Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Reeksamen i Statistik 2

25. august 2016

Eksamen varer 4 timer. Alle hjælpemidler er tilladt under eksamen, også computer, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af tre opgaver med i alt 22 delspørgsmål. De tre opgaver vægtes ens. Data til opgave 3 ligger i filen `benzin.txt` på en USB-stick. Sticken skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den kan altså ikke indgå som en del af besvarelsen.

Opgave 1

1. Betragt fordelingen med tæthed

$$f_p(x) = (x+1)p^x(1-p)^2 \quad \text{for } x \in \mathbb{N}_0$$

mht tællemålet på $\mathbb{N}_0$. Fordelingen afhænger af parameteren $p \in (0,1)$.

Du kan uden bevis benytte at f_p er en tæthed.

Lad $X_1, \ldots, X_n$ være uafhængige og identisk fordelte stokastiske variable med tæthed f_p, med ukendt $p \in (0,1)$.

(a) Reparametriser f_p-fordelingen ved $\theta = \log p$ så det fremgår at det er en eksponentiel familie med kanonisk stikprøvefunktion $t(x) = x$.

Solution: Da $p = e^\theta$ fås

$$\begin{aligned} f_p(x) &= (x+1)p^x(1-p)^2 \\ &= (x+1)e^{\theta x}(1-e^\theta)^2 \end{aligned}$$

Vi får således (se def. 2.13 i lærebogen)

$$\begin{aligned} f_\theta(x) &= (x+1)(1-e^\theta)^2 e^{\theta x} \\ &= \underbrace{(x+1)}_{\text{funktion af } x} \quad \underbrace{(1-e^\theta)^2}_{\text{funktion af } \theta} \quad \underbrace{e^{\theta x}}_{t(x)=x} \end{aligned}$$

(b) Identificer grundmålet. Identificer normeringskonstanten $c(\theta)$.

Solution: Lad μ have tæthed

$$(x+1) \quad \text{for } x \in \mathbb{N}_0$$

med hensyn til tællemålet. Udtrykt ved $\theta = \log p$ har X tæthed

$$(1-e^\theta)^2 e^{\theta x} \quad \text{for } x \in \mathbb{N}_0$$

med hensyn til μ. Der er altså tale om en en-dimensional eksponential familie på $\mathbb{N}_0$ med kanonisk stikprøvefunktion $t(x) = x$, grundmål μ og normeringskonstant

$$c(\theta) = (1-e^\theta)^{-2}.$$

(c) Argumenter for at X_1 har momenter af enhver orden. Find middelværdi og varians af X_1.

Solution: Ifølge Lemma 2.19 har $t(X_1) = X_1$ momenter af enhver orden. Bemærk at Lemma 2.19 udtaler sig om momenter af $t(X)$, ikke om momenter af X. Det er derfor vigtigt at den kanoniske stikprøvefunktion er identitetsfunktionen.

Ifølge (2.15) (eller formlerne lige over, eller Lemma 2.20) er

$$E(t(X_1)) = E(X_1) = \frac{d}{d\theta} \log c(\theta) = \frac{2e^\theta}{1-e^\theta}$$

og

$$\mathrm{Var}(t(X_1)) = \mathrm{Var}(X_1) = \frac{d^2}{d\theta^2} \log c(\theta) = \frac{2e^\theta}{(1-e^\theta)^2}$$

Udtrykt ved $p = e^\theta$ fås (ikke nødvendigt at angive i besvarelsen)

$$E(X_1) = \frac{2p}{1-p}$$

og

$$\mathrm{Var}(X_1) = \frac{2p}{(1-p)^2}$$

(d) Opskriv likelihoodligningen for θ. Find maksimaliseringsestimatoren for θ. Find maksimaliseringsestimatoren for p.

Solution: Likelihoodligningen for θ er givet i (5.6), det følger derfor fra resultaterne i (c) at likelihoodligningen er

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{2e^\theta}{1-e^\theta}$$

Løses denne fås maksimaliseringsestimatoren for θ:

$$\hat{\theta} = \log\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{2+\frac{1}{n}\sum_{i=1}^{n} X_i}\right)$$

Maksimaliseringsestimatoren for p:

$$\hat{p} = e^{\hat{\theta}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{2+\frac{1}{n}\sum_{i=1}^{n} X_i}$$

(e) Er maksimaliseringsestimatoren for θ veldefineret med sandsynlighed 1? Er maksimaliseringsestimatoren for θ asymptotisk veldefineret? Begrund dine svar.

