

SOS3003
**Anvendt statistisk dataanalyse i
samfunnsvitenskap**

Oversikt over

Forelesing 1-12

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Fall 2004

© Erling Berge 2004

1

PENSUM SOS 3003

- Hamilton, Lawrence C. 1992 "Regression with graphics", Belmont, Duxbury, Kap. 1-7 .
- Hardy, Melissa A. 1992 "Regression with dummy variables" Sage University Paper: QASS 93, London, Sage,
- Allison, Paul D. 2002 "Missing Data" Sage University Paper: QASS 136, London, Sage,

Fall 2004

© Erling Berge 2004

2

Mål for kurset

- Kunne lese faglitteratur som drøftar ”kvantitative” data kritisk
 - Vi skal kjenne fallgruvene
- Gjennomføre enkle analysar av samvariasjon i ”kvantitative” data
 - Vi skal demonstrere at vi kjenner fallgruvene

Fall 2004

© Erling Berge 2004

3

VARIABEL: SENTRALTENDENS

- **GJENNOMSNIITT** $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Summen av verdiane på variabelen for alle einingane dividert på talet av einingar
- **MEDIAN**
- Den verdien i ei ordna fordeling som har halvparten av einingane på kvar side
- **MODUS**
- Den typiske verdien. Den verdien i ei fordeling som har høgast frekvens.

Fall 2004

© Erling Berge 2004

4

VARIABEL: SPREDNINGSMÅL I

- MODALPROSENTEN
- Prosent av einingane som har verdi lik modus
- VARIASJONSBREDDA
- Differansen mellom høgaste og lågaste verdi i ei ordna fordeling
- GJENNOMSNETTSKILNADEN
- Gjennomsnittet av absoluttverdien til avviket frå gjennomsnittet
- KVARTILDIFFERENSEN
- Variasjonsbreidda for dei 50% av einingane som ligg rundt medianen ($Q_3 - Q_1$)
- MAD - Median Absolute Deviation
- Medianen til absoluttverdien til skilnaden mellom median og observert verdi: $MAD(x_i) = \text{median } |x_i - \text{median}(x_i)|$

Fall 2004

© Erling Berge 2004

5

VARIABEL: SPREDNINGSMÅL II

- STANARDAVVIKET
- Kvadratrot av gjennomsnittleg kvadrert avvik frå gjennomsnittet

$$s_y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

- VARIANSEN
- Kvadratet av standardavviket

$$s_y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

Fall 2004

© Erling Berge 2004

6

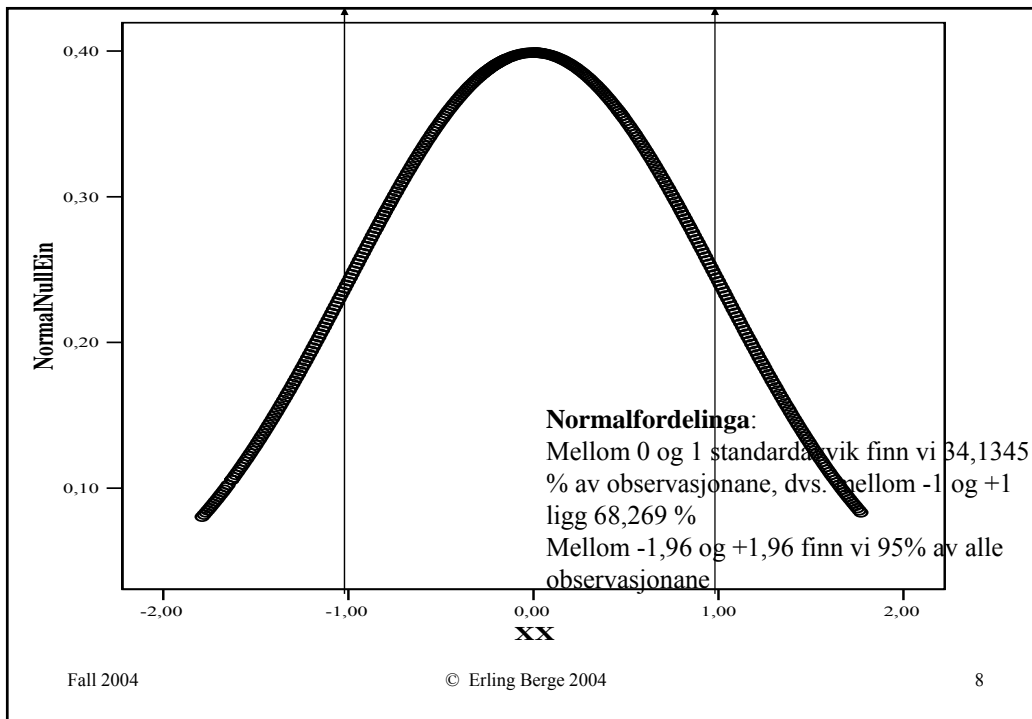
Variabel: fordelingsform I

- Symmetriske fordelingar
- Skeive fordelingar
 - ”Tunge” og ”lette” halar
- Normalfordelingane
 - Er ikkje ”normale”
 - Er eintydig fastlagt av gjennomsnitt (μ) og standardavvik (σ)

Fall 2004

© Erling Berge 2004

7



Fall 2004

© Erling Berge 2004

8

Skeive fordelingar

- Positivt skeive har $\bar{Y} > Md$
- Negativt skeive har $\bar{Y} < Md$
- Symmetriske fordelingar har $\bar{Y} \approx Md$

Symmetriske fordelingar

- Medianen og IQR er resistente mot verknader av ekstreme verdiar. Gjennomsnitt og standardavvik er ikkje resistente
- I normalfordelinga er $s_y \approx IQR/1.35$
- Dersom vi i ei symmetrisk fordeling finn
 - $s_y > IQR/1.35$ er halane tyngre enn i normalfordelinga
 - $s_y < IQR/1.35$ er halane lettare enn i normalfordelinga
 - $s_y \approx IQR/1.35$ er halane omlag slik som i normalfordelinga

Potenstransformasjonar (jfr. H:17-22)

Y^* : les “transformert Y ”
(transformasjon fra Y til Y^*)

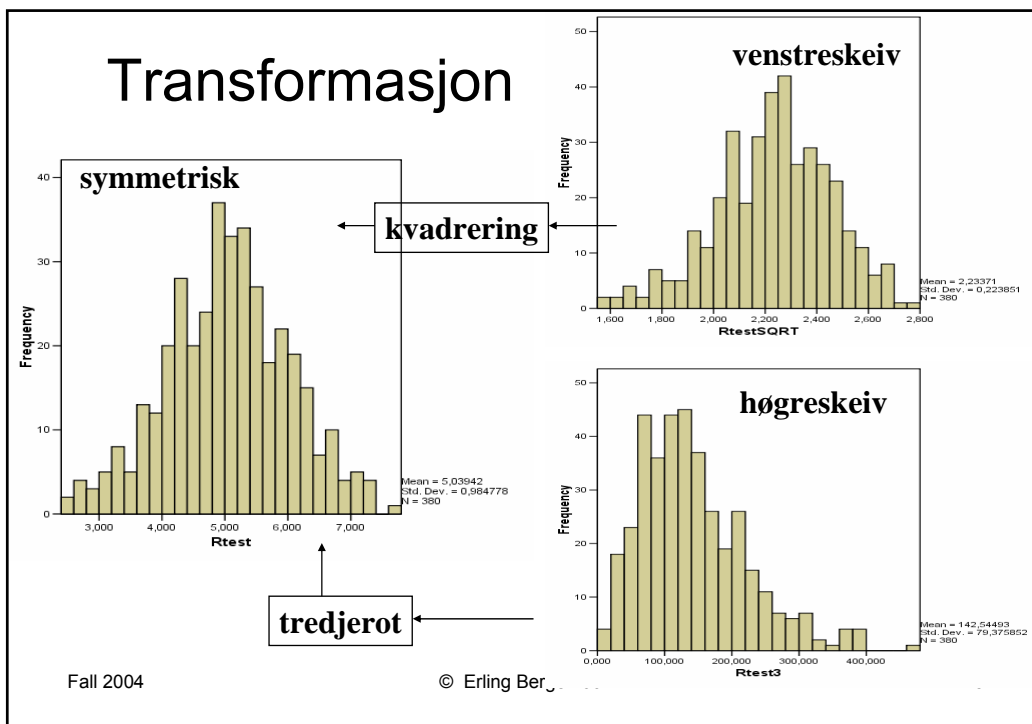
- $Y^* = Y^q$ $q > 0$
- $Y^* = \ln[Y]$ $q = 0$
- $Y^* = - [Y^q]$ $q < 0$

Invers transformasjon
(transformasjon fra Y^* til Y)

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

Potenstransformasjonar: konsekvensar

- $X^* = X^q$
 - $q > 1$ **aukar tyngda** til øvre hale relativt til nedre
 - $q = 1$ gir identitet
 - $q < 1$ **reduserer tyngda** til øvre hale relativt til nedre
- Dersom $Y^* = \ln(Y)$ vil regresjonskoeffisienten for ein intervallskala X variabel kunne tolkast som % endring i Y for ei einings endring i X



Variabel: fordelingsanalyse I

- Boksplott
 - Basert på kvartilverdiane og interkvartilavviket
 - Definerer nærliggjande utliggarar som dei som ligg innanfor intervalla $\langle Q_1 - 1.5IQR, Q_1 \rangle$ og $\langle Q_3, Q_3 + 1.5IQR \rangle$ og fjerntliggjande utliggarar dei som ligg utanfor grensene $\langle Q_1 - 1.5IQR, Q_3 + 1.5IQR \rangle$

Variablar: fordelingsform II

- Kvantilar er ei generalisering av kvartilar og percentilar
- Kvantilverdiane er variabelverdiane som svarar til gitte fraksjonar (kvantilar) av det samla utvalet eller observasjonsmaterialet, t.d.
 - Medianen er 0.5 kvantilen (eller 50% percentilen)
 - Nedre kvartil er 0.25 kvantilen
 - 10% percentilen er 0.1 kvantilen osv.

