Eksamen i Statistisk Dataanalyse 1, 11. november 2020

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

Vi indlæser først data (her fra filen `nov2020opg1.txt`)

```
data1 <- read.table(file = "nov2020opg1.txt", header = T)
```

1. Den statistiske model fittes i R (her er datasættet indlæst som `data1`)

   ```
   model1 <- lm(log(transporttid) ~ studie, data = data1)
   ```

   Lader vi $\mathtt{y}_i$ betegne logaritmen til transporttiden for i-te studerende, og $\mathtt{studie}_i$ den tilhørende studieretning, så kan modellen opskrives som

   $$\mathtt{y}_i = \alpha_{\mathtt{studie}_i} + e_i,$$

   hvor $e_1, \ldots, e_{173}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

   ```
   summary(model1)

   ##
   ## Call:
   ## lm(formula = log(transporttid) ~ studie, data = data1)
   ##
   ## Residuals:
   ##      Min       1Q   Median       3Q      Max
   ## -2.93434 -0.34308  0.06139  0.47790  1.88636
   ##
   ## Coefficients:
   ##             Estimate Std. Error t value Pr(>|t|)
   ```

```
## (Intercept)   2.9343     0.1012  28.990  < 2e-16 ***
## studieHV      0.6327     0.1661   3.808 0.000196 ***
## studieJE     -0.3209     0.2077  -1.545 0.124224
## studieNR     -0.1934     0.1456  -1.328 0.186053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7905 on 169 degrees of freedom
## Multiple R-squared:  0.1478,Adjusted R-squared:  0.1326
## F-statistic: 9.768 on 3 and 169 DF,  p-value: 5.605e-06
```

Residualspredningen estimeres til $\hat{\sigma} = 0.7905$. Det er gruppen `studie = BB` der benyttes som reference gruppe (`Intercept`). Den forventede værdi af logaritmen til transporttiden estimeres til $\hat{\alpha}_{BB} = 2.9343$ og $\hat{\alpha}_{HV} = 2.9343 + 0.6327 = 3.5670$ (for husdyrvidenskabstuderende).

2. Vi ønsker at teste hypotesen om at den forventede værdi af logaritmen til transporttiden er den samme for alle 4 studieretninger. 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 4 studieretninger.

```
fuldmodel <- lm(log(transporttid) ~ studie, data = data1)
nulmodel <- lm(log(transporttid) ~ 1, data = data1)
anova(nulmodel, fuldmodel)

## Analysis of Variance Table
##
## Model 1: log(transporttid) ~ 1
## Model 2: log(transporttid) ~ studie
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    172 123.93
## 2    169 105.62  3    18.314 9.7678 5.605e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusion: Vi finder en F-teststørrelse $F = 9.768$ med en tilhørende p-værdi $p = 5.605e-06$. Der er således forskel på den forventede værdi af logaritmen til transporttiden for studerende på de fire studieretninger som optræder i datasættet.

3. Vi reparametriserer/genfitter modellen med `studie = HV` som reference. Herved kan vi ved at bruge `summary()` og `confint()` direkte aflæse et estimat og et 95 % - konfidensinterval for den forventede forskel i logaritmen til transporttiden.

   Man kan enten benytte R-koden

```
model1ny <- lm(log(transporttid) ~ relevel(factor(studie)
                                           , ref = "HV"), data = data1)
```

eller (for at få lidt pænere output)

```
data1$studie_ny <- relevel(factor(data1$studie), ref = "HV")
model1ny <- lm(log(transporttid) ~ studie_ny, data = data1)
## summary(model1ny)

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.5670340  0.1317574 27.072750 2.376806e-63
## studie_nyBB -0.6326961  0.1661483 -3.808021 1.957420e-04
## studie_nyJE -0.9535807  0.2241710 -4.253810 3.472579e-05
## studie_nyNR -0.8260620  0.1682980 -4.908329 2.149859e-06

confint(model1ny)

