SOS3003
Applied data analysis for social science
Lecture note 08-2009

Erling Berge
Department of sociology and political science
NTNU

Literature

- Fitting Curves
  Hamilton Ch 5 p145-273

Fitting Curves

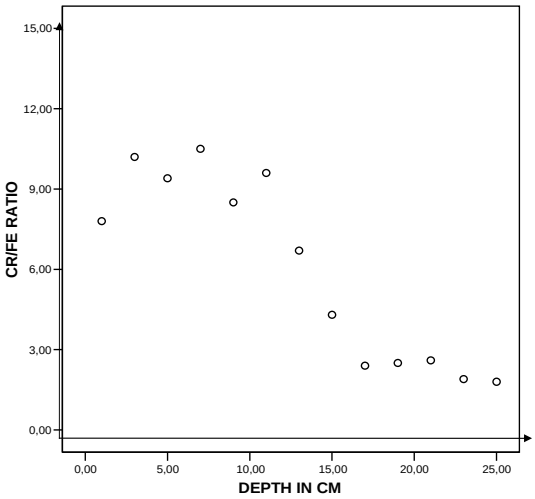
- A correctly specified model require that the function linking x-variables and y-variable is true to what really exist: is the relationship linear?
- Data can be inspected by means of band regression or smoothing
- The theory of causal impact can specify a non-linear relationship
- For phenomena that cannot be represented by a line we shall present some alternatives
  - Curvilinear regression
  - Non-linear regression

Band regression

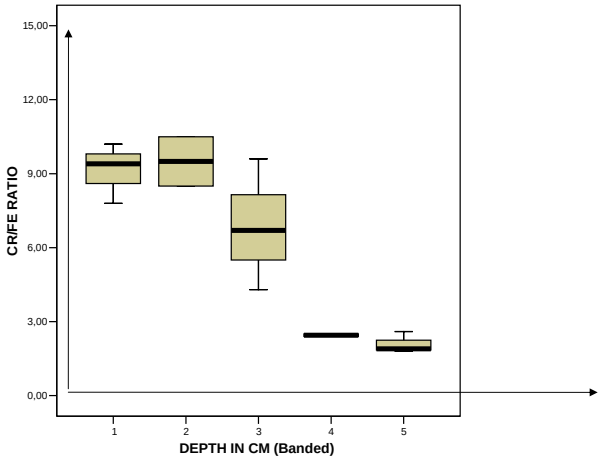
- Can be used to explore how the relationship among the variables actually appears
- If we can see a non-linear underlying trend of the data we must through transformations or use of curves find a form for the function better representing the relationship

Pollution at different depths in sediments outside the coast of NH

- Pollution measured by the ratio chromium/iron at different depths of various sediment samples
- Is the relationship linear?



Medians of 5 bands: rate of chromium/iron in sediments outside the coast of NH



The relationship is obviously non-linear

Transformed variables

- Using transformed variables makes a regression curvilinear. The transformation makes the original curve relationship into a linear relationship
- This is the most important reason for a transformation
- At the same time transformations may rectify several other types of statistical problems (outliers, heteroscedsticity, non-normal errors)
- Procedure:
  - Choose an appropriate transformation and make new trasnformed variables
  - Do a standard regression analysis with the transformed variables
  - To interpret the results one usually will have to transform back to the original measurement scale

The linear model

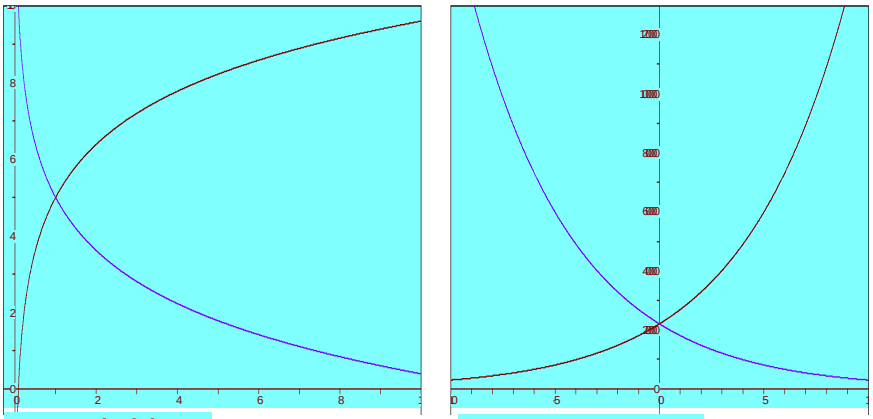
$$y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji} + \varepsilon_i$$

- In the linear model we can transform both x- and y- variables without any consequences for the properties of OLS estimates of the parameters
- As long as the model is linear in the parameters OLS is a valid method

