SOS3003
Applied data analysis for social science
Collected lectures 2009

Erling Berge
Department of sociology and political science
NTNU

Lecture I

- Points of departure
- Goals of this class
- Repeating what you are assumed to know

History

- In the history of civilization there are 2 unrivalled accelerators:
  - The invention of writing about 5-6000 years ago
  - The invention of the scientific method for separating facts from fantasy about 5-600 years ago
- There is no topic more important to learn than the basics of the scientific method
- That does not mean that it is not – at times – rather boring ….

Basics of causal beliefs

- First: doubt what you believe is a causal link until you can give good valid reasons justifying your belief
- Second: there are many types of good valid reasons for believing in a particular causal link
  - For example if the overwhelming majority of certified scientists says that human activities contribute to global warming, then we are justified believing that by changing our activities we could contribute less to global warming
- Third: random conjunctures (“correlation”) are not good valid reasons for believing in a causal link

Causal correlations

- This class will focus on how to distinguish between random conjunctures and that which might be a valid causal correlation
- That which might be a valid causal correlation will need a *causal mechanism* explaining how the cause can produce the effect before we have a valid reason to believe in the causal link

Causal mechanism

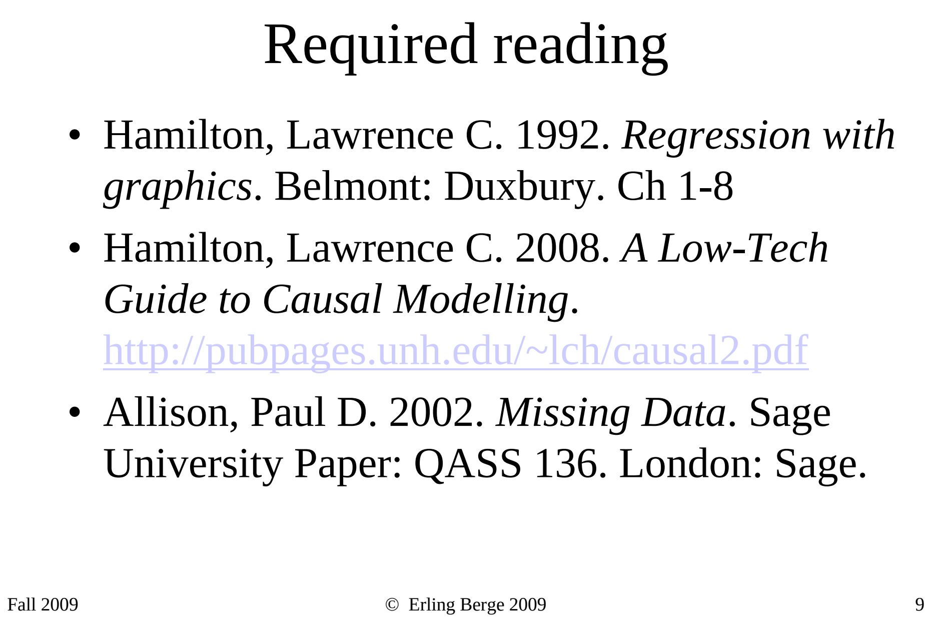
- Elster 2007 *Explaining Social Behaviour*:
- ”mechanisms are frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with indeterminate consequences” (page 36)
- Also sometimes limited to “causal chains”

Primacy of theory

- To say it more bluntly: If you do not have a believable theory (and this may well start as a fantasy) then regression techniques will tell you nothing even if you find a seemingly non-random correlation
- But without a valid and believable regression analysis any believable fantasy will remain just that: a fantasy (assuming you cannot find other valid empirical verification)

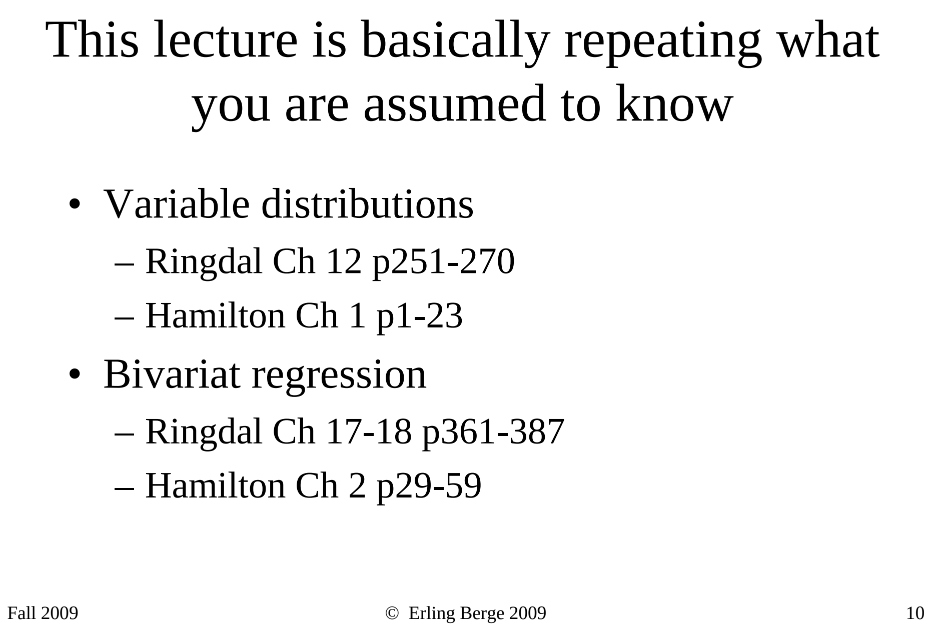
Goals for the class

- The goal is that each of you shall be able to read critically research articles discussing quantitative data. This means
  - You are to know the pitfalls
- You are to learn how to perform straightforward analyses of co-variation in ”quantitative” and ”qualitative” data (nominal scale data in regression anlysis), and in particular:
  - Also here you have to demonstrate that you know the pitfalls



Required reading

- Hamilton, Lawrence C. 1992. *Regression with graphics*. Belmont: Duxbury. Ch 1-8
- Hamilton, Lawrence C. 2008. *A Low-Tech Guide to Causal Modelling*. http://pubpages.unh.edu/~lch/causal2.pdf
- Allison, Paul D. 2002. *Missing Data*. Sage University Paper: QASS 136. London: Sage.



This lecture is basically repeating what you are assumed to know

- Variable distributions
  - – Ringdal Ch 12 p251-270
  - – Hamilton Ch 1 p1-23
- Bivariat regression
  - – Ringdal Ch 17-18 p361-387
  - – Hamilton Ch 2 p29-59

Some basic concepts

- – Cause
- – Model
- – Population
- – Sample
- – Variable: level of measurement
- – Variable: measure of centralization
- – Variable: measure of dispersion

Data analysis

- Descriptive use of data
  - – Developing classifications
- Analytical use of data
  - – Describe phenomena that cannot be observed directly (inference)
  - – Causal links between directly eller indirectly observable phenomena (theory or model development)

Causal analysis: from co-variation to causal connection

- From colloquial speach to theory
  - Fantasy and intuition, established science tradition
- From theory to model
  - Operationalisation
- From observation to generalisation
  - Causal analysis

THREE BASIC DIVISIONS

Observed		Real interest
THEORY/ MODEL	-	REALITY
SAMPLE	-	POPULATION
CO-VARIATION	-	CAUSE

On the one hand we have what we are able to observe and record, on the other hand, we have what we would like to discuss and know more about

Basic sources of error

- Errors in theory / model
  - Model specification: valid tests require a correct (true) model
- Errors in the sample
  - Selection bias
- Measurement problems
  - Missing cases and measurement errors
  - Validity og reliability
- Multiple comparisons
  - Conclusions are valid only for the sample

From population to sample

- POPULATION (all units)

Simple random sampling

- SAMPLE (selected units)

Unit and variable

- A unit, as a carrier of data, will be contextually defined
  - SUPER - UNIT: e.g. the local community
  - UNIT: e.g. household
  - SUB - UNIT: e.g. person
- Variable: empirical concept used to characterize units under investigation. Each unit is characterized by being given a variable value

Data matrix and level of measurement

- Matrix defined by Units * Variables
  - A table presenting the characteristics of all investigated units ordered so that all variable values are listed in the same sequence for all units
- Level of measurement for a variable
  - Nominal scale *classification
  - Ordinal scale *classification and rank
  - Interval scale *classification, rank and distance
  - Ratio scale *classification, rank, distance and absolute zero

Variable analysis

- Description
  - Central tendency and dispersion
  - Form of distribution
  - Frequency distributions and histograms
- Comparing distributions
  - Quantile plots
  - Box plots

VARIABLE: central tendency

- Mean

  sum of all values of the variable for all units divided by the number of units
- MEDIAN

  The variable value in an ordered distribution that has half the units on each side
- MODUS

  The typical value. The value in a distribution that has the highest fequency

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

VARIABLE: measure of dispersion I

- MODAL PERCENTAGE
- The percentage of units with value like the mode
- RANGE OF VARIATION
- The difference between highest and lowest value in an ordered distribution
- QUARTILE DIFFERENCE
- Range of variation of the 50% of units closest to the median (Q_3-Q_1)
- MAD - Median Absolute Deviation
- Median of the absolute value of the difference between median and observed value:
  - $\mathrm{MAD}(x_i) = \mathrm{median}\ |x_i - \mathrm{median}(x_i)|$

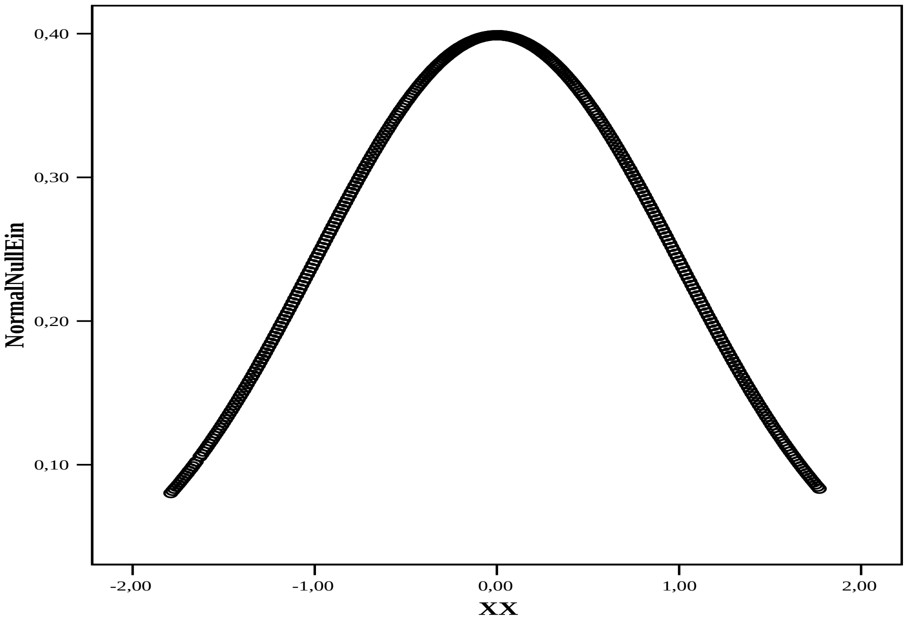
VARIABLE: measure of dispersion II

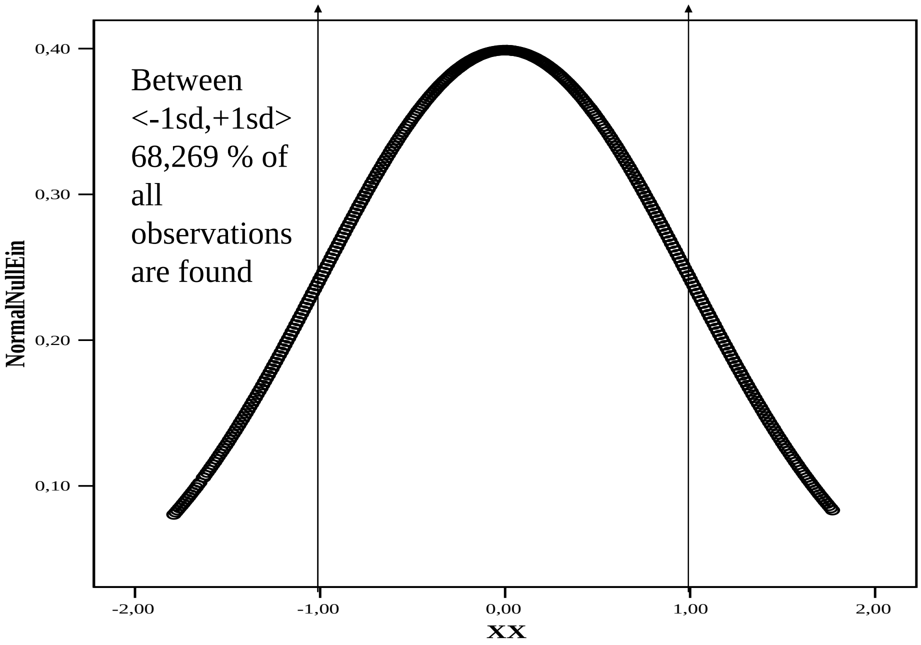
- STANDARD DEVIATION
- Square root of mean squared deviation from the mean
  - $s_y = \sqrt{} [(\Sigma_i(Y_i - \tilde{Y})^2)/(n – 1)]$
- MEAN DEVIATION
- Mean of the absolute value of the deviation from the mean
- VARIANCE
- Standard deviation squared:
  - $s_y^2 = (\Sigma_i(Y_i - \tilde{Y})^2)/(n – 1)$

(ps: here $\tilde{Y}$ is the mean of Y)

Variable: form of distribution I

- Symmetrical distributions
- Skewed distributions
  - ”Heavy” and ”Light” tails
- Normal distributions
  - Are not ”normal”
  - Are unambiguously determined by mean and variance (μ og σ^2)





Between <-1sd,+1sd> 68,269 % of all observations are found

Skewed distributions

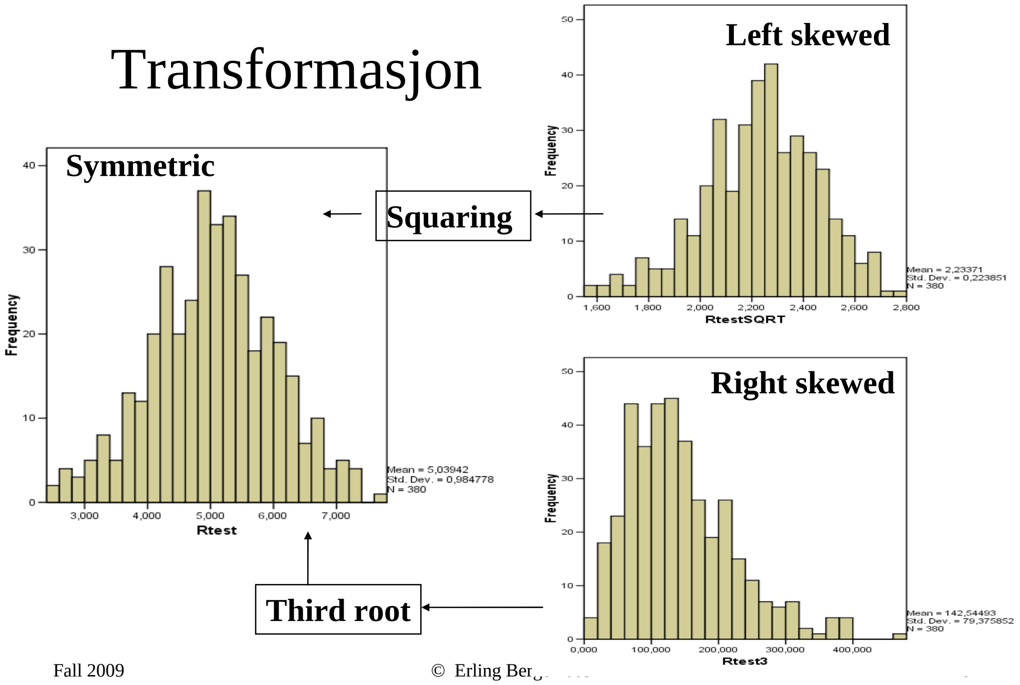
- Positively skewed has Ỹ > Md
- Negatively skewed has Ỹ < Md
- Symmetric distributions has Ỹ ≈ Md

- Ps: Ỹ = mean of Y

Symmetric distributions

- Median and IQR are resistent against the impact of extreme values
- Mean and standard deviation are not
- In the normal distribution (ND) s_y ≈ IQR/1.35
- If we in a symmetric distribution find
  - – s_y > IQR/1.35 then the tails are heavier than in the ND
  - – s_y < IQR/1.35 then the tails are lighter than in the ND
  - – s_y ≈ IQR/1.35 then the tails are about similar to the ND

Transformasjon



Symmetric

Squaring

Left skewed

Right skewed

Third root

Variable: analyzing distributions I

- Boxplot
  - The box is constructed based on the quartile values Q_1 og Q_3 . Observations within < Q_1, Q_3> are in the box-
  - Adjacent large values are defined as those outside the box but inside Q_3 + 1.5*IQR or Q_1 - 1.5*IQR
  - Outliers (seriously extreme values) are those outside of Q_3 + 1.5*IQR or Q_1 - 1.5*IQR

Variables: analyzing distributions II

- Quantiles is a generalisation of quartiles and percentiles
- Quantile values are variable values that correspond to particular fractions of the total sample or observed data, e.g.
  - Median is 0.5 quantile (or 50% percentile)
  - Lower quartile is 0.25 quantile
  - 10% percentile is 0.1 quantile …

Variables: analyzing distributions III

- Quantile plots
  - Quantile values against value of variable
    - The Lorentz curve is a special case of this (it gives us the Gini-index)
- Quantile-Normal plot
  - Plot of quantil values on one vairable against quantil values of Normal distribution with the same mean and standard deviation

Example: Randaberg 1985

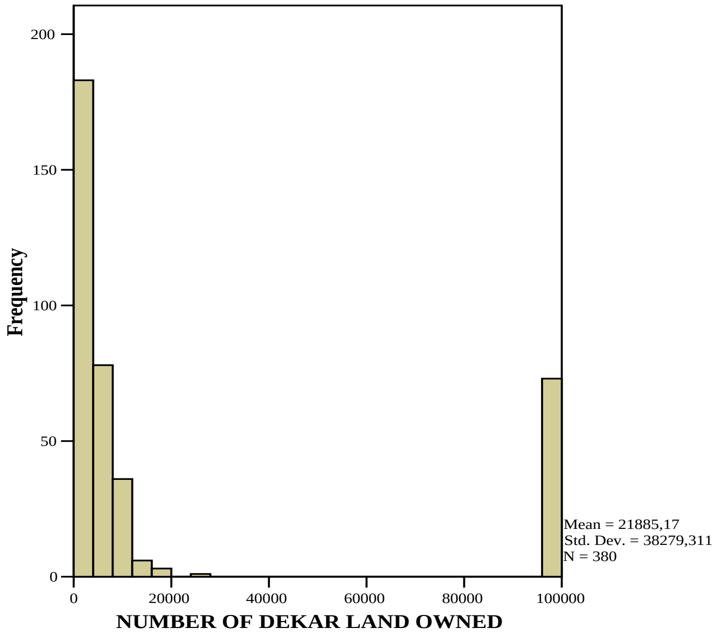
- Questionnaire: (the number of decare land you own / 10 da = 1 ha)

Q: ANTALL DEKAR GRUNN DU eier:_________

(Number of decar you own: ____)

NUMBER OF DEKARE LAND OWNED

	NUMBER OF DEKARE LAND OWNED	Valid N (listwise)
N	380	380
Minimum	0	
Maximum	99900	
Mean	21885.17	
Std. Deviation	38279.311	

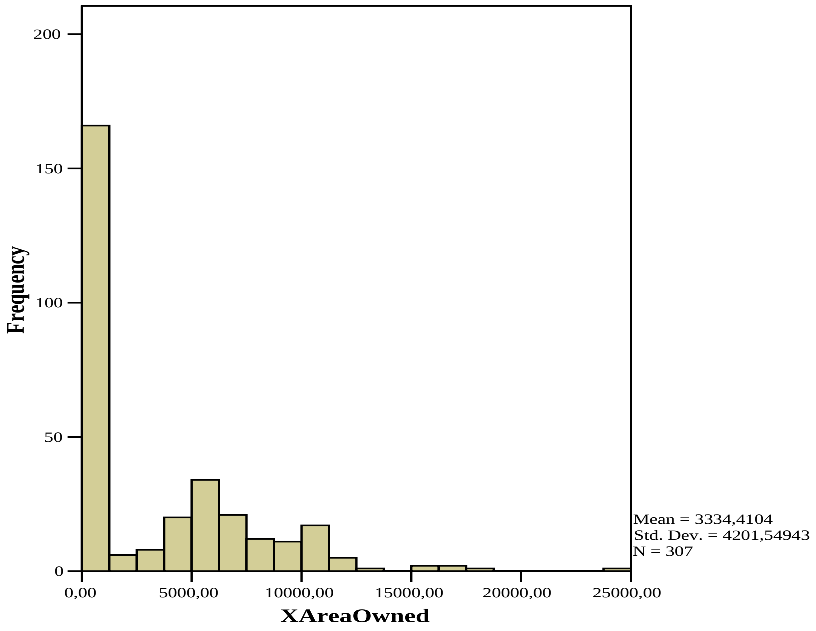


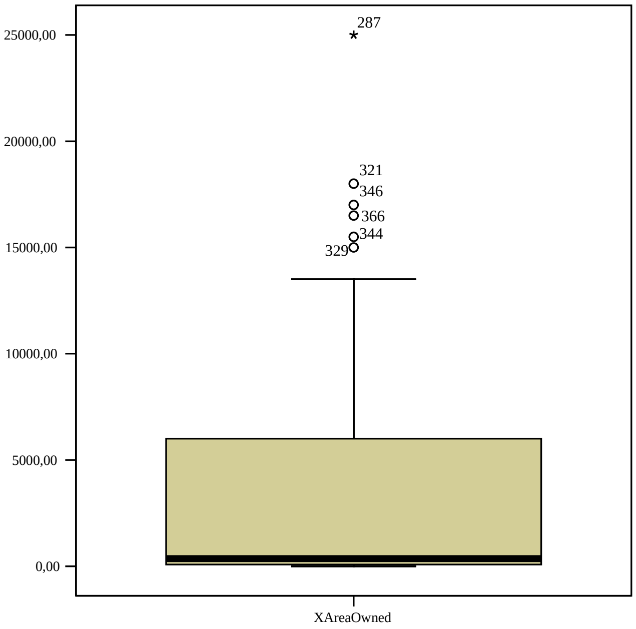
XAreaOwned

(NUMBER OF DEKARE LAND OWNED)

	XAreaOwned	Valid N (listwise)
N	307	307
Minimum	.00	
Maximum	25000.00	
Mean	3334.4104	
Std. Deviation	4201.54943	

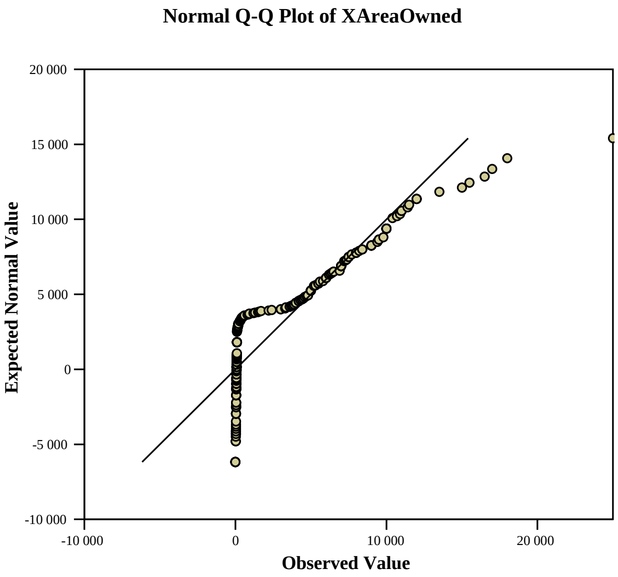
		XAreaOwned	Valid N (listwise)
N	Statistic	307	307
Range	Statistic	25000.00	
Minimum	Statistic	.00	
Maximum	Statistic	25000.00	
Sum	Statistic	1023664.00	
Mean	Statistic	3334.4104	
	Std. Error	239.79509	
Std. Deviation	Statistic	4201.54943	
Variance	Statistic	17653017.596	
Skewness	Statistic	1.352	
	Std. Error	.139	
Kurtosis	Statistic	2.194	
	Std. Error	.277	

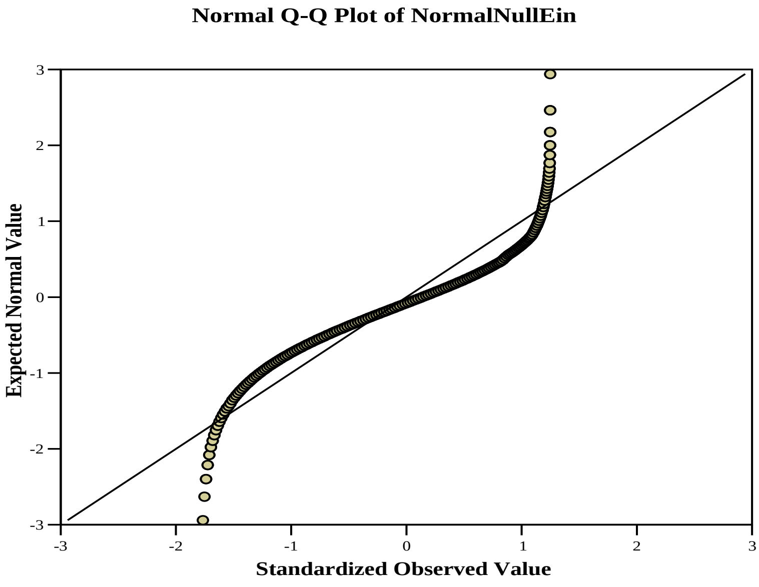

200
150
100
50
0
Frequency
0,00
5000,00
10000,00
15000,00
20000,00
25000,00
XAreaOwned
Mean = 3334,4104
Std. Dev. = 4201,54943
N = 307


25000,00
20000,00
15000,00
10000,00
5000,00
0,00
287
321
346
366
344
329
XAreaOwned

NB

Figures from SPSS are mirrors of figures in Hamilton





Questionnaire:

- **Hvor viktig er det at myndighetene kontrollerer og regulerer bruken av arealer gjennom for eksempel kontroll av**
- av tomtetildelinger (kommunal formidl.)

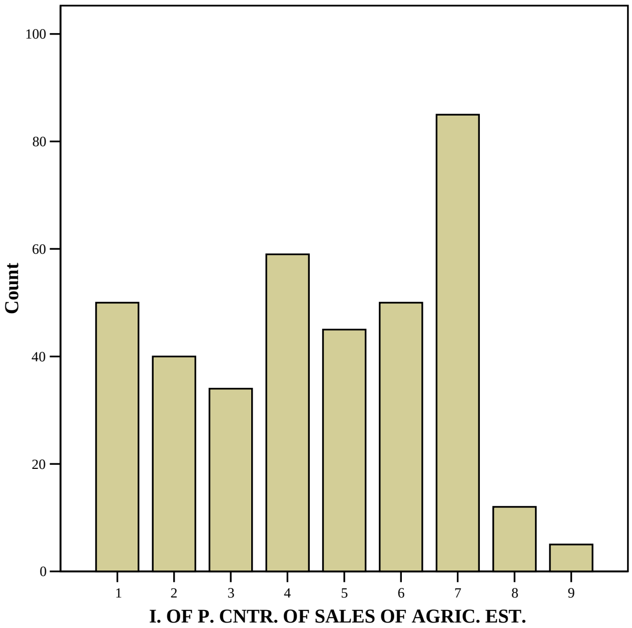
  1 2 3 4 5 6 7 8
- avkjørsler fra hus til vei

  1 2 3 4 5 6 7 8
- kjøp og salg av landbrukseiendommer

  1 2 3 4 5 6 7 8

Importance of public control of sales of agric. estates

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	50	13.2	13.2	13.2
	2	40	10.5	10.5	23.7
	3	34	8.9	8.9	32.6
	4	59	15.5	15.5	48.2
	5	45	11.8	11.8	60.0
	6	50	13.2	13.2	73.2
	7	85	22.4	22.4	95.5
	8	12	3.2	3.2	98.7
	9	5	1.3	1.3	100.0
	Total	**380**	**100.0**	**100.0**	



Questionnaire: coding

Ved utfylling: **sett ring rundt et tall som synes å gi passelig uttrykk for viktigheten når 1 betyr svært lite viktig og 7 særdeles viktig, eller sett et kryss inne i parantesene ( ) som står bak svaret du velger**

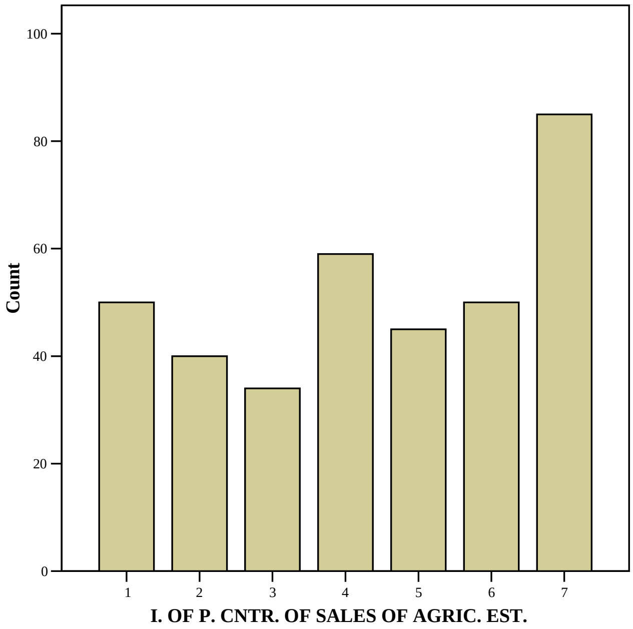
På noen spørsmål kan du krysse av flere svar

	lykkes dårlig/ lite viktig						lykkes godt/ svært viktig	vet ikke
Kodeverdi	1	2	3	4	5	6	7	8

Dei som ikkje kryssar av noko svar vert koda 9 (ie. missing)

I. OF P. CNTR. OF SALES OF AGRIC. EST.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	50	13.2	13.8	13.8
	2	40	10.5	11.0	24.8
	3	34	8.9	9.4	34.2
	4	59	15.5	16.3	50.4
	5	45	11.8	12.4	62.8
	6	50	13.2	13.8	76.6
	7	85	22.4	23.4	100.0
	Total	**363**	**95.5**	**100.0**	
Missing	8	12	3.2		
	9	5	1.3		
	Total	**17**	**4.5**		
Total		**380**	**100.0**		

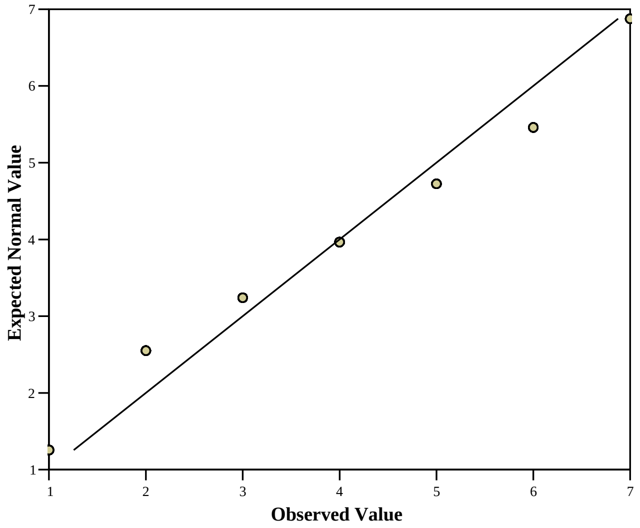


		I. OF P. CNTR. OF SALES OF AGRIC. EST.	YControlSalesAgricEstate Valid N (listwise)
N	Statistic	380	363
Range	Statistic	8	6.00
Minimum	Statistic	1	1.00
Maximum	Statistic	9	7.00
Sum	Statistic	1729	1588.00
Mean	Statistic	4.55	4.3747
	Std. Error	.114	.11045
Std. Deviation	Statistic	2.213	2.10435
Variance	Statistic	4.897	4.428
Skewness	Statistic	-.171	-.234
	Std. Error	.125	.128
Kurtosis	Statistic	-1.148	-1.267
	Std. Error	.250	.255

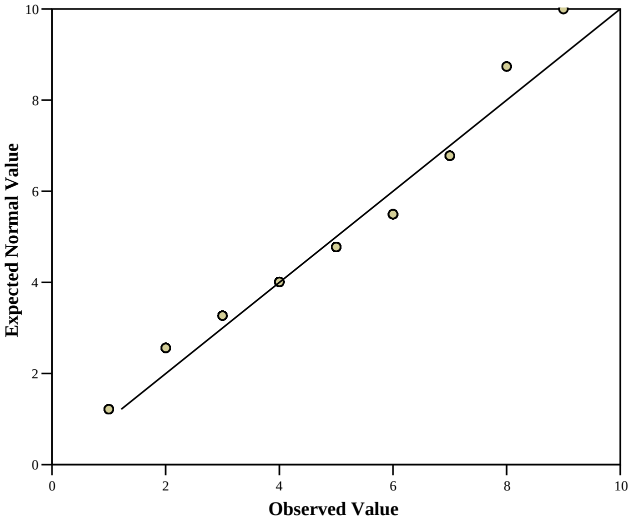
Distributions with or without missing?

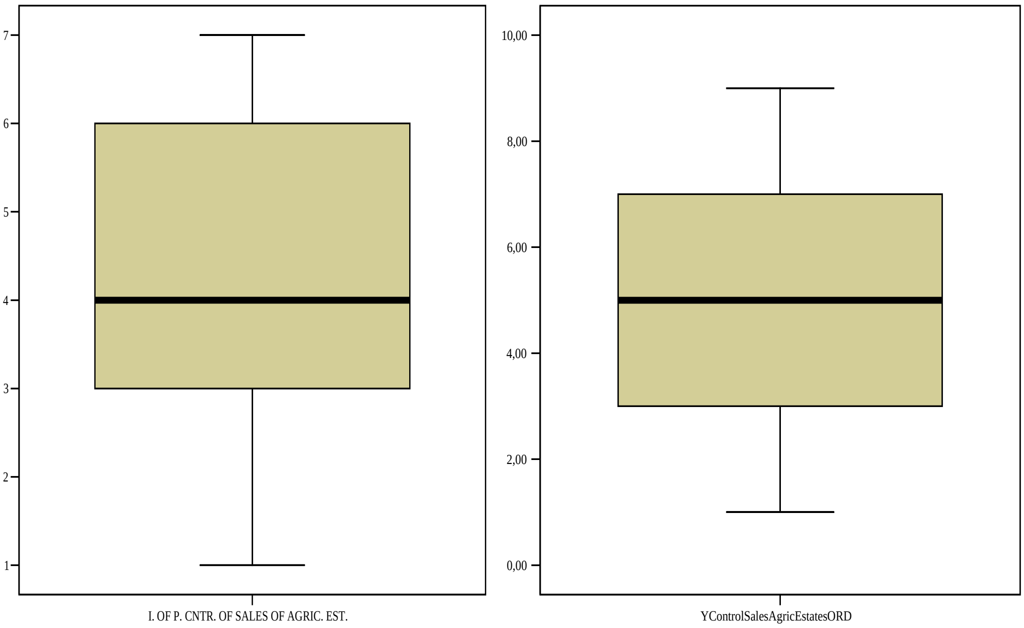
- What difference do the 17 missing observations make in the
  - Quantile-Normal plot?
  - Box plot?

Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.



Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.





Data collection and data quality

- Questions – techniques for asking questions will not be discusssed
- Sample
  - From sampling to final data matrix: selection of cases, refusing to participate, and missing answers on questions
- What is important for the quality of the data?
  - A possible causal link between missing observations and the topic studied
- What can be done if data are faulty?

