Eksamen i Matematisk Statistik, 20. juni 2019

Steffen Lauritzen og Anders Tolver

Vejledende besvarelse, 8. juli 2019

Opgave 1

1. Vi omskriver tætheden som

$$\begin{aligned} f_{(1,\lambda)}(x) &= \exp\left\{\lambda\frac{-(x-1)^2}{2x}+\frac{1}{2}\log\lambda\right\}\frac{1}{\sqrt{2\pi x^3}} \\ &= \exp\left\{-\lambda(x+1/x)/2+\lambda+\frac{1}{2}\log\lambda\right\}\frac{1}{\sqrt{2\pi x^3}} \end{aligned}$$

2. Man kan fx direkte skrive log-likelihood funktionen som

$$\ell(\lambda) = -n\lambda - \frac{n}{2}\log\lambda + \frac{\lambda}{2}\sum_{i=1}^{n}(X_i + 1/X_i)$$

   og differentiere, hvilket giver

$$S(\lambda) = -n - \frac{n}{2\lambda} + \frac{n}{2}(\bar{X}_n + \bar{Y}_n)$$

   hvoraf resultatet følger.

3. Vis at informationen $i_0(\lambda)$ for λ i delfamilien $\mathscr{P}_0$ baseret på en enkelt observation er bestemt som

$$i_0(\lambda) = \frac{1}{2\lambda^2}$$

   og angiv den asymptotiske fordeling af $\hat{\lambda}_n$.

   I en eksponentiel familie som er kanonisk parametriseret, er informationen lig med den anden afledede af kumulantfunktionen

$$i_0(\lambda) = \psi''(\lambda) = I(\lambda) = S'(\lambda) = \frac{1}{2\lambda^2}.$$

   Den asymptotiske fordeling af $\hat{\lambda}_n$ er $N(\lambda, \lambda^2/2n)$.

4. Vi omskriver eksponenten som

$$-\frac{\lambda(x-\mu)^2}{2\mu^2 x} = -\frac{\lambda}{2\mu^2}x - \frac{\lambda}{2x} + \frac{\lambda}{\mu} = \theta_1\frac{-x}{2} + \theta_2\frac{-1}{2x} + \sqrt{\theta_1\theta_2}$$

og tilsvarende faktoren inden eksponentialfunktionen som

$$\sqrt{\frac{\lambda}{2\pi x^3}} = e^{\frac{1}{2}\log\theta_2}\frac{1}{2\pi x^3}$$

hvorfor hele tætheden kan skrives som

$$f_{\mu,\lambda}(x) = \exp\{\theta^T t(x) - \psi(\theta)\}\cdot\frac{1}{2\pi x^3}$$

så grundmålet er

$$d\mu(x) = \frac{1}{2\pi x^3}\mathbf{1}_{(0,\infty)}\,dx$$

hvor dx betegner standard Lebesgue mål.

5. Vi differentierer kumulantfunktionen og finder

$$\mathbf{E}(-X/2) = \frac{\partial}{\partial\theta_1}\psi(\theta) = -\frac{1}{2}\sqrt{\frac{\theta_2}{\theta_1}} = -\frac{\mu}{2}$$

og videre

$$\mathbf{V}(-X/2) = \frac{\partial^2}{\partial\theta_1^2}\psi(\theta) = \frac{1}{4\theta_1}\sqrt{\frac{\theta_2}{\theta_1}} = \frac{1}{4}\frac{\mu^3}{\lambda}$$

hvoraf resultatet følger.

6. I en regulær eksponentiel familie finder vi MLE ved at sætte de kanoniske stikprøvefunktioner lig med deres middelværdi. Vi har allerede fundet $\mathbf{E}_{\mu,\lambda}(-X/2) = -\mu/2$, hvoraf $\tilde{\mu}_n = \bar{X}_n$. Vi skal bruge middelværdien af den anden komponent:

$$\mathbf{E}(-1/(2X)) = \frac{\partial}{\partial\theta_2}\psi(\theta) = -\frac{1}{2}\sqrt{\frac{\theta_1}{\theta_2}} - \frac{1}{\theta_2}$$

hvoraf

$$\mathbf{E}(Y) = \mathbf{E}(\frac{1}{X}) = \frac{1}{\mu} + \frac{1}{\lambda}.$$

Indsættes $\tilde{\mu}_n = \bar{X}_n$ fås derfor

$$\bar{Y}_n = \frac{1}{\bar{X}_n} + \frac{1}{\tilde{\lambda}_n}$$

og resultatet fremkommer nu ved at løse ligningen. Ligningen har en entydig løsning hvis og kun hvis $\bar{X}_n\bar{Y}_n > 1$, hvilket sker med sandsynlighed 1 hvis $n \geq 2$ idet ikke to værdier af X_i så er ens.

