Interne notater

STATISTISK SENTRALBYRÅ

80/30 2. oktober 1980

KORRELASJONSKOEFFISIENTEN - ENDA ENGANG

Av

H.T. Amundsen

INNHOLD

KORRELASJONSKOEFFISIENTEN - ENDA EN GANG

"Disse variablene forklarer 62 prosent av forskjellen mellom mann og kvinne". (Ikke funnet i Byråets publikasjoner.)

1. Innledning

En stor del av de problemene vi søker å belyse ved hjelp av statistiske observasjonsmaterialer går ut på å undersøke om det er samvariasjon mellom to eller flere variable. Hva er sammenhengen, gjennomsnittlig sett, mellom utgiftene til mat i en husholdning og husholdningens størrelse, dens inntekt osv.? Hvordan varierer månedslønn med yrkesgruppe, utdannelse, kjønn, alder m.m.?

I blandt nøyer vi oss ikke med å sette opp observasjonsmaterialet i tabellform, men forsøker å uttrykke og analysere den eventuelle samvariasjonen ved kvantitative metoder, f.eks. ved regresjonsanalyse. Det er blitt enkelt å kjøre ut regresjonslikninger i mange variable, problemet er å tolke resultatene.

I våre forsøk på å forenkle det kompliserte griper vi ofte til ett enkelt tall som uttrykk for den multiple sammenhengen, og da gjerne verdien av (den multiple) korrelasjonskoeffisienten, som EDB-programmet gir som et av beregningsresultatene. Mål som i en viss forstand er beslektet med den multiple korrelasjonskoeffisienten brukes også ved andre dataanalyseprogrammer, f.eks. i Multiple Classification Analysis (MCA) og Automatic Interaction Detection (AID). Navnet determinasjonskoeffisient er mer betegnende i slike tilfeller.

Når vi forsøker å forenkle beskrivelsen av en komplisert sammenheng så drastisk at vi bare bruker ett enkelt statistisk mål, så er det viktig å være klar over hva det er mulig å få frem ved hjelp av dette målet. Bl.a kan tolkingen av resultatet være ulik for forskjellige problemstillinger og forutsetninger om de variable vi analyserer. Vi vil oppdage at korrelasjons = koeffisienten bare belyser én side av saken, og at vi i de fleste situasjoner bør bruke andre kriterier i tillegg til den eller istedenfor den, når vi skal legge frem resultatene av en statistisk analyse. Som hjelpemiddel under beregningene kan vi imidlertid ha nytte av de ulike typer av korrelasjonskoeffisienter, eller kovarianser.

I dette notatet skal vi forsøke å se helt elementært på hva den vanlige korrelasjonskoeffisienten, den såkalte produktmomentkorrelasjonskoeffisienten, kan gi uttrykk for. Vi må i denne sammenheng også trekke inn litt regresjonsanalyse, men bare som bakgrunn for korrelasjonen.

Vi behandler et observasjonsmateriale for to variable i avsnitt 2. Noe av det vi kommer fram til der, gjelder også for den multiple korrelasjonskoeffisienten som vi ellers ser på i avsnitt 3. I avsnitt 4 tar vi opp de to vanlige enkle teoretiske modellene som ofte postuleres ved regresjons-/korrelasjonsanalyse. Dette gir nødvendig bakgrunn for testing av koeffisientene.

I resten av notatet ser vi på situasjoner der en eller flere variable er av spesiell karakter, som binære variable, ordningsvariable eller variable observert med feil. Med et par unntak nevner vi ikke andre typer samvariasjonsmål, siden hensikten i dette notatet primært er å se på korrelasjons=koeffisienten. I siste avsnitt forsøker vi en kortfattet oppsummering av resultatene. En kort litteraturliste, som det vises til i teksten, står til slutt.

Vi blir ofte skuffet når vi får en korrelasjonskoeffisient som ikke ligger meget nær 1 (eller -1). Vi må imidlertid være klar over at korrelasjon nær 1 er et meget strengt krav: observasjonspunktene skal nesten ligge på en rett linje i spredningsdiagrammet for to variable, eller i et plan (3 variable), respektive hyperplan (flere variable). For binære variable finner vi tilsvarende at korrelasjon nær 1 krever at observasjonene er sterkt konsentrert om visse punkter. Vi må tenke etter om det virkelig er dette vi vil ha et mål for.

Den regresjonen vi har valgt, kan være både fornuftig og nyttig selv om den multiple korrelasjonskoeffisienten ikke er så stor. De parametrene vi er interessert i, kan likevel være godt estimert. Som nevnt, må vi i alminnelighet bruke andre kriterier enn bare en enkelt koeffisient for å bedømme situasjonen. Hvilke, må vi avgjøre ut fra det spesielle problemet vi behandler.

2. Produktmomentkorrelasjonskoeffisienten

Anta at vi har et materiale som består av n observasjonspar, (x_1,y_1), (x_2,y_2), $\ldots(x_n,y_n)$. Både x og y er kvantitative variable målt i vanlig forholdstallskala, slik at både differanser og forholdstall har god mening. I tabell 1 har vi et (konstruert) eksempel fra en tidsnyttingsundersøkelse, der x_i er timetall pr. dag brukt til inntektsgivende arbeid og y_i er timetall brukt til egenarbeid, begge for person nr. i.

Tabell 1. Korrelasjon mellom
x: antall timer brukt til inntektsgivende arbeid, og
y: antall timer brukt til egenarbeid,
for n=5 observasjonspar fra en befolkningsgruppe

Person nr. i	Observerte timer		Avvik fra gj.snitt		Avviksprodukt	Kvadrerte avvik	
	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})\ (y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	3,5	0	0	-2	0	0	4
2	4,5	1	1,0	-1	-1,0	1	1
3	4,0	2	0,5	0	0	0,25	0
4	3,0	3	-0,5	1	-0,5	0,25	1
5	2,5	4	-1,0	2	-2,0	1	4
Sum	17,5	10	0	0	-3,5	2,5	10
$\frac{1}{5}$ sum	3,5	2	0	0	-0,7	0,5	2

Vi finner her det aritmetiske gjennomsnittet for x-ene:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\,17{,}5 = \underline{3{,}5},$$

og tilsvarende for y-ene:

$$\bar{y} = \frac{1}{n}\sum_{i} y_i = \frac{1}{5}\,10 = \underline{2}.$$

Så finner vi det empiriske standardavviket for x-ene som

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{5}\cdot 2{,}5} = \sqrt{0{,}5} = \underline{0{,}707},$$

og for y-ene:

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2} = \sqrt{\frac{1}{5}\cdot 10} = \sqrt{2} = \underline{1{,}414}.$$

Videre finner vi den empiriske kovariansen (produktmomentet) mellom x-ene og y-ene ved

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{5}(-3{,}5) = \underline{-0.7}.$$

Endelig får vi den empiriske korrelasjonskoeffisienten (produktmoment -)

$$r_{xy}, \text{ eller kortere, } r = \frac{s_{xy}}{s_x\, s_y} = \frac{-0{,}7}{0.707\cdot 1{,}414} = \underline{-0{,}7}. \qquad (2.1)$$

Formelen for r forutsetter at både

$s_x > 0$ og $s_y > 0$,

dvs. at det finnes minst to ulike x-verdier og to ulike y-verdier blandt observasjonene. Vi skal forutsette dette i det følgende hvis intet annet sies. I alle tilfelle kan vi imidlertid si at

x-ene og y-ene er ukorrelert hvis $s_{xy} = 0$.

Noen merknader:

Det er en tilfeldighet at $r=s_{xy}$ i dette eksemplet.

Vi bruker betegnelsen "empirisk" for å skille disse størrelsene vi regner ut i observasjonsmaterialet fra de tilsvarende teoretiske størrelsene vi skal se på i avsnitt 4.

Vi har brukt n og ikke (n-1) som divisor i standardavvikene. Hvis vi bruker (n-1), må vi også bruke (n-1) i s_{xy} for at r skal bli riktig definert.

Oppstillingen i tabellene er valgt for å illustrere sammenhengene, selve beregningene kan oftest gjøres mer praktisk på annen måte.

I dette eksemplet har vi <u>negativ</u> korrelasjon mellom x- og y-verdiene. Det fremgår av tabellen hva dette kommer av. Tallene er her ordnet etter <u>stigende</u> verdier av den ene variable, nemlig y. Vi ser at bortsett fra første observasjon, så <u>avtar</u> x-ene når y-ene <u>stiger</u>. Dette fører til at de fleste avvikene $(x_i-\bar{x})$ får <u>motsatt fortegn</u> av sin tilsvarende $(y_i-\bar{y})$ og dermed blir de fleste produktene $(x_i-\bar{x})(y_i-\bar{y})$ negative. Summen av dem blir negativ og det gjør også s_{xy} og r. Negativ korrelasjon betyr at det er en tendens til at verdiene av den ene variable avtar når den andre øker, eller omvendt.

I et annet eksempel der x_i-verdiene gjennomgående <u>stiger</u> med <u>stigende</u> y_i-verdier vil de fleste avviksparene ha samme fortegn, enten begge negative eller begge positive. Da blir de fleste produktene positive og det samme gjelder s_{xy} og r. Vi har da <u>positiv</u> korrelasjon mellom de to variable, jfr. tabell 2 nedenfor.

I noen eksempler kan det tenkes at f.eks. ett stort positivt produkt oppveier en rad av negative (eller omvendt). Da kan vi få en s_{xy} nær null. Det får vi også hvis det er liten tendens til samvariasjon mellom x-ene og y-ene, dvs. at produktene $(x_i-\bar{x})(y_i-\bar{y})$ har skiftende fortegn, og dermed en tendens til å oppveie hverandre ved summeringen. Det er <u>liten</u> korrelasjon i materialet.

Vi kan avbilde observasjonene som punkter i et spredningsdiagram, idet vi avsette x_i langs den vanrette aksen og y_i langs den loddrette. Tallverdien på r henger nøye sammen med hvor godt punktene samler seg om en rett linje i diagrammet. Vi skal her se på minste kvadraters regresjonslinjen for y m.h.p. x.

