Københavns Universitet
Det Naturvidenskabelige Fakultet

# Eksamen i Statistik 2

21. juni 2018

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

Eksamenssættet består af fire opgaver og dækker både pensum fra 2017 og 2018. Hvis man ønsker at blive bedømt i forhold til pensum fra 2018 afleverer man besvarelser af opgave 1, 3, og 4. Hvis man ønsker at blive bedømt i forhold til gammelt pensum (2017) afleverer man besvarelser af opgaverne 2, 3, og 4.

*Hvis man ønsker at blive bedømt efter gammelt pensum skal dette angives tydeligt på første side i besvarelsen.*

Hvis intet er angivet, vil besvarelsen blive bedømt i forhold til pensum fra 2018 og kun besvarelser af opgave 1, 3, og 4 vil indgå i bedømmelsen. Angives, at besvarelsen ønskes bedømt i forhold til gammelt pensum, indgår kun besvarelser af opgaverne 2, 3, og 4 i bedømmelsen.

I begge tilfælde vil der være i alt 17 delspørgsmål, som skal bedømmes; alle delspørgsmål vægtes ens i bedømmelsen.

Data til Opgave 3 og 4 ligger på en USB-nøgle i filerne `moral.txt` og `koed.txt`. Nøglen skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den skal altså ikke indgå som del af besvarelsen.

## Opgave 1

*Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2018 (nyt) pensum.*

Lad $X = (X_1, X_2, X_3, X_4)^T \in \mathbb{R}^4$ være normalfordelt $\mathcal{N}(\xi, \Sigma)$, hvor

$$\xi = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \frac{1}{2} \begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

1. Er $X$ regulært eller singulært normalfordelt? Begrund dit svar.

   **Vejledende besvarelse**
   Da $\Sigma$ er singulær er $X$ singulært normalfordelt, jvf. Korollar 9.43.

2. Lad $X = (X_1, X_2, X_3, X_4)^T$ være som ovenfor. Angiv om $Y = (X_1, X_2)^T$, hhv. $Z = (X_3, X_4)^T$ er regulært eller singulært normalt fordelte.

   **Vejledende besvarelse**
   Vi ser at de marginale fordelinger (jvf. Sætning 9.47) er $Y \sim \mathcal{N}(0, \Sigma_Y)$ og $Z \sim \mathcal{N}(0, \Sigma_Z)$, med

   $$\Sigma_Y = \frac{1}{2} \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \quad \text{og} \quad \Sigma_Z = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

   hvoraf $\Sigma_Y^{-1}$ eksisterer, mens $\Sigma_Z$ er singulær. Kovariansmatricerne viser derfor at $Y$ er regulært normalfordelt, mens $Z$ er singulært normalfordelt, jvf. Korollar 9.43.

3. Angiv fordelingen af $(Y+Z)^T \Sigma_{Y+Z}^{-1} (Y+Z)$, hvor $\Sigma_{Y+Z}$ er variansmatricen for $Y+Z$.

   **Vejledende besvarelse**
   Bemærk først at $Y \perp\!\!\!\perp Z$ da $\mathrm{Cov}(Y, Z) = 0$, jvf. Sætning 9.48. Derfor er $Y + Z$ er en sum af uafh. normalfordelte variable hvilket giver at den selv er 2-dimensionalt normal fordelt, med middelværdi $\mathbb{E}Y + \mathbb{E}Z = 0$ og varians

   $$\Sigma_{Y+Z} = \Sigma_Y + \Sigma_Z = \frac{1}{4} \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$$

   Da $\Sigma_{Y+Z}^{-1}$ eksisterer er den regulært normalfordelt.

Da $Y+Z$ er 2-dimensionelt regulært normalfordelt med præcision $\langle \cdot, \cdot \rangle$ givet ved $\Sigma_{Y+Z}^{-1}$ vil $||Y+Z||^2_{\Sigma_{Y+Z}} = (Y+Z)^T \Sigma_{Y+Z}^{-1}(Y+Z)$ være $\chi^2$-fordelt med 2 frihedsgrader, jvf. Sætning 9.29.

Bemærk at de næste delopgaver er uden sammenhæng med de foregående.

Lad nu $W \sim \mathcal{N}(\xi, \sigma^2 I_4)$ være regulært normal fordelt.

4. Opskriv design matricen $A$ for følgende specifikation af $\xi$

$$\xi = A\beta = \begin{pmatrix} \beta_1 + \beta_3 \\ \beta_1 \\ \beta_2 + \beta_3 \\ \beta_2 \end{pmatrix}$$

**Vejledende besvarelse**
Idet der indgår tre parametre $(\beta_1, \beta_2, \beta_3)$ i en 4-dimensional fordeling kan vi se at $A$ skal være $4 \times 3$. Udfra kombinationerne af $\beta$'er i $\xi$ ser vi at

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Med $\beta = (\beta_1, \beta_2, \beta_3)$ ser vi at $A\beta$ giver $\xi$ som beskrevet.

5. Betragt en observation $w = (2,\ 5,\ 3,\ 1)^T$ fra $W$ defineret som i spørgsmål 4).

   Find maksimum likelihood estimatorerne for $\beta$ og $\sigma^2$ med designmatricen fra spørgsmål 4) og angiv standard errors for $\hat{\beta}_1, \hat{\beta}_2$ og $\hat{\beta}_3$.

