Modelkontrol
Statistisk Dataanalyse 2

Anders Tolver

Dagens program

- Hvorfor skal man lave modelkontrol?
- Modelkontrol af en lineær model
  - middelværdistrukturen (systematisk del)
  - normalfordeling (tilfældig del)
  - varianshomogenitet (tilfældig del)
  - uafhængighed (tilfældig del)
- Værktøj til modelkontrol
  - prædikterede værdier og residualer
  - residual plot: varianshomogenitet?
  - Box-Cox (transformation af responsen)
  - qq-plot (normalfordelingsantagelse)
  - residual plot: hvordan skal kovariaten indgå?

Hvorfor modelkontrol?

Spørgsmål Verden er ikke en matematisk model, så hvorfor skal vi tro på resultaterne af en statistisk analyse?

Svar Hvis modelantagelserne er rimelige i forhold til data, kan man måske godt opnå indsigt om interessante sammenhænge

> *Modelvalidering af en statistisk model er jeres argument for, at der er mening med galskaben.*

Vil man benytte en statistisk model til at drage konklusioner, må antagelserne bag modellen efterprøves og diskuteres kritisk.

Hvis man er fremlægger sin model og forholder sig åbent og kritisk til modelantagelserne, så er det svært at argumentere imod, at man kan stole på konklusionerne af en statistisk analyse.

Systematisk del: eksempler

Lineær regression

$$\mathbb{E}Y_i = \alpha + \beta x_i$$

Er den retlinede sammenhæng rimelig?

Ensidet ANOVA

$$\mathbb{E}Y_i = \alpha(\texttt{beh}_i)$$

Ingen antagelser!

Additiv tosidet ANOVA

$$\mathbb{E}Y_i = \alpha(\texttt{beh}_i) + \beta(\texttt{blok}_i)$$

Behandling og blok virker additivt (-ingen vekselvirkning).

Systematisk del: grafisk kontrolmetode

Residual plot

- Plot residualer mod prædikterede værdier
  - prædikteret værdi = middelværdiestimat ($\hat{\mu}_i$)
  - residual = observation - prædikteret værdi ($Y_i - \hat{\mu}_i$)
- Plot residualer mod kovariater
  - kovariat (x_i) = kvantitativ forklarende variabel

For begge plot skal residualerne (y-aksen) udvise samme variation for alle værdier af variablen på x-aksen.

Hvis modellens antagelser er opfyldt, skal der ikke være nogen systematisk sammenhæng mellem de to variable.

Tilfældig del: uafhængighed

Er data fra noternes eksempel 3.2 (klorofylproduktion i vinterhvede) uafhængige?

54 46 44 53 62 65 74 68 76 69 84 74

Spørgsmålet giver ingen mening, da afhængighed skal ses i forhold til noget.

Uafhængighed er et spørgsmål om

- fælles fejlkilder, som kun vedrører dele af datamaterialet
- er der en glemt faktor? (-fx. `kuld`, `blok`, gentagne målinger på samme `person`)

Overvej forsøgsplanen!

Tilfældig del: normalfordelingsantagelsen

Typisk model på SD2

$$Y_i = \mu_i(\text{parametre}) + e_i$$

hvor $e_1, e_2, \ldots, e_N$ er uafhængige $\sim N(0, \sigma^2)$.

Residualerne estimerer $e_1, \ldots, e_N$

$$\hat{e}_i = Y_i - \hat{\mu}_i$$

Derfor undersøges om residualernes fordeling ligner en normalfordeling:

- Man kan lave et histogram over $\hat{e}_1, \ldots, \hat{e}_N$
- Man kan lave et qq-plot over $\hat{e}_1, \ldots, \hat{e}_N$

På histogrammet skal man kigge efter normalfordelingens karakteristiske "klokkeform", mens man på qq-plottet kan nøjes med at kigge efter en ret linje.

Tilfældig del: varianshomogenitet

Restleddende $e_1, \ldots, e_N$ antages at have samme spredning (σ), uanset prædiktorerne (fx. behandling).