Vi skriver linjens likning på formen

$y = a+bx.$

Her er a og b konstanter som vi bestemmer ut fra observasjonsmaterialet på en slik måte at kvadratsummen

$$q = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad (2.2)$$

blir minst mulig. Løsningen av dette gir

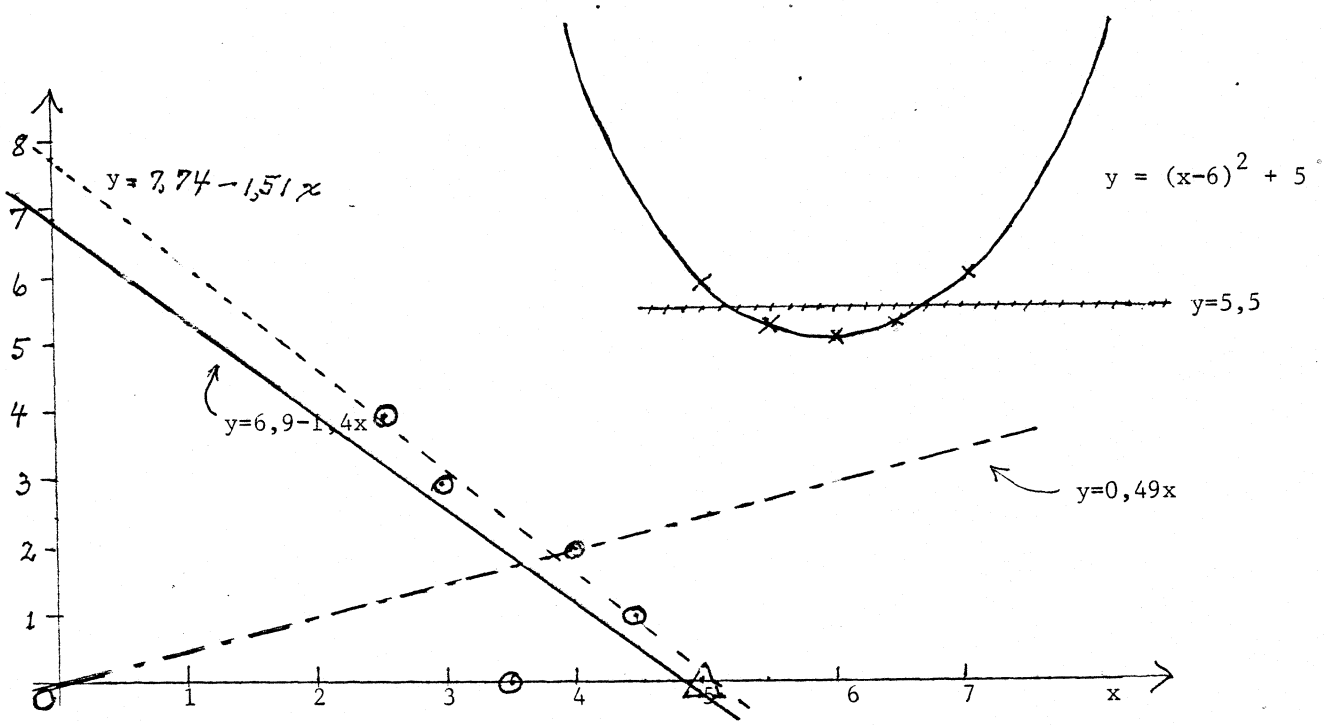
$$b = \frac{s_{xy}}{s_x^2} = \frac{-0,7}{0,5} = \underline{-1,4}$$

og

$a = \bar{y} - b\bar{x} = 2-(-1,4)\cdot 3,5 = \underline{6,9}$, dvs. vi har $y=\bar{y}+b(x-\bar{x})$.

Linjen har altså likningen

$y = 6,9 - 1,4x = 2-1,4(x-3,5)$.



Spredningsdiagram 1. Punktene fra tabell 1 avmerket med sirkler, regresjonslinjen er heltrukket.

Det endrede punktet (5,0) er merket med en trekant, og regresjonslinjen er streket opp for dette andre eksemplet.

Videre er minste kvadraters linjen uten konstantledd stiplet inn.

Eksemplet i tabell 3 er angitt ved kryssene øverst til høyre.

Regresjonskoeffisienten b har altså samme fortegn som s_{xy} og r.

Vi bruker ofte restavviket (residualavviket) s som et mål for observasjonspunktenes spredning rundt regresjonslinjen. Vi definerer restvariansen s^2 ved

$$s^2 = \frac{1}{n} q_o, \tag{2.3}$$

der q_o er minimum av q, dvs. den verdien vi får når vi setter de funne a- og b-verdiene inn i uttrykket for q.

Det viser seg at vi kan forenkle dette uttrykket, så vi får

$$s^2 = s_y^2 (1-r^2), \text{ dvs. } s = s_y \sqrt{1-r^2}. \tag{2.4}$$

Siden hverken s^2 eller s_y^2 kan være negative tall, så viser dette at r^2 ikke kan være større enn 1, dvs. at vi alltid må ha

$$-1 \leqq r \leqq 1.$$

Hvis r = 1 eller r = -1, så er s^2, og q_o, lik null, dvs. at alle punktene (x_i, y_i) ligger på den rette linjen. Hvis r^2 er nær 1, så må s^2 være liten i forhold til s_y^2, og punktene ligger nær linjen.

I eksemplet finner vi

$s^2 = 2(1-0,49) = 1,02$ og $s = 1,01$.

Punktene ligger nokså nær regresjonslinjen.

Hvis vi endrer eksemplet slik at $x_1 = 5$, mens alle de andre tallene er uendret, så finner vi at alle punktene ligger meget nær regresjonslinjen, som nå blir

$$y = 7,74 - 1,51x,$$

jfr. diagrammet. Vi får i dette eksmplet

$\bar{x} = 3,8$, $s_x^2 = 0,86$, $s_{xy} = -1,3$, $r = \underline{-0,99}$,

$b = -1,51$, $a = 7,74$, $s^2 = 0,035$ og $s = \underline{0,187}$.

Hvis r = 0, så er $s = s_y$ og linjen har b = 0, dvs. den går vannrett: $y = \bar{y}$ for alle x.

Av uttrykket (2.4) set vi at vi også kan skrive

$$r^2 = 1 - \frac{s^2}{s_y^2} = \frac{s_y^2 - s^2}{s_y^2} \tag{2.5}$$

I en del lærebøker og EDB-programmer blir "korrelasjonskoeffisienten" definert ved uttrykket til høyre. I nyere litteratur kalles dette for determinasjonskoeffisienten (coefficient of determination). Vi må imidlertid være oppmerksom på at r^2 og $(s_y^2 - s^2)/s_y^2$ ikke behøver å falle sammen hvis regresjonsligningen er en annen enn y=a+bx, eller funnet på annen måte enn ved å minimere kvadratsummen q i (2.2) (vi kunne trukket linjen på øyemål, f.eks.).

Som eksempel på en annen ligning enn y = a + bx, vil vi føye en linje uten konstantledd til data i tabell 1. Linjen har da ligningen

$$y = cx,$$

og vi bestemmer c ved å minimere kvadratsummen

$$u = \sum_{i=1}^{n} (y_i - cx_i)^2.$$

Da finner vi

$$c = \frac{\Sigma y_i x_i}{\Sigma x_i^2} = \frac{31,5}{63,75} = \underline{0,49},$$

og

$$s^2 = \frac{1}{5} \sum_i (y_i - 0,49x_i)^2 = 2,887.$$

Dette gir

$$\frac{s_y^2 - s^2}{s_y^2} = \frac{2-2,887}{2} = -0,44,$$

som ikke har noe med kvadratet av r å gjøre, men som selvsagt indikerer at en linje uten konstantledd passer temmelig dårlig til våre data, jfr. diagram 1.

Vi skal vise tilknytningen mellom r og minste kvadraters regresjonslinjen på enda en måte. For hver observert x_i-verdi kan vi regne ut den y-verdien, y_i^*, som vi får ved å sette inn x_i i regresjonslikningen. For i=1 har vi f.eks. i vårt første eksempel:

$$y_1^* = 6,9 - 1,4 \cdot 3,5 = 2,0,$$

osv., som i tabell 2.

Så vil vi finne produktmomentkorrelasjonen mellom y_i- og y_i^*- verdiene.

Tabell 2. Korrelasjon mellom observerte y_i-verdier og beregnede y_i^*-verdier ut fra regresjonslikningen.

Person nr. i	y_i	y_i^*	$y_i - \bar{y}$	$y_i^* - \bar{y}^*$	$(y_i - \bar{y})(y_i^* - \bar{y}^*)$	$(y_i - \bar{y})^2$	$(y_i^* - \bar{y}^*)^2$
1	0	2	-2	0	0	4	0
2	1	0,6	-1	-1,4	1,4	1	1,96
3	2	1,3	0	-0,7	0	0	0,49
4	3	2,7	1	0,7	0,7	1	0,49
5	4	3,4	2	1,4	2,8	4	1,96
Sum	10	10,0	0	0	4,9	10	4,90
$\frac{1}{5}$ sum	2	2			0,98	2	0,98

Vi finner gjennomsnittet $\bar{y}^* = 2$, $s_y^{2*} = 0,98$, $s_{y,y}^{2*} = 0,98$ og

$$r_{y,y^*} = \frac{s_{y,y^*}}{s_y s_{y^*}} = \frac{0,98}{1,414 \cdot 0,99} = \frac{0.98}{1.4} = 0.7.$$

Den kvadrerte korrelasjonskoeffisienten er altså 0,49 = r^2. Vi kan vise at dette gjelder generelt: korrelasjonen mellom y_i- og y_i^*-verdiene er aldri negativ og har samme tallverdi som r_{xy}.

Vi kan vise at sammenhengene

$$\bar{y}^* = \bar{y},\quad s_{y^*}^2 = s_y^2 - s^2 \quad \text{og} \quad s_{y,y^*} = s_y^2 - s^2 \geq 0, \tag{2.6}$$

gjelder generelt når y_i^* er beregnet ut fra minste kvadraters regresjonslikningen (A.II, s.188). Dette gjelder også for regresjoner med flere x-er (multippel regresjon, se avsnitt 3).

Det er viktig å merke seg at det er tendensen til lineær sammenheng mellom x og y vi eventuelt kan si noe om ved hjelp av r.

I eksemplet i tabell 3 ligger alle punktene på kurven

$$y = (x-6)^2 + 5$$

dvs. at når x er gitt, så kan vi regne ut y nøyaktig, det er funksjonell avhengighet mellom x_i og y_i. Likevel skal vi se at r=0. Jfr. diagram 1.

Tabell 3. Ikke-lineær samvariasjon. Korrelasjon mellom x og y når $y=(x-6)^2 + 5$ for x = 5, 5,5 , 6, 6,5 , 7.

Person nr. i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{y})(y_i-\bar{y})$
1	5	6	-1	0,5	-0,5
2	5,5	5,25	-0,5	-0,25	0,125
3	6	5	0	-0,5	0
4	6,5	5,25	0,5	-0,25	-0,125
5	7	6	1	0,5	0,5
Sum	30	27,50	0	0	0
$\frac{1}{5}$ sum	6	5,5			0

Vi har $s_{xy} = 0$ og dermed r=0 og b=0 samt $a=\bar{y}=5,5$.

