Statistisk Dataanalyse 1
(StatDat1, LMAB10069U)

Eksamen, november 2024

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 50%, 25% og 25% i bedømmelsen. Der er i alt 10 sider i opgavesættet inkl. denne side. Side 10 indeholder kun kildehenvisninger, der ikke skal bruges i besvarelsen af opgaven. Der er 6 delspørgsmål i opgave 1, 3 delspørgsmål i opgave 2 og 6 delspørgsmål i opgave 3.

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. Vær opmærksom på, at der i opgave 3 visse steder er indsat sideskift for at sikre, at alle svarmuligheder står samlet.

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

Opgave 1

Denne opgave vægtes med 50% ved bedømmelsen, og svarene skal begrundes. Data til denne opgave er baseret på data fra McCaffrey m.fl., 2023

I et studie har en gruppe forskere bl.a. indsamlet målinger af køn og fedtmængde for 824 øgler af arten argentinsk teju (Salvator merianae).

Datafilerne `nov24opg1.txt` og `nov24opg1.xlsx` indeholder data til opgaven og kan fx. indlæses med en af følgende R-kommandoer

```
library(readxl)
data1 <- read_excel(path = "nov24opg1.xlsx")
```

eller

```
data1 <- read.table(file = "nov24opg1.txt", header = T)
```

De første linjer i datasættet kan ses her:

```
##   Sex    Fat StorageTime Year
## 1   F  2.780        1579 2012
## 2   F  5.916        1567 2012
## 3   F  8.865           4 2012
## 4   F 95.155         403 2012
## 5   F 48.737          19 2012
```

Hver linje i datasættet er målingerne hørende til samme teju. I datasættet findes variablene `Sex`, der angiver tejuens køn, `Fat`, der angiver tejuens fedtmængde målt i g, `StorageTime`, der angiver opbevaringstid i fryser målt i dage, og `Year`, der er indfangningsåret. I datasættet er tejuer med `StorageTime` $= 0$ blevet frasorteret.

Tejuerne er blevet fanget over en periode på 6 år fra 30/5/2012 til 11/10/2018. Tejuerne er blevet aflivet kort tid efter indfangelsen (0-4 dage), men kan have været opbevaret i en fryser i op til 1847 dage inden de er blevet optøet og obduktionen er foretaget. I delspørgsmål 1.1-1.3 analyser vi, om nedfrysningstiden har indflydelse på den målte fedtmængde, og du skal se bort fra variablen `Sex`.

1.1 Opskriv den lineære regressionsmodel svarende til modellen, der kan fittes med kommandoen

```
mod11 <- lm(log(Fat) ~ log(StorageTime), data=data1)
```

Fit modellen og angiv estimater for samtlige paramete i modellen.

Hvad er fortolkningen af skæringsparameteren i modellen? Hint: Husk, at $\log(1) = 0$.

1.2 Undersøg ved hjælp af modellen `mod11` fra delspørgsmål 1.1, hvilken indflydelse nedfrysningstiden har på den målte fedtmængde. Du skal inddrage et relevant hypotesetest og/eller konfidensinterval i dit svar.

1.3 Forklar, hvordan modellen, der kan fittes med kommandoen

```
mod12 <- lm(log(Fat) ~ log(StorageTime) + factor(Year),
            data=data1)
```

er relateret til modellen `mod11`, der blev fittet i delspørgsmål 1.1.

Benyt modellen `mod12` til at undersøge, om nedfrysningstiden har en indflydelse på den målte fedtmængde, når der også tages højde for indfangningsåret.

I delspørgsmål 1.4-1.6 skal du se bort fra variablen `StorageTime`, og du skal inddrage `Year` som en kategorisk variabel. Vi ønsker at undersøge, hvordan den målte fedtmængde afhænger af indfangningsåret og tejuens køn.

1.4 Forklar, hvorfor dette læggger op til at benytte en tosidet variansanalysemodel med vekselvirkning.

Fit en tosidet variansanalysemodel med vekselvirkning, hvor du bruger `log(Fat)` som responsvariabel og `factor(Year)` og `Sex` som forklarende variable.

Angiv et estimat for forventet log-fedtmængde for en teju af hunkøn i år 2014, og angiv et estimat for forventet log-fedtmængde for en teju af hankøn i år 2018.

