

Regression og geometrisk data analyse (1. del)

▷ Ulf Brinkkjær

*Institut for pædagogisk sociologi, Danmarks Pædagogiske Universitetsskole ved
Århus Universitet*

Introduktion

Artiklen¹ søger med konkrete eksempler 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 forklarende metoder. I denne første del-artikel fokuseres på regressionsanalysen.

I Roux og Rouanet (2003) indlejres såvel den simple som den multiple korrespondanceanalyse i en familie af lignende metoder, der tilsammen benævnes *geometrisk data analyse*. Bogens metoder har bl.a. det til fælles, at de alle benytter geometri og grafiske fremstillinger i analysen af de givne data. Når der i det følgende løbende refereres til geometrisk data analyse er det dermed ikke alene korrespondanceanalysen, men også tilgrænsende metoder som er i fokus.

I det konstruerede eksempel, der skal anvendes i det følgende, indgår 3 variable (Rouanet 1978):

- G (gymnasium)
- K (køn)
- E (eksamen)

Hvor mange af en bys gymnasielever består studentereksamen et bestemt år – lyder det spørgsmål, som undersøges? Svaret findes i tabel 1.1 og tabel 1.2.

¹ Målet med denne tekst er at redegøre for pointerne i Henry Rouanets, Frédéric Lebarons, Viviane le Hays, Werner Ackermanns og Brigitte le Rouxs artikel "Régression et analyse Géométrique des données: Refleksions et suggestions", bragt i *Matématiques et Science humaines* årgang 40, no. 160, 2002, side 13-45. Det er imidlertid samtidig ambitionen at folde deres pointer så meget ud, at de bliver tilgængelige uden egentlig statistikkundskab, når blot man er udstyret med en smule tålmod og nysgerrighed. Artiklen er blevet delt i to, hvor første del følger her og næste bringes i et senere nummer af *Praktiske Grunde*.

Tabel 1.1 og 1.2 Studentereksamen fordelt på køn						
Tabel 1.1 Obs				Tabel 1.2 %		
	Bestå	Dump	e+e'	Bestå	Dump	e+e'
♂	24	36	60	40%	60%	100%
♀	36	24	60	60%	40%	100%

Som det fremgår af tabel 11 er der 60 drenge (øverste række) og 60 piger (nederste række). Af de 60 drenge har 24 bestået studentereksamen, mens resten dvs. 36 er dumpet. Af drengene er det dermed $24/60 = 40\%$, der er bestået. Procentfordelingen fremgår af tabel 1.2.

Tabel 1.2 fortæller os dermed, at 40% af drengene og 60% af pigerne har bestået deres studentereksamen. Sammenligner man kønnene, når man dermed til en forskel i succesrate på 20%-point i pigernes favør.

Grundlaget for tabel 1.1 og 1.2 er samtlige byens drenge og piger i 3.g. Vi betragter således byens to gymnasier under ét. Deles tallene op efter ophavsgymnasium, fremkommer følgende fordelinger.

Tabel 2.1 og 2.3 Abs				Tabel 2.2 og 2.4 Pct.			
G = Romeo				G = Romeo			
2.1	Bestå	Dump		2.2	Bestå	Dump	
♂	15	35	50	♂	30%	70%	100%
♀	1	9	10	♀	10%	90%	100%
2.3 G = Julie				2.4 G = Julie			
♂	9	1	10	♂	90%	10%	100%
♀	35	15	50	♀	70%	30%	100%

Bemærk, at tabellernes struktur er fuldstændig sammenfaldende med tabel 1.1 og 1.2. Det samme talmateriale, som anvendtes i tabel 1.1, fordeles nu i to tabeller, tabel 2.1 og tabel 2.3. Summerer man f.eks. øverste venstre celle i tabel 2.1 med den tilsvarende i tabel 2.3 ($15 + 9$), får man netop de 24, som optræder i øverste venstre hjørne i tabel 1.1 - tilsvarende for tabellernes øvrige celler.

Som det fremgår (tabel 2.2), blev eksamen på gymnasium Romeo bestået af 30% af drengene og 10% af pigerne. Tallene på gymnasium Julie var 90% hhv. 70%. På begge gymnasier var der således en forskel på 20% i *drengenes* favør! Men hov, lige før fandt vi ud af, at det var i pigernes favør – hvordan kan det hænge sammen?

Eksemplet illustrerer et paradoksalt tilfælde på et fænomen, der normalt beskrives som *strukturel effekt*. Faktoren køns effekt på beståelse af eksamen er forskellig afhængig af, om man betragter den på et *globalt* niveau (begge gymnasier under ét) eller på et *marginalt/betinget* niveau (de to gymnasier for sig). Det skyldes, at faktorerne Køn og Gymnasium er *korreleret*, jf. senere. En nærmere redegørelse for dette

og tilstødende fænomener kræver et mindre indblik i den anvendte terminologi. Det skal derfor introduceres i det følgende.

Kort (sic!) ekskurs om betingede sandsynligheder

Vi har et spil slidte kort, så slidte, at der kun er Es og dame tilbage og det endda kun i kulørerne ♥, ♠ og ♦. I første forsøg er målet at trække en ♥.

Sandsynligheden, for at forsøget lykkes, udregnes som antal succesfulde udfald, der i figuren har grå baggrund, delt med antal mulige udfald. Idet vi noterer sandsynligheder med P, gælder det dermed, at:

$$P(\text{succes}) = \frac{\text{antal succesfulde udfald}}{\text{antal mulige udfald}}$$



I eksemplet betyder det, at sandsynligheden $P(\heartsuit) = \text{antal } \heartsuit / \text{antal kort} = 2/6 = 1/3$.

I næste forsøg er målet at trække et Es. Sandsynligheden, for at det lykkes, er dermed: $P(\text{Es}) = \text{antal Es'er} / \text{antal kort} = 3/6 = 1/2$. Igen er de succesfulde udfald på grå baggrund.



Når der er $1/3$ sandsynlighed for at trække en hjerter og $1/2$ for at trække et Es, hvad er mon så sandsynligheden for at trække *enten* en hjerter *eller* et Es? Det ville være nærliggende at lægge sandsynlighederne sammen, dvs. $P(\text{første forsøg}) + P(\text{andet forsøg})$. Kigger man engang mere på de to figurer ovenfor, bliver det dog klart, hvorfor det ikke kan være rigtigt: ♥Es har grå baggrund i begge figurer. At lægge de to sandsynligheder sammen ville indebære at ♥Es tælles dobbelt.

