Helle Sørensen, Anders Tolver (MATH, KU)
Efterår, 2022

Opgaver til Statistisk Dataanalyse 1

Opgave HS.24 (Quizzer)

Der blev stillet quizzer hver uge på Absalon som en del af afviklingen af Statistisk Dataanalyse 1 i 2018. Tabellen nedenfor viser hvor mange studerende der onsdag kl. 15.45 i kursusuge 5 havde svaret på quizzerne til uge 1–4.

Uge	1	2	3	4
Antal svar	93	53	29	20

1. Lav to variable, `uge` og `antal`, med tallene fra tabellen, og lav et punktplot/scatterplot af data. Overvej i den forbindelse hvad der er naturligt at have på x-aksen henholdsvis y-aksen.

2. Fit en lineær regressionsmodel til data, og indtegn den estimerede regressionslinie i grafen med `abline`). Hvad er fortolkningen af hældningskoefficienten?

3. Overvej kort om modellen passer godt til data. Du behøver ikke lave residualplot; du kan nøjes med at kigge på grafen fra spørgsmål 2.

4. Lav et nyt punktplot/scatterplot af data, hvor du bruger `uge` og `log(antal)` som variable. Du skal altså kun transformere den ene variabel.

   Fit den tilhørende lineære regressionsmodel, og indtegn den estimerede regressionslinie i grafen. Hvilken af de to modeller er bedst til data?

5. Gør rede for at regressionen fra spørgsmål 4 svarer til at der er ekponentiel sammenhæng mellem ugenummer og antal svar, altså at sammenhængen kan skrives på formen

$$\text{antal} \approx c \cdot e^{\beta \cdot uge} \tag{1}$$

   NB: Du skal ikke tænke nærmere over approksimationstegnet; det er blot indsat fordi der jo ikke præcist gælder lighedstegn i regressionsmodellen.

   Angiv estimater for c og β.

6. Forklar at sammenhængen i (1) medfører at forholdet mellem antal svar i to på hinanden følgende uger er e^{β}, altså at

$$\frac{\text{antal}''}{\text{antal}'} \approx e^{\beta}$$

   hvor antal$'$ og antal$''$ er antal svar i uge u henholdsvis $u+1$. Dette gælder uanset hvilke uger man kigger på, bare de ligger lige efter hinanden!

   Bestem også et estimat og et 95% konfidensintervaller for forholdet, altså for e^{β}.

7. Lav en tegning med følgende kommandoer, og forklar hvad grafen viser:

```
plot(antal ~ uge, xlim=c(0,5), ylim=c(0,150))
lm(log(antal) ~ uge)
f <- function(x) exp(5.0199) * exp(-0.5214*x)
plot(f, 0,5, add=TRUE)
```

Der er 149 studerende tilmeldt kurset. Prøv nedenstående kommando, og kommentér hvad du ser:

```
points(0,149, col="red")
```

Opgave HS.25 (Effekt af aske og kalk på plantevækst)

Denne opgave benytter data fra samme eksperiment som eksamen fra November 2016, opgave 2, men analysen er helt anderledes og opgaverne kan regnes uafhængigt af hinanden.

I et forsøg undersøgte man effekten på plantevækst af gødskning med aske hhv. kalk på følgende måde: Tredive potter blev inddelt i to lige store grupper. Den ene gruppe blev behandlet med aske i tre forskellige mængder (3 hhv. 9 hhv. 30 t/ha), mens den anden gruppe blev behandlet med kalk i tre forskellige mængder så næringsindholdet skulle modsvare askebehandlingerne.

Data er tilgængelige i filerne `aske-kalk2.xlsx` og `aske-kalk2.txt`. Der er tre variable:

- `behandling` angiver om potten er behandet med aske eller kalk
- `maengde` angiver om mængden er lav (I), mellem (II) eller høj (III)
- `torvaegt` angiver tørvægten af plantermaterialet i potten

1. Forskerne er særligt interesserede i om effekten af gødningsmængden (I/II/III) på plantevæksten afhænger af om der gødes med aske eller kalk. Hvilken type model vil være velegnet til at undersøge dette? Svaret skal begrundes.
2. Fit den foreslåede model, først med `torvaegt` som responsvariabel og derefter med `sqrt(toervagt)` som responsvariabel. Udfør modelkontrol for begge modeller og gør kortfattet rede for at modellen med `sqrt(toervagt)` som responsvariabel er mest fornuftig. Besvarelsen skal indeholde skitser af de relevante figurer.

I de følgende spørgmål bruges `sqrt(toervagt)` som responsvariabel.

3. Undersøg med et hypotesetest om effekten af gødningsmængden på sqrt-tørvægten afhænger af om der gødes med aske eller kalk.
4. Angiv estimater for den forventede værdi af sqrt-tørvægt for følgende kombinationer af behandling og mængde:
   - aske, I
   - aske, III
   - kalk, I
   - kalk, III
5. Angiv estimat og 95% konfidensinterval for den forventede forskel i sqrt-tørvægt mellem mængde III og mængde I for askebehandlede potter.

