Reeksamen i Statistisk Dataanalyse 1, 3. februar 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

Indlæsning af data

```
library(readxl)
data1 <- read_excel(path = "feb2021opg1.xlsx")
head(data1)

## # A tibble: 6 x 4
##   region      kommune_id   dec   jan
##   <chr>            <dbl> <dbl> <dbl>
## 1 Syddanmark         580  5.68  3.57
## 2 Nordjylland        851 11.2   4.28
## 3 Midtjylland        751 17.3   5.23
## 4 Syddanmark         492  1.34  2.68
## 5 Hovedstaden        165 27.7   9.74
## 6 Hovedstaden        201 15.6   5.27
```

1. Der er tale om parrede målinger af incidensen af smittede i to forskellige måneder for de *samme* kommuner. Testet kan udføres som et parret t-test, hvorved man undersøger om den forventede ændring i incidensen kan antages at være lig 0.

```
t.test(data1$dec, data1$jan, paired = TRUE)

##
##  Paired t-test
##
## data:  data1$dec and data1$jan
## t = 13.834, df = 97, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.624328 8.843427
## sample estimates:
## mean of the differences
##                7.733878
```

Vi forkaster hypotesen og konkluderer, at ændringen i incidensen er signifkant forskellig fra nul ($t = 13.834, P < 0.0001$).

2. Vi fitter den statistiske model

$$\texttt{fald}_i = \alpha + \beta \cdot \texttt{dec}_i + e_i,$$

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

```
data1$fald <- data1$dec - data1$jan
linreg <- lm(fald ~ dec, data = data1)
summary(linreg)
```

```
##
## Call:
## lm(formula = fald ~ dec, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.218 -0.736  0.127  1.076  2.863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.47123    0.29622  -4.967 2.96e-06 ***
## dec          0.72060    0.02009  35.877  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 96 degrees of freedom
## Multiple R-squared:  0.9306,Adjusted R-squared:  0.9299
## F-statistic:  1287 on 1 and 96 DF,  p-value: < 2.2e-16
```

Vi aflæser estimatet for interceptet til $\hat{\alpha} = -1.471$ og regressionsparameteren til $\hat{\beta} = 0.721$. Estimatet for residualspredningen er $\hat{\sigma} = 1.466$.

3. Regressionsparameteren β beskriver sammenhængen mellem faldet i antal nye smittede og antallet af nye smittetilfælde fra december. Man kan derfor undersøge hypotesen $\beta = 0$.

   Fra `summary(linreg)` finder vi t-teststørrelsen ($t = 35.9$), og vi konkluderer, at hypotesen forkastes ($P < 0.0001$). Vi ser desuden, at estimatet $\hat{\beta}$ er positivt. Derfor vil faldet i antallet af nye smittetilfælde fra december til januar være størst i kommuner, hvor der var mange smittetilfælde i december.

En alternativ løsning til opgaven består i at udregne konfidensintervallet for β.

```
confint(linreg)

##                  2.5 %     97.5 %
## (Intercept) -2.0592261 -0.8832277
## dec          0.6807329  0.7604714
```

Da et 95 % - konfidensinterval $[0.681 - 0.760]$ *ikke* indeholder værdien 0, så kan vi konkludere, at større smitteantal i december vil medføre et større forventet fald i smitteantallet fra december til januar.

4. Vi benytter `predict()` funktionen i R til at konstruere et 95 % - prædiktionsinterval.

```
newdata <- data.frame(dec = 10)
predict(linreg, newdata, interval = "p")

##        fit      lwr      upr
## 1 5.734795 2.808736 8.660854
```

For en tilfældigt valgt kommune med incidensen 10 i december, så vil *faldet* i incidensen fra december til januar med 95 % sandsynlighed være indholdt i intervallet $[2.81 - 8.66]$. Det forventede fald i incidensen vil være 5.73 tilfælde per 1000 indbyggere.