##                  2.5 %     97.5 %
## (Intercept)  3.3069317  3.8271362
## studie_nyBB -0.9606895 -0.3047028
## studie_nyJE -1.3961168 -0.5110447
## studie_nyNR -1.1582991 -0.4938248
```

Vi aflæser den estimerede forskel mellem de forventede værdier af logaritmen til transporttiden (for JE og HV) til -0.9536 med et tilhørende 95 % - konfidensinterval på [-1.396,-0.511]. **Bemærk:** Dette betyder at den forventede værdi af log-transporttid er 0.9536 højere form HV-studerende.

Hvis vi tilbagetransformerer resultatet til oprindelig skala ved brug af eksponentialfunktionen fås estimatet

```
exp(-0.9536)

## [1] 0.3853513
```

med et tilhørende 95 % - konfidensinterval

```
exp(-1.3961168) ### nedre graense

## [1] 0.2475564

exp(-0.5110447) ### oevre graense

## [1] 0.5998686
```

Fortolkningen af resultatet er, at medianen for transporttiden for jordbrugsøkonomistuderende estimeres til kun at være 38.5 % [95 %-KI: 24.8 - 60.0] af medianen for transporttiden for husdyrvidenskabstuderende. **Bemærk:** Man kan også (i overensstemmelse med opgaveformuleringen) vælgeat besvare opgaven med at sige, at medianen er $1/\exp(-0.9536) = 2.595$ gange højere [95 %-KI: 1.667-4.039] for HV-studerende.

4. Ved et kigge på `summary(model1ny)` ovenfor kan vi se forskellene på den forventede værdi af logaritmen til transporttiden mellem husdyrvidenskabstuderende og de øvrige tre grupper af studerende. Vi aflæser p-værdien for et t-test for, om transporttiden for hver af de øvrige grupper er forskellig fra gruppen af husdyrvidenskabstuderende:

$$\begin{aligned} &\texttt{BB vs HV} && (t = -3.808, P = 0.000196) \\ &\texttt{JE vs HV} && (t = -4.254, P = 3.47e-05) \\ &\texttt{NR vs HV} && (t = -4.908, P = 2.15e-06). \end{aligned}$$

   Det ses, at alle forskelle er statistiske signfikante ($P < 0.05$), og at transporttiden er længere for HV-studerende end for de øvrige grupper.

   Det er muligt at besvare spørgsmålet ved at tage udgangspunkt i modellen `model2` fra opgaveformuleringen i delopgave `1.4`

```
## summary(model2)

