Introducerende Statistik og Dataanalyse med R

Gruppespecifik regression

Jens Ledet Jensen

I dag

Formulering af generel lineær model

Flere regressionslinjer (gruppespecifik regression)

Teori: uafhængighed og χ^2-fordeling

n observationer: $x_1, x_2, \ldots, x_n$, uafhængige

$$\left.\begin{array}{l}\text{alle normalfordelte}\\ \text{alle samme varians}\end{array}\right\} \quad X_i \sim N(\bullet, \sigma^2)$$

$E(X_i) = \xi_i$: Model for ξ_i–erne

M : $\xi_i =$ sum af led

Led: faktor deler ind i grupper, hver gruppe får sit eget bidrag

Led: regression, $\beta \cdot t_i$

Model i dag: β afhænger af gruppe defineret ved faktor

Den formelle definition af en generel lineær model:

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\mathsf{T} \in L \quad L \text{ lineært underrum af } R^n$$

Eksempel 1: $X_i \sim N(\mu, \sigma^2)$

$$\boldsymbol{\xi} = (\mu, \ldots, \mu)^\mathsf{T} = \mu \mathbf{e} \in \text{Span}(\mathbf{e}), \quad \mathbf{e} = (1, 1, \ldots, 1)^\mathsf{T}$$

$$\text{Span}(\mathbf{v}_1, \ldots, \mathbf{v}_k) = \{\boldsymbol{\xi} = z_1\mathbf{v}_1 + \cdots + z_k\mathbf{v}_k | z_j \in R,\ j = 1, \ldots, k\}$$

Eksempel 2: $X_i \sim N(\mu_{G_i}, \sigma^2)$, $G = (1, \ldots, 1, 2, \ldots, 2)$, $n = n_1 + n_2$

$$\boldsymbol{\xi} = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2)^\mathsf{T} = \mu_1\mathbf{e}_1 + \mu_2\mathbf{e}_2 \in \text{Span}(\mathbf{e}_1, \mathbf{e}_2)$$

$$\mathbf{e}_1 = (1, \ldots, 1, 0, \ldots, 0)^\mathsf{T},\ \mathbf{e}_2 = (0, \ldots, 0, 1, \ldots, 1)^\mathsf{T}$$

Eksempel 3: $X_i \sim N(\alpha + \beta t_i, \sigma^2)$

$$\boldsymbol{\xi} = (\alpha + \beta t_1, \ldots, \alpha + \beta t_n)^\mathsf{T} = \alpha \mathbf{e} + \beta \mathbf{t} \in \mathrm{Span}(\mathbf{e}, \mathbf{t})$$

$$\mathbf{e} = (1, \ldots, 1)^\mathsf{T},\ \mathbf{t} = (t_1, \ldots, t_n)^\mathsf{T}$$

Generelt: en faktor G, der deler op i k grupper, giver k vektorer $\mathbf{e}_1, \ldots, \mathbf{e}_k$, hvor den j'te har 1 for de observationsnumre, der ligger i gruppe j, og nul ellers

En regressionsvariabel t giver en vektor $\mathbf{t}$

$$L = \mathrm{Span}(\ldots)$$

"plus" i modelformel giver ekstra vektorer i Span(...)

Faktor G deler ind i g grupper: vektorer $\mathbf{e}_1(G), \ldots, \mathbf{e}_g(G)$

Faktor D deler ind i d grupper: vektorer $\mathbf{e}_1(D), \ldots, \mathbf{e}_d(D)$

Underrum L_G og L_D har ikke-trivielt snit:

$$\mathbf{e}_1(G) + \cdots + \mathbf{e}_g(G) = \mathbf{e}_1(D) + \cdots + \mathbf{e}_d(D)$$

$$\dim(\text{Span}(\mathbf{e}_1(G), \ldots, \mathbf{e}_g(G), \mathbf{e}_1(D), \ldots, \mathbf{e}_d(D))) = g + d - 1$$