1.5 Undersøg, om forskellen i forventet log-fedtmængde mellem tejuer af hankøn og hunkøn afhænger af indfangningsåret.

1.6 Angiv et estimat og et 95% konfidensinterval for forholdet mellem medianfedtmængden for tejuer af hunkøn og medianfedtmængden for tejuer af hankøn i samme år.

Opgave 2

Denne opgave vægtes med 25% ved bedømmelsen, og svarene skal begrundes. Data til denne opgave er baseret på data fra McCaffrey m.fl., 2023

I denne opgave arbejder vi videre med data fra studiet beskrevet i opgave 1. Forskerne har foruden målinger af tejuernes køn og fedtmængde også indsamlet målinger af vægt og længde. Vi fokuserer i denne opgave på målingerne af vægt, længde og fedtmængde for de 130 tejuer af hankøn, der blev fanget i år 2018.

Datafilerne `nov24opg2.txt` og `nov24opg2.xlsx` indeholder data til opgaven og kan fx. indlæses med en af følgende R-kommandoer

```
library(readxl)
data2 <- read_excel(path = "nov24opg2.xlsx")
```

eller

```
data2 <- read.table(file = "nov24opg2.txt", header = T)
```

De første linjer i datasætter kan ses her:

```
##      Fat Weight Length
## 1 54.479   1030   29.4
## 2 21.896   1560   34.5
## 3 10.885    880   29.5
## 4 21.575    990   32.0
## 5  4.804    750   29.1
```

Hver linje i datasættet er målingerne hørende til samme teju. I datasættet findes variablene `Fat`, der angiver tejuens fedtmængde målt i g, `Weight`, der angiver tejuens vægt målt i g, og `Length`, der angiver tejuens længde fra næsespids til kloakåbning målt i cm.

2.1 Betragt modellerne, der kan fittes med kommandoerne:

```
mod21 <- lm(Weight ~ Length, data=data2)
mod22 <- lm(log(Weight) ~  Length, data=data2)
mod23 <- lm(log(Weight) ~ log(Length), data=data2)
```

Forklar grundigt, hvorfor modellen `mod23` er den mest velegnede af de tre modeller til beskrive sammenhængen mellem vægten og længden.

Hint: Du skal udføre modelkontrol og inddrage relevante grafer i din besvarelse.

Forskerne er interesseret i, om det er tilfældet, at sammenhængen mellem tejuerne vægt og længde generelt kan beskrives som

$$\log(\texttt{Weight}) = a - 3\log(\texttt{Length}) \tag{1}$$

2.2 Forklar hvordan modellen `mod23` fra opgave 2.1 kan benyttes til at undersøge om sammenhængen i (1) gælder for populationen af tejuer af hankøn i 2018.

Afgør ved hjælp af et konfidensinterval eller et hypotesetest, om der på baggrund af data er evidens imod at sammenhængen i (1) gælder for populationen af tejuer af hankøn i 2018.

For at fedtmængden kan måles præcist, kræves det, at tejuen aflives og obduceres. Derfor er det af interesse at kunne prædiktere tejuens fedtmængde ved hjælp af tejuens længde og vægt.

2.3 Foreslå en passende lineær regressionsmodel med log(`Fat`) som responsvariabel, der kan benyttes til at prædiktere en hantejus fedtmængde ved hjælp af dens længde og vægt.

Benyt den foreslåede model til at angive et 95% prædiktionsinterval for fedtmængden hos en hanteju med længde 22.6 cm og kropsvægt 386 g.

For at besvare dette spørgsmål fyldestgørende bedes du: Opskrive den statistiske model samt angive R-kode til at fitte modellen og R-kode til at beregne et prædiktionsinterval sammen med angivelsen af prædiktionsintervallet.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 25% i bedømmelsen. For hvert delspørgsmål er der netop et korrekt svar. Der er i alle spørgsmål 5 svarmuligheder A-E. 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 Længden af 12 måneder gamle drenge i Danmark er normalfordelt med middelværdi 75.5 cm og spredning 1.67 cm. Bestem et tal L, så 90% af 12 måneder gamle drenge i Danmark er kortere end L cm.

