Statistisk Dataanalyse 1

Helle Sørensen

Vejledende besvarelse, eksamen januar 2019

Dette er en vejledende besvarelse, incl. den R-kode jeg har benyttet 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 allerførst data fra en af filerne:

```
library(readxl)
ss_gr <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Jan2019/ss_gr.xlsx")
ss_gr2 <- read.table("~/Teaching/Courses/StatDat1/Eksamen/Jan2019/ss_gr.txt", header=TRUE)
```

Spørgsmål 1.1

Tid er kvantitativ og er den naturlige responsvariabel. De andre variable er begge kategoriske og skal benyttes som forklarende variable. Der er altså to kategoriske forklarende variable og en kvantitativ respons, hvilket præcis er sitationen for en tosidet ANOVA.

Hvis datasættet kaldes **ss_gr**, så er den relevante `lm`-kommando (hvor modelfittet er navngivet):

```
medVeksel <- lm(Tid ~ Studie + SidsteSudoko + Studie*SidsteSudoko, data=ss_gr)
```

Residualspredningen aflæses nederst i *summary* til 83.1 sekunder.

```
summary(medVeksel)
```

```
##
## Call:
## lm(formula = Tid ~ Studie + SidsteSudoko + Studie * SidsteSudoko,
##     data = ss_gr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -141.833  -70.333   -5.357   54.000  172.167
##
## Coefficients:
##                                         Estimate Std. Error t value
## (Intercept)                               151.50      23.99   6.315
## StudieMatematik                            35.86      28.67   1.251
## StudieMatOk                                10.41      34.69   0.300
## SidsteSudokoLaengeSiden                    86.20      35.58   2.423
## StudieMatematik:SidsteSudokoLaengeSiden   -28.72      45.70  -0.629
## StudieMatOk:SidsteSudokoLaengeSiden        27.75      53.67   0.517
##                                         Pr(>|t|)
## (Intercept)                              1.8e-08 ***
## StudieMatematik                           0.2150
## StudieMatOk                               0.7650
## SidsteSudokoLaengeSiden                   0.0179 *
```

```
## StudieMatematik:SidsteSudokoLaengeSiden   0.5316
## StudieMatOk:SidsteSudokoLaengeSiden       0.6067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.1 on 74 degrees of freedom
## Multiple R-squared:  0.1993, Adjusted R-squared:  0.1452
## F-statistic: 3.683 on 5 and 74 DF,  p-value: 0.004957
```

Spørgsmål 1.2

Hypotesen er at der *ikke* er vekselvirkning. Den testes fx ved at fitte modellerne med og uden vekselvirkning og sammenligne dem med et F-test. Vi får p-værdien 0.51, så hypotesen kan ikke afvises.

Der er altså ikke tegn på vekselvirkning, så effekten af at have løst sudokoer for nylig synes at være den samme uanset studieretningen: Eller omvendt: Forskelle mellem studier synes at være de samme uanset om man har løæst sudokoer for nylig eller ej.

```r
udenVeksel <- lm(Tid ~ Studie + SidsteSudoko, data=ss_gr)
anova(udenVeksel, medVeksel)
```

```
## Analysis of Variance Table
##
## Model 1: Tid ~ Studie + SidsteSudoko
## Model 2: Tid ~ Studie + SidsteSudoko + Studie * SidsteSudoko
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     76 520404
## 2     74 511009  2    9395.5 0.6803 0.5096
```

Spørgsmål 1.3

Estimat og konfidensinterval forskellen i forventet løsningstid mellem studerende der sidst har løst sudokoer for længe siden og studerende der har løst sudokoer for nylig, aflæses direkte fra *summary* og *confint* da en af parametrene i modellen netop er denne forskel (`ForNylig` er referencegruppe):

$$\text{estimat}: \ 79.44, \quad 95\% \text{ KI}: \ (40.74\,, 118.14)$$

Hvis vi tester hypotesen om at forskellen i forventede værdier er 0, fås p-værdien 0.00011, så forskellen er signifikant. Dette kunne alternativt konkluderes fra KI da nul ikke er indeholdt. Studerende der har løst sudokoer for nylig, er altså hurtigere end andre studerende.

```r
summary(udenVeksel)$coefficients
```

```
##                           Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              154.57287   19.72995 7.8344293 2.282648e-11
## StudieMatematik           26.19523   22.16872 1.1816303 2.410367e-01
## StudieMatOk               20.75614   26.33035 0.7882973 4.329748e-01
## SidsteSudokoLaengeSiden   79.43970   19.43287 4.0879031 1.070651e-04
```

```r
confint(udenVeksel)
```

```
##                               2.5 %    97.5 %
## (Intercept)              115.27727 193.86846
## StudieMatematik          -17.95760  70.34805
## StudieMatOk              -31.68529  73.19757
## SidsteSudokoLaengeSiden   40.73578 118.14362
```

Spørgsmål 1.4

`Aktuar` er referencegruppen, og det ses at man skal addere 26.20 hhv. 20.76 for de andre studier. Estimatet er således laves for de aktuarstuderende.

