Lineære modeller
Statistisk Dataanalyse 2

Anders Tolver

Program for denne uge

I dag: eksempel 4.2, afsnit 4.2, 4.3, dele af 4.4 og 4.5

- Intro: modeller og eksempel 4.2 (hydrolyse)
- Lineære modeller: definition, dimension, estimation, test af hypoteser
- konfidensintervaller ved brug af funktionen `estimable`
- Hydrolyseeksempel: R-program vil blive uploadet i Absalon

Torsdag: eksempel 4.1, dele af afsnit 4.4 og 4.5, samt kapitel 6.

Trykfejl i kompendium side 54 (midt): $\bar{\hat{\beta}} = \frac{1}{6}(\hat{\beta}(1) + \cdots + \hat{\beta}(6))$

Trykfejl i kompendium side 35 (nederst):

$$\mathrm{LSD}_T = 2.0369 \cdot 0.1441 \cdot \sqrt{2/(2 \cdot 6)} = 0.12$$

Intro til lineære modeller

Kender (mindst) to forskellige modeltyper:

- ANOVA (variansanalyse), dvs. faktormodeller
- lineær regression, dvs. modeller med kontinuert/numerisk/kvantitativ forklarende variabel

Klassen af lineære modeller indeholder begge typer samt blandinger af dem (og meget mere).

Kan håndtere manglende obs. og ufuldstændige designs.

Meget veludviklet (matematisk) teori.

Altså yderst anvendelig! Men har også begrænsninger:

- observationer skal være uafhængige
- observationer skal være normalfordelte

Eksempel 4.2: serin-indhold i foderprøver

```
data<-read.table(file="../data/hydrolysis.txt",header=T)
data$hourfac<-factor(data$hour)
data[1:12,]
```

```
##       feed hour serine hourfac
## 1  barley    8   4.47       8
## 2  barley   16   4.34      16
## 3  barley   24   4.22      24
## 4  barley   32   4.10      32
## 5  barley   72   3.48      72
## 6  barley    8   4.46       8
## 7  barley   16   4.30      16
## 8  barley   24   4.19      24
## 9  barley   32   4.08      32
## 10 barley   72   3.53      72
## 11   fish    8   4.23       8
## 12   fish   16   4.09      16
```

```
[  ... more data lines here  ...  ]
```

Eksempel 4.2: en hurtig ANOVA...

Målinger: serin-indholdet for to foderprøver for hver kombination af fem fodertyper [F] og fem hydrolysetider [T].

Vi vil bruge de log-transformerede serin-indhold som respons.

Observationer: $Y_1, Y_2, \ldots, Y_{50}$.

Tænk tilbage til de foregående to uger:

- Hvilken type forsøg er der tale om?
- Hvordan ville du analysere disse data?
- Hvad er konklusionen?

Definition af lineær model

Observationer $Y_1, \ldots, Y_N$.

En statistisk model for $Y_1, \ldots, Y_N$ kaldes lineær hvis

- $Y_1, \ldots, Y_N$ er uafh. og normalford. med samme spredning, σ
- Middelværdien af Y_i er en lineær funktion af parametrene:

$$\mathbb{E}Y_i = c_{i,1}\beta_1 + c_{i,2}\beta_2 + \cdots + c_{i,p}\beta_p$$

Her er

- $\beta_1, \ldots, \beta_p$ parametre, dvs. ukendte tal (som vi vil estimere)
- c'erne kendte tal som afhænger af designet

Skriver også:

$$Y_i = c_{i,1}\beta_1 + c_{i,2}\beta_2 + \cdots + c_{i,p}\beta_p + e_i, \quad e_i \text{ iid } N(0, \sigma^2)$$

Dimension af en lineær model

Lineære modeller kan parametriseres (skrives op) på mange måder!

For eksempel,

$$Y_i = \gamma(\mathtt{F_i}, \mathtt{T_i}) + \mathtt{e_i}$$
$$Y_i = \mu + \alpha(\mathtt{F_i}) + \beta(\mathtt{T_i}) + \gamma(\mathtt{F_i}, \mathtt{T_i}) + \mathtt{e_i}$$

Forskellige parametriseringer kan være hensigtsmæssige ved modelreduktion hhv. estimation.

