Center for Statistik
Handelshøjskolen i København

MPAS
Tue Tjur
September 2004

Notat 1:

"POISSONMODELLEN MED OVERSPREDNING"

Udgangspunktet er en sædvanlig multiplikativ Poissonmodel, som vi kan skrive

$$y_i \sim \text{Poiss}(\lambda_i)$$

hvor

$$\lambda_i = \exp(\alpha_g + \beta x_i + \dots) \text{ eller } \log(\lambda_i) = \alpha_g + \beta x_i + \dots .$$

Her symboliserer "$\alpha_g + \beta x_i + \dots$" en modelformel, der i princippet kan være hvad som helst der kunne stå på højre side af lighedstegnet i en linær normalfordelingsmodel. I projekt 1 kunne det for eksempel være en modelformel der specificerer hovedvirkninger af og 2–faktor vekselvirkninger mellem køn, alder og amt.

Imidlertid viser "goodness–of–fit–testet" (test mod den fulde model, 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 komme videre. Hvis man ignorerer overspredningen og fjerner de vekselvirkninger man ikke bryder sig om får man "falske signifikanser" — de estimerede standardafvigelser på f.eks. hovedvirkningskontraster bliver for små, da de er baserede på Poissonvariansen og ikke tager højde for overspredningen.

Derfor opstilles en ny model, som går ud på at observationerne stammer fra en anden fordeling med større varians end Poissonfordelingen. Middelværdien antages at være den samme, og for variansens vedkommende antages (som den simplest mulige generalisering), at den er proportional med Poissonvariansen. Man kan yderligere lægge det krav på at der er tale om normalfordelinger, i så fald fås modellen

$$y_i \sim N\left(\lambda_i, \varphi\lambda_i\right),$$

hvor λ_i'erne har samme form som ovenfor. $\sqrt{\varphi}$ kaldes i denne forbindelse for *overspredningsparameteren*. Man kan argumentere for, at middelværdiparametrene i denne model skal estimeres præcis som i den multiplikative Poissonmodel, idet det approksimativt er hvad man får ved maximum likelihood, endda eksakt det samme efter en lille modifikation af ML–metoden (iteratively reweighted least squares). Et naturligt estimat for φ 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.

Operationelt er overspredningsmodellen meget nem at håndtere. Analysen af en overspredningsmodel kan foretages på basis af resultaterne fra den tilsvarende multiplikative Poissonmodel, når blot man husker

(1) at senere –2logQ–størrelser eller tilsvarende Pearson–størrelser ved tests for yderligere modelreduktioner skal divideres med $\hat{\varphi}$, 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, 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 $\sqrt{\hat{\varphi}}$.

Det er dog ikke nødvendigt at foretage disse modifikationer i hånden. I ISU kan man foretage beregningerne ved hjælp af kommandoen `FITNONLINEAR`, i SAS bruges `PROC GENMOD` med passende options til `MODEL` specifikationen (vist nok `DIST=POISSON` og `SCALE`).

Modelkontrol foretages primært ved at man tegner et residualplot, hvor normerede residualer $\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\varphi}\lambda_i}}$ plottes mod fittede værdier $\hat{\lambda}_i$. Hvis store residualer forekommer væsentligt oftere i den ene side af tegningen end i den anden er det tegn på at overspredningsmodellens antagelser om variansens afhængighed af middelværdien er forkerte. I så fald kan man (i ISU, og vist nok også i SAS) specificere en anden variansfunktion.

I nogle tilfælde kan man som en simpel approksimation bruge den lineære normalfordelingsmodel

$$\log(y_i) \sim N\left(\alpha_g + \beta x_i + \ldots, \sigma^2\right) .$$

Det er klart at denne model approksimativt har samme middelværdistruktur som overspredningsmodellen. Men variansen (som antages konstant) er naturligvis anderledes. Denne approksimation virker bedst hvis y_i'erne er store og ikke alt for forskellige. Specielt kan den selvfølgelig ikke bruges hvis værdien $y_i = 0$ forekommer.