Reeksamen i Statistisk Dataanalyse 1, 27. januar 2020

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

1. Hvis $\texttt{stof}_i$ betegner tørstofmængden for i-te kar, og $\texttt{nitrat}_i$ det tilhørende nitratniveau, så kan modellen opskrives som

$$\texttt{stof}_i = \alpha_{\texttt{nitrat}_i} + e_i,$$

hvor $e_1, \ldots, e_{16}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

Vi indlæser data og fitter modellen i R

```
salat <- read.table(file = "salat.txt", header = T)
model1 <- lm(stof ~ factor(nitrat), data = salat)
summary(model1)
```

```
##
## Call:
## lm(formula = stof ~ factor(nitrat), data = salat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7675 -1.7831 -0.1088  1.8213  5.9025
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24.465      1.824  13.410 1.39e-08 ***
## factor(nitrat)1    7.635      2.580   2.959 0.011937 *
## factor(nitrat)2   10.143      2.580   3.931 0.001995 **
## factor(nitrat)3   13.073      2.580   5.067 0.000277 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.649 on 12 degrees of freedom
## Multiple R-squared:  0.7021,Adjusted R-squared:  0.6276
## F-statistic: 9.427 on 3 and 12 DF,  p-value: 0.001769
```

Den forventede tørstofmængde svarende til en tilsat nitratmængde på 0.5 (her brugt som referencegruppe) estimeres til 24.47. For nitratmængden 3.0 estimeres den forventede tørstofmængde til $24.47 + 13.07 = 37.54$. Residualspredningen estimeres til $\hat{\sigma} = 3.649$.

2. Vi ønsker at teste hypotesen

$$\alpha_{0.5} = \alpha_{1.0} = \alpha_{2.0} = \alpha_{3.0}$$

om at den forventede tørstofmængde er ens for hvert af de fire niveauer af tilsat nitrat. Testet kan udføres som et F-test

```
nulmodel <- lm(stof ~ 1, data = salat)
anova(nulmodel, model1)

## Analysis of Variance Table
##
## Model 1: stof ~ 1
## Model 2: stof ~ factor(nitrat)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     15 536.25
## 2     12 159.76  3    376.49 9.4265 0.001769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: F-teststørrelsen beregnes til $F = 9.43$ med en tilhørende p-værdi på $p = 0.0018$ (beregnet ud fra en F-fordeling med (3, 12)-frihedsgrader). Der er mao. forskel på den forventede mængde tørstof for de fire tilsatte nitratmængder.

Testet kan også udføres med kommandoen

```
drop1(model1, test = "F")
```

3. `model2` er en lineær regressionsmodel, hvor den forventede mængde tørstof ved nitratmængden x er givet på formen

$$\alpha + \beta \cdot x.$$

```
model2 <- lm(stof ~ nitrat, data = salat)
summary(model2)

##
## Call:
```

```
## lm(formula = stof ~ nitrat, data = salat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7488 -2.2958 -0.1171  2.5724  5.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.683      1.959  12.598    5e-09 ***
## nitrat         4.612      1.038   4.443 0.000557 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.987 on 14 degrees of freedom
## Multiple R-squared:  0.585,Adjusted R-squared:  0.5554
## F-statistic: 19.74 on 1 and 14 DF,  p-value: 0.0005572
```

Svarende til en tilsat nitratmængde på 1 fås ved indsættelse følgende estimat for den forventede tørstofmængde

$$\hat{\alpha} + \hat{\beta} \cdot 1 = 24.68 + 4.61 \cdot 1 = 29.29.$$

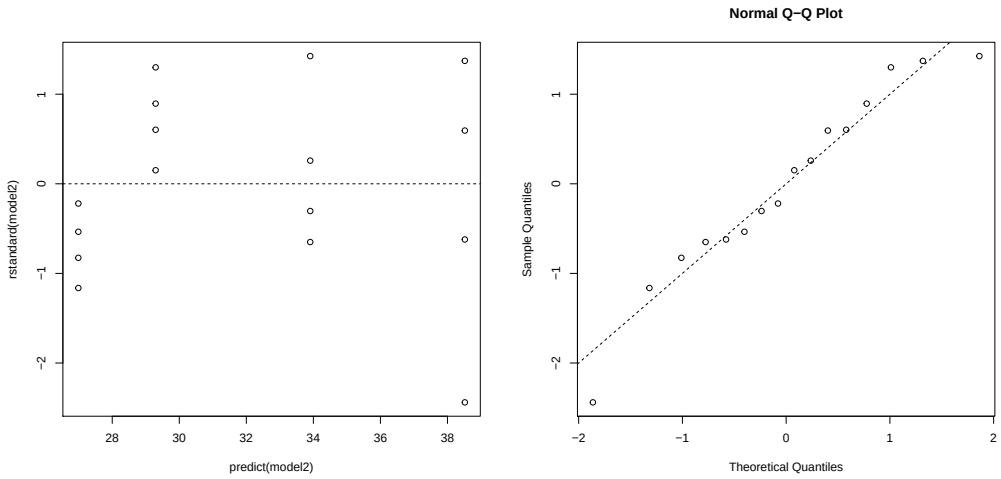
Et 95 % - prædiktionsinterval for tørstofmængden bliver $(20.37 - 38.22)$ for et kar, som tilføres nitratmængden 1. Der er her taget udgangspunkt i følgende R-output

