SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing V

- Kritikk av regresjon I
  Hamilton Kap 4 s109-123

Modellanalysar bygg på føresetnader

- OLS er ein enkel analyseteknikk med gode teoretiske eigenskapar, men
- Eigenskapane er basert på visse føresetnader
- Dersom føresetnadane ikkje held vil dei gode eigenskapane forvitre
- Å undersøkje i kva grad føresetnadane held er den viktigaste delen av analysen

OLS-REGRESJON: føresetnader

- I SPESIFIKASJONSKRAVET
  - Føresetnaden er at modellen er rett
- II GAUSS-MARKOV KRAVA
  - Sikrar at estimata er “BLUE”
- III NORMALFORDELTE RESTLEDD
  - Sikrar at testane er valide

FØRESETNADER: I Spesifikasjonskravet

- Modellen er rett spesifisert dersom
  - Forventa verdi av y, gitt verdien av dei uavhengige variablane, er ein lineær funksjon av parametrane til x-variablane
  - Alle inkluderte x-variablar påverkar forventa y-verdi
  - Ingen andre variablar påverkar forventa y-verdi samtidig som dei korrelerer med inkluderte x-variablar

FØRESETNADER: II Gauss-Markov krava (i)

(1) x er gitt, dvs utan stokastisk variasjon

(2) Feila har ein forventa verdi på 0 for alle i

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

Gitt (1) og (2) vil ε_i vere uavhengig av x_k for alle k.

Da gir OLS **forventningsrette** estimat av β

(unbiased = forventningsrett)

FØRESETNADER: II Gauss-Markov krava (ii)

(3) Feila har konstant varians for alle i

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

(4) Feila er ukorrelerte med kvarandre

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

FØRESETNADER: II Gauss-Markov krava (iii)

Gitt (3) og (4) i tillegg til (1) og (2) får vi

- a. Estimat av standardfeilen til regresjonskoeffesientane er forventningsrette, og
- b. **Gauss-Markov teoremet**:

  OLS estimata har **mindre varians** enn alle andre lineære forventningsrette estimat.

OLS gir "BLUE"

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

FØRESETNADER: II Gauss-Markov krava (iv)

(1) - (4) kallast GAUSS-MARKOV krava

- Gitt (2) - (4) med tillegg av krav om at feila er ukorrelerte med X variablane (jfr. Hardy s5), dvs.:
  - $\text{cov}(x_{ik}, \varepsilon_i) = 0$ for alle i,k

  er koeffesientar og standardfeil **konsistente**

Fotnote 1:

Forventningsrette (unbiased) estimatorar

- Forventningsrett tyder at

  $E[b_k] = \beta_k$
- I det lange løp vil vi treffe populasjonsverdien - β_k - dersom vi trekkjer mange nok utval, reknar ut b_k og tar gjennomsnittet av desse

Fotnote 2:
Konsistente estimatorar

- Estimatoren er konsistent dersom vi har at når utvalsstorleiken (n) veks mot uendeleg går b mot β og s_b mot σ_β
- [b_k er ein konsistent estimator for β_k dersom vi for vilkårleg liten c har

  $\lim_{n? 8} [Pr\{ Ib_k - \beta_k I < c \}] = 1$

Fotnote 3:
Variansminimale estimatorar

- Variansminimale (effisente) estimatorar tyder at

  $var(b_k) < var(a_k)$ for alle estimatorar a ulik b
- Ekvivalent:

  $E[b_k - \beta_k]^2 < E[a_k - \beta_k]^2$ for alle estimatorar a ulik b.