Solution: Nej, maksimaliseringsestimatoren for θ er ikke veldefineret med sandsynlighed 1. Der er positiv sandsynlighed for at $X_i = 0$ for $i = 1, \ldots, n$. I det tilfælde er estimatoren ikke veldefineret, da man ikke kan tage logaritmen til 0.

Ja, maksimaliseringsestimatoren for θ er asymptotisk veldefineret. Sandsynligheden for at $X_i = 0$ for $i = 1, \ldots, n$ går mod 0 når $n \to \infty$, og derfor vil argumentet til logaritmen være et tal mellem 0 og 1 (da $X_i \geq 0$). Da $p \in (0,1)$ er $\theta < 0$, og vi ser at $\hat{\theta} < 0$ med sandsynlighed gående mod 1, og maksimaliseringsestimatoren er således asymptotisk veldefineret.

(f) Gør rede for at $\hat{\theta}$ er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametriseret ved θ.

Solution: Da $\mathrm{Var}(X_1) > 0$ benytter vi direkte nederste formel på side 187, og får at

$$\hat{\theta} \overset{as}{\sim} N\left(\theta, \frac{1}{n}\frac{(1-e^\theta)^2}{2e^\theta}\right)$$

Man kan også gå en lille omvej, og bruge at ifølge CLT, Sætning 5.11, er

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{as}{\sim} N\left(\frac{2e^\theta}{1-e^\theta}, \frac{1}{n}\frac{2e^\theta}{(1-e^\theta)^2}\right)$$

Vi har at

$$\hat{\theta} = f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \quad \text{hvor} \quad f(x) = \log\left(\frac{x}{2+x}\right).$$

For $x > 0$ er f differentiabel, herunder specielt for $x = E(X_1)$. Vi har $f'(x) = \frac{2}{x(2+x)}$ og $f'(E(X_1)) = \frac{(1-e^\theta)^2}{2e^\theta}$. Vi benytter deltametoden, og får

$$\begin{aligned}\hat{\theta} = f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &\overset{as}{\sim} N\left(f\left(\frac{2e^\theta}{1-e^\theta}\right), \frac{1}{n}\left(\frac{(1-e^\theta)^2}{2e^\theta}\right)^2 \frac{2e^\theta}{(1-e^\theta)^2}\right) \\ &= N\left(\theta, \frac{1}{n}\frac{(1-e^\theta)^2}{2e^\theta}\right)\end{aligned}$$

som før.

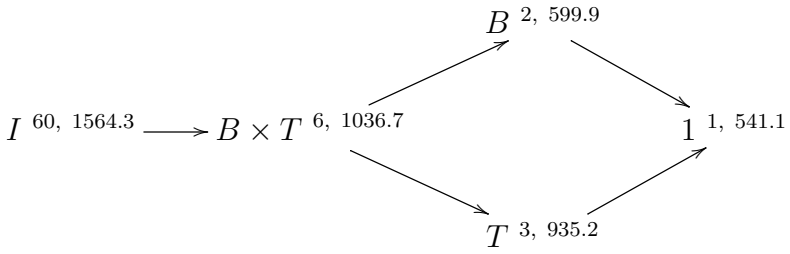
(g) Gør rede for at $\hat{p}$ er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametriseret ved p.

Solution: $\hat{p} = g(\hat{\theta})$ hvor $g(x) = e^x$ er målelig og differentiabel, $g'(x) = e^x$. Deltametoden giver

$$\hat{p} = g(\hat{\theta}) \overset{as}{\sim} N\left(e^\theta, \frac{1}{n}e^{2\theta}\frac{(1-e^\theta)^2}{2e^\theta}\right) = N\left(p, \frac{1}{n}\frac{p(1-p)^2}{2}\right)$$

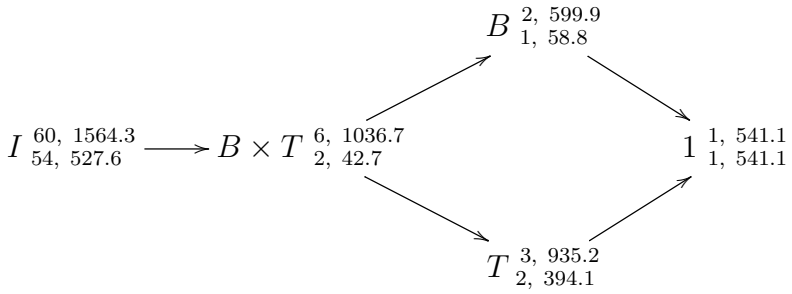
Opgave 2

2. Betragt de surjektive faktorer B og T, der antages at være usammenlignelige og geometrisk ortogonale, og deres tilhørende underrum L_B og L_T. Betragt det annoterede faktorstrukturdiagram:



(a) Udfyld resten af det annoterede faktorstrukturdiagram i den ortogonale dekomposition.

