Matematisk Statistik: Modelbaseret Inferens

Dataanalyse

Jens Ledet Jensen

I dag

Kort oversigt over modeller og modelbaseret analyse

Et par dataeksempler: diamantdata / smartphone eller ej

OL 2002, speed scating 23 parløb, 15 gange startede vinder på yderbane

yder = 15
Model: Yder $\sim$ binom$(23, p)$, $0 \leq p \leq 1$
Hypotese: $p = \frac{1}{2}$
P-værdi: $P(|Yder - \frac{23}{2}| \geq |15 - \frac{23}{2}|) = 0.2100$

95%-konfidensinterval for p (approksimativt):

$$\frac{15+\frac{1.96^2}{2} \pm 1.96\sqrt{15 \cdot (23-15)/23+\frac{1.96^2}{4}}}{23+1.96^2} = [0.45, 0.81]$$

Ud af 36 løb er der 16 hvor vinderen starter på yderbanen

Find p-værdien for test af hypotesen at sandsynligheden for at vinderen starter på yderbanen er $\frac{1}{2}$

Efter bestråling med dosis 50 er der sket 111 skader på 2652 celler

skader = 111
Model: Skader $\sim$ pois$(2652 \cdot 50 \cdot \lambda)$, $\lambda \geq 0$
λ er rate per celle per dosis

95%-konfidensinterval for λ (approksimativt):

$$\frac{111 + \frac{1.96^2}{2} \pm 1.96\sqrt{111 + \frac{1.96^2}{4}}}{2652 \cdot 50} = [0.00070, 0.00101]$$

Dødsfald for hvert regiment for hvert år, 280 målinger
A_j: antal blandt de 280 med bestemt værdi

Index	1	2	3	4	5
Værdi	0	1	2	3	≥ 4
a_j	144	91	32	11	2
e_j	139.04	97.33	34.07	7.95	1.61

Model: $(A_1, \ldots, A_5) \sim \text{multinom}(n, (\pi_1, \ldots, \pi_5))$, $\pi_j \geq 0$, $\pi_1 + \cdots \pi_5 = 1$
Hypotese: $\pi_j = \frac{\lambda^{j-1}}{(j-1)!} e^{-\lambda}$, $j \leq 4$, $\pi_5 = 1 - \pi_1 - \cdots - \pi_4$

$$\hat{\lambda} = \frac{196}{280} = 0.7$$

Teststørrelse: $G = 2 \sum_j a_j \log\left(\frac{a_j}{e_j}\right) = 1.84$ (slår kasse 4 og 5 sammen)

$$p\text{-værdi} = 1 - \chi^2_{\text{cdf}}(1.84, 4 - 1 - 1) = 0.399$$

	Rejse	Spise	Leg	Total
Morgen	6	28	38	72
Aften	13	56	10	79

$(\text{Delf}_{MR}, \text{Delf}_{MS}, \text{Delf}_{ML}) \sim \text{multinom}(72, (\pi_{MR}, \pi_{MS}, \pi_{ML})), \quad \pi_{Mj} \geq 0, \pi_{MR}+\pi_{MS}+\pi_{ML}=1$

$(\text{Delf}_{AR}, \text{Delf}_{AS}, \text{Delf}_{AL}) \sim \text{multinom}(79, (\pi_{AR}, \pi_{AS}, \pi_{AL})), \quad \pi_{Aj} \geq 0, \pi_{AR} + \pi_{AS} + \pi_{AL} = 1$

$$\text{Hypotese:} \quad (\pi_{MR}, \pi_{MS}, \pi_{ML}) = (\pi_{AR}, \pi_{AS}, \pi_{AL})$$

Forventede for (Morgen,Rejse): $72 \cdot 19/151 = 9.06$

Teststørrelse: $G = 2\left\{6 \cdot \log\left(\frac{6}{9.06}\right) + \cdots + 10 \cdot \log\left(\frac{10}{25.11}\right)\right\} = 29.25$

p-værdi $= 1 - \chi^2_{\text{cdf}}\big(29.25, (2-1)(3-1)\big) = 4.5 \cdot 10^{-7}$ (alle forventede er ≥ 5)

15 studerende er udvalgt og preference undersøgt med følgende resultat:

Aargang	Instatgram	Snapchat
Første	2	5
Anden	7	1

Vælg model og hypotese - og lav test

Udledning af likelihood ratio test

Forskellige betingningsargumenter

Betinge med summen i poissonfordelinger

Køkkenvægt: 10 målinger (V_i), forventer en visning på 600

Model: $V_i \sim N(\mu, \sigma^2), \;\; i = 1, \ldots, 10, \;\; (\mu, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+.$

Hypotese: $\mu = 600$

t-teststørrelsen: $t = (581.9 - 600)/(15.022/\sqrt{10}) = -3.81$ (mean(v) og sd(v))

p-værdi $= 2(1 - t_{\mathrm{cdf}}(3.81, 9)) = 0.0042,$ (R: t.test(v,mu=600))

95%-konfidensinterval for μ: $581.9 \pm 2.2622 \cdot 15.022/\sqrt{10} = [571.2,\ 592.6]$

95%-konfidensinterval for σ: `sqrt(df*s^2/qchisq(c(0.975,0.025),df))`
$s = 15.022, \;\; df = 9$

Teste at middelværdi er nul for øget søvnlængde blandt 10 personer:

ekstra=sleep[1:10,1]

Længden af horn af hornede tudseøgle

$\text{Doede}_i \sim N(\mu_1, \sigma^2), \ i = 1, \ldots, 30$

$\text{Levende}_i \sim N(\mu_2, \sigma^2), \ i = 1, \ldots, 154$

$(\mu_1, \mu_2, \sigma_1, \sigma_2) \in \mathbf{R}^2 \times \mathbf{R}_+^2$

Hypotese: $\sigma_1^2 = \sigma_2^2$. Teststørrelse: $F = \frac{s_1^2}{s_2^2} \sim F(29, 153)$ (var.test(doede,levende))

Hypotese $\mu_1 = \mu_2$: t.test(doede,levende,var.equal=TRUE/FALSE),
giver både t-test og konfidensinterval for $\mu_1 - \mu_2$

Universets udvidelse: hastighed og afstand mellem galakser

$\text{Hast}_i \sim N(\alpha + \beta \cdot \text{afstand}_i, \sigma^2), \;\; i = 1, \ldots, 24, \;\; (\alpha, \beta, \sigma) \in \mathbf{R}^2 \times \mathbf{R}_+.$

Analyse i R: `summary(lm(hast afstand))` og `confint(lm(hast afstand))`

giver $\hat{\alpha}, \hat{\beta}$ og konfidensintervaller, samt s_r

Modelkontrol: plot(afstand,lm(hast∼afstand)$residuals), qqnorm()

Find skøn over skæring og hældning i lineær regression af tryk$^{1/8}$ på temperatur:

```
tryk8=pressure[,2]^(1/8)

temp=pressure[,1]
```

Bakterieantal Bakt_i efter 32 håndvask, delt op på 4 metoder

Model: $\text{Bakt}_i \sim N(\mu_{\text{metode}_i}, \sigma^2)$, metode er en faktor

Hypotese $\mu_{\text{antibakspray}} = \mu_{\text{antisaebe}} = \mu_{\text{saebe}} = \mu_{\text{vand}}$

F-test i R: anova(lm(bakt∼1),lm(bakt∼metode))

```
Model 1: bakt ~ 1
Model 2: bakt ~ metode
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     31 69366
2     28 39484  3     29882 7.0636 0.001111 **
```

Teste varianser ens: bartlett.test(bakt∼metode)

Areal af blade på soyaplanten udsat for to lysforhold (lav/høj) og to stressforhold (med og uden mekanisk stress)

Model:

$$\text{Areal}_i \sim N\left(\mu_{\text{lys}_i,\text{stress}_i}, \sigma^2_{\text{lys}_i,\text{stress}_i}\right), \quad i = 1, \ldots, 42$$

$$\text{Hypotese: } \sigma^2_{\text{Hoej,Med}} = \sigma^2_{\text{Hoej,Uden}} = \sigma^2_{\text{Lav,Med}} = \sigma^2_{\text{Lav,Uden}}$$

Ny model:

$$\text{Areal}_i \sim N\left(\mu_{\text{lys}_i,\text{stress}_i}, \sigma^2\right), \quad i = 1, \ldots, 42$$

$$\text{Hypotese: } \mu_{\text{lys}_i,\text{stress}_i} = \zeta_{\text{lys}_i} + \eta_{\text{stress}_i}$$

F-test i R: anova(lm(areal~lys+stress),lm(areal~lys*stress))

