Eksamen i Statistisk Dataanalyse 1, 9. november 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

```
library(readxl)
# data1 <- read.table(file = "nov2022opg1.txt", header = T)
data1 <- read_excel(path = "nov2022opg1.xlsx")
```

1. Den statistiske model er en ensidet ANOVA

$$\texttt{udbytte}_i = \alpha_{\texttt{variety}_i} + e_i,$$

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

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

##                 Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)  6.55000e+00  0.6763490  9.684350e+00 9.161996e-10
## varietyF    -9.25000e-01  0.9565019 -9.670655e-01 3.431561e-01
## varietyG    -3.80739e-15  0.9565019 -3.980536e-15 1.000000e+00
## varietyM    -9.50000e-01  0.9565019 -9.932024e-01 3.305201e-01
## varietyP    -5.00000e-02  0.9565019 -5.227381e-02 9.587432e-01
## varietyR1   -3.05000e+00  0.9565019 -3.188703e+00 3.947127e-03
## varietyRe   -9.75000e-01  0.9565019 -1.019339e+00 3.182080e-01
## varietyV    -2.50000e-01  0.9565019 -2.613691e-01 7.960378e-01
```

Estimatet for det forventede udbytte på områder med sorten `E` aflæses ud for `(Intercept)`: 6.55.

Estimat for det forventede udbytte på områder med sorten `R1`: $6.55 - 3.05 = 3.50$.

2. For at bestemme konfidensintervallerne er det hensigtsmæssigt at fitte modellen uden referencegruppe. Man kan da benytte R-kommandoen `confint()` til at udtrække 95 %-konfidensintervaller

```
mod1alt <- lm(udbytte ~ variety - 1, data = data1)
summary(mod1alt)$coef

##            Estimate Std. Error  t value     Pr(>|t|)
## varietyE      6.550   0.676349 9.684350 9.161996e-10
## varietyF      5.625   0.676349 8.316713 1.576574e-08
## varietyG      6.550   0.676349 9.684350 9.161996e-10
## varietyM      5.600   0.676349 8.279750 1.708417e-08
## varietyP      6.500   0.676349 9.610424 1.061954e-09
## varietyR1     3.500   0.676349 5.174843 2.670835e-05
## varietyRe     5.575   0.676349 8.242786 1.851623e-08
## varietyV      6.300   0.676349 9.314718 1.930105e-09

confint(mod1alt)

##               2.5 %   97.5 %
## varietyE  5.154084 7.945916
## varietyF  4.229084 7.020916
## varietyG  5.154084 7.945916
## varietyM  4.204084 6.995916
## varietyP  5.104084 7.895916
## varietyR1 2.104084 4.895916
## varietyRe 4.179084 6.970916
## varietyV  4.904084 7.695916
```

KI for områder med sorten `E`: $[5.154 - 7.946]$

KI for områder med sorten `R1`: $[2.104 - 4.896]$

Vi kan finde et 95 % - konfidensinterval for den forvente forskel mellem de to sorter `R1` og `E` ved at benytte `confint()` på den version af modellen, som er fitted uden referencegruppe.

```
confint(mod1)

##                 2.5 %     97.5 %
## (Intercept)  5.154084  7.9459156
## varietyF    -2.899123  1.0491228
## varietyG    -1.974123  1.9741228
## varietyM    -2.924123  1.0241228
## varietyP    -2.024123  1.9241228
## varietyR1   -5.024123 -1.0758772
## varietyRe   -2.949123  0.9991228
## varietyV    -2.224123  1.7241228
```

Konfidensintervallet bestemmes til: $[-5.024 - (-1.076)]$.

3. Vi ønsker at test hypotesen

$$H_0 : \alpha_{\mathsf{E}} = \alpha_{\mathsf{F}} = \ldots = \alpha_{\mathsf{V}}$$

om at det forventede udbytte en ens for de otte jordbærsorter. Testet udføres som et F-test

```
mod2 <- lm(udbytte ~ 1, data = data1)
anova(mod2, mod1)

## Analysis of Variance Table
##
## Model 1: udbytte ~ 1
## Model 2: udbytte ~ variety
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     31 73.000
## 2     24 43.915  7    29.085 2.2708 0.06342 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: Med en F-teststørrelse $F = 2.271$ og en tilhørende P-værdi på 0.063 kan vi (på et 5 % niveau) ikke afvise hypotesen om, at der er samme forventede udbytte for alle sorter.