Hvis vi i eksemplet isteden hadde hatt verdier for x = 3,4,5,5,5 og 6, ville vi fått en negativ r. Med verdier fra x = 6 og oppover, ville vi fått en positiv r (men vi kan ikke få $r^2=1$). Hvis vi i dette eksemplet bruker $y = (x-6)^2 + 5$ som regresjonskurve, blir $s^2 = 0$ og dermed $(s_y^2 - s^2)/s_y^2 = 1$, fordi punktene ligger på kurven.

Uttrykket

$$(s_y^2 - s^2)/s_y^2$$

kalles altså determinasjonskoeffisienten. Det gir jo den relative forskjellen mellom den opprinnelige variansen, s_y^2, for y-ene og variansen s^2 målt rundt

regresjonslinjen(-kurven). Tenker vi på x som en "forklaringsvariabel" for y, så synes vi at x "forklarer" mye av variasjonen i y hvis s^2 er liten, mens "forklaringen" er dårlig når s^2 er nær s_y^2. For vanlig lineær regresjon viser (2.4) at s^2 er liten når r^2 er nær 1 og s^2 er nær s_y^2 når r^2 er liten, og omvendt.

En ser ofte utsagn som at "x forklarer 100 r^2 prosent av variasjonen i y". Dette betyr da at variansen for y-ene er redusert med 100 r^2 prosent. Men vi måler jo gjerne variasjonen i standardavvik, tenk f.eks. på utsagn som "gjennomsnittet $\pm$ 2 standardavvik". Hvis vi vil se på hvor mye av standardavviket s_y som er "forklart" ved regresjonen, så må vi se på

$$\frac{s_y - s}{s_y} = \frac{s_y - s_y\sqrt{1-r^2}}{s_y} = 1 - \sqrt{1-r^2}$$

Denne reduksjonen er alltid mindre enn r^2, når $0<r^2<1$, jfr. tabell 4.

Tabell 4. Forholdsvis reduksjon i standardavviket for y for ulike verdier av r^2, (eller $(s_y^2-s^2)/s_y^2$).

Tallverdi $\lvert r\rvert$	1	0,995	0,975	0,95	0,90	0,80	0,70	0,50	0,40	0,30
r^2	1	0,99	0,95	0,90	0,81	0,64	0,49	0,25	0,16	0,09
$1-\sqrt{1-r^2}$	1	0,90	0,78	0,68	0,56	0,40	0,29	0,13	0,08	0,05

Vi ser at reduksjonen i standardavviket kan være nokså moderat selv for store tallverdier av r. I eksemplet i tabell 1 finner vi

$$1 - \sqrt{1-0,49} = 0,29.$$

I det modererte eksemplet med $x_1=5$ finner vi

$$1 - \sqrt{1-0,9825} = 0,87.$$

Både disse tallene og eksemplet med ikke-lineær samvariasjon sier vel noe om hvor mye det skal til for å "forklare" variasjonen i en variabel ved hjelp av en annen, eller for å påvise "god" samvariasjon mellom to variable, i den forstand at r skal være nær 1 i tallverdi. På den annen side så kan samvariasjonen være av betydning selv om den ikke er så sterk. Men her kommer vi på gyngende grunn. Hittil har vi bare sett på og "beskrevet" observasjonsmaterialet på ymse vis. Hvis hensikten med analysen er å trekke slutninger om samvarisjon mellom de variable ut over dette observasjonsmaterialet, så er det ikke nok å legge fram de tallene som vi har funnet. Vi må også undersøke hvilke usikkerheter som hefter ved dem når vi vil tolke dem

som estimater for tilsvarende teoretiske størrelser, eller parametre, i den modell vi velger å stille opp for sammenhengen mellom våre variable. Vi skal se litt på dette i avsnitt 4. Først skal vi i avsnitt 3 se på den multiple korrelasjonskoeffisienten.

3. Den multiple korrelasjonskoeffisienten

Vi har ofte et materiale med observasjoner av flere enn to variable for hver observasjonsenhet. For person nr. i kan vi i eksemplet foran ha observasjonen y_i, $x_{1i},\ldots,x_{ki}$,
der y_i er antall timer brukt til egenarbeid,
x_{1i} er antall timer brukt til inntektsgivende arbeid,
x_{2i} er antall barn
.....
x_{ki} er alder, f.eks.

I tabell 5 har vi utvidet det konstruerte eksemplet i tabell 1. Vi har k = 4, med x_{3i} for bosted (1 for by, 0 for land) og x_{4i} for aldersgruppe (1 for 40 år og over, 0 for under 40 år).

Tabell 5 Konstruert eksempel for n = 5 observasjonssett fra en befolkningsgruppe
y: antall timer brukt til egenarbeid
x_1: antall timer brukt til inntekts = givende arbeid
x_2: antall barn
x_3: bosted
x_4: aldersgruppe

Person nr. i	Observasjoner				
	y_i	x_{1i}	x_{2i}	x_{3i}	x_{4i}
1	0	3,5	0	1	0
2	1	4,5	0	0	0
3	2	4,0	1	0	1
4	3	3,0	1	0	0
5	4	2,5	2	1	0
Gj.snitt	2	3,5	0,8	0,4	0,2
Varians	2	0,5	0,56	0,24	0,16
s_{yxj}		- 0,7	1	0	0
r_{yxj}		- 0,7	0,9	0	0

For å kunne beregne regresjons- og korrelasjonskoeffisientene trenger vi variansene og kovariansene for x-ene,

$$s_{tj} = s_{jt} = \frac{1}{5}\sum_{i=1}^{5}(x_{ji}-\bar{x}_j)(x_{ti}-\bar{x}_t) \text{ for } j = 1,2,3,4 \text{ og } t = 1,2,3,4.$$

Vi samler dem i kovariansmatrisen

$$M = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix} = \begin{bmatrix} 0,5 & -0,4 & -0,2 & 0,1 \\ -0,4 & 0,56 & 0,08 & 0,04 \\ -0,2 & 0,08 & 0,24 & -0,08 \\ 0,1 & 0,04 & -0,08 & 0,16 \end{bmatrix}$$

På tilsvarende måte som for to variable kan vi beregne minste kvadraters regresjonsligningen for y m.h.p. x_1, x_2,...,x_k. Dvs. vi bestemmer konstantene a, b_1, b_2,..,b_k i likningen

$$y = a + b_1x_1 + b_2x_2 \ldots + b_kx_k \qquad (3.1)$$

slik at vi minimerer kvadratsummen

$$Q = \sum_{i=1}^{n}(y_i-a-b_1x_{1i}-b_2x_{2i}-\ldots-b_kx_{ki})^2.$$

Restvariansen kan vi også her sette lik $s^2 = \frac{1}{n}Q_o$, der Q_o er minimum av Q.

Regresjonen for y mhp x_1 og x_2 (altså k = 2) blir for tallene i tabell 5:

$y = 0,3 + 0,067x_1 + 1,833x_2$, med restvarians 0,233 samt $R^2 = 0,89$ og $R = 0,945$. Se nedenfor om R.

Regresjonen mhp x_1,x_2 og x_3 blir:

$y = 3,3 - 0,6x_1 + 1,5x_2 - x_3$, med restvarians 0,08 samt $R^2 = 0,96$ og $R = 0,98$

Og regresjonen mhp x_1,x_2,x_3 og x_4 blir:

$y = 1 + 0x_1 + 2x_2 - x_3 - x_4$, med restvarians 0, samt $R^2 = R = 1$.

Bortsett fra helt spesielle tilfeller av ukorrelerthet mellom y og x-ene, vil vi alltid få s = 0 og R = 1 når vi har like mange koeffisienter i likningen som vi har observasjonssett (her 5).

Ellers ser vi i dette eksemplet hvordan koeffisienten til x_1 endrer seg ettersom vi trekker inn fler variable. I den siste regresjonen har x_2,x_3 og x_4 "overtatt hele forklaringen" fra x_1. Dette er spesielt for dette eksemplet (og er vel et ikke urimelig resultat med den tolkingen vi har gitt de variable). Hvis vi beregner regresjonen for y mhp x_2,x_3 og x_4 alene, finner vi her

$y = 1 + 2x_2 - x_3 - x_4$, med restvarians 0 og R = 1 dvs. at når x_2, x_3 og x_4 er gitt, kan vi beregne y nøyaktig (slik tallene i tabell 5 er).

For å bedømme hvor "god" en regresjon er, kan vi se på restvariansen, men det vanlige er jo å se på den mutiple korrelasjonskoeffisienten, R. Denne kan vi bestemme ved å gå frem som i tabell 2 foran. Vi kan regne ut y_i^*- verdier ved å sette inn $(x_{1i}, x_{2i},\ldots,x_{ki})$ i regresjonslikningen (3.1), og så kan vi beregne produktmomentkorrelasjonen $r_{y.12\ldots k}$, populært kalt R, som blir (jfr. merknadene etter tabell 2 foran):

$$R = r_{y\cdot 12\ldots k} = \frac{s_{yy^*}}{s_y\, s_{y^*}} = \frac{s_y^2 - s^2}{s_y\sqrt{s_y^2 - s^2}} = \sqrt{\frac{s_y^2 - s^2}{s_y^2}} \quad . \tag{3.2}$$

Vi har altså $R \geq 0$, og

$$R^2 = \frac{s_y^2 - s^2}{s_y^2} \quad , \text{ eller } s^2 = s_y^2(1-R^2). \tag{3.3}$$

også i dette tilfelle.

Vi får som i avsnitt 2 at likheten i (3.2) gjelder når y_i^*-verdiene og s^2 er beregnet ut fra minste kvadraters regresjonen (3.1). Det eneste som skiller R fra en "vanlig" korrelasjonskoeffisient er at den ikke kan bli negativ. Ellers blir "tolkingen" av R og s som i avsnitt 2 :

R = 1, dvs. s=0, når alle observasjonssettene passer i likningen (3.1), slik som i eksemplet når både x_2,x_3 og x_4 er med.

R nær 1, s nær 0, når det er små divergenser mellom y_i og y_i^* ved innsetting av $x_{1i}, x_{2i},\ldots,x_{ki}$ i (3.1), som i eksemplet der x_2 og x_3 men <u>ikke</u> x_4, er med.

R = 0, $s = s_y$, når likningen blir $y = \bar{y}$, dvs. alle b-ene lik null, da blir $\frac{1}{n} Q_o = s_y^2$. Dette får vi et eksempel på ved å beregne regresjonen for y mhp x_3 og x_4, som gir y = 2.

Vi ser at enda x_3 og x_4 både hver for seg og sammen er ukorrelert med y, så bidrar de til redusert varians, og øket multipel korrelasjon, når de kommer inn i regresjonen sammen med x_1 og/eller x_2. Dette henger sammen med at x-ene er innbyrdes korrelert. Denne innbyrdes korrelasjonen er også grunnen til at regresjonskoeffisienten til x_1 endrer seg ettersom flere variable trekkes inn.

Av (3.2) ser vi at tolkingen av R^2 som den relative reduksjon i variansen for y, er som i avsnitt 2 og vi kan bedømme reduksjonen i standardavviket på akkurat samme måte.

