Institut for Matematisk Erhvervsøkonomi og Statistik
Handelshøjskolen i København

MPAS
Tue Tjur
September 2000

**Notat 4:**

## OVERSPREDNINGSMODELLEN I PRAKSIS

**Generelt.** Ved forelæsningerne har jeg ret udførligt forklaret hvad en overspredningsmodel går ud på i forbindelse med logistisk regression. Det vil derfor blive gjort kort her.

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 linær 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.

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 er 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 simpel modifikation af ML–metoden. Et naturligt estimat for $\lambda$ er Pearsons $\chi^2$–størrelse for "goodnes–of–fit" divideret med sit frihedsgradsantal. Man kan her bruge deviansen eller –2logQ i stedet for Pearsons $\chi^2$–størrelse, de er jo approksimativt ens.

**Operationelt** er overspredningsmodellen 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) $\chi^2$–fordeling (man kan argumentere for at bruge en F–fordeling i stedet for, i analogi med hvad der kendes fra almindelige lineære normalfordelingsmodeller, men det er uvæsentligt hvis antal frihedsgrader for det oprindelige goodnes–of–fit test er stort).

(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 $\frac{y_i - n_i\hat{p}_i}{\hat{\lambda}\sqrt{n_i p_i(1-p_i)}}$ 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.