4. Modellen skal være en lineær regressionsmodel

$$\text{udbytte}_i = \alpha + \beta \cdot \text{afstand}_i + e_i,$$

hvor $e_1, \ldots, e_{32}$ er uafhængige $\sim N(0, \sigma^2)$.

```
linreg <- lm(udbytte ~ afstand, data = data1)
summary(linreg)$coef

##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 3.7178571 0.43784996 8.491167 1.783441e-09
## afstand     0.4571429 0.08670721 5.272259 1.079811e-05
```

Parameterestimaterne bliver: $\hat{\alpha} = 3.718, \hat{\beta} = 0.457$.

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

Metode A:

Vi fitter en kvadratisk regressionsmodel

$$\text{udbytte}_i = \alpha + \beta \cdot \texttt{afstand}_i + \gamma \cdot \texttt{afstand}_i^2 + e_i$$

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

```
kvadreg <- lm(udbytte ~ afstand + I(afstand^2), data = data1)
summary(kvadreg)$coef

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   2.0794643 0.70800867  2.937060 0.0064302635
## afstand       1.4401786 0.36097254  3.989718 0.0004113887
## I(afstand^2) -0.1092262 0.03915297 -2.789729 0.0092253886
```

Vi findes estimatet $\hat{\gamma} = -0.109$ og testet for hypotesen giver en T-teststørrelse på -2.790 med en tilhørende P-værdi på 0.009. 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 ($F = 7.7826, P = 0.009$)

```
anova(linreg, kvadreg)

## Analysis of Variance Table
##
## Model 1: udbytte ~ afstand
## Model 2: udbytte ~ afstand + I(afstand^2)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 37.891
## 2     29 29.874  1    8.0172 7.7826 0.009225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Metode B:

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

```
ensidet <- lm(udbytte ~ factor(afstand), data = data1)
anova(linreg, ensidet)

## Analysis of Variance Table
##
## Model 1: udbytte ~ afstand
## Model 2: udbytte ~ factor(afstand)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     30 37.891
## 2     24 21.980  6    15.911 2.8956 0.02873 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi forkaster hypotesen om, at der er en lineær sammenhæng ($F = 2.896, P = 0.029$).

6. **Bemærk:** Da der kun er en observation for hver kombination af variablene `variety` og `afstand`, så kan vi ikke fitte modellen med vekselvirkning. Dette er grunden til, at det anbefales at benytte en additiv model for tosidet ANOVA.

   Vi fitter derfor den additive model for tosidet ANOVA med R-koden

```
tosidet <- lm(udbytte ~ variety + factor(afstand), data = data1)
summary(tosidet)$coef

##                      Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)         3.8722619  0.8244530  4.6967651 2.076264e-04
```

```
## varietyF          -1.0194070  0.6181704 -1.6490712 1.174878e-01
## varietyG          -0.4932775  0.6582537 -0.7493730 4.638767e-01
## varietyM          -0.4373292  0.8005569 -0.5462812 5.919707e-01
## varietyP           0.1361836  0.6893556  0.1975520 8.457384e-01
## varietyR1         -2.0439510  0.8942024 -2.2857812 3.537606e-02
## varietyRe         -1.3318523  0.6515310 -2.0441886 5.674383e-02
## varietyV          -0.3263624  0.5692371 -0.5733330 5.739261e-01
## factor(afstand)2   2.1721435  0.8085199  2.6865677 1.560631e-02
## factor(afstand)3   1.9735127  0.7239411  2.7260683 1.437046e-02
## factor(afstand)4   3.0232941  0.8005569  3.7764886 1.505781e-03
## factor(afstand)5   3.3548264  0.7750091  4.3287572 4.558790e-04
## factor(afstand)6   3.3433782  0.5754557  5.8099667 2.090243e-05
## factor(afstand)7   3.5880976  0.7693050  4.6640769 2.225508e-04
## factor(afstand)8   3.2826481  0.8327869  3.9417624 1.051960e-03
```

Estimat for området med sorten E i afstanden 1 m: 3.872 (svarer til intercept).