Fotnote 4:
Skeive estimatorar

- Sjølv om krava til ”BLUE” er oppfylt vil ein stundom kunne finne skeive estimatorar (dvs. dei er ikkje forventningsrette) som har mindre varians, t.d. i
- Ridge Regression

Fotnote 5:
Ikkje-lineære estimatorar

- Det kan finnast ikkje-lineære estimatorar som er forventningsrette og har mindre varians enn ”BLUE” estimatorane

FØRESETNADER: III Normalfordelte restledd

- (5) Dersom alle feila er normalfordelt med forventning 0 og standardavvik på 1, dvs. dersom

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

  – vil ein kunne teste hypoteser om β og σ, og
  – OLS estimata vil ha mindre varians enn estimat fra alle andre forventningsrette estimatorar

OLS gir "BUE"

(Best Unbiased Estimate)

Problem i regresjonsanalysar som ikkje kan testast

- Om alle relevante variablar er med
- Om det er målefeil i x'ane
- Om forventa verdi til feilleddet er 0

  (Dette er der samme som at vi ikkje kan sjekke om korrelasjonen mellom feilledd og x-variabel faktisk er 0. Dette er i prinsippet det samme som første punkt om at modellen er rett spesifisert)

Problem i regresjonsanalysar som kan oppdagast (1)

- Ikkje-lineære samband
- Inkludert irrelevant variabel
- Ikkje-konstant varians hos feilleddet
- Autokorrelasjon hos feilleddet
- Korrelasjonar mellom feilledd
- Ikkje-normale feilledd
- Multikollinearitet

Konsekvensar av problem (Hamilton, s113)

		Uønska eigenskapar ved estimata			
	Problem	Skeive estimat av b	Skeive estimat av SE_b	Ugyldige t&Ftestar	Høg var[b]
Spesifika-sjonskrav	**Ikkje-lineært samband**	X	X	X	-
	Utelatt relevant variabel	X	X	X	-
	Inkludert irrelevant variabel	0	0	0	X
Gauss-Markov krav	**X er målt med feil**	X	X	X	-
	Heteroskedastisitet	0	X	X	X
	Autokorrelasjon	0	X	X	X
	X er korrelert med e	X	X	X	-
Normal-fordeling	e **er ikkje normalfordelt**	0	0	X	X
Multikol-linearitet		0	0	0	X

Problem i regresjonsanalysar som kan oppdagast (2)

- Utliggarar (ekstreme y-verdiar)
- Innflytelse (case med stor innverknad: uvanlege kombinasjonar av y og x-verdiar)
- Leverage (potensiale for innflytelse)

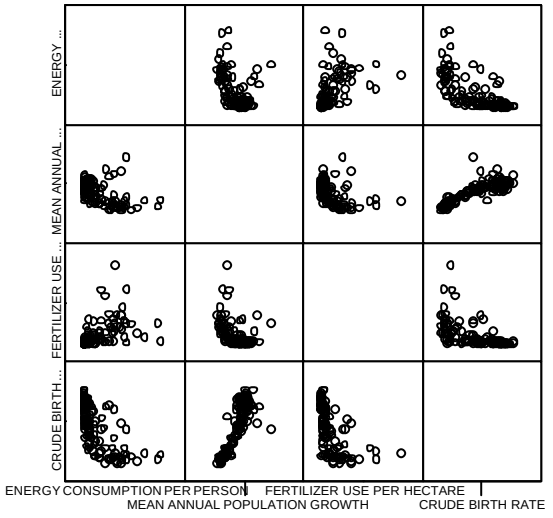
Hjelpemiddel

- Studiar av
  - Ein-variabel fordelingar (frekvens fordeling og histogram)
  - To-variabel samvariasjon (korrelasjon og scatterplott)
  - Residualen (fordeling og i samvariasjon med predikert verdi)

