Statistisk Dataanalyse 1

Helle Sørensen

Vejledende besvarelse, eksamen november 2018

Dette er en vejledende besvarelse, incl. den R-kode jeg har benyttet 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 allerførst data fra en af filerne:

```
library(readxl)
sooer <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Nov2018/soeer.xlsx")
sooer2 <- read.table("~/Teaching/Courses/StatDat1/Eksamen/Nov2018/soeer.txt", header=TRUE)
```

Spørgsmål 1.1

Vi er interesseret i eventuelle forskelle i fosforindholdet mellem de fem områder. Vi skal derfor bruge fosforindholdet, dvs. *fosfor* som respons og områdevariablen, dvs. *sted*, som forklarende variabel. Da *fosfor* er kvantitativ og *sted* er kategorisk svarer dette til en ensidet ANOVA.

Hvis datasættet kaldes **sooer**, så er de relevante `lm`-kommandoer følgende (hvor vi også har givet modellerne navne):

```
model1 <- lm(Fosfor ~ Sted, data=sooer)
model2 <- lm(log(Fosfor) ~ Sted, data=sooer)
```

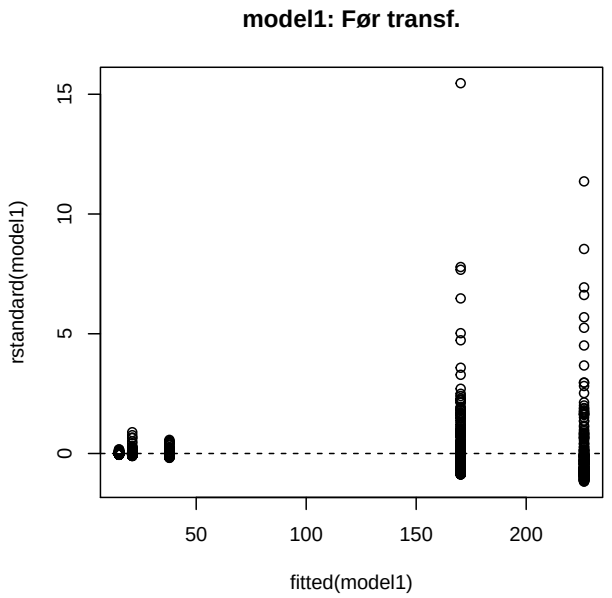
Spørgsmål 1.2

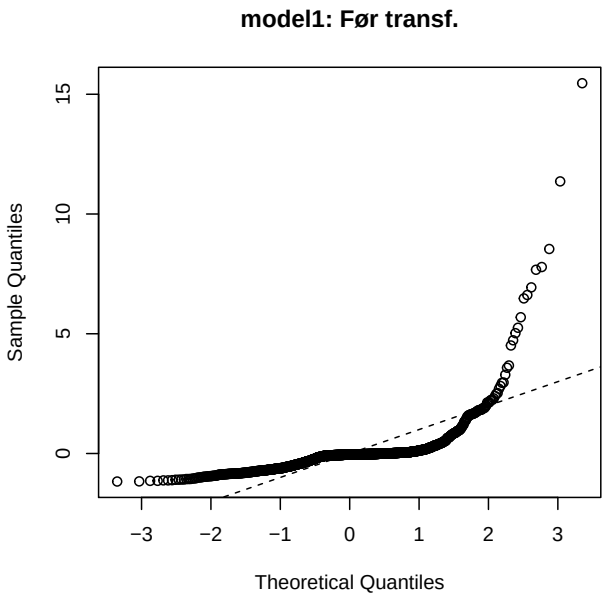
Nedenfor ses residualplot og QQ-plots for hver af de to modeller. Koden er ikke vist i output, men kan ses i Rmd-filen.

