SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 07

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing VII

- Logistisk regresjon I
  Hamilton Kap 7 s217-234

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
- Men

  Kvifor kan ikkje E[Y | X] vere ein lineær funksjon også her?

Fall 2004 © Erling Berge 2004 3

Den lineære sannsynsmodellen: LPM

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

Fall 2004 © Erling Berge 2004 4

Er føresetnadene rette i LPM?

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

Fall 2004 © Erling Berge 2004 5

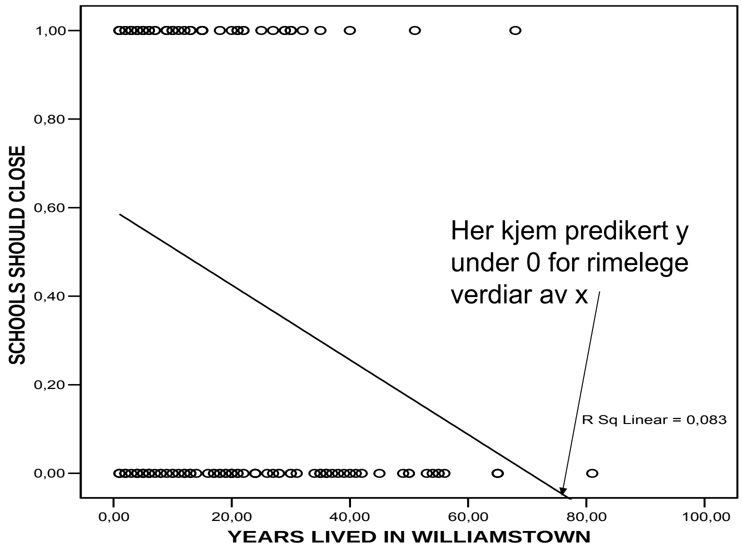
OLS regresjon av dikotom avhengig variabel på variabelen ”år budd i byen”

ANOVA tabell	Sum of Squares	df	Mean Square	F	Sig.
Regression	3,111	1	3,111	13,648	,000(a)
Residual	34,418	151	,228		
Total	37,529	152			

Dependent Variable: SCHOOLS SHOULD CLOSE	B	Std. Error	t	Sig.
(Constant)	,594	,059	10,147	,000
YEARS LIVED IN TOWN	-,008	,002	-3,694	,000

Regresjonen ser heilt OK ut i desse tabellane.

Fall 2004 © Erling Berge 2004 6



Spreiingsplott med regresjonslinje. Figur 7.1 Hamilton

LPM er feil modell

- Vi ser i eksempelet her at ein for rimelege verdiar av x-ane kan får ein verdi av predikert y der
  $E[Y_i \mid \mathbf{X}] > 1$ eller $E[Y_i \mid \mathbf{X}] < 0$,
- Dette kan ein ikkje gjere noko med
- <u>LPM er substansielt sett feil modell</u>
- Det trengst ein modell der ein alltid har
  $0 < E[Y_i \mid \mathbf{X}] < 1$

Den logistiske funksjonen

Den generelle logistiske funksjonen er

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

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

γ fastlegg det horisontale punkt for rask vekst

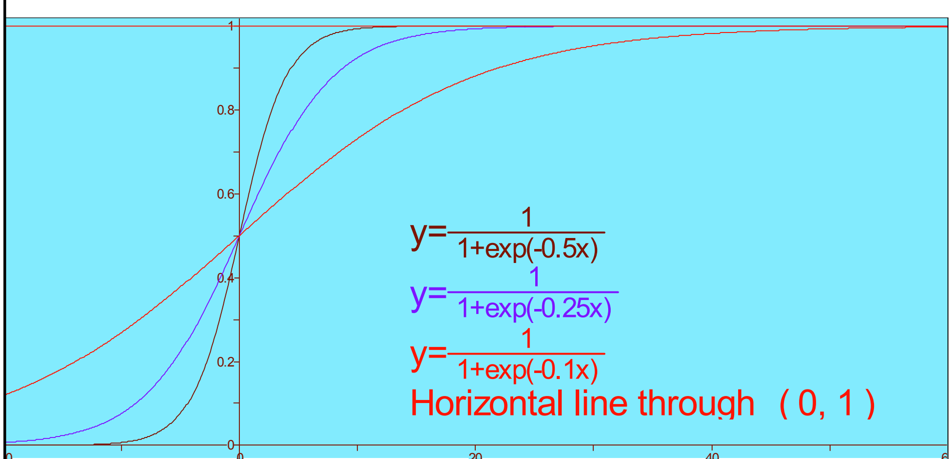
Set ein $\alpha = 1$ og $\gamma = 1$

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

Logistiske kurver for ulik β



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:

$$\mathrm{O}_i\left(y_i=1\right)=\frac{\Pr\left(y_i=1\right)}{1-\Pr\left(y_i=1\right)}=\frac{p_i}{1-p_i}$$

