

Regression og geometrisk data analyse (2. del)

Ulf Brinkkjær

Introduktion

Artiklen søger at vise, hvordan regressionsanalyse og geometrisk data analyse kan integreres. Det er interessant, fordi disse metoder ofte opstilles som modsætninger f.eks. som en modsætning mellem beskrivende og forkla- rende metoder. Artiklens første del bragtes i *Praktiske Grunde* 3-4 / 2007.

Resume af problematikken

		Tabel 1: Vending			Tabel 2: Forsvinder			Tabel 3: Stabilitet			Tabel 4: Opstår		
		Bestå	Dumpe		Bestå	Dumpe		Bestå	Dumpe		Bestå	Dumpe	
Romeo	Dreng	12	28	40	8	32	40	8	32	40	16	24	40
Romeo	Pige	1	9	10	2	8	10	8	2	10	1	9	10
Julie	Dreng	9	1	10	8	2	10	2	8	10	9	1	10
Julie	Pige	28	12	40	32	8	40	32	8	40	24	16	40
		50	50	100	50	50	100	50	50	100	50	50	100

Hver af tabellerne læses på den måde, at eksempelvis den fremhævede celle i tabel 2 fortæller, at der var 8 piger på gymnasiet Romeo, som dumpede til studentereksamen.

Som det fremgår af tabel 1 er der 50 drenge (40 på Romeo i linje 1 og 10 på Julie i linje 3) og 50 piger (10 på Romeo i linje 2 og 40 på Julie i linje 4). Kigger vi på tallene i tabel 1, så er andelen, der har bestået, som følger:

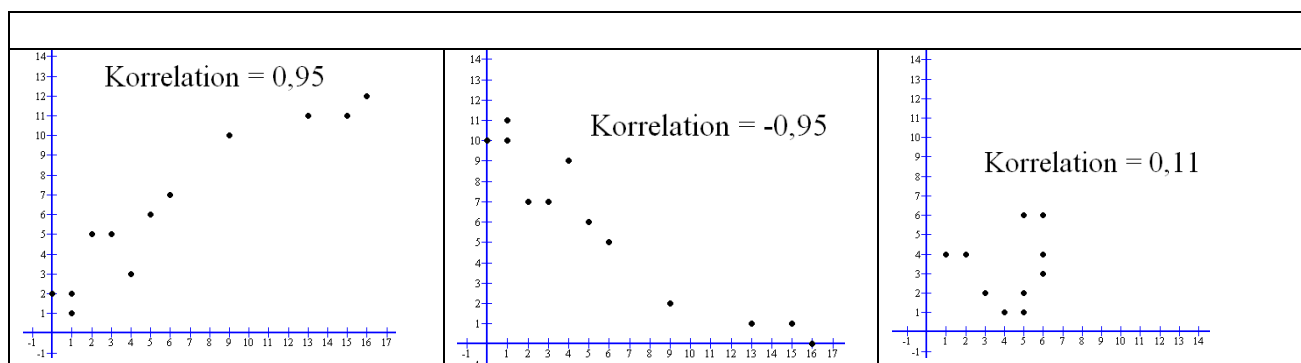
	A	B	C	D
	Tabel 5	Drenge	Piger	Forskel
1	Bestået: Begge gymnasier i alt	$\frac{12 + 9}{40 + 10} = 0,42$	$\frac{1 + 28}{10 + 40} = 0,58$	$0,42 - 0,58 = -0,16$
2	Bestået: Romeo-gymnasiet:	$\frac{12}{40} = 0,3$	$\frac{1}{10} = 0,1$	$0,3 - 0,1 = 0,2$
3	Bestået: Julie-gymnasiet:	$\frac{9}{10} = 0,9$	$\frac{28}{40} = 0,7$	$0,9 - 0,7 = 0,2$

	Tabel 6	Romeo	Julie	Forskel
	A	B	C	D
1	Bestået: Begge køn i alt	$\frac{12 + 1}{40 + 10} = 0,26$	$\frac{9 + 28}{10 + 40} = 0,74$	$0,26 - 0,74 = -0,48$
2	Bestået: Drengene	$\frac{12}{40} = 0,3$	$\frac{9}{10} = 0,9$	$0,3 - 0,9 = -0,6$
3	Bestået: Piger	$\frac{1}{10} = 0,1$	$\frac{28}{40} = 0,7$	$0,1 - 0,7 = -0,6$

Den strukturelle effekt af faktoren køn er dermed -0,16 (Tabel 2 række 1). Dvs. pigerne klarer sig bedre end drengene. Kigger vi i tabel 2 række 2 og 3, ser vi derimod, at drengene på begge gymnasier klarer sig bedre end pigerne. Retningen i den globale effekt af køn, der er -0,16, vendes dermed (til +0,2) i begge de lokale tabeller. Fokus for det følgende er en grafisk repræsentation af forskellige relationer mellem strukturelle effekter i global og marginal sammenhæng.

