Kapitel 5

Normalfordelingsmodeller med hierarkisk variation

Hidtil har vi hovedsageligt kun betragtet modeller med ét niveau af tilfældig variation (gentagelsesvariation). I mange af de situationer, man møder i praksis, vil en sådan modellering imidlertid tydeligt være for stærk en simplificering af den "naturlige variation" i data.

Ofte udtages nogle forsøgsobjekter (personer, produktioner, forsøgsdage, testmateriale etc.) tilfældigt fra en større population af objekter (første trin), og dernæst foretages en række målinger (andet trin) for hvert af disse objekter. Variationen er således *hierarkisk*.

I dette afsnit vil vi diskutere nogle simple normalfordelingsmodeller for hierarkisk variation. Vi vil tage udgangspunkt i en generalisering af envejsmodellen (ensidet variansanalyse) til en model med tilfældig effekt, og beskrive varians- kovariansforhold under denne model (afsnit 5.2.1 - 5.3). Den væsentligste forskel fra den systematiske model er, at under den hierarkiske model er *observationer i samme gruppe korrelerede*.

I afsnit 5.4 betragtes testet for en forsvindende effekt, og i afsnit 5.5 diskuteres estimation af modellens systematiske (faste) parametre, nemlig den fælles middelværdi og de to varianser (hhv variansen indenfor grupper og variansen mellem grupper). I den balancerede situation kan man benytte de simple momentestimater, men hvis data er ubalancerede (forskelligt antal observationer i de forskellige grupper) kompliceres estimationen idet gruppe-

gennemsnittene har forskellige varianser (præcisioner).

I afsnit 5.6 diskuteres estimation (prædiktion) af de tilfældige effekter ved brug af et likelihoodprincip, og begrebet *Bedste lineære centrale prædiktor* (BLUP) indføres.

I afsnit 5.7 genoptages den generelle diskussion af hierarkiske modeller, og begreberne aprioriforde ling og aposteriorifordeling introduceres. Dette begrebsapparat benyttes i afsnit 5.9 til at indføre en såkaldt "empirisk Bayes" løsning til estimation af de tilfældige effekter, og det vises at BLUP-estimatet kan fortolkes som et empirisk Bayesestimat.

Afsnit 5.8 viser eksempler på brug af SAS-procedurer til analyse af envejsmodeller, og endelig giver afsnit 5.10 nogle yderligere simple eksempler på modeller for hierarkisk variation. I kapitel 6 vil vi introducere en generel formulering af lineære normalfordelingsmodeller med hierarkisk variation.

5.1 Betingede, simultane og marginale fordelinger

Vi indleder med en introduktion af nogle begreber fra sandsynlighedsregningen:

Betragt to stokastiske variable X og Y med den *simultane tæthedsfunktion* $f_{xy}(x, y)$. For kontinuerte fordelinger definerer den simultane tæthedsfunktion en *flade* i det tredimensionale rum. Sandsynligheden for at få værdier af X og Y i et område af xy-planen bestemmes som *volumenet* under fladen i det pågældende område. Specielt gælder altså at

$$\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{xy}(x, y) dy dx = 1$$

Tæthederne for de *marginale fordelinger* af X og Y bestemmes ved

$$\begin{aligned} f_X(x) &= \int_{y=-\infty}^{\infty} f_{xy}(x, y) dy \\ f_Y(y) &= \int_{x=-\infty}^{\infty} f_{xy}(x, y) dx \end{aligned}$$

Den marginale fordeling af X angiver sandsynlighederne for at få bestemte værdier af X uden hensyn til værdien af Y (og tilsvarende for den marginale fordeling af Y).

Vi kan nu definere de *betingede fordelinger*.

Den betingede fordeling af Y givet $X = x$ er fordelingen med tæthed

$$h(y|x) = \frac{f_{xy}(x,y)}{f_X(x)} \tag{5.1}$$

Den betingede fordeling udtrykker således de *relative* værdier af $f_{xy}(x,y)$ for fastholdt værdi, x, af første koordinat. Størrelsen $f_X(x)$ i nævneren tjener blot til at sikre at $\int h(y|x)dy = 1$.

Tilsvarende har den betingede fordeling af X givet $Y = y$ tætheden

$$g(x|y) = \frac{f_{xy}(x,y)}{f_Y(y)} \tag{5.2}$$

Af (5.2) fås $f_{xy}(x,y) = f_Y(y)g(x|y)$, der indsat i (5.1) fører til

$$h(y|x) = \frac{g(x|y)f_Y(y)}{f_X(x)} \tag{5.3}$$

hvor

$$f_X(x) = \int_{y=-\infty}^{\infty} g(x|y)f_Y(y)dy$$

I sammenhænge, hvor $f_Y(y)$ udtrykker en såkaldt "subjektiv fordeling" (evt udartet), kaldes resultatet (5.3) for Bayes' sætning, benævnt efter den engelske præst, Thomas Bayes. Hans efternavn udtales [Bæ·js]. I sådanne sammenhænge kaldes fordelingen $f_Y(y)$ af Y for *apriorifordelingen*,og den betingede fordeling med tætheden $h(y|x)$ kaldes *aposteriorifordelingen* efter observation af $X = x$.
Modellen (og sætningen) dukker naturligt op i forbindelse med hierarkiske modeller, hvor den variable X symboliserer en ikke-observerbar tilstand (state) eller parameter, der er knyttet til det udvalgte forsøgsobjekt, mens Y symboliserer de observerede størrelser. I sådanne situationer vil man ofte

kunne beskrive den betingede fordeling af Y for givet tilstand (x), og man vil have observationer af den marginale fordeling af Y. Derimod er det netop ikke muligt at observere X'erne og derfor heller ikke fordelingen $f_X(x)$. I afsnit 5.7 og 5.9 vil vi betragte brugen af (5.3) i sådanne situationer, hvor "apriorifordelingen" $f_Y(y)$ har en frekvensfortolkning.

5.1.1 Lidt om betingede og marginale middelværdier og varianser

Vi minder om formlen, der relaterer betingede og marginale middelværdier og varianser, sætning 1.2.1, side 8:
Lad X, Y og Z være endimensionale stokastiske variable. Da gælder:

$$\begin{aligned} \mathsf{E}[X] &= \mathsf{E}_Y[\,\mathsf{E}[\,X|Y]\,] \\ \mathsf{V}[X] &= \mathsf{E}_Y[\,\mathsf{V}[\,X|Y]\,] + \mathsf{V}_Y[\,\mathsf{E}[\,X|Y]\,] \\ \mathsf{COV}[X,Z] &= \mathsf{E}_Y[\,\mathsf{COV}[X,Z]\,|Y\,] + \mathsf{COV}_Y[\,\mathsf{E}[\,X|Y], \mathsf{E}[\,Z|Y]\,] \end{aligned} \tag{5.4}$$

Der gælder tilsvarende relationer for flerdimensionale fordelinger:

$$\begin{aligned} \mathsf{E}[X] &= \mathsf{E}_Y[\,\mathsf{E}[X|Y]\,] \\ \mathsf{D}[X] &= \mathsf{E}_Y[\,\mathsf{D}[X|Y]\,] + \mathsf{D}_Y[\,\mathsf{E}[X|Y]\,] \\ \mathsf{COV}[X,Z] &= \mathsf{E}_Y[\,\mathsf{COV}[X,Z|Y]\,] + \mathsf{COV}_Y[\,\mathsf{E}[X|Y],\ \mathsf{E}[Z|Y]\,] \end{aligned} \tag{5.5}$$

For senere reference anføres et eksplicit resultat for den flerdimensionale normalfordeling:

Sætning 5.1.1 ***Marginale og betingede fordelinger for flerdimensional normalfordeling***
Betragt den flerdimensionale stokastiske variable $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

og antag, at $\mathbf{X}$ følger en flerdimensional normalfordeling med

$$\mathsf{E}[\mathbf{X}_1] = \boldsymbol{\mu}_1 \quad \text{og } \mathsf{E}[\mathbf{X}_2] = \boldsymbol{\mu}_2$$

og

$$\mathsf{D}\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Da gælder, at de marginale fordelinger af $\mathbf{X}_1$ og $\mathbf{X}_2$ er normalfordelinger,

$$\mathbf{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad \text{og } \mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

og endvidere er den betingede fordeling af $\mathbf{X}_1$ givet $\mathbf{X}_2 = \mathbf{x}_2$ ligeledes en normalfordeling med

$$\begin{aligned} \mathsf{E}[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2] &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \mathsf{D}[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2] &= \boldsymbol{\Sigma}_{11}^{-1} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

hvor $\boldsymbol{\Sigma}_{22}^{-1}$ fortolkes som en generaliseret invers, hvis $\boldsymbol{\Sigma}_{22}$ ikke har fuld rang.

Udtrykket for $\mathsf{E}[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2]$ betegnes undertiden "regressionen af $\mathbf{X}_1$ på $\mathbf{X}_2$.

∇

5.2 Envejsanalyse, tilfældig model

5.2.1 Eksempel

Vi indleder med et eksempel til støtte for den abstrakte gennemgang:

Eksempel 5.2.1 ***Balancerede data, variansanalyseskema***
Data er fra J.M. Cameron: The use of Components of Variance in preparing schedules for sampling of baled wool, [34].
Råuld indeholder varierende mængder af fedt og andre urenheder, der må fjernes før den videre forarbejdning. Prisen - og værdien - af råulden afhænger af den mængde ren uld, der opnås efter omhyggelig rensning. Renheden af råulden udtrykkes som den rene uld i procent af vægten af råulden.
Med henblik på at vurdere forskellige stikprøveplaner til estimation af renheden af et parti bestående af adskillige baller uld har U.S. Customs Laboratory, Boston, blandt andet udtaget 4 prøver tilfældigt fra hver af 7 baller uruguyansk uld.
Prøveresultaterne er angivet i tabel 5.1.

∇

Tabel 5.1: Renheden i % ren uld af 4 prøver fra hver af 7 baller uruguyansk uld.

Prøve	Balle nr.						
	1	2	3	4	5	6	7
1	52.33	56.99	54.64	54.90	59.89	57.76	60.27
2	56.26	58.69	57.48	60.08	57.76	59.68	60.30
3	62.86•	58.20•	59.29•	58.72	60.26	59.58	61.09
4	50.46•	57.35•	57.51•	55.61	57.53	58.08	61.45
Balle gennem-snit	55.48	57.81	57.23	57.33	58.86	58.78	60.78

5.3 Model

Da de enkelte baller blot er tilfældigt udvalgt blandt samlingen af baller, bruger vi en *tilfældig model*:

Vi antager at

$$Y_{ij}|\mu_i \sim \mathrm{N}(\mu_i, \sigma^2) \ , \tag{5.6}$$

hvor vi i modsætning til den *systematiske model*, beskriver gruppeniveauet μ_i som en realisation af en stokastisk variabel.

$$\mu_i \sim \mathrm{N}(\mu_0, \sigma_b^2), \tag{5.7}$$

hvor μ_i'erne antages indbyrdes uafhængige, og Y_{ij} antages at være indbyrdes uafhængige i den betingede fordeling af Y_{ij} for givet μ_i.

I lighed med den systematiske model formuleres modellen ofte som en additiv model for observationerne Y_{ij} ved

$$Y_{ij} = \mu_0 + \alpha_i + \epsilon_{ij} \ , \tag{5.8}$$

med $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ og $\alpha_i \sim \mathrm{N}(0, \sigma_b^2)$, hvor ϵ_{ij} er indbyrdes uafhængige, og α_i ligeledes er indbyrdes uafhængige og endelig er α_i'erne uafhængige af ϵ_{ij}.