Estimatet for forskellen mellem matematik og matematik-økonomi fås som 26.20 - 20.76 = 5.44.

Der skal arbejdes lidt mere for at få konfidensintervallet. Det nemmeste er at skifte referencegruppe til `MatOK` eller `Matematik`. Så genfindes estimatet 5.44 og konfidensintervallet kan aflæses:

$$\text{estimat}: \ 5.44, \quad 95\% \text{ KI}: \ (-41.46\,,\,52.34)$$

```
ss_gr <- transform(ss_gr, Studie2 = relevel(factor(Studie), ref="MatOk"))
udenVeksel2 <- lm(Tid ~ Studie2 + SidsteSudoko, data=ss_gr)
summary(udenVeksel2)$coefficients
```

```
##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)             175.329007   20.91708  8.3820969 2.040223e-12
## Studie2Aktuar           -20.756142   26.33035 -0.7882973 4.329748e-01
## Studie2Matematik          5.439084   23.54957  0.2309632 8.179640e-01
## SidsteSudokoLaengeSiden  79.439697   19.43287  4.0879031 1.070651e-04
```

```
confint(udenVeksel2)
```

```
##                            2.5 %    97.5 %
## (Intercept)             133.66903 216.98899
## Studie2Aktuar           -73.19757  31.68529
## Studie2Matematik        -41.46394  52.34211
## SidsteSudokoLaengeSiden  40.73578 118.14362
```

Spørgsmål 1.5

Hypotesen er at den forventede løsningstid er den samme uanset studieretningen. Testet skal udføres som et samlet test i den tosidede ANOVA. Dette kan fx gøres med *drop1*. Vi får p-værdien 0.49, så der er ingen tegn på forskelle mellem studieretningerne.

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

```
## Single term deletions
##
## Model:
## Tid ~ Studie + SidsteSudoko
##              Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                    520404 710.43
## Studie        2      9780 530184 707.92  0.7141 0.4928879
## SidsteSudoko  1    114427 634831 724.33 16.7110 0.0001071 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Opgave 2

Vi indlæser først data fra en af filerne:

```
library(readxl)
florida <- read_excel("~/Teaching/Courses/StatDat1/Eksamen/Jan2019/florida.xlsx")
florida2 <- read.table("~/Teaching/Courses/StatDat1/Eksamen/Jan2019/florida.txt", header=TRUE)
```

Spørgsmål 2.1

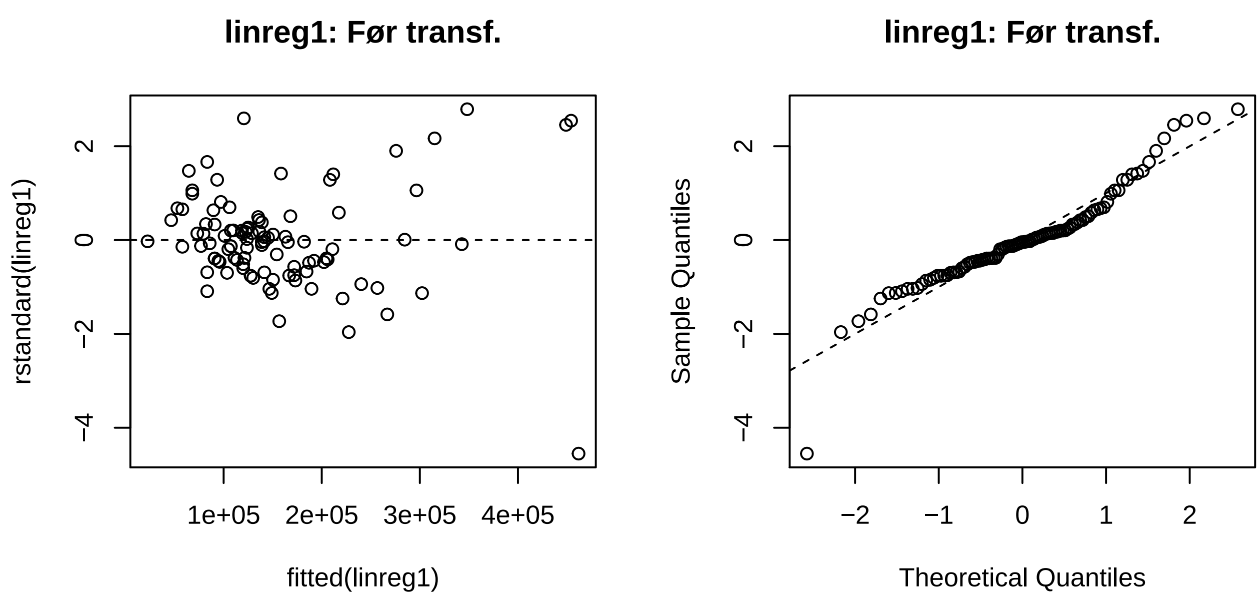
Vi vil gerne forklare salgsprisen ved hjælp af husets størrelse, så pris (eller en funktion af pris) skal benyttes som respons og størrelse (eller en funktion af størrelse) skal benyttes som forklarende variabel. Det er derfor kun **linreg1** og **linreg2** der er relevante.

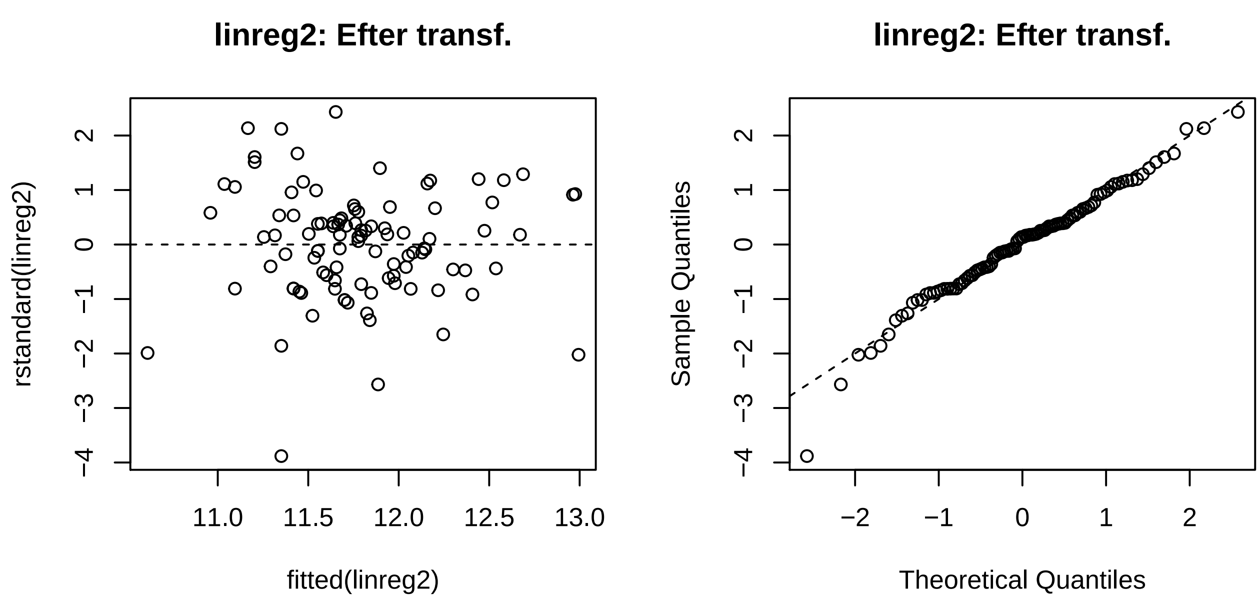
Nedenfor ses residualplot og QQ-plots for hver af de to modeller. Koden er ikke vist i output, men kan ses i Rmd-filen.

- Residualplottene: For *linreg1* (før transformation) ligger de standardiserede residualer ikke så symmetrisk om nul. Dette ser bedre ud for *linreg2* (efter transformation). For *linreg2* er der måske en antydning af at variationen er større for små end for store forventede værdier, men det ser nogenlunde OK ud.
- QQ-plottene ser nogenlunde fornuftige ud for begge modeller, men bedst for *linreg2*, hvor pyunkterne ligger pænt omkring den rette linie med skæring 0 og hældning 1.
- Ekstreme observationer: Det ser ud til at der for begge modeller er et enkelt hus der skiller sig ud og har et standardiseret residual på cirka -4 og altså er blevet solgt for billigt i forhold til hvad man ville forvente). Man kunne undersøge dette nærmere, men det er ikke en del af opgaven.

Samlet set er konklusionen at *linreg2* er den mest relevante og velegnede model til at beskrive data.