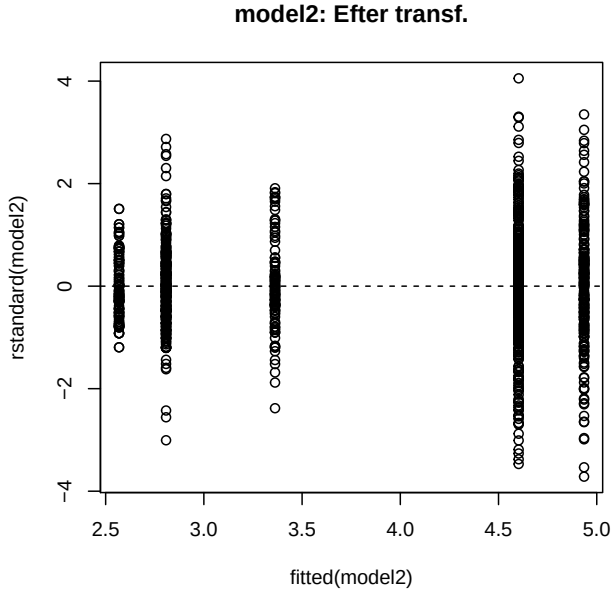
Kommentarer:

- For *model1* (før transformation) er begge plots stærkt problematiske: Residualplottet viser at der er langt større spredning for områder med store værdier end for områder med små værdier, og faktisk også at fordelingen for nogle af områderne er meget skæv. Punkterne i QQ-plottet ligger ikke omkring en ret linie, så residualerne er ikke normalfordelte.
- For *model2* (efter transformation) ser begge figurer langt bedre ud: Der er nogenlunde samme grad af variation for alle fem grupper, og punkterne i residualplottet ligger nogenlunde omkring den rette linie med skæring 0 og hældning 1.

Det er derfor helt oplagt at *model2*, hvor data er log-transformere er bedre til at beskrive fordelingen af data.







Spørgsmål 1.3

Hypotesen er at den forventede log-fosforkoncentration er den samme for alle fem områder. Den kan fx testes vha. `drop1` (vist nedenfor) eller ved at fitte modellen uden områdeeffekt og sammenligne de to modeller med `anova` (prøv selv!).

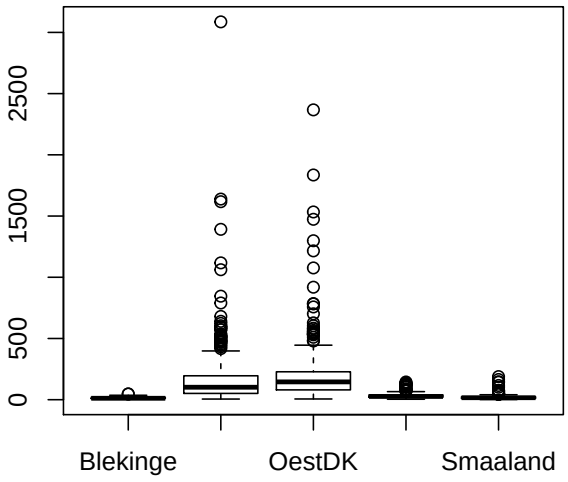
Under alle omstændigheder få en forsvindende lille p-værdi (<2.2e-16). Hypotesen forkastes derfor meget klart, og vi har med stor sikkerhed påvist at der er forskel mellem de fem områder hvad angår fosforkoncentrationen.

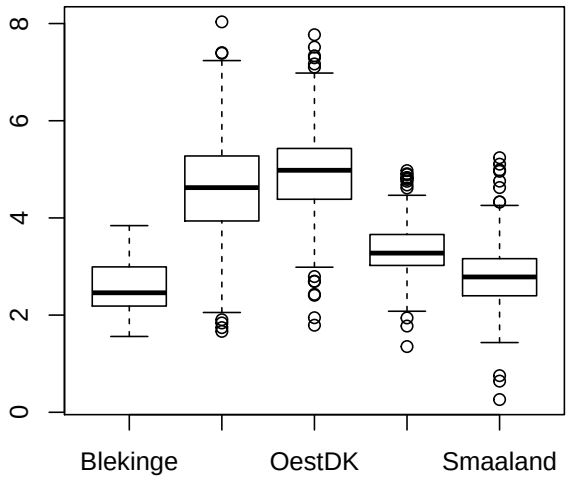
```
drop1(model2, test="F")
```

```
## Single term deletions
##
## Model:
## log(Fosfor) ~ Sted
```

```
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>               889.03 -405.26
## Sted    4    1148.9 2037.95  617.07  399.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusionen er ikke særligt overraskende hvis man betragter boxplots over data, enten på oprindelig skala eller log-skala. Bemærk dog at der ikke er bedt om disse plots i opgaven.





Spørgsmål 1.4

Estimater og 95% konfidensintervaller for forventet log-koncentration for Blekinge og Skåne.

- Blekinge: Dette er referencegruppen, og vi aflæser estimatet til 2.568 og 95% KI til (2.430 , 2.707)

