

Kapitel 6

Generelle lineære mixede modeller

6.0 Introduktion

De hierarkiske lineære normalfordelingsmodeller, vi betragtede i det foregående afsnit, er specialtilfælde af en generel klasse af modeller, de såkaldte generelle lineære mixede modeller (*General linear mixed models*). I det følgende skitseres for referenceformål den generelle teori for denne modelklasse. For en mere udførlig beskrivelse henvises til Pawitan [32] eller Pinheiro og Bates [26].

6.1 Formulering af den generelle lineære mixed model

Modellen

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon} \quad (6.1)$$

hvor $\mathbf{X}$ og $\mathbf{Z}$ er kendte matricer, og hvor $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$ og $\mathbf{U} \sim N(\mathbf{0}, \mathbf{D})$ er uafhængige, kaldes en *generel lineær mixed model*. I det generelle tilfælde antages variansmatricerne $\mathbf{R}$ og $\mathbf{D}$ at afhænge af nogle ukendte parametre, $\boldsymbol{\theta}$, som også skal estimeres.

Parametrene $\boldsymbol{\beta}$ kaldes systematiske parametre (*fixed effects*), mens størrelserne $\mathbf{U}$ kaldes tilfældige effekter (*random effects*).

På grund af uafhængigheden mellem $\mathbf{U}$ og $\boldsymbol{\epsilon}$ har vi

$$\mathbf{D} \left[\begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{U} \end{pmatrix} \right] = \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}$$

Modellen kan også opfattes som en hierarkisk model

$$\mathbf{U} \sim \mathbf{N}(\mathbf{0}, \mathbf{D}) \quad (6.2)$$

$$\mathbf{Y} | \mathbf{U} = \mathbf{u} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}) \quad (6.3)$$

med tæthederne

$$f_U(\mathbf{u}; \boldsymbol{\theta}) = \frac{1}{(\sqrt{2\pi})^q \det(\mathbf{D})} \exp \left[\frac{-1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u} \right] \quad \text{for } \mathbf{u} \sim \mathbb{R}^q \quad (6.4)$$

og

$$f_{Y|u}(\mathbf{y}; \boldsymbol{\beta}) = \frac{1}{(\sqrt{2\pi})^N \det(\mathbf{R})} \exp \left[\frac{-1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right] \quad (6.5)$$

for $\mathbf{y} \sim \mathbb{R}^N$

Det følger da, at den marginale fordeling af $\mathbf{Y}$ er en normalfordeling med

$$\mathbf{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} \quad (6.6)$$

$$\mathbf{D}[\mathbf{Y}] = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}' \quad (6.7)$$

Vi indfører symbolet $\mathbf{V}$ for dispersionsmatricen i den marginale fordeling af $\mathbf{Y}$, dvs

$$\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}' \quad (6.8)$$

Matricen $\mathbf{V}$ kan let svulme op og blive vanskelig at håndtere

6.1.1 Eksempel: Envejs model med tilfældig effekt

Envejsmodellen med tilfældig effekt:

$$Y_{ij} = \mu + U_i + e_{ij}$$

som vi betragtede tidligere, udtrykkes som

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \epsilon \quad (6.9)$$

med

$$\begin{aligned} \mathbf{X} &= \mathbf{1}_N \\ \beta &= \mu \\ \mathbf{U} &= (U_1, U_2, \dots, U_q)' \\ \mathbf{R} &= \sigma^2 \mathbf{I}_N \\ \mathbf{D} &= \sigma_b^2 \mathbf{I}_q \end{aligned}$$

hvor $\mathbf{1}_N$ er en søjle af ettaller. Det i, j 'te element i den $N \times q$ dimensionale matrix $\mathbf{Z}$ er lig med 1, hvis y_{ij} kommer fra den i 'te gruppe, ellers er det nul.

