Statistisk Dataanalyse 1, januar 2018

Helle Sørensen

Vejledende besvarelse

Nedenstående er en vejledende besvarelse. Besvarelsen inkluderer også R-kode og R-output, men det vil en eksamensbesvarelse naturligvis ikke gøre.

Opgave 1

```
library(readxl)
setwd("~/Teaching/Courses/StatDat1/Eksamen/Jan2018")
ssData1 <- read.table("ss2017-18.txt", header=TRUE)
ssData2 <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Jan2018/ss2017-18.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/Europe/Berlin'
```

Spørgsmål 1.1

Det er naturligt at bruge en tosidet ANOVA eftersom begge forklarende variable (studie og køn) er kategoriske.

R-koden til de to ønskede modeller er

```
model1 <- lm(figur2 ~ studie + kon + studie*kon, data=ssData2)
model2 <- lm(log(figur2) ~ studie + kon + studie*kon, data=ssData2)
```

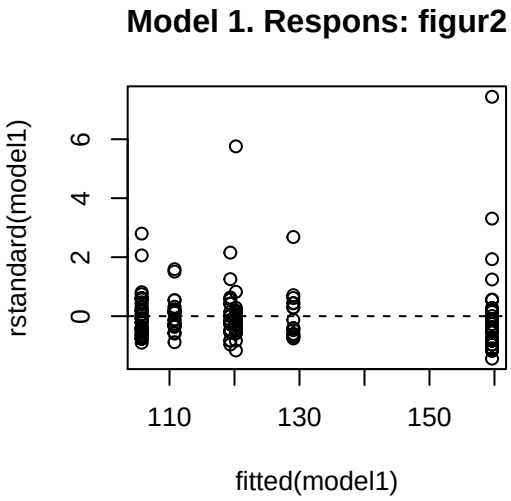
Fra *summary* fra de to modeller (ikke vist) kan man aflæse residualspredningen til 73.71 i modellen med *figur2* som respons og til 0.4342 i modellen med *log(figur2)* som respons.

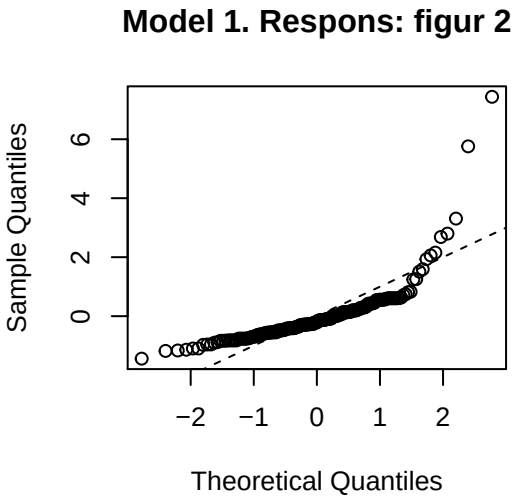
Spørgsmål 1.2

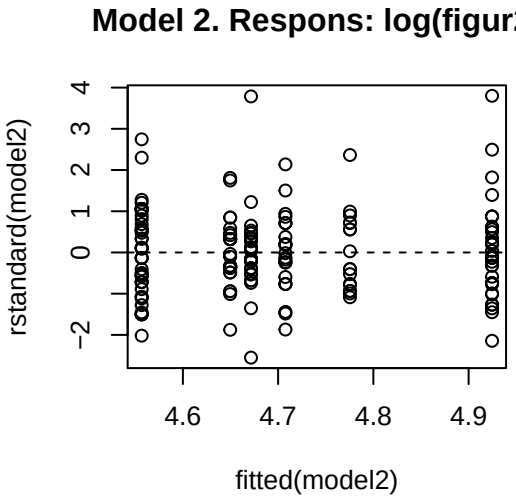
Nedenfor ses residualplot og QQ-plot for standardiserede residualer for de to modeller.

