Dagens program:

Mandag den 25. april
Hierarkiske normalfordelingsmodeller

- Resume af ensidet variansanalysemodel med tilfældig effekt estimation af tilfældige effekter, fortolkning som BLUP estimation, empirisk Bayes-estimation, Steins sætning
- Tilfældige regressionslinier
- Hierarkiske flerdimensionale observationer
- Generelle lineære mixed models
- SAS PROC Mixed
- Begyndelse af hierarkiske modeller for eksponentielle familier

Hierarkisk model, normalfordeling

Først udtages en balle tilfældigt. **Derefter** udtages fire prøver tilfældigt fra **denne balle**.

Y_{ij} j'te prøve fra i'te balle.

For fastholdt balle, uafhængige gentagelser.

$$Y_{ij}|\mu_i \sim \mathsf{N}(\mu_i, \sigma^2)\ ,$$
$$\mathbf{Y}_i|\mu_i \sim \mathsf{N}(\mu_i, \sigma^2\mathbf{I}_4)$$

Ballerenheden μ_i varierer fra balle til balle:

$$\mu_i \sim \mathsf{N}(\mu_0, \sigma_b^2)$$

indbyrdes uafhængige

Modellens parametre

Modellen har parametrene $\mu_0, \sigma^2, \sigma_b^2$.

Vi sætter

$$\gamma = \sigma_b^2/\sigma^2$$

(Signal støj- forhold)

Fordeling af i'te sæt

Lad $\mathbf{Y}_i$ angiver observationer fra i'te udtagne balle.

$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_q$ er uafhængige observationssæt

$$\mathbf{Y}_i \sim \mathsf{N}_{n_i}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I}_{n_i} + \sigma_b^2\mathbf{J}_{n_i})$$

hvor $\mathbf{J}_n$ angiver en $n \times n$ matrix med ettaller.

Dispersionsmatricen for $\mathbf{Y}_i$ er

$$\begin{aligned}\mathbf{V}_i &= \mathbf{D}\,[\mathbf{Y}_i] = \mathsf{E}[(\mathbf{Y}_\mathrm{i} - \boldsymbol{\mu}_0)(\mathbf{Y}_\mathrm{i} - \boldsymbol{\mu}_0)^\mathsf{T}] \\ &= \begin{pmatrix} \sigma_b^2+\sigma^2 & \sigma_b^2 & \ldots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma^2 & \ldots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \ldots & \sigma_b^2+\sigma^2 \end{pmatrix}\end{aligned}$$

Fordeling af hele observationssættet

$\mathbf{Y}$ består af q uafhængige gentagelser med fordeling som $\mathbf{Y}_i$

$$\mathsf{D}\,[\mathbf{Y}] = \mathbf{V} = \text{ Blok diag}\{\mathbf{V}_i\}$$

Estimation (prædiktion) af enkelte ballemiddelværdier

Løsningen tilfredsstiller

$$\widehat{\mu}_i = (1 - w_i(\widehat{\gamma}))\overline{y}_i + w_i(\widehat{\gamma})\widehat{\mu}_0$$

et **vejet gennemsnit** mellem de **individuelle gennemsnit** $\overline{y}_i$ og det **fælles gennemsnit** $\widehat{\mu}_0$ med vægtene $(1 - w_i(\widehat{\gamma}))$ og $w_i(\gamma)$, hvor

$$1 - w = \frac{\mathsf{V}[\mathsf{E}[\widehat{\theta}|\theta]]}{\mathsf{E}[\mathsf{V}[\widehat{\theta}|\theta]] + \mathsf{V}[\mathsf{E}[\widehat{\theta}|\theta]]}$$

$$w = \frac{\mathsf{E}[\mathsf{V}[\widehat{\theta}|\theta]]}{\mathsf{E}[\mathsf{V}[\widehat{\theta}|\theta]] + \mathsf{V}[\mathsf{E}[\widehat{\theta}|\theta]]}$$

netop er de to bidrag (variationen indenfor balle, og variationen mellem baller) fra opspaltningen af den totale varians

$$\mathsf{V}[\widehat{\theta}] = \mathsf{E}[\mathsf{V}[\widehat{\theta}|\theta]] + \mathsf{V}[\mathsf{E}[\widehat{\theta}|\theta]]$$

Estimation (prædiktion) af enkelte ballemiddelværdier

Det hierarkiske maximum likelihood estimat er BLUP-estimat (best linear unbiased predictor), et vejet gennemsnit mellem stikprøveresultat og fælles middelværdi

Shrinkage (krympning) mod fælles gennemsnit.