Da det ofte vil være af interesse at betragte forholdet mellem de to varianser, indfører vi symbolet γ for dette forhold,

$$\gamma = \sigma_b^2/\sigma^2 \tag{5.9}$$

Parameteren γ udtrykker således inhomogeniteten mellem grupper i forhold til gruppernes interne variation. Vi vil ofte bruge betegnelsen signal/støj forholdet for parameteren γ.
Den tilfældige model vil være rimelig i situationer, hvor interessen ikke er indskrænket til alene de betragtede forsøgsomstændigheder (grupper), men hvor disse omstændigheder snarere opfattes som repræsentative for en større samling (population) af varierende forsøgsomstændigheder, principielt udtaget tilfældigt fra denne population.

Vi siger ofte, at vi har hierarkisk variation, eller at vi har en hierarkisk model, svarende til at vi har en tilfældig variation mellem grupper (underordnet den konstante faktor), og gentagelsen inden for grupper er underordnet inddelingen i grupper.
Til illustration af forskellen mellem den tilfældige og den systematiske model bemærker vi yderligere, at analysen af den systematiske model lægger vægt på vurderingen af resultaterne i de enkelte grupper, μ_i, og eventuelle forskelle, $\mu_i - \mu_h$, på resultaterne i specifikke grupper, mens analysen af den tilfældige model i første række sigter mod at beskrive variationen mellem grupperne, $\mathsf{V}[\mu_i] = \sigma_b^2$.

5.3.1 Marginale og simultane fordelinger

Sætning 5.3.1 ***Marginale fordelinger i den tilfældige model for ensidet variansanalyse*** Den marginale fordeling af Y_{ij} er en normal fordeling med

$$\begin{aligned} \mathsf{E}[Y_{ij}] &= \mu_0 \\ \mathsf{COV}[Y_{ij}, Y_{hl}] &= \begin{cases} \sigma_b^2 + \sigma^2 & \text{for } (i,j) = (h,l) \\ \sigma_b^2 & \text{for } i = h, \;\; j \neq l \\ 0 & \text{for } i \neq h \end{cases} \end{aligned} \tag{5.10}$$

∇

Bemærkning 1 ***Observationer fra samme gruppe er korrelerede***
Vi bemærker, at der en positiv kovarians mellem observationer fra samme gruppe. Denne positive kovarians udtrykker netop, at observationer inden for en gruppe vil afvige i samme retning fra den marginale middelværdi μ_0, nemlig i retning mod den pågældende gruppemiddelværdi.

Korrelationskoefficienten

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} = \frac{\gamma}{1+\gamma} \tag{5.11}$$

der beskriver korrelationen indenfor gruppe, benævnes ofte intraklassekorrelationen. ∇

Bemærkning 2 ***Korrelations- og dispersionsmatrix for observationer fra samme gruppe***
Betragter vi observationssættet svarende til den i'te gruppe som en n_i-dimensional søjlevektor,

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

har vi, at korrelationsmatricen i den marginale fordeling af Y_i er en *equikorrelationsmatrix* (se Introduktion til fordelinger med anvendelse i statistik, afsnit 0.2.1) af formen

$$\mathbf{E}_{n_i} = (1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i} = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \tag{5.12}$$

hvor $\mathbf{J}_{n_i}$ er en $n_i \times n_i$-dimensional matrix bestående af lutter ettaller
Observationssættene $Y_i,\ i = 1, 2, \dots, k$ kan således beskrives som k uafhængige observationer af en n_i dimensional variabel $Y_i \sim \mathrm{N}_{n_i}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_{n_i} + \sigma_b^2\mathbf{J}_{n_i})$,

dvs at dispersionsmatricen for Y_i er

$$\begin{aligned}\mathbf{V}_i &= \mathsf{D}[Y_i] = \mathsf{E}[(Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})'] \\ &= \begin{pmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma^2 \end{pmatrix} \end{aligned} \tag{5.13}$$

En sådan matrix betegnes undertiden en sammensat symmetrisk matrix (eng. compound symmetric). ∇

Bemærkning 3 ***Kovariansstruktur for hele observationssættet***
Opstiller vi samtlige observationer i én søjle, organiseret efter grupper, ser vi, at den $N \times N$-dimensionale dispersionsmatrix $\mathsf{D}[\mathbf{Y}]$ er

$$\mathbf{V} = \mathsf{D}[\mathbf{Y}] = \text{ Blok diag}\{\mathbf{V}_i\} \tag{5.14}$$

hvor $\mathbf{V}_i$ er givet ved (5.13)
Tilsvarende finder man, at korrelationsmatricen for hele observationssættet er en $N \times N$-dimensional blokmatrix med matricerne $\mathbf{E}_{n_i}$ i diagonalen, og nuller udenfor, hvilket illustrerer, at observationer fra forskellige grupper er uafhængige, mens observationer fra samme gruppe er korrelerede. ∇

Bemærkning 4 ***Simultan fordeling af gruppegennemsnittene***
Vi bemærker endelig, at den simultane fordeling af gruppegennemsnittene er karakteriseret ved

$$\mathsf{COV}[\overline{Y}_{i\cdot}, \overline{Y}_{h\cdot}] = \begin{cases} \sigma_b^2 + \sigma^2/n_i & \text{for } i = h \\ 0 & \text{ellers} \end{cases} \tag{5.15}$$

Det vil sige, at de k gruppegennemsnit $\overline{Y}_{i\cdot}$, $i = 1, 2, \dots, k$ er indbyrdes uafhængige, og variansen på gruppegennemsnittet,

$$\mathsf{V}[\overline{Y}_{i\cdot}] = \sigma_b^2 + \sigma^2/n_i = \sigma^2(\gamma + 1/n_i)$$

omfatter både variansen på den tilfældige komponent, γ_i, og residualvariansen på gennemsnittet.

En forøgelse af stikprøvestørrelsen i de enkelte grupper vil således forøge præcisionen ved bestemmelse af gruppeforventningsværdien α_i, men variationen mellem de enkelte gruppeforventningsværdier formindskes naturligvis ikke ved denne gennemsnitsdannelse.

∇

5.4 Test af homogenitetshypotese

Vi sætter som sædvanlig

$$\text{sak}_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i))^2 \tag{5.16}$$

$$\overline{\overline{y}} = \sum_i n_i \overline{y}_{i\cdot} \Big/ N \tag{5.17}$$

$$\text{sak}_2 = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{\overline{y}})^2 \tag{5.18}$$

$$\tag{5.19}$$

Endvidere indfører vi

$$w_i(\gamma) = \frac{1}{1 + n_i\gamma} \tag{5.20}$$

(betydningen af $w_i(\gamma)$ fremgår af at $\mathsf{V}[\overline{Y}_{i\cdot}] = \sigma^2/w_i(\gamma)$).

Den hypotese, at de varierende forsøgsomstændigheder er uden påviselig indflydelse på observationerne, formuleres under den tilfældige model som

$$\mathrm{H}_{II} : \sigma_b^2 = 0. \tag{5.21}$$

Hypotesen testes ved at sammenligne varianskvotienten

$$Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \tag{5.22}$$

med fraktilerne i en $\mathrm{F}(k-1, N-k)$-fordeling. Der gælder

Sætning 5.4.1 ***Test af homogenitetshypotese i den tilfældige model***

Under modellen givet ved (5.6) og (5.7) har kvotienttestet for hypotesen (5.21) det kritiske område

$$C = \{z | z > \mathrm{F}(k-1, N-k)_{1-\alpha}\}$$

hvor z er givet ved (5.22)

I det balancerede tilfælde, $n_1 = n_2 = \ldots = n_k = n$, gælder

$$Z \sim (1 + n\gamma)\mathrm{F}(k-1, N-k) \tag{5.23}$$

∇

Bemærkning 1 ***Konfidensinterval for varianskvotienten***
I det balancerede tilfælde, $n_1 = n_2 = \ldots = n_k = n$, kan man benytte (5.23) til at konstruere et konfidensinterval for varianskvotienten γ. Ved benyttelse af (5.23) finder man, at et $1-\alpha$ konfidensinterval for γ, dvs. et interval (γ_L, γ_U), der tilfredsstiller

$$\mathsf{P}[\gamma_L < \gamma < \gamma_U] = 1 - \alpha$$

fås ved at benytte

$$\begin{aligned} \gamma_L &= \frac{1}{n}\left(\frac{Z}{\mathrm{F}(k-1, N-k)_{1-\alpha/2}} - 1\right) \\ \text{og} & \\ \gamma_U &= \frac{1}{n}\left(\frac{Z}{\mathrm{F}(k-1, N-k)_{\alpha/2}} - 1\right) \end{aligned} \tag{5.24}$$

hvor Z er givet ved (5.22).

∇

Eksempel 5.4.1 ***Balancerede data, variansanalyseskema***
Vi udfører nu beregninger til en ensidet variansanalyse for uldballeeksemplet, og får variansanalyseskemaet

Variation	SAK	f	$s^2 = SAK/f$	$\mathsf{E}[S^2]$
Mellem baller	65.9628	6	10.9938	$\sigma^2 + 4\sigma_b^2$
Indenfor baller	131.4726	21	6.2606	σ^2

Udfører vi et test for hypotesen $\mathrm{H}_{II} : \sigma_b^2 = 0$, finder vi teststørrelsen

$$z = \frac{10.9938}{6.2606} = 1.76 < \mathrm{F}_{0.95}(6, 21) = 2.57$$

Det kan således ikke afvises - ved test på et 5 % niveau - at variationen mellem renheden af ballerne ikke overstiger den interne variation i ballernes renhed.

Ønsker vi et 95 % konfidensinterval for varianskvotienten $\gamma = \sigma_b^2/\sigma^2$ finder vi ved benyttelse af (5.24), at intervallet er bestemt som

$$\begin{aligned} \gamma_L &= \frac{1}{4}\left(\frac{1.76}{\mathrm{F}(6,21)_{0.975}} - 1\right) = 0.25 \times \left(\frac{1.76}{3.09} - 1\right) = -0.11 \\ \gamma_U &= \frac{1}{4}\left(\frac{1.76}{\mathrm{F}(6,21)_{0.025}} - 1\right) 0.25 \times (1.76 \times 6.31 - 1) = 2.53\ , \end{aligned}$$

idet $\mathrm{F}(6,21)_{0.025} = 1/\mathrm{F}(21,6)_{0.975}$ ∇

5.5 Estimation af faste parametre

Sætning 5.5.1 ***Maksimum-likelihood estimater for parametrene under den tilfældige model***
Under modellen givet ved (5.6) og (5.7) er maximum-likelihood estimaterne for μ, σ^2 og $\sigma_0^2 = \sigma^2\gamma$ bestemt ved
a) For $\sum_i n_i^2(\overline{y}_{i\cdot} - \overline{\overline{y}}_{\cdot\cdot})^2 < sak_1 + sak_2$ fås

$$\begin{aligned} \widehat{\mu} &= \overline{\overline{y}}_{\cdot\cdot} = \sum_i n_i \overline{y}_{i\cdot} \Big/ N \\ \widehat{\sigma}^2 &= (sak_1 + sak_2)/N \qquad (5.25) \\ \text{og} & \\ \widehat{\gamma} &= 0 \qquad (5.26) \end{aligned}$$

b) For $\sum_i n_i^2(\overline{y}_{i\cdot} - \overline{\overline{y}}_{\cdot\cdot})^2 > sak_1 + sak_2$ bestemmes estimaterne som løsning til

$$\widehat{\mu} = \sum_{i=1}^{k} n_i w_i(\widehat{\gamma}) \overline{y}_{i\cdot} \Big/ W(\widehat{\gamma}) \qquad (5.27)$$

$$\widehat{\sigma}^2 = \frac{1}{N}\{sak_1 + \sum_{i=1}^{k} n_i w_i(\widehat{\gamma})(\overline{y}_{i\cdot} - \mu^*)^2\} \tag{5.28}$$

$$\sum_{i=1}^{k} n_i^2 w_i(\widehat{\gamma})^2(\overline{y}_{i\cdot} - \widehat{\mu})^2 \Big/ W(\widehat{\gamma}) \tag{5.29}$$

$$= \Big\{sak_1 + \sum_{i=1}^{k} n_i w_i(\widehat{\gamma})(\overline{y}_{i\cdot} - \widehat{\mu})^2\Big\}\Big/ N$$

hvor $w_i(\gamma)$ er givet ved (5.20) og

$$W(\gamma) = \sum_{i=1}^{k} n_i w_i(\gamma)\ . \tag{5.30}$$

Løsningen til (5.27) til (5.29) bestemmes ved iteration.

