SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 10

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

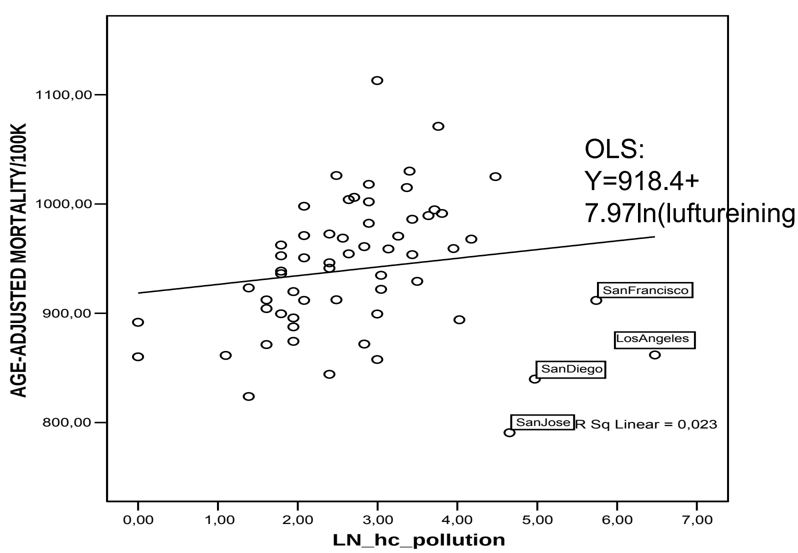
Lecture X

- Robust Regression
  Hamilton Ch 6 p183-212

Robust Regresjon

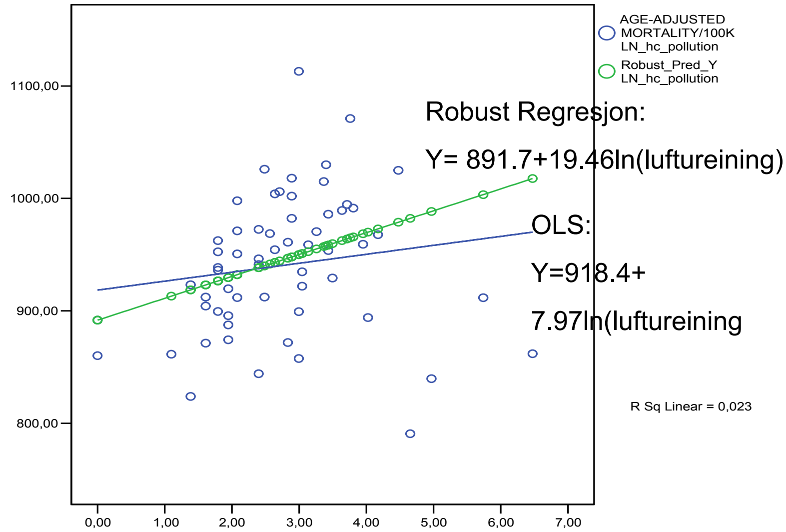
- Er utvikla for å fungere godt i situasjonar der OLS regresjonen bryt saman. Der OLS føresetnadene er oppfylt gir robust regresjon dårlegare resultat enn OLS, men ikkje mye
- Sjølv om robust regresjon høver betre for den som ikkje vil leggje mye arbeid i å teste føresetnader er metodane førebels vanskelege å gjere seg bruk av
- Robust regresjon har fokusert mest på fordelingar av residualar med tunge halar (mange case med stor innverknad på regresjonen)

Regresjon av mortalitet på luftureining



Figur 6.1
Hamilton

Robust regresjon av mortalitet på luftureining



Figur 6.2
Hamilton

Robust regresjon og SPSS

- Det er ikkje noko rutine i SPSS som gjer robust regresjon direkte
- Den kan gjerast ved hjelp av vekta regresjon, men det krev at vi lagar vektfunksjon og går igjennom iterasjonane ein for ein med utrekning av nye vekter for kvar gong
- Framgangsmåte kjem vi til nedanfor

ROBUST OG RESISTENT

- RESISTENTE metodar lar seg ikkje påverke av små feil/ endringar i utvalsdata
- ROBUSTE metodar lar seg ikkje påverke av små avvik frå føresetnadene til modellen
- Dei fleste resistente estimatorane er også robuste i høve til føresetnaden om normalfordeling av residualane
- **OLS er verken ROBUST eller RESISTENT**

