Eksamen i Statistisk Dataanalyse 1, 1. februar 2023

Anders Tolver

Vejledende besvarelse

I denne vejledende besvarelse har jeg inkluderet en del R-kode med tilhørende output. En besvarelse behøver ikke inkludere R-kode og -output medmindre der er bedt eksplicit om det. Jeg har desuden givet korte forklaringer til opgave 3 (multiple choice) hvilket ikke er muligt ved eksamen.

Opgave 1

Vi indlæser først data

```
library(readxl)
# data1 <- read.table(file = "feb2023opg1.txt", header = T)
data1 <- read_excel(path = "feb2023opg1.xlsx")
```

1. Den statistiske model er en ensidet variansanalysemodel og kan opskrives som

$$\texttt{co2}_i = \alpha_{\texttt{day}_i} + e_i,$$

hvor e_i'erne uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

```
mod1 <- lm(co2 ~ factor(day) - 1, data = data1)
summary(mod1)$coef

##                 Estimate Std. Error   t value     Pr(>|t|)
## factor(day)2    1.484848  0.5002325  2.968316 2.501200e-02
## factor(day)4    3.167024  0.5002325  6.331104 7.263105e-04
## factor(day)6    4.730565  0.5002325  9.456733 7.957325e-05
## factor(day)7    6.671267  0.5002325 13.336335 1.099626e-05
## factor(day)9    9.053909  0.5002325 18.099403 1.830721e-06
## factor(day)12 17.914535  0.5002325 35.812420 3.160680e-08
```

Den forventede totale mængde CO_2 efter 2 dage estimeres til 1.48 og efter 12 dage til 17.91.

2.
```r
confint(mod1)
```

```
##                     2.5 %    97.5 %
## factor(day)2   0.2608234  2.708873
## factor(day)4   1.9429989  4.391048
## factor(day)6   3.5065398  5.954589
## factor(day)7   5.4472427  7.895292
## factor(day)9   7.8298842 10.277934
## factor(day)12 16.6905099 19.138559
```

KI for den forventede mængde CO_2 på dag 9 er: $[7.90, 10.28]$

Vi kan finde et 95 % - konfidensinterval for den forvente forskel i mængden af CO_2 på dag 12 og dag 9 ved at benytte `confint()` på den version af modellen, som er fitted med `day = 9` som referencegruppe.

```r
mod1alt <- lm(co2 ~ relevel(factor(day), ref = "9"), data = data1)
confint(mod1alt)
```

```
##                                        2.5 %     97.5 %
## (Intercept)                         7.829884 10.2779337
## relevel(factor(day), ref = "9")2  -9.300093 -5.8380285
## relevel(factor(day), ref = "9")4  -7.617918 -4.1558529
## relevel(factor(day), ref = "9")6  -6.054377 -2.5923121
## relevel(factor(day), ref = "9")7  -4.113674 -0.6516091
## relevel(factor(day), ref = "9")12  7.129593 10.5916581
```

Konfidensintervallet bestemmes til: $[7.13, 10.59]$.

3.
```r
linreg <- lm(log(co2) ~ day, data = data1)
summary(linreg)
```

```
##
## Call:
## lm(formula = log(co2) ~ day, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18674 -0.11240 -0.02879  0.07664  0.23258
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07605    0.09420   0.807    0.438
## day          0.24034    0.01270  18.922 3.69e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.143 on 10 degrees of freedom
## Multiple R-squared:  0.9728,Adjusted R-squared:  0.9701
## F-statistic:   358 on 1 and 10 DF,  p-value: 3.686e-09
```

Estimaterne for modellens parametre bliver: $\hat{\alpha} = 0.076, \hat{\beta} = 0.240, \hat{\sigma} = 0.143$.

Estimatet $\hat{\beta} = 0.240$ beskriver den forventede ændring i $\log(\texttt{co2})$ per dag, hvilket bør omregnes ved at udregne $\exp\hat{\beta} \approx 1.272$. Fortolkningen er at den totale (kumulerede) mængde CO_2 øges med ca. 27.2 % for hver ekstra dag efter forsøgets start.

```
confint(linreg)

##                  2.5 %    97.5 %
## (Intercept) -0.1338325 0.2859378
## day          0.2120412 0.2686430
```

Et 95 % - konfidensinterval for β bliver $[0.212, 0.269]$.

4. Den lineære regressionsmodel kan testes enten mod en kvadratisk regressionsmodel eller mod en ensidet ANOVA (hvor `day` inddrages i modellen som en kategorisk variabel). Begge metoder giver fuldt point.

