SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing VIII

- Kurvetilpasning
  Hamilton kap 5 s145-273

Kurvetilpassing

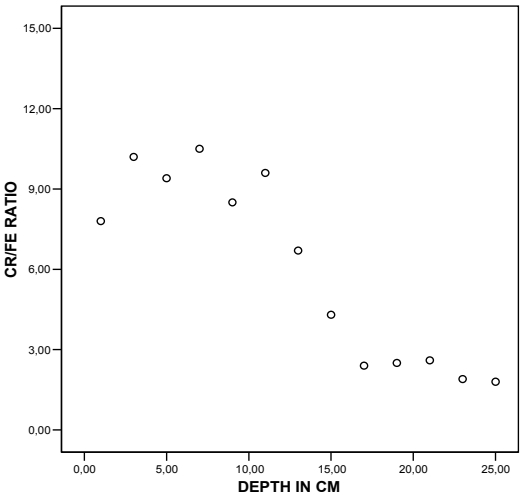
- Ein rett spesifisert modell krev at funksjonen som bind x-variablane og y variabelen saman er i samsvar med røyndomen: er sambandet lineært?
- Data kan granskast gjennom bandregresjon eller glatting
- Teori om kausalsambandet kan spesifisere eit ikkje-lineært samband
- For fenomen som ikkje kan representerast med ei rett linje skal vi sjå på nokre alternativ
  - Kurvelineær regresjon
  - Ikkje-lineær regresjon

Bandregresjon

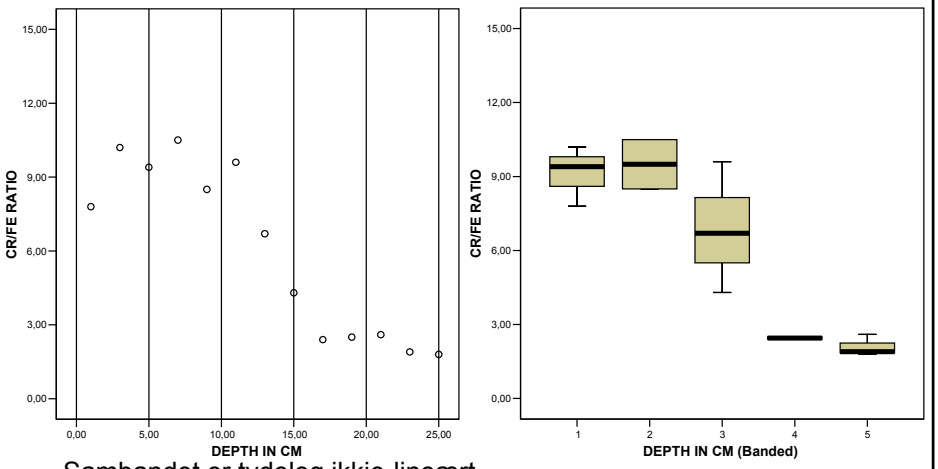
- Kan nyttast til å utforske korleis sambanda mellom variablane ser ut.
- Dersom vi kan sjå at ein underliggjande trend i data er ikkje-lineær, må vi gjennom transformasjonar eller bruk av kurver finne ei form på funksjonen som betre representerer samanhengen

Ureining i ulike djup av sediment på sjøbotnen utanfor NH

- Ureining målt ved raten krom/jern i ulike djup av ulike sedimentprøver
- Er sambandet lineært?



Medianane i 5 band: raten krom/jern i sediment utanfor kysten i NH



Sambandet er tydeleg ikkje-lineært

Transformerte variablar

- Brukar vi transformerte variablar vert regresjonen kurvelineær. Transformasjonen gjer den opphavelege kurvesamanhengen til ein lineær samanheng
- Dette er den viktigaste grunnen til å transformere.
- Samtidig kan transformering ordne opp i ulike typar statistiske problem (utliggjarar, heteroskedastisitet, ikkje-normale feil)
- Framgangsmåte:
  - Vel høveleg transformasjon og lag nye transformerte var.
  - Gjennomfør ein standard analyse med dei transformerte var
  - For tolking bør ein vanlegvis transformere attende til opphaveleg måleskala

Den lineære modellen

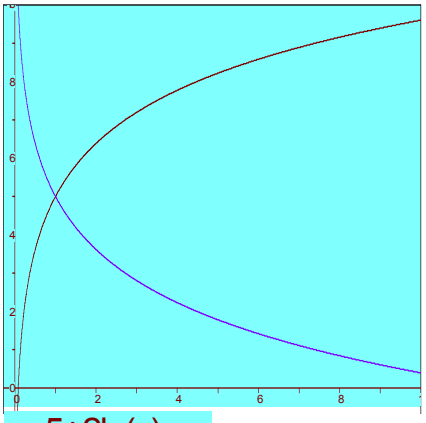
$$y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji} + \varepsilon_i$$