Den grafiske repræsentation

Korrelationen mellem to variable er et mål for i hvor høj grad, de samvarierer linært. Det har en minimumsværdi på minus en og en maksimumsværdi på plus en.



De tre plot i figuren demonstrerer sammenhængen mellem plottede punkter og korrelationskoefficienter. Ligger punkterne tæt på en ret linje, og udtrykker punkterne, at en lille værdi af den ene variabel hører sammen med en lille også på den anden, så er korrelationen stor, det vil sige tæt på +1. Plottet i midten viser, at når punkterne ligger tæt på en ret linje, men hvor store værdier på en variabel hører sammen med små værdier på den anden, så er korrelationen meget lille, det vil sige tæt på -1. Endelig viser plottet til højre, at når punkterne ikke har nogen lineær sammenhæng, så nærmer korrelationskoefficienten sig 0.

Den grafiske fremstilling starter med at tegne en såkaldt korrelationscirkel. Det vil sige en cirkel med centrum i aksernes skæringspunkt (0,0) og en radius på 1 (kaldet en enhedscirkel).

I Figur 1 afsættes inden i cirklen to diametre A og B (Figur 2), der mellem sig har en vinkel som svarer til, at:

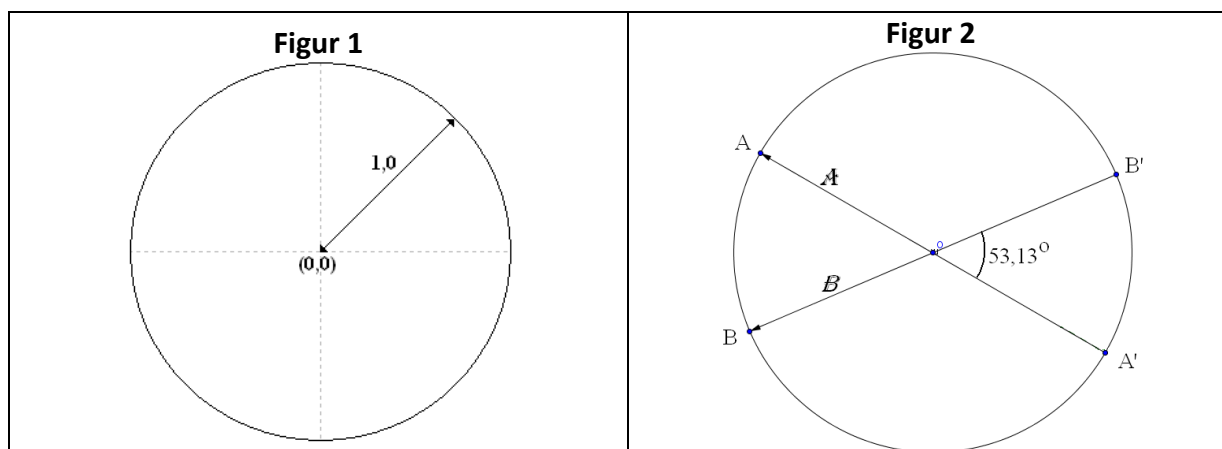
$$\cos^2(\text{vinkel}) = \text{korrelationen mellem variablene gymnasium og køn}$$

Data summeret i tabel 1 kan udregnes at have en korrelation mellem gymnasium og køn på 0,6 hvilket svarer til en vinkel mellem de to indtegnede linjer på ca. 53,13 grader.¹

¹ $\cos^2(\text{vinkel}) = 0,36 = 0,6^2$. Vinklen bliver dermed $\cos^{-1}(0,6) \approx 53,13^\circ$

² $\overrightarrow{OM'_b}$ og $\overrightarrow{OM_b}$ er afstanden fra centrum (punktet O) til punktet M'_b hhv. M_b , regnet med fortegn således at den er negativ i retning væk fra den

Den ene diameter løber mellem punkterne A' (Julie) og A (Romeo), mens den anden løber mellem punkterne B (Dreng) og B' (Piger).



