Dagens program:

Mandag den 2. maj
Hierarkiske generaliserede lineære modeller - Afslutning

- Hierarkisk generaliseret lineær model
- Lidt mere om andre likelihood'er og censurerede observationer
- Lidt om repræsentatitivitet
- Slut

Vi husker en Generel lineær mixed model

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

$\boldsymbol{\epsilon} \sim \mathsf{N}(\mathbf{0}, \mathbf{R})$ og $\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \mathbf{D})$

Fixed effects $\boldsymbol{\beta}$ (designmatrix $\mathbf{X}$)
Random effects $\mathbf{U}$ (designmatrix $\mathbf{Z}$)

$$\mathsf{E}[\mathbf{Y}|\mathbf{u}] = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

Mixed effect ligningerne

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

løses ved iteration - og iterativ bestemmelse af parametre i $\mathbf{R}$ og $\mathbf{D}$

Estimaterne for $\mathbf{u}$ er BLUP-estimater. Varianser estimeres sævanligvis ved REML

Hierarkiske generaliserede lineære modeller

Lee og Nelder, Biometrika (2001) **88**, pp 987-1006

Ingredienser:
Betinget fordeling af Y for givet θ følger en exponentiel dispersionsparameterfamilie på grundform

$$f_Y(y;\theta,\lambda) = c(y,\lambda)\exp[\lambda\{\theta y - \kappa(\theta)\}] \quad \text{for } \theta \in \Omega$$

Middelværdirum $\mathcal{M}$ (for μ) - kanonisk parameterrum Ω (for θ)

$\mu = \kappa'(\theta)$; **Kanonisk link**, $\theta = g(\mu)$

konjugeret fordeling af kanonisk parameter, θ

$$g_\theta(\theta; m, \gamma) = \frac{1}{C(m,\gamma)}\exp\{[\theta m - \kappa(\theta)]/\gamma\}$$

og tilsvarende konjugeret fordeling af middelværdiparameter, μ
Eks: $f_Y(y)$ Poisson, $g_\mu(\mu)$ Gamma, link= log, dvs $g(\mu) = \log(\mu)$

Hierarkiske generaliserede lineære modeller

Vi har responser $y_1, y_2, \ldots, y_k$ ($\in \mathcal{M}$) og **uobserverede** tilstandsvariable $u_1, u_2, \ldots, u_q$ ($\in \mathcal{M}$) med de tilsvarende uobserverede **kanoniske** variable $v_i = g(u_i)$. ($\in \Omega$)

Lineær prædiktor

$$\boldsymbol{\theta} = g(\boldsymbol{\mu}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$$

Fordeling af $\mathbf{v} \in \Omega$ er konjugeret fordeling til kanonisk parameter θ.
Fordeling af $\mathbf{U} \in \mathcal{M}$ er den tilsvarende konjugerede fordeling til middelværdiparameter μ.
$\mathsf{E}[U] = \psi$

og så bliver fordelingen af Y den tilsvarende marginale fordeling.

For $Y|\mu$ Poisson vælges fordeling af v så $u = \ln(v)$ er gamma med middelværdi $\psi = 1$, og så bliver fordelingen af Y en negativ binomialfordeling med parametre bestemt af $\mathbf{X}\boldsymbol{\beta}$ og af $\mathbf{Z}\mathbf{v}$.

Estimation ved en udvidet generaliseret lineær model for y, ψ, se Lee og Nelder.

Estimation i hierarkisk generaliseret lineær model

Estimation af "middelværdiparametre" $\boldsymbol{\beta}$ og $\mathbf{v}$

$$\begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}_0^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}_0^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Sigma}_0^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Sigma}_0^{-1}\mathbf{Z}+\boldsymbol{\Sigma}_1^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{v}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\Sigma}_0^{-1}\mathbf{z}_0 \\ \mathbf{Z}'\boldsymbol{\Sigma}_0^{-1}\mathbf{z}_0+\boldsymbol{\Sigma}_1^{-1}\mathbf{z}_1 \end{pmatrix}$$

hvor $\mathbf{z}_0 = \boldsymbol{\eta}_0 + (\mathbf{y}-\boldsymbol{\mu}_0)(\partial\eta_0/\partial\mu_0)$ og $\mathbf{z}_1 = \mathbf{v}_1 + (\psi_1) - u)(dv_1/du)$ er de justerede afhængige variable hhv for fordelingen af y givet v og og af v.

