SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing XI

- Logistisk regresjon III
  Hamilton Kap 7 s233-242

LOGISTISK REGRESJON: FØRESETNADER

- Modellen er korrekt spesifisert
  - logiten er lineær i parametrane
  - 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

FØRESETNADER som ikkje kan testast

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

To vil teste seg sjølve

- Ikkje perfekt multikollinearitet
- Ikkje perfekt diskriminering

LOGISTISK REGRESJON
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.

Diskriminering Hamilton tabell 7.5

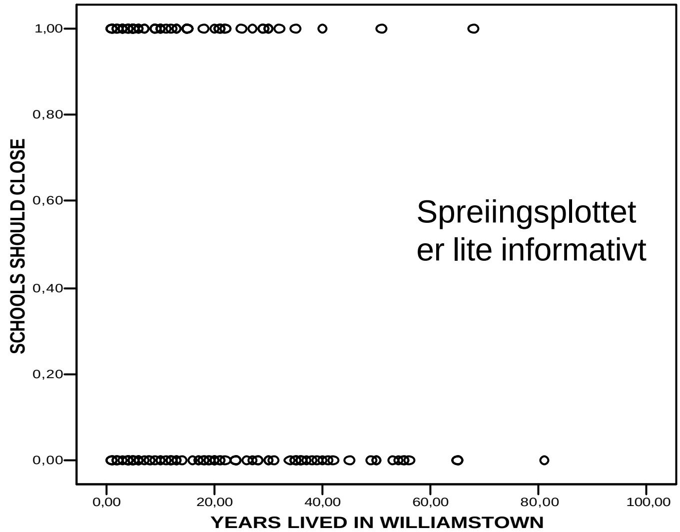
- Odds for svakare krav er 44/202 = 0,218 mellom kvinner utan småbarn
- Odds for svakare krav er 0/79 = 0 mellom kvinner med småbarn
- Oddsraten er 0/0,218 = 0 slik at exp{b_{kvinne}}=0
- Dette tyder at b_{kvinne} = minus uendeleg

	Kvinne utan små barn	Kvinne med små barn
Ikkje svakare krav	202	79
Svakare krav OK	44	0

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

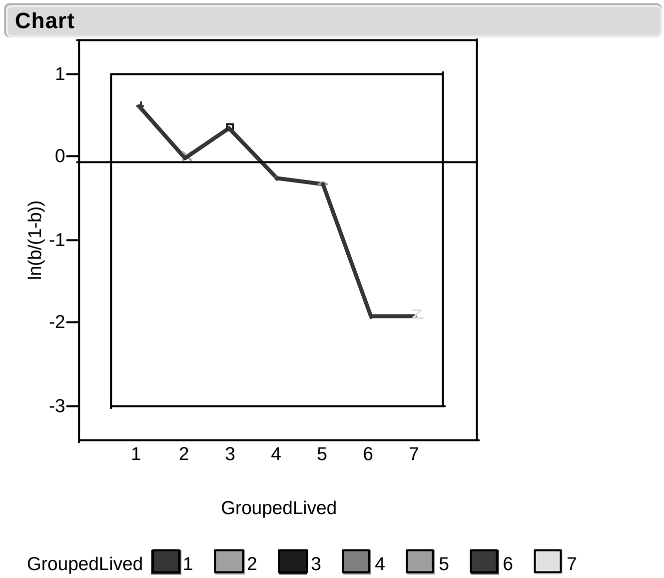
Y=Lukke skolen mot x= år budd i byen



Linearitet i logiten: eksempel

SCHOOLS SHOULD CLOSE		YEARS LIVED IN WILLIAMSTOWN (Banded)						
		<= 3	4-6	7-11	12-22	23-33	34-44	45+
N	OPEN	7	14	7	22	11	13	13
N	CLOSE	13	14	10	17	8	2	2
Within group	Mean (=p)	,65	,50	,59	,44	,42	,13	,13
Logit	Ln(p/(1-p))	0,619	0	0,364	-0,241	-0,323	-1,901	-1,901

Er logiten lineær i ”år budd i byen”?