Solution:



(b) Test den additive hypotese op mod vekselvirkningsmodellen ved brug af det annoterede faktorstrukturdiagram.

Solution: Test af $B + T$ mod $B \times T$:

$$\begin{aligned} F &= \frac{\left(\|P_{B\times T}X\|^2 - \|P_{B+T}X\|^2\right)/(\dim \mathcal{L}_{B\times T} - \dim \mathcal{L}_{B+T})}{\left(\|X\|^2 - \|P_{B\times T}X\|^2\right)/(|I| - \dim \mathcal{L}_{B\times T})} \\ &= \frac{\|Q_{B\times T}X\|^2/\dim(Q_{B\times T})}{\|Q_I\|^2/\dim(Q_I)} = \frac{42.7/2}{527.6/54} = 2.18 \end{aligned}$$

der skal evalueres i en F-fordeling med 2 og 54 frihedsgrader. p-værdien bliver 0.123, og vi accepterer den additive hypotese.

(c) Test om der er en effekt af faktor B mod den additive hypotese ved brug af det annoterede faktorstrukturdiagram.

Solution: For at teste om der er effekt af faktor B testes T mod $B + T$:

$$\begin{aligned} F &= \frac{\left(\|P_{B+T}X\|^2 - \|P_TX\|^2\right)/(\dim \mathcal{L}_{B+T} - \dim \mathcal{L}_T)}{\left(\|X\|^2 - \|P_{B+T}X\|^2\right)/(|I| - \dim \mathcal{L}_{B+T})} \\ &= \frac{\|Q_BX\|^2/\dim(Q_B)}{(\|Q_I\|^2 + \|Q_{B\times T}\|^2)/\dim(Q_I + Q_{B\times T})} = \frac{58.8/1}{(527.6 + 42.7)/(54 + 2)} = 5.78 \end{aligned}$$

der skal evalueres i en F-fordeling med 1 og 56 frihedsgrader. p-værdien er 0.0196, og vi afviser at B ingen effekt har.

(d) Test om der er en effekt af faktor T mod den additive hypotese ved brug af det annoterede faktorstrukturdiagram.

Solution: Test af B mod $B+T$:

$$\begin{aligned} F &= \frac{\left(\|P_{B+T}X\|^2 - \|P_B X\|^2\right)/(\dim \mathcal{L}_{B+T} - \dim \mathcal{L}_B)}{\left(\|X\|^2 - \|P_{B+T}X\|^2\right)/(|I| - \dim \mathcal{L}_{B+T})} \\ &= \frac{\|Q_T X\|^2/\dim(Q_T)}{(\|Q_I\|^2 + \|Q_{B\times T}\|^2)/\dim(Q_I + Q_{B\times T})} = \frac{394.1/2}{(527.6+42.7)/(54+2)} = 19.35 \end{aligned}$$

der skal evalueres i en F-fordeling med 2 og 56 frihedsgrader. p-værdien er $<$ 0.0001, og vi afviser at T ingen effekt har.

(e) Er faktoren $B \times T$ surjektiv?

Solution: Ja. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af $B \times T$ ses at være $6 = 2 \times 3 =$ produktet af dimensionerne af enkeltfaktorerne.

(f) Hvor mange observationer er der i datasættet?

Solution: Der er 60 observationer. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af I er 60.

(g) Hvor mange labels er der for faktoren B?

Solution: Der er 2 labels for faktoren B. Det kan ses ud fra det annoterede faktorstrukturdiagram, hvor dimensionen af B er 2.

(h) Er L_B og L_T ægte ortogonale?

Solution: Nej. Det kan blandt andet ses udfra at summen af deres dimensioner (=5) ikke er lig dimensionen af sumunderrummet (=4). Det kan også ses udfra at $L_B \cap L_T = L_1 \neq \{0\}$.