I data fra tabel 1 (Vending) så vi i artiklens del 1, at regressionsanalyserne resulterede i følgende ligninger:

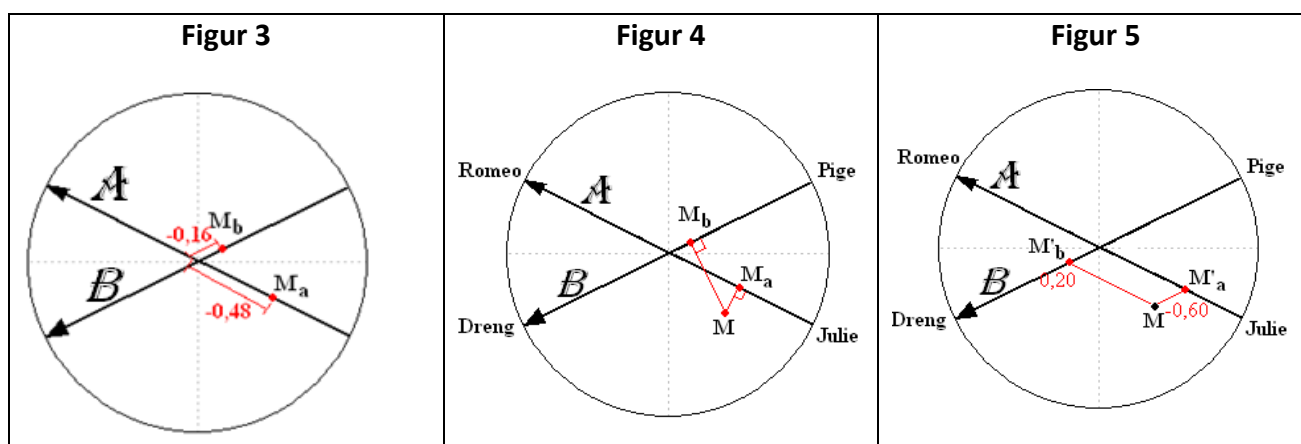
- a) I Simpel regression af x_a på y : $y^a = -0,48x_a + 0,74$
- b) I Simpel regression af x_b på y : $y^b = -0,16x_b + 0,58$
- c) I Multiple regression af x_a og x_b på y : $y^{a+b} = -0,6x_a + 0,2x_b + 0,7$

Som det fremgår er koefficienterne i regressionsligningerne sammenfaldende med:

- a) Værdierne i tabel 5 celle D1 og celle C1
- b) Værdierne i tabel 6 celle D1 og celle C1
- c) Værdierne i tabel 5 celle D3 tabel 3 celle D3 og tabel 2 celle C3.

Dette kan udtrykkes grafisk ved på Akse $\mathcal{A}$ at afsætte punktet M_a med koordinaten -0,48 (Globale effekt af faktor $\mathcal{A}$). På Akse $\mathcal{B}$ afsættes punktet M_b med koordinaten -0,16 (Globale effekt af faktor $\mathcal{A}$). Da begge størrelser er negative bevæger vi os fra centrum mod højre dvs. væk fra de pile, som angiver retningen på $\mathcal{A}$ og $\mathcal{B}$ (Figur 3). En linje, der er vinkelret på $\mathcal{A}$, trækkes nu fra M_a og en anden, der er vinkelret på $\mathcal{B}$, trækkes fra M_b . Punktet M (variablen bestået) findes i skæringen mellem disse to (Figur 4).

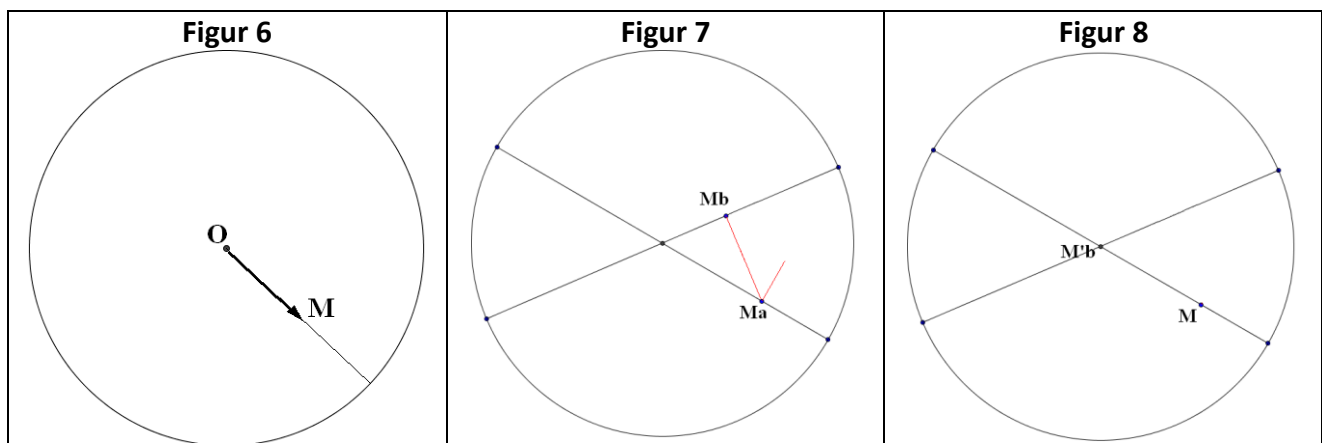
Herefter projeceres punktet M ind på akse $\mathcal{A}$ med en projektion, der løber parallelt med akse $\mathcal{B}$. Herved fremkommer punktet M'a. På tilsvarende vis findes punktet M'b (Figur 5).