Writing up a model

- Defining the elements of the model
  - Variables, error term, population, and sample
- Defining the relations among the elements of the model
  - Sampling procedure, time sequence of the events and observations, the functions that links the elements into an equation
- Specification of the assumptions stipulated to be true in order to use a particular method of estimation
  - Relationship to substance theory (specification requirement)
  - Distributional characteristics of the error term

Elements of a model

- Population
- Sample
- Variables
- Error terms

Relations among elements of a model

- Sampling: biased sample?
- Time sequence of events and observations (important to aid ausal modelling)
- Co-variation (genuine vs spurious co-variation)
  - Conclusions about causal impacts require genuine co-variation
- Equations and functions

Bivariat Regresjon: Modelling a population

- $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
- i=1,...,n n = # cases in the population

- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

Bivariat Regresjon: Modelling a sample

- $Y_i = b_0 + b_1\ x_{1i} + e_i$
- i=1,...,n n = # cases in the sample
- e_i is usually called the residual (mot the error term as in the population model)
- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

An example of a bad regression

- The example following contains a series of errors. If you present such a regression in your term paper you will fail
- Your task is to identify the errors as quickly as possible and then never do the same
- Clue: look again at the distributions of the variables above

Importance of public control of sales of agric. Estates

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.047(a)	.002	.000	2.213

a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED

Importance of public control of sales of agric. Estates

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.145	1	4.145	.846	.358(a)
	Residual	1851.905	378	4.899		
	Total	1856.050	379			

a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED
b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

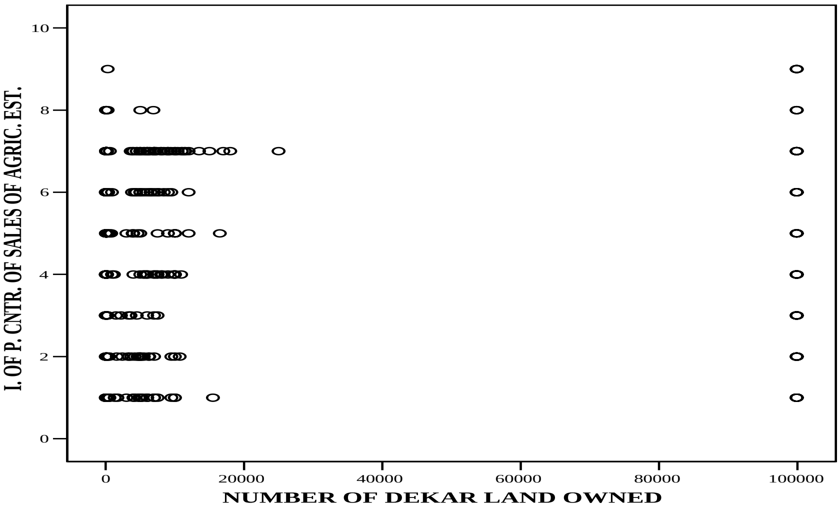
Importance of public control of sales of agric. Estates

Coefficients(a)

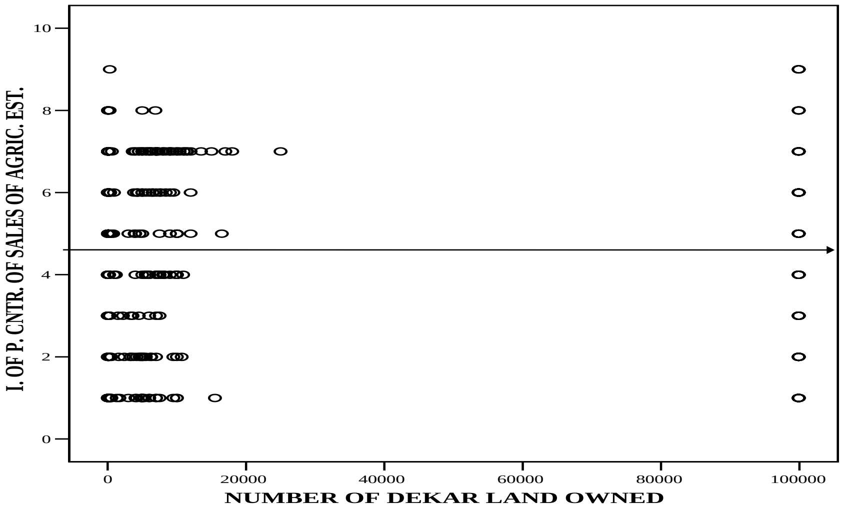
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.610	.131		35.233	.000
	NUMBER OF DEKAR LAND OWNED	.000	.000	-.047	-.920	.358

a Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Scatterplot



Scatterplot with regression line



Assumptions needed for the use of OLS to estimate a regression model

OLS: ordinary least squares (minste kvadrat metoden)

Requirements for OLS estimation of a regression model can shortly be summed up as

- We assume that the linear model is correct (true) with independent, and identical normally distributed error terms (”normal i.i.d. errors”)

Estimation method: OLS

- Model $Y_i = b_0 + b_1 x_{1i} + e_i$

The observed error (the residual) is

- $e_i = (Y_i - b_0 - b_1 x_{1i})$

Squared and summed residual

- $\Sigma_i(e_i)^2 = \Sigma_i (Y_i - b_0 - b_1 x_{1i})^2$

Find b_0 and b_1 that minimizes the squared sum

Relationship sample - population (1)

- A new mathematical operator: E[¤] meaning the expected value of [¤] where ¤ stands for some expression containig at least one variable or unknown parameter, e.g.
- $E[Y_i] = E[b_0 + b_1 x_{1i} + e_i]$

  $= \beta_0 + \beta_1 x_{1i}$
- Note in particular that in our model
  - $E[b_0] = \beta_0$;
  - $E[b_1] = \beta_1$;
  - $E[e_i] = \varepsilon_i$

Relationship sample – population (2)

- Relationship sample - population is determined by the characteristics that the error term has been given in the sampling and observation procedure
- In a simple random sample with complete observation

$E[\ \varepsilon_i\ ] = 0$ for all i, and

var $[\varepsilon_i] = \sigma^2$ for all i

NB: var(¤) is a new mathematical operator meaning ”the procedure that will find the variance of some algebraic expression ”¤”

Complete observation

- Make it possible to make a completely specified model. This means that all variables that causally affects the phenomenon we study (Y) have been observed, and are included in the model equation
- This is practically impossible. Therefore the error term will include also unobserved factors affecting (Y)

Testing hypotheses I

	In reality H_0 is true	In reality H_0 is untrue
We conclude that H_0 is true	Our method gives the correct answer with probability $1 - \alpha$	Error of type II (probability $1 - \beta$)
We conclude that H_0 is untrue	Error of type I The **test level** α is the probability of errors of type I	Our method gives the correct answerwith probability β (= power of the test)

Testing hypotheses II

- A test is always constructed based on the assumption that H_0 is true
- The construction leads to a
  - **Test statistic**
- The test statistic is constructed so that is has a known probability distribution, usually called a
  - **sampling distribution**

T-test and F-test

- Sums of squares
  - TSS = ESS + RSS
  - RSS = $\Sigma_i(e_i)^2 = \Sigma_i(Y_i - \hat{Y}_i)^2$ distance observed- estimated value
  - ESS = $\Sigma_i(\hat{Y}_i - \tilde{Y})^2$ distance estimated value - mean
  - TSS = $\Sigma_i(Y_i - \tilde{Y})^2$ distance observed value – mean
- Test statistic
  - **t = (b - β)/ SE_b** SE = standard error
  - **F = [ESS/(K-1)]/[RSS/(n-K)]** K = number of model parameters

The p-value of a test

- The p-value of a test gives the estimated probability for observing the values we have in our sample or values that are even more in accord with a conclusion that H_0 is untrue; assuming that our sample is a simple random sample from the populasjon where H_0 in reality is true
- Very low p-values suggest that we cannot believe that H_0 is true

Confidence interval for β

- Pick a t_α- value from the table of the t-distribution with n-K degrees of freedom so that the interval

  $< b – t_\alpha(SE_b) , b + t_\alpha(SE_b) >$

  in a two-tailed test gives a probability of α for committing error of type I
- This means that $b – t_\alpha(SE_b) \leq \beta \leq b + t_\alpha(SE_b)$ with probability 1 – α

Coefficient of determination

Coefficient of determination:

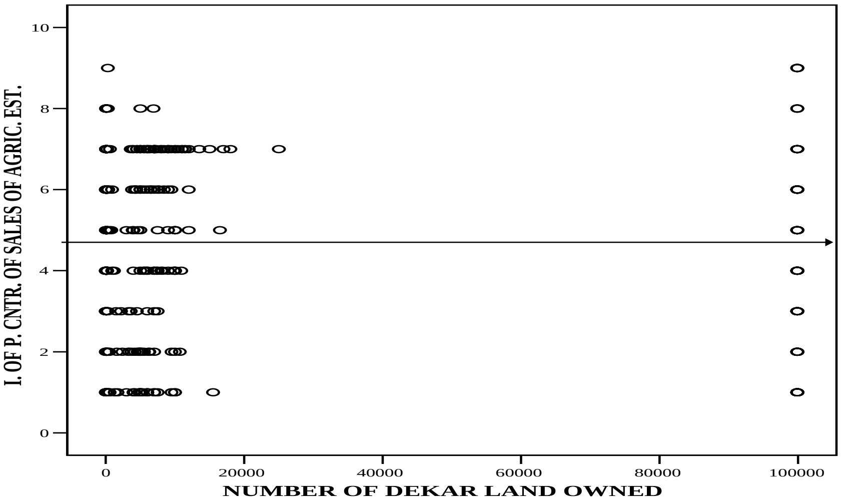
- R^2 = ESS/TSS = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 / \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Tells us how large a fraction of the variation around the mean we can ”explain by” (attribute to) the variables included in the regression ($\hat{Y}_i$ = predicted y)
- In bi-variate regression the coefficient of determination equals the coefficient of correlation: $r_{yu}^2 = s_{yu} / s_y s_u$
- Co-variance: $s_{yu} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(U_i - \bar{U})$

Detecting problems in a regression

- Take a second look at the example presented above where
  - Y = IMPORTANCE OF PUBLIC CONTROL OF SALES OF AGRICULURAL ESTATES
  - **X =** NUMBER OF DEKAR LAND OWNED
  - $Y_i = b_0 + b_1 x_{1i} + e_i$

What was the problem in this example?

What is wrong in this scatterplot with regression line?



In general: what can possibly cause problems?

- Ommitted variables
- Non-linear relationships
- Non-constant error term (heteroskedastisitet)
- Correlation among error terms (autocorrelation)
- Non-normal error terms
- Influential cases

Non-normal errors:

- Regression **DO NOT need assumptions about the distribution of variables**
- But to test hypotheses about the parameters we need to assume thet the **error terms are normally distributed** with the same mean and variance
- **If the model is correct** (true) and n (number of cases) is large the central limit theorem demonstrates that the error terms approach the normal distribution
- **But usually a model will be erroneously or incompletely specified**. Hence we need to inspect and test residuals (observed error term) to see if they actually are normally distributed

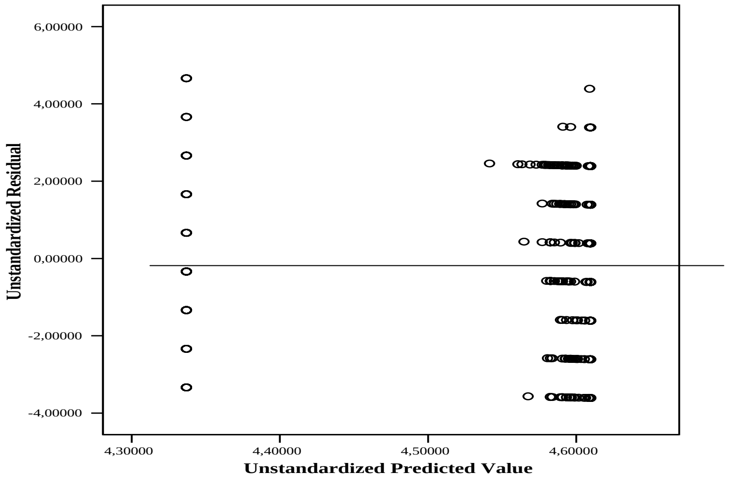
Residual analysis

- This is the most important starting point for diagnosing a regression analysis

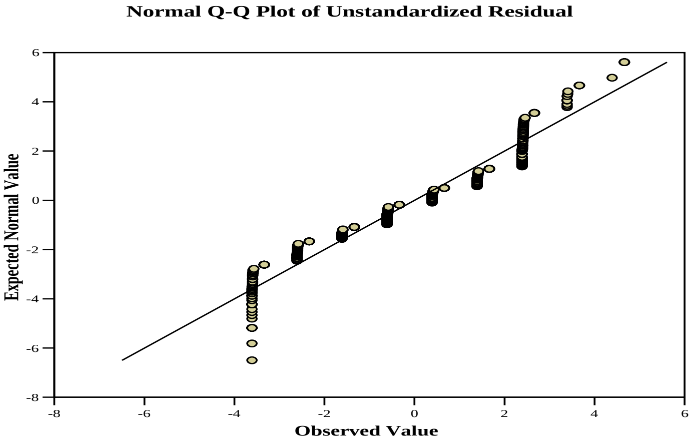
Useful tools:

- Scatterplot
- Plot of residual against predicted value
- Histogram
- Boxplot
- Symmetry plot
- Quantil-normal plot

What went wrong? (1) residual-predicted value plot



What went wrong?
(1) normal-quantile plot



Power transformations

May solve problems related to

- Curvilinearity in the model
- Outliers
- Influential cases
- Non-constant variance of the error term (heteroskedasticity)
- Non-normal error term

NB: Power transformations are used to solve a problem. If you do not have a problem do not solve it.

Power transformations (see H:17-22)

Y* : read "transformed Y"

(transforming Y to Y*)

- $Y^* = Y^q$ q>0
- $Y^* = \ln[Y]$ q=0
- $Y^* = - [Y^q]$ q<0

Inverse transformation

(transforming Y* to Y)

- $Y = [Y^*]^{1/q}$ q>0
- $Y = \exp[Y^*]$ q=0
- $Y = [- Y^*]^{1/q}$q<0

Power transformations: consequences

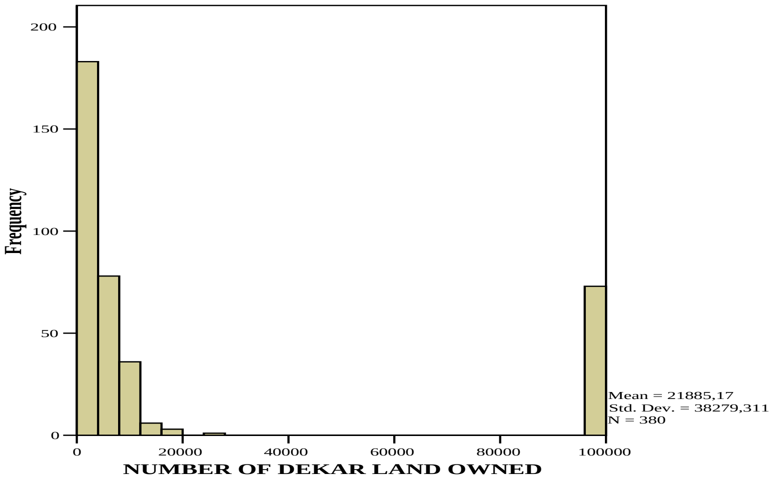
- $X^* = X^q$
  - q > 1 increases the weight of the right hand tail relative to the left hand tail
  - q = 1 produces identity
  - q < 1 redices the weight of the right hand tail relative to the left hand tail
- If $Y^* = \ln(Y)$ the regression coefficient of an interval scale variable X can be interpreted as % change in Y per unit change in X

  E.g. if $\ln(Y)= b_0 + b_1 x + e$

  b_1 can be interpreted as % change in Y pr unit change in X

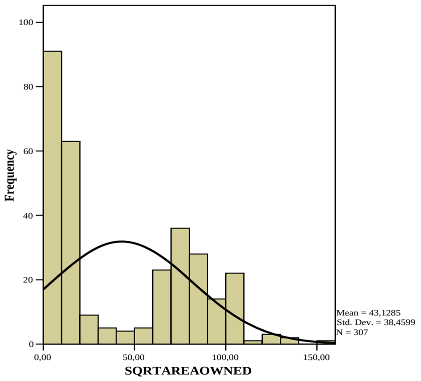
Point of departure

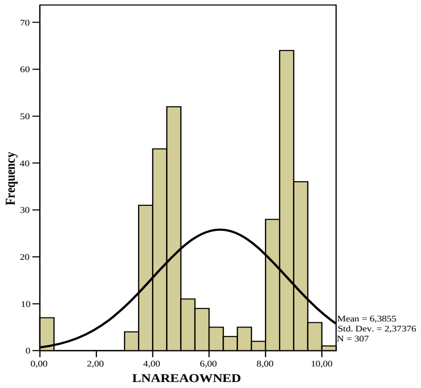
X = NUMBER OF DEKAR LAND OWNED



Power transformed

X = NUMBER OF DEKAR LAND OWNED

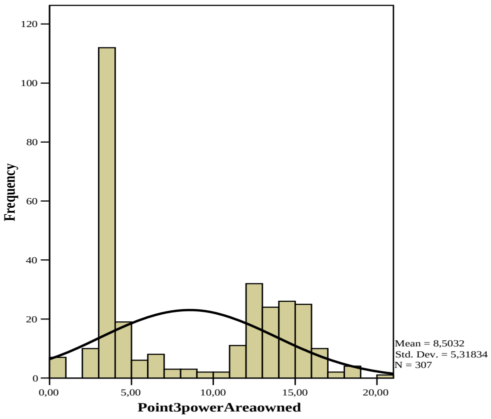




SQRT=square root of areaowned – LN= natural logarithm of (areaowned+1)

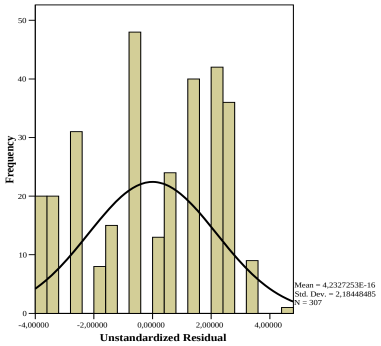
Power transformed

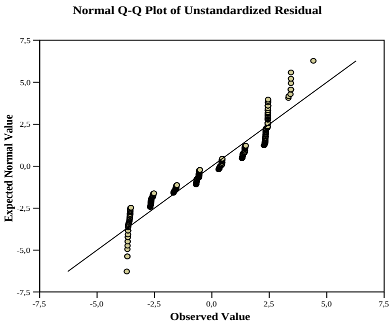
X = NUMBER OF DEKAR LAND OWNED



Point3power = 0,3 power of areaowned

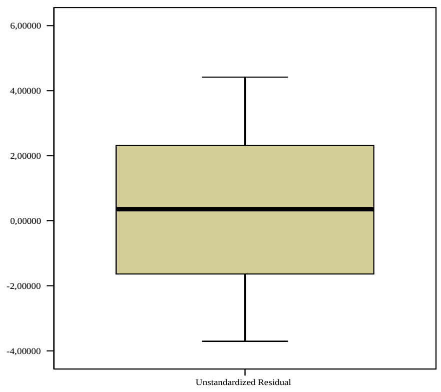
Does power transformation help?





0.3 power-transformation gives lighter tails and no outliers

Box plot of the residual shows approximate symmetry and no outliers



Curviliear regression

- The example above used the variable ”Point3powerAreaowned”, or 0.3 power of number of dekar land owned:
- Point3powerAreaowned = (NUMBER OF DEKAR LAND OWNED)$^{0.3}$

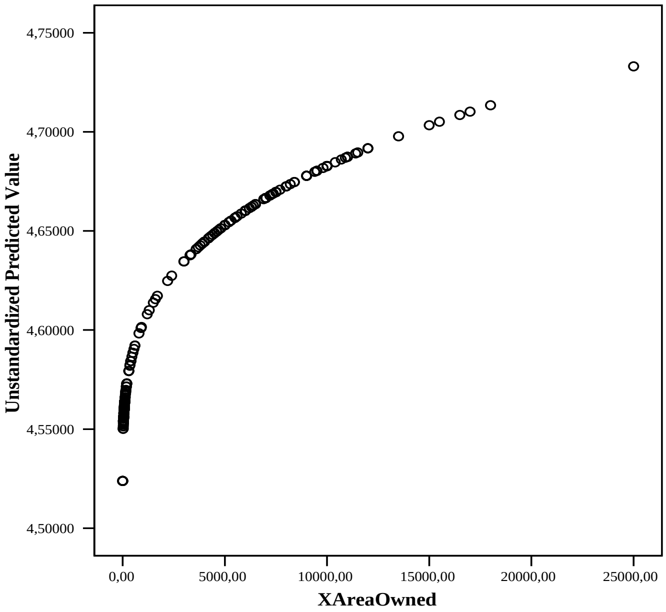
The model estimated is thus

$y_i = b_0 + b_1 (x_i) + e_i$

$y_i = b_0 + b_1$ (Point3powerAreaowned$_i$) + e_i

$\hat{y}_i$ = 4.524 + 0.010*(NUMBER OF DEKAR LAND OWNED$_i$)$^{0.3}$

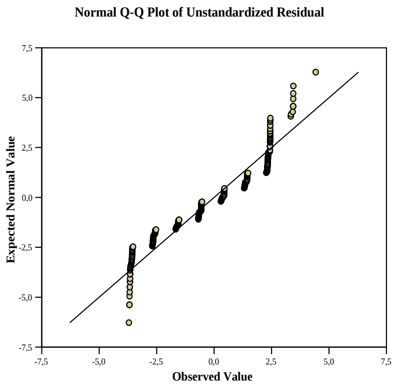
Use of power transformed variables means that the regression is curvilinear

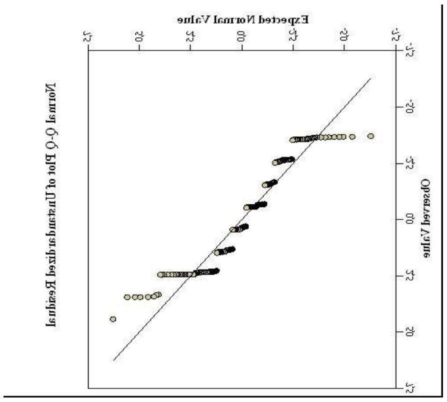


Summary

- In bi-variate regression the OLS method finds the ”best” LINE or CURVE in a two dimensional scatter plot
- Scatter-plot and analysis of residuals are tools for diagnosing problems in the regression
- Transformations are a general tool helping to mitigate several types of problems, such as
  - Curvilinearity
  - Heteroscedasticity
  - Non-normal distributions of residuals
  - Case with too high influence
- Regression with transformed variables are always curvilinear. Results can most easily be interpreted by means of graphs

SPSS printout vs the book (see p16)





Reading printout from SPSS (1)

Descriptive Statistics	Mean	Std. Deviation[1]	N[2]
I. OF P. CNTR. OF SALES OF AGRIC. EST.	4.61	2.185	307
Point3powerAreaowned	8.5032	5.31834	307

Model	R	R Square[3]	Adjusted R Square[4]	Std. Error of the Estimate[5]	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.024(a)	.001	-.003	2.188	.001	.182	1	305	.670

a Predictors: (Constant), Point3powerAreaowned
b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Footnotes to the table above (1)

1. Standard deviation of the mean
2. Number of cases used in the analysis
3. Coefficient of determination
4. The adjusted coefficient of determination (see Hamilton page 41)
5. Standard deviation of the residual

   s_e = SQRT (RSS/(n-K)),

   where SQRT (*) = square root of (*)

Reading printout from SPSS (2)

Model		Sum of Squares[3]	df	Mean Square	F[1]	Sig.[2]
1	Regression	.870	1	.870	.182	.670(a)
	Residual	1460.224	305	4.788		
	Total	1461.094	306			

•Sums of squares: TSS = ESS + RSS

•RSS = $\Sigma_i(e_i)^2 = \Sigma_i(Y_i - \hat{Y}_i)^2$: sum of squared (distance observed – estimated value)

•Mean Square = RSS / df For RSS it is known that df=n-K

K equals number of parameters estimated in the model (b_0 og b_1)

Here we have n=307 and K=2, hence Df = 305

Footnotes to the table above (2)

1. F-statistic for the null hypothesis $\text{beta}_1 = 0$ (see Hamilton p45)
2. p-value of the F-statistic: the probability of finding a F-value this large or larger assuming that the null hypothesis is correct
3. Sums of squares
   1. TSS = ESS + RSS
   2. RSS = $\Sigma_i(e_i)^2 = \Sigma_i(Y_i - \hat{Y}_i)^2$ distance observed value – estimated value
   3. ESS = $\Sigma_i(\hat{Y}_i - \tilde{Y})^2$ distance estimated value – mean
   4. TSS = $\Sigma_i(Y_i - \tilde{Y})^2$ distance observed value – mean

Reading printout from SPSS (3)

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B[1]	Std. Error[2]	Beta[3]	t[4]	Sig.[5]	Lower Bound	Upper Bound
1	(Constant)	4.524	.236		19.187	.000	4.060	4.988
	Point3-powerArea-owned	.010	.024	.024	.426	.670	-.036	.056

Footnotes to the table above (3)

1. Estimates of the regression coefficients b_0 og b_1
2. Standard error of the estimates of b_0 og b_1
3. Standardized regression coefficients: b_1^{st} = b_1*(s_x/s_y) see Hamilton pp38-40
4. t-statistic for the null hypothesis $beta_1$ = 0 (see Hamilton p44)
5. p-value of the t-statistic: the probability of finding a t-value this large or larger assuming that the null hypothesis is correct

SOS3003

Applied data analysis for social science

Lecture notes on

Hamilton Ch 3 p65-101

Basics of multiple regression

Erling Berge
Department of sociology and political science
NTNU

Recall:
Bivariate regression: Modelling a sample

- $Y_i = b_0 + b_1 x_{1i} + e_i$
  - i=1,...,n n = # cases in the sample
- e_i is usually called the residual (mot the error term as in the population model)
- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

Recall:
Bivariate regression: Modelling a population

- $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
  - i=1,...,n n = # cases in the population
  - ε_i is the error term for case no i
- Y and X must be defined unambiguously, and Y must be interval scale (or ratio scale) in ordinary regression (OLS regression)

Summary on bivariate regression

- In bi-variate regression the OLS method finds the ”best” LINE or CURVE in a two dimensional scatter plot
- Scatter-plot and analysis of residuals are tools for diagnosing problems in the regression
- Transformations are a general tool helping to mitigate several types of problems, such as
  - Curvilinearity
  - Heteroscedasticity
  - Non-normal distributions of residuals
  - Case with too high influence
- Regression with transformed variables are always curvilinear. Results can most easily be interpreted by means of graphs

Multiple regression: model (1)

- The goal of multiple regression is to find the net impact of one variable controlled for the impact of all other variables
- Let K= number of parameters in the model (this means that K-1 is the number of variables)
- Then the population model can be written
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1} + \varepsilon_i$

Multiple regression: model (2)

- This can also be written

  $y_i = E[y_i] + \varepsilon_i$,

  this means that

- $E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1}$

  $E[y_i]$ is read as "the expected value of y_i"

Multiple regression: model (3)

- We will find the OLS estimates of the model parameters as the b-values in

  $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{K-1} x_{i,K-1}$

  ($\hat{y}_i$ is read as "estimated" or "predicted" value of y_i)

  That minimizes the squared sum of the residuals

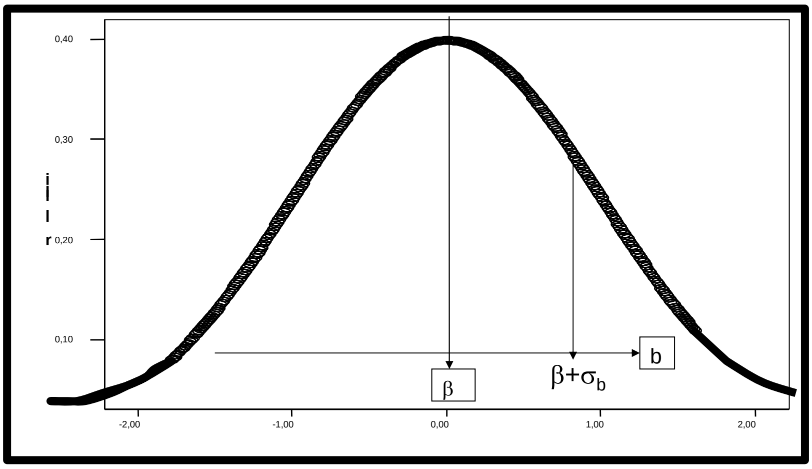
$$RSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} e_i^2$$

Estimation methods

- OLS: parameters are found by minimizing RSS
- But this is not the only method for finding suitable b-values. Two alternatives are:
  - WLS: Weighted least squates
  - ML: maximum likelihood

More on testing hypotheses

- We can draw many samples from a population
- In every new sample we can estimate new values (a new b-value) of the same regression parameter (β)
- If we make a histogram of the many estimates of e.g. b_1 we will see that b_1 has a distribution. This distribution is called the sampling distribution of β_1
- Different types of parameters have different types of sampling distributions
- Regression parameters (β-as) have a t-distribution



Sampling distribution of the regression parameter b:

E[b]= β

On partial effects (1)

- Example with 2 variables
- If we estimate a model

  $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$

  it will in principle involve 3 different correlations:
  - Between y and x_1
  - Between y and x_2
  - Between x_1 and x_2

On partial effects (2)

- This might have been represented by 3 different bivariate regressions where the third variable was kept constant

  (1) $y = a_{yIx1} + b_{yIx1}x_1 + e_{yIx1}$ x_2 constant

  (2) $y = a_{yIx2} + b_{yIx2}x_2 + e_{yIx2}$ x_1 constant

  (3) $x_1 = a_{x1Ix2} + b_{x1Ix2}x_2 + e_{x1Ix2}$ y constant

  the index ”yIx1” is read ”from the regression of y on x1”

- Equations (2) and (3) can be rewritten as:

On partial effects (3)

(2) $e_{yIx2} = y - (a_{yIx2} + b_{yIx2}x_2)$

(3) $e_{x1Ix2} = x_1 - (a_{x1Ix2} + b_{x1Ix2}x_2)$

We may interprete this as a removal of the effect of x_2 from y and from x_1

We also see that e_{yIx2} and e_{x1Ix2} become new

variables where the effect of x_2 has been removed

On partial effects (4)

- If we based on this make a new regression

  $\hat{e}_{yIx^2} = a + b\, e_{x^1Ix^2}$

  we find that

  a = 0

  b = b_1 from the regression

  $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$

- b_1 is in other words the effect of x_1 on y after we have removed the effect of x_2

Experiments and partial effects

- Experiments investigate the causal connection between two variables controlled for all other causal impacts
- Multiple regression is a kind of half-way replication of experiments – the next best solution – and is a close relative of quasi-experimental research designs

Partial effects

A leverage plot for y and x_k is a plot where

- y-axis is the residual from the regression of y on all x-variables except x_k , and
- x-axis is the residual from regression of x_k on all the other x-variables

The regression line in such a plot will always go through y=0 and will ahve a slope coefficient equal to b_k

An example with 2 independent variables

Table 2.2 Dependent: **Summer 1981 Water Use**	**B**	Std. Error	t	Sig.
(Constant)	**1201.124**	123.325	9.740	.000
Income in Thousands	**47.549**	4.652	10.221	.000

Table 3.1 Dependent: **Summer 1981 Water Use**	**B**	Std. Error	t	Sig.
(Constant)	**203.822**	94.361	2.160	.031
Income in Thousands	**20.545**	3.383	6.072	.000
Summer 1980 Water Use	**.593**	.025	23.679	.000

From the table 2.2 (p46) and 3.1 (p68) in Hamilton. In the tables in the book the constant is on the last line. SPSS put it on the first line.

Question: What does it mean that the coefficient of income declines when we add a new variable?

On the addition of new variables

- It is not common that existing theory will give precise prescriptions for what variables to include in a model. Usually there is an element of trial and error in developing a model
- When new variables are added to a model several things happen
  - The explanatory force increase: R^2 increase, but will the increase be significant?
  - The coefficient of the regression shows the effect on y. Is this effect significantly different from 0?
  - If the coefficient is significantly different from 0, is it also so big that it is of substantial interest?
  - Spurious coefficients can decline. Do the new variable change the interpretation of the effect of the other variables?

Parsimony

- Parsimony is what might be called an aesthetic criterion of a good model. We want to explain as much as possible of the variation in y by means of as few variables as possible
- The adjusted coefficient of determination, Adjusted R^2, is based on parsimony in the sense that it takes into consideration the complexity of the data relative to the complexity of the model by the difference between n and K

  (n-K is the degrees of freedom in the residual,

  n = number of observations, K = number of estimated parameters)

Irrelevant variable

- Including irrelevant variables
  - A variable is irrelevant if the real effect (β) is 0; or more pragmatically, if it si so small that it has no substantive interest
  - **Inclusion of an irrelevant variable** makes the model unnecessarily complex and will have the consequence that coefficient estimates on all variables have larger variance (coefficients varies more form sample to sample)
- Including an irrelevant variable is probably the least damaging error we can do

Relevant variable

- A variable is relevant if
  1. Its real effect (β) is significantly different from 0, and
  2. Large enough to have substantive interest, and
  3. Is **correlated with other included x-variables**
- If we exclude a relevant variable all results from our regression will be unreliable. The model is unrealistically simple
- Not including a relevant variable is the most damaging error we can do. But consider requirement 2 and 3. This makes it a lot easier to avoid this problem.

Sample specific results?

- Choice of variables is a trade-off among risks. Which risk is worse depends on the purpose of the study and the strength of relations
- With a test level of 0.05 one may easily find sample specific results. In about 5% of all samples a coefficient that show up as not significantly different from 0 will in ”reality” be different from 0 ($\beta \neq 0$) and vice versa for those we find to be significantly different from 0
- The best defence against this is the theoretical argument for finding an effect different from 0

Hamilton (s74) example

y_i	Postshortage water use (1981)
x_{i1}	Household income, in thousands of dollars
x_{i2}	Preshortage water use, in cubic feet (1980)
x_{i3}	Education of household head, in years
x_{i4}	retirement (coded 1 if household head is retired and 0 otherwise)
x_{i5}	Number of people living in household at time of water shortage (summer 1981)
x_{i6}	Change in number of people, summer 1981 minus summer 1980

Table 3.2 (Hamilton p74)

Dependent Variable: Summer 1981 Water Use	B	Std. Error	t	Sig.	Beta
(Constant)	242.220	206.864	1.171	.242	
Income in Thousands	20.967	3.464	6.053	.000	.184
Summer 1980 Water Use	.492	.026	18.671	.000	.584
Education in Years	-41.866	13.220	-3.167	.002	-.087
Head of house retired?	189.184	95.021	1.991	.047	.058
# of People Resident, 1981	248.197	28.725	8.641	.000	.277
Increase in # of People	96.454	80.519	1.198	.232	.031

How do we interprete the coefficient of ”Increase in # of People” ?

What leads to less water use after the crisis?

Standardized coefficients

- Standardized variables (z-scores)

  $z_{ix} = (x_i - \bar{x})/s_x$

  (means unit of measurement is standad deviation)

- Standardized regression coefficients (beta-weights, or path coefficients)

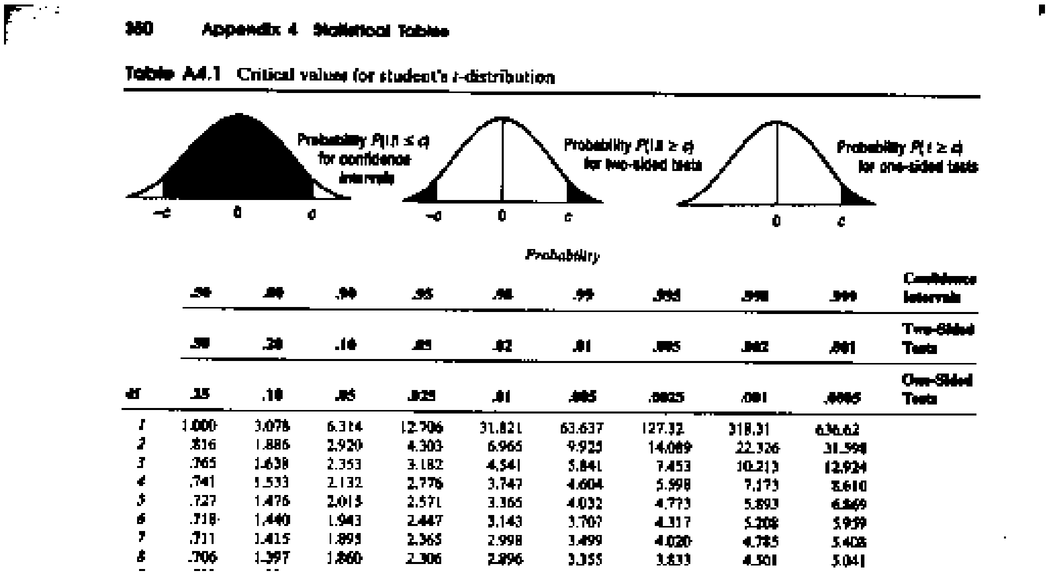
  $b_k^s = b_k(s_k/s_y)$ (varies between -1 and +1)

- Predicted standard score of y_i (z_{iy} with hat) =

  $0.18z_{i1} + 0.58z_{i2} - 0.09z_{i3} + 0.06z_{i4} + 0.28z_{i5} + 0.03z_{i6}$

t-test

- The difference between the observed coefficient (b_k) and the unobserved coefficient (β_k) standardized by the standard deviation of the observed coefficient (SE_{b_k}) will usually be very close to zero if the observed b_k is close to the population value. This means that we in the formula
- $t = (b_k - \beta_k)/ SE_{bk}$ substitutes H_0: $\beta_k = 0$ and find that ”t” is small we will believe that the population value β_k in reality equals 0
- How big ”t” has to be before we stop believing that $\beta_k = 0$ we can find from knowing the sampling distribution of b_k and SE_{b_k}



380 Appendix 4 Statistical Tables

Table A4.1 Critical values for student's t-distribution

Probability $P(|t| \le c)$ for confidence intervals
$-c$
0
c
Probability $P(|t| \ge c)$ for two-sided tests
$-c$
0
c
Probability $P(t \ge c)$ for one-sided tests
0
c

Probability

	.50	.80	.90	.95	.98	.99	.995	.998	.999	Confidence Intervals
	.50	.20	.10	.05	.02	.01	.005	.002	.001	Two-Sided Tests
df	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	One-Sided Tests
1	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62	
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598	
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924	
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610	
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	
7	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041	

”t” has a sampling distribution called the t-distribution The t-distribution varies with the number of degrees of freedom (n-K) and is listed according to level of significance α

Confidence interval for β

- Chose a t_α-value from the table of the t-distribution with n-K degrees of freedom
- Then if $H_0 : \beta_k = b_k$ is correct, a two tailed test will have a probability of α to reject H_0 when H_0 in reality is correct (type I error)
- This means that there is a probability of α that β_k in reality is outside the interval

$$< \mathbf{b_k – t_\alpha(SE_{b_k}) , b_k + t_\alpha(SE_{b_k})} >$$

- This is equivalent to saying that

$$\mathbf{b_k – t_\alpha(SE_{b_k}) \le \beta_k \le b_k + t_\alpha(SE_{b_k})}$$

is correct with probability 1 – α

F-test: big model against small

RSS{*} = residual sum of squares with index {*} where

* stands for number of parameters in the model

  - Big model: RSS{K}
  - Small model: RSS{K-H}
  - H equals the difference in the number of parameters in the two models

- Define:

$$F^H_{n-K} = \frac{(RSS\{K\text{-}H\} – RSS\{K\})/H}{(RSS\{K\})/(n\text{-}K)} = F[H, n\text{-}K]$$

F[H, n-K] will have the sampling distribution called the F-distribution with H and n-K degrees of freedom

Example (Hamilton table 3.1 and 3.2)

Liten modell Table 3.1	**Sum of Squares**	df	Mean Square	F	Sig.
Regression **(Model)** **(Explained)**	671025350.237	2	335512675.119	391.763	.000(a)
Residual	422213359.440	493	856416.551		
Total	1093238709.677	495			

Stor model Tabell 3.2	Sum of Squares	df	Mean Square	F	Sig.
Regression	740477522.059	**K - 1 = 6**	123412920.343	171.076	.000(a)
Residual	352761187.618	**n - K = 489**	721393.022		
Total	1093238709.677	**n - 1 = 495**			

Test if the big model (7 parameters) is better than the small (3 parameters)

Notes to the example

- K = number of parameters of the big model (6 variablar pluss konstant) = 7
- H = K – [number of parameters in the small model (2 variables plus constant)] = 7 – 3 = 4
- RSS{K-H} = 422213359.440
- RSS{K} = 352761187.618
- n = 496
- n – K = 496 – 7 = 489
- (RSS{K-H} – RSS{K})/H = (422213359.440 - 352761187.618)/4 = 17363042.9555
- RSS{K)/(n-K) = 352761187.618/489 = 721393.0217

Testing all parameters in one test

- If the big model has K parameters and we let the small model be as small as possible with only 1 parameter (the constant = the mean) our test will have H=K-1. Inserting this into our formula we have

$$F_{\{K-1,\, n-K\}} = \frac{ESS/(K-1)}{RSS/(n-K)}$$

This is the F-value we find in the ANOVA tables from SPSS

Multicollinearity (1)

- Multicollinearity only involves the x-variables, not y, and is about linear relationships between two or more x-variables
- If there is a perfect correlation between 2 explanatory variables, e.g. x and w (r_{xw} = 1) the multiple regression model breaks down
- The same will happen if there is perfect correlation between two groups of x-variables

Multicollinearity (2)

- Perfect correlation is rarely a practical problem
- But high correlations between different x-variables or between groups of x-variables will make estimates of their effect unreliable. The regression coefficients will have a very large standard deviation and t-tests will practically speaking have no interest whatsoever
- F-tests of groups of variables will not be affected by this

Search strategies

- There are methods for authomatic searches for explanatory variables in a large set of data
- The best advice to give on this is to avoid using it
- One problem is that the p-values of the tests from such searches are wrong and too ”kind” (the problem of multiple comparisons)
- Another problem is that such searches do not work well if the variables are highly correlated

Dummy variables: group differences

- Dichotomous variables taking the values of 0 or 1 are called dummy variables
- In the example in table 3.2 (p74) x_{i4} is (Head of house retired?) a dummy variable
- First put into the equation $x_{i4} = 1$ then $x_{i4} = 0$