Grafisk modelkontrol for varianshomogenitet via residualplot:

- estimer parametrene i den systematiske del af modellen og udregn prædikterede værdier, $\hat{\mu}_i$
- udregn residualerne $\hat{e}_i = Y_i - \hat{\mu}_i$
- udregn standardiserede residualer $r_i = \frac{\hat{e}_i}{sd(\hat{e}_i)}$
- plot standardiserede residualer mod prædikterede værdier

Punkterne på figuren bør udvise omtrent konstant variation i lodret retning (hen over x-aksen). Da spredningen på de standardiserede residualer, r_i, er 1, bør de fleste punkter ($\approx 95\%$) ligge i intervallet $[-2, 2]$.

Udregning af prædikterede værdier, residualer og standardiserede residualer for en lineær `model` udføres i `R` med kommandoerne:

`predict(model)`, `resid(model)` og `rstandard(model)`.

Eksempel: hjertevolumen for raske hunde

For 97 hunde har man målt arealet af venstre forkammer i hjertet (`maxLA`).

Hundene repræsenterer 5 forskellige racer (`race`), og man har desuden registreret hundenes vægt i kg (`wgt`).

Data er venligst stillet til rådighed af Miriam Höllmer.

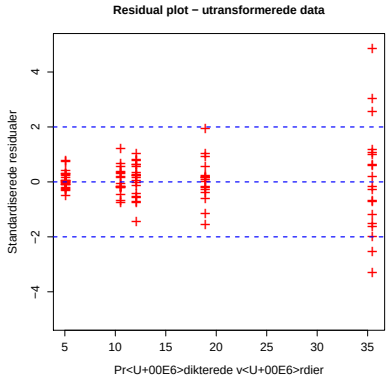
```
##             race  wgt maxLA
## 1 Border_Terrier  9.2  6.07
## 2 Border_Terrier  6.9  4.67
## 3 Border_Terrier  7.7  4.40
## 4 Border_Terrier 11.0  4.48
## 5 Border_Terrier  6.3  2.78
## 6 Border_Terrier  6.7  3.60
```

Spørgsmål som ønskes besvaret ved statistiske analyser

1. Hvordan afhænger `maxLA` af `race`? [ensidet ANOVA]
2. Hvordan afhænger `maxLA` af vægt (`wgt`)? [regression]

Eksempel: hjertevolumen for raske hunde

1-sidet ANOVA til analyse af sammenhæng ml. `maxLA` og `race`.

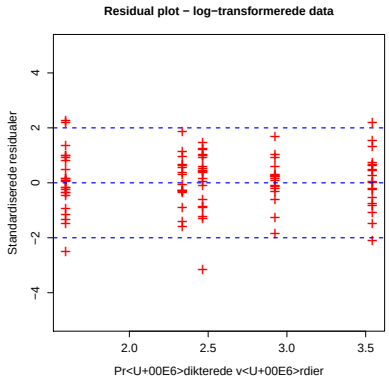


Varianshomogenitet?

Nej ... Øget variation for store prædikterede værdier.

Eksempel: hjertevolumen for raske hunde

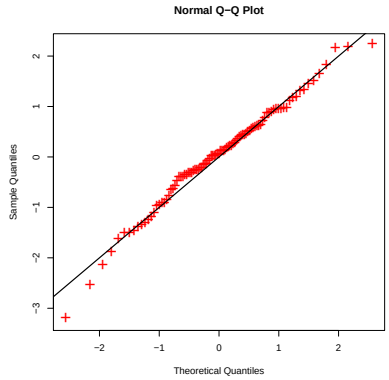
1-sidet ANOVA til analyse af smh. ml. log(`maxLA`) og `race`.



Varianshomogenitet?

Ja ... variation uafhængig af prædikterede værdier.