Fall 2004 © Erling Berge 2004 11

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

Fall 2004 © Erling Berge 2004 12

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

Fall 2004 © Erling Berge 2004 13

MODELL (4)

I modellen $Y_i = E[Y_i \mid \mathbf{X}] + \varepsilon_i$ er feilen enten

- $\varepsilon_i = 1 - E[Y_i \mid \mathbf{X}]$ med sannsyn $E[Y_i \mid \mathbf{X}]$

  (sidan $\Pr(Y_i = 1) = E[Y_i \mid \mathbf{X}]$),

eller feilen er

- $\varepsilon_i = - E[Y_i \mid \mathbf{X}]$ med sannsyn $1 - E[Y_i \mid \mathbf{X}]$

- **mao** feilen har ei fordeling kjent som binomialfordelinga med $p_i = E[Y_i \mid \mathbf{X}]$

Fall 2004 © Erling Berge 2004 14

Estimering

- 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 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

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

Maximum Likelihood (2)

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

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

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

Fall 2004 © Erling Berge 2004 17

Logistisk modell i staden for LPM

Iteration	-2 Log Likelihood	Coefficients	
		Constant	Lived in town
Step 0	209,212	-,275	0
1	195,684	,376	-,034
2	195,269	,455	-,041
3	195,267	,460	-,041
4	195,267	,460	-,041

Dependent: Schools should close	B	S.E.	Wald	df	Sig.	Exp(B)
Lived in town	-,041	,012	11,399	1	,001	,960
Constant	,460	,263	3,069	1	,080	1,584

Fall 2004 © Erling Berge 2004 18

Fotnotar til tabellen

- Step 0: Utgangspunktet er ein modell med konstantledd og ingen variablar
- **Iterativ estimering**
  - Estimeringa vart avslutta ved iterasjon nr 4 sidan parameterestimata endra seg med mindre enn 0,001
- Observatoren Wald som SPSS gir oss er lik kvadratet av den “t” som Hamilton (og STATA) gir.

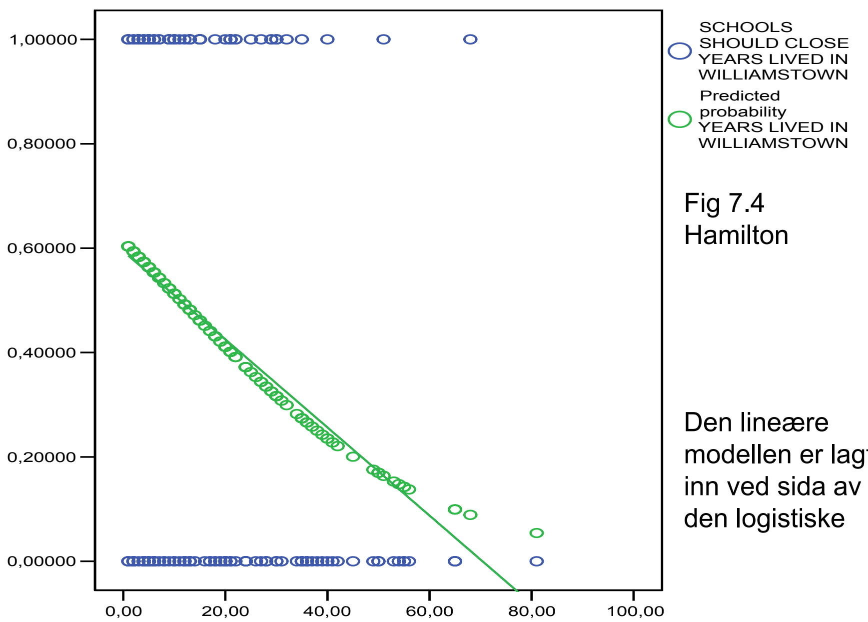


Fig 7.4 Hamilton

Den lineære modellen er lagt inn ved sida av den logistiske

TESTING

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

Tolkning (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 einigar
- 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 $x_{.j}$.

Fall 2004 © Erling Berge 2004 23

Tolkning (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 endre seg med ein faktor som er påverka av alle elementa i logiten

Fall 2004 © Erling Berge 2004 24

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

Fall 2004 © Erling Berge 2004 29

Logistisk regresjon

- Dersom føresetnadene er korrekte vil logistisk regresjon gi oss normalfordelte, forventningsrette og variansminimale estimat av parametrane

Fall 2004 © Erling Berge 2004 30