Løses ved iteration

Generaliseret lineær mixed model

En hybrid

$$\boldsymbol{\theta} = g(\boldsymbol{\mu}|\mathbf{v}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$$

Fordeling af $\mathbf{v} \in \Omega$ modelleres som en normalfordeling .

Indskud: Lidt mere om likelihood'er

Likelihoodfunktion defineret ved simultan tæthed af observationerne som funktion af parametre.
Udtrykker grad af overensstemmelse mellem parameter og data.
Der kan vises en række gode egenskaber for maksimum-likelihood estimat.

Vi opfandt quasi-likelihood til situationer, hvor vi minimerede devians (defineret ved en variansfunktion), og så bort fra normeringskonstanten

og når vi havde overdispersion (udtrykt ved en **dispersionsparameter**)

Når man opfinder en variant af likelihoodbegrebet skal man principielt eftervise at en række gode egenskaber (konvergens mod "sand værdi" for voksende stikprøvestørrelse) stadig gælder

Indskud: nok en likelihoodvariant

Eksempel tælleprocesser (levetider)

91 komponenter under test. Efter et år er 8 af komponenterne fejlede, mens de resterende 83 stadig virker.

dvs vi har 8 levetider, $t_1, t_2, \ldots, t_8$ og viden om at de resterende 83 levetider er større end et år.

Censurerede observationer. Giver anledning til en anden udvidelse af likelihoodbegrebet, **partial likelihood**

Vi vil ikke diskutere dette begreb, men vi vil lige se nogle begreber, der optræder i forbindelse med levetidsmodeller

Fordelingen af den mindste blandt 91 er anderledes end fordelingen af en tilfældig af de 91.

Tælleprocesser, hazardfunktion

Positiv stokastisk variabel, T.
Tæthedsfunktion $f_T(t)$, kumuleret fordeling

$$F_T(t) = \mathsf{P}[T \leq t] = \int_0^t f_T(x)dx$$

Overlevelsesfunktion (survival function)

$$R_T(t) = 1 - F_T(t) = \mathsf{P}[T > t]$$

Hazardfunktion

$$\lambda_T(t) = \frac{f_T(t)}{1 - F_T(t)} = \frac{-d \log R_T(t)}{dt}$$

ssh for at dø netop til tid t, når man har overlevet op til t

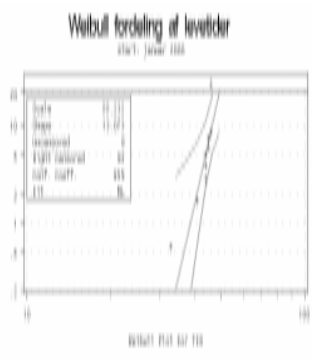
$$R_T(t) = \exp\left(-\int_0^t \lambda_T(u)du\right) = \exp\Big(-\Lambda(t)\Big)$$

hvor $\Lambda(t)$ angiver den **kumulerede hazard**

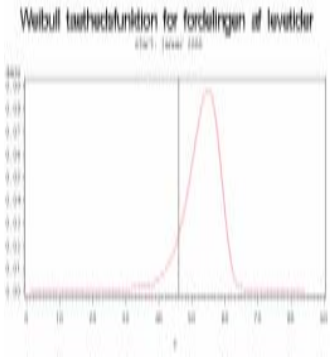
$$\Lambda(t) = \int_0^t \lambda_T(u)du$$