Tja, kanskje det.

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

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

Diagnosegrafar

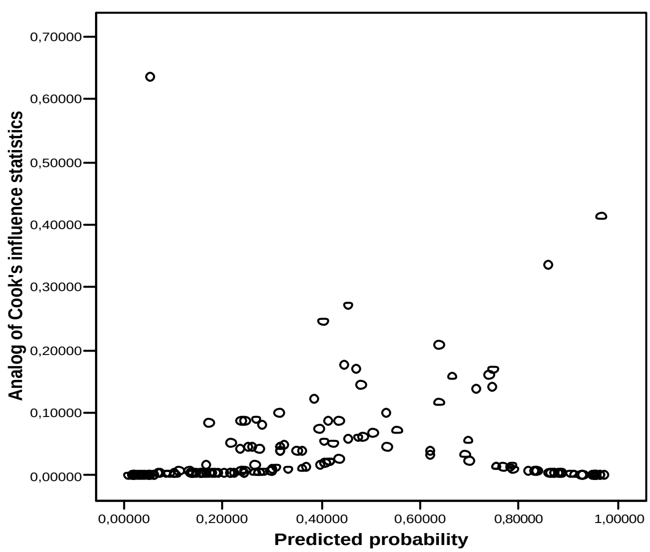
Utliggjarplott kan baserast på plott av estimert sannsyn for $Y_i=1$ (estimert P_i) mot

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

SPSS output

- **Cook's = delta B hos Hamilton**
  - The logistic regression analog of Cook's influence statistic. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients.
- **Leverage Value = h hos Hamilton**
  - The relative influence of each observation on the model's fit.
- **DfBeta(s)** er ikkje nytta av Hamilton i logistisk regresjon
  - The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.

Delta B



SPSS output frå ”Save” (1)

- **Unstandardized Residuals.**
  - The difference between an observed value and the value predicted by the model.
- **Logit Residual**. (ukjen tolking)
  - Den storleiken programmet rapporterer er ikkje i samsvar med dokumentasjonen

SPSS output frå ”Save” (2)

- **Standardized = Pearson residual**
  - På komandoen ”standardized” vil SPSS skrive ut noko som vert kalla **ZRE_1** (normaized residual)
  - Dette er det same som Pearson residualen hos Hamilton
- **Studentized = [SQRT(delta avvikskjikvadrat)]**
  - På komandoen ”Studentized” vil SPSS skrive ut noko det kallar **SRE_1** (standard residual)
  - Dette er det same som kvadratrota av ”delta

Utrekning av $\Delta\chi^2_{P(i)}$

- Med utgangspunkt i dei storleikane SPSS gir oss kan vi rekne ut ”delta Pearson-kjikvadratet”
- Der det står r_j i formelen set vi inn ZRE_1 og der det står h_j set vi inn LEV_1

$$\Delta c^2_{P(j)} = \frac{r_j^2}{\left(1 - h_j\right)}$$

Utrekning av $\Delta\chi^2_{D(i)}$

Med utgangspunkt i dei storleikane SPSS gir oss kan vi rekne ut ”delta Avviks-kjikvaratet”