   Du må her gerne benytte maksimaliserings-estimatoren for $\hat{\sigma}^2$, fremfor den centrale estimator til at beregne standard errors for $\hat{\beta}$.

   **Vejledende besvarelse**
   Fra Korollar 10.21 har vi at

$$\hat{\beta} = (A^T A)^{-1} A^T W$$
$$\hat{\sigma}^2 = \frac{||W - A\hat{\beta}||^2}{N}$$

Med $A$ som i forrige spørgsmål er

$$A^TA = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

hvilket giver

$$(A^TA)^{-1} = \frac{1}{4}\begin{pmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix}$$

Dvs. når vi indsætter $w$ får vi

$$\hat{\beta} = (A^TA)^{-1}A^Tw = \frac{1}{4}\begin{pmatrix} 1 & 3 & -1 & 1 \\ -1 & 1 & 1 & 3 \\ 2 & -2 & 2 & -2 \end{pmatrix}\begin{pmatrix} 2 \\ 5 \\ 3 \\ 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 15 \\ 9 \\ -2 \end{pmatrix}$$

Indsætter vi $A\hat{\beta}$ i variansestimatoren får vi

$$\begin{aligned} \hat{\sigma}^2 = \frac{||W - A\hat{\beta}||^2}{N} = \frac{W^TW - (A\hat{\beta})^TA\hat{\beta}}{N} \\ = \frac{39 - 32.75}{4} = 1.5625 \quad = \frac{25}{16} \end{aligned}$$

Endeligt bliver de estimerede standard errors for $\hat{\beta}$ $\sqrt{}$-roden af diagonalen i matricen

$$\hat{\sigma}^2(A^TA)^{-1} = \frac{25}{64}\begin{pmatrix} 3 & 1 & -2 \\ 1 & 3 & -2 \\ -2 & -2 & 4 \end{pmatrix} = \begin{pmatrix} 1.17 & 0.39 & -0.78 \\ 0.39 & 1.17 & -0.78 \\ -0.78 & -0.78 & 1.56 \end{pmatrix}$$

dvs.

$$\text{SE}(\hat{\beta}) = (\sqrt{1.17}, \sqrt{1.17}, \sqrt{1.56})^T = (1.08, 1.08, 1.25)^T \quad = \frac{5}{8}(\sqrt{3}, \sqrt{3}, 2)^T$$

Teknisk set bør vi bruge den centrale estimator $\tilde{\sigma}^2$ til standard errors for $\hat{\beta}$, men det undlader vi her, som efterspurgt. Hvis $\tilde{\sigma}^2$ er benyttet, er tælleren i estimatoren $N - k = 4 - 3 = 1$, dvs. $\tilde{\sigma}^2 = 4\hat{\sigma}^2$ og derfor bliver $\tilde{\sigma} = 2\hat{\sigma}$. Det giver så standard errors for $\hat{\beta}$'erne der er dobbelt så store som angivet ovenfor

$$\text{SE}(\hat{\beta}) = (2.17, 2.17, 2.50)^T = \quad \frac{5}{4}(\sqrt{3}, \sqrt{3}, 2)^T$$

## Opgave 2

*Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2017 (gammelt) pensum.*

Betragt fordelingen med tæthed

$$f_\alpha(x) = \left(\frac{1}{\sqrt{2\pi}}\right) x^{-3/2} \exp\left\{-\frac{(x-\alpha)^2}{2\alpha^2 x}\right\} \quad \text{for } x > 0$$

med hensyn til Lebesgue-målet. Fordelingen afhænger af parameteren $\alpha > 0$.

Du kan uden bevis benytte at $f_\alpha$ er en tæthed.

1. Reparametriser $f_\alpha$-fordelingen ved $\theta = -\frac{1}{2\alpha^2}$ så det fremgår at det er en eksponentiel familie med kanonisk stikprøvefunktion $t(x) = x$.

   **Vejledende besvarelse**
   Vi får

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

   hvor $\theta = -\frac{1}{2\alpha^2}$. Vi får således (se def. 2.13 i lærebogen)

$$f_\alpha(x) = \underbrace{\frac{1}{\sqrt{2\pi}} x^{-3/2} e^{-\frac{1}{2x}}}_{\text{funktion af } x} \quad \underbrace{e^{\sqrt{-2\theta}}}_{\text{funktion af } \theta} \quad \underbrace{e^{\theta x}}_{t(x) = x}$$

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

   **Vejledende besvarelse**
   Lad $\mu$ have tæthed

$$\frac{1}{\sqrt{2\pi}} x^{-3/2} e^{-\frac{1}{2x}} \quad \text{for } x > 0$$

med hensyn til Lebesgue målet. Udtrykt ved $\theta = -\frac{1}{2\alpha^2}$ har $X$ tæthed

$$e^{\sqrt{-2\theta}} e^{\theta x} \quad \text{for } x \in \mathbb{R}^+$$

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

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

3. Lad $X$ have tæthed $f_\alpha$. Argumenter for at $X$ har momenter af enhver orden. Find middelværdi og varians af $X$, både som funktion af $\theta$ og som funktion af $\alpha$.

   **Vejledende besvarelse**
   Ifølge Lemma 2.19 har $t(X) = X$ 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

   $$E(t(X)) = E(X) = \frac{d}{d\theta} \log c(\theta) = \frac{1}{\sqrt{-2\theta}} = \alpha$$

   og

   $$\mathrm{Var}(t(X)) = \mathrm{Var}(X) = \frac{d^2}{d\theta^2} \log c(\theta) = \frac{1}{(\sqrt{-2\theta})^3} = \alpha^3$$

4. Opskriv likelihoodligningen for $\theta$. Find maksimaliseringsestimatoren for $\theta$. Find maksimaliseringsestimatoren for $\alpha$.

