SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing X

- Logistisk regresjon II
  Hamilton Kap 7 s223-233

LOGISTISK REGRESJON ESTIMERING

- ML (Maximum likelihood) metoden finn dei parametrane i Logit likninga som maksimerer den naturlege logaritmen til Likelihooden,

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

- **LogLikelihooden = $\mathcal{LL}$ = ln($\mathcal{L}$) = $\log_e(\mathcal{L})$ =**

  $\Sigma_i \{Y_i \log_e P_i + (1-Y_i)\log_e(1-P_i)\}$, i= 1, 2, 3, ... , 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

Utgangspunktet er ein modell med konstantledd

LOGISTISK REGRESJON: TESTING (1)

To testar er aktuelle

- (1) Sannsynsratetesten

  ”Likelihood ratio test” Denne kan nyttast analogt med F-testen

- (2) Wald testen

LOGISTISK REGRESJON: TESTING (2)

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

LOGISTISK REGRESJON: TESTING (3)

Sannsynsrate test-observatoren

- 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) kji-kvadratfordelt 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 Kji-kvadratet 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)

Wald-testen

- 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 Kji-kvadratfordelt 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

LOGISTISK REGRESJON
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 normal-fordelinga 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 koeffesientane β tyder

TOLKING: 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 $x_{.j.}$

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 ?

LOGISTISK REGRESJON
Oddsraten: eksempel

- 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 reknereglane for potensar

LOGISTISK REGRESJON
Oddsraten: eksempel

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

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

Hugs reknereglane for potensar

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

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

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%)

LOGISTISK REGRESJON 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

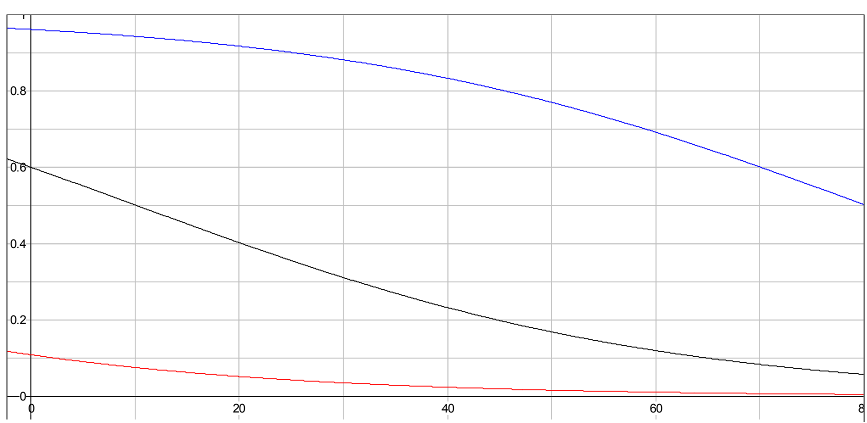
	B	S.E.	Wald	df	Sig.	Exp(B)	Minimum	Maximum	Mean
lived	-,040	,015	6,559	1	,010	,961	1,00	81,00	19,2680
educ	-,197	,093	4,509	1	,034	,821	6,00	20,00	12,9542
contam	1,299	,477	7,423	1	,006	3,664	,00	1,00	,2810
hsc	2,279	,490	21,591	1	,000	9,763	,00	1,00	,3072
nodad	-1,731	,725	5,696	1	,017	,177	,00	1,00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

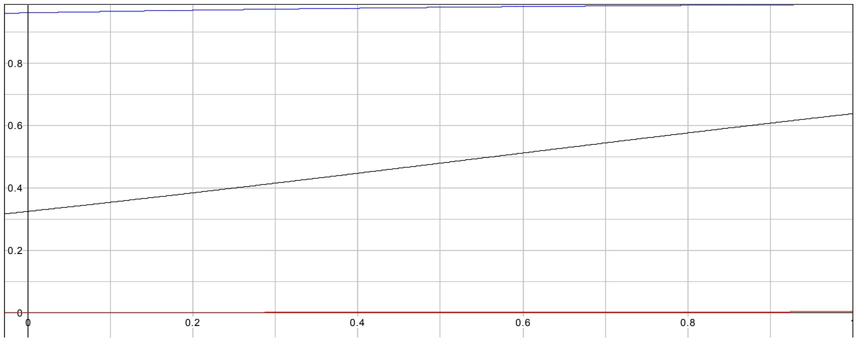


$y=1/(1+\exp(-(2.182-0.04x-0.197\times 12.95+1.299\times 0.28+2.279\times 0.31-1.731\times 0.17)))$

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

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

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)))

Determinasjonskoeffesientar

- I logistiske regresjonsmodellar finst ikkje mål tilsvarande determinasjons-koeffesienten 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	***