Empirisk Bayes fortolkning (estimeret aprioridfordeling).

Empirisk Bayes estimat har mindre estimationsfejl end ML-estimat i systematisk model

Stein's sætning vedrørende simultan estimation:
(Bemærkning 1, side 246.)

Ved en **systematisk** (fixed effects) model for de enkelte balleniveauer estimeres hver balle ved sit gennemsnit, $\overline{y}_i$, med varians σ^2/n.

Betragt den **gennemsnitlige estimationsfejl**

$$R(\boldsymbol{\mu}, d(.)) = \frac{1}{k}\,\mathsf{E}\left[\sum\left(\mathsf{d_i}(\mathbf{y}) - \mu_i\right)^2\right]$$

hvor $d_i(\mathbf{y})$ angiver estimatoren $\widehat{\mu}_i$ for det i'te balleniveau.

Stein's sætning

Betragt opspaltningen af den gennemsnitlige estimationsfejl i varians og bias:

$$\begin{aligned} R(\boldsymbol{\mu}, d(.)) &= \frac{1}{k}\,\mathsf{E}\left[\sum\left(\mathsf{d_i}(\mathbf{y}) - \mu_\mathsf{i}\right)^2\right] \\ &= \frac{1}{k}\,\mathsf{E}\left[\sum\left\{(\mathsf{d_i}(\mathbf{y}) - \mathsf{E}[\mathsf{d_i}])^2 + (\mathsf{E}[\mathsf{d_i}] - \mu_\mathsf{i})^2\right\}\right] = \frac{1}{\mathsf{k}}\sum\left(\mathsf{V}[\mathsf{d_i}] + \delta_\mathsf{i}^2\right) \end{aligned}$$

hvor $\delta_i = \mathsf{E}[\mathsf{d_i}] - \mu_\mathsf{i}$ angiver **bias** for estimatoren $d_i(\mathbf{y})$

Såfremt man vælger maximum-likelihood estimatoren $d_i^{ML}(\mathbf{y}) = \overline{y}_i$, finder man

$$R(\boldsymbol{\mu}, d^{ML}(.)) = \frac{1}{k}\sum \mathsf{V}[\overline{\mathsf{y}}_\mathsf{i}] = \sigma^2/\mathsf{n}$$

idet estimatoren er **central** (ikke har nogen bias).

Man kan vise (Stein 1955), at såfremt man benytter empirisk Bayes estimatoren

$$d_i^{EB}(\mathbf{y}) = w(\mathbf{y})\overline{\overline{y}} + (1 - w(\mathbf{y}))\overline{y}_i$$

gælder for et vilkårligt parametersæt $\boldsymbol{\mu}$, at

$$R(\boldsymbol{\mu}, d^{EB}(.)) < R(\boldsymbol{\mu}, d^{ML}(.))$$

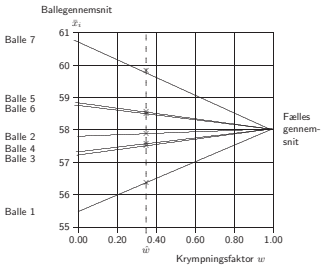
Intuitivt

For $w < 1$ vil $\mathsf{E}[\mathsf{w}\overline{\mathsf{y}}_\mathsf{i}] = \mathsf{w}\mu_\mathsf{i} < \mu_\mathsf{i}$ og $\mathsf{V}[\mathsf{w}\overline{\mathsf{y}}_\mathsf{i}] = \mathsf{w}^2\mathsf{V}[\overline{\mathsf{y}}_\mathsf{i}] < \sigma^2/\mathsf{n}$

Ved brug af empirisk Bayes-estimatet får vi altså reduceret variansen på bekostning af en lille bias $\delta_i = (1 - w)\mu_i$.

$$R(\boldsymbol{\mu}, d^{EB}(.)) \approx \frac{1}{k}\sum[w^2\sigma^2/n + (1 - w^2)\mu_i^2] < \frac{1}{k}\sum \sigma^2/n$$

Empirisk Bayes estimat som shrinkage estimat (vejet gennemsnit mellem fælles middelværdi og individuelle gennemsnit)

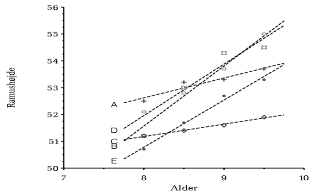


Andre eksempler på hierarkiske modeller, og bestemmelse af aposteriorifordelinger

1. Tilfældige regressionslinier
2. Todimensionale målinger

Varierende regressionslinier

Ramushøjder for 5 drenge ved fire forskellige aldre.



