Nedenstående opgaver omhandler modelkontrol for lineære modeller, herunder

- hvilke figurer der indgår som standard som modelkontrol for en lineær model
- hvordan man grafisk undersøger antagelsen om varianshomogenitet
- hvordan man grafisk undersøger antagelsen om normalfordelte fejl
- hvordan man ser efter stærkt afvigende observationer (outliers)
- begreberne prædikteret værdi, residual og standardiseret residual
- hvordan man kan undersøge antagelsen om en retlinet sammenhæng med en kontinuert prædiktor (kovariat)
- hvordan modeller omregnes ved transformation af data

Øvelser til torsdag i uge 2

Opgave 2.A.

Løs opgave 5.1 i kompendiet (BMS) (-modelkontrol for den additive model). Datasættet `organic.txt` kan hentes via link fra ugeplan 2 i Absalon.

Opgave 2.B.

NB: Denne opgave giver et eksempel på, hvordan R kunne tænkes brugt i en eksamenslignende opgave!

Et forsøg havde til formål at undersøge effekten af dosis (i milligram) af et bestemt medikament på en persons reaktionstid (i millisekunder) på en given stimulus. Forsøget blev udført sådan at 15 personer blev randomiseret til en af fem doser (0.5, 1.0, 1.5, 2.0, 2.5) sådan at der var tre personer i hver dosisgruppe.

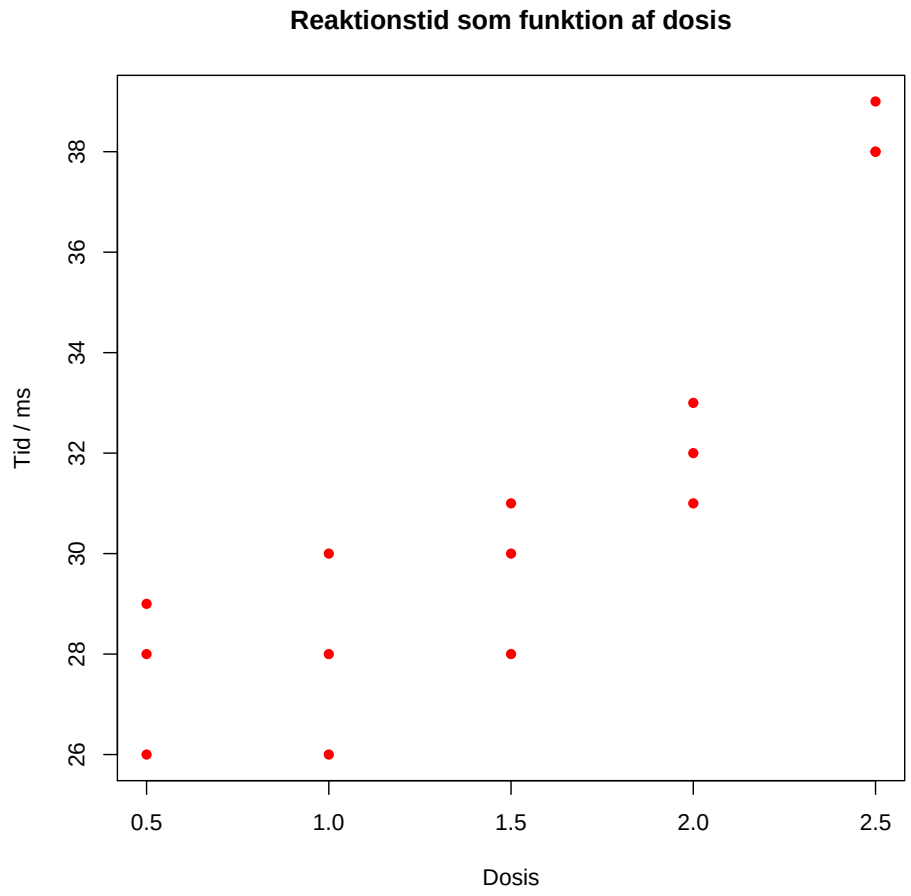
Data til denne opgave kan findes i tekstfilen `opg2B.txt`. Variablene `dosis` og `tid` angiver henholdsvis dosis og reaktionstid for personerne.

Start med indlæse data og plotte reaktionstiden som funktion af dosis. Du kan f.eks. benytte følgende R-kode.

```r
data<-read.table(file = "../data/opg2B.txt",header=T)
data
```

```
##    dosis tid
## 1    0.5  26
## 2    0.5  28
## 3    0.5  29
## 4    1.0  28
## 5    1.0  26
## 6    1.0  30
## 7    1.5  28
## 8    1.5  30
## 9    1.5  31
## 10   2.0  32
## 11   2.0  33
## 12   2.0  31
## 13   2.5  38
## 14   2.5  39
## 15   2.5  38
```

```r
plot(data$dosis, data$tid
     , xlab = "Dosis", ylab = "Tid / ms"
     , main = "Reaktionstid som funktion af dosis"
     , pch = 16, col = "red")
```



Besvar dernæst følgende 5 spørgsmål.

a) Følgende R-kode fitter to modeller til data

```
data$dosis2<-data$dosis*data$dosis
modelA<-lm(tid~dosis+dosis2,data=data)
modelB<-lm(tid~dosis,data=data)
```

Opskriv de statistiske modeller svarende til `modelA` og `modelB` i ovenstående R-kørsel. Beskriv (i ord) forskellen på de to modeller.

b) Lav en statistisk analyse med henblik på at undersøge, hvordan reaktionstiden afhænger af dosis. Angiv hvilken slutmodel du kommer frem til.