En praktisk løsning på det problem, der kan fremstå lidt Molbo-agtig for en umiddelbar betragtning, er at tælle ♥Es med to gange og trække det fra en gang. Det vil sige, at

$$P(\heartsuit \text{ eller Es}) = P(\heartsuit) + P(\text{Es}) - P(\heartsuit \text{ og Es}).^2$$

² Sat på generel formel ser udtrykket således ud:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Sættes tal ind, ser det således ud:

$$P(\heartsuit \text{ eller } Es) = (1/3 + 1/2 - 1/6) = (2/6 + 3/6 - 1/6) = 4/6 = 2/3.$$

Med kun 6 kort er beregningens resultat næppe noget chok! Det lille antal kort gør det nemt at nå resultatet ved at tænke hvert enkelt udfald igennem for sig. Som det fremgår af figur 2, fordeler de 6 mulige udfald sig på fire, som opfylder succeskriteriet om at være enten hjerter eller Es. Sandsynligheden for succes er dermed: $P(\heartsuit \text{ eller } Es) = 4 \text{ ud af } 6 = 4/6 = 2/3$.

♥ eller Es?					
♥ Es	♠ Es	♦ Es	♥ D	♠ D	♦ D
Ja	Ja	Ja	Ja	Nej	Nej

Sorterer man nu alle damerne fra, sidder man tilbage med 3 kort, der alle er Es'er. Sandsynligheden for at trække ♥Es er dermed én ud af tre dvs. $1/3$. Det kan man tænke på, som sandsynligheden for at trække ♥Es *på betingelse af* at antallet af mulige udfald begrænses til de tre Es'er. Denne sandsynlighed benævnes **den betingede sandsynlighed** for at få en ♥ *givet*, at det bliver et Es, og den kan beregnes med følgende sammenhæng:³

$$P(\heartsuit \text{ givet } Es) = \frac{P(\heartsuit \text{ og } Es)}{P(Es)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Der er altså en vigtig forskel mellem $P(\heartsuit \text{ og } Es)$, der er en såkaldt simultan sandsynlighed, og $P(\heartsuit \text{ givet } Es)$, der er en såkaldt betinget sandsynlighed. Det kan fremstå lidt teknisk, men forskellen på simultane sandsynligheder og betingede er vigtig. Forveksling af de to er grundlag for meget vās bl.a. i pressen. Forskellen betyder nemlig bl.a., at:

$$\begin{array}{llll} \text{Mens:} & P(\heartsuit \text{ og } Es) & = & P(Es \text{ og } \heartsuit) \\ \text{Så gælder det at:} & P(\heartsuit \text{ givet } Es) & \neq & P(Es \text{ givet } \heartsuit) \end{array}$$

De forskellige typer af sandsynligheder kan slutteligt illustreres ved at geninddrage tabel 1.1 i forskellige manipuleringer:

³ På formel ser det således ud: $P(Hj\ddot{e}rter \mid Es) = \frac{P(Hj\ddot{e}rter \cap Es)}{P(Es)} = \frac{1}{3}$

Tabel 3.2 Tabel-procenter

	Bestået	Dumpet		Bestået (+)	Dumpet (-)	
♂	20%	30%	50%	♂	$P(\text{♂ og +})$	$P(\text{♂ og -})$ $P(\text{♂})$
♀	30%	20%	50%	♀	$P(\text{♀ og +})$	$P(\text{♀ og -})$ $P(\text{♀})$
	50%	50%	100%		$P(+)$	$P(-)$

I tabel 3.2 har vi beregnet tabelprocenter på tabel 1.1. F.eks. øverste venstre celle er dermed beregnet således: $(24 \text{ beståede drenge}) / (60 \text{ drenge} + 60 \text{ piger i alt}) = 24/120 = 0,20 = 20\%$. Udtrykt i sandsynlighedstermer: Ved udvalg af én af de 120 personer er sandsynligheden for, at det er en dreng og en, der har bestået = 20%. Ved at beregne *tabel*procenter finder vi dermed simultane sandsynligheder.

Ønsker vi i stedet at finde ud af, hvor stor en andel af drengene henholdsvis af pigerne, der har bestået, har vi behov for rækkeprocenter, som dem der ses i tabel 1.2. Beregningen for første celle bliver nu $(24 \text{ beståede drenge}) / (60 \text{ drenge i alt}) = 40\%$. I den beregning er udgangspunktet eller betingelsen, at det er drenge, vi fokuserer på, idet vi netop beregner andelen *af drenge*. Rækkeprocenterne bliver dermed betingede sandsynligheder for bestået *givet*, at det er en dreng.

Tabel 3.3 Række-procenter

	Bestået	Dumpet			Bestået (+)	Dumpet (-)
Drenge	40%	60%	100%	♂	P(+ givet ♂)	P(- givet ♂)
Piger	60%	40%	100%	♀	P(+ givet ♀)	P(- givet ♀)

Vi kan også beregne kolonneprocenterne (tabel 3.4) og dermed få oplyst, hvor stor en andel af de, som er bestået, som er drenge henholdsvis piger. Her er udgangspunktet den gruppe, som er bestået og beregningens resultat dermed sandsynligheden for f.eks. dreng *givet* bestået.

Tabel 3.4 Kolonne-procenter

	Bestået	Dumpet		Bestået (+)	Dumpet (-)
Drenge	41%	59%	♂	$P(\text{♂ givet } +)$	$P(\text{♂ givet } -)$
Piger	59%	41%	♀	$P(\text{♀ givet } +)$	$P(\text{♀ givet } -)$
	100%	100%			

En strukturel effekt

Før ekskursen fremgik det, hvordan en global sammenhæng i pigernes favør blev vendt til en marginal eller betinget sammenhæng i drengenes favør. Det blev afslut-

ningsvis kort berørt, at det var et paradoksalt tilfælde på et fænomen, der normalt beskrives som *strukturel effekt*.