Bevis:

Resultatet følger ved at betragte logaritmen til likelihoodfunktionen baseret på den marginale fordeling af observationssættet. Denne bliver - på nær en additiv konstant -

$$\begin{aligned} l(\mu, \sigma^2, \gamma) \;\; = \;\; & -\frac{sak_1}{2\sigma^2} \; - \; \frac{1}{2\sigma^2}\sum_{i=1}^{k} n_i w_i(\gamma)(\overline{y}_{i\cdot} - \mu)^2 \\ & -\frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{k}\log(w_i(\gamma)) \end{aligned} \tag{5.31}$$

∇

Bemærkning 1 ***Maksimum-likelihood estimatet $\widehat{\mu}$ er et vejet gennemsnit af gruppegennemsnittene***

Vi ser af (5.27), at $\widehat{\mu}$ er et vejet gennemsnit af gruppegennemsnittene, $\overline{y}_{i\cdot}$, med estimaterne for de marginale præcisioner

$$\sigma^2 n_i w_i(\gamma) = \sigma^2/\mathsf{V}[\overline{Y}_{i\cdot}]$$

som vægte. Der gælder nemlig

$$\mathsf{V}[\overline{Y}_{i\cdot}] = \sigma_0^2 + \sigma^2/n_i = \frac{\sigma^2}{n_i}(1 + n_i\gamma) = \sigma^2/\{n_i w_i(\gamma)\}$$

Hvis eksperimentet er balanceret, d.v.s. $n_1 = n_2 = \cdots = n_k$, bliver alle vægtene ens, og man får det simple resultat, at $\widehat{\mu}$ blot er det simple gennemsnit af gruppegennemsnittene. ∇

Bemærkning 2 ***Estimatet for σ^2 udnytter også variationen mellem grupper***
Vi bemærker, at estimatet for σ^2 ikke kun er baseret på variationen indenfor grupper, sak_1 , men estimatet udnytter desuden kendskabet til variationen imellem grupper, idet nemlig

$$\mathsf{E}[(\overline{Y}_{i\cdot} - \mu)^2] = \mathsf{V}[\overline{Y}_{i\cdot}] = \sigma^2/\{n_i w_i(\gamma)\}$$

hvorfor leddene $(\overline{y}_{i\cdot} - \mu)^2$ indeholder information, såvel om σ^2, som information om γ. ∇

Bemærkning 3 ***Estimatet for σ^2 er ikke nødvendigvis centralt***
Vi bemærker endvidere, at - som vanligt ved ML-estimatet- er estimatet for σ^2 ikke nødvendigvis centralt.
I stedet for maksimum-likelihood estimatet benytter man derfor undertiden et estimat baseret på likelihoodfunktionen svarende til fordelingen af residualerne, det såkaldte REML-estimat, svarende til at man benytter profillikelihood'en (efter bortmaximering af μ_0) og korrigerer profillikelihood'en med den observerede Fisher information for μ_0. Se fx. Pawitan [24] side 440 ff. Da estimatet således fås ved at betragte fordelingen af *residualerne* kaldes det et ***RE****sidual* ***M****aximum* ***L****ikelihood*-estimat. Nogle forfattere bruger betegnelsen *Restricted Maximum Likelihood*-estimat. ∇

Sætning 5.5.2 ***Momentestimater (variansanalyseestimater) i den tilfældige model***
Under modellen givet ved (5.6) og (5.7) finder man momentestimaterne for parametrene μ_0, σ^2 og σ_0^2 ved

$$
\begin{aligned}
\widetilde{\mu}_0 &= \overline{\overline{Y}}_{..} \\
\widetilde{\sigma^2} &= SAK_1/(N-k) \qquad (5.32) \\
\widetilde{\sigma_0^2} &= \frac{SAK_2/(k-1) - SAK_1/(N-k)}{n_0} = \frac{SAK_2/(k-1) - \widetilde{\sigma^2}}{n_0}
\end{aligned}
$$

hvor den vægtede gennemsnitlige gruppestørrelse n_0 er givet ved

$$
n_0 = \frac{\sum_1^k n_i - (\sum_1^k n_i^2 / \sum_1^k n_i)}{k-1} = \left(N - \frac{\sum_i n_i^2}{N}\right) \Big/ (k-1) \qquad (5.33)
$$

Bevis:

$$
\begin{aligned}
\mathsf{E}[SAK_1/(N-k)] &= \sigma^2 \\
&\text{og} \\
\mathsf{E}[SAK_2/(k-1)] &= \sigma^2 + n_0\sigma_0^2
\end{aligned}
$$

∇

Bemærkning 1 ***Centrale estimatorer for varianskvotient i balanceret tilfælde***
I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, kan vi angive eksplicitte centrale estimatorer for γ og $w(\gamma) = 1/(1+n\gamma)$. Der gælder

$$
\begin{aligned}
\widetilde{w} &= \frac{SAK_1}{k(n-1)} \Big/ \frac{SAK_2}{k-3} \\
&\text{og} \qquad (5.34) \\
\widetilde{\gamma} &= \frac{1}{n} \left\{ \frac{SAK_2}{k-1} \Big/ \frac{SAK_1}{k(n-1)-2} - 1 \right\} \qquad (5.35)
\end{aligned}
$$

er centrale estimatorer for henholdsvis $w(\gamma) = 1/(1+n\gamma)$ og for $\gamma = \sigma_b^2/\sigma^2$

Bevis:
Følger ved at bemærke, at $SAK_1 \sim \sigma^2\chi^2(k(n-1))$ og
$SAK_2 \sim \{\sigma^2/w(\gamma)\}\chi^2(k-1)$ er indbyrdes uafhængige med

$$\begin{aligned}
\mathsf{E}[SAK_1/\{k(n-1)\}] &= \sigma^2 \\
\mathsf{E}[\{k(n-1)-2\}/SAK_1] &= 1/\sigma^2 \\
\mathsf{E}[SAK_2/(k-1)] &= \sigma^2/w(\gamma) \\
\mathsf{E}[(k-3)/SAK_2] &= w(\gamma)/\sigma^2
\end{aligned}$$

∇

Bemærkning 2 ***Maximum-likelihood-estimater i det balancerede tilfælde***
I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k$ afhænger vægtene

$$w_i(\gamma) = \frac{1}{1+n\gamma}$$

ikke af i, og (5.27) bliver

$$\widehat{\mu} = \sum_{i=1}^{k} \overline{y}_{i+}/k = \overline{\overline{y}}_{++} \ ,$$

altså netop momentestimatet.
Hvis $(n-1)sak_2 > sak_1$ svarer maksimum-likelihood estimatoren til et indre punkt i parameterrummet.
Ligningerne (5.28) og (5.29) til bestemmelse af σ^2 og $\gamma = \sigma_0^2/\sigma^2$ bliver

$$\begin{aligned}
N\sigma^2 &= sak_1 + \frac{1}{1+n\gamma}\, sak_2 \\
N\,\frac{n}{1+n\gamma}\,\frac{sak_2}{k} &= sak_1 + \frac{1}{1+n\gamma}\, sak_2
\end{aligned}$$

med løsningen

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{sak_1}{N-k} \\
\widehat{\gamma} &= \frac{1}{n}\left[\frac{sak_2}{k\widehat{\sigma}^2} - 1\right]
\end{aligned}$$

dvs.

$$\widehat{\sigma}_0^2 = \frac{sak_2/k - \widehat{\sigma}^2}{n}$$

∇

5.6 Estimation af de tilfældige effekter, BLUP-estimation

Det kan undertiden være af interesse at bestemme et estimat for de tilfældige effekter (ballemiddelværdierne).

Hvis vi også opfatter de tilfældige effekter som parametre, kan vi opskrive likelihood'en svarende til alle parametrene som

$$L(\mu_0, \sigma^2, \gamma, \boldsymbol{\mu}) = f(\mathbf{y}|\boldsymbol{\mu}) f(\boldsymbol{\mu})$$

$$L = \frac{1}{\sigma^N} \exp\left\{ - \frac{sak_1 + \sum_i (\overline{y}_i - \mu_i)^2}{2\sigma^2} \right\} \frac{1}{(\gamma\sigma^2)^{k2/2}} \exp\left\{ - \frac{\sum_i (\mu_i - \mu_0)^2}{2\gamma\sigma^2} \right\}$$

hvorfor vi får log-likelihood'en

$$\ell(\mu_0, \sigma^2, \gamma, \boldsymbol{\mu}) = -(N/2)\log(\sigma^2) - \frac{sak_1 + \sum_i (\overline{y}_i - \mu_i)^2}{2\sigma^2} - (k/2)\log(\gamma\sigma^2) - \frac{\sum_i (\mu_i - \mu_0)^2}{2\gamma\sigma^2}$$

Differentieres med hensyn til μ_i finder vi

$$\frac{\partial}{\partial \mu_i} \ell(\mu_0, \sigma^2, \gamma, \boldsymbol{\mu}) = \frac{n_i(\overline{y}_i - \mu_i)}{\sigma^2} - \frac{\mu_i - \mu}{\gamma\sigma^2} \quad ,$$

der er nul for

$$\mu_i \left(\frac{n_i}{\sigma^2} + \frac{1}{\gamma\sigma^2} \right) = \frac{n_i}{\sigma^2} \overline{y}_i + \frac{1}{\gamma\sigma^2} \mu_0 \tag{5.36}$$

Ligningerne skal løses sammen med (5.27) til (5.29) ved iteration.

Det ses, at løsningen tilfredsstiller

$$\mu_i \frac{n_i\gamma + 1}{\gamma\sigma^2} = \frac{n_i\gamma\overline{y}_i + \mu_0}{\gamma\sigma^2}$$

dvs

$$\mu_i = \frac{n_i\gamma}{n_i\gamma + 1} \overline{y}_i + \frac{1}{n_i\gamma + 1} \mu_0 \tag{5.37}$$

altså

$$\mu_i = \Big(1 - w_i(\gamma)\Big) \overline{y}_i + w_i(\gamma)\mu_0$$

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

$$w_i(\gamma) = \frac{1}{1 + n_i\gamma} \tag{5.38}$$

Løsningen kaldes den bedste lineære centrale prædiktor, BLUP (best linear predictor).
For en anvendelse af dette estimat og en populær forklaring se fx
http://swbzone.com/blup/blup2002.htm

Betegnelsen "Best linear unbiased" skyldes at estimatoren minimerer

$$\mathsf{E}\big[(\widehat{B}_i - \mathsf{E}[B_i])^2\big]$$

blandt alle lineære (i observationerne) estimatorer, $\widehat{B}_i$, der er centrale (unbiased) for $\mathsf{E}[B_i]$, dvs som opfylder $\mathsf{E}[\widehat{B}_i] = \mathsf{E}[B_i]$.