Fall 2004 © Erling Berge 2004 7

Utliggjarar er eit problem for OLS

Utliggjarar verkar inn på estimat av

- Parametrar
- Standardfeil (standardavvik til parameter)
- Determinasjonskoeffisient
- Testobservatorar
- Og mange andre observatorar

Robust regresjon freistar verne mot dette ved å gi mindre vekt til slike case,
<u>ikkje ved å ekskludere dei</u>

Fall 2004 © Erling Berge 2004 8

Hjelp mot IKKJE-NORMALE residualar

Robuste metodar kan vere til hjelp når

- Halane i residualfordelinga er "tunge" dvs. når det er "for mange" utliggjarar i høve til normalfordelinga
- Uvanlege X-verdiar gir påverknad (leverage) problem

Ved andre årsaker til ikkje-normalitet hjelper dei ikkje.

Fall 2004 © Erling Berge 2004 9

Estimeringsmetodar for robust regresjon

- M-estimering (maximum likelihood) minimerer ein vekta sum av residualane. Kan tilnærmast med vekta minste kvadrat metoden (WLS)
- R-estimering (basert på rang) minimerer ein sum der ein vekta rang inngår. Metoden er vanskelegare å bruke enn M-estimeringa
- L-estimering (basert på kvantilar) brukar lineære funksjonar av utvalsordningsobservatorane (kvantilane)

Fall 2004 © Erling Berge 2004 10

IRLS- Iterativt Revekta Minste Kvadrat

M-estimat ved hjelp av IRLS treng

1. Startverdiar frå OLS. Ta vare på residualane.
2. Bruk OLS residualane til å finne vekter. Til større residual, til mindre vekt
3. Finn nye parameterverdiar og residualar med WLS
4. Gå til 2 og finn nye vekter frå dei nye residualane, fortsett til steg 3 og 4, heilt til endringane i parametrane vert små

Iterasjon: å gjenta ein sekvens av operasjonar

IRLS

- IRLS er i teorien ekvivalent med M-estimering
- For å nytte metoden treng vi å rekne ut
- Skalerte residualar, u_i , og ein
- Vektfunksjon, w_i , som gir minst vekt til dei største residualane

Skalering av residualar I

- Skalert residual u_i
  - s er skaleringsfaktoren og e_i residualen
- Skaleringsfaktoren i OLS er estimatet av standardfeilen til residualen: nb! s_e er ikkje resistent
- Eit resistent alternativ er basert på MAD, "median absolute deviation"

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

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

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

Fall 2004 © Erling Berge 2004 13

Skalering av residualar II

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

Skaleringsfaktoren (standardfeilen i fordelinga) vert med bruk av eit resistent estimat

- s = MAD/ 0.6745 = 1.483MAD

og den skalerte residualen

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

I ei normalfordeling vil s= MAD/ 0.6745 estimere standardfeilen korrekt slik som s_e

Ved ikkje-normale feil vil s= MAD/ 0.6745 vere betre.
Det er eit resistent estimat, s_e er ikkje resistent

Fall 2004 © Erling Berge 2004 14

Vektfunksjonar I

- Eigenskapane vert målt i høve til OLS på normalfordelte feil. Metoden skal vere "nesten like god" som OLS ved normalfordelte feil og mye bedre når feila er ikkje-normale
- Eigenskapane vert fastlagt ved ein "kalibreringskonstant" (c, i formlane)

Vektfunksjonar II

- **OLS-vekter**: $w_i = 1$ for alle i
- **Huber-vekter**: vektar ned når den skalerte residualen er større enn c, c=1,345 gir 95% av OLS sin effektivitet på normalfordelte feil
- **Tukey's bivekta** estimat får 95% av OLS sin effektivitet på normalfordelte feil ved gradvis nedvekting av skalerte feil opp til $|u_i| \leq c = 4.685$ og ved å droppe case der residualen er større.

