Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Reeksamen - Statistik 2
4 timer skriftlig eksamen
24. august 2017

4 timers skriftlig prøve. Alle hjælpemidler tilladt (inkl. computer uden netforbindelse). Opgavesættet består af tre opgaver med 15 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Til besvarelsen af opgave 3 har du fået udleveret et datasæt på en USB-nøgle.

Opgave 1

Betragt sandsynlighedsmålet givet ved fordelingsfunktionen

$$F_\alpha(y) = \begin{cases} 0, & y < 0 \\ y^\alpha, & 0 \leq y \leq 1 \\ 1, & y > 1 \end{cases}$$

hvor $\alpha > 0$ er en ukendt parameter. Lad $Y_1, \ldots, Y_n$ være uafhængige og identisk fordelte med fordelingsfunktion F_α. Sæt $X_i = 1(Y_i \leq 1/2)$.

1. Argumenter for, at $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ er asymptotisk normalfordelt og bestem de asymptotiske parametre.

2. Vis at $\tilde{\alpha}_n = -\frac{\log\left(\frac{1}{n}\sum_{i=1}^n X_i\right)}{\log(2)}$ er konsistent for α og asymptotisk normalfordelt med asymptotisk varians $\frac{1}{n}\frac{2^\alpha - 1}{(\log(2))^2}$.

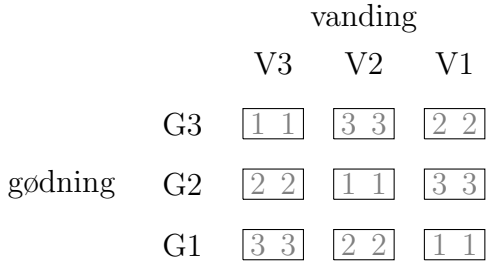
3. Vis at fordelingen af Y_1 kan opfattes som en eksponentiel familie. Angiv normeringskonstanten og den kanoniske stikprøvefunktion.

4. Opskriv likelihoodfunktion, loglikelihoodfunktion og scorefunktion for α svarende til uafhængige og identisk fordelte observationer $Y_1, \ldots, Y_n$.

5. Gør rede for, at der findes en entydig maksimaliseringsestimator for α, og angiv dens asymptotiske fordeling.

Opgave 2

Vi betragter et dyrkningsforsøg med tomater udført i 9 plantekasser, som hver indeholder 2 tomatplanter. Plantekasserne er organiseret som anført på figuren



Plantekasser i samme vandrette række modtager samme dosis gødning, mens plantekasser i samme lodrette søjle modtager samme daglige vandmængde. Der indgår tre niveauer af faktoren `G` (gødning) og tre niveauer af faktoren `V` (vanding). Endelig indgår 3 forskellige tomatsorter i forsøget givet ved faktoren `T` (tomat). De to planter i samme plantekasse er altid af samme sort. Tomatsorten `T` er angivet med de grå tal på figuren. Således vil tomatplanterne i kassen placeret i øverste venstre hjørne på figuren ovenfor begge være af sort `1` og modtage gødning svarende til niveau `G3` og daglig vandmængde svarende til niveau `V3`.

Det samlede udbytte `y` (nettovægt i gram) fra hver af de 18 tomatplanter angives med vektoren $X = (X_i)_{i \in I}$. Observationerne X_i antages at være uafhængige og normalfordelte $\mathcal{N}(\xi_i, \sigma^2)$ med ukendt varians $\sigma^2 > 0$,

1. Brug R-udskriften sidst i opgaven til udføre et test for, om der er en vekselvirkning mellem faktorerne `G` og `V`.

2. Estimer parametrene i den additive model, hvor alle tre faktorer indgår, og angiv estimatorernes simultane fordeling. Forklar hvordan parametrene i R-udskriften skal fortolkes. Du bedes i dette delspørgsmål kun forholde dig til de parametre, der indgår i beskrivelsen af middelværdistrukturen.

3. Angiv fordelingen af maksimaliseringsestimatoren $\hat{\sigma}^2$ for variansen under modellen fra delopgave 2.

Året efter det oprindelige forsøg udføres en simplere variant af dyrkningsforsøget med tomater. Der anvendes nu kun en tomatsort og der plantes kun en tomat i hver plantekasse (dvs. 9 tomatplanter i alt). Endelig er der nu en del af planterne, som slet ikke modtager gødning/vanding (angivet ved niveauet `ingen`) og må klare sig med regnvand og næringsstoffer fra jorden de står i. De øvrige planter modtager enten høj (`=hoj`) eller lav (`=lav`) dosis. Det fulde datasæt ses nedenfor

```
tomat_data_ny
```

```
##       G     V   y
## 1 ingen ingen 286
## 2 ingen ingen 248
## 3 ingen ingen 268
## 4   lav   lav 456
## 5   lav   hoj 548
## 6   lav   hoj 561
## 7   hoj   lav 501
## 8   hoj   hoj 600
## 9   hoj   hoj 593
```

4. Find minimum ($\wedge$) mellem faktorerne `G` (gødning) og `V` (vanding). Afgør desuden om faktorerne `G` og `V` er geometrisk ortogonale.