```
linreg1 <- lm(Price ~ Size, data=florida)
linreg2 <- lm(log(Price) ~ log(Size), data=florida)
linreg3 <- lm(Size ~ Price, data=florida)
linreg4 <- lm(log(Size) ~ log(Price), data=florida)
```





Spørgsmål 2.2

Den lineære regressionsmodel er

$$\log(\text{Price})_i = \alpha + \beta \cdot \log(\text{Size})_i + e_i, \quad i = 1, \dots, 15$$

hvor $e_1, \dots, e_{15}$ er iid $N(0, \sigma^2)$-fordelte. Hældningsparameteren hedder β, og estimat og KI er følgende:

$$\text{estimat}: \ 1.225, \quad 95\% \text{ KI}: \ (1.035\,,\, 1.416)$$

```
summary(linreg2)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.814491 0.70414483  3.997034 1.243282e-04
## log(Size)   1.225488 0.09598689 12.767241 1.443328e-22
```

```r
confint(linreg2)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.417138 4.211843
## log(Size)   1.035005 1.415971
```

Spørgsmål 2.3

For to huse (A og B) med et areal på hhv. 1000 square feet og 2000 square feet, har vi

$$\log(\text{Size}_B) - \log(\text{Size}_A) = \log(2000) - \log(1000) = \log(2) = 0.6931.$$

Forskellen i forventet log-pris er derfor

$$\hat{\beta} \cdot \log(2) = 1.225 \cdot 0.6931 = 0.849.$$

Dette er derfor også estimatet for logaritmen til forholdet mellem de to priser, da

$$\log\Big(\frac{\text{Price}_B}{\text{Price}_A}\Big) = \log(\text{Price}_B) - \log(\text{Price}_A) = 0.849,$$

og et fornuftigt estimat for forholdet mellem priserne er derfor $e^{0.849} = 2.337$. Vi estimerer altså at det store hus vil være en faktor 2.337 dyrere end det lille. (Dette gælder faktisk ved enhver fordobling fordi sammenhængen mellem de ikke-transformerede variable er en potenssammenhæng.)

Spørgsmål 2.4

Den prædikterede værdi for et hus på 3000 square feet er

$$\exp\big(\hat{\alpha} + \hat{\beta} \cdot \log(3000)\big) = e^{12.626} = 304429 \text{ USD}$$

For at undersøge om den givne salgspris på 215000 USD er usædvanlig, skal vi bruge et prædiktionsinterval.

Prædiktionsintervallet bliver (150839, 614413), så dette er intervallet som med sandsynlighed 95% vil indeholde den nye salgspris hvis huset er sammenligneligt med de andre huse. Værdien 215000 ligger i intervallet, så det er ikke en usædvanlig pris i forhold til de 100 huse i datasættet.

```r
newData <- data.frame(Size=3000)
predict(linreg2, newData, interval="p")
```

```
##       fit      lwr      upr
## 1 12.6262 11.92397 13.32842
```

```r
exp(predict(linreg2, newData, interval="p"))
```

```
##        fit    lwr      upr
## 1 304429.8 150839 614413.3
```

Spørgsmål 2.5

Sammenhængen mellem variablene i den fittede multiple regressionsmodel er givet ved følgende ligning:

$$\log(\text{Price}) = \hat{\alpha} + \hat{\beta}_1 \cdot \log(\text{Price}) + \hat{\beta}_2 \cdot \text{Baths} = 3.2177 + 1.153 \cdot \log(\text{Price}) + 0.0717 \cdot \text{Baths}.$$

At antallet af badeværelser ikke har effekt på salgsprisen, for fastholdt areal, svarer til at regressionskoefficienten β_2 er nul. Vi tester altså hypotesen $H_0 : \beta_2 = 0$. Fra *summary* aflæses p-værdien 0.38, så hypotesen kan ikke afvises. Der synes altså ikke at være en ekstra effekt af antal badeværelser, når der er taget højde for arealet af huset.

