STATISTIK 1
INSTITUT FOR MATEMATISKE FAG, KU

HELLE SØRENSEN
JANUAR 2017

Supplerende opgave til Statistisk 1

Dette dokument indeholder supplerende opgaver til kurset Statistik 1 (Institut for Matematiske Fag, KU). Dokumentet opdateres i løbet af kurset. Når der i det følgende refereres til EH, menes der til *Introduktion til Matematisk Statistisk* af Ernst Hansen (3. udgave).

Kommentarer til HS.1–HS.4

Opgaverne HS.1–HS.4 handler alle om det samme datasæt. I opgave HS.1 introduceres data og der laves nogle R-datasæt der skal bruges i HS.2–HS.4 til diverse analyser.

Der er flere formål med opgaverne:

- At repetere stof fra SS; både teori, fortolkninger og R
- Selv at køre analyserne i R. Mange kommandoer er givet i opgaverne, men husk hver gang at sikre dig at du forstår hvad der foregår!
- Så småt at tilpasse opskrivningen af normalfordelingsmodellerne til den måde vi vil bruge på Stat1, og indse at samtlige modeller kan fittes med R-funktionen `lm`.

Opgaverne er forøvrigt omskrivninger af to gamle eksamensopgaver fra SS (opgave 2 fra januar 2012 og opgave 3 fra juni 2010).

Opgave HS.1

I et forsøg indgik 144 katte, nemlig 97 hanner og 47 hunner. Hver kat blev vejet og derefter aflivet. Hjertet blev taget ud og vejet. Datasættet `cats` i R-pakken `MASS` indeholder tre variable:

- `Bwt`: Kropsvægt i kg
- `Hwt`: Vægten af hjertet i gram
- `Sex`: Køn (`F` for hunner og `M` for hanner)

1. Gør datasættet tilgængeligt i R og forstå strukturen af datasættet med følgende kommandoer:

```
library(MASS)
cats
head(cats)
dim(cats)
summary(cats)
```

2. Brug funktionen `transform` til at lave en ny variabel i datasættet som angiver hvor stor en procentdel af kropsvægten der udgøres af hjertet. Variablen skal hedde `pct`. Check efterfølgende at datasættet er blevet lavet korrekt, fx med `dim`, `head` og/eller `summary`.

   Vink: Se side 4 i notatet *R i Statistik 1*.

3. Brug funktionen `subset` til at lave to nye datasæt: `maleData` med data fra hanner og `femaleData` med data for hunner. Check efterfølgende at datasættene er blevet lavet korrekt, fx med `dim`, `head` og/eller `summary`.

   Vink: Se side 5 i notatet *R i Statistik 1*.

Opgave HS.2

Formålet med denne opgave er at repetere lineær regression fra SS, se evt. kapitel 6 i *Introduktion til Statistik* af Ditlevsen og Sørensen.

I denne opgave skal du kun bruge deldatasættet `maleData` fra opgave HS.1.

1. Lav et (`Bwt`, `Hwt`) scatterplot. Opskriv derefter en lineær regressionsmodel (med papir og blyant), og fit til sidst modellen i R. Du kan bruge kommandoerne

```
with(maleData, plot(Bwt, Hwt))
linreg <- lm(Hwt~Bwt, data=maleData)
```

2. Brug de standardiserede residualer fra modellen til at udføre modelkontrol. Husk at de fittede værdier og de standardiserede residualer kan udtrækkes med `fitted(linreg)` og `rstandard(linreg)`.

   Vær eksplicit omkring hvordan du kontrollerer de enkelte forudsætninger i modellen. Kan alle antagelserne kontrolleres ved hjælp af residualerne? Er der samlet set grund til bekymring vedrørende modellens egnethed til at beskrive data?

3. Brug kommandoen `summary(linreg)`, og angiv estimater for samtlige parametre i modellen. Angiv også et estimat for forskellen i forventet hjertevægt for to katte der har en forskel i kropsvægt på 0.5 kg.

4. Bestem et 95% konfidensinterval for hældningsparameteren vha. formlen

$$\text{estimat} \pm t\text{-fraktil} \cdot \text{SE}.$$

   Vink: Husk at `qt` kan bruges til at finde fraktiler i t-fordelingen. Hvilken fraktil og hvor mange frihedsgrader skal du bruge?

5. Prøv kommandoen `confint(linreg)`, og genfind konfidensintervallet fra spørgsmål 4.

6. Hvis x_i er kropsvægten og Y_i er den stokastiske variabel svarende til vægten af hjertet for kat i, så er en af modelantagelserne at $\mathrm{E}(Y_i) = \alpha + \beta x_i$. Betragt nu $\xi \in \mathbb{R}^{97}$ defineret ved

$$\xi = \begin{pmatrix} \mathrm{E}(Y_1) \\ \vdots \\ \mathrm{E}(Y_{97}) \end{pmatrix}$$

   Opskriv en matrix A således at ξ kan skrives som $\xi = A \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. Matricen A kaldes for en *designmatrix*.

Opgave HS.3

Formålet med denne opgave er at repetere analyse af to stikprøver fra SS, se evt. kapitel 5 i *Introduktion til Statistik* af Ditlevsen og Sørensen.