Korrelasjon og scatterplott

		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

Korrelasjon og scatterplott



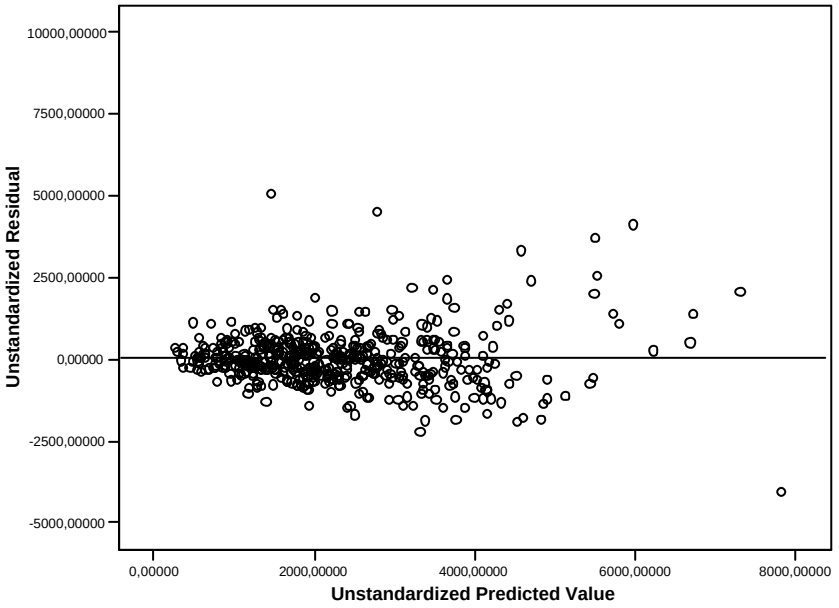
Heteroskedastisitet,

(ikkje-konstant varians hos feilleddet) kan skuldast

- Målefeil (t.d. y meir nøyaktig ved større x)
- Utliggarar
- At ε_i innheld eit viktig ledd som varierer saman med x og y (spesifikasjonsfeil)
- Spesifikasjonsfeil er det samme som feil modell og gir heteroskedastisitet
- Eit viktig diagnoseverktøy er plott av residual mot predikert verdi ($\hat{y}$)

Eksempel: Hamilton tabell 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

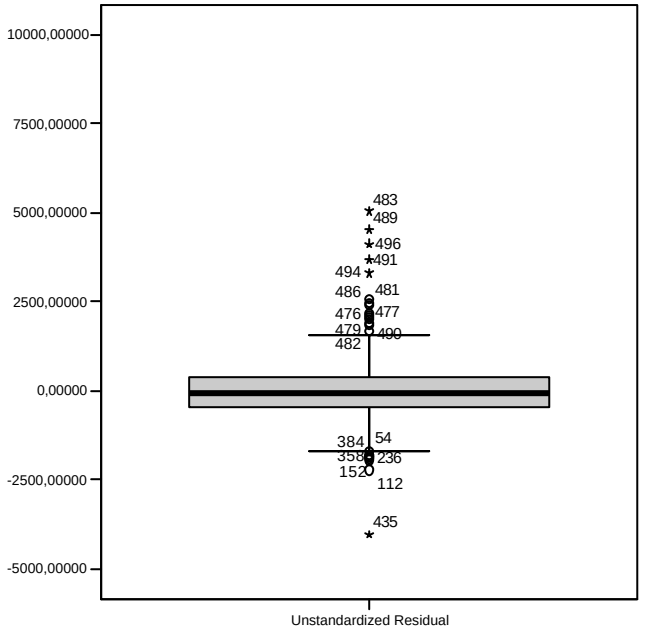


Basert på regresjonen i tabell 3.2 i Hamilton

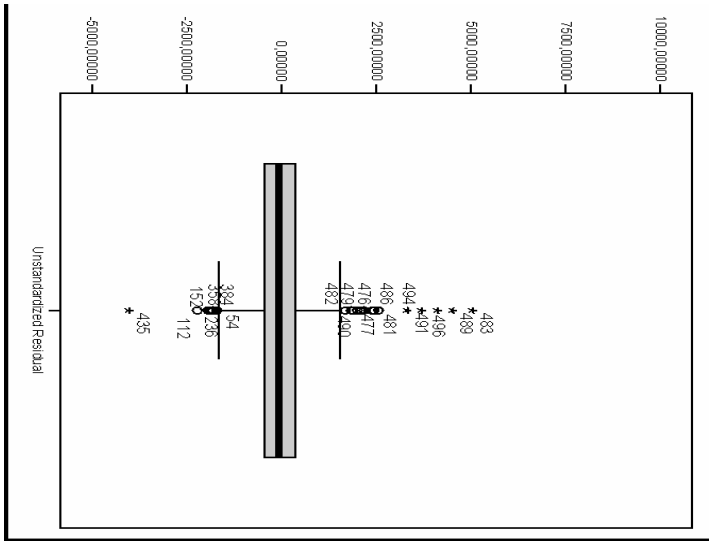
Box-plott av residualen viser

- Tunge halar
- Mange utliggarar
- Svakt positiv skeiv fordeling

Vil nokon av utliggarane påverke regresjonen?