- I den lineære modellen kan vi transformere x-ane og y-ane utan at det har noko å seie for eigenskapane til OLS-estimata i seg sjølv.
- Så lenge modellen er lineær i parametrane er OLS ein lovleg metode

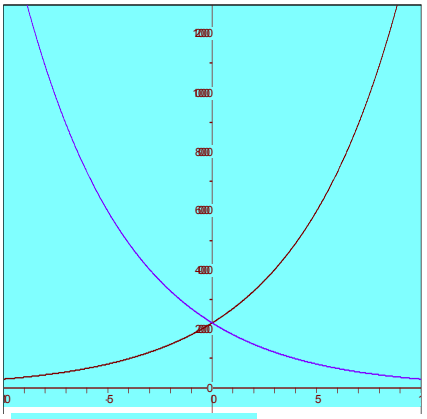
Kurvelineære Modellar

- Dette er i praksis regresjon med transformerte variablar
- Vi skal sjå på korleis ulike transformasjonar gir ulik form på sambanda
  - Semilogaritmiske kurver
  - Log-Log kurver
  - Log-resiproke kurver
  - Polynom (2 og 3 orden)

Semilogaritmiske kurver Fig 5.2 i Hamilton

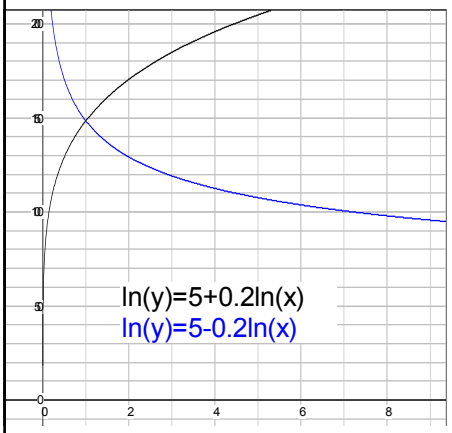


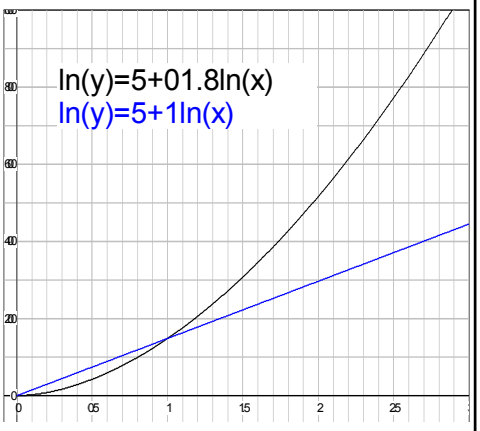
y=5+2ln(x)
y=5-2ln(x)



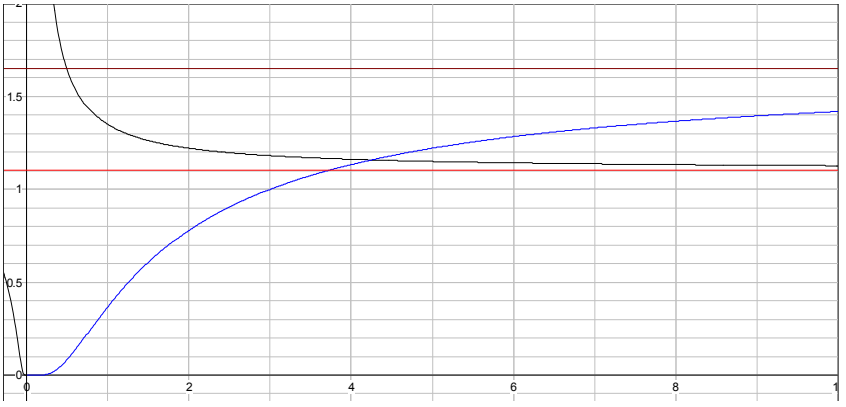
ln(y)=10+0.2x
ln(y)=10-0.2x

Log-log kurver Fig 5.3 i Hamilton





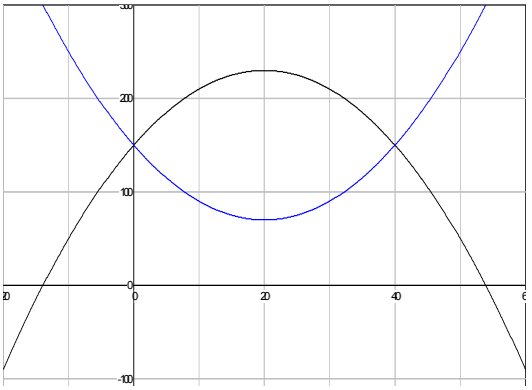
Log-resiproke kurver Fig 5.4 Hamilton



ln(y)=0.1+0.2/x
ln(y)=0.5-1.5/x
Horizontal line through (0, 1.105)
Horizontal line through (0, 1.649)