   Metode A:

   Vi fitter en kvadratisk regressionsmodel

   $$\log(\texttt{co2})_i = \alpha + \beta \cdot \texttt{day}_i + \gamma \cdot \texttt{day}_i^2 + e_i$$

   og tester hypotesen $H_0 : \gamma = 0$. Dette test fremgår direkte af et summary af den kvadratiske model

```
kvadreg <- lm(log(co2) ~ day + I(day^2), data = data1)
summary(kvadreg)$coef

##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -0.217646747 0.147622865 -1.474343 1.744878e-01
## day          0.347375641 0.046829183  7.417931 4.026238e-05
## I(day^2)    -0.007633773 0.003254109 -2.345888 4.359780e-02
```

   Vi findes estimatet $\hat{\gamma} = -0.0076$ og testet for hypotesen giver en T-teststørrelse på -2.346 med en tilhørende P-værdi på 0.0436. På et 5 % - niveau må vi altså forkaste hypotesen $H_0 : \gamma = 0$ som svarer til den lineære regressionsmodel.

   Vær opmærksom på, at vi også kun udføre testet som et F-test, som giver samme resultat $(F = 5.5032, P = 0.0436)$

```
anova(linreg, kvadreg)

## Analysis of Variance Table
##
## Model 1: log(co2) ~ day
## Model 2: log(co2) ~ day + I(day^2)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     10 0.20435
## 2      9 0.12681  1  0.077541 5.5032 0.0436 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Metode B:

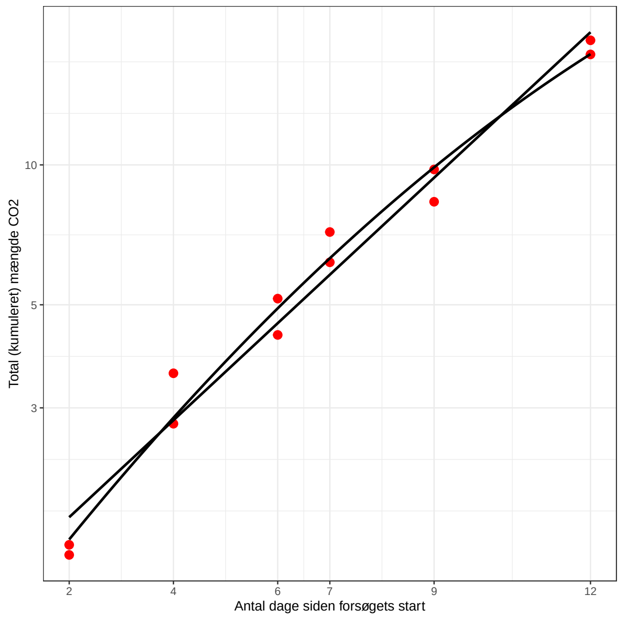
Da vi har flere målinger for hver værdi af variablen `day`, så har vi også mulighed for at lave en ensidet ANOVA, hvor `day` opfattes som en kategoriske variabel. Vi kan derefter teste den lineære regressionsmodel op imod den ensidede variansanalysemodel ved et F-test.

```
ensidet <- lm(log(co2) ~ factor(day), data = data1)
anova(linreg, ensidet)

## Analysis of Variance Table
##
## Model 1: log(co2) ~ day
## Model 2: log(co2) ~ factor(day)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     10 0.20435
## 2      6 0.07520  4   0.12915 2.5762 0.1443
```

Vi kan her ikke forkaste hypotesen om, at der er en lineær sammenhæng ($F = 2.5762, P = 0.1443$).

Løsningerne A og B giver lidt forskellig konklusion. Figuren nedenfor antyder (måske), at man med en krum/kvadrisk funktion (og kun een ekstra parameter) kan opnå en væsentlig bedre approksimation til målepunkterne end med den rette linje. Forbedringen i modelfittet når man går fra en lineær regressionmodel til en ensidet variansanalysemodel er derimod ret begrænset, når man man tager i betragtning, at førstnævnte model benytter fire ekstra parametre til at beskrive middelværdistrukturen.




5. ```
newdata <- data.frame(day = 10)
predict(linreg, newdata, interval = "conf")

fit   lwr   upr
1 2.479474 2.34774 2.611207
```
```

Vi benytter `predict()`-funktionen til at bestemme et estimat og et 95 % - konfidensinterval for den forventede værdi af $\log(CO_2)$ i prøver taget 10 dage efter forsøgets start: 2.479 [KI: $2.348, 2.611$].

Estimat og konfidensinterval bør tilbagetransformeres til en oprindelige skala, hvorved tallene snarere bør fortolkes som medianer: 11.935 [KI: $10.462, 13.615$].

6. Vi benytter igen `predict()`-funktionen. Husk at resultatet skal tilbagetransformeres til oprindelig skala!