5.7 Lidt mere om hierarkiske modeller

5.7.1 Generelt

Definition 5.7.1 ***Apriori- aposteriorifordeling***

Betragt en statistisk model for observationssættet $X_1, X_2, \ldots, X_n$ givet ved familien af tætheder

$$\{f(x_1, x_2, \ldots, x_n|\theta)\}_{\theta\sim\Theta} \tag{5.39}$$

hvor $\Theta \subset \mathsf{R}^k$.

Antag at fordelingen af $X_1, X_2, \ldots, X_n$ for en fastholdt værdi af θ har tætheden

$$f(x_1, x_2, \ldots, x_n|\theta)$$

og antag yderligere, at parameteren θ kan opfattes som en stokastisk variabel med tætheden

$$w(\theta)$$

Undertiden kaldes fordelingen af θ med tætheden $w(\cdot)$ for strukturfordelingen, eller evt. apriorifordelingen af θ.

Den betingede fordeling af θ givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ kaldes aposteriorifordelingen af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Fordelingen kan bestemmes ved Bayes' sætning (5.3). ∇

I mange situationer vil det være sådan, at at $X_1, X_2, \ldots, X_n$ er betinget uafhængige af θ i den simultane fordeling af $X_1, X_2, \ldots, X_n$ og θ. I de simpleste af disse vil det yderligere være sådan, at for fastholdt værdi af θ vil $X_1, X_2, \ldots, X_n$ være uafhængige identisk fordelte gentagelser. I sådanne tilfælde gælder at den simultane fordeling af $X_1, X_2, \ldots, X_n$ for fastholdt θ har tætheden

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} g(x_i|\theta) \tag{5.40}$$

hvor $g(\cdot|\theta)$ angiver tætheden for den i'te observation.

Sådanne tilfælde kan f.eks. være situationer, hvor parameteren θ varierer i en population i overensstemmelse med aprioritætheden $w(\theta)$, og hvor en måling, X af θ er behæftet med en målefejl svarende til fordelingen $g(x|\theta)$. Modellen (5.40) svarer da til at der foretages n uafhængige målinger $X_1, X_2, \ldots, X_n$ af en forelagt population.

Modellen (5.40) kan også beskrive en situation, hvor en proces genererer en serie af produktenheder hvor den i'te produktenhed karakteriseres ved X_i. De enkelte produktenheder genereres uafhængigt af hinanden i overensstemmelse med en fordeling med tæthed $g(x|\theta)$, hvor parameteren θ varierer fra produktion til produktion i overensstemmelse med aprioritætheden $w(\theta)$.

I begge disse situationer kan det være af interesse - ikke blot at beskrive aposteriorifordelingen af θ efter observation af en stikprøve $x_1, x_2, \ldots, x_n$, men også at beskrive fordelingen af fremtidige observationer fra den betragtede population.

Definition 5.7.2 ***Prædiktiv fordeling***
Lad den simultane fordeling af θ, $X_1, X_2, \ldots, X_n$ og $X'_1, X'_2, \ldots, X'_r$ være sådan, at for givet θ er

$$X_1, X_2, \ldots, X_n, X'_1, X'_2, \ldots, X'_r$$

er indbyrdes uafhængige og identisk fordelte med en tæthed $g(\cdot|\theta)$, dvs. de variable $X'_1, X'_2, \ldots, X'_r$ er frembragt af det samme θ som de variable $X_1, X_2, \ldots, X_n$.

Antag at θ er en stokastisk variabel med tætheden $w(\theta)$.
Den prædiktive fordeling af X'_j for givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, er da fordelingen med tæthed

$$g_1(x'|x_1, x_2, \ldots, x_n) = \int_\Theta g(x'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \nu\{d\theta\} \tag{5.41}$$

hvor $w_1(\theta|x_1, x_2, \ldots, x_n)$ angiver aposterioritætheden af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.
Lad $Z' = \sum_{j=1}^r X'_j$, og lad den betingede fordeling af Z' for fastholdt θ have tætheden $g^{(r)}(z'|\theta)$. Den prædiktive fordeling af $Z' = \sum_{j=1}^r X'_j$ for givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ er da fordelingen med tæthed

$$g_1^{(r)}(z'|x_1, x_2, \ldots, x_n) = \int_\Theta g^{(r)}(z'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) d\nu\{\theta\} \tag{5.42}$$

hvor $w_1(\theta|x_1, x_2, \ldots, x_n)$ som før angiver aposterioritætheden af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.
På tilsvarende måde defineres den prædiktive fordeling af $Y' = \overline{X'}_\cdot = Z'/r$ for givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ som fordelingen med tæthed

$$g_1^{[r]}(y'|x_1, x_2, \ldots, x_n) = \int_\Theta g^{[r]}(y'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \nu\{d\theta\} \tag{5.43}$$

hvor $g^{[r]}(y'|\theta)$ angiver tætheden i den den betingede fordeling af $Y' = \sum_{j=1}^r X'_j/r$ for fastholdt θ.

∇

Den prædiktive fordeling af X' for givet $x_1, x_2, \ldots, x_n$ er således den marginale fordeling af X' svarende til at man bruger aposteriorifordelingen af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. som strukturfordeling ("apriorifordeling") af θ .

5.7.2 Aposteriorifordeling svarende til normal-normal fordeling

For modellen (5.6) kan man vise, at hvis vi kender μ_0, σ^2 og γ, så er den betingede fordeling af $\mu_i = \mu_0 + u_i$ efter observation af gennemsnittet $\bar{y}_i$ fra

den i'te gruppe en normalfordeling (*aposteriorifordelingen*) med middelværdi

$$\mathsf{E}[\mu_i \mid \overline{Y}_i = \bar{y}_i] = \frac{\mu_0/\sigma_b^2 + n_i\bar{y}_i/\sigma^2}{1/\sigma_b^2 + n_i/\sigma^2}$$

altså et vejet gennemsnit af apriorimiddelværdi μ_0 og stikprøvegennemsnit, med de tilsvarende *præcisioner* (reciprokke varianser) som vægte.

Det kan også skrives:

$$\mathsf{E}[\mu_i \mid \overline{Y}_i = \bar{y}_i] = \frac{\mu_0/\gamma + n_i\bar{y}_i}{1/\gamma + n_i} = w_n m + (1 - w_n)\bar{x}$$

med $w = 1/(1 + n\gamma)$.

Variansen i aposteorifordelingen er

$$\sigma^2_{apost} = \mathsf{V}[\mu_i \mid \overline{Y}_i = \bar{y}_i] = \frac{1}{\dfrac{1}{\sigma_b^2} + \dfrac{n_i}{\sigma^2}} =$$

Og den er meget nemmere at gennemskue, hvis man skriver den som præcisionen

$$\frac{1}{\sigma^2_{apost}} = \frac{1}{\sigma_b^2} + \frac{n_i}{\sigma^2}$$

Altså:
aposterioripræcision er summen af aprioripræcision og stikprøvepræcision

Eller, udtrykt ved signal/støjforhold, γ, hvor $\gamma_{apriori} = \sigma_b^2/\sigma^2$ og $\gamma_{aposteriori} = \sigma^2_{aposteriori}/\sigma^2$

$$\frac{1}{\gamma_{aposteriori}} = \frac{1}{\gamma_{apriori}} + n_i$$

og

$$\mu_{aposteori} = w_n\mu_{apriori} + (1 - w_n)\bar{y}_i \tag{5.44}$$

med

$$w_n = \frac{1}{1 + n\gamma_{apriori}}$$

analogt til BLUP-estimatet (5.37).

Bemærkning 1 ***Uddybende bemærkninger***
På grund af normalfordelingens udbredte anvendelse, vil vi uddybe fortolkningen af aposteriorifordelingen og den prædiktive fordeling.

Vi lader

- $x_1, x_2, \ldots, x_n$ betegne stikprøveresultatet fra en given gruppe
- (μ, σ^2) betegne middelværdi og varians for observationer fra den pågældende gruppe
- $X'_1, X'_2, \ldots, X'_r$ betegne observationer fra en fremtidig stikprøve fra denne gruppe

Vi antager, at $X_1, X_2, \ldots, X_n$ og $X'_1, X'_2, \ldots, X'_r$ for fastholdt μ er indbyrdes uafhængige $\mathrm{N}(\mu, \sigma^2)$-fordelt med samme μ, og at $\mu \sim \mathrm{N}(m, \sigma_b^2)$ med $\sigma_b^2 = \gamma\sigma^2$.
Samvariationen mellem X_i ***og*** X_j
Vi minder om, at der gælder (Sætning 5.3.1)

$$\begin{aligned} \mathsf{COV}[X_i, X_j] &= \mathsf{E}[\mathsf{COV}[X_i, X_j \mid \mu]] + \mathsf{COV}[\mathsf{E}[X_i \mid \mu], \mathsf{E}[X_j \mid \mu]] \\ &= 0 + \mathsf{V}[\mu] = \sigma_b^2 \end{aligned}$$

hvorfor vi har intraklyngekorrelationen

$$\rho_{X_i, X_j} = \frac{1}{1 + \sigma^2/\sigma_b^2} = \frac{\gamma}{1+\gamma}$$

Samvariationen mellem $\overline{X}$ ***og*** μ
Idet $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, har vi den betingede forventningsværdi og varians

$$\mathsf{E}[\overline{X} \mid \mu] = \mu; \quad \mathsf{V}[\overline{X} \mid \mu] = \sigma^2/n$$

Den marginale forventningsværdi og varians er

$$\mathsf{E}[\overline{X}] = \mu_0; \quad \mathsf{V}[\overline{X}] = \sigma_b^2 + \sigma^2/n$$

og kovariansen mellem $\overline{X}$ og μ er

$$\mathsf{COV}[\overline{X}, \mu] = \mathsf{E}[\mathsf{COV}[\overline{X}, \mu] \mid \mu] + \mathsf{COV}[\mathsf{E}[\overline{X} \mid \mu], \mu] = 0 + \mathsf{COV}[\mu, \mu] = \sigma_b^2$$

hvorfor vi har korrelationskoefficienten mellem stikprøvegennemsnit, $\overline{X}$ og gruppeparameter, μ

$$\rho_{\overline{X}, \mu} = \frac{1}{\sqrt{1 + \sigma^2/(n\sigma_b^2)}} = \frac{1}{\sqrt{1 + 1/(n\gamma)}} = \sqrt{1 - w_n}$$

med $w_n = 1/(1+n\gamma)$
For $n \to \infty$ vil $\rho_{\overline{X},\mu} \to 1$.
Aposteriorifordelingen af μ givet $\overline{X} = \bar{x}$ karakteriseres ved regressionen

$$\begin{aligned} \mathsf{E}[\mu \mid \overline{X} = \bar{x}] &= m + (1 - w_n)(\bar{x} - m) \\ \mathsf{V}[\mu \mid \overline{X} = \bar{x}] &= \frac{1}{1/\sigma_b^2 + n/\sigma^2} = \frac{\sigma^2}{n}(1 - w_n) \end{aligned}$$

For $n \to \infty$ vil $\mathsf{E}[\mu \mid \overline{X} = \bar{x}] \to \bar{x}$,, og $\mathsf{V}[\mu \mid \overline{X} = \bar{x}] \to 0$.
Samvariationen mellem $\overline{X}$ og $\overline{X'}$

I modellen symboliserer $\overline{X'}$ gennemsnittet af r fremtidige observationer, hidrørende fra den samme gruppe, som de allerede kendte værdier af X (d.v.s. med den samme ukendte værdi af μ). Efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ beskrives vores viden om μ ved aposteriorifordelingen af μ som beskrevet ovenfor, hvorfor den prædiktive fordeling af totalen $Z'_+ = X'_1 + X'_2 + \cdots + Z'_r$ er en $\mathrm{N}(rm_1, r\sigma^2 + r^2\sigma_1^2)$-fordeling, nemlig den marginale fordeling af summen af r observationer under hensyntagen til den opdaterede viden om μ.
Korrelationen mellem $\overline{X}$ og $\overline{X'}$ i den simultane fordeling af $\overline{X}$ og $\overline{X'}$ er

$$\begin{aligned} \rho_{\overline{X},\overline{X'}} &= \frac{1}{\sqrt{1 + \sigma^2/(n\sigma_b^2)}\sqrt{1 + \sigma^2/(r\sigma_b^2)}} = \frac{1}{\sqrt{1 + 1/(n\gamma)}\sqrt{1 + 1/(r\gamma)}} \\ &= \sqrt{1 - wn}\sqrt{1 - w_r} \end{aligned}$$

Den prædiktive fordeling af $\overline{X'}$ svarende til $\overline{X} = \bar{x}$ kan da fortolkes som regressionen af $\overline{X'}$ på $\bar{x}$ i den simultane fordeling af $\overline{X}$ og $\overline{Y}$ (hvor vi har bortintegreret μ).

$$\begin{aligned} \mathsf{E}[\overline{X'} \mid \overline{X} = \bar{x}] &= m + (1 - w_n)(\bar{x} - m) \\ \mathsf{V}[\overline{X'} \mid \overline{X} = \bar{x}] &= \frac{\sigma^2}{p} + \frac{1}{1/\sigma_b^2 + n/\sigma^2} = \frac{\sigma^2}{p} + \frac{\sigma^2}{n}(1 - w_n) \end{aligned}$$

∇

5.8 I SAS

SAS-INSIGHT kan ikke klare tilfældige modeller, så vi må i gang med procedurerne.

