Statistikgruppen
Handelshøjskolen i København

MPAS
Tue Tjur
November 2001

Notat 2:

Betinget logistisk regression.

Udgangspunktet er en logistisk regressionsmodel for binære data, som vi kan skrive

$$P(y_{gi} = 1) = p_{gi},\ g = 1, \ldots, G,\ i = 1, \ldots, n_g$$

hvor

$$p_{gi} = \frac{\exp(\alpha_g + \beta x_{gi} + \ldots)}{1 + \exp(\alpha_g + \beta x_{gi} + \ldots)} \text{ eller } \operatorname{logit}(p_{gi}) = \alpha_g + \beta x_{gi} + \ldots .$$

Her symboliserer "$\beta x_i + \ldots$" en modelformel, der i princippet kan være hvad som helst der kan stå på højre side af lighedstegnet i (f.eks.) en lineær normalfordelingsmodel.

Vi har her valgt en dobbelt indicering af observationerne y_{gi} med et gruppenummer g og et løbende nummer i inden for gruppen, for at markere den særlige rolle som faktoren "gruppe" har i det følgende. Bemærk at modellen involverer en hovedvirkning af denne faktor. I praksis må man forestille sig, at det drejer sig om en faktor med mange niveauer, som kun er medtaget af nød, idet det man primært er interesseret i ikke er forskelle mellem grupper, men de øvrige parametre β, ...

For at slippe af med parametrene α_g kan man betinge med summerne $y_{g.}$ inden for grupper. Denne betingning vil ikke fjerne væsentlig information, idet gruppesummerne først og fremmest indeholder information om parametrene α_g.

SÆTNING. *I den betingede fordeling af y_{gi}'erne, givet summerne $y_{g.}$, bliver likelihoodfunktionen*

$$\prod_{g=1}^{G} \left(\frac{\exp\left(\sum_{i:y_i=1}(\beta x_{gi} + \ldots)\right)}{\sum_M \exp\left(\sum_{i \in M}(\beta x_{gi} + \ldots)\right)} \right)$$

hvor summen i nævneren skal tages over de delmængder M af $\{1, \ldots, n_g\}$ der netop har $y_{g.}$ elementer.

Beviset er forholdsvis elementært, men meget besværligt at skrive op. Man ser, at når man danner den betingede sandsynlighed for at få en bestemt konfiguration af ettaller inden for en gruppe, givet at der netop

skal være $y_{g\cdot}$ af dem, så er der en forfærdelig masse ting der forkorter ud. Først og fremmest forkorter alle nævnerne i udtrykkene for de enkelte sandsynligheder og deres komplementære ud. Derefter forkorter parametrene α_g ud, fordi de forekommer netop $y_{g\cdot}$ gange i tælleren (under exp–tegnet) og også netop $y_{g\cdot}$ gange i hvert af leddene i nævneren. Beviset overlades i øvrigt til læseren.

I ISUW kan kommandoen FITCLOGIT håndtere disse modeller, se ISUW–eksemplet CLOGIT.ISU til opgave 2, MPAS efteråret 2001. Jeg har ikke kendskab til andre programmer som kan håndtere disse modeller. Hvis man bruger SAS må man nøjes med at fitte den ubetingede model.