```
linreg <- lm(log(co2) ~ day, data = data1)
newdata <- data.frame(day = 10)
exp(predict(linreg, newdata, interval = "p"))

##        fit      lwr      upr
## 1 11.93498 8.455277 16.84673
```

Prædiktionsintervallet aflæses til: $[8.455, 16.847]$

Opgave 2

Vi indlæser først data

```
library(readxl)
# data2 <- read.table(file = "feb2023opg2.txt", header = T)
data2 <- read_excel(path = "feb2023opg2.xlsx")
```

1.
```
mod2 <- lm(log(co2) ~ factor(treat) * factor(day) , data = data2)
summary(mod2)$coef

##                                 Estimate Std. Error      t value     Pr(>|t|)
## (Intercept)                       -1.010 0.08502451 -11.87892814 5.413954e-08
## factor(treat)raj                   1.405 0.12024281  11.68469035 6.502515e-08
## factor(day)4                       0.615 0.12024281   5.11465093 2.555144e-04
## factor(day)6                       0.995 0.12024281   8.27492306 2.655958e-06
## factor(day)7                       1.370 0.12024281  11.39361265 8.598996e-08
## factor(day)9                       1.810 0.12024281  15.05287511 3.733124e-09
## factor(day)12                      2.120 0.12024281  17.63099184 6.045246e-10
## factor(treat)raj:factor(day)4      0.135 0.17004901   0.79388876 4.426766e-01
## factor(treat)raj:factor(day)6      0.160 0.17004901   0.94090520 3.653093e-01
## factor(treat)raj:factor(day)7      0.130 0.17004901   0.76448547 4.593478e-01
## factor(treat)raj:factor(day)9     -0.005 0.17004901  -0.02940329 9.770263e-01
## factor(treat)raj:factor(day)12     0.370 0.17004901   2.17584327 5.026652e-02
```

Estimat for prøve med `treat = raj` taget på dag 9: $-1.010 + 1.405 + 1.810 - 0.005 = 2.200$. Omregnes til oprindelig skala: $\exp(2.200) \approx 9.025$ (fortolkes som median).

Estimat for prøve med `treat = kontrol` taget på dag 9: $-1.010 + 1.810 = 0.800$.

Estimat for ml. `treat = raj` og `treat = kontrol` på dag 9: $2.200 - 0.800 = 1.400$. Kan tilbageregnes til $\exp(1.400) \approx 4.055$. Dette fortolkes som om, at medianværdien af den totale mængde CO_2 er ca. 4.4 gange højere på dag 9 for prøver med rajgræs i forhold til for kontrolprøver.

2. I praksis skal man blot udføre et test for, om der er vekselvirkning mellem de to faktorer `treat` og `day`. Vi tester derfor modellen `mod3` (om at der *ikke* er vekselvirkning) imod modellen `mod2` (*med* vekselvirkning) ved et F-test

```
mod3 <- lm(log(co2) ~ factor(treat) + factor(day) , data = data2)
anova(mod3, mod2)

## Analysis of Variance Table
##
## Model 1: log(co2) ~ factor(treat) + factor(day)
## Model 2: log(co2) ~ factor(treat) * factor(day)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 0.26713
## 2     12 0.17350  5  0.093633 1.2952 0.3287
```

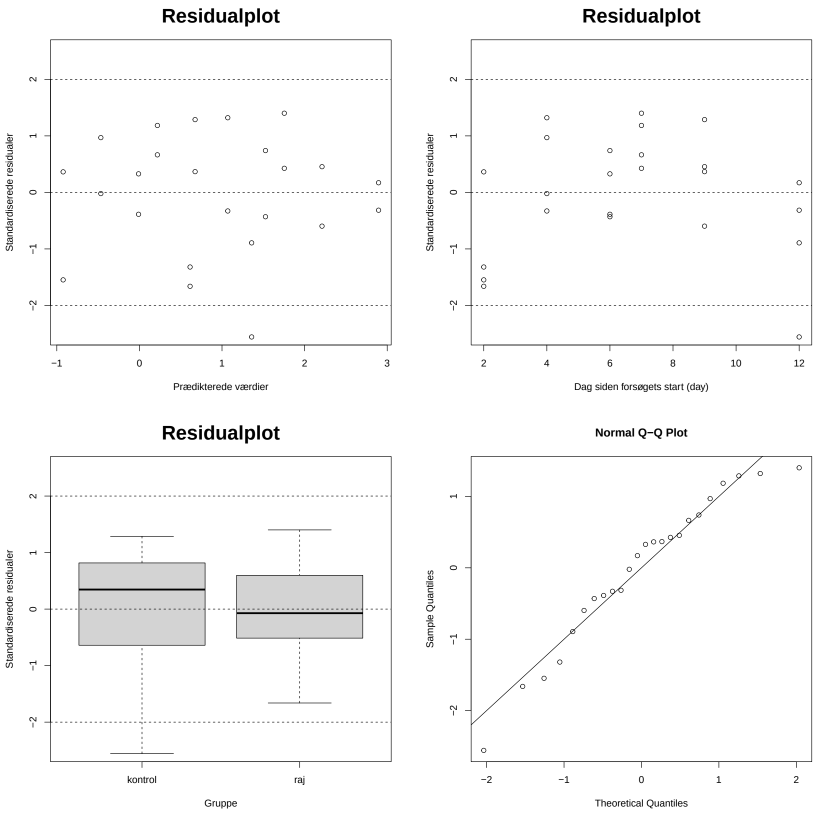
Med et F-teststørrelse på 1.2952 og en tilhørende P-værdi på 03287 kan vi ikke afvise hypotesen om, at der ikke er vekselvirkning.