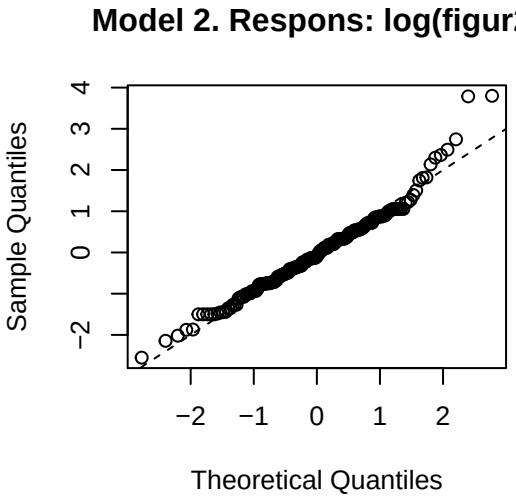
```
par(mfrow=c(2,2))
plot(fitted(model1), rstandard(model1), main="Model 1. Respons: figur2")
abline(h=0, lty=2)
qqnorm(rstandard(model1), main="Model 1. Respons: figur 2")
abline(0,1, lty=2)

plot(fitted(model2), rstandard(model2), main="Model 2. Respons: log(figur2)")
abline(h=0, lty=2)
qqnorm(rstandard(model2), main="Model 2. Respons: log(figur2)")
abline(0,1, lty=2)
```









- Modellen med *figur2* som respons: QQ-plottet udviser kraftig afvigelse fra den rette linie. Der er fx en del residualer der er alt for store i forhold til det forventede (dette ses også i residualplottet). Bortset fra disse store residualer, udviser residualplottet nogenlunde symmetri (godt) og ikke tydelige tegn på variansinhomogenitet.
- Modellen med *log(figur2)* som respons: QQ-plottet ser meget bedre ud selvom der stadig er en del residualer der er større end man ville forvente. Residualplottet viser fin symmetri og ingen klare tegn på variansinhomogenitet.
- Udfra ovenstående fremgår det at modellen med *log(figur2)* klart er at foretrække.

Spørgsmål 1.3

Vi skal teste hypotese om at der ikke er vekselvirkning mellem køn og studie. Vi fitter derfor modellen uden vekselvirkning og sammenligner de to modeller med et F-test:

```
model3 <- lm(log(figur2) ~ studie + kon, data=ssData2)
anova(model3, model2)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: log(figur2) ~ studie + kon
## Model 2: log(figur2) ~ studie + kon + studie * kon
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    178 34.171
## 2    176 33.185  2   0.98604 2.6147 0.07603 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi får $F = 2.62$ og p-værdien 0.076. Vi forkaster således *ikke* hypotesen: Der er ikke evidens for vekselvirkning i data. Ingen vekselvirkning betyder at forskellen mellem mænd og kvinder i forventet værdi af *log(figur2)* er ens for de tre studier (eller omvendt at forskellen mellem studierne ikke afhænger af kønnet).

Spørgsmål 1.4

```
summary(model3)
```

```
##
## Call:
## lm(formula = log(figur2) ~ studie + kon, data = ssData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00521 -0.27651 -0.01292  0.23857  1.71835
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.828164   0.077444  62.344  < 2e-16 ***
## studieMatematik  0.004568   0.081390   0.056  0.95530
## studieMatOk     -0.027752   0.089537  -0.310  0.75696
## konMand         -0.211680   0.065883  -3.213  0.00156 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4381 on 178 degrees of freedom
## Multiple R-squared:  0.05515,    Adjusted R-squared:  0.03923
## F-statistic: 3.463 on 3 and 178 DF,  p-value: 0.01753
```

```
confint(model3)
```

```
##                       2.5 %      97.5 %
## (Intercept)      4.6753382  4.98098933
## studieMatematik -0.1560457  0.16518215
## studieMatOk     -0.2044418  0.14893801
## konMand         -0.3416929 -0.08166759
```

Referencegruppen for køn er kvinder, og referencegruppen for studier er aktuar, så interceptet svarer til kvindelige aktuarstuderende. Estimatet for forventet værdi af *log(figur2)* for kvindelige aktuarstuderende estimeres således til 4.828 med 95% konfidensinterval (4.675 , 4.981).

Der er 142 punkter i plottet, og $\log(142) = 4.956$. Denne værdi ligger lige netop indenfor konfidensintervallet, så de kvindelige aktuarstuderende gætter altså ikke signifikant for højt.

Spørgsmål 1.5

Den forventede forskel mellem kvinder og mænd i forventet værdi af *log(figur2)* estimeres til 0.2117.

Dette svarer til at kvinder gætter en faktor $e^{0.2117} = 1.24$ højere end mænd.

Hypotesen om at der ikke er forskel er mænd og kvinders gæt, giver p-værdien 0.0016, så der *er* signifikant forskel på mænd og kvinder. Altså: Kvindeer gætter højere end mænd.

Spørgsmål 1.6

Hypotesen er at studerende fra alle tre studier har samme forventede værdi af *log(figur2)*. Vi fitter modellen uden effekt af studie og sammenligner de to modeller. Dette giver p-værdien 0.91, så hypotesen kan ikke forkastes. Der er altså ikke tegn på forskel mellem studieretningerne.