A. $L \approx 72.8$ cm

B. $L \approx 73.4$ cm

C. $L \approx 77.6$ cm

D. $L \approx 78.2$ cm

E. $L \approx 78.8$ cm

3.2 Der er 25% kort af en vilkårlig kulør (spar, klør, hjerter, ruder) i et almindeligt kortsæt med 52 spillekort. Hvad er sandsynligheden for at trække mindst 4 kort af samme kulør, hvis man trækker 10 kort ved at trække ét kort fra 10 forskellige omhyggeligt blandede almindelige kortsæt?

A. Ca. 1.97%

B. Ca. 7.81%

C. Ca. 22.4%

D. Ca. 77.6%

E. Ca. 92.2%

3.3 Dette spørgsmål ligger i forlængelse af delspørgsmål 1.3, man kan besvares uafhængigt af dette. Følgende kode er blevet kørt i R

```
mod12 <- lm(log(Fat) ~ log(StorageTime) + factor(Year),
            data=data1)
mod13 <- lm(log(Fat) ~ log(StorageTime) + Year,
            data=data1)
anova(mod13,mod12)
```

```
## Analysis of Variance Table
##
## Model 1: log(Fat) ~ log(StorageTime) + Year
## Model 2: log(Fat) ~ log(StorageTime) + factor(Year)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    754 2002.5
## 2    749 1951.3  5    51.248 3.9343 0.00158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvilket af følgende udsagn er korrekt?

A. Der er ved hjælp af et F-test testet for, om en model hvor `Year` indgår lineært beskriver data lige så godt som modellen `mod12`. $p < 0.05$, så vi konkluderer at log(`Fat`) ikke afhænger lineært af `Year`.

B. Der er ved hjælp af et F-test testet for, om en model hvor `Year` indgår lineært beskriver data lige så godt som modellen `mod12`. $p > 0.05$, så vi konkluderer at log(`Fat`) ikke afhænger lineært af `Year`.

C. Der er ved hjælp af et F-test testet for, om en model hvor `Year` indgår lineært beskriver data lige så godt som modellen `mod12`. $p < 0.05$, så vi konkluderer at log(`Fat`) afhænger lineært af `Year`.

D. Der er ved hjælp af et F-test testet for, om en model hvor `Year` indgår lineært beskriver data lige så godt som modellen `mod12`. $p > 0.05$, så vi konkluderer at log(`Fat`) afhænger lineært af `Year`.

E. Der er benyttet et F-test til at sammenligne to modeller, der ikke er nestede, så vi kan ikke konkludere noget på baggrund af kørslen af denne kode.

I et studie (Wild m.fl., 2024) har tyske forskere været interesseret i udbredelsen af to lungeparasitter (Dictyocaulus capreolus og Varestrongylus capreoli) hos rådyr (Capreolus capreolus) i forskellige områder af Tyskland. Delspørgsmål 3.4-3.6 omhandler forskellig aspekter af data fra studiet.

3.4 I 2018 har man i Haßberge skudt 14 voksne rådyr af hankøn, som ikke

havde lungeparasitter. Man har bestemt den gennemsnitlige vægt uden indvolde for de 14 rådyr uden lungeparasitter til at være 15.56 kg og spredningen til at være 2.04 kg. Rådyrenes vægt uden indvolde kan antages at være uafhængige og normalfordelte med samme middelværdi og spredning.

Bestem et 95% konfidensinterval for den forventede vægt uden indvolde for raske, voksne rådyr af hankøn i Haßberge i 2018.

A. $(12.20, 18.92)$ kg

B. $(11.25, 19.99)$ kg

C. $(14.59, 16.53)$ kg

D. $(14.38, 16.74)$ kg

E. $(14.34, 16.78)$ kg

3.5 I 2017 har man i Haßberge skudt 22 voksne rådyr af hunkøn hvoraf 6 rådyr har lungeparasitter.

Du skal i denne opgave forestille dig, at vægten uden indvolde for de 16 rådyr uden lungeparasitter er indlæst i vektoren `udenlp` i R, og at vægten uden indvolde for de 6 rådyr med lungeparasitter er indlæst i vektoren `medlp` i R.

Vægten uden indvolde er målt i kg. Det kan antages, at observationerne af rådyrenes vægt uden indvolde er uafhængige og normaltfordelte med samme spredning. Det kan også antages, at middelværdien af vægt uden indvolde er ens for alle rådyr af hunkøn med lungeparasitter, og at middelværdien af vægt uden indvolde er ens for alle rådyr af hunkøn uden lungeparasitter.

Vi er interesserede i, om vægten uden indvolde hos rådyr af hunkøn i Haßberge i 2017 er påvirket af, om rådyret har lungeparasitter.

Hvad kan du konkludere på baggrund af følgende udskrift?

