Reeksamen i Statistik 2, 24. august 2017

Vejledende besvarelse

Opgave 1

1. Variablene X_i er uafhængige og identisk Bernouilli-fordelte med $P(X_i = 1) = (1/2)^{\alpha}$. Det følger specialt af Den Centrale Grænseværdisætning ... og EH eksempel ... , at $\hat{\theta}_n \overset{as}{\sim} \mathcal{N}((1/2)^{\alpha}, \frac{(1/2)^{\alpha}(1-(1/2)^{\alpha}}{n})$.

2. Benyttes Deltametoden med $f(y) = -\frac{\log(y)}{\log(2)}$ fås, at $\tilde{\alpha}_n = f(\hat{\theta}_n)$ er asymptotisk normalfordelt med middelværdi $f((1/2)^{\alpha}) = \alpha$ og asymptotisk varians $\frac{(1/2)^{\alpha}(1-(1/2)^{\alpha})}{(-(1/2)^{\alpha}\log(2))^2}$.

3. Tætheden for fordelingen af Y_1 kan skrives på formen

$$f_{\alpha}(y) = \alpha y^{\alpha-1} = \frac{1}{1/\alpha}\exp(\alpha \cdot \log(y)) \cdot \frac{1}{y},$$

hvoraf det ses, at tætheden er en eksponentiel familie. Den kanoniske stikprøvefunktion er $\theta = \frac{1}{\alpha}$ og normeringskonstanten er $c(\theta) = \frac{1}{\theta}$.

4. Likelihoodfunktion

$$\begin{aligned} L_{Y_1,\ldots,Y_n}(\alpha) &= \prod_{i=1}^{n} \alpha Y_i^{\alpha-1} \\ &= \alpha^n \left(\prod_{i=1}^{n} Y_i^{\alpha-1} \right) \end{aligned}$$

(Minus) loglikelihoodfunktion

$$l_{Y_1,\ldots,Y_n}(\alpha) = c - n\log(\alpha) - \alpha \cdot \sum_i \log(Y_i)$$

Scorefunktion

$$l'_{Y_1,\ldots,Y_n}(\alpha) = -\frac{n}{\alpha} - \sum_i \log(Y_i)$$

5. Den observerede information

$$l''_{Y_1,\ldots,Y_n}(\alpha) = \frac{n}{\alpha^2}$$

er strengt positiv, hvorfor en eventuel løsning til likelihoodligningen vil være et globalt minimum for $l_{Y_1,\ldots,Y_n}(\alpha)$. Løses likelihoodligningen fås følgende udtryk for maksimaliseringsestimatoren $\hat{\alpha}_n = -\frac{n}{\sum_i \log(Y_i)}$.

Den asymptotiske fordeling af MLE kan bestemmes enten ved at kombinere Den Centrale Grænseværdisætning (anvendt på $\frac{\sum_i \log(Y_i)}{n}$)) med Deltametoden eller vha. Cramér's sætning (EH: Sætning 5.23). Vi konkluderer, at $\hat{\alpha}_n \overset{as}{\sim} \mathcal{N}(\alpha, \frac{\alpha^2}{n})$.

Opgave 2

1. Den eneste model i R udskriften, der indeholder den relevante vekselvirkning beskriver $X = (X_i)_{i\in I}$ som regulært normalfordelt på $\mathbb{R}^I$ med middelværdi $\xi \in L_T + L_{G\times V}$ og varians $\sigma^2 I$. Den additive hypotese, $H_0 : \xi \in L_T + L_G + L_V$ kan testes ved F-teststørrelsen $F = 0.7249$, der under H_0 følger en F-fordeling med $(2,9)$-frihedsgrader. P-værdien ses at være 0.5106, hvorfor vi accepterer nulhypotesen. Det er ikke et krav at model og hypotese opskrives, men angivelse af teststørrelse, P-værdi og konklusion anses som minimum for en fuldstændig besvarelse. Resultaterne er aflæst i R-udskriften efter `anova(mod2, mod1)`.

2. MLE for parametrene i middelværdistrukturen er givet ved formlen $\hat{\beta} = (A^T A)^{-1} A^T X$, hvor A er designmatricen for den valgte parametrisering af den additive model med alle tre faktorer (`mod2`. Det følger af EH korollar 10.21, at $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$, hvor $A^T A$ og dens inverse er anført i R-udskriften.