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   2.8093841 0.06776116 41.460094 3.890679e-91
## studie_hvTRUE 0.7576499 0.14854316  5.100537 8.925819e-07
```

   Heraf ses, at forskellen i den forventede værdi af logaritmen til transporttiden mellem husdyrvidenskabstuderende og de øvrige tre studieretninger estimeres til 0.758. Denne forskel er signfikant forskellig fra nul (her: > 0) ($t = 5.101, p < 0.00001$), så studerende på husdyrvidenskab har længere transporttid.

   Ikke påkrævet (og giver ikke fuldt point alene): Hvis man vælger den sidste fremgangsmåde, så er det naturlig først at teste en hypotese om, at transporttiden er ens for studerende på de øvrige 3 studieretninger, eller formuleret som hypotese i modellens parametre

$$\texttt{H0} : \alpha_{\texttt{BB}} = \alpha_{\texttt{JE}} = \alpha_{\texttt{NR}}.$$

   R-koden laver en ny variabel `studie_hv` med to niveauer / grupper, som holder styr på om observationen stammer fra en studerende på husdyrvidensskab eller ej. Derfor kan testet udføres som et F-test mod den ensidede variansanalyse model (`model1`), hvor der er en parameter/gruppemiddelværdi for hvert af de 4 studier. Vi får en F-teststørrelse på 1.559 med en tilhørende P-værdi på 0.2134. Der er således ikke forskel på transporttiden for studerende på de øvrige tre studieretninger (`BB, JE, NR`).

5. Der er tale om en såkaldt *blandet* model, hvor variationen i transporttiden forklares ud fra både `studie` og `alder`. Modellen kan skrives som

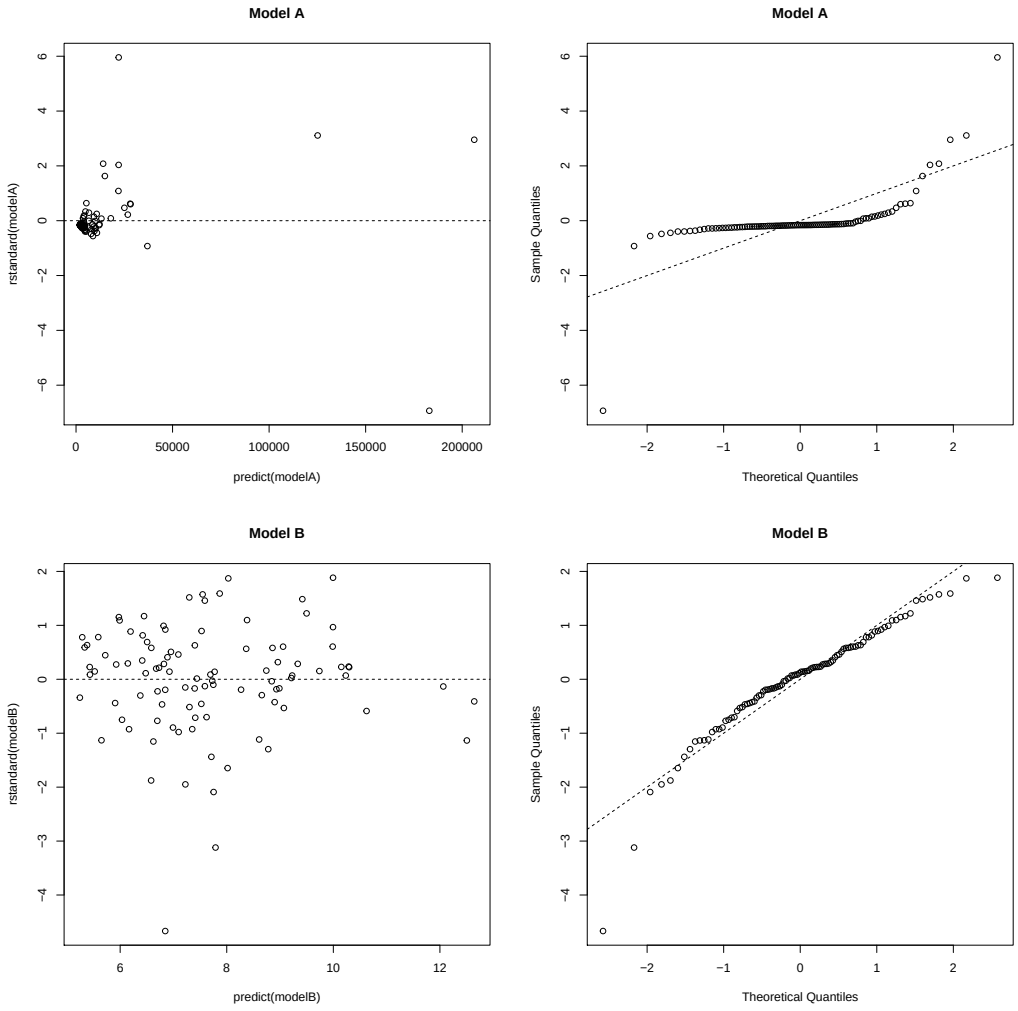
$$\log(\texttt{transporttid}_i) = \delta(\texttt{studie}_i) + \beta \cdot \texttt{alder}_i + e_i,$$

   hvor e_i'erne er uafhængige og normalfordelte $\sim N(0, \sigma^2)$. Parameteren β beskriver forskellen den forventede ændring af logaritmen til transporttiden for to studerende fra samme studie, som har en alderforskel på 1 år. Vi finder at $\hat{\beta} = -0.03769$. Hypotesen H_0 : $\beta = 0$ om at der ikke er sammenhæng mellem `alder` og `transporttid` kan ikke afvises ($t = -1.428, p = 0.155$).

Opgave 2

2.1 Vi fitter de to modeller i R og optegner både residualplot og QQ-plot.

