Dagens program, Mandag den 4. april :

udvidelse til modeller med overdispersion, quasi-likelihood, dvs alle knapperne i Metode menuen
Hierarkiske normalfordelingsmodeller

- Generaliserede lineære modeller, afslutning
  - Ordnet kategorisk respons, Data buscum, busmult
  - Overdispersion, Data beetles
  - Quasi-likelihood
- Hierarkiske normalfordelingsmodeller
  - Model
  - Varians- kovariansforhold
  - Fordeling af gruppegennemsnit
  - Test af homogenitetshypotese , variansanalyseskema
  - Maksimum-likelihood estimation
  - Estimation af tilfældige effekter
  - Fortolkning af estimater , empiriske Bayes

Test for modeltilpasning i generaliserede lineære modeller

Størrelsen Error Deviance i Analysis of Deviance bruges til at teste goodness of fit før evt reduktion af modellen.
Model: dod = sta myaar sta∗myaar

Hvis data viser stærk uoverensstemmelse med modellen, dvs hvis
"Error" deviansen er signifikant i χ^2-fordelingen, må man finde flere forklarende variable - eller bruge en model med **overdispersion**

Overdispersion, dispersionsparameter:

Eksempel
Datasæt Beetles indeholder data fra et eksperiment, der havde til formål at vurdere sammenhængen mellem indblæst koncentration af ethylenoxid og andel døde biller i et prøvemedium.

"konc" angiver koncentrationen, "ekspon" antallet af eksponerede, "doede" antallet af døde, og "andel" angiver andelen af døde.

Grafisk vurdering:
Scatterplot af andel mod koncentration, ser logistisk ud

Billedødelighed:

Fit menu:
doede $\longrightarrow$ Y
konc $\longrightarrow$ X
Method:
Response Dist: Binomial
ekspon $\longrightarrow$ Binomial
Link: Logit

Teststørrelsen for Goodness of fit er stærkt signifikant, så vi må afvise antagelsen om binomialfordeling og logistisk dosis-respons.

Der er ingen systematik i residualerne, så vi vælger at fastholde den logistiske model, men der er åbenbart større varians, end forklaret ved binomialfordelingen (fx pga klumpning).

Dispersionsparameter:

Vi vælger at indføre en **dispersionsparameter** σ^2 ved $\lambda_i = w_i/\sigma^2$ (jvf fremstillingen (4.5) side 106) som en fast faktor, σ^2, til variansen $\mu_i(1-\mu_i)/w_i$ på en observeret binomial **andel** $Z_i = Y_i/n_i$.

$$\mathrm{V}[\mathrm{Z_i}] = \sigma^2 \ \frac{\mu_\mathrm{i}(1-\mu_\mathrm{i})}{\mathrm{n_i}}$$

idet $w_i = n_i$.

Bemærk, at her indgår

- Enhedsvariansfunktionen $V(\mu) = \mu(1-\mu)$
- Den kendte vægt $w_i = n_i$ (indføres i proceduren via angivelse af nbinomial).
- dispersionsparameteren σ^2

Estimation af dispersionsparameter:

Vi ved ikke, hvordan tætheden ser ud for en binomialfordeling med påhæftet dispersionsparameter. Dispersionsparameteren indgår i normeringskonstanten på en eller anden måde.
Så vi kan ikke uden videre lave en maximum-likelihood estimation.

Men vi ved, at residualdeviansen $D\big(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\big) \sim \sigma^2\chi^2$, og så kan vi da i lighed med normalfordelingssituationen estimere σ^2 ved

$$\widehat{\sigma}^2 = \frac{D\big(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\big)}{k-m}$$

eller vi kunne have brugt Pearson-størrelsen på tilsvarende måde.
I SAS kaldes kvadratroden, σ, af dispersionsparameteren σ^2 for Scale. [Dette gælder for Normal, Invers Gauss, Binomial og Poisson-fordeling. For gammafordelingen er Scale netop dispersionsparameteren.]

Tastetryk ved estimation af dispersionsparameter :

Ved Fit menu:
I Method menuen vælges
Scale: Deviance
endvidere vælges Quasi-likelihood

Det bevirker, at der estimeres en Scale ved

$$\widehat{\sigma} = \frac{D\big(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\big)}{k-m}$$

men iøvrigt får man de samme størrelser som før (dog ændres standardiserede deviansresidualer mv).

Tastetryk ved estimation af dispersionsparameter :

Ved PROC GENMOD:

```
PROC GENMOD data = beetles;
model doede/ekspon = konc /DIST = Bin
SCALE=D OBSTATS Type1 ;
RUN; QUIT;
```

Scale = D (DEVIANCE) bevirker at scale estimeres som kvadratroden af devians divideret med frihedsgrader.