6.2 Estimation

6.2.1 Estimation af fixed effects og variansparametre

Fixed-effects parametrene β og variansparametrene θ estimeres fra den marginale fordeling af $\mathbf{Y}$. Log-likelihood'en for fixed-effects parametrene bliver

$$\ell(\beta, \theta; \mathbf{y}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

Så for fastholdt θ bliver scorefunktionen mht β jvf (3.17) på side 46

$$\ell'_\beta(\beta; \theta) = \mathbf{X}' [\mathbf{V}^{-1} \mathbf{y} - \mathbf{V}^{-1} \mathbf{X}\beta] \quad (6.10)$$

Profillikelihood'en for variansparameteren θ er da

$$\ell(\theta) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{Y} - \mathbf{X}\hat{\beta})' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\beta}) \quad (6.11)$$

hvor $\hat{\beta}$ er bestemt ved at sætte (6.10) lig med nul. Den observerede Fisher information for β er

$$I(\hat{\beta}) = \mathbf{X}' \mathbf{V}^{-1} \mathbf{X} \quad (6.12)$$

hvorfra man kan bestemme et estimat for dispersionsmatricen for $\hat{\beta}$

For at bestemme estimer for variansparametrene $\boldsymbol{\theta}$ modificeres profillikelihood'en (6.11) for $\boldsymbol{\theta}$ for at tilgodese estimationen af $\boldsymbol{\beta}$. Den *modificerede profillikelihood* bliver

$$\ell_m(\boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (6.13)$$

Modifikationen svarer til at bestemme REML (residual maximum likelihood estimatet) for variansparametrene $\boldsymbol{\theta}$.

Hvis $\hat{\boldsymbol{\beta}}$ afhænger af $\boldsymbol{\theta}$ er det nødvendigt at bestemme løsningen ved iteration. (Det kan undertiden være ganske kompliceret at bestemme et analytisk udtryk for (6.13)). Pawitan [25] angiver i afsnit 17.5 en simplere fremgangsmåde, der ikke kræver bestemmelse af $\mathbf{V}$.

6.2.2 Estimation af de tilfældige effekter

Log-likelihood'en for alle parametrene bygger på den simultane tæthed for $(\mathbf{Y}, \mathbf{U})$, der fremkommer som

$$f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\theta}) = f_{Y|u}(\mathbf{y}; \boldsymbol{\beta}) f_U(\mathbf{u}; \boldsymbol{\theta})$$

hvor $f_{Y|u}(\mathbf{y}; \boldsymbol{\beta})$ og $f_U(\mathbf{u}; \boldsymbol{\theta})$ er givet ved (6.4) og (6.5). Man får derfor

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) &= -\frac{1}{2} \log |\mathbf{R}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ &\quad - \frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u} \end{aligned} \quad (6.14)$$

hvorfor

$$\frac{\partial}{\partial \mathbf{u}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) = \mathbf{Z}' \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) - \mathbf{D}^{-1} \mathbf{u} \quad (6.15)$$

Ved at sætte den afledede lig nul og løse med hensyn til $\mathbf{u}$ finder man, at estimatet $\hat{\mathbf{u}}$ er løsning til

$$(\mathbf{Z}' \mathbf{R}^{-1} \mathbf{Z} + \mathbf{D}^{-1}) \mathbf{u} = \mathbf{Z}' \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (6.16)$$

hvor man indsætter estimatet $\hat{\boldsymbol{\beta}}$ i stedet for $\boldsymbol{\beta}$. Løsningen til (6.16) er BLUP-estimatet (best linear unbiased predictor).

Den anden afledede af log-likelihood'en med hensyn til $\mathbf{u}$ er

$$\frac{\partial^2}{\partial \mathbf{u} \partial \mathbf{u}'} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) = -\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{D}^{-1}$$

hvorfor man har, at den observerede Fisher information mht $\mathbf{b}$ er

$$\mathbf{I}(\hat{\mathbf{u}}) = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})$$

der kan benyttes til bestemmelse af usikkerheden på $\hat{\mathbf{u}}$.

6.2.3 Samtidig estimation af $\boldsymbol{\beta}$ og $\mathbf{u}$

Estimaterne for $\boldsymbol{\beta}$ og for $\mathbf{u}$ er netop de værdier, der samtidigt maksimerer $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u})$ (6.14) for et fast $\boldsymbol{\theta}$.

Man har nemlig af (6.14)

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}) = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})$$

Kombineres dette med (6.15) og sættes lig med nul får man de såkaldte *mixed model equations*

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \quad (6.17)$$

Estimaterne fra disse simultane ligninger er netop de estimater, der fremkommer fra (6.10) og (6.16). Ligningerne gør det muligt at estimere $\boldsymbol{\beta}$ og $\mathbf{u}$ uden at beregne den marginale varians $\mathbf{V}$, eller dens inverse.