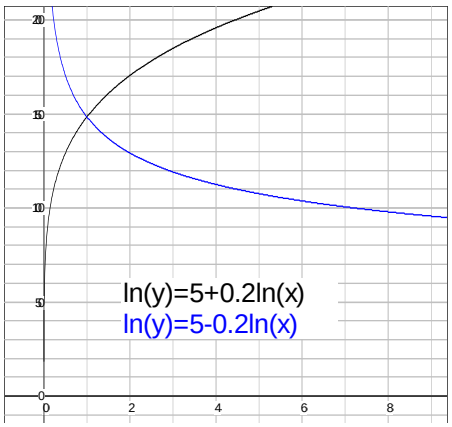
Curvilinear Models

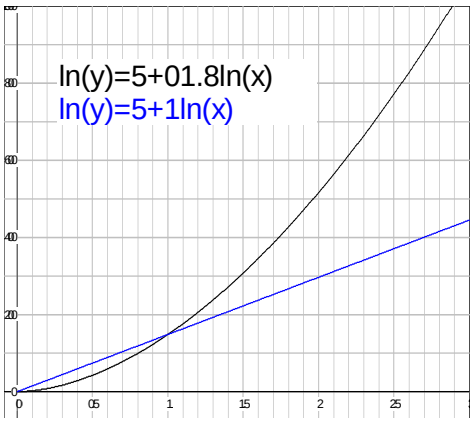
- Practically speaking this is regression with transformed variables
- We shall take a look at how different transformations provide different forms for the variable relations
  - Semi-logarithmic curves
  - Log-Log curves
  - Log-reciprocal curves
  - Polynomials (2 and 3 order)

Semilog curves Fig 5.2 in Hamilton

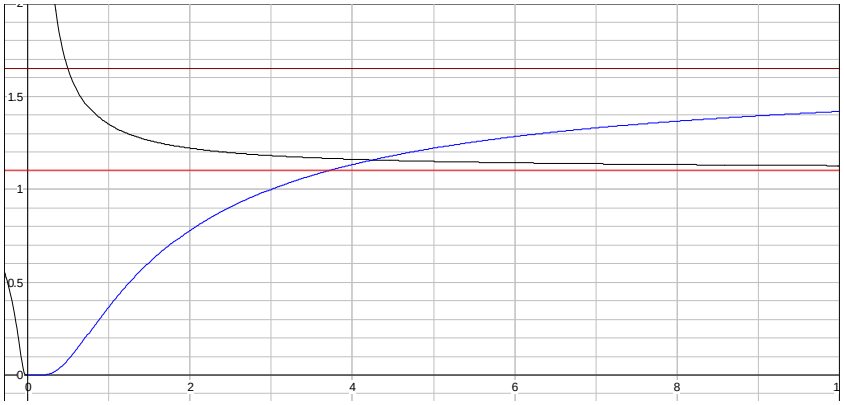


Log-log curves Fig 5.3 in Hamilton





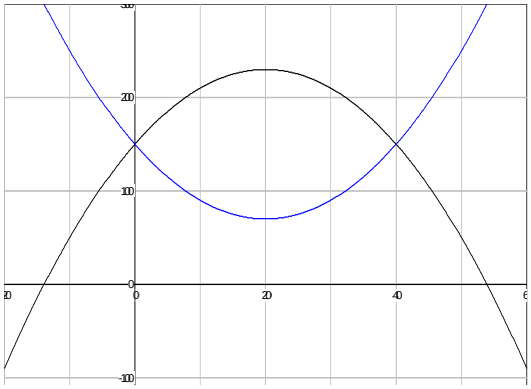
Log-reciprocal curves Fig 5.4 in Hamilton



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)

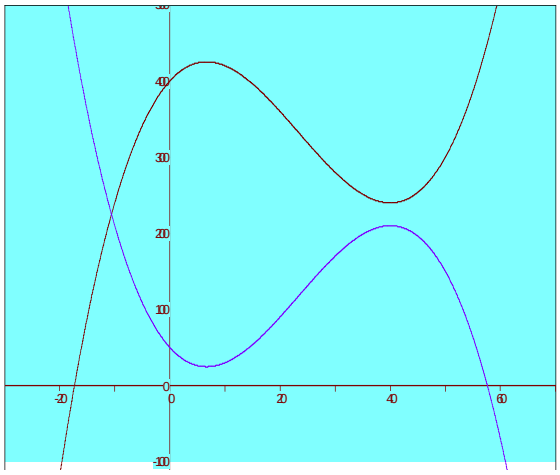
The horizontal lines give the value of y when x grows towards infinity: the asymptote for y