Ved GLM

Eksempel 5.8.1 ***Balancerede data, beregning ved SAS-proceduren GLM***

Betragt situationen i eksempel 5.2.1 og antag at data er indlæst i de variable renh, balle og prove, ialt 28 observationer. SAS programmet

```
PROC GLM ;
CLASS balle prove ;
MODEL renh = balle ;
RANDOM balle ;
RUN;
```

definerer de variable balle og prove som klassifikationsvariable, og angiver modelformlen svarende til en balleeffekt:
MODEL renh = balle
Endelig i sætningen
RANDOM balle ;
erklæres balleeffekten som tilfældig.
Programmet resulterer i følgende udskrift:

```
          General Linear Models Procedure          p.1
         Class Level Information

 Class  Levels  Values

 BALLE       7  1 2 3 4 5 6 7

 PROVE       4  1 2 3 4

 Number of observations in data set = 28
```

General Linear Models Procedure p.2

Dependent Variable: RENH

Source	DF	Sum of Squares	F Value	Pr > F
Model	6	65.96264286	1.76	0.1573
Error	21	131.47220000		
Corrected Total	27	197.43484286		

R-Square	C.V.	RENH Mean
0.334098	4.311284	58.0364286

Source	DF	Type I SS	F Value	Pr > F
BALLE	6	65.96264286	1.76	0.1573

General Linear Models Procedure p.3

Source	Type I Expected Mean Square
BALLE	Var(Error) + 4 Var(BALLE)

Man genfinder det sædvanlige variansanalyseskema i udskriftens p.2. (Da der kun optræder én forklarende variabel, er der ingen forskel på type I og type III kvadratafvigelser).
Rubrikken Model refererer til kvadratafvigelsessummen sak_2 svarende til den betragtede model (som her kun indeholder faktoren Balle). Rubrikken `Error` refererer til variationen indenfor baller (sak_1).
Under overskriften `RENH Mean` er angivet gennemsnittet af renhedsmålingerne.
Testet for hypotesen $\sigma_b^2 = 0$ er det samme som testet for balleeffekt i den systematiske model. Denne teststørrelse er anført såvel i variansanalyseskemaet, som på en separat linie længere nede i udskriften.
Alle disse størrelser ville også blive udskrevet, selv om man ikke havde angivet specifikationen RANDOM balle.
Udskriftens side 3 er et resultat af denne specifikation. Udskriften viser, at den gennemsnitlige variation hidrørende fra baller ($sak_2/(k-1)$) har forventningsværdien

$$\mathsf{E}[SAK_2/(k-1)] = \sigma^2 + 4\sigma_b^2$$

(jvf beviset for sætning 5.5.2).
Herved kan σ_b^2 estimeres som i eksempel 5.2.1. ∇

Eksempel 5.8.2 ***Ubalancerede data, beregning ved SAS-proceduren GLM***
For sætttet af ubalancerede data vil programmet

```
PROC GLM ;
CLASS balle prove ;
MODEL renh = balle ;
RANDOM balle ;
RUN;
```

give udskriften:

```
         General Linear Models Procedure          p.2

 Dependent Variable: RENH

 Source                 DF  Sum of Squares  F Value     Pr > F

 Model                   6     72.55028636     4.36      0.0096
 Error                  15    41.573700000
 Corrected Total        21   114.123986366

                 R-Square            C.V.            RENH Mean
                 0.635715        2.862840           58.1522727

 Source                 DF       Type I SS  F Value     Pr > F
 BALLE                   6     72.55028636     4.36      0.0096
         General Linear Models Procedure          p.3

 Source  Type I Expected Mean Square

 BALLE   Var(Error) + 3.0909 Var(BALLE)
```

Udskriftens struktur adskiller sig ikke fra udskriften fra de balancerede data. Vi bemærker specielt, at udskriftens side 3 viser, at den anførte gennemsnitlige variation hidrørende fra baller ($sak_2/(k-1)$) har forventningsværdien

$$\mathsf{E}[SAK_2/(k-1)] = \sigma^2 + 3.0903\sigma_b^2$$

(jvf beviset for sætning 5.5.2), dvs $n_0 = 3.0903$.
Idet $\widetilde{\sigma^2} = 41.5737/15 = 2.77158$ får man derfor momentestimatet
$\widetilde{\sigma_b^2} = (12.091714 - 2.771580)/3.0909 = 3.0153$ ∇

5.8.1 Mixed

Proceduren er direkte rettet mod analyse af "blandede normalfordelingsmodeller", dvs lineære modeller, der både indeholder systematiske og tilfældige komponenter.

Proceduren kan derfor specielt benyttes til at udføre en ensidet variansanalyse i den tilfældige model.

Proceduren giver mulighed for at vælge mellem maksimum-likelihood estimatorerne (Sætning 5.5.1) og estimatorer bestemt ved maksimering af likelihoodfunktionen svarende til residualerne, (REML)-estimaterne (se side 226)

Eksempel 5.8.3 ***Balancerede data, maksimum-likelihood estimation ved SAS-proceduren MIXED***
Programmet:

```
PROC MIXED METHOD=ML ASYCOV ;
CLASS balle prove;
MODEL renh =  ;
RANDOM balle ;
RUN;
```

kalder procedure MIXED. Nøgleordet METHOD =ML angiver, at man ønsker maksimum-likelihood estimaterne, og ordet ASYCOV angiver, at man ønsker den asymptotiske varians-kovariansmatrix for estimaterne.

I procedure MIXED specificeres kun de systematiske (såkaldte "fixed") effekter i modelformlen. Det eneste systematiske led i den tilfældige model for den ensidede variansanalyse er interceptleddet, som altid er underforstået i modelspecifikationen. Modelformlen bliver derfor blot MODEL renh=, der angiver, at den afhængige variable er renh.

Sætningen RANDOM balle angiver de tilfældige komponenter i modellen. Her altså blot den variable balle.

Proceduren giver anledning til følgende udskrift:

```
              The MIXED Procedure

           Class Level Information

        Class     Levels  Values

        BALLE          7   1 2 3 4 5 6 7
        PROVE          4   1 2 3 4
```

til kontrol af de indlæste specifikationer.
Endvidere udskrives iterationsforløbet:

```
                    The MIXED Procedure

                ML Estimation Iteration History

      Iteration  Evaluations     Objective     Criterion

              0            1   82.68971508
              1            1   82.22198159    0.00000000

                  Convergence criteria met.
```

Da data er balancerede, behøves kun én iteration.
Derefter udskrives estimaterne for σ_b^2 og σ^2:

```
  Covariance Parameter Estimates (MLE)

  Cov Parm      Estimate

  BALLE       0.79066344
  Residual    6.26058095
```

De to varianser σ_b^2 og σ^2 kaldes kovariansparametre, da modellen jo er karakteriseret ved dispersionsmatricen, $\mathbf{V}$, (5.14)
Udskriftens første søjle Ratio angiver det estimerede forhold $\widehat{\gamma}$ mellem variansen svarende til den tilfældige effekt og residualvariansen.
I søjlen Estimate angives estimaterne $\widehat{\sigma}_0^2$ (svarende til BALLE) og $\widehat{\sigma}^2$ (svarende til residualvariansen).
Søjlen Z angiver

$$\frac{\widehat{\sigma}_0^2}{\sqrt{\widehat{\mathrm{V}}[\widehat{\sigma}_0^2]}}$$

og den tilsvarende størrelse for residualvariansen.
Såfremt den sande varians er nul, vil Z approximativt følge en standardiseret normalfordeling, og man kan derfor bruge Z til et approximativt test for en hypotese $\sigma_b^2 = 0$, og evt. også en hypotese $\sigma^2 = 0$.
Størrelsen Pr > |Z| angiver netop testsandsynligheden svarende til dette test. Hypotesen forkastes for små værdier af Pr > |Z|.
Som en konsekvens af ordren ASYCOV udskrives:

```
 Asymptotic Covariance Matrix of Estimates

Cov Parm    Row         COVP1         COVP2

BALLE         1    1.81896982   -0.93321128
Residual      2   -0.93321128    3.73284513
```

Estimatet er den inverse observerede informationsmatrix. Den inverse observerede informationsmatrix beregnes som 2 gange den inverse Hessianmatrix i maksimumspunktet.
Endelig udskrives et resume af tilpasningen,

```
                The MIXED Procedure

          Model Fitting Information for RENH

Description                          Value

Observations                       28.0000
Log Likelihood                    -66.8413
Akaike's Information Criterion    -68.8413
Schwarz's Bayesian Criterion      -70.1735
-2 Log Likelihood                 133.6825
```

Vi noterer således, at værdien af log likelihood, og tilsvarende af -2 log likelihood udskrives. Disse værdier kan eventuelt bruges i mere komplicerede situationer, hvor man har en række hierarkisk organiserede hypoteser. ∇

Eksempel 5.8.4 ***Balancerede data, REML estimation ved SAS-proceduren MIXED***
Såfremt vi i stedet havde benyttet residual-likelihood funktionen til estimationen ved procedurekaldet

```
PROC MIXED METHOD=REML ASYCOV ;
CLASS balle prove;
MODEL renh =  ;
RANDOM balle ;
RUN;
```

havde vi fået estimaterne

```
Covariance Parameter Estimates (REML)

Cov Parm       Estimate

BALLE        1.18329821
Residual     6.26058095
```

med de tilsvarende varians-kovariansestimater:

```
 Asymptotic Covariance Matrix of Estimates

Cov Parm    Row         COVP1         COVP2

BALLE         1    2.75128329   -0.93321128
Residual      2   -0.93321128    3.73284513
```

Vi ser, at REML-estimatet for σ^2 er det samme, som ML-estimatet, men REML-estimatet for σ_b^2 er større end ML-estimatet, svarende til at ML-estimatet ikke korrigerer for frihedsgraderne i estimationen af σ_b^2. ∇

5.9 Empirisk Bayes estimation, BLUP estimation

Bestemmelsen af aposteriorifordelingen i afsnit 5.7.2 er baseret på en fuldt specificeret apriorifordeling. Såfremt apriorifordelingen blot afspejler beslutningstagerens subjektive tiltro til "chancen"for forekomst af forskellige parameterværdier, kan man naturligvis fortolke apriorifordelingen som indeholdende al tilgængelig information om aprioriusikkerheden på parameteren, men til gengæld vil aposteriorifordelingen også kun have en subjektiv fortolkning.