5. Den statistiske model er en såkaldt *blandet model*

$$\texttt{fald}_i = \alpha(\texttt{region}_i) + \beta \cdot \texttt{dec}_i + e_i,$$

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

En mulighed er at teste hypotesen om, at vi kan se bort fra variablen `region` i den blandede model. Dette kan testes ved et F-test

```
model1 <- lm(fald ~ region + dec, data = data1)
anova(linreg, model1)

## Analysis of Variance Table
##
## Model 1: fald ~ dec
## Model 2: fald ~ region + dec
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     96 206.20
## 2     92 185.01  4    21.187 2.6339 0.03907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

På et 5 % signifikansniveau vil man forkaste hypotesen ($F = 2.634, P = 0.039$) og konkludere, at vi ikke kan udelade `region` fra modellen.

Opgave 2

Vi indlæser først data (her fra filen `feb2021opg2.txt`)

```
data2 <- read.table(file = "feb2021opg2.txt", header = T)
```

1. Vi fitter en ensidet variansanalysemodel

```
ensidet <- lm(udbytte ~ variety, data = data2)
summary(ensidet)

##
## Call:
## lm(formula = udbytte ~ variety, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1147 -0.9467  0.0570  0.9396  3.4926
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.3588     0.7495  67.192  < 2e-16 ***
## varietyLami    3.5709     1.0599   3.369  0.00185 **
## varietyLofa   -2.4678     1.0599  -2.328  0.02580 *
## varietySalka   1.0540     1.0599   0.994  0.32683
## varietyZita    0.8166     1.0599   0.770  0.44619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.12 on 35 degrees of freedom
## Multiple R-squared:  0.4895,Adjusted R-squared:  0.4311
## F-statistic: 8.389 on 4 and 35 DF,  p-value: 7.429e-05
```

   Det er sorten `Blanding` der anvendes som referencegruppe i R-outputtet. Det forventede udbytte på et plot, hvor man anvender blandingssorten bliver således 50.36. Estimatet for residualspredningen aflæses til $\hat{\sigma} = 2.12$.

2. Vi ønsker at teste hypotesen om at den forventede værdi af udbyttet er den samme for alle fem sorter. Testet udføres som et F-test enten med drop1 eller ved at fitte en nulmodel svarende til hypotesen om, at der er samme forventede værdi for alle sorter.

```
nulmodel <- lm(udbytte ~ 1, data = data2)
anova(nulmodel, ensidet)

## Analysis of Variance Table
##
```

```
## Model 1: udbytte ~ 1
## Model 2: udbytte ~ variety
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     39 308.08
## 2     35 157.28  4     150.8 8.3895 7.429e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: Med en F-teststørrelse $F = 8.3895$ og en tilhørende P-værdi < 0.0001 konkluderer vi, at der ikke kan antages at være samme forventede udbytte af alle fem sorter.

3. Vi reparametriserer/genfitter modellen med `variety = Salka` som reference. Herved kan vi ved at bruge `summary()` og `confint()` direkte aflæse et estimat og et 95 % - konfidensinterval for forskelle i det forventede udbytte.