Second order polynomials Fig 5.5 in Hamilton



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

Third order polynomials Fig 5.6 in Hamilton



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

Choice of transformation

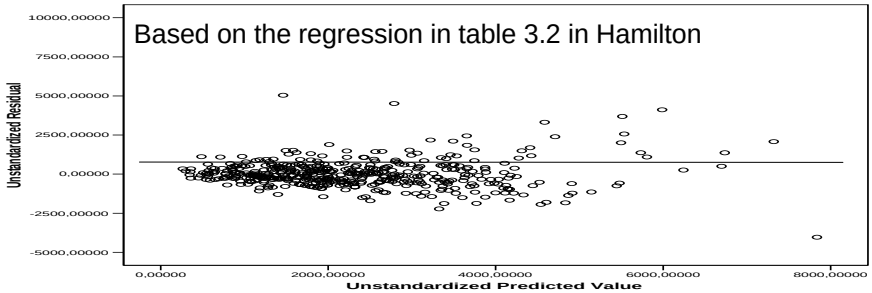
- Scatter plot or theory may provide advice
- Otherwise: transformation to symmetry gives the best option
- The regression reported in table 3.2 in Hamilton proved to be problematic
- Regression with transformed variables can reduce the problems

Choice of transformation in table 3.2 in Hamilton

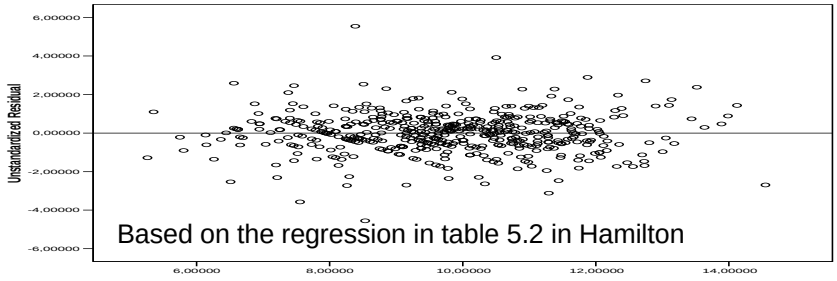
Y = Water use 1981	$Y^*=Y^{0.3}$ provides approximate symmetry
X_1 = Income	$X_1^*= X_1^{0.3}$ provides approximate symmetry
X_2 = Water use 1980	$X_2^*= X_2^{0.3}$ provides approximate symmetry
X_3 = Education	Transformations are inappropriate
X_4 = Pensioner	Transformations do not work for dummies
X_5 = # people in 1981	$X_5^*= \ln(X_5)$ provides approximate symmetry
X_6 = Change in # people	$X_6 = X_5 – X_0$ (= # people in 1980)
X_7 = Relative change in #people	$X_7^* = \ln (X_5/X_0)$

Regression with transformed variables
Tab 5.2 in Hamilton

Dependent Variable: (Wateruse81)$^{0.3}$	B	Std. Err	t	Sig.
(Constant)	1,856	,385	4,822	,000
Income$^{0.3}$	,516	,130	3,976	,000
Wateruse80$^{0.3}$	,626	,029	21,508	,000
Education in Years	-,036	,016	-2,257	,024
Retired?	,101	,119	,852	,395
Ln(# of people81)	,715	,110	6,469	,000
Ln(people81/people80)	,916	,263	3,485	,001



Residual against predicted Y



Other consequences of the transformations

- Two cases with large influence on the coefficient for income (large DFBTAS) do not have such influence (fig 4.11 and 5.9)
- One case with large influence on the coefficient for water use in 1980 do not have that large influence (fig 4.12 and 5.10)
- Transformation to symmetrical distributions will often solve many problems – but not always

Interpretation

- The model estimate now looks like this

$$y_i^{0.3} = 1.856 + 0.516x_{1i}^{0.3} + 0.626x_{2i}^{0.3} - 0.036x_{3i}$$
$$+0.101x_{4i} + 0.715\ln(x_{5i}) + 0.916\ln(\frac{x_{5i}}{x_{0i}})$$

- The interpretation of the coefficients are not so straightforward any more. For example: the measurement units of the parameters have been changed
- The simplest way of interpreting is to use conditional effect plots

Conditional effect plot

- Should be used to study the relationship between the dependent variable and one x-variable with the rest of the x-variables given fixed values
- Typically we are interested in the relationship x-y when the other variables are given values that
  - Maximizes y
  - Are averages values of of the x-variables
  - Minimizes y