Vi skal analysere variablen `pct` fra opgave HS.1.

1. Opskriv en statistisk model (med papir og blyant) der gør det muligt at undersøge om der er forskel på den procentdel som hjertets vægt udgør af kropsvægten for hunkatte og hankatte.

2. Bestem estimater for samtlige parametre i modellen. Husk at der kun er en variansparameter i modellem. Du får brug for R-funktionerne `mean` og `var`.

3. Undersøg om modellens forudsætninger med rimelighed kan antages at være opfyldt. Du får blandt andet brug for R-funktionen `qqnorm`.

4. Bestem et estimat og et 95% konfidensinterval for den forventede forskel mellem hankatte og hunkatte i procentdelen som hjertets vægt udgør af den samlede kropsvægt. Du må gerne bruge funktionen `t.test`, men husk at modellen antager at varianserne er ens i de to grupper.

   Er der belæg for at sige at procentdelen som hjertets vægt udgør af den samlede kropsvægt er forskellig for hankatte og hunkatte?

5. Prøv følgende kommandoer, og find de relevante størrelser i outputtet:

   ```
   fit <- lm(pct ~ Sex, data=cats)
   summary(fit)
   confint(fit)
   ```

   Find desuden „`Residual standard error`“ i outputtet fra `summary` og kvadrér tallet. Sammenlign med spørgsmål 2.

   Tænk på `fit` på følgende måde: Kommandoen med `lm` fitter en model til `pct` hvor middelværdien afhænger af værdien af `Sex` — næsten som i lineær regression. Men eftersom `Sex` er en kategorisk variabel, svarer dette nu til situationen med to stikprøver. Vi kommer tilbage til dette senere.

6. Definer $Z_1, \ldots, Z_{144}$ som de stokastiske variable svarende til `pct` 144 katte. Vi har antaget at $\mathrm{E}(Z_i) = \alpha_1$ for $i = 1, \ldots, 97$ (hannerne) og $\mathrm{E}(Z_i) = \alpha_2$ for $i = 98, \ldots, 144$ (hunnerne). Opskriv en designmatrix C så

$$\begin{pmatrix} \mathrm{E}(Z_1) \\ \vdots \\ \mathrm{E}(Z_{144}) \end{pmatrix} = C \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Opgave HS.4

Formålet med denne opgave er at repetere analyse af en enkelt stikprøve fra SS, evt. kapitel 4 i *Introduktion til Statistik* af Ditlevsen og Sørensen.

Du skal bruge datasættet `femaleData` fra opgave HS.1.

Lad U_i være den stokastiske variabel svarende til procentdelen for kat i, $i = 1, \ldots, 47$. I modellen for en enkelt stikprøve antages det at alle U_i'erne er uafhængige og at alle $U_i \sim N(\mu, \sigma^2)$.

1. Bestem estimaterne for μ og σ, og bestem derefter et 95% konfidensinterval for μ ved at indsætte tallene i den relevante formel. Du får brug for R-funktionerne `mean`, `sd` og `qt`.

2. Beregn de samme størrelser mere "automatisk" vha. nedenstående kommandoer. Du skal have lavet datamanipulationerne i opgave HS.1 for at de virker. Husk at tænke over "hvad der sker" undervejs.

   - Brug kommandoen `t.test(femaleData$pct)` og genfind de relevante størrelser.
   - Brug i stedet kommandoerne

     ```
     oneSample <- lm(pct ~ 1, data=femaleData)
     summary(oneSample)
     confint(oneSample)
     ```

     og genfind de relevante størrelser.

3. Opskriv en designmatrix B således at

$$\begin{pmatrix} \mathrm{E}(U_1) \\ \vdots \\ \mathrm{E}(U_{144}) \end{pmatrix} = B\mu.$$

Opgave HS.5

Lad X_1 og X_2 være uafhængige og identisk fordelte reelle stokastiske variable med

$$\mathrm{E}X_1 = \mathrm{E}X_2 = 0, \quad \mathrm{V}X_1 = \mathrm{V}X_2 = 1$$

Lad desuden $\rho \in (-1,1)$ være en konstant, og definer

$$Y_1 = \frac{1}{\sqrt{1-\rho^2}} X_1, \quad Y_2 = \frac{\rho}{\sqrt{1-\rho^2}} X_1 + X_2,$$

1. Opskriv variansmatricen for $X = (X_1, X_2)^T$.

2. Opskriv matricen B, således at $Y = (Y_1, Y_2)^T$ kan skrives som $Y = BX$. Bestem derefter variansmatricen for Y.

   Vink: Brug sætning 19.31 fra MI. Som det fremgår af sætningen, gælder regnereglen ikke kun normalfordelte variable!

Korrelationen mellem to stokastiske variable Z og U der begge har andetmoment og begge har varians der ikke er nul, defineres som

$$\mathrm{Corr}(Z,U) = \frac{\mathrm{Cov}(Z,U)}{\sqrt{\mathrm{V}Z \cdot \mathrm{V}U}} = \frac{\mathrm{E}\big((Z-EZ)(U-EU)\big)}{\sqrt{\mathrm{V}Z \cdot \mathrm{V}U}},$$

se evt. definition 19.21 i MI. Der gælder altid $-1 \leq \mathrm{Corr}(Z,U) \leq 1$, se MI side 474.