Fall 2004

© Erling Berge 2004

15

Variabel: fordelingsanalyse II

- Kvantilplott
 - Plott av kvantilverdi mot variabelverdi
 - Lorentzkurva er ein spesialvariant av dette (gir oss Gini-indeksen)
- Kvantil-Normalplott
 - Plott av kvantilverdiane på ein variabel mot kvantilverdiane i ei normalfordeling med same gjennomsnitt og spreieing

Fall 2004

© Erling Berge 2004

16

Formulering av modellar

- Definisjon av elementa i modellen
 - **variablar**, feilledd, populasjon og utval
- Definisjon av relasjonar mellom elementa
 - **likninga som bind elementa saman**, utvalsprosedyre, tidsrekkefølgje av hendingar og observasjonar,
- Presisering av føresetnader for bruk av gitt estimeringsmetode
 - tilhøve til substanssteori (**spesifikasjon**)
 - **fordeling og eigenskapar ved feilledd**

Fall 2004

© Erling Berge 2004

17

Elementa i modellen

- **Variablar:** fenomenet vi ønskjer å studere må kunne observerast og seiast å ha ulike tilstandar eller uttrykksformer i ulike einingar i den populasjonen vi observerer. Vi må finne variasjon.
- **Feilledd:** feilleddet er ein abstrakt sekk som inneheld alle dei mange aspekta av populasjonen som vi ikkje er i stand til å observere og inkludere i modellen.
- **Populasjon:** kven eller kva er det vi ønskje å seie noko om?
- **Utval:** idealet er eit reint tilfeldig utval, om vi ikkje kan få det må vi vite nøyaktig korleis utvalsmetoden er knytt opp mot den avhengige variabelen (fenomenet) vi ønskjer å studere

Fall 2004

© Erling Berge 2004

18

Relasjoner mellom elementa

- **Likninga:** relasjonar mellom variablar
- **Utvalsprosedyre:** skeive (biased) utval pga seleksjon og manglande data
- Tidsrekkefølgje av hendingar og observasjonar: kausal retning
- Samvariasjon, genuin/ spuriøs samvariasjon
 - Konklusjonar om kausalsamband krev genuin samvariasjon

Multipel regresjon: modell (1)

- Sett K = talet på parametrar i modellen (dvs. $K-1$ er talet på variablar).
Da kan (populasjons) modellen skrivast
- $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_{K-1} x_{i,K-1} + \varepsilon_i$$

Multippel regresjon: modell (2)

Modellen kan skrivast

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

Dette tyder at

- $E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_{K-1} x_{i,K-1}$
 $E[y_i]$ les vi som forventa verdi av y_i
- Målet med multippel regresjon er å finne nettoeffekten av ein forklaringsvariabel, dvs. effekten til variabel x etter at vi har kontrollert for variasjonen i alle dei andre forklaringsvariablane

Fall 2004

© Erling Berge 2004

21

Multippel regresjon: modell (3)

- Vi finn OLS estimata av modellen som dei b-verdiane i

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

(\hat{y}_i les vi som estimert eller "predikert" verdi av y_i)

som minimerer kvadratsummen av residualane

$$RSS = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

Fall 2004

© Erling Berge 2004

22

Fullstendig observasjon

- Ville gjere det mogeleg å sette opp ein fullstendig spesifisert modell. Dette tyder at alle variablar som kausalt verkar på det fenomenet vi studerer (Y) er observert, dvs. inkludert i likninga
- Dette er i praksis umogeleg. Derfor nyttar vi feilleddet til å samle opp uobserverte faktorar

Fall 2004

© Erling Berge 2004

23

Eksperiment og partielle effektar

- I eksperiment granskar ein kausalsamband mellom to variable med kontroll for alle mogelege andre kausale faktorar
- Multippel regresjon er ei form for etterlikning av eksperimentet – ei nest beste løysing - og ligg nært opp til det som heiter kvasi-eksperimentel forskingsdesign
- Ein finn netto effekt (partiell effekt) av ein variabel ved å fjerne effekten av andre variablar

Fall 2004

© Erling Berge 2004

24

Konklusjonar om populasjonen

- Dersom vi trekkjer utval etter utval frå same populasjonen og estimerer same modellen på alle utvala vil parametrane varierer frå gong til gong. Fordelinga vi finn kallast samplingfordelinga til parameteren.
- Statistisk testteori er basert på at samplingfordelingane til dei parametrane vi er interessert i er kjente slik at vi kan gjere oss opp ein meining om skilnaden mellom ei teoretisk hypotese (H_0) og observasjonane er rimeleg

Samplingfordelingar

- Lagar vi eit histogram over ulike estimerte verdiar av t.d. β_1 vil vi sjå at b_1 har ei fordeling (ei samplingfordeling)
- Ulike parametrar og observatorar har ulike samplingfordelingar
- Regresjonsparametrane (b 'ane) er t-fordelt
- Gjennomsnittet er normalfordelt

Hypotesetesting I

- Ein test er alltid konstruert ut frå føresetnaden at H_0 er rett
- Testkonstruksjonen fører fram til ein
 - **testobservator** (t-testen, F-testen)
- Testobservatoren er konstruert slik at den har ei kjent sannsynsfordeling, ei
 - **Samplingfordeling** (t-fordelinga, F-fordelinga)
- Dersom føresetnaden gir verdiar av observatoren som er lite sannsynlege, har vi liten grunn til å tru at føresetnaden er rett

Fall 2004

© Erling Berge 2004

27

Testen sin p-verdi

- Testen sin p-verdi gir oss det estimerte sannsynet for å observere dei verdiane vi har i utvalet eller verdiar som er enno meir gunstige ut frå teorien om at H_0 er gal dersom utvalet vårt er reint tilfeldig trekt frå ein populasjon det H_0 er rett
- Svært låge p-verdiar gjer at vi ikkje kan tru at H_0 er rett

Fall 2004

© Erling Berge 2004

28

Hypotesetesting II

	I røynda er H_0 sann	I røynda er H_0 usann
Vi konkluderer med at H_0 er sann	Metoden gir rett konklusjon med sannsyn $1 - \alpha$	<u>Feil av type II</u> (sannsyn $1 - \beta$)
Vi konkluderer med at H_0 er usann	<u>Feil av type I</u> Testnivået α er sannsynet for feil av type I	β = styrken til testen

Fall 2004

© Erling Berge 2004

29

T-test og F-test

- Kvadratsummar
 - $TSS = ESS + RSS$
 - $RSS = \sum_i (e_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2$ avstand observert – estimert verdi
 - $ESS = \sum_i (\hat{Y}_i - \tilde{Y})^2$ avstand estimert verdi – gjennomsnitt
 - $TSS = \sum_i (Y_i - \tilde{Y})^2$ avstand observert verdi – gjennomsnitt
- Testobservator
 - $t = (\mathbf{b} - \boldsymbol{\beta}) / \mathbf{SE}_b$ $SE =$ standard error
 - $F = [ESS / (\mathbf{K} - 1)] / [RSS / (\mathbf{n} - \mathbf{K})]$ $K =$ talet av parametar

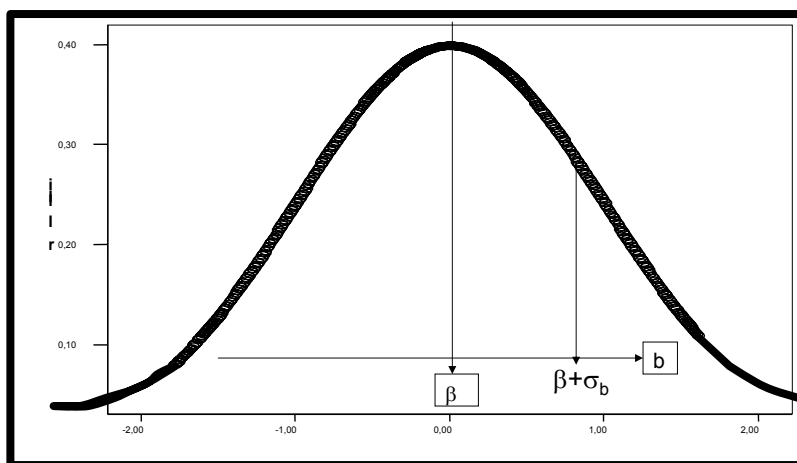
Fall 2004

© Erling Berge 2004

30

t-test

- Skilnaden mellom observert koeffisient (b_k) og uobservert koeffisient (β_k) standardisert med standardavviket til den observerte koeffisienten (SE_{b_k}) vil normalt vere svært nær null dersom den observerte b_k ligg nær populasjonsverdien (uobservert). Dette tyder at dersom vi i formelen
- $t = (b_k - \beta_k) / SE_{b_k}$ set inn $H_0: \beta_k = 0$ og finn at "t" er liten vil vi tru at populasjonsverdien β_k egentleg er lik 0. Kor stor "t" må vere for at vi skal slutte å tru at $\beta_k = 0$ kan vi finne ut frå kunnskap om samplingfordlingane til b_k og SE_{b_k}



Sampling fordeling for regresjonsparameteren b: $E[b] = \beta$

Konfidensintervall for β

- Vel ein t_α -verdi frå tabellen over t-fordelinga med $n-K$ fridomsgrader slik at dersom $H_0 : \beta_k = b_k$ er rett vil ein tosidig test ha eit sannsyn på α for å forkaste H_0 når H_0 eigentleg er rett (feil av type I)
dvs det er eit sannsyn α for at β_k eigentleg ligg utanfor intervallet $< b_k - t_\alpha(SE_{b_k}), b_k + t_\alpha(SE_{b_k}) >$
- Dette er det same som at påstanden
 $b_k - t_\alpha(SE_{b_k}) \leq \beta_k \leq b_k + t_\alpha(SE_{b_k})$
er rett med sannsyn $1 - \alpha$

Fall 2004

© Erling Berge 2004

33

F-testen: stor mot liten modell

Symbolbruk:

RSS = residual sum of squares med indeks $\{*\}$:

RSS $\{K\}$ = RSS i modellen med K parametar

RSS $\{K-H\}$ = RSS i modellen med $K-H$ parametar

(H er lik skilnaden i talet på parametar i to modellar)

Testobservatoren er

$$F_{n-K}^H = \frac{\frac{RSS\{K-H\}}{H}}{\frac{RSS\{K\}}{n-K}}$$

der F_{n-K}^H vil vere F-fordelt med H og $n-K$ fridomsgrader

Fall 2004

© Erling Berge 2004

34

Test av alle parametrene under eitt

- Dersom den store modellen har K parametrar og vi lar den vesle modellen vere så liten som mogeleg med berre 1 parameter (gjennomsnittet) vil testen vår ha $H = K - 1$.
- Set ein inn i formelen ovanfor får vi

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

Dette er F-verdien vi finn i ANOVA tabellane frå SPSS

Determinasjonskoeffisienten

Determinasjonskoeffisienten:

- $R^2 = \text{ESS}/\text{TSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$
 - Fortel kor stor del av variasjonen rundt gjennomsnittet vi ”forklarer” ved hjelp av variablane vi nyttar i regresjonen ($\hat{Y}_i = \text{predikert } y$)
- I bivariat regresjon er determinasjonskoeffisienten lik korrelasjonskoeffisienten: $r_{yu}^2 = s_{yu} / s_y s_u$
- Kovariansen $s_{yu} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})(U_i - \bar{U})$

Om val av uavhengige variablar

- Det er sjeldan eksisterande teori gir oss presise råd om kva for variablar vi skal inkludere i ein modell. Det vil som regel vere eit element av prøving og feiling i arbeidet med å utvikle ein modell.
- Når vi legg til nye variablar skjer fleire ting:
 - Forklaringskrafta aukar: R^2 vert større, men er auken signifikant?
 - Regresjonskoeffisienten viser effekten på y . Er effekten signifikant ulik 0 og så stor at den har substansiell verknad?
 - Spuriøse koeffisientar kan minke. Endrar dei nye variablane tolkinga av dei andre variablane sine effektar?