Huber-vekter

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

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

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

Tukey vekter

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

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

$$\forall = \ for\ alle$$

- Tukey vekting i IRLS er sensitiv for startverdiane av parametrane (ein kan finne lokale minimum)

Standardfeil og testar ved IRLS

- WLS program vil ikkje estimere standardfeil og testobservatorar rett ved IRLS
- Ein prosedyre som fungerer er gitt av Hamilton på side 198-199

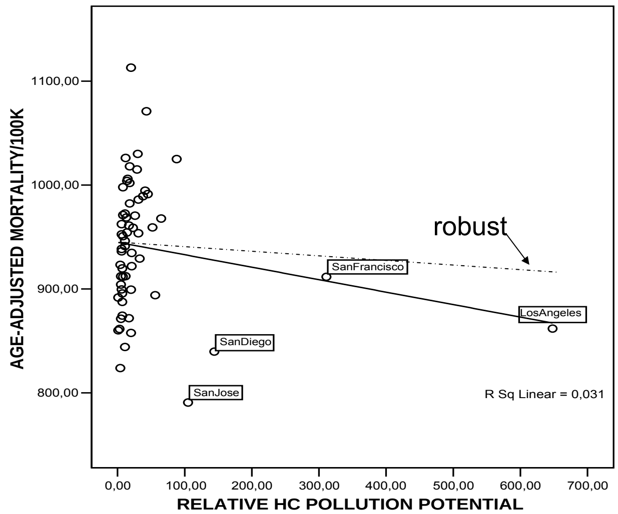
Bruk av Robust Estimering

- Dersom OLS estimat og Robuste estimat er ulike tyder det at utliggjarar verka inn på OLS slik at vi ikkje kan stole på resultata
- Robuste predikerte verdiar reflekterer betre hovudmassen av data
- Robuste residualar vil derfor betre avsløre kva som er uvanlege case
- Vektene frå den robuste regresjonen vil vise kva for case som er utliggjarar
- OLS og RR kan stø kvarandre

Fig 6.9 Hamilton: OLS og RR på data utan transformasjon

Mortalitet på luftureining

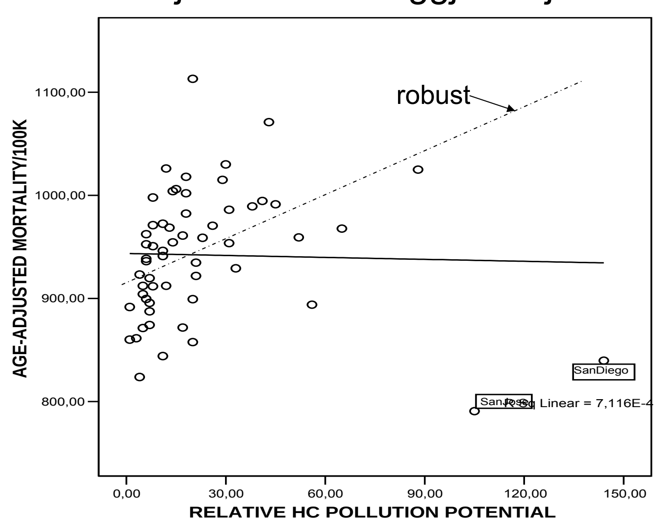
Effekt av høg leverage



Fall 2004 © Erling Berge 2004 21

Fig 6.10 Hamilton: OLS og RR på data utan transformasjon med to utliggjarar fjerna

Mortalitet på luftureining



Fall 2004 © Erling Berge 2004 22

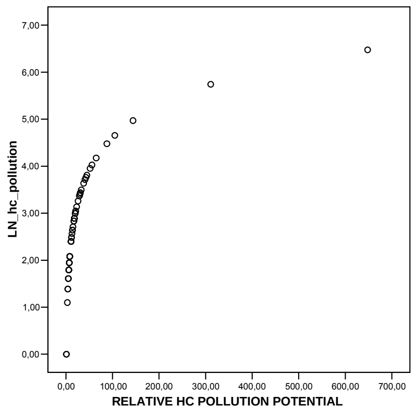
RR vernar ikkje mot leverage

- RR med M-estimering vernar mot uvanlege y-verdiar (utliggjarar) men ikkje nødvendigvis mot uvanlege x-verdiar (leverage)
- Innsats på testing og diagnose trengst framleis (heteroskedastisitet er td. problematisk ved IRLS)
- Studiar av datamaterialet og symmetri-transformasjon reduserer sjansen for at problem dukkar opp
- Ingen metode er "trygg" om den blir brukt utan omtanke og studiar av data