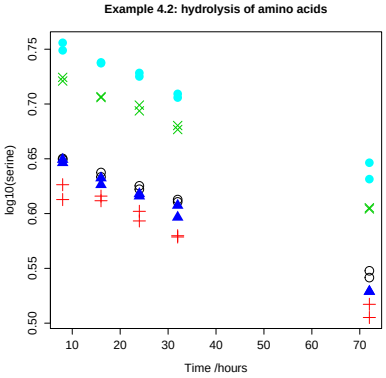
Dimensionen, d, af en given model er *det mindste antal parametre der skal til for at beskrive modellen*

Hvis der er flere parametre i modellen end nødvendigt, så siger vi at modellen er overparametriseret.

Eksempel 4.2: en ny indgangsvinkel

Lad os gentænke situationen...

- Kan vi bruge `hours` på en anden måde?
- Hvilke forskellige modeller kan vi tænke os for disse data?



Eksempel 4.2: diverse modeller

Model med vekselvirkning (1), uden vekselvirkning (2), med parallelle rette linier (3), forskellige hældninger (6):

$$\begin{aligned}
1: Y_i &= \gamma(\mathtt{F_i}, \mathtt{T_i}) + \mathtt{e_i} \\
2: Y_i &= \alpha(\mathtt{F_i}) + \beta(\mathtt{T_i}) + \mathtt{e_i} \\
3: Y_i &= \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i} \\
6: Y_i &= \alpha(\mathtt{F_i}) + \beta(\mathtt{F_i}) \cdot \mathtt{T_i} + \mathtt{e_i}
\end{aligned}$$

For hver af disse modeller:

- Hvilke parametre er der (i den givne parametrisering)?
- Hvad er middelværdien af Y_1 (barley, 8 hours)?
- Hvad er dimensionen af modellen?

Estimation af middelværdiparametre

Lineær model: $Y_i = c_{i,1}\beta_1 + c_{i,2}\beta_2 + \cdots + c_{i,p}\beta_p + e_i$

Parametre: $\beta_1, \beta_2, \ldots, \beta_p$ og σ.

Estimation: Bestem de (tal)værdier af parametrene der "*får modellen til passe bedst muligt med data*". Mindste kvadrater (least squares).

Estimater: $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ og $s = \hat{\sigma}$.

- De forventede værdier $\hat{\mu}_i = c_{i,1}\hat{\beta}_1 + c_{i,2}\hat{\beta}_2 + \cdots + c_{i,p}\hat{\beta}_p$ entydigt bestemt
- Hvis modellen ikke er overparametriseret, så er parameterestimaterne $\hat{\beta}_1, \ldots, \hat{\beta}_p$ også entydigt bestemt.
- Estimaterne har pæne egenskaber!

Estimation af spredningen

Residualer (observeret minus forventet):

$$\hat{e}_i = Y_i - \hat{\mu}(\hat{\beta}_1, \ldots, \hat{\beta}_p) = Y_i - \left(c_{i,1}\hat{\beta}_1 + c_{i,2}\hat{\beta}_2 + \cdots + c_{i,p}\hat{\beta}_p\right)$$

Residualkvadratsum

$$\mathrm{SS}_e = \sum_{i=1}^{N} \hat{e}_i^2 = \hat{e}_1^2 + \ldots + \hat{e}_N^2$$

Estimat for variansen på e_i:

$$s^2 = \hat{\sigma}^2 = \frac{\mathrm{SS}_e}{\mathrm{DF}_e} = \frac{\mathrm{SS}_e}{N-d}; \quad s = \sqrt{s^2}$$

Antal frihedsgrader, $\mathrm{DF}_e = N - d$: *antal obs. minus modeldimension.*

Analyse i R med lm

Eksempel 4.2: fit af modeller i R

```
model1 <- lm(log10(serine) ~ feed:hourfac,data=data)
model2 <- lm(log10(serine) ~ feed + hourfac,data=data)
model3 <- lm(log10(serine) ~ feed + hour,data=data)
model6 <- lm(log10(serine) ~ feed + hour + feed:hour,data=data)
```

Brug `summary`, `confint` og `anova` som sædvanlig.

Hvis modellen er overparametriseret sætter R visse parametre til 0 (referenceniveauer). Disse optræder så ikke i `summary`.