Systematiske modeller: Vi kunne lave én fælles regressionslinie for alle 5 drenge, eller vi kunne estimere 5 individuelle regressionslinier (en for hver dreng),

Varierende regressionslinier, tilfældig (hierarkisk) model

Data opfattes som en **stikprøve** af drenge

$$Y_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,\ i = 1, 2, \ldots, k$$

hvor

$$\boldsymbol{\beta}_i \sim \mathsf{N}_2(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma}), \quad \boldsymbol{\epsilon}_i \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)\ ,$$

$\boldsymbol{\beta}_i, \boldsymbol{\beta}_j$ er indbyrdes uafhængige for $i \neq j$,

Kovariansmatricen $\boldsymbol{\Gamma}$ angiver kovariansmatricen i populationsfordelingen af intercepter og hældninger (med målestøjen σ^2 trukket ud som en faktor).

Aposteriorifordeling af regressionskoefficienter

Y en $n \times 1$ dimensional vektor af observationer, og $\mathbf{X}$ en $n \times p$ dimensional matrix af kendte koefficienter.

Lad $Y \mid \boldsymbol{\beta} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ og lad aprioridfordelingen af $\boldsymbol{\beta}$ være

$$\boldsymbol{\beta} \sim \mathsf{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Lambda})$$

Aposteriorifordelingen af $\boldsymbol{\beta}$ efter observation af $Y = \mathbf{y}$ er

$$\boldsymbol{\beta} \mid Y = \mathbf{y} \sim \mathsf{N}_p(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Gamma}_1)$$

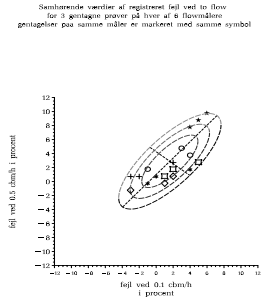
hvor

$$\boldsymbol{\beta}_1 = \mathbf{W}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{W})\widehat{\boldsymbol{\beta}}$$

med $\mathbf{W} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}$ og $\widehat{\boldsymbol{\beta}}$ er den sædvanlige mindste kvadraters estimator

Flerdimensional normalfordeling

3 gentagne kalibreringer af 6 flowmålere, der er udtaget af en større målerpopulation. De 6 målere blev hver kalibreret ved de samme to flow, henholdsvis 0.1 [m^3 /h] og 0.5 [m^3 /h].

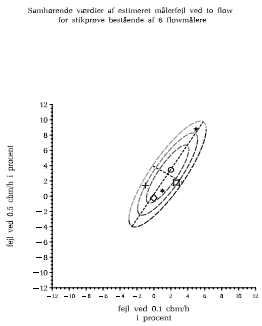


Flerdimensional normalfordeling, hierarkisk model

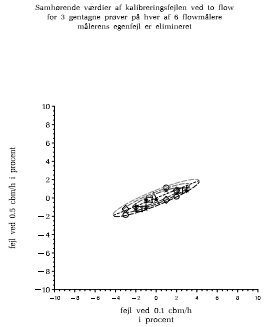
$$\mathbf{X}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}, \;\; i = 41, 42, ..., 46; \;\; j = 1, 2, 3.$$

hvor $\boldsymbol{\alpha}_i$ er uafhængige, $\boldsymbol{\alpha}_i \sim \mathsf{N}_2(0, \boldsymbol{\Sigma}_0)$

Flerdimensional normalfordeling, variation mellem målere



Flerdimensional normalfordeling, variation indenfor målere



Flerdimensional normalfordeling, aposteriorifordeling

Lad $\mathbf{X} \mid \boldsymbol{\mu} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, og lad $\boldsymbol{\mu} \sim \mathsf{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$,