```r
multipel <- lm(log(Price) ~ log(Size) + Baths, data=florida)
summary(multipel)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 3.21774536 0.84160293 3.8233533 2.326798e-04
## log(Size)   1.15125075 0.12804905 8.9907010 2.024861e-14
## Baths       0.07177453 0.08181785 0.8772478 3.825196e-01
```

Opgave 3

Spørgsmål 3.1: B

Vi skal beregne $P(X = 3)$ i binomialfordelingen med antalsparameter 5 og sandsynlighedsparameter 1/3. Det viser sig at være 0.165.

```r
dbinom(3, size=5, prob=1/3)
```

```
## [1] 0.1646091
```

Spørgsmål 3.2: D

Vi skal finde antalsparameteren n i binomialfordelingen så sandsynligheden for at få gevinst på en eller flere kalendere er 0.95 eller større, dvs. at sandsynligheden for at der slet ikke er gevinst på nogle kalendere er mindre end 0.05, altså at $P(X = 0) < 0.05$.

Ved at prøve sig frem kan man se at betingelsen er opfyldt for $n = 8$, men ikke for $n = 7$.

```r
dbinom(0, size=7, prob=1/3)
```

```
## [1] 0.05852766
```

```r
dbinom(0, size=8, prob=1/3)
```

```
## [1] 0.03901844
```

Spørgsmål 3.3: B

Sandsynligheden for at en tilfældig person scorer mere end 130 er 0.023. Således optager Mensa personer hvis de er blandt de 2.3% (cirka 2%) af befolkningen der score højst i testen.

```r
pnorm(130, mean=100, sd=15)
```

```
## [1] 0.9772499
```

```r
1 - pnorm(130, mean=100, sd=15)
```

```
## [1] 0.02275013
```

Spørgsmål 3.4: A

Testet i tabellen udføres med *chisq.test* uden kontinuitetskorrektion. Dette giver p = 0.07. Der er således ikke signikant forskel mellem andelen af rygere de to år.

```r
A <- matrix(c(1106, 1158, 4018, 3859), 2,2)
A
```

```
##      [,1] [,2]
## [1,] 1106 4018
## [2,] 1158 3859
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  A
## X-squared = 3.2752, df = 1, p-value = 0.07033
```

Spørgsmål 3.5: C

Fejl af type 1 betyder at man *forkaster en sand hypotese*. I dette tilfælde er hypotesen at der ikke er forskel, og en fejl af type 1 er derfor den situation at man konkluderer at der *er* forskel, selvom der ikke er det.

Spørgsmål 3.6: C

Der er tale om en enkelt stikprøve med $n = 30$, $\bar{y} = 57.38$ og $s = 3.16$. konfidensintervallet beregnes til

$$\bar{y} \pm t_{0.975,29} \cdot \frac{s}{\sqrt{n}} = 57.38 \pm 2.045 \cdot \frac{3.16}{\sqrt{30}} = (56.20\,,\,58.56)$$

```r
qt(0.975, df=29)
```

```
## [1] 2.04523
```

```r
57.38 - qt(0.975, df=29) * 3.16 / sqrt(30)
```

```
## [1] 56.20004
```

```r
57.38 + qt(0.975, df=29) * 3.16 / sqrt(30)
```

```
## [1] 58.55996
```

Spørgsmål 3.7: D

Den lineære regression er

$$\text{rodlængde} = \alpha + \beta \cdot \text{antal uger} + e = \alpha + \beta \cdot \frac{7 \cdot \text{antal uger}}{7} + e = \alpha + \frac{\beta}{7} \cdot \text{antal dage} + e$$

hvor e'erne er normalfordelt med middelværdi 0 og spredning σ.

Ovenfor er β hældningen når tiden måles i uger, så vi ved at $\hat{\beta} = 2.95$. Vi kan se fra ligningen at hældningen i modellen med antal dage er $\beta/7$, så det nye estimat er $2.95/7 = 0.421$.

Til gengæld er restleddet det samme i begge tilfælde, så residualspredningen er uændret 5.13.

Spørgsmål 3.8: E

Der indgår kun mænd i forsøget, så det giver ikke mening at inddrage køn i analysen. Smertetærskel er en kategorisk variabel, så der er tale om variansanalyse snarere (ikke lineær regression). Der er tre niveauer af smertetærskel, altså mere end to grupper, så det er hverken to uafhængige eller to parrede stikprøver. Med andre ord: Ensidet variansanalyse.