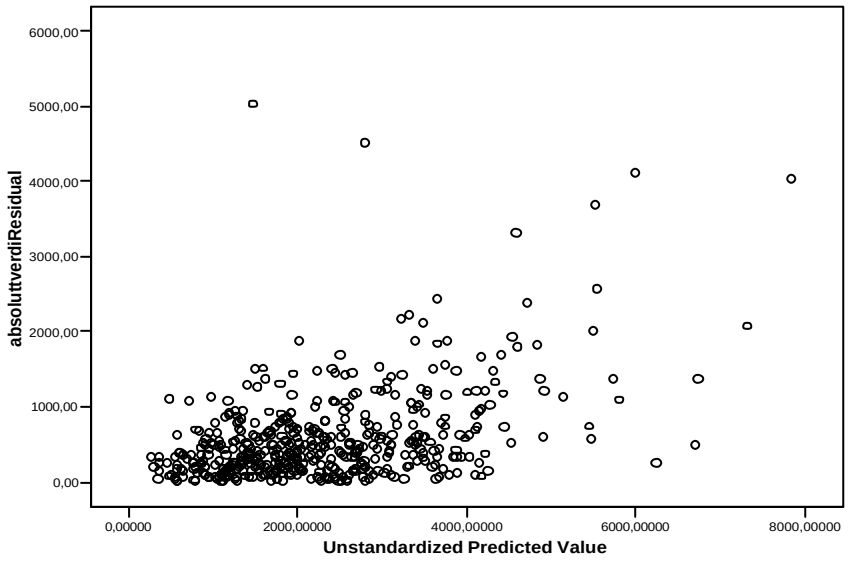


Fordelinga sett frå ein annan vinkel



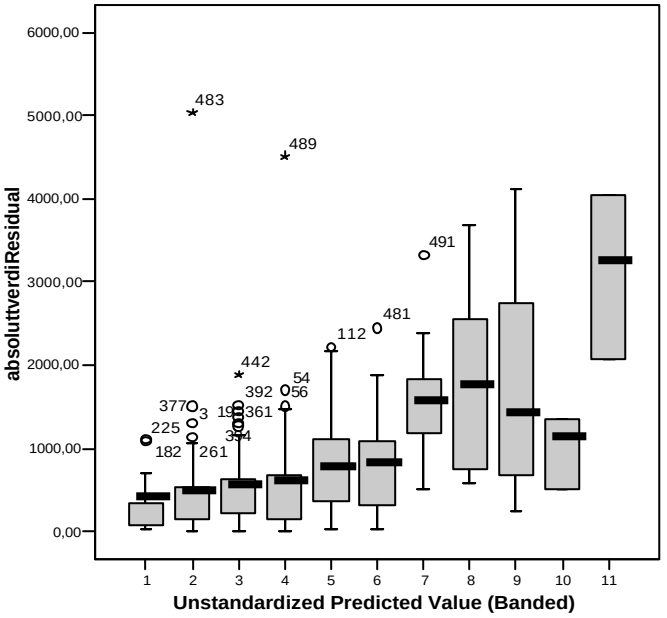
Band-regresjon

- Homoskedastisitet tyder at medianen (og gjennomsnittet) til absoluttverdien av residualen, dvs: median{|e|}, skal vere om lag den same for alle verdiar av predikert y
- Dersom vi finn at medianen av $|e_i|$ for gitt predikert verdi av y endrar seg systematisk med verdien av predikert y indikerer det heteroskedastisitet
- Slike analysar kan lett gjerast i SPSS



Absoluttverdien av ę (Basert på regresjonen i tabell 3.2 i Hamilton)

Tilnærma bandregresjon (jfr figur 4.4 i Hamilton)



Autokorrelasjon (1)

- Korrelasjon mellom variabelverdiar på same variabel over ulike case

  (t.d. mellom ε_i og ε_{i-1})
- Autokorrelasjon gir større varians og skeive estimat av standardfeil slik som heteroskedastisitet
- Når vi har enkelt tilfeldig utval frå ein populasjon, er autokorrelasjon usannsynleg