Betragter vi situationen efter placering af punktet M, gælder det således, at den globale effekt af faktor B (køn) fremkommer ved en *vinkelret* projektion af punktet M på akse **A** ($M_a = -0,48$) mens den globale effekt af faktor A fremkommer ved en *vinkelret* projektion af punktet M på akse **B** ($M_b = -0,16$). Tilsvarende kan de lokale effekter af faktor A og B findes ved *parallelle* projektioner ($M'_a = -0,6$ og $M'_b = 0,2$) (Figur 5).

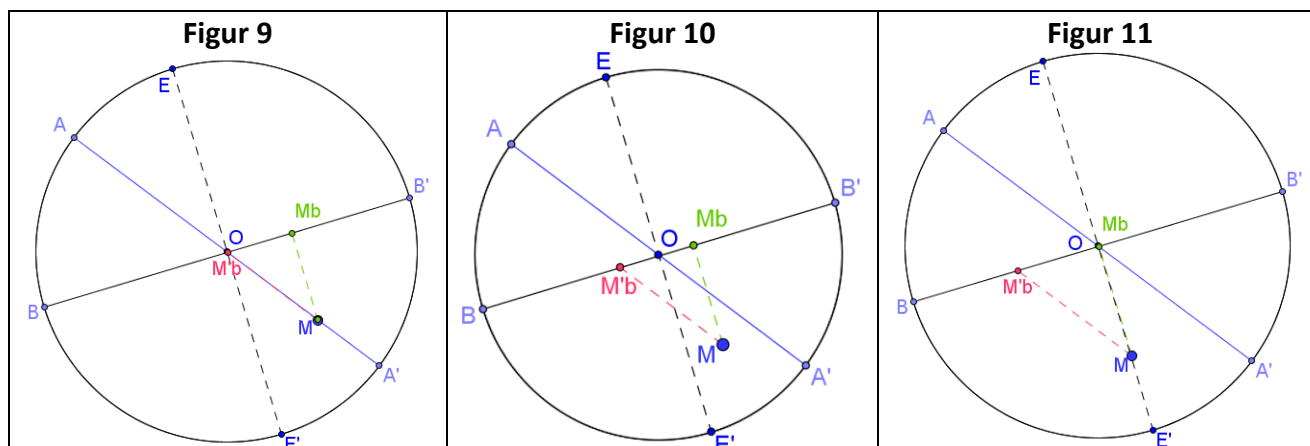
En sammenligning af den lokale effekt og den globale effekt svarer til at sammenligne projektionerne af den lokale effekt $(\overrightarrow{OM'_b})^2$ og den globale effekt $(\overrightarrow{OM_b})$. I det konkrete tilfælde altså $\frac{\overrightarrow{OM'_b}}{\overrightarrow{OM_b}} = \frac{0,20}{-0,16} = -1,25$. Dette forhold mellem $\overrightarrow{OM'_b}$ og $\overrightarrow{OM_b}$ kommer grafisk til udtryk som *en bestemt retning*. På en bestemt radius, f.eks. den som er angivet i Figur 6, vil forholdet $\frac{\overrightarrow{OM'_b}}{\overrightarrow{OM_b}}$ dermed være den samme, uanset hvor langt fra centrum punktet M ligger. Afstanden fra centrum er derimod et grafisk udtryk for, hvor godt den lineære regression beskriver sammenhængen mellem data, normalt symboliseret ved symbolet R^2 . I praksis er R^2 et mål for, hvor stor en del af variationen i responsvariablen (Y), som beskrives (eller forklares) af den lineære sammenhæng, som er output af den lineære regressionsanalyse. opstillede linære sammenhæng. Denne værdi svarer til produktet af de skrå og de lige koordinater dvs: $R^2 = (-0,48 * -0,60) + (-0,16 * 0,2) = 0,256$.

Kvadratroden af dette tal svarer til længden af vektoren $\overrightarrow{OM}$ ($R = \sqrt{0,256} = 0,506$) og dermed afstanden fra centrum af cirklen til punktet M. Jo længere punktet M ligger fra cirkelns centrum, jo bedre beskriver den linære funktion fra regressionsanalysen dermed variationen i responsvariablen (Y).