   **Vejledende besvarelse**
   Likelihoodligningen for $\theta$ er givet i (5.6), det følger derfor fra resultaterne i (c) at likelihoodligningen er

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

   Løses denne fås maksimaliseringsestimatoren for $\theta$:

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

   Maksimaliseringsestimatoren for $\alpha$:

   $$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

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

**Vejledende besvarelse**

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

$$\hat{\theta} \overset{as}{\sim} N\left(\theta, \frac{1}{n}(\sqrt{-2\theta})^3\right)$$

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

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{as}{\sim} N\left(\frac{1}{\sqrt{-2\theta}}, \frac{1}{n(\sqrt{-2\theta})^3}\right)$$

Vi har at

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

For $x > 0$ er $f$ differentiabel, herunder specielt for $x = E(X_1)$. Vi har $f'(x) = \frac{1}{x^3}$ og $f'(E(X_1)) = (\sqrt{-2\theta})^3$. Vi benytter deltametoden, og får

$$\begin{aligned}\hat{\theta} = f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &\overset{as}{\sim} N\left(f\left(\frac{1}{\sqrt{-2\theta}}\right), \frac{1}{n}\left((\sqrt{-2\theta})^3\right)^2 \frac{1}{(\sqrt{-2\theta})^3}\right) \\ &= N\left(\theta, \frac{1}{n}(\sqrt{-2\theta})^3\right)\end{aligned}$$

som før.

## Opgave 3

I et studie[1] undersøgte man om der var forskel på mænd og kvinders moral i den amerikanske kystvagt. Specifikt noterede man en score for hver person udfra den såkaldte *Rest's Defining Issues Test (DIT)*, hvor en højere score indikerer en højere moral.

Studiet involverede 225 personer fordelt på 4 grupper efter køn og rang. Tabel 1 viser

| | | Rang | |
|---|---|---|---|
| | | Officer | Menig |
| Køn | Mand | 60 | 120 |
| | Kvinde | 15 | 30 |

Tabel 1: Antalstabel for studiet omkring mænd og kvinders moral i US Coast Guard.

antallet af personer i hver af de fire grupper. Data til besvarelse af opgaven findes på USB nøglen med filnavnet `moral.txt`.

Betragt faktorerne $K = \{\text{mand}, \text{kvinde}\}$ og $R = \{\text{officer}, \text{menig}\}$ med hver to niveauer.

1. Gør rede for at de to faktorer $K$ og $R$ er geometrisk ortogonale og angiv hvorvidt designet

   $$\mathbb{G} = \{K, R, K \times R\}$$

   er ortogonalt og minimums-stabilt.

   **Vejledende besvarelse**
   Tabel 1 er præcis antals-tabellen for de to faktorer $K$ og $R$. Indsætter vi række- og søjle-sum har vi

| | | Rang | | |
|---|---|---|---|---|
| | | Officer | Menig | |
| Køn | Mand | 60 | 120 | 180 |
| | Kvinde | 15 | 30 | 45 |
| | | 75 | 150 | 225 |

[1] Kilde: R.D. White, Jr. (1999). "*Are Women More Ethical? Recent Findings on the Effects of Gender Under Moral Development*,"Journal of Public Administration Research and Theory, Vol. 9, #3, pp.459-471

og da

$$\frac{180 \cdot 75}{225} = 60$$
$$\frac{180 \cdot 150}{225} = 120$$
$$\frac{45 \cdot 75}{225} = 15$$
$$\frac{45 \cdot 150}{225} = 30$$

ser vi at balance-ligningen er opfyldt (her kan henvises til enten Lemma 13.11 da $K \wedge R = 1$, eller den mere generelle Sætning 14.8). Dermed er $K \underset{G}{\perp} R$. Bemærk også at $R, K \leq K \times R$, dvs. $K, R \underset{G}{\perp} K \times R$.

Designet $\mathbb{G}$ er derfor ortogonalt (de indbyrdes faktorer er geometrisk ortogonale), men det er *ikke* minimums-stabilt. Der mangler en minimiumsfaktor: $K \wedge R = 1$, dvs.

$$\mathbb{G}' = \{1, K, R, K \times R\}$$

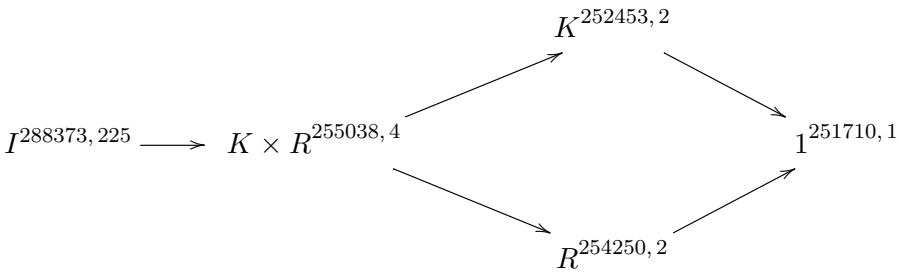
er både $\wedge$-stabilt (og ortogonalt).