Example based on the regression in table 3.2 in Hamilton

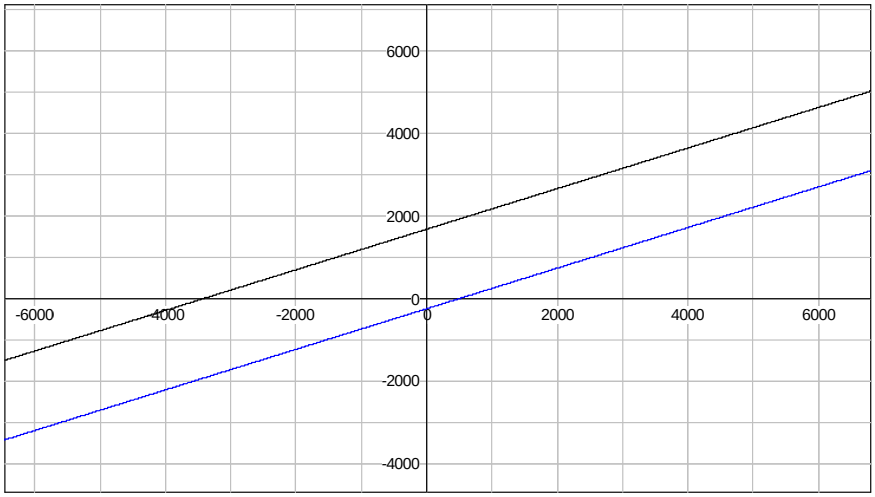
Dependent Variable: Summer 1981 Water Use	Unstandardized Coefficients			
	B	Std. Error	t	Sig.
(Constant)	242,220	206,864	1,171	,242
Summer 1980 Water Use	,492	,026	18,671	,000
Income in Thousands	20,967	3,464	6,053	,000
Education in Years	-41,866	13,220	-3,167	,002
head of house retired?	189,184	95,021	1,991	,047
# of People Resident, 1981	248,197	28,725	8,641	,000
Increase in # of People	96,454	80,519	1,198	,232

To produce conditional effect plot it is useful to have a table of minimum, maximum and average variable values

	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 household retired?	496	0	1	,29
# of people resident, 1981	496	1	10	3,07
Relative increase in # of people	496	-3	3	-,04
# People living in 1980	496	1	10	3,11

The equation

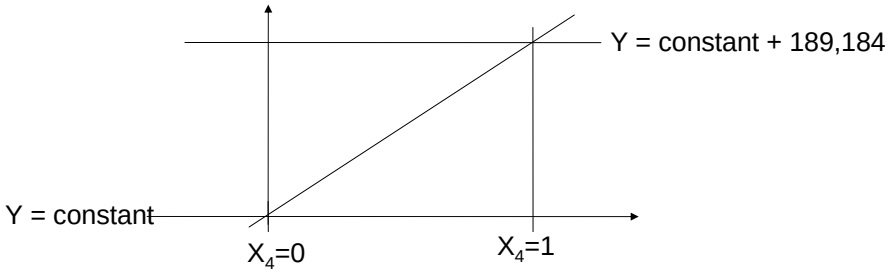
- Estimated Y = 242,22 + 0,492X_1 + 20,967X_2 - 41,866X_3 + 189,184X_4 + 248,197X_5 + 96,454X_6
- Maximizing the effect of X_1 on Y require maximum of X_2 , X_4 , X_5 , X_6 and minimum of X_3
- Average values of the effect of X_1 on Y is obtained by inserting average values of X_2 , X_3 , X_4 , X_5 , X_6
- Minimizing the effect of X_1 on Y require minimum of X_1 , X_2 , X_4 , X_5 , X_6 and maximum of X_3



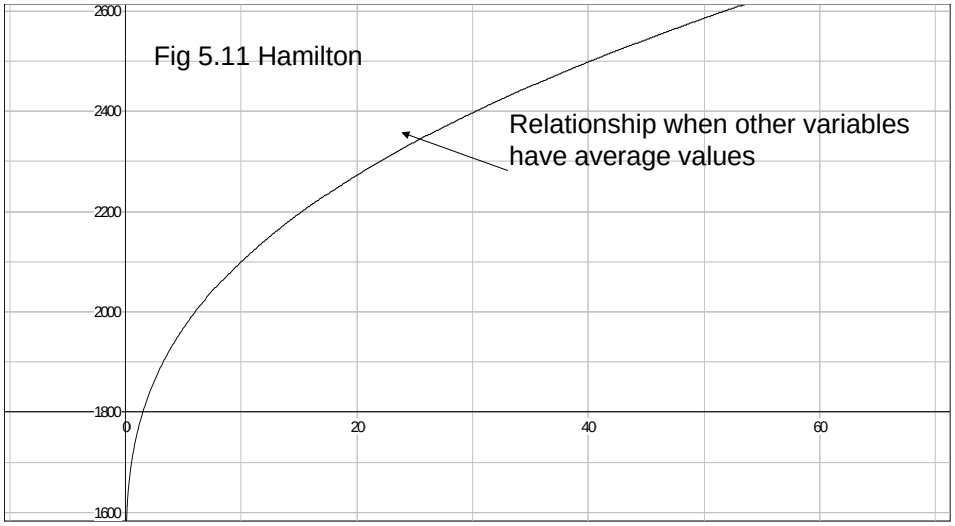
Y = 242.22 + 0.492X + 20.967×10 - 41.866×7 + 189.184×1 + 248.197×5 + 96.454×1
Y = 242.22 + 0.492X + 20.967×1 - 41.866×18 + 189.184×0 + 248.197×1 + 96.454×0