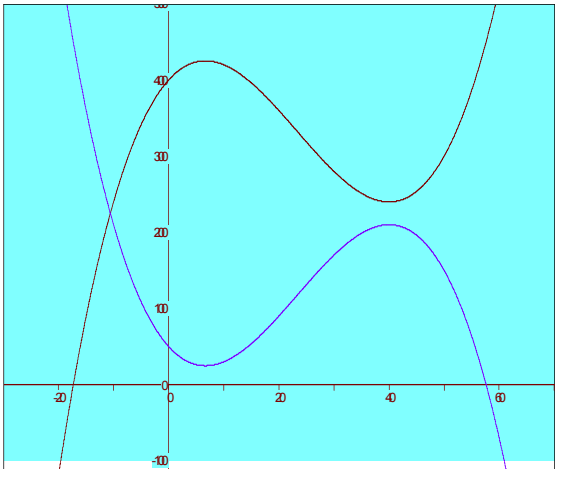
Dei horisontale linjene gir verdien av y når x veks mot uendeleg: asymptoten for y

Andregrads polynom Fig 5.5 Hamilton



y=150+8x-0.2x^2
y=150-8x+0.2x^2

Tredjegrads polynom Fig 5.6 Hamilton



y=400+8x-0.7x^2+0.01x^3
y=50-8x+0.7x^2-0.01x^3

Val av transformasjon

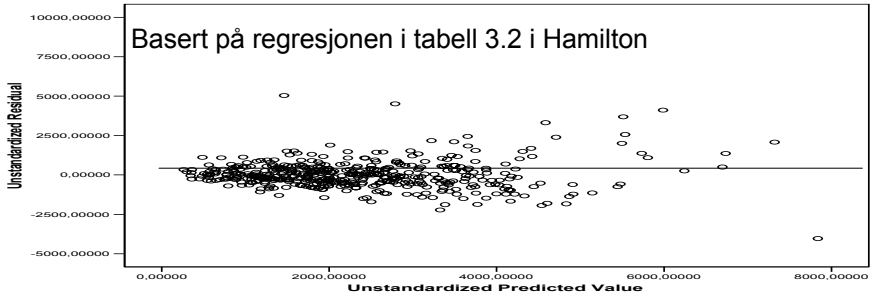
- Spreiingsplott eller teori kan gi råd
- Elles er transformasjon til symmetri det beste utgangspunktet
- Regresjonen rapportert i tabell 3.2 i Hamilton viste seg problematisk
- Regresjon med transformerte variablar kan redusere problema

Val av transformasjon i tab 3.2 Hamilton

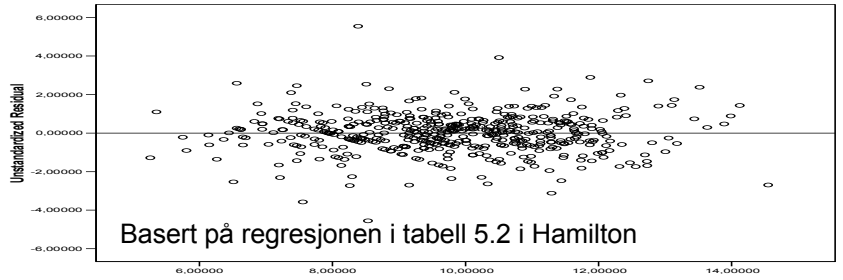
Y	$Y^*=Y^{0.3}$ gir tilnærma symmetri	Vassforbruk 1981
X_1	$X_1^*= X_1^{0.3}$ gir tilnærma symmetri	Inntekt
X_2	$X_2^*= X_2^{0.3}$ gir tilnærma symmetri	Vassforbruk 1980
X_3	Transformasjonar kan gjere lite	Utdanning
X_4	Transformasjon påverkar ikkje dummyvar	Pensjonist
X_5	$X_5^*= \ln(X_5)$ gir tilnærma symmetri	# menneskje i 1981
X_6	$X_6 = X_5 – X_0$ (= # menneskje i 1980)	Endring i # menneskje
X_7	$X_7^* = \ln (X_5/X_0)$	Relativ endring i # m

Regresjon med transformerte variable
Tab 5.2 Hamilton

Dependent Variable: wtr81_3	B	Std. Error	t	Sig.
(Constant)	1,856	,385	4,822	,000
inc_3	,516	,130	3,976	,000
wtr80_3	,626	,029	21,508	,000
Education in Years	-,036	,016	-2,257	,024
head of house retired?	,101	,119	,852	,395
logpeop	,715	,110	6,469	,000
clogpeop	,916	,263	3,485	,001



Residual mot predikert Y



Andre verknader av transformasjonane

- To case med stor innverknad på koeffesienten for inntekt (store DFBTAS) har no ikkje slik innverknad (fig 4.11 og 5.9)
- Eit case med stor innverknad på koeffesienten for vassforbruk i 1980 har no ikkje så stor innverknad (fig 4.12 og 5.10)
- Transformasjonar som gjer fordelingar symmetriske vil ofte løyse mange problem – men ikkje alltid!