Hvis vi måler s^2 ut fra en annen likning enn (3.1), vil vi derimot ikke uten videre kunne tolke determinasjonskoeffisienten

$$\frac{s_y^2 - s^2}{s_y^2} \tag{3.4}$$

som den multiple korrelasjonskoeffisienten mellom y på den ene siden og settet $(x_1, x_2, \ldots, x_k)$ på den annen. Med s^2 beregnet ut fra denne andre likningen, kan vi selvsagt nytte (3.4) eller $(s_y - s)/s_y$ som et mål for redusert usikkerhet hvis vi ønsker det, men vi må i tilfelle undersøke nærmere om det er mulig å tolke (3.4) som en korrelasjonskoeffisient mellom observerte og beregnede y-verdier (hvis det er av interesse).

Det er blitt vanlig å beregne flere multiple regresjoner for y m.h.p. ulike sett av x-variable fra samme observasjonsmateriale for å finne det x-settet som "best forklarer" variasjonen i y-ene. En velger da det x-settet som gir størst verdi av den multiple korrelasjonskoeffisienten (dvs. det som gir minst restvarians). I tillegg til spørsmålet om usikkerhet (jfr. avsnitt 4), og om det har mening ut fra vårt problem å velge x-ene slik, så er det også noen algebraiske sammenhenger å være oppmerksom på. Sett at vi beregner følgende regresjoner etter tur, og kaller restvariansene henholdsvis $s_1^2, s_2^2, \ldots, s_k^2$ og koeffisientene $R_1^2, R_2^2, \ldots R_k^2$. Regresjonen for y er altså henholdsvis m.h.p. x_1 alene, x_1 og x_2, x_1, x_2 og x_3, ..., x_1, x_2, x_3, ..., x_k. Da er, som vi også ser i eksemplet ovenfor,

$$s_1^2 \geqq s_2^2 \geqq s_3^2 \geqq \ldots\ldots \geqq s_k^2$$

og

$$R_1^2 \leqq R_2^2 \leqq R_3^2 \leqq \ldots\ldots \leqq R_k^2.$$

Vi får altså i alminnelighet bedre tilpassing jo flere "forklaringsvariable" vi tar med (se A.II.s.190).

(I enkelte dataprogrammer ser det ut som om ulikhetene ikke holder. Det skyldes gjerne at $\frac{Q_o}{n-k-1}$ er brukt istedenfor $s^2 = \frac{Q_o}{n}$. Siden nevneren da avtar når k vokser, kan vi få stigende restvarianser hvis Q_o blir lite eller ikke redusert når vi trekker inn flere x-er i regresjonen.)

Vi ser av eksemplet at vi ikke kan bedømme "godheten" av den multiple regresjonen (3.1) ved å se på korrelasjonskoeffisientene mellom y og hver enkelt av de x-variable. La oss kalle dem r_{y1} for y og x_1, r_{y2} for y og x_2, osv. r_{yk} for y og x_k.

Vi kan da vise at

$$R^2 = r^2_{y\,.123\ldots k} \leqq r_{y1}^2 + r_{y2}^2 + \ldots + r_{yk}^2.$$

Likhetstegnet gjelder når og bare når alle x-variablene er parvis ukorrelert med hverandre. Hvis de er det, så vil "virkningen" av dem være additiv i den forstand at

$$R^2 = r_{y1}^2 + r_{y2}^2 + \ldots + r_{yk}^2.$$

Dette vil meget sjelden være tilfelle i praksis, medmindre vi transformerer de variable så vi får en såkalt ortogonal regresjon.

I eksemplet har vi

$$r_{y_1}^2 + r_{y_2}^2 + r_{y_3}^2 + r_{y_4}^2 = 0{,}49 + 0{,}81 + 0 + 0 = 1{,}3$$

Det er en spesiell fare ved å sammenlikne regresjoner fra ulike observasjonsmaterialer ved hjelp av R-verdiene, dette forsøker vi å forklare i avsnitt 4b nedenfor.

4. Er koeffisientene signifikante?

Når vi vil trekke slutninger om samvariasjon mellom de variable ut over observasjonsmaterialet, må vi bygge på en modell. For å kunne bedømme usikkerheten ved våre utsagn tyr vi til sannsynlighets- (stokastiske) modeller. Vi kan ikke gi noen innføring i regresjonsteorien her, men viser til lærebøker, som f.eks. AI, kap. 7 (for to variable) og AII, kap. 12. Vi skal bare antyde visse grunnleggende trekk i teorien.
Vi tenker oss at hver y_i er en observasjon av en teoretisk variabel som har en sannsynlighetsfordeling. Denne fordelingen kan ha visse parametre, f.eks. teoretisk gjennomsnitt (forventning) og teoretisk varians, respektive

$$\mu_y \text{ og } \sigma_y^2,$$

som bestemmer den nærmere. Disse parametrene kan vanligvis tolkes som tall som henholdsvis $\bar{y}$ og s_y^2 vil nærme seg mot ved beregning ut fra stadig flere observasjoner, når modellen gjelder.

Videre vil vi her være interessert i om fordelingen for y kan avhenge av x-ene, vi tenker oss f.eks. at x-ene kan være parametre i fordelingen. Vi skal her bare se på den vanlige, enkle regresjonsmodellen, dvs. at forventning og varians for y for gitte x-er kan skrives

$$\mu_{y|x_1,x_2,\ldots,x_k} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad (4.1a)$$

$$\text{var}(y|x_1, x_2, \ldots x_k) = \sigma^2, \text{ (dvs. den samme for enhver x-kombinasjon).} \qquad (4.1b)$$

Forutsatt at ikke noe annet følger av den måten observasjonene er kommet fram på, vil vi dessuten postulere at fordelingen for hver y_i er uavhengig av de andre y-ene vi har med, og at (4.1) gjelder for hver av dem.

Når det gjelder x-ene kan vi ha to ulike situasjoner (eller en kombinasjon)

- x-ene kan betraktes som ikke-stokastiske, de opptrer som observerbare parametre i problemet. Dvs. at x-ene på en eller annen måte kan betraktes som "forutbestemt", enten av oss selv eller av andre, når y-ene (og x-ene) observeres.
- x-ene kan betraktes som stokastiske. Da har vi for hver i at settet $(y_i, x_{1i}, x_{2i},\ldots,x_{ki})$ er observert fra en simultan (k+1-dimensjonal) sannsynlighetsfordeling. Dette behøver ikke være så vanskelig som det høres ut, for vi postulerer også her at (4.1) gjelder (nå som såkalt betinget forventning og varians). Det er først og fremst (4.1a) som interesserer oss når vi vil fram til utsagn om samvariasjonen mellom y og x-ene.

I virkeligheten er det denne siste modellen som er den greieste når det gjelder å tolke korrelasjonskoeffisienter, noe vi skal se nedenfor.

De vanlige testmetodene som brukes i regresjonsanalyser, som t-tester og F-tester, forutsetter enda mer om y-fordelingen, nemlig at

(4.1c) den teoretiske fordelingen (eventuelt betingede fordelingen gitt x-ene) som y_i er trukket fra er en normal (gaussisk) sannsynlighetsfordeling. Denne forutsetningen er spesielt viktig for mindre observasjonsmaterialer. For stor n (hvor stor, avhenger av hva vi tester) vil som oftest testene være tilnærmet riktige selv uten normal fordeling for y-ene.

Når(4.1a-c) er oppfylt og y-ene er uavhengige,kan vi vise matematisk at størrelser som t i (4.3) nedenfor har en sannsynlighetsfordeling som kalles t-fordelingen, med en parameter som kalles "antall frihetsgrader" og som her er lik (n-2). Vi kan også vise at F i (4.5) og i (4.7) følger F-fordelingen. Denne har to parametre: "antall frihetsgrader for telleren og for nevneren", som er henholdsvis 1 og (n-2) for (4.5) og k og (n-k-1) for (4.7). Disse fordelingene er tabulert, og programmer for dem er lagt inn i regresjonsprogrammene.

4a. Teoretisk korrelasjon når alle variable er stokastiske

Når y og x har en simultan sannsynlighetsfordeling, kan vi definere

den teoretiske kovariansen σ_{xy}

samt

de teoretiske standardavvikene σ_x og σ_y,

og

den teoretiske korrelasjonskoeffisienten

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.2}$$

helt analogt med de empiriske størrelsene i avsnitt 2. (Vi bruker nå forventningsoperatoren istedenfor å summere over i og dele med n).

Vi vil finne at

$-1 \leq \rho \leq 1$,

og vi kan vise at det er eksakt lineær sammenheng mellom x og y hvis og bare hvis $\rho=1$. Vi vil si at x og y er ukorrelert hvis $\sigma_{xy} = 0$, dvs. $\rho = 0$.

Når betingelsene (4.1) er oppfylt, kan vi teste hypotesen $\rho = 0$ mot alternativet $\rho \neq 0$ ved å teste om regresjonskoeffisienten β til x er lik null.

Vi bruker t-testen og forkaster nullhypotesen hvis den observerte t-verdien avviker mye fra null. Forventningen i t-fordelingen er nemlig null når $\rho = 0$. Med valgt sannsynlighetsnivå 5 prosent, forkaster vi nullhypotesen hvis vår observerte t enten er mindre enn 2,5-prosent fraktilen (den t-verdien som det er 2,5- prosent sannsynlighet for å komme under hvis $\rho = 0$) eller større enn 97,5-prosentfraktilen (den t-verdien det er 2,5 prosent sannsynlighet for å komme over). Vi beregner

$$t = \frac{b}{s} s_x \sqrt{n-2} = \frac{-1.4}{1.01} \, 0{,}707 \, \sqrt{3} = -1{,}70. \tag{4.3}$$

Vi finner her at denne ikke er mindre enn $2\frac{1}{2}$-prosentfraktilen -3,18 i t-fordelingen med 3 frihetsgrader, dvs. vi kan ikke forkaste hypotesen $\rho = 0$ i dette tilfelle. (Hvis vi hadde hatt 102 observasjoner, men ellers samme verdi av b, s og s_x, ville vi fått t = -9,8 og forkasting av null-hypotesen).

Det er lett å se at t-testen har noe med r og ρ å gjøre hvis vi setter inn for b og s fra formlene i avsnitt 2. Da finner vi

$$b = r \cdot \frac{s_y}{s_x}, \quad s = s_y \sqrt{1-r^2}$$

og

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} = \frac{-0.7}{\sqrt{0{,}51}} \sqrt{3} = -1{,}70. \tag{4.4}$$

Istedenfor t-testen kunne vi her brukt en F-test som gir nøyaktig samme resultat. Vi har nemlig at $F = t^2$ er F-fordelt med 1 frihetsgrad for telleren og n-2 for nevneren. Vi har

$$F = t^2 = \frac{r^2}{1-r^2}(n-2) = \frac{0{,}49}{0{,}51} \cdot 3 = 2{,}88 \tag{4.5}$$

En stor F-verdi vil indikere at nullhypotesen er gal, vi forkaster nullhypotesen hvis F er større enn 5-prosent fraktilen i F-fordelingen med 1 frihetsgrad for telleren og 3 for nevneren, dvs. 10,13 (lik $(-3,18)^2$). Dette er altså ikke oppfylt her. (Det er litt unøyaktighet i siste desimal p.g.a. avrunding).

