Eksamen i Statistisk Dataanalyse 1, 10. november 2021

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`)

```
data1 <- read.table(file = "nov2021opg1.txt", header = T)
```

1. Den statistiske model er en lineær regressionsmodel

$$\texttt{tid}_i = \alpha + \beta \cdot \texttt{puls}_i + e_i,$$

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

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

```
mod1 <- lm(tid ~ puls, data = data1)
summary(mod1)$coef

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 819.18562 36.1136836  22.68352 2.437125e-36
## puls         -2.54362  0.2178002 -11.67869 6.944555e-19
```

   Estimaterne for modellens parametre bliver

$$\begin{aligned} \text{(Intercept)} \quad \hat{\alpha} &= 819.186 \\ \text{(Hældning)} \quad \hat{\beta} &= -2.544 \\ \text{(Residual spredning)} \quad \hat{\sigma} &= 17.14 \end{aligned}$$

2. Hypotesen $H_0 : \beta = 0$ udtrykker, at der *ikke* er sammenhæng mellem puls og omgangstid. Fra `summary()` af modellen aflæses T-teststørrelsen til -11.68, og den tilhørende P-værdi er praktisk talt 0. Hypotesen forkastes og vi konkluderer (ikke overraskende), at der er sammenhæng mellem puls og omgangstid.

Estimatet for hældningen $\hat{\beta} = -2.544$ angiver den forventede ændring i *omgangstiden*, når pulsen øges med 1 slag. Et 95 %-konfidensinterval for hældningen bliver [-2.977, -2.110] jf. nedenstående output

```
confint(mod1)

##                  2.5 %   97.5 %
## (Intercept) 747.303127 891.0681
## puls         -2.977141  -2.1101
```

Da hver omgang er 1.460 km, så kan estimat og konfidensinterval omregnes til den forventede ændring i *kilometertiden* ved en forøgelse af pulsen på 1 slag, ved at dividere med 1.46

$$\text{Estimat:} \quad -1.742 \quad [95\% - KI : (-2.039) - (-1.445)].$$

Da konfidensintervallet indeholder værdien -2, så er datasættet i overensstemmelse med en påstand om, at den forventede kilometertid falder med 2 sekunder, når pulsen øges med 1 slag.

En **alternativ løsning** består i, at lave et formelt test af hypotesen $H_0 : \beta = -2 \cdot 1.46$. T-teststørrelsen beregnes til

$$T = \frac{\hat{\beta} - (-2 \cdot 1.46)}{SE(\hat{\beta})} = \frac{-2.54362 - (-2 \cdot 1.46)}{0.2178} = 1.728.$$

Under H_0 er teststørrelse t-fordelt med 79 frihedsgrader, hvorfor p-værdien beregnes til

```
2 * (1 - pt(1.728, df = 79))

## [1] 0.08789542
```

Benyttes et 5 % signifikansniveau konkluderes, at vi ikke kan afvise hypotesen.

3. Vi benytter `predict()`-funktionen til beregning af et 95 %-prædiktionsinterval

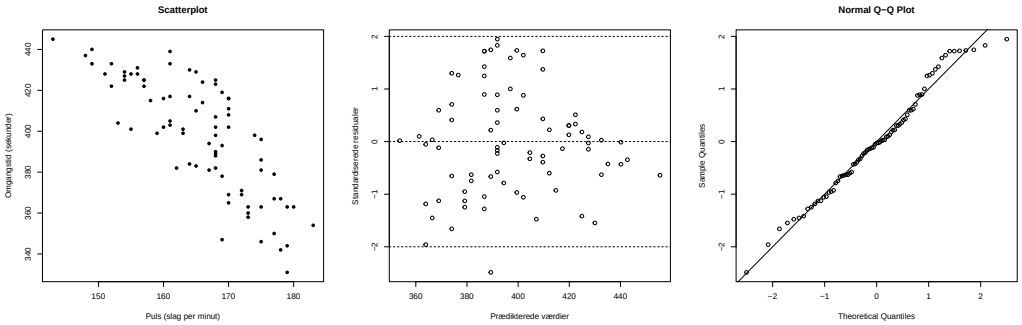
```
newData <- data.frame(puls = 160)
predict(mod1, newData, interval = "p")

##        fit      lwr      upr
## 1 412.2064 377.7876 446.6252
```

Estimat + 95 % - PI: 412.2 [377.8 – 446.6] sekunder.

