

Notat 1:

Vægtede regressionsmodeller.

Lad y være n -dimensionalt normalfordelt med middelværdi $\mu = X\beta$ givet på sædvanlig lineær form, og med kovariansmatrix af formen λW^{-1} , hvor W er en diagonalmatrix med kendte diagonalelementer $w_1, \dots, w_n$.
Altså

$$y \sim N_n(X\beta, \lambda W^{-1}).$$

Likelihooden for denne model bliver åbenbart

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi \frac{\lambda}{w_i}}} \exp\left(-\frac{w_i}{2\lambda}(y_i - \mu_i)^2\right).$$

Heraf følger, at maksimering af likelihooden m.h.t. β for fast λ er ækvi-valent med minimering af kvadratsummen

$$\sum_{i=1}^n w_i (y_i - \mu_i)^2,$$

hvilket er begrundelsen for, at man kalder de givne størrelser w_i for “vægte” og taler om “vægtet mindste kvadraters metode”.

Det beregningsmæssige problem kan føres tilbage til en sædvanlig varianshomogen model ved transformation. Hvis vi med $\sqrt{W}$ betegner diagonalmatricen med elementer $\sqrt{w_1}, \dots, \sqrt{w_n}$ fås åbenbart

$$\sqrt{W}y \sim N\left(\sqrt{W}X\beta, \lambda I_n\right),$$

hvilket er en sædvanlig variansanalysemodel med modelmatrix $\sqrt{W}X$, varians λ og datavektor $\sqrt{W}y$. Ved at indsætte i den sædvanlige formel $\hat{\beta} = (X'X)^{-1}X'y$ fås

$$\hat{\beta} = ((\sqrt{W}X)'(\sqrt{W}X))^{-1}(\sqrt{W}X)'\sqrt{W}y = (X'WX)^{-1}X'Wy.$$

I ISU's FITLINEARNORMAL kommando kan vægte specificeres ved, at navnet på den variate der indeholder vægtene skrives bag en skråstreg lige efter modelformlen. For eksempel

`FITLINEAR Y=1+X/W.`

I SAS angives “WEIGHT” vektoren ved en option til PROC GLM.

Eksempel 1.

Betragt en model for uafhængige positive observationer y_i , hvor middelværdi og varians er specificeret på formen

$$\mu_i = E y_i = \beta x_i \quad \sigma_i^2 = \text{var}(y_i) = \lambda x_i^2.$$

Altså en model for en responsvariabel y , som i sin middelværdi følger en forklarende variabel x proportionalt, og hvis standardafvigelse $\sqrt{\lambda}x_i$ (idet vi antager $x_i > 0$) ligeledes følger x_i proportionalt. Dvs. variationskoefficienten σ_i/μ_i er konstant.

Hvis y 'erne antages normalfordelte har vi åbenbart

$$\frac{y_i}{x_i} \sim N(\beta, \lambda).$$

Altså en model hvor forholdene y_i/x_i opfattes som identisk fordelte. Dette er den model som i ISUW-programmet MODULISU blev benyttet for de gennemsnitlige passagervægte $q_i = v_i/\text{pts}_i$. Vi kunne altså have opnået det samme ved at benytte en regressionsmodel hvor vægtene afhænger lineært af passagertallene, med afskæring 0 og en varians der følger de kvadrerede passagertal proportionalt.

En anden model, som approksimativt fører til samme resultater hvis variationskoefficienten σ/μ er lille (så vi for eksempel er nogenlunde sikre på at alle y 'er er positive) går ud på følgende. Betragt de log-transformerede observationer $z_i = \log(y_i)$. For disse gælder approksimativt

$$E(z_i) \approx \log(\beta x_i) = \log(\beta) + \log(x_i)$$

og

$$\text{var}(z_i) \approx \left(\left[\frac{d}{dy} \log(y) \right]_{y=\beta x_i} \right)^2 \lambda x_i^2 = \left(\frac{1}{\beta x_i} \right)^2 \lambda x_i^2 = \frac{\lambda}{\beta^2}$$

hvilket også kan formuleres således: De transformerede observationer $\log(y_i/x_i) = \log y_i - \log x_i$ er identisk fordelte med middelværdi $\log \beta$ og varians λ/β^2 . Denne model kan være at foretrække hvis de log-transformerede forhold $\log(y_i/x_i)$ ser mere "normale" ud end selve forholdene y_i/x_i .

Eksempel 2: Poissonmodel med overspredning.

Betragt en sædvanlig multiplikativ Poissonmodel

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

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 lineær model.

Imidlertid viser det sig (antager vi) at “goodness-of-fit-testet” (test mod den fulde model v.h.a. Pearson eller $-2\log Q$) ikke godkender modellen. Det betyder at der er “overspredning” — y ’ernes varianser er større end de skulle være i henhold til modellen. I så fald er det ikke tilrådeligt at gå videre med denne model. Hvis man ignorerer overspredningen får man typisk for mange “falske signifikanser”; de estimerede standardafvigelser på parameterestimerer bliver for små, da de er baserede på Poissonvariansen og ikke tager højde for overspredningen.

I stedet kan man opstille en *overspredningsmodel*, som i sin simpleste version går ud på at observationerne stammer fra en 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

$$y_i \sim N(\lambda_i, \varphi \lambda_i), \text{ hvor } \lambda_i = \exp(\alpha_g + \beta x_i + \dots).$$

φ 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 til det der kaldes *iteratively reweighted least squares* (kendes blandt andet fra generaliserede lineære modeller).

Et naturligt estimat for φ er Pearsons χ^2 -størrelse for “goodnes-of-fit” divideret med sit frihedsgradsantal. Man kan også bruge $-2\log Q$ (deviansen) i stedet for, de er jo approksimativt ens.

Overspredningsmodellen er 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 $-2\log Q$ -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 parameterestimerer skal korrigeres tilsvarende, ved multiplikation med $\sqrt{\hat{\varphi}}$.

Det er dog ikke nødvendigt at foretage disse modifikationer i hånden. I ISU kan man f.eks. 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

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

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.

En lineær model for $\log y$ 'erne som er lidt mere loyal mod den oprindelige variansstruktur går ud på at

$$\log(y_i) \sim N\left(\alpha_g + \beta x_i + \dots, \frac{\varphi}{w_i}\right) ,$$

hvor

$$w_i = \exp(\alpha_g + \beta x_i + \dots) .$$

Dette giver ganske vist ikke umiddelbart mening, da vi ikke kender parametrene og derfor ikke kan beregne w_i . Men man kan bruge metoden rekursivt med indsættelse af estimater fra forrige trin. For det meste konvergerer den hurtigt.

Se demonstrationsprogrammet V-REG.ISU.