Den teoretiske multiple korrelasjonskoeffisienten kan vi definere på tilsvarende måte som i tabell 2 ved å ta den teoretiske korrelasjonskoeffisienten mellom y og tilsvarende y(x) fra regresjonen (4.1a).

Vi vil da finne analogt med (3.2)

$$\rho_{y \cdot 12\ldots k} = \frac{\sigma_{y,y(x)}}{\sigma_y \sigma_{y(x)}} = \sqrt{\frac{\sigma_y^2 - \sigma^2}{\sigma_y^2}} \tag{4.6}$$

Vi kan også vise at vi kan skrive

$$\rho^2_{y \cdot 12\ldots k} = \frac{1}{\sigma_y} \sum_{j=1}^{k} \beta_j \sigma_{yj},$$

der σ_{yj} er de teoretiske kovariansene mellom y og x_j-variablene.

Å teste $\rho_{y \cdot 12\ldots k} = 0$ er det samme som å teste om samtlige β-er er null. Under forutsetningene (4.1) kan vi gjøre dette ved hjelp av

$$F = \frac{R^2}{1-R^2} \frac{n-k-1}{k} \tag{4.7}$$

Vi forkaster nullhypotesen hvis den observerte F er større enn (f.eks.) 5 prosent fraktilen i F-fordelingen med k frihetsgrader for telleren og (n-k-1) frihetsgrader for nevneren.

NB: Hvis vi vil bruke determinasjonskoeffisienten fra andre likninger enn (4.1a) for å teste sammenhenger på tilsvarende måte, kan vi ikke uten videre bruke F som angitt ovenfor. Vi må i tilfelle først undersøke fordelingen av (4.7) ut fra den problemstilling vi har.

I tabell 6 gir vi noen eksempler på hvor stor R^2 (og R) må være for å være "signifikant forskjellig fra null" på 5 prosent sannsynlighetsnivå, dvs. for at vi skal kunne forkaste hypotesen: $\rho = 0$. Vi trenger stor R-verdi for å gjøre dette i små utvalg, (men manglende forkasting innebærer ikke nødvendigvis at $\rho = 0$). For middels og store utvalg kan også "middels" og små verdier være signifikante.

Tabell 6. Hvor stor må R^2 (og R) minst være for å være signifikant forskjellig fra null på 5 prosent nivå?. Forutsetninger: (4.1).

Antall fr.gr. nevner	Antall frihetsgrader teller: 1			5			10			24		
	$F_{0.95}$	R^2	R	$F_{0.95}$	R^2	R	$F_{0.95}$	R^2	R	$F_{0.95}$	R^2	R
5	6,61	0,57	0,75	5,05	0,83	0,91	4,74	0,90	0,95	4,53	0,96	0,98
20	4,35	0,18	0,42	2,71	0,40	0,64	2,35	0,54	0,74	2,08	0,56	0,75
40	4,08	0,09	0,30	2,45	0,23	0,48	2,08	0,34	0,58	1,79	0,52	0,72
60	4,00	0,06	0,25	2,37	0,16	0,41	1,99	0,25	0,50	1,70	0,40	0,64
120	3,92	0,03	0,18	2,29	0,09	0,30	1,91	0,14	0,37	1,61	0,24	0,49
ca.1000	3,82	0.004	0,06	2,20	0,01	0,10	1,84	0,02	0,13	1,50	0,03	0,19

4b. "Teoretisk korrelasjon" når x-ene ikke er stokastiske

Når y er stokastisk, mens x-ene må betraktes som "gitte tall", eller observerbare parametre, kan vi ikke definere teoretiske korrelasjonskoeffisienter ved (4.2) eller (4.6). Likevel har jo regresjonen (4.1a) god mening og vi kan ha normalt fordelte y-er, dvs. en såkalt lineærnormal modell. Vi kan da teste $\beta = 0$ ved testen (4.3) og $\beta_1=\beta_2=\ldots=\beta_k=0$ ved testen (4.7).

Hvis vi ønsker å gi mening til begrepet "teoretisk korrelasjonskoeffisient" i dette tilfelle, kan vi definere den ved

$$\rho^2_{y\cdot 12\ldots k} = 1 - \frac{\sigma^2}{\sigma^2_y}$$

der vi nå setter den "marginale" variansen lik

$$\sigma^2_y = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{n}\Big(\sum_{j=1}^{k}\beta_j(x_{ij}-\bar{x})\Big)^2.$$

(Dette er forventningsverdien av $\frac{1}{n-1}\sum_i(y_i-\bar{y})^2$.)

Vi ser at vi har som i 4a,

$$\rho^2_{y\cdot 12\ldots k} = 1 \text{ når og bare når } \sigma^2 = 0$$

$$\rho^2_{y\cdot 12\ldots k} = 0 \text{ når og bare når } \beta_j = 0 \text{ for } j = 1,2,\ldots k.$$

Vi ser imidlertid at størrelsen på $\rho_{y\cdot 12\ldots k}$ avhenger av valget av x_{ij}-verdiene. Vi kan nok si at det er god korrelasjon når koeffisienten er stor, men det er vanskelig å sammenlikne korrelasjonen i to materialer med forskjellig spredning i x-verdiene. Tolkningen av den multiple korrelasjonskoeffisienten er altså tvilsom når x-ene er valgte tall.

La oss se på et eksempel med k = 1. Anta at vi har n = 5, samt den teoretiske regresjonslinjen med $\alpha = 2-\beta\bar{x}$, $\beta=1$, $\sigma^2 = 0{,}02$. Da er ligningen

$$y = 2 + (x-\bar{x})$$

og

$$\sigma_y^2 = 0{,}02 + \frac{1}{4}\sum_{i=1}^{5}(x_i-\bar{x})^2, \quad \rho^2 = 1 - \frac{0{,}02}{0.02 + \frac{1}{4}\Sigma(x_i-\bar{x})^2}$$

Vi velger 3 ulike sett av x_i-verdier, nemlig

i) 0, 0,1 0,2, -0,1 og-0,2, som gir $\sum_{i=1}^{5}(x_i-\bar{x})^2 = 0.10$

ii) 0, 1, 2, -1 og -2, som gir $\Sigma(x_i-\bar{x})^2 = 10$

iii) 0, 100, 200, -100 og -200, som gir $\Sigma(x_i-\bar{x})^2 = 100.000$

Vi finner da

i) $\rho^2 = 1 - \frac{0{,}02}{0{,}045} = 0{,}556$

ii) $\rho^2 = 1 - \frac{0{,}02}{2.52} = 0{,}99206$

iii) $\rho^2 = 1 - \frac{0{,}02}{25.000{,}02} = 0{,}9999992$

Vi kan få lignende resultater for R ut fra observasjonene. Vi ser at jo større spredning vi velger i x_i-verdiene, jo større gjør vi ρ^2 eller R^2, tiltross for at regresjonskoeffisienten er den samme hele tiden. Så ille som her vil det sjelden se ut i praksis, men vi skal være oppmerksom på at valget av x-er kan påvirke korrelasjonskoeffisienten. Vi bør helst ikke legge for mye vekt på den i denne typen modeller. I alle fall kan sammenlikninger for ulike observasjonsmaterialer være tvilsomme.

5. Korrelasjon når x-ene er binære (dummy) variable

Hvis y er en forholdstalls-variabel, mens x-ene er binære variable, som bare antar verdiene 0 og 1, og betingelsene i (4.1) forøvrig er oppfylt, kan vi regne regresjoner og teste hypoteser akkurat som foran.

Men hvordan ter korrelasjonskoeffisientene seg? Vi skal se at uttrykkene i avsnitt 2 kan skrives på spesiell form i dette tilfellet.

Vi gir et konstruert eksempel i tabell 7.

Tabell 7. Korrelasjon mellom
y: antall timer brukt til egenarbeid, og
x: utdanningsnivå x = 1 for "høy utdanning, x = 0 for "lav" utdanning

Person nr. i	Observert y_i	x_i
1	0	1
2	1	0
3	2	1
4	3	0
5	4	0
Sum	10	2
Gj.snitt	2	0,4

$s_y^2 = 2$, $s_x^2 = 0,24$

$s_y = 1,41$ $s_x = 0,49$

$n_o = 3$, $\bar{y}_0 = \frac{1+3+4}{3} = 2.667$, $s_0^2 = \frac{14}{9} = 1,556$

$n_1 = 2$, $\bar{y}_1 = \frac{0+2}{2} = 1$, $s_1^2 = 1$

$s_{xy} = -\frac{2}{5} = -0,4$

$r_{xy} = -0,577$

$b = -\frac{0,4}{0,24} = -1,667$ $(=\bar{y}_1-\bar{y}_0)$

Regresjonslinje: $y = 2 - 1,667(x-0,4) = 2,667 - 1,667x$.

Anta at vi har n observasjonspar, (x_i, y_i). For n_1 av disse er $x_i = 1$, for de $n_o = n-n_1$ resterende er $x_i = 0$. Vi har da

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n}, \quad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{n_1}{n})^2 = \frac{n_1}{n} - (\frac{n_1}{n})^2 = \frac{n_1 n_o}{n^2}.$$

Videre kan vi sette, jfr. eksemplet,

$$\bar{y}_1 = \frac{1}{n_1}\sum_{\text{for i der } x_i=1} y_i \quad \text{og} \quad \bar{y}_o = \frac{1}{n_o}\sum_{\text{for i der } x_1=0} y_i$$

Vi har altså

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{n_1\bar{y}_1+n_o\bar{y}_o}{n} \quad (=\frac{2+8}{5}=2 \text{ i eksemplet}).$$

Dessuten kan vi innføre s_1^2 som den empiriske variansen for y_i-ene for de n_1 y_i-verdiene der $x_i = 1$, og s_o^2 tilsvarende for de y_i der $x_i = 0$. Vi har da sammenhengen

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}(n_1 s_1^2 + n_o s_o^2 + \frac{n_1 n_o}{n}(\bar{y}_1-\bar{y}_o)^2).$$

$$= \frac{1}{5}(2\cdot 1+3\cdot\frac{14}{9}+\frac{2\cdot 3}{5}(3-\frac{4}{3})^2) = 2$$

Vi finner, ved litt regning

$$s_{xy} = \frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y}) = \frac{n_1}{n}(\bar{y}_1-\bar{y}) = \frac{n_1 n_o}{n^2}(\bar{y}_1-\bar{y}_o)$$

og

$$r = \frac{s_{xy}}{s_x s_y} = \sqrt{\frac{n_1 n_o}{n}} \cdot \frac{\bar{y}_1-\bar{y}_o}{\sqrt{n_1 s_1^2 + n_o s_o^2 + \frac{n_1 n_o}{n}(\bar{y}_1-\bar{y}_o)^2}} . \qquad (5.1)$$

Se eksemplet.

