Statistisk Dataanalyse 1, november 2017

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/Nov2017")
gunData1 <- read.table("guns.txt", header=TRUE)
gunData2 <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Nov2017/guns.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'default/Europe/
## Berlin'
```

Spørgsmål 1.1

Det er naturligt at bruge selvmordsraten som respons (y) og våbenudbredelsen som forklarende variabel (x). Tegningen er lavet nedenfor. Sammenhængen er tilnærmelsesvis lineær, så vi bruger en lineær regressionsmodel:

$$\text{SuicideRate}_i = \alpha + \beta \cdot \text{GunOwnerPct}_i + e_i$$

hvor $e_1, \ldots, e_{50}$ er uafhængige og normalfordelte med middelværdi 0 og spredning σ.

Modellen fittes med *lm*, og vi får estimaterne

$$\hat{\alpha} = 6.436, \quad \hat{\beta} = 0.244, \quad \hat{\sigma} = 3.297.$$

```
plot(SuicideRate ~ GunOwnerPct, data=gunData1)
```

```
linreg <- lm(SuicideRate ~ GunOwnerPct, data=gunData1)
summary(linreg)
```

```
##
## Call:
## lm(formula = SuicideRate ~ GunOwnerPct, data = gunData1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9787 -2.2854 -0.9087  1.8910  8.5868
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.43598    1.40527   4.580 3.32e-05 ***
## GunOwnerPct  0.24417    0.03525   6.928 9.49e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.297 on 48 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4896
## F-statistic: 47.99 on 1 and 48 DF,  p-value: 9.488e-09
```

Spørgsmål 1.2

At der *ikke* er sammenhæng mellem våbenudbredelse og selvmordsrate svarer til hypotesen om at hældningen er nul, altså $H_0 : \beta = 0$. Fra *summary* aflæses $T = 6.93$ og p-værdien 9.49e-09, så hypotesen forkastes med et brag, og vi konkluderer at der *er* sammenhæng, og at denne sammenhæng er sådan at våbenudbredelse og selvmordsrate er stor/lille samtidig.

For en stat med våbenudbredelse på 45% fås prædiktionen

$$\hat{\alpha} + \hat{\beta} \cdot 45 = 6.436 + 0.244 \cdot 45 = 17.4$$

Det tilhørende konfidensinterval beregnes nemmest med *predict* (som også giver prædiktionen). Vi får 95% KI (16.3 , 18.5).

```
newData <- data.frame(GunOwnerPct=45)
predict(linreg, newData, interval="c")
```

```
##        fit      lwr      upr
## 1 17.42375 16.34992 18.49758
```

Spørgsmål 1.3

Den angivne model er en model med to parallelle rette linier: en linie for stater *med* love og en for stater *uden love*. De estimerede linier er

$$\begin{aligned} \text{Med love} : \text{SuicideRate} &= (9.68794 - 2.38339) + 0.18433 \cdot \text{GunOwnerPct} \\ &= 7.30455 + 0.18433 \cdot \text{GunOwnerPct} \\ \text{Uden love} : \text{SuicideRate} &= 9.68794 + 0.18433 \cdot \text{GunOwnerPct} \end{aligned}$$

For en fiktiv stat med mindst en lov og en våbenudbredelse på 27.15% får vi således estimatet

$$7.30455 + 0.18433 \cdot 27.15 = 12.31,$$

og for en fiktiv stat uden love og en våbenudbredelse på 45.19% får vi estimatet

$$9.68794 + 0.18433 \cdot 45.19 = 18.02.$$

Prædiktionerne kan også beregnes med *predict* (se nedenfor).

```
model <- lm(SuicideRate ~ GunOwnerPct + Law, data=gunData1)
summary(model)
```

```
##
## Call:
## lm(formula = SuicideRate ~ GunOwnerPct + Law, data = gunData1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9212 -1.7932 -0.8349  2.1256  7.5478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.68794    2.17902   4.446 5.32e-05 ***
## GunOwnerPct  0.18433    0.04638   3.974 0.000242 ***
## LawYes      -2.38339    1.24327  -1.917 0.061324 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.209 on 47 degrees of freedom
## Multiple R-squared:  0.5362, Adjusted R-squared:  0.5165
## F-statistic: 27.17 on 2 and 47 DF,  p-value: 1.439e-08
```

```
newData2 <- data.frame(Law="Yes", GunOwnerPct=27.15)
predict(model, newData2)
```