For udfald $(z_1, \ldots, z_n)$ og $(u_1, \ldots, u_n)$ defineres den empiriske korrelation som

$$r = \frac{\frac{1}{n-1}\sum (z_i - \bar{z})(u_i - \bar{u})}{\sqrt{s_z^2 \cdot s_u^2}}$$

hvor alle summer løber fra 1 til n; $\bar{z}$ og $\bar{u}$ er de almindelige gennemsnit, og s_x^2 og s_u^2 er de sædvanlige empiriske varianser.

3. Bestem $\text{Corr}(X_1, X_2)$ og $\text{Corr}(Y_1, Y_2)$.

Antag nu yderligere at X_1 og X_2 er normalfordelte.

4. Lad $\rho = 0.5$. Simulér 200 udfald af (Y_1, Y_2). Giv vektorerne med de simulerede værdier navnene `y1` og `y2`, og lav et (`y1`,`y2`)-scatterplot. Beregn desuden den empiriske korrelation med kommandoen `cor(y1,y2)`.

   Vink: Simulér først udfald af X_1 og X_2 vha. `rnorm`.

5. Gentag spørgsmål 4 med følgende værdier af ρ: 0.8, -0.8, 0.2, 0.99.

Opgave HS.6

Lad $X_1, \ldots, X_{10}$ være uafhængige stokastiske variable, der alle er $N(0,1)$ fordelt, og lad $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ være gennemsnittet.

1. Angiv fordelingen af $\bar{X}$.

2. "Eftervis" ved simulation at fordelingsresultatet faktisk er sandt. Dermed menes at du skal simulere et stort antal udfald af $\bar{X}$, og undersøge om de simulerede værdier ser ud til at komme fra den angivne fordeling.

   Du kan benytte skitsen til R-kode nedenfor, hvor *** skal erstattes med relevante kode. Vær sikker på at du forstår konstruktionen og de enkelte skridt i koden, og overvej omhyggeligt hvad du kan konkludere fra simulationerne.

```
### Initialisering
gns <- rep(NA,5000)

### Selve simulationerne
for (i in 1:5000)
  {
    x <- ***         # Simuler 10 N(0,1)-udfald
    gns[i] <- ***    # Beregn gennemsnittet af x-værdierne
  }

### Undersøgelse af de simulerede værdier
hist(gns, prob=TRUE)                        # Normeret histogram
f <- function(x) dnorm(x, mean=0, sd=***)   # Den relevante tæthed
plot(f,-1,1,add=TRUE)                       # Tilføj tæthed til graf
qqnorm(gns)                                 # N-QQ plot
abline(***,***)                             # Den relevante rette linie
```

3. Lad nu $\tilde{X}$ betegne medianen af $X_1, \ldots, X_{10}$. Undersøg fordelingen af $\tilde{X}$ ved simulation. Specielt:

   - Kan fordelingen af $\tilde{X}$ approksimeres med en normalfordeling?
   - Beregn estimater for middelværdi og varians af $\tilde{X}$. Har $\bar{X}$ eller $\tilde{X}$ størst varians?
   - Lav et scatterplot af de simulerede værdier af gennemsnit og median, og beregn et estimat for $\text{Corr}(\bar{X}, \tilde{X})$ ved hjælp af funktionen `cor`.

Opgave HS.7

Lad X være en to-dimensional stokastisk variabel med middelværdivektor og variansmatrix givet ved

$$\mathrm{E}X = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathrm{V}X = \begin{pmatrix} 4 & 4 \\ 4 & 16 \end{pmatrix},$$

1. Angiv $\mathrm{E}X_1$, $\mathrm{E}X_2$, $\mathrm{V}X_1$, $\mathrm{V}X_2$, $\mathrm{Cov}(X_1, X_2)$ og $\mathrm{Corr}(X_1, X_2)$.

Definer nu den to-dimensionale stokastiske variabel Y ved $Y_1 = X_1 + 2$ og $Y_2 = X_2 - X_1 + 1$.

2. Bestem middelværdivektoren og variansmatricen for Y.
3. Kan du sige noget om hvorvidt Y_1 og Y_2 er uafhængige?
4. Antag nu desuden at den oprindelige stokastiske variabel X er regulært normalfordelt (jf. definition 18.22 fra MI). Hvad er så fordelingen af Y? Kan du nu sige noget om hvorvidt Y_1 og Y_2 er uafhængige?

Opgave HS.8

Lad V være vektorrummet af andengradspolynomier på $(-1,1)$, og udstyr det med det indre produkt der giver L^2-normen:

$$\mathscr{P}_2 = \{f : (-1,1) \to \mathbb{R}^2 \,|\, \exists a_0, a_1, a_2 \in \mathbb{R} \text{ så } f(x) = a_2x^2 + a_1x + a_0\}$$
$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,dx, \quad f, g \in \mathscr{P}_2$$

Opgaven går ud på at simulere fra den regulære normalfordeling på $\mathscr{P}_2$ med centrum 0 og præcision $\langle\cdot,\cdot\rangle$. Vi skal bruge EH sætning 9.28 og har derfor brug for en ortonormal basis.