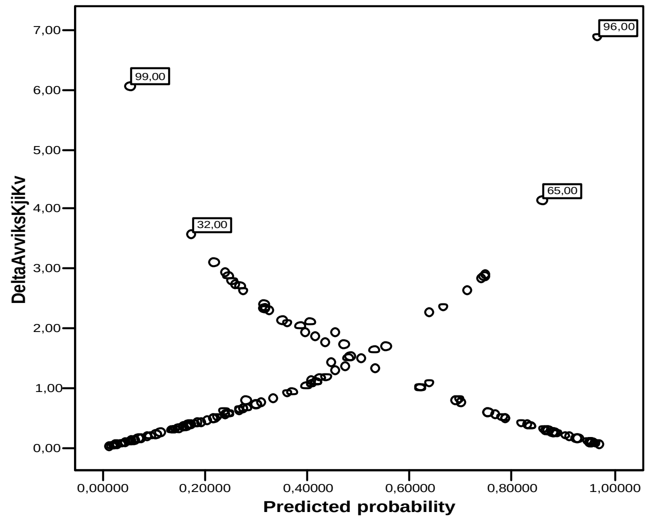
1. For å finne ”delta avvikskjikvadratet ” kvadrerer vi SRE_1

$$\Delta c^2_{D(j)} = SRE\_1 * SRE\_1$$

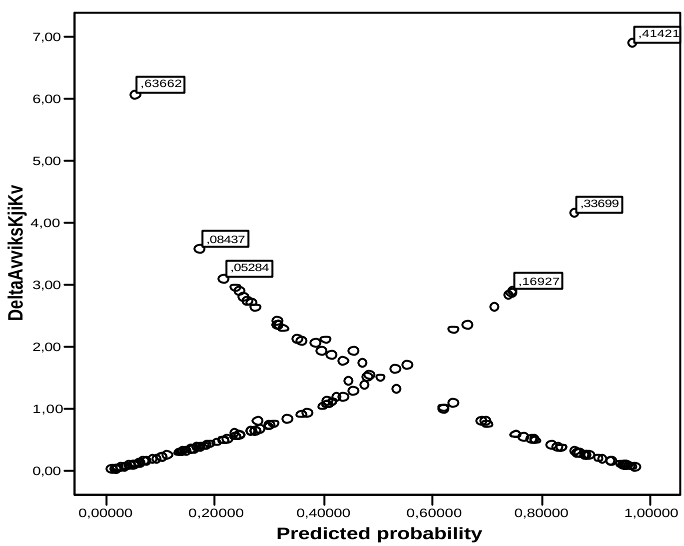
2. Alternativt set vi inn d_j=DEV_1 og h_j=LEV_1 i formelen

$$\Delta c^2_{D(j)} = \frac{d_j^2}{\left(1 - h_j\right)}$$

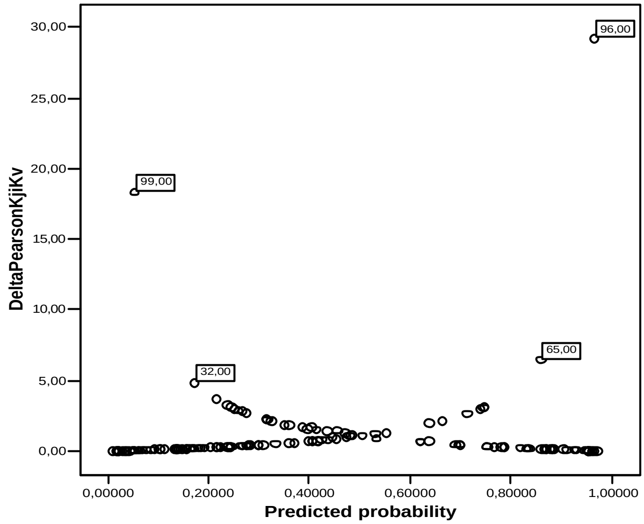
DeltaAvviksKvikvadrat (m/CaseNO)



