$ *samt tværsnit af linien F.*

Af fig. 1b ses det, at enhver projektion af observationerne i $X_1x_2x_3$ rummet ned på planet x_2x_3 vil kunne henføres til den rette linie L. Ved forklaringen af X_1 mistes derfor én dimension. Det bedste skøn over X_1 er ikke et plan, men den rette linie F. Da regressionskoefficienterne stadig kan beregnes sålænge det lineære sammenhæng er ufuldstændigt, medfører den indbyrdes korrelation, at man må opgive at fortolke b'erne hver for sig.

Hertil vil nogle økonometrikere stadig hævde, at estimatorerne kan anvendes til forudsigelsesformål, selvom en partiel fortolkning er udelukket som følge af multicollinearitet. De vil sige, at så længe man holder sig til P_1 (se fig. 1b), opstår der intet problem. En sådan konklusion må dog ses i lyset af, at dette kun

er tilfældet, hvis sammenhænget mellem x_2 og x_3 ikke er tilfældigt men systematisk koncentreret omkring linien L i hele forudsigelsesperioden.

I modsætning til fig. 1a, hvor planets placering i rummet var stabil, er det nu principielt muligt at lægge uendelig mange regressionsplaner gennem F. Planet P kan derfor frit rotere omkring F afhængig af minimaliseringsretningen. Man kan derfor konkludere, at et plan i tilfælde af multicollinearitet ikke kan gives nogen fornuftig fortolkning.

Et naturligt spørgsmål må derfor blive: Har vi i de to tekniske størrelser, determinationsgraden og spredningen på de estimerede parametre, der jo traditionels opfattes som indikatorer for den estimerede regressionslignings pålidelighed, en sikkerhed for, at »nonsensestimater« altid bliver opdaget?

Lad os først se på determinationsgraden defineret ved R^2. Hvis målsætningen med den estimerede regressionsligning er at kunne anlægge en partiel fortolkning, må det rent intuitivt være klart, at R^2 er et upålideligt mål. Hvis man nemlig til et ukorreleret sæt af forklarende variable $\mathbf{x} = (x_2, \ldots, x_i, \ldots, x_k)$ tilføjer en ny variabel x_{k+1}, der på nær afrundingsfejl og lignende kan beskrives som en lineær funktion af $\mathbf{x}$, vil R^2 trods eksistensen af næsten fuldkommen multicollinearitet ikke aftage, idet den forklarede variation ikke vil falde.

Man må derfor håbe, at standardafvigelserne på de estimerede regressionskoefficienter kan give et klart signal.

Estimatet for standardafvigelsen, SD, til den j'te regressionskoefficient skrives som:

$$[SD(b_{1j})]^2 = \frac{1}{T-3} \frac{(1-R^2)}{(1-r^2_{23})} \frac{s^2_1}{s^2_j} \text{ for } j = 2,3 \qquad (2)$$

hvor s^2_1 og s^2_j er de empiriske varianser for x_1 og x_j hvor $j = 2,3$.
r^2_{23} er kvadratet på den simple empiriske korrelationskoefficient mellem x_2 og x_3.

I jo højere grad x_2 og x_3 er korrelerede jo nærmere vil r^2_{23} være ved 1, således at nævneren i (2) er næsten 0. Dette resulterer ceteris paribus i at SD går mod uendelig. De estimerede regressionskoefficienter vil derfor ikke blive fundet signifikant forskellig fra nul og en fejlagtig partiel fortolkning vil blive undgået. Under forudsætning af, at antallet af frihedsgrader er positivt, vil det ofte vise sig, at de to forklarende variable (x_2,x_3) simultant giver en god beskrivelse af x_1, således at determinationsgraden R^2 ligeledes er næsten 1. Herved vil også tælleren i (2) være næsten 0. Dette forhold trækker altså i retning af at gøre SD mindre. (2) kan da blive en upålidelig indikator for, om det er forsvarligt at anlægge en partiel fortolkning. Helt grelt bliver det, hvis man hypotetisk an-

tager, at fejl – eller direkte regnefejl skjuler denne korrelation. Herved bestemmes den absolutte størrelse af standardafvigelserne udelukkende af afrundingsfejlene.