When x is dummy coded

- Estimated Y = 242,22 + 0,492X_1 + 20,967X_2 - 41,866X_3 + 189,184X_4 + 248,197X_5 + 96,454X_6
- Estimated Y = constant + 189,184X_4
  - X_4 can take the values of 0 or 1



Water usage according to income controlled for the effect of other variables



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

Which plots might be of interest?

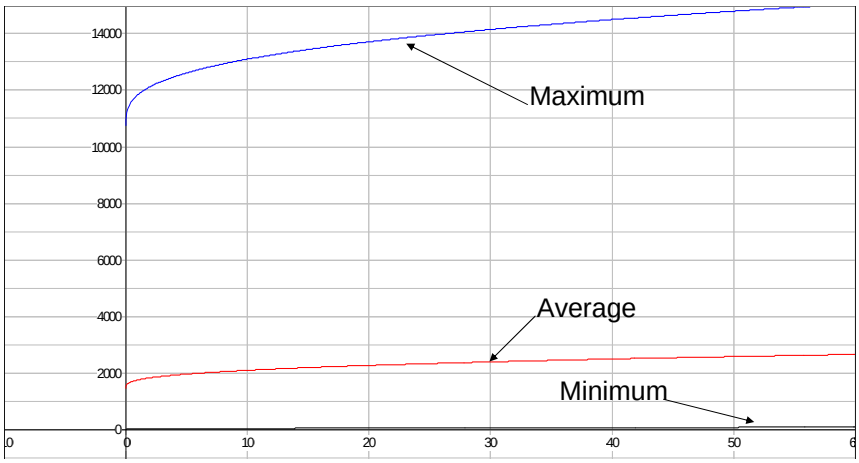
- The relationship between water usage and income controlled for the effect of other variables
  - Those minimizing water usage
  - Those maximizing water usage
  - Average values

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

Comparing three types of usage



Relationship between water usage and income Fig 5.12 in Hamilton

The role of the constant in the plot

- The only difference between the three curves is the constant
  - In the maximum curve (konst) = 14.046
  - In the minimum curve (konst) = 4.204
  - In the average curve (konst) = 8.507

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

- The effect of income varies with the value of (konst)
- When we transform dependent variable all relationships become interaction effects

Comparing effects

- For some relationships the standardized regression coefficient can be used to compare effects, but it is sensitive for biased estimates of the standard error
- A more general method is to compare conditional effect plots where the scaling of the y-axis is kept constant

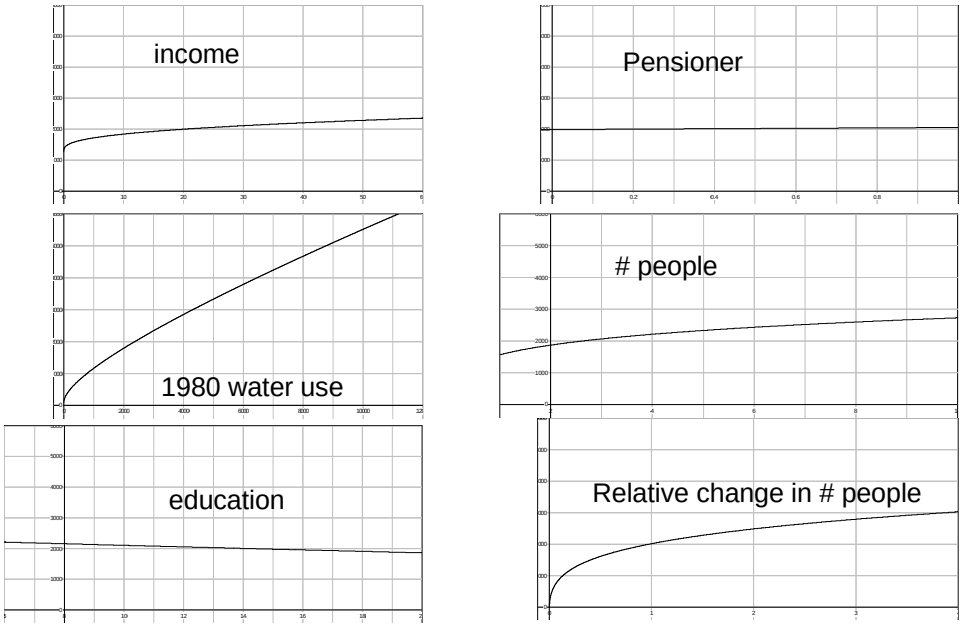
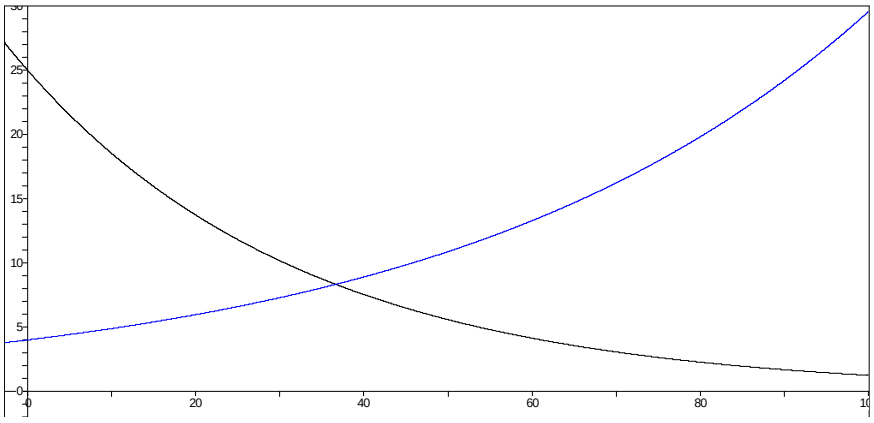


