SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 11

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing XI

- Logistisk regresjon II
  Hamilton Kap 7 s217-235

LOGIT REGRESJON eller LOGISTISK REGRESJON

- **Skal nyttast når avhengig variabel er på nominalnivå**
- Føreset at Y har verdiane 0 eller 1
- Modellen av den betinga forventninga til Y, E[Y | X], nyttar den logistiske funksjonen
- Den lineære sannsynsmodellen (LPM) er substansielt sett feil modell

Den logistiske funksjonen

Den generelle logistiske funksjonen er

- $Y_i = \alpha/(1+\gamma*\exp[-\beta X_i])$

$\alpha>0$ gir den øvre grensa for Y, dvs vi har at $0<Y<\alpha$

γ fastlegg det horisontale punkt for rask vekst

Set ein $\alpha = 1$ og $\gamma = 1$

Vil ein alltid ha

- $0 < 1/(1+\exp[-\beta X_i]) < 1$

Den logistiske funksjonen vil for alle verdiar av x liggje mellom 0 og 1

MODELL (1)

Definisjonar

- Sannsynet for at person i skal ha verdien 1 på variabelen Y skriv vi $Pr(Y_i=1)$. Da er $Pr(Y_i \neq 1) = 1 - Pr(Y_i=1)$
- Oddsen for at person i skal ha verdien 1 på variabelen Y_i, her kalla O_i , er tilhøvet mellom to sannsyn:

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

MODELL (2)

Definisjonar:

- 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:
- $L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$, der j=1,...,K-1, og i=1,...,n

MODELL (3)

- Sett X = (samlinga av alle x_j). Da er sannsynet for at Y_i = 1 for person nr i

$$\Pr(y_i = 1) = E\left[y_i \mid X\right] = \frac{1}{1+\exp\left(-L_i\right)} = \frac{\exp(L_i)}{1+\exp(L_i)}$$

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

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

Estimering ved ML metoden

- Metoden brukt for å estimere parametrane i modellen heiter 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 $\mathcal{L}$
- Kriteriet for å velje regresjonsparametrar er at likelihooden ($\mathcal{L}$) skal vere størst mogeleg

Maximum Likelihood (1)

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

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

Maximum Likelihood (2)

- For lettare å kunne maksimere sannsynet $\mathcal{L}$ tar ein den naturlege logaritmen til $\mathcal{L}$:

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

- Den naturlege logaritmen til $\mathcal{L}$ kallar vi LogLikelihooden, Vi skriv det som $\mathcal{LL}$.
- $\mathcal{LL}$ har ei sentral rolle i logistisk regresjon.

Maximum Likelihood (3)

- $\mathcal{LL}$ er alltid negativ.
- Maksimering av $\mathcal{LL}$ er derfor likeverdig med minimering av den positive LogLikelihooden dvs. minimering av -$\mathcal{LL}$
- Å finne parameterverdiar som minimerer -$\mathcal{LL}$ kan gjerast berre ved ”prøving og feiling”, gjennom ein iterativ prosedyre

Fall 2004 © Erling Berge 2004 11

Iterativ estimering

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

From 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

Utgangspunktet er ein modell med konstantledd

Legg merke til kolonnen med tittel -2 LogLikelihood

Fall 2004 © Erling Berge 2004 12

Tolking (1)

- Skilnaden mellom den lineære modellen og den logistiske er stor i nærleiken av 0 og 1
- LPM er lett å tolke: $Y_i = \beta_0$ når $x_{1i} = 0$, og når x_{1i} veks med ei eining veks Y_i med β_1 einingar
- Logitmodellen er vanskelegare å tolke. Den er ikkje-lineær både i høve til oddsen og sannsynet.

ODDS og ODDSRATER

- Logiten, L_i, ($L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$) 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.

Tolking (2)

- Når alle x er lik 0 er $L_i = \beta_0$ Det tyder at oddsen for at $y_i = 1$ i det høvet er $\exp\{\beta_0\}$
- Dersom ein held alle x-ane fast (set dei lik ein konstant) medan x_1 aukar med 1 vil oddsen for at $y_i = 1$ verte multiplisert med $\exp\{\beta_1\}$ Det tyder at den vil endre seg med $100(\exp\{\beta_1\} - 1)$ %
- Sannsynet $\Pr\{y_i = 1\}$ vil derimot endre seg med ein faktor som er påverka av alle elementa i logiten, alle effektar er interaksjonseffektar

FØRESETNADER

1. Modellen er korrekt spesifisert
   1. logiten er lineær i parametrane
   2. alle relevante variablar er med
   3. ingen irrelevante er med
2. x-variablane er målt utan feil
3. 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

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

LOGISTISK REGRESJON: TESTING (1)

To testar er aktuelle

- (1) Sannsynsratetesten ”Likelihood ratio test”
  - Denne kan nyttast analogt med F-testen
- (2) Wald testen
  - Kvadratrota av denne kan nyttast analogt med t-testen