Hvis man mindre ambitiøst ønsker at anlægge en svag fortolkning af de estimerede regressionskoefficienter, dvs. blot er interesseret i fortegnene til b_2 og b_3, kan eksistensen af multicollinearitet også i dette tilfælde lede til fejlkonklusioner. Det kan vises, jvf. Toft-Nielsen (1974), at risikoen for forkerte fortegn er til stede, hvis f. eks. x_2's procentvise andel af den samlede forklaringsgrad er mindre end den indre forklaringsgrad. Ligeledes er det vist, at både b_2 og b_3 i denne situation kan blive fundet signifikante, og det til trods for at estimationen må betegnes som direkte vildledende. At problemet ikke blot er af teoretisk interesse vil et konkret eksempel senere i artiklen klart demonstrere.

Da således de traditionelle teststørrelser ikke med sikkerhed giver et klart signal i de situationer, hvor en partiel fortolkning bør undgås, bør man supplere den klassiske regressionsanalyse med en metode til at afsløre og afhjælpe forekomsten af multicollinearitet.

En sådan blev allerede i 1930'erne konstrueret af Ragnar Frisch, nemlig hans konfluensanalyse (Frisch 1934), der efter min mening må anses som værende yderst relevant i økonomiske relationsanalyser.

Af øvrige arbejder skal her blot nævnes Farrar og Glauber (1967) hvor forskellige tests for multicollinearitet gennemgås, Theil (1963) og Theil og Goldberger (1961) anvendende a priori information, Silvey (1969) anvendende supplerende observationssæt samt Neeleman (1973) anvendende Penrose-generaliserede inverse ved estimation af simultane modeller. Vedrørende de matematiske aspekter ved multicollinearitet henvises specielt til Silvey (1969), Klein og Nakamura (1962) samt Toft-Nielsen (1974).

Ragnar Frisch's konfluensanalyse

Betragtes istedet for eksemplet i (1) en generel model med k-1 forklarende variable, er ideen i konfluensanalysen, at hver variabel er sammensat af en systematisk komponent x_i' og en forstyrrende komponent x_i'', således at

$$x_i = x_i' + x_i'' \qquad \text{for } i = 1,2,\ldots,k \tag{3}$$

Frisch postulerer herefter, at der eksisterer en eksakt lineær relation mellem de systematiske komponenter, dvs.

$$\alpha_1 x_1' + \ldots + \alpha_k x_k' = 0 \tag{4}$$

$$x_1' = \sum_{i=2}^{k} \left(-\frac{\alpha_i}{\alpha_1} x_i' \right)$$

Ud fra formel (3) og (4), skulle man nu være i stand til at se forskellen mellem den klassiske metode og konfluensanalysen. At *alle* de økonomiske variable tillades at indeholde index-afrundings- eller direkte målefejl ifølge (3) må siges i mange praktiske situationer at være en realistisk udvidelse af den klassiske metodes forudsætning. Imidlertid lider konfluensanalysen også af visse mangler. Af (4) fremgår det, at Frisch går ud fra en *fuldt specificeret* relation hvor samtlige systematiske variable er med i sættet, hvorved antallet af observationer bliver uden betydning for vurderingen af usikkerheden på estimaterne.

Den restriktive forudsætning om en eksakt lineær relation mellem de systematiske komponenter medfører ligeledes, at det ikke er nødvendigt at specificere et kausalt sammenhæng. *Alle* indgående variable behandles ens. Ser vi endnu engang på fig. 1a medfører relation (4), at man uanset minimaliseringsretning skulle bestemme det »rigtige« plan på nær en vilkårlig proportionalitetsfaktor. Hvis derimod planets placering er ustabilt grundet eksistensen af multicollinearitet, måtte man forvente (se fig. 1b), at planerne ville rotere afhængigt af minimaliseringsretningen. Konfluensanalysen kan derfor opdeles i to faser. En estimationsfase, der ikke vil blive omtalt her, samt en metode til visuel afklaring af de estimerede regressionsplaners placering i rummet i forhold til hinanden. Denne fase kaldes *bunch-map* analysen, og må nok anses for at være Frisch's væsentligste bidrag.

Da de klassiske forudsætninger og Frisch's forudsætninger i højere grad er komplementære end alternative foreslås det, at den traditionelle regressionsanalyse suppleres med bunch-map analysen for at undgå nonsensestimationer.