4. Vi laver et scatterplot, et residualplot og et QQ-plot

```
plot(data1$puls, data1$tid, xlab = "Puls (slag per minut)"
     , ylab = "Omgangstid (sekunder)", main = "Scatterplot", pch = 16)
plot(predict(mod1), rstandard(mod1), xlab = "Prædikterede værdier"
     , ylab = "Standardiserede residualer")
abline(h = c(-2, 0, 2), lty = 2)
qqnorm(rstandard(mod1))
abline(0,1)
```



På scatterplottet kigger vi efter, om det er rimeligt at beskrive sammenhængen mellem puls og omgangstid ved en lineær funktion. Det ser ikke helt urimeligt ud, med det er svært at afgøre om punkterne snarere ligger omkring en kurve, som krummer lidt.

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

- der er en tendens til at residualerne er overvejende negative for meget små eller meget store prædikterede værdier
- residualerne har nogenlunde samme variation for både store og små prædikterede værdier, men det er faktisk svært at vurdere, fordi residualerne ikke ligger pænt omkring 0

Punkterne på QQ-plottet ligger pænt omkring en ret linje med skæring 0 og hældning 1, så residualerne virker til at være pænt normalfordelte.

Analysen af figurer giver anledning til en smule bekymring omkring modellens egnethed til at beskrive variationen i data. En fuldstændig besvarelse af spørgsmålet bør i det mindste stille lidt spørgsmålstegn ved, om middelværdistrukturen er god beskrevet ved en lineær funktion.

Det er endnu flottere, hvis man forslår et alternativ/forbedring til `mod2`. Man fx. teste om sammenhængen er væsentlig bedre beskrevet ved en kvadratisk funktion

$$\alpha + \beta \cdot \texttt{puls} + \gamma \cdot \texttt{puls}^2$$

end ved en ret linje.

```
modkvad <- lm(tid ~ puls + I(puls^2), data = data1)
summary(modkvad)$coef

##                  Estimate   Std. Error    t value   Pr(>|t|)
## (Intercept) -540.04313286 561.67596923 -0.9614852 0.33927947
## puls          14.06142007   6.85161360  2.0522786 0.04349725
## I(puls^2)     -0.05056641   0.02085491 -2.4246767 0.01763509
```

Vi ser af output, at estimatet hørende til det kvadratiske led bliver $\hat{\gamma} = -0.051$. 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 $(T = -2.425, P = 0.0176)$ eller som et F-test

```r
anova(mod1, modkvad)

## Analysis of Variance Table
##
## Model 1: tid ~ puls
## Model 2: tid ~ puls + I(puls^2)
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     79 23219
## 2     78 21591  1    1627.4 5.8791 0.01764 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypotesen (om en lineær sammenhæng) forkastes rent faktisk på et 5 % -signifikansniveau til fordel for en kvadratisk model, der kan *fange* den (lidt) krumme struktur, som antydes på scatterplottet.

5. Modellen `mod2` er en blandet model der kan skrives på formen

$$\text{tid}_i = \alpha(\text{tid\_paa\_dagen}_i) + \beta \cdot \text{puls}_i + e_i,$$

hvor e_i'erne uafhængige og normalfordelte $\sim N(0, \sigma^2)$. Vi fitter modellen i R og ser på et `summary()`

```r
mod2 <- lm(tid ~ puls + tid_paa_dag, data = data1)
summary(mod2)$coef

##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)          858.911617 36.3656046  23.618791 2.773914e-37
## puls                  -2.740837  0.2151495 -12.739223 9.556617e-21
## tid_paa_dagformiddag -12.185905  3.8127695  -3.196077 2.012024e-03
```

Modellen udtrykker, at forskellen mellem den forventede omgangstid på løbeture formiddag og eftermiddag er uafhængig af, hvilken puls der løbes ved. Forskellen estimeres til $\hat{\alpha}(\text{formiddag}) - \hat{\alpha}(\text{eftermiddag}) = -12.19$ sekunder. Forskellen er signfikant forskellig fra 0 ($T = -3.20, P = 0.0020$), så vi konkluderer at løbetiderne er hurtigere om formiddagen. En **alternativ** løsning består i at angive et 95 %-konfidensinterval for forskellen $[(-19.78) - (-4.60)]$, og bemærke at dette interval *ikke* indeholder værdien 0.

