# Reeksamen i Statistisk Dataanalyse 1, 2. februar 2022

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 (her fra filen `nov2021opg1.txt`)

```
library(readxl)
# data1 <- read.table(file = "feb2022opg1.txt", header = T)
data1 <- read_excel(path = "feb2022opg1.xlsx")
```

1. Modellen fittes i R

```
m1 <- lm(tid ~ type, data = data1)
summary(m1)
```

```
##
## Call:
## lm(formula = tid ~ type, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44688 -0.22469 -0.06563  0.15438  0.63562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.20688    0.07246  71.862  < 2e-16 ***
## typesmirna   -0.28250    0.10247  -2.757   0.0084 **
## typesymphony -0.77125    0.10247  -7.527  1.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2898 on 45 degrees of freedom
## Multiple R-squared:  0.5631,Adjusted R-squared:  0.5437
## F-statistic:    29 on 2 and 45 DF,  p-value: 8.096e-09
```

og vi aflæser residualspredningen til $\hat{\sigma} = 0.2898$. Referencegruppen i R-output er netop `type = napoli`, hvorfor vi aflæser estimatet for den forventede spiringstid for `type = symphony` til at $5.207 - 0.771 = 4.436$ dage.

2. Vi ønsker at teste hypotesen om at den forventede værdi af spiringstiden er den samme for alle tre arter. Testet udføres som et F-test enten med `drop1()` eller ved at fitte en nulmodel svarende til hypotesen om, at der er samme forventede værdi for alle arter.

```
m2 <- lm(tid ~ 1, data = data1)
anova(m2, m1)

## Analysis of Variance Table
##
## Model 1: tid ~ 1
## Model 2: tid ~ type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     47 8.6520
## 2     45 3.7799  2    4.8721 29.001 8.096e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Konklusion:** Med en F-teststørrelse $F = 29.001$ og en tilhørende P-værdi $< 0.0001$ konkluderer vi, at der ikke kan antages at være samme forventede spiringstid af alle tre sorter.

3. Den letteste løsning består i at reparametrisere modellen, så man fx. benytter `type = smirna` som referencegruppe. Dette kan fx. gøres således

```
data1$typeny <- relevel(factor(data1$type), ref = "smirna")
m1alt <- lm(tid ~ typeny, data = data1)
summary(m1alt)$coef

##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     4.924375 0.07245622 67.963450 5.309908e-47
## typenynapoli    0.282500 0.10246858  2.756943 8.398528e-03
## typenysymphony -0.488750 0.10246858 -4.769755 1.971957e-05

confint(m1alt)

##                      2.5 %     97.5 %
## (Intercept)     4.77844067  5.0703093
## typenynapoli    0.07611769  0.4888823
## typenysymphony -0.69513231 -0.2823677
```

**Konklusion:** Den forventede spiringstid for frø af typen `symphony` estimeres til at være 0.489 dage kortere end for frø af typen `smirne`. Forskellen er statistisk signfikant ($T = -4.770, P < 0.0001$).

Dette kan også ses ved at bemærke, at et 95 % - konfidensinterval for forskellen i den forventede spiringstid er (-0.695)-(-0.282) dage, og at dette interval ikke indeholder 0.

## Opgave 2

Vi indlæser først data (her fra filen `feb2022opg2.txt`)

```
# data2 <- read.table(file = "feb2022opg2.txt", header = T)
data2 <- read_excel(path = "feb2022opg2.xlsx")
```

1. Den statistiske model er en lineær regressionsmodel

$$\texttt{length}_i = \alpha + \beta \cdot \texttt{omkreds}_i + e_i,$$

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

Et udpluk af `summary()` af modellen ses her

```
mod1 <- lm(length ~ omkreds, data = data2)
summary(mod1)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.9602306  0.6525882 -1.471419 1.460029e-01
## omkreds      1.1904812  0.1007707 11.813764 7.789958e-18
```

Estimaterne for modellens parametre bliver

$$\begin{aligned} \text{(Intercept)} \quad \hat{\alpha} &= -0.960 \\ \text{(Hældning)} \quad \hat{\beta} &= 1.190 \\ \text{(Residual spredning)} \quad \hat{\sigma} &= 1.200 \end{aligned}$$

2. Estimatet kan let beregnes i hånden, men for at bestemme et prædiktionsinterval benyttes `predict()`-funktionen.