Konstruktion af et bunch-map

I den klassiske flerdimensionale regressionsanalyse blev estimaterne af hældningskoefficienterne fundet ved at benytte mindste kvadraters metode. Mere eksakt minimaliserede man residualernes afstand til hyperplanet i x_1's retning med x_1 som endogen variabel. Generelt er det naturligvis muligt at minimalisere i enhver af de k x'ers retning. Hvis x_1 fastholdes som endogen variable, kan man få i alt k estimater for hver af de k-1 regressionskoefficienter ved at minimalisere i hver retning.

Uden tab af generalitet antages det i det følgende at alle x'erne er på normaliseret form, d.v.s. at gennemsnittet er 0 og spredningen SD 1. Den centrale matrise bliver derfor korrelationsmatrisen **R**. For $k = 3$ bliver **R**

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

og man får eksempelvis 3 skøn over regressionskoefficienten til den forklarende variabel x_2

$$\overset{1}{b}_{12(123)} = \frac{r_{23}r_{31} - r_{21}}{1 - r^2_{23}} = -\frac{|\mathbf{R}_{12}|}{|\mathbf{R}_{11}|}, \quad \overset{2}{b}_{12(123)} = \frac{1 - r^2_{13}}{r_{13}r_{32} - r_{12}} = -\frac{|\mathbf{R}_{22}|}{|\mathbf{R}_{21}|}$$

$$\overset{3}{b}_{12(123)} = \frac{r_{13}r_{21} - r_{23}}{1 - r^2_{12}} = -\frac{|\mathbf{R}_{32}|}{|\mathbf{R}_{31}|}$$

Som det ses, svarer det første skøn til det klassiske[1]. Det andet skøn svarer til koefficienten fundet i den forklarende variabels egen retning. Hvis vi betegner disse skøn som værende de to ledende (markeret ved cirkler), kan de 3 regressionskoefficienter fremstilles grafisk i et såkaldt bunch-map (se fig. 2).

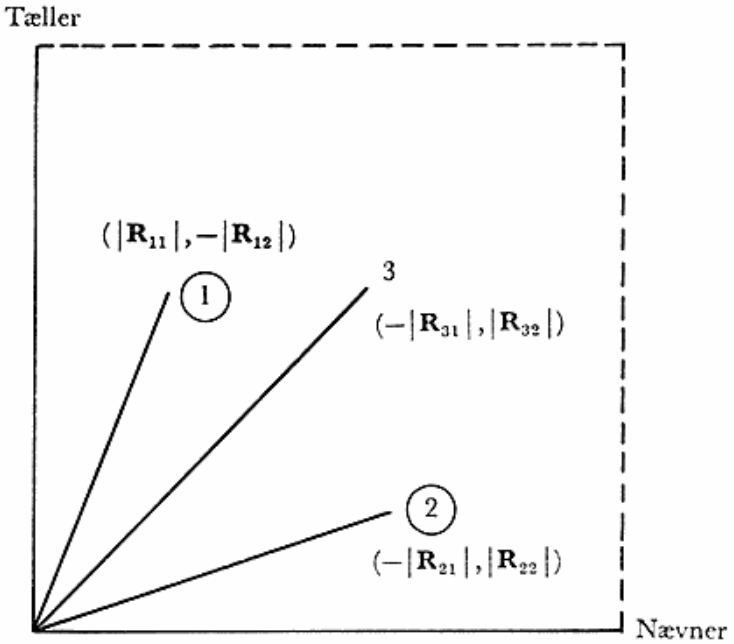

FIG. 2. *Et bunch-map for b* $\overset{1}{\underset{2}{\cdot}}\,3$ *i sættet* (123)

1 *og* 2 *er ledende stråler.*

1. $|\mathbf{R}_{ij}|$ er defineret som kofaktoren til r_{ij} i $\mathbf{R}$.

De tre rette linier i fig. 2 betegnes *stråler,* og alle strålerne under et kaldes et *strålebundt.* Hvis planet er stabilt må man forvente, at strålerne ligger tæt, og strålebundtet kaldes derfor lukket.

Bunch-map analysen, samt vejledende regler til klassifikation af nytilførte variable

Hovedtanken i bunch-map analysen er herefter, at man forsøger at konstruere et bunch-map for enhver tænkelig kombination af det givne antal variable. Ikke kun i det fuldstændige sæt med k variable, men også for alle undersæt. Lad m være antallet af variable i undersættet. Man lader da m gå fra 2 til k. For $k = 4$ er proceduren fremstillet i fig. 3.