Teste den additive model i tosidet variansanalyse

Eksempel 4.3 i afsnt 4.6 i webbogen: Beregninger i R

- benyt anova til at teste additive model

Estimation = projektion

Fordeling af parameterskøn: vektornormalfordeling: middelværdi og varians

Uafhængighed: spaltningssætningnen

Herfra: nogle nye dataeksempler

Pris på diamant som funktion af klarhed, farve og karat

nummer	klarhed	farve	karat	pris
1	VS	E	0.31	1555
2	VS	F	0.31	1427
3	IFV	G	0.31	1427
4	VS	H	0.31	1126
5	VS	F	0.32	1468
⋮	⋮	⋮	⋮	⋮
96	VS	G	0.60	3421
97	IFV	H	0.60	3925
98	IFV	H	0.61	3616
99	IFV	H	0.64	3785
100	IFV	H	0.66	4300

Diamond Clarity

FL, IF: Flawless, Internally Flawless: No internal or external flaws.

VVS1, VVS2: Very, Very Slightly Included: Very difficult to see inclusions under 10x magnification.

VS1, VS2: Very Slightly Included: Inclusions are not typically visible to the unaided eye.

E: Colorless. Only minute traces of color can be detected by an expert gemologist.
F: Colorless. Slight color detected by an expert gemologist, but still considered a "colorless" grade.

G-H: Near-colorless. Color noticeable when compared to diamonds of better grades.

1 carat $= 0.2$ gram

Johannesbrødtræet: "Seed size variability: from carob to carats", 2006

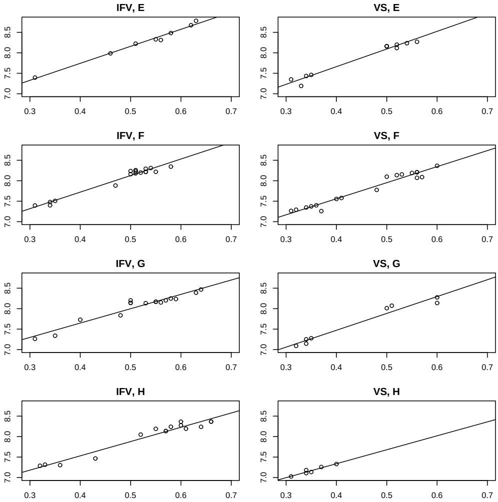
Klarhed: faktor med to niveauer: IFV og VS (forkortes K)

Farve: faktor med 4 niveauer: E, F, G, H (forkortes F)

Klarhed*Farve deler op i 8 undergrupper

karat: forklarende variabel der angiver vægt (v)

$Lpris = \log(\text{pris})$: respons


IFV, E
VS, E
IFV, F
VS, F
IFV, G
VS, G
IFV, H
VS, H
7.0
7.5
8.0
8.5
0.3
0.4
0.5
0.6
0.7

Model M_0: $\text{Lpris}_i \sim N(\alpha_{K_i,F_i} + \beta_{K_i,F_i} v_i, \sigma^2_{K_i,F_i})$

Hver af de 8 undergruppe har sin egen linære sammenhæng og sin egen varians

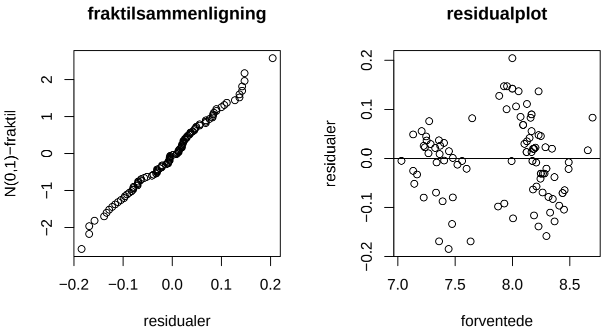
Teste varianser ens: Bartletts test (vis i R)

