Eksamen Matematisk Statistik F 2021

Lucas Bagge 202004972

11. Juni 2021

Opgave 1

(1.a)

Allerførst skal vi se på at de kritiske område skal være $[C, \infty)$ fremfor (0, C).

Det må være at den mindste værdi er den ventetid de påstår eller mindre nemlig C. Der er ikke en mulighed for at det kun vil tage 0 min. Der er ikke noget som siger at vi har en grænse for hvornår vi kan forvente hjælpen. Derfor kan det tag · e uendelig lang tid.

Grunden til vi vil foretage en ensidet test er at vi kun er interresert i at se hvornår vi overstiger den reklameret middeltid, som er 4. Vi vil ikke se på hvornår den er under 4.

(1.b)

Her skal vi tage fordelingsfunktionen (cdf) og sætte den lig med p og løse for x, som vi erstatter med q.

Her bruger jeg et cas værktøj og finder frem til det angivet resultat.

(1.c)

Beregn C:

Her kan vi bruge formlen som er i mSRR bogen side 95.

```
EX = 3/2 * 4
sd = sqrt( (5/4) * 4^2)
```

```
critical_value <- (EX - 4) / sd
critical_value
```

```
## [1] 0.4472136
```

Som er den kritiske værdier.

(1.d)

Ved moment metoden sætter man den teoretiske middelværdi lig med gennemsnittet af observationerne.

$$\bar{x} = \frac{3}{2}\theta$$

$$\hat{\theta} = \frac{2}{3}\bar{x}$$

Udregningen i r.

```
x = c(12, 5, 15, 20, 8)
(2/3) * mean(x)
```

```
## [1] 8
```

Den estimerede stikprøve er 3.33.

(1.e)

(1.f)

```
# Funktion rmeanwait:
# simulation af stikprøvegennemsnit fra ventetidsfordelingen i opgave 1.
#
# Argumenter:
# n = antal simulerede stikprøver,
# size = stikprøvestørrelse,
# theta = fordelingsparameter. Default værdi: theta = 1.
# Eksempel:
# rmeanwait(3, 5, 8)
# simulerer 3 stikprøver af størrelse 5 fra ventetidsfordelingen med theta = 8
# og returnerer en vektor med de 3 stikprøvegennemsnit.
rmeanwait <- function(n, size = 1, theta = 1) {
  replicate(n, theta * mean(-log(1 - sqrt(runif(size)))))
}
```

Opgave 2

(2.a)

LAd (a_1, a_2, a_3, a_4) være antallet:

Modellen bliver således

$$(A_1, A_2, A_3, A_4) \sim multinom(97, (\pi_1, \pi_2, \pi_3, \pi_4))$$

$$\pi_j \geq 0, \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

(2.b)

```
prop.test(43, 97, conf.level = 0.95)$conf.int
```

```
## [1] 0.3436332 0.5475056
## attr(,"conf.level")
## [1] 0.95
```

For at en person ligge i G4 vil sandynligheden spænde mellem 0.344 og 0.548.

(2.c)

Når vi ser på observationer så fra A1 til A2 stiger det med 1, altså med n. Derimod fra A1 til A3 er det ikke med antallet n, men der stiger den med den ukendte parameter θ.

Derfor når vi går op til en anden model vil det passe med at vi skal antage at A1+A2 følger en binomial model med parameter n og rho, mens for A1 + A3 følger erstatter vi rho permeter med theta, da theta fanger at antallet observered stiger på en anderledes måde end fra A1 til A2.

(2.d)

Fra bogen i afsnit 1.3 hvor der forklares om G størrelsen, å snakkes der om at denne kan bruges til enten at teste ens hypotese eller teste om vi kan reducerer modellen.

Af denne begrundelse bruger jeg G testen.

```
a <- c(16,    17,    21,    43)
ex <- c(12.59, 20.41, 24.42, 39.58)

G = 2 * sum(a * log(a / ex))
c(G, 1- pchisq(G, 4 - 1 - 1))
```

```
## [1] 2.2447423 0.3255071
```

Da pværiden er høj siger vi at data ikke strider mod hypotesen, så vi kan beskrive data med den multiplikaive model.

Opgave 3

(3.a)

Her skal vi opstille den multiple regressionsmodel.

Selve den formelle model skrives op således:

$$oktan_i \sim N(\alpha + \beta_{T1} \cdot T_1 + \beta_{T2} \cdot T_2 + \beta_{T3} \cdot T_3 + \beta_{T4} \cdot T_4 + \beta_{T5} \cdot T_5, \sigma^2)$$

Hvor vi vil forklare oktan niveauet ud fra de 5 forklarende variabeler, som er lysmålinger.

(3.b)

Til at starte med indlæse jeg data:

```
oktan <- read.csv("oktan.csv")
head(oktan)
```

```
##         T1       T2       T3       T4       T5 oktan
## 1 0.292567 0.466098 0.044695 0.017254 0.311237 85.30
## 2 0.316834 0.516133 0.056740 0.020912 0.342958 85.25
## 3 0.276645 0.504049 0.055821 0.022470 0.352530 88.45
## 4 0.319126 0.497873 0.052995 0.022399 0.325323 83.40
## 5 0.251579 0.432696 0.045830 0.018535 0.320695 87.90
## 6 0.295713 0.478875 0.048868 0.019856 0.320822 85.50
```

NU skal vi så til at benytte os af backward selektion til at få signifikante test.