Fig 5.13 Hamilton

Non-linear models

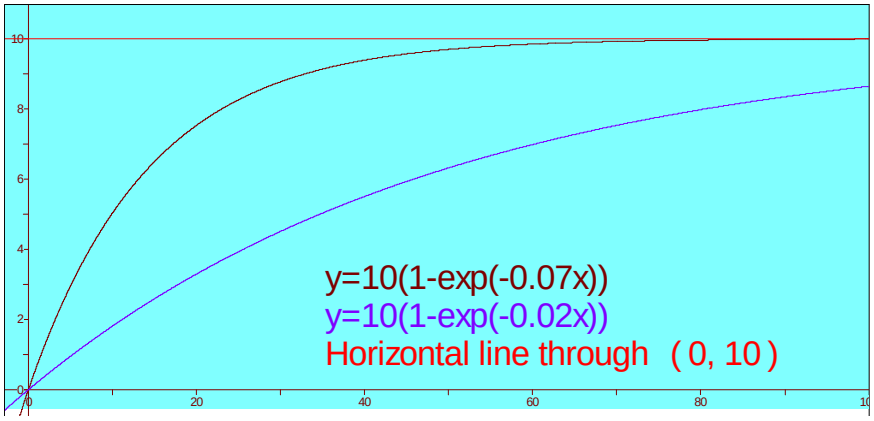
- If we do not have a model that is linear in the parameters other techniques than OLS are needed to estimate the parameters
- One may find two types of arguments for such models
  - Theory about the causal mechanism may say so
  - Inspection of the data may point towards one particular type of model
- We shall take a look at
  - Exponential models
  - Logistic models
  - Gompertz models

Exponential growth and decay
Fig 5.14 in Hamilton

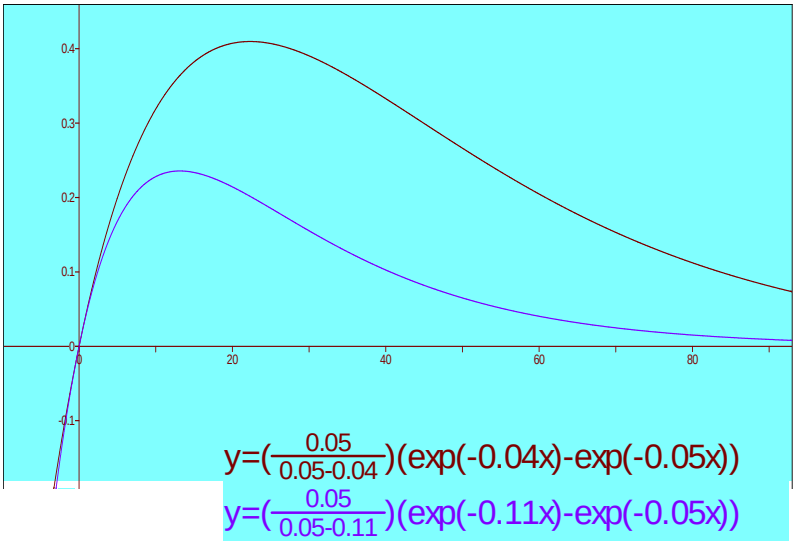


y=25exp(-0.03x)
y=4exp(0.02x)