$y_i = 242 + 21x_{i1} + 0.49x_{i2} - 42x_{i3} + 189x_{i4} + 248x_{i5} + 96x_{i6}$ og

- Explain what the two equations tell us

Interaction

- There is interaction between two variables if the effect of one variable changes or varies depending on the value of the other variable

Interaction effects in regression (1)

- If we do a non-linear transformation of y all estimated effects will implicitly be interaction effects
- Simple additive interaction effects can be included in a linear model by means of product terms where two x-variables are multiplied
- $\hat{y}_i = b_0 + b_1x_i + b_2w_i + b_3x_iw_i$
- Conditional effect plots will be able to illustrate what interaction means

Interaction effects in regression (2)

- An interaction effect involving x and w can be included in a regression model by means of an auxiliary variable equal to the product of the two variables, i.e.
- Auxiliary variable H=x*w
- $y_i = b_0 + b_1*x_i + b_2*w_i + b_3*H_i + e_i$
- $y_i = b_0 + b_1*x_i + b_2*w_i + b_3*x_i *w_i + e_i$

Example from Hamilton(p85-91)

Let

- y = natural logarithm of chloride concentration
- x = depth of well (1=deep, 0=shallow)
- w = natural logarithm of distance from road
- xw = interaction term between distance and depth (product x*w). Then
- $\hat{y}_i = b_0 + b_1 x_i + b_2 w_i + b_3 x_i w_i$

First take a look at the simple models without interaction

Figures 3.3 and 3.4 (Hamilton p85-86)

Figure 3.3 is based on

Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	3.775	.429		8.801	.000
x= BEDROCK OR SHALLOW WELL?	-.706	.477	-.205	-1.479	.145

Figure 3.4 is based on

Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	4.210	.961		4.381	.000
w= lnDistanceFromRoad	-.091	.180	-.071	-.506	.615
x= BEDROCK OR SHALLOW WELL?	-.697	.481	-.202	-1.449	.154

Figure 3.3

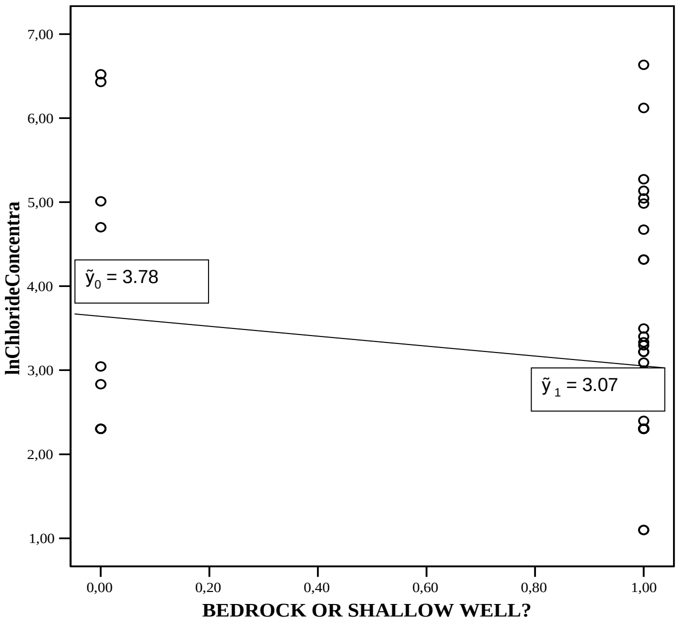
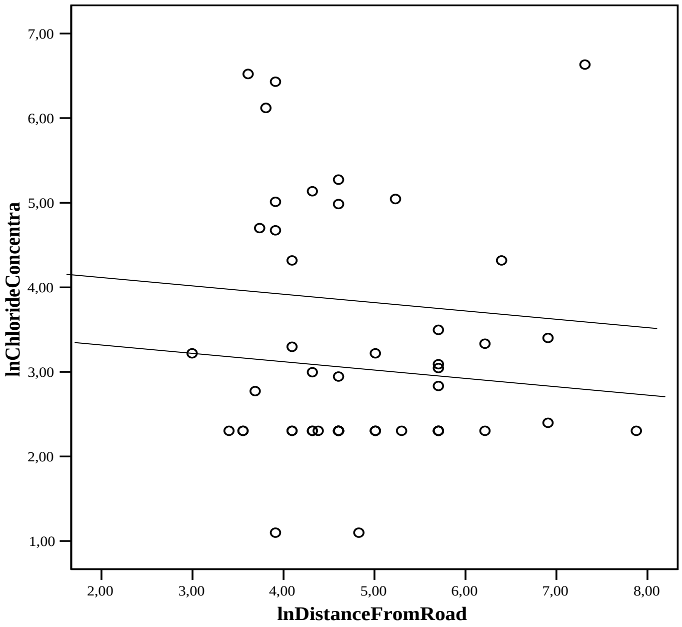


Figure 3.4



Figures 3.5 and 3.6 (Hamilton p89-91)

Figure 3.5 is based on

Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	3.666	.905		4.050	.000
w= lnDistanceFromRoad	-.029	.202	-.022	-.144	.886
x*w= lnDroadDeep	-.081	.099	-.128	-.819	.417

Figure 3.6 is based on

Also see Table 3.4 in Hamilton p90 Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	9.073	1.879		4.828	.000
w= lnDistanceFromRoad	-1.109	.384	-.862	-2.886	.006
x= BEDROCK OR SHALLOW WELL?	-6.717	2.095	-1.948	-3.207	.002
x*w= lnDroadDeep	1.256	.427	1.979	2.942	.005

Figure 3.5

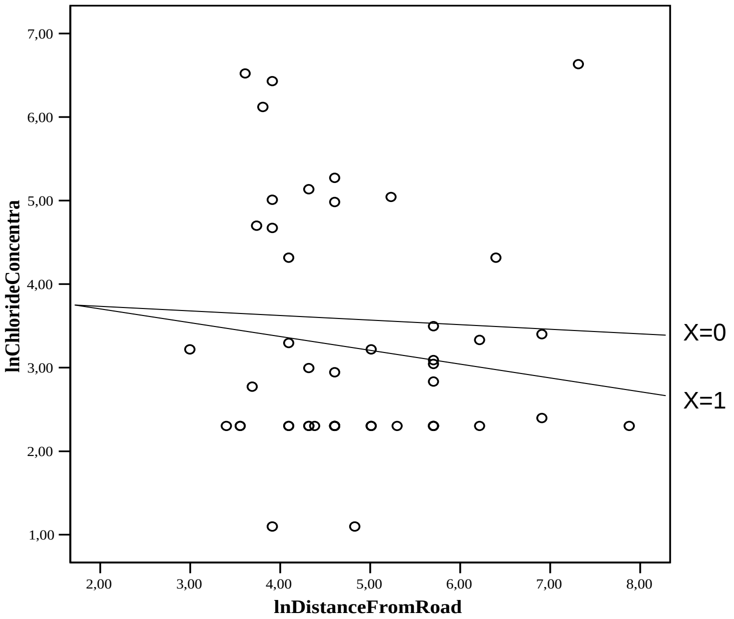
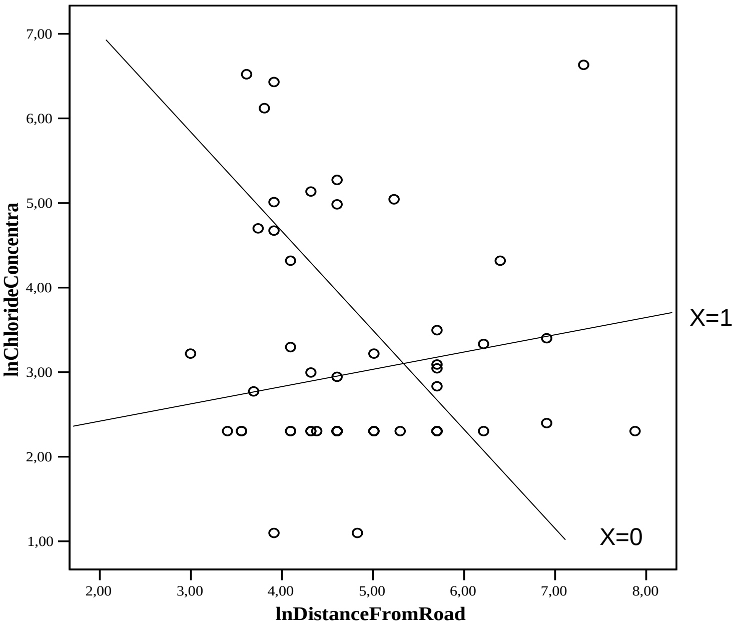


Figure 3.6



Multicollinearity

- Taking all three variables, x, w, and x*w will introduce an element of multicollinearity. This means that we cannot trust tests of single coefficients
- But as shown in the previous example one can not drop any one of the variables without dropping a relevant variable
- F-test of e.g. w and z*w simultaneously circumvents the test problem, and with some experimentation with different models one may see if excluding w or x*w changes the relations substantially

Nominal scale variable

- Can be included in regression models by the use of new auxiliary variables: one for each category of the nominal scale variable. J categories implies H(j), j=1,…,J new auxiliary variables
- If the dependent variable is interval scale and the the only independent variable is nominal scale analysis of variance (ANOVA) is the most common approach to analysis
- By introducing auxiliary variables the same type of analysis can be done in a regression model

Analysis of variance - ANOVA

- Analysing an interval scale dependent variable with one or more nominal scale independent variables, often called factors
  - One way ANOVA uses one nominal scale variable
  - Two way ANOVA uses two nominal scale variable
  - And so on …
- Tests of differences between groups are based on an evaluation of whether the variation within a group (defined by the ”factors”) is large compared to the variation between groups

Nominal scale variables in regression (1)

- If the nominal scale has J categories a maximum of J-1 auxiliary variables can enter the regression

  If H(j), j=1, ... , J-1 are included H(J) have to be excluded
- The excluded auxiliary variable is called the **reference category** and is the most important category in the interpretation of the results from the regression

Nominal scale variables in regression (2)

Dummy coding of a nominal scale variable

- The auxiliary variable H(j) is coded 1 for a person if the person belongs to category j on the nominal scale variable, it is coded 0 if theperson do not belong to category j
- NB: The mean of a dummy coded variable is the proportion in the sample with value 1 (i.e. that belongs in the category)

Nominal scale variables in regression (3)

The reference category

(the excluded auxiliary variable)

- The chosen reference category ought to be large and clearly defined
- The estimated effect of an included auxiliary variable measures the effect of being in the included category relative to being in the reference category

Nominal scale variables in regression (4)

- This means that the regression parameter for an included dummy coded auxiliary variable tells us about additions or subtractions from the expected Y-value a person gets by being in this category rather than in the reference category

Nominal scale variables in regression (5)

Testing I

- Testing if a regression coefficient for an included auxiliary variable equals 0 answers the question whether the persons in this group have a mean Y value different from the mean value of the persons in the reference category

Nominal scale variables in regression (6)

Testing II

- Testing whether a Nominal scale variable contributes significantly to a regression model have to be done by testing if all auxiliary variables in sum contributes significantly to the regression
- For this we use the F-test, applying formula 3.28 in Hamilton (p80)

Nominal scale variables in regression (7)

Interaction

- When dummy coded nominal scale variables are entered into an interaction all included auxiliary variables have to be multiplied with the variable suspected of interacting with it

On terminology (1)

- **Dummy coding** of nominal scale variables are called different names in different textbooks. For example it is
  1. Dummy coding in Hamilton, Hardy, and Weisberg
  2. Indicator coding in Menard (and also Weisberg)
  3. Reference coding or partial method in Hosmer&Lemeshow

On terminology (2)

- To reproduce results from the analysis of variance (ANOVA) by means of regression techniques Hamilton introduces a coding of the auxiliary variables he calls effect coding. Other authors call it differently:
  - It is called effect coding by Hardy
  - It is called deviance coding by Menard
  - It is called the marginal method or deviance method by Hosmer&Lemeshow
- To highlight particular group comparisons Hardy (Ch5) introduces a coding scheme called contrast coding

Ordinal scale variables

- Can be included as an interval scale if the unobserved theoretical dimension is continuous and distance mesures seems resonable
- Also it may be used directly as dependent variable if the program allows ordinal dependent variables
  - In that case parameters are estimated for every level above the lowest as cumulative effects relative to the lowest level

Nominal scale variables

TYPE OF GROUP	Frequency	Percent	Valid Percent	Cumulative Percent
POL	48	12.6	12.6	12.6
FARMER	132	34.7	34.7	47.4
O. PEOPLE	200	52.6	52.6	100.0
Total	**380**	**100.0**	**100.0**	

Example of dummy coding

Nominal scale			Auxiliary	variables	H (*)	
Type of group	Code	N	H(1)= Pol	H(2)= Farmer	H(3)= People	
Politicians	1	48	1	0	0	
Farmers	2	132	0	1	0	
Other People	3	200	0	0	1	**Reference category**

A variable with 3 categories leads to 2 dummy coded variables in a regression with the third used as reference

Example of effect coding

Nominal scala			Auxiliary variable			
Type of group	Code	N	H(1)= Pol	H(2)= Farmer		
Politicians	1	48	1	0		
Farmers	2	132	0	1		
Other People	3	200	-1	-1		Reference category

In effect coding the reference category is coded -1. Effect coding make if possible to duplicate all F-tests of ordinary ANOVA analyses.

Contrast coding

- Is used to present just those comparisons that are of the highest theoretical interest
- Contrast coding requires
  - That with J categories there have to be J-1 contrasts
  - The values of the codes on each auxiliary variable have to sum to 0
  - The values of the codes on any two auxiliary variables have to be orthogonal (their vector product has to be 0)

Use of dummy coded variables(1)

Dependent Variable: I. of political contr. of sales of agric. est.	B	Std. Error	Beta	t	Sig.
(Constant)	4.106	.152		26.991	.000
Pol	.914	.337	.147	2.711	.007
Farmer	.421	.240	.096	1.758	.080

- The constant shows the mean of the dependent variable for those who belong to the reference category
- The mean of the dependent variable for politicians are 0.91 opinion score points above the mean of the reference category
- The mean on the dependent variable for farmers are 0.42 opinion score points above the mean of the reference category

Use of dummy coded variables (2)

Dependent Variable: I. of political control of sales of agricultural estates	B	Std. Error	t	Sig.
(Constant)	4.264	.186	22.954	.000
Number of dekar land Owned	.000	.000	2.176	.030
Pol	.566	.382	1.482	.139
Farmer	-.309	.338	-.913	.362

Compare this table with the previous. What has changed?

How do we interpret the coefficient on ”Pol” and ”Farmer”?

Recall:

Multiple regression: modell

Let K = number of parameters in the model

(then K-1 = number of variables)

Population model

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1} + \varepsilon_i$

  i = 1, ... ,N; where N = number of case in the population

Sample model

- $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{K-1} x_{i,K-1} + e_i$

  i = 1, ... ,n; where n = number of case in the sample

Conclusions (1)

- Linear regression can easily be extended to use 2 or more explanatory variables
- If the assumptions of the regression is satisfied (that the error terms are normally distributed with independent and identically distributed errors – normal I.i.d. errors) the regression will be a versatile and strong tool for analytical studies of the connection between a dependent and one or more independent variables

Conclusions (2)

- The most common method of estimating coefficients for a regression model is called OLS (ordinary least squares)
- Coefficients computed based on a sample are seen as estimates of the population coefficient
- Using the t-test we can judge how good such coefficient estimates are
- Using the F-test we may evaluate several coefficient estimates in one test

Conclusions (3)

- Dummy variables are useful in several ways
  - A single dummy coded x-variable will give a test of the difference in means for two groups (0 and 1 groups)
  - Nominal scale variables with more than 2 categories can be recoded by means of dummy coding and included in regression anlysis
  - By using effect coding we can perform analysis of variance of the ANOVA type

SOS3003
Applied data analysis for social science

Lecture notes on

Hamilton Ch 8 p249-282 Factor analysis

and

A Low-Tech Guide to Causal Modelling.

Erling Berge
Department of sociology and political science
NTNU

Literature

- Hamilton, Lawrence C. 2008. A Low-Tech Guide to Causal Modelling. http://pubpages.unh.edu/~lch/causal2.pdf
- Principal components and factor analysis
  - Hamilton Ch 8 p249-282
- Also see:

  Winship, Chrisopher, and Stephen L. Morgan 1999 “The Estimation of Causal Effects from Observational Data”, Annual Review of Sociology Vol 25: 659-707

Causal analysis

- Experiment
  - Randomized causal impacts (”treatment”) provide precise causal conclusions about effects (”response”) if there is significant differences in means
  - This can be impossible to achieve due to
    - Practical conditions
    - Economic constraints
    - Ethical judgements
- Instead one tries to obtain quasi-experiments
  - Using for example regression analysis

Model of causal effects Ref.:

- Research using observations utilize concepts from experimental design
  - “Treatment”, “Stimulus”
  - “Effect”, “Outcome”

Ref.:

Winship, Chrisopher, and Stephen L. Morgan 1999 “The Estimation of Causal Effects from Observational Data”, Annual Review of Sociology Vol 25: 659-707

Experiments allocate ”cases” randomly to one of two groups:

- **TREATMENT (T)**

  with observation
  - before treatment
  - after treatment

- **CONTROLL (C)**

  with observation
  - before non-treatment
  - after non-treatment

The counterfactual hypothesis for the study of causality

- Individual “i” can a priori be assumed selected for one of two groups
  - Treatment group, T, or control group, C.
- Treatment, t, as well as non-treatment, c, can a priori be given to individuals both in the T- and C-group
- In reality we are able to observe t only in the T-group and c in the C-group

Modelling of causal effects:

The counterfactual hypothesis (1)

- There are for each individual ”i” four possible outcomes
  - **$\mathbf{Y_i}$(c,C)** or Y_i(t,C); if allocated to a control group
  - Y_i(c,T) or **$\mathbf{Y_i}$(t,T)** ; if allocated to a treatment group

  - Only **$\mathbf{Y_i}$(c, given that ”i” is a member of C)** or
  - **$\mathbf{Y_i}$(t, given that ”i” is a member of T)** can be observed for any particular individual

Modelling of causal effects:

The counterfactual hypothesis (2)

More formally one may write the possible outcomes for person no i:

	Treatment: t	Non-treat.: c
T-group	$\mathbf{Y^t_i \in T}$	$Y^c_i \in T$
C-group	$Y^t_i \in C$	$\mathbf{Y^c_i \in C}$

Modelling of causal effects:

The counterfactual hypothesis (3)

- Then the causal effect for individual i is
- $\delta_i = Y_i\ (t) - Y_i\ (c)$
- Only one of these two quantities can be observed for any given individual
- This leads to the “counterfactual hypothesis”

The counterfactual hypothesis: concluding

- “The main value of this counterfactual framework is that causal inference can be summarized by a single question: Given that the δ_i cannot be calculated for any individual and therefore that Y^t_i and Y^c_i can be observed only on mutually exclusive subsets of the population, what can be inferred about the distribution of the δ_i from an analysis of Y_i and T_i ?” (Winship and Morgan 1999:664)

Modelling of causal effects: from individual effects to population averages

- We can observe
  Y_i (c |i∈C), but not Y_i (t |i∈C)
- The problem may be called a problem of missing data
- Instead of individual effects we can estimate average effects for the total population

Modelling of causal effects (1)

- Average effects can be estimated, but usually it involves difficulties
- One assumption is that the effect of the treatment will be the same for any given individual independent of which group the individual is allocated to
- This, however, is not self-evident

Modelling of causal effects (2)

The counterfactual hypothesis assumes:

- That changing the treatment group for one individual do not affect the outcome of other individuals (no interaction)
- That treatment in reality can be manipulated (e.g. sex can not be manipulated)

Modelling of causal effects (3)

- One problem is that in a sample the process of allocating person no i to a control or treatment group may affect the estimated average effect (the problem of selection)
- In some cases, however, the interesting quantity is the average effect for those who actually receive the treatment

Modelling of causal effects (4)

- It can be shown that there are two sources of bias for the estimates of the average effect

1. An established difference between the C- and T- groups
2. The treatment works in principle differently for those allocated to the T-group compared to those in the C-group
   - To counteract this one has to develop models of how people get into C- and T-groups (selection models)

Modelling of causal effects (5)

- A general class of methods that may be used to estimate causal effects are the regression models
- These are able to “control for” observable differences between the C- and T- groups, but not for unequal response to treatment

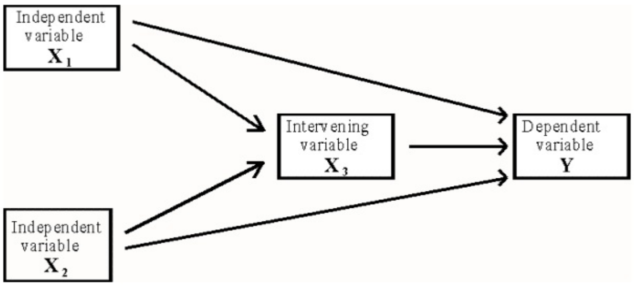
Causal modelling

- “path analysis” or “structural equations modelling” go back to the 60ies
- Jöerskog and Sörbom: LISREL
  - Use maximum likelihood to estimate model parameters maximising fit to the variance-covariance matrix
  - Commonly available in statistical packages
    - Covariance structural modelling
    - Structural equation modelling
    - Full information maximum likelihood estimation

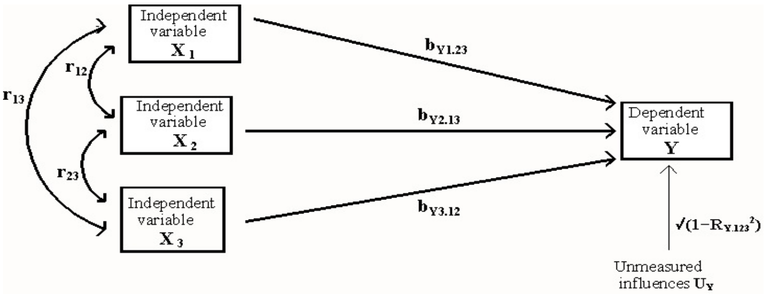
Low-Tech approach

- Uses OLS to do simple versions of the structural equations models
- The key assumption is the causal ordering of variables. In survey data this ordering is supplied by theory
- The causal diagram visualize the order of causation:
  - Causality flows from left to right
  - Intervening variables give rise to indirect effects
  - “reverse causation” creates problems

Low-Tech causal modelling
Figure 1


Independent variable X_1
Intervening variable X_3
Dependent variable Y
Independent variable X_2

Multiple regression as a causal model
Figure 2


Independent variable X_1
r_{12}
r_{13}
Independent variable X_2
r_{23}
Independent variable X_3
$b_{Y1.23}$
$b_{Y2.13}$
$b_{Y3.12}$
Dependent variable Y
$\sqrt{(1-R_{Y.123}^2)}$
Unmeasured influences U_Y

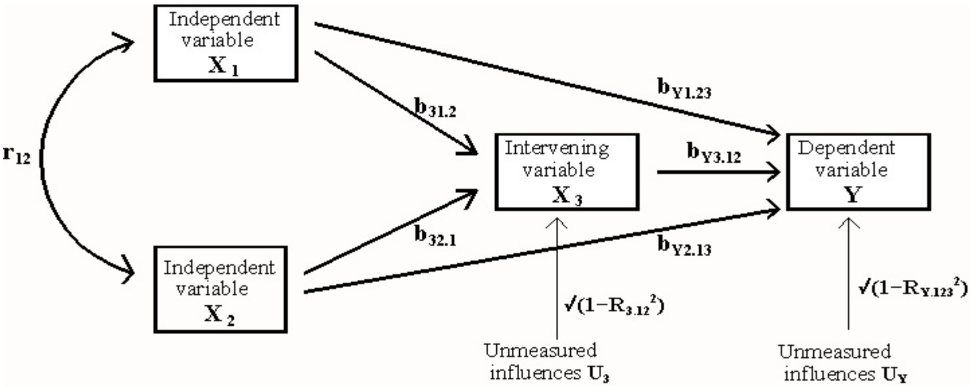
Quantities in the diagram

r_{12}, r_{13}, r_{23}	Pearson correlations among x-variables
$b_{Y1.23}$, etc.	Usually a standardized regression coefficient (“beta weight”) taken from the regression of Y on X_1, and “.” means controlled for X_2, X_3
$R_{Y.123}^2$	Coefficient of determination R^2 from the regression of Y on X_1, X_2, X_3
$\sqrt{1-R_{Y.123}^2}$	Is an estimate of unmeasured influences called error term or disturbance

Multiple regression

- All assumptions and all problems apply as before
  - – Note in particular that error terms must be uncorrelated with included x-variables (no relevant variable has been omitted)
- If some of the X-es are intervening in figure 2 the model is too simple, but it matters only if we are interested in causality

Path coefficients
Figure 3


Independent variable X_1
r_{12}
$b_{31.2}$
$b_{Y1.23}$
Intervening variable X_3
$b_{Y3.12}$
Dependent variable Y
$b_{32.1}$
$b_{Y2.13}$
Independent variable X_2
$\sqrt{(1-R_{3.12}^2)}$
$\sqrt{(1-R_{Y.123}^2)}$
Unmeasured influences U_3
Unmeasured influences U_Y

New elements in figure 3

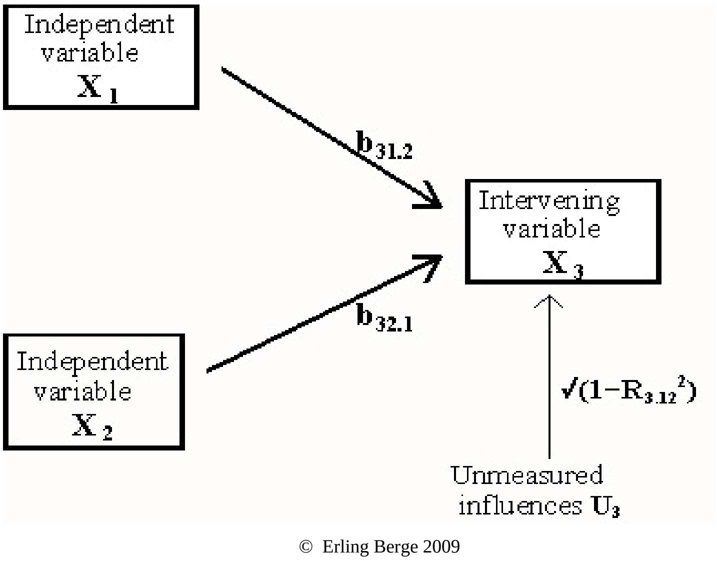
$b_{31.2}$, $b_{32.1}$	Standardized regression coefficients ("beta weight") from the regression of X_3 on X_1 controlled for X_2 and from the regression of X_3 on X_2 controlled for X_1
$R_{3.12}^2$	Coefficient of determination (R^2) from the regression of X_3 on X_1 and X_2
$\surd\{1-R_{3.12}^2\}$	The error term from the regression of X_3 on X_1 and X_2

The structural model of figure 3

- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $\hat{X}_3 = b_{31.2}X_1 + b_{32.1}X_2$

- In structural equations variables and coefficients are standardized
- That means that variables have an average of 0 and a standard deviation of 1 and that coefficients vary between -1 and +1

Figure 5: the regression of X_3 on X_1 and X_2



Direct, Indirect and Total Effects

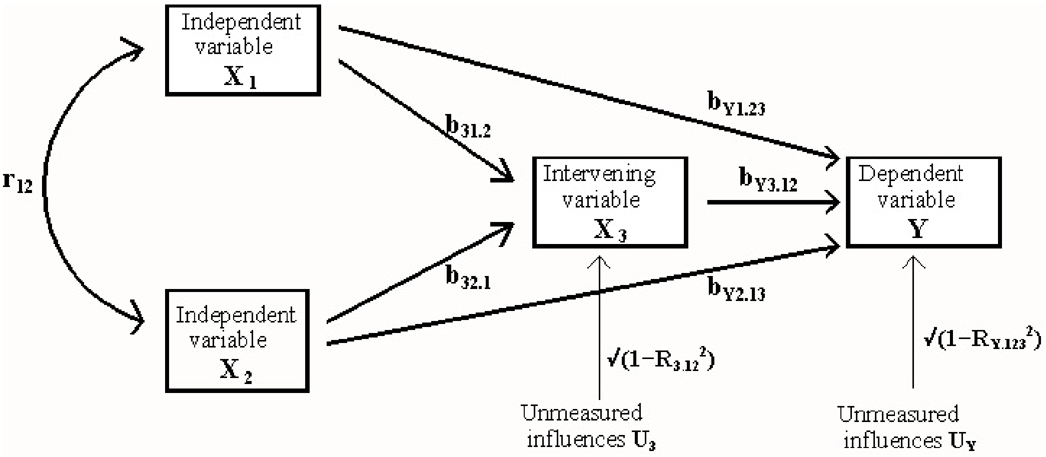
- *Indirect effects* equal the product of coefficients along any series of causal paths that link one variable to another
- *Total effects* equal the sum of all direct and indirect effects linking two variables

Indirect effects as products of path coefficients

- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $\hat{X}_3 = b_{31.2}X_1 + b_{32.1}X_2$
- Means that we have
- $\hat{Y} = b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}X_3$
- $= b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.1\hat{2}}(b_{31.2}X_1 + b_{32.1}X_2)$
- $= b_{Y1.23}X_1 + b_{Y2.13}X_2 + b_{Y3.12}b_{31.2}X_1 + b_{Y3.12}b_{32.1}X_2$
- $= (b_{Y1.23} + b_{Y3.12}b_{31.2})X_1 + (b_{Y2.13} + b_{Y3.12}b_{32.1})X_2$

- Compare compound coefficients to the diagram

Structural model


Independent variable X_1
Independent variable X_2
r_{12}
$b_{31.2}$
$b_{32.1}$
$b_{Y1.23}$
$b_{Y2.13}$
$b_{Y3.12}$
Intervening variable X_3
Dependent variable Y
$\sqrt{(1-R_{3.12}^2)}$
$\sqrt{(1-R_{Y.123}^2)}$
Unmeasured influences U_3
Unmeasured influences U_Y

Path Coefficients

- X_1 to Y: $\mathbf{b_{Y1.23}}$ (regression coefficient of Y on X_1, controlling for X2 and X3)
- X_2 to Y: $\mathbf{b_{Y2.13}}$ (regression coefficient of Y on X_2, controlling for X_1 and X3)
- X_3 to Y: $\mathbf{b_{Y3.12}}$ (regression coefficient of Y on X_3, controlling for X_1 and X_2)
- X_1 to X_3: $\mathbf{b_{31.2}}$ (regression coefficient of X_3 on X_1, controlling for X_2)
- X_2 to X_3: $\mathbf{b_{32.1}}$ (regression coefficient of X_3 on X_2, controlling for X_1)

Direct effects

X_1 to Y: $\mathbf{b_{Y1.23}}$	regression coefficient of Y on X1, controlling for X2 and X3
X_2 to Y: $\mathbf{b_{Y2.13}}$	regression coefficient of Y on X2, controlling for X1 and X3
X_3 to Y: $\mathbf{b_{Y3.12}}$	regression coefficient of Y on X3, controlling for X1 and X2
X_1 to X_3: $\mathbf{b_{31.2}}$	regression coefficient of X3 on X1, controlling for X2
X_2 to X_3: $\mathbf{b_{32.1}}$	regression coefficient of X3 on X2, controlling for X1

Indirect and total effects

Indirect effects	
X_1 to Y, through X_3:	$\mathbf{b_{31.2} \times b_{Y3.12}}$
X_2 to Y, through X_3:	$\mathbf{b_{32.1} \times b_{Y3.12}}$
Total effects	
X_1 to Y:	$\mathbf{b_{Y1.23}} + (\mathbf{b_{31.2} \times b_{Y3.12}})$
X_2 to Y:	$\mathbf{b_{Y2.13}} + (\mathbf{b_{32.1} \times b_{Y3.12}})$

Additions to multiple regressions

- We learn something new if the indirect effects are large enough to have substantial interest
- More than two steps of causation tends to become very weak
  - 0.3*0.3*0.3 = 0.027
  - 0.3 standard deviation change in causal variables leads to a 0.027 standard deviation change in the dependent variable

Variables and measurement

- All interval scale variables used in multiple regression (including non-linear transformed variables and interaction terms) can be included in structural equations models
- But interpretation becomes tricky when variables are complex. Conditional effect plots are very useful
- Robust, quantile, logit, and probit regression should not be used
- Categorical variables should not be used as intervening variables
- Scales or index variables can be used as usual in OLS regression

Concluding on structural equations modelling

- Including factors from factor analysis as explanatory variables make it possible to approximate a LISREL type analysis
- If assumptions are true LISREL will perform a much better and more comprehensive estimation, but too often assumptions are not true then the low-tech approach has access to the large toolkit of OLS regression for diagnostics and exploratory methods testing basic assumptions and discovering unusual data points
- Simple diagnostic work sometimes yields the most unexpected, interesting and replicable findings from our research

Principal components and factor analysis

- Principal components and factor analysis are both methods for data reduction
- They seek underlying dimensions that are able to account for the pattern of variation among a set of observed variables
- Principal components analysis is a transformation of the observed data where the idea is to explain as much as possible of the observed variation with a minimum number of components

Factor analysis

- Estimates coefficients on - and values of - unobserved variables (Factors) to explain the co-variation among an observed set of variables
- The assumption is that a small set of the unobserved factors are able to explain most of the co-variation
- Hence factor analysis can be used for data reduction. Many variables can be replaced by a few factors

Factor analysis

- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kJ}F_J + u_k$
  - $k = 1, 2, 3, \dots, K$
- Symbols
  - K observed variables, Z_k ; $k=1, 2, 3, \dots, K$
  - J unobserved factors, F_j ; $j=1, 2, 3, \dots, J$ where $J<K$
  - For each variable there is a unique error term, u_k, also called unique factors while the F factors are called common factors
  - For each factor there is a **standardized** regression coefficient, ℓ_{kj}, also called factor loading; k refers to variable no, j refers to factor no. An index denoting case no has been omitted here.

Correlation of factors

- Factors my be correlated or uncorrelated
  - Uncorrelated: they are then called **orthogonal**
  - Correlated: they are then called **oblique**
- Factors may be rotated
  - Oblique rotations create correlated factors
  - Orthogonal rotations create uncorrelated factors

Principal components

- Represents a simple transformation of variables. There are as many principal components as there are variables
- Principal components are uncorrelated
- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kK}F_K$
- If the last few principal components explain little variation we can retain J<K components. Thus Principal Components also can be used to reduce data.
- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kJ}F_J + v_k$

  where J<K and

  the residual v_k has small variance and consist of the discarded principal components

Principal components vs factor analysis

- Principal components analysis attempts to explain the observed variation of the variables
- Factor analysis attempts to explain their intercorrelations
- Use principal components to generate a composite variable that reproduce the maximum variance of observed variables
- Use factor analysis to model relationships between observed variables and unobserved latent variables and to obtain estimates of latent variable values
- The choice between the two is often blurred, to some degree it is a matter of taste

The number of principal components

- K variables yield K principal components
- If the first few components account for most of the variation, we can concentrate on them and discard the remaining
- The eigenvalues of the standardized correlation matrix provides a guide here
- Components are raked according to eigenvalues
- A principal component with an eigenvalue $\lambda<1$ accounts for less variance than a single variable
- Thus we discard components with eigenvalues below 1
- Another criterion for keeping components is that each component should have substantive meaning

Eigenvalues and explained variance

- In a covariance matrix the sum of eigenvalues equals the sum of variances.
- In a correlation matrix this = K (the number of variables) since each standardized variable has a variance of 1
- Thus the sum of eigenvalues of the principal components
- $\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_K = K$ and
- λ_j / K = proportion of variance explained by component no j

Uniqueness and communality

- If K-J components are discarded and we have only J factors
- $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \ldots + \ell_{kj}F_j + \ldots + \ell_{kJ}F_J + v_k$
- And an error term v_k
- The variance of the error term is called the uniqueness of the variable
- Communality is the proportion of a variable’s variance shared with the components
- Communality = h_k^2 = 1 - Uniqueness = $\Sigma_j \lambda_{kj}^2$, j=1,…, J ; k = variable number

Rotation to simple structure

- The idea is to transform (rotate) the factors so that the loadings on each components make it easier to interpret the meaning of the component
- If the loading are close either to 1 or -1 on one factor and close to 0 on all others the structure is simpler to interpret: we rotate to “simple structure”. The rotated factors fit data equally well but are simpler to interpret
- Rotations may be
  - Orthogonal (method typically: varimax)
  - Oblique (method typically: oblimin, promax)

Why rotate?

- Underlying unobserved dimensions may in theory be seen as correlated
- Allowing correlated factors may provide even simpler structure than uncorrelated factors, thus easier to interpret
- All rotations fit data equally well
- Hence the one chosen depends on a series of choices done by the analyst
- Try different methods to see if results differ

SPSS output

- For rotated factor solutions with correlated factors SPSS provides two matrixes for interpretation
- The pattern matrix provides the direct regression of the variables on the factors. The coefficients tells about the direct contribution of a factor in explaining the variance of a variable. Due to the correlations of the factors there are also indirect contributions
- The structure matrix provides the correlations between the variables and the factors

Factor scores

- Both principal components and factor analysis may be used to compute composite scores called factor scores
- Recall that variables and factors are assumed to be related like
  - $Z_k = \ell_{k1}F_1 + \ell_{k2}F_2 + \dots + \ell_{kj}F_j + \dots + \ell_{kK}F_K$
- Then it is possible to find values c_{ij} making
  - $\hat{F}_j = c_{1j}Z_1 + c_{2j}Z_2 + \dots + c_{kj}Z_j + \dots + c_{Kj}Z_K$
- The coefficients c_{ij} are the factor score coefficients. They come from the regression of the factor F_j on the variables

Methods for extracting factors

- Principal factor analysis
  - The original correlation matrix **R** is replaced by **R*** where the original 1-values of the diagonal has been replaced by estimates of the communality (the shared variance)
  - The factors extracted tries to explain the co-variance or correlations among the variables.
  - The unexplained variance is attributed to a unique factor (error term). The uniqueness may reflect measurement error or something that this variable measure that no other variable measure
  - The most common estimate of communality is R_k^2 the coefficient of determination from the regression of Z_k on all other variables

How many factor should we retain?

- In principal component analysis factors with eigenvalues above 1 is recommended
- In principal factor analysis factors with eigenvalues above 0 is recommended
- Procedure:
  - Extract initial factors or components
  - Rotate to simple structure
  - Decide on how many factors to retain
  - Obtain and use scores for the retained factors, ignoring discarded factors

Concluding (1)

- Principal components
  - transformation of the data, not model based. Appropriate if goal is to compactly express most of the variance of k variables. Minor components (perhaps all except the first) may be discarded and viewed as a residual.
- Factor analysis
  - Estimates parameters of a measurement model with latent (unobserved) variables.