Som led i at folde ovenstående fænomen yderligere ud, bliver der i det følgende talt om *faktorer*, hvis to ting er opfyldt:

- ▷ Variablen er en variabel, der *anvendes til at forklare*, modsat de variable som *skal forklares*. Ofte benævnes førstnævnte uafhængige variable og sidstnævnte afhængige variable, men da ordet uafhængig i denne sammenhæng skal bruges til noget andet, vil der blive talt, om de variable, som *forklares*, henholdsvis dem, der *skal forklares*.
- ▷ Variablen er kodet i et endeligt antal kategorier. Dreng/pige er således to kategorier. Det ville dermed være anderledes, hvis vi målte alder med et uendeligt antal decimaler. I et sådan tilfælde vil der ikke være et *endeligt* antal kategorier.⁴

Forklaringen på den forskel, vi fandt i ovenstående analyser på globalt henholdsvis marginalt/betinget niveau, er, at faktorerne Køn og Gymnasium er *korreleret*. Var det ikke tilfældet, ville der ikke være forskel på sammenhængen på *globalt* og på *marginalt/betinget* niveau. I det konkrete eksempel hænger det sammen med, at Romeo stort set er et drengegymnasium, mens Julie stort set er et pigegymnasium, det vil sige, at faktorerne køn og gymnasium ikke er uafhængige⁵. Mere generelt kan man sige, at afhængigt af forholdet mellem de to faktorer kan de betingede analyser opføre sig på en af følgende fire måder:

Sammenhængen kan:

1. *Ændre retning* (= fortegn) på den globale sammenhæng, således som det var tilfældet ovenfor, hvor en global forskel i pigernes favør ændres til en tilsvarende i drengenes.
2. *Forsvinde* det vil sige en global sammenhæng, men ingen marginale sammenhænge.
3. *Stabiliseres* samme tendens findes i både den globale og den marginale analyse.
4. *Opstå* en sammenhæng, som ikke fandtes globalt, optræder i den marginale analyse.

⁴ Konkret i denne sammenhæng anvendes faktor-begrebet så endvidere som distinktion til det begreb om *faktoriel variabel*, som indgår i forbindelse med geometriske analyser, men herom senere.

⁵ Der tales i det følgende en del om uafhængighed eller modsat om sammenhæng. *Verbalt* kan det udtrykkes på den måde, at når det i tabel 1.1 er 40% af drengene og 60% af pigerne som har bestået, så har det den konsekvens, at hvis man 'får udleveret' et af de unge mennesker, uden at vide om vedkommende har bestået eller ej, så vil blot det at se vedkommendes køn give os bedre mulighed for at 'gætte', om vedkommende har bestået. Sandsynligheden for at vedkommende har bestået *hænger dermed sammen med* vedkommendes køn. (Bemærk det er en statistisk, ikke en moralsk betragtning! ☺). *Matematisk* taler man om uafhængighed, hvis $P(A \text{ og } B) = P(A) * P(B)$. Det er kun opfyldt når værdien i samtlige celler er lige med de forventede værdier. De forventede værdier er jo netop værdier under den hypotese, at der er uafhængighed mellem de to variable.

Interesserer man sig for den strukturelle effekt, som variabelen K(øn) har for beståelse af studentereksamen, starter man med at betragte den *globale effekt*. Det vil sige forskellen på succesrater for drenge og piger ($E(\text{ksamen}) \cdot K(\text{øn})$) som de fremgår af tabel 1.1, og den *betingede effekt*, som forskellene mellem succesraterne givet gymnasium ($E(\text{ksamen}) \cdot K(\text{øn})$ givet G(ymnasium)), som de fremgår af tabellerne 2.1 og 2.3 ovenfor.

I det følgende skal tallene manipuleres rundt for at illustrere de fire situationer, som kan opstå mellem globale og betingede analyser jf. tabel 3. For at fastholde fokus på de strukturelle effekter har samtlige eksempler to fælles træk. For det første er den generelle succesrate alle steder 50%, og for det andet er forskellen mellem drenge og pigers succesrate den samme på begge gymnasier.

En strukturel effekt, der ændrer retning

Vi arbejder med to faktorer (E og K) hver bestående af ($e_{\text{bestået}}$, e_{dumpe}) hhv. (K_{dreng} , K_{pige}). Fokus er på K, hvormed den globale effekt (af K) bliver forskellen: $\text{Beståelse}_{\text{Dreng}} - \text{Beståelse}_{\text{Piger}}$, mens den marginale effekt af K er kønsforskellen i $\text{Beståelse}_{\text{Dreng}} - \text{Beståelse}_{\text{Piger}}$ betinget af gymnasium.

Talgrundlag for dette første eksempel:

	Tabel 4.1: Romeo			Tabel 4.2: Julie			Tabel 4.3: Begge gymnasier		
	Bestå	Dumpe	Sum	Bestå	Dumpe	Sum	Bestå	Dumpe	Sum
Dreng	12	28	40	9	1	10	21	29	50
Pige	1	9	10	28	12	40	29	21	50
Sum	13	37	50	37	13	50	50	50	100

Den globale effekt af K(øn)

Denne effekt fortæller om forskellen på beståelsesandel for drenge og piger dvs. $P_{\text{Bestå givet Dreng}} - P_{\text{Bestå givet Piger}}$, hvilket er det samme som at skrive:

$$\text{Global effekt af køn} = \frac{P_{\text{bestå OG dreng}}}{P_{\text{dreng}}} - \frac{P_{\text{bestå OG pige}}}{P_{\text{pige}}}, \text{ fordi}$$

$$P_{\text{bestå GIVET dreng}} = \frac{P_{\text{bestå OG dreng}}}{P_{\text{dreng}}}$$

Tallene for begge gymnasier under ét, dvs. tabel 4.3 er udgangspunktet for at undersøge den globale effekt. Beregnes tabel-procenter vil tabellen oplyse os de simultane sandsynligheder jf. tidligere, dvs.:

Tabel 4.4: K(øn) * E(ksamen)						
	$e_{\text{bestå}}$	e_{dumpe}	=	$e_{\text{bestå}}$	e_{dumpe}	Sum
k_{dreng}	$ek_{\text{bestå OG dreng}}$	$ek_{\text{dumpe OG dreng}}$		21%	29%	50%
K'_{pige}	$ek_{\text{bestå OG pige}}$	$ek_{\text{dumpe OG pige}}$		29%	21%	50%
Sum				50%	50%	100%

Det betyder, at den globale effekt af køn kan findes med følgende beregning:

$$\begin{aligned}
 P_{\text{Bestå givet Dreng}} &= P_{\text{Bestå og Dreng}} / P_{\text{Dreng}} = (21\%/50\%) = 0,42 \\
 P_{\text{Bestå givet Pige}} &= P_{\text{Bestå og Pige}} / P_{\text{Pige}} = (29\%/50\%) = 0,58 \\
 \text{Global effekt af K(øn)} &= \underline{-0,16}
 \end{aligned}$$

Hvorvidt man beregner den globale effekt af K(øn) som differencen $P_{\text{Bestå givet Dreng}} - P_{\text{Bestå givet Pige}} = -0,16$ eller omvendt som $P_{\text{Bestå givet Pige}} - P_{\text{Bestå givet Dreng}} = 0,16$ beslutter man selv. Forskellen er, at i det første tilfælde har vi gjort pigerne til basis og dermed beskrevet drenge ved deres forskel til pigerne, mens man i den modsatte situation gøre drengene til basis og beskrive pigerne ved deres forskel hertil.