Tolking

- Estimatet av modellen ser no slik ut

$$y_i^{0.3} = 1.856 + 0.516x_{1i}^{0.3} + 0.626x_{2i}^{0.3} - 0.036x_{3i}$$

$$+0.101x_{4i} + 0.715\ln(x_{5i}) + 0.916\ln(\frac{x_{5i}}{x_{0i}})$$

- Tolking av koeffesientane er ikkje lenger så enkelt (t.d.: måleeiningane på parametrane er endra)
- Den enklaste måten å tolke på er å nytte betinga effekt plott

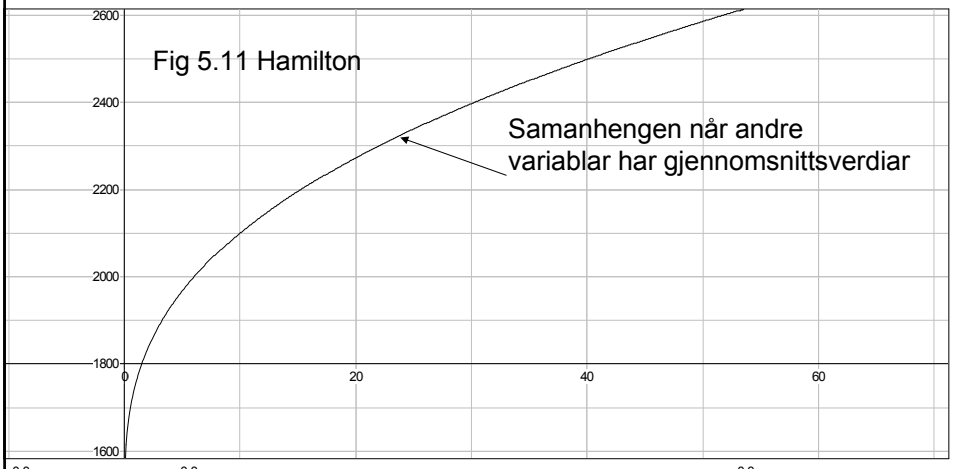
For betinga effekt plott er det nyttig med ein tabell over minimum, maximum og gjennomsnitt

	N	Minimum	Maximum	Mean
Summer 1981 Water Use	496	100	10100	2298,39
Summer 1980 Water Use	496	200	12700	2732,06
Income in Thousands	496	2	100	23,08
Education in Years	496	6	20	14,00
head of house retired?	496	0	1	,29
# of People Resident, 1981	496	1	10	3,07
Increase in # of People	496	-3	3	-,04
# people living in 1980	496	1	10	3,11

Variabelverdiar som maksimerer eller minimerer predikert verdi

	Koeff (+/-)	Minimere y	Maximere y	Mean
Summer 1981 Water Use		100	10100	2298,39
Summer 1980 Water Use	0.626	200	12700	2732,06
Income in Thousands	0.516	2	100	23,08
Education in Years	-0.036	20	6	14,00
head of house retired?	0.101	0	1	,29
# of People Resident, 1981	0.715	1	10	3,07
Increase in # of People	0.916	-3	3	-,04
# people living in 1980		10	1	3,11

Vassforbruk etter inntekt kontrollert for andre variablar



$y^{0.3}=1.856+0.626(2732)^{0.3}-0.036(14)+0.101(0.294)+0.715\ln(3.07)+0.916(\ln(3.07)-\ln(3.11))+0.516(x)^{0.3}$

Kva er interessant å plotte?

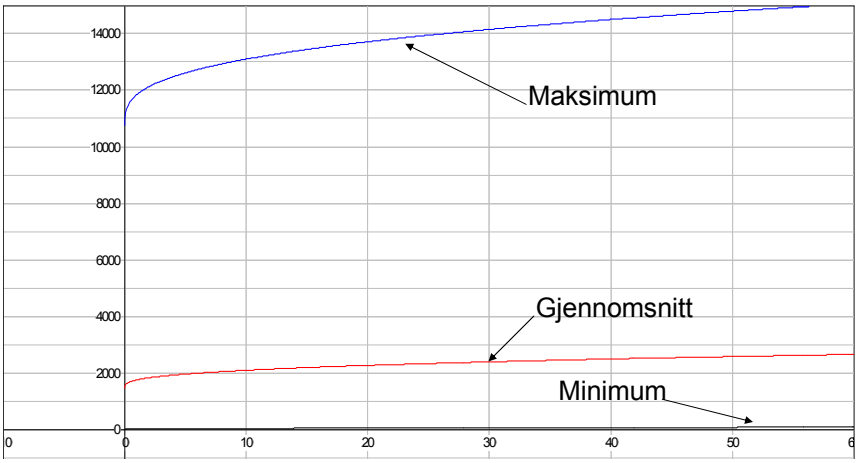
- Samanhengen inntekt vassforbruk kontrollert for ulike kombinasjonar av andre variabelverdiar
  1. Dei som minimerer vassforbruket
  2. Dei som maksimerer vassforbruket
  3. Gjennomsnittverdiane