Model M_1: $\text{Lpris}_i \sim N(\alpha_{K_i,F_i} + \beta_{K_i,F_i} v_i, \sigma^2)$

modelformel: Lpris = K*F + K*F*v

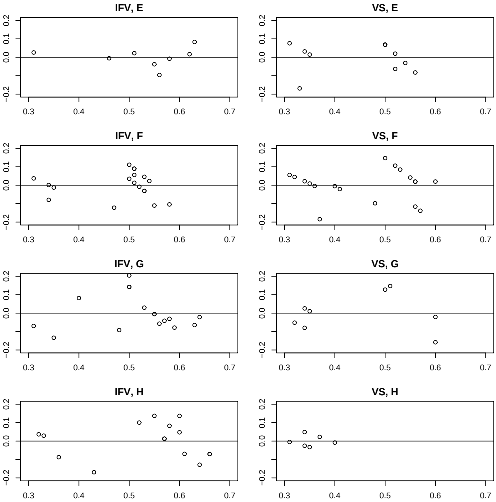
$\text{df}(M_1) = n - d(M_1) = 100 - 16 = 84$

Forventede værdier: $\hat{\xi}_i(M_1) = \hat{\alpha}_{K_i,F_i} + \hat{\beta}_{K_i,F_i} v_i$

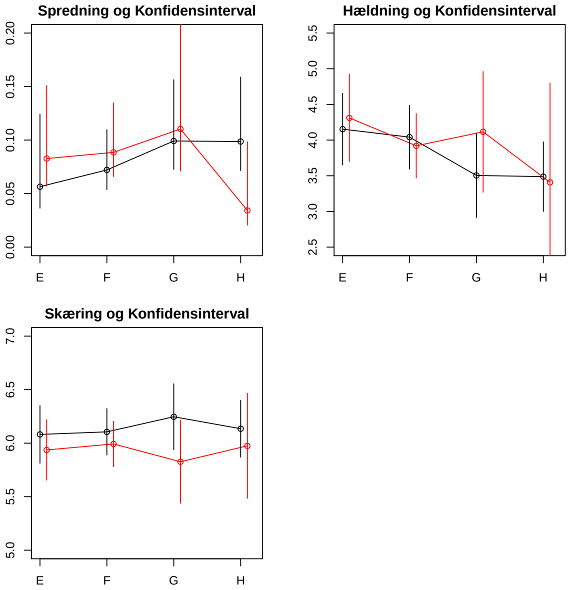


residual: $r_i = \text{Lpris}_i - \hat{\xi}_i(M_1)$, forventede: $\hat{\xi}_i(M_1) = \hat{\alpha}_{k_i,f_i} + \hat{\beta}_{k_i,f_i} v_i$

Modelkontrol: subplots af residualer



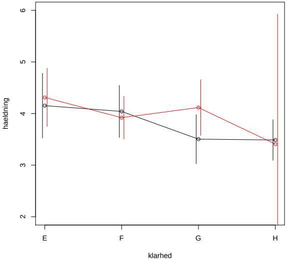
Figur med spredninger



Data og model er præsenteret

Model M_1: $\quad \text{Lpris}_i \sim N(\alpha_{K_i,F_i} + \beta_{K_i,F_i} v_i, \sigma^2)$

Hypotese: $\beta_{IFV,f} = \beta_{VS,f}$ for $f = E,\ F,\ G,\ H$



M_1: Lpris = K*F + K*F*v $\qquad \xi_i = \alpha_{K_i,F_i} + \beta_{K_i,F_i} v_i$

Teste at hældning ikke afhænger af *klarhed*:

$$\begin{cases} \beta_{IFV,E} = \beta_{VS,E} \\ \beta_{IFV,F} = \beta_{VS,F} \\ \beta_{IFV,G} = \beta_{VS,G} \\ \beta_{IFV,H} = \beta_{VS,H} \end{cases}$$

M_2: Lpris = K*F + F*v $\qquad \xi_i = \alpha_{K_i,F_i} + \tilde{\beta}_{F_i} v_i$

I et *t*-test "fjerner" vi 1 parameter

Her: ønsker test for simultant at fjerne 4 parametre

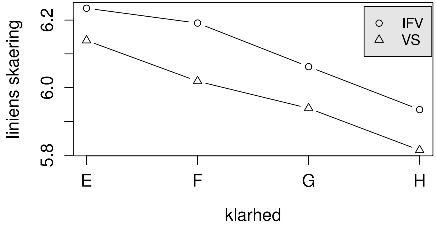
```
anova(lm(Lpris~K*F+F*v),lm(Lpris~K*F+F*K*v))
Model 1: Lpris ~ K*F + F*v
Model 2: Lpris ~ K*F + F*K*v
  Res.Df       RSS Df  Sum of Sq      F Pr(>F)
1     88 0.6548373
2     84 0.6312998  4  0.0235375  0.7830 0.5394

anova(lm(Lpris~K*F+v),lm(Lpris~K*F+F*v))
Model 1: Lpris ~ K*F + v
Model 2: Lpris ~ K*F + F*v
  Res.Df       RSS Df  Sum of Sq      F Pr(>F)
1     91 0.7120019
2     88 0.6548373  3  0.0571646  2.5607 0.0600
```