7. Delfamilien svarende til H_0 har dimension 1 og er en krum delfamilie af $\mathscr{P}$ med parametriseringen $(\theta_1,\theta_2) = (\beta,\beta)$ idet $\mu = 1 \iff \theta_1 = \theta_2$. Derfor vil Λ_n være asymptotisk χ^2-fordelt med $2-1=1$ frihedsgrad.

Opgave 2

1. Ifølge EH Korollar 9.43 er X regulært normalfordelt hvis og kun hvis variansen Σ er invertibel. Det ses at determinanten af Σ er lig med

$$2\cdot1\cdot2+1\cdot1\cdot1+1\cdot1\cdot1-2\cdot1\cdot1-1\cdot1\cdot2-1\cdot1\cdot1=1,$$

hvorfor Σ er invertibel. Vi konkluderer at X er regulært normalfordelt.

2. Det følger af EH Lemma 9.47 at $Y = (Y_1, Y_2)^T$ er normalfordelt med middelværdi

$$EY = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

og varians

$$\Sigma_Y = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \Sigma \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Dette viser, at Y_1 og Y_2 er uafhængige og standardnormalfordelte. Kvadratet Y_2^2 på en standardnormalfordelt variabel er per definition χ^2-fordelt med 1 frihedsgrad.

Da $Y_1 \sim N(0,1)$ og uafhængig af Y_2^2 som er χ_1^2-fordelt, så er forholdet $\frac{Y_1}{\sqrt{Y_2^2}}$ t-fordelt med 1 frihedsgrad. Denne konstruktion har fx. været benyttet i forbindelse med udledning af konfidensintervaller og prædiktionsintervaller i den lineære normale model, eller mere direkte i HS opgave 9. Resultatet er også kendt fra MI eksempel 20.27.

Opgave 3

1. Vi ser at faktorerne `G` og `time` opfylder EH Sætning 14.8, hvorfor de er geometrisk ortogonale. I praksis er dette en konsekvens af, at der er foretaget to målinger for *alle* børn.

```
data <- read.table(file = "MatStatJuni2019.txt", header = T)
table(data$G, data$time)

##
##           after before
##   B/C        41     41
##   Control    36     36
##   N          42     42
##   P/S        39     39
##   S/J        30     30
```

   Faktoren `subj` er finere end `G`, hvorfor $\texttt{subj} \wedge \texttt{G} = \texttt{subj}$.

2. Modellen kan udtrykkes ved at vektoren $X = (X_i)_{i \in I}$ bestående af målinger af børnenes `liking` er normalfordelt på $\mathbb{R}^{376}$ med $\xi = EX \in L_{\texttt{G} \times \texttt{time}}$ og varians $VX = \sigma^2 + v_1^2 B_1$, hvor B_1 er effektmatricen hørende til parret $(\texttt{subj}, \texttt{1})$.

   Kovariansmatricen for de 2 målinger på samme barn kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 \end{pmatrix}.$$

3. Modellen fittes i R med følgende kode (angivelse af kode ikke påkrævet)

```
library(lme4)

## Error in library(lme4): there is no package called 'lme4'

mod1 <- lmer(liking ~ G * time + (1|subj), data = data)

## Error in lmer(liking ~ G * time + (1 | subj), data = data): could not
find function "lmer"

mod1

## Error in eval(expr, envir, enclos): object 'mod1' not found
```

   Estimaterne (REML!) for variansparametrene er $\hat{\sigma} = 1.4249$ og $\hat{v}_1 = 0.9154$. Den forventede liking for børn i gruppen $\texttt{G} = \texttt{B/C}$ bliver 4.9268 (after) og $4.9268 - 1.0244$ (before).

4. Modellen fra foregående delspørgsmål genfittes nu med en parametrisering, hvor man direkte kan aflæse ændring i liking (`before` til `after`) for hver af de 5 eksponeringsgrupper. På baggrund af estimater og konfidensintervaller for tilvæksterne kan man nu konkludere på effekten af interventionen (dvs. eksponering til pågældende snackbar) på ændring i `liking`.

```
mod1refit <- lmer(liking ~ G + G : time - 1 + (1|subj), data = data)

## Error in lmer(liking ~ G + G:time - 1 + (1 | subj), data = data): could
not find function "lmer"

confint(mod1refit)

## Error in confint(mod1refit): object 'mod1refit' not found
```

Vi ser, at der i eksponeringsgrupperne `B/C`, `P/S` og `S/J` sker signfikante stigninger i `liking` (baseret på et 5 % signifikansniveau).