```
t.test(medlp, udenlp, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  medlp and udenlp
## t = -1.3205, df = 20, p-value = 0.2016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.9892487  0.6717487
## sample estimates:
## mean of x mean of y
```

```
##  14.17000  15.32875
```

A. $p = 0.2016$. Der er derfor signifikant forskel på vægten uden indvolde for rådyr af hunkøn med lungeparasitter og vægten uden indvolde for rådyr af hunkøn uden lungeparasitter.

B. Et 95%-konfidensinterval for forskellen i vægt uden indvolde for rådyr af hunkøn med og uden lungeparasitter er $(14.17, 15.33)$ kg. Der er derfor signifikant forskel på vægten uden indvolde for rådyr af hunkøn med lungeparasitter og vægten uden indvolde for rådyr af hunkøn uden lungeparasitter.

C. Et 95%-konfidensinterval for forskellen i vægt uden indvolde for rådyr med og uden lungeparasitter er $(14.17, 15.33)$ kg. Der er derfor ikke signifikant forskel vægten uden indvolde for rådyr af hunkøn med lungeparasitter og vægten uden indvolde for rådyr af hunkøn uden lungeparasitter.

D. Et 95%-konfidensinterval for forskellen i vægt uden indvolde for rådyr af hunkøn med og uden lungeparasitter er $(-2.989, 0.672)$ kg. Der er derfor signifikant forskel på vægten uden indvolde for rådyr af hunkøn med lungeparasitter og vægten uden indvolde for rådyr af hunkøn uden lungeparasitter.

E. Et 95%-konfidensinterval for forskellen i vægt uden indvolde for rådyr af hunkøn med og uden lungeparasitter er $(-2.989, 0.672)$ kg. Der er derfor ikke signifikant forskel på vægten uden indvolde for rådyr af hunkøn med lungeparasitter og vægten uden indvolde for rådyr af hunkøn uden lungeparasitter.

3.6 I 2017 har man i Ruhpolding observeret følgende forekomster af lungeparasitter hos rådyr af forskellig køn:

```
lungparasite_table
```

```
##             Sex
## Lungparasite Female Male
##        FALSE     14    9
##        TRUE       9    9
```

Hvad kan du konkludere fra følgende udskrift?

```
chisq.test(lungparasite_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lungparasite_table
## X-squared = 0.14359, df = 1, p-value = 0.7047
```

A. 50% af rådyrene har lungeparasitter uanset køn.

B. Forekomsten af lungeparasitter er den samme blandt rådyr af hankøn og rådyr af hunkøn.

C. 70.47% af rådyrene har lungeparasitter uanset køn.

D. Forekomsten af lungeparasitter er ikke den samme blandt rådyr af hankøn og rådyr af hunkøn.

E. Forekomsten af lungeparasitter er ikke uafhængig af rådyrets køn.

Kilder

McCaffrey, K. R., Balaguera-Reina, S. A., Falk, B. G., Gati, E. V., Cole, J. M., & Mazzotti, F. J. (2023). How to estimate body condition in large lizards? Argentine black and white tegu (Salvator merianae, Duméril and Bibron, 1839) as a case study. *PLOS ONE*, *18*(2), 1–19. https://doi.org/https://doi.org/10.1371/journal.pone.0282093

Wild, T., Ehrmantraut, C., Dahl, S.-A., Langer, F., Kiess, E., Simon, K., Meissner, M., & König, A. (2024). Why are our roe deer short of breath? – prevalence and promotive factors of lung parasites in roe deer Capreolus capreolus in south-eastern Germany. *Wildlife Biology*, e01275. https://doi.org/https://doi.org/10.1002/wlb3.01275