SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing VI

- Kritikk av regresjon II
  Hamilton Kap 4 s109-137

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 som ikkje kan testast

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

Dei viktigaste føresetnadane som kan oppdagast

- Ikkje-lineære samband
- Heteroskedastisitet
- Autokorrelasjon
- Ikkje-normale feilledd

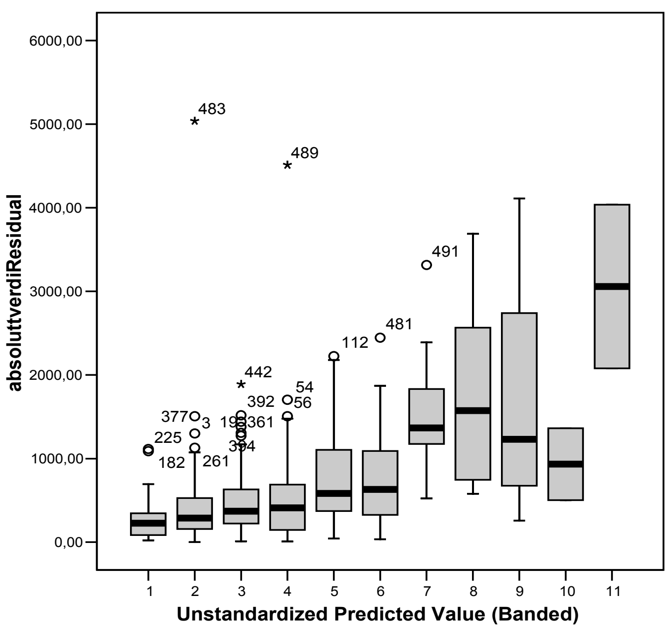
Heteroskedastisitet

- Har vi når variansen til feilleddet varierer med storleiken til x-verdiane
- Predikert y er ein indikator på storleiken av x-verdiane
- Heteroskedastisitet kan komme av
  - Målefeil (t.d. y vert målt meir nøyaktig ved større x)
  - Utliggarar
  - Feil modell (spesifikasjonsfeil) som t.d.
    - Feil funksjonsform eller
    - Når ε_i innheld eit viktig ledd som korrelerer med ein eller fleire x-ar og y (Feilleddet ε_i er ikkje uavhengig av x-ane slik at Gauss-Markov krava 1 og 2 ikkje kan vere korrekte)

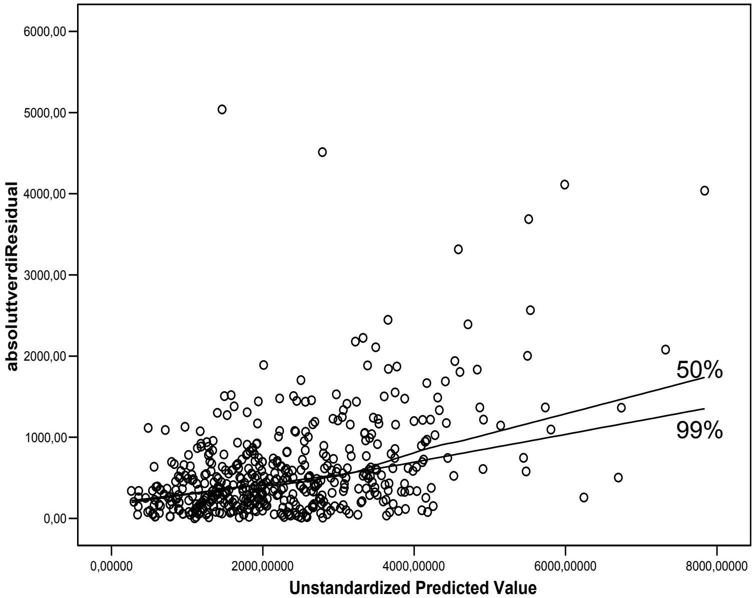
Indikatorar på heteroskedastisitet

- Inspeksjon av spreiingsplott for predikert y mot residual
- Bandregresjon i spreiingsplottet
- Lokalt vekta/ ”glidande” regresjon i den sentrale delen av utvalet