Bemærk, at nu kan vi ikke teste goodness of fit - og vi kan ikke umiddelbart bruge teststørrelser $D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\widehat{\boldsymbol{\gamma}}))/df$ i type 1 og type 3 tests, men at de skal divideres med kvadratet på scale parameteren resulterende i et approksimativt F-test (i lighed med normalfordelingssituationen).

I PROC GENMOD foretages denne skalering automatisk.

Pseudo-likelihood :

I princippet er estimaterne for middelværdiparametrene ikke maksimum-likelihood estimater, for likelihoodfunktionen indeholder også normeringskonstantens afhængighed af σ.

Den funktion

$$D^*(\mathbf{y};\boldsymbol{\mu}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{k} w_i d(y_i,\mu_i)$$

som vi har maksimeret, kaldes en **quasi-loglikelihood** (quasi= som om).

Estimationsligningen (estimation equation) til bestemmelse af middelværdiparametrene fås netop ved at sætte quasi-scoren lig med nul.

Hvad angår middelværdiparametrene opfører den sig jo som den "rigtige" likelihood. Estimationen af middelværdiparametrene ændres ikke ved at indføre denne overdispersion.

PROC GENMOD, Scale option

Noscale	Scale= talværdi	
markeret	markeret	scale fastholdt på *talværdi*
markeret	ej markeret	scale fastholdt på værdien 1
ej markeret	ej markeret	scale estimeres ved ML
ej markeret	markeret	iteration starter ved *talværdi*

NOSCALE holder skalaparameteren fast.
ML-estimation kan udføres for normal, invers gauss, negativ binomialfordeling og gamma.

SCALE = DEVIANCE eller SCALE = PEARSON bruges til estimation af overdispersion.

Overdispersion:

Overdispersion kan manifesteres på to forskellige måder:

Heterogenitetsfaktor:
Konstant faktor, σ^2 til variansfunktionen $V(\mu)$

$$\mathsf{V}[\mathsf{Y}] = \sigma^2\mathsf{V}(\mu)$$

Resultat af "random effects"
Det kan ske, at μ varierer tilfældigt inden for et forsøg fx kan det forekomme ved musefosterforsøgene, at der er en tilfældig variation af dødssandsynligheden p fra kuld til kuld.

Vi vil senere (kap 7) vende tilbage til modellering af sådanne tilfældige effekter.

Tilbage til normalfordelingsmodeller

Eksempel, uldballer
Renhed 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

Formål at beskrive variation med henblik på tilrettelæggelse af stikprøveudtagning fra tilsvarende sendinger.

Hierarkisk model

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)

Alternativ formulering:

$$Y_{ij} = \mu_0 + A_i + \epsilon_{ij} \ ,$$

$\epsilon_{ij} \sim \mathsf{N}(0, \sigma^2)$ uafhængige

$A_i \sim \mathsf{N}(0, \sigma_b^2)$ uafhængige, og uafhængige af ϵ

Marginale fordelinger

Den marginale fordeling af Y_{ij} er en normal fordeling

$\mathsf{E}[\mathsf{Y}_{ij}] = \mu_0$

$$\mathsf{COV}[\mathsf{Y}_{ij}, \mathsf{Y}_{hl}] = \begin{cases} \sigma_b^2 + \sigma^2 & \text{for } (i,j) = (h,l) \\ \sigma_b^2 & \text{for } i = h, \ \ j \neq l \\ 0 & \text{for } i \neq h \end{cases}$$

Vi husker nemlig

$$\mathsf{E}[\mathsf{X}] = \mathsf{E}_{\mathsf{Y}}[\,\mathsf{E}[\mathsf{X}|\mathsf{Y}]\,]$$

$$\mathbf{D}\,[X] = \mathsf{E}_{\mathsf{Y}}[\,\mathbf{D}\,[\mathsf{X}|\mathsf{Y}]\,] + \mathbf{D}_{\mathsf{Y}}[\,\mathsf{E}[\mathsf{X}|\mathsf{Y}]\,]$$

Egenskaber ved modellen

Observationer fra samme balle er korrelerede

Korrelationskoefficienten

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

kaldes **intraklassekorrelationen**

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

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

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

Dispersionsmatricen for $\mathbf{Y}_i$ er

$$\mathbf{V}_i = \mathbf{D}\,[Y_i] = \mathsf{E}[(\mathsf{Y_i} - \boldsymbol{\mu})(\mathsf{Y_i} - \boldsymbol{\mu})^{\mathsf{T}}]$$

$$= \begin{pmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \ldots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \ldots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \ldots & \sigma_b^2 + \sigma^2 \end{pmatrix}$$

Fordeling af hele observationssættet

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

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

Fordeling af gruppegennemsnit

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

Gruppegennemsnit $\overline{Y}_{i\cdot}$, $i = 1, 2, \ldots, q$ er indbyrdes uafhængige.