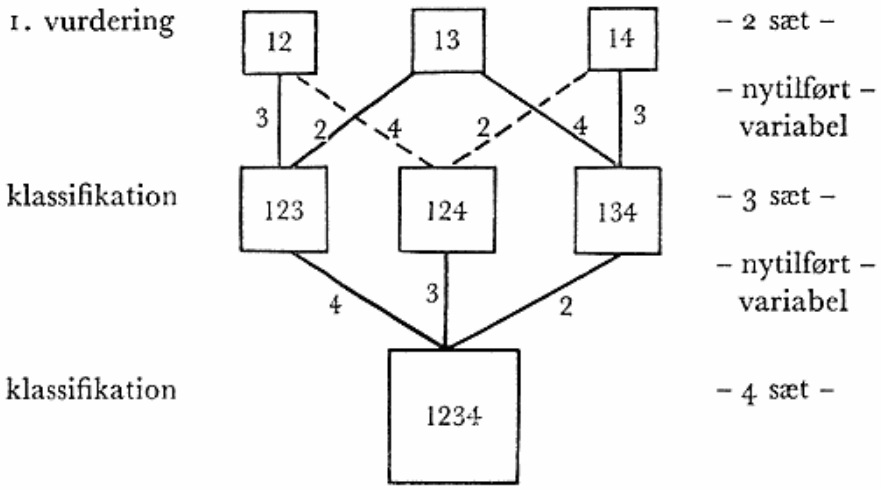

FIG. 3. *Grafisk fremstilling af bunch-map analyse proceduren for $k = 4$ variable.*

Ved at studere de mulige bunch-maps udvælger man det sæt, der giver den bedste beskrivelse, dvs. det tætteste strålebundt. For at få et indblik i hvilken betydning en nytilført variabel har for et givet undersæt, eksempelvis variabel nr. 3 til sættet 12, har jeg valgt kort at give en oversigt over forskellige vejledende regler, der kan benyttes ved klassifikationen af en nytilført variabel som værende A: nyttig, B: overflødig eller C: skadelig. Alle punkterne behøver ikke at være opfyldt samtidigt[2].

2. For en mere omfattende diskussion henvises til Toft-Nielsen (1971).

A: En nyttig variabel – ★

1. Strålebundtet indsnævres, dvs. at vinklen mellem de to yderste stråler formindskes ved at inkludere en ny variabel i sættet.
2. Strålen repræsenterende den nye variabel falder indenfor det *gamle* strålebundt.
3. Strålebundtets hældning ændres generelt.

B: En overflødig variabel – ○

1. Strålebundtet indsnævres ikke.
2. Hældningen ændres ikke markant.
3. Den *nye* stråle falder udenfor det gamle strålebundt.
4. Den nye stråle er væsentlig kortere end de øvrige.
5. De øvrige stråler forkortes ikke nævneværdigt.

Til punkt B4 og B5 bør det bemærkes, at størrelsen af regressionskoefficienten fundet i den nye variabels retning, vil være bestemmende for det generelle niveau af de øvrige koefficienter i ligningen. En given stråle i bundtet vil derfor generelt blive kortere jo mere fuldstændig de øvrige variable i sættet er lineært afhængige. Da alle de indgående variable er på normeret form og derfor sammenlignelige, kan man prioritere deres indflydelse efter strålens længde. En meget kort stråle betyder derfor, at den pågældende variabel er uden betydning – en overflødig variabel.

C: En skadelig variabel – △

1. Strålebundtet eksploderer. Denne adfærd forventes, hvis der opstår multicollinearitet i sættet, da planets hældning bestemmes af den tilfældige variation.
2. Strålebundtet udvides.
3. Den ledende stråle for den endogene variabel (her nr. 1) reduceres kraftigt. Den kausale sammenhæng bliver domineret af et stærkere lineært bånd mellem de exogene variable, hvorved de klassiske regressionskoefficienter bliver indeterminerede.

Ved hjælp af de opstillede regler skulle man nu være i stand til at gennemføre bunch-map analyse og således sikre, at de estimerede regressionsligninger får et meningsfyldt indhold.

Da den skitserede metode imidlertid er ret tidkrævende, selv for en trænet

bunch-map analytiker, kan det ikke anbefales at benytte den ved hver eneste regressionsanalyse.