Model M_1: $\xi_i = \alpha_{K_i,F_i} + \beta_{K_i,F_i} v_i$
Model M_2: $\xi_i = \alpha_{K_i,F_i} + \beta_{F_i} v_i$
Model M_3: $\xi_i = \alpha_{K_i,F_i} + \beta v_i$

Konklusion: data strider ikke mod samme hældning

$$\xi_i = \alpha_{k_i,f_i} + \beta v_i$$



Teste reduktion fra $\xi_i = \alpha_{K_i,F_i} + \beta v_i$ til $\xi_i = \eta_{K_i} + \zeta_{K_i} + \beta v_i$

eller teste: $\alpha_{K_i,F_i} = \eta_{K_i} + \zeta_{F_i}$

```
anova(lm(Lpris~K+F+v),lm(Lpris~K*F+v))
Model 1: Lpris ~ K*F + v
Model 2: Lpris ~ K*F + F*v
  Res.Df       RSS Df  Sum of Sq       F Pr(>F)
1     94 0.7326882
2     91 0.7120019  3  0.0206863  0.8813 0.4539
```

$\alpha_{K_i,F_i} = \eta_{K_i} + \zeta_{F_i}$:

uanset farve er der samme forskel mellem IFV og VS

uanset klarhed er der samme forskel mellem farve E og farve G

```
summary(lm(Lpris~v+F+K))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.27235    0.05246 119.569  < 2e-16
v            3.83641    0.08988  42.685  < 2e-16
FF          -0.08607    0.02514  -3.424 0.000916
FG          -0.19242    0.02750  -6.998 3.79e-10
FH          -0.31957    0.02854 -11.197  < 2e-16
KVS         -0.13787    0.01942  -7.100 2.34e-10
Residual standard error: 0.08829 on 94 degrees of freedom
```

Prisen stiger med en faktor $e^{0.14}$ for at gå fra IFV til VS klassen

Prisen stiger med en faktor $e^{0.10}$ for at gå en farveklasse op

Prisen stiger med en faktor $e^{0.1 \cdot 3.84}$ for at øge vægten med 0.1 karat

Data er analyseret i artiklen "Application of near-infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit dough", Journal of the Science of Food and Agriculture, 1984

Ønsker at kunne prædiktere mængden af vand i dej ud fra gennemlysning med lys

NIR: near infrared reflectance spectroscopy

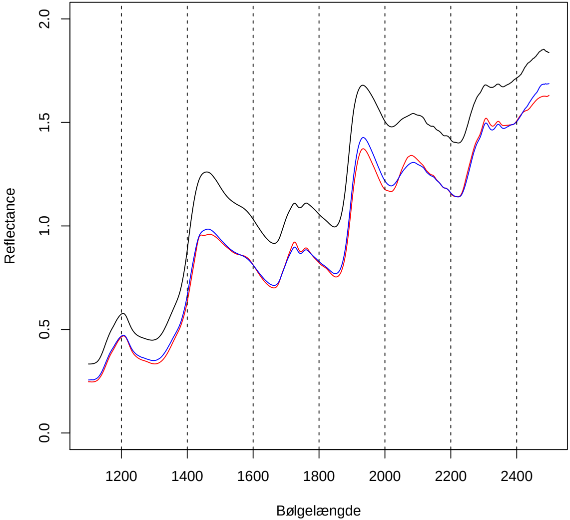
måler reflektion ved 700 bølgelænger i intervallet 1100-2500 nm

I dag: benytter kun bølgelængderne:

1200, 1400, 1600, 1800, 2000, 2200, 2400

Tidligere: se på alle 700 bølgelængder

Størst, mindst og mellem-værdi af vandindhold (sort, rød, blå)



Data: 40 dejprøver

Respons: x_i = mængden af vand i dejen

Forklarende variable: $t_1, \ldots, t_7$: mængden af "reflekteret" lys ved 7 bølgelængder