Autokorrelasjon (2)

- Autokorrelasjon kjem frå feilspesifikasjon av modellen
- Ein finn det typisk i tidsseriar og ved geografisk ordna case
- Testar (t.d. Durbin-Watson) er basert på sorteringsrekkefølgja av casa. Derfor:
- Ei hypotese om autokorrelasjon må spesifisere korleis casa skal sorterast

Durbin-Watson testen (1)

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

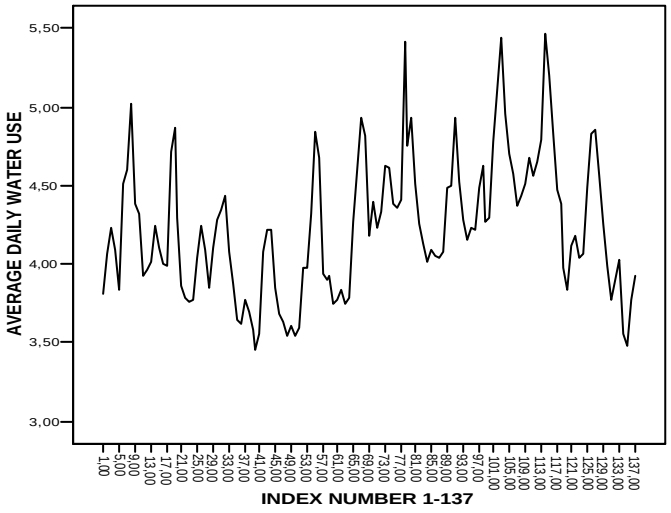
Bør ikkje nyttast for autoregressive modellar, dvs. modellar der y-variabelen også finst som forklaringsvariabel (x-variabel) jfr tabell 3.2

Durbin-Watson testen (2)

- Samplingfordelinga til d-observatoren er kjent og tabellert som $\mathbf{d_L}$ og $\mathbf{d_U}$ (tabell A4.4 i Hamilton), talet av fridomsgrader baserer seg på n og K-1
- Testregel:
  - Forkast dersom $d<d_L$
  - Forkast ikkje dersom $d>d_U$
  - Dersom $d_L < d < d_U$ kan det ikkje konkluderast
- d=2 tyder ukorrelerte residualar
- Positiv autokorrelasjon gir d<2
- Negativ autokorrelasjon gir d>2

Dagleg vassforbruk, gjennomsnitt pr månad

Eksempel:



Vanleg OLS-regresjon der caset er månad

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 for autokorrelasjon

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

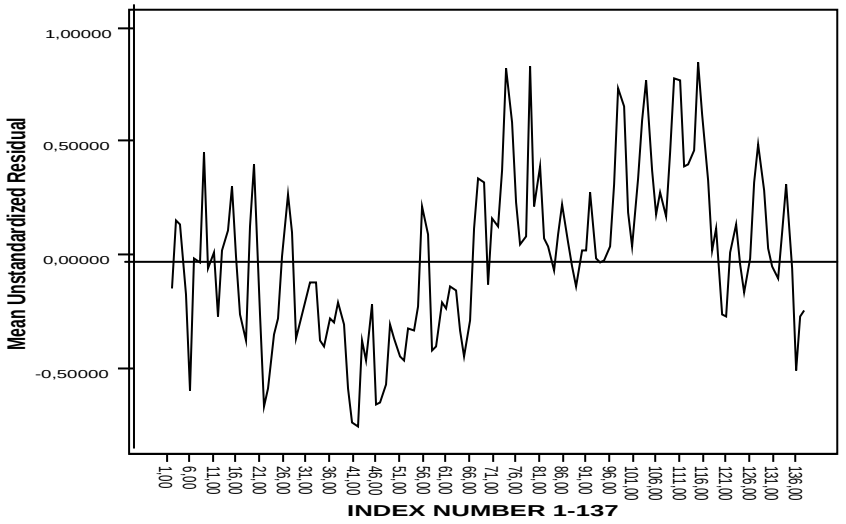
Finn grensene for å forkaste / akseptere nullhypotesa om ingen autorkorrelasjon med nivå 0,05

