Eksamen i Statistisk Dataanalyse 1, 6. november 2019

Anders Tolver

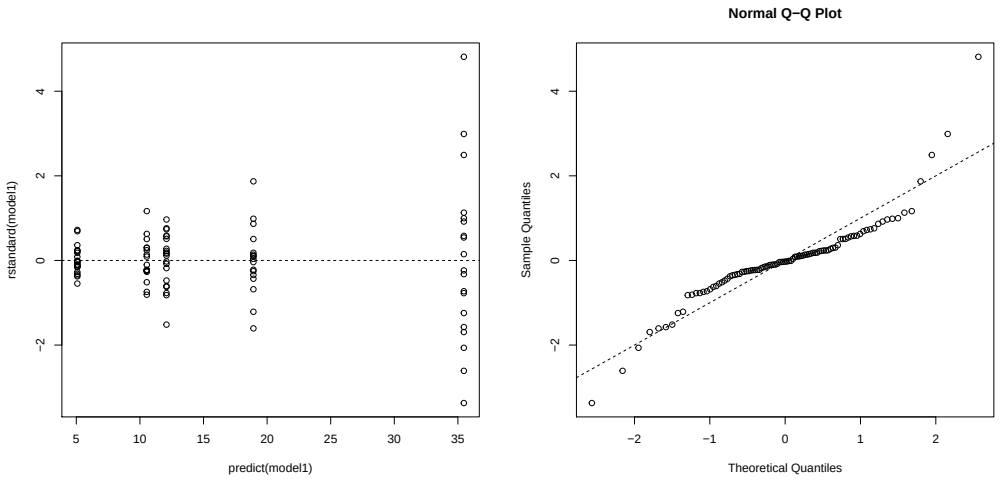
Vejledende besvarelse

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. Nedenfor ses residualplot og QQ-plots for `model1`

```
library(tidyverse)
hunde <- read.table(file = "hunde.txt", header = T)
hunde <- as_tibble(hunde)
model1 <- lm(maxLA ~ race, data = hunde)
plot(predict(model1), rstandard(model1))
abline(h = 0, lty = 2)
qqnorm(rstandard(model1))
abline(0, 1, lty = 2)
```



Vi bemærker at

- residualernes variation vokser med størrelsen af de prædikterede værdier
- det er således ikke rimeligt at antage, at variansen er ens inden for alle grupper/racer

- QQ-plottet for de standardiserede residualer afviger systematisk fra den rette linje, så residualerne er ikke normalfordelte (dette er dog mindre væsentligt, når der er allerede er problemer med varianshomogeniteten)

Det er derfor ikke at rimelig at benytte den ensidede variansanalysemodel fittet som `model1` til beskrivelse af målingerne af hjertevolumen.

2. Modellen kan fittes i R med følgende kode (her er datasættet indlæst som `hunde`).

```
lm(log(maxLA) ~ race, data = hunde)
```

Lader vi $\texttt{maxLA}_i$ betegne målingen af hjertevolumen for den i-te hund, og $\texttt{race}_i$ den tilhørende race, så kan modellen opskrives som

$$\log(\texttt{maxLA}_i) = \alpha_{\texttt{race}_i} + e_i,$$

hvor $e_1, \ldots, e_{97}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

```
summary(lm(log(maxLA) ~ race, data = hunde))

##
## Call:
## lm(formula = log(maxLA) ~ race, data = hunde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71922 -0.09587  0.00793  0.13221  0.49884
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.59549    0.04675   34.13   <2e-16 ***
## raceGrand_Danois  1.94684    0.07033   27.68   <2e-16 ***
## raceLabrador      1.32955    0.07260   18.31   <2e-16 ***
## racePetit_Basset  0.86945    0.06934   12.54   <2e-16 ***
## raceWhippet       0.74124    0.07260   10.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.229 on 92 degrees of freedom
## Multiple R-squared:  0.9007,Adjusted R-squared:  0.8964
## F-statistic: 208.7 on 4 and 92 DF,  p-value: < 2.2e-16
```

Residualspredningen estimeres til $\hat{\sigma} = 0.229$ og en forventede værdi af log-transformeret hjertevolumen for racen `Whippet` estimeres til $\hat{\alpha}_{\texttt{Whippet}} = 1.595 + 0.741 = 2.336$.