```
predict(model2, newdata = data.frame(nitrat = 1), interval = "p")

##        fit      lwr      upr
## 1 29.29508 20.37182 38.21835
```

4. En korrekt besvarelse bør i hvert tilfælde indeholde residualplot og QQ-plot for `model2`.

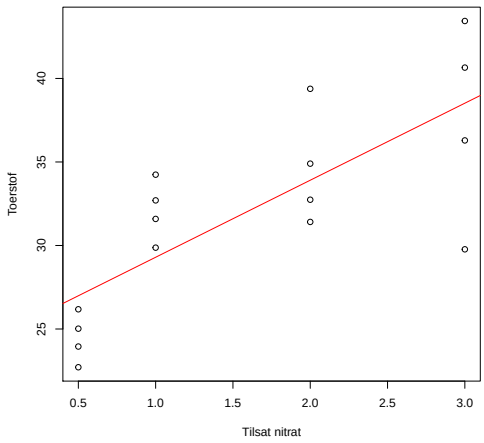
```
plot(predict(model2), rstandard(model2))
abline(h = 0, lty = 2)
qqnorm(rstandard(model2))
abline(0, 1, lty = 2)
```



På baggrund af residualplottet ser det ud som om, at der er problemer med middelværdistrukturen: alle residualer med den mindste prædikterede værdi (`nitrat = 0.5`) er negative, mens alle residualer hørende til næstmindste prædikterede værdi (`nitrat = 1.0`) er positive.

Et scatterplot af `nitrat` mod tørstofmængde kan alternativt benyttes til at illustrere problemerne med linearitetsantagelsen. På denne figur er den estimerede regressionslinje yderligere indtegnet. Sammenhængen er ikke godt beskrevet ved regressionslinjen. Der er desuden en antydning af, at residualvariationen lader til at vokse lidt med størrelsen af målingerne (variansinhomogenitet).

```
plot(salat$nitrat, salat$stof, xlab = "Tilsat nitrat"
     , ylab = "Toerstof")
abline(coef(model2), col = "red")
```



5. Da den lineære regressionsmodel (`model2`) er en delmodel af den ensidede variansanalysemodel (`model1`), kan de to modeller faktisk testes imod hinanden ved et F-test.

```
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: stof ~ nitrat
## Model 2: stof ~ factor(nitrat)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 222.53
## 2     12 159.76  2    62.771 2.3575 0.1369
```

Vi finder, at p-værdien for dette test bliver 0.14. På et signfikansniveau på 5 % er der derfor ikke tilstrækkelig evidens i data for at forkaste hypotesen/modellen der udtrykker, at der er en lineær sammenhæng mellem tilsat nitratmængde og mængden af tørstof.

Opgave 2

1. Responsvariablen (`Kd`) er kontinuert, og vi har to kategoriske forklarende variable (`Lokation`, `Treat`). Det er derfor oplagt, at tage udgangspunkt i en tosidet variansanalysemodel.

   Vi fitter modellen (med vekselvirkning) i `R`

```
pestgolf <- read.table(file = "pestgolf.txt", header = T)
head(pestgolf)