```
summary(mod3)$coef

##                    Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       -1.075833 0.06769911 -15.891395 1.234582e-11
## factor(treat)raj   1.536667 0.05117572  30.027261 3.597937e-16
## factor(day)4       0.682500 0.08863895   7.699776 6.119829e-07
## factor(day)6       1.075000 0.08863895  12.127852 8.554819e-10
## factor(day)7       1.435000 0.08863895  16.189272 9.171997e-12
## factor(day)9       1.807500 0.08863895  20.391714 2.179165e-13
## factor(day)12      2.305000 0.08863895  26.004371 3.949334e-15
```

Det er ikke et krav at man kvantificerer forskellen mellem mængden af CO_2 i prøver med rajgræs i forhold til kontrolprøver. Man kan dog bemærke at medianværdien estimeres til at være $\exp 1.536667 \approx 4.649$ gange højere for prøver med rajgræs (uanset hvilken dag man måler den totale kumulerede mængde CO_2).

3. Det mest oplagte er at lave modelkontrol, hvor man ser på residualplot og QQ-plot over de standardiserede residualer.



Det mest bemærkelsesværdige er, at middelværdien af residualerne ikke lader til at være nul uanset værdien af `day`. Residualerne hørende til målinger taget på dag 4-9 er overvejende positive. Det tyder på, at modellens middelværdistruktur ikke er korrekt. I lyset af resultatet fra delopgave **1.4** bør man også have en formodning om, at en lineær funktion ikke giver en helt optimal beskrivelse af sammenhængen mellem middelværdien af responsen (log(`co2`)) og antallet af dage (`day`) siden forsøgets start.

En alternativ løsning består i at teste den blandede model fx. imod den tosidede variansanalysemodel uden vekselvirkning (denne model kunne ikke forkastes jf. svar på delopgave **2.2**). Vælges denne løsning bør man i princippet først lave modelkontrol for model `mod3` (eller model `mod2`). Den blandede model forkastes på niveau 5 % ($F = 4.38, P = 0.013$)!

```
anova(mod4, mod3)

## Analysis of Variance Table
##
## Model 1: log(co2) ~ treat + day
## Model 2: log(co2) ~ factor(treat) + factor(day)
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     21 0.54264
## 2     17 0.26713  4   0.27551 4.3833 0.01286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Opgave 3

3.1 Korrekt svar E.

```r
pnorm(21, mean = 19.664, sd = 0.929)  - pnorm(18, mean = 19.664, sd = 0.929)

## [1] 0.8881652
```

3.2 Korrekt svar E.

```r
qnorm(0.05, mean = 19.664, sd = 0.929)

## [1] 18.13593
```

3.3 Korrekt svar C.

```r
summary(lm(after - before ~ 1, data = my_data))

##
## Call:
## lm(formula = after - before ~ 1, data = my_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.987 -6.737 -3.237  3.112 18.413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.313      3.272  -1.929    0.095 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.254 on 7 degrees of freedom
```

3.4 Korrekt svar B.

```r
1 - pbinom(24, size = 100, prob = 1/6)

## [1] 0.02170338
```

3.5 Korrekt svar C.

3.6 Korrekt svar B.

```r
my_table <- matrix(2, 2, data = c(15, 10, 60, 20))
my_table
```

```
##      [,1] [,2]
## [1,]   15   60
## [2,]   10   20
```

```r
chisq.test(my_table, correct = FALSE)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  my_table
## X-squared = 2.1, df = 1, p-value = 0.1473
```