Tilnærma bandregresjon (jfr figur 4.4 i Hamilton)



”Glidande” tilpassa linje ved hjelp av lokalt vekta OLS regresjon



Autokorrelasjon

- Korrelasjon mellom variabelverdiar på same variabel over ulike case
- Autokorrelasjon gir større varians og skeive estimat av standardfeil
- Autokorrelasjon kjem frå feilspesifikasjon av modellen
- Ein finn det typisk i tidsseriar og ved geografisk ordna case
- Testar er basert på sorteringsrekkefølgja av casa og hypoteser om autokorrelasjon må spesifisere korleis casa skal sorterast

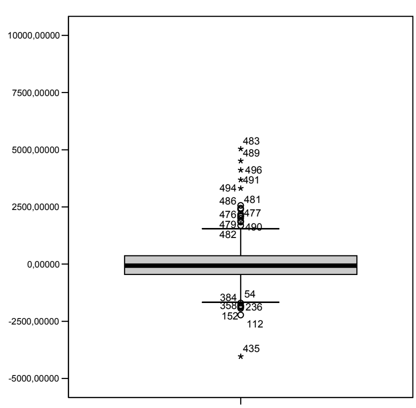
Ikkje-normale residualar

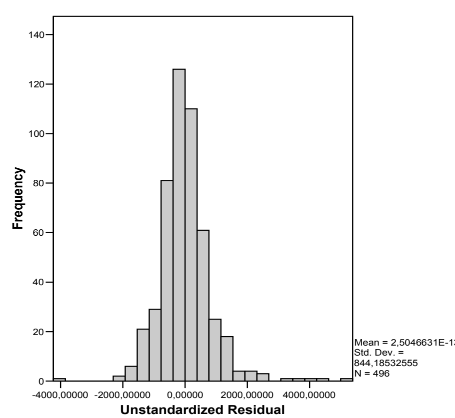
- Gjer at vi ikkje kan nytte t- og F-testar
- Sidan OLS-estimata av parametrane er lett påverkeleg av utliggjarar vil tunge halar i fordelinga av feila indikere stor variasjon i estimata frå utval til utval
- Vi kan sjekke føresetnaden om normalfordeling gjennom å sjå på fordelinga av residualen
  - Histogram, boxplott eller kvantil-normal plott

Diagram av residualen viser:

Tunge halar, mange utliggarar og svakt positiv skeiv fordeling

BOXPLOTT HISTOGRAM

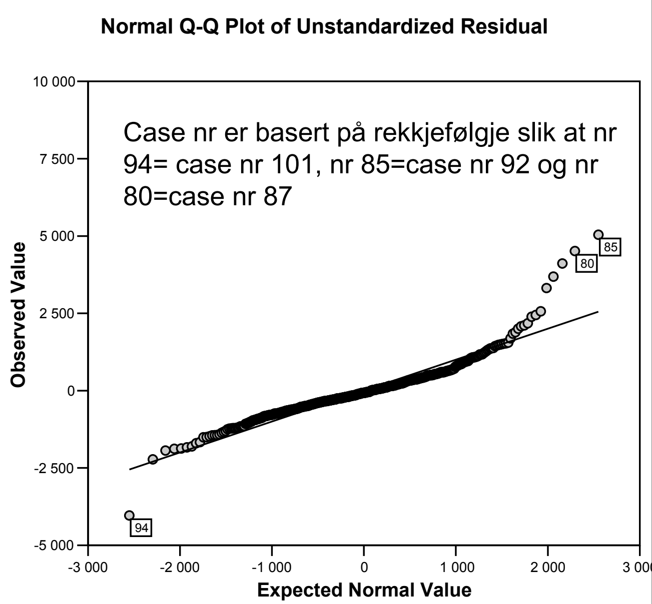




Skeiv fordeling av residualen

- Sidan gjennomsnittet av residualane (e_i) alltid er lik 0, vil fordelinga vere skeiv dersom medianen er ulik 0
- Vi veit at i normalfordelinga er standardavviket (eller standardfeilen) lik IQR/1.35
- Dersom fordelinga av residualen er symmetrisk kan vi samanlikne SE_e med IQR/1.35. Dersom
  - SE_e > IQR/1.35 er halane tynger enn i normalfordelinga
  - $SE_e \approx$ IQR/1.35 er halane tilnærma lik normalfordelinga
  - SE_e < IQR/1.35 er halane lettare enn i normalfordelinga