##   Treat Lokation    Kd
## 1   T04      KNY 0.347
## 2   T04      KNY 0.689
## 3   T04      KNY 0.652
## 4   T05      KNY 0.638
## 5   T05      KNY 0.542
## 6   T05      KNY 0.660

modelVeksel <- lm(Kd ~ Treat * Lokation, data = pestgolf)
summary(modelVeksel)

##
## Call:
## lm(formula = Kd ~ Treat * Lokation, data = pestgolf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21567 -0.03192  0.02250  0.04475  0.12633
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.848667   0.057735  14.699  4.9e-09 ***
## TreatT05                 0.315000   0.081649   3.858 0.002277 **
## LokationHONE            -0.459333   0.081649  -5.626 0.000111 ***
```

```
## LokationKNY           -0.286000   0.081649  -3.503 0.004359 **
## TreatT05:LokationHONE -0.008333   0.115470  -0.072 0.943656
## TreatT05:LokationKNY  -0.264333   0.115470  -2.289 0.040991 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1 on 12 degrees of freedom
## Multiple R-squared:  0.8998,Adjusted R-squared:  0.858
## F-statistic: 21.55 on 5 and 12 DF,  p-value: 1.299e-05
```

Estimatet for den ønskede gruppe (`Lokation = HONE, Treat = T05`) estimeres til

$$0.849 + 0.315 - 0.459 - 0.008 = 0.696.$$

2. Vi fitter den additive model for tosidet variansanalyse

```
modelAdd <- lm(Kd ~ Treat + Lokation, data = pestgolf)
```

og tester modellerne imod hinanden ved et F-test.

```
anova(modelAdd, modelVeksel)

## Analysis of Variance Table
##
## Model 1: Kd ~ Treat + Lokation
## Model 2: Kd ~ Treat * Lokation
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     14 0.18774
## 2     12 0.12000  2  0.067739 3.387 0.0682 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: På et signfikansniveau på 5 % kan vi (lige akkurat) ikke afvise hypotesen om, at en additiv model kan benyttes til at beskrive variationen i sorptionskoefficienten for jordprøverne (p-værdi $p = 0.068$). Den additive model udtrykker, at forskellen i den forventede værdi af `Kd` mellem `Treat = T05` og `Treat = T04` er ens, for hver af de tre steder (`Lokation`er) i datasættet.

3. Et `summary` fra den additive model (se nedenfor) viser, at den forventede sorptionskoefficient ligger 0.46 lavere på jordprøver fra `Lokalitet = HONE` end fra `Lokalitet = DYR` (reference).

```
summary(modelAdd)

##
## Call:
## lm(formula = Kd ~ Treat + Lokation, data = pestgolf)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15806 -0.06435 -0.02086  0.06978  0.21306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.89411    0.05459  16.379 1.58e-10 ***
## TreatT05      0.22411    0.05459   4.105  0.00107 **
## LokationHONE -0.46350    0.06686  -6.933 6.95e-06 ***
## LokationKNY  -0.41817    0.06686  -6.255 2.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1158 on 14 degrees of freedom
## Multiple R-squared:  0.8432,Adjusted R-squared:  0.8096
## F-statistic: 25.09 on 3 and 14 DF,  p-value: 6.803e-06
```

Følgende R-output viser, at et 95 % - konfidensinterval for den forventede forskel bliver $((-0.61) - (-0.32))$.

```
confint(modelAdd)
```

```
##                   2.5 %     97.5 %
## (Intercept)   0.7770291  1.0111931
## TreatT05      0.1070291  0.3411931
## LokationHONE -0.6068956 -0.3201044
## LokationKNY  -0.5615622 -0.2747711
```

4. En simpel løsning består i at ændre reference-gruppen for `Lokation` til `HONE` og genfitte den additive model for tosidet variansanalyse.

```
modelAddNy <- lm(Kd ~ Treat + relevel(Lokation, ref = "HONE"), data = pestgolf)
summary(modelAddNy)
```

```
##
## Call:
## lm(formula = Kd ~ Treat + relevel(Lokation, ref = "HONE"), data = pestgolf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15806 -0.06435 -0.02086  0.06978  0.21306
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.43061    0.05459   7.888 1.61e-06
```

```
## TreatT05                              0.22411    0.05459   4.105  0.00107
## relevel(Lokation, ref = "HONE")DYR  0.46350    0.06686   6.933 6.95e-06
## relevel(Lokation, ref = "HONE")KNY  0.04533    0.06686   0.678  0.50879
##
## (Intercept)                          ***
## TreatT05                             **
## relevel(Lokation, ref = "HONE")DYR ***
## relevel(Lokation, ref = "HONE")KNY
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1158 on 14 degrees of freedom
## Multiple R-squared:  0.8432,Adjusted R-squared:  0.8096
## F-statistic: 25.09 on 3 and 14 DF,  p-value: 6.803e-06
```