3. Vi ønsker at teste hypotesen om at den forventede værdi af logaritmen til hjertevolumen er den samme for alle 5 racer. 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 de 5 racer.

```
fuldmodel <- lm(log(maxLA) ~ race, data = hunde)
nulmodel <- lm(log(maxLA) ~ 1, data = hunde)
anova(nulmodel, fuldmodel)

## Analysis of Variance Table
##
## Model 1: log(maxLA) ~ 1
## Model 2: log(maxLA) ~ race
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     96 48.617
## 2     92  4.826  4    43.792 208.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: Vi finder et F-teststørrelsen $F = 208.72$ med en tilhørende p-værdi $p = 2.2e-16$. Der er således (ikke overraskende) forskel på den forventede værdi af logaritmen til hjertevolumen for de fem racer i datasættet.

4. Vi reparametriserer modellen med `race = Labrador` som reference gruppe, så vi direkte kan få et estimat for forskellen mellem de to ønskede racer med en tilhørende t-teststørrelse og en p-værdi for test af hypotesen om, at forskellen er lig nul.

```
hunde$race <- relevel(factor(hunde$race), ref = "Labrador")
fuldmodelny <- lm(log(maxLA) ~ race, data = hunde)
round(summary(fuldmodelny)$coef, digits = 4)

##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          2.9250     0.0555  52.6592        0
## raceBorder_Terrier  -1.3295     0.0726 -18.3130        0
## raceGrand_Danois     0.6173     0.0765   8.0734        0
## racePetit_Basset    -0.4601     0.0756  -6.0899        0
## raceWhippet         -0.5883     0.0786  -7.4891        0

confint(fuldmodelny)

##                         2.5 %     97.5 %
## (Intercept)         2.8147168  3.0353573
## raceBorder_Terrier -1.4737372 -1.1853530
## raceGrand_Danois    0.4654361  0.7691465
## racePetit_Basset   -0.6101493 -0.3100458
## raceWhippet        -0.7443183 -0.4322855
```

Vi aflæser den estimerede forskel mellem de forventede værdier af logaritmen til hjertevolumen (for `Petit_Basset` og `Labrador`) til -0.46010 med et tilhørende 95 % -konfidensinterval på [-0.610,-0.310].

5. Vi laver et 95 %- prædiktionsinterval for hjertevolumen af venstre forkammer for racen

`Labrador`. Den letteste løsning er at benytte `predict` funktionen hvorefter resultatet tilbagetransformeres til oprindelig (dvs. *ikke* log-skala).

```
newdata <- data.frame(race = "Labrador")
predict(fuldmodel, newdata, interval = "p")

##        fit      lwr      upr
## 1 2.925037 2.456988 3.393086

exp(predict(fuldmodel, newdata, interval = "p"))

##        fit      lwr      upr
## 1 18.63492 11.66961 29.75765
```

Vi aflæser prædiktionsintervallet (på oprindelig skala) til at være $[18.63, 29.76]$. En `Labrador` med et hjertevolumen på 32 mL ligger således uden for et 95 % - prædiktionsinterval. Det kunne tyde på, at hunden har et unormalt stort hjerte.

Opgave 2

Vi indlæser først data (her fra `training.txt`)

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

1. Datasættet består af parrede målinger (before / after) af konditionen for de samme forsøgspersoner. Vi udfører derfor et parret t-test med henblik på at se, om den gennemsnitlige ændring i konditionen kan antages at være 0. Testet kan fx. udføres med R kommandoen `t.test()`

   ```
   t.test(training$after, training$before, paired = T)

   ##
   ##  Paired t-test
   ##
   ## data:  training$after and training$before
   ## t = 14.085, df = 66, p-value < 2.2e-16
   ## alternative hypothesis: true difference in means is not equal to 0
   ## 95 percent confidence interval:
   ##  0.3249671 0.4323164
   ## sample estimates:
   ## mean of the differences
   ##               0.3786418
   ```

   T-teststørrelsen bliver 14.085 svarende til en p-værdi som er 0! Vi konkluderer, at der er sker en ændring i konditionen hen over træningsperioden.

   R-koden nedenfor svarer til to andre metoder, hvorpå man kan udføre samme test

```r
t.test(training$after - training$before)
summary(lm(after - before ~ 1, data = training))
```

2. På baggrund af R-outputtet ovenfor estimeres den gennemsnitlig ændring i konditionen til 0.379 med et 95 % - konfidensinterval på [0.325, 0.432]. Da konfidensintervallet for ændringen *ikke* indeholder værdien nul konkluderes, at træningsprogrammet i gennemsnit ændrer konditionen (der sker en stigning).

   Delopgaven kan også besvares ved at opfatte tilvæksterne i konditionen (`forskel`) som et enstikprøveproblem, hvor man ønsker at bestemme et 95%-konfidensinterval for populationsgennemsnittet.

3. Lader vi $\texttt{forskel}_i$ betegne ændringen i konditionen og lader vi $\texttt{sex}_i$ samt $\texttt{age}_i$ være køn og alder for forsøgsperson i, så udtrykker modellen at

$$\texttt{forskel}_i = \alpha_{\texttt{sex}_i} + \beta \cdot \texttt{age}_i + e_i,$$

   hvor $e_1, \ldots, e_{67}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$. Vi har på kurset kaldt dette for en *blandet model* (både kontinuerte og kategoriske forklarende variable).