2. Find $\dim(V_G)$ og $||Q_G X||^2$ for hver faktor $G$ i faktorstrukturdiagrammet hvor $V_G$ er den ortogonale dekomposition

$$V_G = L_G \ominus \sum_{G' \in \mathbb{G}, G' < G} L_{G'}, \quad \text{for hvert } G \in \mathbb{G}$$

med tilhørende projektion $Q_G$, og angiv $\dim L_K + L_R$.

Værdierne i diagrammet (1) kan benyttes.



(1)

**Vejledende besvarelse**

Sætning 14.21 (ortogonal dekomposition) benyttes til at finde $V_G$ og $\dim(V_G)$, hvor

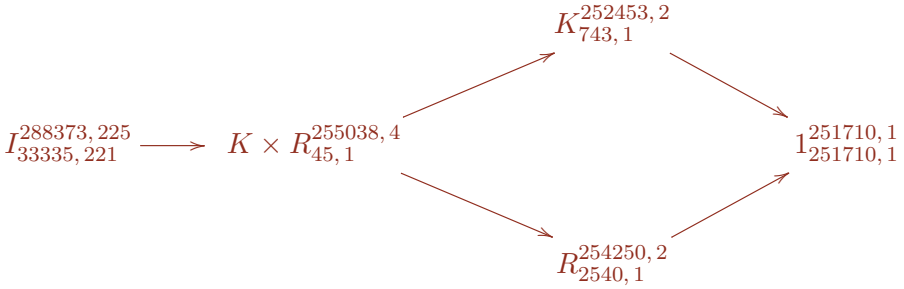
der startes bagfra:

$$\begin{aligned}
\dim(V_1) &= \dim(L_1) \\
\dim(V_K) &= \dim(L_K) - \dim(V_1) \\
\dim(V_R) &= \dim(L_R) - \dim(V_1) \\
\dim(V_{K\times R}) &= \dim(L_{K\times R}) - \dim(V_K) - \dim(V_R) - \dim(V_1) \\
\dim(V_I) &= \dim(L_I) - \dim(V_{K\times R}) - \dim(V_K) - \dim(V_R) - \dim(V_1)
\end{aligned}$$

tilsvarende for projektionerne $||Q_G X||^2$:

$$\begin{aligned}
||Q_1 X||^2 &= ||X||^2 \\
||Q_K X||^2 &= ||P_K X||^2 - ||Q_1 X||^2 \\
||Q_R X||^2 &= ||P_R X||^2 - ||Q_1 X||^2 \\
||Q_{K\times R} X||^2 &= ||P_{K\times R} X||^2 - ||Q_K X||^2 - ||Q_R X||^2 - ||Q_1 X||^2 \\
||Q_I X||^2 &= ||P_I X||^2 - ||Q_{G\times R} X||^2 - ||Q_G X||^2 - ||Q_R X||^2 - ||Q_1 X||^2
\end{aligned}$$

heraf får vi




Derudover ser vi at $\dim L_{K+R} = \dim L_R + \dim L_R - \dim L_{K\wedge R} = 2 + 2 - 1 = 3$.

3. Idet vi antager en lineær normal model

$$X \sim \mathcal{N}(\xi, \sigma^2 I_{225})$$

hvor $I_{225}$ betegner den 225-dimensionale identitetsmatrix, udfør $F$-testet der sammenligner modellerne $\xi \in L_{K\times R}$ og $\xi \in L_K + L_R$, udfra projektionerne $Q_G$ fra spørgsmål 2).

**Vejledende besvarelse**

Vi stiller $F$-testet op som i formel (10.31) og får

$$\begin{aligned} F_{\text{test}} &= \frac{||P_{G\times R} - P_{G+R}||^2/(\dim L_{G\times R} - \dim L_{G+R})}{\big(||x||^2 - ||P_{G\times R} - ||^2\big)/(N - \dim L_{G\times R})} \\ &= \frac{||Q_{K\times R}X||^2/(\dim V_{K\times R})}{||Q_I X||^2/\dim(V_I)} \\ &= \frac{45/1}{33335/221} \sim 0.298 \sim F_{1,221} \end{aligned}$$

Det giver en $p$-værdi på `1-pf(0.298,1,221)= 0.586`.

4. Angiv de estimerede middelværdier for de fire grupper, estimeret udfra modellen

$$\begin{gathered} X \sim \mathcal{N}(\xi, \sigma^2 I_{225}), \\ \text{hvor } \xi \in L_{K\times R} \end{gathered}$$

og undersøg om residualerne kan antages at følge en normal-fordeling med konstant varians.