LOGISTISK REGRESJON: TESTING (2)

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

LOGISTISK REGRESJON: TESTING (3)

Sannsynsrate testobservatoren

- χ^2_H **= -2[$\mathcal{LL}$(modell1) - $\mathcal{LL}$(modell2)]**
  vil, dersom nullhypotesa om ingen skilnad mellom modellane er rett, vere tilnærma (for store n) kjikvadratfordelt med fridomsgrader lik differansen i talet på parametrar i dei to modellane (H)

Eksempel på Sannsynsratetest

- Modell 1: berre konstant
- Modell 2: konstant pluss ein variabel
- χ^2_H **= -2[$\mathcal{LL}$(modell1) - $\mathcal{LL}$(modell2)]**
- Finn verdien av Kjikvadratet og talet på fridomsgrader
- Eks.: LogLikelihood (mod1) = 209,212/(-2)
- LogLikelihood (mod2) = 195,267/(-2)

Frå Tab 7.1: -2 Log likelihood
209,212
195,684
195,269
195,267
195,267

LOGISTISK REGRESJON: TESTING (4)

Waldtesten

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

Utdrag frå Hamilton Tabell 7.2

Iterasjon	**-2 Log likelihood**					
0	209,212					
1	152,534					
2	149,466					
3	149,382					
4	149,382					
5	149,382					
Variables	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Lived	-,046	,015	9,698	1	,002	,955
Educ	-,166	,090	3,404	1	,065	,847
Contam	1,208	,465	6,739	1	,009	3,347
Hsc	2,173	,464	21,919	1	,000	8,784
Constant	1,731	1,302	1,768	1	,184	5,649

Konfidensintervall for parameterestimat

- Konfidensintervall for parameterestimat kan konstruerast ut frå at kvadratrota av Wald-observatoren med 1 fridomsgrad er tilnærma normalfordelt (sjå bilde 9)
- $b_k - t_\alpha * SE(b_k) < \beta_k < b_k + t_\alpha * SE(b_k)$ der t_α er tabellverdien teken frå **normalfordelinga** med signifikansnivå α

Konfidensintervall basert på t-fordelinga (1)

- I mangel av tabellar over normalfordeling kan ein gjere seg nytte av at **t-fordelinga** er tilnærma lik normalfordelinga ved store n-K (t.d. n-K > 120)

Utdrag frå Hamilton Tabell 7.3

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	lived	-,047	,017	7,550	1	,006	,954
	educ	-,206	,093	4,887	1	,027	,814
	contam	1,282	,481	7,094	1	,008	3,604
	hsc	2,418	,510	22,508	1	,000	11,223
	female	-,052	,557	,009	1	,926	,950
	kids	-,671	,566	1,406	1	,236	,511
	nodad	-2,226	,999	4,964	1	,026	,108
	Constant	2,894	1,603	3,259	1	,071	18,060

Meir om Hamilton Tabell 7.3

Iteration		-2 Log likelihood	Coefficients							
			Const	lived	educ	contam	hsc	female	kids	nodad
Step0		209,212	-0,276							
Step1	1	147,028	1,565	-,027	-,130	,782	1,764	-,015	-,365	-1,074
	2	141,482	2,538	-,041	-,187	1,147	2,239	-,037	-,580	-1,844
	3	141,054	2,859	-,046	-,204	1,269	2,401	-,050	-,662	-2,184
	4	141,049	2,893	-,047	-,206	1,282	2,418	-,052	-,671	-2,225
	5	141,049	2,894	-,047	-,206	1,282	2,418	-,052	-,671	-2,226

Er modellen i tabell 7.3 betre enn modellen i tabell 7.2 ?

- $\mathcal{LL}$**(modell i 7.3) = 141,049/(-2)**
- $\mathcal{LL}$**(modell i 7.2) = 149,382/(-2)**

- χ^2_H = -2[$\mathcal{LL}$(modell i 7.2) - $\mathcal{LL}$(modell i 7.3)]
- Finn χ^2_H verdien
- Finn H
- Slå opp i tabellen over kjikvadratfordelinga

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

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

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 tyder

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 ?

Oddsraten: eksempel I

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

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

Hugs rekneregleane for potensar

Oddsraten: eksempel II

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

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

Hugs reknereglane for potensar

Fall 2004 © Erling Berge 2004 35

Eksempel frå Hamilton tabell 7.2

- Kva er oddsraten for å gå inn for å stengje skolen ved eitt års auke i skolegangen?
- Oddsraten er kvotienten mellom to odds der den eine er oddsen for den som har eitt år meir utdanning

$$\frac{e^{b_0+b_1*ÅrBuddIByen+b_2*(Utdanning+1)+b_3*UreiningEigEigedom+b_4*MangeHSCmøter}}{e^{b_0+b_1*ÅrBuddIByen+b_2*Utdanning+b_3*UreiningEigEigedom+b_4*MangeHSCmøter}} = e^{b_2}$$