² $\overrightarrow{OM'_b}$ og $\overrightarrow{OM_b}$ er afstanden fra centrum (punktet O) til punktet $M'b$ hhv. Mb , regnet med fortegn således at den er negativ i retning væk fra den retning, som er angivet på de to diametre.

Retningens betydning



Med fokus på effekten af B (køn) kan korrelationscirklen deles op i 6 rum eller lagkagestykker med delinger langs akserne A og B , samt en (stiplet) linje, der står vinkelret på akse B , som i det følgende benævnes $B_{\perp}$ (Figur 9). Det fremgår af (Figur 9, Figur 10 og Figur 11), at når punktet M bevæger sig rundt som viseren på et ur, vil projektionerne Mb og $M'b$ og forholdet $\frac{OM'b}{OMb}$ naturligvis også ændre sig. Bevæger man derimod punktet M

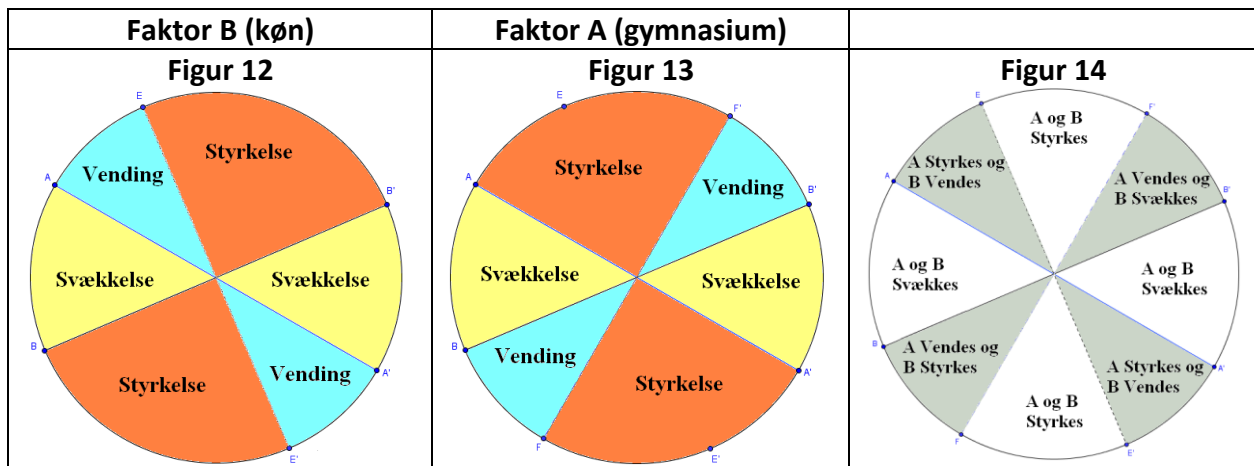
længere fra eller tættere på, vil projektionerne Mb og $M'b$ ændres sig, men forholdet $\frac{OM'b}{OMb}$ vil som nævnt ovenfor være uændret. Forestiller man sig, at vi kører punktet M hele vejen rundt i cirklen, men hele tiden med en afstand fra centrum på 0,7, vil projektionerne og deres forhold undervejs antage følgende værdier.

	A'	A' - E'	E'	E' - B	B	B - A	A	A - E	E	E - B	B'	B' - A'
$OM'b$	0	→	0,53	→	0,7	→	0	→	-0,53	→	-0,7	→
OMb	-0,42	→	0	→	0,7	→	0,42	→	0	→	-0,7	→
$\frac{OM'b}{OMb}$	0	→ -		+ →	1	→	0	→		→	1	→

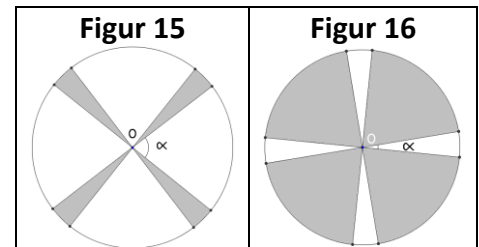
De to tomme felter indikere, at man ikke kan dividere med 0.

Man kan selv udforske forholdet mellem M og dets lige og parallelle projektion i filen [ventilator](#)

Udtrykt grafisk kan vi dermed dele cirklen op i 6 felter med forskelligt forhold mellem den globale og den lokale effekt (Figur 12). Et tilsvarende ræsonnement kan gennemføres for faktoren A (gymnasium), og igen vil cirklen kunne deles op i 6 rum (Figur 13).