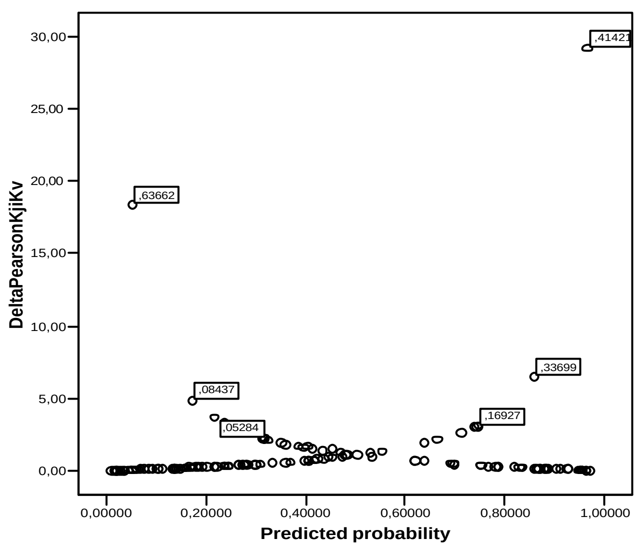
DeltaAvviksKvikvadrat (m/delta B)



Delta Pearson KjiKvadrat (m/CaseNO)



Delta Pearson KiiKvadrat (m/ delta B)



Case med stor innverknad

Variables	Case No 96	Case No 65	Case No 99
Y=close	**1,00**	**,00**	**,00**
lived	68,00	40,00	1,00
educ	12,00	12,00	12,00
contam	,00	1,00	1,00
hsc	,00	1,00	1,00
nodad	,00	,00	,00
PRE_1	**,05**	**,86**	**,97**
COO_1	,64	,34	,41
RES_1	,95	-,86	-,97
SRE_1	2,46	-2,04	-2,62

Variables	Case No 96	Case No 65	Case No 99
ZRE_1	4,21	-2,48	-5,36
DEV_1	2,42	-1,98	-2,61
DFB0_1	-,32	,01	-,36
DFB1_1	,01	,00	,00
DFB2_1	,02	,01	,02
DFB3_1	-,08	-,15	-,18
DFB4_1	-,06	-,17	-,19
DFB5_1	-,08	,16	,14
DeltaPearsonKjiKv	18,34	6,47	29,20
DeltaAvviksKjiKv	6,07	4,14	6,89

Frå Case til Mønster

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

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-variabel verdiar 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 parameterestimat

Påverknad: Mønster i x-verdiar

- Predikert verdi, og dermed residualen vil vere lik for alle case som har same mønster
- Påverknad frå mønster j kan finnast ved hjelp av
  - Frekvensen til mønsteret
  - Pearsonresidualen
  - Avviksresidualen
  - Leverage: dvs. observatoren h_j

Finne X-Mønster ved hjelp av SPSS

- I ”Data” – menyen finn vi kommandoen ”Identify duplicate cases”
- Marker dei x-variablane som skal nyttast i modellen og flytt dei til ”Define matching cases by”
- Kryss av for ”Sequential count of matching cases in each group” og ”Display frequencies for created variables”
- Dette lagar to nye variablar. Den eine, ”MatchSequence”, nummerer sekvensielt 1, 2, … der fleire mønster er identiske. Der mønsteret er unikt får variabelen verdien 0.
- Den andre variabelen ”Primary…” har verdien 0 for duplikat og 1 for unike mønster

X-Mønster i SPSS i Hamilton s238-242

	Frequency	Percent	Valid Percent	Cumulative Percent
Duplicate Case	21	13,7	13,7	13,7
Primary Case	132	86,3	86,3	100,0
Total	**153**	**100,0**	**100,0**	

Sequential count of matching cases	Frequency	Percent	Valid Percent	Cumulative Percent
0 [115 mønster med 1 case]	115	75,2	75,2	75,2
1 [17 mønster med 2 eller 3 case]	17	11,1	11,1	86,3
2 [17 – 4=13 mønster med 2 case]	17	11,1	11,1	97,4
3 [4 mønster med 3 case]	4	2,6	2,6	100,0
Total	**153**	**100,0**	**100,0**	