I mange tilfælde vil apriorifordelingen imidlertid have en frekvensfortolkning, og parametrene i apriorifordelingen kan estimeres fra statistiske data, således som det er illustreret i de foregående afsnit. Det er klart, at parametrene i apriorifordelingen er behæftet med en vis estimationsusikkerhed, og man kan derfor overveje, i hvilket omfang denne estimationsusikkerhed overføres til aposteriorifordelingen.

Det følger af resultatet i afsnit 5.7.2, at såfremt antallet af observationer svarende til en bestemt parameterværdi vokser mod uendeligt, da vil aposteriorifordelingen snævre sig sammen omkring denne parameterværdi, såfremt blot apriorifordelingen har tillagt positiv sandsynlighed til denne værdi.

Når bare antallet af observationer svarende til en bestemt parameterværdi er stort nok, er det åbenbart ikke særlig betydningsfuldt, om apriorifordelingen er nøjagtigt bestemt, da en klassisk estimator i dette tilfælde vil give et nøjagtigt skøn over parameterværdien.

For at illustrere situationen svarende til et mere begrænset antal observationer vil vi betragte den simultane estimation af gruppeparametrene i den simple normalfordelingssituation.

Antag at vi har observationerne $X_{ij}, i = 1, 2, .., k; \quad j = 1, 2..., n$, hvor index i angiver hvilken gruppe, der har frembragt observationen, og index j angiver gentagelsen indenfor gruppe. Vi antager at observationerne indenfor en gruppe er indbyrdes uafhængige og identisk fordelt, $X_{ij}|\theta_i \sim \mathrm{N}(\theta_i, \sigma^2)$,

svarende til den sædvanlige model

$$X_{ij} = \theta_i + \epsilon_{ij} \tag{5.45}$$

hvor ϵ_{ij} er indbyrdes uafhængige, $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

Vi ønsker nu at estimere de k parametre θ_i ved en funktion

$$d(\cdot) = (d_1(\cdot), d_2(\cdot), \ldots, d_k(\cdot))'$$

af observationssættet $\mathbf{X}$. Vi indfører betegnelsen

$$L(\boldsymbol{\theta}, d(\mathbf{X})) = \frac{1}{k\sigma^2} \sum_{i=1}^{k} (\theta_i - d_i(\mathbf{X}))^2 \tag{5.46}$$

for den gennemsnitlige estimationsfejl, når de sande værdier er $(\theta_1, \theta_2, \ldots \theta_k)'$, når observationssættet antager værdien $\mathbf{X}$, og man benytter estimatoren $d(\cdot) = (d_1(\cdot), d_2(\cdot), \ldots, d_k(\cdot))'$. Vi har normeret med σ^2 for at gøre målet for estimationsfejlen invariant under skalatransformationer af $\mathbf{X}$.

Det er naturligt at udjævne effekten af den tilfældige variation af $\mathbf{X}$ for fastholdte værdier af $\boldsymbol{\theta}$ ved at betragte den forventede gennemsnitlige estimationsfejl

$$R(\boldsymbol{\theta}, d) = \mathsf{E}_{\mathbf{X}|\boldsymbol{\theta}}[L(\boldsymbol{\theta}, d(\mathbf{X}))] \tag{5.47}$$

Såfremt det yderligere antages, at $\boldsymbol{\theta}$ varierer tilfældigt, nemlig at θ_i er uafhængige og $\theta_i \sim \mathrm{N}(\mu, \gamma\sigma^2)$, kan man passende betragte den apriori forventede gennemsnitlige estimationsfejl

$$\eta((\mu, \gamma), d) = \mathsf{E}_{\boldsymbol{\theta}}[R(\boldsymbol{\theta}, d)] \,, \tag{5.48}$$

der angiver den forventede gennemsnitlige estimationsfejl under hensyntagen såvel til den tilfældige variation af $\mathbf{X}$ som af $\boldsymbol{\theta}$.

Såfremt parametrene μ, γ og σ^2 var kendt, ville man vælge en estimator, $d(\cdot)$ således at $\eta((\mu, \gamma), d)$ blev mindst mulig.

Der gælder

Sætning 5.9.1 ***Bayesløsning ved simultan estimation***
Lad fordelingen af observationssættet $\mathbf{X}$ være givet ved (5.45), hvor $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ er uafhængige, og $\theta_i \sim \mathrm{N}(\mu, \gamma\sigma^2)$ er indbyrdes uafhængige og uafhængige af ϵ_{ij}. Da antager $\eta((\mu, \gamma), d)$ sit minimum for $d = d^B$, hvor

$$d_i^B(\mathbf{X}) = \widetilde{\theta}_i^B = w\mu + (1-w)\overline{X}_{i\cdot} \quad i = 1, 2, \ldots, k \tag{5.49}$$

med

$$w = \frac{1}{1+n\gamma} \quad \text{og} \quad \overline{X}_{i\cdot} = \sum_{j=1}^{n} X_{ij}/n; \qquad i = 1, 2, \ldots, k \tag{5.50}$$

Der gælder

$$\eta((\mu, \gamma), d^B) = \frac{1}{n} \; \frac{n\gamma}{1+n\gamma} = \frac{1}{n}\,(1-w) \tag{5.51}$$

∇

Til sammenligning betragter vi nu den estimator, man ville benytte under modellen med systematiske gruppeeffekter, nemlig maksimum likelihood estimatoren $\theta_i^{ML} = \overline{X}_{i\cdot}$ for gruppeparameteren θ_i,

$$d_i^{ML}(\mathbf{X}) = \widehat{\theta}_i^{ML} \; = \overline{X}_{i\cdot} \quad i = 1, 2, \ldots, k \tag{5.52}$$

Der gælder

Sætning 5.9.2 ***Aprioriestimationsfejlen for ML-estimatoren***
Under antagelserne fra Sætning 5.9.1 er

$$\eta((\mu, \gamma), d^{ML}) = R(\boldsymbol{\theta}, d^{ML}) = \frac{1}{n}$$

hvor maksimum likelihood estimatoren d^{ML} er givet ved (5.52).

∇

Under den systematiske model har ML-estimatoren, $\widehat{\theta}_i^{ML}$, de egenskaber, der sædvanligvis kræves af en god estimator: estimatoren er central og variansminimal, men alligevel (eller måske netop derfor) er aprioriforventningen af den gennemsnitlige estimationsfejl større end for estimatoren $\widetilde{\theta}_i^B$.

Den stokastiske sammenhæng mellem de enkelte θ_i, som er specificeret i aprioriforde lingen af $\boldsymbol{\theta}$, udtrykker at parameterværdierne ikke varierer vilkårligt, men at de i gentagne stikprøver vil gruppere sig omkring μ i overenstemmelse med hyppighederne i en normalfordeling med varians $\gamma\sigma^2$. Bayesestimatoren

$$\widetilde{\theta}_i^B = w\mu + (1-w)\widehat{\theta}_i^{ML}$$

udnytter denne information ved at trække de fundne gruppegennemsnit $\widehat{\theta}_i^{ML}$ ind mod tyngdepunktet μ. Det kan vises, at "krympningsfaktoren", $1-w$, er bestemt som koefficienten til θ i regressionen af θ på $\widehat{\theta}_i^{ML}$ i den simultane fordeling af θ og $\widehat{\theta}_i^{ML}$. Vi har nemlig

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

Det kan vises, at

$$\eta((\mu,\gamma),d^B) = (1-w)\eta((\mu,\gamma),d^{ML}) \tag{5.53}$$

der udtrykker, at reduktionen i den apriori forventede gennemsnitlige estimationsfejl ved at benytte Bayesestimatoren d^B i stedet for maksimum likelihood estimatoren d^{ML} er proportional med krympningsfaktoren $(1-w)$. Jo stærkere krympning, dvs jo mindre værdi af $(1-w)$, desto større er effekten ved at benytte Bayesestimatoren.

Dette er ikke overraskende, da krympningsfaktoren $(1-w)$ jo netop er lille, når variansen mellem parameterværdierne θ er lille i forhold til måleusikkerheden σ^2/n ved bestemmelsen af de enkelte parameterværdier.

Såfremt man ikke kender apriorifordelingens parametre (μ,γ), kan man udnytte hele observationssættet $\mathbf{X}$ til estimation af μ,γ og σ^2, således som det er beskrevet i afsnit 5.5. Man har

$$\widehat{\mu} = \overline{X}_{\cdot\cdot} = \frac{1}{k}\sum_{i=1}^{k}\overline{X}_{i\cdot}, \quad \widehat{\sigma}^2 = SAK_1/\{k(n-1)\}$$

med

$$SAK_1 \sim \sigma^2\chi^2(k(n-1)) \quad \text{og} \quad SAK2 \sim \sigma^2(1+n\gamma)\chi^2(k-1)\ .$$

Idet SAK_1 og SAK_2 er uafhængige, og

$$\mathsf{E}\left[\frac{k-3}{SAK_2}\right] = \frac{1}{\sigma^2(1+n\gamma)}$$

finder vi, at

$$\mathsf{E}\left[\frac{\widehat{\sigma}^2}{SAK_2/(k-3)}\right] = \frac{1}{1+n\gamma} = w \ .$$

Vi vil imidlertid indføre estimatoren

$$\widetilde{\sigma}^2 = \frac{SAK_1}{k(n-1)+2} \tag{5.54}$$

og benytte

$$\widehat{w} = \frac{\widetilde{\sigma}^2}{SAK_2/(k-3)} \tag{5.55}$$

Vi bemærker, at $\widehat{w}$ kan udtrykkes ved den sædvanlige F-teststørrelse som

$$\widehat{w} = \frac{k-3}{k-1}\ \frac{k(n-1)}{k(n-1)+2}\ \frac{1}{F} \ .$$

Indsættes $\widehat{\mu}$ og $\widehat{w}$ i udtrykket

$$\mathsf{E}[\theta_i|\overline{X}_{i.}] = w\mu + (1-w)\overline{X}_{i.}$$

for aposterioriforventningsværdien fås estimatoren

$$d_i^{EB}(\mathbf{X}) = \widetilde{\theta}_i^{EB} = \widehat{w}\widehat{\mu} + (1-\widehat{w})\overline{X}_{i.} \tag{5.56}$$

Estimatoren kaldes en empirisk Bayes-estimator, da den udnytter et estimat for apriorifordelingen baseret på den empiriske (den observerede) fordeling af $\overline{X}_{i.}$.