**Vejledende besvarelse**
Data kan indlæses i `R` med kommandoen

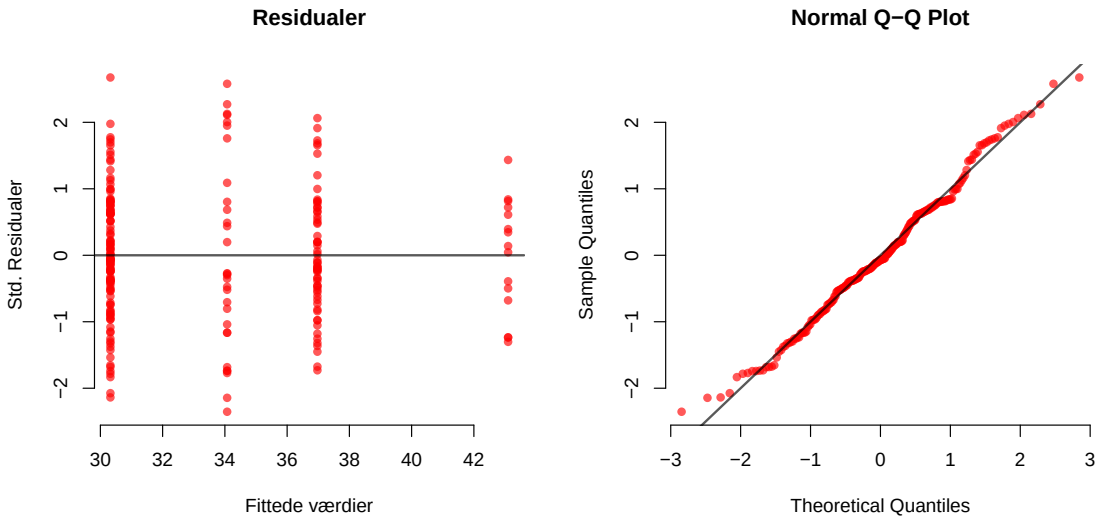
```
read.delim("moral.txt", header=TRUE)
```

Herefter kan modellen fittes i `R` med kommandoen

```
m0 = lm(score~koen*rang, data=moral))
```

og `summary(m0)` giver nu parametrene, mm.

Nedenfor ses de standardiserede residualer plottet mod de fittede værdier, $\hat{\xi}$, og et fraktil-plot (mod standard normal-fordelingen). De giver ingen anledning til bekymring, dog ses at variansen er en smule mindre for kvindelige officerer, men det er ikke noget vi vil bekymre os mere om da der her også er færre observationer til at estimere variationen. Fraktil-plottet er overbevisende pænt i forhold til identitetslinien. Alt efter hvordan modellen er parametriseret og hvilke grupper er valgt som reference, vil sammensætningen af outputtet variere. Her benyttes referencerne 'kvinde' og 'menig' og modellen er parametriseret med kontraster, dvs. inklusiv `(intercept)`-parameteren. `summary`-kommandoen giver flg. estimater for modellen med interaktion $(K \times R)$:




| Parameter | Estimat |
|---|---|
| (Intercept) | 34.073 |
| koenmand | -3.751 |
| rangofficer | 9.027 |
| koenmand:rangofficer | - 2.376 |

Heraf fås (med de angivne referencegrupper) de estimerede gruppe-middelværdier:

$$\begin{aligned}
\alpha_{\text{kvinde,menig}} &= 34.073 && = 34.073 \\
\alpha_{\text{kvinde,officer}} &= 34.073 + 9.027 && = 43.100 \\
\alpha_{\text{mand,menig}} &= 34.073 - 3.751 && = 30.322 \\
\alpha_{\text{mand,officer}} &= 34.073 - 3.751 + 9.027 - 2.376 && = 36.973
\end{aligned}$$

5. Test om det kan antages at der er vekselvirkning mellem faktorerne $K$ og $R$ med et signifikans niveau på 5% og forklar i ord hvad resultatet fortæller.

   **Vejledende besvarelse**
   Der kan benyttes en af 3 fremgangsmetoder:

   (a) Fra `summary(m0)` aflæses $t$-testet for at fjerne interaktionen til -0.547 med tilhørende $p$-værdi: 0.585.

   (b) Alternativt fittes modellen under $H_0$ i R med kommandoen `m1 = lm(score~koen+rang,`

`data=moral)` hvorved `anova(m1,m0)` kan bruges til at finde

$$F_{\text{test}} = 0.2994 \sim F_{1,221}$$

hvilket giver en $p$-værdi på 0.585 (tilsvarende $t$-testet ovenfor).

(c) Endeligt kan man benytte $F$-testet fra spørgsmål 3) hvor man evt. mangler at beregne $p$-værdien: `1-pf(0.298,1,221)= 0.586`. Der er en mindre forskel til `anova(...)` og $t$-testet pga. afrunding.

Vi ser at $H_0$ ej forkastes og vi kan derfor antage at der ikke er interaktion mellem $K$ og $R$. Altså indeholder modellen effekter af både køn og rang, men effekterne kan antages uafhængige af hinanden.

6. Antag nu modellen $\xi \in L_K + L_R$, dvs. modellen under $H_0$ i forrige spørgsmål, uanset din konklusion for $H_0$. Undersøg hvorvidt denne model kan reduceres yderligere.

**Vejledende besvarelse**
Ved at benytte `summary(m1)`, hvor `m1` er modellen estimeret som ovenfor, ses at alle parametrene er signifikante. Dog er $p$-værdien for at fjerne $K$ lig med 0.027, dvs. parameteren er signifikant på et 5% signifikans niveau, men havde man valgt et signifikansniveau på 1% ville man have kunnet fjerne den. Det er altså tale om en borderline-case.

7. Angiv konfidensintervaller for parametrene i modellen hvor $\xi \in L_K + L_R$ og beskriv hvad den fortæller om moralen for mænd og kvinder, samt menige og officerer i den amerikanske kystvagt.