```
summary(model2)$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  2.5682349 0.07064677 36.353182 1.926734e-197
## StedJylland  2.0342215 0.08260982 24.624451 2.796712e-109
## StedOestDK   2.3671874 0.08971767 26.384851 4.475100e-122
## StedSkaane   0.7940588 0.10680788  7.434459  1.954321e-13
## StedSmaaland 0.2393105 0.08362364  2.861756  4.283949e-03
```

```
confint(model2)
```

```
##                    2.5 %    97.5 %
## (Intercept)  2.42963420 2.7068357
## StedJylland  1.87215065 2.1962924
## StedOestDK   2.19117176 2.5432030
## StedSkaane   0.58451421 1.0036035
## StedSmaaland 0.07525062 0.4033703
```

- Skåne: Estimatet beregnes til $2.568 + 0.794 = 3.362$. Konfidensintervallet er lidt sværere: Man kan enten skifte referencegruppe vha. `relevel` (se nedenfor) eller benytte at SE for estimatet er givet ved

$$\text{SE} = \frac{s}{\sqrt{n_{\text{Skåne}}}} = \frac{0.8478}{\sqrt{112}} = 0.0801$$

  Her har vi benyttet at der er 112 søer fra Skåne (se output fra `table` nedenfor). 95% konfidensintervallet bliver så

$$3.362 \pm 1.962 \cdot 0.0801 = (3.205\,,\, 3.519)$$

hvor 1.962 er 97.5% fraktilen i t-fordelingen med 1242 - 5 = 1237 frihedsgrader.

```
# Manuelt
est <- 2.5682349 + 0.7940588
est
```

```
## [1] 3.362294
```

```
table(sooer$Sted)
```

```
##
## Blekinge  Jylland   OestDK   Skaane Smaaland
##      144      392      235      112      359
```

```
SE <- 0.8478 / sqrt(112)
SE
```

```
## [1] 0.08010957
```

```
qt(0.975, df=1237)
```

```
## [1] 1.961884
```

```
est - qt(0.975, df=1237) * SE
```

```
## [1] 3.205128
```

```
est + qt(0.975, df=1237) * SE
```

```
## [1] 3.519459
```

```
# Med relevel:
sooer <- transform(sooer, Sted3 = relevel(factor(Sted), ref="Skaane"))
model3 <- lm(log(Fosfor) ~ Sted3, data=sooer)
confint(model3)
```

```
##                     2.5 %     97.5 %
## (Intercept)    3.2051353  3.5194522
## Sted3Blekinge -1.0036035 -0.5845142
## Sted3Jylland   1.0619617  1.4183636
## Sted3OestDK    1.3821570  1.7641001
## Sted3Smaaland -0.7347602 -0.3747365
```

Spørgsmål 1.5

Estimat og 95% konfidensinterval for forskellen mellem Blekinge og Skåne på log-skala aflæses direkte fra den oprindelige model til

$$\text{estimat}: \ 0.7941, \quad 95\%\ \text{KI}: \ (0.5845\,,\, 1.0036)$$

Bemærk at den forventede værdi er højere i Skåne end i Blekinge.

Ovenstående er den absolutte forskel på log-skala. Hvis vi tager eksponentialfunktionen til estimatet og grænserne i KI, får vi i stedet estimat og 95% KI for den faktor som fosforkoncentationen er højere i Skåne i forhold til Blekinge:

$$\text{estimat for faktor}: \ 2.212, \quad 95\%\ \text{KI for faktor}: \ (1.794\,,\, 2.728)$$