Kvantil-Normal plott av residual frå regresjon i tabell 3.2 i Hamilton



Tiltak ved ikkje-normalitet

- Sjekk om vi har funne rette funksjonsforma
- Sjekk om vi har utelate ein viktig variabel
  - Dersom vi ikkje kan forbetre modellen substansielt kan vi freiste å transformere den avhengige variabelen så den blir symmetrisk
- Sjekk om manglande normalitet skuldast utliggjarar eller påverknadsrike case
  - Dersom vi har utliggjarar kan transformasjon hjelpe

Påverknad (1)

- Eit case (eller ein observasjon) har påverknad dersom regresjonsresultatet endrar seg når case blir utelate
- Somme case har uvanleg stor påverknad på grunn av
  - Uvanleg stor y-verdi (utliggjar)
  - Uvanleg stor verdi på ein x-variabel
  - Uvanlege kombinasjonar av variabelverdiar

Påverknad (2)

- Vi ser om eit case har påverknad ved å samanlikne regresjonar med og utan eit bestemt case. Ein kan t.d.
- Sjå på skilnaden mellom b_k og $b_{k(i)}$ der case nr i er utelate i estimeringa av den siste koeffesienten.
- Denne skilnaden målt relativt til standardfeilen til $b_{k(i)}$ vert kalla $DFBETAS_{ik}$

$DFBETAS_{ik}$

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

$s_{e(i)}$ er residualen sitt standardavvik når case nr i er utelate frå regresjonen

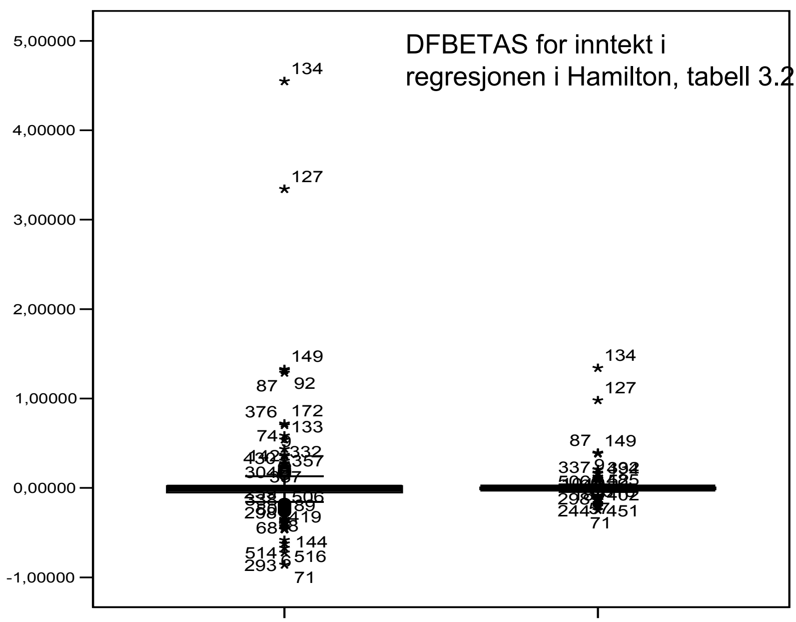
RSS_k er Residual Sum of Squares frå regresjonen av x_k på alle dei andre x-variablane

