Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2022

Fire timers skriftlig prøve. Alle hjælpemidler tilladt, herunder computer, men du må ikke tilgå internettet bortset fra i forbindelse med udlevering og aflevering af eksamensopgaven.

Der er 3 opgaver, som vægtes med henholdsvis 45 %, 30 % og 25 % i bedømmelsen.

Husk at mange delspørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 bliver gjort tilgængelige sammen med selve eksamensopgaven. Alle svar i opgave 1 og 2 skal begrundes. Husk at det ikke er tilstrækkeligt at aflevere R-kode med tilhørende output. Du skal derimod svare på det du bliver spurgt om, typisk ved at du skriver svaret i almindelig tekst, hvori du inddrager relevante tal fra R-outputtet.

Opgave 3 består af multiple choice spørgsmål. For hvert multiple choice spørgsmål er der netop et korrekt svar, og din besvarelse skal udelukkende bestå af bogstavet for dit valgte svar. Du kan altså ikke begrunde svaret. Hvis et multiple choice spørgsmål besvares forkert, ikke besvares, eller flere svar bliver givet, så gives 0 point.

Din besvarelse skal afleveres elektronisk. Du skal følge de officielle instrukser du har modtaget omkring aflevering af opgaven.

Opgave 1

Denne opgave vægtes med 45 % ved bedømmelsen, og svarene skal begrundes.

Ved et dyrkningsforsøg har man inddelt en mark i 32 områder, som hvert er blevet beplantet med en jordbærsort. Der indgår 8 jordbærsorter i forsøget, og hver sort er plantet på fire områder. Formålet med forsøget er at sammenligne udbyttet af de 8 jordbærsorter. Der var en hæk i højre siden af forsøgsmarken, og man har desuden registreret afstanden fra hvert jordområde til hækken.

Datafilerne `nov2022opg1.txt` og `nov2022opg1.xlsx` indeholder data fra forsøget og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "nov2022opg1.xlsx")
```

eller

```
data1 <- read.table(file = "nov2022opg1.txt", header = T)
```

De første linjer i datasættet kan ses her

```
##   variety afstand udbytte
## 1       G       8     5.8
## 2       V       7     6.3
## 3      R1       6     4.9
## 4       F       5     6.5
## 5      Re       4     4.5
## 6       M       3     5.2
```

Variablen `udbytte` angiver udbyttet på hver af de i alt 32 områder med jordbærplanter målt i kg. Jordbærsorten (`variety`) kan antage værdierne `G, V, R1, F, Re, M, E, P`. Variablen `afstand` angiver afstanden (målt i m) fra området til hækken.

Ved besvarelsen af delopgaverne **1.1-1.3** skal du *ikke* benytte variablen `afstand`

1. Forklar hvilken statistisk model som fittes med koden

   ```
   mod1 <- lm(udbytte ~ variety, data = data1)
   ```

   Angiv estimater for det forventede udbytte på områder beplantet med hver af jordbærsorterne `E` og `R1`.

2. Angiv 95 % - konfidensintervaller for det forventede udbytte på områder beplantet med hver af jordbærsorterne `E` og `R1`. Angiv desuden et 95 % - konfidensinterval for forskellen i det forventede udbytte for de to jordbærsorter `E` og `R1`.

3. Undersøg om udbyttet kan antages at være ens for alle 8 jordbærsorter i forsøget.

Da hækken i den ene side af marken skygger for solen en del af dagen, så er det nærliggende at tro, at udbyttet af jordbærplanterne kan afhænge af afstanden til hækken. Ved besvarelsen af delopgaverne **1.4-1.5** skal du *ikke* inddrage variablen `variety`.

4. Opskriv (i din besvarelse) den lineære regressionsmodel, der beskriver en lineær sammenhæng mellem det forventede udbytte og afstanden til hækken.

   Fit modellen i R og angiv estimater for samtlige parametre i modellen.

5. Udfør et test for, om det er rimeligt at antage, at der er en lineær sammenhæng mellem forventet udbytte og afstand til hækken.

   Hint: Der er flere korrekte løsninger på dette spørgsmål. Det er vigtigt at du forklarer, hvilken model du tager udgangspunkt i, for at teste hypotesen.

I delopgave **1.6** skal du inddrage begge variablene `variety` og `afstand` som *kategoriske* variable i modellen.

6. Angiv R-koden til at fitte en additiv model for tosidet variansanalyse, hvor begge variablene inddrages.

   Fit modellen i R og angiv et estimat for det forventede udbytte på et område med sorten `E` som ligger i afstanden 1 m fra hækken.

   Angiv desuden et estimat for forskellen i det forventede udbytte på to områder som ligger henholdvis 2 m og 4 m fra hækken.

I delopgave **1.7** indfører vi en ny variabel $\texttt{x} = \frac{1}{\texttt{afstand}}$ der angiver den *reciprokke værdi af afstanden til hækken.*

7. Opskriv (i din besvarelse) den statistiske model, som fittes med R-koden