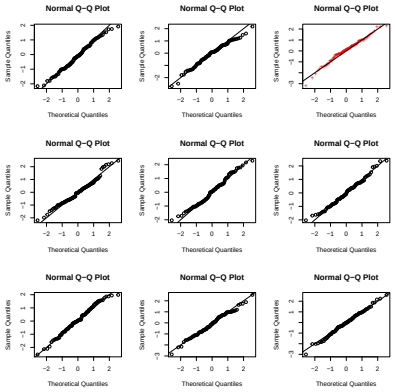
Eksempel: hjertevolumen for raske hunde



Residualer på qq-plot ligger omkring en ret linje.

Men hvor "pænt" bør det egentlig være for at være OK?

Eksempel: hjertevolume for raske hunde



Tegn eventuelt nogle qqplot for simulerede data fra normalfordeling med samme antal observationer (N=97):

`qqnorm(rnorm(N))`

Box-Cox transformationer

Box-Cox metoden er en automatiseret process til at finde en transformation af responsen, som er et kompromis mht. de systematiske del af modellen, varianshomogenitet og normalfordelingsantagelsen.

Konkret sammenlignes flg. transformationer (for varierende λ):

$$Z = \begin{cases} (Y^\lambda - 1)/\lambda & ,\lambda \neq 0 \\ \log(Y) & ,\lambda = 0 \end{cases}$$

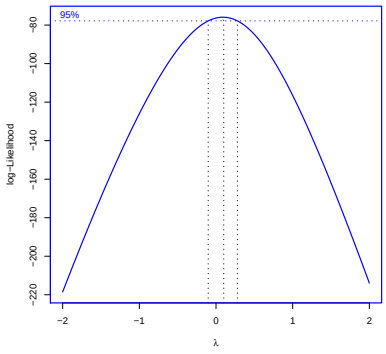
Box-Cox transformationer i R:

```
model=lm(maxLA~race,hunde)
library(MASS)
boxcox(model)
```

Specielt almindelige er:

$$\log(Y) \quad (\sim \lambda = 0) \text{ og } \sqrt{Y} \quad (\sim \lambda = 1/2)$$

Eksempel: hjertevolumen for raske hunde



Konf.int. for λ indeholder *ikke* 1, så responsen bør transformeres!

Et oplagt valg af transformation vil være $\log(Y)$ (-svarende til $\lambda = 0$), fordi tallet 0 ligger tæt på maksimum af Box-Cox plottet.

Eksempel: hjertevolumen for raske hunde

Vi ønsker at undersøge, hvordan maxLA afhænger af vægt (wgt).

Jf. slide 9-14 virker det rimeligt at bruge log(maxLA) som respons.

Forslag til statistiske modeller

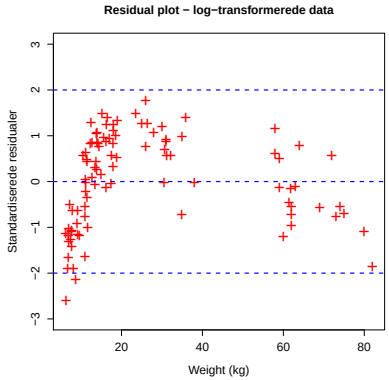
$$\begin{aligned} \log(\texttt{maxLA}_i) &= \alpha + \beta \cdot \texttt{wgt}_i + e_i \quad \text{lineær regression} \\ \log(\texttt{maxLA}_i) &= \alpha + \beta \cdot \texttt{wgt}_i + \gamma \cdot \texttt{wgt}_i^2 + e_i \quad \text{kvadratisk regression} \\ \log(\texttt{maxLA}_i) &= \alpha + \beta \cdot \log(\texttt{wgt}_i) + e_i \quad \text{log-log lineær regression} \end{aligned}$$

Stat. analyse bør baseres på model, hvor antagelserne er opfyldt.

Som før bør standardiserede residualer optegnes mod prædikterede værdier ligesom der bør laves et qq-plot for at checke normalfordelingsantagelsen (-ikke vist på følgende slides).

Desuden plottes residualerne mod de kontinuerte forklarende variable (=kovariaterne) i modellen og der undersøges for varianshomogenitet.