Fall 2004

© Erling Berge 2004

37

Parsimonitet

- Parsimonitet er det vi kan kalle eit estetisk kriterium på ein god modell. Vi ønskjer å forklare mest mogeleg av variasjonen i y ved hjelp av færrest mogelege variablar
- Den justerte determinasjonskoeffisienten Adjusted R^2 er basert på parsimonitet i den forstand at den tar omsyn til kompleksiteten i data relativt til modellen gjennom differansen $n-K$ (residualen sine fridomsgrader)
(n = talet på observasjonar, K = talet på estimerte parametar)

Fall 2004

© Erling Berge 2004

38

Irrelevant variabel

- Inkludere ein irrelevant variabel
 - Ein variabel er irrelevant dersom den verkelege effekten (β) ikkje er signifikant ulik 0, eller meir pragmatisk, dersom effekten av variabelen er for liten til å ha substansiell interesse.
 - **Inklusjon av ein irrelevant variabel** gjer modellen unødige kompleks og vil føre til at koeffisientestimata på alle variablane får større varians (varierer meir frå utval til utval)

Fall 2004

© Erling Berge 2004

39

Relevant variabel

- Ein variabel er relevant dersom den
 1. verkelege effekten (β) er signifikant ulik 0, og stor nok til å ha substansiell interesse og
 2. er **korrelert med andre inkluderte x-variablar**
- Dersom vi **utelet ein relevant variabel** vil alle resultat frå regresjonen verte upålitelege. Modellen er ei urealistisk forenkling.

Fall 2004

© Erling Berge 2004

40

Utvalsspesifikke resultat?

- Å velje variablar er ei avveging mellom ulike riskar. Kva for ein risk som er verst er avhengig av formålet med studien og styrken i relasjonane.
- Resultata vi finn kan godt vere utvalsspesifikke: i omlag 5% av alle utval vil ein koeffisient som ikkje er signifikant ulik null "eigentleg" vere signifikant ($\beta \neq 0$) (og tilsvarande for dei som er signifikant ulik null)
- Sjansen for å finne slike resultat aukar med talet på variablar som vert testa ("**multiple comparison fallacy**")

Fall 2004

© Erling Berge 2004

41

Nominalskalavariabel

- Kan inkluderast i regresjonsmodellar ved å lage nye hjelpevariablar: ein for kvar kategori i nominalskalavariabelen
- Dersom vi har ein intervallskala avhengig variabel og ein nominalskala uavhengig variabel vil ein tradisjonelt analysere den ved hjelp av variansanalyse (ANOVA)
- Ved introduksjon av hjelpevariable kan vi utføre same analysane i regresjonsmodellen

Fall 2004

© Erling Berge 2004

42

Referansekategori

- Dersom den kategoriske variabelen har J kategoriar kan vi maksimalt ta med J-1 hjelpevariablar i regresjonen
($H(j)$, $j=1, \dots, J-1$)
- Den utelatne hjelpevariabelen kallar vi referansekategorien
- Ein bør velje ein stor og eintydig definert kategori som referansekategori
- Den estimerte effekten av inkluderte hjelpevariablar måler effekten av å vere i den inkluderte kategorien relativt til å vere i referansekategorien

Fall 2004

© Erling Berge 2004

43

Interaksjonseffektar i regresjon

- Det er interaksjon mellom to variablar dersom effekten av den eine variabelen varierer etter kva verdi den andre variabelen har
- Enkle additive interaksjonseffektar kan vi inkludere i ein lineær modell ved hjelp av produktledd der vi multipliserer to x-variablar med kvarandre
- $\hat{y}_i = b_0 + b_1x_i + b_2w_i + b_3x_iw_i$
- Dersom vi foretar ein ikkje-lineær transformasjon av y vil alle estimerte effektar implisitt vere interaksjonseffektar
- Betinga effekt plott vil kunne illustrere kva interaksjon tyder

Fall 2004

© Erling Berge 2004

44

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
 - Ingen irrelevante variablar er inkludert: Alle inkluderte x -variablar påverkar forventa y -verdi
 - Alle relevante variablar er inkludert: Ingen andre variablar påverkar forventa y -verdi samtidig som dei korrelerer med inkluderte x -variablar

Fall 2004

© Erling Berge 2004

47

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

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

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

Da gir OLS **forventningsrette** estimat av β
(unbiased = forventningsrett)

Fall 2004

© Erling Berge 2004

48

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 $i \neq j$

FØRESETNADER: II Gauss-Markov krava (iii)

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

- a. Estimat av standardfeilen til regresjonskoeffisientane er forventningsrette, og
- b. **Gauss-Markov teoremet:**
OLS estimata har **mindre varians** enn alle andre lineære forventningsrette estimat.

OLS gir "BLUE"

(Best Linear Unbiased Estimat)

FØRESETNADER: III Normalfordelte restledd

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

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

- vil ein kunne teste hypotesar om β og σ , og
- OLS estimata vil ha mindre varians enn estimat frå alle andre forventningsrette estimatorar

**OLS gir “BUE”
(Best Unbiased Estimant)**

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 det same som at vi ikkje kan sjekke om korrelasjonen mellom feilledd og x-variabel faktisk er 0. Dette er i prinsippet det same 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

Fall 2004

© Erling Berge 2004

53

Konsekvensar av problem (Hamilton, s113)

	Problem	Uønska eigenskapar ved estimata			
		Skeive estimat av b	Skeive estimat av SE_b	Ugyldige t&Ftestar	Høg var[b]
Spesifikasjonskrav	Ikkje-lineært samband	X	X	X	-
	Utelaten 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 ϵ	X	X	X	-
Normalfordeling	ϵ er ikkje normalfordelt	0	0	X	X
Multikollinearitet		0	0	0	X

Fall 2004

© Erling Berge 2004

54

Andre problem i regresjonsanalysar som kan oppdagast (2)

- Utliggjarar (ekstreme y-verdiar)
- Innverknad (case med stor innverknad: uvanlege kombinasjonar av y og x-verdiar)
- Leverage (potensiale for innverknad)

Fall 2004

© Erling Berge 2004

55

Heteroskedastisitet (ikkje-konstant varians hos feilledet) kan skuldast

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

Fall 2004

© Erling Berge 2004

56

Autokorrelasjon (1)

- Korrelasjon mellom variabelverdier 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

Fall 2004

© Erling Berge 2004

57

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 sorteringsrekkefølga av casa. Derfor:
- Ei hypotese om autokorrelasjon må spesifisere korleis casa skal sorterast

Fall 2004

© Erling Berge 2004

58

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) jf. 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 $d > 2$

Konsekvensar av autokorrelasjon

- Hypotesetestar og konfidensintervall er upålitelege. Regresjon kan likevel gi ei god skildring av utvalet. Parametrane er forventningsrette

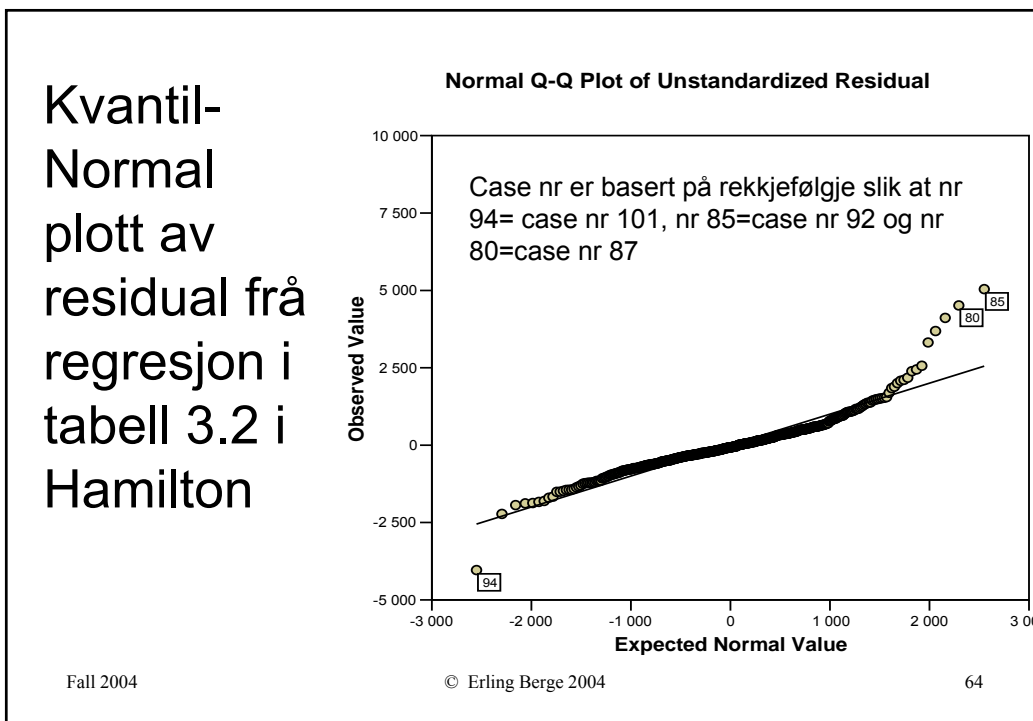
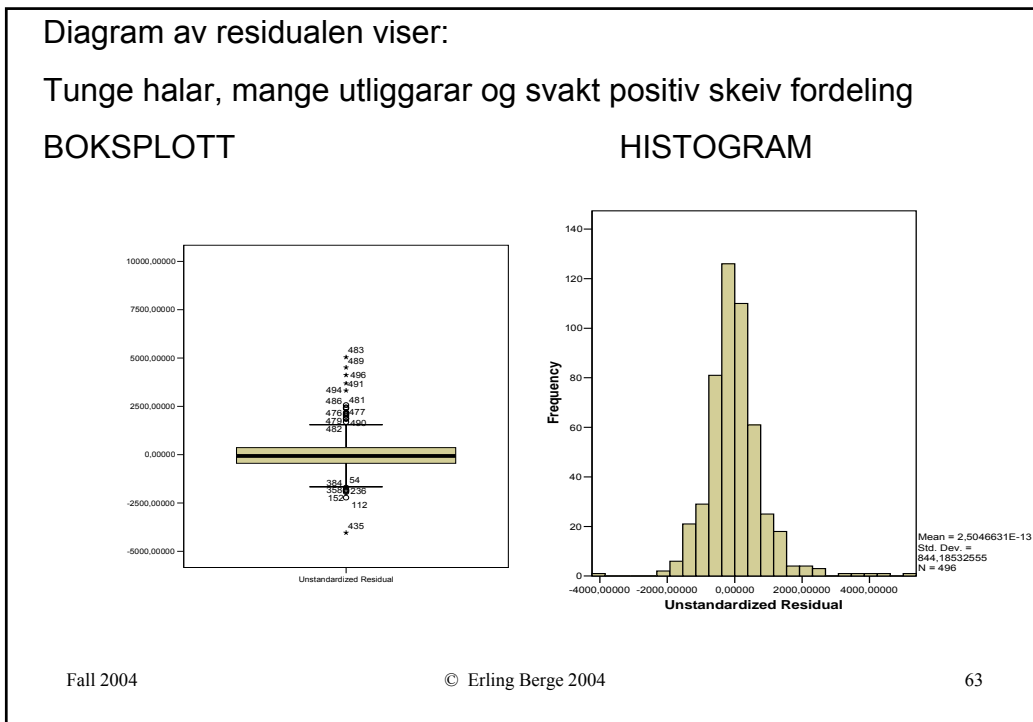
Kva kan gjerast meir?