```
newdata <- data.frame(omkreds = 3.14 * 2)
predict(mod1, newdata, interval = "predict")
```

```
##        fit      lwr      upr
## 1 6.515991 4.101626 8.930357
```

Den forventede længde af en gulerod med diameter 2 cm estimeres til 6.52 cm med et 95 % - prædiktionsinterval som er givet ved 4.10 - 8.93 cm. Det vil derfor ikke være usædvaneligt at finde en gulerod med diameter 2 cm og en længde på 8 cm.

3. Med udgangspunkt i den lineære regressionsmodel `mod1`, så kan tommelfingerregelen formuleres som hypotesen, $H_0 : \alpha = 0, \beta = 3/\pi \approx 3/3.14 = 0.955$.

**Der gives fuldt point** hvis man med udgangspunkt i `mod1` tester hver af de to hypoteser $\alpha = 0$ og $\beta = 3/3.14$.

For hypotesen $\alpha = 0$ kan vi direkte fra `summary(mod1)` aflæse en t-teststørrelse på $-1.471$ med tilhørende P-værdi 0.146. Vi kan derfor ikke afvise, at skæringen (`Intercept`) kan være 0.

For at teste hypotesen $\beta = 3/3.14$ konstrueres t-teststørrelsen

$$T = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{1.1904812 - 3/3.14}{0.1007707} \approx 2.332694.$$

Under hypotesen om at $\beta = 3/3.14$, så vil t-teststørrelsen være $t$-fordelt med 65 frihedsgrader, hvorfor vi finder en p-værdi

```
Tobs <- (1.1904812-3/3.14)/0.1007707
Tobs

## [1] 2.332694

Pvalue <- 2 * (1 - pt(Tobs, df = 67 - 2))
Pvalue

## [1] 0.02277106
```

Benyttes et signfikansniveau på 5 %, så må vi konkludere, at det *ikke* er rimeligt at antage, at længden er 3 gange så stor som diameteren.

**Der gives også fuldt point** hvis man starter med at teste hypotesen $\alpha = 0$ som beskrevet ovenfor efterfulgt af et test for $\beta = 3/3.14$ i den reducerede model. Det betyder blot, at t-teststørrelse og P-værdi for det andet test skal beregnes med udgangspunkt i følgende R-output

```
mod2c <- lm(length ~ omkreds - 1, data = data2)
summary(mod2c)$coef

##         Estimate Std. Error  t value     Pr(>|t|)
## omkreds 1.045995 0.02283673 45.80319 9.526988e-52
```

T-teststørrelse bliver her

```
Tobs2 <- (1.045995-3/3.14)/0.02283673
Tobs2

## [1] 3.96646

Pvalue2 <- 2 * (1 - pt(Tobs, df = 66 - 1))
Pvalue2

## [1] 0.02277106
```

og den tilhørende P-værdi er 0.023. Benyttes et signfikansniveau på 5 %, så må vi altså også med denne metode konkludere, at det *ikke* er rimeligt at antage, at længden er 3 gange så stor som diameteren.

**En elegant løsning som også giver fuldt point** består i at få R til fitte en model som repræsenterer begge restriktioner $\alpha = 0$ og $\beta = 3/3.14$ og så lave et $F$-test for den samlede

hypotese. En variant af denne løsning har været illustreret i et eksempel undervisningen vha. en multipel lineær regression til beskrivelse af sammenhængen mellem volumen, omkreds og højde af kirsebærtræer. For begge de to metoder illustreret nedenfor fås en $F$-teststørrelse på 9.0878 med tilhørende P-værdi på 0.000331. Vi må derfor konkludere, at datasættet ikke understøtter tommelfingerregelen (hypotesen) om, at den forventede længde af en gulerød er lig med 3 gange omkredsen.

**Metode A:**

```
mod1a <- lm(length - 3/3.14 * omkreds ~ omkreds, data = data2)
mod2a <- lm(length - 3/3.14 * omkreds ~ 0, data = data2)
anova(mod2a, mod1a)

## Analysis of Variance Table
##
## Model 1: length - 3/3.14 * omkreds ~ 0
## Model 2: length - 3/3.14 * omkreds ~ omkreds
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     67 119.770
## 2     65  93.598  2    26.172 9.0878 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Metode B:**

```
mod2b <- lm(length ~ offset(3/3.14 * omkreds) - 1, data = data2)
anova(mod2b, mod1)

## Analysis of Variance Table
##
## Model 1: length ~ offset(3/3.14 * omkreds) - 1
## Model 2: length ~ omkreds
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     67 119.770
## 2     65  93.598  2    26.172 9.0878 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Den foreslåede model er en kvadratisk regressionsmodel

$$\texttt{length}_i = \alpha + \beta \cdot \texttt{omkreds}_i + \gamma \cdot \texttt{omkreds}_i^2 + e_i,$$

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

