

SOS3003
**Applied data analysis for
social science**
Lecture note 05-2009

Erling Berge
Department of sociology and political
science
NTNU

Fall 2009

© Erling Berge 2009

1

Literature

- Regression criticism I
Hamilton Ch 4 p109-123

Fall 2009

© Erling Berge 2009

2

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

Fall 2009

© Erling Berge 2009

3

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

Fall 2009

© Erling Berge 2009

4

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

Fall 2009

© Erling Berge 2009

5

II GAUSS-MARKOV REQUIREMENTS (i)

- (1) x is known, without stochastic variation
- (2) Errors have an expected value of 0 for all i

$$\bullet E(\varepsilon_i) = 0 \quad \text{for all } i$$

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

Fall 2009

© Erling Berge 2009

6

II GAUSS-MARKOV REQUIREMENTS (ii)

(3) Errors have a constant variance for all i

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

This is called homoscedasticity

(4) Errors are uncorrelated with each other

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

This is called no autocorrelation

Fall 2009

© Erling Berge 2009

7

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)

Fall 2009

© Erling Berge 2009

8

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:
 - $\text{cov}(X_{ik}, \varepsilon_i) = 0$ for all i, k

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

Fall 2009

© Erling Berge 2009

9

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, calculate b_k and average these

Fall 2009

© Erling Berge 2009

10

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

$$\lim_{n \rightarrow \infty} [\Pr\{|b_k - \beta_k| < c\}] = 1$$

Fall 2009

© Erling Berge 2009

11

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

- Minimal variance or efficient estimator means that

$$\text{var}(b_k) < \text{var}(a_k)$$
 for all estimators a different from b
- Equivalent:

$$E[b_k - \beta_k]^2 < E[a_k - \beta_k]^2$$
 for all estimators a unlike b

Fall 2009

© Erling Berge 2009

12

Footnote 4: Biased estimators

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

Fall 2009

© Erling Berge 2009

13

Footnote 5: Non-linear estimators

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

Fall 2009

© Erling Berge 2009

14

III NORMALLY DISTRIBUTED ERROR TERM

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

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{for all } i$$

- 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)

Fall 2009

© Erling Berge 2009

15

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)

Fall 2009

© Erling Berge 2009

16

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

Fall 2009

© Erling Berge 2009

17

Consequences of problems (Hamilton, p113)

Requirement	Problem	Unwanted properties of estimates			
		Biased estimate of b	Biased estimate of SE _b	Invalid t&F-tests	High var[b]
Specification	Non-linear relationship	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

Fall 2009

© Erling Berge 2009

18

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)

Fall 2009

© Erling Berge 2009

19

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)

Fall 2009

© Erling Berge 2009

20

Correlation and scatter plot

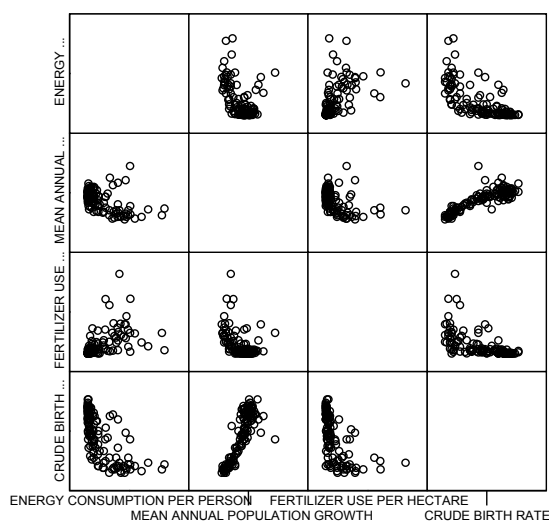
Data from 122 countries		ENERGY CONSUMPTION PER PERSON	MEAN ANNUAL POPULATION GROWTH	FERTILIZER USE PER HECTARE	CRUDE BIRTH RATE
ENERGY CONSUMPTION PER PERSON	Pearson Correlation	1	-.505	.533	-.689
	N	125	122	125	122
MEAN ANNUAL POPULATION GROWTH	Pearson Correlation	-.505	1	-.469	.829
	N	122	125	125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

Fall 2009

© Erling Berge 2009

21

Correlation and scatter plot



Fall 2009

© Erling Berge 2009

22

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})