Definer først $p_0, p_1, p_2 : (-1,1) \to \mathbb{R}$ ved

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = 3x^2 - 1.$$

1. Overbevis dig selv om at $\{p_0, p_1, p_2\}$ en basis for $\mathscr{P}_2$. Du behøver ikke lave argumenterne formelt, men gør dig nogle fornuftige overvejelser.
2. Overbevis dig selv om at p_0, p_1, p_2 er parvis ortogonale. Du må gerne gøre det numerisk, fx. vha. følgende kommandoer i R:

```
p0 <- function(x) 1
p1 <- function(x) x
p2 <- function(x) 3*x^2 - 1
p1p2 <- function(x) p1(x)*p2(x)
integrate(p1,-1,1)
integrate(p2,-1,1)
integrate(p12,-1,1)
```

Det viser sig at

$$\|p_0\|^2 = 2, \quad \|p_1\|^2 = \frac{2}{3}, \quad \|p_2\|^2 = \frac{8}{5}$$

således at

$$e_0(x) = \frac{1}{\sqrt{2}}, \quad e_1(x) = \sqrt{\frac{3}{2}}x, \quad e_2(x) = \sqrt{\frac{5}{8}}(3x^2 - 1).$$

udgør en ortonormal basis for $\mathscr{P}_2$.

3. Overvej hvordan du kan bruge sætning 9.28 til at simulere udfald fra den regulære normalfordeling på $\mathscr{P}_2$ med centrum 0 og præcision $\langle\cdot,\cdot\rangle$. Simuler et enkelt udfald fra fordelingen (altså et andengradspolynomium), og tegn grafen for dette udfald.
4. Simuler 25 udfald fra fordelingen og tegn alle polynomierne i samme graf. Du kan fx. bruge følgende kode, hvor `***` skal erstattes af relevante kode:

```
plot(0,0, type="n", xlim=c(-1,1), ylim=c(-3,3), xlab="x", ylab="f(x)")
for (i in 1:25)
  {
   f <- function(x) ***
   plot(f,-1,1, add=TRUE)
  }
```

5. Antag at X er regulært normalfordelt på $\mathscr{P}_2$ med centrum 0 og præcision $\langle\cdot,\cdot\rangle$. Betragt desuden afbildningen $t : \mathscr{P}_2 \to \mathbb{R}$ givet ved $t(f) = f(0)$, og definer en ny stokastisk variabel $Z = t(X)$. Bestem fordelingen af Z.

Opgave HS.9

Betragt den lineære normale model på $\mathbb{R}^N$, dvs. antag at $X \sim N(\xi, \sigma^2 I)$ med ukendte parametre $(\xi, \sigma^2) \in L \times (0, \infty)$ hvor L er et underrum af $\mathbb{R}^N$. Antag at $\dim(L) = k$ og at $N \times k$ matricen A er en designmatrix for modellen således at $L = \{A\beta | \beta \in \mathbb{R}^k\}$.

Lad desuden C være en $1 \times k$ matrix (ikke 0-matricen), og betragt hypotesen

$$H : C\beta = 0.$$

Hypotesen udtrykker at en linearkombination af elementerne i β, svarende til vektoren C, er 0. Hypotesen kan også skrives som $\xi \in L^*$ hvor

$$L^* = \{A\beta | \beta \in \mathbb{R}^k \text{ og } C\beta = 0\}$$

1. Gør rede for at L^* er et underrum, at $L^* \subset L$, og at $\dim(L^*) = k - 1$.

Lad $\hat{\beta}$ være MLE for β og $\tilde{\sigma}^2$ være den centrale estimator for σ^2 i modellen (altså ikke under hypotesen), og definer

$$T = \frac{C\hat{\beta}}{\tilde{\sigma}\sqrt{C(A^TA)^{-1}C^T}}$$

2. Find fordelingen af $C\hat{\beta}$ under hypotesen.
3. Vis at T under hypotesen er t-fordelt med $N - k$ frihedsgrader.

   Vink: Du kan benytte at hvis X_1 og X_2 er uafhængige, $X_1 \sim N(0,1)$ og $X_2 \sim \chi^2_f$ (f frihedsgrader), så er $\frac{X_1}{\sqrt{X_2/f}}$ t-fordelt med f frihedsgrader. Dette følger efter nogle overvejelser fra MI, eksempel 20.27, men du må gerne bruge resultatet direkte.
4. Forklar hvordan T kan bruges til at teste hypotesen. Mere specifikt: Hvilke værdier af T er kritiske? Hvordan beregnes p-værdien?

Husk F-teststørrelsen for hypotesen, formel (10.31) fra bogen. Man kan vise at $T^2 = F$.

5. Gør det nogen forskel om vi vælger at teste hypotesen ved hjælp af F-testet eller t-testet, dvs. får vi den samme p-værdi med de to test? For hvilken slags hypoteser har vi valget mellem F-test og t-test?

De sidste par spørgsmål går ud på faktisk at vise at $T^2 = F$ og er lidt regnetunge...