1 $y^{0.3}=(1.856+0.626(200)^{0.3}-0.036(20)+0.101(0)+0.715\ln(1)+0.916(\ln(1)-\ln(10))+0.516(x)^{0.3})$

2 $y^{0.3}=(1.856+0.626(12700)^{0.3}-0.036(6)+0.101(1)+0.715\ln(10)+0.916(\ln(10)-\ln(1))+0.516(x)^{0.3})$

3 $y^{0.3}=(1.856+0.626(2732)^{0.3}-0.036(14)+0.101(0.29)+0.715\ln(3.07)+0.916(\ln(3.07)-\ln(3.11))+0.516(x)^{0.3})$

Samanlikning av tre typar brukssituasjon



Samanhengen mellom inntekt og vassforbruk Fig 5.12 Hamilton

Konstanten si rolle i plottet

- Den einaste skilnaden mellom dei tre kurvene er konstanten
  - I maksimumskurva er (konst) = 14.046
  - I minimumskurva er (konst) = 4.204
  - I gjennomsnittskurva er (konst) = 8.507

$$y_i^{0.3} = \left(konst\right) + 0.516 x_{1i}^{0.3}$$

- Effekten av inntekt varierer med verdien av (konst)
- Når vi transformerer avhengig variabel vert **alle** samanhengar til interaksjonseffektar

Samanlikning av effektar

- I somme samanhengar kan ein nytte den standardiserte regresjonskoeffesienten til å samanlikne effektar, men den er sensitiv for skeive estimat av standardfeilen
- Ein meir generell metode er å samanlikne betinga effekt plott der skaleringa på y-aksen er halden konstant

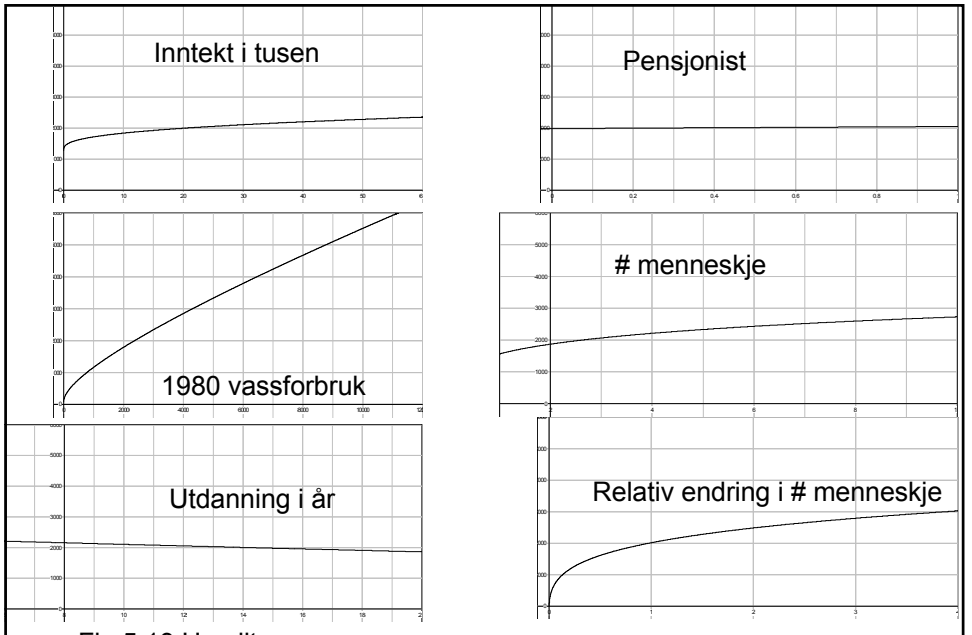
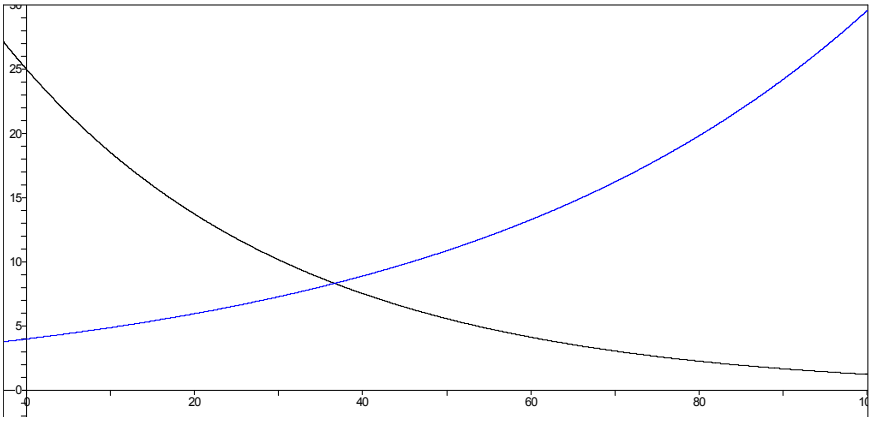