Lad os først på på den fulde model

```
model_fuld <- lm(oktan ~ T1 + T2 + T3 + T4 + T5, data = oktan)
```

```
summary(model_fuld)
```

```
##
## Call:
## lm(formula = oktan ~ T1 + T2 + T3 + T4 + T5, data = oktan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70559 -0.16640  0.00369  0.14092  0.69024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   91.225      2.364  38.584  < 2e-16 ***
```

```
## T1           -89.771      6.123 -14.661  < 2e-16 ***
## T2            30.505     10.096   3.021  0.00384 **
## T3           -37.670     40.097  -0.939  0.35168
## T4            28.279     29.113   0.971  0.33570
## T5            22.732     16.322   1.393  0.16941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2732 on 54 degrees of freedom
## Multiple R-squared:  0.9708, Adjusted R-squared:  0.9681
## F-statistic: 359.2 on 5 and 54 DF,  p-value: < 2.2e-16
```

Vi ser at VI skal fjerne T3:T5

```
model_reduc_b <- lm(oktan ~ T1 + T2, data = oktan)
summary(model_reduc_b)
```

```
##
## Call:
## lm(formula = oktan ~ T1 + T2, data = oktan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78638 -0.16527  0.01215  0.17415  0.58995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  92.8562     0.9047  102.64   <2e-16 ***
## T1          -97.6559     2.5076  -38.94   <2e-16 ***
## T2           44.5063     2.6802   16.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.276 on 57 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9675
## F-statistic: 878.3 on 2 and 57 DF,  p-value: < 2.2e-16
```

Det tyder på at kun T1 og T2 er med til at forklare oktan niveauet.

(3.c)

Jeg laver en anova test mellem den fulde model og model_reduc_b

```
anova(model_fuld, model_reduc_b)
```

```
## Analysis of Variance Table
##
```

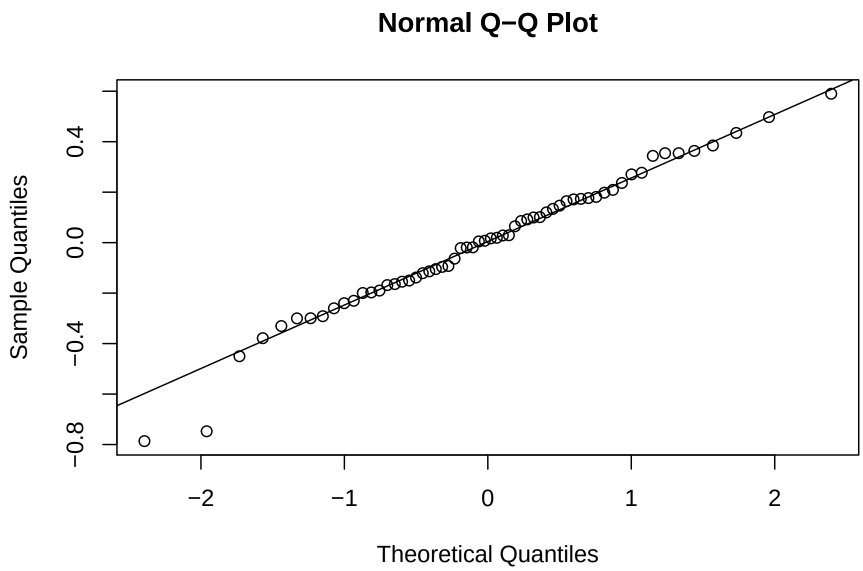
```
## Model 1: oktan ~ T1 + T2 + T3 + T4 + T5
## Model 2: oktan ~ T1 + T2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     54 4.0313
## 2     57 4.3414 -3   -0.3101 1.3846 0.2574
```

Vi ser med den høje p værdi kan vi godt reducer den multiple model.

(3.d)

NU skal vi se på residualer i den reducerede model.

```
SUMmodel_reduc_b <- summary(model_reduc_b)
qqnorm(SUMmodel_reduc_b$residuals)
qqline(SUMmodel_reduc_b$residuals)
```



De ser meget normalt fordelt ud pånår nedre del af plottet hvor vi har to observationer er i øjefaldende.

(3.e)

I den sidste del af opgaven skal vi angive et parameterskøn, som fås ved vores summary funktion.

```
summary(model_reduc_b)
```

```
##
## Call:
## lm(formula = oktan ~ T1 + T2, data = oktan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78638 -0.16527  0.01215  0.17415  0.58995
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  92.8562     0.9047  102.64   <2e-16 ***
## T1          -97.6559     2.5076  -38.94   <2e-16 ***
## T2           44.5063     2.6802   16.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.276 on 57 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9675
## F-statistic: 878.3 on 2 and 57 DF,  p-value: < 2.2e-16
```

Lad os skrive modellen op:

$$oktan \sim 92.8562 - 97.6559 \cdot T1 + 44.5063 \cdot T2$$

Hvor vi har vores koefficienter skøn til - 97.6559 and 44.5063.

Herefter skal vi angive konfidensintervallerne.

```
confint(model_reduc_b)
```

```
##                  2.5 %    97.5 %
## (Intercept)   91.04452  94.66783
## T1          -102.67741 -92.63447
## T2            39.13918  49.87338
```

Her kan vi se i hvilket interval de forklarende variabler er i.

Vi bliver ikke bedt om at angive skønnet over spredning, men det kan findes sådan.

```
SUMmodel_reduc_b$sigma
```

```
## [1] 0.2759809
```