Estimationen kan foretages ved en *iterativ backfitting algoritme* som følger:

1. Start med et estimat for $\boldsymbol{\beta}$, for eksempel det sædvanlige mindste kvadraters estimat

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

og iterer mellem nedenstående trin 2 og 3 indtil der er opnået konvergens

2. Beregn en korrigeret observation

$$\mathbf{y}^{kor} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

og estimer $\mathbf{u}$ fra en tilfældig model $\mathbf{y}^{kor} = \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ ved

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{u} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}^{kor}$$

3. Genbereg en korrigeret observation

$$\mathbf{y}^{kor} = \mathbf{y} - \mathbf{Z}\mathbf{u}$$

og estimer $\boldsymbol{\beta}$ fra en fixed effects model $\mathbf{y}^{kor} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ved

$$(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}^{kor}$$

Det er imidlertid et problem, at variansparametrene sædvanligvis indgår i matricerne $\mathbf{R}$ og $\mathbf{D}$. Det vil derfor være nødvendigt at bestemme estimater $\hat{\mathbf{R}}$ og $\hat{\mathbf{D}}$ (fx ved REML), og derefter løse mixed model ligningerne:

$$\begin{pmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{D}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{pmatrix}$$

i stedet for (6.17). Derefter bestemmes reviderede estimater $\hat{\mathbf{R}}$ og $\hat{\mathbf{D}}$. Processen gentages indtil der er opnået konvergens.

6.2.4 Empirisk Bayes-fortolkning

Det ses af (6.16)

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})\mathbf{u} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

at estimatet $\hat{\mathbf{u}}$ for de tilfældige effekter er “krympet” mod nul, idet det er et vejet gennemsnit mellem det direkte estimat $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ og “apriorimiddelværdien” $\mathbf{0}$, hvor vægten er præcisionen $\mathbf{D}^{-1}$ i fordelingen af $\mathbf{U}$.

6.3 SAS-proceduren PROC MIXED

Proceduren er dedikeret til behandling af modellen (6.1)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon}$$

hvor $\mathbf{X}$ angiver “designmatricen” for de systematiske (“fixed”) effekter, og $\mathbf{Z}$ tilsvarende angiver “designmatricen” for de tilfældige (“random”) effekter. Proceduren giver mulighed for en række forskellige varians-kovarians forhold $\mathbf{R}$ og $\mathbf{D}$ for $\boldsymbol{\epsilon}$ og $\mathbf{U}$.

Sætningerne er i en vis udstrækning analoge til sætningerne i PROC GLM. Således specificeres modellen i MODEL, RANDOM og REPEATED sætninger. Der er dog den forskel, at i PROC MIXED beskriver MODEL sætningen alene fixed effects.

MODEL sætningen definerer $\mathbf{X}$ matricen. Strukturen for $\mathbf{R}$ er som default en enhedsmatrix. Ved brug af en Repeated sætning kan man imidlertid frembringe blokmatricer med en hel række forskellige strukturer. Subject = definerer blokkene i $\mathbf{R}$ og Type= definerer strukturen i blokkene.

RANDOM sætningen definerer $\mathbf{Z}$ - matricen og benyttes endvidere til at angive strukturen for $\mathbf{D}$. I RANDOM sætningen konstrueres “design matricen” for de tilfældige effekter. Optionen SUBJECT = identificerer de *uafhængige gentagelser* af de tilfældige effekter. Således bevirker SUBJECT = at $\mathbf{D}$ matricen bliver en blok-diagonal matrix med identiske blokke, en for hvert subject. Brug af SUBJECT= optionen indebærer at $\mathbf{Z}$ -matricen evt bliver modificeret for at respektere blok-diagonaliteten, det vil sige at eventuelle andre tilfældige effekter bliver underordnet subject-effekten.

Strukturen for $\mathbf{D}$ bestemmes som en option TYPE= i RANDOM sætningen (efter en slash). Default optionen er *varianskomponent*, dvs en diagonalmatrix med forskellige værdier for de forskellige effekter.

6.3.1 Eksempel på brug af PROC MIXED

Programstumpen

```

data ramus ;
  input dreng$    alder ramus @@;
  ald = alder - 8.75;
  cards;
  A  8  52.5  A   8.5  53.2  A   9  53.3  A   9.5  53.7
  B  8  51.2  B   8.5  53.0  B   9  54.3  B   9.5  54.5
  C  8  51.2  C   8.5  51.4  C   9  51.6  C   9.5  51.9
  D  8  52.1  D   8.5  52.8  D   9  53.7  D   9.5  55.0
  E  8  50.7  E   8.5  51.7  E   9  52.7  E   9.5  53.3
;

```

indlæser ramushøjderne ved aldrene 8, 8.5, 9, 9.5 år for fem drenge, A,B,C,D og E.