Estimat for forskellen mellem udbyttet på områder i afstand 4 m og 2 m fra hækken: $3.023 - 2.172 = 0.851$.

Ønskes yderligere forklaring, så kan man eventuelt bemærke, at da modellen er additiv, så er forskellen i udbyttet for afstand 4 m og 2 m uafhængigt af sorten der dyrkes på området. Derfor kan man lave et beregningseksempel, hvor man regner på udbyttet på områder som beplantes med sorten E.

Estimat for sort E i afstand 2 m: $3.872 + 2.172$

Estimat for sort E i afstand 4 m: $3.872 + 3.023$

Forskel i udbytte i afstand 4 m og 2 m:

$$(3.872 + 3.023) - (3.872 + 2.172) = 3.023 - 2.172 = 0.851$$

.

7. Modellen der fittes er en *blandet model*

$$\texttt{udbytte}_i = \alpha_{\texttt{variety}_i} + \beta \cdot \frac{1}{\texttt{afstand}_i} + e_i,$$

hvor $e_1, \ldots, e_{32}$ er uafhængige $\sim N(0, \sigma^2)$. Det er også ok, hvis man opskriver modellen ved at skrive $\texttt{x}_i$ i stedet for $\frac{1}{\texttt{afstand}_i}$.
Vi beregner det ønskede prædiktionsinterval ud fra modellen

```
data1$x <- 1/data1$afstand
mod2 <- lm(udbytte ~ x + variety, data1)
newdata <- data.frame(x = 1/4.2, variety = "P")
predict(mod2, newdata, interval = "predict")

##        fit      lwr      upr
## 1 7.137794 5.455567 8.820021
```

Estimat med tilhørende 95 % - prædiktionsinterval bliver: 7.138 $[5.456 - 8.820]$.

Opgave 2

Vi indlæser først data

```
library(readxl)
# data2 <- read.table(file = "nov2022opg2.txt", header = T)
data2 <- read_excel(path = "nov2022opg2.xlsx")
```