c) Angiv parameterestimater og konfidensintervaller for slutmodellen.

d) Prædiktér reaktionstiden for personer som har fået dosis 1.8. Gør det samme for dosis 4.0. Hvad er forudsætningerne for de to prædiktioner, og er de lige troværdige for de to tilfælde?

e) Hvis man ønsker at konstruere et 95 %-konfidensinterval for de prædikterede værdier fra spørgsmål **d)**, så kan man med fordel bruge funktionen `estimable` fra R-pakken `gmodels` (-pakken skal hentes ned på din computer først!). Prøv f.eks. at køre følgende R-kode og diskuter outputtet

```
library(gmodels)
est18<-c(1,1.8,1.8*1.8)
est4<-c(1,4,4*4)
est=rbind(est18,est4)
estimable(modelA,est,conf.int=0.95)
```

Ekstra opgave til hjemmebrug

Opgave 2.C: Box-Cox transformation og Poissonfordeling.

For at undersøge koncentrationen af nematoder (*feltiae*) i jord blev der udtaget 3 prøver (samples) som hver fortyndedes med vand i samme blandingsforhold. Fra hver opløsning udtoges der 15 del-prøver (sub-samples) på hver 40 $\mu\ell$ fra sample 1, mens de indeholdt hver 20 $\mu\ell$ fra sample 2 og 3. I hver del-prøve blev antallet af nematoder opgjort ved tælling under mikroskop. Følgende tabel viser antallet af nematoder fundet i hver af de 45 del-prøver.

Sample	Sub-sample size	The number of nematodes
1	40 $\mu\ell$	31 28 33 38 28 32 39 27 28 39 21 39 45 37 41
2	20 $\mu\ell$	14 16 18 9 21 21 14 12 13 13 14 20 24 15 24
3	20 $\mu\ell$	18 13 19 14 15 16 14 19 25 16 16 18 9 10 9

Data kan eventuelt hentes fra filen `feltiae.txt` ved at følge linket fra ugesiden. Alle de følgende spørgsmål vedrører gennemførelsen af en analyse af dette datamateriale med det mål at kunne estimere koncentrationen af nematoder i prøverne.

1) Opstil (altså opskriv!) modellen for ensidet variansanalyse for data.

2) Tegn residualplottet (ved brug af R) hvor du tegner standardiserede (Studentized) residualer op mod de prædikterede værdier. Hvad slutter du om forudsætningen om varianshomogenitet?

3) Brug Box-Cox metoden til at lave en figur svarende til kompendiets figur 5.9. Hvilken transformation af data er ud fra denne analyse optimal? Er kvadratrodstransformationen rimelig? (Kvadratrodstransformationen er begrundet i en forestilling om Poisson-fordelte antal).

Uanset svaret ovenfor skal du i resten af opgaven benytte kvadratrodstransformationen

$$Z = \sqrt{Y}$$

hvor Y er antal nematoder (count) i del-prøven.

4) Tegn residualplottet (som i spørgsmål 2) fra modellen for ensidet variansanalyse for de transformerede data og vurder det i forhold til residualplottet fra spørgsmål 2.

5) Benyt funktionen `estimable()` i R til at undersøge om der er signifikant forskel på sample 2 og 3.

Det forventede antal nematoder i en del-prøve på 40 $\mu\ell$ er $40 \cdot c$ hvor c er koncentrationen målt i antal pr. $\mu\ell$, og tilsvarende for en del-prøve på 20 $\mu\ell$. Derved er den forventede værdi af Z lig med $\sqrt{40}\sqrt{c}$ henholdsvis $\sqrt{20}\sqrt{c}$ i de to typer af prøver, hvilket svarer til modellen givet ved

$$\mathrm{E}Z = b\sqrt{\mathrm{size}}$$

med $b = \sqrt{c}$.

6) Tilpas denne model som en lineær model i R og test denne model mod modellen for ensidet variansanalyse. (Vejledning: du skal bruge en variabel `x= sqrt(size)`).

7) Beregn, stadig ved brug af R, et estimat og en øvre og nedre konfidensgrænse for parameteren b.

8) Benyt resultat fra spørgsmål 7 til at beregne et estimat og en øvre og nedre konfidensgrænse for koncentrationen af nematoder (c). (Aflæs eventuelt blot resultaterne fra spørgsmål 7, og indtast dem i R til brug for omregningen.)

(Opgaven er bygget over samme data og problemstilling som kompendiets Exercise 5.3.)

Hjælp til visse af spørgsmålene

Opg2.A,1 Hvis `model` er resultatet fra funktionen `lm(..)`, giver kommandoerne `predict(model)` og `resid(model)` henholdsvis prædikterede værdier og residualer. Standardiserede residualer fås med `stdres(model)` eller `rstandard(model)`. (Hvis du er på toppen af denne uges pensum, kan du også prøve en Box-Cox analyse.)

Opg2.B,a Modellen opskrives på formen $Y_i = \ldots$ med forklaring på hvad ingredienserne betyder. Sørg for at du ved hvad de enkelte led i modellen betyder og hvorfor de skrives som de gør.

Hjælp 2: Der findes et eksempel i kompendiet hvor de samme modeller benyttes.

Opg2.B,b Den helt korrekte analyse starter med en tredje model, der er mere general end begge modellerne `modelA` og `modelB`. Opskriv hypoteser svarende til mulige modelreduktioner. Bemærk hvad der er forudsætning og hvad der er konklusion i de enkelte trin.

Opg2.B,d Her skal du indsætte estimaterne i den ligning du opskrev i svaret på spørgsmål (b).