Programmet

```

PROC MIXED DATA=ramus          ;
CLASS DRENG;
  MODEL ramus = ald /solution outp=pred;
  RANDOM int    ald  /subject = dreng solution type=un;
run;

```

analyserer den model for tilfældige regressionslinier, der blev betragtet i afsnit 5.10.2, side 251.

Sætningen `MODEL ramus = ald /solution outp=pred;` beskriver den systematiske del (fixed effekt), nemlig et intercept og en koefficient til alderen, og nøgleordet `solution` bevirker at disse koefficienter bliver udskrevet. Endelig bevirker nøgleordet `outp=pred` at de prædikterede værdier for hver observation (under den tilfældige model) udskrives i datasættet `pred`.

Sætningen `RANDOM int ald /subject = dreng solution type=un;` beskriver den tilfældige del (random effekt), nemlig et tilfældigt intercept og en tilfældig hældning, der begge varierer omkring nul (det er derfor vi også havde brug for at specificere det systematiske bidrag i `MODEL` sætningen). Nøgleordet `subject = dreng` angiver at observationer fra forskellige drenge er uafhængige. Da øvrige tilfældige effekter er underordnet `subject` indebærer det, at det tilfældige intercept og hældning, der er knyttet til den enkelte dreng, opfattes som en todimensional variabel for den enkelte dreng, mens

gentagelser (en ny dreng) er uafhængige. Nøgleordet `type = uns` (unstructured) angiver, at der ikke er påtvunget nogen kovariansstruktur for denne todimensionale variabel.

Kapitel 7

Hierarkisk variation for exponentielle dispersionsfamilier

7.0 Indledning, modellering af overdispersion

En karakteristisk egenskab ved de generaliserede lineære modeller var, at variansen, $V[Y]$ i fordelingen af responset var en kendt funktion, $V(\mu)$, der alene afhænger af middelværdien μ , nemlig

$$V[Y_i] = \lambda_i V(\mu) = \frac{\sigma^2}{w_i} V(\mu)$$

hvor w_i udtrykker en kendt *vægt*, knyttet til den i 'te observation, og hvor σ^2 udtrykker en *dispersionsparameter* fælles for alle observationer.

Dispersionsparameteren σ^2 benyttes som beskrevet i afsnit 4.12 til at udtrykke *overdispersion* i situationer, hvor residualdeviansen er større end hvad der kan forklares ved variansfunktionen $V(\mu)$ og kendte vægte w_i .

I dette afsnit vil vi beskrive en alternativ fremgangsmåde til modellering af overdispersion, nemlig ved hierarkiske modeller analoge til normalfordelingsmodellerne i afsnit 5. Se også Lee og Nelder, Two ways of modelling overdispersion in non-normal data. *Applied Statistics* **49**, (2000), pp 591-598, [30] for en diskussion af de to forskellige fremgangsmåder.

Ved en hierarkiske modellering tager man udgangspunkt i at fordelingen af “støjen” modelleres ved en eksponentiel dispersionsfamilie (binomial, Poisson, etc) og det drejer sig da om at vælge en passende (apriori) fordeling af middelværdiparameteren μ . Det vil være naturligt at vælge en fordeling, hvis støtte netop er middelværdirummet $\mathcal{M}$, snarere end at vælge en normalfordeling, hvis støtte jo er hele den reelle akse. I dette kapitel vil vi indføre de såkaldte *konjugerede* fordelingsfamilier til eksponentielle dispersionsfamilier til modellering af variationen af middelværdiparameteren μ .

Ved gennemgangen vil vi først beskrive nogle eksempler, Poisson-gamma, Binomial Beta, etc. i afsnittene 7.1 til 7.4

I afsnit 7.5 indføres den mere generelle formulering, og bestemmelsen af aposteriorifordeling beskrives generelt i afsnit 7.6 og eksemplificeres for Poissonfordelingssituationen i afsnit 7.7.

7.1 Hierarkisk Poisson Gamma model

7.1.1 Eksempel, Variation mellem episoder af tordenvejr ved Cape Kennedy

Nedenstående tabel (efter Williford et al. (1974)) viser fordelingen af antal daglige tordenvejrsepisoder ved Cape Kennedy, Florida i månederne juni, juli og august for 10-års perioden 1957-1966, ialt 920 døgn.

