Introducerende Statistik og Dataanalyse med R

Generelle Linære Model: F-test

Jens Ledet Jensen



l dag

- Twoway anova
- Model, F-test
- Teste regressionsmodellen

Dagens spørgsmål:

Modeller og modelformler

$$X_i \sim N(\xi_i, \sigma^2)$$

One way anova: respons ~ Gruppe (Gruppe er en faktor)

$$\xi_i = \mu_{\mathsf{Gruppe}_i}$$

Two way anova: respons \sim K*A (K og A er faktorer)

$$\xi_i = \mu_{K_i, A_i}$$

additive model: respons \sim K+A, $\xi_i = \eta_{\mathcal{K}_i} + \zeta_{\mathcal{A}_i}$

Lineær regression: respons \sim t (t er en regressionsvariabel)

$$\xi_i = \alpha + \beta t_i$$

To-sidet variansanalyse: Alanin i lymfevæske

kan	art	Ala
kqn		
han	1	21.5
han	1	19.6
han	1	20.9
han	1	22.8
han	2	14.5
han	2	17.4
han	2	15.0
han	2	17.8
han	3	16.0
han	3	20.3
han	3	18.5
han	3	19.3
hun	Ĭ	14.8
hun	1	15.6
hun	ī	13.5
hun	1	16.4
hun	2	12.1
hun	2	11.4
hun	2	12.7
	2	14.5
hun	2	
hun	1 1 2 2 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2	14.4
hun	3	14.7
hun	3	13.8
hun	3	12.0

kqn: køn: 2 niveauer

art:: 3 niveauer

Ala: koncentration af Alanin

Model: Ala_i ~ $N(\mu_{\mathsf{kqn}_i,\mathsf{art}_i},\sigma^2)$, $i=1,\ldots,n$

Parametre: $\mu_{\text{han},1}, \mu_{\text{han},2}, \dots, \mu_{\text{hun},3}, \sigma^2$

Seks grupper med hver sin middelværdi

Additive model: $\mu_{\mathsf{kqn,art}} = \eta_{\mathsf{kqn}} + \zeta_{\mathsf{art}}$

Bidrag fra køn plus bidrag fra art: 2+3-1 parametre

To faktorer

kqn, art: begge faktorer: Se webbog afsnit 4.1

Modelformel: kqn*art giver ny faktor der deler op efter både kqn og art Direkte i kommandovindue: kqn:art

R modelformel:

kqn*art samme som kqn+art+kqn*art samme som kqn+art+kqn:art

Parametrisering i R:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} = \begin{bmatrix} a & a & a \\ a & a & a \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ d-a & d-a & d-a \end{bmatrix} + \begin{bmatrix} 0 & b-a & c-a \\ 0 & b-a & c-a \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & e-b-d+a & f-c-d+a \end{bmatrix}$$

Intercept kqnhun art2 art3 kqnhun:art2 kqnhun:art3

R: kqn + art + kqn:art

```
lm
```

summary($lm(Ala \sim kqn*art)$

```
Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.2000 0.7267 29.175 < 2e-16 ***

kqnhun -6.1250 1.0276 -5.960 1.22e-05 ***

art2 -5.0250 1.0276 -4.890 0.000118 ***

art3 -2.6750 1.0276 -2.603 0.017983 *

kqnhun:art2 2.6250 1.4533 1.806 0.087631 .

kqnhun:art3 1.3250 1.4533 0.912 0.373967
```

Residual standard error: 1.453 on 18 degrees of freedom

Gennemsnit for (han,art1): 21.20, (hun,art2): 21.20-6.125-5.0250+2.625=12.675

Skal vi acceptere additive model svarende til: kqnhun:art2=0 og kqnhun:art3=0 ? lm

```
anova(Im(Ala\sim kqn+art),Im(Ala\sim kqn*art))
```

```
Model 1: Ala ~ kqn + art

Model 2: Ala ~ kqn * art

Res.Df RSS Df Sum of Sq F Pr(>F)

1 20 44.9083

2 18 38.0175 2 6.8908 1.6313 0.22331
```

Model 2 (kalder jeg
$$M_1$$
): Ala_i $\sim N(\mu_{kqn_i,art_i}, \sigma^2)$

Model 1 (kalder jeg
$$M_2$$
): Ala_i $\sim N(\eta_{\mathsf{kqn}} + \zeta_{\mathsf{art}}, \sigma^2)$

Teste additive model: teste reduktion fra M_1 til M_2 : F=1.6313, p-værdi=0.22331, data strider ikke mod reduktionen

Sammenligne "variation indenfor gruppe" med "variation mellem grupper" ?