Concluding (2)

- Types of factor analysis
  - Principal factoring – principal components of a modified correlation matrix R* in which communality estimates (R_k^2) replace one's on the main diagonal
    - Principal factoring without iteration
    - Principal factoring with iteration
  - Maximum likelihood estimation – significance tests regarding number of factors and other hypotheses, **assuming multivariate normality**

Concluding (3)

- Rotation
  - If we retain more than one factor rotation simplifies structure and improves interpretability
    - Orthogonal rotation (varimax) maximum polarization given uncorrelated factors
    - Oblique rotation (oblimin, promax) further polarization by permitting interfactor correlations. The results may be more interpretable and more realistic than uncorrelated factors
- Scores
  - Factor scores can be calculated for use in graphs and further analysis, based on rotated or unrotated factors and principal components

Concluding (4)

- Factor analysis is based on correlations and hence as affected by non-linearities and influential cases as in regression
  - Use scatter plots to check for outliers and non-linearities
  - In maximum likelihood estimation this becomes even more important since it assumes multivariate normality making it even less robust than principal factors

Principal components of trust in Malawi

- Survey of 283 households in 18 villages in Malawi, 2007
- There are 8 related questions asked in one group
- Are there 1, 2 or more underlying dimensions shaping the attitudes expressed?
- The questions:

M3 Would you say you trust all, most, some or just a few people in the following groups? (All=1 – None=5)

a	Your family members	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
b	Your relatives	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
c	Your village	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
d	People from outside the village	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
e	People of same ethnic group	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
f	People from outside ethnic group	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
g	People from same church/mosque	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*
h	People *not* from same church/mosque	*All*	*Most*	*Some*	*Only a few*	*None*	*Do not know*

Descriptive Statistics

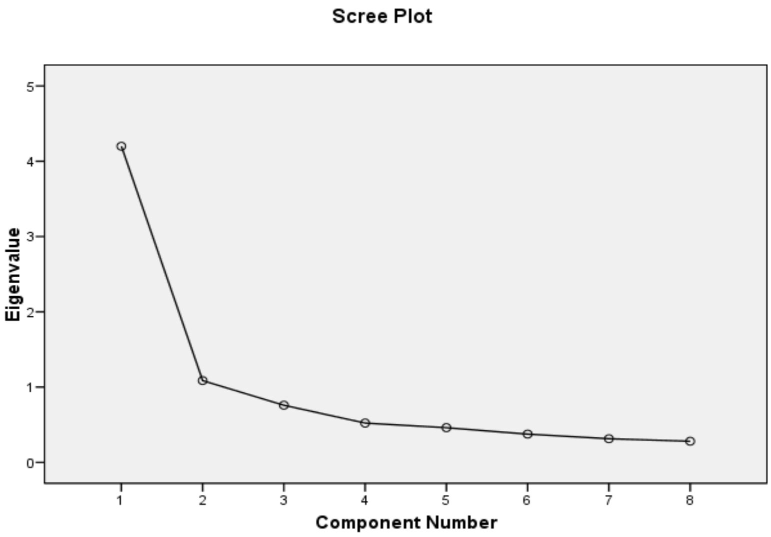
	Mean	Std. Deviation	Analysis N
M3.a. Trust in family members	1.60	.935	266
M3.b. Trust in relatives	2.12	1.136	266
M3.c. Trust in people in own village	2.69	1.090	266
M3.d. Trust in people outside the village	3.28	1.118	266
M3.e. Trust in people of same ethnic group	2.90	1.082	266
M3.f. Trust in people outside ethnic group	3.26	1.098	266
M3.g. Trust in people from same church/mosque	2.39	1.062	266
M3.h. Trust in people not from same church/mosque	3.02	1.197	266

Trust in Malawi: correlation of variables

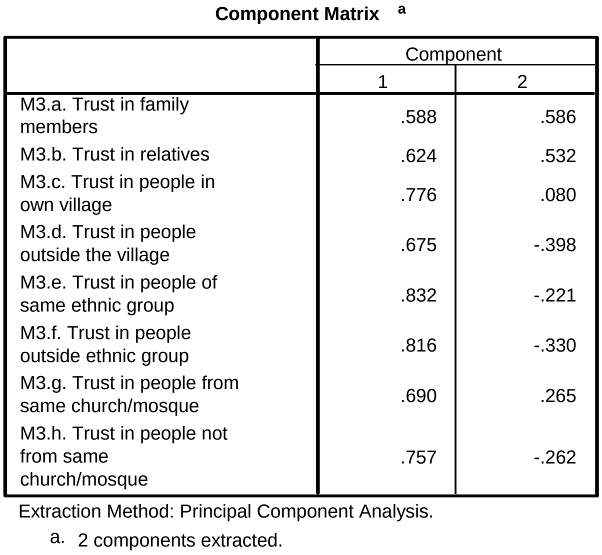
Correlation Matrix

	M3.a. Trust in family members	M3.b. Trust in relatives	M3.c. Trust in people in own village	M3.d. Trust in people outside the village	M3.e. Trust in people of same ethnic group	M3.f. Trust ir people outside ethnic group	M3.g. Trust in people from same church/mosque	M3.h. Trust in people not from same church/mosque
M3.a. Trust in family members	1.000	.500	.416	.236	.370	.316	.422	.305
M3.b. Trust in relatives	.500	1.000	.496	.315	.363	.353	.424	.292
M3.c. Trust in people in own village	.416	.496	1.000	.482	.588	.573	.465	.430
M3.d. Trust in people outside the village	.236	.315	.482	1.000	.526	.610	.233	.469
M3.e. Trust in people of same ethnic group	.370	.363	.588	.526	1.000	.702	.504	.643
M3.f. Trust in people outside ethnic group	.316	.353	.573	.610	.702	1.000	.430	.618
M3.g. Trust in people fror same church/mosque	.422	.424	.465	.233	.504	.430	1.000	.536
M3.h. Trust in people not from same church/mosque	.305	.292	.430	.469	.643	.618	.536	1.000

Trust in Malawi: number of factors



Trust in Malawi: factor/ component matrix



Component Matrix [a]

	Component	
	1	2
M3.a. Trust in family members	.588	.586
M3.b. Trust in relatives	.624	.532
M3.c. Trust in people in own village	.776	.080
M3.d. Trust in people outside the village	.675	-.398
M3.e. Trust in people of same ethnic group	.832	-.221
M3.f. Trust in people outside ethnic group	.816	-.330
M3.g. Trust in people from same church/mosque	.690	.265
M3.h. Trust in people not from same church/mosque	.757	-.262

Extraction Method: Principal Component Analysis.

[a] 2 components extracted.

Trust in Malawi: orthogonal factors

Rotated component matrix	Unrotated components		Orthogonal varimax	
Variables	**F1**	**F2**	**F1**	**F2**
M3.a. Trust in family members	.588	.586	.117	.821
M3.b. Trust in relatives	.624	.532	.178	.800
M3.c. Trust in people in own village	.776	.080	.572	.531
M3.d. Trust in people outside the village	.675	-.398	.779	.089
M3.e. Trust in people of same ethnic group	.832	-.221	.798	.324
M3.f. Trust in people outside ethnic group	.816	-.330	.850	.228
M3.g. Trust in people from same church/mosque	.690	.265	.391	.627
M3.h. Trust in people not from same church/mosque	.757	-.262	.762	.246

Trust in Malawi: communalities

Communalities

	Extraction
M3.a. Trust in family members	.689
M3.b. Trust in relatives	.671
M3.c. Trust in people in own village	.609
M3.d. Trust in people outside the village	.614
M3.e. Trust in people of same ethnic group	.741
M3.f. Trust in people outside ethnic group	.774
M3.g. Trust in people from same church/mosque	.546
M3.h. Trust in people not from same church/mosque	.641

Extraction Method: Principal Component Analysis.

Trust in Malawi: explained variance

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.199	52.487	52.487	3.071	38.387	38.387
2	1.087	13.582	66.069	2.215	27.681	66.069

Extraction Method: Principal Component Analysis.

Trust in Malawi: oblique factors pattern matrix

Rotated component matrix	**varimax (orthogonal)**		**oblimin**		**promax**	
Variables	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
M3.a. Trust in family members	.117	**.821**	-.087	**.868**	-.145	**.901**
M3.b. Trust in relatives	.178	**.800**	-.014	**.826**	-.067	**.855**
M3.c. Trust in people in own village	**.572**	**.531**	**.493**	**.414**	**.476**	**.409**
M3.d. Trust in people outside the village	**.779**	.089	**.838**	-.133	**.864**	-.170
M3.e. Trust in people of same ethnic group	**.798**	.324	**.797**	.120	**.806**	.093
M3.f. Trust in people outside ethnic group	**.850**	.228	**.881**	-.001	**.899**	-.036
M3.g. Trust in people from same church/mosque	.391	**.627**	.268	**.573**	.237	**.582**
M3.h. Trust in people not from same church/mosque	**.762**	.246	**.779**	.045	**.792**	.016

Trust in Malawi: oblique factors structure matrix

Rotated component matrix	**varimax**		**oblimin**		**promax**	
Variables	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
M3.a. Trust in family members	.117	.821	.327	.826	.351	.821
M3.b. Trust in relatives	.178	.800	.380	.819	.403	.817
M3.c. Trust in people in own village	.572	.531	.690	.649	.702	.671
M3.d. Trust in people outside the village	.779	.089	.775	.267	.771	.306
M3.e. Trust in people of same ethnic group	.798	.324	.854	.500	.857	.537
M3.f. Trust in people outside ethnic group	.850	.228	.880	.419	.880	.460
M3.g. Trust in people from same church/mosque	.391	.627	.541	.700	.557	.712
M3.h. Trust in people not from same church/mosque	.762	.246	.800	.416	.801	.452

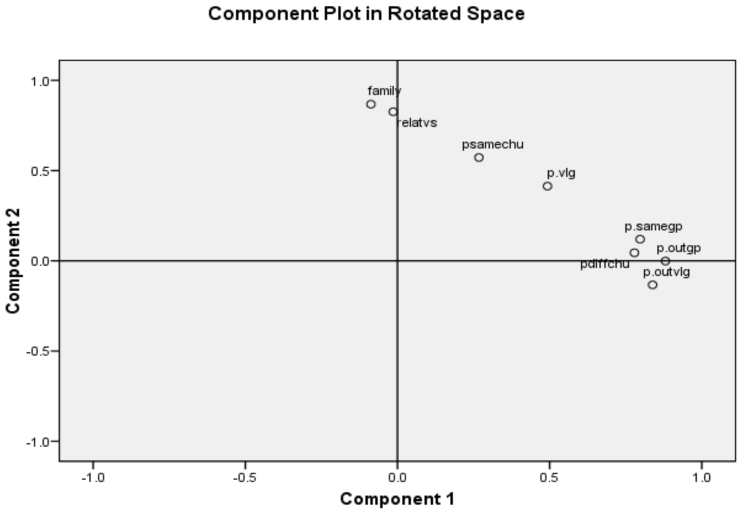
Trust in Malawi: correlation of components

Component Correlation Matrix

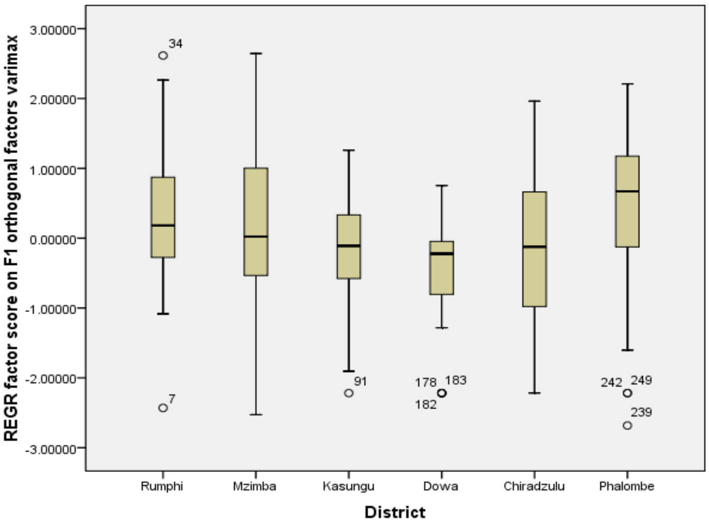
Component	1	2
1	1.000	.477
2	.477	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

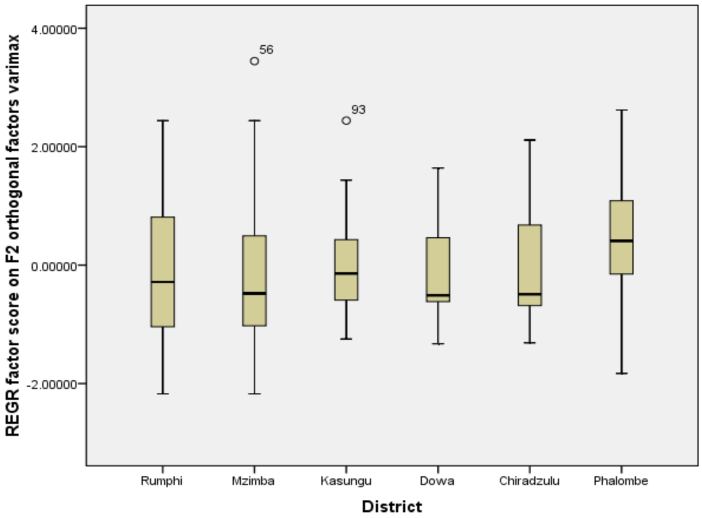
Trust in Malawi: variables in component plot



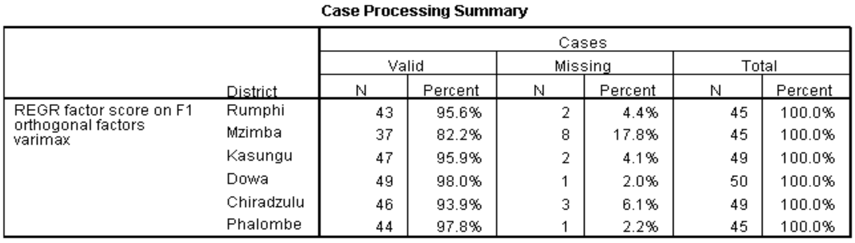
Trust in Malawi: Orthogonal Factor 1 by district



Trust in Malawi: Orthogonal Factor 2 by district

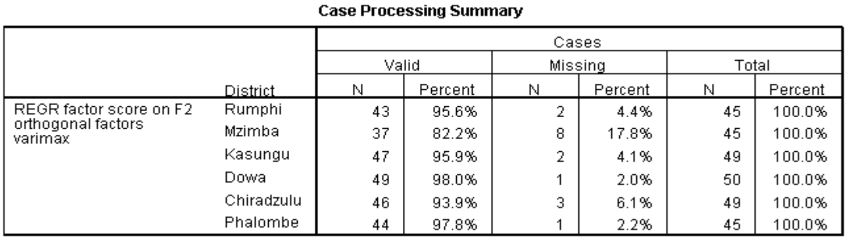


Trust in Malawi: Orthogonal factors by district



Case Processing Summary

		Cases					
		Valid		Missing		Total	
	District	N	Percent	N	Percent	N	Percent
REGR factor score on F1 orthogonal factors varimax	Rumphi	43	95.6%	2	4.4%	45	100.0%
	Mzimba	37	82.2%	8	17.8%	45	100.0%
	Kasungu	47	95.9%	2	4.1%	49	100.0%
	Dowa	49	98.0%	1	2.0%	50	100.0%
	Chiradzulu	46	93.9%	3	6.1%	49	100.0%
	Phalombe	44	97.8%	1	2.2%	45	100.0%



Case Processing Summary

		Cases					
		Valid		Missing		Total	
	District	N	Percent	N	Percent	N	Percent
REGR factor score on F2 orthogonal factors varimax	Rumphi	43	95.6%	2	4.4%	45	100.0%
	Mzimba	37	82.2%	8	17.8%	45	100.0%
	Kasungu	47	95.9%	2	4.1%	49	100.0%
	Dowa	49	98.0%	1	2.0%	50	100.0%
	Chiradzulu	46	93.9%	3	6.1%	49	100.0%
	Phalombe	44	97.8%	1	2.2%	45	100.0%

SOS3003
Applied data analysis for social science
Lecture notes on
Allison: Missing data

Erling Berge
Department of sociology and political science
NTNU

Literature

- Allison, Paul D 2002 “Missing Data”, Sage University Paper: QASS 136, London, Sage,

There is a missing case in the sample

- If one person
  - Refuses to answer
  - Are not at home
  - Has moved away
  - Etc.
- The problem of missing data belong to the study of biased samples. In general biased samples is a more severe problem than the fact that we are missing answers for a few variables on some cases (see Breen 1996 ”Regression Models: Censored, Sample Selected, or Truncated Data”, QASS Paper 111, London, Sage)
- But the problems are related

There are missing answers for a few variables if

- Persons refuse to anser certain questions
- Persons forget or do n ot notice some question or the interviewer does it
- Persons do not know any answer to the question
- The question is irrelevant
- In administrative registers some documents may have been lost
- In research designs where variables with measurement problems may have been measured only for a minority of the sample

Missing data entail problems

- There are practical problems due to the fact that all statistical procedures assumes complete data matrices
- It is an analytical problem since missing data as a rule produce biased parameter estimates
- It is important to distinguish between data missing for random causes and those missing from systematic causes

The simple solution: remove all cases with missing data

- Listwise/ casewise removal of missing data means to remove all cases missing data on one or more variables included in the model
- The method has good properties, but may in some cases remove most of the cases in the sample
- Alternatives like pairwise removal or replacement with average variable value has proved not to have good properties
- More recently developed methods like ”maximum likelihood” and ”multiple imputation” have better properties but are more demanding
- In general it pays to do good work in the data collection stage

Types of randomly missing

- MCAR: missing completely at random
  - Means that missing data for one person on the variable y is uncorrelated with the value on y and with the value on any other variable in the data set (however, internal case by case the value of missing may of course correlate with the value missing on other variables)
- MAR: missing at random
  - Means that missing data for person i on the variable y do not correlate with the value on y if one control for the variation of other variables in the model
  - More formally:

$\text{Pr}(Y_i = \text{missing l } Y_i,X_i) = \text{Pr}(Y_i = \text{missing l } X_i)$

Process resulting in missing

- Is ignorable if
  - The result is MAR and the parameters governing the process are unrelated to the parameters that are to be estimated
- Is non-ignorable if
  - The result is not MAR. Estimation of the model will then require a separate model of the missing process
  - See Breen 1996 ”Regression Models: Censored, Sample Selected, or Truncated Data”, QASS Paper 111, London, Sage
- Here the situation with MAR will be discussed

Conventional methods

Common methods in cases with MAR data:

- Listwise deletion
- Pairwise deletion
- Dummy variable correction
- Imputation (guessing a value for the missing)

Of the commonly available methods

listwise deletion is the best

Listwise deletion (1)

- Can always be used
- If data are MCAR we have a simple random subsample of the original sample
- Smaller n entails large variance estimates
- In the case of MAR data and the missing values on an x-variable are independent of the value on y, listwise deletion will produce unbiased estimates

Listwise deletion (2)

- In logistic regression listwise deletion may cause problems only if missing is related both to dependent and independent variables
- If missing depends only on the values of the independent variable listwise deletion is better than maximum likelihood and multiple imputation

Pairwise deletion

- Means that all computations are based on all available information seen pairwise for all pairs of variables included in the anlysis
- One consequence is that different parameters will be estimated on different samples (we see variation in n from statistic to statistic)
- Then all variance estimates are biased
- Common test statistics provides biased estimates (e.g. t-values and F-values)
- DO NOT USE PAIRWISE DELETION !!

Dummy variable correction

If data is missing for the independent variable x

- Let $x^*_i = x_i$ if x_i is not missing and
  $x^*_i = c$ (an arbitrary constant) if x_i is missing
- Define D_i=1 if x_i is missing, 0 otherwise
- Use x^*_i and D_i in the regression instead of x_i
- In nominal scale variables missing can get its own dummy

  Investigations reveal that even if we have MCAR data parameter estimates will be biased

Do not use dummy variable correction!

Imputation

- The goal is to replace missing values with reasonable guesses about what the value might have been before one do an analysis as if this were real values; e.g.
  - Average of valid values
  - Regression estimates based on many variables and case with valid observations
- Parameter estimates are consistent, but estimates of variances are biased (consistently to small), and the test statistics are too big
- Avoid if possible the simple form of imputation

Concluding on conventional methods for missing data

- Conventional methods of correcting for missing data make problems of inference worse
- Be careful in the data collection so that the missing data are as few as possible
- Make an effort to collect data that may help in modelling the process resulting in missing
- If data is missing use listwise deletion if not maximum likelihood or multiple imputation is available

New methods for ignorable missing data (MAR data): Maximum Likelihood (ML)

- Conclusions
  - Based on the probability for observing just those values found in the sample
  - ML provides optimal parameter estimates in large samples in the case of MAR data
  - But ML require a model for the joint distribution of all variables in the sample that are missing data, and it is difficult to use for many types of models

ML-method: example (1)

- Observing y and x for 200 cases
- 150 distributed as shown
- For 19 cases with Y=1 x is missing and for 31 cases with Y=2 x is missing
- We want to find the probabilities p_{ij} in the population

	Y=1	Y=2
X=1	52	21
X=2	34	43

	Y=1	Y=2
X=1	p_{11}	p_{12}
X=2	p_{21}	p_{22}

ML-method: example (2)

- In a table with I rows and J columns, complete information on all cases and with n_{ij} cases in cell ij the Likelihood is

$$\mathcal{L} = \prod_{i,j} \left(p_{ij} \right)^{n_{ij}}$$

That is the product of all probabilities for every table cell taken to the power of the cell frequency

ML-method: example (3)

For a fourfold table the Likelihood will be

$$\mathcal{L} = (p_{11})^{n_{11}} (p_{12})^{n_{12}} (p_{21})^{n_{21}} (p_{22})^{n_{22}}$$

For the 150 cases in the table above where we have all observations the Likelihood will be

$$\mathcal{L} = (p_{11})^{52} (p_{12})^{21} (p_{21})^{34} (p_{22})^{43}$$

ML-method: example (4)

- For tables the ML estimator is $p_{ij} = n_{ij}/n$
- This provides good estimates for the table where we do not have missing data (listwise deletion)
- How can one use the information about y for the 50 cases where x is missing?
- Since MAR is assumed to be the case the 50 extra cases with known y should follow the marginal distribution of y
- Pr(Y=1) = $(p_{11} + p_{21})$ og Pr(Y=2) = $(p_{12} + p_{22})$

ML-method: example (5)

- Taking into account all that is known about the 200 cases the Likelihood becomes

$$\mathcal{L} = (p_{11})^{52}(p_{12})^{21}(p_{21})^{34}(p_{22})^{43}(p_{11}+p_{21})^{19}(p_{11}+p_{21})^{31}$$

- The ML-estimators will now be

$$\widehat{p}_{ij} = \widehat{p}(x = i \mid \mathrm{y = j})\,\widehat{p}(y = j)$$

ML-method: example (6)

- Taking into account the information we have about cases with missing data, parameter estimates change

Estimate of	Missing deleted	Missing included
p_{11}	0.346	0.317
p_{21}	0.227	0.208
P_{12}	0.140	0.156
p_{22}	0.287	0.319

The ML-method

- For the general case with missing data there are two approaches
  - The EM method, a two stage method where one starts out with the expected value of the missing data and use these to obtain parameter estimates that again will be used to provide better estimates of the missing values and so on …

    (this method provides biased estimates of standard errors)
  - Direct ML estimates are better but can be provided only for linear and log-linear models

New methods for ignorable missing data (MAR data): Multiple Imputation

- Conclusions
  - Is based on a random component added to estimates of the missing data values
  - Has as good properties as the ML method and is easier to implement for all kinds of models
  - But it gives different results every time it is used

Multiple Imputation (1)

- MI have the same optimal properties as the ML method. It can be used on all kinds of data and with all kind of models. In principle it can be done with the ordinary analytical tools
- The use of MI can be rather convoluted. This makes it rather easy to commit errors. And even if it is done correctly one will never have the same result twice due to the random component in the imputed variable value

Multiple Imputation (2)

- Use of data from a simple imputation (with or without a random component) will underestimate the variance of parameters. Conventional techniques are unable to adjust for the fact that data have been generated by imputation
- The best way of doing imputation with a random component is to repeat the process many times and use the observed variation of parameter estimates to adjust the estimates of the parameter variances
- Allison, p.30-31, explaines how this can be done

Multiple Imputation (3)

- MI requires a model that can be used to predict values of missing data. Usually there is an assumption of normally distributed variables and linear relationships. But models can be tailored to each problem
- MI can not handle interactions
- MI model should contain all variables of the anlysis model
- (including the dependent variable)
- MI works only for interval scale variables. If nominal scale variables are used special programs are needed
- Testing of several coefficients in one test is complicated

When data are missing systematically

- Will usually require a model of how the missing cases came about
- ML and MI approaches can still be used, but with much stronger restrictions and the results are very sensitive for deviations from the assumptions

Summary

- If listwise deletion leaves enough data this is the simples solution
- If listwise deletion do not work one should test out multiple imputation
- If there is a suspicion that data are not MAR one needs to create a model of the process creating missing. This can then be used together with ML or MI. Good results require that the model for missing is correct

SOS3003
Applied data analysis for social science
Lecture notes on

Hamilton Ch 4 p109-123

Regression criticism I

Erling Berge
Department of sociology and political science
NTNU

Analyses of models are based on assumptions

- OLS is a simple technique of analysis with very good theoretical properties. But
- The good properties are based on certain assumptions
- If the assumptions do not hold the good properties evaporates
- Investigating the degree to which the assumptions hold is the most important part of the analysis

OLS-REGRESSION: assumptions

- I SPECIFICATION REQUIREMENT
  - The model is correctly specified
- II GAUSS-MARKOV REQUIREMENTS
  - Ensures that the estimates are “BLUE”
- III NORMALLY DISTRIBUTED ERROR TERM
  - Ensures that the tests are valid

I SPECIFICATION REQUIREMENT

- The model is correctly specified if
  - The expected value of y, given the values of the independent variables, is a linear function of the parameters of the x-variables
  - All included x-variables have an impact on the expected y-value
  - No other variable has an impact on expected y-value at the same time as they correlate with included x-variables

II GAUSS-MARKOV REQUIREMENTS (i)

(1) x is known, without stochastic variation

(2) Errors have an expected value of 0 for all i

- $E(\varepsilon_i)=0$ for all i

Given (1) and (2) ε_i will be independent of x_k for all k and OLS provides **unbiased estimates** of β

(unbiased = forventningsrett)

II GAUSS-MARKOV REQUIREMENTS (ii)

(3) Errors have a constant variance for all i

- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i

This is called homoscedasticity

(4) Errors are uncorrelated with each other

- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)=0$ for all $i \neq j$

This is called no autocorrelation

II GAUSS-MARKOV REQUIREMENTS (iii)

Given (3) and (4) in addition to (1) and (2) provides:

- a. Estimates of standard errors of regression coefficients are unbiased and
- b. The **Gauss-Markov theorem**:

OLS estimates have **less variance** than any other linear unbiased estimate (including ML estimates)

OLS gives "BLUE"

(**B**est **L**inear **U**nbiased **E**stimate)

II GAUSS-MARKOV REQUIREMENTS (iv)

(1) - (4) are called the GAUSS-MARKOV requirements

- Given (2) - (4) with an additional requirement that errors are uncorrelated with x-variables:
  - cov (x_{ik}, ε_i)=0 for all i,k

The coefficients and standard errors are consistent (converging in probability to the true population value as sample size increases)

Footnote 1:
Unbiased estimators

- Unbiased means that

  $E[b_k] = \beta_k$
- In the long run we are bound to find the population value - β_k - if we draw sufficiently many samples, calculates b_k and average these

Footnote 2:
Consistent estimators

- An estimator is consistent if we as sample size (n) grows towards infinity, find that b approaches β and s_b approaches σ_β
- [b_k is a consistent estimator of β_k if we for any small value of c have

  $\lim_{n\to\infty}$ [Pr{ Ib_k - β_kI < c }] = 1

Footnote 3: In BLUE ”Best” means minimal variance estimator

- Minimal variance or efficient estimator means that

  var(b_k) < var(a_k) for all estimators a different from b
- Equvalent:

  E[b_k - β_k]^2 < E[a_k - β_k]^2 for all estimators a unlike b

Footnote 4: Biased estimators

- Even if the requirements ensuring that our estimates are BLUE one may at times find biased estimators with less varaince such as in
- Ridge Regression

Footnote 5: Non-linear estimators

- There may be non-linear estimators that are unbiased and with less variance than BLUE estimators

III NORMALLY DISTRIBUTED ERROR TERM

- (5) If all errors are normally distributed with expectation 0 and standard deviation of σ^2 , that is if

  $\varepsilon_i \sim N(0, \sigma^2)$ for all i

  - Then we can test hypotheses about β and σ, and
  - OLS estimates will have less variance than estimates from all other unbiased estimators
  - **OLS results in "BUE"**

(Best Unbiased Estimate)

Problems in regression analysis that cannot be tested

- If all relevant variables are included
- If x-variables have measurement errors
- If the expected value of the error is 0
- (This means that we are unable to check if the correlation between the error term and x-variables actually is 0 and actually the same as the first point that we are unable to test if the model is correctly specified)

Problems in regression analysis that can be tested (1)

- Non-linear relationships
- Inclusion of an irrelevant variable
- Non-constant error of the error term (heteroscedasticity)
- Autocorrelation for the error term
- Correlations among error terms
- Non-normal error terms
- Multicollinearity

Consequences of problems (Hamilton, p113)

		Unwanted properties of estimates			
Require ment	**Problem**	Biased estimate of b	Biased estimate of SE_b	Invalid t&F-tests	High var[b]
Specification	**Non-linear reltionship**	X	X	X	-
-”-	**Excluded relevant variable**	X	X	X	-
-”-	**Included irrelevant variable**	0	0	0	X
Gauss-Markov	**X with measurement error**	X	X	X	-
-”-	**Heteroscedasticity**	0	X	X	X
-”-	**Autocorrelation**	0	X	X	X
-”-	**X correlated with ε**	X	X	X	-
Normal distribution	**ε not normally distributed**	0	0	X	X
... no requirement	**Multicollinearity**	0	0	0	X

Problems in regression analysis that can be discovered (2)

- Outliers (extreme y-values)
- Influence (cases with large influence: unusual combinations of y and x-values)
- Leverage (potential for influence)

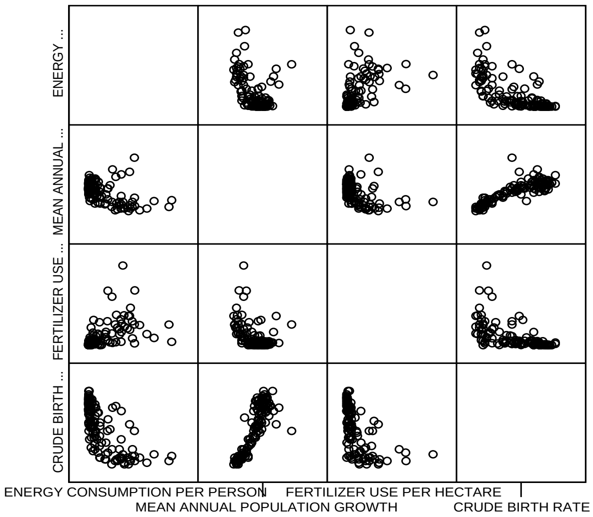
Tools for discovering problems

- Studies of
  - One-variable distributions (frequency distributions and histogram)
  - Two-variable co-variation (correlation and scatter plot)
  - Residual (distribution and covariation with predicted values)

Correlation and scatter plot

Data from 122 countries		ENERGY CONSUMP TION PER PERSON	MEAN ANNUAL POPULATION GROWTH	FERTILIZER USE PER HECTARE	CRUDE BIRTH RATE
ENERGY CONSUMPTION PER PERSON	Pearson Correlation	1	-,505	,533	-,689
	N	125	122	125	122
MEAN ANNUAL POPULATION GROWTH	Pearson Correlation	**-,505**	1	-,469	,829
	N	122	125	125	125
FERTILIZER USE PER HECTARE	Pearson Correlation	**,533**	**-,469**	1	-,589
	N	125	125	128	125
CRUDE BIRTH RATE	Pearson Correlation	**-,689**	**,829**	**-,589**	1
	N	122	125	125	125

Correlation and scatter plot

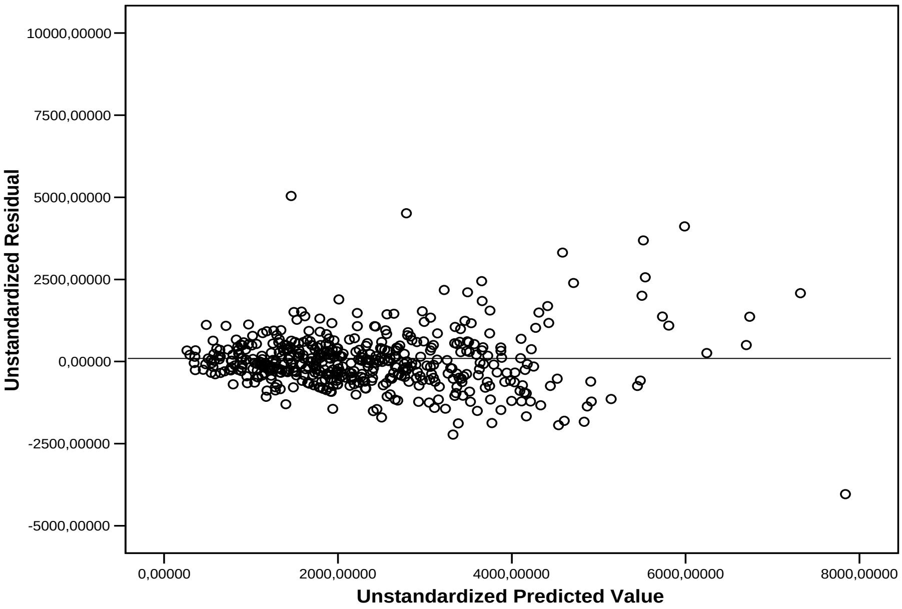


Heteroscedasticity (non-constant variance of error term) can arise from:

- Measurement error (e.g. y more accurate the larger x is)
- Outliers
- If ε_i contain an important variable that varies with both x and y (specification error)
- Specification error is the same as the wrong model and may cause heteroscedasticity
- An important diagnostic tool is a plot of the residual against predicted value ($\hat{Y}$)

Example: Hamilton table 3.2

Dependent Variable: Summer 1981 Water Use	Unstandardized Coefficients			
	B	Std. Error	t	Sig.
(Constant)	242,220	206,864	1,171	,242
Income in Thousands	20,967	3,464	6,053	,000
Summer 1980 Water Use	,492	,026	18,671	,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



From the regression reported in table 3.2 in Hamilton

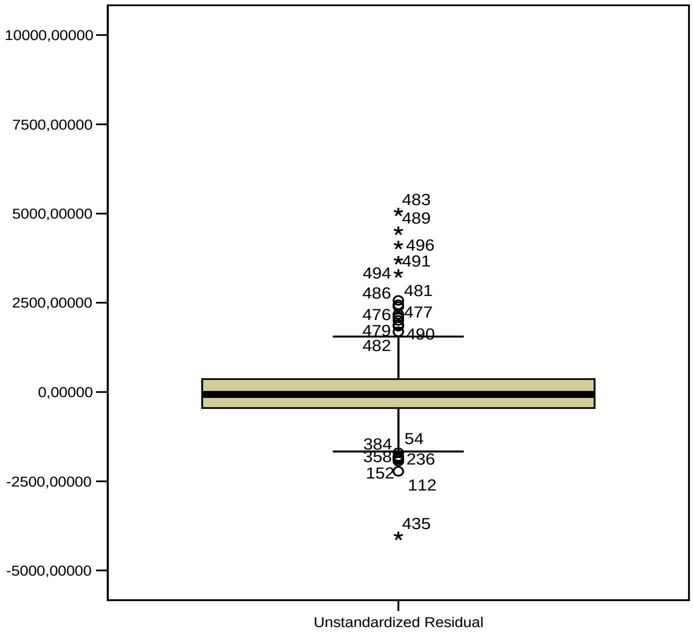
Footnote for the previous figure

- There is heteroscedasticity if the variation of the residual (variation around a typical value) varies systematically with the value of one or more x-variables
- The figure shows that the variation of the residual increses with increasing predicted y: $\hat{y}$
- Predicted Y ($\hat{Y}$) is in this case an index showing high average x-values
- When the variation of the residual varies systematically with the values of the x-variables like this, we conclude with heteroscedasticity

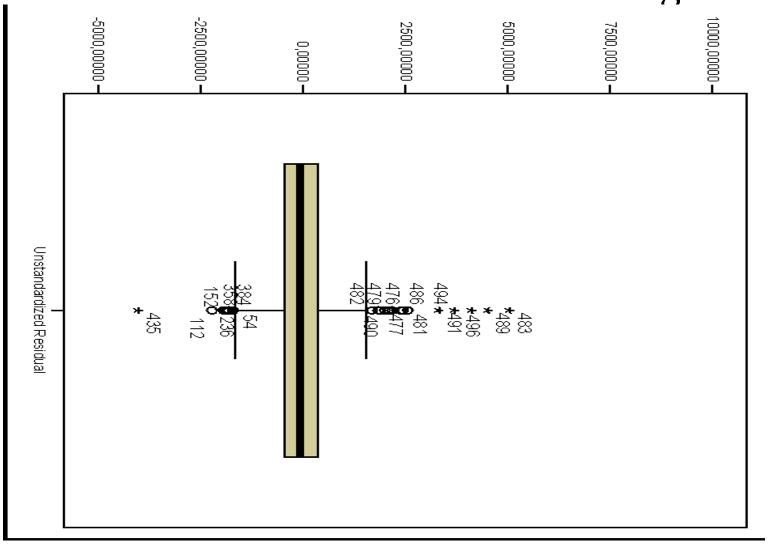
Box-plot of the residual shows

- Heavy tails
- Many outliers
- Weakly positively skewed distribution

Will any of the outliers affect the regression?