Hamilton tabell 7.6 Symbolbruk

J	Talet av unike mønster av x-verdiar i data (J<=n)
m_j	Talet av case med mønsteret j (m>=1)
$\hat{P}_j$	Predikert sannsyn for Y=1 for case med mønster j
Y_j	Sum av y-verdiar for case med mønster j (=talet av case med mønster j og y=1)
r_j	Pearsonresidual for mønster j
c_P^2	Pearsonkjikvadrat observator
d_j	Avviksresidual for mønster j
c_D^2	Avvikskjikvadrat observator
h_i	Leverage for case i
h_j	Leverage for mønster j

Nye verdiar for $\Delta\chi^2_{P(i)}$ og $\Delta\chi^2_{D(i)}$

- Ved ”Compute” kan ein no rekne ut Pearson residualen (formel 7.19 i Hamilton) og delta Pearson kjikvadratet (formel 7.24 i Hamilton) på nytt og vil da finne korrigerte verdiar
- Tilsvarande kan gjerast for avviksresidualen (formel 7.21) og delta avvikskjikvadratet (formel 7.25a)

Leverage og residualar (1)

- Leverage til eit mønster finn vi som talet av case med mønsteret gonger leverage for eit av casa med mønsteret. Leverage for eit case er det same som i OLS regresjonen
- $h_j = m_j * h_i$
- Pearson residualen finn vi som

$$r_j = \frac{Y_j - m_j \hat{P}_j}{\sqrt{m_j \hat{P}_j \left(1 - \hat{P}_j\right)}}$$

Leverage og residualar (2)

- Avviksresidualen finn vi som

$$d_j = \pm \sqrt{\left\{ 2 \left[Y_j \ln\left(\frac{Y_j}{m_j \hat{P}_j} \right) + \left(m_j - Y_j\right) \ln\left(\frac{m_j - Y_j}{m_j \left(1 - \hat{P}_j\right)} \right) \right] \right\}}$$

To Kji-kvadrat observatorar

- Pearson Kji-kvadrat observatoren

$$c_P^2 = \sum_{j=1}^{J} r_j^2$$

- Avviks kjikvadrat observatoren

$$c_D^2 = \sum_{j=1}^{J} d_j^2$$

- Formlane er dei same anten vi reknar case eller mønster

Kjikvadrat observatorane

Begge kjikvadrat observatorane

1. Pearson-kjikvadratet χ^2_P og
2. Avviks-kjikvadratet χ^2_D

- Kan lesast som ein test av nullhypotesa om ingen skilnad mellom den estimerte modellen og ein "metta modell", dvs. ein modell med like mange parametrar som case / mønster

Meir om mål for påverknad

- Mål for påverknad ved endring (Δ) i observator/ parameterverdi pga utelatne case med mønster j
  - ΔB_j "delta B" - analog til Cook's D
  - $\Delta\chi^2_{P(i)}$ "delta Pearsonkjikvadrat"
  - $\Delta\chi^2_{D(i)}$ "delta Avvikskjikvadrat"

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"

ΔB_j "delta B"

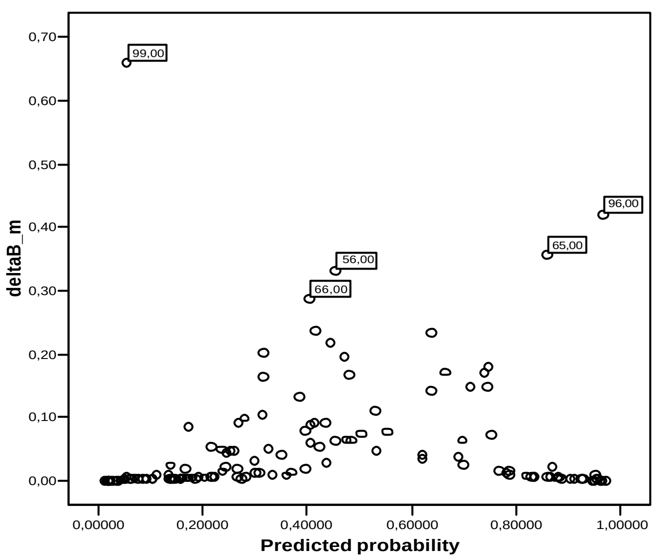
- Måler den standardiserte endringa i dei estimerte parametane (b_k) som oppstår når ein utelet alle case med eit gitt mønster j

$$\Delta B_j = \frac{r_j^2 h_j}{\left(1 - h_j\right)^2}$$

Til større verdi til meir påverknad.