```
##        1
## 12.30901
```

```
newData3 <- data.frame(Law="No", GunOwnerPct=45.19)
predict(model, newData3)
```

```
##        1
## 18.01765
```

Spørgsmål 1.4

Vi kan teste hypotesen om regressionslinierne for stater med og uden ekstra love har samme skæring. Hvis skæringerne kaldes α_{yes} og α_{no}, er hypotesen at $\alpha_{yes} = \alpha_{no}$. Hypotesen svarer til at variablen *Law* kan fjernes fra modellen uden at det modellen giver dårligere tilpasning til data.

Hypotesen kan testes med et t-test, som angoívet i *summary* fra modellen: Vi får $T = -1.92$ og p-værdien 0.06. Vi accepterer altså lige netop hypotesen og konkluderer at love i sig ikke har en effekt udover den effekt de har på våbenudbredelsen.

Hypotesen kunne også testes med et F-test, fx vha. funktionen *drop1*, med samme p-værdi (som altid).

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

```
## Single term deletions
##
## Model:
```

```
## SuicideRate ~ GunOwnerPct + Law
##             Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                   483.99 119.50
## GunOwnerPct  1   162.641 646.64 131.99  15.794 0.0002417 ***
## Law          1    37.844 521.84 121.27   3.675 0.0613240 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Opgave 2

Spørgsmål 2.1

Vi ser på de øverste QQ-plots. For de log-transformerede data ligger datapunkterne fint omkring en ret linie, mens det overhovedet ikke er tilfælde for de rå data. Det er således mere rimeligt at antage at de log-transformere værdier er normalfordelt.

Vi antager således at de 26 værdier fra kontrolskyerne er normalfordelt med middelværdi μ og spredning σ, og beregner et 95% konfidensinterval for μ. Hvis $y_1, \ldots, y_{26}$ er de log-værdierne, får vi følgende:

$$\bar{y} \pm t_{0.975,25} \cdot \frac{s}{\sqrt{n}} = 3.990 \pm 2.059 \cdot \frac{1.642}{\sqrt{26}} = 3.990 \pm 0.663 = (3.327\,,\, 4.653)$$

```
qt(0.975, df=25)
```

```
## [1] 2.059539
```

```
# qt(0.975, df=25) * 1.642 / sqrt(26)
# 3.990 - qt(0.975, df=25) * 1.642 / sqrt(26)
# 3.990 + qt(0.975, df=25) * 1.642 / sqrt(26)
```

Spørgsmål 2.2

Vi skal bruge modellen for to uafhængige stikprøver for de log-transformerede data. Dette svarer til en ensidet ANOVA med to grupper. Formelt kan modellen skrives

$$\log(\text{regn})_i = \alpha_{\text{beh}_i} + e_i$$

hvor e_i'erne er uafhængige og $N(0, \sigma^2)$-fordelt for $i = 1, \ldots, n$.

Vi er interesseret i hypotesen

$$H_0 : \alpha_{\text{sølviodid}} = \alpha_{\text{kontrol}}$$

Ingredienserne til testet kan finde i R-outputtet eftersom anden linie i koefficienttabellen netop vedrører forskellen mellem de to α'er:

$$T_{\text{obs}} = \frac{\hat{\alpha}_{\text{sølviodid}} - \hat{\alpha}_{kontrol}}{\text{SE}(\hat{\alpha}_{\text{sølviodid}} - \hat{\alpha}_{kontrol})} = \frac{1.1438}{0.4495} = 2.54$$

Teststørrelsen skal vurderes i t-fordelingen med $52 - 2 = 50$ frihedsgrader. Dette giver p-værdien

$$p = P(|T| \geq 2.54) = 0.014$$

Vi afviser hypotesen og konkluderer at behandling med sølviodid øger log-regnmængden fra skyer, og dermed også regnmænden skyer.

```
1.1438/0.4495
```

```
## [1] 2.544605
```

```r
2 * pt(-2.54, df=50)
```

```
## [1] 0.01423897
```

Spørgsmål 2.3

Fra *summary* og *confint* ser vi at sølviodid øger forventet log-regnmængde med 1.1438 med tilhørende 95% KI (0.241 , 2.047).

Dette svarer til at medianen af regn (ikke-transformeret) øges med en faktor exp(1.1438) = 3.138 med tilhørende konfidensinterval (exp(0.241) , exp(2.047)) = (1.272 , 7.742).

Dette svarer til slut til en procentvisforøgelse på 214% med 95% KI på (27% , 674%).