- 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)

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, boksplott eller kvantil-normal plott



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

Fall 2004

© Erling Berge 2004

65

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

Fall 2004

© Erling Berge 2004

66

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 koeffisienten.
- Denne skilnaden målt relativt til standardfeilen til $b_{k(i)}$ vert kalla $DFBETAS_{ik}$

Fall 2004

© Erling Berge 2004

67

$DFBETAS_{ik}$

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

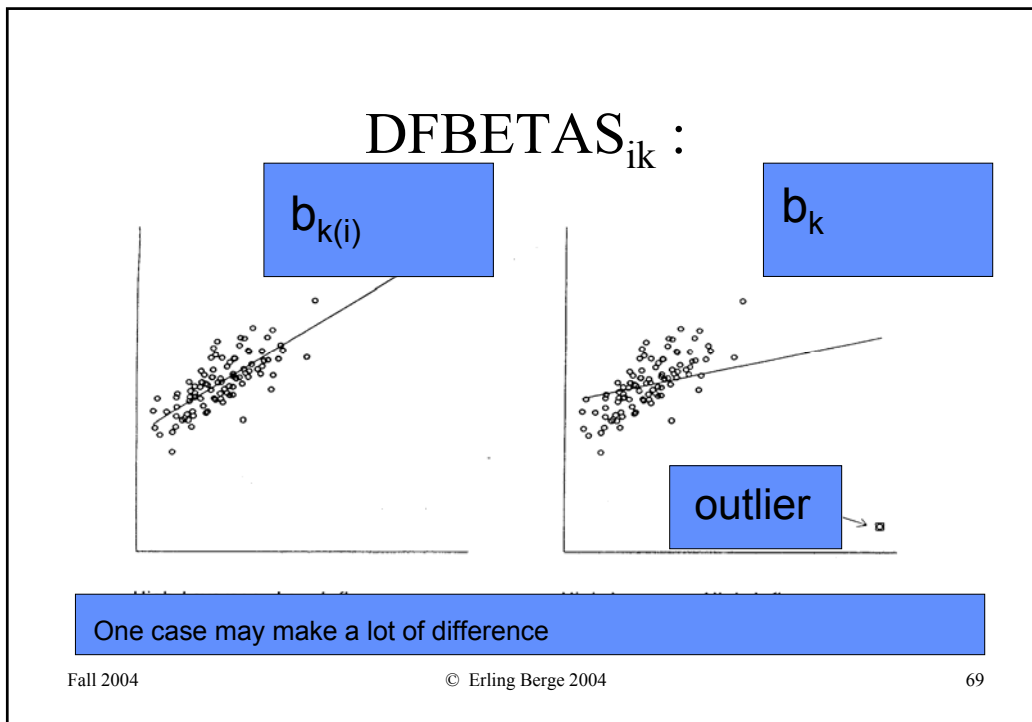
$s_{e(i)}$ er residualen sitt standardavvik når case nr i er utelate frå regresjonen

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

Fall 2004

© Erling Berge 2004

68



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: finn dei 5% største DFBETAS ved $\mathbf{IDFBETAS}_{ik} > 2/\sqrt{n}$
 - Intern skalering:
 - $Q_1 - 1.5IQR < \mathbf{IDFBETAS}_{ik} < Q_3 + 1.5IQR$
(alvorleg utliggjar i boksplott av DFBETAS_{ik})
 - Gap i fordelinga av DFBETAS_{ik}

Konsekvensar av case med stor påverknad

- Om vi oppdagar påverknadsrike case skal vi ikkje nødvendigvis ta dei ut av analysen
- Rapportar 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

Fall 2004

© Erling Berge 2004

71

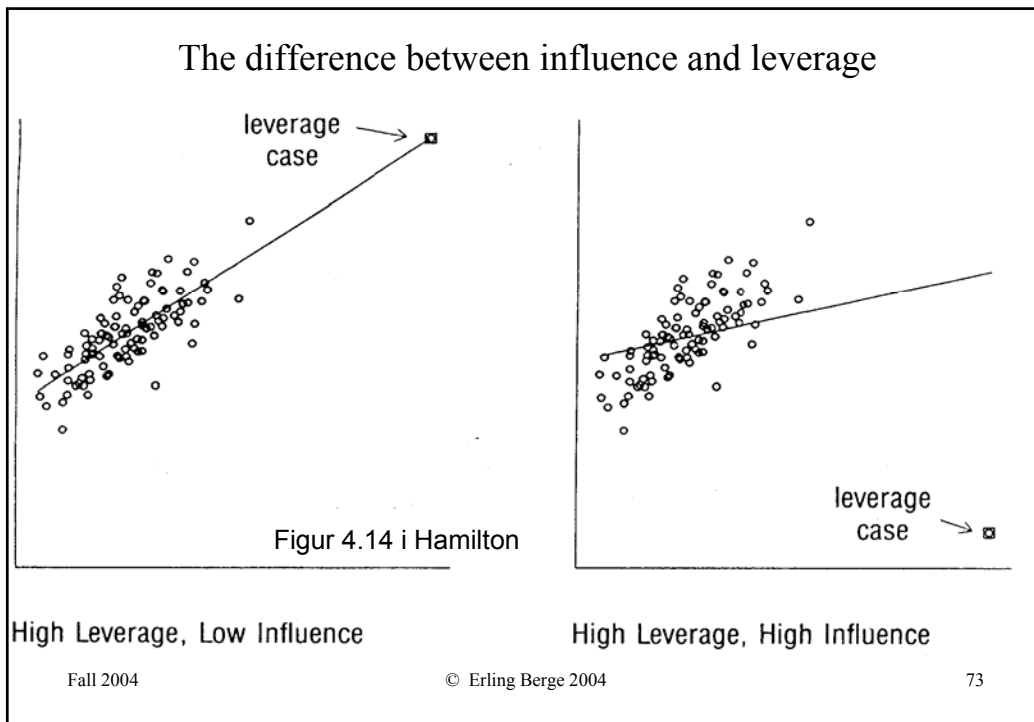
Potensiell påverknad: leverage

- 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 skriv ut $h_i - K/n$)
 - Dersom $h_i > 2K/n$ finn vi dei ca 5% største h_i ; alternativt
 - Dersom $\max(h_i) \leq 0.2$ har vi ikkje problem
 - Dersom $0.2 \leq \max(h_i) \leq 0.5$ er der ein viss risiko for problem
 - Dersom $0.5 \leq \max(h_i)$ har vi truleg eit problem

Fall 2004

© Erling Berge 2004

72



Total påverknad: Cook's D_i

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

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

der z_i er den standardiserte residualen
og h_i er hatt observatoren (leverage)

Fall 2004 © Erling Berge 2004 74

Kva er ein stor D_i ?

- Det kan vere verd å sjå på alle
 - $D_i > 1$ alternativt vil
 - $D_i > 4/n$ gir dei ca 5% største D_i
- Sjølv om eit case har låg D_i kan det likevel vere slik at det verkar inn på storleiken til einskildkoeffisientar (har stor $DFBETAS_{ik}$)

Fall 2004

© Erling Berge 2004

75

Påverknad: 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

Fall 2004

© Erling Berge 2004

76

Multikollinearitet (1)

- Multikollinearitet involverer berre x-variablane, ikkje y, og dreiar seg om lineære samband mellom to eller fleir x-variablar
- Dersom det er perfekt korrelasjon mellom to forklaringsvariablar t.d. x og w (dvs. $r_{xw} = 1$) vil den multiple regresjonsmodellen bryte saman
- Tilsvarande skjer dersom der er perfekt korrelasjon mellom to grupper av forklaringsvariable

Fall 2004

© Erling Berge 2004

77

Multikollinearitet (2)

- Perfekt multikollinearitet er svært sjeldan eit praktisk problem
- Men høge korrelasjonar mellom ulike x-variablar eller ulike grupper av x-variablar vil gjere at estimata av effekten deira blir svært usikker. Dvs. regresjonskoeffisienten vil ha svært stor standardfeil og t-testane blir i praksis uinteressante
- F-testen av ei gruppe variablar er ikkje påverka

Fall 2004

© Erling Berge 2004

78

Multikollinearitet (3)

- Når modellen omfattar kurvelineære element eller interaksjonsledd vil vi også introdusere eit visst element av multikollinearitet. Vi kan ikkje stole på testane av einskildkoeffisientar.
- Vi kan ikkje fjerne slike element utan fare for å droppe ein relevant variabel
- F-test av t.d. w og z^*w under eitt unngår testproblemet (stor varians på einskildparameter), og litt eksperimentering med ulike modellar vi vise om utelating av w eller x^*w endrar samanhengane substansielt

Fall 2004

© Erling Berge 2004

79

Toleranse

- Mengda av variasjon i ein variabel x_k som er unik for variabelen vert kalla toleransen til variabelen
- La R^2_k vere determinasjonskoeffisienten i regresjonen av x_k på dei andre x -variablane. Dei andre x -variablane forklarar 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)

Fall 2004

© Erling Berge 2004

80

Kva er for låg toleranse?

Når $R_k^2 > 0,9$ er toleransen $< 0,1$ og $VIF > 10$

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

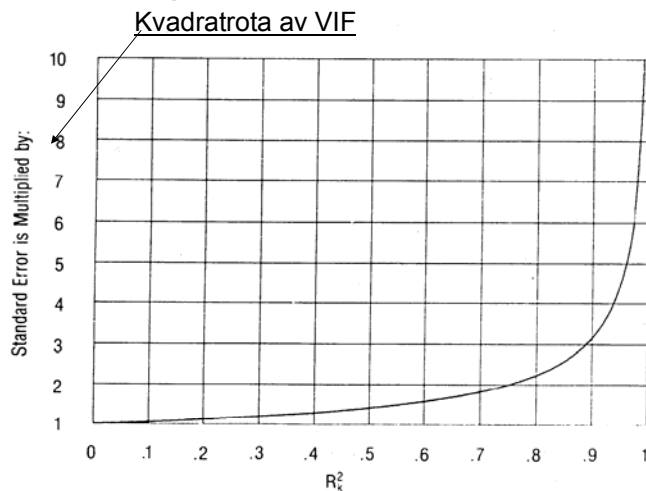


Figure 4.15 Effect of multicollinearity on standard errors (simplified).

Indikatorar for multikollinearitet

- sjekk korrelasjonar mellom parameterestimater
- sjekk om toleransen (den delen av variasjonen i x som ikkje er felles med andre variablar) er mindre enn t.d. 0,1
- $VIF = \text{variansinflasjonsfaktor} = 1/\text{toleranse}$

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
- Det er ikkje eit problem når det skuldast at to variablar måler same omgrep. Då 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)

Fall 2004

© Erling Berge 2004

83

Manglande Data

Data manglar av mange grunnar

- Personar nektar å svar
- Personar gløymer eller overser nokre spørsmål
- Personar veit ikkje noko svar
- Spørsmålet er irrelevant
- I administrative register kan somme dokument ha gått tapt
- I forskingsdesign for vanskeleg målbare variablar

Fall 2004

© Erling Berge 2004

84

Manglande data fører til problem

- Det er eit praktisk problem sidan alle statistiske prosedyrar føreset fullstendige datamatriser
- Det er eit analytisk problem sidan manglande data som regel gir skeive estimat av parametrane
- Det er eit viktig skilje mellom data som manglar av tilfeldige årsaker og dei som manglar av systematiske årsaker

Fall 2004

© Erling Berge 2004

85

Den enkle løysinga: fjern alle case med manglande data

- Listewis (Listwise/ casewise) fjerning av manglande data tyder at ein fjernar alle case som manglar data på ein eller fleire variablar inkludert i modellen
- Metoden har gode eigenskapar, men kan i somme høve ta ut av analysen mesteparten av casa
- Vanlege alternativ, som parvis ("pairwise") fjerning, har vist seg å vere dårlegare
- Nyare metodar som "maximum likelihood" og "multiple imputation" har betre eigenskapar men er krevjande
- Det løner seg å gjere god arbeid i datainnsamlinga

Fall 2004

© Erling Berge 2004

86

Konvensjonelle metodar

Vanlege metodar ved MAR (missing at random) data:

- Listewis utelating (Listwise deletion)
- Parwis utelating (Pairwise deletion)
- Dummy variabel korreksjon
- Innsetjing av verdi (Imputation)

Ingen av dei vanleg brukte metodane er tydeleg betre enn listewis utelating. Bruk dei ikkje!