Multiple regressionsmodel:

$$X_i \sim N(\alpha + \beta_1 t_{i1} + \cdots + \beta_7 t_{i7}, \sigma^2)$$

Backward selection: fjerner successivt led: t4, t1, t7, t3

Slutmodel:

$$E(X_i) = \alpha + \beta_2 \cdot \text{t2}_i + \beta_5 \cdot \text{t5}_i + \beta_6 \cdot \text{t6}_i$$

fra $s(M_{\text{fulde}}) = 0.368$ til $s(M_B) = 0.384$ (LOOCV: 0.415)

F-test for reduktion fra fulde model til slutmodel: p-værdi $= 0.17$

Forward: t5, t6, t2, samme som backward model i dette tilfælde

FWcrossval(T,x,5): 0.855 0.476 0.415 0.4057 0.451

Backward med 700 variable: Tager 6 med, $s_{\text{cv}} = 0.217$

Multipel regression er vist

Næste: ægte testsæt

Vi har benyttet dej-data med 40 observationer

I det oprindelige datasæt var der 72 observationer, de 32 har jeg "gemt" som et nyt testssæt

Metode: Model vælges og parametre estimeres ud fra oprindelige 40 observationer

baseret på parameterskøn laves prædikterede værdier for de 32 testdata

Prædiktionsspredning: $s_P = \sqrt{\frac{1}{32}\sum_{i=1}^{32}\left(z_i - \hat{\xi}(i;M)\right)^2}$

z_i: vandindhold i den i'te testprøve

$\hat{\xi}(i;M)$ prædikterede værdi for den i'te testprøve under den estimerede model M

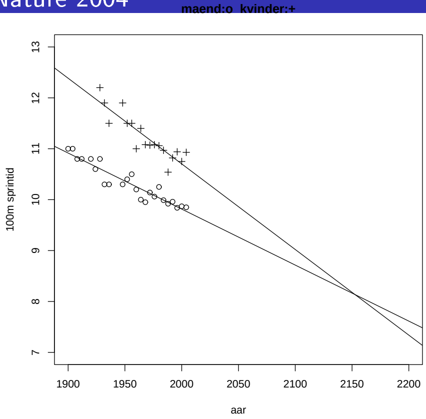
Sammenligner forward baseret på 7 variable og forward baseret på alle 700 variable

7 variable (tager 3 med) $s_P = 0.719$

700 variable (tager 6 med) $s_P = 0.253$

Multipel regression vist

Næste: 100 m sprinttider



Model: $X_{1i} \sim N(\alpha_1 + \beta_1 t_{1i}, \sigma_1^2) \quad i = 1, \ldots, n_1$

$$X_{2i} \sim N(\alpha_2 + \beta_2 t_{2i}, \sigma_2^2) \quad i = 1, \ldots, n_2$$

Tester $\sigma_1^2 = \sigma_2^2$:

Ny model: $X_i \sim N(\alpha_{K_i} + \beta_{K_i} t_i, \sigma^2)$

Tester $\beta_1 = \beta_2$:

summary(lm(x~K+K*t))

Konklusion: data strider mod hypotesen $\beta_1 = \beta_2$

Tidspunkt t_* hvor linjer skærer hinanden:

$$\alpha_1 + \beta_1 t_* = \alpha_2 + \beta_2 t_*$$

Konfidensinterval for t_* ?

Teste en værdi t_*: lave regressions på $t - t_*$ og teste samme skæring

```
tt=t-2073
sumUD=summary(lm(x~K:tt+K))
sumUD$coefficients[2,4]
```

95%-konfidensinterval for tidspunkt: [2073, 2641]

Kan findes ved at løse 2.grads-ligning

Sceneskift

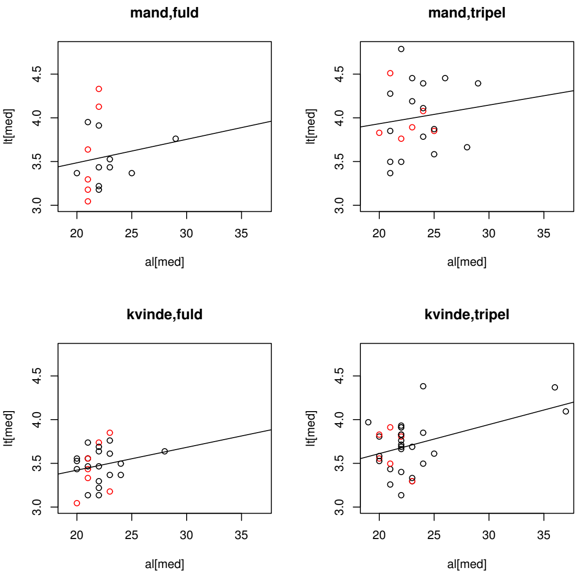
Skæringstidspunkt mellem to linjer er vist