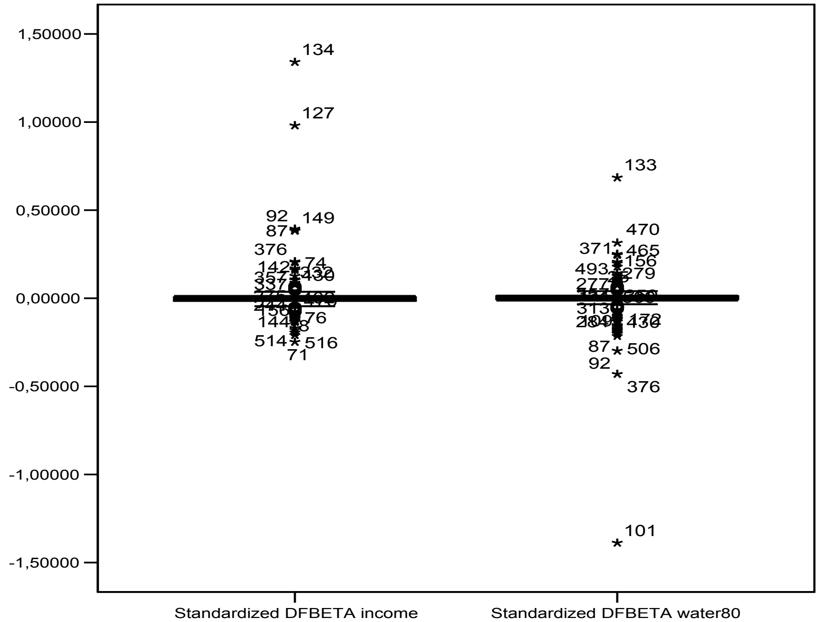
Kva er ein stor DFBETAS?

- $DFBETAS_{ik}$ vert rekna ut for kvar uavhengig variabel og kvart einaste case. Vi kan ikkje inspisere alle verdiane
- Tre kriterium for å finne dei store verdiane vi treng sjå på (ingen av dei treng vere problematiske)
  – Ekstern skalering: $|DFBETAS_{ik}| > 2/\sqrt{n}$
  – Intern skalering:

    $Q_1 - 1.5IQR < |DFBETAS_{ik}| < Q_3 + 1.5IQR$

    (alvorleg utliggjar i box-plott av $DFBETAS_{ik}$)
  – Gap i fordelinga av $DFBETAS_{ik}$





Rekkjefølgje i datafila og case nr er ikkje det same. Case nr er fast.

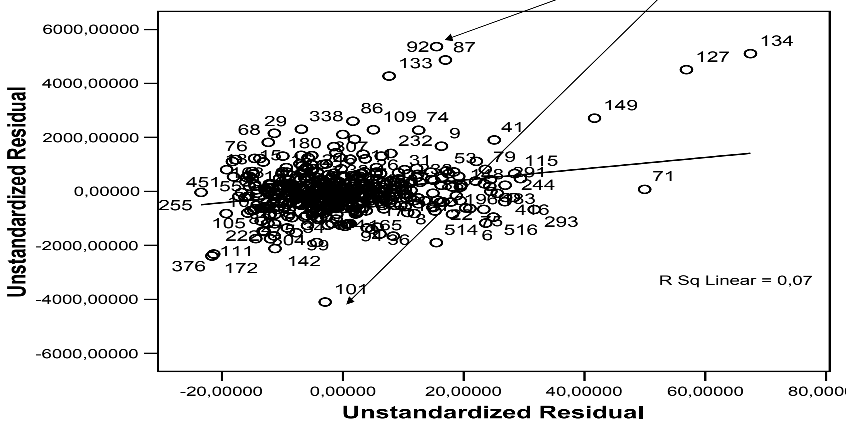
Rekkje nr	Case nr	water81	water80	water 79	educat	retire	peop 81	cpeop
91	98	1500	1300	1500	16	0	2	0
92	99	3500	6500	5100	14	0	6	0
93	100	1000	1000	2700	12	1	1	0
94	101	3800	12700	4800	20	0	5	0
95	102	4100	4500	2600	20	0	5	0
96	103	4200	5600	5400	16	0	5	-1
97	104	2400	2700	800	16	0	6	0
98	105	1600	2300	2200	14	0	4	0
99	107	2300	2300	3100	16	0	4	-2

Leverage plott for vassforbruk og inntekt (sjå Hamiton side 69-72 om partielle regressionsplott)