6. Vis at ortogogonalprojektionen af $\mathbb{R}^N$ på L^* er givet ved $p^*(x) = P^*x$ hvor

$$P^* = A(A^TA)^{-1}A^T - A(A^TA)^{-1}C^T\left(C(A^TA)^{-1}C^T\right)^{-1}C(A^TA)^{-1}A^T$$

   Vink: Brug Sætning 10.7.
7. Vis at $T^2 = F$.

   Vink: Regn løs på $\|PX - P^*X\|^2$ hvor P^* er givet i spørgsmål 6, og genkend at visse dele har noget med $\mathrm{V}(C\hat{\beta})$ at gøre.

Opgave HS.10

Lad $Y_1, \ldots, Y_3$ være uafhængige stokastiske variable der hver især er standard normalfordelte, $Y_i \sim N(0,1)$. Definer $X = CY$ hvor

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ -2 & -1 & -1 \end{pmatrix},$$

1. Vis at X er singulært normalfordelt på $\mathbb{R}^3$.
2. Betragt afbildning givet ved C, altså $s : \mathbb{R}^3 \to \mathbb{R}^3$, $s(x) = Cx$. Gør rede for at billedmængden er $U = \{x \in \mathbb{R}^3 | x_1 + x_2 = -x_3\}$, og at X derfor er regulært normalfordelt på U.

Opgave HS.11

Denne opgave er et supplement til EH 10.5, og kan kun laves hvis EH 10.5 er lavet i forvejen.

1. Lav designmatricen i R, og kald den `A`. Lav også en vektor `x` der indeholder observationerne.
2. Prøv nedenstående kommandoer, og genkend estimaterne for α_1, α_2 og σ:

```
model <- lm(x ~ A-1)
model.matrix(model)
summary(model)
```

3. Prøv kommandoen `vcov(model)`. Hvad er sammenhængen mellem denne matrix og matricen $(A^T A)^{-1}$?
4. Hvad er sammenhængen mellem diagonalelementerne i `vcov(model)` og de standard errors som angives i `summary(model)`?

Opgave HS.12

Funktionen `rnorm` kan bruges til at simulere fra uafhængige normalfordelinger. Funktionen tager tre argumenter:

- `n`: antal observationer
- `mean`: middelværdivektoren
- `sd`: vektor af spredninger

Hvis middelværdierne er ens, kan man nøjes med at angive den fælles værdi. Tilsvarende for spredningerne. For eksempel giver `rnorm(n=2, mean=c(1,2), sd=c(3,4))` to uafhængige normalfordelte observationer, den ene fra $N(1,9)$ og den anden fra $N(2,16)$, mens kommandoen `rnorm(3,0,4)` giver tre uafhængige observationer fra $N(0,16)$. Bemærk specielt at det er spredningerne — ikke varianserne — der gives som argument til `sd`.

1. Simuler 1000 uafhængige observationer fra $N(1,0.01)$. Tegn derefter et histogram og et QQ-plot over de simulerede værdier.
2. Simuler — med et enkelt kald til `rnorm` — 500 uafhængige observationer fra $N(0,1)$ og 500 uafhængige observationer fra $N(2,0.25)$. Tegn derefter et histogram og et QQ-plot over de simulerede værdier. Forklar hvad du ser på figurerne!

   Vink: Du kan sikkert have nytte af at bruge `c(rep(0,500), rep(2,500))`.

Opgave HS.13

Denne opgave består af diverse simulationseksperimenter med udgangspunkt i eksempel 11.10 fra EH. Den vigtigste mission med opgaven er (så småt) at få jer til at indse at simulationseksperimenter er et fantastisk redskab til at opnå indsigter som ville være vanskelige at opnå med papir og blyant.

Det er vigtigt at du skriver koden i et R-program i stedet for direkte ved prompten. Du får sikkert brug for lignende konstruktioner en anden gang.

Del 1 Først skal vi fitte den kvadratiske regressionsmodel til data, simulere datasæt fra den fittede model, og bruge de simulerede datasæt til at vurdere om modellen egner sig til at beskrive data. Data er tilgængelige på Absalon i filen `paddy.txt`.

1. Indlæs data, og fit den kvadratiske regressionsmodel med kommandoerne

   ```
   daysSqr <- days^2
   kvadReg <- lm(yield ~ days + daysSqr)
   ```

   Check at du får samme estimater som i eksempel 11.10.
2. Lav det tilhørende residualplot.

Lad os nu lade som om de sande værdier for parametrene er lig estimaterne. Vi antager således at $X \sim N(\xi_0, \sigma_0^2 I)$ hvor ξ_0 og σ_0^2 er estimaterne fra spørgsmål 1. Vi skal nu simulere fra denne fordeling.

3. Kør følgende kommandoer og forklar hvorfor `simYield` indeholder en simulation fra modellen:

```
xi0 <- fitted(kvadReg)
sigma0 <- summary(kvadReg)$sigma
simYield <- rnorm(16, xi0, sigma0)
```

   Lav derefter et (`days`,`simYield`) plot, og se om det ser fornuftigt ud (dvs. ligner det tilsvarende plot for de rigtige data).

4. Fit den kvadratiske regressionsmodel til de simulerede data, og lav det tilhørende residualplot.

5. Gentag succesen nogle gange, dvs. simuler data, fit modellen og tegn residualplottet. Ser residualplottet baseret på de rigtige data (fra spørgsmål 1) værre ud end for de simulerede data? Hvad er konklusionen vedrørende modellens egnethed til at beskrive de rigtige data?