ΔBj >= 1 må i alle fall reknast som "stor påverknad"

delta B (m/caseNO)

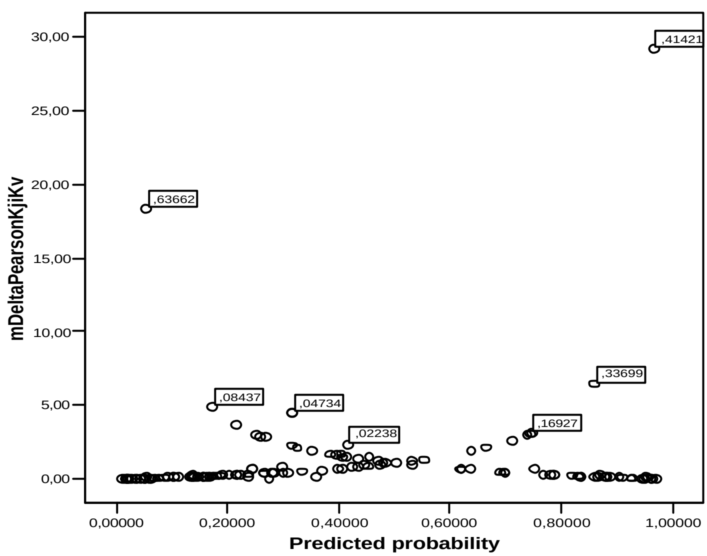


$\Delta\chi^2_{P(i)}$ “delta Pearsonkjikvadrat”

- Måler minken i Pearson χ^2 som følgjer av utelating av alle case med mønster j

$$\Delta c^{2}_{P(j)} = \frac{r_j^2}{\left(1-h_j\right)}$$

delta Pearsonkjikvadrat (m/delta B)



$\Delta\chi^2_{D(i)}$ "delta Avvikskjikvadrat"

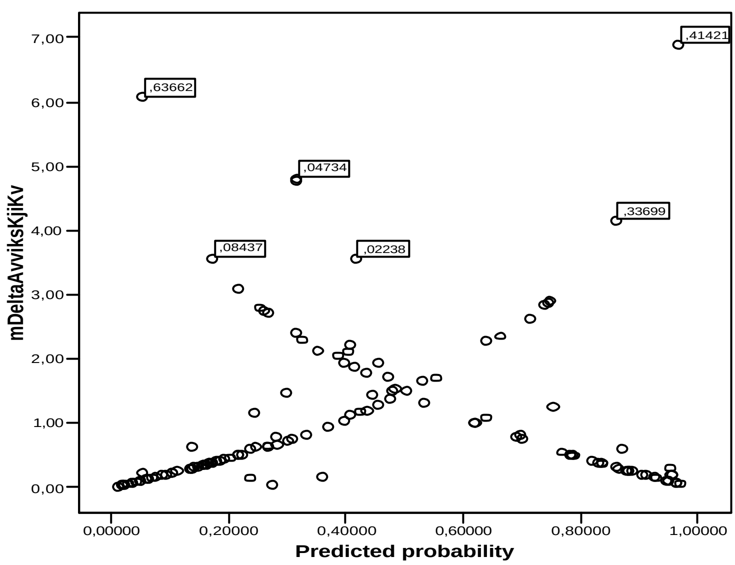
- Måler endring i avvik som følgje av å utelate alle case med mønster j
- Dette er ekvivalent med

$$\Delta c^2_{D(j)} = \frac{d_j^2}{\left(1-h_j\right)}$$

$$\Delta c^2_{D(j)} = -2\left[LL_{\mathrm{K}} - LL_{\mathrm{K(j)}}\right]$$

LL_K er LogLikelihooden for ein modell med K parametrar estimert på heile utvalet og $LL_{K(j)}$ er frå estimatet av same modellen etter at alle case med mønster j er utelatne