Dessuten blir $b = \frac{s_{xy}}{s_x^2} = \bar{y}_1-\bar{y}_o$

og regresjonsligningen $y = \bar{y}_o + (\bar{y}_1-\bar{y}_o)x$,

dvs. den går gjennom punktene $(0,\bar{y}_o)$ og $(1,\bar{y}_1)$, jfr. diagram 2. Dette er altså en mer innviklet måte å estimere forventet y-verdi ved hjelp av gjennomsnittet i hver av gruppene !

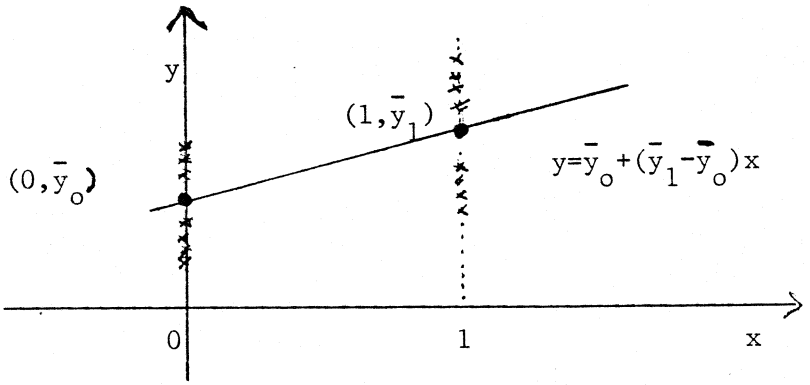


Diagram 2. Regresjon for y m.h.p. x når x er binær. (I eksemplet er $\bar{y}_o > \bar{y}_1$, så regresjonslinjen heller nedover mot høyre).

Vi ser av (5.1) at $r = 0$ når $\bar{y}_0 = \bar{y}_1$

$r = 1$ når $s_0^2 = s_1^2 = 0$ og $\bar{y}_1 > \bar{y}_0$

$r = -1$ når $s_o^2 = s_1^2 = 0$ og $\bar{y}_1 < \bar{y}_o$.

Vi kan altså bare ha $r^2=1$ når alle observerte y_i for $x_i=0$ er like store og dessuten alle observerte y_i for $x_i=1$ er lik hverandre. For å få r^2 nær 1 må både s_o^2 og s_1^2 være små, dvs. at y-verdiene må ligge meget nær hverandre for hver av de to x-verdiene.

For multiple regresjoner med flere binære x-variable vil vi få tilsvarende resultater. For hver kombinasjon av nuller og enere for x-ene må spredningen i y-ene være liten hvis vi skal ha "god korrelasjon", dvs. at y-verdiene må ligge godt konsentrert om de punktene i observasjonsrommet som har x-koordinater 0 og/eller 1.

I regresjoner med både binære og vanlige x-variable, blir tolkingen av R^2 stort sett som i avsnitt 3 og 4 med hensyn tatt til det spesielle ved binær-variablene.

6. Korrelasjon når y er en binær variabel, men ikke x

Situasjonen når vi har en binær y, med verdier 0 og 1, samt en "vanlig" x, er analog med den vi hadde i avsnitt 5 når det gjelder r, men ikke når det gjelder regresjonen. I tabell 8 har vi et eksempel.

Tabell 8. Korrelasjon mellom
y: nivå for egenarbeid, y_i=0 for 1 time eller mindre, y_i=1 for mer enn 1 time
x: antall timer brukt til inntektsgivende arbeid

Person nr. i	Observert y_i	x_i
1	0	3,5
2	0	4,5
3	1	4,0
4	1	3,0
5	1	2,5
Sum	3	17,5
Gj.snitt	0,6	3,5

$s_y^2 = 0,24 = \frac{2 \cdot 3}{25}$, $s_x^2 = 0,5$

$s_y = 0,49$, $s_x = 0,71$

$n_o = 2$, $\bar{x}_o = \frac{3,5+4,5}{2} = 4$, $s_o^2 = 0,25$

$n_1 = 3$, $\bar{x}_1 = \frac{4+3+2,5}{3} = 3,167$, $s_1^2 = 0,389$

$s_{xy} = \frac{-1}{5} = -0,2 = \frac{2 \cdot 3}{25}(3,167-4)$

$r_{xy} = -0,577$

$b = -\frac{0,2}{0,5} = -0,4$

Regresjonslinjen blir: $y = 0,6 - 0,4(x-3,5) = 2 - 0,4x$

Lar vi n_1 være antall observasjoner med y = 1, og $n_o = n-n_1$ antallet med y = 0, så er

$$\bar{y} = \frac{1}{n}\sum_i y_i = \frac{n_1}{n} \text{ og } s_y^2 = \frac{n_1 n_o}{n^2}.$$

Videre er

$$s_{xy} = \frac{n_1}{n}(\bar{x}_1-\bar{x}) = \frac{n_o n_1}{n^2}(\bar{x}_1-\bar{x}_o), \text{ der } \bar{x}_1 = \frac{1}{n_1}\sum x_i \text{ for } y_i=1$$

$$\bar{x}_o = \frac{1}{n_o}\sum x_i \text{ for } y_i=0$$

og

$$r = \frac{s_{xy}}{s_x s_y} = \sqrt{\frac{n_1 n_o}{n}} \; \frac{\bar{x}_1-\bar{x}_o}{\sqrt{n_1 s_1^2+n_o s_o^2 + \frac{n_1 n_o}{n}(\bar{x}_1-\bar{x}_o)^2}}$$

jfr. (5.1). Her er s_1^2 og s_o^2 de empiriske variansene for x-ene i de to gruppene for $y_i=1$ og $y_i=0$. (Vi setter opp dette uttrykket for å lette analysen. Selve beregningen kan vi gjøre etter formlene i avsnitt 2, jfr. eksemplet)

Vi får r = 0 hvis $\bar{x}_1 = \bar{x}_o$

Vi får også r-verdier nær null når s_1^2 og s_2^2 er store i forhold til $(\bar{x}_1-\bar{x}_o)^2$.

Videre er $r^2=1$ hvis $s_1^2 = s_o^2 = 0$,

dvs. at alle x-verdiene for y = 1 må være like store, og alle x-verdiene for y = 0 må være like. Så langt har vi analogi med avsnitt 5. Men regresjonslinjen blir nå

$$y = \bar{y} + b(x-\bar{x}) = \frac{n_1}{n} + \frac{n_1 n_o}{n} \; \frac{\bar{x}_1-\bar{x}_o}{n_1 s_1^2+n_o s_o^2 + \frac{n_1 n_o}{n}(\bar{x}_1-\bar{x}_o)^2}(x-\bar{x}). \qquad (6.1)$$

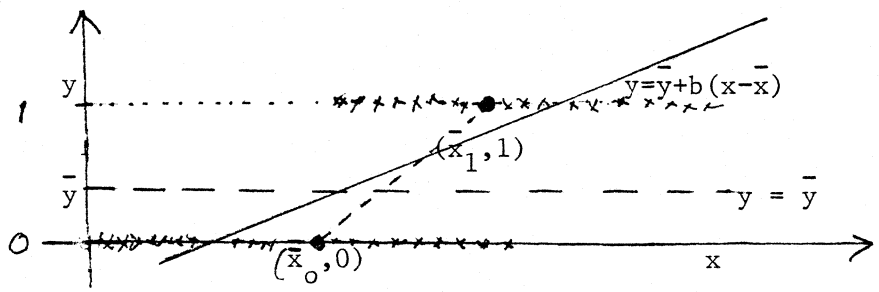


Diagram 3. Regresjon for binær y m.h.p. ikke-binær x .

Her vil regresjonslinjen gå gjennom punktene $(\bar{x}_o,0)$ og $(\bar{x}_1,1)$ bare når $s_1^2 = s_o^2 = 0$, dvs. r=1. Regresjonskoeffisienten b er da $1/(\bar{x}_1-\bar{x}_o)$.

Hvis det er spredning i x-verdiene, blir tallverdien av b mindre enn tallverdien av $1/(\bar{x}-\bar{x}_o)$. I eksemplet er b = -0,4 og $1/(\bar{x}-\bar{x}_o)$ =-1,2.

Vi har

$$0 < b \leqq \frac{1}{\bar{x}_1 - \bar{x}_o} \text{ for } \bar{x}_1 > \bar{x}_o$$

$$0 > b \geqq \frac{1}{\bar{x}_1 - \bar{x}_o} \text{ for } \bar{x}_1 < \bar{x}_o, \text{ jfr. at } -0,4 > -1,2 .$$

Det er betenkelig å postulere en lineær regresjon av en binær variabel y m.h.p. andre variable og spesielt ikke-binære, se f.eks. Henrik Dahl (1978). Vi risikerer bl.a. å få y^*-verdier utenfor intervallet (0,1). For hvilke x vil dette finne sted ved regresjonen (6.1)?
Vi finner at for $\bar{x}_1 > \bar{x}_o$ så er

$$y = \frac{n_1}{n} + b(x-\bar{x}) \leqq 1$$

når

$$x \leqq \bar{x} + \frac{1}{b}(1 - \frac{n_1}{n}) = \bar{x}_1 + \frac{n_1 s_1^2 + n_o s_o^2}{n_1(\bar{x}_1 - \bar{x}_o)}$$

og

$$y = \frac{n_1}{n} + b(x-\bar{x}) \geqq 0$$

når

$$x \geqq \bar{x} - \frac{n_1}{nb} = \bar{x}_o - \frac{n_1 s_1^2 + n_o s_o^2}{n_o(\bar{x}_1 - \bar{x}_o)} .$$

For $\bar{x}_1 < \bar{x}_o$ må vi bytte om de to grensene. I eksemplet får vi at

$0 < y < 1$ for $2,5 < x < 5$.

Jo større s_1^2 og s_o^2 er i forhold til $\bar{x}_1 - \bar{x}_o$, dvs. jo nærmere r er til null, jo "tryggere" er vi på å ha y-verdier i det "tilatte" intervallet [0,1]. Også forholdet mellom n_o og n_1 spiller en rolle her. Men i alle tilfelle er det altså dårlig korrelasjon mellom x og y.
Formel (2.5) gjelder fremdeles her, vi har

$$r^2 = 1 - \frac{s^2}{s_y^2}$$

og

$$s^2 = s_y^2(1-r^2).$$

Når det gjelder testing og hypoteseprøving kan vi imidlertid ikke bruke de vanlige metodene, fordi forutsetningene (4.1b og c) ikke gjelder. Når y bare kan anta verdiene 0 og 1, så er den ikke normalt fordelt. Videre er den teoretiske regresjonen

$$\alpha + \beta x$$

lik sannsynligheten for å få y = 1 når x er gitt. Dette medfører at den teoretiske variansen for y er

$$(\alpha+\beta x)\ (1-\alpha-\beta x)$$

dvs. den varierer med x .