- Hvordan testes hypotesen om, at indholdet af `log(serine)` afhænger lineært af hydrolysetiden?
- Hvordan kan vi finde dimensionen af modellen i R?

Test af lineær hypotese

To lineære modeller (A) og (B). Antag at model (B) er indeholdt i model (A), dvs. at model (B) er et specialtilfælde af model (A).

Vil teste model (B) mod model (A).

Teststørrelse:

$$F_{AB} = \frac{\mathrm{MS}_{AB}}{\mathrm{MS}_e^A} = \frac{(\mathrm{SS}_e^B - \mathrm{SS}_e^A)/(\mathrm{DF}_e^B - \mathrm{DF}_e^A)}{\mathrm{SS}_e^A/\mathrm{DF}_e^A}$$

- Hvorfor er dette en rimelig teststørrelse?
- Er store eller små værdier kritiske?

F_{AB} er $F(\mathrm{DF}_e^B - \mathrm{DF}_e^A, \mathrm{DF}_e^A)$-fordelt under model (B), så F_{AB} skal vurderes i denne fordeling.

Eksempel 4.2: test for linearitet

Additiv faktormodel og model med parallelle rette linier:

$$2: Y_i = \alpha(\mathtt{F_i}) + \beta(\mathtt{T_i}) + \mathtt{e_i}$$
$$3: Y_i = \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i}$$

Test for linearitet (`model3` mod `model2`):

- Hvorfor er `model3` indeholdt i `model2`?
- Hvad er hypotesen, udtrykt ved parametrene?

Fra `deviance(model2)` og `summary(model2)` (-samt tilsvarende for `model3`) fås

$$F_{23} = \frac{(0.000802 - 0.000767)/(44 - 41)}{0.000767/41} = 0.624; p = 0.60.$$

Hvad er konklusionen omkring testet? Kunne vi have fået teststørrelse og p-værdi frem på andre måder?

Eksempel 4.2: flere test

Kunne altså ikke afvise modellen med lineær sammenhæng.

$$3: Y_i = \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i}$$

To naturlige hypoteser/modeller herfra:

$$4: Y_i = \alpha(F_i) + e_i$$
$$7: Y_i = \alpha + \beta \cdot T_i + e_i$$

Hvad udtrykker disse modeller (hvilke hypoteser svarer de til)?

- Test for model 4 mod model 3: $F_{34} = 4258$, $p = 0$
- Test for model 7 mod model 3: $F_{35} = 1654$, $p = 0$

Eksempel 4.2: konklusion

Model 3 er vores slutmodel: $Y_i = \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i}$.

- Effekt af hydrolysetiden? Forskel på fodertyperne?

Analysen kan sammenfattes i

- variansanalyseskema (tabel 4.5).
- tabel over estimater (tabel 4.6: NB: vi bruger 10-tals-logaritme som i kompendiet) og fx. konfidensintervaller

```
model3 <- lm(log10(serine) ~ feed + hour,data=data)
summary(model3)
```

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  0.666851638 1.581941e-03 421.540016 5.476254e-81
## feedfish    -0.028801269 1.909863e-03 -15.080279 5.374849e-19
## feedmais     0.068617527 1.909863e-03  35.927981 3.029453e-34
## feedmeat    -0.008050510 1.909863e-03  -4.215229 1.220410e-04
## feedsoy      0.099552215 1.909863e-03  52.125313 3.415175e-41
## hour        -0.001768393 2.710076e-05 -65.252521 1.973932e-45
```

```
Residual standard error: 0.004271 on 44 degrees of freedom
```

Hvordan skal vi besvare følgende spørgsmål?

1. Hvad er fortolkningen af estimatet med navn `hour`?
2. Estimér ændringen i log-serin indhold når hydrolysetiden øges med 10 timer. Konfidensinterval?
3. Hvad sker der med serin-indholdet (ikke log) når hydrolysetiden øges med 10 timer. Konfidensinterval?
4. Estimér forskellen i log-serin indhold mellem byg og fisk. Afhænger forskellen af hydrolysetiden? Konfidensinterval?
5. Estimér forskellen i log-serin indhold mellem majs og fisk. Konfidensinterval?
6. Estimér det forventede serin-indhold for byg ved en hydrolysetid på 16 timer. Konfidensinterval?
7. Hvordan kunne vi have testet for linearitet hvis den additive faktormodel (model 2) var blevet afvist?

Hydrolyseeksempel: konfidensintervaller (I)

Slutmodel

$$Y_i = \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i}, \quad \mathtt{e_i} \sim \mathtt{N}(0, \sigma^2).$$