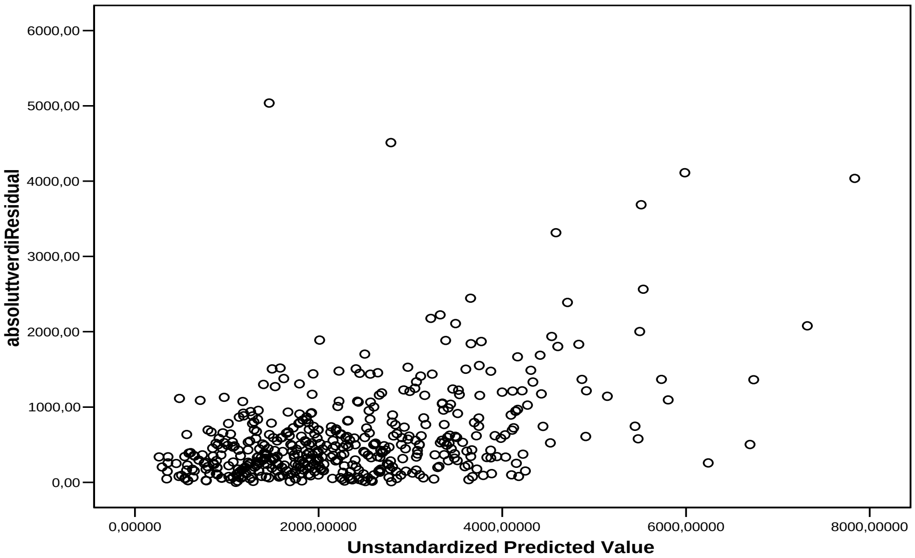


The distribution seen from another angle



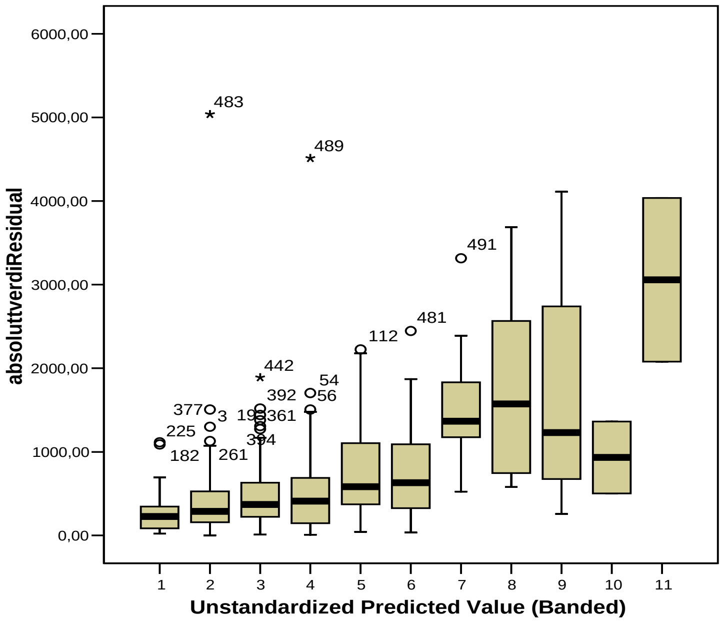
Band-regression

- Homoscedasticity means that the median (and the average) of the absolute value of the residual, i.e.: median{Ie$_i$I}, should be about the same for all values of the predicted y$_i$
- If we find that the median of Ie$_i$I for given predicted values of y$_i$ changes systematically with the value of predicted y$_i$ it indicates heteroscedasticity
- Such analyses can easily be done in SPSS



Absolute value of e$_i$ (Based on regression in table 3.2 in Hamilton)

Approximate band regression (cpr figure 4.4 in Hamilton)



Band regression in SPSS

- Start by saving the residual and predicted y from the regression
- Compute a new variable by taking the absolute value of the residual (Use “compute” under the “transform” menu)
- Then partition the predicted y into bands by using the procedure ”Visual bander” under the ”Transform” menu
- Then use ”Box plot” under ”Graphs” where the absolute value of the residual is specified as variable and the band variable as category axis

Autocorrelation (1)

- Correlation among variable values on the same variable across different cases

  (e.g. between ε_i and $\varepsilon_{i\ -1}$)
- Autocorrelation leads to larger variance and biased estimates of the standard error - similar to heteroscedasticity
- In a simple random sample from a population autocorrelation is improbable

Autocorrelation (2)

- Autocorrelation is the result of a wrongly specified model
- Typically it is found in time series and geographically ordered cases
- Tests (e.g. Durbin-Watson) is based on the sorting of the cases. Hence:
- A hypothesis about autocorrelation needs to specify the sorting order of the cases

Durbin-Watson test (1)

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^{\,2}}$$

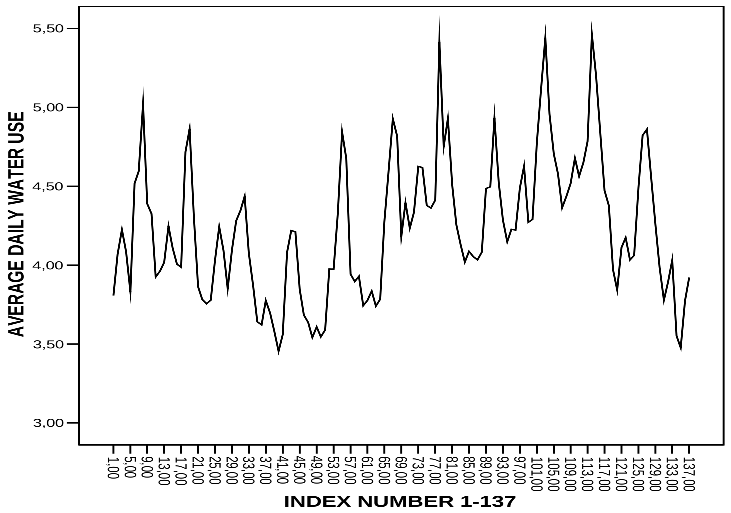
Should not be used for autoregressive models, i.e. models where the y-variable also is an x-variable, see table 3.2

Durbin-Watson test (2)

- The sampling distribution of the d-statistic is known and tabled as **d_L** and **d_U** (table A4.4 in Hamilton), the number of degrees of freedom is based on n and K-1
- Test rule:
  - Reject if $d<d_L$
  - Do not reject if $d>d_U$
  - If $d_L < d < d_U$ the test is inconclusive
- d=2 means uncorrelated residuals
- Positive autocorrelation results in d<2
- Negative autocorrelation results in d>2

Daily water use, average pr month

Example:



Ordinary OLS-regression where the case is month

Dependent Variable: AVERAGE DAILY WATER USE	Unstandardized Coefficients		t	Sig.
	B	Std. Error		
(Constant)	3,828	,101	38,035	,000
AVERAGE MONTHLY TEMPERATURE	,013	,002	7,574	,000
PRECIPITATION IN INCHES	-,047	,021	-2,234	,027
CONSERVATION CAMPAIGN DUMMY	-,247	,113	-2,176	,031

Predictors: (Constant), CONSERVATION CAMPAIGN DUMMY, AVERAGE MONTHLY TEMPERATURE, PRECIPITATION IN INCHES

Test of autocorrelation

Dependent Variable: AVERAGE DAILY WATER USE	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,572(a)	,327	,312	,36045	**,535**

Predictors: (Constant), CONSERVATION CAMPAIGN DUMMY, AVERAGE MONTHLY TEMPERATURE, PRECIPITATION IN INCHES

N = 137, K-1 = 3

Find limits for rejection / acceptance of the null hypothesis of no autocorrelation with level of significance 0,05

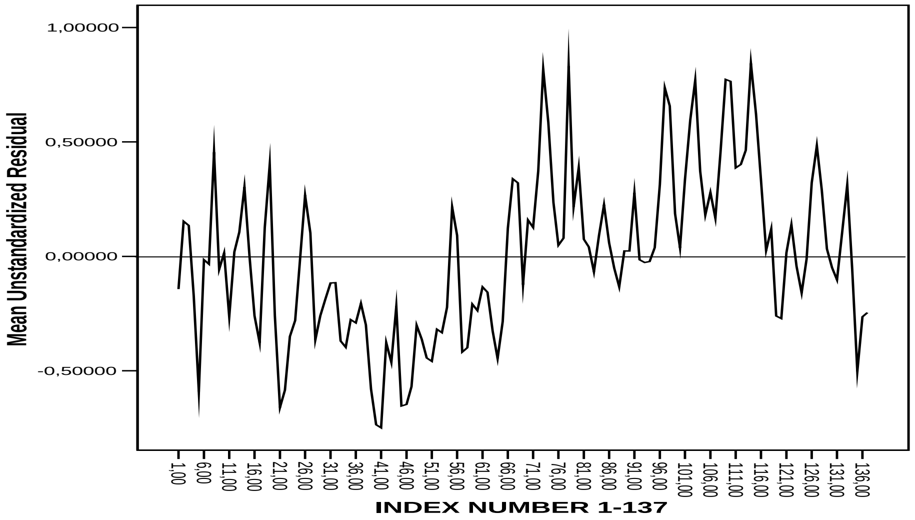
Tip: Look up table A4.4 in Hamilton, p355

Autocorrelation coefficient

m-th order autocorrelation coefficient

$$r_m = \frac{\sum_{t=1}^{T-m}\left(e_t - \bar{e}\right)\left(e_{t+m} - \bar{e}\right)}{\sum_{t=1}^{T}\left(e_t - \bar{e}\right)^2}$$

Residual ”Daily water use”, month



Smoothing with 3 points

- Sliding average

$$e_t^* = \frac{e_{t-1} + e_t + e_{t+1}}{3}$$

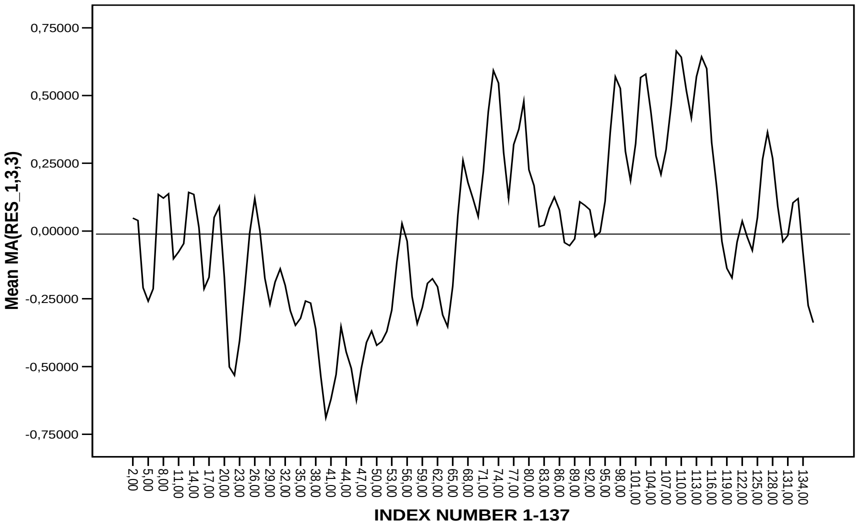
- ”Hanning”

$$e_t^* = \frac{e_{t-1}}{4} + \frac{e_t}{2} + \frac{e_{t+1}}{4}$$

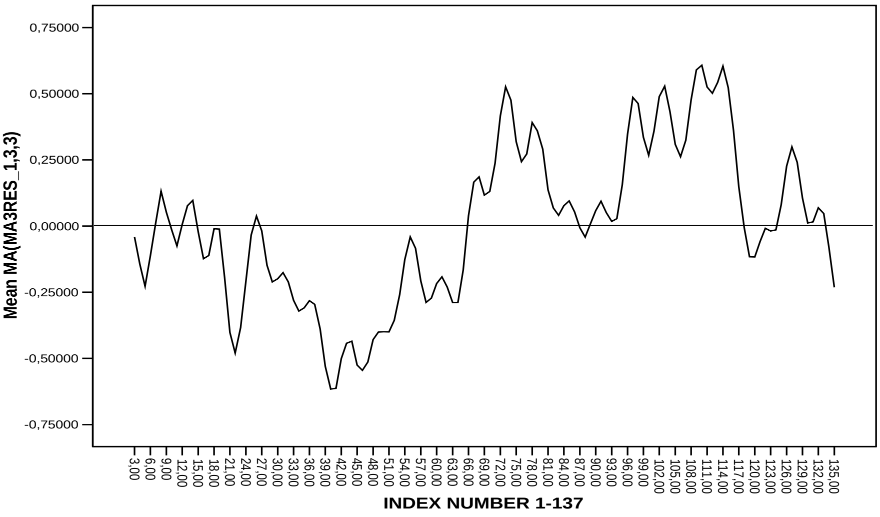
- Sliding median

$$e_t^* = median\left\{e_{t-1}, e_t, e_{t+1}\right\}$$

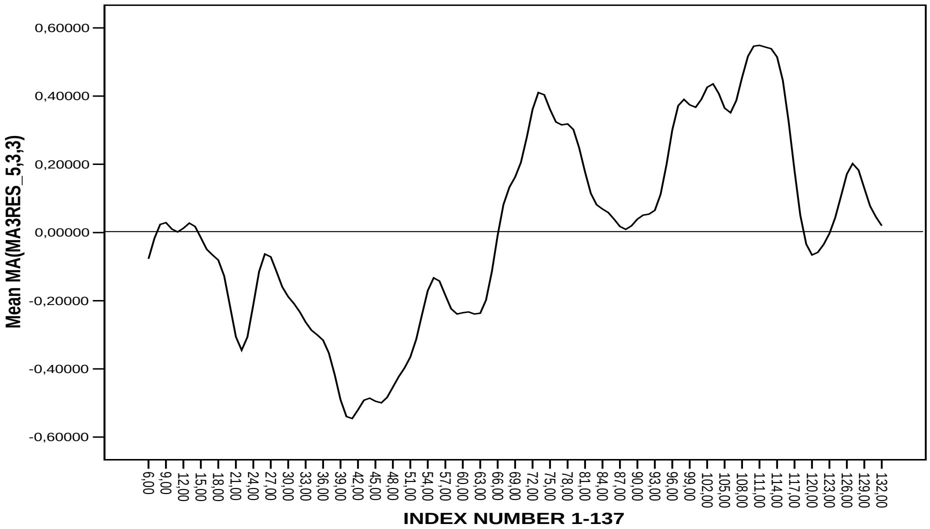
Residual, smoothing once



Residual, smoothing twice



Residual, smoothing five times



Consequences of autocorrelation

- Tests of hypotheses and confidence intervals are unreliable. Regressions may nevertheless provide a good description of the sample. Parameters are unbiased
- Special programs can estimate standard errors consistently
- Include in the model variables affecting neighbouring cases
- Use techniques developed for time series analysis (e.g.: analyse the difference between two points in time, Δy)

SOS3003
Applied data analysis for social science

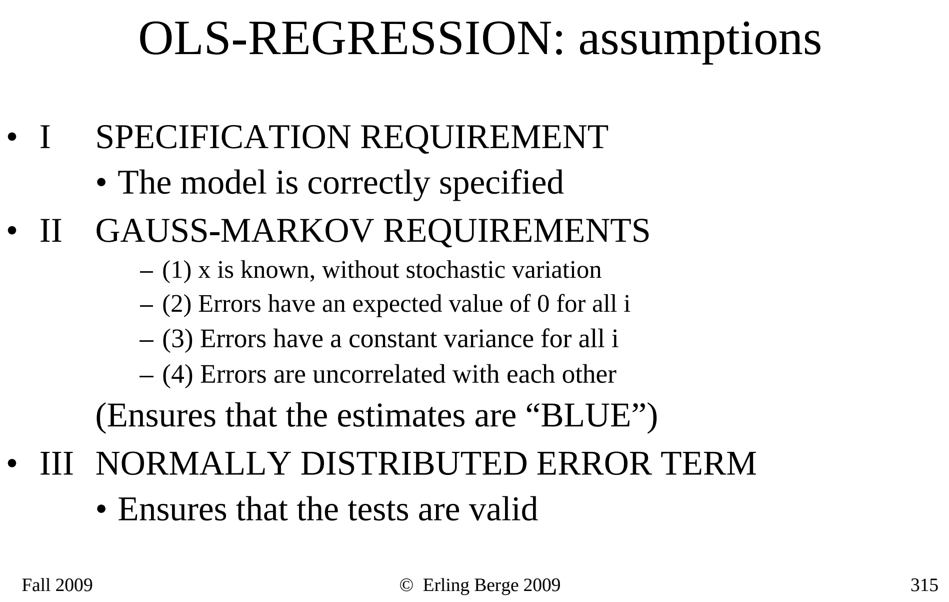
Lecture notes on

Hamilton Ch 4 p109-137
Regression criticism II

Erling Berge
Department of sociology and political science
NTNU

Analyses of models are based on assumptions

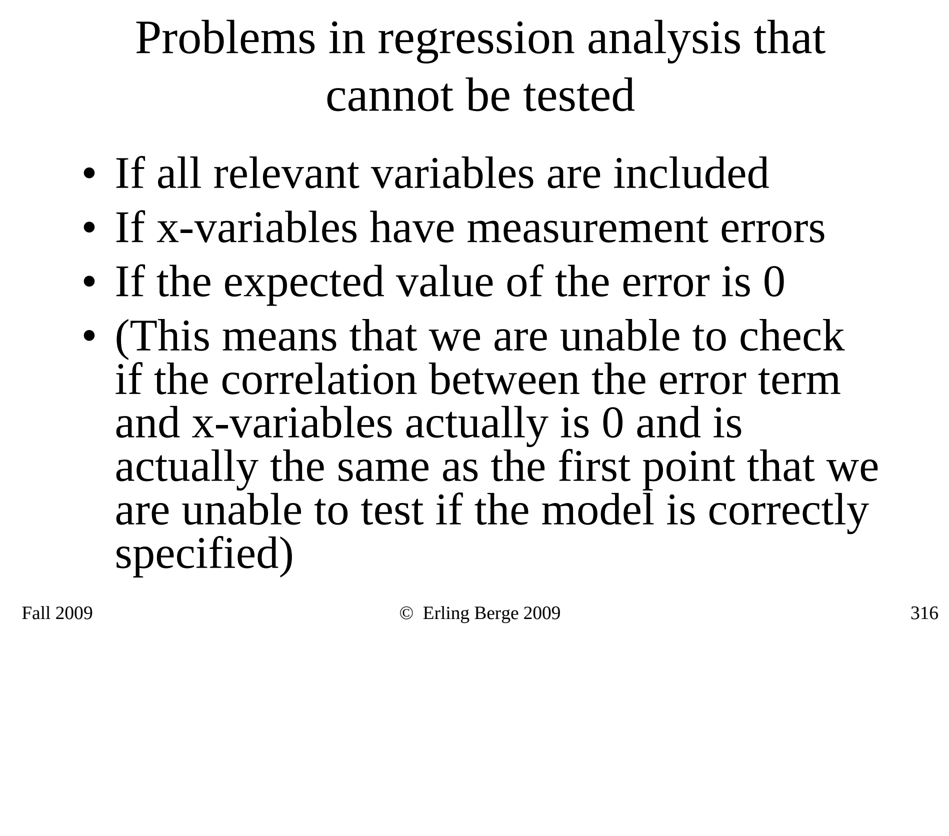
- OLS is a simple technique of analysis with very good theoretical properties. But
- The good properties are based on certain assumptions
- If the assumptions do not hold the good properties evaporates
- Investigating the degree to which the assumptions hold is the most important part of the analysis



OLS-REGRESSION: assumptions

- I SPECIFICATION REQUIREMENT
  - The model is correctly specified
- II GAUSS-MARKOV REQUIREMENTS
  - (1) x is known, without stochastic variation
  - (2) Errors have an expected value of 0 for all i
  - (3) Errors have a constant variance for all i
  - (4) Errors are uncorrelated with each other

  (Ensures that the estimates are “BLUE”)
- III NORMALLY DISTRIBUTED ERROR TERM
  - Ensures that the tests are valid



Problems in regression analysis that cannot be tested

- If all relevant variables are included
- If x-variables have measurement errors
- If the expected value of the error is 0
- (This means that we are unable to check if the correlation between the error term and x-variables actually is 0 and is actually the same as the first point that we are unable to test if the model is correctly specified)

The most important problems in regression analysis that can be tested

- Non-linear relationships
- Non-constant error of the error term (heteroscedasticity)
- Autocorrelation for the error term
- Non-normal error terms

Heteroscedastisity

- Is present if the variance of the error term varies with the size of x-values
- Predicted y is an indicator of the size of x-values (hence scatter plot of residual against predicted y)
- Heteroscedasticity (non-constant variance of error term) can arise from
  - Measurement error (e.g. y more accurate the larger x is)
  - Outliers
  - The wrong functional form
  - If ε_i contain an important variable that varies with one or more x and y. The error term ε_i is not independent of the x-es. Hence the Gauss-Markov requirements 1 and 2 cannot be correct.

Indicators of heteroscedastisity

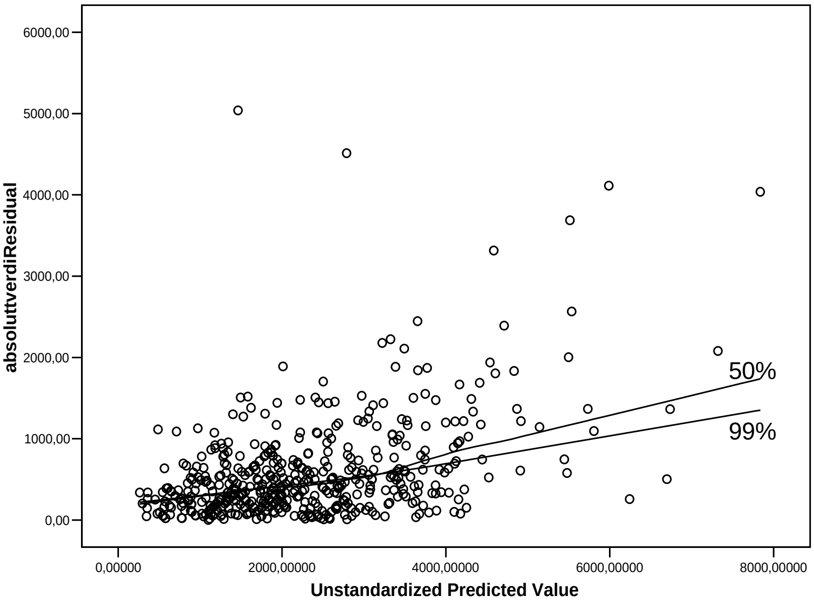
- Inspection of the scatter plot of residual against predicted value of y
- Band regression of the scatter plot

An interesting option here is:

- Locally weighted / ”sliding” regression on the central part of the sample

”Sliding” adapted line by means of locally weighted OLS regression

The procedure is called LOESS (see next slide)



A footnote: SPSS explains

Fit Lines

- In a fit line, the data points are fitted to a line that usually does not pass through all the data points. The fit line represents the trend of the data. Some fit lines are regression based. Others are based on iterative weighted least squares.
- Fit lines apply to scatter plots. You can create fit lines for all of the data values on a chart or for categories, depending on what you select when you create the fit line.

Loess

- Draws a fit line using iterative weighted least squares. At least 13 data points are needed. This method fits a specified percentage of the data points, with the default being 50%. In addition to changing the percentage, you can select a specific kernel function. The default kernel (probability function) works well for most data.

Autocorrelation

- Correlation among variable values on the same variable across different cases (e.g. between ε_i and $\varepsilon_{i -1}$)
- Autocorrelation leads to larger variance and biased estimates of the standard error - similar to heteroscedasticity
- Autocorrelation is the result of a wrongly specified model
- Typically it is found in time series and geographically ordered cases. In a simple random sample from a population autocorrelation is improbable
- Tests (e.g. Durbin-Watson) is based on the sorting of the cases. Hence: hypotheses about autocorrelation need to specify the sorting order of the cases

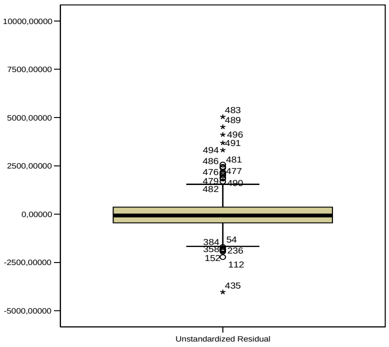
Non-normal residuals

- Imply that t- and F-tests cannot be used
- Since OLS estimates of parameters are easily affected by outliers, heavy tails in the distribution of the residual will indicate large variation in estimates from sample to sample
- We can test the assumption of normally distributed error term by inspecting the distribution of the residual, e.g. by inspecting
  - Histogram, box plot, or quantile-normal plot
  - There are also more formal tests (but not very useful) based on skewness and kurtosis

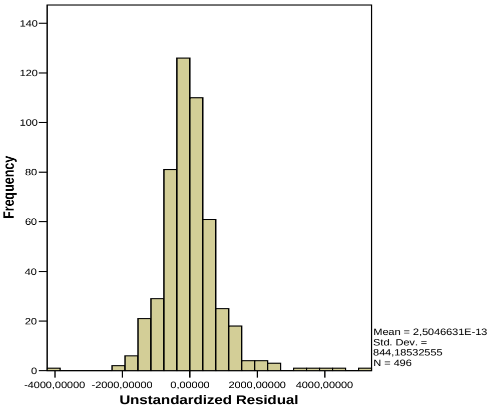
Diagram of the residual shows:

Heavy tails, many outliers, and weakly positively skewed distribution

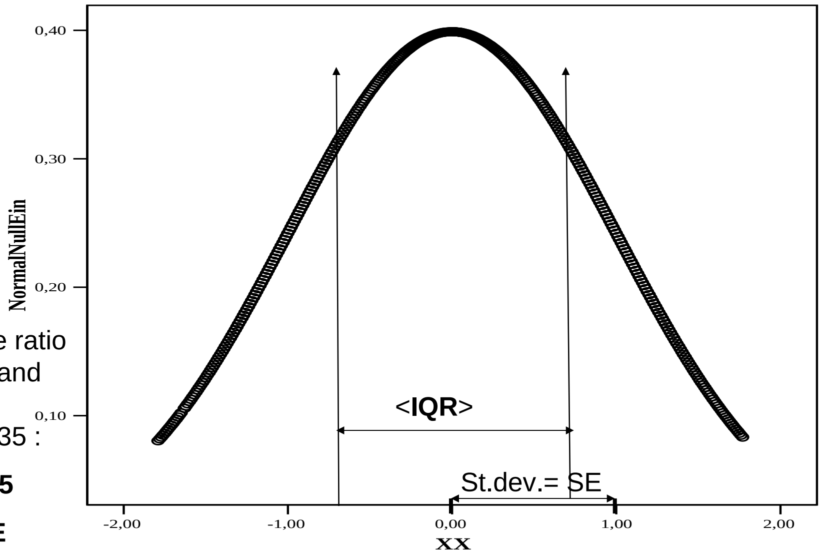
BOX PLOT



HISTOGRAM



Skewed distribution of the residual (1)



In the normal distribution the ratio between IQR and the standard deviation is 1.35 :

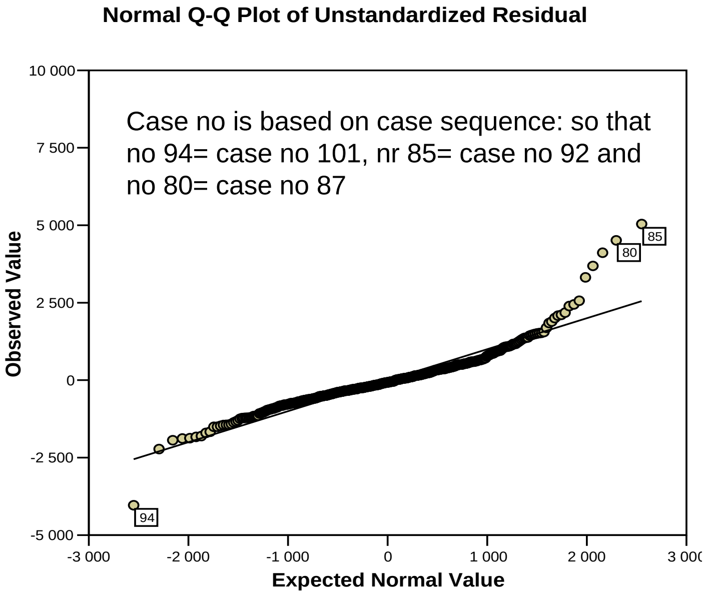
IQR/ SE = 1.35

IQR/1.35 = SE

Skewed distribution of the residual (2)

- Since the average of the residuals (e_i) always equals 0, the distribution will be skewed if the median is unequal to 0
- It is known that for the normal distribution the standard deviation (or the standard error) equals approximately IQR/1.35
- If the distribution of the residual is symmetric we can compare SE_e to IQR/1.35. If
  - SE_e > IQR/1.35 the tails are heavier than the normal distribution
  - $SE_e \approx$ IQR/1.35 the tails are approximately equal to the normal distribution
  - SE_e < IQR/1.35 the tails are lighter than the normal distribution

Quantile-Normal plot of residual from regression in table 3.2 in Hamilton



Case no is based on case sequence: so that no 94= case no 101, nr 85= case no 92 and no 80= case no 87

Options if non-normality is found

- Test out if the right function has been used
- Test out if some important variable has been excluded
  - If the model cannot be improved substantially, we may try transforming the dependent variable to symmetry
- Test out if lack of normality is caused by outliers or influential cases
  - If there are outliers, transforming of the variable where the case is outlier may help

Influence (1)

- A case (or observation) has influence if the regression result changes when the case is excluded
- Some cases have unusually large influence because of
  - Unusually large y-value (outliers)
  - Unusually large value on an x-variable
  - Unusual combinations of variable values

Influence (2)

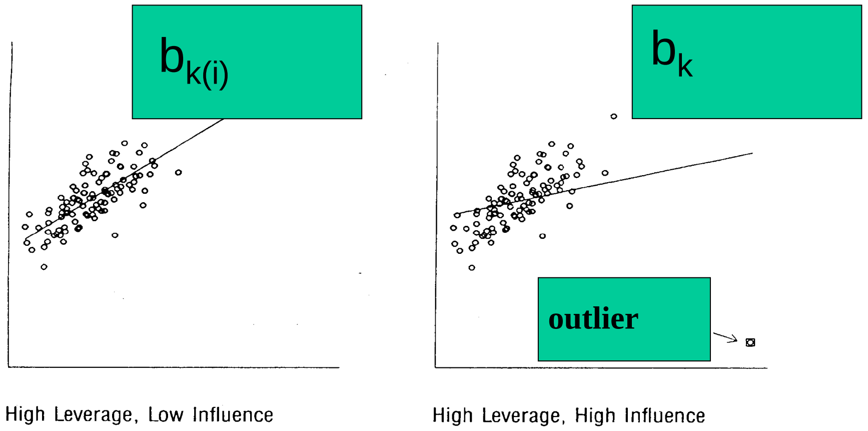
- We can see if a case has influence by comparing regressions with and without a particular case. One may for example
- Inspect the difference between b_k and $b_{k(i)}$ where case no i has been excluded in the estimation of the last coefficient
- This difference measured relative to the standard error of $b_{k(i)}$ is called DFBETAS_{ik}

DFBETAS$_{ik}$

$$DFBETAS_{ik} = \frac{b_k - b_{k(i)}}{\dfrac{s_{e(i)}}{\sqrt{RSS_k}}}$$

$s_{e(i)}$ is the standard deviation of the residual when case no i has been exclude from the analysis RSS$_k$ is Residual Sum of Squares from the regression of x$_k$ on all other x-variables

DFBETAS$_{ik}$:



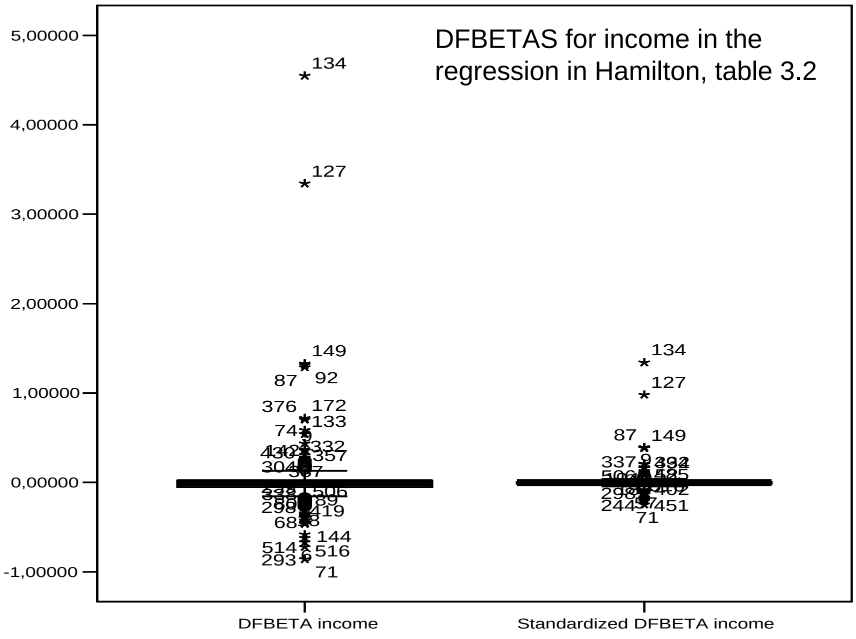
One case may make a lot of difference

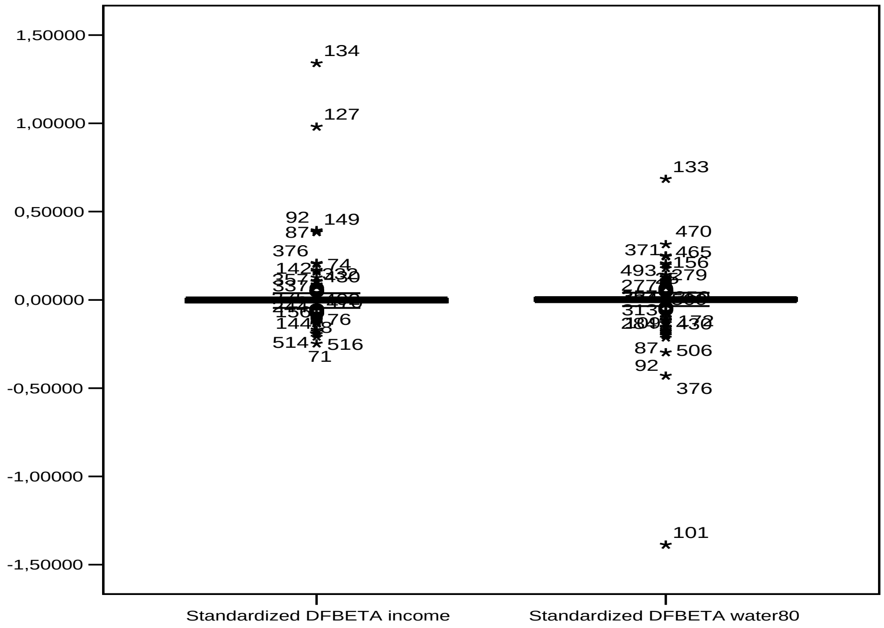
What is a large DFBETAS?

- DFBETAS$_{ik}$ is calculated for every independent variable for every case. We do not want to inspect all values for it
- Three criteria for finding large values we need to inspect are
  - External scaling. **l**DFBETAS$_{ik}$**l** > 2/ SQRT(n) $\sqrt{n}$
  - Internal scaling. Look for **severe outliers** in the box plot of DFBETAS$_{ik}$:

    DFBETAS$_{ik}$< Q$_1$-3IQR

    Q$_3$ + 3IQR < DFBETAS$_{ik}$
  - Gap in the distribution of DFBETAS$_{ik}$
- None of the DFBETAS$_{ik}$ needs to be problematic



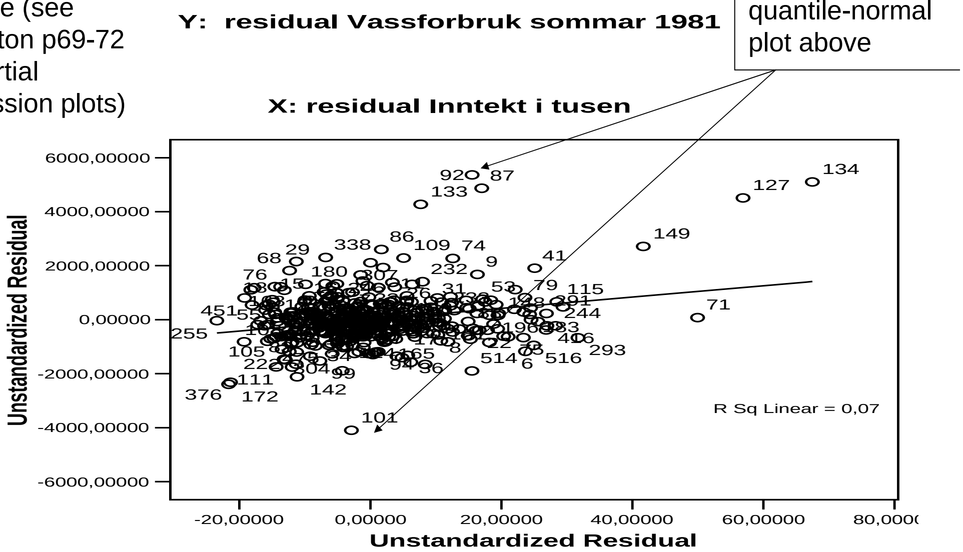


Sequence in the data set and case no is not the same. Case no is fixed. Variable values.

Sequence no	Case nr	water81	water80	water79	educat	retire	peop81	cpeop
91	98	1500	1300	1500	16	0	2	0
92	99	3500	6500	5100	14	0	6	0
93	100	1000	1000	2700	12	1	1	0
94	101	3800	12700	4800	20	0	5	0
95	102	4100	4500	2600	20	0	5	0
96	103	4200	5600	5400	16	0	5	-1
97	104	2400	2700	800	16	0	6	0
98	105	1600	2300	2200	14	0	4	0
99	107	2300	2300	3100	16	0	4	-2

Leverage plot for water use and income (see Hamilton p69-72 on partial regression plots)

Look at the quantile-normal plot above



Consequences of case with large influence

- If we discover cases with large influence we should not necessarily remove them from the analysis
- Report results both with and without the cases
- Take a careful look at influential cases, maybe there are measurement errors
- When influential cases are outliers their influence can be reduced by transformation
- Use robust regression not so easily affected as OLS regression

Potential influence: leverage

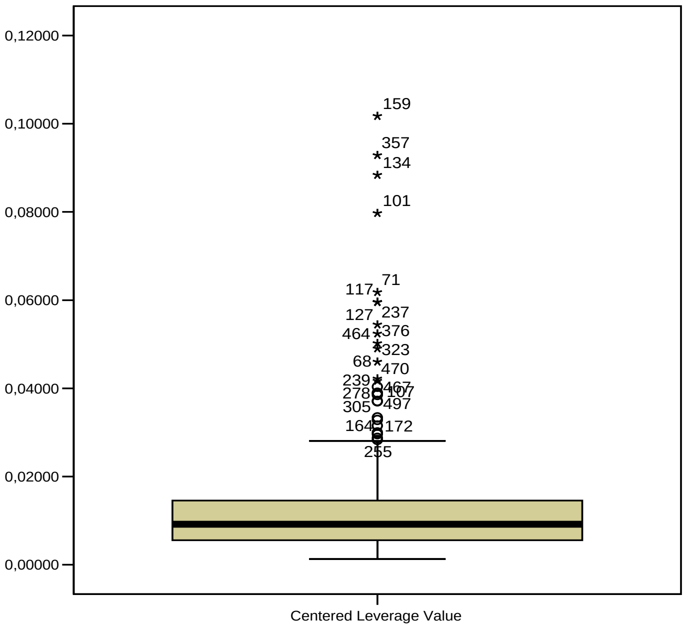
- The potential for influence of a case from a particular combination of x-values is measured by the hat statistic h_i
- h_i varies from 1/n to 1. It has an average of K/n (K = # parameters)
- SPSS reports the centred h_i
  - i.e. $(h_i – K/n)$, we may call this for h^c_i

What is a large value of leverage?