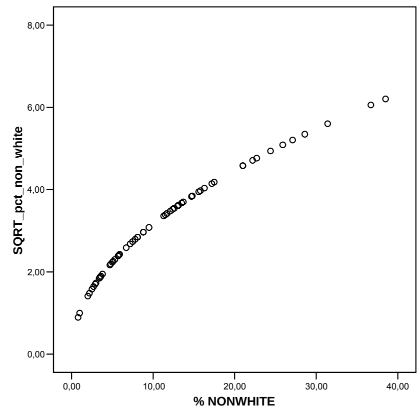
Robust Multippel Regresjon

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, %	

Multippel OLS regresjon med transformerte variable: effekten av transformasjon



ln av luftureining



Kvadratrot av % ikkje-kvite

OLS med baklengs eliminering gir

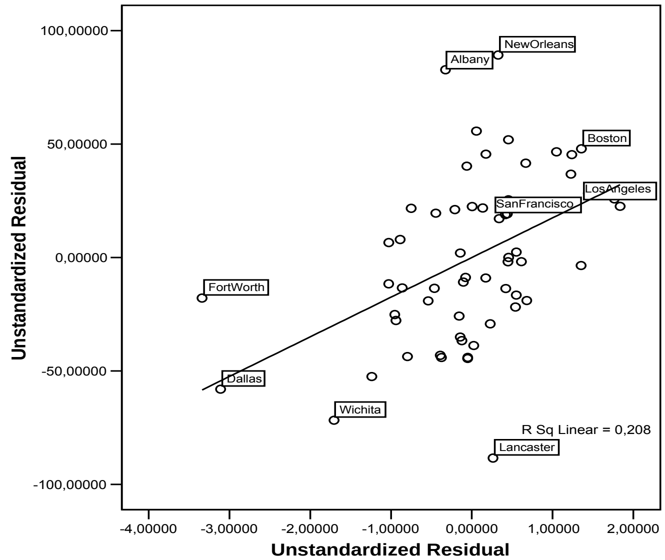
Dependent Variable: AGE-ADJUSTED MORTALITY/100K	B	Std. Error	t	Sig.
(Constant)	986,261	82,674	11,929	,000
LN_hc_pollution	17,469	4,636	3,768	,000
AVG. YEARLY PRECIP. INCHES	2,352	,640	3,677	,001
AVG. JANUARY TEMPERATURE, F	-2,132	,504	-4,228	,000
MEDIAN EDUCATION OF POP 25+	-17,958	6,204	-2,895	,005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

- Robust regresjon gir predikert y

$= 1001.8+17.77x_{1i}+2.32x_{2i}-2.11x_{3i}-19.1x_{4i}+26.2x_{5i}$

Multippel OLS regresjon med transformerte variable

Leverage plott av residual av mortalitet (y) og residual av ln_luftureining (x)



Los Angeles og San Francisco er ikkje lenger utliggjarar

Fall 2004 © Erling Berge 2004 27

Fire estimat av samanhengen mortalitet - luftureining

Effekten av luftureining

	OLS	Robust
1 variabel	7.97	19.46
5 variablar	17.47	17.77

- Legg merke til at RR i den bivariate regresjonen kjem ganske nær resultatet i den multivariate regresjonen

- I fem-variabel modellen er det nye case som verkar inn på regresjonslina
- Fjerning av dei 5 casa som har høgast leverage parameter (h_i) gir ikkje substansielle endringar i koeffisientane

Fall 2004 © Erling Berge 2004 28

Robust Regresjon vs Avgrensa Påverknad Regresjon

- Robust Regresjon vernar mot effekt av utliggarar (Uvanlege y-verdiar) dersom dei ikkje kjem saman med uvanlege x-verdiar
- Bounded Influence Regression
  (Regresjon med avgrensa påverknad)
  - vernar i tillegg mot påverknad (uvanlege kombinasjonar av x-verdiar)

BI (bounded influence) - Avgrensa påverknad regresjon

- BI-metodane er laga for å avgrense verknaden av stort potensiale for påverknad (stor h_i - leverage)
- Den aller enklaste tilnærminga til problemet er å modifisere Huber-vektene eller Tukey-vektene med ein faktor basert på leverage observatoren