Det er også muligt at konstruere et overordnet likelihoodtest for hypotesen om, at $\xi \in L_G$ svarende til at der ikke sker ændringer i `liking` for nogle af eksponeringsgrupperne.

```
mod2 <- lmer(liking ~ G + (1|subj), data = data)

## Error in lmer(liking ~ G + (1 | subj), data = data): could not find function
"lmer"

anova(mod2, mod1)

## Error in anova(mod2, mod1): object 'mod2' not found
```

Hypotesen forkastes med et brag. P-værdi < 0.0001 baseret på et opslag i en χ^2-fordeling med 5 frihedsgrader.

Opgave 4

1. Modellen er en ensidet variansanalysemodel, så middelværdiunderrummet er $L_{\texttt{race}}$ og har dimension 5 (=antal forskellige racer i datasættet). I R udskriften er benyttet en parametrisering med middelværdien for `race = Border_Terrier` og forskellene i forhold til denne gruppe. Estimaterne for parametrene til beskrivelse af middelværdivektoren bliver

$$\begin{aligned}
\hat{\alpha}_{\texttt{Border_Terrier}} &= 5.0812 \\
\hat{\alpha}_{\texttt{Grand_Danios}} &= 5.0812 + 30.3866 \\
\hat{\alpha}_{\texttt{Labrador}} &= 5.0812 + 13.8429 \\
\hat{\alpha}_{\texttt{Petit_Basset}} &= 5.0812 + 7.0097 \\
\hat{\alpha}_{\texttt{Whippet}} &= 5.0812 + 5.4599
\end{aligned}$$

   og (det centrale) variansestimat er $\tilde{\sigma} = 4.304$. MLE for variansen er $\hat{\sigma}^2 = \frac{92}{97}\tilde{\sigma}^2$. Lader vi $\xi = A\beta$, så giver EH Korollar 10.21 at $\hat{\beta} \sim N(\beta, \sigma^2(A^TA)^{-1})$ og at $\hat{\sigma}^2$ er χ^2-fordelt med 92 frihedsgrader og skalaparameter $\sigma^2/97$.

2. Ved at betragte residualplottet for `modA` ses, at variansen for de standardiserede residualer ser ud til vokse med størrelsen på de fittede værdier. Det lader derfor ikke til at antagelsen om, at residualerne er normalfordelte med samme varians er opfyldt, når man laver en regressionsanalyse af `maxLA` på `wgt`.

   På residualplottet for `modB` ses et mønster fx. med hensyn til residualernes placering i forhold til 0. Residualerne hørende til observationer med små fittede værdier er oftest negative, mens residualer knyttet til lidt større fittede værdier har en tendens til at være positive. Dette strider imod en antagelse om, at residualerne skal have samme middelværdi.

   Residualplottet for `modC` giver ikke umiddelbart anledning til at anfægte antagelserne om, at residualerne er normalfordelte med samme varians.

   QQ-plottet af de standardiserede residualer ligger for alle modeller rimelig pænt omkring en ret linje med hældning 1 og skæring 0.

   Regressionsmodellen `modC` udtrykker, at middelværdien af $\log(\texttt{maxLA}_i)$ (for den i-te hund i datasættet) er givet ved $\alpha + \beta \cdot \log(\texttt{wgt}_i)$.

3. Tager vi udgangspunkt i formel (10.42) fra EH Eksempel 10.31 med $\phi = (1, log(25))^T$ så kan vi beregne et 95 % - prædiktionsinterval for $\log(\texttt{maxLA})$ for en hund på 25 kg på baggrund af R-udskriften i opgaven.

```r
beta_hat <- c(-0.11931, 0.89163)
sigma_hat <- 0.2344
ATAinv <- matrix(nrow = 2, ncol = 2, byrow = T
                 , data = c(0.170702, -0.054205, -0.054205,  0.018319))
phi <- matrix(nrow = 2, ncol = 1, data = c(1, log(25)))
z_alpha <- qt(1 - 0.05/2, 97 - 2)
pred_low <- t(phi) %*% beta_hat - z_alpha *
  sqrt(sigma_hat^2 * (1 + t(phi) %*% ATAinv %*% phi))
pred_up <- t(phi) %*% beta_hat + z_alpha *
```

```
  sqrt(sigma_hat^2 * (1 + t(phi) %*% ATAinv %*% phi))
c(pred_low, pred_up)

## [1] 2.282714 3.218759
```

Dette kan omregnes til et 95 % - prædiktionsinterval for `maxLA` (målt i mL) ved tilbagetransformation med eksponentialfunktionen

```
exp(c(pred_low, pred_up) )

## [1]  9.803249 24.997072
```