Oddsraten = Exp{b_2} = exp(-0,166) = 0,847

Eitt ekstra år utdanning fører til at oddsen vert redusert med ein faktor 0,847

Ein kan og seie at oddsen ”aukar” med 100(0,847-1)% = -15,3% (dvs. minkar med 15,3%)

Fall 2004 © Erling Berge 2004 36

BETINGA EFFEKT PLOTT

- Gi faste verdiar 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[-konst - $b_k x_k$])
  for rimelege verdiar av x_k
  "konst" er konstanten ein får ved innsetting i logiten av dei valde faste variabelverdiane

Utdrag frå Hamilton Tabell 7.4

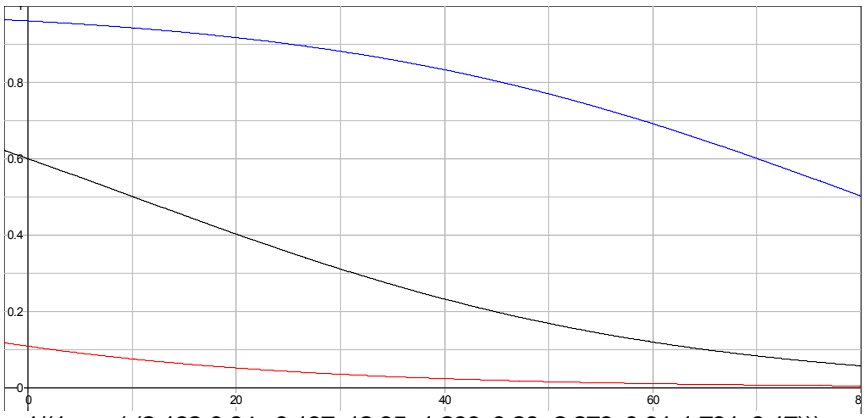
	B	S.E.	Wald	df	Sig.	Exp(B)	Minimum	Maximum	Mean
lived	-,040	,015	6,559	1	,010	,961	1,00	81,00	19,2680
educ	-,197	,093	4,509	1	,034	,821	6,00	20,00	12,9542
contam	1,299	,477	7,423	1	,006	3,664	,00	1,00	,2810
hsc	2,279	,490	21,591	1	,000	9,763	,00	1,00	,3072
nodad	-1,731	,725	5,696	1	,017	,177	,00	1,00	,1699
Constant	2,182	1,330	2,692	1	,101	8,866			

Logiten:

L = 2.182 -0.04*lived -0.197*educ +1.299*contam +2.279*hsc -1.731*nodad

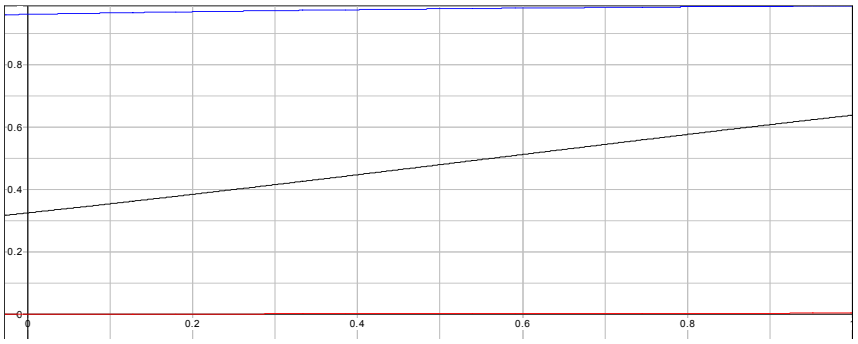
Her lar vi "lived" variere og set inn høveleg valde verdiar for dei andre

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-0.197×12.95+1.299×0.28+2.279×0.31-1.731×0.17)))
y=1/(1+exp(-(2.182-0.04x-0.197×12.95+1.299×1+2.279×1-1.731×0)))
y=1/(1+exp(-(2.182-0.04x-0.197×12.95+1.299×0+2.279×0-1.731×1)))

Betinga effekt plott frå Hamilton tabell 7.4 (fig7.6) effekten av ureining på eigen eigedom



y=1/(1+exp(-(2.182-0.04×19.27-0.197×12.95+1.299x+2.279×0.31-1.731×0.17)))
y=1/(1+exp(-(2.182-0.04×1-0.197×6+1.299x+2.279×1-1.731×0)))
y=1/(1+exp(-(2.182-0.04×81-0.197×20+1.299x+2.279×0-1.731×1)))

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

Ulike pseudo R^2 i SPSS

- SPSS rapporterer Cox og Snell, Nagelkerke, og i multinomisk logistisk regresjon også McFadden sine framlegg til R^2
- Aldrich og Nelson sitt kan vi rekne ut sjølv

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	***	***	***

Pseudo R-Square	
Cox and Snell	***
Nagelkerke	***
McFadden	***