Fordelingen af antal dage med 0,1,2 eller flere episoder af tordenvejr ved Cape Kennedy

Antal episoder z_i	Antal dage $\# i$	Poisson forventet
0	803	791.85
1	100	118.78
2	14	8.91
3+	3	0.46

Alle observationsperioder er på $n_i = 1$ dag.

For Poissonfordelingen er variansfunktionen $V(\mu) = \mu$, og såfremt observationerne er uafhængige og identisk fordelte skal variansen således være lig med middelværdien. Man finder det gennemsnitlige antal tordenvejr $\bar{y}_+ = 0.15$ og den fundne varians er $s^2 = 0.1769$, der er noget større end gennemsnittet. Det ses også, at den observerede fordeling har *tykkere haler* end Poissonfordeling. Man kunne således få en mistanke om overdispersion. En sådan overdispersion kunne eventuelt forklares ved at antallet af tordenvejr på en given dag varierer som en Poisson fordelt variabel, men at Poisson-intensiteten varierer mellem dagene.

I det følgende vil vi beskrive en sådan model med varierende intensitet.

7.1.2 Formulering af hierarkisk model

Simpel hierarkisk model

Der gælder

Lad fordelingen af Y givet μ være en Poissonfordeling $Y \sim P(\mu)$, og lad fordelingen af μ være en Gammafordeling $G(\alpha, \beta)$, da er den marginale fordeling af Y en negativ binomialfordeling, $Y \sim \text{NB}(\alpha, 1/(1+\beta))$ fordeling.

Bevis

For illustrationens skyld anføres beviset:

Vi har frekvensfunktionen for fordelingen af Y givet μ

$$f_{Y|\mu}(y; \mu) = \frac{\mu^y}{y!} \exp(-\mu) \quad (7.1)$$

og tæthedsfunktionen for fordelingen af μ

$$f_{\mu}(\mu; \alpha, \beta) = \frac{1}{\beta \Gamma(\alpha)} \left(\frac{\mu}{\beta} \right)^{\alpha-1} \exp(-\mu/\beta) \quad (7.2)$$

Da bestemmes frekvensfunktionen for den *marginale* fordeling af Y af

$$\begin{aligned} g_Y(y; \alpha, \beta) &= \int_{\mu=0}^{\infty} f_{Y|\mu}(y; \mu) f_{\mu}(\mu; \alpha, \beta) d\mu \\ &= \int_{\mu=0}^{\infty} \frac{\mu^y}{y!} \exp(-\mu) \frac{1}{\beta \Gamma(\alpha)} \left(\frac{\mu}{\beta}\right)^{\alpha-1} \exp(-\mu/\beta) d\mu \\ &= \frac{1}{\beta^\alpha y! \Gamma(\alpha)} \int_{\mu=0}^{\infty} \mu^{y+\alpha-1} \exp(-\mu(\beta+1)/\beta) d\mu \end{aligned} \quad (7.3)$$

Det ses, at integranden netop er *kernen* for tæthedsfunktionen for en gammafordeling, $G(y+\alpha, \beta/(\beta+1))$. Da tætheden skal integrere til et ettal, har vi altså ved at betragte normeringskonstanten, at

$$\int_{\mu=0}^{\infty} \mu^{y+\alpha-1} \exp(-\mu/[\beta/(\beta+1)]) d\mu = \frac{\beta^{y+\alpha} \Gamma(y+\alpha)}{(\beta+1)^{y+\alpha}}$$

hvorfor vi får

$$g_Y(y; \alpha, \beta) = \frac{\Gamma(y+\alpha)}{y! \Gamma(\alpha)} \frac{\beta^y}{(\beta+1)^{y+\alpha}} = \binom{y+\alpha-1}{y} \frac{1}{(\beta+1)^\alpha} \left(\frac{\beta}{\beta+1}\right)^y \quad (7.4)$$

hvor vi har benyttet konventionen

$$\binom{z}{y} = \frac{\Gamma(z+1)}{\Gamma(z+1-y) y!}$$

for z reel og y heltallig.

Det ses at (7.4) netop er frekvensfunktionen for en $NB(\alpha, 1/(1+\beta))$ fordeling jvf oversigtstabellen side 23.