Da er aposteriorifordelingen af $\boldsymbol{\mu}$ efter observation af $\mathbf{X} = \mathbf{x}$ givet ved

$$\boldsymbol{\mu} \mid X = \mathbf{x} \sim \mathsf{N}_p(\mathbf{W}\boldsymbol{\mu}_0 + (\mathbf{I} - \mathbf{W})\mathbf{x}, (\mathbf{I} - \mathbf{W})\boldsymbol{\Sigma})$$

med

$$\mathbf{W} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1} \;\; \text{og} \;\; \mathbf{I} - \mathbf{W} = \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1}$$

Lad $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$ betegne den generaliserede kvotient mellem variansen mellem grupper og variansen inden for grupper, da kan matricerne $\mathbf{W}$ og $\mathbf{I} - \mathbf{W}$ udtrykkes ved

$$\mathbf{W} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1} \;\; \text{og} \;\; \mathbf{I} - \mathbf{W} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}$$

Generel lineær mixed model

$$\mathbf{Y}_{n,1} = \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1} + \mathbf{Z}_{n,m}\mathbf{U}_{m,1} + \boldsymbol{\epsilon}_{n,1}$$

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

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

$$\mathsf{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$$

$$\mathsf{D}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$$

$\mathbf{R}$ og $\mathbf{D}$ kendt struktur, men eventuelt med ukendte parametre.

Marginal fordeling

Marginal fordeling af $\mathbf{Y}$

$$\begin{aligned}\mathsf{E}[\mathbf{Y}] &= \mathbf{X}\boldsymbol{\beta}\\ \mathsf{D}[\mathbf{Y}] &= \mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}'\end{aligned}$$

De hierarkiske modeller som generelle lineære mixede modeller

Uldballer:

$$\mathbf{X}_{28,1} = \begin{pmatrix}1\\1\\\dots\\1\end{pmatrix}, \quad \mathbf{Z}_{28,7} = \begin{pmatrix}1&0&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 1&0&0&\cdots&0\\ 0&1&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&0&\cdots&1\end{pmatrix}$$

$$\mathbf{R} = \sigma^2\mathbf{I}_{28}\,, \quad \mathbf{D} = \sigma_0^2\mathbf{I}_7$$

idet nemlig

$$\mathbf{u} = \begin{pmatrix}u_1\\u_2\\\vdots\\u_7\end{pmatrix}$$

angiver det tilfældige bidrag fra hvert af de syv balleniveauer.

Ramushøjder:

$$\mathbf{X}_{20,2} = \begin{pmatrix}1&8.0\\1&8.5\\1&9.0\\1&9.5\\1&8.0\\\vdots&\vdots\\1&9.5\end{pmatrix}, \quad \mathbf{Z}_{20,5} = \begin{pmatrix}1&8.0&0&0&\cdots&0\\1&8.5&0&0&\cdots&0\\1&9.0&0&0&\cdots&0\\1&9.5&0&0&\cdots&0\\0&0&1&8.0&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&9.5\end{pmatrix}$$

$$\mathbf{R} = \sigma^2\mathbf{I}_{20}\,, \quad \mathbf{D} = \boldsymbol{\Lambda}_{2,2}\otimes\mathbf{I}_5 = \text{ Blok diag}\boldsymbol{\Lambda}_{2,2}$$

idet nemlig

$$\mathbf{u} = \begin{pmatrix}u_1\\u_2\\\vdots\\u_{10}\end{pmatrix}$$

angiver det tilfældige bidrag til intercept, $u_{2\nu-1}$, og hældning, $u_{2\nu}$, fra hver af de fem drenge, $\nu = 1, 2, \dots, 5$.

Generel lineær mixed model

Er en meget rig klasse af lineære normalfordelingsmodeller

Forstyrrende tilfældige effekter (blokke i forsøgsplanlægning)
Interessante tilfældige effekter (variation mellem personer)

Kan også løses i ubalancerede situationer.

Generel lineær mixed model, estimation af effekter

Mixed effect ligningerne

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

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

Estimaterne for $\mathbf{u}$ er BLUP-estimater

PROC MIXED

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

Udfører estimation af varians-kovarians matrix for intercept og hældning, samt BLUP-estimation af de tilfældige effekter.

PROC MIXED

SUBJECT = specificerer de uafhængige blokke

TYPE = specificerer strukturen af $\mathbf{R}$ og $\mathbf{D}$

Mange muligheder for matricerne $\mathbf{R}$ og $\mathbf{D}$