I de tilfælde, hvor de forklarende variable er parvis uafhængige dvs. ukorrelerede, vil bunch-map analysen være overflødig. Hvis derimod en eller flere af de forklarende variable er eksakt lineært afhængige er bunch-map analysen en nødvendighed. I praksis vil den observerede determinant til korrelationsmatrisen ligge mellem o og 1. Den naturligste måde at løse problemet på, er da ved at teste en hypotese om, at de forklarende variable er parvis uafhængige. Hvis hypotesen forkastes gennemføres bunch-map analysen, da eksistensen af *M*ulti *C*ollinearitet kan give nonsensestimationer. Jeg har valgt at kalde testet MC1-test.

MC1-test for multicollinearitet

Lad testhypotesen være at de forklarende variable er uafhængige, svarende til at den sande korrelationsmatrise er lig en enhedsmatrise af orden k-1.

Hvis de forklarende variable $(x_2, \ldots, x_k)$ følger en flerdimensional normalfordeling, kan det vises[3], at

$$MC1 = -T\left(1 - \frac{2k+9}{6T}\right) ln\,|\mathbf{R}_{11}| \qquad (5)$$

under uafhængighedsforudsætningen er tilnærmet χ^2-fordelt med $\frac{1}{2}$ (k-1)(k-2) frihedsgrader. Hvis hypotesen forkastes, forekommer der multicollinearitet og bunch-map analysen bør gennemføres.

Et konkret eksempel

Modellen og observationsmaterialet til dette eksempel er hentet fra Erling Olsens disputats fra 1971 *International Trade Theory and Regional Income Differences,* kapitel 6. Allerede ved forsvaret af disputatsen påpegede P. Nørregaard Rasmussen[4] i forbindelse med visse reestimationer: »Det kunne jo være, at begge estimeringer led af en fælles svaghed, som måske er særlig relevant i dette tilfælde. Problemet er multicollinearitet – et begreb, som Frisch indførte for næsten 40 år siden, men mærkelig nok overhovedet ikke nævnes af forfatteren«.

For at undersøge denne påstand har jeg valgt rent statistisk at vurdere én af Erling Olsens regressionsligninger, nemlig relation 6.14 p. 139, der angiver væksten i den i'te regions arbejdsstyrke i forhold til hele landet forklaret ved ind-

3. T. W. Anderson (1958) hvor øvrige referencer er omtalt.

komsten pr. capita, den relative løn og det relative økonomiske befolkningspotentiel[5].

$$\frac{L_i^{t+20}}{L_i^{t}\,e^{20\Theta}} = \left(\frac{y^t}{y_i^t}\right)^{20H} \left(\frac{w_i^t}{w^t}\right)^{20\lambda} \left(\frac{iV^t}{V^t}\right)^{20\mu} \tag{6}$$

Ved at tage logaritmen til (6) bringes relationen på lineær form

$$x_1 = 20H \cdot x_2 + 20\lambda\, x_3 + 20\mu\, x_4 \tag{7}$$

I det oprindelige observationsmateriale er der 18 datasæt, 9 regionssæt fra 1880 og 9 fra 1900. Dette antal reduceres til 13, hvilket øger den multiple korrelationskoefficient.

TABEL 1. *Resultatet af den flerdimensionale regressionsanalyse. For t=1880 og 1900 på relation (7).*

t = 1880 og 1900	20H	20λ	20μ
uafhængig variabel	x_2	x_3	x_4
regressionskoefficienten	−7,129	−7,073	−0,569
standardafvigelse	2,18	2,17	0,09
beregnet t-værdi	−3,28	−3,26	−5,96
den multiple korrelationskoefficient R	0,935		

ANM.: Tabellen svarer i princippet til Erling Olsen (1971, Tabel 6.11 p. 143), hvor bl. a. $\mathbf{b}$ = (−0,025, 0,286, −0,272), $\mathbf{t}$ = (−0,13, 1,03, −3,85) og R = 0,91.

Resultatet ses i Tabel 1[6]. Det ses, at ud fra et statistisk synspunkt, må estimationen siges at være tilfredsstillende. Korrelationskoefficienten er 0,935 og alle regressionskoefficienterne er fundet stærkt signifikante. En partiel fortolkning skulle derfor være mulig. Ud fra et økonomisk synspunkt stiller sagen sig anderledes. μ er stadig negativ og med Erling Olsen (1971, p. 143) kunne man sige »Again, our impression is that the explanation of the negative μ-value should be found outside conventional economic theory. But it is also our impression that it

4. *Nationaløkonomisk Tidsskrift* (1971 p. 61–72 (især p. 69–70)).
5. Jævnfør Erling Olsen (1971, p. 90–91).
6. Resultatet er ikke sammenfaldende med Olsen (1971, p. 143) da logaritmetransformationen der er udeladt.

is more than difficult to find a reasonable one[7]«. At H og λ er negative og signifikante måtte også have undret, især da Erling Olsen (p. 91) forventer, at både H og λ er positive.