Listewis utelating (1)

- Kan alltid nyttast
- Dersom data er MCAR gir det eit enkelt tilfeldig utval av det opphavslege utvalet
- Mindre n gir sjølvsagt større variansestimater
- Også når data er MAR og missing på x -variablar er uavhengig av verdien på y vil listewis utelating gi forventingsrette estimat

Listevis utelating (2)

- I logistisk regresjon er listevis utelating problematisk berre dersom missing er relatert både til avhengig og uavhengige variablar
- Når missing berre er avhengig av den uavhengige variabelen sine egne verdiar er listevis betre enn maximum likelihood og multiple imputation

Fall 2004

© Erling Berge 2004

89

Oppsummering om konvensjonelle metodar for manglande data

- Vanlege metodar utanom listevis utelating for korleksjon av manglande data gjer problema verre
- Ver nøye med datainnsamlinga slik at det er eit minimum av manglande data
- Prøv å samle inn data som kan hjelpe til med å modellere prosessen som fører til missing
- Der data manglar **bruk listevis utelating** dersom ikkje maximum likelihood eller multiple imputasjon er tilgjengeleg

Fall 2004

© Erling Berge 2004

90

Nye metodar for ignorerbare manglande data (MAR data): Maximum Likelihood (ML)

- Konklusjonar
 - Baserer seg på sannsynet for å observere nett dei variabelverdiane vi har funne i utvalet
 - ML gir optimale parameterestimat i store utval når data er MAR
 - Men ML krev ein modell for den felles fordelinga av alle variablane i utvalet som manglar data, og den er vanskeleg å bruke for mange typar modellar

Fall 2004

© Erling Berge 2004

91

Nye metodar for ignorerbare manglande data (MAR data): Multippel Imputasjon (MI)

- Konklusjonar
 - Baserer seg på ein tilfeldig komponent som vert lagt til estimat av dei einskilde manglande opplysningane
 - Har like gode eigenskapar som ML og er enklare å implementere for alle slags modellar.
 - Men den gir ulike resultat for kvar gong den blir brukt

Fall 2004

© Erling Berge 2004

92

Data som manglar systematisk

- Krev som regel ein modell av korleis fråfallet oppstår
- ML og MI tilnærmingane kan framleis nyttast, men med mye strengare restriksjonar og resultatane er svært sensitive for brot på føresetnadene

Fall 2004

© Erling Berge 2004

93

Manglande data: Oppsummering

- Dersom nok data vert igjen er listevise utelating den enklaste løysinga
- Dersom listevise utelating ikkje fungerer bør ein freiste med multippel imputasjon
- Dersom ein har mistanke om at data ikkje er MAR må ein lage ein modell for prosessen som skaper missing. Denne kan eventuelt nyttast saman med ML eller MI. Gode resultat krev at modellen for missing er korrekt

Fall 2004

© Erling Berge 2004

94

Kurvetilpassing

- Ein rett spesifisert modell krev at funksjonen som bind x-variablane og y variabelen saman er i samsvar med røyndomen: er sambandet lineært?
- Data kan granskast gjennom bandregresjon eller glatting
- Teori om kausalsambandet kan spesifisere eit ikkje-lineært samband
- For fenomen som ikkje kan representerast med ei rett linje har vi alternativa
 - Kurvelineær regresjon
 - Ikkje-lineær regresjon

Fall 2004

© Erling Berge 2004

95

Transformerte variablar

- Brukar vi transformerte variablar vert regresjonen kurvelineær. Transformasjonen gjer den opphavelige kurvesamanhengen til ein lineær samanheng
- Dette er den viktigaste grunnen til å transformere.
- Samtidig kan transformering ordne opp i ulike typar statistiske problem (utliggjarar, heteroskedastisitet, ikkje-normale feil)

Fall 2004

© Erling Berge 2004

96

Den lineære modellen

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

- I den lineære modellen kan vi transformere x-ane og y-ane utan at det har noko å seie for eigenskapane til OLS estimata i seg sjølv.
- Så lenge modellen er lineær i parametranne er OLS ein lovleg metode

Fall 2004

© Erling Berge 2004

97

Val av transformasjon

- Spreiingsplott eller teori kan gi råd
- Elles er transformasjon til symmetri det beste utgangspunktet
- Regresjonen rapportert i tabell 3.2 i Hamilton viste seg problematisk
- Regresjon med transformerte variablar kan redusere problema

Fall 2004

© Erling Berge 2004

98

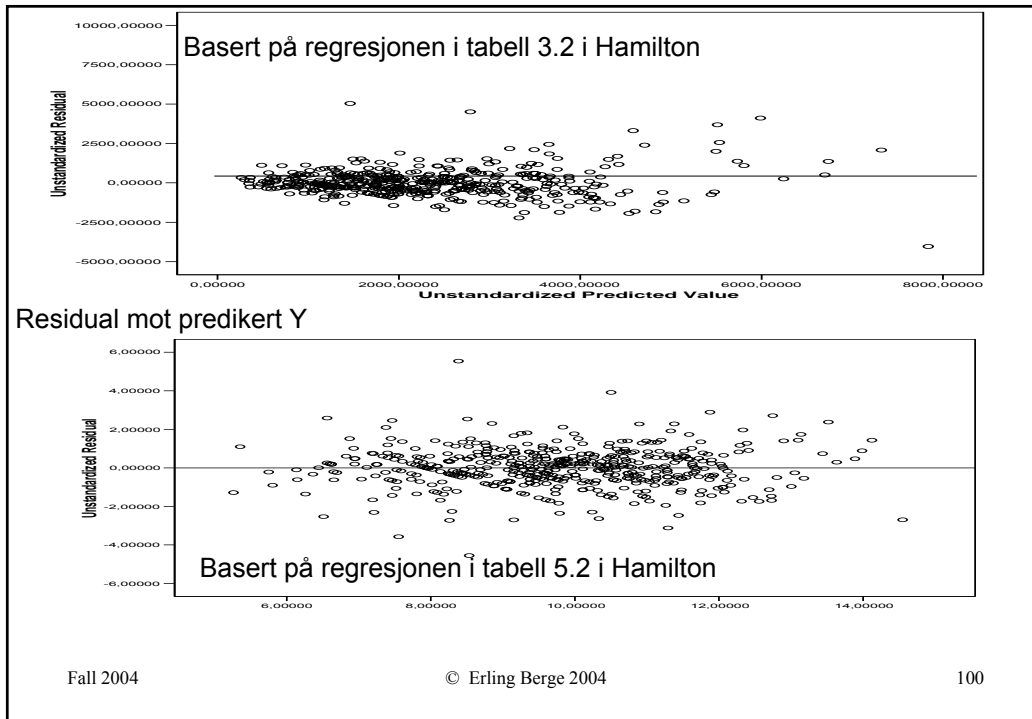
Regresjon med transformerte variable Tab 5.2 Hamilton

Dependent Variable: wtr81_3	B	Std. Error	t	Sig.
(Constant)	1,856	,385	4,822	,000
inc_3	,516	,130	3,976	,000
wtr80_3	,626	,029	21,508	,000
Education in Years	-,036	,016	-2,257	,024
head of house retired?	,101	,119	,852	,395
logpeop	,715	,110	6,469	,000
clogpeop	,916	,263	3,485	,001

Fall 2004

© Erling Berge 2004

99



Fall 2004

© Erling Berge 2004

100

Andre verknader av transformasjonane

- To case med stor innverknad på koeffisienten for inntekt (store DFBTAS) har no ikkje slik innverknad (fig. 4.11 og 5.9)
- Eit case med stor innverknad på koeffisienten for vassforbruk i 1980 har no ikkje så stor innverknad (fig. 4.12 og 5.10)
- Transformasjonar som gjer fordelingar symmetriske vil ofte løyse mange problem – men ikkje alltid!

Fall 2004

© Erling Berge 2004

101

Tolking

- Estimaten av modellen ser no slik ut

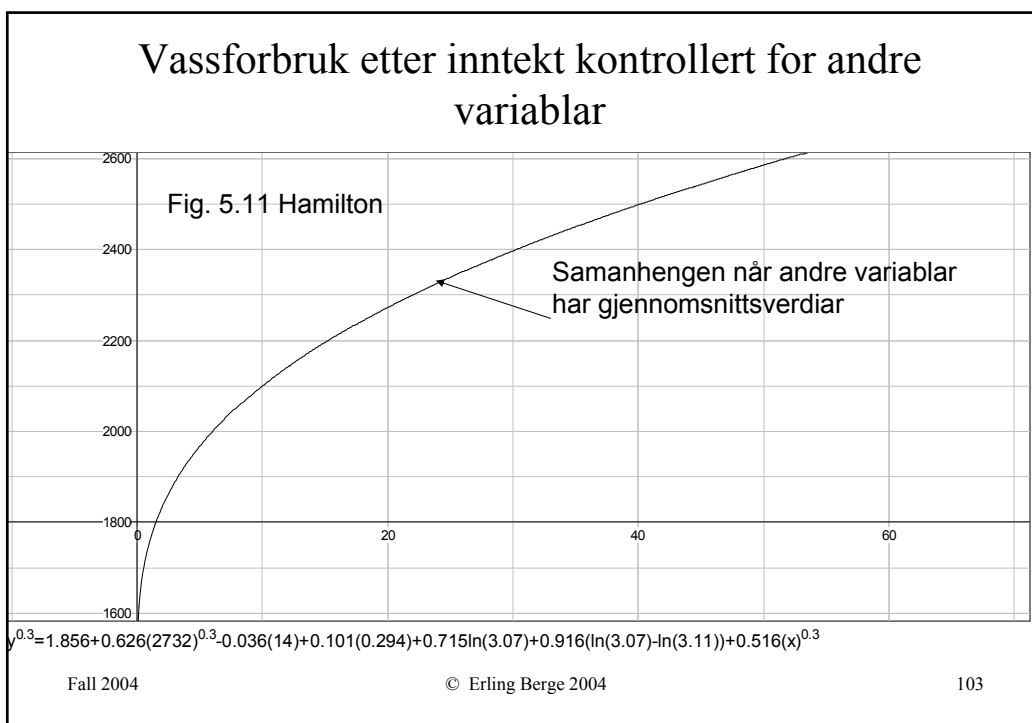
$$y_i^{0.3} = 1.856 + 0.516x_{1i}^{0.3} + 0.626x_{2i}^{0.3} - 0.036x_{3i} \\ + 0.101x_{4i} + 0.715 \ln(x_{5i}) + 0.916 \ln\left(\frac{x_{5i}}{x_{0i}}\right)$$

- Tolking av koeffisientane er ikkje lenger så enkelt (t.d.: måleiningane på parametrane er endra)
- Den enklaste måten å tolke på er å nytte betinga effekt plott

Fall 2004

© Erling Berge 2004

102



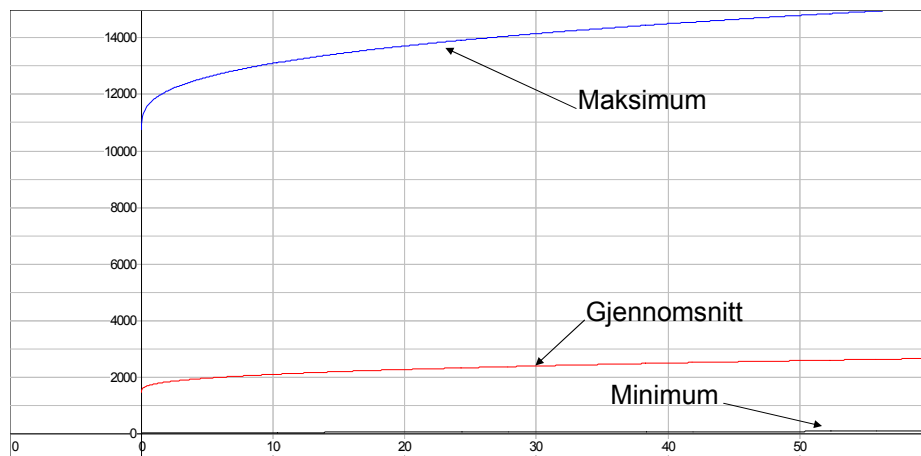
Kva er interessant å plote?