```
data1$x <- 1/data1$afstand
mod2 <- lm(udbytte ~ x + variety, data = data1)
```

   Benyt modellen til at bestemme et estimat og et 95 % - prædiktionsinterval for udbyttet på et område som ligger i afstanden 4.2 m fra hækken og som beplantes med jordbærsorten `P`.

Opgave 2

Denne opgave vægtes med 30 % ved bedømmelsen, og svarene skal begrundes. Data er venligst stillet til rådighed af Emma Anemone Kofoed Lauridsen.

Ved planlægning af operationer er det relevant at kunne forudsige længden af operationen ud fra patienternes kliniske data. I denne opgave interesserer vi os for operationstiden for patienter, som har fået fjernet en svulst i hjernen (hjernetumor). Vi betragter et datasæt som indeholder sammenhørende værdier af alder (`age` i år), størrelse af hjernetumor (`volume` i mL) og operationstid (`duration` i minutter) for 71 patienter.

Datafilerne `nov2022opg2.txt` og `nov2022opg2.xlsx` indeholder data og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data2 <- read_excel(path = "nov2022opg2.xlsx")
```

eller

```
data2 <- read.table(file = "nov2022opg2.txt", header = T)
```

De første linjer i datasættet kan ses her

```
##        age  volume duration
## 1 54.71422  39.950      165
## 2 32.29897 139.500       91
## 3 57.37010 100.053      248
## 4 41.75582  31.850      292
## 5 76.42867  36.800      143
## 6 38.19654  54.400      151
```

1. Antag at alderen (`age`) for patienter, som skal opereres for en hjernetumor, kan beskrives ved en normalfordeling $N(\mu, \sigma^2)$ med middelværdi μ og spredning σ.

   Benyt de 71 værdier af `age` i datasættet til at angive et estimat og et 95 % - konfidensinterval for μ.

2. Opskriv den statistiske model svarende til `model3` nedenfor.

   ```
   model1 <- lm(duration ~ volume + age, data = data2)
   model2 <- lm(log(duration) ~ volume + age, data = data2)
   model3 <- lm(log(duration) ~ log(volume) + age, data = data2)
   ```

   Diskuter grundigt hvorfor man bør foretrække `model3` fremfor `model1` og `model2` til at beskrive sammenhængen mellem operationstiden og størrelsen af tumor samt alder.

3. Udfør et test for, om der er sammenhæng mellem patientens alder og operationstiden.

Ved besvarelsen af delopgave **2.4** skal variablen `age` ikke benyttes.

4. Benyt R til at fitte modellen

$$\log(\texttt{duration}_i) = \alpha + \beta \cdot \log(\texttt{volume}_i) + e_i,$$

hvor $e_1, \ldots, e_{71}$ er uafhængige $\sim N(0, \sigma^2)$.

Benyt modellen til at angive hvor meget længere operationstiden må forventes at være, for en patient med tumorstørrelse på 100 mL i forhold til for en patient med tumorstørrelse på 50 mL.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 25 % i bedømmelsen. For hvert delspørgsmål er der netop et korrekt svar. Din besvarelse skal udelukkende bestå af bogstavet for dit valgte svar; du kan altså ikke begrunde svaret. Hvis et spørgsmål besvares forkert, ikke besvares eller flere svar bliver givet, gives 0 point. Du må naturligvis gerne bruge R til opgaven.

3.1 For en lineær regressionsmodel med y som respons og x som forklarende variabel interesserer vi os for hypotesen om, at hældningen er nul. Hvad er en fejl af type II i dette tilfælde?

- a) Vi konkluderer at der ikke er nogen sammenhæng mellem x og y, selvom der i virkeligheden er en sammenhæng
- b) Vi konkluderer at der er en sammenhæng mellem x og y selvom der i virkeligheden ikke er en sammehæng
- c) Vi konkluderer at der ikke er en lineær sammenhæng mellem x og y selvom der er i virkeligheden er en lineær sammenhæng
- d) Vi konkluderer at der er en lineær sammenhæng mellem x og y selvom der er i virkeligheden er en kvadratisk sammenhæng
- e) Vi konkluderer at middelværdien for y og x er forskellige, selvom de to middelværdier i virkeligheden er ens

3.2 For en lineær regressionsmodel baseret på 37 observationer har man testet hypotesen om, at hældningen kan antages at være nul. Der beregnes en t-teststørrelse på -1.732. Hvad bliver den tilhørende P-værdi?