Opgave 3

```r
rodData1 <- read.table("rodlaengde.txt", header=TRUE)
rodData2 <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Nov2017/rodlaengde.xlsx")
```

Spørgsmål 3.1

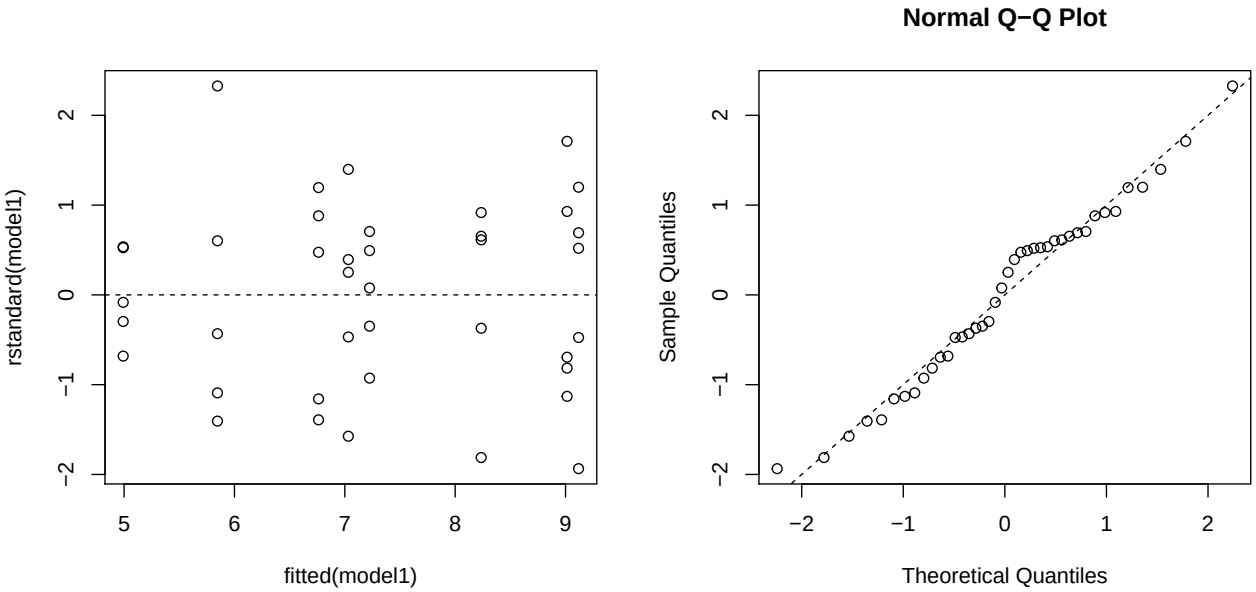
Den tosidede ANOVA med vekselvirkning fittes med kommandoen

```
lm(lgd ~ dosis + sted + dosis*sted, data=rodData1)
```

Residualplottet set fint ud eftersom de standardiserede resdualer fordeler sig nogenlunde symmetrisk omkring 9 og med cirka samme lodrette variation i alle dele af plottet. Der er heller ikke nogle ekstremt store/små residualer.

QQ-plottet ser ligeledes fint ud eftersom punkterne fordelinger sig omkring den rette linie med skæring 0 og hældning 1.

```r
par(mfrow=c(1,2))
model1 <- lm(lgd ~ dosis + sted + dosis*sted, data=rodData1)
plot(fitted(model1), rstandard(model1))
abline(h=0, lty=2)
qqnorm(rstandard(model1))
abline(0,1,lty=2)
```



Spørgsmål 3.2

Forskel i dosiseffekt mellem sted A og sted B er det samme som vekselvirkning mellem dosis og sted. Vi tester derfor hypotesen om at der *ikke* er vekselvirkning mellem dosis og sted.

Hypotesen testes med F-testet der sammenligner modellerne med og uden vekselvirkning. Vi får F=0.55 og p=0.65, sp hypotesen accepteres. Der er altså ikke umiddelbart tegn på at dosiseffekten er forskellig på sted A og sted B.

```
model2 <- lm(lgd ~ dosis + sted, data=rodData1)
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: lgd ~ dosis + sted
## Model 2: lgd ~ dosis + sted + dosis * sted
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     35 40.880
## 2     32 38.867  3    2.0123 0.5523 0.6503
```

Spørgsmål 3.3

Vi tager nu udgangspunkt i modellen uden vekselvirkning, dvs. model2.