Estimation = projektion

Estimere parametre i middelværdimodel = finde $\boldsymbol{\xi} \in L$

Minimere $\sum_i (x_i - \xi_i(M))^2$ samme som at finde projektion af $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T}$ på L:

$$\sum_i (x_i - \xi_i)^2 = \|\mathbf{x} - \boldsymbol{\xi}\|^2 = \|\mathbf{x} - \mathbf{P}\mathbf{x} + \mathbf{P}\mathbf{x} - \boldsymbol{\xi}\|^2 = \|\mathbf{x} - \mathbf{P}\mathbf{x}\|^2 + \|\mathbf{P}\mathbf{x} - \boldsymbol{\xi}\|^2$$

$\mathbf{P}$ er projektionsmatricen, $\mathbf{x} - \mathbf{P}\mathbf{x}$ er vinkelret på alt i L, $\hat{\boldsymbol{\xi}} = \mathbf{P}\mathbf{x}$

$L = \mathrm{Span}(\mathbf{h}_1, \ldots, \mathbf{h}_k)$, $\mathbf{h}_1, \ldots, \mathbf{h}_k$ lineært uafhængige

$$\mathbf{P} = \mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}$$

$\mathbf{h}_1, \ldots, \mathbf{h}_k$ er søjlerne i $\mathbf{H}$

Vise at $\mathbf{P} = \mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}$ er projektionsmatricen:

Typisk vektor i L: $\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \mathbf{R}^k$

vise vinkelret på $\mathbf{x} - \mathbf{P}\mathbf{x}$

$$(\mathbf{x} - \mathbf{P}\mathbf{x})^\mathsf{T}\mathbf{H}\boldsymbol{\theta} = \mathbf{x}^\mathsf{T}\mathbf{H}\boldsymbol{\theta} - \mathbf{x}^\mathsf{T}\mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}\mathbf{H}\boldsymbol{\theta}$$

$$= \mathbf{x}^\mathsf{T}\mathbf{H}\boldsymbol{\theta} - \mathbf{x}^\mathsf{T}\mathbf{H}\boldsymbol{\theta} = 0$$

Eksempel: projektion på 1.aksen i $\mathbf{R}^2$, $\mathbf{H} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$$\binom{1}{0}((1,0)\binom{1}{0})^{-1}(1,0) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

M_2 er en undermodel af M_1: $L_2 \subset L_1$

Vise: $\mathrm{SSD}(M_2) = \mathrm{SSD}(M_1) + \mathrm{SSD}(M_1, M_2)$

$\mathrm{SSD}(M_1, M_2) = \sum_i (\hat{\xi}_(M_1) - \hat{\xi}_i(M_2)^2$

$\mathbf{P}_1$, $\mathbf{P}_2$ er projektionsmatricerne

$$\mathrm{SSD}(M_2) = \|\mathbf{x} - \mathbf{P}_2\mathbf{x}\|^2 = \|\mathbf{x} - \mathbf{P}_1\mathbf{x} + \mathbf{P}_1\mathbf{x} - \mathbf{P}_2\mathbf{x}\|^2$$