▽

På tilsvarende måde kan man vise:

Såfremt $Z|\mu \sim P(n\mu)$, og fordelingen af μ er en $G(\alpha, \beta)$ fordeling, da er den marginale fordeling af Z en $NB(\alpha, 1/(1+n\beta))$ fordeling.

I stedet for at parametrisere fordelingen af μ ved formparameter α og skalaparameter β for gammafordelingen kan man vælge at parametrisere ved den reciprokke formparameter, og middelværdi $m = \alpha\beta$, dvs

$$k = 1/\alpha \quad (7.5)$$

$$m = \alpha\beta \quad (7.6)$$

Resultatet kan da udtrykkes som

Såfremt $Z|\mu \sim P(n\mu)$, og fordelingen af μ er en $G(1/k, mk)$ fordeling, da er den marginale fordeling af Z en $NB(1/k, 1/(1 + nmk))$ fordeling.

Vi har

$$\begin{aligned} E[Z|\mu] &= n\mu \\ V[Z|\mu] &= n\mu \\ E[\mu] &= m \\ V[\mu] &= m^2k \end{aligned}$$

hvorfor

$$\begin{aligned} E[Z] &= nE[\mu] = nm \\ V[Z] &= E_{\mu}[V[Y|\mu]] + V_{\mu}[E[Y|\mu]] \\ &= E[n\mu] + V[n\mu] = nE[\mu] + n^2V[\mu] \\ &= nm + n^2m^2k \end{aligned}$$

hvorfor vi har for $Y = Z/n$

$$\begin{aligned} E[Y] &= m \\ V[Y] &= \frac{1}{n^2} V[Z] = \frac{m}{n} + m^2k \end{aligned}$$

Indfører vi “signal/støj” forholdet

$$\gamma = \frac{V[\mu]}{E[V_P(\mu)]} = \frac{V[\mu]}{E[\mu]} = \frac{m^2k}{m} = mk \quad (7.7)$$

ses, at vi kan udtrykke formparameteren α som

$$\alpha = \frac{1}{k} = \frac{m}{\gamma}$$

Parametriseringen ved $k = 1/\alpha$ indebærer at parameterområdet $0 < \alpha < \infty$ føres over i området $0 < k < \infty$, og specielt har vi for $\alpha \rightarrow \infty$ vil $k \rightarrow 0$ og $G(1/k, mk)$ fordelingen vil nærme sig en etpunktsfordeling med hele massen

i $\mu = m$ og tilsvarende vil $NB(1/k, 1/(1 + nm k))$ fordelingen nærme sig en Poissonfordeling med middelværdi nm . Tilsvarende vil signal/støj forholdet γ nærme sig nul. Med disse fortolkninger er formlerne således også gyldige for $k = 0$, og man kan teste for overdispersion ved at teste om $k = 0$.

7.1.3 SAS PROC GENMOD

SAS proceduren PROC GENMOD har fordelingsoptionen `DIST = NEGBIN` med logaritmisk link som tilhørende default linkfunktion.

I lighed med de eksponentielle dispersionsmodeller benyttes notationen $Y \sim NB(1/k, 1/(1 + k\mu))$ med

$$E[Y] = \mu \quad V[Y] = \mu(1 + k\mu)$$

svarende til frekvensfunktionen

$$f_Y(y; \mu, k) = \frac{\Gamma(y + 1/k)}{y! \Gamma(1/k)} \frac{(k\mu)^{1/k}}{(1 + k\mu)^{y+1/k}} \quad \text{for } y = 0, 1, 2, \dots$$

I SAS-outputtet kaldes parameteren k *dispersionsparameter* selv om k ikke er dispersionsparameter i en generaliseret lineær model, idet enhedsvariansfunktionen for den negative binomialfordeling jo er $V_{NB}(\mu) = \mu(1 + \mu)$ jvf Tabel 4.1 på side 115.

Opfattet som marginal fordeling af Y svarende til den hierarkiske model i det foregående afsnit benyttes parameteren m i stedet for μ , og parameteren n sættes lig 1, sådan at man har

$$\begin{aligned} E[Y] &= m \\ V[Y] &= m + m^2 k \end{aligned}$$

7.1.4 Eksempel, tordenvejr igen

Ved numerisk maksimering af likelihoodfunktionen finder man $\hat{m} = 0.1489$ [tordenvejr/dag] og $k = 1.3022$, dvs $\alpha = 1/k = 0.7679$ og signal/støjforhold $\gamma = m/\alpha = mk = 0.1939$