```
mod2 <- lm(length ~ omkreds + I(omkreds^2), data = data2)
summary(mod2)$coef

##                Estimate Std. Error   t value  Pr(>|t|)
## (Intercept)  0.32185329 2.26691259 0.1419787 0.8875428
## omkreds      0.80177300 0.66564772 1.2045005 0.2328317
## I(omkreds^2) 0.02791819 0.04725219 0.5908338 0.5567127
```

Estimaterne for modellens parametre bliver

$$\begin{aligned} \text{(Intercept)} \quad \hat{\alpha} &= 0.322 \\ \text{(Hældning/lineær led)} \quad \hat{\beta} &= 0.802 \\ \text{(Kvadratisk led)} \quad \hat{\gamma} &= 0.028 \\ \text{(Residual spredning)} \quad \hat{\sigma} &= 1.206 \end{aligned}$$

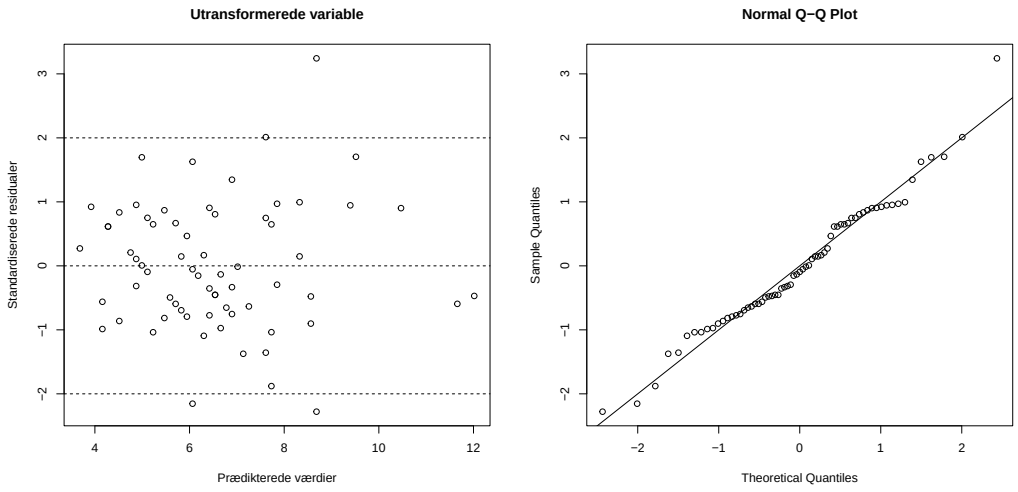
Vi ser af output, at estimatet hørende til det kvadratiske led bliver $\hat{\gamma} = 0.028$. Et test af hypotesen $H_0 : \gamma = 0$ svarer til at undersøge, om data er godt beskrevet ved en lineær sammenhæng. Testet kan udføres som et $T$-test $(T = 0.591, P = 0.557)$ eller som et F-test

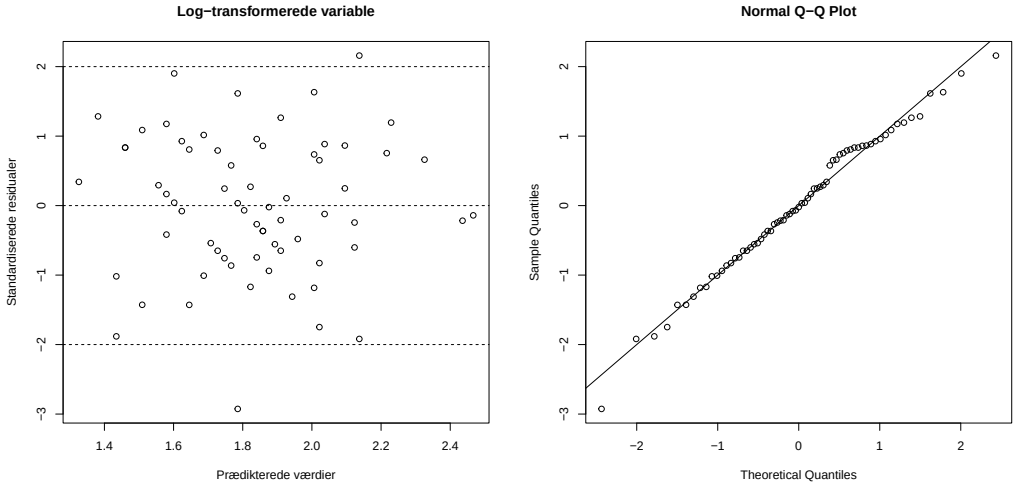
```
anova(mod2, mod1)