Y: residual Vassforbruk sommar 1981

X: residual Inntekt i tusen

Sjå tilbake på kvantil-normal plottet ovanfor



Konsekvensar av case med stor påverknad

- Om vi oppdagar påverknadsrike case skal vi ikkje nødvendigvis ta dei ut av analysen
- Rapporter resultat med og utan casa
- Sjekk påverknadsrike case nøye, kanskje er der målefeil
- Når påverknadsrike case er utliggjarar kan ein minske innverknaden ved transformasjon
- Bruk robust regresjon som ikkje er så lett påverkeleg som OLS regresjon

Potensiell påverknad: leverage

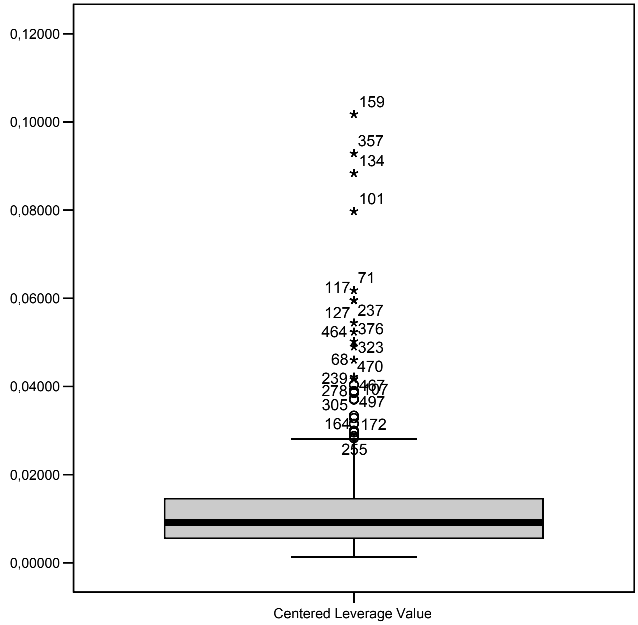
- Den samla påverknaden frå ein bestemt kombinasjon av x-verdiar på eit case måler vi med h_i ”hatt-observatoren”
- h_i varierer frå 1/n til 1. Den har eit gjennomsnitt på K/n (K = # parametrar)
- SPSS rapporterer den sentrerte h_i dvs. (h_i – K/n), vi kan kalle denne for h^c_i

Kva er stor verdi av leverage?

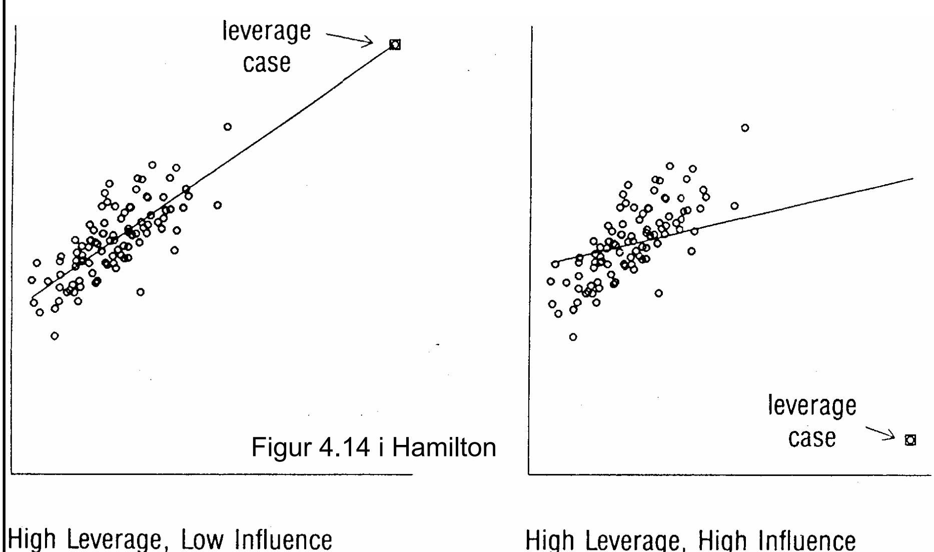
- Slik som med DFBETAS kan det stillast opp alternative kriterium. Dei er alle avhengig av utvalsstorleiken n.
  - Dersom h_i > 2K/n (eller h^c_i > K/n) finn vi dei ca 5% største h_i ; alternativt
    - Dersom max (h_i) ≤ 0.2 har vi ikkje problem
    - Dersom 0.2 ≤ max (h_i) ≤ 0.5 er der ein viss risiko for problem
    - Dersom 0.5 ≤ max (h_i) har vi truleg eit problem