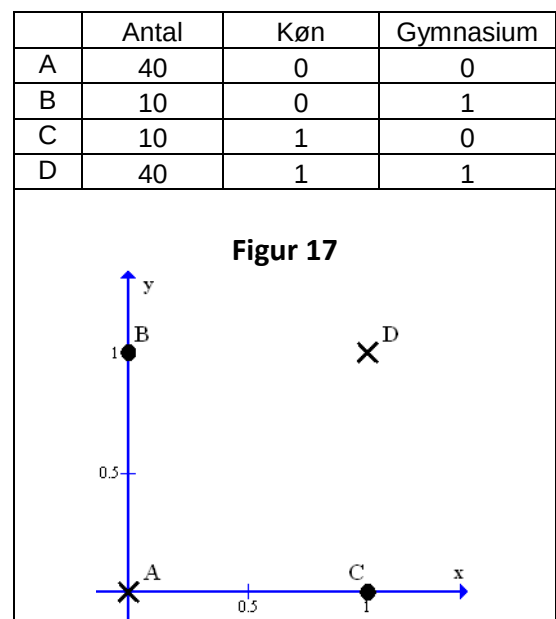
Kobles de to cirkler sammen (Figur 14) fremgår det, hvordan cirklen kan deles op i rum, hvor de to faktorer bevæger sig parallelt, fordi begge styrkes eller svækkes, og i andre rum hvor den ene styrkes eller svækkes, og den anden vendes. Størrelsen af de forskellige rum vil variere med hvor svag eller stærk sammenhæng, der er mellem de to faktorer A og B, der er udgangspunktet. I det gennemgåede eksempel var korrelationen $= 0,36 = 0,6^2$. Jo stærkere sammenhæng mellem de to faktorer, jo større korrelation og jo mindre vinkel mellem A og B ³. Sammenhængen og korrelationen er dermed stor i Figur 16, derfor bliver vinklen mellem A og B lille, og de grå områder, hvor sammenhængen vendes, store jf. Figur 14. Omvendt ser vi i Figur 15 en svag sammenhæng mellem de to faktorer og derfor også en stor vinkel mellem A og B , hvormed de grå områder er små. Det vil sige, stor korrelation mellem faktor A og B gør de grå områder, hvor sammenhængen vendes for en af faktorerne, relativt store.



Regression og Principal Komponentanalyse

Vælger vi som her at gennemføre en Principal komponentanalyse, vil analysens første punkt være at omkonvertere inputvariablene, således at materialets forskelle er bedre repræsenteret på de to akser. Ser man på forandringen fra Figur 17 til Figur 19, bliver det tydeligt, hvordan det i praksis fører til, at den første (vandrette akse) kommer til at modstille Romeo+dreng og Julie+piger. De to punkter i Figur 19, som ligger på den vandrette akse, er dermed dem, som i Figur 17 er markeret med kryds (punkterne A og D), men hvorfor nu det?

Kigger man i tabellen over Figur 17, fremgår det, at i punkt A og punkt D, der er de to punkter markeret med kryds i Figur 17, er antallet af personer 40, mens det i de to øvrige punkter kun er 10. Det betyder, at teoretisk set ligger der 40 kryds ovenpå hinanden i både punkt A og D, mens der kun ligger 10 prikker oven i hinanden i punkterne B og C.