5. Angiv dimensionen af det additive underrum $L_G + L_V$.

```
### data til besvarelse af opgave 2, delopgave 1-3
tomat_data
```

```
##     G  V tomat   y
## 1  G3 V3     1 136
## 2  G3 V3     1  98
## 3  G3 V2     3 158
## 4  G3 V2     3  96
## 5  G3 V1     2 248
## 6  G3 V1     2 261
## 7  G2 V3     2 151
```

```
## 8  G2 V3     2 200
## 9  G2 V2     1 433
## 10 G2 V2     1 479
## 11 G2 V1     3 309
## 12 G2 V1     3 339
## 13 G1 V3     3 131
## 14 G1 V3     3 158
## 15 G1 V2     2 404
## 16 G1 V2     2 470
## 17 G1 V1     1 488
## 18 G1 V1     1 533
```

```
### fit af modeller
mod1 <- lm(y ~ tomat + G * V, data = tomat_data)
mod2 <- lm(y ~ tomat + G + V, data = tomat_data)
mod3 <- lm(y ~ tomat + V, data = tomat_data)

### test
anova(mod3, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ tomat + V
## Model 2: y ~ tomat + G + V
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 139268
## 2     11  10440  2    128827 67.866 6.483e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ tomat + G + V
## Model 2: y ~ tomat + G * V
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     11 10440
## 2      9  8992  2    1448.4 0.7249 0.5106
```

```r
### estimater
summary(mod1)

##
## Call:
## lm(formula = y ~ tomat + G * V, data = tomat_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.00 -21.62   0.00  21.62  33.00
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  510.500     22.351  22.840 2.81e-09 ***
## tomat2       -79.833     25.808  -3.093 0.012861 *
## tomat3      -148.667     25.808  -5.760 0.000273 ***
## GG2          -37.833     25.808  -1.466 0.176714
## GG3         -176.167     25.808  -6.826 7.68e-05 ***
## VV2            6.333     36.499   0.174 0.866081
## VV3         -217.333     18.249 -11.909 8.21e-07 ***
## GG2:VV2      -23.000     54.748  -0.420 0.684259
## GG3:VV2      -65.000     54.748  -1.187 0.265515
## GG2:VV3           NA         NA      NA       NA
## GG3:VV3           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.61 on 9 degrees of freedom
## Multiple R-squared:  0.9769,Adjusted R-squared:  0.9565
## F-statistic: 47.67 on 8 and 9 DF,  p-value: 1.81e-06
```

```r
summary(mod2)
```

```
##
## Call:
## lm(formula = y ~ tomat + G + V, data = tomat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.889 -18.514   2.111  18.236  42.778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   522.39      19.21  27.191 1.94e-11 ***
## tomat2        -72.17      17.79  -4.057  0.00189 **
## tomat3       -162.67      17.79  -9.145 1.79e-06 ***
## GG2           -45.50      17.79  -2.558  0.02661 *
## GG3          -197.83      17.79 -11.122 2.53e-07 ***
## VV2           -23.00      17.79  -1.293  0.22248
## VV3          -217.33      17.79 -12.219 9.66e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.81 on 11 degrees of freedom
## Multiple R-squared:  0.9732,Adjusted R-squared:  0.9586
## F-statistic: 66.66 on 6 and 11 DF,  p-value: 5.235e-08
```

```r
A <- model.matrix(mod2)
solve(t(A) %*% A) ### resultat lettere afrundet ...
```

```
##             (Intercept) tomat2 tomat3    GG2    GG3    VV2    VV3
## (Intercept)       0.389 -0.167 -0.167 -0.167 -0.167 -0.167 -0.167
## tomat2           -0.167  0.333  0.167  0.000  0.000  0.000  0.000
## tomat3           -0.167  0.167  0.333  0.000  0.000  0.000  0.000
## GG2              -0.167  0.000  0.000  0.333  0.167  0.000  0.000
## GG3              -0.167  0.000  0.000  0.167  0.333  0.000  0.000
## VV2              -0.167  0.000  0.000  0.000  0.000  0.333  0.167
## VV3              -0.167  0.000  0.000  0.000  0.000  0.167  0.333
```

Opgave 3

I forbindelse med et konditionsforsøg er 14 motionsløbere blevet bedt om at løbe en bestemt rute (omgang) på 1460 m. Hver løber har gennemført ruten 5 gange, men der er altid kun løbet en tur på en given dag. Omgangstiden `tid` (i sekunder) og den gennemsnitlige `puls` (slag per minut) er blevet registreret for hver omgang ved brug af et pulsur. Desuden har hver løber enten løbet alle sine ture om morgenen eller alle sine ture senere på dagen. Denne oplysning er i datasættet registreret med faktoren `morgen` med niveauerne `ja` og `nej`.

Data til opgaven er gjort tilgængelige i filen `stat2aug2017opg3.txt` som er udleveret på vedlagte USB nøgle. I datasættet indgår faktorerne `person` og `morgen` samt pulsmålingerne `puls` og omgangstiderne `tid`.

1. Tegn et faktorstrukturdiagram der indeholder faktorerne `person` og `morgen` samt den identiske og den konstante faktor. (Du skal ikke tilføje $||P_G X||^2$ og $||Q_G X||^2$!)

Personerne i forsøget er blev instrueret i at løbe forskellig belastning på hver af deres fem ture, og formålet med forsøget er primært at undersøge sammenhængen mellem omgangstid og puls. Man ønsker at drage generelle konklusioner, der ikke kun udtaler sig om lige præcis de 14 personer, som deltager i dette eksperiment.

2. Opstil en varianskomponentmodel der udtrykker, at der er en lineær sammenhæng mellem omgangstid (`tid`) og `puls`, og hvor `person` inddrages som tilfældig effekt (intercept). Du bedes se helt bort fra faktoren `morgen` ved besvarelse af denne delopgave.

3. Estimer parametrene i modellen fra 2. i R og angiv estimater for samtlige parametre der indgår i beskrivelsen af middelværdi- og variansstruktur.

4. Giv et forslag til hvordan man kan teste, om sammenhængen mellem omgangstid og puls er anderledes ved morgenløb end ved løb senere på dagen. Dette kræver at du udvider din model fra 2. Udfør et eller flere relevante test i R og angiv en konklusion.

5. Angiv kovariansmatricen for de 5 målinger af omgangstiden der stammer fra person nummer `P1`.