```
model3 <- lm(log10(serine) ~ feed + hour,data=data)
confint(model3)

##                    2.5 %       97.5 %
## (Intercept)  0.663663444  0.670039831
## feedfish    -0.032650345 -0.024952192
## feedmais     0.064768451  0.072466604
## feedmeat    -0.011899586 -0.004201433
## feedsoy      0.095703138  0.103401291
## hour        -0.001823011 -0.001713775
```

En del af konfidensintervallerne fra spørsmål 2.-6. kan vi aflæse af ovenstående `R`-udskrift ... men hvilke?

Hydrolyseeksempel: konfidensintervaller (II)

Man vil ofte være interesserede i konfidensintervaller for andet end blot parametrene i `R`-udskriften.

- Hvordan kan vi få konfidensintervallet for forskellen mellem `barley` og `fish`? [-spørgsmål 4. fra slide 17!]
- Hvordan kan vi få konfidensintervallet for estimatet hørende til `feed`=fish?
- Hvordan kan vi få konfidensintervallet for forskellen mellem `mais` og `fish`? [-spørgsmål 5. fra slide 17!]
- Hvordan kan vi få konfidensintervallet for det forventede log(`serine`)-indhold svarende til `feed`=barley og hydrolysetiden `T`=16 timer? [-spørgsmål 6. fra slide 17!]

De første tre spm. har du baggrund for at kunne besvare selv, men lad os lige repetere ...

Til besvarelse af sidste spørgsmål kan benyttes funktionen `estimable` i pakken `gmodels` som skal "loades" og evt. installeres først.

Hydrolyseeksempel: forskel ml. `mais` og `fish`

Se på `summary` af modellen:

```
##                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  0.666851638 1.581941e-03 421.540016 5.476254e-81
## feedfish    -0.028801269 1.909863e-03 -15.080279 5.374849e-19
## feedmais     0.068617527 1.909863e-03  35.927981 3.029453e-34
## feedmeat    -0.008050510 1.909863e-03  -4.215229 1.220410e-04
## feedsoy      0.099552215 1.909863e-03  52.125313 3.415175e-41
## hour        -0.001768393 2.710076e-05 -65.252521 1.973932e-45
```

Find relevant linearkombination af koefficienterne, her

$$0 \cdot (\text{Intercept}) - 1 \cdot \text{feedfish} + 1 \cdot \text{feedmais} + 0 \cdot \text{feedmeat} + 0 \cdot \text{feedsoy} + 0 \cdot \text{hour}$$

svarende til vektoren `c(0,-1,1,0,0,0)`.

```
library(gmodels)
mf<-c(0,-1,1,0,0,0)
est=rbind(mf)
```

```
estimable(model3,est,conf.int=0.95)