Det fremgår nu direkte at R-outputtet, at forskellen mellem i den forventede sorption for jordprøver fra KNY og HONE estimeres til 0.05. Ligeledes ses, at et t-test for, om forskellen kan antages at være 0 giver p-værdien 0.51. Der er således ikke belæg for at hævde, at der er forskel på sorptionen i jord fra KNY og HONE.

En alternativ løsning består i at teste hypotesen ved et F-test, med udgangspunkt i en ny version af faktoren Lokation, hvor niveauerne KNY og HONE slås sammen. Det kan fx. udføres som vist her

```
pestgolf$NyLokation <- pestgolf$Lokation
levels(pestgolf$NyLokation)

## [1] "DYR"  "HONE" "KNY"

levels(pestgolf$NyLokation) <- c("DYR", "HONE-KNY", "HONE-KNY")
modelAdd2 <- lm(Kd ~ Treat + NyLokation, data = pestgolf)
anova(modelAdd2, modelAdd)

## Analysis of Variance Table
##
## Model 1: Kd ~ Treat + NyLokation
## Model 2: Kd ~ Treat + Lokation
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     15 0.19390
## 2     14 0.18774  1 0.0061653 0.4598 0.5088
```

Bemærk: Der fås samme p-værdi ($p = 0.51$) ved begge de to fremgangsmåder.

Opgave 3

3.1 Korrekt svar A.

```r
pnorm(25, mean = 24.1, sd = 2.0) - pnorm(20, mean = 24.1, sd = 2.0)
```

```
## [1] 0.6534626
```

3.2 Korrekt svar E.

Estimatet er uændret ($= 0.2$), men da der er indsamlet svar fra fire gange så mange personer, så vil længden af konfidensintervallet blive præcis halvt så stort. Man finder derfor umiddelbart, at konfidensintervallet bliver $0.2 \pm \frac{0.078}{2}$ dvs. (0.161-0.239). Alternativt kan man sætte ind i formlen $\hat{p} \pm 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$, hvor $\hat{p} = 0.2$ og $n = 400$.

3.3 Korrekt svar E.

Benyt formlen for et 95 % - konfidensinterval for en stikprøve, men sample gennemsnit 3043 og spredning 233.

```r
3043 - qt(0.975, 79 - 1) * 233 / sqrt(79) # lower
```

```
## [1] 2990.811
```

```r
3043 + qt(0.975, 79 - 1) * 233 / sqrt(79) # upper
```

```
## [1] 3095.189
```

3.4 Korrekt svar A.

Data udgør en antalstabel, hvor de 366 hanmink er inddelt efter to inddelingskriterier. Der er udført et χ^2-test for, om der er uafhængighed mellem inddelingskriterierne. Hypotesen om uafhængighed forkastes med en p-værdi på $p = 0.0003$.

3.5 Korrekt svar D.

Hvis X er antallet af seksere ved slag med k terninger, så antager vi at $X \sim \text{bin}(k, 1/6)$. Vi udregner sandsynligheden $P(X = 0)$ for *ikke* at få nogle seksere for $k = 6, 7, \ldots, 10$.

```r
dbinom(0, 6, 1/6)
```

```
## [1] 0.334898
```

```r
dbinom(0, 7, 1/6)
```

```
## [1] 0.2790816
```

```r
dbinom(0, 8, 1/6)
```

```
## [1] 0.232568
```

```r
dbinom(0, 9, 1/6)
```

```
## [1] 0.1938067

dbinom(0, 10, 1/6)

## [1] 0.1615056
```

Vi ser at først med 9 terninger, så bliver sandsynligheden for *ikke* at slå en sekser under 20 % (her: 0.194). Dermed kræves mindst 9 terninger for at sikre, at der er mindst 80 % sandsynlighed for at få mindst en sekser.

3.6 Korrekt svar D.

Der skal tages højde for, at målingerne er parrede, da der måles flere gange på hver person.

3.7 Korrekt svar B.