

1 Flere rotter

Datasættet **Metal** indeholder resultaterne fra en forsøgsserie, der havde til formål at bestemme toxiciteten af et metalsalt (Nikkel).

Ved forsøget benyttede man fem grupper af hunrotter, hver bestående af 25-28 rotter, der igennem en periode blev indgivet dette metalsalt oralt i doser af hhv 0, 1, 2.5, 5 og 10 mg/kg/dag igennem 10 uger før parring og igennem drægtighedsperioden. Efter fødslen opgjorde man for hvert dyr kuldstørrelsen og antallet af døde unger i kullet (indtil 4 dage efter fødslen).

Man ønsker nu at vurdere den eventuelle sammenhæng mellem den indgivne dosis og andelen af døde unger.

Så man prøver den sædvanlige logistiske regression - men goodness of fit er ikke tilfredsstillende. Der er overdispersion. En sådan overdispersion kan eventuelt forklares ved at der - selv for samme dosis - er en vis variation mellem dødssandsynligheden for de forskellige kuld.

Foretag en analyse af sammenhængen under hensyn til denne overdispersion.

2 Opgave i hierarkiske normalfordelingsmodeller

Opgaver vedrører de uldballer, der blev betragtet i notens afsnit om hierarkiske normalfordelingsmodeller.

Formålet med opgaven er at opnå en vis grad af fortrolighed med sådanne simple hierarkiske modeller, og deres behandling ved SAS PROC MIXED.

Opgaven skal således opfattes som en regneøvelse, snarere end en opgave i dataanalyse.

Data ligger i datasættet `uldopg`. Data er en anden ubalanceret delmængde af det fulde datasæt `uld`, end den delmængde, der blev betragtet i noten. Nu er nemlig de to sidste prøver fra balle 5 og den sidste prøve fra balle 6 blevet væk.

Vi betragter modellen for j 'te prøve fra i 'te balle.

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \quad (1)$$

hvor $A_i \sim N(0, \sigma_b^2)$ er uafhængige, og $\epsilon_{ij} \sim N(0, \sigma^2)$ er indbyrdes uafhængige, og uafhængige af A_i .

Vi vil bruge symbolet γ for forholdet $\gamma = \sigma_b^2 / \sigma^2$.

2.1 PROC GLM

Spørgsmål:

Bestem estimater for σ^2 og σ_b^2 ved brug af momentmetoden (som netop er den, der lægges op til i PROC GLM).

Man kan bruge programstumpen:

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

2.2 PROC MIXED

Spørgsmål:

Bestem estimater for σ^2 og σ_b^2 ved brug af residual maximum likelihoodmetoden (som netop er standardvalget i PROC MIXED, `METHOD = REML`) og sammenlign med estimatet fra PROC GLM.

Spørgsmål:

Og bestem den approximative varians for disse estimater og deres kovarians,

ASYCOV

Spørgsmål:

Hvorfor er der mon en kovarians imellem disse to variansestimater ?

Man kan bruge programstumpen:

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

Prøv også at estimere forholdet γ RATIO ved at bruge programstumpen

```
PROC MIXED DATA = DROP METHOD=REML RATIO ;  
  CLASS balle prove;  
  MODEL renh = ;  
  RANDOM balle ;  
  RUN;
```

2.3 Estimation af ballemiddelværdier

2.3.1 En ny balle

Optionen OUTPM = `predmean` under MODEL-sætningen bestemmer den præsikterede middelværdi for hver observation svarende til værdien af de systematiske effekter, der indgår i modelsætningen. Her er der kun intercept-leddet, så det er jo samme præsiktion for alle observationer. De præsikterede værdier placeres i datasættet `predmean` sammen med diverse informative størrelser.

Kør programstumpen:

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

og kig på datasættet `predmean`

Spørgsmål:

Er den prædikterede middelværdi blot gennemsnittet af alle observationer ?
Hvorfor er den næsten, men ikke helt ?

Datasættet indeholder en variabel `Std Err Pred`, som angiver den estimerede spredning på dette “fælles gennemsnit”.

Når bortses fra ubalancen har vi et gennemsnit af ballegennemsnittene fra hver af de syv baller med ca. 4 prøver i hver. Disse ballegennemsnit er uafhængige med variansen, $\sigma_b^2 + \sigma^2/n_i$, dvs variansen på gennemsnittet af de syv ballegennemsnit er

$$(\sigma_b^2 + \sigma^2/4)/7$$

Spørgsmål:

Verificer, at `Std Err Pred` svarer til denne varians.

Datasættet indeholder en variabel `Resid`, som angiver afvigelsen mellem observationen og den prædikterede interceptværdi.

Spørgsmål:

Tegn et Q-Q plot for disse residualer.

Er residualerne fra de syv baller spredt tilfældigt over plottet ?

Hvorfor ikke ?

2.3.2 En af de undersøgte baller

Optionen `OUTP = pred` under `MODEL`-sætningen bestemmer den prædikteret værdi for hver observation svarende såvel til værdien af de systematiske effekter, der indgår i modelsætningen, som de tilfældige effekter, der indgår i `RANDOM`-sætningen. Her altså de prædikterede balleværdier. De prædikterede værdier placeres i datasættet `pred` sammen med diverse informative størrelser.

Kør programstumpen:

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

RANDOM balle ;
RUN;

og kig på datasættet `pred`

Spørgsmål:

Er de præsikterede balleværdier blot gennemsnittet af prøverne fra ballen ?

Er den nærmere, eller fjernere, end det fælles gennemsnit ?

Efterprøv, at den stort set er et vejet gennemsnit mellem $\hat{\mu}$ og $\bar{y}_i$ med vægtene hhv

$$w = \frac{1}{1 + n_i \gamma}$$

på $\hat{\mu}$ og $1 - w$ på $\bar{y}_i$ (jvf side 26 midt).

Prøv evt at sammenligne prædiktions (og usikkerhed) for baller med 2, 3 og fire prøver.

Spørgsmål:

Sammenlign `Std Err Pred` med spredningen på ballegennemsnittet,

$$\sqrt{\sigma^2/n_i}$$

og spredningen bestemt udfra aposteriorivariansen σ_{apost}^2 midt på side 233,

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

Spørgsmål:

Tegn et Q-Q plot for residualerne svarende til disse præsikterede balleværdier
Er residualerne fra de syv baller spredt tilfældigt over plottet ?

Hvorfor mon ?