```
exp(0.79406)
```

```
## [1] 2.21236
```

```r
exp(0.58451421)
```

```
## [1] 1.794119
```

```r
exp(1.0036035)
```

```
## [1] 2.728095
```

Spørgsmål 1.6

Hypotesen er at de forventede værdier for log-koncentration er den samme for Blekinge, Skåne og Småland.

Hypotesen testes ved (i) at lave en ny kategorisk variabel der har et fælles niveau for de tre svenske områder, men stadig forskellige niveauer for Jylland Østdanmark, (ii) at benytte den nye variabel som forklarende variabel i en ny ensidet ANOVA, og (iii) sammenligne de en nye og den oprindelige model med et F-test vha. funktionen `anova`. Dette giver mening fordi de to modeller er nestede.

F-testet giver p-værdien 5.675e-13, så hypotesen forkastes. Det er meget klart påvist at der er forskel på niveauet af log-fosforkoncentrationen i de tre svenske områder.

```r
Sted4 <- sooer$Sted
Sted4[Sted4 == "Blekinge"] <- "Sverige"
Sted4[Sted4 == "Smaaland"] <- "Sverige"
Sted4[Sted4 == "Skaane"] <- "Sverige"
table(Sted4)
```

```
## Sted4
## Jylland  OestDK Sverige
##     392     235     615
```

```r
model4 <- lm(log(Fosfor) ~ Sted4, data=sooer)
anova(model4, model2)
```

```
## Analysis of Variance Table
##
## Model 1: log(Fosfor) ~ Sted4
## Model 2: log(Fosfor) ~ Sted
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1239 930.50
## 2   1237 889.03  2    41.469 28.85 5.675e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Eftersom man i spørgsmålet eksplicit bliver bedt om at benytte hele datasættet, gives der ikke fuldt pointtal hvis man i stedet har testet hypotesen ved kun at benytte data fra de svenske områder - men det er bestemt ikke en dum ting at gøre.

De to danske områder er i øvrigt også forskellige (p=2.157e-06).

Opgave 2

Vi indlæser først data fra en af filerne:

```r
library(readxl)
elections <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Nov2018/elections.xlsx")
elections2 <- read.table("~/Teaching/Courses/StatDat1/Eksamen/Nov2018/elections.txt", header=TRUE)
```

Spørgsmål 2.1

Den lineære regressionsmodel er

$$\text{expenditures}_i = \alpha + \beta \cdot \text{approval}_i + e_i, \quad i = 1, \ldots, 15$$

hvor $e_1, \ldots, e_{15}$ er iid $N(0, \sigma^2)$-fordelte.

Vi fitter modellen og tester hypotesen $H_0 : \beta_0$. Vi får p-værdien 0.0011, så hypotesen afvises. Vi har med andre ord påvist at en sammenhæng mellem popularitet og kampagneudgifter. Estimatet for β er negativt, så mere populær kandidaten er ved valgkampens start, jo mindre mindre bruger han/hun på kampagnen - hvilket jo giver god mening.

```
linreg <- lm(expenditures ~ approval, data=elections)
summary(linreg)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 178.017194 18.8660394  9.435854 3.514224e-07
## approval     -1.525383  0.3657189 -4.170916 1.097505e-03
```

Spørgsmål 2.2

Vi fitter den multiple regressionsmodel med koden givet i opgaven og får følgende estimerede sammenhæng:

$$\begin{aligned}\text{performance} &= \hat{\alpha} + \hat{\beta}_1 \cdot \text{approval} + \hat{\beta}_2 \cdot \text{expenditures} \\ &= 0.1279 + 0.7986 \cdot \text{approval} + 0.0994 \cdot \text{expenditures}\end{aligned}$$