Konfidensintervaller i additive model

Intercept:
$$\eta_{\sf han}+\zeta_1$$
 kqnhun = $\eta_{\sf hun}-\eta_{\sf han}$, art2 = $\zeta_2-\zeta_1$, art3 = $\zeta_3-\zeta_1$

Uanset art så er middelværdien for han 4.8 større end middelværdi for hun Uanset køn så er middelværdien for art1 2.0 større end middelværdi for art3

Prøv selv i webbog

Gå til webbog afsnit 4.7 skjulte punkt "Beregninger i R"

Fjern list(...) og:

Inskriv kode til beregning af F-test fra additive model til model hvor alle kombinationer af lys og stress har samme middelværdi

Sceneskift

Data, Im og Testtabel (anova) er vist

Næste: forstå testtabel og F-test generelt

F-test generelt

Model: $X_i \sim N(\xi_i, \sigma^2)$, $i = 1, \ldots, n$, uafhængige

alle har samme varians

Lave (lineære) modeller for middelværdierne ξ_1, \ldots, ξ_n

Til hver model M knytter vi:

skøn
$$\hat{\xi}(M)=(\hat{\xi}_1(M),\ldots,\hat{\xi}_n(M))$$

$$SSD(M) = \sum_{i} (x_i - \hat{\xi}_i(M))^2 = \text{sum af kvadrerede afvigelser}$$

df(M) = n minus antal frie parametre i middelværdi

$$s^2(M) = \frac{SSD(M)}{df(M)} \sim \sigma^2 \chi^2(df(M))/df(M)$$
, skøn over varians σ^2

F-test generelt

To modeller M_1 og M_2 : M_2 er en reduktion af M_1

f.eks. nogle parametre sættes lig med nul

eller nogle parametre antages ens

$$x-\hat{\xi}(M_1)$$
: afvigelse mellem model M_1 og data

$$x-\hat{\xi}(\mathit{M}_{2})$$
: afvigelse mellem model M_{2} og data

$$= (x - \hat{\xi}(M_1)) + (\hat{\xi}(M_1) - \hat{\xi}(M_2))$$

Testprincip: hvis den ekstra afvigelse $(\hat{\xi}(M_1) - \hat{\xi}(M_2))$ er for stor i forhold til $(x - \hat{\xi}(M_1))$, så strider data mod hypotesen

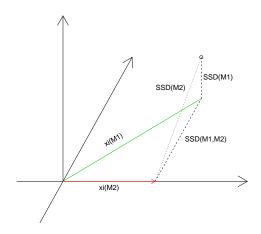
$$SSD(M_2) = SSD(M_1) + \sum_i (\hat{\xi}_i(M_1) - \hat{\xi}_i(M_2))^2$$
 (Pythagoras)
= $SSD(M_1) + SSD(M_1, M_2)$
er $SSD(M_1, M_2)$ for stor i forhold til $SSD(M_1)$?

Sammenligner "variation mellem grupper" med "variation indenfor grupperne"

$$SSD(M_1, M_2) = SSD(M_2) - SSD(M_1) \sim \sigma^2 \chi^2(df)/df$$

Næste slide: grafisk illustration

Tegning med $SSD(M_1, M_2)$



F-testet

Teststørrelse:

$$F = \frac{SSD(M_1, M_2) / (d(M_1) - d(M_2))}{SSD(M_1) / df(M_1)} = \frac{(SSD(M_2) - SSD(M_1)) / (df(M_2) - df(M_1))}{SSD(M_1) / df(M_1)}$$
$$F \sim F(df(M_2) - df(M_1), df(M_1))$$

store værdier er kritiske

$$F = \frac{V_1}{V_2}, \ V_1 \sim \chi^2(f_1)/f_1, \ V_2 \sim \chi^2(f_2)/f_2$$

$$p$$
-værdi = $1 - F_{cdf}(F, df(M_2) - df(M_1), df(M_1))$

Teori:
$$Q = \frac{\max_{M_2 \text{ Likelihood}}}{\max_{M_1 \text{ Likelihood}}} = \left(1 + \frac{d_1 - d_2}{n - d_1}F(x)\right)^{-n/2}$$

små værdier af Q = store værdier af F(x)

Navn: F for Fisher

To-sidet variansanalyse: Alanin i lymfevæske

Køn:	han	han	han	hun	hun	hun
Art:	1	2	3	1	2	3
	21.5	14.5	16.0	14.8	12.1	14.4
	19.6	17.4	20.3	15.6	11.4	14.7
	20.9	15.0	18.5	13.5	12.7	13.8
	22.8	17.8	19.3	16.4	14.5	12.8