Fall 2009

© Erling Berge 2009

23

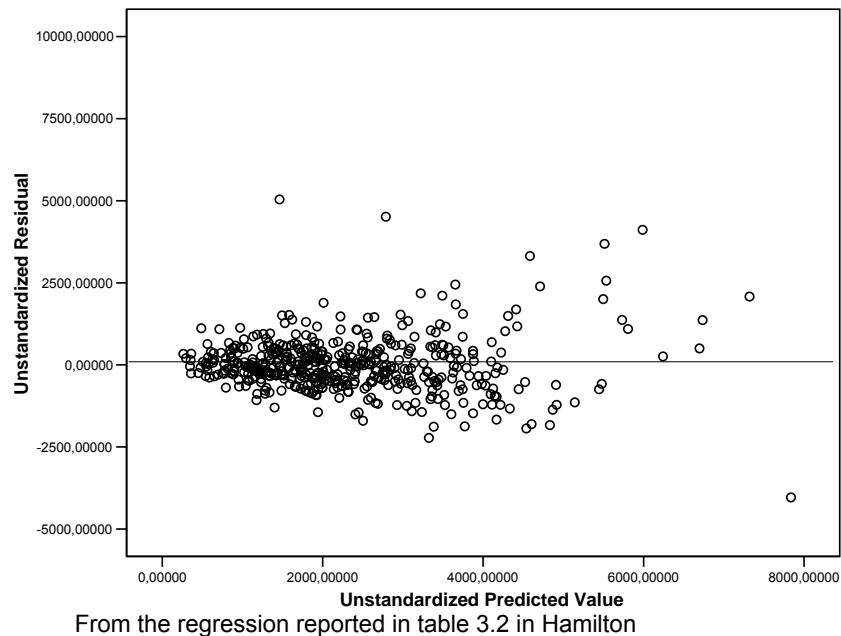
Example: Hamilton table 3.2

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

Fall 2009

© Erling Berge 2009

24



Fall 2009

© Erling Berge 2009

25

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 increases 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

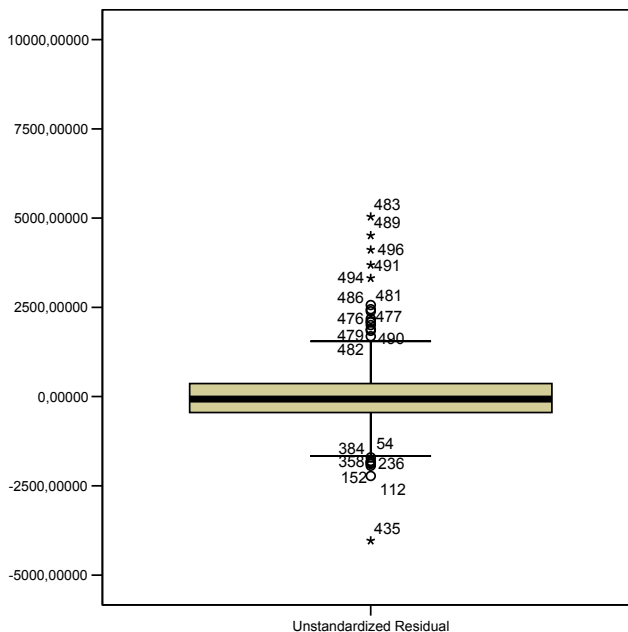
Fall 2009

© Erling Berge 2009

26

- Box-plot of the residual shows
- Heavy tails
 - Many outliers
 - Weakly positively skewed distribution

Will any of the outliers affect the regression?