Vurdering af relationen

Af Erling Olsens kommentarer til de estimerede regressionskoefficienter må man slutte, at målsætningen med den estimerede relation er at kunne tolke **b** partielt. For at undersøge, om en sådan er forsvarlig udføres først et MC1-test. Da $k = 4$, $T = 13$ og $|\mathbf{R}_{11}|$ (iflg. appendix) er 0,006982 haves følgende teststørrelse iflg. formel (5)

$$MC1 = 50{,}5$$

Da 99 % fraktilen i χ^2-fordelingen med 3 frihedsgrader er 11,3, må hypotesen om, at de forklarende variable er parvis ukorrelerede forkastes. Vi kan derfor konkludere, at der er afhængighed, og man må specielt være på vagt over for multicollineære sæt. Midlet hertil var bunch-map analysen.

Kommentarer til bunch-map analysen

Lad os først betragte undersættene bestående af den endogene variabel nr. 1 og én af de exogene. Lad os betegne dem 2-sættene. Af fig. 4 ses det, at sættet (14) giver det tætteste strålebundt svarende til, at variabel nr. 4 giver den største forklaring. Ved at tilføre nr. 2 til sættet ses det, at denne må betegnes som en overflødig variabel. Den nye stråle (nr. 2) er kortere end de øvrige, ligesom strålebundtet (14) og (12) ikke indsnævres i (124).

Tilsvarende gælder for variabel nr. 3. Strålebundterne i (13) og (14) indsnævres ikke i (134) ligesom stråle 3 er væsentlig kortere end nr. 1 og 4.

Går vi fra sættet (134) resp. (124) til (1234) ses det, at alle strålebundterne indsnævres betydeligt. Dette er generelt set et godt tegn. Imidlertid reduceres stråle nr. 1 og 4 kraftigt. Nr. 1 reduceres næsten til o-vektoren. Dette er et af tegnene på »multicollinearitet«. De klassiske regressionskoefficienter bliver næsten fuldstændig indeterminerede. Nr. 2 resp. nr. 3 må derfor betegnes som skadelige variable.

Lad os herefter starte med (12) resp. (13). Ved at tilføre 3 resp. 2 ses det, at strålebundtet i (123) indsnævres betydeligt. Medens regressionskoefficienterne havde modsat fortegn i (12) og (13), *som man iflg. Erling Olsens model skulle*

7. Da μ i hele eksemplet forbliver negativ, kunne forklaringen være, at befolkningsbevægelsen i U.S.A. fra 1880–1900 netop var rettet mod Vest. Dette ville give et $\mu < 0$. Se f. eks. Faulkner: *American Economic History*, London 1964, kap. 18.