Model: Ala_i
$$\sim N(\mu_{\mathsf{kqn}_i,\mathsf{art}_i},\sigma^2)$$
, $i=1,\ldots,n$

Teste additivitet: $\mu_{
m kqn,art} = \eta_{
m kqn} + \zeta_{
m art}$

$$anova(Im(Ala \sim kqn + art), Im(Ala \sim kqn * art))$$

```
Model 1: Ala ~ kqn + art

Model 2: Ala ~ kqn * art

Res.Df RSS Df Sum of Sq F Pr(>F)

1 20 44.9083

2 18 38.0175 2 6.8908 1.6313 0.22331 *
```

Teste reduktion fra kgn*art til kgn+art:

$$F = \frac{\left(\frac{44.908 - 38.017}{20 - 18}\right)}{38.0175/18} = 1.6313$$
 p-værdi = $1 - F_{cdf}(1.6313, 2, 18) = 0.22331$

Teste reduktion fra kqn+art til kqn. Hypotese: ingen effekt af art

```
anova(Im(Ala{\sim}kqn),Im(Ala{\sim}kqn{+}art))
```

```
Model 1: Ala ~ kqn

Model 2: Ala ~ kqn + art

Res.Df RSS Df Sum of Sq F Pr(>F)

1 22 100.1692

2 20 44.9083 2 55.2609 12.3053 0.00033 ***
```

$$F = 12.305$$
, p-værdi = $1 - F_{cdf}(12.305, 2, 20) = 0.00033$

Prøv selv i R

Tilvækst i tænder på grise, vitamindosis, indtagelsesmetode

len = Tooth Growth[, 1]

M=ToothGrowth[,2]

D = factor(ToothGrowth[,3])

Antag den additive model beskriver data. Lav test for ingen effekt af dosis

Likelihoodratio test

Finde Q ved brug af: $SSD(M_2) = SSD(M_1) + SSD(M_1, M_2)$

Likelihoodfunktion
$$L(\xi, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\sum_i (x_i - \xi_i)^2/(2\sigma^2)}$$

Maksimum over ξ i model M: $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-SSD(M)/(2\sigma^2)}$

Maksimum over
$$\sigma^2$$
: $\frac{1}{\sqrt{2\pi \cdot SSD(M)/n}}e^{-n/2}$

$$Q = \frac{\max_{\xi \in M_2, \sigma^2} L(\xi, \sigma)}{\max_{\xi \in M_1, \sigma^2} L(\xi, \sigma)} = \frac{\frac{1}{\sqrt{2\pi \cdot \text{SSD}(M_2)/n^n}} e^{-n/2}}{\frac{1}{\sqrt{2\pi \cdot \text{SSD}(M_1)/n^n}} e^{-n/2}} = \left(\frac{1}{1 + \text{SSD}(M_1, M_2)/\text{SSD}(M_1)}\right)^{n/2}$$

$$= \left(\frac{1}{1 + \frac{\mathrm{df}(M_2) - \mathrm{df}(M_1)}{\mathrm{df}(M_2)} \cdot F}\right)^{n/2}$$

Store værdier af F-teststørrelse = små værdier af likelihood ratio

Sceneskift : Ronald A. Fisher

Teststrategi

Starter med en "fuld" model

Reducerer model successivt ved at fjerne et led ad gangen

F-test bruges til at acceptere reduktion

```
Reduktioner: M_1 \to M_2 \to M_3 \to \cdots
anova(lm(M_{i+1}),lm(M_i))
tester reduktion fra M_i til M_{i+1} (har accepteret M_i)
```

$$\mathsf{anova}(\mathsf{Im}(M_{i+1}),\mathsf{Im}(M_i))$$

Sceneskift

Teststrategi er omtalt

Næste: teste for lineær sammenhæng

Artikel

Engineering quantum dot calibration standards for quantitative fluorescent profiling

Journal of Materials Chemistry B

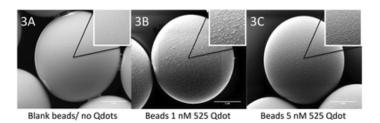
Fluorescence-based tools, in particular optimized fluorophores, offer useful approaches to map cellular heterogeneity. Applying quantum dot (Qdot) technology towards heterogeneity profiling would be a novel approach for characterizing cellular dispersion and requires sensitive calibration standards.