Auto-korrelasjonskoeffesient

m-te ordens autokorrelasjonskoeffesient

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

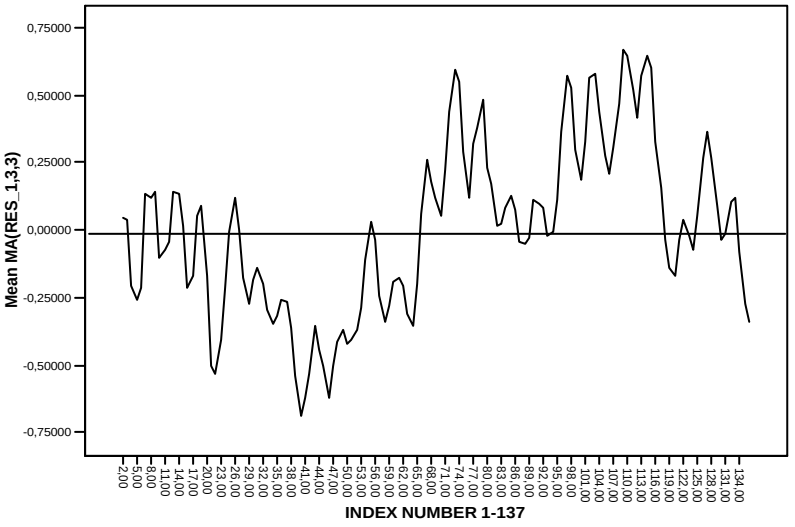
Residual ”Dagleg vassforbruk”, månad



Glatting med 3 punkt

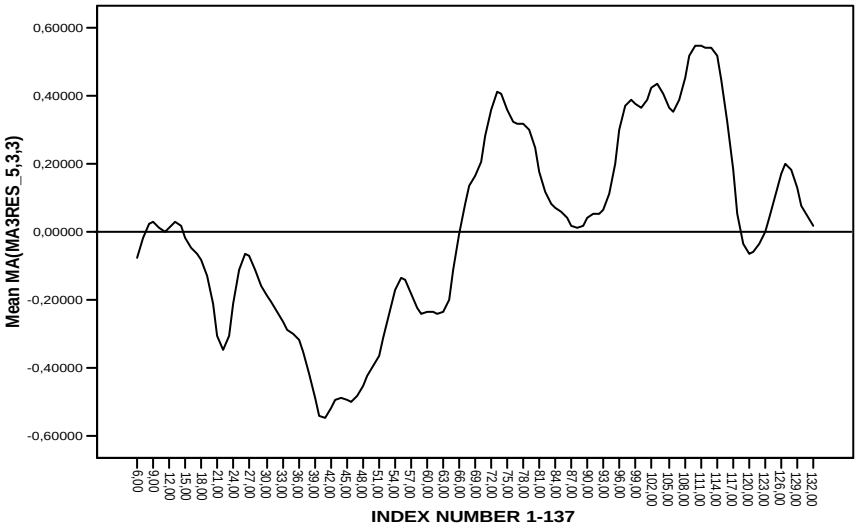
- Glidande gjennomsnitt $$e_t^* = \frac{e_{t-1} + e_t + e_{t+1}}{3}$$
- ”Hanning” $$e_t^* = \frac{e_{t-1}}{4} + \frac{e_t}{2} + \frac{e_{t+1}}{4}$$
- Glidande median $$e_t^* = median\{e_{t-1}, e_t, e_{t+1}\}$$

Residual, glatta ein gong



Residual, glatta to gonger

Residual, glatta fem gonger



Konsekvensar av autokorrelasjon

- Hypotesetestar og konfidensintervall er upålitelege. Regresjon kan likevel gi ein god beskrivelse av utvalet. Parametrane er forventningsrette
- Spesialprogram kan estimere standardfeil konsistent
- Ta inn i analysen variablar som påverkar ”hosliggjande” case
- Ta i bruk teknikkar frå tidsserieanalyse (t.d.: analyser differansen mellom to tidspunkt) (Δy)