Egenskaberne ved den empiriske Bayes-estimator fremgår af

Sætning 5.9.3 ***Apriori estimationsfejlen for den empiriske Bayes-estimator***
Under antagelserne fra sætning 5.9.1 er

$$\eta((\mu,\gamma), d^{EB}) = \frac{1}{n}\left(1 - \frac{2(k-3)}{\{k(n-1)+2\}(1+n\gamma)}\right)$$

∇

Bemærkning 1 ***Den forventede estimationsfejl for den empiriske Bayesestimator er mindre end for maksimum likelihood estimatoren***
Ved sammenligning med sætning 5.9.2 ser vi, at såfremt $k > 3$, vil den empiriske Bayesestimator d^{EB} have en mindre apriori forventet estimationsfejl, end maksimum likelihood estimatoren d^{ML}. Jo mindre værdi af γ, desto større er forskellen mellem de apriori forventede estimationsfejl for de to estimatorer. Det kan iøvrigt vises, at også den forventede estimationsfejl $R(\boldsymbol{\theta}, d^{EB}) < R(\boldsymbol{\theta}, d^{ML})$ se f.eks. Efron og Morris (1972 a).
Vi bemærker endelig, at estimatet $\widehat{w}$ for w ikke nødvendigvis er mindre end 1. Estimatoren d^{EB} kan derfor forbedres ved at tvinge estimatet for w til at være mindre end 1. Der gælder således, at

$$\eta((\mu, \gamma), \widetilde{d}^{EB}) < \eta((\mu, \gamma), d^{EB})$$

hvor $\widetilde{d}^{EB}$ fremkommer af d^{EB} ved at erstatte $\widehat{w}$ med

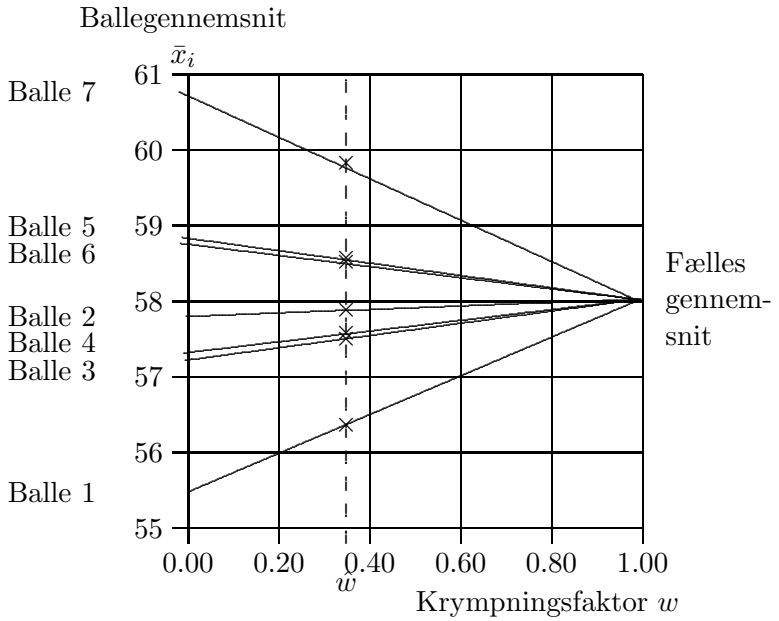
$$\widehat{w}^{+} = \begin{cases} \widehat{w} & \text{for } \widehat{w} < 1 \\ 1 & \text{for } \widehat{w} \geq 1 \end{cases}$$

∇

Eksempel 5.9.1 ***Simultan estimation af ballemiddelværdier***
Til illustration af de foregående betragtninger vil vi betragte data fra eksempel 5.2.1 og antage at formålet er at estimere renheden i hver af de 7 undersøgte baller.
I eksemplet fandt vi $sak_1 = 131.47220$, $sak_2 = 65.9626$, $\widetilde{\sigma}^2 = sak_1/23 = 5.7162$; $sak_2/4 = 16.4907$.
Ved indsættelse i (5.55) fås

$$\widehat{w} = 5.7162/16.4907 = 0.3466$$

De 7 estimater for ballernes renhed kan da bestemmes ved (5.56).
Figuren illustrererer betydningen af "krympningsfaktoren" w for de 7 individuelle renhedsestimater. For $w = 0$ føres man til at bruge de 7 individuelle ballegennemsnit, for $w = 1$ føres man til det fælles gennemsnit.



Vi har $sak_1 = 131.47220$, $sak_2 = 65.9626$, $\tilde{\sigma}^2 = sak_1/23 = 5.7162$ $sak_2/4 = 16.4907$ hvorfor $\hat{w} = 5.7162/16.4907 = 0.3466$. ∇

BLUP I PROC MIXED

I PROC MIXED beregnes de såkaldte "Empiriske Best Linear Unbiased Predictors", EBLUP, for de tilfældige effekter, γ, (som i det væsentlige er empirisk Bayes estimater) ved i RANDOM sætningen at vælge optionen (efter en slash) SOLUTION, se hjælp funktionen.
Man skal bare lige passe på, hvis man har flere faktorer af random effects, så man hellere læse manualen om Mixed Linear Models.

5.10 Flere eksempler på hierarkisk variation for normalfordelingsmodeller

Som nævnt oplever man ofte at dataindsamlingen inducerer en hierarkisk variation i data. En sådan ekstra variation kan optræde som en forstyrrende effekt, eller man kan netop være interesseret i at beskrive denne effekt. I det følgende gives eksempler på begge situationer.

5.10.1 Analyse af blokforsøg

Eksempel på forstyrrende tilfældig effekt

Antag nu, at vi i balleeksemplet havde analyseret de to første prøver (nr 1 og 2) med én metode (A), og de sidste prøver (nr. 3 og 4, hvis de ellers var der) med en anden metode (B), og at hele undersøgelsen gik ud på at undersøge forskellen mellem de to metoder.

Hvis nu der bare havde været en prøve med hver metode fra hver balle havde det jo været det parvise t-test fra statistik 1. Så havde vi bare beregnet differensen mellem de to prøver fra hver balle, og undersøgt om denne differens kunne tænkes at stamme fra en normalfordeling med middelværdi nul (ved et t-test på differenserne, $D_i \sim \mathrm{N}(\delta, \sigma_d^2)$). Herved havde vi jo elimineret variationen mellem baller ved vurdering af differensen. (Vi havde brugt hver balle som sin egen kontrol).

I forsøgsplanlægningslitteraturen siger man, at ballen udgør en *blok*.

Datasættet `uldmetode` indeholder en variant af ulddatasættet, nemlig samme ubalancerede udpluk som `uldopg`, men nu er prøverne antaget analyseret ved to metoder, og værdierne for metode B er fremkommet ved at addere 2 til de oprindelige målinger.

Programstumpen

```
PROC MIXED DATA= uldmetode  ASYCOV ;
 CLASS BALLE PROVE metode ;
 MODEL renh = metode / s ;
 RANDOM BALLE         ;
 RUN;
```

udfører en analyse svarende til modellen

$$Y_{ijk} = \mu_0 + u_i + \delta_k + \epsilon_{ijk} \ , \tag{5.57}$$

hvor δ_k angiver metodeeffekten.

Optionen "s" for *solution* i modellen for den systematiske effekt bevirker, at metodestimaterne udskrives.

Teststørrelsen bruger netop estimatet over σ^2 som nævner.
Tilsvarende udføres et type-3 test for forskellen mellem metoder. I denne situation med kun to metoder giver det samme resultat, som testet for effekten af metode A. Hvis vi havde brugt flere metoder, ville type 3 testet vurdere forskellen under ét, mens testene under solution ville vurdere afvigelsen af hver metode fra referencemetoden for sig.

Vi kunne også have bedt om estimaterne for balleeffekterne ved at skrive "s" i random sætningen

```
PROC MIXED DATA= uldmetode  ASYCOV ;
 CLASS BALLE PROVE metode ;
 MODEL renh = metode / s ;
 RANDOM BALLE  / s        ;
 RUN;
```

Og vi kunne selvfølgelig også bede om de prædikterede værdier (der nu tager hensyn til den benyttede metode)

```
PROC MIXED DATA= uldmetode  ASYCOV ;
 CLASS BALLE PROVE metode ;
 MODEL renh = metode / s outp = pred ;
 RANDOM BALLE  / s        ;
 RUN;
PROC PRINT DATA=pred;
RUN;
```

Der er mange varianter af sådanne "blokstrukturer", split-plot, repeated measurements jvf hjælp siderne og bogen om PROC MIXED.

5.10.2 Varierende regressionslinier

Eksempel på interessant varianskomponent

I den oprindelige variant af balleeksemplet (uden de to metoder A og B) var vi netop interessseret i at beskrive den tilfældige variation σ_b^2 mellem baller.

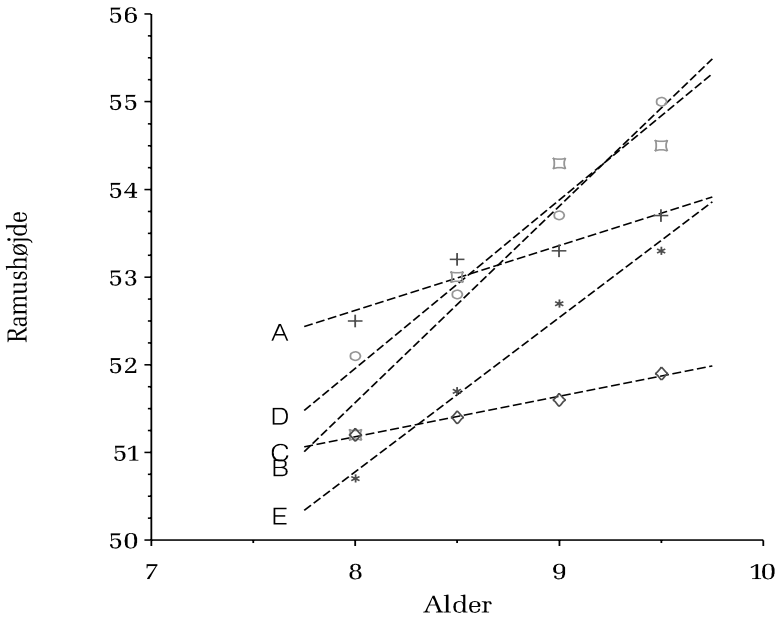
Datasættet `ramus` indeholder vi samhørende værdier af alder og ramushøjde for fem drenge. Data er udvalgte data fra Elston og Grizzle (1962).

Nedenstående tabel viser samhørende værdier af højden af kæbebenet (ramus) i [mm] og alderen i [år] hos 5 drenge i aldersgruppen 8-10 år. For hver dreng er ramushøjden registreret fire gange, nemlig ved alder 8, 8 1/2 , 9 og 9 1/2 år.

Tabel over samhørende værdier af ramus højde (i mm.) og alder for 5 drenge.

	Alder i år					
	8	8 1/2	9	9 1/2		
Dreng	reduceret alder x_i = alder -8.75				β_{i1}	β_{i2}
	-0.75	-0.25	0.25	0.75		
A	52.5	53.2	53.3	53.7	53.175	0.74
B	51.2	53.0	54.3	54.5	53.250	2.24
C	51.2	51.4	51.6	51.9	51.525	0.46
D	52.1	52.8	53.7	55.0	53.400	1.92
E	50.7	51.7	52.7	53.3	52.100	1.76
snit	51.54	52.42	53.12	53.68	52.690	1.424

Data er vist grafisk i nedenstående figur



For eksemplets skyld vil vi modellere disse data med en regressionsmodel (selv om en model, der tilgodeså, at tilvæksten kun kan være positiv måske ville være mere passende).

Vi formulerer modellen

$$Y_{ij} = \beta_{i1} + x_{ij}\beta_{i2} + \epsilon_{ij},\ i = 1, 2, \ldots, 5;\ j = 1, 2, 3, 4,$$

hvor Y_{ij} angiver den registrerede ramushøjde i [mm] hos den i'te dreng ved den reducerede alder x_{ij}, og hvor ϵ_{ij} antages uafhængige identisk normaltfordelte $\mathrm{N}(0, \sigma^2)$.

Vi har altså modelmatricen

$$\mathbf{X} = \begin{pmatrix} 1.0 & -0.75 \\ 1.0 & -0.25 \\ 1.0 & 0.25 \\ 1.0 & 0.75 \end{pmatrix} \quad \text{med} \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 1.25 \end{pmatrix}$$

hvorfor

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix}$$

Skønnene over de individuelle regressionkoefficienter $\widehat{\boldsymbol{\beta}}_i$ beregnet ved $\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i$ er anført i tabellen.

Hvis vi kun havde været interesseret i disse fem drenge, kunne vi vælge en *systematisk model* med fem regressionslinier med hver sin intercept og hver sin hældning, dvs

```
RAMUS = DRENG ALDER DRENG*ALDER
```

og det giver da stærk mistanke om at disse fem drenge er forskellige (altså forskellige hældninger og forskellige intercepter).