   Angiv estimat og 95% konfidensinterval for den forventede forskel i sqrt-tørvægt mellem mængde III og mængde I for kalkbehandlede potter.

   Vink til konfidensintervallet i andet delspørgsmål: Der er flere måder at gøre det på, men det kan fx være nyttigt at bruge `relevel`. Hvis du har indlæst data fra Excel-filen skal du huske at gøre variablen til en faktor inden du bruger `relevel`; dette gøre med funktionen `factor`.

Opgave HS.26 (Jellyfish)

Dette er en omskrivning af opgave 8.10 fra bogen.

I et eksperiment langs Hawkesbury River i New South Wales (Australien) har man indsamlet data fra 46 gopler (jellyfish). Goplerne blev indsamlet to forskellige steder, Dangar Island og Salamander Bay, og længde og bredde blev målt for alle gopler. Data er vist på side 267 i bogen, og vi kan se at sammenhængen ser ud til at være nogenlunde lineær.

Data ligger som **jellyfish** i *isdals*-pakken.

1. Fit en model der svarer til figuren i bogen. Brug koden

   ```
   lm(Length ~ Width + Location, data=jellyfish)
   ```

2. Lav et summary af modellen, og forklar hvordan de enkelte estimater skal fortolkes.

3. Bestem et estimat for den forventede længde af en gople fra Dangar Island med en bredde på 14. Bestem også et estimat for den forventede længde af en gople fra Salamander Bay med en bredde på 14.

4. I figuren (Figure 8.12 fra Opgave 8.10 i lærebogen) ses det at gopler fra Salamander Bay typisk er større end gopler fra Dangar Island, men der er dog et overlap. Angiv et estimat og et 95% konfidensinterval for forskellen i forventet længde mellem to gopler med samme længde, men fra Dangar Island henholdsvis Salamander Bay. Er forskellen statistisk signifikant?

Opgave HS.27 (Styrke af hypotesetest)

Den statistiske styrke — eller blot styrken — ved et hypotesetest er sandsynligheden for at forkaste en falsk hypotese. Vi vil altså gerne have en høj styrke, og hvis styrken er høj, er sandsynligheden lille for at begå fejl af type II. Styrken afhænger af flere ting, og formålet med denne opgave er at få en fornemmelse for disse sammenhænge.

Vi betragter situationen med en enkelt normalfordelt stikprøve med n observationer. Vi antager således at vi har data $y_1, \ldots, y_n$ der alle kommer fra normalfordelingen med middelværdi δ og spredning σ. Middelværdien betegnes δ i denne opgave (i stedet for μ) pga. syntaksen i R-koden nedenfor.

Vi vil hele tiden interessere os for hypotesen $H_0 : \delta = 0$ (mod det sædvanlige alternativ $H_A : \delta \neq 0$).

Funktionen `power.t.test` kan beregne styrken i denne situation:

```
> power.t.test(n=25, delta=0.5, sd=1, sig.level=0.05,
               type="one.sample", strict=TRUE)

     One-sample t test power calculation

              n = 25
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.6697077
    alternative = two.sided
```

Outputtet fortæller os at hvis vi bruger signifikansniveau 0.05 og har 25 observationer fra $N(0.5, 1)$ — specielt er hypotesen altså falsk — så er der 67% sandsynlighed for at forkaste hypotesen og korrekt konkludere at δ ikke er lig 0.

1. Hvordan tror du styrken ændres hvis antallet af observationer øges: Bliver den større eller mindre? Tænk først, og prøv derefter at ændre `n` i ovenstående kommando, fx til 50.

2. Hvordan tror du styrken ændres hvis δ gøres større? Tænk først, og prøv derefter at ændre `delta` i ovenstående kommando, fx til 0.75.

   Vink: Husk at δ er den sande middelværdi, mens hypoteseværdien er 0, således at vi kan tænke på δ som et udtryk for „hvor falsk" hypotesen er.

3. Hvordan tror du styrken ændres hvis σ gøres større? Tænk først, og prøv derefter at ændre `sd` i ovenstående kommando, fx til 2.

4. Hvordan tror du styrken ændres hvis vi bruger et mindre signifikansniveaet når vi tester, altså fx kun afviser hypotesen hvor p-værdien er mindre 0.01. Tænk først, og prøv derefter at ændre `sig.level` i ovenstående kommando til 0.01.

5. Prøv at ændre `delta` til 0 i kommandoen ovenfor. Hvilken styrke, dvs. sandsynlighed for at forkaste hypotesen, får du så? Kunne du have gennemskuet på forhånd at du ville få netop denne sandsynlighed?

6. `power.t.test` kan også bruges til at beregne hvor mange observationer der skal til for at opnå en given styrke, hvis man kender δ, σ og signifikansniveauet.

   Prøv fx kommandoen

   ```
   power.t.test(power=0.8, delta=0.5, sd=1, sig.level=0.05,
                type="one.sample", strict=TRUE)
   ```

   Hvor mange observationer skal man bruge?

   Bemærk at man skal kende både δ og σ for at kunne bestemme stikprøvestørrelsen.