Del 2 Risdyrkerne er interesseret i at estimere det optimale høsttidspunkt og det maksimale udbytte, samt at sige noget om usikkerheden på disse estimater. Vi skal nu se hvordan vi kan bruge simulationer til at sige noget begavet om den sag.

6. Brug parameterestimaterne fra fittet i spørgsmål 1 til komme med fornuftige estimater for det optimale høsttidspunkt og det maksimale udbytte.

   Vink: Brug for eksempel følgende kommandoer (forklar hvad der foregår!):

```
est <- coef(kvadReg)
optDay <- -est[2]/2/est[3]
optYield <- - (est[2]^2 - 4*est[1]*est[3]) / 4 / est[3]
```

Det var jo ikke så svært, men vi vil også gerne have en ide om fordelingen af de tilhørende estimatorer, samt kunne bestemme konfidensintervaller.

7. Forklar hvorfor teorien fra EH, kapitel 9 og 10 ikke umiddelbart giver os fordelingen af estimatoren for det optimale høsttidspunkt.

8. Heldigvis kan vi simulere! Først laver vi en enkelt simulation:

   - Simuler data efter den kvadratiske regressionsmodel, præcis som i spørgsmål 3.
   - Fit den kvadratiske regressionsmodel til de simulerede data. Kald fx modelobjektet for `simModel`.
   - Træk estimaterne ud, fx med følgende kommando: `est <- coef(simModel)`.
   - Beregn estimater for det optimale høsttidspunkt og det optimale udbytte baseret på disse estimater.

9. Dette vil vi gerne gøre mange gange. Prøv følgende kommandoer. Kør følgende programstump, hvor du i linierne med # indsætter kode for forrige spørgsmål. Vær sikker på at du forstår hvert enkelt trin i koden.

```
res <- matrix(NA, 1000, 2)   # Initialiering af matrix
for (i in 1:1000)            # Løkke
  {
```

```
  # simuler et datasæt
  # fit den kvadratiske regressionsmodel
  # træk estimaterne ud
  # beregn det optimale høsttidspunkt
  # beregn det optimale udbytte
  # res[i,1] <- optimalt høsttidsp.  # Læg værdien på plads (i,1)
  # res[i,2] <- optimalt udbytte     # Læg værdien på plads (i,2)
}
```

10. Tegn histogrammer med kommandoerne `hist(res[,1])` og `hist(res[,2])`. Hvilken type fordeling ser det ud til at estimatorerne har?

11. Overvej hvordan du kan lave et (approksimativt) konfidensinterval for det optimale høsttidspunkt.

Del 3 Lad os endelig se hvad der kan ske hvis vi fitter modeller hvor der er problemer med antagelserne om middelværdi eller varians.

12. Lad os først se hvad der sker når der er problemer med middelværdien:

    - Simuler data efter den kvadratiske regressionsmodel, præcis som i spørgsmål 3.
    - Fit den simple lineære regressionsmodel — altså uden det kvadratiske led — til de simulerede data, og tegn residualplottet. Gentag evt. proceduren nogle gange.
    - Forklar hvorfor residualplotttet ser ud som det gør.

13. Lad os så se hvad der sker når der er problemer med variansen:

    - Kør følgende kommandoer og forklar hvad der sker:

      ```
      newSD <- 25*(15-abs(days-31))
      newSD
      simYield <- rnorm(16, xi0, newSD)
      plot(days, simYield)
      points(days, xi0, type="l")
      ```

    - Fit den kvadratiske regressionsmodel til de simulerede data, og tegn residualplottet. Gentag evt. proceduren nogle gange.
    - Forklar hvorfor residualplottet ser ud som det gør.

Opgave HS.14

I et eksperiment med 20 kvinder mellem 25 og 34 år har man for hver kvinde målt fedtprocenten i kroppen, tykkelsen af hudfolden ved triceps (overarmen), samt omkreds af arm og lår. Data er tilgængelige på Absalon i filen `bodyfat.txt`. Se evt. slides fra uge 3, torsdag, for en lignende analyse (eksemplet om kirsebærtræer).

1. Man er hovedsageligt i at kunne prædiktere fedtprocent vha. de øvrige variable. Overvej hvilken model der vil være velegnet til dette formål.

2. Indlæs data i et R-datasæt, kald det fx `bodyfat`. Prøv kommandoerne `plot(bodyfat)` og `cor(bodyfat)`, og overvej hvad grafen og outputtet fortæller dig.

3. Fit følgende model:

```
m1 <- lm(Fat ~ Triceps + Thigh + Midarm, data=bodyfat)
```

   Er det den samme model som du nåede frem til i spørgsmål 1? Udfør modelkontrol, og overvej om modellen er egnet til at beskrive data.

4. Lav et `summary` på modellen. I `summary` rapporteres p-værdierne for en række t-tests. Gør rede for hvilke hypoteser der testes, og hvad vi kan konkludere fra testene.