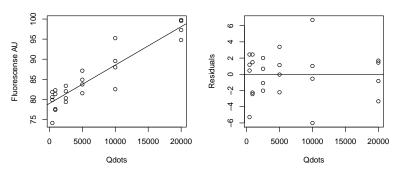
Artikel

Fig. 4 shows a representative plot of Qdot 525 fluorescence levels observed at increasing concentrations of Qdots.

The x-axis shows the number of Qdots per bead, QDSites. This represents a theoretical value.

The y-axis shows the empirical measurements of the fluorescence of each mixture.





Model M_1 : hver QDOTgruppe (450, 900, ..., 20000) har sin egen middelværdi

Model M2: middelværdi afhænger lineært af Qdots

Hvad synes I: er model M_2 god?

Teste lineær sammenhæng

Dobbeltindexnotation:

Model
$$M_1$$
: $X_{ji} \sim N(\mu_j, \sigma^2)$, $j = 1, ..., k$, $i = 1, ..., n_j$
Model M_2 : $X_{ji} \sim N(\alpha + \beta t_j, \sigma^2)$, $j = 1, ..., k$, $i = 1, ..., n_j$

Model
$$M_1$$
: $\hat{\xi}_{ji}(M_1) = \hat{\mu}_j = \bar{X}_j$, gennemsnit i gruppe j Model M_2 : $\hat{\xi}_{ji}(M_2) = \hat{\alpha} + \hat{\beta}t_j$

$$SSD(M_1, M_2) = \sum_{j} \sum_{i} (\hat{\mu}_j - (\hat{\alpha} + \hat{\beta}t_j))^2 = SSD(M_2) - SSD(M_1)$$

$$s^2(M_1) = \sum_{j} \sum_{i} (X_{ji} - \bar{X}_j)^2 / (n - k) = SSD(M_1) / (n - k)$$

Beregnet fra data:

$$SSD(M_1) = 178.6701$$
, $SSD(M_2) = 181.1265$
 $df(M_1) = 24 - 6 = 18$, $df(M_2) = 24 - 2 = 22$

Teste lineær sammenhæng

Model	SSD(M)	df (M)
M_1 : FakQ	178.6701	18
M_2 : Qdots	181.1265	22

$$F = \frac{\left(\frac{ss_{D(M_2)} - ss_{D(M_1)}}{df(M_2) - df(M_1)}\right)}{s^2(M_1)}$$

Lave F-test for reduktion fra model M_1 til model M_2 :

$$F = \frac{\frac{(181.1265 - 178.6701)/(22 - 18)}{178.6701/18}}{178.6701/18} = 0.0619$$

$$pval = 1 - F_{cdf}(0.0619, 4, 18) = 0.992$$

Konklusion: ifølge dette test strider data ikke mod lineær sammenhæng mellem middelværdi af *Fluor* og *Qdots*

Direkte udregning af test i R:

```
anova(Im(Fluor \sim Qdots), Im(Fluor \sim FakQ))
```

```
Model 1: Fluor ~ Qdots

Model 2: Fluor ~ FakQ

Res.Df RSS Df Sum of Sq F Pr(>F)

1 22 181.1265

2 18 178.6701 2 2.4564 0.0619 0.9923
```

Konfidensintervaller

```
confint(lm(Fluor~Qdots))
```

	2.5 %	97.5 %
(Intercept)	77.30	80.64
Qdots	0.00078	0.00113

Hvad er Qdots når Fluor måles til 95:

inversReg(Im(Fluor \sim Qdots),95): [10368, 23884]

Stor usikkerhed!

Prøv selv i R

Plasma concentrations of indometacin

Lav test for lineær sammenhæng mellem log(tid) og log(plasma)

```
t=log(Indometh[,2])
fakT=factor(Indometh[,2])
```

x=log(Indometh[,3])

Sceneskift

Test for at en sammenhæng er lineær er vist

Næste: "regne opgave"

Plantgrowth

Results obtained from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment condition.