```
library(tidyverse)
data2 <- read.table(file = "nov2020opg2.txt", header = T)
modelA <- lm(deaths ~ cases, data = data2)
modelB <- lm(log(deaths) ~ log(cases), 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)
plot(predict(modelB), rstandard(modelB), main = "Model B")
abline(h = 0, lty = 2)
qqnorm(rstandard(modelB), main = "Model B")
abline(0, 1, lty = 2)
```



På baggrund af figurerne for `Model A` bemærker vi at

- Residualernes variation vokser med størrelsen af de prædikterede værdier. Særligt

for observationer hørende til en prædikteret værdi på under 10.000 er der meget lille variation i de tilhørende residualer.

- Det er således ikke rimeligt at antage, at variansen (af residualerne) er uafhængig af de prædikterede værdier.
- QQ-plottet for de standardiserede residualer afviger systematisk fra den rette linje, så residualerne er ikke standardnormalfordelte med middelværdi 0 og varians 1 (dette er dog mindre væsentligt, når der er allerede er problemer med varianshomogeniteten).

På baggrund af figurerne for `Model B` bemærker vi at

- Residualernes variation virker til at være uafhængig af de prædikterede værdier. I hvert tilfælde i langt mere udpræget grad end på residualplottet hørende til `Model A`.
- Man kan eventuelt bemærke, at der er et par af de standardiserede residualer, som er meget store. Dette tyder for, at der er enkelte observationer som ikke er i særlig godt overensstemmelse med `Model B`.
- QQ-plottet for de standardiserede residualer lægger sig pænere op ad den rette linje, end det var tilfældet for `Model B`. Hvis man simulerer flere datasæt bestående af 173 perfekt normalfordelte variable, så ses måske en svag tendens til, at det tilhørende QQ-plot lægger sig tættere op ad en ret linje, end det er tilfældet for det rigtige datasæt.

På baggrund af kommentarer til residualplot og QQ-plot konkluderes, at `Model B` er mest velegnet til at beskrive sammenhængen mellem antal Corona tilfælde og antallet af dødsfald relateret til Corona.

2.2 R-koden til at fitte **Model B** er anført i svaret under delspørgsmål **2.1**. Vi benytter `confint()` til at bestemme konfidensintervallet.

```
confint(modelB)

##                 2.5 %    97.5 %
## (Intercept) -7.052799 -4.484091
## log(cases)   1.041017  1.259308
```

Et 95 % - konfidensinterval for parameteren β bliver $[1.041 - 1.259]$.

2.3 Et 95 % - prædiktionsinterval for antallet af dødsfald kan bestemmes ved brug af `predict()`-kommandoen. For at angive facit på den oprindelige skala, så skal prædiktionsintervallet tilbagetransformeres ved brug af eksponentialfunktionen.