**Vejledende besvarelse**

Vi ender i en model med additiv virkning af faktoerne $K$ og $R$. Denne model kan ikke reduceres yderligere på et 5% konfidens niveau. Vi ender derfor med 3 parametre: `(intercept)` (for referencegruppen: {kvinde, menig}) samt to kontrast parametre: en for køn en for rang.

| | Fit med intercept | | | Fit uden intercept | | |
|---|---|---|---|---|---|---|
| Parameter | Estimat | 2.5 % | 97.5 % | Estimat | 2.5 % | 97.5 % |
| (Intercept) | 34.71 | 30.93 | 38.48 | 34.71 | 30.93 | 38.48 |
| koenmand | -4.54 | -8.57 | -0.52 | 30.16 | 28.03 | 32.29 |
| rangofficer | 7.13 | 3.71 | 10.54 | 7.13 | 3.71 | 10.54 |

Modellen fortæller os at den gennemsnitlige score for en menig kvinde er 34.71 ([30.93; 38.48]), svarende til `(intercept)`-parameteren. Derudover ser vi at officerer scorer noget højere, gennemsnitligt 7.13 ([3.71; 10.54]) point mere end menige. Endeligt ser vi at mænd generelt scorer en forskel på -4.54 ([$-8.57$; $-0.52$]) point i forhold til kvinder. Vi kan derfor konkludere på baggrund af modellen at der rent faktisk *er* en signifikant (på 5%) forskel i mænd og kvinders moral i den amerikanske kystvagt.

## Opgave 4

I et eksperiment med grisekød har man undersøgt hvordan kødets farve (intensitet af rød) falmer over tid som følge af to forskellige opbevaringer: $B$={lyst, mørkt}. Da farven på kød indikerer hvor attraktivt kødet vurderes af forbrugerne, er man interesseret i hvordan de to behandlinger påvirker kødets farve, og specielt hvorvidt der er forskel på de to behandlinger.

I eksperimentet målte man på 10 slagtede grise. Fra hver gris tog man 6 stykker kød, hvoraf 3 blev lagret lyst i hhv. 1,4 og 6 dage og 3 blev lagret mørkt i hhv. 1,4 og 6 dage. Dette giver i alt $10 \times 3 \times 2 = 60$ observationer. Tabel 2 viser designet for hver gris.

| Opbevaring | 1 dag | 4 dage | 6 dage |
|---|---|---|---|
| Lyst | Stykke 1 | Stykke 2 | Stykke 3 |
| Mørkt | Stykke 4 | Stykke 5 | Stykke 6 |

Tabel 2: Fordeling af kødstykker fra hver gris.

Figur 1 viser et plot af observationerne fordelt på de to opbevaringsmetoder.

Som det fremgår af Tabel 3 er der ialt 3 faktorer i spil.

| Faktor | Betydning | Niveauer |
|---|---|---|
| $G$ | Gris | $\{1, \ldots, 10\}$ |
| $T$ | Tid | $\{1, 4, 6\}$ |
| $B$ | Opbevaring | {lyst, mørkt} |

Tabel 3: Faktorer i eksperimentet med grisekød.

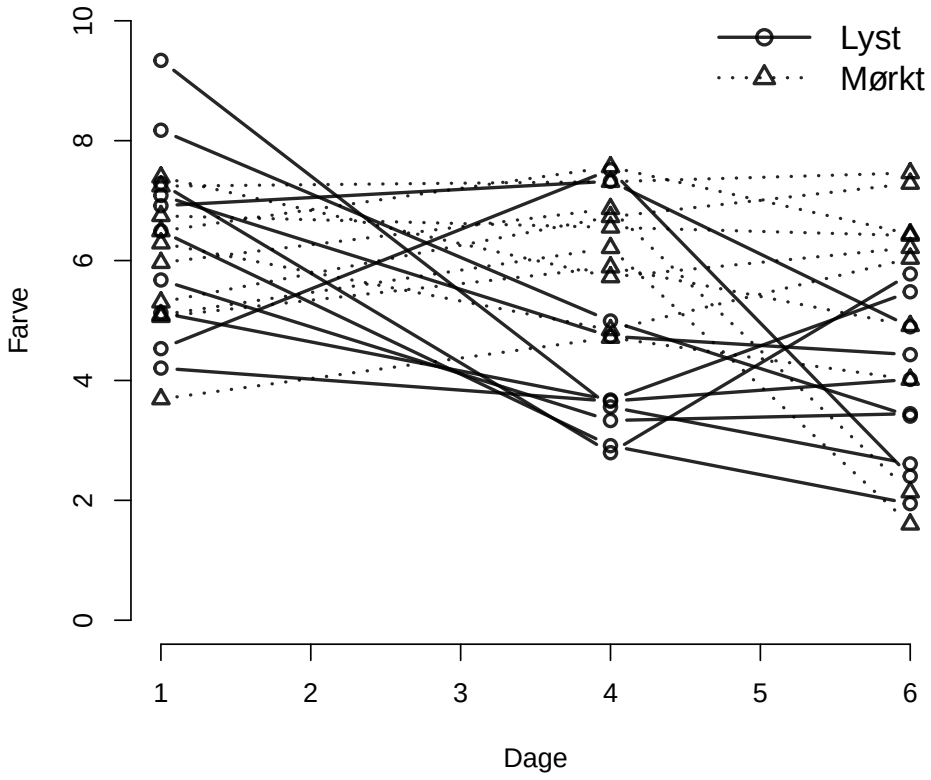
Betragt $G$ som tilfældig effekt og $T, B$ som faste effekter. Betragt derudover $T$ (tid) som faktor og ej numerisk.

Data til besvarelse af opgaven findes på USB nøglen med filnavnet `koed.txt`.