Fig 5.13 Hamilton

Ikkje-lineære modellar

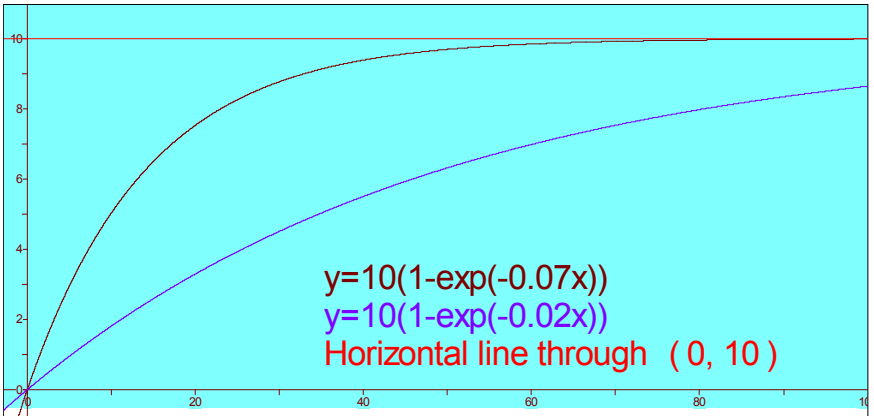
- Dersom vi ikkje har modellar som er lineære i parametrane vil ein trenge andre teknikkar for å estimere parametrane
- Det kan vere to typar argument for slike modellar
  - Teori om den kausale mekanismen kan diktere ein slik modell
  - Inspeksjon av data kan peike på ein bestemt type modell
- Vi skal sjå på
  - Eksponentielle modellar
  - Logistiske modellar
  - Gompertz modellar

Eksponentiell vekst og fall Fig 5.14 Hamilton

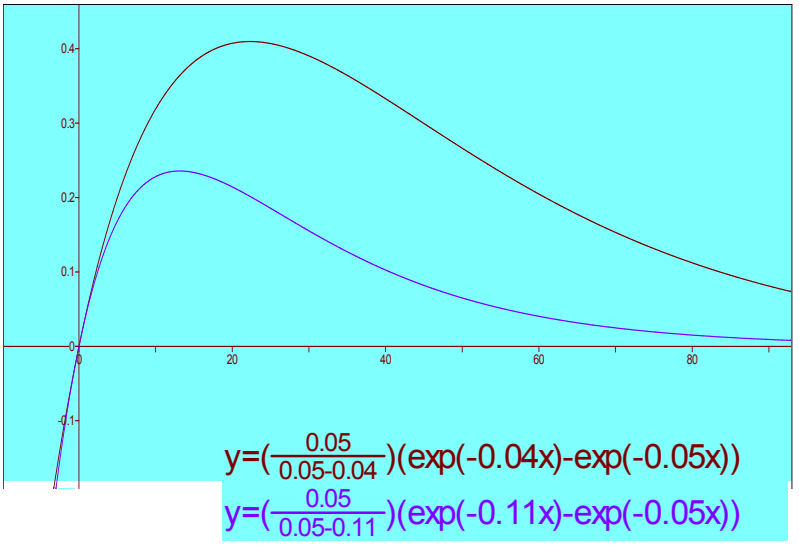


y=25exp(-0.03x)
y=4exp(0.02x)