Eksempel: hjertevolumen for raske hunde

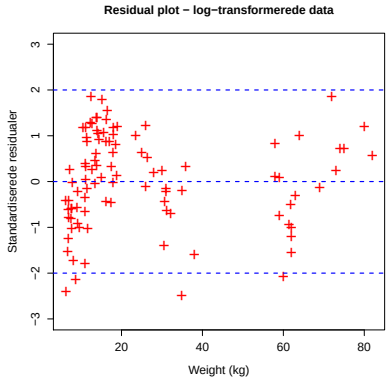


Varianshomogenitet? Nej ... afvigende mønster for modellen

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \texttt{wgt}_i + e_i$$

Model ikke velegnet som udgangspunkt for statistisk analyse!

Eksempel: hjertevolumen for raske hunde



Varianshomogenitet? Ser noget pænere ud for modellen

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \texttt{wgt}_i + \gamma \cdot \texttt{wgt}_i^2 + e_i$$

der således er mere velegnet som udgangspunkt for stat. analyse!

Eksempel: hjertevolumen for raske hunde

Hypotesen $H_0 : \gamma = 0$ i modellen for kvadratisk sammenhæng

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \texttt{wgt}_i + \gamma \cdot \texttt{wgt}_i^2 + e_i$$

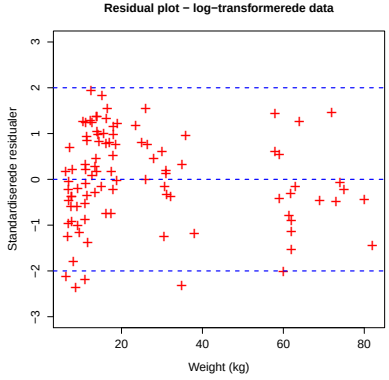
svarer til en lineær sammenhæng ml. $\log(\texttt{maxLA})$ og $\texttt{wgt}$.

```
modelkvad<-lm(log(maxLA)~wgt+I(wgt^2),hunde)
summary(modelkvad)$coef


##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  1.1964403377 8.148342e-02 14.683237 4.574875e-26
## wgt          0.0791836157 6.157811e-03 12.859052 1.965028e-22
## I(wgt^2)    -0.0006466304 7.701925e-05 -8.395699 4.689441e-13


modellin<-lm(log(maxLA)~wgt,hunde)
```

```
anova(modelkvad,modellin)
```

```
##   Res.Df       RSS Df Sum.of.Sq        F       Pr..F.
## 1     94  6.922573 NA        NA       NA           NA
## 2     95 12.113601 -1 -5.191028 70.48776 4.689441e-13
```


Eksempel: hjertevolumen for raske hunde



Varianshomogenitet? Ser også fornuftigt ud for modellen

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \log(\texttt{wgt}_i) + e_i$$

Eksempel: hjertevolumen for raske hunde

Kvadratisk regression givet ved modellen

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \texttt{wgt}_i + \gamma \cdot \texttt{wgt}_i^2 + e_i$$

svarer ved tilbageregning til følgende sammenhæng ml. `wgt` og medianen(!) af `maxLA`

$$\text{median}(\texttt{maxLA}) = \exp(\alpha + \beta \cdot \texttt{wgt} + \gamma \cdot \texttt{wgt}^2)$$

Regression på log-log-skala givet ved modellen

$$\log(\texttt{maxLA}_i) \quad = \quad \alpha + \beta \cdot \log(\texttt{wgt}_i) + e_i \qquad (1)$$

svarer ved tilbageregning til følgende sammenhæng ml. `wgt` og middelværdien af `maxLA`

$$\begin{aligned} \text{median}(\texttt{maxLA}) &= \exp(\alpha + \beta \cdot \log(\texttt{wgt})) \\ &= \delta \cdot \texttt{wgt}^{\beta} \quad \text{hvor } \delta = \exp(\alpha) \end{aligned}$$

Let(tere) fortolkning af parametre i modellen (1).