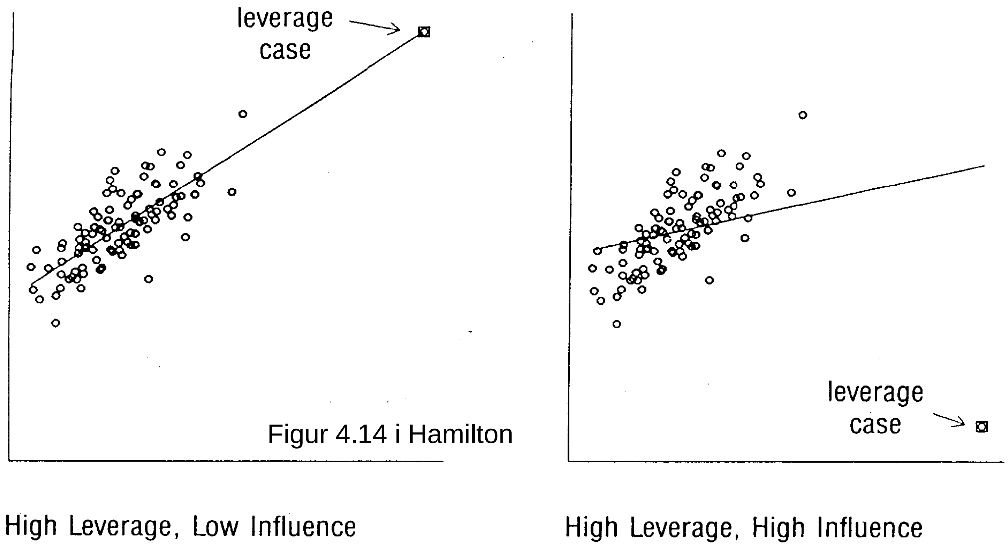
- As for DFBETAS different criteria can be suggested. They all depend on the sample size n
  - If $h_i > 2K/n$ (or $h^c_i > K/n$) we find the ca 5% largest h_i ; alternatively
    - If max $(h_i) \leq 0.2$ there is no problem
    - If $0.2 \leq$ max $(h_i) \leq 0.5$ there is some risk for a problem
    - If $0.5 \leq$ max (h_i) probably there is a problem

Centred leverage (h^c_i) from the regression in table 3.2 in Hamilton

Max av h^c_i er 0.102



The difference between influence and leverage



Figur 4.14 i Hamilton

The leverage statistic is found in many other case statistics

- Variance of the i-th residual $\mathrm{var}[e_i] = s_e^2[1-h_i]$
- Standardized residual (*ZRESID in SPSS) $$z_i = \frac{e_i}{s_e\sqrt{1-h_i}}$$
- Studentized residual (*SRESID in SPSS) $$t_i = \frac{e_i}{s_{e(i)}\sqrt{1-h_i}}$$
- And remember that the standard deviation of the residual is $$s_e = \sqrt{RSS/(n-K)}$$

Total influence: Cook's D_i

- Cook's distance D_i measure influence on the model as a whole, not on a specific coefficient as $\mathrm{DFBETAS}_{ik}$

$$D_i = \frac{z_i^2 h_i}{K\left(1-h_i\right)}$$

where z_i is the standardized residual

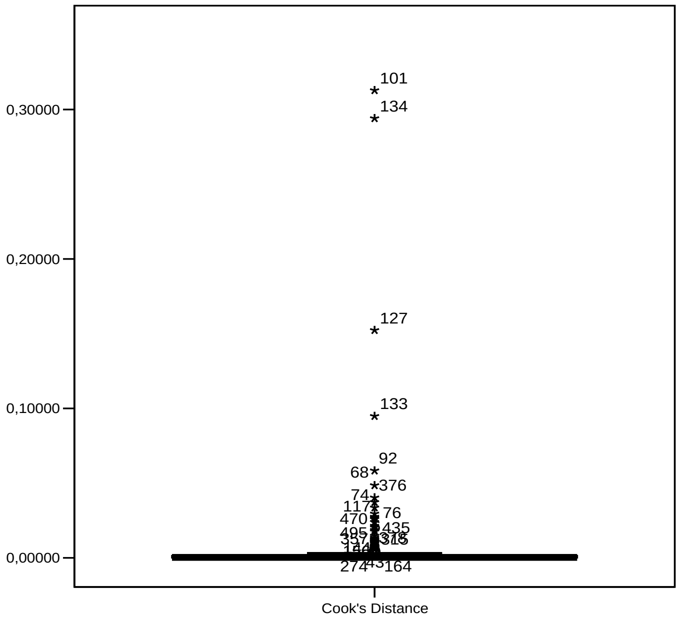
and h_i is the hat statistic (leverage)

What is a large D_i ?

- One might want to take a look at all
  - – $D_i > 1$ or
  - – $D_i > 4/n$ these are about the 5% largest D_i
- Even if a case has low D_i it may still be the case that it affects the size of single coefficients (it has a large $DFBETAS_{ik}$)

Cook’s distance D_i from the regression in table 3.2 in Hamilton

Also see table 4.4 (p133) in Hamilton



Summarizing

What can be done with outliers and cases with large influence? We can

- Investigate if data are erroneous. If data are wrong the case can be removed from the analysis
- Investigate if transformation to symmetry helps
- Report two equations: with and without cases with unreasonable large influence
- Get more data

Multicollinearity

- Means very high intercorrelations among x-variables
- Check if parameter estimates are correlated
- Check if tolerance (the part of the variation of x that is not shared with other variables) is less than say 0.1. If so there may be a problem
- VIF = variance inflation factor = 1/tolerance
- If multicollinearity is caused by squaring of variables or interaction terms it should not be seen as problematic

Tolerance

- The amount of variation in a variable x_k unique to that variable is called the tolerance of the variable
- Let R^2_k be the coefficient of determination in the regression of x_k on all the rest of the x-variables. The other x-variables explain the proportion R^2_k of the variation in x_k.
- Then 1- R^2_k is the unique variation: tolerance= 1- R^2_k
- Perfect multicollinearity means that
  - $R^2_k = 1$ and tolerance = 0
- Low values of tolerance make regression results less precise (larger standard errors)

Variance Inflation Factor (VIF)

- The standard error of the regression coefficient b_k can be written

$$SE_{b_k} = \frac{s_e}{\sqrt{RSS_k}} = \frac{s_e}{\sqrt{\left(1 - R_k^2\right)TSS_k}} = \sqrt{VIF}\,\frac{s_e}{\sqrt{TSS_k}}$$

- 1/tolerance = 1/(1-R^2_k) = VIF
- Other things being equal lower tolerance (larger VIF) for x_k will give higher standard error for b_k [SE increase with a factor equal to square root of VIF]

Indicators of multicollinearity

- The best indicators is tolerance or VIF (both are based on R^2_k)
- Other indicators are
  - – Correlation among singe variables (not reliable)
  - – Inclusion/ exclusion of single variables give large changes in the effect of other variables
  - – Unexpected signs on the effects of some variable
  - – Standardized regression coefficients larger than1 or less than -1
  - – Correlation among parameter estimates

Tolerance and VIF from regression in table 3.2 in Hamilton

Dependent Variable: Summer 1981 Water Use	Unstandardized Coefficients		t	Sig.	**Collinearity Statistics**	
	B	Std. Error			**Tolerance**	**VIF**
(Constant)	242,220	206,864	1,171	,242		
Summer 1980 Water Use	,492	,026	18,671	,000	,675	1,482
Income in Thousands	20,967	3,464	6,053	,000	,712	1,404
Education in Years	-41,866	13,220	-3,167	,002	,873	1,145
head of house retired?	189,184	95,021	1,991	,047	,776	1,289
# of People Resident, 1981	248,197	28,725	8,641	,000	,643	1,555
Increase in # of People	96,454	80,519	1,198	,232	,957	1,045

What is low tolerance?

When $R^2_k > 0,9$ tolerance is < 0,1 and VIF > 10

Factor of multiplication for the standard error is the square root of VIF (ca 3.2 for $R^2_k = 0,9$)

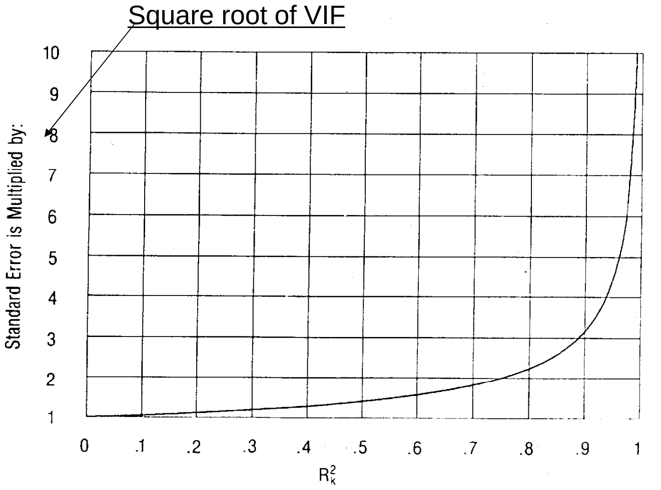

Square root of VIF

Figure 4.15 Effect of multicollinearity on standard errors (simplified).

When is multicollinearity a problem?

- It is not a problem if the reason is curvilinearity or interaction terms in the model. But in testing we need to take account of the fact that if VIF is high parameter estimates are imprecise (high standard errors). They are tested as a group by the F-test
- If the reason is that two variables measure the same concept one of them should be dropped, or they can be combined in an index
- It is a problem if we need estimates of the separate effects of two highly correlated variables (if a test of their joint effect is not sufficient)

Summarizing (1)

- When errors are independent and identically normally distributed OLS estimates are as good or better than other possible estimates
- But the assumptions are rarely satisfied completely, we have to test the degree to which they are satisfied
- Many problems can be corrected if we learn about them
- Check early on if curvilinearity, outliers or heteroscedasticity are problems (for example by use of scatter plots)

Summarizing (2)

- Do more exact investigations using residual/predicted Y plots and leverage plots
  - Curvilinearity (leverage plot, residual vs predicted Y plot)
  - Heteroscedasticity (leverage plot, [absolute value of residual] against predicted Y plot)
  - Non-normal residuals (quantile-normal plot, box-plot with analysis of median and IQR/1.35
  - Influence (check DFBETAS and Cook’s D)
  - When we do not find serious problems we can have more confidence in our conclusions

SOS3003
Applied data analysis for social science

Lecture notes on

Hamilton Ch 5 p145-273
Fitting Curves

Erling Berge
Department of sociology and political science
NTNU

Fitting Curves

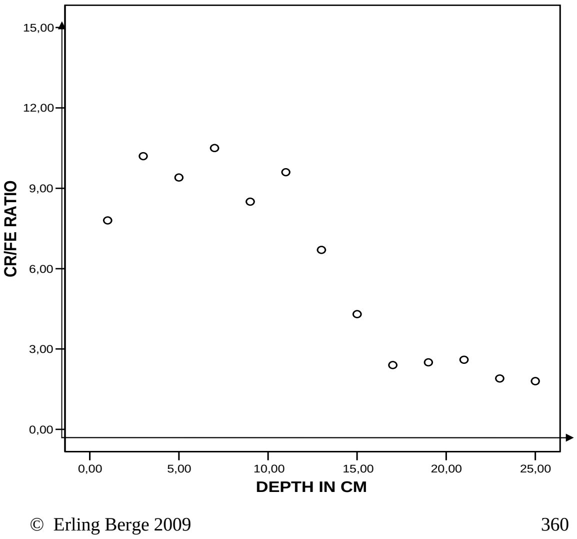
- A correctly specified model require that the function linking x-variables and y-variable is true to what really exist: is the relationship linear?
- Data can be inspected by means of band regression or smoothing
- The theory of causal impact can specify a non-linear relationship
- For phenomena that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

Band regression

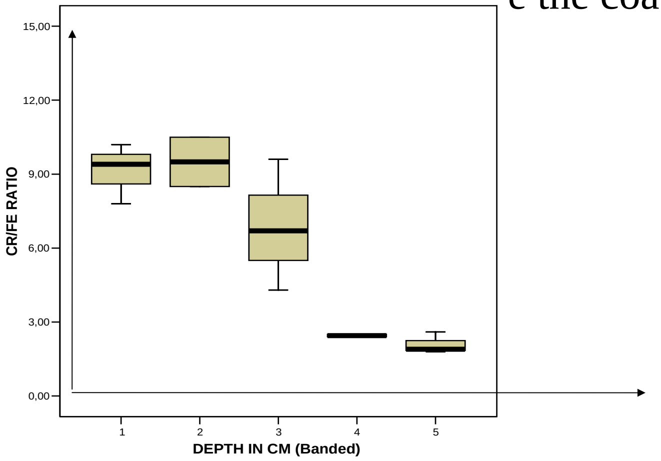
- Can be used to explore how the relationship among the variables actually appears
- If we can see a non-linear underlying trend of the data we must through transformations or use of curves find a form for the function better representing the relationship

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

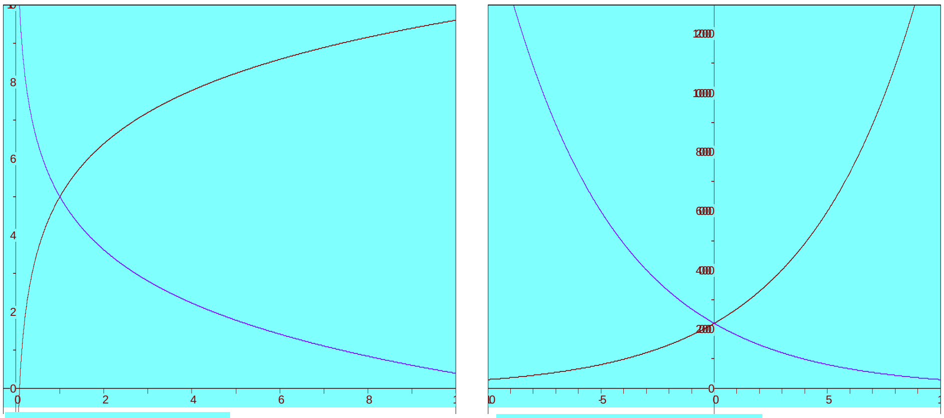
$$y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji} + \varepsilon_i$$

- 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)

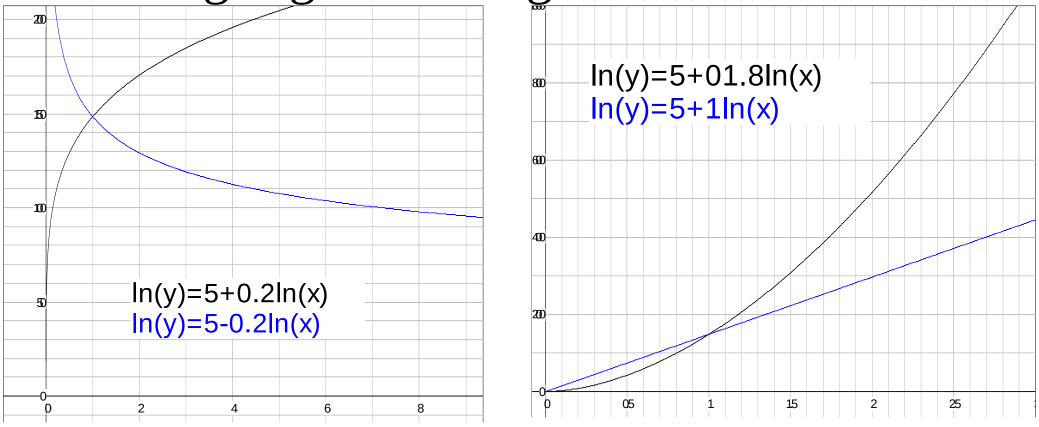
Semilog curves Fig 5.2 in Hamilton



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

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

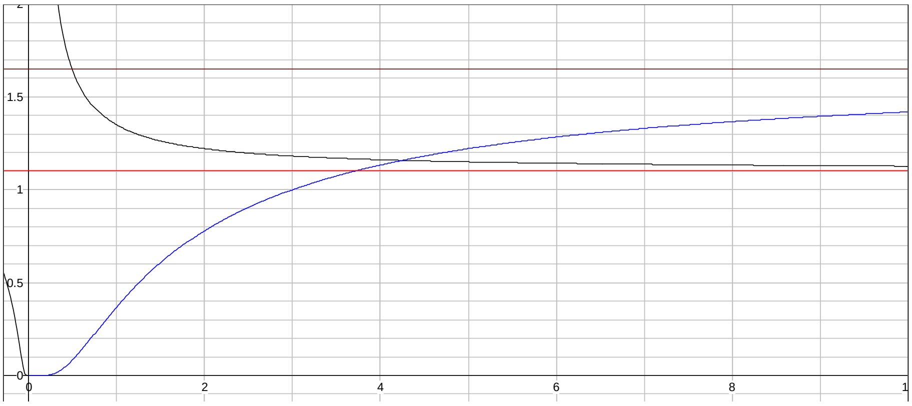
Log-log curves Fig 5.3 in Hamilton



ln(y)=5+0.2ln(x)
ln(y)=5-0.2ln(x)

ln(y)=5+01.8ln(x)
ln(y)=5+1ln(x)

Log-reciprocal curves Fig 5.4 in Hamilton



ln(y)=0.1+0.2/x

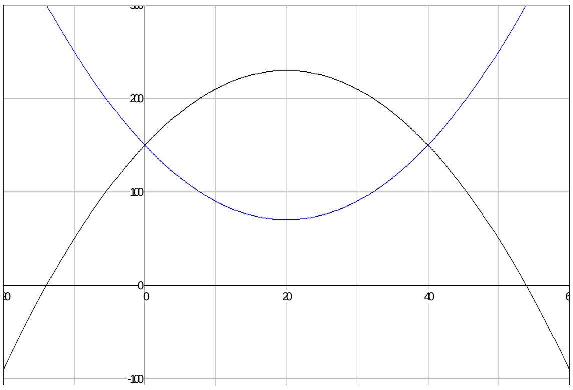
ln(y)=0.5-1.5/x

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

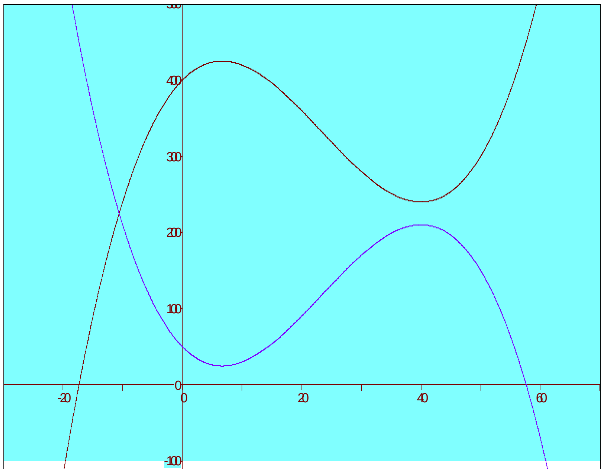
Second order polynomials Fig 5.5 in Hamilton



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

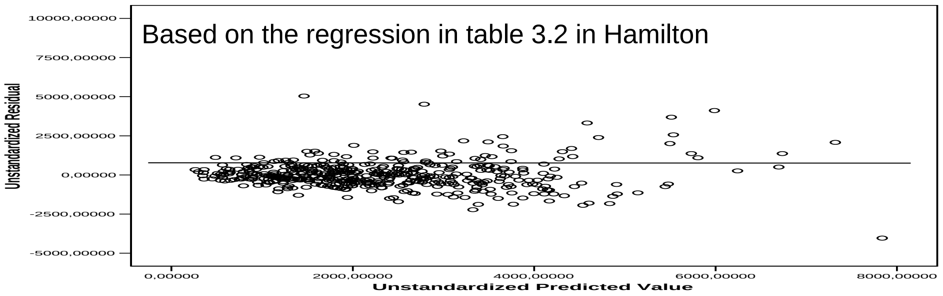
- Scatter plot or theory may provide advice
- Otherwise: transformation to symmetry gives the best option
- The regression reported in table 3.2 in Hamilton proved to be problematic
- Regression with transformed variables can reduce the problems

Choice of transformation in table 3.2 in Hamilton

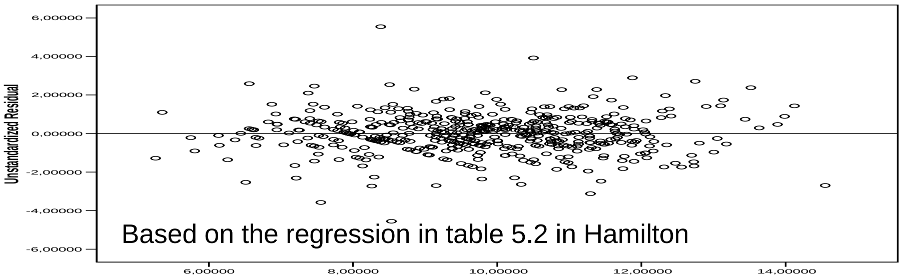
Y = Water use 1981	$Y^*=Y^{0.3}$ provides approximate symmetry
X_1 = Income	$X_1^*= X_1^{0.3}$ provides approximate symmetry
X_2 = Water use 1980	$X_2^*= X_2^{0.3}$ provides approximate symmetry
X_3 = Education	Transformations are inappropriate
X_4 = Pensioner	Transformations do not work for dummies
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

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

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

- 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

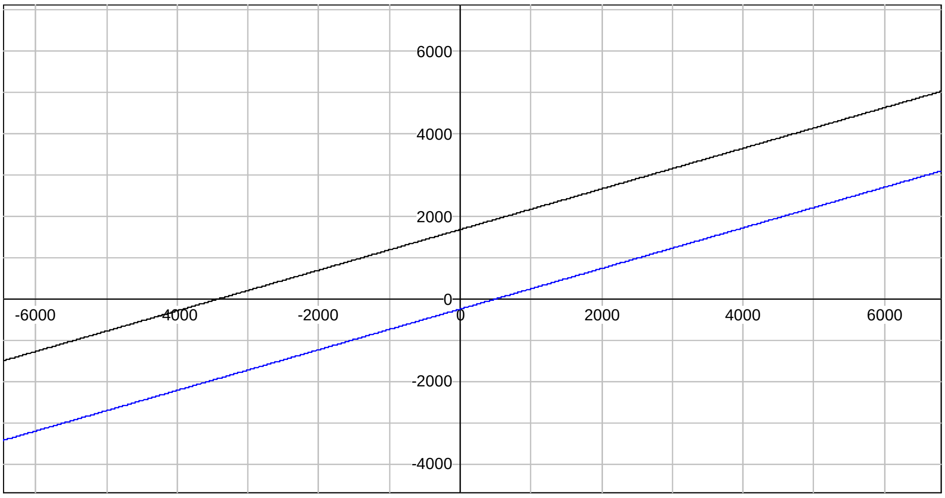
Dependent Variable: Summer 1981 Water Use	Unstandardized Coefficients			
	B	Std. Error	t	Sig.
(Constant)	242,220	206,864	1,171	,242
Summer 1980 Water Use	,492	,026	18,671	,000
Income in Thousands	20,967	3,464	6,053	,000
Education in Years	-41,866	13,220	-3,167	,002
head of house retired?	189,184	95,021	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

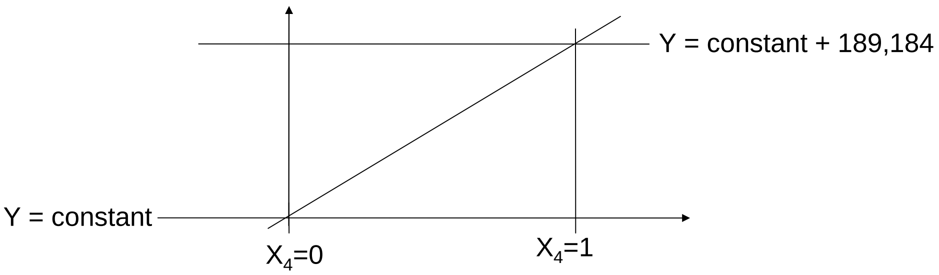


Y = 242.22 + 0.492X + 20.967×10 - 41.866×7 + 189.184×1 + 248.197×5 + 96.454×1

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

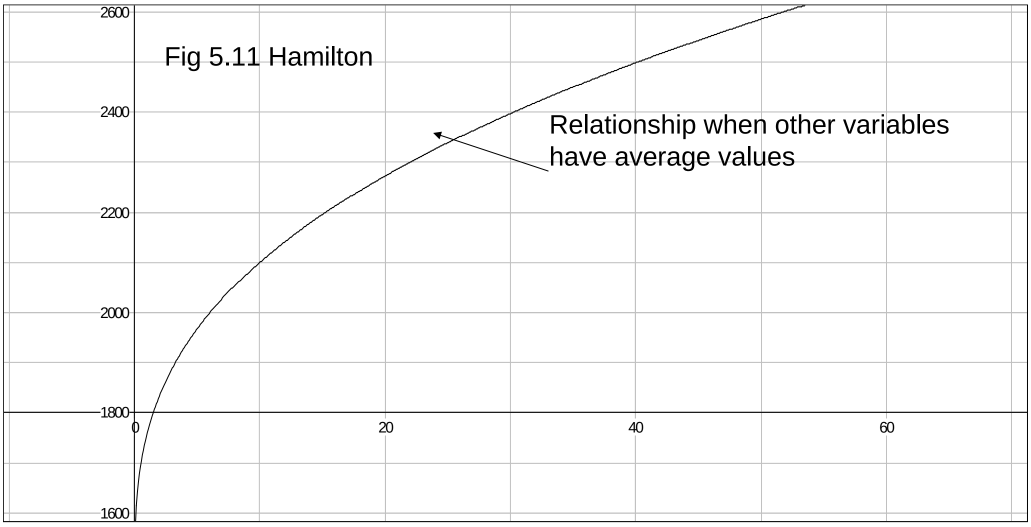


Fig 5.11 Hamilton

Relationship when other variables have average values

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

Which plots might be of interest?

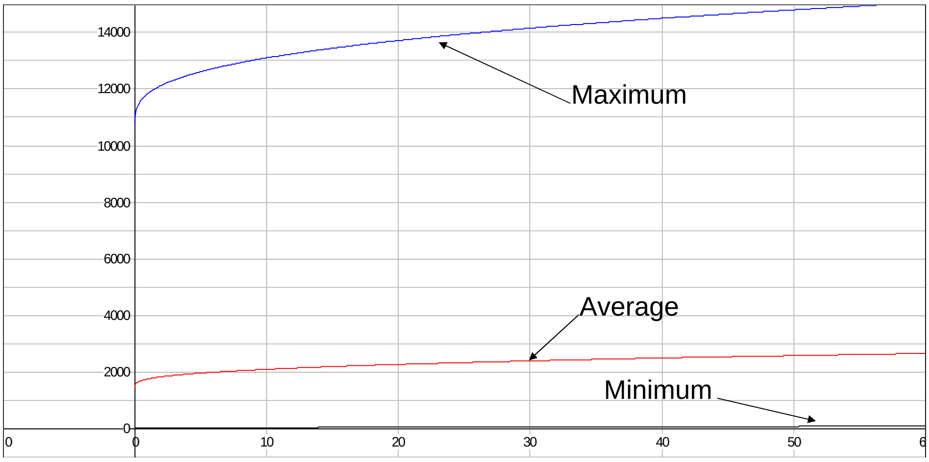
- The relationship between water usage and income controlled for the effect of other variables
  - Those minimizing water usage
  - Those maximizing water usage
  - Average values

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

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

- 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

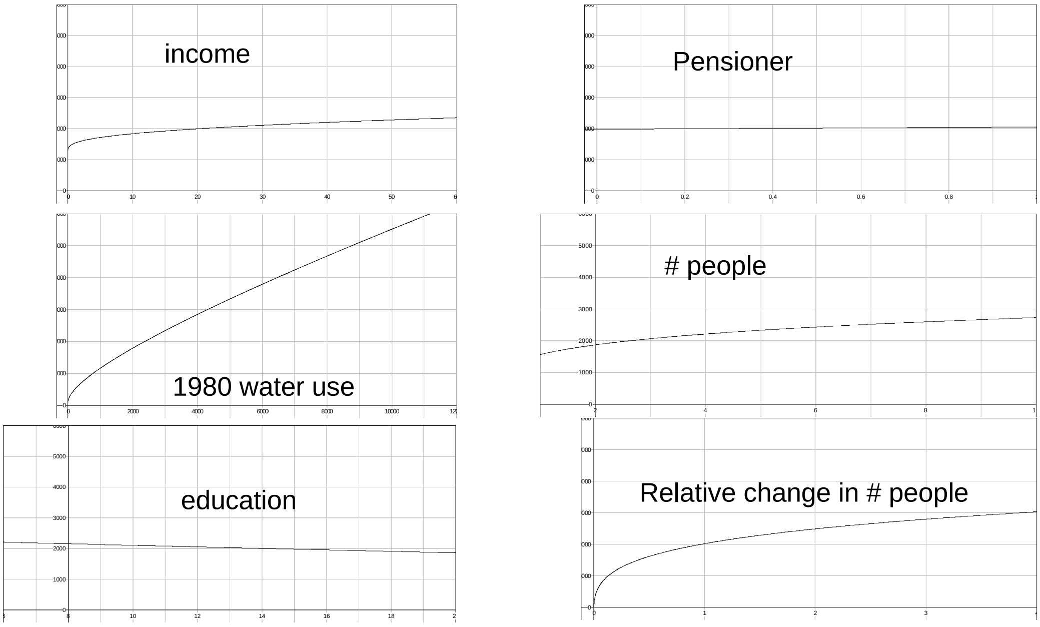
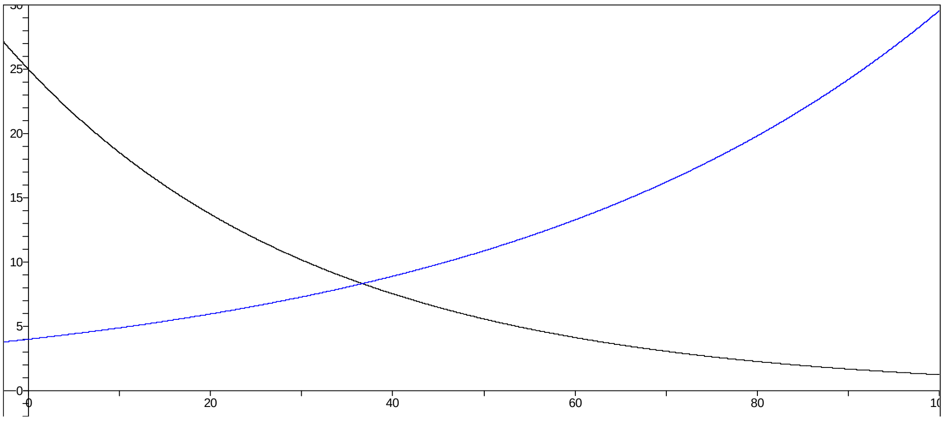


Fig 5.13 Hamilton

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

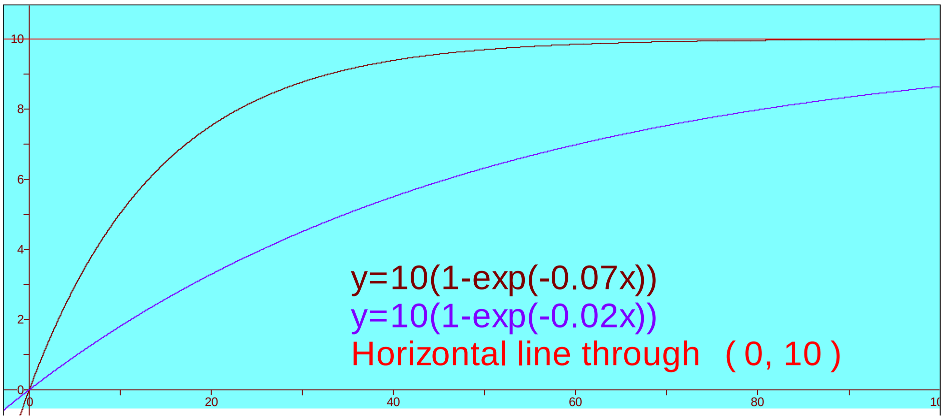
Exponential growth and decay
Fig 5.14 in Hamilton



y=25exp(-0.03x)

y=4exp(0.02x)

Negative exponential curves Fig 5.15 in Hamilton

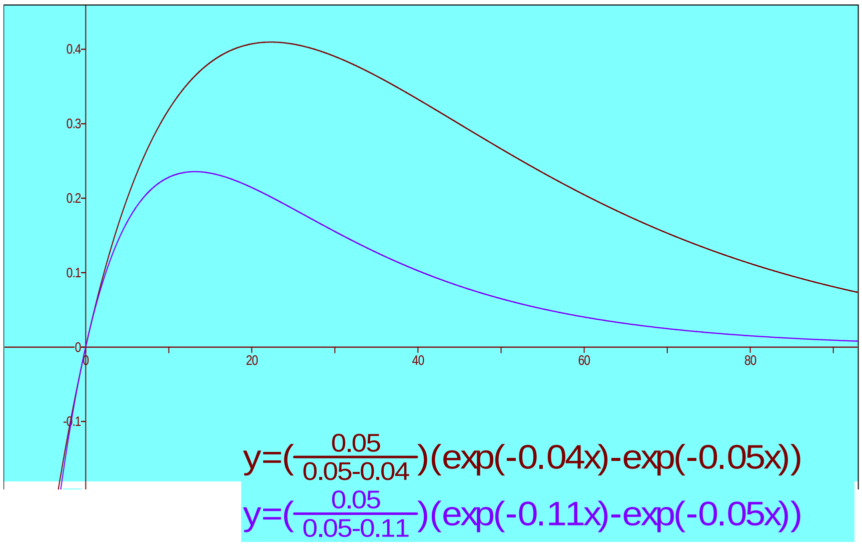


y=10(1-exp(-0.07x))

y=10(1-exp(-0.02x))

Horizontal line through (0, 10)

To-term exponential curves Fig 5.16 in Hamilton



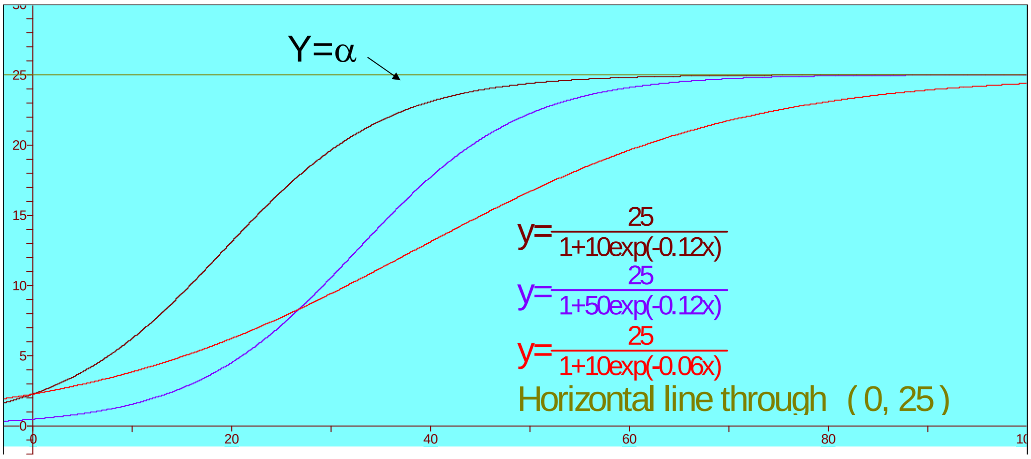
Logistic models

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

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

- 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 α=γ=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

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

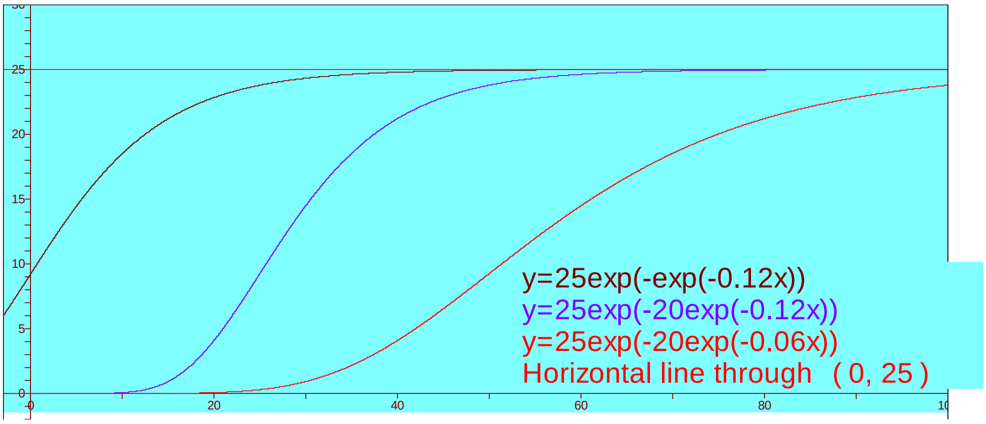
Gompertz curves

- 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



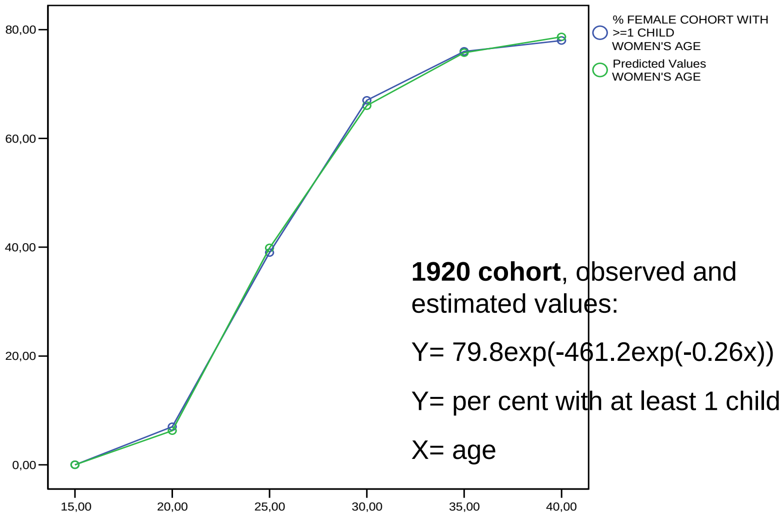
Estimation of non-linear models

- 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)



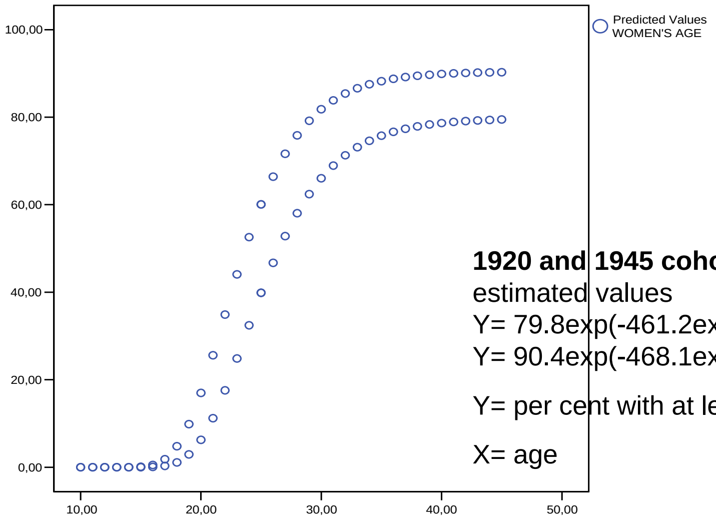
1920 cohort, observed and estimated values:

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

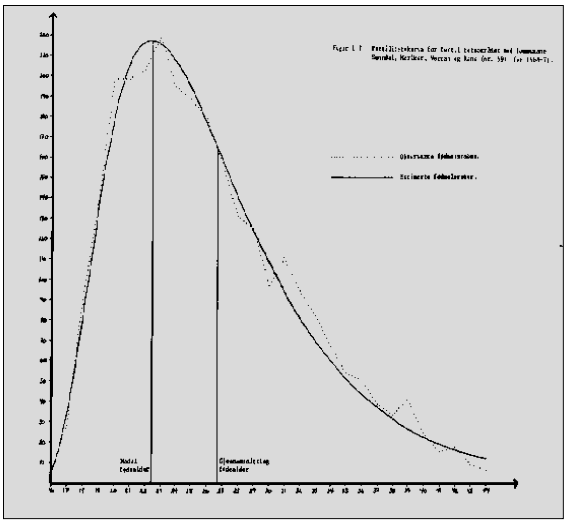
Model estimation and fit

- 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

SOS3003
Applied data analysis for social science
Lecture notes on

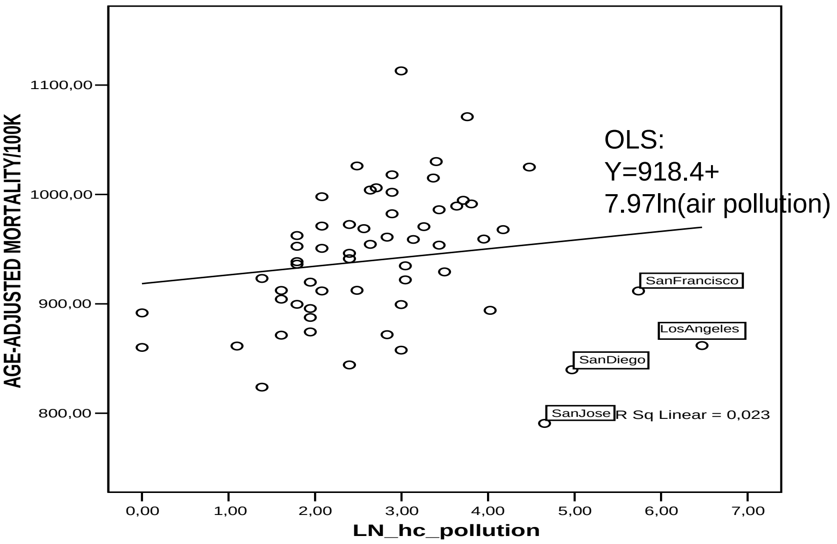
Hamilton Ch 6 p183-212
Robust Regression

Erling Berge
Department of sociology and political science
NTNU

Robust Regression

- Has been developed to work well in situations where OLS breaks down. Where the OLS assumptions are satisfied robust regression are not as good as OLS, but not by very much
- Even if robust regression is better suited for those who do not want to put much effort into testing the assumptions, it is so far difficult to use
- Robust regression has focused on residuals with heavy tails (many cases with high influence on the regression)

Regression of mortality on air pollution



OLS:
Y=918.4+
7.97ln(air pollution)

Figure 6.1
Hamilton

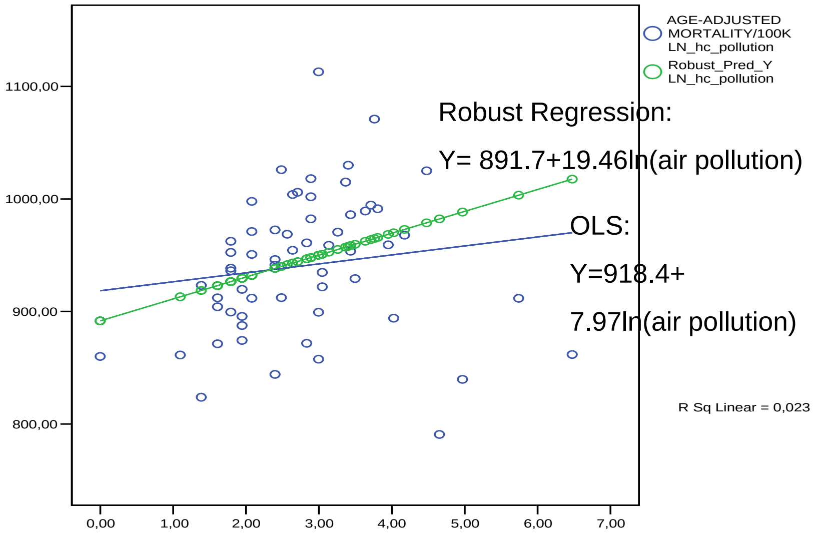


Figure 6.2
Hamilton

Robust regression and SPSS

- SPSS do not have a particular routine that performs robust regression
- It can possibly be done within the Generalized linear models procedure <but I have not tested it our>
- It can be done by weighted OLS regression, but then it is required that we make the weight functions and go through the iterations one by one including computation of weights every time
- This procedure will be outlined below

ROBUST AND RESISTANT

- RESISTANT methods are not affected by small errors or changes in the sample data
- ROBUST methods are not affected by small deviations from the assumptions of the model
- Most resistant estimators are also robust in relation to the assumption about normally distributed residuals
-
- **OLS is neither ROBUST nor RESISTANT**

Outliers is a problem for OLS

Outliers affect the estimates of

- Parameters
- Standard errors (standard deviation of parameters)
- Coefficient of determination
- Test statistics
- And many other statistics

Robust regression tries to protect against this by giving less weight to such cases,
not by excluding them

Protection against NON-NORMALE residuals

Robust methods can help when

- the tails in the distribution of the residuals are heavy, i.e. when it is too many outliers compared to the normal distribution
- Unusual X-values have leverage and may cause problems

But for other causes of non-normality

robust methods will not help

Estimation methods for robust regression

- M-estimation (maximum likelihood) minimizes a weighted sum of the residuals. This can be approximated by the weighted least squares method (WLS)
- R-estimation (based on rank) minimizes a sum where a weighted rank is included. The method is more difficult to use than M-estimation
- L-estimation (based on quantiles) uses linear functions of the sample order statistics (quantiles)

IRLS- Iterated Reweighted Least Squares

M-estimation by means of IRLS needs

1. Start values from OLS. Save the residuals
2. Use OLS residuals to find weights. Larger residuals gives less weight
3. Find new parameter values and residuals with WLS
4. Go to step 2 and find new weights from the new residuals, go on to step 3 and 4, until changes in the parameters become small

Iteration: to repeat a sequence of operations

IRLS

- IRLS is in theory equivalent to M-estimation
- To use the method we need to compute
- Scaled residuals, u_i , and a
- Weight function, w_i ,that gives least weight to the largest residuals

Scaling of residuals I

- Scaled residual u_i
  - s is the scale factor and e_i residual
- The scale factor in OLS is the estimate of the standard error of the residual: <u>nb! s_e is not resistant</u>
- A resistant alternative is based on MAD, "median absolute deviation"

$$u_i = \frac{e_i}{s}$$

$$s_e = \sqrt{\frac{RSS}{n-K}}$$

$$MAD = median \mid e_i - median\left(e_i\right) \mid$$

Scaling of residuals II

$$MAD = median \mid e_i - median\left(e_i\right) \mid$$

The scale factor (standard error of the distribution)

Using a resistant estimate will be

- s = MAD/ 0.6745 = 1.483MAD

and the scaled residual

- u_i = [e_i / s] = (0.6745*e_i)/MAD

In a normal distribution s= MAD/ 0.6745 will estimate the standard error correctly like s_e

In case of non-normal errors s= MAD/ 0.6745 will be better.

This is a resistant estimate, s_e is not resistant

Weight functions I

- Properties is measured in relation to OLS on normally distributed errors.
- The method should be “almost as good” as OLS on normally distributed errors and much better when the errors are non-normal
- Properties are determined by a “calibration constant” (c in the formulas)

Weight functions II

- **OLS-weights**: w_i = 1 for all i
- **Huber-weights**: weights down when the scaled residual is larger than c, c=1,345 gives 95% of the efficiency of OLS on normally distributed errors
- **Tukey’s bi-weighted** estimates get 95% of the efficiency of OLS on normally distributed errors by gradually weighting down scaled errors until $|u_i| \leq c = 4.685$ and by dropping cases where the residual is larger

Huber-weights

$$w_i = 1 \;\forall\; |\,\mathrm{u_i}\,| \le \mathrm{c}$$

$$w_i = \frac{c}{u_i} \forall\; |\,\mathrm{u_i}\,| > \mathrm{c}$$

$$\forall = \text{for alle}$$

Tukey weights

$$w_i = \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2 \forall\; |\,\mathrm{u_i}\,| \le c$$

$$w_i = 0 \;\forall\; |\,\mathrm{u_i}\,| > c$$

$$\forall = \; for \;\; alle$$

- Tukey weighting in IRLS is sensitive for start values of the parameters (one may end up at local minima)

Standard errors and tests in IRLS

- The WLS program cannot estimate standard errors and test statistics correctly by IRLS
- A procedure that works is described by Hamilton on page 198-1999

Use of Robust Estimation

- If OLS and Robust estimates are different it means that outliers have influence on the OLS results making them unreliable. Results cannot be trusted
- Robust predicted values will better portray the bulk of the data
- Robust residuals will better at discovering which cases are unusual
- Weights from the robust regression will show which cases are outliers
- OLS and RR can support each other

Fig 6.9 Hamilton: OLS and RR on untransformed

Mortality regressed on air pollution

Effect of high leverage

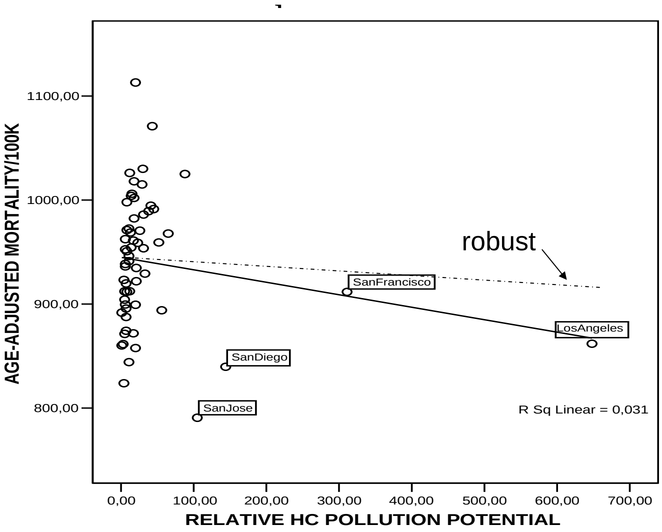
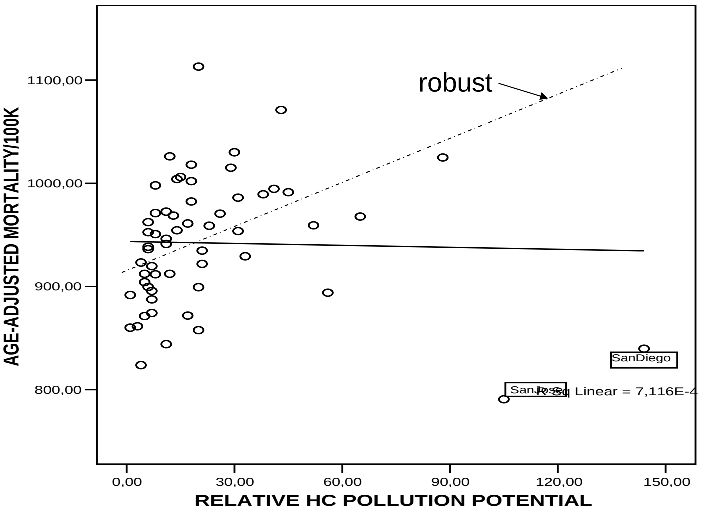


Fig 6.10 Hamilton: OLS and RR on untransformed

Mortality regressed on air pollution



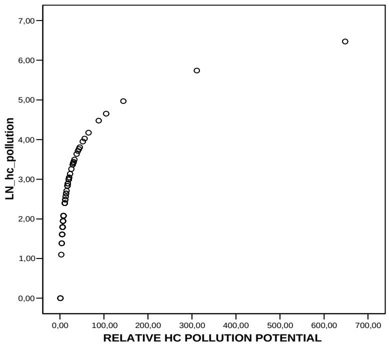
RR do not protect against leverage

- RR with M-estimation protects against unusual y-values (outliers) but not necessarily against unusual x-values (leverage)
- Efforts to test and diagnose are still needed (heteroscedasticity is still a problem for IRLS)
- Studies of the data and transformation to symmetry will reduce the risk of problems appearing
- No method is “safe” if it is used without forethought and diagnostic studies of data

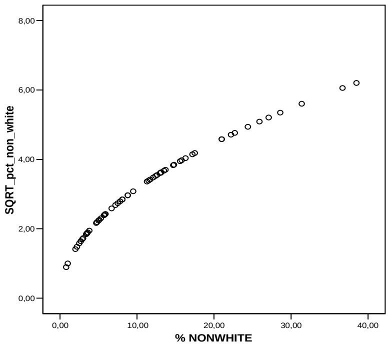
Robust Multippel Regresjon

X_1	RELATIVE HC POLLUTION POTENTIAL (natural log)
X_2	AVG. YEARLY PRECIP. INCHES
X_3	AVG. JANUARY TEMPERATURE, F
X_4	MEDIAN EDUCATION OF POP 25+
X_5	% NON-WHITE (square root)
X_6	POPULATION PER HOUSEHOLD
X_7	% 65 AND OVER
X_8	% SOUND HOUSING UNITS
X_9	PEOPLE PER SQUARE MILE (natural log)
X_{10}	AVG. JULY TEMPERATURE, F
X_{11}	% WHITE COLLAR EMPLOYMENT
X_{12}	% FAMILIES WITH INCOME<$3000 (negative reciprocal root)
X_{13}	AVG RELATIVE HUMIDITY, %

Multiple OLS regression with transformed variables: effect of transformation



ln of air pollution



Square root of % non-white

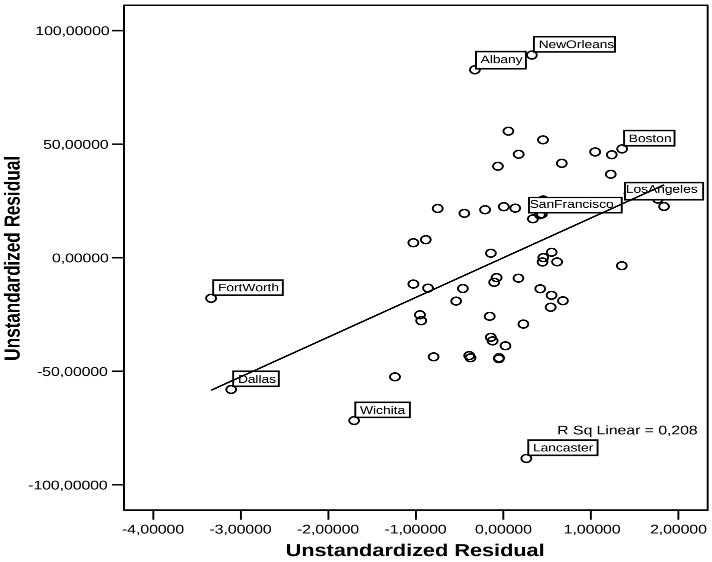
OLS with backward elimination gives

Dependent Variable: AGE-ADJUSTED MORTALITY/100K	B	Std. Error	t	Sig.
(Constant)	986,261	82,674	11,929	,000
LN_hc_pollution	17,469	4,636	3,768	,000
AVG. YEARLY PRECIP. INCHES	2,352	,640	3,677	,001
AVG. JANUARY TEMPERATURE, F	-2,132	,504	-4,228	,000
MEDIAN EDUCATION OF POP 25+	-17,958	6,204	-2,895	,005
SQRT_pct_non_white	27,335	4,398	6,215	,000

- Robust regression gives predicted y:
- $Y = 1001.8+17.77x_{1i}+2.32x_{2i}-2.11x_{3i}-19.1x_{4i}+26.2x_{5i}$

Multiple OLS regression with transformed variables

Leverage plot of residual from mortality (y) and residual of ln_air_pollution (x)



Los Angeles and San Francisco are no longer outliers

Four estimates of the relationship mortality – air pollution

Effect of air pollution

	OLS	Robust
1 variable	7.97	19.46
5 variables	17.47	17.77

- Note that in RR the bivariate regression comes pretty close to the result of the multivariate regression
- In the five-variable model there are new cases with influence on the line of regression
- Removing the 5 cases that have the highest leverage parameter (h_i) do not give substantial changes in the coefficients

Robust Regression vs Bounded Influence Regression

- Robust Regression protect against the effect of outliers (unusual y-values) if these do not go together with unusual x-values

- Bounded Influence Regression is designed to protect against influence from unusual combinations of x-values

BI - Bounded Influence Regression

- BI-methods are made to limit the influence of high leverage cases (large h_i = high leverage)
- The simplest way of doing this is to modify the Huber-weights or Tukey-weights in the IRLS procedure for RR (robust regression) with a factor based on the leverage statistic

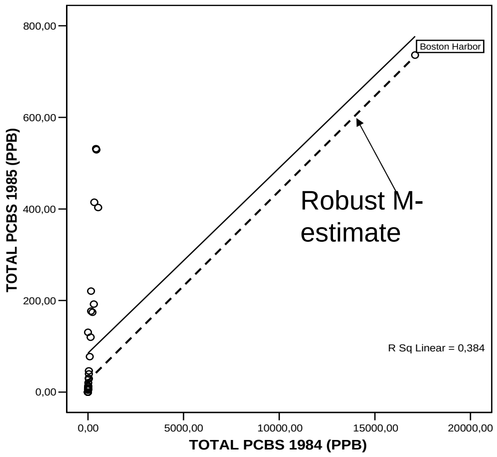
Bounded influence: modification of weights

- Expand the weight function with a weight based on the leverage statistic h_i
- $w^H_i = 1$ if $h_i \leq c^H$
- $w^H_i = (c^H/ h_i)$if $h_i > c^H$
- c^H is often set to the 90% percentile in the distribution of h_i
- Then the IRSL weight becomes $w_i\, w^H_i$ where w_i is either the Tukey- or Huber-weight that changes from iteration to iteration while w^H_i is constant

Bounded influence as a diagnostic tool

- Estimation of standard errors and test statistics becomes even more complicated than for the M-estimators mentioned above
- We can use BI estimates as a descriptive tool to check up on other estimates
- One (somewhat) extreme example: PCB pollution in river mouths in 1984 and 1984 (Hamilton table 6.4)

Fig 6.15 and 6.16 Hamilton



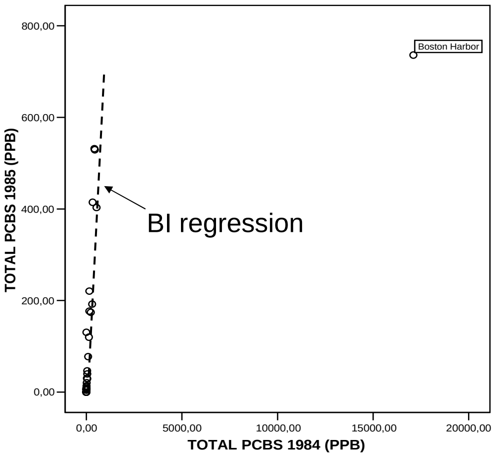
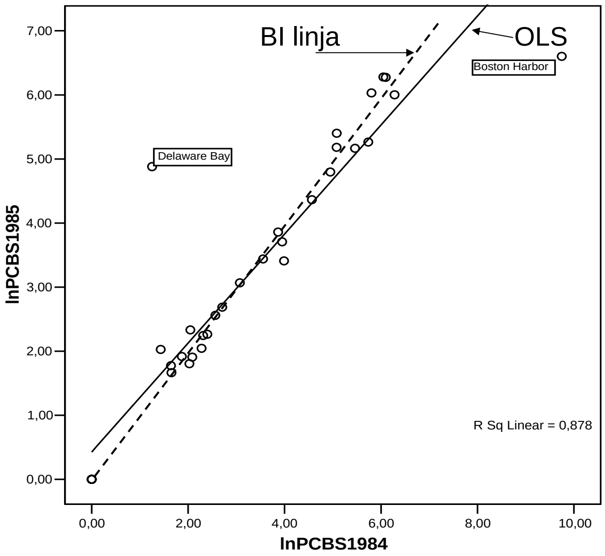


Fig 6.17 Hamilton

OLS and BI estimates with transformed variables give about the same result



Conclusions

- When data have many outliers robust methods will have better properties than OLS
  - They are more effective and give more accurate confidence intervals and tests of significance
- Robust regression can be used as a diagnostic tool
  - If OLS and RR agree we can have more confidence to the OLS results
  - If they disagree we will
    - Know that a problem exist
    - Have a model that fits the data better and identifies the outliers better
- Robust methods does not protect against problems that are due to curvilinear or non-linear models, heteroscedasticity, and autocorrelation

SOS3003
Applied data analysis for social science

Lecture notes on

Hamilton Ch 7 p217-234

Logistic regression I

Erling Berge
Department of sociology and political science
NTNU

LOGIT REGRESSION

- **Should be used if the dependent variable (Y) is a nominal scale**
- Here it is assumed that Y has the values 0 or 1
- The model of the conditional probability of Y, E[Y | X], is based on the logistic function

  (E[Y | X] is read “the expected value of Y given the value of X”)
- But

  Why cannot E[Y | X] be a linear function also in this case?

The linear probability model: LPM

- The linear probability model (LPM) of Y_i when Y_i can take only two values (0, 1) assumes that we can interpret E[Y_i | **X**] as a probability
- $E[Y_i \mid \mathbf{X}] = b_0 + \Sigma_j b_j x_{ji} = \Pr[Y_i = 1]$
- This leads to severe problems:

Are the assumptions of a linear regression model satisfied for the LPM?

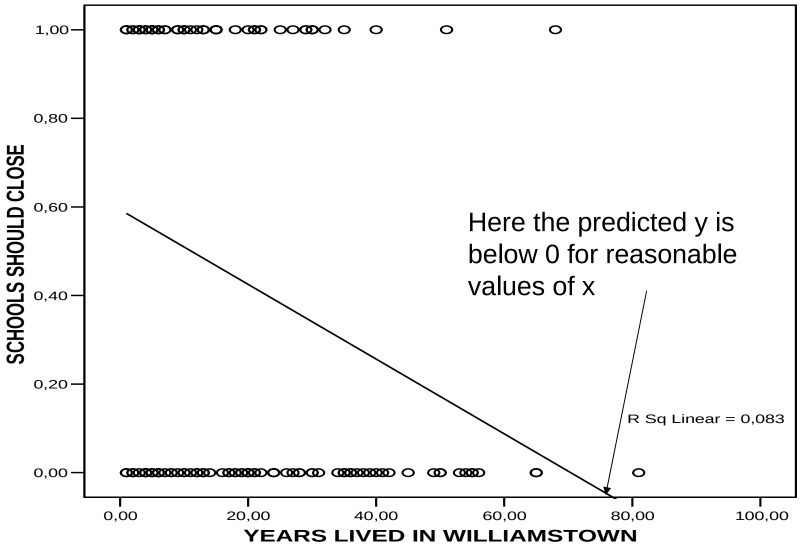
- One assumptions of the LPM is that the residual, e_i satisfies the requirements of OLS
- The the residual must be either
  - $e_i = 1 – (b_0 + \Sigma_j\, b_j\, x_{ji})$ or
  - $e_i = 0 – (b_0 + \Sigma_j\, b_j\, x_{ji})$
- This means that there is heteroscedasticity (the residual varies with the size of the values on the x-variables)
- There are estimation methods that can get around this problem (such as 2-stage weighted least squares method)
- One example of LPM:

OLS regression of a binary dependent variable on the independent variable ”years lived in town”

ANOVA tabell	Sum of Squares	df	Mean Square	F	Sig.
Regression	3,111	1	3,111	13,648	,000(a)
Residual	34,418	151	,228		
Total	37,529	152			

Dependent Variable: SCHOOLS SHOULD CLOSE	B	Std. Error	t	Sig.
(Constant)	,594	,059	10,147	,000
YEARS LIVED IN TOWN	-,008	,002	-3,694	,000

The regression looks OK in these tables



Here the predicted y is below 0 for reasonable values of x

Scatter plot with line of regression. Figure 7.1 Hamilton

Conclusion: LPM model is wrong

- The example shows that for reasonable values of the x variable we can get values of the predicted y where $E[Y_i \mid \mathbf{X}] > 1$ or $E[Y_i \mid \mathbf{X}] < 0$,
- For this there is no remedy
- <u>LPM is for substantial reasons a wrong model</u>
- We need a model where we always will have $0 \le E[Y_i \mid \mathbf{X}] \le 1$
- The logistic function can provide such a model

The logistic function

The general logistic function is written

- $Y_i = \alpha/(1+\gamma*\exp[-\beta X_i]) + \varepsilon_i$

$\alpha>0$ provides an upper limit for Y

this means that $0<Y< \alpha$

γ determines the horizontal point for rapid growth

If we determines that $\alpha = 1$ and $\gamma = 1$

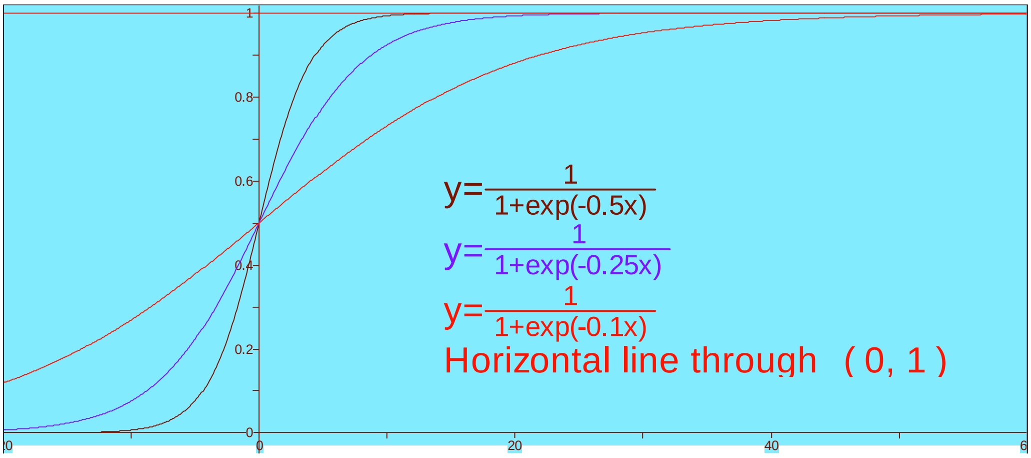
One will always find that

- $0 < 1/(1+\exp[-\beta X_i]) < 1$

The logistic function will for all values

of x lie between 0 and 1

Logistic curves for different β



β determines how rapidly the curve grows

MODEL (1)

Definitions:

- The probability that person no i shall have the value 1 on the variable Y will be written Pr(Y_i =1). Then Pr(Y_i ≠ 1) = 1 - Pr(Y_i=1)
- The odds that person no i shall have the value 1 on the variable Y, here called O_i, is the ratio between two probabilities

$$\mathrm{O}_i\left(y_i=1\right)=\frac{\Pr\left(y_i=1\right)}{1-\Pr\left(y_i=1\right)}=\frac{p_i}{1-p_i}$$

MODEL (2)

Definitions:

- The LOGIT, L_i , is the natural logarithm of the odds, O_i , for person no i:

  L_i = ln(O_i)
- The model assumes that L_i is a linear function of the explanatory variables x_j ,
- i.e.:
- $L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$, where j=1,..,K-1, and i=1,..,n

MODEL (3)

- Let X = (the collection of all x_j), then the probability of Y_i = 1 for person no i

$$\Pr(y_i = 1) = E\left[y_i \mid \mathrm{X}\right] = \frac{1}{1+\exp\left(-L_i\right)} = \frac{\exp(L_i)}{1+\exp(L_i)}$$

$$\text{where } L_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji}$$

The graph of this relationship is useful for the interpretation what a change in x means

MODEL (4)

In the model $Y_i = E[Y_i \mid X] + \varepsilon_i$ the error is either

- $\varepsilon_i = 1 - E[Y_i \mid X]$ with probability $E[Y_i \mid X]$

  (since $\Pr(Y_i = 1) = E[Y_i \mid X]$),

or the error is

- $\varepsilon_i = - E[Y_i \mid X]$ with probability $1 - E[Y_i \mid X]$

- Meaning that the error has a distribution known as the binomial distribution with

  $p_i = E[Y_i \mid X]$

Estimation by the ML method

- The method used to estimate the parameters in the model is Maximum Likelihood
- The ML-method gives us the parameters that maximize the Likelihood of finding just the observations we have got
- This likelihood we call $\mathcal{L}$
- The criterion for choosing regression parameters is that the likelihood becomes as large as possible

Maximum Likelihood (1)

- The Likelihood equals the product of the probability of each observation. For a dichotomous variable where Pr(Y_i = 1)=P_i this can be written

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ P_i^{Y_i} \left(1 - P_i\right)^{(1-Y_i)} \right\}$$

Maximum Likelihood (2)

- It is easier to maximize the likelihood $\mathcal{L}$ if one uses the natural logarithm of $\mathcal{L}$:

$$\ln\left(\mathcal{L}\right) = \sum_{i=1}^{n} \left\{ y_i \ln P_i + \left(1 - y_i\right) \ln\left(1 - P_i\right) \right\}$$

- The natural logarithm of $\mathcal{L}$ is called the LogLikelihood, It may be called $\mathcal{LL}$.
- $\mathcal{LL}$ has a central role in logistic regression.

Maximum Likelihood (3)

- The LogLikelihood $\mathcal{LL}$ will always be negative
- Maximizing $\mathcal{LL}$ is the same as minimizing the positive LogLikelihood; i.e. minimizing -$\mathcal{LL}$
- Finding parameter values that minimizes - $\mathcal{LL}$ can be done only by ”trial and error”, using an iterative procedure

Iterative estimation

			Coefficients	
From Hamilton Tabell 7.1	Iteration	**-2 Log Likelihood**	Constant	lived
Initial	0	209,212	-,276	
Step	1	195,684	,376	-,034
	2	195,269	,455	-,041
	3	195,267	,460	-,041
	4	195,267	,460	-,041

Note the column titled -2 LogLikelihood

Footnotes to the table

- Step 0: Point of departure is a model with a constant and no variables
- **Iterative estimation**
  - Estimation ends at iteration no 4 since the parameter estimates changed less than 0.001
- The Wald statistic that SPSS provides equals the square of the “t” that Hamilton (and STATA) provides (Wald = t^2)

Logistic model instead of LPM

OLS regression (slide 6 above)

Dependent Variable: SCHOOLS SHOULD CLOSE	B	Std. Error	t	Sig.
(Constant)	,594	,059	10,147	,000
YEARS LIVED IN TOWN	-,008	,002	-3,694	,000

Logistic regression

Dependent: Schools should close	B	S.E.	Wald	df	Sig.	Exp(B)
Lived in town	-,041	,012	11,399	1	,001	,960
Constant	,460	,263	3,069	1	,080	1,584

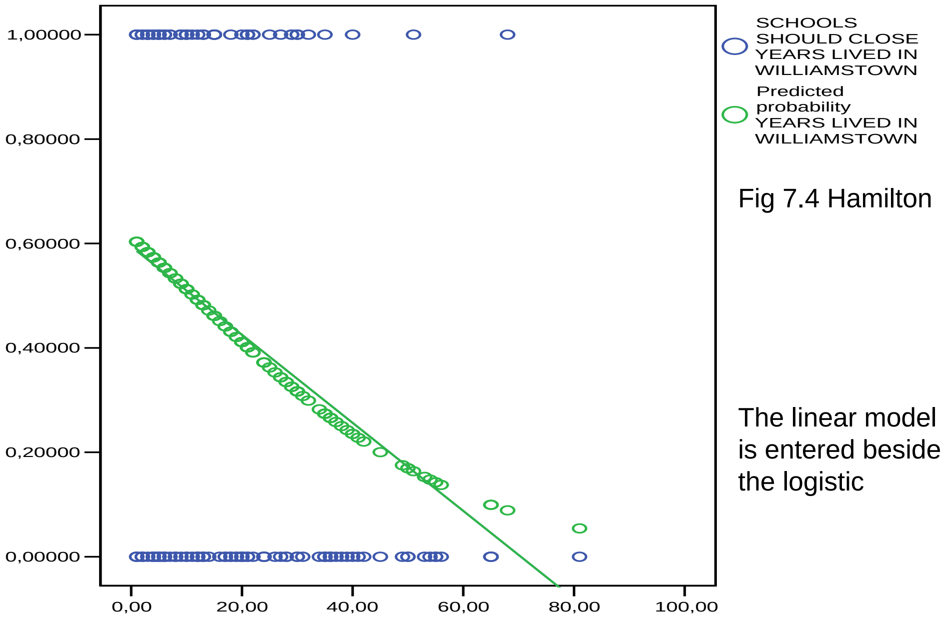


Fig 7.4 Hamilton

The linear model is entered beside the logistic

TESTING

Two tests are useful

- (1) The Likelihood ratio test
  - This can be used analogous to the F-test
- (2) Wald test
  - The square root of this can be used analogous to the t-test

Interpretation (1)

- The difference between the linear model and the logistic is large in the neighbourhood of 0 and 1
- LPM is easy to interpret: $Y_i = \beta_0$ when x_{1i}=0, and when x_{1i} increases with one unit Y_i increases with β_1 units
- The logistic model is more difficult to interpret. It is non-linear both in relation to the odds and the probability

ODDS and ODDS RATIOS

- The Logit, L_i, ($L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$) is defined as the natural logarithm of the odds

This means that

- odds = $O_i (Y_i=1) = \exp(L_i) = e^{L_i}$

and

- **Odds ratio**= $O_i (Y_i=1| L_i') / O_i (Y_i=1| L_i)$
  - where L_i' and L_i have different values on only one variable $x_{.j.}$

Interpretation (2)

- When all x equals 0 then $L_i = \beta_0$ This means that the odds for $y_i = 1$ in this case is $\exp\{\beta_0\}$
- If all x-variables are kept fixed (they sum up to a constant) while x_1 increases with 1, the odds for $y_i = 1$ will be multiplied by $\exp\{\beta_1\}$
- This means that it will change with

  $100(\exp\{\beta_1\} – 1)$ %
- The probability $\Pr\{y_i = 1\}$ will change with a factor affect by all elements in the logit

Logistic regression: assumptions

- The model is correctly specified
    - The logit is linear in its parameteres
    - All relevant variables is included
    - No irrelevant variables are included
- x-variables are measured without error
- Observations are independent
- No perfect multicollinearity
- No perfect discrimination
- Sufficiently large sample

Assumptions that cannot be tested

- Model specification
    - All relevant variables are included
- x-variables are measured without error
- Observations are independent

Two will be tested automatically.

If the model can be estimated there is

- No perfect multicollinearity and
- No perfect discrimination

LOGISTIC REGRESSION
Statistical problems may be due to

- Too small a sample
- High degree of **multicollinearity**
  - Leading to large standard errors (imprecise estimates)
  - M is discovered and treated in the same way as in OLS regression
- High degree of **discrimination** (or separation)
  - Leading to large standard errors (imprecise estimates)
  - Will be discovered automatically by SPSS

Discrimination/ separation

- Problems with discrimination appear when we for a given x-value get almost perfect prediction of the y-value (nearly all with a given x-value have the same y-value)
- In SPSS it may produce the following message:

Warnings

• There is possibly a quasi-complete separation in the data. Either the maximum likelihood estimates do not exist or some parameter estimates are infinite.
• The NOMREG procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

Discrimination in Hamilton table 7.5

- Odds for weaker requirements is 44/202 = 0,218 among women without small children
- Odds for weaker requirement is 0/79 = 0 among women with small children
- Odds rate is 0/0,218 = 0 hence exp{b_{woman}}=0
- This means that b_{woman} = minus infinity

	Women without small children	Women with small children
No weaker requirements	202	79
Weaker requirements OK	44	0

Logistic regression

- If the assumptions are satisfied logistic regression will provide normally distributed, unbiased and efficient (minimal variance) estimates of the parameters

SOS3003
Applied data analysis for social science

Lecture notes on

Hamilton Ch 7 p217-242

Logistic regression II

Erling Berge

Department of sociology and political science

NTNU

Definitions I

Definitions:

- The probability that person no i shall have the value 1 on the variable Y will be written $\Pr(Y_i = 1)$. Then $\Pr(Y_i \neq 1) = 1 - \Pr(Y_i = 1)$
- The odds that person no i shall have the value 1 on the variable Y, here called O_i, is the ratio between two probabilities

$$\mathrm{O}_i\left(y_i = 1\right) = \frac{\Pr\left(y_i = 1\right)}{1 - \Pr\left(y_i = 1\right)} = \frac{p_i}{1 - p_i}$$

Definitions II

Definitions:

- The LOGIT, L_i , is the natural logarithm of the odds, O_i , for person no i:

  $L_i = \ln(O_i)$
- The model assumes that L_i is a linear function of the explanatory variables x_j ,
- i.e.:
- $L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$, where j=1,..,K-1, and i=1,..,n

Logistic regression: assumptions

- The model is correctly specified
  - The logit is linear in its parameters
  - All relevant variables are included
  - No irrelevant variables are included
- x-variables are measured without error
- Observations are independent
- No perfect multicollinearity
- No perfect discrimination
- Sufficiently large sample

Assumptions that cannot be tested

- Model specification
  - All relevant variables are included
- x-variables are measured without error
- Observations are independent

Two will be tested automatically.