Hvis vi har stor n og β nær null, så behøver ikke de vanlige metodene gi så gale resultater.

Imidlertid er det ikke så sikkert at regresjonen (6.1) og minste kvadraters metode er det lureste valg av metode for variable av denne typen.

Henrik Dahl (1978) har undersøkt dette for en spesiell modell, som bl.a. sikrer at vi ikke kommer utenfor intervallet [0,1] for y. Se også Fridstrøm (1980?).

For multiple regresjoner med y binær kan vi i prinsippet beregne minste kvadraters regresjonslikning og multippel korrelasjonskoeffisient eller determinasjonskoeffisient. Problemet med å holde y innenfor [0,1] blir imidlertid fort påtrengende, og situasjonen lite oversiktlig.

Rent teoretisk må vi kunne tenke oss problemstillinger der sannsynligheten for y = 1 varierer som en lineær funksjon av en rekke x-er, men i praksis er det vanskelig å avgrense en slik klasse av problemer. En vil gjerne velge andre modeller og metoder enn lineær regresjon i slike situasjoner, f.eks. logit analyse, se f.eks. Cox (1970), Fridstrøm (1980).

7. Både y og x-ene binære variable

Når både y og x bare antar verdiene 0 og 1, kan det ha mening å bruke lineær regresjon, selv om vi må gi avkall på metodene fra avsnitt 4. Vi kan da sette opp observasjonsmaterialet i en 2 x 2- tabell:

y \ x	0	1	Sum
0	n_{oo}	n_{o1}	$n_{o.}$
1	n_{10}	n_{11}	$n_{1.}$
Sum	$n_{.0}$	$n_{.1}$	n

f.eks.

y \ x	0	1	Sum
0	526	144	670
1	469	307	776
Sum	995	451	1446

Her er $n_{oo} = 526$, antall observasjoner med x = 0 og y = 0, og $n_{o.} = n_{oo} + n_{o1} = 670$, osv.

Vi finner her, jfr. avsnitt 5,

$$\bar{x} = \frac{n_{.1}}{n} = \frac{451}{1446} = 0{,}31, \quad s_x^2 = \frac{n_{.1}n_{.0}}{n^2} = \frac{451 \cdot 995}{1446 \cdot 1446} = 0{,}215$$

$$\bar{y} = \frac{n_{1.}}{n} = \frac{776}{1446} = 0{,}54, \quad s_y^2 = \frac{n_{1.}n_{o}}{n^2} = \frac{776 \cdot 670}{1446 \cdot 1446} = 0{,}249$$

$$s_{xy} = \frac{n_{11}}{n} - \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n} = \frac{307}{1446} - \frac{451}{1446} \cdot \frac{776}{1446} = 0{,}0449$$

og

$$r = \frac{n\, n_{11} - n_{1.} n_{.1}}{\sqrt{n_{1.} n_{0.} n_{.1} n_{.0}}} = \frac{n_{11} n_{00} - n_{10} n_{01}}{\sqrt{n_{1.} n_{0.} n_{.1} n_{.0}}} = \frac{0{,}0449}{0{,}463 \cdot 0{,}499} = \underline{0{,}194}$$

Vi vil ha $r = 0$ når $\frac{n_{11}}{n} = \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n}$, som kan skrives $\frac{n_{11}}{n_{.1}} = \frac{n_{1.}}{n}$.
Dessuten er da $\frac{n_{10}}{n_{.0}} = n_{1.}$. Det er altså samme relative hyppighet av $y_1=1$ for hvert av observasjonssettene med x=0 og med x=1.
Dette svarer til stokastisk uavhengighet hvis vi erstatter de relative hyppighetene med sannsynligheter, dvs. $p_{11} = p_{1.}\, p_{.1}$.

Videre finner vi at

$r = 1$ når $n_{10} = n_{01} = 0$ (dvs. at y=0 når x=0 og y=1 når x=1)

og

$r = -1$ når $n_{00} = n_{11} = 0$ (y=0 når x = 1 og omvendt).

Vi har altså $r^2=1$ når vi bare har tall $\neq 0$ i en <u>diagonal</u> i tabellen.
Regresjonslikningen blir:

$$y = \frac{n_{10}}{n_{.0}} + \left(\frac{n_{11}}{n_{.1}} - \frac{n_{10}}{n_{.0}}\right) x = 0{,}47 + (0{,}68-0{,}47)x.$$

Regresjonslinjen er her redusert til to punkter: $(0, \frac{n_{10}}{n_{.0}})$ og $(1, \frac{n_{11}}{n_{.1}})$.

Vi kommer ikke utenom intervallet [0,1] for y. Likningen viser at vi estimerer den betingede sannsynligheten for y = 1 gitt x=0 ved den relative hyppigheten $\frac{n_{10}}{n_{.0}}$, og tilsvarende $\frac{n_{11}}{n_{.1}}$ for x = 1. Vi kan dermed bruke test- og estimeringsmetoder bygget på den binomiske fordeling istedenfor metodene i avsnitt 4, som ikke gjelder her (bortsett fra tilnærmet i stort utvalg og når det er liten forskjell på de betingede sannsynlighetene).

For tilfellet med flere binære x-er er situasjonen behandlet i A (1974). Hvis vi tar med like mange koeffisienter i regresjonslikningen som vi har grupper, dvs. kombinasjoner av x-variable i observasjonsmaterialet (en såkalt mettet modell), får vi resultater analoge med eksemplet foran, dvs. det er den relative hyppigheten av y = 1 i hver gruppe som kommer ut. Men hvis vi har færre koeffisienter enn grupper blir resultatene noe anderledes.

Vi skal se på den multiple korrelasjonskoeffisienten for en mettet modell, med s grupper. Vi kan da skrive, se s. 64 op.cit.,

$$R^2 = 1 - \frac{n}{n_{1.}n_{0.}} \sum_{r=1}^{s} \frac{n_{1,r}\, n_{0,r}}{n_{.r}}$$

Her er $n_{1.}$ antall observasjoner med y = 1, $n_{0.} = n - n_{1.}$, s antall koeffisienter i likningen, og $\frac{1}{n_{.r}} n_1$ er den relative hyppigheten av y = 1 i den gruppen på $n_{.r}$ observasjoner som har felles x-vektor nr. r.

Vi ser at

$R^2 = 0$ hvis $\frac{n_{1,r}}{n_{.r}} = \frac{n_{1.}}{n}$ og dermed

$$\frac{n_{0.r}}{n_{.r}} = \frac{n_0}{n} \quad \underline{\text{for alle r}}.$$

dvs. at den relative hyppigheten av y = 1 er den samme i alle gruppene.

Vi får

$R^2 = 1$ hvis enten $n_{1,r} = 0$ eller $n_{0,r} = 0$ <u>for alle r,</u> dvs. at for hver r finnes det bare en verdi av y. Dette svarer til "diagonalbetingelsen" ovenfor i 2 x 2-tabellen.

Hvis vi tenker oss alle tallene i gruppene gitt unntatt ett, nemlig $n_{1,r}$ (og dermed $n_{0,r}$), så vil R^2 avta hvis $n_{1,r}$ øker, så lenge $\frac{n_{1,r}}{n_{.r}} \leq 0{,}5$, men R^2 øker med voksende $n_{1.r}$ hvis $\frac{n_{1r}}{n_{.r}} > 0{,}5$.

Vi har altså "dårlig" korrelasjon for hyppigheter nær 0,5 og "god" korrelasjon for hyppigheter nær null eller en.

I prinsippet vil altså R^2 kunne gi en ide om hvor nær regresjonslikningen føyer seg til observasjonssettene også her.
For <u>stor n,</u>, og store $n_{.r}$, kan den vanlige test for å teste $\rho^2=0$, dvs. uavhengighet mellom y og <u>hele settet</u> av x-er, gi noenlunde god tilnærmelse.

Dette kan være nyttig bl.a. når alle x-ene utelukker hverandre innbyrdes, som når de angir en gruppering av observasjonene i (k+1) ulike grupper av en underliggende variabel, som f.eks. inntekt.

<u>Estimering v.hj.a.gruppegjennomsnittene</u> (de relative hyppighetene i gruppe nr. r) for y-ene, istedenfor enkeltobservasjonene, vil føre til de samme regresjonskoeffisientene, mens restvariansen og dermed korrelasjonen får en annen form. Vi får korrelasjonen mellom de <u>observerte relative</u> hyppighetene og <u>de estimerte</u> ut fra regresjonen, og denne blir ofte stor, ja, den vil gå mot 1 når n vokser, så sant modellen er riktig, se f.eks. Fridstrøm (1980). Dette betyr <u>at estimeringen av sannsynligheten</u> $p_{y/r}$ er god, men <u>sier ikke noe om</u> at $p_{y/r}$ <u>er en lineær funksjon av x-ene</u> i problemet. I en mettet modell får vi R = 1, uansett om vi har greid eller ikke greid å velge x-ene slik at hver x-vektor, x_r, bestemmer verdien av y helt ut.

8. Korrelasjon når x, y, eller begge, er ordningsvariable

Anta først at x er en ordningsvariabel som angir en rangordning av grupper, f.eks. etter utdanningsnivå, etter sosial status e.l. Anta at x er gitt verdiene 1,2,..., m. Vi ønsker å si noe om y som funksjon av x-nivå, og det er fristende å bruke lineær regresjon av y m.h.p. x. Hva får vi ut av dette?

Vi tenker oss observasjonsparene ordnet etter stigende verdier av x, som i tabell 9, og uttrykker en del av de størrelsene vi trenger ved hjelp av gruppegjennomsnittene.