Negative exponential curves Fig 5.15 in Hamilton



To-term exponential curves Fig 5.16 in Hamilton



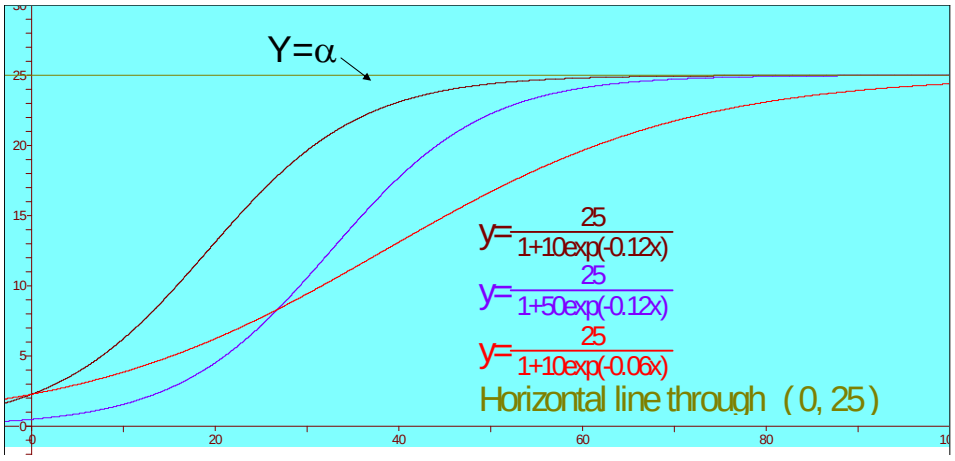
Logistic models

- The logistic function is written
- As x grows towards infinity y will approach α
- When x declines towards minus infinity y will approach 0

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

- Logistic models are appropriate for many phenomena
  - Growth of biological populations
  - Scattering of rumours
  - Distribution of illnesses

Logistic curves Fig 5.17 in Hamilton



- γ determines where growth starts
- β determines how fast the growth is

Logistic probability model

- If it is determined that $\alpha=\gamma=1$ y will vary between 0 and 1 as x goes from minus infinity to plus infinity
- Logistic curves can then be used to model probabilities

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

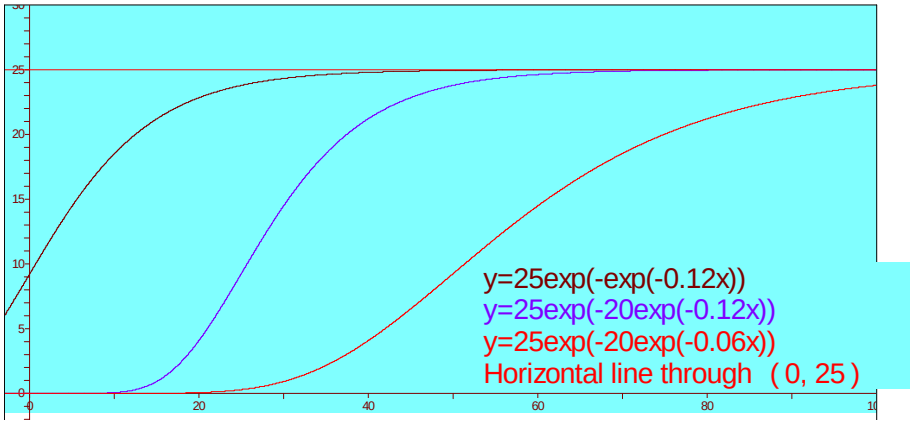
Gompertz curves

- Gompertz curves are sigmoid curves like the logistic, but growth increase and growth reduction occur at different rates. Hence they are not symmetric