##     Estimate  Std. Error  t value DF Pr(>|t|)   Lower.CI  Upper.CI
## mf 0.0974188 0.001909863 51.00826 44        0 0.09356972 0.1012679
```

Hydrolyseeksempel: plenumopgave

På tilsvarende måde kan man finde estimat og konfidensinterval for forventet log-serin i byg efter 16 timers hydrolyse.

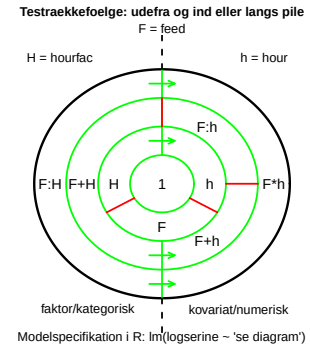
- Hvordan?
- Konfidensinterval for forventet serin (ikke log)?

Koefficienterne og modellen skal passe sammen:

Det er ligegyldigt, hvilken opskrivning af modellen man bruger — bare koefficienterne passer sammen med den.

Et godt råd: beregn først *estimatet* i hånden, så er du (mere) sikker på, at `R` faktisk estimerer, det du ønsker.

Hydrolyseeksempel: modeloversigt



- Lær modellerne i forb. med eksempel 4.2 godt at kende.
- Hvordan kunne vi have testet for linearitet, hvis den additive faktormodel (`model2`) var blevet afvist?

Hydrolyseeksempel 4.2: lineær model

Her nåede vi frem til flg. slutmodel

$$Y_i = \alpha(\mathtt{F_i}) + \beta \cdot \mathtt{T_i} + \mathtt{e_i}, \quad \mathtt{e_i} \sim \mathtt{N}(0, \sigma^2).$$

Hvordan skal vi fortolke slutmodellen grafisk?

Bemærk, at dette er en lineær model!

$$\begin{aligned}
\mathbb{E}Y_1 &= 1\cdot\alpha(\mathtt{barley}) + 0\cdot\alpha(\mathtt{fish}) + \ldots + 0\cdot\alpha(\mathtt{soy}) + 8\cdot\beta \\
\mathbb{E}Y_2 &= 1\cdot\alpha(\mathtt{barley}) + 0\cdot\alpha(\mathtt{fish}) + \ldots + 0\cdot\alpha(\mathtt{soy}) + 16\cdot\beta \\
&\vdots \\
\mathbb{E}Y_{50} &= 0\cdot\alpha(\mathtt{barley}) + 0\cdot\alpha(\mathtt{fish}) + \ldots + 1\cdot\alpha(\mathtt{soy}) + 72\cdot\beta
\end{aligned}$$

Middelværdien er for alle 50 obs. en (kendt) lineær funktion af modellens 6 parametre.

Lineære modeller i perspektiv

En model er lineær hvis

- observationerne er uafhængige og normalfordelte
- middelværdierne er lineære i parametrene

Stor klasse af modeller. Kan håndtere

- faktorer og/eller kontinuerte variable (kovariater) som forklarende variable
- manglende observationer
- ufuldstændige designs

Fuldstændig teori om estimation, test, estimerbarhed osv.

Vigtigt at kontrollere om modellerne passer på data
→ modelkontrol (torsdag i uge 2)

Lineære modeller i perspektiv (2)

Lineære modeller kan ikke bruges hvis

- der er en ikke-lineær sammenhæng mellem respons og en kovariat (medmindre vi kan transformere til noget lineært) → ikke-lineære modeller (ikke på SD2)
- responsen ikke er normalfordelt (medmindre vi kan transformere og opnå normalitet), fx.
  - død/levende, spirer/spirer ikke, rask/syg/død osv.

  → logistisk regression mm. (ikke på SD2)
- observationer ikke er uafhængige, fx.
  - flere målinger per individ
  - flere dyr fra samme kuld.

  → tilfældige effekter, gentagne målinger (kap. 7–10)

Repetition af lineær modeller

- Hvad er en lineær model?
- Eksempler: ANOVA-modeller, regressionsmodeller og kombinationer
- Dimension af lineær model, overparametrisering
- Test af to modeller mod hinanden: F-testet
- Modelreduktioner og test af hypoteser: modelhjulet
- Jonglering med modeller: fortolkning, specifikation i `R`
- Forklarende variable som faktorer og/eller kovariater
- Test for reduktion af forklarende variabel fra faktor til kovariat (test for lineær sammenhæng)
- Relevante `R`-funktioner: `lm, anova, summary, confint`
- Opstilling af og estimation af interessante kontraster og andre funktioner af parametrene
- `R`-funktionen `estimable` fra `gmodels`-pakken

Øvelse: nok mest til hjemmbrug!

Observationer $Y_1, \ldots, Y_N$ med tilhørende kendte, kontinuerte målinger $x_1, \ldots, x_N$.

Antag at $Y_1, \ldots, Y_N$ er uafhængige og normalfordelte med spredning σ, og at

$$\mathbb{E}Y_i = \alpha \cdot x_i^{\beta}$$

- Hvilke parametre er der i modellen?
- Er det en lineær model?
- Hvis nej, kan du forestille dig en transformation af data der giver en lineær model? Hvad er dimensionen af denne?

Hint Tænk på eksemplet med hjertevolumen for raske hunde fra forelæsningen d. 14/9-2017!