```r
confint(mod2)

##                             2.5 %     97.5 %
## (Intercept)          786.513260 931.309975
## puls                  -3.169166  -2.312507
## tid_paa_dagformiddag -19.776546  -4.595264
```

6. Vi kan benytte `predict()`-funktion til beregning af estimat (og 95 %-konfidensinterval)

```r
newData <- data.frame(puls = 160, tid_paa_dag = "formiddag")
predict(mod2, newData, interval = "c")

##        fit      lwr     upr
## 1 408.1918 403.2547 413.129
```

Estimatet for den forventede løbetid på en omgang der løbes om formiddagen med puls 160 bliver 408.2 sekunder (med et 95 %-KI: $[403.3 - 413.1]$).

Opgave 2

Vi indlæser først data (her fra filen `nov2021opg2.txt`)

```
data2 <- read.table(file = "nov2021opg2.txt", header = T)
```

1. Modellen fittes i R

```
m1 <- lm(week12 - week0 ~ treat, data = data2)
summary(m1)
```

```
##
## Call:
## lm(formula = week12 - week0 ~ treat, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5719  -2.0472  -0.1719   1.6022  18.0269
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.4281     0.6937  -0.617   0.5396
## treatintervention   2.4012     1.0361   2.318   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.924 on 56 degrees of freedom
## Multiple R-squared:  0.08752,Adjusted R-squared:  0.07123
## F-statistic: 5.371 on 1 and 56 DF,  p-value: 0.02415
```

og vi aflæser residualspredningen til $\hat{\sigma} = 3.924$. Referencegruppen i R-output er netop standardbehandlingen, hvorfor vi aflæser den estimere ændring i håndgrebsstyrken til at være -0.428 dvs. et fald på 428 gram.

2. Vi tester hypotesen om, at der er samme ændring i håndgrebsstyrken for de to behandlingsgrupper. Dette kan gøre enten ved et F-test

```
m2 <- lm(week12 - week0 ~ 1, data = data2)
anova(m2, m1)
```

```
## Analysis of Variance Table
##
## Model 1: week12 - week0 ~ 1
## Model 2: week12 - week0 ~ treat
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     57 945.02
```

```
## 2     56 862.32  1    82.709 5.3712 0.02415 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

eller ved et T-test, som kan aflæses i under `summary()` for modellen `m1` ovenfor. De to metoder giver samme P-værdi (0.0242). Benyttes et signfikansniveau på 5%, så kan vi konkludere, at der er forskel på *ændringen i håndgrebsstyrken* for de to behandlingsgrupper. Vi ser fra `summary()`, at der er ændringen er størst i interventionsgruppen (estimeret forskel: 2.401), så interventionen lader til at have en (positiv) effekt.

3. Vi ønsker at beskrive, hvordan ændringen i håndgrebsstyrken (kontinuert respons) afhænger af de to kategoriske variable `treat` og `diagnose`. Det er derfor naturligt at benytte en tosidet variansanalysemodel. Vi ønsker at tillade, at behandlingseffekten kan være forskellig for de to diagnosegrupper, så vi bør benytte en model med vekselvirkning. (Rent teknisk kan vi kun benytte modellen med vekselvirkning, hvis der er gentagelser/flere målinger hørende til de forskellige kombinationer af `treat` og `diagnose`, hvilket også er tilfældet i dette datasæt).