Lad $i = 1, 2$, $j = 1, 2, 3$, $k = 1, \ldots, 10$ angive $i$'te behandling (opbevaring) til $j$'te tidspunkt for den $k$'te gris.

1. Opskriv en varianskomponentmodel for rødheden af kødet, med $T \times B$ (vekselvirkning mellem $T$ og $B$) som fast effekt og $G$ som tilfældigt intercept. Angiv variansmatricen for en vilkårlig gris udtrykt ved hjælp af de teoretiske parametre i modellen.

   **Vejledende besvarelse**




Figur 1: Effekt af 2 forskellige opbevaringer af grisekød til 3 forskellige tidspunkter.

Modellen har interaktion mellem $T \times B$ der har 6 niveauer. Vi indekserer som flg.

$$\begin{aligned} \text{Opbevaring: } & i = 1, 2 \\ \text{Tid: } & j = 1, 2, 3 \\ \text{Gris: } & k = 1, \ldots, 10 \end{aligned}$$

og skriver da

$$\begin{aligned} X_{ijk} &= \alpha_{ij} + Y_k + \varepsilon_{ij} \\ Y_k &\sim \mathcal{N}(0, \nu^2) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

hvor $\alpha_{ij}$ betegner kombinationen af den $i$'te behandling og $j$'te tidspunkt. $Y_k$ angiver det tilfældige intercept for den $k$'te gris.

Idet vi lader $1_n$ og $1_{m \times n}$ betegne hhv. en $n$-dimensional vektor at 1-taller og en $m \times n$ matrix at 1-taller, kan vi skrive effektmatricen tilhørende effektparret $(G, 1)$, dvs. den

tilfældige effekt på interceptet, som

$$B = \begin{pmatrix} 1_6 & 0 & \dots & 0 \\ 0 & 1_6 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_6 \end{pmatrix}$$

hvilket giver at

$$BB^T = \begin{pmatrix} 1_{6\times 6} & 0 & \dots & 0 \\ 0 & 1_{6\times 6} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_{6\times 6} \end{pmatrix}$$

Vi ser derfor at for en vilkårlig gris bliver kovariansstrukturen en $6 \times 6$ matrix på formen

$$\mathrm{Cov}(X_{ijk}, X_{i'j'k'}) = \sigma^2 I_6 + \nu^2 1_{6\times 6} = \begin{pmatrix} \sigma^2+\nu^2 & \nu^2 & \dots & \nu^2 \\ \nu^2 & \sigma^2+\nu^2 & \dots & \nu^2 \\ \vdots & \vdots & \ddots & \vdots \\ \nu^2 & \nu^2 & \dots & \sigma^2+\nu^2 \end{pmatrix}, \quad \text{for } k = k'$$

2. Angiv estimaterne for variansparametrene i modellen fra spørgsmål 1). Husk at variablerne `gris` og `tid` skal betragtes som faktorer og ej numeriske!

   **Vejledende besvarelse**
   Data indlæses med `read.delim("data/grise.txt")`. Husk at lave variablene `gris` og `tid` om til faktorer

   ```
   koed$gris = as.factor(koed$gris)
   koed$tid = as.factor(koed$tid)
   ```

   herefter kan modellen fittes ved

   ```
   m0 = lmer(farve~opbevaring*tid+(1|gris),data=koed)
   ```

   `summary(m0)` angiver parametrene, heraf fås

   $$\sigma^2 = 1.9502$$
   $$\nu^2 = 0.2727$$

3. Angiv de estimerede middelværdier for rødheden af kød for de tre $B \times T$-niveauer:

   {lyst, 1 dag }
   {mørkt, 1 dag }
   {mørkt, 6 dage }

   udfra estimaterne fra modellen fra spørgsmål 1). Bemærk at der spørges til estimerede middelværdier for niveauerne i produktfaktoren, snarere end parameterestimater. Sidstnævnte afhænger af den parametrisering du har valgt, men det gør middelværdierne ikke!

   **Vejledende besvarelse**
   Som før angiver `summary(m0)` parametrene (husk $ij$ betegner den $i$'te behandling til $j$'te tidspunkt. Her er reference grupperne for de to faktorer $B$ og $T$ hhv. {lyst} og {1}, dvs. `(intercept)` svarer til $B \times T$ gruppen {lyst,1}. Heraf ses flg. niveauer

$$\begin{aligned}
\alpha_{11} &= 6.5885 && = 6.5885 \\
\alpha_{12} &= 6.5885 - 2.1508 && = 4.4377 \\
\alpha_{13} &= 6.5885 - 2.7813 && = 3.8072 \\
\alpha_{21} &= 6.5885 - 0.6933 && = 5.8952 \\
\alpha_{22} &= 6.5885 - 0.6933 - 2.1508 + 2.4769 && = 6.2213 \\
\alpha_{23} &= 6.5885 - 0.6933 - 2.7813 + 2.1438 && = 5.2577
\end{aligned}$$

   De tre der spørges til er hhv. $\alpha_{11} = 6.5885, \alpha_{21} = 5.8952$ og $\alpha_{23} = 5.2577$.

4. Opskriv konfidensintervaller for middelværdierne tilhørende de faste effekter i modellen med vekselvirkning, dvs. hvor $\xi \in L_{T \times B}$. Kommenter på effekten af de to behandlinger.