```
model4 <- lm(log(figur2) ~ kon, data=ssData2)
anova(model4, model3)
```

```
## Analysis of Variance Table
##
## Model 1: log(figur2) ~ kon
## Model 2: log(figur2) ~ studie + kon
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    180 34.207
## 2    178 34.171  2  0.036024 0.0938 0.9105
```

Opgave 2

```
library(readxl)
setwd("~/Teaching/Courses/StatDat1/Eksamen/Jan2018")
johnsonData1 <- read.table("johnson-fatpct.txt", header=TRUE)
johnsonData2 <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Jan2018/johnson-fatpct.xlsx")
```

Spørgsmål 2.1

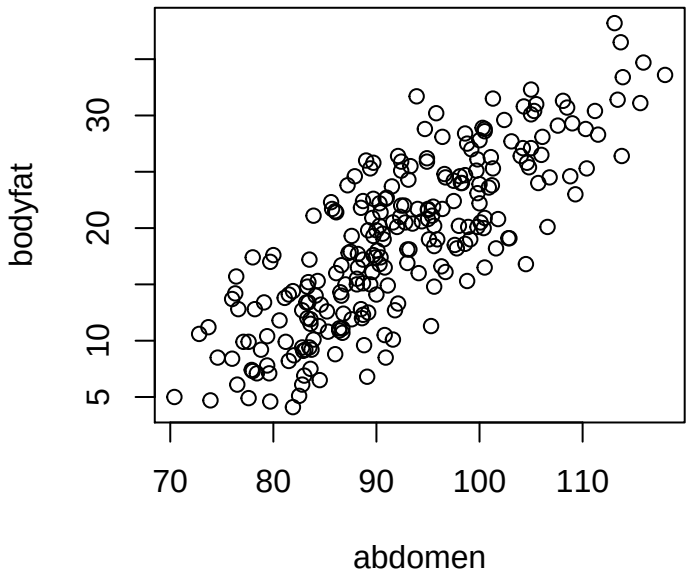
Det er mest naturligt at benytte maveomkreds som forklarende variabel (x) og fedtprocent som responsvariabel (y). Den tilhørende tegning er vist nedenfor.

Sammenhængen ser ud til at være lineær, så vi vil bruge en lineær regressionsmodel:

$$\text{bodyfat}_i = \alpha + \beta \cdot \text{abdomen}_i + e_i$$

hvor $e_1, \ldots, e_{243}$ er uafhængige og $N(0, \sigma^2)$-fordelte.

```
plot(bodyfat ~ abdomen, data=johnsonData2)
```



Spørgsmål 2.2

Modellen fittes med *lm* og estimaterne fås med *summary*. Vi får

$$\hat{\alpha} = -37.65, \quad \hat{\beta} = 0.6126, \quad \hat{\sigma} = 4.301$$

Betragt to mænd med maveomkreds på 110 cm hhv. 100 cm. Forskellen i maveomkreds er 10 cm, så den forventede forskel i fedtprocent er $10 \cdot \beta$, der estimeres til 6.126.

NB: Det er naturligvis fint at beregne estimaterne for hver af de to mænd (29.737 hhv. 23.611) og trække dem fra hinanden.