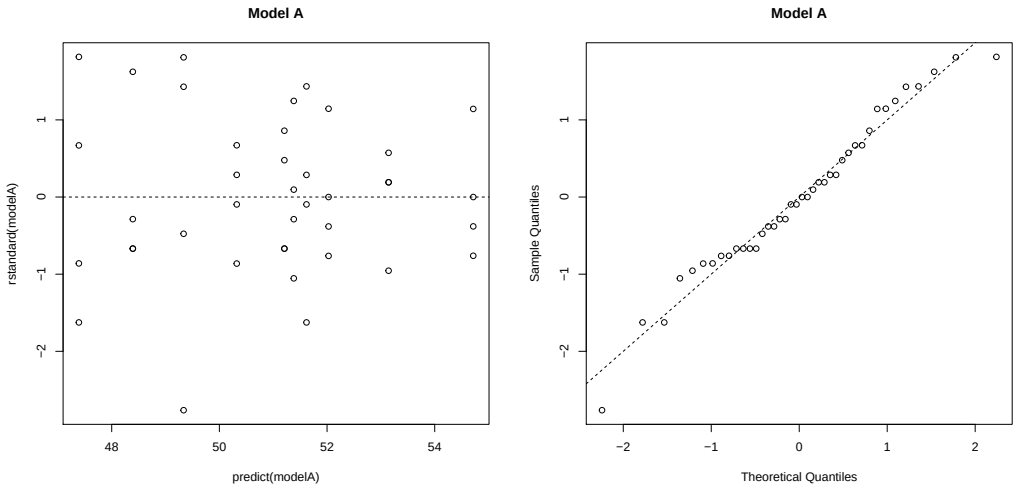
```
data2$varietyny <- relevel(factor(data2$variety), ref = "Salka")
ensidetny <- lm(udbytte ~ varietyny, data = data2)
## summary(ensidetny)

##                     Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       51.4128625  0.7494817 68.5978868 6.642349e-39
## varietynyBlanding -1.0540375  1.0599272 -0.9944433 3.268329e-01
## varietynyLami      2.5168375  1.0599272  2.3745380 2.318486e-02
## varietynyLofa     -3.5218125  1.0599272 -3.3226928 2.096645e-03
## varietynyZita     -0.2373875  1.0599272 -0.2239658 8.240861e-01

confint(ensidetny)

##                       2.5 %    97.5 %
## (Intercept)       49.8913337 52.934391
## varietynyBlanding -3.2058042  1.097729
## varietynyLami      0.3650708  4.668604
## varietynyLofa     -5.6735792 -1.370046
## varietynyZita     -2.3891542  1.914379
```

Vi aflæser den estimerede forskel til -0.237 med et 95 % - konfidensinterval $[(-2.389) - 1.914]$. Da konfidensintervallet for forskellen indeholder værdien nul, så er der *ikke* forskel på det forventede udbytte på plots der dyrkes med sorten `Zita` og på plots dyrket med `Salka`.

4. Vi ønsker at forklare værdien af en numerisk responsvariabel (`udbytte`) ud fra to kategoriske variable. Da vi har flere målinger (her 4 gentagelser) for hver kombination af `variety` og `bayleton`, så er det nærliggende at tage udgangspunkt i tosidet variansanalysemodel med vekselvirkning.

   For at kontrollere modellens antagelser laver vi et residualplot og et QQ-plot.

```r
modelA <- lm(udbytte ~ variety * bayleton, data = data2)
plot(predict(modelA), rstandard(modelA), main = "Model A")
abline(h = 0, lty = 2)
qqnorm(rstandard(modelA), main = "Model A")
abline(0, 1, lty = 2)
```



På residualplottet kigger vi efter om den lodrette variation i størrelsen af de standardiserede residualer ligger omkring 0 og er nogenlunde konstant uanset størrelsen af de prædikterede værdier. Dette lader til at være tilfældet (i hvert tilfælde ses ikke systematiske afvigelser).

På QQ-plottet ser vi efter, om punkterne ligger tilnærmelsesvist omkring et ret linje. Dette virker helt rimeligt her.