Den marginale effekt af køn

Ligesom ovenfor har vi her fokus på effekten af køn, nu blot den marginale effekt. Grundlaget er tabellerne for Køn*Eksamen – én for hvert gymnasium. Vi kan dermed beregne en kønsforskel for hvert gymnasium. Den marginale effekt af faktoren køn findes herefter som gennemsnittet. Det vil sige gennemsnittet af effekterne af køn givet gymnasium - en beregning parallel til den tidligere, men nu på de to deltabeller tabel 4.1 og 4.2. Vi får således:

Tabel 4.5 Marginal effekt af Køn givet Gymnasium

	Romeo				Julie		
	Bestå	Dumpe	sum		Bestå	Dumpe	sum
Dreng	12	28	40	Dreng	9	1	10
Pige	1	9	10	Pige	28	12	40
sum	13	37	50	sum	37	13	50
Rækkeandele (nu ikke i %)							
Dreng	30%	70%	100%	Dreng	90%	10%	100%
Pige	10%	90%	100%	Pige	70%	30%	100%
Marginal effekt = forskel	20%				20%		

Det vil sige, at den marginale effekt af køn på gymnasiet Romeo er $(30\% - 10\% = 20\% = 0,2)$, mens effekten af køn på gymnasiet Julie er $(90\% - 70\% = 20\% = 0,2)$.

Gymnasiet Julie (tabel 4.6)

$$\begin{aligned}
 P_{\text{Bestå givet Dreng}} &= P_{\text{Bestå og Dreng}} / P_{\text{Dreng}} = (21\%/50\%) = 0,9 \\
 P_{\text{Bestå givet Pige}} &= P_{\text{Bestå og Pige}} / P_{\text{Pige}} = (29\%/50\%) = 0,7 \\
 \text{Marginal effekt for tabel 4.1} &= \underline{0,2}
 \end{aligned}$$

Den marginale effekt af K(øn) givet E(ksamen) findes herfter som et gennemsnit. Den betingede effekt af faktor K(øn) givet faktor G(ymnasium) er, som det fremgår af tabellen ovenfor, den samme for begge gymnasier nemlig $= 0,2$. Det vil sige den betin-

gede effekt af K(øn) givet G(ymnasium) er, at en større andel af drengene består end tilfældet er for pigerne – sammenhængen er dermed *modsat* af den globale sammenhæng, præcist som det var tilfældet i indledningseksemplet.

Simpel og multipel regression

En klassisk måde at studere globale og betingede effekter på er ved hjælp af regressionsanalysen, der består i at beskrive en variabel (kaldet responsvariablen), der normalt noteres som y eller $f(x)$, ved hjælp af *en* (simpel) eller *flere* (multipel) såkaldte forklarende variable (fx. Kreiner 1999, side 214; Andersen m.fl. 2001, side 367 ff). Der kan være grund til at understrege, at denne klassiske sondring mellem en variabel som *respons* variabel og en anden hhv. andre som *forklarende* ikke fortæller noget om variablene eller de fænomener, som søges beskrevet i disse, men alene om den rolle analytikeren tilskriver dem.

I en klassisk terminologi er responsvariablen dermed den variabel, som man sætter sig for at forklare, og forklaringen søges i de(n) forklarende variabel/variable.

Vi ønsker i det følgende at give svar på spørgsmålene:

1. Hvordan påvirker køns-effekten den andel, som består studentereksamen? Her kan svaret findes med en simpel regression.
2. Hvordan påvirker køns-effekten og Gymasie-effekten den andel, som består studentereksamen? Her kræver svaret en multipel regression.

Til dette formål definerer vi tre nye variable:

Variabel udtryk	Variabel navn
$Y_{\text{bestå}}$	'Har vedkommende bestået', der kodes med ja=1 og nej=0
X_{romeo} :	'Har vedkommende gået på Romeo-gymnasiet', der kodes med ja, dvs. Romeo = 1 og Nej, dvs. Julie = 0.
X_{dreng} :	'Er det en Dreng', der kodes med ja, dvs. dreng = 1 og nej, dvs. pige = 0.

Simpel lineær regression

Når spørgsmålet om *Hvordan køns-effekten påvirker den andel som består studentereksamen?* skal besvares med en simpel regressionsanalyse, består det i at beskrive andelen med bestået studentereksamen (y) som en funktion af én variabel. Og hvis man som her arbejder med *lineær* regression, ønskes beståelse af studentereksamen beskrevet som en lineær dvs. retlinet funktion af køn.

Den generelle formel for en simpel lineær regression kan noteres forskelligt⁶, men pointen er den samme. Vi beskriver responsvariablen som resultat af en koefficient til

⁶ Udtryk 1: $\tilde{y}^b = \alpha + \beta_0$ (Roux&Rouanet 2002, side 17). Udtryk 2: $y = \beta_1 x + \beta_0 + e$ dvs.

Udtryk 2.1: $y = \beta_1 x + \beta_0$ (Andersen m.fl. 2001, p. 368), (Aczel 2002, p 438)

Det generelle udtryk for regressionsligningen er $y = \beta_1 x + \beta_0 + e$. Heri optræder et sidste led e , som er udeladt i denne sammenhæng. Det led tematiserer den del af variationen i responsvari-

den forklarende variabel + en konstant. Oversat til dagligdagssprog kan formelen dermed beskrives som:

$$\text{Andel beståede} = (\text{Et tal}) * \text{'er du dreng'} + (\text{Et andet tal}) \quad \text{Sætning 1}$$

Første eksempel på lineær regression

Da variablene x_{dreng} (er du en dreng) og y (har du bestået) fordeler sig, som det sås i tabel 4.3, fås følgende opstilling:

Variabel x_{dreng} og y				Antal	x_{dreng}	Y
21	stk	Dreng + Bestå	=	21	1	1
29	stk	Pige + Bestå		29	0	1
29	stk	Dreng + Dumpe		29	1	0
21	stk	Pige + Dumpe		21	0	0
Gennemsnit, x $(21*1+29*0+29*1+21*0)/100 = 0,21+0,29 =$					0,5	
Gennemsnit, y $(21*1+29*1+29*0+21*0)/100 = 0,21+0,29 =$						0,5