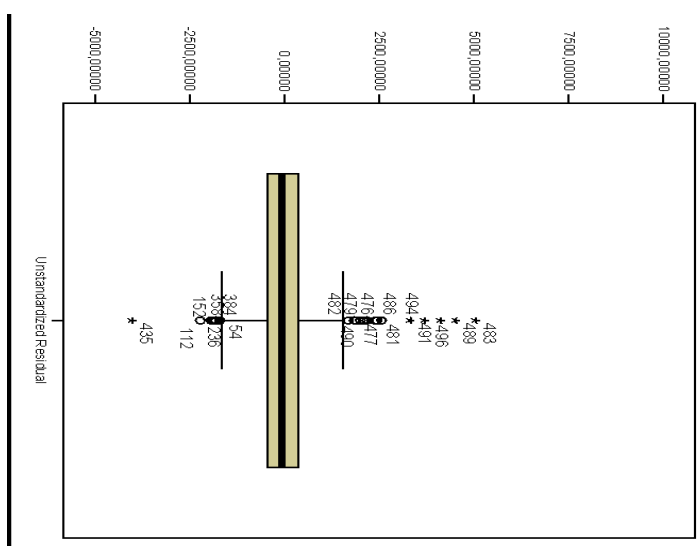


Fall 2009

© Erling Berge 2009

27

The distribution seen from another angle



Fall 2009

© Erling Berge 2009

28

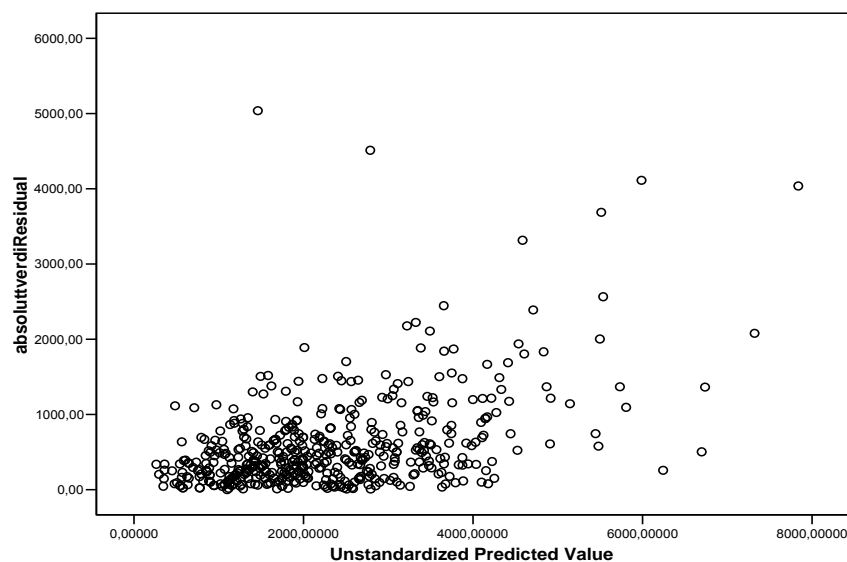
Band-regression

- Homoscedasticity means that the median (and the average) of the absolute value of the residual, i.e.: $\text{median}\{|e_i|\}$, should be about the same for all values of the predicted y_i
- If we find that the median of $|e_i|$ 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

Fall 2009

© Erling Berge 2009

29

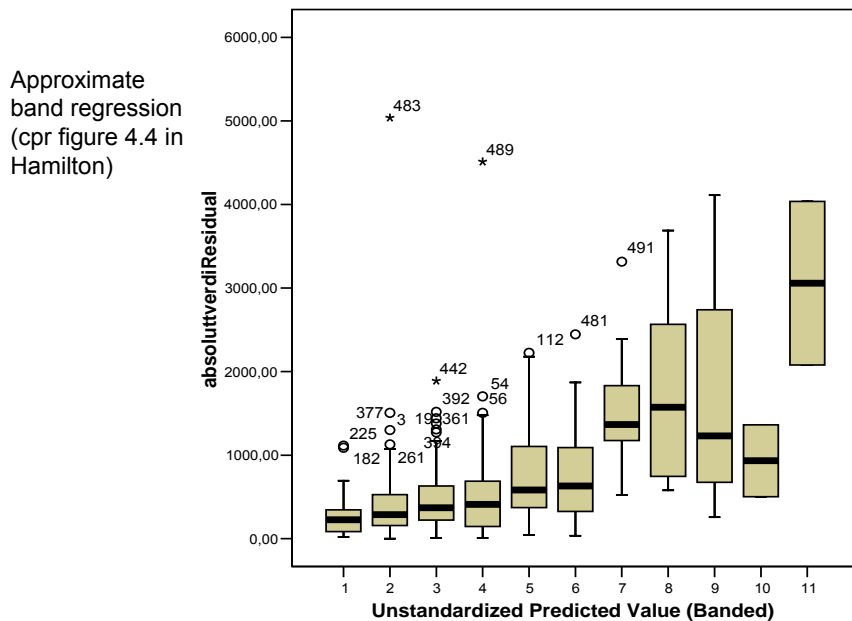


Absolute value of e_i (Based on regression in table 3.2 in Hamilton)