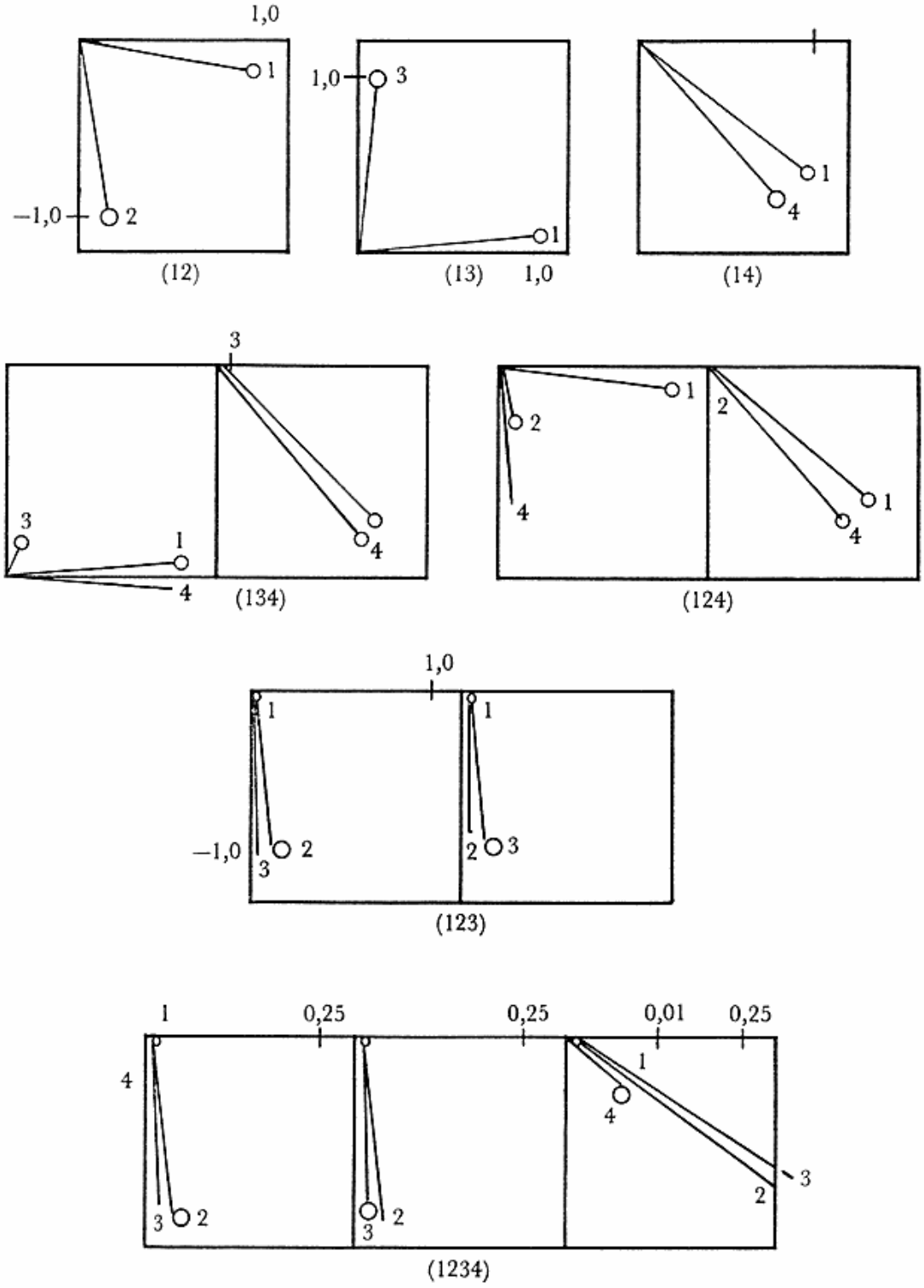

FIG. 4. *Bunch-map analyse på Erling Olsens data. Konstrueret ud fra appendix.*

forvente[8], bliver de begge negative i (123). Ved at betragte (23) ses det tydeligt, at interkorrelationen mellem (23) slår igennem sættet. Da 2 resp. 3 reducerer den ledende stråle nr. 1 kraftigt, må de begge karakteriseres som skadelige.

TABEL 2. *Oversigtsskema til bunch-map analysen i fig. 4.*

Endogen: nr. 1	1 2 3	1 2 4	1 3 4	1 2 3 4
Begyndelsessæt 12	△	○		△ ○
Begyndelsessæt 13	△		○	△ ○
Begyndelsessæt 14	○		○	△ △

ANM.: ★ angiver en nyttig variabel
○ angiver en overflødig variabel
△ angiver en skadelig variabel

Vi kan herefter konkludere, at selvom den klassiske flerdimensionale regressionsanalyse i det reestimerede Erling Olsen eksempel klart indicerede signifikante regressionskoefficienter, samt en teknisk determinationsgrad på 0,935, må den fundne relation alligevel siges at have karakter af nonsens. Skulle man rent statistisk anbefale et bedre sæt, må det blive (14), dvs. at væksten i den i'te regions arbejdsstyrke skulle forklares udelukkende ved det relative, økonomiske befolkningspotentiel. At regressionskoefficienten a priori må forventes at blive negativ i perioden 1880 til 1900 skyldes sikkert »frontierbevægelsen«, der gjorde sig gældende i U.S.A. i dette tidsrum.