   Vi tester hypotesen om at der er vekselvirkning mellem `treat` og `diagnose`

```
m3 <- lm(week12 - week0 ~ treat * diagnose, data = data2)
m4 <- lm(week12 - week0 ~ treat + diagnose, data = data2)
anova(m4, m3)

## Analysis of Variance Table
##
## Model 1: week12 - week0 ~ treat + diagnose
## Model 2: week12 - week0 ~ treat * diagnose
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     55 842.52
## 2     54 841.92  1    0.5928 0.038 0.8461
```

   Vi kan ikke afvise hypotesen om, at der ikke er vekselvirkning ($F = 0.038, P = 0.846$). Derfor kan effekten af behandlingen antages at være den samme for begge diagnosegrupper (hvilket udtrykkes ved den additive struktur i modellen `m4`).

4. Baseret på modellen `m4` estimeres den forventede ændring i håndgrebsstyrken i interventionsgruppen til

$$\begin{aligned} \text{(For)}\texttt{diagnose = bugspytkirtel} &: -1.016 + 2.311 = 1.295 \\ \text{(For)}\texttt{diagnose = lunge} &: -1.016 + 2.311 + 1.175 = 2.470 \end{aligned}$$

5. En **simpel** løsning er blot at reparametrisere modellen `m1` uden intercept

```
m1alt <- lm(week12 - week0 ~ treat - 1, data = data2)
confint(m1alt)

##                       2.5 %   97.5 %
## treatcontrol      -1.8177490 0.961499
## treatintervention  0.4314275 3.514726
```

Af R-outputtet kan vi aflæse 95 %-konfidensintervaller (KI) for ændringen for `control`gruppen til $[(-1.818) - 0.961]$ og for `intervention`sgruppen til $[0.431, 3.515]$. Vi kigger på om KI indeholder værdien 0 (svarende til ingen ændring). Der er i data *ikke* evidens for at konkludere, at der sker en ændring i håndgrebsstyrken for kontrolgruppen, mens der sker en signifikant ændring (her: stigning) i interventionsgruppen. Man kan også vælge at lave T-test for om størrelsen af ændringerne er lig med 0, hvilket giver resultatet $T = -0.617, P = 0.540$ (`control`) og $T = 2.564, P = 0.013$ (`intervention`) jf. nedenstående output

```
summary(m1alt)$coef

##                    Estimate Std. Error    t value   Pr(>|t|)
## treatcontrol      -0.428125  0.6936880 -0.6171723 0.53962251
## treatintervention  1.973077  0.7695777  2.5638436 0.01306317
```

En **mindre elegant** (men fuldt korrekt) metode er at opdele datasættet i to og lave et parret t-test for, om der sker en ændring over tid i hver af de to behandlingsgruppe. Testet kan udføres i R som beskrevet nedenfor

```
library(tidyverse)
data_control <- filter(data2, treat == "control")
t.test(data_control$week0, data_control$week12, paired = TRUE)

##
##  Paired t-test
##
## data:  data_control$week0 and data_control$week12
## t = 0.75285, df = 31, p-value = 0.4572
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7316963  1.5879463
## sample estimates:
## mean of the differences
##                0.428125

data_intervention <- filter(data2, treat == "intervention")
t.test(data_intervention$week0, data_intervention$week12, paired = TRUE)

##
##  Paired t-test
##
## data:  data_intervention$week0 and data_intervention$week12
## t = -2.1617, df = 25, p-value = 0.04041
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.85289989 -0.09325396
## sample estimates:
## mean of the differences
##               -1.973077
```

Vi konkluderer, at ændringen i interventionsgruppen er statistisk signifikant på et 5 % niveau ($T = -2.162, P = 0.040$).

En **mere kompliceret** (men også korrekt) løsning består i at udtrække relavante estimater og test fra den tosidede ANOVA `m4`. *Bemærk:* Da man fra `summary(m4)` i opgaveformuleringen ser, at forskellen mellem ændringen i de to diagnosegrupper *ikke* er statistisk signifikant ($T = 1.137, P = 0.261$), så har man faktisk et stærkt argument for at se bort fra `diagnose`, sådan som vi gjorde i den simple løsning).