Negative eksponential kurver Fig 5.15 Hamilton



To-ledds eksponentialkurver Fig 5.16 Hamilton



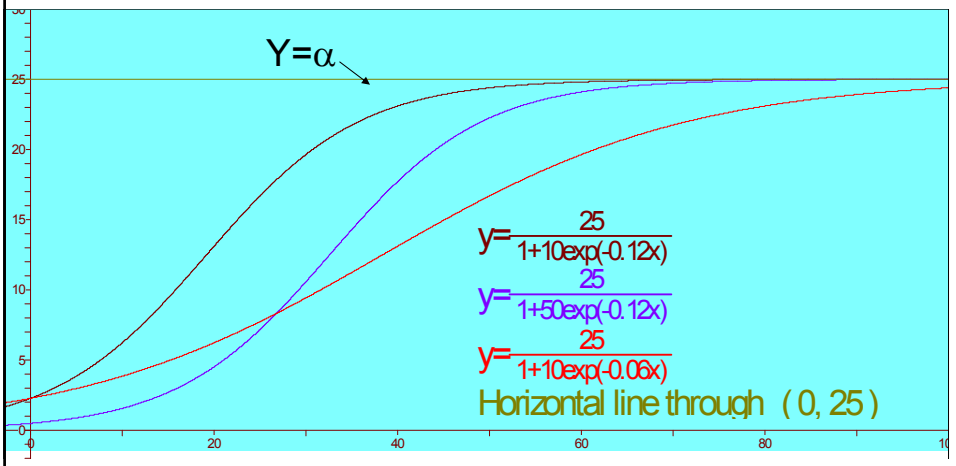
Logistiske modellar

- Den logistiske funksjonen skriv ein
- Når x veks mot uendeleg vil y nærme seg α
- Når x minkar mot minus uendeleg vil y nærme seg 0

$$y = \frac{\alpha}{1 + \gamma \exp(-\beta x)} + \varepsilon$$

- Logistiske modellar passar til mange fenomen
  - Vekst i biologiske populasjonar
  - Spreiing av rykte
  - Spreiing av sjukdom

Logistiske kurver Fig 5.17 Hamilton



- γ fastset kvar veksten startar, β kor rask veksten er

Logistisk sannsynsmodell

- Dersom ein set $\alpha=\gamma=1$ vil y variere mellom 0 og 1 når x varierer mellom minus uendeleg og pluss uendeleg.
- Logistiske kurver kan da nyttast til å modellere sannsyn

$$y_i = \frac{1}{1+\exp\left(-\beta x_i\right)} + \varepsilon_\iota$$

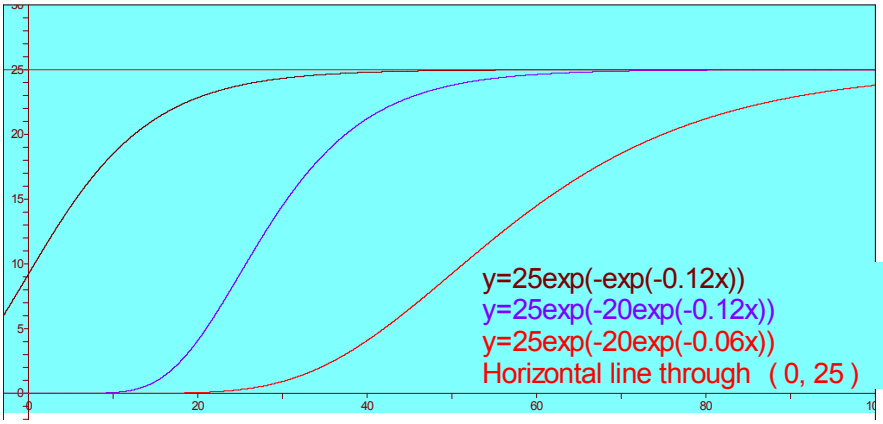
Gompertzkurver

- Gompertzkurver er sigmoidkurver slik som den logistiske, men tilvekst og vekstreduksjon skjer i ulik takt, så dei er ikkje symmetriske

$$y = \alpha e^{-\gamma e^{-\beta x}} + \varepsilon$$

- Parametrane α, γ og β har den same tolkinga som i den logistiske modellen

Gompertzkurver Fig 5.18 Hamilton



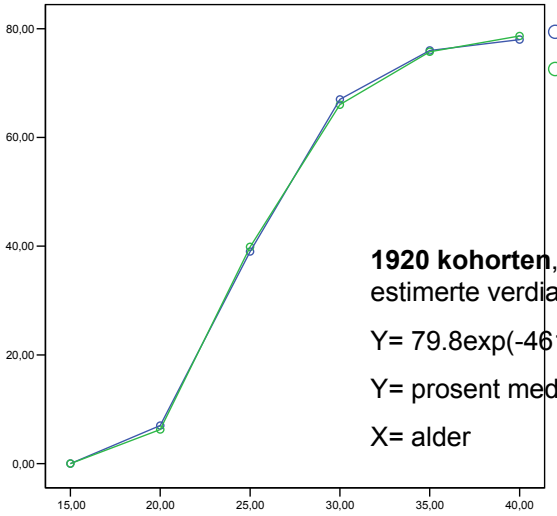
Estimering av ikkje-lineære modellar

- Kriteriet på tilpassing er framleis minimum RSS
- Ein kan sjeldan finne analytiske uttrykk for parametrane. Ein må gjette på ein startverdi og gå igjennom fleire iterasjonar for å finne kva parameterverdi som gir den minste RSS verdien
- Gode startverdiar er som regel nødvendig og alt frå teori til inspeksjon av data vert brukt for å finne dei

Prosent kvinner med minst 1 barn etter kvinnas alder og fødselsår (England og Wales)