Næste: SMS-eksperiment

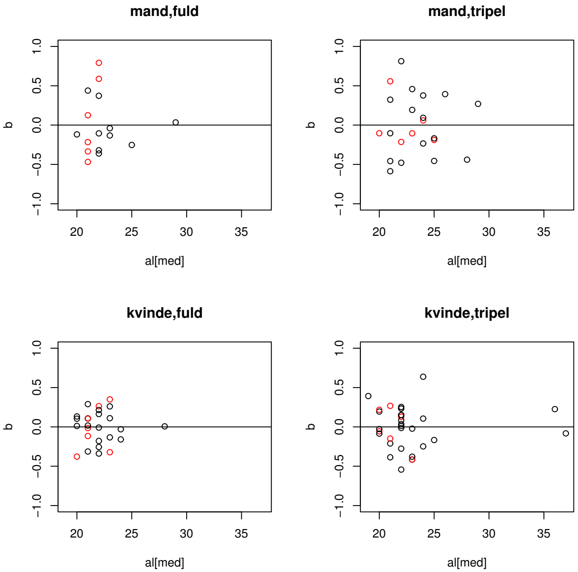
Dansk jomfru på Ærø kyler halvsexet quizbog ned i wc

The quick brown fox jumps over the lazy dog

ln(tid) mod alder

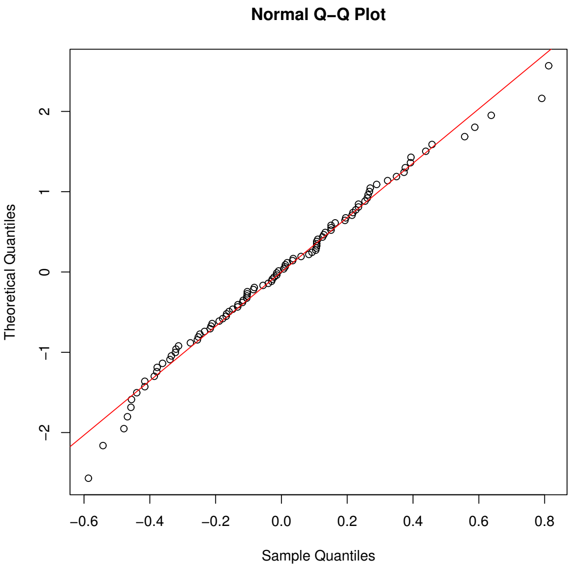


Residualer



mand,fuld
mand,tripel
kvinde,fuld
kvinde,tripel
b
al[med]
−1.0
−0.5
0.0
0.5
1.0
20
25
30
35

Fraktilsammenligning (qqplot)



Højde har ingen betydning

```
anova(lm(LT~K*S*Al),lm(LT~K*S*Al+K*S*Ho))
```

Teste fælles hældning

```
anova(lm(LT~K*S+Al),lm(LT~K*S*Al))
```

Teste additivitet af K og S

```
anova(lm(LT~K+S+Al),lm(LT~K*S+Al))
```

Har køn betydning

```
anova(lm(LT~S+Al),lm(LT~K+S+Al))
```

Parameterskøn

```
summary(lm(LT~Al+S+K))
```

Estimater

```
lm(formula = LT ~ Al + S + K)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.04787    0.26441  11.527  < 2e-16
Al           0.03045    0.01140   2.671  0.00891
SSm         -0.28964    0.06295  -4.601 1.31e-05
KMa          0.18817    0.06352   2.962  0.00387
Residual standard error: 0.305 on 94 degrees of freedom

confint(lm(LT~Al+S+K))
                  2.5 %      97.5 %
(Intercept)  2.52288250  3.57284758
Al           0.00781671  0.05308931
SSm         -0.41463141 -0.16465431
KMa          0.06204998  0.31428951
```

Mænd bruger 20% mere tid.   Smartphone bruger 25% mindre tid

10 år øger tidsforbruget med 35%

Slut med gennemgang af det modelbaserede

Næste: regne eksamensopgaver