1. Vi benytter formlen for et 95 %-konfidensinterval for middelværdien for en enkelt stikprøve.

   Estimat for middelværdi: $\hat{\mu} = 58.96$

   Estimat for spredning (ikke påkrævet): $\hat{\sigma}^2 = 13.87$

   Beregning af 95 %-konfidensinterval

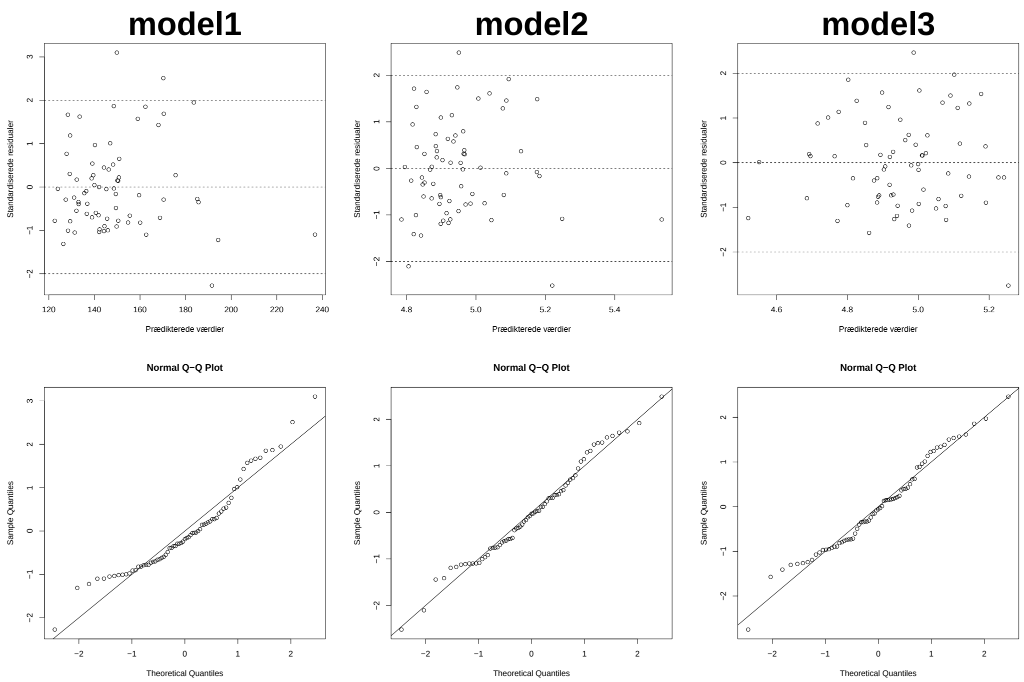
   ```
   yhat <- mean(data2$age)
   yhat + c(-1, 1) * qt(0.975, df = 71 - 1) * sd(data2$age)/sqrt(71)

   ## [1] 55.67549 62.24191
   ```

2. Modellen svarende til `model3` er en *multipel lineær regressionsmodel* med to forklarende variable

   $$\log(\texttt{duration}_i) = \alpha + \beta_1 \cdot \log(\texttt{volume}_i) + \beta_2 \cdot \texttt{age}_i + e_i,$$

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

   Vi laver residualplot og QQ-plot for de tre modeller



   Vi konstaterer at:

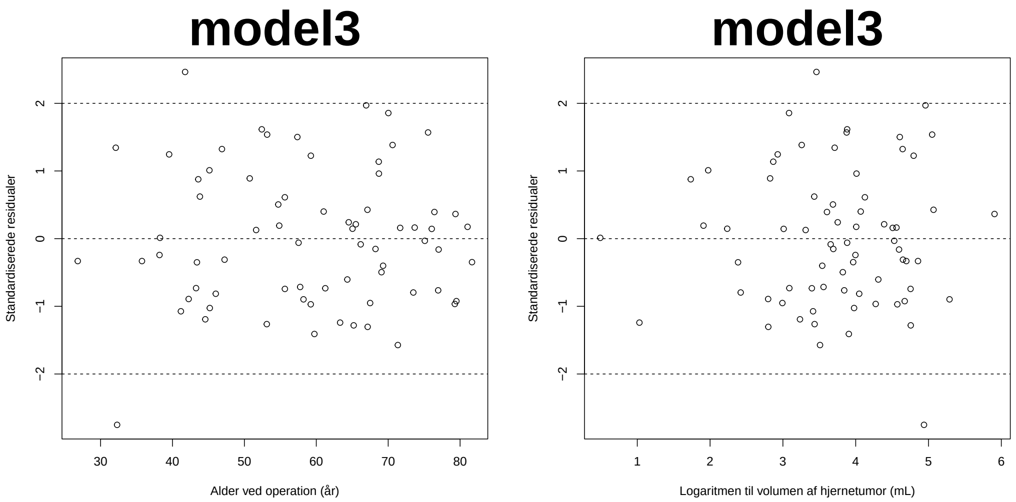
model1: Residualerne ligger ikke symmetrisk omkring 0. Der er fx. flere positive residualer i intervallet [1.5-2.5] end i intervallet [-2.5-(-1.5)]. Desuden har QQ-plottet for de standardiserede residualer en systematisk afvigelse (S-form) omkring den rette linje.

model2: Her ses ligesom for `model1` en tendens til at residualerne ikke ligger symmetrisk omkring 0. QQ-plottet for de standardiserede residualer ligger dog pænt omkring den rette linje.

model3: For denne model er residualerne i højere grad symmetrisk fordelt omkring 0 end de var tilfældet for de to øvrige modeller. QQ-plottet for de standardiserede residualer ligger pænt omkring den rette linje.