Fall 2009

© Erling Berge 2009

30



Fall 2009

© Erling Berge 2009

31

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

Fall 2009

© Erling Berge 2009

32

Autocorrelation (1)

- Correlation among variable values on the same variable across different cases (e.g. between ε_i and ε_{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

Fall 2009

© Erling Berge 2009

33

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

Fall 2009

© Erling Berge 2009

34

Durbin-Watson test (1)

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

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

Fall 2009

© Erling Berge 2009

35

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 $d > 2$

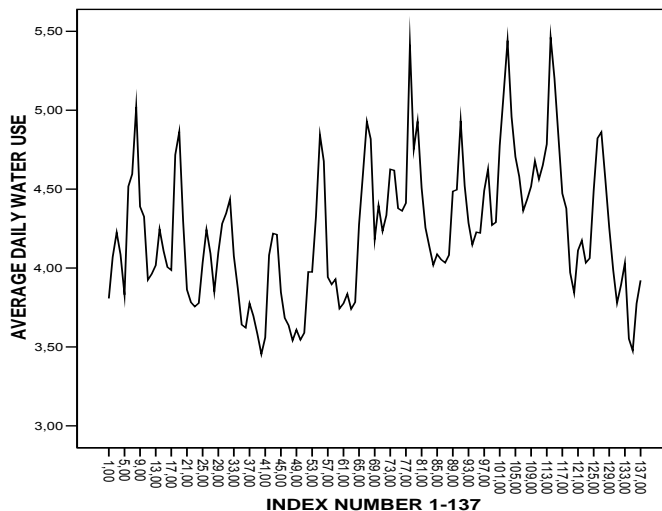
Fall 2009

© Erling Berge 2009

36

Daily water use, average pr month

Example:



Fall 2009

© Erling Berge 2009

37

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

Fall 2009

© Erling Berge 2009

38

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

Fall 2009

© Erling Berge 2009

39

Autocorrelation coefficient

m-th order autocorrelation coefficient

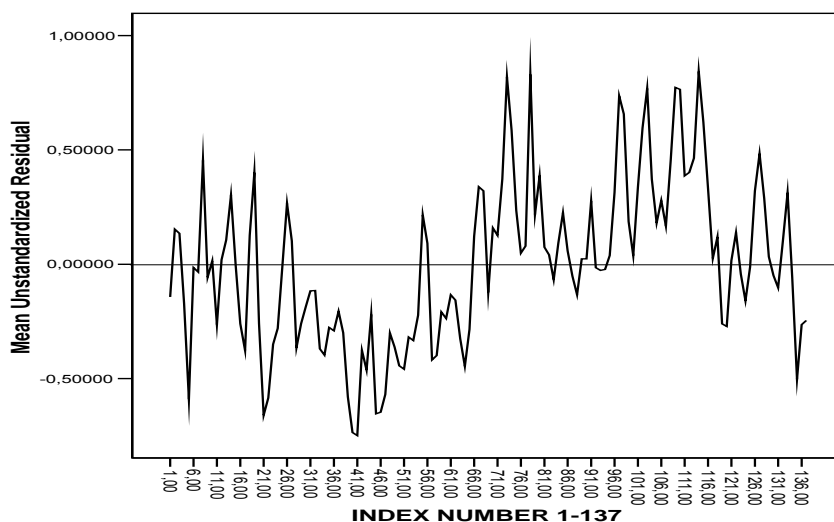
$$r_m = \frac{\sum_{t=1}^{T-m} (e_t - \bar{e})(e_{t+m} - \bar{e})}{\sum_{t=1}^T (e_t - \bar{e})^2}$$