If the model can be estimated there is

- No perfect multicollinearity and
- No perfect discrimination

LOGISTIC REGRESSION Statistical problems may be due to

- Too small a sample
- High degree of **multicollinearity**
  - Leading to large standard errors (imprecise estimates)
  - Multicollinearity is discovered and treated in the same way as in OLS regression
- High degree of **discrimination** (or separation)
  - Leading to large standard errors (imprecise estimates)
  - Will be discovered automatically by SPSS

Assumptions that can be tested

- Model specification
  - logit is linear in the parameters
  - no irrelevant variables are included
- Sufficiently large sample
  - What is "sufficiently large" depends on the number of different patterns in the sample and how cases are distributed across these
- Testing implies an assessment of whether statistical problems leads to departure from the assumptions

LOGISTIC REGRESSION: TESTING (1)

Two tests are useful

- (1) The Likelihood ratio test
  - This can be used analogous to the F-test
- (2) Wald test
  - The square root of this can be used analogous to the t-test

LOGISTIC REGRESSION: TESTING (2)

- The LikeLihood Ratio test :
- The ratio between two Likelihoods equals the difference between two **Log**Likelihoods
- The difference between the **LogLikelihood** ($\mathcal{LL}$) of two **nested** models, estimated on **the same data**, can be used to test which of two models fits the data best, just like the F-statistic is used in OLS regression
- The test can also be used for singe regression coefficients (single variables). In small samples it has better properties than the Wald statistic

LOGISTIC REGRESSION: TESTING (3)

The LikeLihood Ratio test statistic

- χ^2_H **= -2[**$\mathcal{LL}$**(model1) -** $\mathcal{LL}$**(model2)]**

will, if the null hypothesis of no difference between the two models is correct, be distributed approximately (for large n) as the chi-square distribution with number of degrees of freedom equal to the difference in number of parameters in the two models (H)

Example of a Likelihood Ratio test

- Model 1: just constant
- Model 2: constant plus one variable

- $\chi^2_H = -2[\mathcal{LL}(\text{model1}) - \mathcal{LL}(\text{model2})]$
  $= -2\mathcal{LL}(\text{model1}) + 2\mathcal{LL}(\text{model2})$
- Find the value of the ChiSquare and the number of degrees of freedom
- e.g.: LogLikelihood (mod1) = 209,212/(-2)
- LogLikelihood (mod2) = 195,267/(-2)

From Tab 7.1: **-2 Log Likelihood**
209,212
195,684
195,269
195,267
195,267

LOGISTIC REGRESSION: TESTING (4)

The Wald test

- The Wald (or chisquare) test statistic provided by SPSS = $t^2 = (b_k / SE(b_k))^2$ (where t is the t used by Hamilton) can be used for testing single parameters similarly to the t-statistic of the OLS regression
- If the null hypothesis is correct, t will (for large n) in logistic regression be approximately normally distributed
- If the null hypothesis is correct, the Wald statistic will (for large n) in logistic regression be approximately chisquare distributed with df=1

Excerpt from Hamilton Table 7.2

Iterasjon	**-2 Log likelihood**					
0	209,212					
1	152,534					
2	149,466					
3	149,382					
4	149,382					
5	149,382					
Variables	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Lived	-,046	,015	9,698	1	,002	,955
Educ	-,166	,090	3,404	1	,065	,847
Contam	1,208	,465	6,739	1	,009	3,347
Hsc	2,173	,464	21,919	1	,000	8,784
Constant	1,731	1,302	1,768	1	,184	5,649

Confidence interval for parameter estimates

- Can be constructed based on the fact that the square root of the Wald statistic approximately follows a normal distribution with 1 degree of freedom
- b_k - t_α*SE(b_k) < β_k < b_k + t_α*SE(b_k)

  where t_α is a value taken from the table of the **normal distribution** with level of significance equal to α

Can be constructed based on the t-distribution (1)

- If a table of the normal distribution is missing one may use the **t-distribution** since the t-distribution is approximately normally distributed for large n-K (e.g. for n-K > 120)

Excerpt from Hamilton Table 7.3

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	lived	-,047	,017	7,550	1	,006	,954
	educ	-,206	,093	4,887	1	,027	,814
	contam	1,282	,481	7,094	1	,008	3,604
	hsc	2,418	,510	22,508	1	,000	11,223
	female	-,052	,557	,009	1	,926	,950
	kids	-,671	,566	1,406	1	,236	,511
	nodad	-2,226	,999	4,964	1	,026	,108
	Constant	2,894	1,603	3,259	1	,071	18,060

More from Hamilton Table 7.3

Iteration		-2 Log likelihood	Coefficients							
			Const	lived	educ	contam	hsc	female	kids	nodad
Step0		209,212	-0,276							
Step1	1	147,028	1,565	-,027	-,130	,782	1,764	-,015	-,365	-1,074
	2	141,482	2,538	-,041	-,187	1,147	2,239	-,037	-,580	-1,844
	3	141,054	2,859	-,046	-,204	1,269	2,401	-,050	-,662	-2,184
	4	141,049	2,893	-,047	-,206	1,282	2,418	-,052	-,671	-2,225
	5	141,049	2,894	-,047	-,206	1,282	2,418	-,052	-,671	-2,226

Is the model in table 7.3 better than the model in table 7.2 ?

- **$\mathcal{LL}$(model in 7.3) = 141,049/(-2)**
- **$\mathcal{LL}$(model in 7.2) = 149,382/(-2)**

- χ^2_H = -2[$\mathcal{LL}$(model 7.2) - $\mathcal{LL}$(model 7.3)]
- Find χ^2_H value
- Find H
- Look up the table of the chisquare distribution

The model of the probability of observing y=1 for person i

$$\Pr(y_i = 1) = E\left[y_i \mid x\right] = \frac{1}{1+\exp\left(-L_i\right)} = \frac{\exp(L_i)}{1+\exp(L_i)}$$

where the logit $L_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji}$ is a linear function of the explanatory variables

It is not easy to interpret the meaning of the β coefficients just based on this formula

The odds ratio

- The odds ratio, **O**, can be interpreted as the relative effect of having one variable value rather than another
- e.g. if x_{ki} = t+1 in L_i’ and x_{ki} = t in L_i
- **O** = O_i (Y_i=1| L_i’)/ O_i (Y_i=1| L_i)
  = exp[L_i’]/ exp[L_i]
  = exp[β_k]
- Why β_k ?

The odds ratio : example I

- The Odds for answering yes =
  $e^{b_0+b_1*\text{Alder}+b_2*\text{Kvinne}+b_3*\text{E.utd}+b_4*\text{Barn i HH}}$

- The odds ratio for answering yes between women and men =

$$\frac{e^{b_0+b_1*Alder+b_2*1+b_3*E.utd+b_4*Barn\_i\_HH}}{e^{b_0+b_1*Alder+b_2*0+b_3*E.utd+b_4*Barn\_i\_HH}}=e^{b_2}$$

Remember the rules of power exponents

The odds ratio : example II

- The Odds for answering yes given one year of extra education

$$\frac{e^{b_0+b_1*Alder+b_2*Kvinne+b_3*\left(E.utd+1\right)+b_4*Barn\_i\_HH}}{e^{b_0+b_1*Alder+b_2*Kvinne+b_3*E.utd+b_4*Barn\_i\_HH}}=e^{b_3}$$

Remember the rules of power exponents

Example from Hamilton table 7.2

- What is the odds ratio for yes to closing the school from one year extra education?
- The odds ratio is the ratio of two odds where one odds is the odds for a person with one year extra education

$$\frac{e^{b_0 + b_1 * ÅrBuddIByen + b_2 * (Utdanning + 1) + b_3 * UreiningEigEigedom + b_4 * MangeHSCmøter}}{e^{b_0 + b_1 * ÅrBuddIByen + b_2 * Utdanning + b_3 * UreiningEigEigedom + b_4 * MangeHSCmøter}}$$

$$= \frac{e^{b_2 * (Utdanning + 1)}}{e^{b_2 * Utdanning}} = e^{b_2}$$

Example from Hamilton table 7.2 cont.

- Odds ratio = Exp{b2} = exp(-0,166) = 0,847
- One extra year of education implies that the odds is reduced with a factor of 0.847
- One may also say that the odds has increased with a factor of

  100(0,847-1)% = -15,3%
- Meaning that it has declined with 15,3%

Conditional Effect Plot

- Set all x-variables except x_k to fixed values and enter these into the equation for the logit
- Plot Pr(Y=1) as a function of x_k i.e.
- $P = 1/(1+\exp[-L]) = 1/(1+\exp[-konst - b_k x_k])$

  for all reasonable values of x_k ,

  “konst” is the constant obtained by entering into the logit the fixed values of variables other than x_k

Excerpt from Hamilton Table 7.4

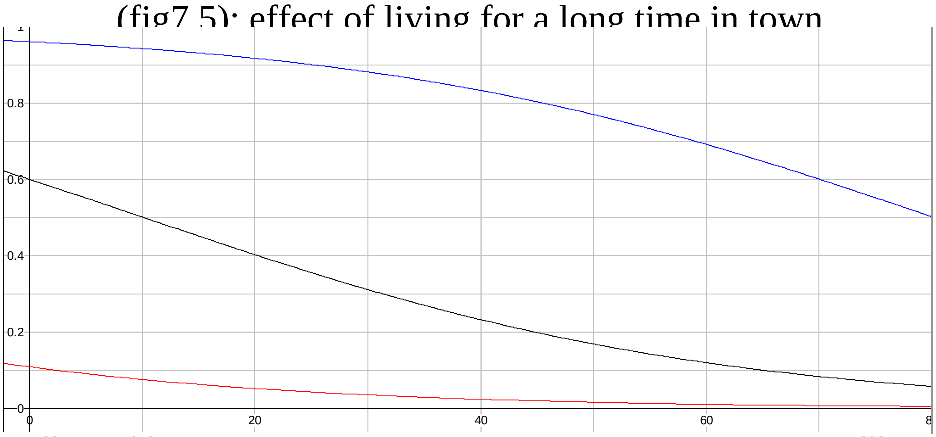
	B	S.E.	Wald	df	Sig.	Exp(B)	Minimum	Maximum	Mean
lived	-,040	,015	6,559	1	,010	,961	1,00	81,00	19,2680
educ	-,197	,093	4,509	1	,034	,821	6,00	20,00	12,9542
contam	1,299	,477	7,423	1	,006	3,664	,00	1,00	,2810
hsc	2,279	,490	21,591	1	,000	9,763	,00	1,00	,3072
nodad	-1,731	,725	5,696	1	,017	,177	,00	1,00	,1699
Constant	2,182	1,330	2,692	1	,101	8,866			

Logit:

L = 2.182 -0.04*lived -0.197*educ +1.299*contam +2.279*hsc -1.731*nodad

Here we let ”lived” vary and set in reasonable values for other variables

Conditional effect plot from Hamilton table 7.4 (fig7.5): effect of living for a long time in town

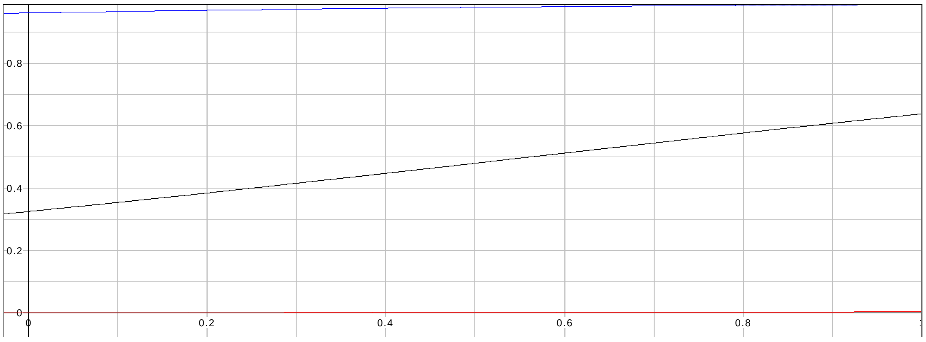


y=1/(1+exp(-(2.182-0.04x-0.197×12.95+1.299×0.28+2.279×0.31-1.731×0.17))) Mean

y=1/(1+exp(-(2.182-0.04x-0.197×12.95+1.299×1+2.279×1-1.731×0))) Max

y=1/(1+exp(-(2.182-0.04x-0.197×12.95+1.299×0+2.279×0-1.731×1))) Min

Conditional effect plot from Hamilton table 7.4 (fig7.6): effect of pollution on own land



y=1/(1+exp(-(2.182-0.04×19.27-0.197×12.95+1.299x+2.279×0.31-1.731×0.17))) Mean

y=1/(1+exp(-(2.182-0.04×1-0.197×6+1.299x+2.279×1-1.731×0))) Max

y=1/(1+exp(-(2.182-0.04×81-0.197×20+1.299x+2.279×0-1.731×1))) Min

Coefficients of determination

- Logistic regression does not provide measures comparable to the coefficient of determination in OLS regression
- Several measures analogous to R^2 have been proposed
- They are often called pseudo R^2
- Hamilton uses Aldrich and Nelson’s

  pseudo $R^2 = \chi^2/(\chi^2+n)$

  where χ^2 = test statistic for the test of the whole model against a model with just a constant and n= the number of cases

Some pseudo R^2 in SPSS

- SPSS reports Cox and Snell, Nagelkerke, and in multinomial logistic regression also McFadden’s proposal for R^2
- Aldrich and Nelson’s pseudo R^2 can easily be computed by ourselves [pseudo $R^2 = \chi^2/(\chi^2+n)$]

Model Summary

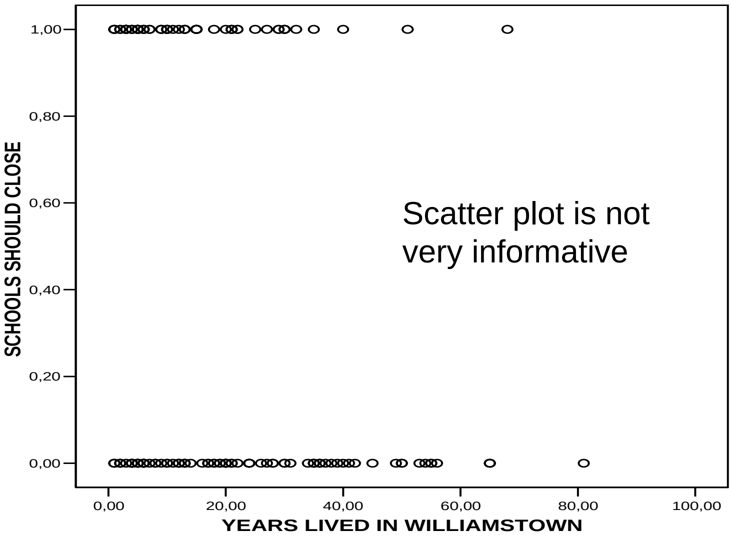
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	***	***	***

Pseudo R-Square	
Cox and Snell	***
Nagelkerke	***
McFadden	***

Statistical problem: linearity of the logit

- Curvilinearity of the logit can give biased parameter estimates
- Scatter plot for y - x is not informative since y only has 2 values
- To test if the logit is linear in an x-variable one may do as follows
  - Group the x variable
  - For every group find average of y and compute the logit for this value
  - Make a graph of the logits against the grouped x

Y=”Closing school” vs. x= ”Years lived in town”

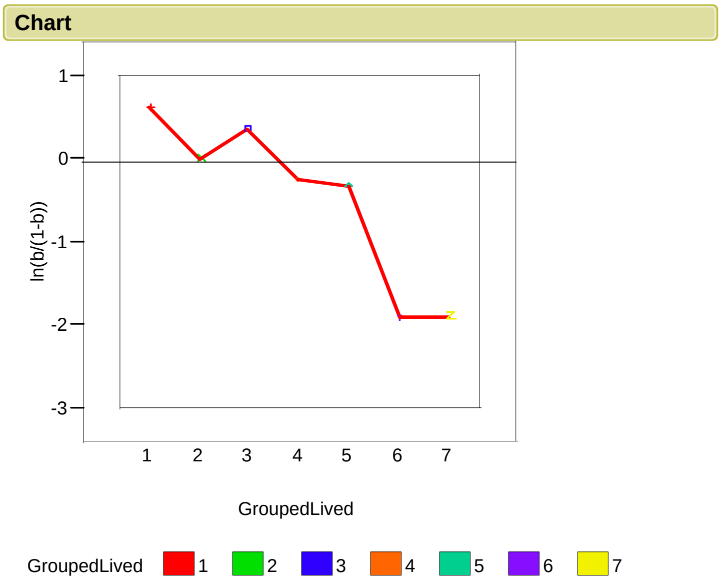


Linearity in logit: example

SCHOOLS SHOULD CLOSE		YEARS LIVED IN WILLIAMSTOWN (Banded)						
		<= 3	4-6	7-11	12-22	23-33	34-44	45+
N	OPEN	7	14	7	22	11	13	13
N	CLOSE	13	14	10	17	8	2	2
Within group	Mean (=p)	,65	,50	,59	,44	,42	,13	,13
Logit	Ln(p/(1-p))	0,619	0	0,364	-0,241	-0,323	-1,901	-1,901

Is the logit linear in ”years lived in town”?

Maybe!



In case of curvilinearity the odds ratio is non-constant

Assume the logit is curvilinear in education. Then the odds ratio for answering yes, adding one year of education, is:

$$\frac{e^{b_0+b_a*Alder+b_k*Kvinne+b_{utd}*(E.utd+1)+b_{utd\,2}*(E.utd+1)^2}}{e^{b_0+b_a*Alder+b_k*Kvinne+b_{utd}*E.utd+b_{utd\,2}*E.utd^2}} =$$

$$\frac{e^{b_{utd}+b_{utd\,2}*(E.utd^2+2E.utd+1)}}{e^{b_{utd\,2}*E.utd^2}} = \frac{e^{b_{utd}+b_{utd\,2}*(2E.utd+1)}}{e^0} = e^{b_{utd}+b_{utd\,2}*(2E.utd+1)}$$

Statistical problems: influence

- Influence from outliers and unusual x-values are just as problematic in logistic regression as in OLS regression
- Transformation of x-variables to symmetry will minimize the influence of extreme variable values
- Large residuals are indicators of large influence

Influence: residuals

- There are several ways to standardize residuals
  - ”Pearson residuals”
  - ”Deviance residuals”
- Influence can be based on
  - Pearson residual
  - Deviance residual
  - Leverage (potential for influence): i.e. the statistic h_j

Diagnostic graphs

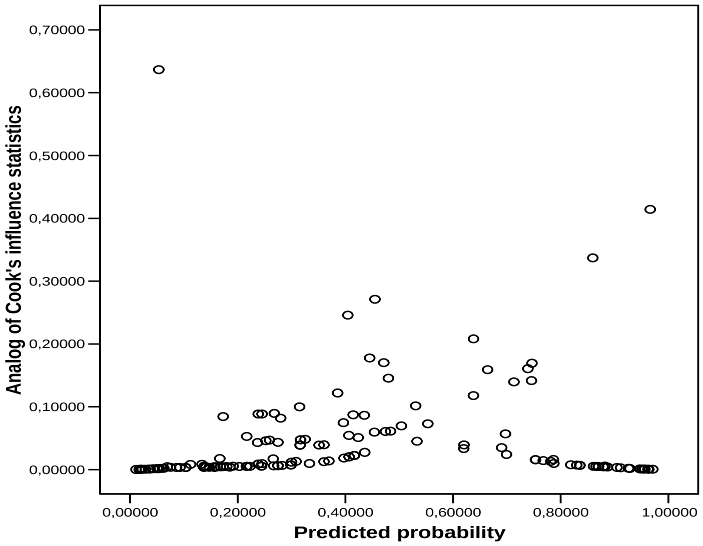
Outlier plots can be based on plots of estimated probability of Y_i=1 (estimated P_i) against

- Delta B , ΔB_j , or
- Delta Pearson Chisquare, $\Delta \chi^2_{P(j)}$, or
- Delta Deviance Chisquare, $\Delta \chi^2_{D(j)}$

SPSS output

- **Cook's = delta B in Hamilton**
  - The logistic regression analogue of Cook's influence statistic. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients.
- **Leverage Value = h in Hamilton**
  - The relative influence of each observation on the model's fit.
- **DfBeta(s)** is not used by Hamilton in logistic regression
  - The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.

Delta B



SPSS output from ”Save” (1)

- **Unstandardized Residuals**
  - The difference between an observed value and the value predicted by the model.
- **Logit Residual**

$$\tilde{e}_i = \frac{e_i}{\hat{\pi}_i(1-\hat{\pi}_i)}; where\ e_i = y_i - \hat{\pi}_i$$

π_i is the probability that $y_i = 1$; the “hat” means estimated value

SPSS output from ”Save” (2)

- **Standardized = Pearson residual**
  - The command ”standardized” will make SPSS write a variable called ZRE_1 nad labelled “Normalized residual”
  - This is the same as the Pearson residual in Hamilton
- **Studentized = [SQRT(delta deviance chisquare)]**
  - The command ”Studentized” will make SPSS write a variable called **SRE_1** and labelled “Standardized residual”
  - This is the same as the square root of ”delta Deviance chisquare” in Hamilton, i.e. ”delta Deviance chisquare” = $(\text{SRE\_1})^2$
- **Deviance = Deviance residual**
  - The command ”Deviance” will make SPSS write a variable called **DEV_1** and labelled “Deviance value”
  - This is the same as the deviance residual in Hamilton

Computation of $\Delta\chi^2_{P(i)}$

- Based on the quantities provided by SPSS we can compute ”delta Pearson chisquare”
- Where it says r_j in the formula we put in ZRE_1 and where it says h_j we put in LEV_1

$$\Delta\chi^2_{P(j)} = \frac{r_j^2}{(1-h_j)}$$

Computation of $\Delta\chi^2_{D(i)}$

Based on the quantities provided by SPSS we can compute ”Delta Deviance Chisquare”

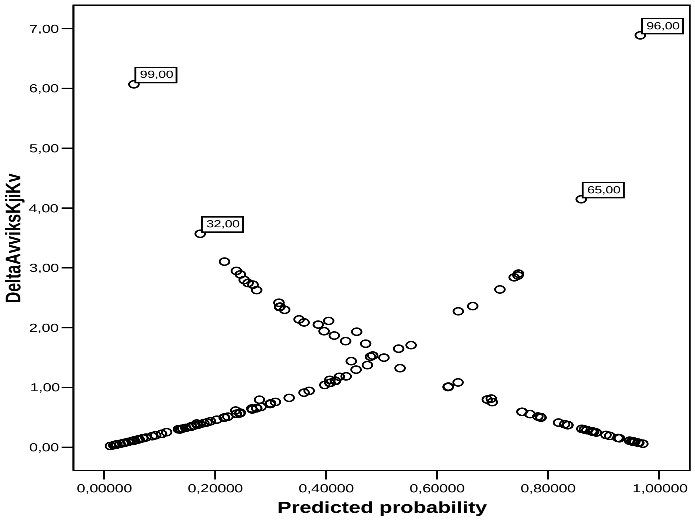
1. To find ”delta deviance chisquare” we square SRE_1

$$\Delta\chi^2_{D(j)} = SRE\_1 * SRE\_1$$

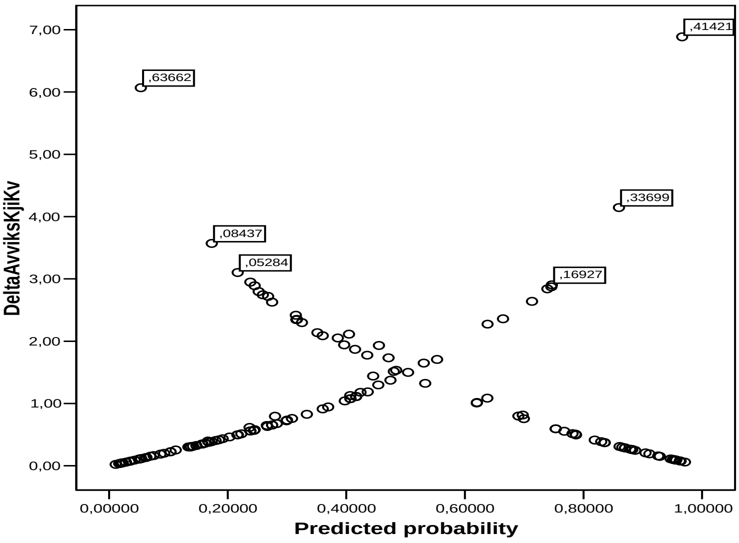
2. Alternatively we put in d_j=DEV_1 and h_j=LEV_1 in the formula

$$\Delta\chi^2_{D(j)} = \frac{d_j^2}{(1-h_j)}$$

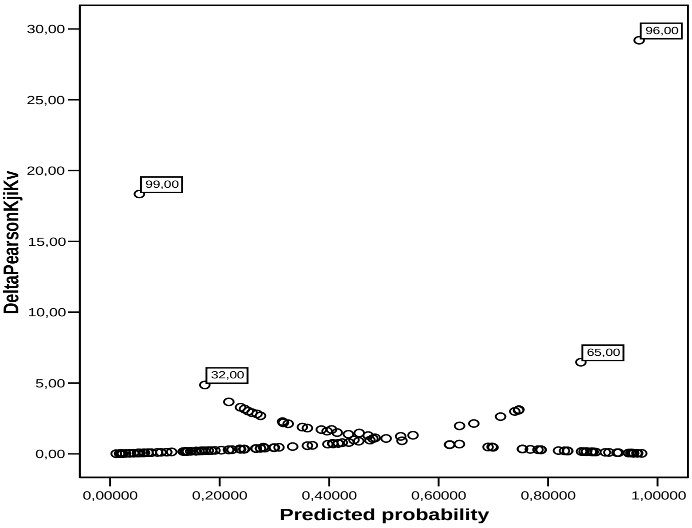
DeltaDevianceChisquare (with/CaseNO)



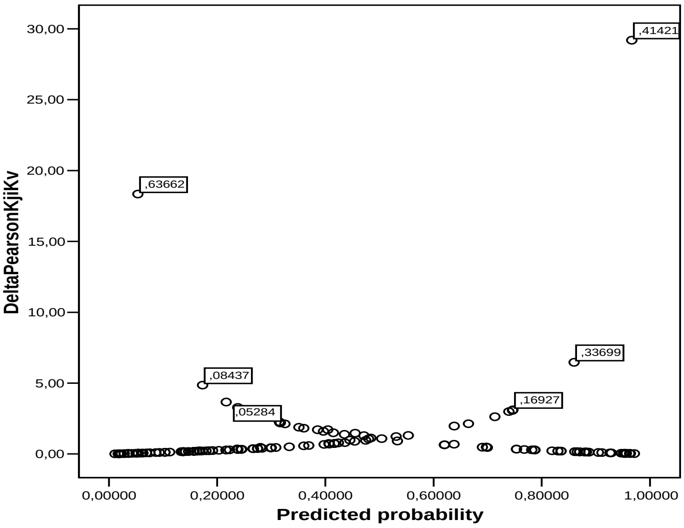
DeltaDevianceChisquare (with/delta B)



Delta Pearson Chisquare (with/CaseNO)



Delta Pearson Chisquare (with/ delta B)



Cases with large influence

Variables	CaseNo 96	CaseNo 65	CaseNo 99
Y=close	**1,00**	**,00**	**,00**
lived	68,00	40,00	1,00
educ	12,00	12,00	12,00
contam	,00	1,00	1,00
hsc	,00	1,00	1,00
nodad	,00	,00	,00
PRE_1	**,05**	**,86**	**,97**
COO_1	,64	,34	,41
RES_1	,95	-,86	-,97
SRE_1	2,46	-2,04	-2,62

Variables	CaseNo 96	CaseNo 65	CaseNo 99
ZRE_1	4,21	-2,48	-5,36
DEV_1	2,42	-1,98	-2,61
DFB0_1	-,32	,01	-,36
DFB1_1	,01	,00	,00
DFB2_1	,02	,01	,02
DFB3_1	-,08	-,15	-,18
DFB4_1	-,06	-,17	-,19
DFB5_1	-,08	,16	,14
DeltaPearsonKjiKv	18,34	6,47	29,20
DeltaAvviksKjiKv	6,07	4,14	6,89

From Cases to Patterns

- The figures shown previously are not identical to those you see in Hamilton
- Hamilton has corrected for the effect of identical patterns

Influence from a shared pattern of x-variables

- In a logistic regression with few variables many cases will have the same value on all x-variables. Every combination of x-variable values is called a pattern
- When many cases have the same pattern, every case may have a small influence, but collectively they may have unusually large influence on parameter estimates
- Influential patterns in x-values can give biased parameter estimates

Influence: Patterns in x-values

- Predicted value, and hence the residual will be the same for all cases with the same pattern
- Influence from pattern j can be found by means of
  - The frequency of the pattern
  - Pearson residual
  - Deviance residual
  - Leverage: i.e. the statistic h_j

Finding X-pattern by means of SPSS

- In the ”Data” – menu find the command ”Identify duplicate cases”
- Mark the x-variables that are used in the model and move them to ”Define matching cases by”
- Cross for ”Sequential count of matching cases in each group” and ”Display frequencies for created variables”
- This produces two new variables. One, ”MatchSequence”, numbers cases sequentially 1, 2, … where several patterns are identical. If the pattern is unique this variable has the value 0.
- The other variable, ”Primary…”, has the value 0 for duplicates and 1 for unique patterns

X-patterns in SPSS; Hamilton p238-242

	Frequency	Percent	Valid Percent	Cumulative Percent
Duplicate Case	21	13,7	13,7	13,7
Primary Case	132	86,3	86,3	100,0
Total	**153**	**100,0**	**100,0**	

Sequential count of matching cases	Frequency	Percent	Valid Percent	Cumulative Percent
0 [115 patterns with 1 case]	115	75,2	75,2	75,2
1 [17 patterns with 2 or 3 cases]	17	11,1	11,1	86,3
2 [17–4=13 patterns with 2 cases]	17	11,1	11,1	97,4
3 [4 patterns with 3 cases]	4	2,6	2,6	100,0
Total	**153**	**100,0**	**100,0**	

Hamilton table 7.6 Symbols

J	# unique patterns of x-values in the data (J<=n)
m_j	# cases with the pattern j (m>=1)
$\hat{P}_j$	Predicted probability of Y=1 for case with pattern j
Y_j	Sum of y-values for cases with pattern j (= # cases with pattern j and y=1)
r_j	Pearson residual for pattern j
χ^2_P	Pearson Chisquare statistic
d_j	Deviance residual for pattern j
χ^2_D	Deviance Chisquare statistic
h_i	Leverage for case i
h_j	Leverage for pattern j

New values for $\Delta\chi^2_{P(i)}$ and $\Delta\chi^2_{D(i)}$

- By ”Compute” one may calculate the Pearson residual (equation 7.19 in Hamilton) and delta Pearson chisquare (equation 7.24 in Hamilton) once more. This will provide the correct values
- The same applies for deviance residual (equation 7.21) and delta deviance chisquare (equation 7.25a)

Leverage and residuals (1)

- Leverage of a pattern is obtained as number of cases with the pattern times the leverage of a case with this pattern. The leverage of a case is the same as in OLS regression
- $h_j = m_j{*}h_i$
- Pearson residual can be found from

$$r_j = \frac{Y_j - m_j \hat{P}_j}{\sqrt{m_j \hat{P}_j \left(1 - \hat{P}_j\right)}}$$

Leverage and residuals (2)

- Deviance residual can be found from

$$d_j = \pm\sqrt{\left\{2\left[Y_j \ln\left(\frac{Y_j}{m_j \hat{P}_j}\right) + \left(m_j - Y_j\right)\ln\left(\frac{m_j - Y_j}{m_j\left(1 - \hat{P}_j\right)}\right)\right]\right\}}$$

Two Chi-square statistics

- Pearson Chi-square statistics

$$\chi_P^2 = \sum_{j=1}^{J} r_j^2$$

- Deviance Chi-square statistics

$$\chi_D^2 = \sum_{j=1}^{J} d_j^2$$

- Equations are the same for both cases and patterns

The Chisquare statistics

Both Chisquare statistics:

1. Pearson-Chisquare χ^2_P and
2. Deviance-Chisquare χ^2_D

- Can be read as a test of the null hypothesis of no difference between the estimated model and a “saturated model”, that is a model with as many parameters as there are cases/ patterns

Large values of measures of influence

- Measures of influence based on changes (Δ) in the statistic/ parameter value due to excluded cases with pattern j
  - ΔB_j “delta B” - analogue to Cook’s D
  - $\Delta\chi^2_{P(i)}$ “delta Pearson-Chisquare”
  - $\Delta\chi^2_{D(i)}$ “delta Deviance-Chisquare”

What is a large value of $\Delta\chi^2_{P(i)}$ and $\Delta\chi^2_{D(i)}$

- Both $\Delta\chi^2_{P(i)}$ and $\Delta\chi^2_{D(i)}$ measure how badly the model fits the pattern j. Large values indicates that the model would fit the data much better if all cases with this pattern were excluded
- Since both measures are distributed asymptotically as the chisquare distribution, values larger than 4 indicate that a pattern affects the estimated parameters “significantly”

ΔB_j “delta B”

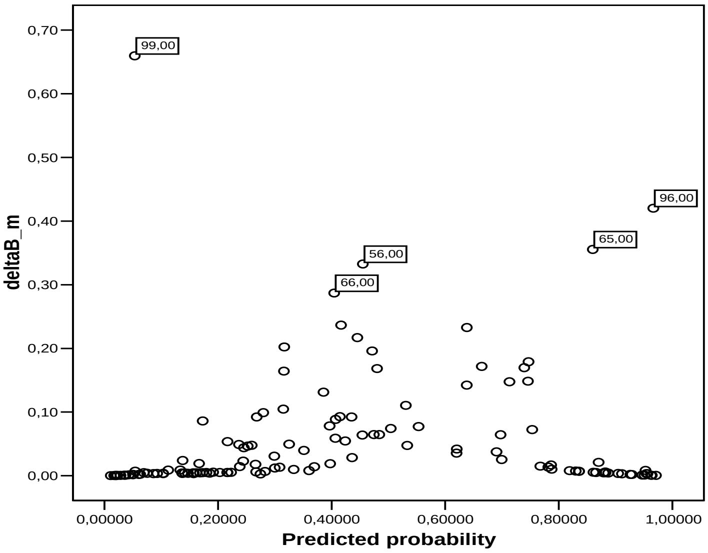
- Measures the standardized change in the estimated parameters (b_k) that obtain when all cases with a given pattern j are excluded

$$\Delta B_j = \frac{r_j^2 h_j}{\left(1 - h_j\right)^2}$$

Larger values means larger influence

ΔBj >= 1 must in any case be seen as ”large influence”

delta B (with caseNO)

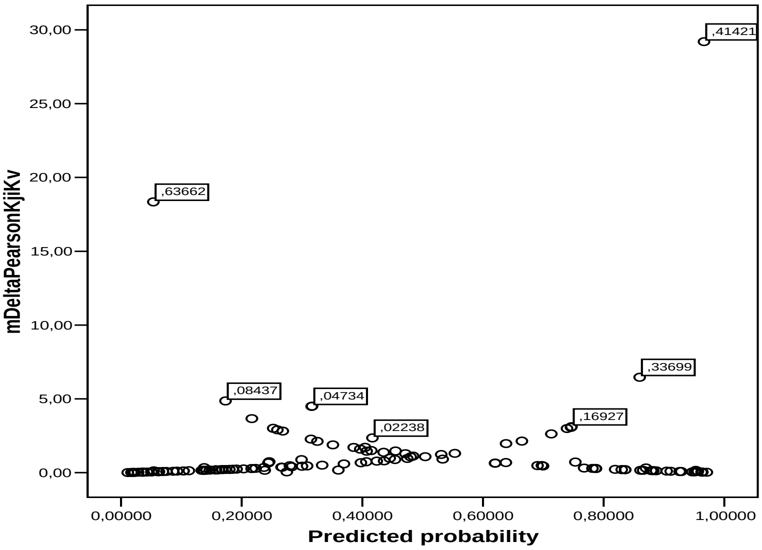


$\Delta\chi^2_{P(i)}$ “Delta Pearson Chisquare”

- Measures the reduction in Pearson χ^2 that obtains from excluding all cases with pattern j

$$\Delta\chi^2_{P(j)} = \frac{r_j^2}{\left(1 - h_j\right)}$$

Delta Pearson Chisquare (with delta B)



$\Delta\chi^2_{\text{D(i)}}$ “Delta Deviance Chisquare”

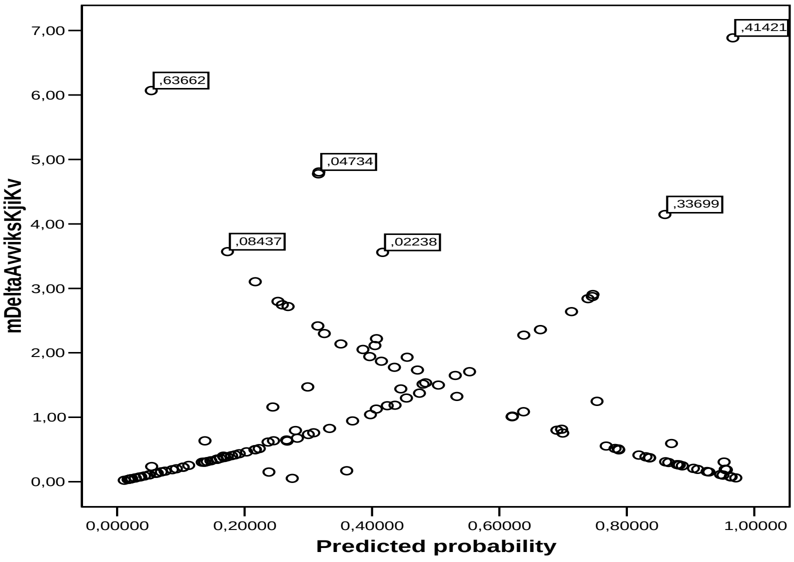
- Measures changes in deviance that obtains from excluding all cases with pattern j
- This is equivalent to

$$\Delta\chi^2_{D(j)} = \frac{d_j^2}{\left(1-h_j\right)}$$

$$\Delta\chi^2_{\mathcal{D}(j)} = -2\left[\mathcal{LL}_{\text{K}} - \mathcal{LL}_{\text{K(j)}}\right]$$

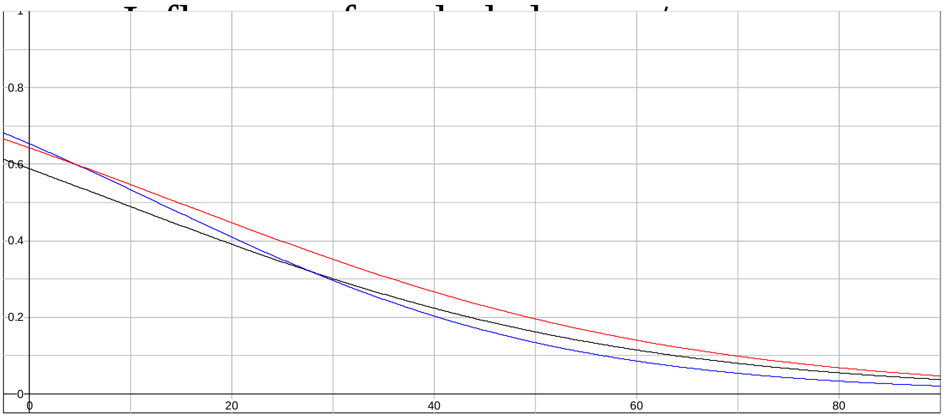
$\mathcal{LL}_{\mathbf{K}}$ is the LogLikelihood of a model with K parameters estimated on the whole sample and $\mathcal{LL}_{\mathbf{K(j)}}$ is from the estimate of the same model when all cases with pattern j are excluded

Delta Deviance Chisquare (with delta B)



Influence of excluded cases/patterns

	Logit coefficient		
Variables in the model	Sample	Excluding case 99 Δχ2P(i) =18,34	Excluding case 96 Δχ2P(i) =29,20
lived	-,040	-,045	-,052
educ	-,197	-,224	-,214
contam	1,299	1,490	1,382
hsc	2,279	2,492	2,347
nodad	-1,731	-1,889	-1,658
Constant	2,182	2,575	2,530
2*ℒℒ(modell)	**-142,652**	**-135,425**	**-136,124**



$y=1/(1+\exp(-(2.18-0.04x-0.2\times13+1.3\times0.28+2.28\times0.31-1.73\times0.17)))$

$y=1/(1+\exp(-(2.53-0.05x-0.21\times13+1.38\times0.28+2.35\times0.31-1.65\times0.17)))$

$y=1/(1+\exp(-(2.58-0.04x-0.22\times13+1.49\times0.28+2.49\times0.31-1.89\times0.17)))$

Conclusions (1)

Ordinary OLS do not work well for

dichotomous dependent variables since

- It is impossible to obtain normally distributed errors or homoscedasticity, and since
- The model predicts probabilities outside the interval [0-1]

The Logit model Is better

- Likelihood ratio tests statistic can be used to test nested models analogous to the F-statistic
- In large samples the chisquare distributed Wald statistic [or the normally distributed t=SQRT(Wald)] will be able to test single coefficients and provide confidence intervals
- There is no statistic similar to the coefficient of determination

Conclusions (2)

- Coefficient of estimated models can be interpreted by
  1. Log-odds (direct interpretation)
  2. Odds
  3. Odds ratio
  4. Probability (conditional effect plot)
- Non-linearity, case with influence, and multicollinearity leads to the same kinds of problems as in OLS regression (inaccurate or uncertain parameter values)
- Discrimination leads to problems of uncertain parameter values (large variance estimates)
- Diagnostic work is important