Den overordnede konklusion er, at blandt de tre modeller så opfylder `model3` bedst antagelserne for en multipel lineær regressionsmodel.

Det er ikke et krav at man også plotter de standardiseret residualer op imod de to forklarende variable i modellen, men for fuldstændighedens skylds vises disse figurer her.



Ingen af figurerne giver anledning til at stille spørgsmålstegn ved antagelserne bag `model3`.

3. Ud fra `summary(model3)` kan vi aflæse et T-test for hypotesen $H_0: \beta_2 = 0$ $(T = -1.803, P = 0.076)$. Benyttes et signifikansniveau på 5 %, så kan vi mao. ikke afvise hypotesen om, at der ikke er sammenhæng mellem operationstid og alder.

```
summary(model3)$coef

##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  4.648321643 0.181664761 25.587360 1.274583e-36
## log(volume)  0.152174060 0.034962645  4.352476 4.635662e-05
## age         -0.004503921 0.002497603 -1.803297 7.577177e-02
```

4. Vi fitter modellen i R

```
model4 <- lm(log(duration) ~ log(volume), data = data2)
summary(model4)$coef

##               Estimate Std. Error   t value     Pr(>|t|)
```

```
## (Intercept) 4.4247810 0.13494817 32.788745 8.266484e-44
## log(volume) 0.1409212 0.03495811  4.031144 1.409927e-04

confint(model4)

##                  2.5 %    97.5 %
## (Intercept) 4.15556674 4.6939952
## log(volume) 0.07118165 0.2106607
```

Vi finder dernæst estimatet for den forventede værdi af log(`duration`) for operation af hjernetumorer på 50 mL og 100 mL. Ved at tage eksponentialfunktionen kan disse tilbagetransformeres til den oprindelige skala, blot er det mere korrekt at fortolke de tilbagetransformerede værdier som medianer.

```
newdata <- data.frame(volume = c(50, 100))
log_est <- predict(model4, newdata)
log_est

##        1        2
## 4.976068 5.073747

exp(log_est)

##        1        2
## 144.9035 159.7719
```

Medianværdi for operationstider ved volumen på 50 mL: 144.9 min

Medianværdi operationstider ved volumen på 100 mL: 159.7 min

Forskellen mellem median operationstiderne er 159.7-144.9 = 14.8 min.

Det forventes ikke at man kommenterer på, at vi mere generelt kan kvantificere den relative forøgelse af median operationstiden ved en fordobling af tumorstørrelsen ved tallet

```
exp(0.1409*log(2))

## [1] 1.102593
```

Dvs.: en fordobling af tumorstørrelsen vil forøge median for fordelingen af operationstiden med ca. 10.3 %.

Opgave 3

3.1 Korrekt svar A.

3.2 Korrekt svar A.

```
2*(1-pt(1.732, df = 37 - 2))

## [1] 0.09207971
```

3.3 Korrekt svar E.

```
qnorm(0.25, mean = 42.9, sd = 12.3)

## [1] 34.60378
```

3.4 Korrekt svar C.

```
my_tab <- matrix(c(25, 24, 44, 37), 2, 2)
my_tab

##      [,1] [,2]
## [1,]   25   44
## [2,]   24   37

chisq.test(my_tab, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  my_tab
## X-squared = 0.13354, df = 1, p-value = 0.7148
```

3.5 Korrekt svar C.

```
phat <- (25 + 24) / 130
phat

## [1] 0.3769231

se <- sqrt(phat * (1 - phat) / 130)
phat + c(-1, 1) * 1.96 * se

## [1] 0.2936161 0.4602301
```

3.6 Korrekt svar E.