```
newdata <- data.frame(cases = 10000)
pred1000 <- predict(modelB, newdata, interval = "p")
pred1000

##        fit      lwr      upr
## 1 4.824943 3.289999 6.359887
```

```
exp(pred1000)

##        fit      lwr      upr
## 1 124.5793 26.84283 578.1807
```

Antallet af dødsfald for et (nyt) lande med 10.000 Corona tilfælde vil med 95 % sandsynlighed falde inden for intervallet $[26.8 - 578.2]$.

2.4 Vi ønsker at teste hypotesen $H_0 : \beta = 1$. Da vores 95 % - konfidensinterval fra delspørgmsål 2.3 *ikke* indeholder værdi 1, så kan vi umiddelbart konkludere, at hypotesen forkastes. Alternativt kan vi bruge R-outputtet fra opgaveformuleringen til at konstruere en t-teststørrelse til test af hypotesen H_0. Vi får at

```
Tobs <- (1.150162 - 1) / 0.0549998
Tobs

## [1] 2.730228

pvalue <- 2 * (1 - pt(Tobs, df = 100 - 2))
pvalue

## [1] 0.007505098
```

Med en p-værdi på 0.0075 forkaster vi hypotesen om, at $\beta = 1$.

Opgave 3

3.1 Korrekt svar D.

```
pnorm(30, mean = 39.4, sd = 16.0)

## [1] 0.278434
```

3.2 Korrekt svar A. Løses ved at bestemme en passende fraktil i den relevante normalfordeling. Vær dog opmærksom på, at hvis vi skal være sikre på at 90 % kan koncentrere sig, så skal vi finde 10 % - fraktilen.

```
qnorm(0.10, mean = 39.4, sd = 16.0)

## [1] 18.89517
```

3.3 Korrekt svar E. Hvis Y betegner antallet blandt de 10 studerende, som *ikke* kan lide rosiner, så antager vi at $Y \sim \text{bin}(10, 0.647)$. Vi beregner $P(Y > 8) = P(Y \geq 9)$.

```r
1 - pbinom(8, 10, 0.647)

## [1] 0.08298565
```

3.4 Korrekt svar D. Vi har at gøre med to parrede stikprøver. Opgaven løses ved at betragte de 10 forskelle (`diff`) som en enkelt stikprøve, hvorefter vi tester hypotesen om at middelværdien i denne stikprøve er nul. Der er 10 observationer, hvorfor antallet af frihedsgrader der benyttes ved udregning af t-teststørrelsen bliver 10-1=9.

```r
tobs <- 0.3992268 / (1.621325 / sqrt(10))
tobs

## [1] 0.7786631

pvalue <- 2 * (1 - pt(tobs, df = 10 - 1))
pvalue

## [1] 0.4561619
```

3.5 Korrekt svar B. Da der er foretaget 3 målinger for hver dosis, så har vi mulighed for at vælge, om dosis bør indgår som en kategorisk variabel (via `grp`) eller som en kontinuert variable (via `dose`). I praksis svarer dette til at vi sammenligner en ensidet variansanalysemodel (`mod1`) med en lineær regressionsmodel (`mod2`) ved et F-test. Hypotesen om at sammenhængen mellem `udbytte` og `dose` kan beskrives ved en lineær funktion forkastes ($F = 13.112, P = 0.0008$).

3.6 Korrekt svar D. Antalstabellen er fremkommet ved, at man har inddelt de 173 efter to inddelingskriterier. Hverken rækkesummer eller søjlesummer er således kendt på forhånd. Testet er derfor et test for, om der er uafhængighed mellem de to inddelingskriterier. På baggrund af p-værdier (0.3048) konkluderer vi, at der ikke er sammenhæng mellem den foretrukne undervisningsform og villigheden til at inkludere andre studerende.

3.7 Korrekt svar D. Modellen er en tosidet variansanalysemodel med vekselvirkning. Estimatet under navnet `(Intercept)` svarer til gruppen, hvor konserveringsmidlet hverken tilføjes hos blomsterhandleren eller hos kunden (dvs. `hander = ikke-tilsat, kunde = ikke-tilsat`). Estimatet for gruppen, hvor konserveringsmidlet tilsættes både hos handler og kunde findes ved at lægge alle fire estimater sammen

```r
9.900 + 0.733 + 1.083 + 2.133

## [1] 13.849
```