log-lineære modeller, modeller med proportionale hazards

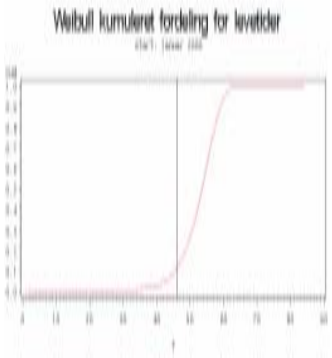
Levetider for 8 samlemuffer



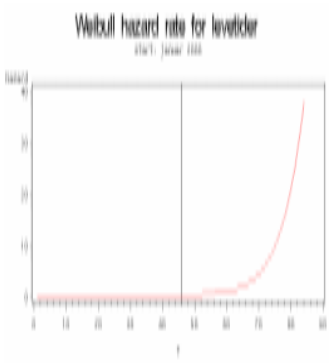
Levetider for 8 samlemuffer



Levetider for 8 samlemuffer



Levetider for 8 samlemuffer



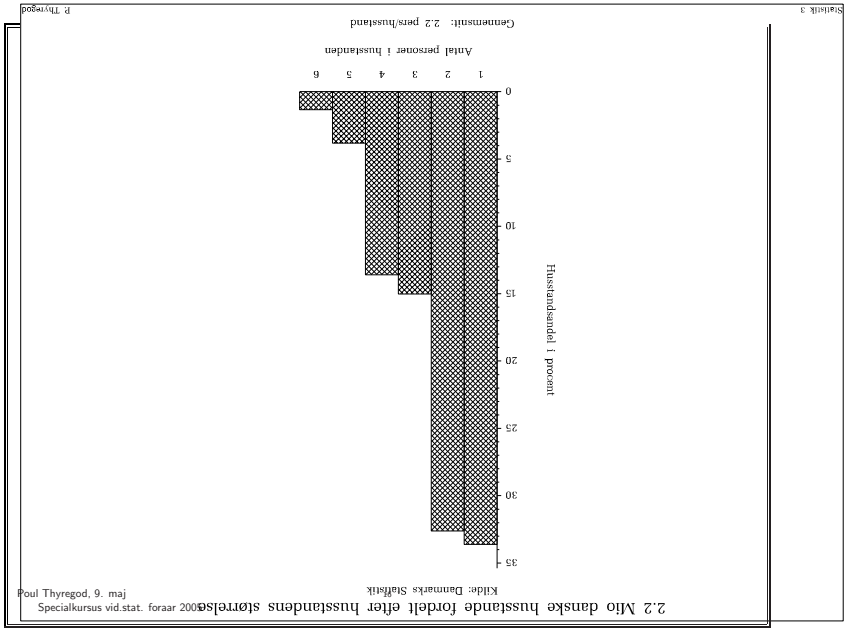
Konkurrerende risici

Eksempel: opholdstid på DTU

Lidt om repræsentativitet

Der udvælges en stikprøve tilfældigt blandt personer i Danmark, og man spørger, hvor mange personer, der bor i deres husstand.

Man får et gennemsnit på 3.8 personer. Er dette udtryk for at den gennemsnitlige husstandsstørrelse er på ca. 3.8 personer ??



2.2 Mio danske husstande fordelt efter husstandens størrelse

Kilde: Danmarks Statistik

DTU

Oversigt

Udgangspunkt generel lineær model (normalfordeling, konstant varians)

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

Likelihood metode svarer til mindste kvadraters metode, projektioner i vektorrum (pga linearitet)

Udvidelse, generalisering i to retninger:

- til andre responsfordelinger, generaliserede lineære modeller
  varians afhænger af middelværdi. Ved estimation vægtes observationer med deres præcision
- til hierarkiske modeller, generel lineær mixed model
  Observationer korrelerede, BLUP-estimation, empirisk Bayes, "krympning" mod fælles gennemsnit

DTU

Lidt om repræsentativitet

Analoge situationer

- Rejsehyppighed/længde for buspassagerer
- liggedage på hospitaler
- Fiberlængde i garn