## Analysis of Variance Table
##
## Model 1: length ~ omkreds + I(omkreds^2)
## Model 2: length ~ omkreds
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     64 93.090
## 2     65 93.598 -1  -0.50776 0.3491 0.5567
```

Vi kan derfor ikke afvise hypotesen om, at den forventede længde af en gulerod kan beskrives ved en linæer funktion af omkredsen.

5. Vi laver et scatterplot, et residualplot og et QQ-plot for både `mod1` og `mod2`

```
mod3 <- lm(log(length) ~ log(omkreds), data = data2)
plot(predict(mod1), rstandard(mod1), xlab = "Prædikterede værdier"
     , ylab = "Standardiserede residualer", main = "Utransformerede variable")
abline(h = c(-2, 0, 2), lty = 2)
qqnorm(rstandard(mod1))
abline(0,1)
plot(predict(mod3), rstandard(mod3), xlab = "Prædikterede værdier"
     , ylab = "Standardiserede residualer", main = "Log-transformerede variable")
abline(h = c(-2, 0, 2), lty = 2)
qqnorm(rstandard(mod3))
abline(0,1)
```






På residualplottet kigger vi efter, om residulerne ligger omkring 0, og om der er nogenlunde samme variation for både små og store prædikterede værdier. For begge modeller konkluderer vi, at

- residualerne ligger rimelig symmetrisk omkring 0 for både små og store prædikterede værdier
- residualerne har nogenlunde samme variation for både store og små prædikterede værdier

Punkterne på QQ-plottet ligger pænt omkring en ret linje med skæring 0 og hældning 1, så residualerne virker til at være pænt normalfordelte. Dette gælder for begge de foreslåede modeller.

Det virker som om, at modelantagelserne med god tilnærmelse er opfyldt for begge de to modeller. Der er derfor ikke grund til at foretrække den ene model frem for den anden. Men vi skal naturligvis huske at foretolke resultaterne korrekt i lyset af, hvilken model der vælges.

6. En mulighed er at lave en blandet model

$$\texttt{length}_i = \alpha(\texttt{variety}_i) + \beta \cdot \texttt{omkreds}_i + e_i,$$

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

```
mod4 <- lm(length ~ variety + omkreds, data = data2)
summary(mod4)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)   -0.1410516  0.8445838 -0.1670073 8.678992e-01
## varietyorange -0.4816657  0.3428633 -1.4048330 1.649826e-01
## varietyroed   -0.6296165  0.4924079 -1.2786483 2.057113e-01
## omkreds        1.1146433  0.1135093  9.8198401 2.533941e-14
```

Estimaterne for modellens parametre bliver

$$\begin{aligned}
\text{(Intercept/skæring for referencegruppe)} \quad \hat{\alpha}(\texttt{gul}) &= -0.1411 \\
\text{(Forskel ml. skæringer)} \quad \hat{\alpha}(\texttt{orange}) - \hat{\alpha}(\texttt{gul}) &= -0.4817 \\
\text{(Forskel ml. skæringer)} \quad \hat{\alpha}(\texttt{roed}) - \hat{\alpha}(\texttt{gul}) &= -0.6296 \\
\text{(Hældning)} \quad \hat{\beta} &= 1.1146 \\
\text{(Residual spredning)} \quad \hat{\sigma} &= 1.206
\end{aligned}$$

7. **Løsning 1:** Af R-output ovenfor kan vi også aflæse resultatet af t-test til undersøgelse af, om skæringen er forskellig fra skæringen i referencegruppen for hver af de to sorter `orange` ($T = -1.405, P = 0.165$) og `roed` ($T = -1.279, P = 0.206$). Dette tyder på, at der ikke er forskel på hældningerne for de tre sorter i forsøget. (Man kunne være endnu mere grundig og reparametrisere modellen med `relevel()` for at få t-test for sammenligning mellem sorterne `orange` og `roed`).

   **Løsning 2:** Man kan også blot udføre et F-test for hypotesen om, at der *ikke* er forskel på skæringsparametrene hørende til de tre sorter.