Undersøg hvordan plantevæksten afhænger af gruppen (ctrl, trt1, trt2)

```
w=PlantGrowth[,1]
Gruppe=PlantGrowth[,2]
```

(a) Undersøgelser for at se om data kan beskrives med en normalfordelingsmodel boxplot(w~Gruppe) qqnormFlere(w,Gruppe) (Indsæt figurer fra R)

QQplots snor sig om rette linjer så ok med model $W_i \sim N(\mu_{\text{Gruppe}_i}, \sigma_{\text{Gruppe}_i}^2)$ med de 6 parametre $\mu_{\text{ctrl}}, \mu_{\text{trt1}}, \mu_{\text{trt2}}, \sigma_{\text{ctrl}}, \sigma_{\text{trt1}}, \sigma_{\text{trt2}}$

Både boxplot og qqplots kan tyde på samme varians, boxplot peger måske på forskelllige middelværdier

(a) Undersøgelser for at se om der kan antages samme varians

I den opstillede model vil vi teste hypotesen $\sigma_{\rm ctrl}=\sigma_{\rm trt1}=\sigma_{\rm trt2}.$ Vi laver Bartletts test som beskrevet i afsnit 4.5. Fra udregningerne i R finder vi, at

Ba=2.8786,
$$p - vardi = 1 - \chi_{cdf}(2.8786, 2) = 0.2371$$

Da p-værdien er langt over 0.05 siger vi at data ikke strider mod hypotesen om samme varians. Vi betragter derfor nu modellen

$$W_i \sim N(\mu_{\mathsf{Gruppe}_i}, \sigma^2)$$
 med de 4 parametre $\mu_{\mathsf{ctr}|}, \mu_{\mathsf{trt}1}, \mu_{\mathsf{trt}2}, \sigma$

Plantgrowth: samme middelværdi?

(a) Undersøgelser for at se om der kan antages samme middelværdi

I den opstillede model vil vi teste hypotesen $\mu_{\rm ctrl}=\mu_{\rm trt1}=\mu_{\rm trt2}.$ Vi laver det generelle F-test beskrevet i afsnit 6.7. Fra udregningerne i R finder vi, at

$$F = 4.846$$
, p-værdi = $1 - F_{cdf}(4.846, 2, 27) = 0.0159$

Da *p*-værdien er noget under 0.05 siger vi at data strider mod samme middelværdi

Plantgrowth: konfidensintervaller

For modellen med hver sin middelværdi i de tre grupper er parametertabellen

```
Estimate Std. Error t value Pr(>|t|) (Intercept) 5.0320 0.1971 25.527 <2e-16 *** Gruppetrt1 -0.3710 0.2788 -1.331 0.1944 Gruppetrt2 0.4940 0.2788 1.772 0.0877 . Residual standard error: 0.6234 on 27 degrees of freedom
```

Vi ser her at ingen af treatment grupperne afviger markant fra control, men da de går i hver sin retning afviger de to treatments fra hinanden

Hvis vi laver t-test for at de to treatments har samme middelværdi fås $t=(0.4940-(-0.3710))/(0.6234\sqrt{1/10+1/10})=3.103$ med tilhørende p-værdi fra en t(27)-fordeling på 0.0045

Plantgrowth: konfidensintervaller

```
Konfidensintervaller for \mu_{\rm ctrl}, \mu_{\rm trt1}-\mu_{\rm ctrl}, \mu_{\rm trt2}-\mu_{\rm ctrl} 2.5~\%~~97.5~\% (Intercept) 4.62752600~5.4364740 Gruppetrt1 -0.94301261~0.2010126 Gruppetrt2 -0.07801261~1.0660126
```

Plantgrowth: konfidensinterval for $\mu_{\mathrm{trt2}} - \mu_{\mathrm{trt1}}$

```
Gruppe=relevel(Gruppe,"trt1")

confint(lm(w~Gruppe))

2.5 % 97.5 %

(Intercept) 4.2565260 5.0654740

Gruppectrl -0.2010126 0.9430126

Gruppetrt2 0.2929874 1.4370126
```

Sceneskift

Slut på "opgave"

Næste forelæsning: gruppespecifik regression og matematikken bag den generelle lineære model

Slut for i dag, eller hvis tid: parret t-test

Parret t-test som two-way anova

Høstu dbytt e			
Mark	Ny Såmaskine	Gængs Såmaskine	Forskel <i>d</i>
1	8.0	5.6	2.4
2	8.4	7.4	1.0
3	8.0	7.3	0.7
9	5.6	5.5	0.1
10	6.2	5.5	0.7

$$E(X_i) = \xi_i = \eta_{\mathsf{Mark}_i} + \zeta_{\mathsf{Maskine}_i}$$
, Hypotese: $\delta = \zeta_{\mathsf{Ny}} - \zeta_{\mathsf{Gængs}} = 0$

Parret *t*-test:
$$t = \frac{\bar{d}}{s_d/\sqrt{10}} = 3.2143$$
, *p*-værdi = 0.0106

Two-way anova: $summary(Im(Hoest \sim Mark + Maskine))$

Sceneskift

Slut for i dag