Statistikgruppen
Handelshøjskolen i København

MPAS
Tue Tjur
November 2001

Notat 1:

Logistisk regression med overspredning.

Approksimation af en logistisk regressionsmodel med en vægtet lineær normal model.

SÆTNING. *Lad X være binomialfordelt med antalsparameter n og sandsynlighedsparameter p. Hvis n er stor og p ikke er for tæt på 0 eller 1 gælder at*

$$Y = \operatorname{logit}\left(\frac{X}{n}\right) = \log\left(\frac{\frac{X}{n}}{1-\frac{X}{n}}\right) = \log\left(\frac{X}{n-X}\right)$$

er approksimativt normalfordelt med

$$\mathrm{E}Y = \operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) \qquad \textit{og} \qquad \operatorname{var}(Y) = \frac{1}{np(1-p)}\ .$$

BEVIS. Ved rækkeudvikling til første orden af funktionen logit omkring p fås

$$\operatorname{logit}\left(\frac{X}{n}\right) \approx \operatorname{logit}(p) + \operatorname{logit}'(p)\left(\frac{X}{n} - p\right)$$

hvor

$$\operatorname{logit}'(p) = \frac{d}{dp}\left(\log p - \log(1-p)\right) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

Altså

$$\operatorname{logit}\left(\frac{X}{n}\right) \approx \operatorname{logit}(p) + \frac{1}{p(1-p)}\left(\frac{X}{n} - p\right).$$

På grund af den sædvanlige approksimation af binomialfordelingen med en normalfordeling kan vi skrive

$$X \approx np + \sqrt{np(1-p)}U,$$

hvor U er normeret normalfordelt. Indsættes dette får vi

$$\operatorname{logit}\left(\frac{X}{n}\right) \approx \operatorname{logit}(p) + \frac{1}{p(1-p)}\left(\frac{np + \sqrt{np(1-p)}U}{n} - p\right)$$

$$= \text{logit}(p) + \frac{U}{\sqrt{np(1-p)}}.$$

Heraf følger påstanden umiddelbart.

Dette resultat kan bruges til approksimation af en logistisk regressionsmodel med en vægtet lineær regressionsmodel. Som vægte bruges så de inverse varianser $n_i p_i(1-p_i)$. Men eftersom man ikke kender p_i'erne er man tvunget til enten at bruge de vægte man får ved at indsætte de relative hyppigheder i stedet for (det virker bedst hvis n_i'erne er store), eller dem man får ved blot at antage p_i uafhængig af i (svarende til vægte n_i, da vi jo regner på nær proportionalitet). Som responser bruges de logit–transformerede relative hyppigheder. Hvis der er mange nuller og ettaller blandt disse må man modificere definitionen af disse ved f.eks. at addere 0.5 til antal og komplementære antal, ellers får man responser $\pm\infty$ i så mange tilfælde at disse ikke med rimelighed kan ekskluderes fra analysen.

Fordelen ved at gøre det på den måde er, at man herved får mulighed for at udvide modellen med en skalaparameter, der kan tage højde for en eventuel overspredning. Modellen med ukendt skalaparameter (varians) kan opfattes som en approksimation til den model der behandles i det følgende.

Overspredningsmodellen.

Udgangspunktet er en logistisk regressionsmodel, som vi kan skrive

$$y_i \approx \text{bin}(n_i, p_i)$$

hvor

$$p_i = \frac{\exp(\alpha_g + \beta x_i + \dots)}{1 + \exp(\alpha_g + \beta x_i + \dots)} \text{ eller } \text{logit}(p_i) = \alpha_g + \beta x_i + \dots.$$

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

Imidlertid viser testet for "goodness of fit" (test mod den fulde model, enten Pearson eller –2logQ) at modellen ikke kan godkendes. Det betyder at der er "overspredning" — y'ernes varianser er større end de skulle være i henhold til modellen. I så fald kan man ikke direkte bruge modellen. Hvis man ignorerer overspredningen og bare går videre vil man typisk finde alt for mange alt for signifikante effekter, også effekter af faktorer eller regressionsvariable som er helt irrelevante for problemstillingen.

Derfor opstilles en ny model, som går ud på at observationerne stammer fra en anden fordeling med større varians end binomialfordelingen. Middelværdien antages at være den samme, og for variansens vedkommende

antages (som den simplest mulige generalisering), at den er proportional med binomialvariansen. Man kan yderligere lægge det krav på at der (approksimativt) skal være tale om normalfordelinger. I så fald fås modellen

$$y_i \approx N\left(n_i p_i, \lambda n_i p_i(1-p_i)\right),$$

hvor p_i'erne har samme form som ovenfor. $\sqrt{\lambda}$ kaldes i denne forbindelse for *overspredningsparameteren*.

Man kan argumentere for, at middelværdiparametrene i denne model skal estimeres præcis som i den logistiske regressionsmodel, idet det approksimativt er hvad man får ved maximum likelihood, endda eksakt det samme efter en mindre modifikation af ML–metoden. Et naturligt estimat for λ er Pearsons χ^2–størrelse for "goodnes–of–fit" divideret med sit frihedsgradsantal. Man kan her bruge deviansen eller –2logQ i stedet for Pearsons χ^2–størrelse, de er jo approksimativt ens.

Overspredningsmodellen udmærker sig ved at være meget nem at håndtere. Analysen af en overspredningsmodel kan foretages på basis af resultaterne fra den tilsvarende logistiske regressionsmodel, når blot man husker

(1) at senere –2logQ–størrelser eller tilsvarende Pearson–størrelser ved tests for ydeligere modelreduktioner skal divideres med $\hat{\lambda}$, før de vurderes i deres (sædvanlige) χ^2–fordeling. Man kan argumentere for at bruge en F–fordeling i stedet for, i analogi med hvad der kendes fra almindelige lineære normalfordelingsmodeller. Forskellen er uvæsentlig hvis der er få parametre og mange observationer.

(2) standardafvigelser for parameterestimater skal korrigeres tilsvarende, ved multiplikation med $\sqrt{\hat{\lambda}}$.

Modelkontrol foretages primært ved at man tegner et residualplot, hvor normerede residualer $\frac{y_i - n_i\hat{p}_i}{\sqrt{\hat{\lambda} n_i p_i(1-p_i)}}$ plottes mod fittede værdier $n\hat{p}_i$. Hvis store residualer forekommer væsentligt oftere i den ene side af tegningen end i den anden er det tegn på at overspredningsmodellens antagelser om variansen afhængighed af middelværdien er for simpel. Metoder til afhjælpning af sådanne problemer vil vi dog ikke omtale her.

Programmer til håndtering af overspredningsmodeller:

I SAS bruges PROC GENMOD, se eksempel fra MPAS efteråret 2000.

I ISUW bruges kommandoen FITNONLINEAR, se ISUW–eksemplet OVERSPR.ISU til opgave 2, MPAS efteråret 2001.