- Samanhengen inntekt vassforbruk kontrollert for ulike kombinasjonar av andre variabelverdier
 1. Dei som minimerer vassforbruket
 2. Dei som maksimerer vassforbruket
 3. Gjennomsnittverdiane

- 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 $y^{0.3} = (1.856 + 0.626(2732)^{0.3} - 0.036(14) + 0.101(0.29) + 0.715\ln(3.07) + 0.916(\ln(3.07) - \ln(3.11)) + 0.516(x)^{0.3})$

Fall 2004
© Erling Berge 2004
104

Samanlikning av tre typar brukssituasjon



Samanhengen mellom inntekt og vassforbruk Fig. 5.12 Hamilton

Fall 2004

© Erling Berge 2004

105

Konstanten si rolle i plottet

- Den einaste skilnaden mellom dei tre kurvene er konstanten
 - I maksimumskurva er (konst) = 14.046
 - I minimumskurva er (konst) = 4.204
 - I gjennomsnittskurva er (konst) = 8.507

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

- Effekten av inntekt varierer med verdien av (konst), dvs. verdien av dei andre variablane
- Når vi transformerer avhengig variabel vert **alle** samanhengar til interaksjonseffektar

Fall 2004

© Erling Berge 2004

106

Samanlikning av effektar

- I somme samanhengar kan ein nytte den standardiserte regresjonskoeffisienten til å samanlikne effektar, men den er sensitiv for skeive estimat av standardfeilen
- Ein meir generell metode er å samanlikne betinga effekt plott der skaleringa på y-aksen er halden konstant

Fall 2004

© Erling Berge 2004

107

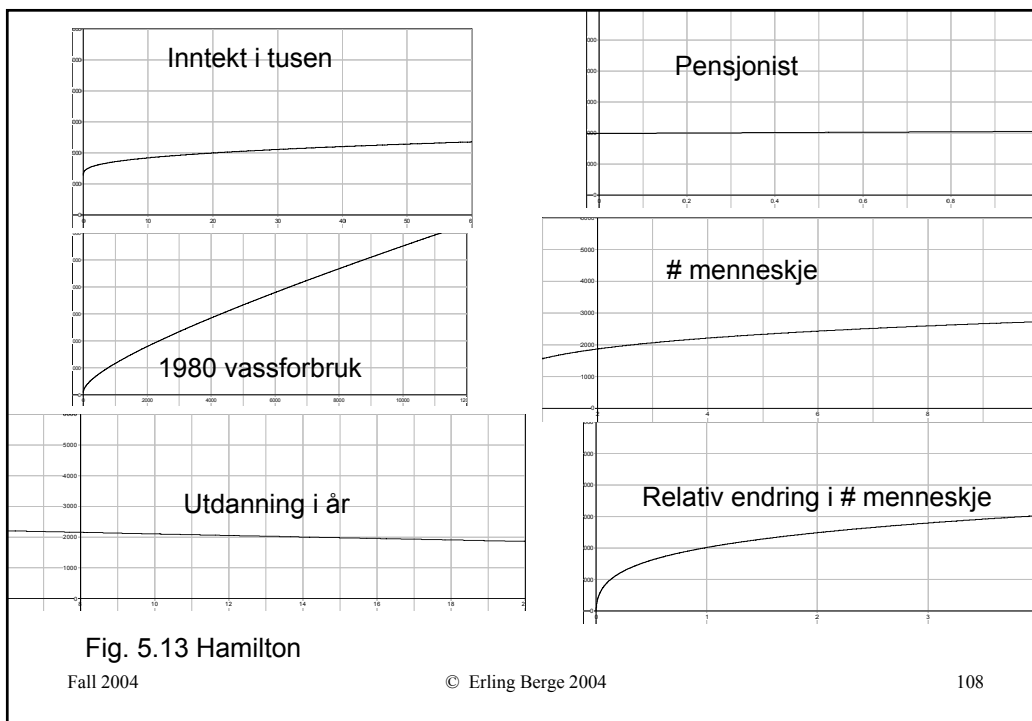


Fig. 5.13 Hamilton

Fall 2004

© Erling Berge 2004

108

Ikkje-lineære modellar

- Dersom vi ikkje har modellar som er lineære i parametrane vil ein trenge andre teknikkar for å estimere parametrane
- Det kan vere to typar argument for slike modellar
 - Teori om den kausale mekanismen kan diktere ein slik modell
 - Inspeksjon av data kan peike på ein bestemt type modell

Fall 2004

© Erling Berge 2004

109

Logistiske modellar

- Den logistiske funksjonen skriv ein
$$y = \frac{\alpha}{1 + \gamma \exp(-\beta x)} + \varepsilon$$
- Når x veks mot uendeleg vil y nærme seg α
- Når x minkar mot minus uendeleg vil y nærme seg 0
- Logistiske modellar passar til mange fenomen
 - Vekst i biologiske populasjonar
 - Spreiing av rykte
 - Spreiing av sjukdom

Fall 2004

© Erling Berge 2004

110

Logistiske modellar: parameterertolking

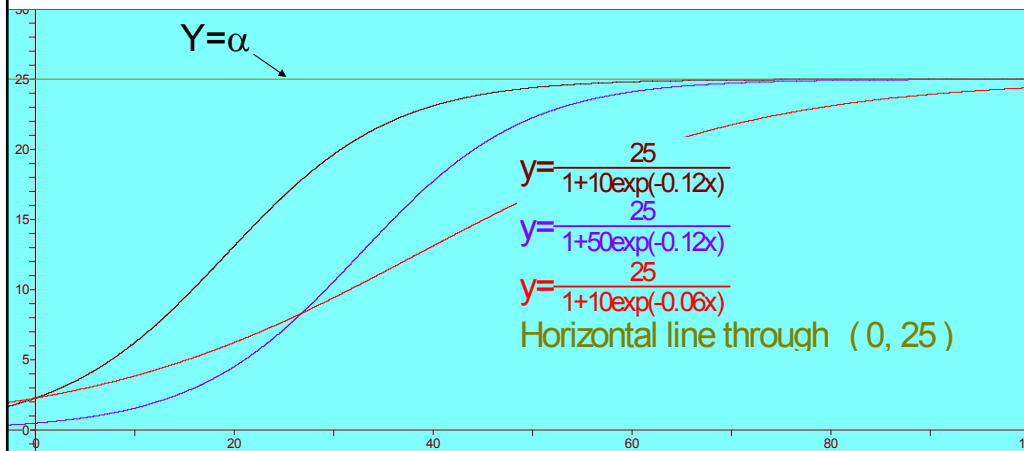
- γ fastset kvar veksten startar,
- β avgjer kor rask veksten er,
- Skifte av forteikn i eksponenten (dersom eksponenten er positiv, $+\beta x$) vil y minke med aukande x

Fall 2004

© Erling Berge 2004

111

Logistiske kurver Fig. 5.17 Hamilton



Fall 2004

© Erling Berge 2004

112

Logistisk sannsynsmodell

- Dersom ein set $\alpha=\gamma=1$ vil y variere mellom 0 og 1 når x varierer mellom minus uendeleg og pluss uendeleg.
- Logistiske kurver kan da nyttast til å modellere sannsyn

$$y_i = \frac{1}{1 + \exp(-\beta x_i)} + \varepsilon_i$$

Fall 2004

© Erling Berge 2004

113

Estimering av ikkje-lineære modellar

- Kriteriet på tilpassing er framleis minimum RSS
- Ein kan sjeldan finne analytiske uttrykk for parametrane. Ein må gjette på ein startverdi og gå igjennom fleire iterasjonar for å finne kva parameterverdi som gir den minste RSS verdien
- Gode startverdiar er som regel nødvendig og alt frå teori til inspeksjon av data vert brukt for å finne dei

Fall 2004

© Erling Berge 2004

114

Konklusjonar (1)

- Dataanalyse startar ofte med lineære modellar. Dei er enklast.
- Teori eller utforskande dataanalyse (bandregresjon, glatting) kan seie oss om kurvelineære eller ikkje-lineære modellar trengst
- Transformasjon av variable gir kurvelineær regresjon. Dette kan motverke fleire problem
 - Kurvelinearitet i samanhengane
 - Case med stor påverknad
 - Ikkje-normale feil
 - Heteroskedastisitet

Fall 2004

© Erling Berge 2004

115

Konklusjonar (2)

- Ikkje-lineær regresjon nyttar iterative prosedyrar for å finne parameterestimat.
- Prosedyrane treng initialverdiar og er ofte sensitive for initialverdiane.
- Tolking av parametrar kan vere vanskeleg. Grafar som viser sambanda for ulike parameterverdiar vil hjelpe mye

Fall 2004

© Erling Berge 2004

116

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)

Fall 2004

© Erling Berge 2004

117

Utliggjarar er eit problem for OLS

Utliggjarar verkar inn på estimat av

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

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

Fall 2004

© Erling Berge 2004

118

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.

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)

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

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 betre når feila er ikkje-normale
- Eigenskapane vert fastlagt ved ein "kalibreringskonstant" (c , i formlane)

Fall 2004

© Erling Berge 2004

123

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.

Fall 2004

© Erling Berge 2004

124

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

Fall 2004

© Erling Berge 2004

125

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 t.d. 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

Fall 2004

© Erling Berge 2004

126

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

Fall 2004

© Erling Berge 2004

127

Avgrensa påverknad: vektmodifikasjon

- Vi utvidar vektfunksjonen med ei vekt basert på innverknad (leverage) observatoren h_i
- Påverknadsfaktoren i vektinga kan t.d. setjast til
- $w_i^H = 1$ dersom $h_i \leq c^H$
 - $w_i^H = (c^H / h_i)$ dersom $h_i > c^H$
 - c^H vert ofte sett lik 90% percentilen i fordelinga av h_i
 - IRSL vekta vert da $w_i w_i^H$ der w_i er enten Tukey eller Huber vektar som endrar seg frå iterasjon til iterasjon medan w_i^H er konstant

Fall 2004

© Erling Berge 2004

128

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

LOGISTISK (LOGIT) REGRESJON

- **Skal nyttast når avhengig variabel er på nominalnivå**
- Den enklaste modellen føreset at Y har verdiane 0 eller 1
- Modellen av den betinga forventninga til Y, $E[Y | X]$, nyttar den logistiske funksjonen
- Men kvifor kan ikkje $E[Y | X]$ vere ein lineær funksjon også her?