På det grundlag kan 'et tal' (sætning 1) beregnes til:

$$\frac{\text{Summen af afvigelseernes produkt}_x}{\text{Summen af afvigelseernes kvadrat}_{xy}} = \frac{-4}{25} = -0,16 \quad (\text{jf. bilag 1})$$

Tilsvarende kan 'et andet tal' (sætning 1) beregnes til:

$$y_{\text{gennemsnit}} - \text{'et tal'} * x_{\text{gennemsnit}} = 0,58$$

Svaret på det stillede spørgsmål er dermed, at køns effekten påvirker andelen af beståede studentereksamen på følgende måde:

$$\text{Andelen af beståede} = -0,16 * \text{'er det en dreng'} + 0,58^7 \quad \text{Sætning 2}$$

Responsvariablen $y_{\text{bestå}}$ har værdien 0,58 for piger (da x_{dreng} her er nul, hvorfor første led forsvinder), dvs. den samme som andelen af piger som har bestået studentereksamen (29 bestået ud af 50). Det er ikke så mærkeligt, fordi valget af Dreng=1 og Piger=0 er det samme som at vælge pigerne som basis. Både piger og drenges beståelse sættes hermed i relation til pigerne.

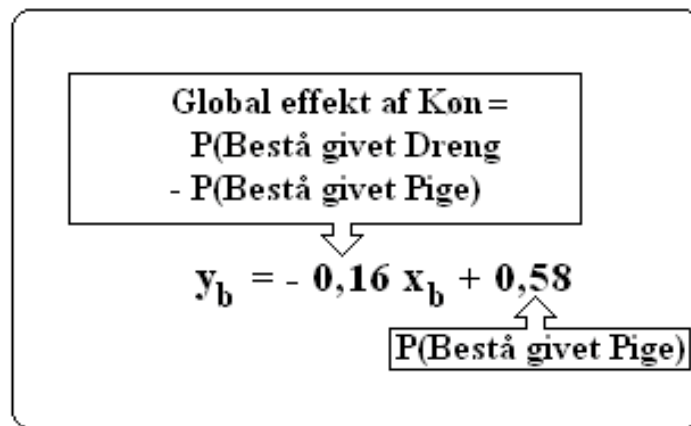
Ligningens 'et andet tal' (koefficienten til x_{dreng}), der her er -0,16, er den globale effekt af faktoren Køn) jf. tidligere var der 42% beståede drenge og 58% beståede piger, forskellen er dermed $42\%-58\% = -16\% = -0.16$. Det er dermed ikke tilfældigt,

ablen, som ikke lader sig beskrive ved en lineær funktion af den forklarende variabel/variable. I bearbejdning af empiriske data vil man normalt betragte e som en slags måleusikkerhed.

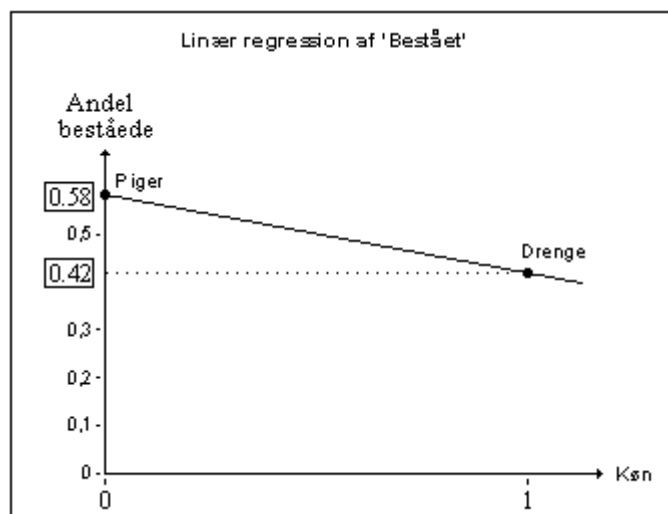
⁷ Dvs. $y_i = -0,16 x_{\text{dreng}} + 0,58$

at værdien er sammenfaldende med den globale effekt, vi fandt med udgangspunkt i tabel 4.4 ovenfor.

Relationen mellem de betingede sandsynligheder og værdierne fra regressionsanalysen kan dermed fremstilles som i figur 1:



Plottes resultaterne af y_{dreng} i en graf fremkommer nedenstående figur.⁸ Figuren viser, hvordan den lineære sammenhæng symboliseret ved den indtegnede linje skærer y-aksen ved værdien 0,58, dvs y_{piger} , fordi vi kodede piger som 0 i variabelen x_{dreng} . y_{dreng} ændrer sig fra 0,58 for $x_{\text{dreng}} = 0$ til 0,42 for $x_{\text{dreng}} = 1$. Linjens hældning er -0,16⁹.



⁸ I Linær regression er den uafhængige variabel normalt af formatet intervalskala. I eksemplet opfatter vi dermed 'køn' som en sådan, hvilket konkret betyder afstanden dreng - pige = 1 skulle give mening, hvad den naturligvis ikke gør, men det leger vi.

⁹ Hældningen er dermed $(Y_{\text{dreng}} - Y_{\text{pige}})/(X_{\text{dreng}} - X_{\text{pige}}) = (0,42 - 0,58)/(1-0) = -0,16$.

Hvordan påvirker køns-effekten og Gymasie-effekten den andel som består studentereksamen? Multipel regression.

Det multiple ligger i at svaret (responsvariabelen) ikke længere beskrives/forklares som funktion af én variabel, men som funktion af flere. I det her gennemgåede eksempel anvendes konkret to variable x_{Romeo} og x_{dreng} . I dagligdagssprog kan formlerne for den multiple lineære regression beskrives som¹⁰:

Andel beståede = et tal * 'er du dreng' + et andet tal * 'Romeo' + et tredje tal

Sætning 3

I denne runde arbejdes derfor med en responsvariabel $y_{\text{Romeo,Dreng}}$ to forklarende variable x_{Romeo} (har du gået på Romeo?) og x_{dreng} (er du dreng?) samt tre tal (koefficienter): 'et tal', 'et andet tal' og 'et tredje tal'¹¹. Udregningerne, der er lidt mere komplicerede end i den simple lineære regression, findes i bilag 2. Resultatet bliver at 'et tal' = 0,7, 'et andet tal' = -0,60, mens 'et tredje tal' = 0,20. Det betyder, altså, at:

Andelen af beståede (y) = $-0,60 * \text{'gået på Romeo'} + 0,20 * \text{'Dreng'} + 0,70$

'Et tal'

Koefficienten u_{Romeo} på -0,60 er den *lokale* Gymnasie-effekt *givet* kønnet. Det skal forstås som gennemsnittet af de marginale beståelsessandsynligheder:

Bestå givet køn

Dreng				Pige			
	Bestå	Dumpe	Sum		Bestå	Dumpe	Sum
Romeo	12	28	40	Romeo	1	9	10
Julie	9	1	10	Julie	28	12	40
Sum	21	29	50	sum	29	21	50
Rækkeandele (ikke i %)							
Romeo	0,3	0,7	1	Romeo	0,1	0,9	1
Julie	0,9	0,1	1	Julie	0,7	0,3	1
Forskel	-0,6			Forskel	-0,6		

$\text{Bestå}_{\text{Romeo givet dreng}} - \text{Bestå}_{\text{Julie givet dreng}}$ (0,3 – 0,9), som det er markeret i indramningen *til venstre*. Og $\text{Bestå}_{\text{Romeo givet pige}} - \text{Bestå}_{\text{Julie givet pige}}$ (0,1 – 0,7), som det er markeret i indramningen *til højre*. Da begge her bliver -0,60 er også gennemsnittet naturligvis -0,60.

¹⁰ Sat på formel kan det f.eks. se således ud $\tilde{y}^{a+b} = u_a x_a + u_b x_b + u_0$. Ruanet m.fl (2002).

¹¹ Det vil sige: u_a , u_b og u_0 .

'Et andet tal'

Den partielle regressions koefficient u_{dreng} på 0,20 er den *lokale* køns-effekt givet Gymnasium, hvilket igen vil sige gennemsnit af de marginale beståelsessandsynligheder:

Romeo				Julie			
	Bestå	Dumpe	Sum		Bestå	Dumpe	Sum
Dreng	12	28	40	Dreng	9	1	10
Pige	1	9	10	Pige	28	12	40
Sum	13	37	50	sum	37	13	50
Rækkeandele (nu ikke i %)							
Dreng	0,3	0,7	1	Dreng	0,9	0,1	1
Pige	0,1	0,9	1	Pige	0,7	0,3	1
Forskel	0,2			Forskel	0,2		

$\text{Bestå}_{\text{Dreng givet Romeo}} - \text{Bestå}_{\text{Pige givet Romeo}} (0,3 - 0,1)$, jf. indramningen *til venstre*, og $\text{Bestå}_{\text{dreng givet Julie}} - \text{Bestå}_{\text{pige givet Julie}} (0,9 - 0,7)$ jf. indramningen *til højre*. Da begge her bliver 0,20 er også gennemsnittet = 0,20.

'Et tredje tal'

Koefficienten på 0,70 er Responsvariablen for $\text{Bestå}_{\text{Julie og Pige}}$. Når netop den indgår handler det igen om, at ved at kode både x_{Julie} og x_{pige} som nul er pige og Julie gjort til basis for beregningen.

Effekt af k(øn)		Effekt af G(ymnasium)	
Global effekt af k(øn) =	-0,16	Global effekt af G(ymnasium) =	-0,48
Marginal effekt af k(øn) =	0,20	Marginal effekt af G(ymnasium) =	-0,60

Global Køns-effekt er dermed -0,16, mens den marginale er 0,20 og den globale Gymnasie-effekt er -0,48, mens den marginale er -0,60. Køns-effekten ændrer sig dermed fra en global effekt på -0,16 til en betinget effekt på 0,20. Det vil sige sammenhængen vendes og ændrer styrke.

Andet eksempel - sammenhænge, der forsvinder

Ovenstående eksempel viste en situation, hvor sammenhængen ændrede retning og styrke, når man bevægede sig fra den globale analyse til den marginale/betingede. I det følgende ændres fordelingen af data en smule, fordi der nu skal vises et eksempel, hvor den globale sammenhæng forsvinder i de marginale/betingede analyser.

De anvendte værdier fordeler sig som det fremgår af figur 5.1 – 5.3.

Tabel 5.1 Romeo

	Bestået	Dumpet	
♂	8	32	40
♀	2	8	10
	10	40	50

Tabel 5.2 Julie

	Bestået	Dumpet	
♂	8	2	10
♀	32	8	40
	40	10	50

Tabel 5.3 Begge gymnasier

	Bestået	Dumpet	
♂	16	34	50
♀	34	16	50
	50	50	100

Succesrate for Dreng	(tab 5.3, række 1):	= 16/50	= 0,32
Succesrate for Piger	(tab 5.3, række 2):	= 34/50	= 0,68
Generel succesrate	(tab 5.3, række 3):	= 50/100	= 0,50

Global effekt af B (køn) er dermed: $\text{Bestå}_{\text{dreng}} - \text{Bestå}_{\text{pige}} = 0,32 - 0,68 = -0,36$. At pigernes succesrate er større end drengenes medfører, at den globale effekt bliver negativ og signalerer, at globalt set klarede pigerne sig bedst.

Kigges herefter på de marginale relationer fås:

Succesrate for Dreng _{givet Romeo}	(tab 5.1, række 1):	= 8/40	= 0,20
Succesrate for Piger _{givet Romeo}	(tab 5.1, række 2):	= 2/10	= 0,20
Marginalt effekt af Køn _{givet Romeo}	= forskellen		= 0,0

Tilsvarende fås for Julie:

Succesrate for Dreng _{givet Julie}	(tab 5.2, række 1):	= 8/10	= 0,80
Succesrate for Piger _{givet Julie}	(tab 5.2, række 2):	= 32/40	= 0,80
Marginalt effekt af Køn _{givet Julie}	= forskellen		= 0,0

Konklusion:

Der er en global effekt, der peger på, at pigerne klarer sig bedre end drengene, men der er ikke nogen marginal/betinget effekt idet $P(\text{Bestå givet Romeo}=1) = P(\text{Bestå givet Romeo}=0)$.