Fall 2009

© Erling Berge 2009

40

Residual "Daily water use", month



Fall 2009

© Erling Berge 2009

41

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

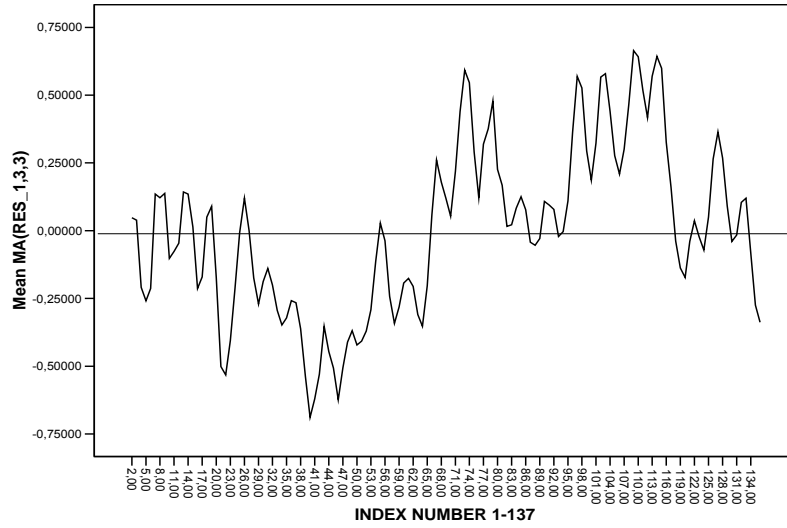
$$e_t^* = \text{median}\{e_{t-1}, e_t, e_{t+1}\}$$

Fall 2009

© Erling Berge 2009

42

Residual, smoothing once

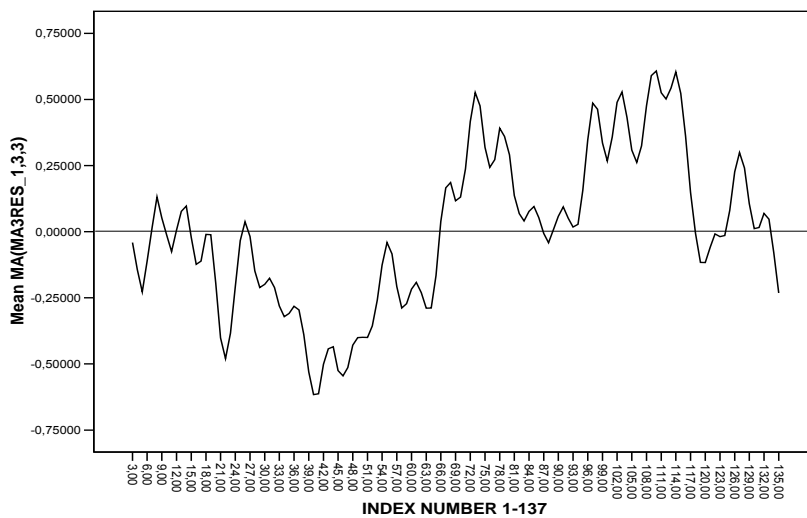


Fall 2009

© Erling Berge 2009

43

Residual, smoothing twice

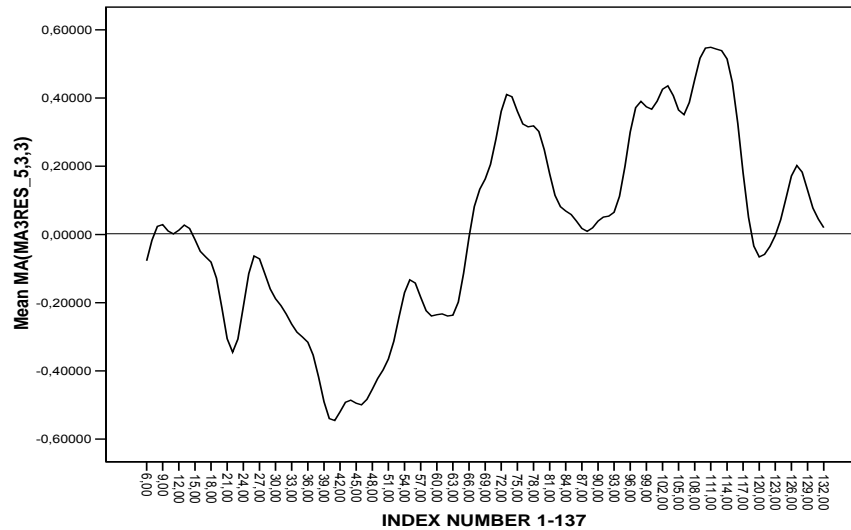


Fall 2009

© Erling Berge 2009

44

Residual, smoothing five times



Fall 2009

© Erling Berge 2009

45

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)

Fall 2009

© Erling Berge 2009

46