Sentrert leverage (h^c_i) frå regresjonen i tabell 3.2 i Hamilton

Max av h^c_i er 0.102



Skilnad mellom leverage og påverknad



Figur 4.14 i Hamilton

Leverage observatoren finst i mange andre case observatorar

– Variansen til den i-te residualen $\operatorname{var}[e_i] = s_e^2[1-h_i]$

– Standardisert residual (*ZRESID i SPSS) $z_i = \dfrac{e_i}{s_e\sqrt{1-h_i}}$

– Studentifisert residual (*SRESID i SPSS) $t_i = \dfrac{e_i}{s_{e(i)}\sqrt{1-h_i}}$

– og hugs at standardavviket til residualen er $s_e = \sqrt{RSS/(n-K)}$

Total påverknad: Cook's D_i

- Cook's distanse D_i måler påverknad på heile modellen, ikkje på dei einskilde koeffesientane slik som $DFBETAS_{ik}$

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

der z_i er den standardiserte residualen

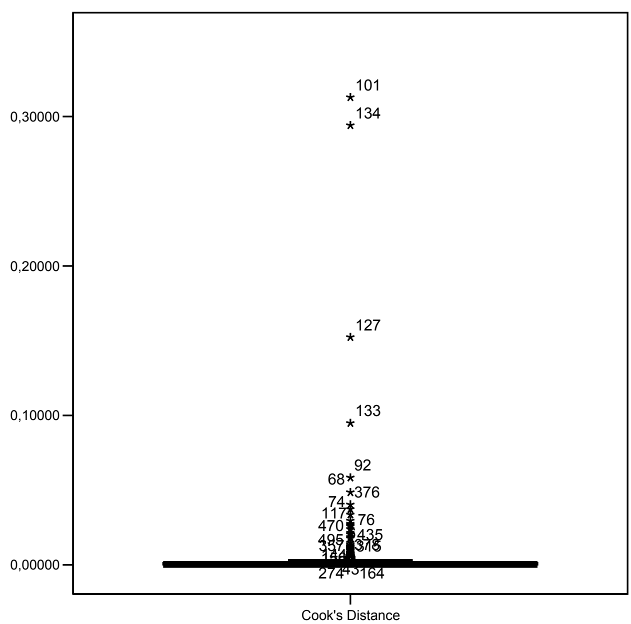
og h_i er hatt observatoren (leverage)

Kva er ein stor D_i ?

- Det kan vere verd å sjå på alle
  - $D_i > 1$ alternativt
  - $D_i > 4/n$
- Sjølv om eit case har låg D_i kan det likevel vere slik at det verkar inn på storleiken til einskildkoeffesientar (har stor $DFBETAS_{ik}$)

Cook's distanse D_i frå regresjonen i tabell 3.2 i Hamilton

Sjå også tabell 4.4 (s133) i Hamilton



Oppsummering

Kva kan gjerast med utliggjarar og case med stor påverknad? Vi kan

- undersøkje om det er feil i data. Ved feil i data kan case fjernast frå analysen
- undersøkje om transformasjon til symmetri hjelper
- rapportere to likningar: med og utan casa som påverkar urimeleg mye
- skaffe meir data

Multikollinearitet

- svært høge korrelasjonar mellom x-variablar
- sjekk korrelasjonar mellom parameterestimat
- sjekk om toleransen (den delen av variasjonen i x som ikkje er felles med andre variablar) er mindre enn t.d. 0,1
- VIF= variansinflasjonsfaktor= 1/toleranse
- dersom multikollinearitet skuldast kvadrering av variablar eller interaksjonsledd er det ikkje problematisk

Toleranse

- Mengda av variasjon i ein variabel x_k som er unik for variabelen vert kalla toleransen til variabelen
- La R^2_k vere determinasjonskoeffesienten i regresjonen av x_k på dei andre x-variablane. Dei andre x-variablane forklarer andelen R^2_k av variasjonen i x_k.
- Da er 1- R^2_k den unike variasjonen, dvs. Toleransen = 1- R^2_k
- Ved perfekt multikollinearitet vil R^2_k = 1 og toleransen = 0
- Låge verdiar av toleransen gjer regresjonsresultata mindre presise (større standardfeil)