Men nu opfatter vi dem bare som en stikprøve af drenge, så vi vil modellere dem ved en *tilfældig model*.

Observationssættet for den i'te modelleres ved

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

hvor

$$\boldsymbol{\beta}_i \sim \mathrm{N}_2(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma}), \ \ \boldsymbol{\epsilon}_i \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \ , \tag{5.58}$$

og hvor $\boldsymbol{\beta}_i, \boldsymbol{\beta}_j$ er indbyrdes uafhængige for $i \neq j$, og $\boldsymbol{\epsilon}_i$ og $\boldsymbol{\epsilon}_j$ er indbyrdes uafhængige for $i \neq j$, og endvidere $\boldsymbol{\beta}_i$ og $\boldsymbol{\epsilon}_j$ er uafhængige.

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

Marginal fordeling
Den marginale fordeling af $\mathbf{Y}_i$ er givet ved

$$\mathbf{Y}_i \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2\{\mathbf{I}_n + \mathbf{X}\boldsymbol{\Gamma}\mathbf{X}'\}) \ ,$$

og den marginale fordeling af estimaterne, $\widehat{\boldsymbol{\beta}}_i$ for regressionskoefficienterne er

$$\widehat{\boldsymbol{\beta}}_i \sim \mathrm{N}_2(\boldsymbol{\beta}_0, \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + \boldsymbol{\Gamma}]) \tag{5.59}$$

Den afspejler både designmatricen $\mathbf{X}$ og kovariansmatricen $\boldsymbol{\Gamma}$ i fordelingen af $\boldsymbol{\beta}$, men vi kender jo $\mathbf{X}$, så vi må kunne få fat i et estimat for $\boldsymbol{\Gamma}$.

Prøv med PROC MIXED

```
PROC MIXED DATA=ramus;
     CLASS DRENG;
      MODEL ramus = alder /s  ;
      RANDOM int   alder  / subject=dreng type= un ;
      RUN;
```

`MODEL` sætningen specificerer den systematiske del af modellen (en "gennemsnitlig" regressionslinie), og `RANDOM` sætningen angiver nu, at interceptleddet og hældningen varierer tilfældigt fra dreng til dreng (`subject 0 dreng`). Optionen `type=un` angiver at $\boldsymbol{\Gamma}$ er *ustruktureret*, dvs vi lader data bestemme formen af den.

Ud kommer så et estimat for parametrene i den simultane fordeling af intercept og hældning:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	dreng	35.5984
UN(2,1)	dreng	-4.3426
UN(2,2)	dreng	0.5364
Residual		0.08640

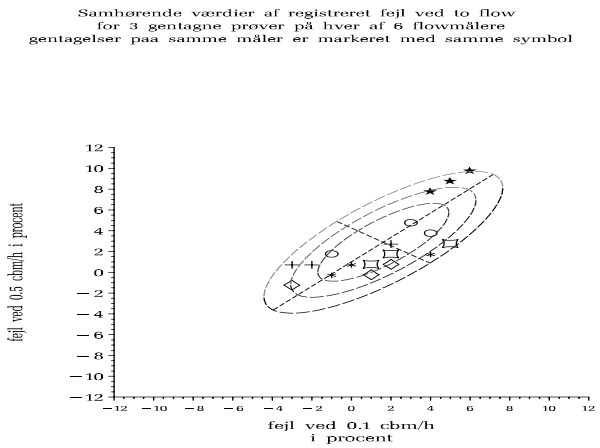
5.10.3 Flerdimensional normalfordeling

Variation mellem målefejl for flowmålere

Nedenstående tabel viser resultaterne af 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].

	Gentagelse					
	1		2		3	
	flow		flow		flow	
Maaler	0.1	0.5	0.1	0.5	0.1	0.5
41	-2.0	1.0	2.0	3.0	2.0	2.0
42	5.0	3.0	1.0	1.0	2.0	2.0
43	2.0	1.0	-3.0	-1.0	1.0	0.0
44	4.0	4.0	-1.0	2.0	3.0	5.0
45	4.0	2.0	0.0	1.0	-1.0	0.0
46	5.0	9.0	4.0	8.0	6.0	10.0

Målingerne er illustreret i nedenstående figur



Der ses at være en klar sammenhæng mellem gentagne kalibreringer på samme måler med en forholdsvist lille gentagelsesvariation - og en noget større variation mellem målerne.

Med henblik på beskrivelse af variationen af fejlvisningen i målerpopulationen opstiller man følgende model:

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

hvor $\boldsymbol{\alpha_i}$ er uafhængige, $\boldsymbol{\alpha}_i \sim \mathrm{N}_2(0, \boldsymbol{\Sigma}_0)$, og hvor $\boldsymbol{\epsilon}_{ij}$ er indbyrdes uafhængige, $\boldsymbol{\epsilon} \sim \mathrm{N}_2(0, \boldsymbol{\Sigma})$.

Situationen er analog til balleeksemplet, blot todimensionalt, en måler svarer til en "balle".
Analyse ved en systematisk model kunne fx foretages ved PROC GLM

Antag, at data fra eksemplet er indlæst i de variable lbnr, fejl1 og fejl2. SAS-programmet

```
PROC GLM ;
CLASS lbnr;
MODEL fejl1 fejl2  = lbnr /E1 ;
MANOVA H=lbnr/PRINTH PRINTE HTYPE=1 ETYPE=1;
RANDOM lbnr;
RUN;
```

definerer i ordren CLASS lbnr ;, at den variable lbnr som klassifikationsvariabel. Modelformlen MODEL fejl1 fejl2 = lbnr / E1 ; angiver, at vi betragter en model svarende til en løbenummereffekt (samt et intercept) for de to variable fejl1 og fejl2. Valget E1 i modelformlen angiver, at vi ønsker en såkaldt type-I kvadratafvigelsessum.
I sætningen

```
MANOVA H=lbnr/PRINTH PRINTE ;
```

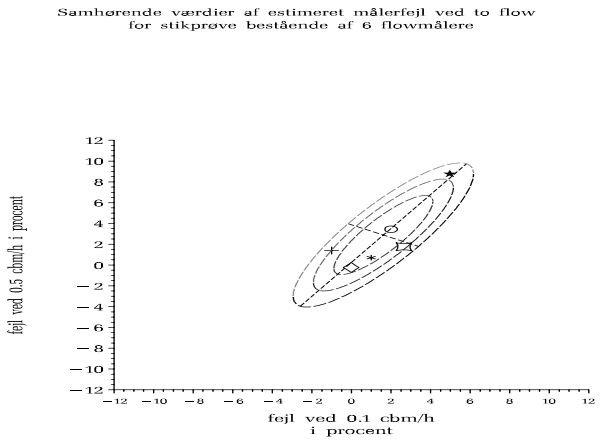
angiver nøgleordet MANOVA, at man ønsker at opfatte de to variable fejl1 og fejl2 på venstre side i modelformlen som en todimensional observation.
Ordren H=lbnr angiver, at man ønsker at teste effekten svarende til lbnr. Ordren er efterfulgt af en række options, nemlig PRINTH, der bevirker, at **SAK**-matricen svarende til lbnr effekten udskrives (dvs. $\mathbf{sak}_2$ matricen), og optionen PRINTE, der bevirker, at **SAK**-matricen svarende til residualerne udskrives (dvs. $\mathbf{sak}_1$-matricen).
Endelig angiver ordren RANDOM lbnr;, at effekten fra lbnr skal opfattes som tilfældig.
Proceduren udfører først de endimensionale analyser på de variable fejl1 og fejl2 hver for sig. Derefter udskrives diverse størrelser svarende til en multivariat variansanalyse

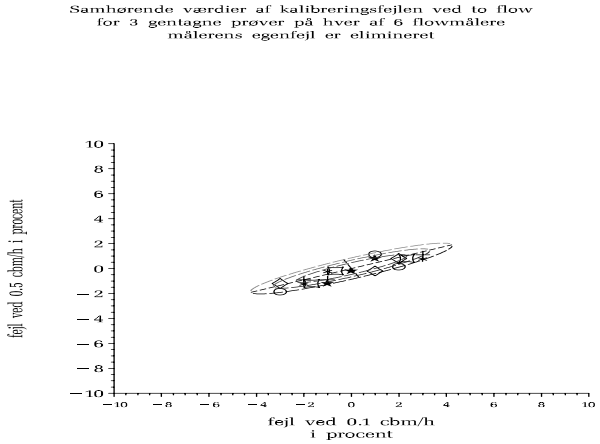
$\mathbf{sak}_1$ matricen udskrives under overskriften E = Error SS&CP Matrix

$\mathbf{sak}_2$-matricen udskrives under overskriften H = Type I SS&CP Matrix for LBNR og endelig udskrives en række teststørrelser for test af hypotesen om forsvindende effekt af lbnr i den systematiske model.
Som et resultat af ordren RANDOM lbnr udskrives udtrykket for

$$\mathsf{E}[\mathbf{sak}_2/(k-1)] = \mathbf{\Sigma} + n_0\mathbf{\Sigma}_0$$

Opspaltningen er illustreret i følgende figurer:





5.11 Aposteriorifordelinger for flerdimensionale normalfordelinger

Lad $\mathbf{X} \mid \boldsymbol{\mu} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ og lad $\boldsymbol{\mu} \sim \mathrm{N}_p(\mathbf{m}, \boldsymbol{\Sigma}_0)$, hvor $\boldsymbol{\Sigma}$ og $\boldsymbol{\Sigma}_0$ har fuld rang, p. Da er aposteriorifordelingen af $\boldsymbol{\mu}$ efter observation af $\mathbf{X} = \mathbf{x}$ givet ved

$$\boldsymbol{\mu} \mid X = \mathbf{x} \sim \mathrm{N}_p(\mathbf{W}\mathbf{m} + (\mathbf{I} - \mathbf{W})\mathbf{x}, (\mathbf{I} - \mathbf{W})\boldsymbol{\Sigma}) \tag{5.60}$$

med

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

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

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

Aposteriorifordeling i regressionsmodel

Lad Y angive en $n \times 1$ dimensional vektor af observationer, og lad X angive en $n \times p$ dimensional matrix af kendte koefficienter. Antag at $Y \mid \boldsymbol{\beta} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ og at apriorifordelingen af $\boldsymbol{\beta}$ er

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

hvor $\boldsymbol{\Lambda}$ har fuld rang.
Da er aposteriorifordelingen af $\boldsymbol{\beta}$ efter observation af $Y = \mathbf{y}$ givet ved

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

hvor

$$\begin{aligned} \boldsymbol{\beta}_1 &= \mathbf{W}\boldsymbol{\beta}_0 + \mathbf{W}\boldsymbol{\Lambda}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad (5.61) \\ \text{med} & \\ \mathbf{W} &= (\mathbf{I}+\boldsymbol{\Gamma})^{-1} \\ \text{og} & \\ \boldsymbol{\Gamma} &= \boldsymbol{\Lambda}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \\ \text{og hvor} & \\ \boldsymbol{\Lambda}_1 &= (\mathbf{I}+\boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda} = \mathbf{W}\boldsymbol{\Lambda} \qquad (5.62) \end{aligned}$$

Aposteriorimiddelværdien udtrykt som et vægtet gennemsnit

Såfremt $\mathbf{X}$ har fuld rang, kan $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ inverteres, og vi får

$$\beta_1 = \mathbf{W}\beta_0 + (\mathbf{I}-\mathbf{W})\hat{\beta}$$

hvor $\hat{\beta}$ angiver den sædvanlige mindste kvadraters estimator (**??**) for β,

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