

Notat 3:

Modeller i forbindelse med projekt 2.

Som responsvariabel tages naturligt

$$a_i = \text{andel}_i = \frac{\text{sumindf}_i}{\text{samobl}_i}$$

Følgende overvejelser fører til en model, der beskriver a_i 'erne som normalfordelte med middelværdier der afhænger logit-lineært af de baggrundsvariable vi har (i særdeleshed en eller anden version af "gevinsten"), med en varians af samme form som variansen i en binomialfordelingsmodel, korrigeret med en overspredningsparameter.

Lad s betegne den gennemsnitlige lånstørrelse. For en meget grov betragtning har vi så

$$a_i = \frac{sy_i}{sn_i} = \frac{y_i}{n_i}$$

hvor n_i betegner antal lån i den pågældende serie det pågældende år og y_i er antal indfrielse. Her forekommer det naturligt at antage y_i binomialfordelt med antalsparameter n_i og sandsynlighedsparameter

$$p_i = \frac{\exp(\alpha + \beta g_i + \dots)}{1 + \exp(\alpha + \beta g_i + \dots)}$$

som afhænger logit-lineært af baggrundsvariable som gevinsten g_i osv. Vi får så

$$\mu_i = E a_i = p_i$$

og

$$\text{var}(a_i) = \frac{p_i(1-p_i)}{n_i} = \frac{\mu_i(1-\mu_i)}{\text{samobl}_i/s} = \lambda \times \frac{\mu_i(1-\mu_i)}{\text{samobl}_i}.$$

Her gælder umiddelbart $\lambda = s$; men på grund af den ret grove approksimation der ligger bag modellen (i særdeleshed antagelsen om at alle lån er af ca. samme størrelse s) forekommer det mere rimeligt at udnævne λ til en "overspredningsparameter", svarende til variansen i en sædvanlig lineær normalfordelingsmodel, hvis fortolkning vi ikke kan sige så meget om, og som vi måske heller ikke er så interesserede i. I dette tilfælde kan man sige, at vi har taget højde for hvordan variansen naturligt afhænger af det samlede antal n_i af lån og sandsynlighederne p_i for at et givet lån konverteres. Men den variation der givetvis er fra serie til serie af den

gennemsnitlige lånstørrelse, og også selve variationen af lånenes størrelse indenfor en serie, er her udnævnt til at være “tilfældig variation”. Derfor forventer vi en varians der er større end binomialvariansen. Men vi antager altså — som det bedste bud vi kan give — at den følger binomialvariansen proportionalt.

En ikke-lineær normalfordelingsmodel med denne middelværdi- og variansstruktur kommer meget tæt på “Svend Jakobsens model”, som nævnt i Mette og Anne’s bachelorprojekt. Men da vi ikke har beskæftiget os med den vil forklaring ikke blive givet her.

Fordelen ved modellen er, at den tager højde for at variansen på a_i er større i en lille serie end i en stor serie (omvendt proportional med samobl). Desuden tillader modellen at man tager de mange værdier $a_i = 0$ ind i modellen uden at give dem “særbehandling”. En linearisering af denne model (altså en lineær model for $\log \frac{a_i}{1-a_i}$, som vi har talt om tidligere) fører jo til et problem med håndteringen af værdien $a_i = 0$. Ganske det samme problem som man har i en almindelig logistisk regression, hvis man forsøger at linearisere den ved at logit-transformere de relative hyppigheder.

En ikke-lineær model af ovennævnte type kan i ISUW fittes med en kommando af formen

```
FITNONLINEAR andel=1+gev+.../samobl \
exp.)/(1+exp(.)) exp.)/sqr(1+exp(.))  .*(1-.)
```

For at forstå syntaxen er man nødt til at læse beskrivelsen af kommandoen FITNONLINEAR. I SAS kan man bruge PROC GENMOD med (så vidt jeg husker) en option WEIGHT SAMOBL til specifikation af at der ønskes “vægtning” med SAMOBL.

Angående baggrunden for betegnelsen “vægtning” i denne forbindelse:

Ved en vægtet regression forstår man i almindelighed en regressionsmodel hvor den sædvanlige antagelse om konstant varians er erstattet med en antagelse om, at varianserne følger værdierne af en given variabel proportionalt. Altså (for eksempel)

$$y_i \approx N\left(\alpha + \beta x_i, \frac{\lambda}{w_i}\right)$$

hvor λ er den ukendte proportionalitetsfaktor (som overtager variansens rolle) og værdierne af den givne variable er $1/w_i$. Dens inverse værdier, w_i ’erne, kaldes i denne sammenhæng for *vægte*. Grunden til dette er, at den sædvanlige mindste kvadraters metode der går ud på at minimere (i dette tilfælde)

$$\sum (y_i - (\alpha + \beta x_i))^2$$

for at finde maksimum-likelihood estimatorerne, kan bruges i dette mere generelle tilfælde på modificeret form. Modifikationen går ud på at man

i stedet skal minimere den *vægtede* kvadratsum

$$\sum w_i (y_i - (\alpha + \beta x_i))^2.$$