Vi ser på estimaterne fra den tosidede variansanalysemodel ...

```r
summary(modelA)
```

```
##
## Call:
## lm(formula = udbytte ~ variety * bayleton, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0915 -1.2318 -0.0881  1.2339  3.3472
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               51.3820     1.0638  48.301   <2e-16 ***
## varietyLami                1.7613     1.5044   1.171   0.2509
## varietyLofa               -3.9920     1.5044  -2.654   0.0126 *
## varietySalka               0.2361     1.5044   0.157   0.8763
## varietyZita               -1.0594     1.5044  -0.704   0.4867
## bayletonNej               -2.0464     1.5044  -1.360   0.1839
```

```
## varietyLami:bayletonNej    3.6191     2.1276   1.701   0.0993 .
## varietyLofa:bayletonNej    3.0485     2.1276   1.433   0.1622
## varietySalka:bayletonNej   1.6359     2.1276   0.769   0.4480
## varietyZita:bayletonNej    3.7522     2.1276   1.764   0.0880 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.128 on 30 degrees of freedom
## Multiple R-squared:  0.5592,Adjusted R-squared:  0.427
## F-statistic: 4.229 on 9 and 30 DF,  p-value: 0.00131
```

Estimatet for det forventede udbytte på et plot dyrket med sorten `Lofa`, hvor der `ikke` blev sprøjtet med bayleton bliver

$$51.38 - 3.99 - 2.05 + 3.05 = 48.39.$$

Bemærk at estimatet kan udtrækkes direkte, hvis man genfitter modellen med følgende R-kode

```
modelAny <- lm(udbytte ~ variety:bayleton - 1, data = data2)
```

5. Vi udfører et test for, om der er vekselvirkning

```
modelB <- lm(udbytte ~ variety + bayleton, data = data2)
anova(modelB, modelA)

## Analysis of Variance Table
##
## Model 1: udbytte ~ variety + bayleton
## Model 2: udbytte ~ variety * bayleton
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     34 155.95
## 2     30 135.79  4    20.157 1.1132 0.3686
```

Konklusion: Vi kan ikke forkaste hypotesen om, at der *ikke* er vekselvirkning ($F = 1.113, P = 0.369$). Dette betyder, at effekten af sprøjtning med bayleton på det forventede udbytte *ikke* afhænger af, hvilken sort der dyrkes på plottet / jordlodden.

Opgave 3

3.1 Korrekt svar B.

```
pnorm(35, mean = 31.0, sd = 5.2) - pnorm(25, mean = 31.0, sd = 5.2)

## [1] 0.6548402
```

3.2 Korrekt svar D. Løses ved at bestemme en passende fraktil i den relevante normalfordeling.

```
qnorm(0.95, mean = 31.0, sd = 5.2)

## [1] 39.55324
```

3.3 Korrekt svar D. Benyt formlen for et 95 % - konfidensinterval for en stikprøve med gennemsnit 10158 og spredning 2736.

```
10158 - qt(0.975, 63 - 1) * 2736 / sqrt(63) # lower

## [1] 9468.947

10158 + qt(0.975, 63 - 1) * 2736 / sqrt(63) # upper

## [1] 10847.05
```

3.4 Korrekt svar C. Testes kan udføres med brug af `prop.test()`.

```
prop.test(c(13, 20), c(59, 163))

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(13, 20) out of c(59, 163)
## X-squared = 2.5375, df = 1, p-value = 0.1112
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03104347  0.22632266
## sample estimates:
##    prop 1    prop 2
## 0.2203390 0.1226994
```

Nedenfor udføres testet med `chisq.test()`. Bemærk, at her skal tabellen indtastes, således at det er antallet på intensiv og antallet som ikke ligger på intensiv som fremgår.

```
##      [,1] [,2]
## [1,]   46  143
## [2,]   13   20
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabel1alt
## X-squared = 2.5375, df = 1, p-value = 0.1112
```

3.5 Korrekt svar E. Her testes en ensidet variansanalysemodel (med `dosis` inddraget som en faktor) imod modellen, hvor der er en lineær sammenhæng mellem `dosis` og `udbytte`.

3.6 Korrekt svar A. Antallet Y af *grønne* biler i stikprøven kan beskrives ved modellen $Y \sim \text{bin}(10, 0.2733)$. Derfor kan den ønskede sandsynlighed beregnes ud fra en binomialfordeling.

```
dbinom(5, 10, prob = 0.2733)

## [1] 0.07787078
```