Andre obligatoriske oppgave i STK3100/4100
Høst 2011

Utlevering: Torsdag 25. Oktober

Innleveringsfrist: Torsdag 8. november, kl. 14:30

Besvarelsen innleveres ved ekspedisjonen i 7. etasje, Niels Henrik Abels hus

Dette er det andre settet med obligatoriske innlevering i STK3100/4100 høsten 2012. Oppgavesettet består av 2 oppgaver. For STK3100 studenter er kun oppgave 1 nødvendig å levere inn. For STK4100 studenter *må* begge oppgaver leveres inn. Det er dog mulig (og anbefalt) for STK3100 studenter også å gjøre oppgave 2.

Det er valgfritt om du vil skrive besvarelsen for hånd eller om du vil bruke et tekstbehandlingsprogram. Der du bruker **R** (eller et annet program), må utskrifter legges ved eller limes inn. Hvis flere studenter samarbeider om å løse oppgavene, må likevel hver student levere sin selvstendige besvarelse. Det må gå fram av besvarelsen hvem du har samarbeidet med. Se ellers "Regelverk for obligatoriske oppgaver" som er gitt på kursets hjemmeside.

Oppgave 1

I denne oppgaven skal vi se på et datasett som måler vekt av barn fra Nepal. Datasettet, som kan leses inn i **R** med følgende kommando

```
nepali = read.table("nepali.txt",header=T)
```

har følgende variable:

- `id` Et tall som angir identitet for barnet
- `sex` 1 = male; 2 = female
- `wt` Barns vekt målt i kilogram (vår responsvariabel)
- `mage` Mors alder i år
- `lit` Indikator av mors lese/skrivekyndighet: 0 = no; 1 = yes
- `died` Antall barn mor har hatt som døde
- `alive` Antall barn mor har født levende
- `age` Alder på barn

Vi vil i denne oppgaven se hvordan blandede modeller kan benyttes for å ta hensyn til avhengigheter i data innen individer.

Det kan være en god ide å se på `RIKZ.R` koden fra forelesning for å løse denne oppgaven.

(a) Start med en modell som inneholder alle forklaringsvariable og så mange interaksjonsledd som mulig (som ikke gir feilmeldinger). Prøv deretter ut følgende modeller for tilfeldige effekter

(i) Ingen tilfeldige effekter (utenom det vanlige residual-leddet)

(ii) Tilfeldig effekt i konstantledd

(iii) Tilfeldig effekt i konstaktledd og i stigningskoeffisient for `age`

Velg den struktur på de tilfeldige effekter som ser ut til å passe best med data.

(b) Finn så den optimale struktur for de faste effekter.

(c) Gjør en validering av din endelige modell.

Oppgave 2 (ANOVA and LMM models)
Consider a single factor ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, ..., I, j = 1, ..., J \qquad (*)$$

where for identifiability we impose the constraints $\sum_{i=1}^{I} \alpha_i = 0$. Important quantities when analyzing such models are

$$\text{SSTr} = J \sum_{i=1}^{I} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{i\cdot})^2$$

where $\bar{y}_{i\cdot} = \frac{1}{J} \sum_{j=1}^{J} y_{ij}$ and $\bar{y}_{\cdot\cdot} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$.

(a) By looking at textbooks from earlier courses or by your own calculations, find the expectations of SSTr and SSE under model (*).

Explain how these quantities can be used to test the hypothesis H_0 : $\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$.

(b) An alternative random effects formulation of the model above is

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, ..., I, j = 1, ..., J \quad (**)$$

where $A_i \overset{iid}{\sim} N(0, \sigma_A^2)$ and all A_i's independent of all ε_{ij}'s.

Show that this is a special case of a linear mixed model (LMM).

(c) Find the expectations of SSTr and SSE under model (**).

Hint: For finding the expectation of SSE, show that this quantity does not depend on the α_i's and argue why you then can use results from (a). For finding the expectation of SSTr, show first that $\overline{Y}_{i\cdot}$ are iid and Gaussian.

(d) Show that

$$\hat{\sigma}^2 = \frac{\text{SSE}}{I(J-1)}$$
$$\hat{\sigma}_A^2 = \frac{\text{SSTr}}{J(I-1)} - \frac{\hat{\sigma}^2}{J}$$

are unbiased estimates for σ^2 and σ_A^2, respectively.

We will now consider a simulation study, where we generate data according to model (**) and explore the behavior of the unbiased estimates as well as the ML estimates (obtained by using `lme`). For this part, there is an **R** script `sim_raneff.R` available from the course home-page which performs the simulations for you.

(e) Discuss the two plots you obtain. In particular, comment on the cases where $\hat{\sigma}_A^2 < 0$. Based on these plots, give an approximate relationship between the unbiased estimates and the ML estimates.

(f) Now modify the script so that $\sigma_A^2 = 0$. How many of the simulations result in that $\hat{\sigma}_A^2 < 0$. Up to 4 digits of precision, what are the values of -2LR for these simulations?

(g) Make a histogram of -2LR for those simulations corresponding to $\hat{\sigma}_A^2 > 0$. Compare this histogram with a χ_1^2 density.

(h) Discuss these results related to the general result that -2LR for testing $H_0 : \sigma_A^2 = 0$ is approximately a mixture of a χ_0^2 and a χ_1^2 distribution (with equal weight on each), where here χ_0^2 is a distribution putting all weight in 0.

Modify the script to include a test on H_0 based on LR. How many times is H_0 rejected?

(i) An alternative to likelihood ratio tests in this case is to use an F test directly through the use of SSE and SSTr. Devore and Beck (2007) state that under H_0,

$$F = \frac{SSTR/(I-1)}{SSE/(I(J-1))}$$

follows an F-distribution with $I-1$ and $I(J-1)$ degrees of freedom.

Include in the script a test based on this approach and compare with the LR test.

Comment on the results.

Remarks: Although the F test is easier to use in this case, such a test will not be possible to use in more general settings such as unbalanced designs and/or nonlinear models. In such cases, likelihood ratio tests needs to be used.