```
anova(mod1, mod4)

## Analysis of Variance Table
##
## Model 1: length ~ omkreds
## Model 2: length ~ variety + omkreds
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     65 93.598
## 2     63 90.170  2    3.4282 1.1976 0.3087
```

   Vi kan ikke afvise hypotesen ($F = 1.1976, P = 0.3087$), så der er ikke noget som tyder på, at sorten har betydning for sammenhængen mellem længde og omkreds af en gulerod.

# Opgave 3

3.1 Korrekt svar E.

```
pnorm(8, mean = 6.31, sd = 1.47) - pnorm(3, mean = 6.31, sd = 1.47)

## [1] 0.8626874
```

3.2 Korrekt svar D. Vi finder at 70 % fraktilen er ca. 7.08, hvorfor 30 % af gulerødderne har en længde på over 7.08 cm.

```
qnorm(0.70, mean = 6.31, sd = 1.47)

## [1] 7.080869
```

3.3 Korrekt svar C. Bemærk at vi har tre metoder til at udføre testet i R, men at de alle giver en P-værdi på under 5 %, hvorfor hypotesen bør forkastes på et 5 % niveau.

```
binom.test(20, 56, p = 0.5)

##
##  Exact binomial test
##
## data:  20 and 56
## number of successes = 20, number of trials = 56, p-value = 0.04405
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2335548 0.4964069
## sample estimates:
## probability of success
##              0.3571429

prop.test(20, 56, correct = FALSE, p = 0.5)

##
##  1-sample proportions test without continuity correction
##
## data:  20 out of 56, null probability 0.5
## X-squared = 4.5714, df = 1, p-value = 0.03251
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2445657 0.4880611
## sample estimates:
##         p
## 0.3571429

prop.test(20, 56, correct = TRUE, p = 0.5)

##
##  1-sample proportions test with continuity correction
##
## data:  20 out of 56, null probability 0.5
## X-squared = 4.0179, df = 1, p-value = 0.04502
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2368611 0.4970247
## sample estimates:
##         p
## 0.3571429
```

3.4 Korrekt svar B. Vi finder en P-værdi på 0.0047 så hypotesen forkastes. Da der er flere positive svar (`Ja`) fra de studerende i 2021, så kan man konkludere, at hypotesen forkastes fordi de studerende har mere positive forventninger i 2021. Det er selvfølgelig her væsentligt, at der (tilfældigvis) viser sig at være 177 studerende som svarer på sprøgsmålet i både 2020 og 2021 (ellers skulle antal først omregnes til procent).

```r
my_table <- matrix(2, 3, data = c(115, 86, 47, 62, 15, 29))
my_table

##      [,1] [,2] [,3]
## [1,]  115   47   15
## [2,]   86   62   29

chisq.test(my_table)

##
##  Pearson's Chi-squared test
##
## data:  my_table
## X-squared = 10.703, df = 2, p-value = 0.004741
```

3.5 Korrekt svar B.

```r
data3 <- read.table(file = "feb2022opg3.txt", header = T)
summary(lm(y - x ~ treat, data = data3))

##
## Call:
## lm(formula = y - x ~ treat, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7522 -0.5292  0.1456  0.4761  1.7971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8959     0.2123   4.220 0.000232 ***
## treatB        0.4253     0.3002   1.417 0.167592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8222 on 28 degrees of freedom
## Multiple R-squared:  0.06689,Adjusted R-squared:  0.03356
## F-statistic: 2.007 on 1 and 28 DF,  p-value: 0.1676
```

Her laves en ensidet ANOVA, hvor output vil indeholde estimat og P-værdi for test af, om den forventede ændring er ens i hver af de to grupper. (Hvis man fx. benytter metoden i D, så vil output indeholde estimat for ændringen i hver gruppe, men der vil ikke være en P-værdi for sammenligning af størrelsen af ændringerne på tværs af grupperne).

3.6 Korrekt svar C. Da *P*-værdier er over 5 % Vi kan ikke afvise hypotesen om, at der *ikke* er vekselvirkning. Dette fortolkes som, at forskellen mellem effekten af de to behandlinger er den samme for de to diagnosegrupper.