Den lineære sannsynsmodellen: LPM

- Den lineære sannsynsmodellen (LPM) brukt på Y_i når Y_i berre kan ta to verdier (0,1) føreset at vi kan tolke $E[Y_i | \mathbf{X}]$ som eit sannsyn
- $E[Y_i | \mathbf{X}] = b_0 + \sum_j b_j x_{ji} = \Pr[Y_i = 1]$
- Dette fører til problem

Fall 2004

© Erling Berge 2004

131

Er føresetnadene rette i LPM?

- Ein føresetnad i LPM er at residualen e_i stettar krava til OLS
- Residualen er anten $e_i = 1 - (b_0 + \sum_j b_j x_{ji})$ eller $e_i = 0 - (b_0 + \sum_j b_j x_{ji})$
- Dette tyder heteroskedastisitet (residualen varierer med storleiken på x-variablane)
- Det finst estimeringsmetodar som kan komme rundt dette problemet (2-steps vekta minste kvadrats metode til dømes)

Fall 2004

© Erling Berge 2004

132

LPM er feil modell

- Eit anna problem, at ein for rimelege verdiar av x -ane kan får ein verdi av predikert y der $E[Y_i | \mathbf{X}] > 1$ eller $E[Y_i | \mathbf{X}] < 0$, kan ein ikkje gjere noko med
- LPM er substansielt sett feil modell
- Det trengst ein modell der ein alltid har $0 < E[Y_i | \mathbf{X}] < 1$

Fall 2004

© Erling Berge 2004

133

Oddsens: definisjon

Definisjonar

- Sannsynet for at person i skal ha verdien 1 på variabelen Y skriv vi $\Pr(Y_i=1)$. Da er $\Pr(Y_i \text{ ulik } 1) = 1 - \Pr(Y_i=1)$
- Oddsens for at person i skal ha verdien 1 på variabelen Y_i , her kalla O_i , er tilhøvet mellom dei to sannsyna

$$O_i(y_i = 1) = \frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)} = \frac{p_i}{1 - p_i}$$

Fall 2004

© Erling Berge 2004

134

Logiten: definisjon

- LOGITEN , L_i , er den naturlege logaritmen til oddsen, O_i , for person i:

$$L_i = \ln(O_i)$$

- Modellen føreset at L_i er ein lineær funksjon av forklaringsvariablane x_j , dvs:
- $L_i = \beta_0 + \sum_j \beta_j x_{ji}$, der $j=1, \dots, K-1$, og $i=1, \dots, n$

Fall 2004

© Erling Berge 2004

135

Modellen av sannsynet

- Sett \mathbf{X} = (samlinga av alle x_j), da er sannsynet for at $Y_i = 1$ for person nr i

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

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

Grafen til dette sambandet er nyttig for tolkinga av kva ei endring i x tyder

Fall 2004

© Erling Berge 2004

136

LOGISTISK REGRESJON: Estimering

- Metoden brukt for å estimere parametrane i modellen er Maximum Likelihood
- ML-metoden gir oss dei parametrane som maksimerer sannsynet (Likelihood) for å finne dei observasjonane vi faktisk har
- Dette sannsynet skal vi kalle \mathcal{L}

Fall 2004

© Erling Berge 2004

137

Likelihood: definisjon

- Likelihooden er lik produktet av sannsynet for kvar einskild observasjon. For ein dikotom variabel der $\Pr(Y_i = 1) = P_i$ kan dette skrivast

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

Fall 2004

© Erling Berge 2004

138

LogLikelihood: definisjon

- For lettere å kunne maksimere likelihooden, \mathcal{L} , tar ein den naturlege logaritmen til \mathcal{L} :

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

- Den naturlege logaritmen til \mathcal{L} kallar vi LogLikelihooden, eller \mathcal{LL}
- \mathcal{LL} har ei sentral rolle i logistisk regresjon

Maksimering av LogLikelihood

- **LogLikelihooden = $\mathcal{LL} = \ln(\mathcal{L}) = \log_e(\mathcal{L}) = \sum_i \{ Y_i \log_e P_i + (1 - Y_i) \log_e(1 - P_i) \}$, $i = 1, 2, 3, \dots, n$,**
- \mathcal{LL} er alltid negativ. Maksimering av \mathcal{LL} er derfor likeverdig med minimering av den positive LogLikelihooden (dvs. $-\mathcal{LL}$)

Iterativ estimering

Estimeringa vart avslutta ved iterasjon nr 4 sidan parameterestimata endra seg med mindre enn 0,001.

Utdrag frå Hamilton Tabell 7.1	Iteration	-2 Log Likelihood	Coefficients	
			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

Iterasjon 0: Utgangspunktet er ein modell med berre konstantledd

Testobservatorane

To testobservatorar er aktuelle

- (1) Sannsynsrateobservatoren er høvetalet mellom to sannsyn (Likelihoodar)
”Likelihood ratio test”
Denne kan nyttast analogt med F-testen i OLS
- (2) Wald observatoren kan nyttast til å lage testar analogt med t-observatoren i OLS regresjon

Sannsynsratetesten (1)

- Sannsynsratetesten :
- Differansen mellom **LogLikelihooden** (\mathcal{LL}) til to modeller estimert på same datamateriale kan nyttast til å teste to nesta modeller mot kvarandre omlag som F observatoren i OLS regresjon
- Testen kan og nyttast på einskildkoeffisientar. I små utval er den betre enn Wald testen

Fall 2004

© Erling Berge 2004

143

Sannsynsratetesten (2)

Sannsynsrate test observatoren

$$\chi^2_H = -2[\mathcal{LL}(\text{modell1}) - \mathcal{LL}(\text{modell2})]$$

vil, dersom nullhypotesen om ingen skilnad mellom modellane er rett, vere tilnærma (for store n) kjikvadratfordelt med fridomsgrader lik differansen i talet på parametar i dei to modellane (H)

Fall 2004

© Erling Berge 2004

144

Wald testen

Wald testen

- Wald (kvikvadrat) observatoren (oppgitt av SPSS) = $t^2 = (b_k / SE(b_k))^2$ (slik t er brukt av Hamilton)
- Observatoren $t = b_k / SE(b_k)$ vil kunne nyttast til testing av einstilte parametrar omlag som t-observatoren i OLS regresjon
- Gitt at nullhypotesen er rett vil t (for store n) i logistisk regresjon vere tilnærma normalfordelt
- Gitt at nullhypotesen er rett vil Wald observatoren (for store n) i logistisk regresjon vere tilnærma Kvikvadratfordelt med $df=1$

Konfidensintervall for parameterestimater

- Konfidensintervall for parameterestimater kan konstruerast ut frå at kvadratrot av Wald observatoren med 1 fridomsgrad er tilnærma normalfordelt
- $b_k - t_\alpha * SE(b_k) < \beta_k < b_k + t_\alpha * SE(b_k)$
der t_α er tabellverdien teken frå normalfordelinga med signifikansnivå α
- I mangel av tabellar over normalfordeling kan ein gjere seg nytte av at **t-fordeling** er tilnærma lik normalfordelinga ved store n-K (t.d. $n-K > 120$)

Modellen av sannsynet for at vi skal observere $y=1$ for person i

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

der logiten $L_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji}$ er ein lineær funksjon av forklaringsvariablane

Ut frå formelen er det ikkje lett å tolke kva koeffisientane β_j tyder

Fall 2004

© Erling Berge 2004

147

TOLKING: ODDS og ODDSRATER

- Logiten, L_i , ($L_i = \beta_0 + \sum_j \beta_j x_{ji}$) er definert som den naturlege logaritmen til oddsen.

Det tyder at

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

og

- **oddsraten** = $O_i(Y_i=1 | L_i') / O_i(Y_i=1 | L_i)$
– der L_i' og L_i har ulik verdi for ein x_j .

Fall 2004

© Erling Berge 2004

148

Oddsraten

- Oddsraten, **O**, kan tolkast som den **relative effekten** av å ha **ein** variabelverdi heller enn ein annan
- t.d. dersom $x_{ki} = t+1$ i L_i' og $x_{ki} = t$ i L_i
- $O = O_i(Y_i=1 | L_i') / O_i(Y_i=1 | L_i)$
= $\exp[L_i'] / \exp[L_i]$
= $\exp[\beta_k]$
- Kvifor β_k ?

Fall 2004

© Erling Berge 2004

149

Oddsraten: eksempel¹ (1)

- Odds for å svare ja =
 $e^{b_0 + b_1 * Alder + b_2 * Kvinne + b_3 * E. utd + b_4 * Barn \ i \ HH}$
- Oddsraten for å svare ja mellom kvinner og menn =

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

¹ Hugs reknereglane for potensar

Fall 2004

© Erling Berge 2004

150

Oddsraten: eksempel (2)

- Oddsraten for å svare ja for eitt års tilvekst i utdanning

$$\frac{e^{b_0 + b_1 * Alder + b_2 * Kvinne + b_3 * (E.utd + 1) + b_4 * Barn_i_HH}}{e^{b_0 + b_1 * Alder + b_2 * Kvinne + b_3 * E.utd + b_4 * Barn_i_HH}} = e^{b_3}$$

Fall 2004

© Erling Berge 2004

151

BETINGA EFFEKT PLOTT

- Gi faste verdier til alle x variablar unnateke ein, t.d. variabel x_k og set desse inn i likninga for logiten
- Plott $\Pr(Y=1)$ som funksjon av x_k , dvs
- $P = 1/(1 + \exp[-L]) = 1/(1 + \exp[-\text{konst} - b_k x_k])$
for rimelege verdier av x_k
- “konst” er konstanten ein får ved å setje inn i logiten dei valde faste variabelverdiane og summere dei

Fall 2004

© Erling Berge 2004

152

Utdrag frå Hamilton Tabell 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			

Logiten:

$$L = 2.182 - 0.04 \cdot \text{lived} - 0.197 \cdot \text{educ} + 1.299 \cdot \text{contam} + 2.279 \cdot \text{hsc} - 1.731 \cdot \text{nodad}$$

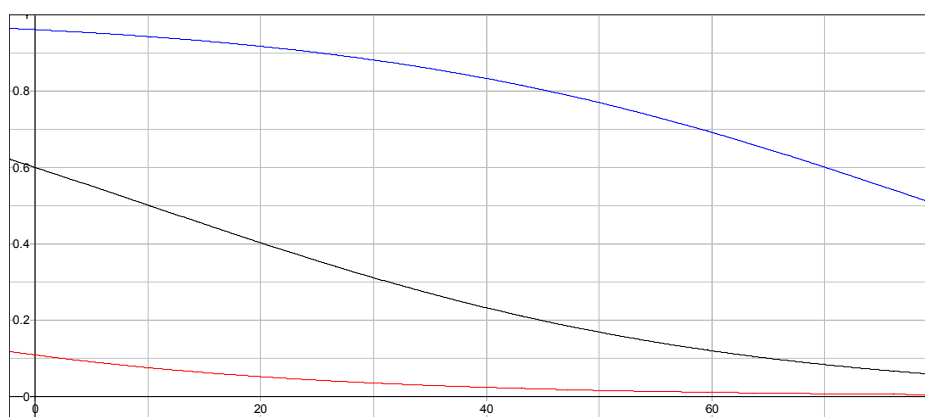
Her lar vi "lived" variere og set inn høveleg valde verdier for dei andre