Residualspredningen estimeres til 0.0464.

```
multipel <- lm(performance ~ approval + expenditures, data=elections)
summary(multipel)
```

```
##
## Call:
## lm(formula = performance ~ approval + expenditures, data = elections)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07247 -0.03957  0.01305  0.03230  0.05869
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1278558  0.1131222    1.13     0.28
## approval     0.7986354  0.0011969  667.26   <2e-16 ***
## expenditures 0.0994301  0.0005936  167.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04637 on 12 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.556e+05 on 2 and 12 DF,  p-value: < 2.2e-16
```

Spørgsmål 2.3

Estimatet hørende til `expenditures` er $\hat{\beta}_2 = 0.0994$. Enheden for `expenditures` er 1000 dollars, så vi har følgende fortolkning: Hvis to siddende borgmestre (A og B) er lige populære ved kampagnens start, og

kandidat A bruger 1000 dollars mere end kandidat B på kampagnen, så vil man forvente at kandidat A får 0.0994 procentpoint flere stemmer end kandidat B.

Vi tester så hypotesen om at det ikke hjælper på valgresultatet at øge kampagneudgifterne, dvs. hypotesen $H_0 : \beta_2 = 0$. Vi aflæser p-værdien til <2e-16 så der er overvældende evidens for at det hjælpe at øge kampagneudgifterne.

Spørgsmål 2.4

Der er tale om prædiktion af en ny observation, så vi beregner et et 95% prædiktionsinterval svarende til at `approval=40` og `expenditures=110`.

Vi får prædiktionsintervallet (42.9 , 43.1), så dette er intervallet som med sandsynlighed 95% vil indeholde det nye valgresultat.

```
newData <- data.frame(approval=40, expenditures=110)
predict(multipel, newData, interval="p")
```

```
##        fit      lwr      upr
## 1 43.01058 42.90467 43.11649
```

Opgave 3

Spørgsmål 3.1: E

Vi antager at $Y \sim N(1060, 195^2)$ beregner sandsynligheden $P(Y < 880)$.

```
pnorm(880, mean=1060, sd=195)
```

```
## [1] 0.1779836
```

Spørgsmål 3.2: A

Vi skal bruge 75% fraktilen i $N(1060, 195^2)$-fordelingen.

```
qnorm(0.75, 1060, 195)
```

```
## [1] 1191.526
```

Spørgsmål 3.3: F

Hypotesen om uafhængighed afvises, dvs. der er evidens for at der er en sammenhæng.

```
tabel <- matrix(c(44,5,32,19),2,2)
tabel
```

```
##      [,1] [,2]
## [1,]   44   32
## [2,]    5   19
```

```
chisq.test(tabel, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabel
## X-squared = 10.025, df = 1, p-value = 0.001544
```

Spørgsmål 3.4: D

Hvis Y betegner antallet af venstrehåndede i stikprøven, så er $Y \sim \text{bin}(25, 0.1)$. Vi beregner $P(Y \leq 4)$.

```
pbinom(4, size=25, prob=0.1)
```

```
## [1] 0.9020064
```

Spørgsmål 3.5: E

Dosis og metode er begge kategoriske variable og bør begge indgå i modellen. At effekten af dosis ikke afhænger af metoden og at dette ikke skal undersøges nærmere, betyder at vekselvirkningen mellem dosis og metode ikke skal inddrages i modellen.

Spørgsmål 3.6: A

En type II fejl består i at man *accepterer en falsk hypotese*. Eftersom hypotesen her er at tilslutningen til Alternativet er uændret, betyder en type II fejl at man konkluderer at tilslutningen er uændret, selvom sandheden er at den *har* ændret sig.

Spørgsmål 3.7: B

Der er tale om en enkelt normalfordelt stikprøve.

```
3487 - qt(0.975, df=31) * 443 / sqrt(32)
```

```
## [1] 3327.281
```

```
3487 + qt(0.975, df=31) * 443 / sqrt(32)
```

```
## [1] 3646.719
```

Spørgsmål 3.8: C

Konfidensintervallet udtaler sig om den forventede værdi, altså om gennemsnittet i populationen (ikke om enkeltpersoner).