Den simple regression

Vi koder fortsat kønnet: $x_{\text{dreng}} = 1$ (dreng), 0 (pige). Vi får dermed:

	x_{dreng}	$y_{\text{bestå}}$	antal	Ak_x	Ak_y	AP_{xy}
Dreng og Bestå	1	1	16	4	4	4
Pige og Bestå	0	1	34	8,5	8,5	-8,5
Dreng og Dumpet	1	0	34	8,5	8,5	-8,5
Pige og Dumpet	0	0	16	4	4	4
Middelværdi	0,5	0,5	100			
Sum af afvigelses kvadrat (SAK)				25	25	
Sum af afvigelses produkt (SAK)						-9

Beregninger:

Middelværdi (af x) = $(16*1 + 34*0 + 34*1 + 16*0)/100 = 0,5$. Ikke så overraskende da der er lige mange drenge og piger, hvorfor *halvdelen* kodes med 1 og resten med 0. Beregning af y 's middelværdi følger sammen system.

Sum af Afvigelsenes **Kvadrat** (forkortet SAK_x) beregnes ved for hver linje at beregne kvadratet på afgivelsen $(x - x_{\text{gennemsnit}})$ dvs. $(x - \bar{x})^2$ og herefter summere de fire linjer. Resultat = 25. Tilsvarende for SAK_y

Sum af Afvigelsenes **Produkt** (forkortet SAP_{xy}) beregnes ved for hver linje at beregne produktet på afgivelserne $(x - x_{\text{gennemsnit}}) * (y - y_{\text{gennemsnit}})$ dvs. $(x - \bar{x}) * (y - \bar{y})$ og herefter summere de fire linjer. Resultat = -9.

Ligesom tidligere (side 14) kan vi herefter udregne

'et tal' som: $SAP_{xy}/SAK_x = -9/25 = -0,36$ og

'et andet tal' som: $y_{\text{gn.snit}} - \text{'et tal'} * x_{\text{gn.snit}} = 0,5 - (0,5 * (-0,36)) = 0,68$

Den simple regression kan dermed sammenfattes som: $y_{\text{bestå}} = -0,36x_{\text{dreng}} + 0,68$

Den multiple regression giver jf. udregninger i bilag:

$$y_{\text{bestå}} = -0,60x_{\text{Romeo}} + 0x_{\text{dreng}} + 0,80, r^2 = 0,360$$

Tredje eksempel – sammenhænge, der stabiliseres

Tabel 16.1 Romeo

	Bestået	Dumpet	
♂	10	40	50
♀	40	10	50
	50	50	100

Tabel 16.1 Romeo

	Bestået	Dumpet	
♂	8	32	40
♀	8	2	10
	16	50	

Tabel 16.2 Julie

	Bestået	Dumpet	
♂	2	8	10
♀	32	8	40
	34	50	

Global effekt af B (køn):

$$10/50 - 40/50 = -0,30/50 = -0,60. \text{ Dvs. pigerne lykkes bedre end drengene}$$

Betinget/marginal effekt af B (Køn):

$$8/40 - 8/10 = -0,60 \text{ og } 2/8 - 32/10 = -0,60. \text{ Dvs. også i de marginale effekter lykkes pigerne bedre end drengene}$$

Den simple regression:

$$\tilde{y}^b = -0,60x_b + 0,80$$

Den multiple regression,:

$$\tilde{y}^{a+b} = 0x_a - 0,60x_b + 0,80, r^2 = 0,360$$

Fjerde eksempel - sammenhæng, der opstår

Tabel 16.1 Romeo

	Bestået	Dumpet	
♂	25	25	50
♀	25	25	50
	50	50	100

Tabel 16.1 Romeo

	Bestået	Dumpet	
♂	16	24	40
♀	1	9	10
	17	50	

Tabel 16.2 Julie

	Bestået	Dumpet	
♂	9	1	10
♀	24	16	40
	33	50	

Global effekt af B (køn): $25/50 - 25/50 = 0$. Dvs. pigerne og drenge lykkes lige godt

Betinget/marginal effekt af B (Køn): $16/40 - 1/10 = 0,30$ og $9/10 - 24/40 = 0,30$.
Dvs. drengene lykkes bedre end pigerne

Den simple regression: $\tilde{y}^b = 0x_b + 0,50$

Den multiple regression,: $\tilde{y}^{a+b} = -0,50x_a + 0,30x_b + 0,60, r^2 = 0,160$

Forudsigelser

Rouanet m.fl fortæller herom, at det drivende for regressionsanalysen er forudsigelsen. Med udgangspunkt i det man allerede ved, det vil sige de(n) forklarende variable, søger man at udtale sig om det man gerne vil vide – nemlig responsvariabelen y . Det kan dreje sig om at forudsige, hvorvidt en elev lykkes på grundlag af oplysninger om køn, gymnasium, socioøkonomisk miljø m.v. Når det kommer til den slags forudsigelser, er listen over forklarende variable ikke for begrænset (eller bør ikke være det) og forbindelser mellem disse, f.eks. i form af kvasi-kolinearitet, er kun en mindre vanskelighed. Det der derimod er vigtigt, er responsvariabelen y , "i alle dens nuancer", defineret ved den samlede mængde af forklarende variable og naturligvis ved koefficienten R (eller R^2), der som kvalitetsmål på den matematiske analyse skal være så stor som mulig. Netop i det perspektiv vil man sjældent være hjulpet af den populære idé vedrørende at betragte en marginal/betinget effekt som en "rigtig effekt – alt andet lige" (Rouanet m.fl side 19).

Regressionsanalysen sonderer, som det er fremgået, mellem en række forklarende variable og en responsvariabel. Det bør stå lysende klart, at den statistiske procedure som sådan er neutral vis-à-vis denne *forklaringsmodel*. Udpegning af hvilke variable der optræder hvor i modellen er ikke en beregningsøvelse, men en konsekvens af den opstillede model. Hermed fremgår det også, at opdeling af variable i forklarende variable henholdsvis responsvariabel nødvendigvis rummer en – erkendt eller ikke-erkendt – forklaringsmodel. Man kan, som det nævnes et sted: "Foretage en regression af en metalstangs længde som funktion af temperaturen og dermed inkludere længdeudvidelsen i forklaringsmodellen, men man kan lige så vel foretage en regression af temperaturen som funktion af metalstangens længde". (Ibid, side 20)

► Referencer

- Erling B. Andersen m.fl. (2001): *Teoretisk statistik for økonomer*, Akademisk Forlag, København.
- Svend Kreiner (1999): *Statistisk problemløsning, præmisser, teknik og analyse*, Djøfs forlag, København.
- Amir D. Aczel (2002): *Complete business statistics*, McGraw-Hill Higher Education, New York.
- Henry Rouanets, Frédéric Lebarons, Viviane le Hays, Werner Ackermanns og Brigitte le Roux: "Régression et analyse Géométrique des données: Refleksions et suggestions", *Matématiques et Science humaines* årgang 40, no. 160, 2002, side 13-45.