APPENDIX

Den simple empiriske korrelationskoefficientmatrise i Erling Olsen eksemplet

$$\mathbf{R} = ((r_{ij})) = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \left(\begin{array}{cccc} 1{,}00 & -0{,}15208 & 0{,}09939 & -0{,}85101 \\ -0{,}15208 & 1{,}00 & -0{,}99617 & 0{,}12596 \\ 0{,}09939 & -0{,}99617 & 1{,}00 & -0{,}10220 \\ -0{,}85101 & 0{,}12596 & -0{,}10220 & 1{,}00 \end{array} \right) \end{array}$$

Determinanten $|R_{11}| = 0{,}006982$

8. Nr. 2 angiver indkomst pr. capital som i modellen virker modsat og nr. 3 angiver den relative løn.

Litteratur

ANDERSON, T. W. 1958. *An Introduction to Multivariate Statistical Analysis.* London.

FARRAR, D. E. og R. R. GLAUBER. 1967. Multicollinearity in Regression Analysis: The problem revisitet. *Review of Economics and Statistics* 49:92–107.

FRISCH, RAGNAR. 1934. *Statistical Confluence Analysis by Means of Complete Regression System.* Universitetets Økonomiske Institut, Oslo.

KLEIN & NAKAMURA. 1962. Singularity in the Equation Systems of Econometrics: Some Aspects of the problem of Multicollinearity. *International Economic Review* 3.

NEELEMAN, D. 1973. *Multicollinearity in linear economic models.* Holland.

NØRREGAARD RASMUSSEN, P. 1971. En disputats om interregionale udligningsmekanismer. *Nationalokonomisk Tidsskrift* 109:61–72

OLSEN, ERLING. 1971. *International Trade Theory and Regional Income Differences.* Amsterdam.

SILVEY, S. D. 1969. Multicollinearity and Imprecise Estimation. *Journal of Royal Stat. Soc., Series B.*

THEIL, H. og GOLDBERGER. 1961. On Pure and Mixed Statistical Estimation in Economics. *International Economic Review* 2.

THEIL, H. 1963. On the use of incomplete prior information in regression analysis. *Journal of Am. Stat. Ass.* 58.

TOFT-NIELSEN, P. 1971. Flerdimensional regressionsanalyse med specielt henblik på undersøgelse af multicollinearitet. Stor opgave ved politstudiet.

TOFT-NIELSEN, P. 1974. Nogle OLS estimatorers længde og hældning under varierende MC-grader. Universitetets Statistiske Institut, nr. 23, grå serie.

forvente[8], bliver de begge negative i (123). Ved at betragte (23) ses det tydeligt, at interkorrelationen mellem (23) slår igennem sættet. Da 2 resp. 3 reducerer den ledende stråle nr. 1 kraftigt, må de begge karakteriseres som skadelige.

TABEL 2. *Oversigtsskema til bunch-map analysen i fig. 4.*

Endogen: nr. 1	1 2 3	1 2 4	1 3 4	1 2 3 4
Begyndelsessæt 12	△	○		△ ○
Begyndelsessæt 13	△		○	△ ○
Begyndelsessæt 14	○		○	△ △

ANM.: ★ angiver en nyttig variabel
○ angiver en overflødig variabel
△ angiver en skadelig variabel

Vi kan herefter konkludere, at selvom den klassiske flerdimensionale regressionsanalyse i det reestimerede Erling Olsen eksempel klart indicerede signifikante regressionskoefficienter, samt en teknisk determinationsgrad på 0,935, må den fundne relation alligevel siges at have karakter af nonsens. Skulle man rent statistisk anbefale et bedre sæt, må det blive (14), dvs. at væksten i den i'te regions arbejdsstyrke skulle forklares udelukkende ved det relative, økonomiske befolkningspotentiel. At regressionskoefficienten a priori må forventes at blive negativ i perioden 1880 til 1900 skyldes sikkert »frontierbevægelsen«, der gjorde sig gældende i U.S.A. i dette tidsrum.

APPENDIX

Den simple empiriske korrelationskoefficientmatrise i Erling Olsen eksemplet

$$\mathbf{R} = ((r_{ij})) = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \left(\begin{array}{cccc} 1{,}00 & -0{,}15208 & 0{,}09939 & -0{,}85101 \\ -0{,}15208 & 1{,}00 & -0{,}99617 & 0{,}12596 \\ 0{,}09939 & -0{,}99617 & 1{,}00 & -0{,}10220 \\ -0{,}85101 & 0{,}12596 & -0{,}10220 & 1{,}00 \end{array} \right) \end{array}$$

Determinanten $|R_{11}| = 0{,}006982$

8. Nr. 2 angiver indkomst pr. capital som i modellen virker modsat og nr. 3 angiver den relative løn.