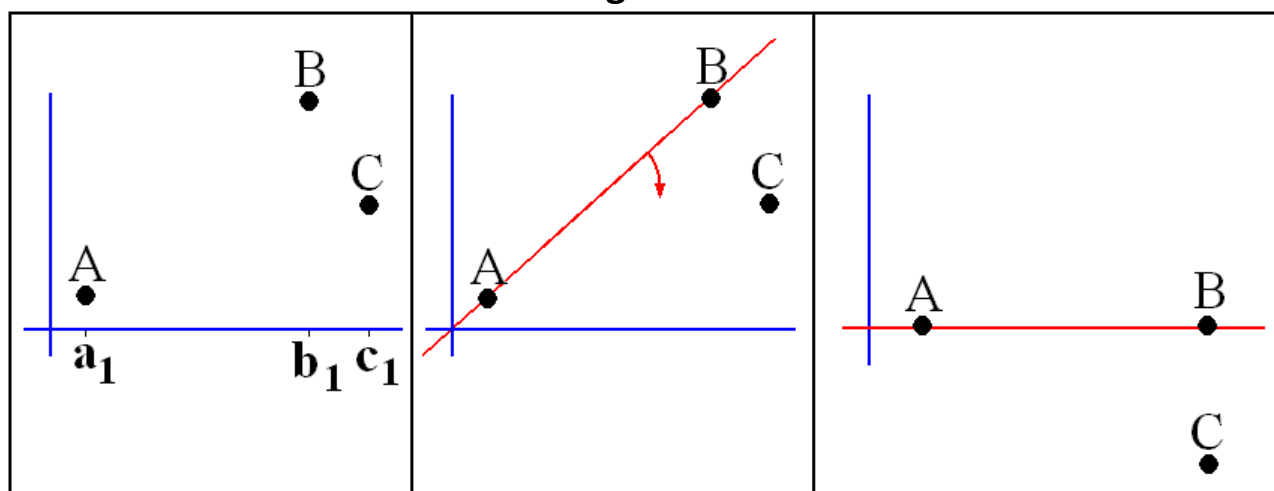


³ $\cos^{-1}(\theta) = 1 \Leftrightarrow \angle\theta = 0^\circ$, mens $\sin^{-1}(\theta) = 0 \Leftrightarrow \angle\theta = 90^\circ$

Når det kommer til at spille en rolle, hænger det sammen med ønsket om en afbildning, hvori førsteaksen opfanger en maksimal del af afstandene mellem punkterne, det vil sige af variansen. Man kan tænke på det på den måde, at der er flere punkter i position A og D, og derfor optræder afstanden mellem $A \rightarrow B$ mange gange, de afstande har dermed stor vægt. Måden at tage højde for det på er at lægge begge punkterne på den nye førsteakse, hvormed disse afstande bliver fuldstændigt afbildet på førsteaksen.

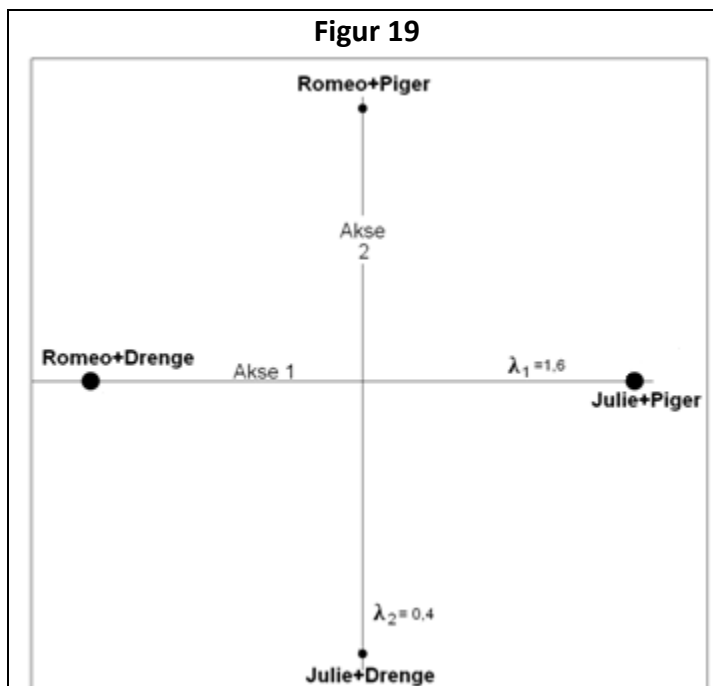
Pointen i den betragtning kan ses i Figur 18, der viser to afstande af samme længde ($A \rightarrow B = A \rightarrow C$). Som det fremgår på grafen til venstre, opfanger den vandrette akse stykket $a_1 \rightarrow b_1$ af afstanden $A \rightarrow B$, men stykket $a_1 \rightarrow c_1$ af afstanden $A \rightarrow C$. Når stykket $a_1 \rightarrow c_1$ er længere end stykket $a_1 \rightarrow b_1$, handler det ikke om forskel på afstandene mellem punkterne, afstanden $A \rightarrow B$ var jo netop den samme som afstanden $A \rightarrow C$. Valgte vi nu at trække en linje gennem både A og B (midterste graf) og gøre det til den nye førsteakse (grafene til højre), så ville afstanden fra A til B blive opfanget fuldstændigt på førsteaksen. Det ville imidlertid kun være delvist tilfældet for afstanden fra A til C, det vil sige, nu ville det gælde, at $A \rightarrow B = a_1 \rightarrow b_1$, og at $a_1 \rightarrow b_1$ er længere end $a_1 \rightarrow c_1$.