```
linreg <- lm(bodyfat ~ abdomen, data=johnsonData2)
summary(linreg)
```

```
##
## Call:
## lm(formula = bodyfat ~ abdomen, data = johnsonData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1339  -3.2678   0.1175   2.8329  11.8257
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -37.64724    2.64472  -14.23   <2e-16 ***
## abdomen       0.61258    0.02851   21.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.301 on 241 degrees of freedom
## Multiple R-squared:  0.657,  Adjusted R-squared:  0.6556
## F-statistic: 461.7 on 1 and 241 DF,  p-value: < 2.2e-16
```

Spørgsmål 2.3

Lineær regression hvor hofteomkredsen er den eneste forklarende variabel: Regressionskoefficienten estimeres til 0.726.

Multipel lineær regression med både maveomkred og hofteomkreds som forklarende variable: Regressionskoefficienten hørende til hofteomkreds estimeres til -0.3318.

I modellen hvor hofte er eneste forklarende variabel, er regressionskoefficienten positiv svarende til at større hofteomkreds og højere fedtprocent typisk hører sammen. I den multiple regression er regressionskoefficienten hørende til hofteomkreds negativ. Dette betyder at *for givet maveomkreds* hører større hofteomkreds typisk sammen med lavere fedtprocent.

Hvis man betragter en mand med en given maveomkreds vil en stor hofteomkreds tyde på at manden er stor (uden nødvendigvis at være tyk), mens en lille hofteomkreds vil tyde på at manden er tyk. Dette argument holder kun for mænd, hvorimod kvinder med høj fedtprocent typisk også vil være store om hofterne.

```
lm(bodyfat ~ hip, data=johnsonData2)
```

```
##
## Call:
## lm(formula = bodyfat ~ hip, data = johnsonData2)
##
## Coefficients:
## (Intercept)          hip
##     -53.450        0.726
```

```
lm(bodyfat ~ abdomen + hip, data=johnsonData2)
```

```
##
## Call:
## lm(formula = bodyfat ~ abdomen + hip, data = johnsonData2)
##
## Coefficients:
## (Intercept)      abdomen          hip
##    -20.8824       0.7892      -0.3318
```

Spørgsmål 2.4

Konfidensintervallet beregnes til (12.92 , 14.43), og prædiktionsintervallet beregnes til (5.43 , 21.92). Eftersom værdien 17 ligger i prædiktionsintervallet, er det ikke en usædvanlig værdi.

```
newData <- data.frame(abdomen=85, hip=98)
multipel <- lm(bodyfat ~ abdomen + hip, data=johnsonData2)
predict(multipel, newData, interval="c")
```

```
##        fit      lwr      upr
## 1 13.67509 12.92283 14.42734
```

```
predict(multipel, newData, interval="p")
```

```
##        fit      lwr      upr
## 1 13.67509 5.426149 21.92402
```

Opgave 3

Spørgsmål 3.1

Stikprøverne fra 2012 og 2017 er parrede fordi de stammer fra de samme plantearter (ikke 73 plantearter det ene år og 73 andre plantearter det andet år).

Vi analyserer således forskellen som en enkelt stikprøve. Estimatet for den forventede forskel er lig gennemnittet, dvs. 3.3536, og et 95% konfidensintervalt beregnes til

$$\bar{y} \pm t_{0.975,72} \cdot \frac{s}{\sqrt{73}} = 3.3536 \pm 1.993 \cdot \frac{8.3039}{\sqrt{73}} = 3.3536 \pm 1.9375 = (1.4161\,,\, 5.2910)$$

Spørgsmål 3.2

Vi udfører et uafhængighedstest. Hypotesen er at udbredelse og fredning er uafhængige.

Med *chisq.test* får vi $X^2 = 9.41$ og p-værdien 0.0022 (og $X^2=7.96 og p-værdien 0.0048 hvis vi laver kontinuitetskorrektion). Vi afviser således hypotesen og konkluderer at fredning og udbredelse ikke er uafhængige.

```
A <- matrix(c(18,15,8,32), 2,2)
A
```

```
##      [,1] [,2]
## [1,]   18    8
## [2,]   15   32
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  A
## X-squared = 9.4104, df = 1, p-value = 0.002158
```

```
chisq.test(A)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  A
## X-squared = 7.9642, df = 1, p-value = 0.004771
```

Spørgsmål 3.3

De estimerede betingede sandsynligheder beregnes udfra tabellen:

$$P(\text{udbredelse vokset} \,|\, \text{fredet}) = \frac{32}{40} = 0.8$$
$$P(\text{udbredelse vokset} \,|\, \text{ikke fredet}) = \frac{15}{33} = 0.454$$

Fredning fremmer således udbredelsen (og dette er statistisk signifikant jf. spørgsmål 3.2).