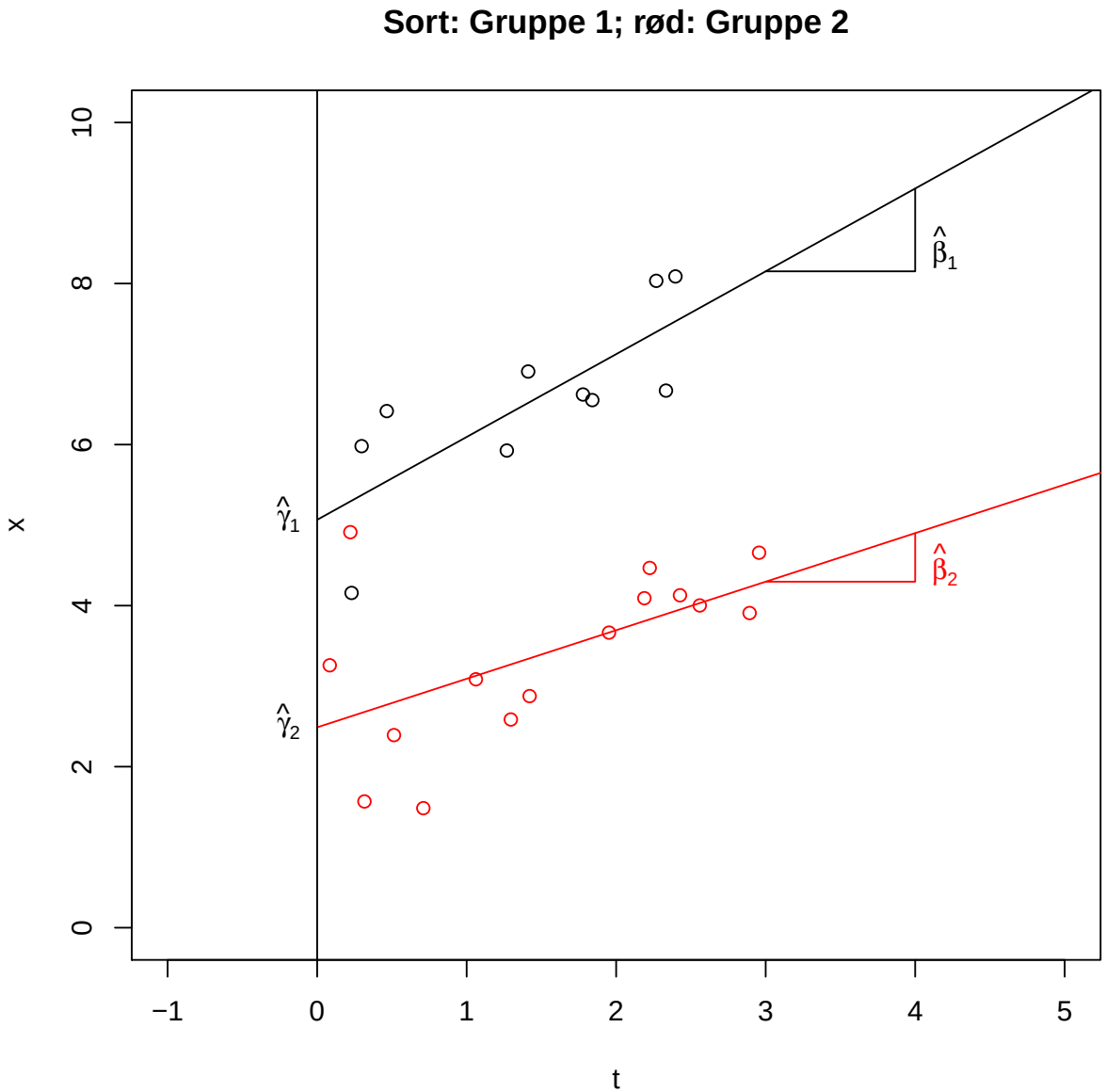
$$= \|\mathbf{x} - \mathbf{P}_1\mathbf{x}\|^2 + \|\mathbf{P}_1\mathbf{x} - \mathbf{P}_2\mathbf{x}\|^2$$

idet $\mathbf{P}_1\mathbf{x} - \mathbf{P}_2\mathbf{x} \in L_1$

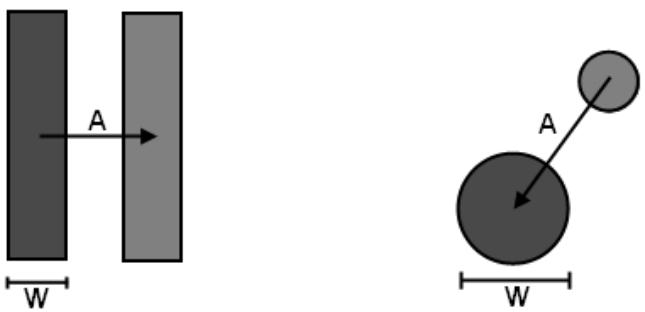
Den generelle lineære model er indført via underrum af R^n