Fall 2004

© Erling Berge 2004

153

Betinga effekt plott frå Hamilton tabell 7.4 (fig7.5) effekten av å bu lenge i byen



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

$$y = 1 / (1 + \exp(-(-2.182 - 0.04x - 0.197 \times 12.95 + 1.299 \times 1 + 2.279 \times 1 - 1.731 \times 0)))$$

$$y = 1 / (1 + \exp(-(-2.182 - 0.04x - 0.197 \times 12.95 + 1.299 \times 0 + 2.279 \times 0 - 1.731 \times 1)))$$

Fall 2004

© Erling Berge 2004

154

Determinasjonskoeffisientar

- I logistiske regresjonsmodellar finst ikkje mål tilsvarande determinasjonskoeffisienten i OLS regresjon
- Fleire analoge mål har vore foreslått
- Dei er vert ofte kalla pseudo R^2
- Hamilton nyttar Aldrich og Nelson sitt
pseudo $R^2 = \chi^2 / (\chi^2 + n)$
der χ^2 = testobservatoren for testen av heile modellen mot ein modell med berre konstant, og n = er talet på case

Fall 2004

© Erling Berge 2004

155

LOGISTISK REGRESJON: FØRESETNADER

- Modellen er korrekt spesifisert
 - logiten er lineær i parametranne
 - alle relevante variablar er med
 - ingen irrelevante er med
- x-variablane er målt utan feil
- Observasjonane er uavhengige
- Ikkje perfekt multikollinearitet
- Ikkje perfekt diskriminering
- (Stort nok utval)

Fall 2004

© Erling Berge 2004

156

FØRESETNADER som ikkje kan testast

- Modellen er korrekt spesifisert
 - Om alle relevante variablar er med
- x-variablane er målt utan feil
- Observasjonane er uavhengige

To føresetnader vil teste seg sjølve

- Ikkje perfekt multikollinearitet
- Ikkje perfekt diskriminering

Statistiske problem kan komme av

- For lite utval
- Høg grad av **multikollinearitet**
 - Fører til store standardfeil (usikre estimat)
 - Vert oppdaga og handtert på same måten som i OLS regresjon
- Høg grad av **diskriminering** (eller separasjon)
 - fører til store standardfeil (usikre estimat)
 - Vert oppdaga automatisk av SPSS

Diskriminering/ separasjon

- Problem med diskriminering dukkar opp når vi for ein gitt x-verdi får nesten perfekt prediksjon av y-verdien (nesten alle med ein gitt x-verdi har same y-verdi)
- I SPSS kan dette gi følgjande melding:

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.

Fall 2004

© Erling Berge 2004

159

Statistiske problem: linearitet i logiten?

- Kurvelinearitet i logiten kan gi skeive parameterestimat
- Spreiingsplott for y-x er lite informative sidan y berre har to verdiar
- For å teste om Logiten er lineær i ein x-variabel kan vi gjere følgjande
 - Gruppere x-variabelen
 - For kvar gruppe finne y-gjennomsnitt og rekne det om til logit
 - Lag ein graf av logitane mot gruppert x

Fall 2004

© Erling Berge 2004

160

Statistiske problem: påverknad

- Påverknad frå utliggjarar og uvanlege x-verdiar er like problematisk i logistisk regresjon som i OLS regresjon
- Transformasjon av x-variablar til symmetri vil minimere innverknaden til ekstreme variabelverdiar
- Store residualar er indikator på stor innverknad

Fall 2004

© Erling Berge 2004

161

Påverknad: residualar

- Det finst ulike måtar å standardisere residualar på:
 - ”Pearsonresidualar” og
 - ”Avviksresidualar”
- Påverknad kan baserast på
 - Pearsonresidualen
 - Avviksresidualen
 - Leverage (potensiale for påverknad): dvs. observatoren h_j

Fall 2004

© Erling Berge 2004

162

To mål på innverknad

- Pearson Kji-kvadrat observatoren basert på personresidualen r
$$\chi_P^2 = \sum_{j=1}^J r_j^2$$
- Avviks kjikvadrat observatoren basert på avviksresidualen d
$$\chi_D^2 = \sum_{j=1}^J d_j^2$$
- Formlane er dei same anten vi reknar case eller mønster

Fall 2004

© Erling Berge 2004

163

Kjikvadrat observatorane

Begge kjikvadrat observatorane

1. Pearson Kjikvadratet χ_P^2 og
 2. Avviks Kjikvadratet χ_D^2
- kan lesast som ein test av nullhypotesen om ingen skilnad mellom den estimerte modellen og ein "metta modell", dvs. ein modell med like mange parametrar som case / mønster

Fall 2004

© Erling Berge 2004

164

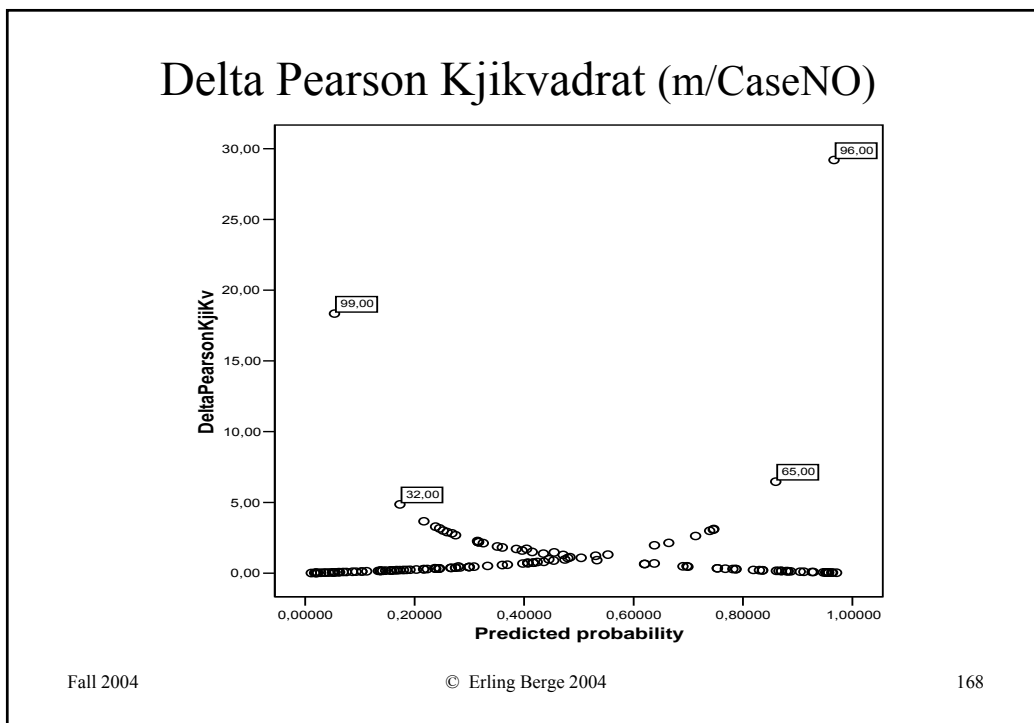
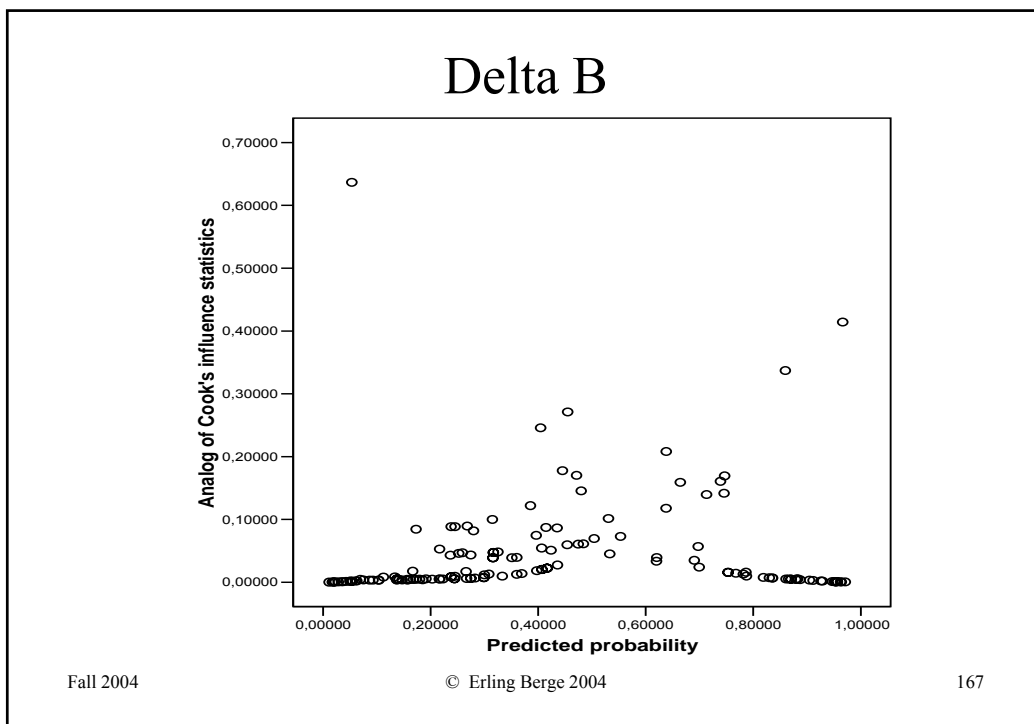
Diagnosegrafar

Utliggjarplott kan baserast på plott av estimert sannsyn for $Y_i=1$ (estimert P_i) mot

- Delta B , ΔB_j , eller
- Delta Pearson Kjikvadratet, $\Delta \chi^2_{P(j)}$, eller
- Delta Avviks Kjikvadratet, $\Delta \chi^2_{D(j)}$

Kva er store verdiar av $\Delta \chi^2_{P(i)}$ og $\Delta \chi^2_{D(i)}$

- $\Delta \chi^2_{P(i)}$ og $\Delta \chi^2_{D(i)}$ måler begge kor dårleg modellen passar med mønsteret j. Store verdiar indikerer at modellen ville passe til data mye betre dersom alle case med mønsteret vart utelatne.
- Sidan begge måla er asymptotisk kjikvadratfordelt vil verdiar større enn 4 indikere at eit mønster påverkar parameterestimata ”signifikant”



Frå Case til Mønster

- Figurane ovanfor er ikkje identisk med Hamilton sine figurar
- Hamilton har korrigert for effekten av identiske mønster

Fall 2004

© Erling Berge 2004

169

Påverknad ved felles mønster av x-variablar

- I logistisk regresjon med få variablar vil mange case ha dei same verdiane på alle x-variablane. Kvar kombinasjon av x-variabelverdiar kallar vi eit mønster.
- Når mange case har same mønster, kan kvart case ha liten innverknad medan dei samla kan ha uvanleg stor innverknad på parameterestimata
- Påverknadsrike mønster i x verdiane kan dermed gi skeive parameterestimata

Fall 2004

© Erling Berge 2004

170