	1920	1930	1940	1945	1950	1955	1960	1965
15	0	0	0	0	0	0	0	0
20	7	9	13	17	19	18	13	11
25	39	48	59	60	53	45	39	-
30	67	75	82	82	75	68	-	-
35	76	83	87	88	83	-	-	-
40	78	86	89	90	-	-	-	-
45	-	86	89	-	-	-	-	-

Estimering av Gompertz-modellar for kohortar (1)



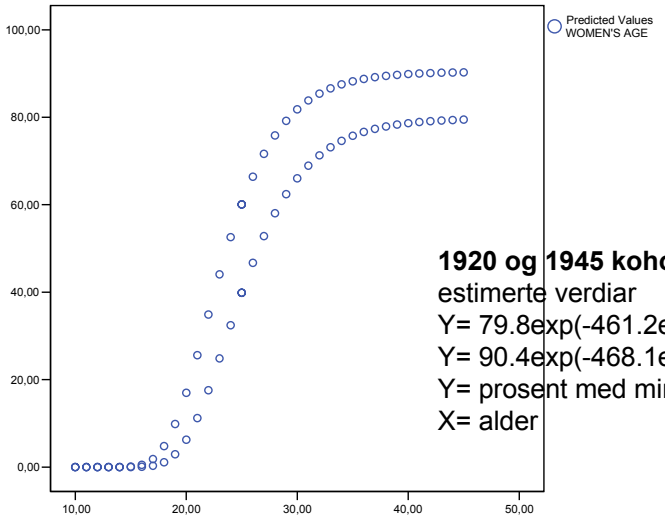
1920 kohorten, observerte og estimerte verdiar:

Y= 79.8exp(-461.2exp(-0.26x))

Y= prosent med minst 1 barn

X= alder

Estimering av Gompertz-modellar for kohortar (2)



1920 og 1945 kohortane, estimerte verdiar
Y= 79.8exp(-461.2exp(-0.26x))
Y= 90.4exp(-468.1exp(-0.28x))
Y= prosent med minst 1 barn
X= alder

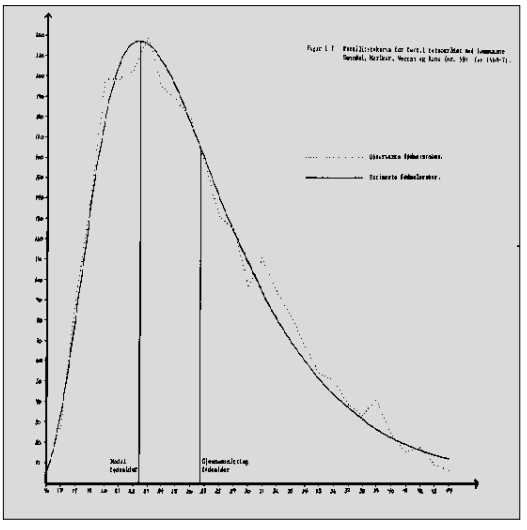
Modelltilpassing

- For å evaluere ein teoretisk utleda modell
- For prediksjon av y innan eller ut over variasjonsområdet for x
- Substansiell eller komparativ vurdering av parmeterverdiar
  - Her kan vi nytte modellen på kohortar som enno ikkje er ferdig med fødslane sine (prediksjon ut over observerte x verdiar)
  - Vi kan nytte modellen til å samanlikne parameterverdiane til ulike kohortar

Parametertolking Tab 5.6 Hamilton

Kohort	α = øvre grense	γ = ?	β = vekstfart
1920	79.8	461.2	0.26
1930	86.5	538.0	0.27
1940	89.1	942.0	0.31
1945	90.4	468.1	0.28
1950	87.5	144.9	0.23
1955	88.9	60.3	0.18

Fødselsrater i Sunndal, Meråker, Verran og Rana 1968-71



- Modellert med Hadwiger funksjonen
- Ref.: Erling Berge 1981 «The Social Ecology of Human Fertility in Norway 1970», Ph.D dissertation, Boston University

Konklusjonar i kapittel 5 (1)

- Dataanalyse startar ofte med lineære modellar. Dei er enklast.
- Teori eller utforskande dataanalyse (bandregresjon, glatting) kan seie oss om kurvelineære eller ikkje-lineære modellar trengst
- Transformasjon av variable gir kurvelineær regresjon. Dette kan motverke fleire problem
  - Kurvelinearitet i samanhengane
  - Case med stor påverknad
  - Ikkje-normale feil
  - Heteroskedastisitet

Konklusjonar i kapittel 5 (2)

- Ikkje-lineær regresjon nyttar iterative prosedyrar for å finne parameterestimat.
- Prosedyrane treng initialverdiar og er ofte sensitive for initialverdiane.
- Tolking av parametrar kan vere vanskeleg. Grafar som viser sambanda for ulike parameterverdiar vil hjelpe mye