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

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

Forøgelse af stikprøvestørrelse

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

Test af homogenitetshypotese

Ingen forskel på baller, svarer til

$$\mathsf{H}_{II} : \sigma_b^2 = 0.$$

Hypotesen testes ved at sammenligne varianskvotienten

$$Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \tag{0.1}$$

med fraktilerne i en $\mathsf{F}(k-1, N-k)$-fordeling

Som vanligt:

$$\begin{aligned} \mathsf{sak}_1 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2 \\ \overline{\overline{y}} &= \sum_i n_i \overline{y}_{i\cdot} \Big/ N \\ \mathsf{sak}_2 &= \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{\overline{y}})^2 \end{aligned}$$

Konfidensinterval for varianskvotienten γ

I balanceret tilfælde, $n_1 = n_2 = \ldots = n_k = n$

$$\mathsf{P}[\gamma_{\mathsf{L}} < \gamma < \gamma_{\mathsf{U}}] = 1 - \alpha$$

fås ved at benytte

$$\gamma_L = \frac{1}{n}\left(\frac{Z}{\mathsf{F}(k-1, N-k)_{1-\alpha/2}} - 1\right)$$

og

$$\gamma_U = \frac{1}{n}\left(\frac{Z}{\mathsf{F}(k-1, N-k)_{\alpha/2}} - 1\right)$$

Resultatet er eksakt

Forventningsværdi af SAK'er

$$\mathsf{E}[\mathsf{SAK}_1/(\mathsf{N}-\mathsf{k})] = \sigma^2$$

og

$$\mathsf{E}[\mathsf{SAK}_2/(\mathsf{k}-1)] = \sigma^2 + n_0\sigma_0^2$$

hvor

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

er den vægtede gennemsnitlige gruppestørrelse (i balanceret tilfælde er $n_0 = n$).

Forventningsværdi af SAK'er

Kan fås af PROC GLM ved at skrive en sætning
RANDOM balle;

Variation	SAK	f	$s^2 = SAK/f$	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

Maksimum- likelihood estimation

Estimation af μ_0:

I balanceret tilfælde bare det fælles gennemsnit

I ubalanceret tilfælde løsning til

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

hvor

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

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

Maksimum- likelihood estimation

Estimation af σ^2 og σ_0^2

Sædvanligvis benyttes REML-estimation (residual maximum likelihood), baseret på residualerne, - og som tager hensyn til at man har estimeret μ_0.

PROC MIXED

Proceduren MIXED er specielt rettet mod modeller med "tilfældige effekter"

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


Estimation (prædiktion) af enkelte ballemiddelværdier

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

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

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

der er nul for

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

Estimation (prædiktion) af enkelte ballemiddelværdier

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

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

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

$$w_i(\gamma) = \frac{1}{1 + n_i\gamma} \tag{0.2}$$

Løsningen kaldes BLUP-estimatet (best linear unbiased predictor) se fx
http://swbzone.com/blup/blup2002.htm

BLUP estimat som empirisk Bayes estimat

Indskud om betingede og marginale fordelinger
Vi starter med marginal fordeling af Y (baller), $f_Y(y)$ og betinget fordeling af X (målinger) for givet $Y = y$, $g(x|y)$.

Simultan tæthed for X og Y,

$$f_{xy}(x, y) = f_Y(y)g(x|y)$$

Marginal fordeling af X

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

Den betingede fordeling af Y givet $X = x$ har tætheden

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

$f_Y(y)$ kaldes **aprioritfordeling**, $g(x|y)$ for stikprøvefordeling, og $h(y|x)$ **aposterioritfordeling**

Apriori, aposteriori for normalfordeling

$\mu_i \sim \mathsf{N}(\mu_0, \sigma_b^2)$ og $Y_{ij}|\mu_i \sim \mathsf{N}(\mu_i, \sigma^2)$

Da er aposteriorifordeling af μ_i efter observation af $\bar{y}_i$ en normalfordeling med:

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

og

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

Apriori, aposteriori for normalfordeling

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

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

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

aposterioripræcision er summen af aprioripræcision og stikprøvepræcision

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

Empirisk Bayes

Hvis vi indsætter estimater for "aprioritordelingens parametre", μ_0, σ^2 og σ_b^2 kaldes løsningen en **Empirisk Bayes løsning**. Den svarer altså til BLUP-løsningen.

I PROC MIXED

Kodeordet OUTP i modelsætningen angiver, at man får udskrevet estimaterne (prædikterede værdier) for de enkelte baller i det valgte datasæt.

```
PROC MIXED DATA = DROP METHOD=REML ASYCOV ;
CLASS balle prove;
MODEL renh = / OUTP = pred ;
RANDOM balle ;
RUN;
```