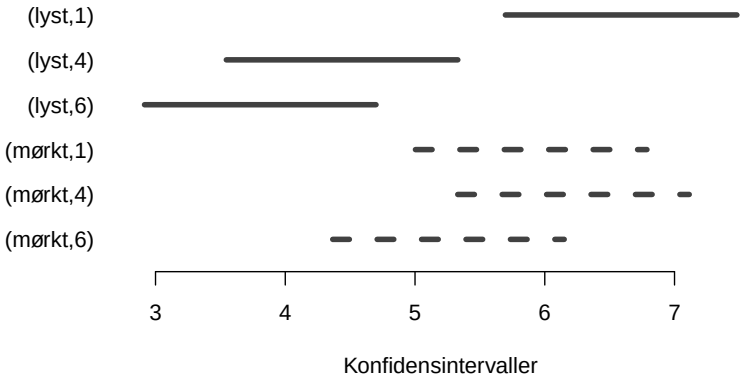
   **Vejledende besvarelse**
   Her benyttes `confint(...)` til at estimere konfidensintervallerne. Vi omparametriserer modellen så de faste effekter svarer til niveauerne i de 6 grupper af $B \times T$ faktoren:

```
m0.1 = lmer(farve~opbevaring:tid-1+(1|gris),data=koed)
```

   Det giver flg. konfidensintervaller for de faste effekter (med `confint(m0.1)`):

| | 2.5 % | 97.5 % |
|---|---|---|
| opbevaringlyst:tid1 | 5.70 | 7.48 |
| opbevaringlyst:tid4 | 3.55 | 5.33 |
| opbevaringlyst:tid6 | 2.92 | 4.70 |
| opbevaringmoerkt:tid1 | 5.00 | 6.79 |
| opbevaringmoerkt:tid4 | 5.33 | 7.11 |
| opbevaringmoerkt:tid6 | 4.37 | 6.15 |

Det ses at intervallerne for $B \times T$ grupperne $\{\text{mørkt}, i\}$, $i = 1, 2, 3$ overlapper hinanden, hvorimode$\{\text{lyst}, i\}$, $i = 1, 2, 3$ ikke gør: de to intervaller for grupperne til tid 4 og 6 er signifikant lavere end tid 1. Plotter vi konfidens intervallerne er tendensen endnu tydeligere




Det konkluderes derfor at kød der opbevares lyst ser ud til at forringes i farve kvalitet over de tre tidsperioder, mens kød der opbevares mørkt ikke viser samme trend.

5. Angiv et forslag til at teste hvorvidt tid har nogen effekt når kødet er opbevaret mørkt. Du behøver ikke at udføre testet.

**Vejledende besvarelse**

Vi kan definere en ny faktor der er hhv. 0 for kød der er opbevaret mørkt og 1 for kød der er opbevaret lyst. Benytter vi faktoren som numerisk (dummy-variabel), kan vi derfor lave interaktionen med `tid`, således at den kun har effekt for gruppen af kød der er opbevaret lyst.

Vi kan lave dummy variablen ved følgende kommando i `R`:

```
koed$opbevaring2 = 1*(koed$opbevaring == "lyst")
```

og kan nu fitte en modellen med den numeriske dummy `opbevaring2` som

```
m2 = lmer(farve~tid:opbevaring2+(1|gris),data=koed)
```

Denne model har interaktion mellem `tid` og `opbevaring2`, men da `opbevaring2` er en dummy variabel, og {mørkt} gruppen bliver reference i $B$ faktoren vil `(intercept)` parameteren svare til niveauet for {mørkt} gruppen. Interaktionen gør at kontrasterne til {lyst} gruppen er forskellige til hvert tidspunkt. Modellen har derfor 4 parametre for de faste effekter:

| Parameter | Gruppe | Estimat |
|---|---|---|
| `(intercept)` | $\{\text{mørkt}, i\}$, $i = 1, 2, 3$ | 5.7914 |
| `tid1:opbevaring2` | $\{\text{lyst}, 1\}$ | 0.7971 |
| `tid4:opbevaring2` | $\{\text{lyst}, 2\}$ | -1.3537 |
| `tid6:opbevaring2` | $\{\text{lyst}, 3\}$ | -1.9842 |

Bemærk at parametrene til {lyst} gruppen er kontraster, dvs. `(intercept)` parameteren skal lægges til hver af de tre {lyst} grupper, dvs. {mørkt} gruppen har et konstant niveau på 5.79, mens {lyst} gruppens niveauer aftager over tid.

Et `anova(m2,m0)` test (dvs. i forhold til startmodellen) giver en $p$-værdi på 0.2638. Ser vi på konfidensintervallerne for de faste parametre, ses det at det kan antages at kød der opbevares 1 dag kan antages at have samme niveau for de to opbevaringsmetoder, eftersom `tid1:opbevaring2` ikke signifikant forskellig fra 0, der ligger i intervallet.

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 5.19 | 6.39 |
| tid1:opbevaring2 | -0.20 | 1.79 |
| tid4:opbevaring2 | -2.35 | -0.36 |
| tid6:opbevaring2 | -2.98 | -0.99 |

Man kunne evt. undersøge nærmere om der er forskel på kød der har ligget lyst i hhv. 4 og 6 dage. I det tilfælde må det antages at farven aftager ikke-lineært over tid, men med kun 3 tidspunkter kan det være svært at afgøre en mere konkret effekt af kombinationen lyst opbevaret kød og tid.