   Der er i R-udskriften angivet 7 parameterestimater svarende til middelværdistrukturen. Det første estimat (=522.39) svarer til referencegruppen `tomat = 1, G = G1, V = V1`, mens det øvrige 6 estimater angiver forskellen på det forventede udbytte i forhold til referencegruppen, hvis man ændrer en af de tre faktor `tomat, G` eller `V`.

3. Fordelingen af MLE for variansen er ifølge EH korollar 10.21 givet ved $\hat{\sigma}^2 \sim \chi^2-$ fordelt med $N - k = 18 - 7 = 11$ frihedsgrader og skalaparameter σ^2/N.

4. En designgraf viser, at der er ikke-trivielt minimum mellem faktorerne gødning (`G`) og vanding (`V`), som har to niveauer. Faktoren angiver, om plantekassen har modtaget en behandling eller ej. Ved at krydstabellere datasættet mht. de to faktorer ses, at hvor af de fremkomne diagonalblokke opfylder balanceligningen. Dermed er de to faktorer geometrisk ortogonale.

5. Den simple løsning består i at anvende formel (13.3) i EH, hvor man bemærker, at der er to sammenhængskomponenter i designgrafen for de to faktorer.

   Alternativt kan bemærkes, at mængden

   $$\mathbb{G} = \{V, G, V \wedge G, 1\}$$

   udgør et geometrisk ortogonalt design, som er afsluttet over for dannelse af minimum. Dimensionerne af V_G-rummene, $G \in \mathbb{G}$, fra sætningen om den ortogonale dekomposition (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

   $$\begin{aligned} dim(L_V + L_G) &= dim(V_V) + dim(V_G) + dim(V_{V\wedge G}) + dim(V_1) \\ &= 1 + 1 + 1 + 1 = 4. \end{aligned}$$

   Det kan her være nyttigt at støtte sig op ad et faktorstrukturdiagram, for at holde styr på ordningen af faktorerne

$$
\begin{tikzcd}
 & \mathtt{V}_1^3 \arrow[dr] & & \\
\mathtt{I}_5^9 \arrow[ur] \arrow[dr] & & \mathtt{G} \wedge \mathtt{V}_1^2 \arrow[r] & \mathtt{1}_1^1 \\
 & \mathtt{G}_1^3 \arrow[ur] & &
\end{tikzcd}
$$

Opgave 3

1. Et faktorstrukturdiagram ser ud som følger

$$\mathtt{I}_{56}^{70} \longrightarrow \mathtt{person}_{12}^{14} \longrightarrow \mathtt{morgen}_1^2 \longrightarrow \mathtt{1}_1^1$$

2. Modellen kan udtrykkes ved at $X = (X_i)_{i \in I}$ er normalfordelt på $\mathbb{R}^{70}$ med

$$\xi_i = EX_i = \alpha + \beta \cdot \mathtt{puls}_i$$

og $VX = \sigma^2 I + \nu^2 BB^T$. Her er B effektmatricen hørende til effektparret $(\mathtt{person}, 1)$.

Da der i opgaveformuleringen blot lægges op til, at man skal modellere sammenhængen mellem `tid` og `puls`, så er det helt ok, hvis man ved besvarelsen af delopgave 2.-4. i stedet benytter `puls` som responsvariabel og `tid` som forklarende variabel. Bemærk dog, at det i delopgave 5. implicit fremgår, at det gennem hele opgaven er tanken, at `tid` skal benyttes som responsvariabel.

3. Parametrene i middelværdistrukturen fremgår af

```r
data <- read.table("Stat2aug2017opg3.txt", header = T)
library(lme4)
```

```r
head(data) ## NB: variablen 'lap' skal ikke benyttes!

##   morgen lap puls tid person
## 1     ja   1  143 445     P1
## 2     ja   2  156 431     P1
## 3     ja   3  156 428     P1
## 4     ja   4  165 383     P1
## 5     ja   5  163 401     P1
## 6     ja   1  154 429     P2

m1 <- lmer(tid ~ puls + (1|person), data = data)
summary(m1)$coefficients

##               Estimate Std. Error   t value
## (Intercept) 902.387649 26.5355840  34.00670
## puls         -3.058166  0.1592837 -19.19949
```

og disse estimeres til $\hat{\alpha} = 902.4$ og $\hat{\beta} = -3.058$.