5. Prøv kommandoerne

```
m2 <- lm(Fat ~ Midarm, data=bodyfat)
anova(m2, m1)
```

   Hvilken hypotese er det der testes, og hvad er konklusionen? Hvorfor er dette ikke i modstrid med konklusionerne fra t-testene fra spørgsmål 4?

I en multipel regressionsmodel "reducerer" (simplificerer) man ofte modellen ved at fjerne ikke-signifikante forklarende variable en ad gangen, indtil alle forklarende variable er signifikante. Mere detaljeret:

- Hvis alle de forklarende variable har en signifikant effekt, siger vi at modellen ikke kan reduceres.
- Hvis der er netop en forklarende variabel der ikke er signifikant, fittes en ny model. Den nye model er identisk med den forrige, bortset fra at den ikke indeholder den ikke-signifikante variabel.
- Hvis der er flere variable der ikke er signifikante, skal man vælge hvilken en man vil tage ud af modellen. Sommetider er der en variabel hvis effekt man er særligt interesseret i. Så beholdes denne variabel i modellen så længe som muligt. Hvis man ikke er særligt interesseret i nogle variable fremfor andre, vælger man ofte at tage den variabel med den højeste p-værdi ud af modellen.

Processen gentages, således at modellen indeholder færre og færre led. Når alle de tilbagværende forklarende variable har en signifikant effekt, stoppes processen, og vi er nået frem til "slutmodellen".

6. Udfør modelreduktion, og indse at man derved når frem til modellen

```
m3 <- lm(Fat ~ Triceps + Midarm, data=bodyfat)
```

   Modelreduktionsprocessen er i dette tilfælde altså ikke ret lang!

7. Kig på figuren fra spørgsmål 2 igen. Undrer det dig at `Midarm` men ikke `Thigh` er med i modellen? Kan du forklare hvad årsagen er?

8. Giver fortegnene på parameterestimaterne i modellen `m3` mening? Husk at der er tale om effekter af en variabel hvis den anden forklarende variabel er uændret.

9. Brug `m3` til at beregne et prædiktion og et 95% prædiktionsinterval for fedtprocenten for en kvinde med en hudfold ved triceps på 25 og en omkreds af armen på 28.

Opgave HS.15

I 1857 indsamlede den tyske statistiker Ernst Engel data fra 235 belgiske husholdninger. Han registrerede indkomsten samt udgiften til mad for hver husstand. Data er tilgængelige på Absalon i filen `engel.txt`. Vi vil lave lineær regression.

1. Overvej hvilken variabel der er naturlig at bruge som responsvariabel (t) hhv. forklarende variabel (x). Lav derefter et scatterplot af data, og overvej hvad du kan se fra grafen.
2. Opskriv den relevante lineære regresssionsmodel, og fit modellen med `lm`. Udfør modelkontrol. Ser det ud til at modelantagelserne er fornuftige?
3. Transformer begge variable med logaritmefunktionen. Opskriv den lineære regressionsmodel for den lineære regressionsmodel for de transformerede variable. Fit modellen og udfør modelkontrol. Ser det ud til at denne model er fornuftig til Engels data?
4. Angiv estimater for samtlige parametre i modellen.

Lad β betegne hældningsparameteren i modellen for de logaritmetransformerede data, dvs. i modellen fra spørgsmål 3. Økonomerne er særligt interesserede i værdien 1 for denne parameter.

5. Forklar hvorfor $\beta = 1$ svarer til at madudgifterne (på nær tilfældig variation) er proportionale med indkomsten — altså at det er en fast procentdel af indkomsten som går til mad, uanset hvor stor eller lille indkomsten er.
6. Angiv et 95% konfidensinterval for β. Er β signifikant forskellig fra 1 på 5% signifikansniveau?
7. Vi vil nu interessere os for hypotesen $H : \beta = 1$. Gør rede for at hypotesen er en affin hypotese i modellen fra spørgsmål 3.
8. Betragt variablen
$$Z = \log(\texttt{foodexp}) - \log(\texttt{income}).$$
Angiv fordelingen af Z i den statistiske model (ikke under hypotesen), samt under hypotesen. Forklar hvordan hypotesen kan testes som et F-test i den statistiske model for Z.
9. Udfør F-testet i R. Husk at konkludere på testet.

Opgave HS.16

I et eksperiment har man inddelt 60 bænkebidere tilfældigt i tre grupper. Bænkebiderne i den ene gruppe blev brugt som kontroller, mens bænkebifderne i de to øvrige grupper blev udsat for stærkt lys henholdsvis fugt. Alle bænkebidere blev sat til at bevæge sig en strækning på 6 tommer, og det blev registreret hvor lang tid hver enkelt bænkebider var om at tilbagelægge de 6 tommer. Det videnskabelige(?) spørgsmål er om lys og fugt påvirker resultatet.

Data ligger på Absalon i filen `pillbug.txt`.

1. Forklar hvorfor eksperimentet lægger op til at lave en ensidet variansanalyse. Gør rede for antagelserne i modellen, og opskriv herunder en designmatrix for modellen.

2. Prøv kommandoen `boxplot(time ~ group)`, og forklar hvad du ser. Fit derefter modellen for ensidet variansanalyse, og lav det tilhørende residualplot. Overvej udfra figurerne om modellen er velegnet til at beskrive data.