Opgave 3

3. Ved en undersøgelse af virkningen af forskellige dæktyper på benzinforbruget af offentlige busser blev følgende forsøg gennemført: 2 busser, A og B, gennemkørte hver 30 gange samme rundstrækning på ca. 10 km. I hver kørsel brugte de en af tre forskellige dæktyper, K, L eller M, således at hver kombination af bus og dæktype blev testet 10 gange, og benzinforbruget i milliliter blev målt. Der var 10 chauffører til at køre de ialt 60 ture.

   Data er tilgængelige i filen `benzin.txt` og består af variablene `bus`, `daek`, `cha` og `benzin`, hvor den sidste angiver benzinforbruget.

   Delopgaverne (b), (e), (f) og (g) skal løses i R, og det er nok at angive værdier fundet i output fra analyserne.

   (a) Opstil en varianskomponentmodel med en fast effekt af `bus` og en fast effekt af `dæk`, således at de indgår additivt, og en tilfældig effekt af `cha`. Alle de forklarende variable skal indgå som faktorer.

Solution: Vi skriver B, D og C for faktorerne `bus`, `dæk` og `cha`. Statistisk model for data:
Lad $I = \{1, \ldots, 60\}$ være indeksmængden for observationerne. X er regulært normalfordelt på $\mathbb{R}^I$ med middelværdivektor $\xi = (\xi_i)_{i\in I} \in L_{B+D}$ og kovariansmatrix $\sigma^2\Sigma = \sigma^2(I + \lambda BB^T)$, hvor B er effektmatricen hørende til effektparret $(C, 1)$. Middelværdiunderrummet har dimension 4. Kovariansmatricen er givet ved

$$\sigma^2\Sigma = \sigma^2 \begin{pmatrix} \Sigma_1 & \cdots & 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_1 & 0 & \cdots & \Sigma_2 \\ \Sigma_2 & \cdots & 0 & \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_2 & 0 & \cdots & \Sigma_1 \end{pmatrix}$$

hvor

$$\Sigma_1 = \begin{pmatrix} 1+\lambda & \lambda & \lambda \\ \lambda & 1+\lambda & \lambda \\ \lambda & \lambda & 1+\lambda \end{pmatrix}; \quad \Sigma_2 = \begin{pmatrix} \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda \end{pmatrix}$$

Alternativt:

$$X = A\beta + Z + \varepsilon$$

hvor A er designmatrix for middelværdiunderrummet L_{B+D}, $\beta \in \mathbb{R}^4$ er middelværdiparametervektoren, $Z = BY \sim \mathcal{N}(0, \nu^2 BB^T)$ er den tilfældige effekt, hvor B er effektmatricen hørende til effektparret $(C, 1)$. Derudover er $Y \sim \mathcal{N}(0, \nu^2 I_{10})$ og $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{60})$. Bemærk at $\nu^2 = \lambda\sigma^2$.

(b) Estimer parametrene i modellen, både ved ML-princippet og ved REML-princippet.

Solution: Her benyttes R til at estimere parametrene (β, ν^2, σ^2).

Estimater med ML-princippet:

R-kode (behøves ikke at vedlægges besvarelsen):

```
require(lme4)
data <- read.table("benzin.txt",head=TRUE)
# Med intercept:
m1ml <- lmer(benzin ~ bus+daek+(1|cha), data=data, REML=FALSE)
summary(m1ml)
# Uden intercept:
m1ml <- lmer(benzin ~ bus+daek-1+(1|cha), data=data, REML=FALSE)
summary(m1ml)
```

Resultat:

$$\begin{aligned}
\hat{\beta} &= (2772.6, 2.7, 330.1, -191.3) \quad \text{(parametriseret med intercept)} \\
\hat{\beta} &= (2772.6, 2775.2, 330.1, -191.3) \quad \text{(parametriseret uden intercept)} \\
\hat{\nu}^2 &= 88934; \quad \hat{\nu} = 298.2 \\
\hat{\sigma}^2 &= 10510; \quad \hat{\sigma} = 102.5
\end{aligned}$$

Den studerende behøver kun at angive en af parametriseringerne, det gælder også i de følgende besvarelser.