- a) $P = 0.092$
- b) $P = 0.083$
- c) $P = 0.042$
- d) $P = 0.046$
- e) $P = 0.954$

3.3 Den maksimale tid som studerende kan fastholde deres koncentration ved en forelæsning kan antages at være normalfordelt med en middelværdi på 42.9 minutter og en spredning på 12.3 minutter.

Hvilket af følgende udsagn er korrekt?

a) 25 % af de studerende kan højst koncentrere sig i 7.3 min.
b) 25 % af de studerende kan højst koncentrere sig i 17.9 min.
c) 25 % af de studerende kan højst koncentrere sig i 18.3 min.
d) 25 % af de studerende kan højst koncentrere sig i 30.6 min.
e) 25 % af de studerende kan højst koncentrere sig i 34.6 min.

Ved forelæsningen i Statistisk Dataanalyse 1 d. 7/9-2022 svarede 130 studerende blandt andet på, om de var enige i følgende to udsagn:

- *"Ingen uddannelser må have et karakterkrav (gennemsnit) på over 9 som optagelseskrav"*
- *"Der stilles større krav til unge mennesker idag end før i tiden"*

Resultaterne er opsummeret i følgende tabel, som skal benyttes ved besvarelse af delopgaverne **3.4-3.5**

```
##
##                             Højst 9 i snit Ingen grænse på karakterkrav
##   Ikke større krav til unge             25                           44
##   Større krav til unge                  24                           37
```

3.4 Udfør et test for, om der er uafhængighed mellem de studerendes holdning til de to spørgsmål. Hvad er P-værdien og konklusionen på dette test? Der skal ikke benyttes kontinuitetskorrektion.

a) $P = 0.8539$ så der er ingen sammenhæng ml. svar på de to spørgsmål.
b) $P = 0.6102$ så der er ingen sammenhæng ml. svar på de to spørgsmål.
c) $P = 0.718$ så der er ingen sammenhæng ml. svar på de to spørgsmål.
d) $P = 0.718$ så der er sammenhæng ml. svar på de to spørgsmål.
e) $P = 0.8539$ så der er sammenhæng ml. svar på de to spørgsmål.

3.5 Bestemt et estimat og et simpelt 95 % - konfidensinterval for andelen (p) af studerende som går ind for en øvre karaktergrænse på 9 for optagelseskrav på uddannelser.

a) $\hat{p} = 0.605$ med 95 % - konfidensinterval $0.498 - 0.711$
b) $\hat{p} = 0.377$ med 95 % - konfidensinterval $0.334 - 0.419$
c) $\hat{p} = 0.377$ med 95 % - konfidensinterval $0.294 - 0.460$
d) $\hat{p} = 0.377$ med 95 % - konfidensinterval $0.295 - 0.467$
e) $\hat{p} = 0.605$ med 95 % - konfidensinterval $0.551 - 0.659$

3.6 Vi betragter igen datasættet fra **Opgave 1**. Vi indfører en ny variabel `variety2`, der antager værdien `EGV` for jordbærsorterne `E, G, V` og værdien `ikkkEGV` for de øvrige jordbærsorter (`R1,F,Re,M,P`). Følgende R-kode viser, hvordan man kan konstruere variablen `variety2`

```
data1 <- read.table(file = "nov2022opg1.txt", header = T)
data1$variety2 <- ifelse(data1$variety %in% c("E", "G", "V"), "EGV", "ikkeEGV")
head(data1)

##   variety afstand udbytte variety2
## 1       G       8     5.8      EGV
## 2       V       7     6.3      EGV
## 3      R1       6     4.9  ikkeEGV
## 4       F       5     6.5  ikkeEGV
## 5      Re       4     4.5  ikkeEGV
## 6       M       3     5.2  ikkeEGV
```

Hvad kan vi konkludere på baggrund af følgende test

```
mod1 <- lm(udbytte ~ variety, data = data1)
nymodel <- lm(udbytte ~ variety2, data = data1)
anova(nymodel, mod1)

## Analysis of Variance Table
##
## Model 1: udbytte ~ variety2
## Model 2: udbytte ~ variety
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 63.815
## 2     24 43.915  6      19.9 1.8126  0.139
```

a) Det forventede udbytte er ens for alle de 8 jordbærsorter.

b) Det forventede udbytte er ens for sorterne `E, G, V` men det forventede udbytte er ikke ens for de øvrige sorter `R1,F,Re,M,P`.

c) Det forventede udbytte er 0 for sorterne `R1,F,Re,M,P`.

d) Det forventede udbytte for sorterne `E, G, V` er ikke det samme som det forventede udbytte for de øvrige sorter `R1,F,Re,M,P`.

e) Det forventede udbytte er ens for sorterne `E, G, V` og det forventede udbytte er ens for de øvrige sorter `R1,F,Re,M,P`.