► Bilag

Bilag 1

Den globale effekt af G(ymnasium)

x_a	y	Freq	Sum af Afvigelsesnes Kvadrat (SAK_{x_a}) $\sum (x_a - x_{a_{snit}})^2$		Sum af Afvigelsesnes Produkt (SAP_{xay}) $\sum (x_a - x_{a_{snit}}) * (y - y_{snit})$	
Romeo	Bestå	13	0,25	3,25	0,25	3,25
Romeo	Dumpe	37	0,25	9,25	-0,25	-9,25
Julie	Bestå	37	0,25	9,25	-0,25	-9,25
Julie	Dumpe	13	0,25	3,25	0,25	3,25
Middel		25				
Sum				25		-12
Global effekt =				(-12)/ 25=		-0,48

Den globale effekt af K(øn)

X_b			SAKxb $\sum (x_a - x_{a_{snit}})^2$		SAPxby $\sum (x_a - x_{a_{snit}}) * (y - y_{snit})$	
Dreng	Bestå	21	0,25	5,25	0,25	5,25
Dreng	Dumpe	29	0,25	7,25	-0,25	-7,25
Pige	Bestå	29	0,25	7,25	-0,25	-7,25
Pige	Dumpe	21	0,25	5,25	0,25	5,25
Middel		25				
Sum				25		-4
Global effekt =				(-4)/25 =		-0,16

Koefficienterne u_0 eller u_1 (=et tal) hhv. u_0 eller u_1 (=et andet tal) beregnes således:

$$u_1 = u_1 = \frac{\sum (x - x_{\text{middel}}) * (y - y_{\text{middel}})}{\sum (x - x_{\text{middel}})^2} \quad u_0 = u_0 = \frac{y_{\text{middel}} - \beta * x_{\text{middel}}}{1} = y_{\text{middel}} - u_1 * x_{\text{middel}}$$

$$u_1 = u_1 = -4/25 \text{ jf. ovenfor} \quad u_0 = u_0 = 0,5 - (-0,16) * 0,5 = 0,58$$

Bilag 2

X_{Romeo} er her noteret x_a , mens x_{dreng} her noteres x_b .

Koefficienterne u_0 , u_1 og u_2 gives ved følgende tre ligninger (med de tre koefficienter som ubekendte):

Ligning 1: $\sum y = nu_0 + u_a \sum x_a + u_b \sum x_b$

Ligning 2: $\sum x_a * y = u_0 \sum x_a + u_a \sum x_a^2 + u_b \sum x_a * x_b$

Ligning 3: $\sum x_b * y = u_0 \sum x_b + u_a \sum x_a * x_b + u_b \sum x_b^2$

hvor n = (antal kategorier i x_a * antal kategorier i x_b), da både x_a og x_b er kodet i to kategorier (0 eller 1) bliver $n = 2 * 2 = 4$

Sættes tal ind for de eksisterende værdier af x_{Romeo} x_{Dreng} fås:

Tabel 16 Indsætning af tal for x_a og x_b i den fundne løsning				
		X_{Romeo}	X_{dreng}	$(y) = -0,60 * x_a + 0,20 * x_b + 0,70$
Julie	Pige	0	0	$(y) = -0,60 * 0 + 0,20 * 0 + 0,70 = 0,70$
Romeo	Pige	1	0	$(y) = -0,60 * 1 + 0,20 * 0 + 0,70 = 0,10$
Julie	Dreng	0	1	$(y) = -0,60 * 0 + 0,20 * 1 + 0,70 = 0,90$
Romeo	Dreng	1	1	$(y) = -0,60 * 1 + 0,20 * 1 + 0,70 = 0,30$

Samtlige disse værdier fremgår af tabel 9.2, der alle er beregninger på tallene fra tabel 4.1 og 4.2:

		x_a (Romeo?)	x_b (dreng?)	Y_{abs} (Bestået)	Y_{tot} (i alt)	$y =$ $y_{\text{abs}}/y_{\text{tot}}$	$x_a * x_b$	x_a^2	x_b^2	$y * x_a$	$y * x_b$
Julie	Pige	0	0	28	40	0,7	0	0	0	0,0	0,0
Romeo	Pige	1	0	1	10	0,1	0	1	0	0,1	0,0
Julie	Dreng	0	1	9	10	0,9	0	0	1	0,0	0,9
Romeo	Dreng	1	1	12	40	0,3	1	1	1	0,3	0,3
SUM		2	2			2	1	2	2	0,4	1,2

Efter indsætning af tal, ser ligningssystemet ud som de første 3 linjer i tabel 11. Da løsningsstrategien bl.a. består i at trække ligninger fra hinanden, for at få elimineret nogle af elementerne, er ligning 2 og ligning 3 multipliceret med 2, hvilket blot består i at fordoble samtlige led på begge sider af lighedstegnene:

Tabel 11 - ligningssystemet	
Ligning 1:	$2 = 4u_0 + 2u_a + 2u_b$
Ligning 2:	$0,4 = 2u_0 + 2u_a + 1u_b$
Ligning 3:	$1,2 = 2u_0 + 1u_a + 2u_b$
2*Ligning 2:	$0,8 = 4u_0 + 4u_a + 2u_b$
2*Ligning 3:	$2,4 = 4u_0 + 2u_a + 4u_b$

Tabel 14 Kalkulation af u_a , u_b og u_0	
Lign 1 – 2Lign2	$1,2 = 0u_0 - 2u_a + 0u_b \Leftrightarrow 1,2 = -2u_a \Leftrightarrow$ $u_a = 1,2/(-2) = -0,60$ $-0,4 = 0u_0 + 0u_a - 2u_b \Leftrightarrow -0,4 = -2u_b$ $u_b = (-0,4)/(-2) = 0,20$
Lign 1 – 2Lign3	$2 = 4u_0 + 2u_a + 2u_b \Leftrightarrow 4u_0 = 2 - 2u_a - 2u_b \Leftrightarrow u_0 = (2 - 2u_a - 2u_b)/4 \Leftrightarrow$ $u_0 = [2 - 2*(-0,6) - 2*0,2]/4 \Leftrightarrow u_0 = (2 + 1,2 - 0,4)/4 \Leftrightarrow u_0 = (2,8)/4$ $u_0 = 0,70$
Ligning 1	

Tabel 15 De fundne værdier af u_0 , u_a og u_b
$u_0 = 0,70$
$u_a = -0,60$
$u_b = 0,20$