Estimater med REML-princippet:

```
# Med intercept:
m1reml <- lmer(benzin ~ bus+daek+(1|cha), data=data)
summary(m1reml)
# Uden intercept:
m1reml <- lmer(benzin ~ bus+daek-1+(1|cha), data=data)
summary(m1reml)
```

$$\begin{aligned}
\hat{\beta} &= (2772.6, 2.7, 330.1, -191.3) \quad \text{(parametriseret med intercept)} \\
\hat{\beta} &= (2772.6, 2775.2, 330.1, -191.3) \quad \text{(parametriseret uden intercept)} \\
\hat{\nu}^2 &= 98898; \quad \hat{\nu} = 314.5 \\
\hat{\sigma}^2 &= 11181; \quad \hat{\sigma} = 105.7
\end{aligned}$$

(c) Diskuter forskelle/ligheder i resultaterne mellem de to metoder.

Solution: Vi ser at middelværdiestimaterne er det samme for de to metoder (de er ikke altid nøjagtig det samme, men tæt på hinanden).

Vi ser at begge variansestimater er højere for REML end for ML. REML er netop en metode til at korrigere for at variansen underestimeres med MLE, især de tilfældige effekter underestimeres.

(d) Når der testes hypoteser vedrørende faste effekter med et kvotienttest, skal man så bruge ML-princippet eller REML-princippet?

Solution: Man skal altid benytte ML-princippet, da resultatet ellers afhænger af den valgte parametrisering af middelværdiunderrummet.

(e) Test om dæktypen påvirker benzinforbruget.

Solution: Der testes ved at fitte modellen kun med `bus` i middelværdien overfor den additive model og benytte anova i R.

R-kode:

```
m2ml <- lmer(benzin ~ bus+(1|cha), data=data, REML=FALSE)
anova(m1ml,m2ml)
```

Bemærk ovenfor at man ikke behøver at fitte med `REML=FALSE` da R automatisk refitter når man kører `anova` på modellerne. Her fås at $-2\log Q = 91.99$ med 2 frihedsgrader, hvilket giver en p-værdi på under 0.0001, og vi afviser derfor hypotesen om at der ikke er forskel på de forskellige dæktypers benzinforbrug. Vi fortsætter med den additive model.

(f) Undersøg om der er en signifikant forskel på de to bussers benzinforbrug.

Solution: Der testes ved at fitte modellen kun med `daek` i middelværdien overfor den additive model og benytte anova i R.

R-kode:

```
m3ml <- lmer(benzin ~ daek+(1|cha), data=data, REML=FALSE)
anova(m1ml,m3ml)
```

Her fås at $-2\log Q = 0.0101$ med en frihedsgrad, hvilket giver en p-værdi på 0.92, og vi accepterer derfor hypotesen om at der ikke er forskel på de to bussers benzinforbrug. Den endelige model bliver derfor at middelværdien kun afhænger af dæktypen.

(g) I den endelige model skal du kun beholde de faste effekter, der var signifikante. Angiv estimaterne i den endelige model. Hvilken bus og dæktype vil du anbefale?

Solution: Her kan man evt teste den konstante model overfor en model med kun dæk i middelværdien, men det er ikke noget krav. R-kode:

```
m4ml <- lmer(benzin ~ 1+(1|cha), data=data, REML=FALSE)
anova(m3ml,m4ml)
```

Hvis det gøres fås at $-2\log Q = 92$ med 2 frihedsgrader, hvilket giver en meget lille p-værdi under 0.0001, og vi afviser igen at dæk ikke har nogen effekt på benzinforbruget. Den endelige model bliver derfor at middelværdien kun afhænger af dæktypen.

Estimater i den endelige model findes i R (her med REML, da disse er estimaterne der bør rapporteres).

R-kode:

```
m3reml <- lmer(benzin ~ daek+(1|cha), data=data)
summary(m3reml)
m3reml <- lmer(benzin ~ daek-1+(1|cha), data=data)
summary(m3reml)
```

$$\begin{aligned}
\hat{\beta} &= (2773.9, 330.1, -191.3) \quad \text{(parametriseret med intercept)} \\
\hat{\beta} &= (2773.9, 3104.0, 2582.6) \quad \text{(parametriseret uden intercept)} \\
\hat{\nu}^2 &= 98937; \quad \hat{\nu} = 314.5 \\
\hat{\sigma}^2 &= 10951; \quad \hat{\sigma} = 104.6
\end{aligned}$$

Det er således dæktype M der bruger mindst benzin. Anbefalingen er derfor at det er ligegyldigt hvilken bus, man vælger, men man bør vælge dæktype M.