Tabell 9. Korrelasjon når x er en ordningsvariabel

Verdi av x	Antall obs.par med x = j	Gruppesummer av y_{ji}	Gruppesummer av x_i	Gruppesummer av x_i^2	Gruppesummer av $j(\bar{y}_j-\bar{y})n_j$
1	n_1	$n_1\bar{y}_1$	n_1	n_1	$(\bar{y}_1-\bar{y})n_1$
2	n_2	$n_2\bar{y}_2$	$2n_2$	$4n_2$	$(\bar{y}_2-\bar{y})n_2$
. . .					
j	n_j	$n_j\bar{y}_j$	jn_j	j^2n_j	$(\bar{y}_j-\bar{y})n_j$
. . .					
m	n_m	$n_m\bar{y}_m$	mn_m	m^2n_m	$(\bar{y}_m-\bar{y})n_m$
Sum	n	$\sum_{i=1}^{n} y_i = \sum_{j=1}^{m} n_j\bar{y}_j$	$\sum_{j=1}^{m} jn_j$	$\sum_j j^2n_j$	$\sum_j jn_j(\bar{y}_j-\bar{y})$
$\frac{1}{n}$ Sum	1	$\bar{y}$	$\bar{x}$	$s_x^2+\bar{x}^2$	s_{xy}

Dessuten har vi som vanlig $s_y^2 = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ji}-\bar{y})^2$

Ved å omskrive formlene i avsnitt 2 ved hjelp av gruppesummene, finner vi at vi kan skrive

$$r = \frac{s_{xy}}{s_x s_y} = \frac{1}{n}\sum_j jn_j(\bar{y}_j-\bar{y})/s_x s_y \ .$$

Vi ser at hvis alle gruppegjennomsnittene, $\bar{y}_j$, er like, dvs lik $\bar{y}$, så er r = 0, som rimelig kan være.

Regresjonslikningen kan skrives

$$y = \bar{y} + \frac{\sum_j n_j(\bar{y}_j - \bar{y})}{\frac{1}{n_j}\sum_j n_j - \bar{x}^2}(x-\bar{x})$$

Vanskeligheten er å se hva slags mening en kan tillegge gjennomsnittet $\bar{x}$ og avvikene $x_i - \bar{x}$, samt s_{xy} når x er en rangordningsvariabel, dvs. at differensene mellom x-verdiene innbyrdes ikke har noen mening.

Denne vanskeligheten unngår vi ved isteden å innføre (m-1) binære variable for å angi hvilken gruppe observasjonene tilhører.

Hvis vi har et problem der y er rangordnet, mens x er en vanlig variabel, så er det rent algebraisk ikke noe problem å regne regresjon og korrelasjon, men tolkingen av resultatene er ikke så enkel. Vi kan i prinsippet skrive om problemet til (m-1) likninger i (m-1) binære y-er, men her kommer vi i de samme vanskeligheter som i avsnitt 6. Det vil avhenge av problemets art hvilken metode vi bør velge for analysen.

Hvis begge variable er rangordnet, kan vi også regne regresjon hvis vi synes det har noen mening. For det spesialtilfelle at vi har n observasjonssett, der x og y hver for seg har verdiene 1,2,...,n,kan vi skrive r på formen nedenfor. Vår r blir gjerne kalt Spearmans rangkorrelasjonskoeffisient for rangordnede observasjoner.

$$r = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} d_i^2, \text{ der } d_i = x_i - y_i.$$

Vi finner

$r = 1$ når $d_i = 0$ for i=1,2,...,n, dvs. at x og y gir samme rangordning

$r = -1$ når $d_i = n-1, n-3, n-5$, osv., dvs. at x-ene er rangordnet i nøyaktig motsatt retning av y-ene.

$r = 0$ når $\sum_i d_i^2 = \frac{n(n-1)(n+1)}{6}$. Dette kan vi f.eks. få hvis halvparten av observasjonene har samme rangordning i de to variable, mens de øvrige går i nøyaktig motsatt retning.

Hvis vi vil teste signifikans, trenger vi fordelingen av r for $\rho=0$. Denne er tabulert for små n. For større n må en ty til tilnærmet normalitet.

Hvis det er flere observasjoner som har samme rang, gjelder formelen og tabellen ikke, da må det spesielle beregninger til.

Vi skal ikke gå nærmere inn på metoder for rangordnede variable her, det vises til tekster om ikke-parametriske metoder.

9. Observasjoner med tilfeldige feil

I det foregående har vi forutsatt at vi faktisk har observasjoner av nettopp de variable vi er interessert i. Vi kan imidlertid også ha situasjoner der vi ikke kan observere disse variable direkte, men må nøye oss med at de observeres med feil.

Vi kan f.eks. anta at istedenfor

y_i observerer vi $u_i = y_i + d_i$

og istedenfor x_i observerer vi

$v_i = x_i + e_i$.

Vi er interessert i korrelasjonen mellom y_i og x_i, men vi kan bare regne ut korrelasjonen mellom u_i og v_i.

Vi har

$$\bar{u} = \frac{1}{n}\sum_i u_i = \frac{1}{n}\Sigma y_i + \frac{1}{n}\Sigma d_i = \bar{y}+\bar{d} \text{ og } \bar{v} = \bar{x}+\bar{e},$$

$$s_u^2 = \frac{1}{n}\Sigma(u_i-\bar{u})^2 = \frac{1}{n}\Sigma(y_i+d_i-\bar{y}-\bar{d})^2 = \frac{1}{n}\Sigma(y_i-\bar{y})^2 +$$

$$+ \frac{2}{n}\Sigma(y_i-\bar{y})(d_i-\bar{d}) + \frac{1}{n}\Sigma(d_i-\bar{d})^2 = s_y^2+2s_{yd}+s_d^2,$$

$$s_v^2 = s_x^2+2s_{xd}+s_e^2,$$

$$s_{uv} = s_{xy}+s_{xd}+s_{ye}+s_{de}.$$

Hvis vi nå kan anta at feilleddene d_i og e_i er ukorrelert med hverandre, og med x_i og y_i, så har vi i et observasjonsmateriale tilnærmet

$$s_u^2 = s_y^2+s_d^2, \quad s_v^2 = s_x^2+s_e^2 \quad \text{og} \quad s_{uv} = s_{xy}.$$

De tilsvarende teoretiske relasjonene vil gjelde eksakt.

Vi får i dette tilfelle korrelasjonskoeffisienten

$$r_{uv} = \frac{s_{uv}}{s_u s_v} = \frac{s_{xy}}{s_x s_y}\,\frac{s_x s_y}{\sqrt{(s_y^2+s_d^2)(s_x^2+s_e^2)}} = r_{xy}\,\frac{s_x s_y}{\sqrt{(s_y^2+s_d^2)(s_x^2+s_e^2)}}.$$

Vi har $r_{uv} = r_{xy}$ bare hvis $s_d = s_e = 0$, dvs. at det ikke er noen tilfeldige feil (eller at det er nøyaktige samme feil for alle y_i og for alle x_i).

Ellers er korrelasjonen mellom de observerte verdier lavere i tallverdi enn r_{xy}.

En tilsvarende reduksjon har vi for de teoretiske korrelasjonene, og for multippel korrelasjon, når vi kan anta at det er ukorrelerthet mellom feilene og mellom feilene og de ikke-observerte variablene.

Er feilene korrelert med hverandre eller de variable, kan vi få andre resultater. Av formelen for s_{uv} ovenfor ser vi f.eks. at vi kan få $s_{uv} \neq 0$ selv om $s_{xy} = 0$.

10. Sammendrag

Korrelasjonskoeffisienten er et mål for hvor nær observasjonspunktene føyer seg til en rett linje/etlineærutryk for y mhp x-ene. For øvrig har vi at

- R, resp $r^2 = 1$ bare når alle observasjonene tilfredsstiller den lineære regresjons=ligningen, (s. 6, 12, 21, 23, 26, og 29). For binære variable betyr dette at alle observasjonene konsentreres i visse punkter.
- R (r^2) nær 1 bare når det er små avvik mellom observasjonene og likningen
- Ved å øke antall x-variable i regresjonen kan vi i alminnelighet få større verdier på R (s.13) Særlig kan vi få store R-verdier når n er liten og k stor
- Valgte (ikke-stokastiske) x-verdier kan influere på størrelsen av R (s. 19). Sammenlikning av R-ene for ulike observasjonsmaterialer kan derfor være villedende
- Tilfeldige feil i de observerte variable kan redusere R (r^2) eller på annen måte influere på størrelsen av den (s.31)
- Det avhenger av antall observasjoner og antall variable om en R av en gitt størrelse er signifikant forskjellig fra null (tabell 6, s. 18)
- En R som er 0,5 eller mindre (i tallverdi) gir en reduksjon i standardavviket for y som er 13% eller mindre (tabell 4, s.9)
- R = 0 viser at det ikke er lineær samvariasjon mellom y og x-ene i observasjonsmaterialet
- Selv om R = 0 kan det være ikke-lineær sammenheng (Tabell 3, s. 8)
- Selv om en x-variabel er ukorrelert med y, kan den bidra til å øke samvariasjonen i en multippel regresjon (jfr. x_3 og x_4 i tabell 5, se s. 12)

Ut fra ovenstående kan det være lett å forstå ønskemålet om å finne en R nær 1. Da vet vi at det er god lineær samvariasjon i observasjons=materialet. Men det er et meget sterkt krav å forlange at vi skal ha greid å finne frem til og observere x-variable som er slik at praktisk talt all variasjon i y kan uttrykkes ved den lineære regresjonen i x-ene. I alminnelighet bør vi nok være tilfreds med lavere R-verdier. Hvor lave kan vi ikke si generelt, det vil avhenge av det problemet vi analyserer. Enhver R som er signifikant større enn null viser jo at det er en viss lineær samvariasjon mellom y og det valgte sett

av x-er. Dette kan være interessant selv om samvariasjonen ikke er sterk, målt med R.

En ting er sikkert: vi bør i alminnelighet ikke la korrelasjonskoeffisienten alene avgjøre om en regresjon er "god" eller ikke. I tillegg til usikkerheten som ligger i de punktene som er nevnt ovenfor, kan det være andre hensyn. Hvilke kriterier vi skal trekke inn i tillegg til (eller istedenfor) R, vil avhenge av den problemstillingen vi har. Vi bør antagelig se på de enkelte regresjonskoeffisienter, vurdere de ulike x-variable, se på prediksjonsfeil m.m. Kanskje er ikke lineær regresjon en god modell for vårt problem i det hele tatt.

For binære variable må vi bl.a. vurdere om det er x-enes evne til å "forklare" y lineært, eller om det er estimering av sannsynlighetene vi er mest interessert i (jfr. slutten av avsnitt 7, s.27). I mange tilfeller vil log-lineære modeller være et bedre analyseverktøy for binære variabler, Haldorsen (1977), AII, kap. 14.

For ordningsvariable vil i alminnelighet ikke-parametriske metoder være å foretrekke.

L I T T E R A T U R

AI: Amundsen, H.T. (1972) Statistisk metodelære. En elementær innføring. Oslo

AII: Amundsen, H.T. (1978). Statistisk metodelære II. Tolking av data, modeller og metoder. Oslo

A(1974): Amundsen H.T. (1974) Binary variable multiple regressions. Scandinavian Journal of Statistics. Vol 1.

Cox, D.R. (1970). The anaysis of binary data

Fridstrøm, L.E. (1980?). Under arbeid

Haldorsen, T. (1977). Om log-lineær analyse av flerveistabeller. Arbeidsnotat 77//46