Figur 18



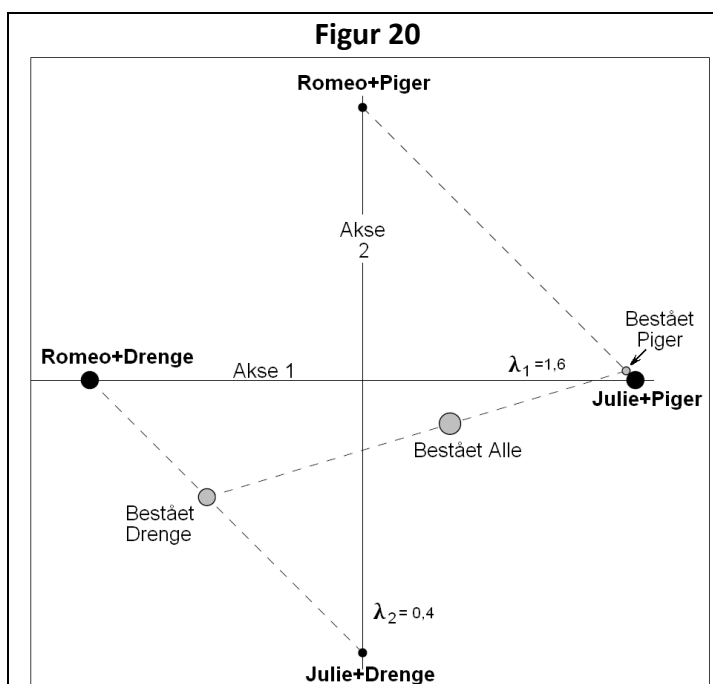
På Figur 19 er det illustreret ved, at de to punkter på førsteaksen er større end de to øvrige.

Med den gennemførte variablekonvertering opnås en førsteakse, der repræsenterer 80% af materialets interti, der er et mål for de samlede afstande mellem punkterne, ($\lambda_1 = 1,6$ af totalt 2,0 dvs. $\frac{1,6}{2,0} = 80\%$), og en anden akse, der repræsenterer de sidste 20%. Der fremgår af Figur 19, at førsteaksen modstiller drenge på gymnasiet Romeo (venstre) og piger på gymnasiet Julie (til højre), mens akse nummer to modstiller drenge på gymnasiet Julie (bund) og Piger på gymnasiet Romeo (top). Fordelingen af drenge og piger på de to gymnasier er skæv. Gymnasiet Romeo, der har 80% drenge, kan derfor betragtes som et drengegymsium, mens gymnasiet Julie, der har 80% piger, tilsvarende kan betragtes som et pigegymsium. I det lys modstiller førsteaksen drenge på drengegymsiet (venstre) og piger på pigegymsiet (højre), men anden aksens modstiller drenge på pigegymsiet (bund) og piger på drengegymsiet (top). Den første principale variable eller komponent er dermed i overensstemmelse mellem gymnasium og køn, mens anden principale variabel eller komponent er i uoverensstemmelse mellem gymnasium og køn.



Vi vælger nu at indlægge den tredje variabel (bestået/ikke-bestået) som såkaldt supplementær variabel (modsat køn og gymnasium, der var aktive variable), så får vi afbildet den tredje variabel på den eksisterende graf.⁴

Når variabelen Bestået indlægges, vises tre yderligere punkter: Det gennemsnitlige punkt for *alle* der har bestået, det gennemsnitlige punkt for *piger* der har bestået og det gennemsnitlige punkt for *drenge* der har bestået. Det gennemsnitlige punkt for *alle* der har bestået, orienterer sig på førsteaksen tættere på Julie+piger end på Romeo+Drenge, fordi der er flere Julie+piger, der har bestået end der er Romeo+Drenge. På den lodrette anden akse er punktet tættere på Julie+Drenge end det er på Romeo+Piger, hvilket igen afspejler, at der er færrest af sidstnævnte. Taler vi om effekter kan man summerer resultatet således:



⁴ Havde vi i stedet valgt at medtage den tredje variable som aktiv, ville den være inddraget i aksernes placering, hvorfor de andre punkter ville ændre position.

- Den første principale variabel (overensstemmelse mellem gymnasium og køn) kan fortolkes som udtryk for, at flere piger er bestået på pigegymnasiet end drenge er bestået på drengegymnasiet.
- Den anden principale variabel (uoverensstemmelse mellem gymnasium og køn), kan fortolkes som udtryk for at flere drenge består på pigegymnasiet end piger gør på drengegymnasiet.

Ovenstående er resultatet, når et fænomen som vending udtrykkes i geometriske termer. Det bør bemærkes at selvom de to principale variable ikke korrelerer (de to akser står vinkelret på hinanden (Figur 20)) så omfatter ”effekten” af en principal variabel på den afhængige variabel både en global effekt og betingede effekter på de øvrige principale variable.

Afrunding

Som det er fremgået, kan regressions-procedurer forbindes med geometrisk data analyse. En geometrisk fremstilling kan udtrykke såvel de betingede som de globale effekter af regressionsanalysens variable.

Ulf Brinkkjær, Institut for pædagogik, Danmarks Pædagogiske Universitetsskole ved Århus Universitet