Avgrensa påverknad: vektmodifikasjon

- Vi utvidar vektfunksjonen med ei vekt basert på leverage observatoren h_i

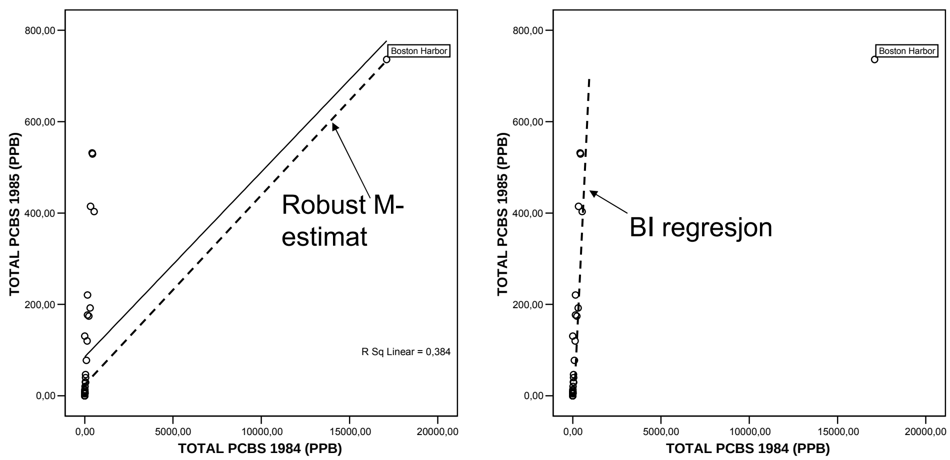
Påverknadsfaktoren i vektinga kan t.d. setjast til

- $w^H_i = 1$ hvis $h_i \leq c^H$
- $w^H_i = (c^H / h_i)$ hvis $h_i > c^H$
- c^H vert ofte sett lik 90% percentilen i fordelinga av h_i
- IRSL vekta vert da $w_i w^H_i$ der w_i er enten Tukey- eller Huber-vekter som endrar seg frå iterasjon til iterasjon medan w^H_i er konstant

Avgrensa påverknad som diagnoseverktøy

- Estimering av standardfeil og testobservatorar vert no enno meir komplisert enn for dei M-estimatorane vi omtala ovanfor
- Vi kan bruke BI estimat som deskriptivt verktøy for å sjekke opp ande estimat
- Eit (litt) ekstremt eksempel: PCB ureining i elvemunningar i 1984 og 1984 (Hamilton tabell 6.4)

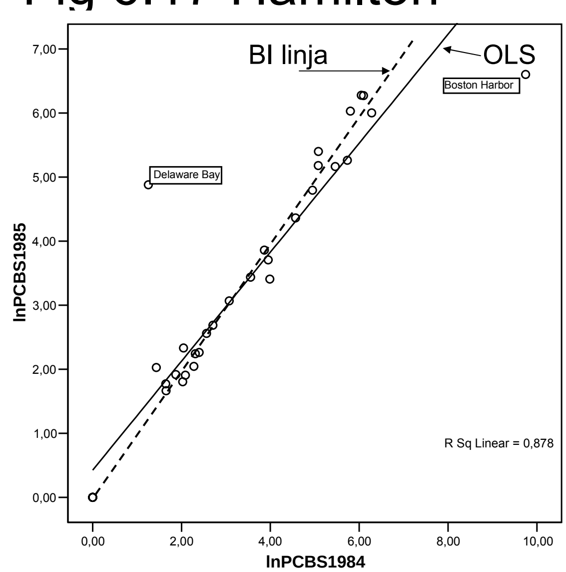
Fig 6.15 og 6.16 Hamilton



Fall 2004 © Erling Berge 2004 33

Fig 6.17 Hamilton

OLS og BI estimat med transformerte variablar gir om lag same resultat



Fall 2004 © Erling Berge 2004 34

Konklusjonar

- Når data har mange utliggjarar vil robuste metodar ha betre eigenskapar enn OLS.
  – Dei er meir effektive og gir meir nøyaktige konfidensintervall og testar
- Robust regresjon kan brukast som diagnoseverktøy.
  – Er OLS og RR einige kan vi ha større tiltru til OLS resultata
  – Er dei ueinige vil vi
    - Vere merksame på at eit problem eksisterer
    - Ha ein modell som passar betre med data og identifiserer utliggjarar betre
- Robuste metodar verkar ikkje mot problem som skuldast kurvelineære eller ikkje-lineære modellar, heteroskedastisitet og autokorrelasjon

Fall 2004 © Erling Berge 2004 35