Fra følgende R-output kan man aflæse estimerede ændringer for begge diagnosegrupper

```
m4A <- lm(week12 - week0 ~ treat + diagnose - 1, data = data2)
summary(m4A)$coef

##                    Estimate Std. Error   t value  Pr(>|t|)
## treatcontrol      -1.015524  0.8635080 -1.176045 0.2446423
## treatintervention  1.295308  0.9718903  1.332772 0.1881017
## diagnoselunge      1.174799  1.0333322  1.136903 0.2605078

confint(m4A)

##                         2.5 %    97.5 %
## treatcontrol      -2.7460332 0.7149843
## treatintervention -0.6524033 3.2430199
## diagnoselunge     -0.8960452 3.2456430

data2$diagnoseny <- relevel(factor(data2$diagnose), ref = "lunge")
m4B <- lm(week12 - week0 ~ treat + diagnoseny - 1, data = data2)
summary(m4B)$coef

##                             Estimate Std. Error    t value   Pr(>|t|)
## treatcontrol               0.1592745  0.8635080  0.1844505 0.85433851
## treatintervention          2.4701072  0.8833445  2.7963125 0.00710696
## diagnosenybugspytkirtel -1.1747989  1.0333322 -1.1369034 0.26050781

confint(m4B)

##                                2.5 %    97.5 %
## treatcontrol              -1.5712343 1.8897832
## treatintervention          0.6998453 4.2403692
## diagnosenybugspytkirtel -3.2456430 0.8960452
```

Vi aflæser følgende estimerede ændringer (95 %-KI og resultater af T-test også angivet)

$$
\begin{aligned}
\text{bugspyt, control} &: -1.016[(-2.746) - 0.715] \quad \text{(Test for ændring lig 0: P=0.245)} \\
\text{bugspyt, intervention} &: 1.295[(-0.652) - 3.243] \quad \text{(Test for ændring lig 0: P=0.188)} \\
\text{lunge, control} &: 0.159[(-1.571) - 1.890] \quad \text{(Test for ændring lig 0: P=0.854)} \\
\text{lunge, intervention} &: 2.470[(0.700 - 4.240] \quad \text{(Test for ændring lig 0: P=0.007)}
\end{aligned}
$$

Denne analyse antyder, at det primært er for patienter med lungekræft, at interventionen giver en signifikant forøgelse af håndgrebsstyrken. Dette kan forklares ved at der generelt er en større tilbagegangen i håndgrebsstyrken for patienter med kræft i bugspytkirtlen, og at interventionen samlet set blot ophæver tilbagegangen.

Opgave 3

3.1 Korrekt svar B.

```
pnorm(9, mean = 8.07, sd = 2.10) - pnorm(7, mean = 8.07, sd = 2.10)

## [1] 0.3658729
```

3.2 Korrekt svar B.

```
qnorm(0.90, mean = 8.07, sd = 2.10)

## [1] 10.76126
```

3.3 Korrekt svar C. Benyt formlen for et 95 %-konfidensinterval for en stikprøve.

```
my_KI <- 5.8858 + c(-1, 1) * qt(0.975, df = 179-4) * 0.1131
my_KI ### på log-skala

## [1] 5.662584 6.109016

exp(my_KI) ### på oprindelig skala

## [1] 287.8917 449.8956
```

3.4 Korrekt svar E. Analysen er lavet på log-skala, så tallet 2.3400 angiver den estimerede forskel på middelværdien for logaritmen til gæt på antal punkter for JØ studerende og BB studerende. Blandt de mulige svarmuligheder er det korrekt at tage eksponentialfunktionen til 0.234 og fortolke resultatet som en *relativ* ændring for medianen. Tallet $\exp(0.2340) = 1.263$ angiver at medianen estimeres til at være ca. 26.3 % højere for JØ-studerendes gæt end BB-studerendes gæt.

3.5 Korrekt svar C.

```
pbinom(2, 6, 1/5)

## [1] 0.90112
```

3.6 Korrekt svar D. De 179 observationer (studerende) er blevet inddelt efter svaret på to forskellige spørgsmål. Testet skal fortolkes som et test for uafhængighed mellem svarene på de to spørgsmål. Vi forkaster hypotesen ($P = 0.008$) og konkluderer, at der *er* en sammenhæng mellem om de studerende gætter på forelæseren kan lide at hækle og at plukke kantareller.