Variansparametrene estimeres til $\hat{\sigma}^2 = 9.4572^2 = 89.4$ og $\hat{\nu}^2 = 14.7736^2 = 218.3$ hvilket fremgår af udskriften

```
VarCorr(m1)

##  Groups   Name        Std.Dev.
##  person   (Intercept) 14.7736
##  Residual              9.4572
```

4. Regressionsmodellen kan udvides ved at ændre middelværdistrukturen således at skæring og hældning tillades at afhænge af, om personen har løbet sine ture om morgenen eller om eftermiddagen svarende til

$$\xi_i = EX_i = \alpha(\texttt{morgen}_i) + \beta(\texttt{morgen}_i) \cdot \texttt{puls}_i.$$

Modellen kan testes direkte imod modellen fra delopgave 2., hvorved vi finder at

```
m3 <- lmer(tid ~ morgen + morgen:puls + (1|person) - 1, data = data)
anova(m1, m3)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## m1: tid ~ puls + (1 | person)
## m3: tid ~ morgen + morgen:puls + (1 | person) - 1
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m1  4 555.27 564.27 -273.64   547.27
## m3  6 558.52 572.01 -273.26   546.52 0.7542      2     0.6858
```

Likelihoodratioteststørrelsen bliver $LRT = 0.7542$ der ved et opslag i en tabel over χ^2-fordelingen med 2 frihedsgrader giver et P-værdi på 0.6858. Det lader således ikke til, at der er forskel på sammenhængen mellem `tid` og `puls` for personer der løber om morgenen og senere på dagen.

Testet kan alternativt udføres i to trin, hvor man først tester om hældning og dernæst om skæringen kan antages at være ens for de to niveauer af faktoren `morgen`. R kode og resultater fremgår nedenfor (den overordnede konklusion ændres ikke!)

```
m2 <- lmer(tid ~ puls + (1|person) + morgen, data = data)
anova(m3, m2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
```

```
## m2: tid ~ puls + (1 | person) + morgen
## m3: tid ~ morgen + morgen:puls + (1 | person) - 1
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m2  5 556.68 567.92 -273.34   546.68
## m3  6 558.52 572.01 -273.26   546.52 0.1586      1     0.6905

anova(m1, m2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## m1: tid ~ puls + (1 | person)
## m2: tid ~ puls + (1 | person) + morgen
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m1  4 555.27 564.27 -273.64   547.27
## m2  5 556.68 567.92 -273.34   546.68 0.5956      1     0.4402
```

5. Kovariansmatricen for de 5 målinger af omgangstiden fra person `P1` er givet ved

$$\begin{pmatrix} \sigma^2+\nu^2 & \nu^2 & \nu^2 & \nu^2 & \nu^2 \\ \nu^2 & \sigma^2+\nu^2 & \nu^2 & \nu^2 & \nu^2 \\ \nu^2 & \nu^2 & \sigma^2+\nu^2 & \nu^2 & \nu^2 \\ \nu^2 & \nu^2 & \nu^2 & \sigma^2+\nu^2 & \nu^2 \\ \nu^2 & \nu^2 & \nu^2 & \nu^2 & \sigma^2+\nu^2 \end{pmatrix}$$

Supplerende R kode til løsning af opgave 3

Følgende R kode kan benyttes til analyserne, hvis `puls` benyttes som responsvariablen i delopgave 2.-4.

```
l1 <- lmer(puls ~ tid + (1|person), data = data)
summary(l1)$coefficients

##                Estimate Std. Error   t value
## (Intercept) 274.8088601 5.85128704  46.96554
## tid          -0.2762089 0.01431843 -19.29045

VarCorr(l1)

##  Groups   Name        Std.Dev.
##  person   (Intercept) 4.5444
##  Residual             2.8268

l3 <- lmer(puls ~ morgen + morgen:tid + (1|person) - 1, data = data)
anova(l1, l3)
```

```
## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l1: puls ~ tid + (1 | person)
## l3: puls ~ morgen + morgen:tid + (1 | person) - 1
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## l1  4 386.95 395.94 -189.47   378.95
## l3  6 389.37 402.87 -188.69   377.37 1.5761      2     0.4547

l2 <- lmer(puls ~ tid + (1|person) + morgen, data = data)
anova(l3, l2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l2: puls ~ tid + (1 | person) + morgen
## l3: puls ~ morgen + morgen:tid + (1 | person) - 1
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## l2  5 387.92 399.16 -188.96   377.92
## l3  6 389.37 402.87 -188.69   377.37 0.5436      1     0.4609

anova(l1, l2)

## refitting model(s) with ML (instead of REML)

## Data: data
## Models:
## l1: puls ~ tid + (1 | person)
## l2: puls ~ tid + (1 | person) + morgen
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## l1  4 386.95 395.94 -189.47   378.95
## l2  5 387.92 399.16 -188.96   377.92 1.0325      1     0.3096
```