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

- Parameters α, γ, and β have the same interpretation as in the logistic model

Gompertz curves Fig 5.18 Hamilton



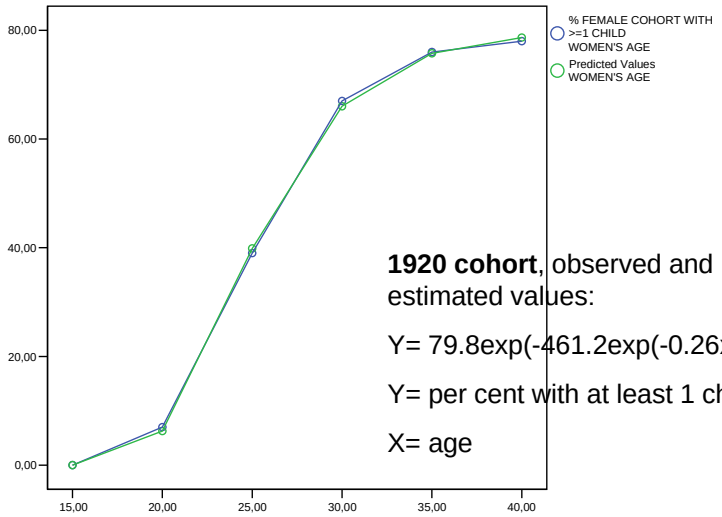
Estimation of non-linear models

- The criterion of fit is still minimum RSS
- It is uncommon to find analytical expressions for the parameters. One has to guess at a start value and go through several iterations to find which parameter value will give minimum RSS
- Good starting values are as a rule necessary, and everything from theory to inspection of data are used to find them

Per cent women with at least 1 child according to the woman's age and year of birth (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	-	-	-	-	-

Estimating Gompertz-models for cohorts (1)



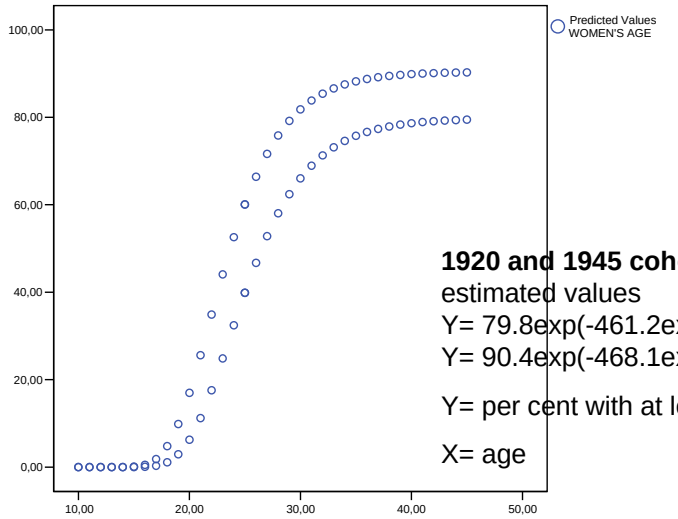
1920 cohort, observed and estimated values:

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

Y= per cent with at least 1 child

X= age

Estimating Gompertz-models for cohorts (2)



1920 and 1945 cohorts, estimated values
Y= 79.8exp(-461.2exp(-0.26x))
Y= 90.4exp(-468.1exp(-0.28x))

Y= per cent with at least 1 child

X= age

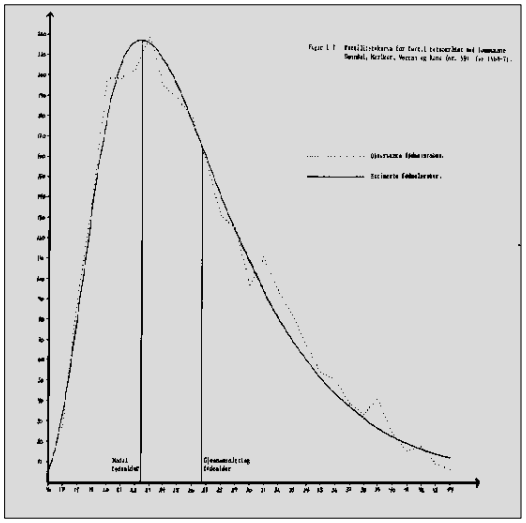
Model estimation and fit

- To evaluate a theoretically developed model
- To predict y within or outside the observed range of variation for x
- Substantial or comparative interpretation of the parameters of the model
  - On cohorts that are not finished with their births (thus predicting outside the observed range of x)
  - We can use the model to compare parameter values of different cohorts

Parameter interpretation
Table 5.6 Hamilton

Cohort	α = upper limit	γ = ?	β = growth speed
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

Birth rates in Sunndal, Meråker, Verran, and Rana 1968-71



- Estimated with a Hadwiger function
- Ref.: Berge, Erling. 1981. The Social Ecology of Human Fertility in Norway 1970. Ph.D. Dissertation. Boston: Boston University.

Conclusions of chapter 5 (1)

- Data analysis often starts with linear models. They are the simplest.
- Theory or exploratory data analysis (band regression, smoothing) can tell us if curvilinear or non-linear models are needed
- Transformation of variables give curvilinear regression. This can counteract several problems:
  - Curvilinear relationships
  - Case with large influence
  - Non-normal errors
  - Heteroscedasticity

Conclusions of chapter 5 (2)

- Non-linear regression use iterative procedures to find parameter estimates
- The procedures need initial values and are often sensitive for the initial values
- The interpretation of the parameters may be difficult. Graphs showing the relationship for different parameter values will provide valuable help for the interpretation