3. Gentag spørgsmål 2 hvor du i stedet for at bruge `time` som responsvariabel bruger `log(time)` som responsvariabel. Ser modellen mere fornuftig ud nu? Overvej hvorfor det hjælper at log-transformere.

4. Opstil og test en hypotese om at fordelingen af løbstiden hverken påvirkes af fugt eller lys. Husk at overveje hvilken responsvariabel du bør bruge til at besvare spørgsmålet.

5. Angiv estimaterne for parametrene i modellen fra spørgsmål 3. Parametriseringen skal være den samme som du brugte i spørgsmål 1 da du skrev designmatricen op. Overvej nøje hvordan estimaterne bør fortolkes.

6. Er der en signifikant effekt af fugt på løbstiden? Er der en signifikant effekt af lys på løbstiden.

Opgave HS.17

Denne opgave handler om gammafordelingen. Fordelingen er gennemgået i MI-bogen, blandt andet Eksempel 15.6, og der står ikke noget i opgaven som ikke står i MI-bogen. Formålet med opgaven er at få repeteret nogle centrale egenskaber.

Lad $\lambda > 0$ og definer funktionen

$$f(x) = \frac{1}{\Gamma(\lambda)} x^{\lambda-1} e^{-x} \cdot 1_{(0,\infty)}(x)$$

hvor gammafunktionen Γ er defineret ved $\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x}\, dx$ således at f integrerer til 1 og kan bruges som tæthed. Fordelingen der har tæthed f mht. lebesguemålet kaldes *gammafordelingen med formparameter* λ.

Ved partiel integration kan man indse at gammafunktionen opfylder

$$\Gamma(\lambda+1) = \lambda\Gamma(\lambda), \quad \lambda > 0.$$

1. Lad X være gammafordelt med formparameter λ. Bestem EX og VX.

2. Lad $\beta > 0$, og bestem tætheden for gammafordelingen med formparameter λ og skalaparameter β, dvs. for $Y = \beta X$. Angiv også middelværdi og varians i fordelingen.

Vi skriver sommetider $Y \sim \Gamma(\lambda, \beta)$ hvis Y er gammafordelt med formparameter λ og skalaparameter β.

Man kan vise *foldningsegenskaben* for gammafordelingen, se evt. MI Eksempel 20.1: Hvis Y_1 og Y_2 er uafhængige og $Y_1 \sim \Gamma(\lambda_1, \beta)$ og $Y_2 \sim \Gamma(\lambda_2, \beta)$, så er $Y_1 + Y_2 \sim \Gamma(\lambda_1 + \lambda_2, \beta)$. Bemærk at skalaparametrene skal være ens. Resultatet generaliserer til n variable ved induktion.

3. Lad Z_1 og Z_2 være iid. eksponentialfordelte med middelværdi μ. Hvad er fordelingen af $Z_1 + Z_2$?

   Vink: Hvilken gammafordeling svarer eksponentialfordelingen med middelværdi μ til?

4. Funktionen `rgamma` kan bruges til at simulere udfald fra gammafordelingen. Brug hjælpefunktionen for funktionen til undersøge præcis hvordan (kommando: `?rgamma`).

   Lav derefter passende simulationer hvor du „checker“ dine formler fra spørgsmål 2 vedr. middelværdi og varians for $\lambda = 0.5$ og $\beta = 1.5$.

5. Lav passende simulationer som „checker“ foldningsegenskaben for $\lambda_1 = 0.5$, $\lambda_2 = 1.5$ og $\beta = 2$.

   Vink: Der er forskellige muligheder for at undersøge om givne tal $z_1, \ldots, z_n$ med rimelighed kan antages at komme fra en specifik fordeling.

   - Én mulighed er at transformere med fordelingsfunktionen for den foreslåede fordeling og sammenligne med ligefordelingen (hvorfor?) Funktionen `pgamma` kan være nyttig i denne forbindelse.
   - En anden mulighed er at simulere udfald fra den foreslåede fordeling og sammenligne fraktilerne for de to sæt af simulerede udfald. Denne sammenligning foretages nemt med funktionen `qqplot`.

Opgave HS.18

Betragt fordelingen der har tæthed

$$f_\beta(x) = \beta\, x^{\beta-1} \cdot 1_{(0,1)}(x)$$

mht. lebesuemålet på $\mathbb{R}$. Fordelingen er bestemt af parameteren $\beta > 0$.

Lad $X_1, \ldots, X_n$ være uafhængige reelle stokastiske variable, der alle har fordeling givet ved tætheden f_β med ukendt $\beta > 0$.

1. Vis at $-\log X_i$ er eksponentialfordelt med middelværdi $1/\beta$. Bestem derefter fordelingen af $-\sum_{i=1}^n \log(X_i)$.

   Vink: Brug transformationssætningen og derefter foldningsegenskaben for gammafordelingen.

2. Opskriv likelihoodfunktionen, log-likelihoodfunktionen, scorefunktionen og den observerede informationsfunktion. Angiv også Fisherinformationen.

3. Eftervis direkte at Bartletts identiteter gælder. Du skal altså regne de relevante middelværdier og varianser ud og checke at de opfører sig som Bartlett siger.