VariansInflasjonsFaktoren (VIF)

- standardfeilen til regresjonskoeffesienten b_k kan skrivast

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

- Her er 1/toleransen = 1/(1-R^2_k) = VIF
- Om alt anna er likt vil lågare toleranse (større VIF) hos x_k gi høgare standardfeil for b_k [den aukar med ein faktor lik kvadratrota av (VIF)]

Indikatorar på multikollinearitet

- Beste indikatoren er toleransen eller VIF (denne er basert på R^2_k)
- Andre indikatorar er
  - Korrelasjon mellom einskildvariable (upåliteleg)
  - Inklusjon / eksklusjon av einskildvariablar gir store endringar i effektane til andre variablar
  - Uventa forteikn til effekten av ein variabel
  - Standardiserte regresjonskoeffesientar større enn 1 eller mindre enn -1
  - Korrelasjon mellom parameterestimat

Toleranse og VIF frå regresjonen i tabell 3.2 i 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

Kva er for låg toleranse?

Når $R^2_k > 0{,}9$ er toleransen < 0,1 og VIF > 10

Multiplikatoren for standardfeilen er da kvadratrota av VIF (ca 3.2)

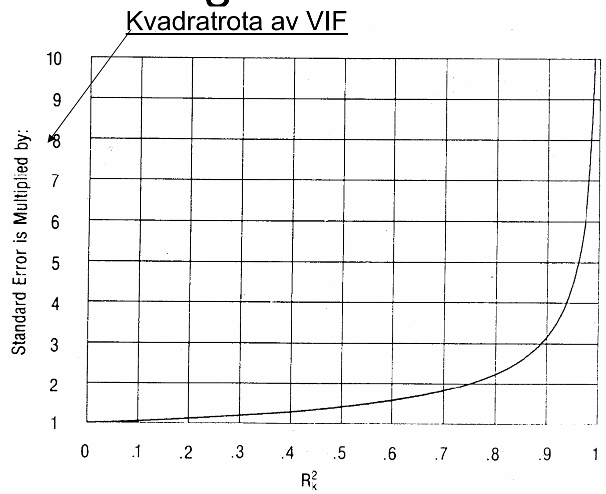


Figure 4.15 Effect of multicollinearity on standard errors (simplified).

Når er multikollinearitet eit problem?

- Det er ikkje eit problem dersom årsaka er kurvelinearitet eller interaksjonsledd i modellen. Men vi må i testinga ta omsyn til at parameterestimat for variablar med høg VIF er upresise. Vi testar dei som gruppe med F-testen
- Når det skuldast at to variablar måler same omgrep kan den eine droppast eller dei kan kombinerast til ein indeks
- Det er eit problem dersom vi treng estimat av variablane sine separate effektar (når kunnskap om deira samla effekt ikkje er nok)

Oppsummering (1)

- Når vi har normalfordelte og identisk uavhengig feil er OLS estimata betre eller like gode som andre mogelege estimat
- Men føresetnadene er sjeldan oppfylt fullt ut, vi må sjekke i kva grad dei er oppfylt
- Mange problem kan rettast opp dersom vi veit om dei
- Sjekk tidleg om det er problem med kurvelinearitet, utliggjarar eller heteroskedastisitet (t.d. gjennom spreiingsdiagram)

Oppsummering (2)

- Gjer meir nøyaktige granskingar gjennom residualplott og leverage plott
  - Kurvelinearitet (leverage plott, residual mot predikert Y plott)
  - Heteroskedastisitet (leverage plott, [absolutt verdi av residual] mot predikert Y plott)
  - Ikkje-normale residualar (kvantil-normal plott, box-plott med analyse av median og IQR/1.35)
  - Påverknad (sjekk DFBETAS og Cook's D)
- Når vi ikkje kan oppdage alvorlege problem vil vi ha større tiltru til konklusjonane