Example 5.6: flowering of arracacha

Nogle R-kommandoer:

```
arracacha<-read.table(file="../data/BMS_ex5_6.txt",header=T)
head(arracacha)
```

```
##   clone treat flowers
## 1    AP     1      18
## 2    BA     1       5
## 3     C     1      15
## 4    CE     1       8
## 5    CM     1       3
## 6     M     1      12
```

```
model1 <- lm(flowers ~ clone + treat,arracacha) ## the additive two-factor model.
pred <- predict(model1)
stres <- rstandard(model1)
```

```
plot(pred, stres,ylim=c(-3,3),main="Residualplot - utransformerede data"
,ylab="Standiserede residualer",xlab="Praedikterede vaerdier"
,cex.lab=1.5,cex.main=1.5,pch="+",col="red",cex=2)
abline(h=c(-2,0,2),col=4,lty=2) ## adds horizontal lines (visual guidance)
```

Example 5.6: flowering of arracacha (hjemme)

To faktorer

- 7 behandlinger
- 8 kloner

Respons

Antal blomstrende planter ud af 20 mulige - en observation per kombination af behandling og klon.

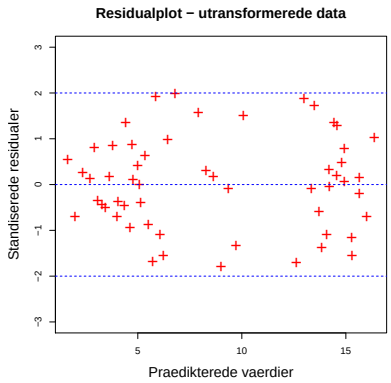
Statistisk model

$$Y_i = \alpha(\texttt{beh}_i) + \beta(\texttt{klon}_i) + e_i \quad e_i \text{ uafh. } \sim N(0, \sigma^2)$$

dvs. den additive model for blokforsøg.

Lad os undersøge om modelantagelserne er opfyldt.

Example 5.6: flowering of arracacha (hjemme)



Varianshomogenitet? Størst variation i midten! Dette skyldes (-til dels) forsøgsdesignet, hvor responsen er begrænset til intervallet fra 0 til 20.

Hvilken transformation kan råde bod på dette?

Example 5.6: flowering of arracacha (hjemme)

Respons

$$Y = \text{antal blomstrende planter ud af 20}$$

Nærliggende at modellere Y som binomialfordelt $b(20, p)$.

I så fald er

- middelværdien: $\mathbb{E}Y = np$
- variansen: $\mathbb{V}Y = np(1-p)$

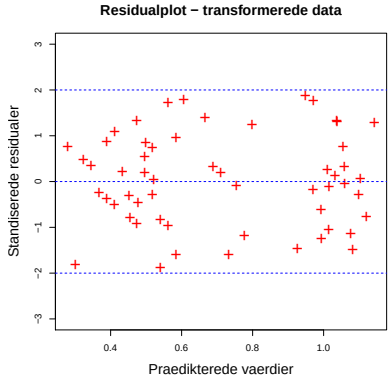
og standardtransformationen er

$$Z = \sin^{-1}\left(\sqrt{Y/20}\right)$$

Nogle R-kommandoer

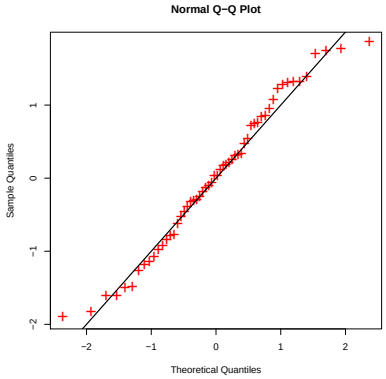
```
> z = asin(sqrt(flowers/20)) ## The transformed response
> modelz = lm(z ~ clone + treat)  ## the additive 2-factor model
```


Example 5.6: flowering of arracacha (hjemme)



Efter transformation er der varianshomogenitet.

Example 5.6: flowering of arracacha (hjemme)



De standardiserede residualer følger en normalfordeling.