```r
training$forskel <- training$after - training$before
model2 <- lm(forskel ~ sex + age, data = training)
summary(model2)$coef

##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  0.665019289 0.126464410  5.258549 1.788185e-06
## sexM         0.249397594 0.070207724  3.552281 7.230486e-04
## age         -0.006673213 0.002597188 -2.569399 1.252804e-02
```

   Modellens parametre estimeres som

$$\hat{\alpha}_{\texttt{F}} = 0.6650, \quad \hat{\alpha}_{\texttt{M}} = 0.6650 + 0.2494, \quad \hat{\beta} = -0.0067.$$

   Det anses som ok, selvom man ikke angiver estimatet for residualspredingen på $\hat{\sigma} = 0.1999$.

4. Hypotesen om at forsøgspersonens alder ikke har indflydelse på den forventede ændring kan udtrykkes ved $H_0 : \beta = 0$. Hypotesen kan testes ved et t-test (-se output ovenfor): $T_{\texttt{obs}} = -2.569$ med en tilhørende p-værdi på 0.012.

5. Den estimerede ændring for en mandlig forsøgsperson på 40 år bliver $\hat{\alpha}_{\texttt{M}} + \hat{\beta} \cdot 40 = 0.647$.

   Frivilligt: Ved brug af `predict()` (-se detaljer nedenfor) bestemmes et 95 % - konfidensinterval for den forventede ændring (`forskel`) til $[0.503, 0.792]$.

```r
newdata <- data.frame(sex = "M", age = 40)
predict(model2, newdata, interval = "c")

##         fit       lwr       upr
## 1 0.6474884 0.5030984 0.7918783
```

Opgave 3

3.1 Korrekt svar D.

```
pnorm(70, mean = 66.1, sd = 7.7) - pnorm(60, mean = 66.1, sd = 7.7)
```

```
## [1] 0.4796251
```

3.2 Korrekt svar A. Løses ved at bestemme 90 % - fraktilen i den relevante normalfordeling.

```
qnorm(0.9, mean = 66.1, sd = 7.7)
```

```
## [1] 75.96795
```

3.3 Korrekt svar C. Da et 95 % - konfidensinterval indeholder værdien 0, så kan vi ikke afvise hypotesen H_0 på et 5 % signifikansniveau. Den tilhørende p-værdi er over 5 %.

3.4 Korrekt svar B. Det eneste man direkte kan aflæse fra R-outputtet er, at den estimerede effekt (på den forventede tilvækst i `wi`) er 0.03, hvis der tilsættes 40 mg B_{12} uden at der tilsættes antibiotika. Denne effekt er ikke signifikant forskellig fra 0 ($p = 0.56$).

3.5 Korrekt svar B. Hvis Y er antallet af Biologi-Bioteknologistuderende i stikprøven, så antager vi at $Y \sim \texttt{bin}(10, 0.4)$. Vi beregner $P(Y \leq 5)$.

```
pbinom(5, size = 10, prob = 0.4)
```

```
## [1] 0.8337614
```

3.6 Korrekt svar C. Benyt den generelle formel $T_{\texttt{obs}} = \frac{\hat{\theta} - \theta_0}{\widehat{SE(\theta)}}$ for, hvordan man kan teste en hypotese af formen $H_0 : \theta = \theta_0$ for en parameter i en (lineær) model.

```
Tobs <- (0.8916347 - 1)/0.03172285
Tobs
```

```
## [1] -3.416001
```

3.7 Korrekt svar C. Benyt fx. et homogenitetstest til at sammenligne andelen af patienter som får infektion i hver af de to grupper. Husk at bruge `correct = F` for *ikke* at komme til at lave kontinuitetskorrektion. Pas på med at få indlæst antalstabellen korrekt i R: det er totalantal og ikke antal ikke-inficerede patienter, som er opgivet i opgaveteksten.

```
tab1 <- matrix(c(6,5,4,8), 2, 2)
tab1
```

```
##      [,1] [,2]
## [1,]    6    4
## [2,]    5    8
```

```r
chisq.test(tab1, correct = F)
```

```
## Warning in chisq.test(tab1, correct = F): Chi-squared approximation
may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 1.0508, df = 1, p-value = 0.3053
```

Hvis du i stedet benytter `prop.test()`, så er der mindre risiko for at taste data forkert ind

```r
prop.test(c(5,6), c(13, 10), correct = F)
```

```
## Warning in prop.test(c(5, 6), c(13, 10), correct = F): Chi-squared
approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(5, 6) out of c(13, 10)
## X-squared = 1.0508, df = 1, p-value = 0.3053
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.6180446  0.1872754
## sample estimates:
##    prop 1    prop 2
## 0.3846154 0.6000000
```