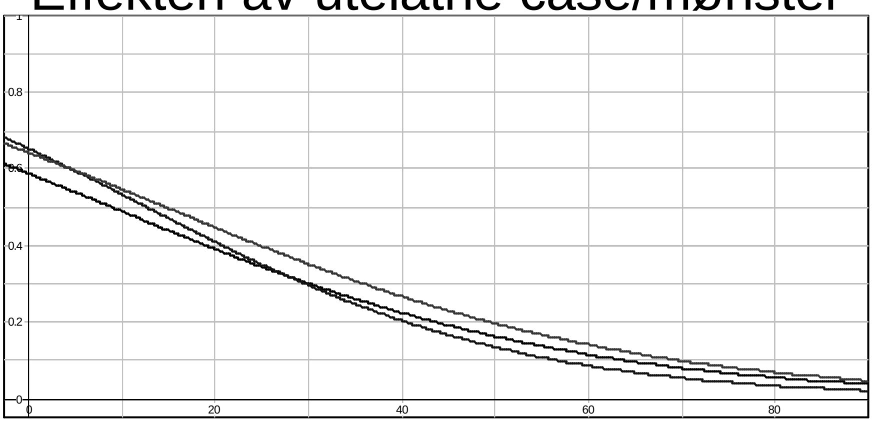
delta Avvikskjikvadrat (m/delta B)



Effekten av utelatne case/mønster

	Logit koeffesient		
Variable i modellen	Heile utvalet	Minus case 99 $\Delta\chi 2P(i)$ =18,34	Minus case 96 $\Delta\chi 2P(i)$ =29,20
lived	-,040	-,045	-,052
educ	-,197	-,224	-,214
contam	1,299	1,490	1,382
hsc	2,279	2,492	2,347
nodad	-1,731	-1,889	-1,658
Constant	2,182	2,575	2,530
2*LL(modell)	**-142,652**	**-135,425**	**-136,124**

Effekten av utelatne case/mønster



y=1/(1+exp(-(2.18-0.04x-0.2×13+1.3×0.28+2.28×0.31-1.73×0.17)))
y=1/(1+exp(-(2.53-0.05x-0.21×13+1.38×0.28+2.35×0.31-1.65×0.17)))
y=1/(1+exp(-(2.58-0.04x-0.22×13+1.49×0.28+2.49×0.31-1.89×0.17)))

Konklusjonar (1)

Vanleg OLS verkar dårleg ved dikotome avhengige variablar sidan vi

- umogeleg kan får normalfordelte feil eller homoskedastisitet, og sidan
- modellen gir sannsyn utanfor intevallet 0-1

Logit-modellen er betre

- Likelihoodrateobservatoren kan teste nesta modellar omlag som F-observatoren
- I store utval vil Wald-observatoren [eller t=SQRT(Wald)] kunne teste einskildkoeffesientar og konfidensintervall kan konstuerast
- Det finst ikkje nokon determinasjonskoeffesient

Konklusjonar (2)

- Koeffesientane i estimerte modellar kan tolkast ved
  1. Log-odds (direkte tolking)
  2. Odds
  3. Oddsrater
  4. Sannsyn (betinga effekt plott)
- Ikkje-linearitet, case med innverknad, og multikollinearitet gir same typen problem som i OLS regresjon (usikre parameterverdiar)
- Diskriminering gir problem av same typen (høge variansestimat, dvs. usikre parameterverdiar)
- Diagnosearbeid er viktig