Den forventede rodlængde for en plante med dosis mellem2 fra sted B estimeres til

$$8.8120 - 2.1690 + 0.5080 = 7.151.$$

Referencegruppen er høj dosis fra sted A. Det ses at estimaterne (kontrasterne) er negative for lav, mellem2 og mellem2, så høj dosis giver størst forventet rodlængde. Estimatet(kontrasten) for B er positiv, så sted B giver størst forventet rodlængde. Da modellen er additiv, giver dette tilsammen nødvendigvis at kombinationen høj dosis fra sted B giver højst forventet rodlængde.

```
summary(model2)
```

```
##
## Call:
```

```
## lm(formula = lgd ~ dosis + sted, data = rodData1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6710 -0.9247  0.0075  0.8275  2.4670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.8120     0.3821  23.062  < 2e-16 ***
## dosislav      -3.6470     0.4833  -7.546 7.65e-09 ***
## dosismellem1  -1.3360     0.4833  -2.764  0.00904 **
## dosismellem2  -2.1690     0.4833  -4.488 7.46e-05 ***
## stedB          0.5080     0.3418   1.486  0.14612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.081 on 35 degrees of freedom
## Multiple R-squared:  0.6398, Adjusted R-squared:  0.5986
## F-statistic: 15.54 on 4 and 35 DF,  p-value: 2.12e-07
```

Spørgsmål 3.4

Forskellen i forventet rodlængde mellem planter der får høj dosis og lav dosis estimeres til 3.647 med 95% KI (0.355 , 2.317).

Hypotesen om at forventet rodlængde er ens på sted A og sted B, testes med et t-test. Fra *summary* aflæses T=1.486 og p=0.15. Hypotesen accepteres, og der er altså ikke tegn for at rodlængden er forskellig de to steder.

```
confint(model2)
```

```
##                   2.5 %     97.5 %
## (Intercept)   8.0362997  9.5877003
## dosislav     -4.6281919 -2.6658081
## dosismellem1 -2.3171919 -0.3548081
## dosismellem2 -3.1501919 -1.1878081
## stedB        -0.1858074  1.2018074
```

Opgave 4

Spørgsmål 4.1

Der er tale om et homogenitetstest. Formelt har vi modellen for to uafhængige binomalfordelinger med antalsparametre 223 hhv. 446 og sandsynlighedsparametre p_1 og p_2 (eller p_{11} og p_{21}).

Hypotesen er $H_0 : p_1 = p_2$, og den testes med at χ^2-test.

Med *chisq.test* får vi $X^2 = 4.89$ og p-værdien 0.027 (og 0.035 hvis vi laver kontinuitetskorrektion). Vi afviser således hypotesen og konkluderer at sandsynligheden for at have været eksponeret er større hvis man har har sygdommen (estimat $54/223 = 0.242$) end hvis man ikke har sygdommen (estimat $76/446 = 0.170$).

```
A <- matrix(c(54,76,169,370),2,2)
A
```

```
##      [,1] [,2]
## [1,]   54  169
## [2,]   76  370
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  A
## X-squared = 4.8884, df = 1, p-value = 0.02704
```

```r
chisq.test(A, correct=TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  A
## X-squared = 4.4408, df = 1, p-value = 0.03509
```

Spørgsmål 4.2

Vi identificerer først de givne sandsynligheder:

$$P(\text{eksponeret} \mid \text{syg}) = 0.242, \quad P(\text{eksponeret} \mid \text{ikke syg}) = 0.170, \quad P(\text{syg}) = 0.005, \quad P(\text{ikke syg}) = 0.995$$

Vi bruger loven om total sandsynlighed og får

$$\begin{aligned} P(\text{eksponeret}) &= P(\text{eksponeret} \mid \text{syg}) \cdot P(\text{syg}) + P(\text{eksponeret} \mid \text{ikke syg}) \cdot P(\text{ikke syg}) \\ &= 0.242 \cdot 0.005 + 0.170 \cdot 0.995 \\ &= 0.17036 \end{aligned}$$

Vi bruger derefter Bayes' lov og får

$$P(\text{syg} \mid \text{eksponeret}) = \frac{P(\text{eksponeret} \mid \text{syg}) \cdot P(\text{syg})}{P(\text{eksponeret})} = \frac{0.242 \cdot 0.005}{0.17036} = 0.0071$$

Sygdommen er altså sjælden, selv hvis man har været eksponeret for ingrediensen.

```r
0.242 *0.005 + 0.170 *0.995
```

```
## [1] 0.17036
```

```r
0.242 *0.005 / 0.17036
```

```
## [1] 0.007102606
```