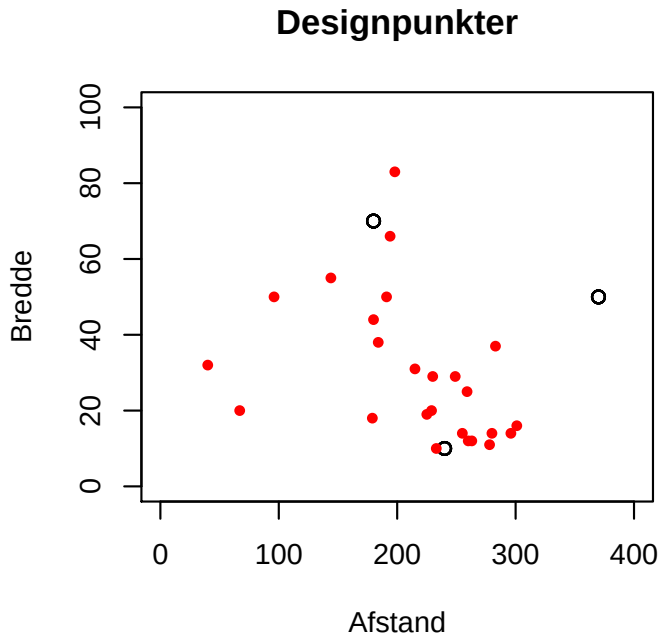
Næste: Model for gruppespecifik regression

Regression delt op i grupper



Berkeley data





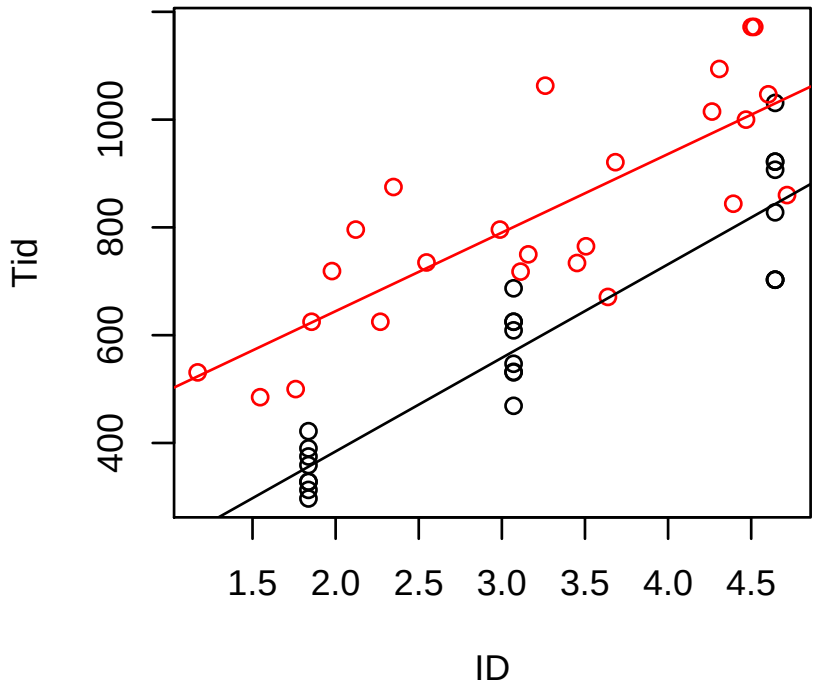
Fitts lov: Tid til at flytte mus er lineær i ID $= \log_2\left(\frac{A+W}{W}\right)$

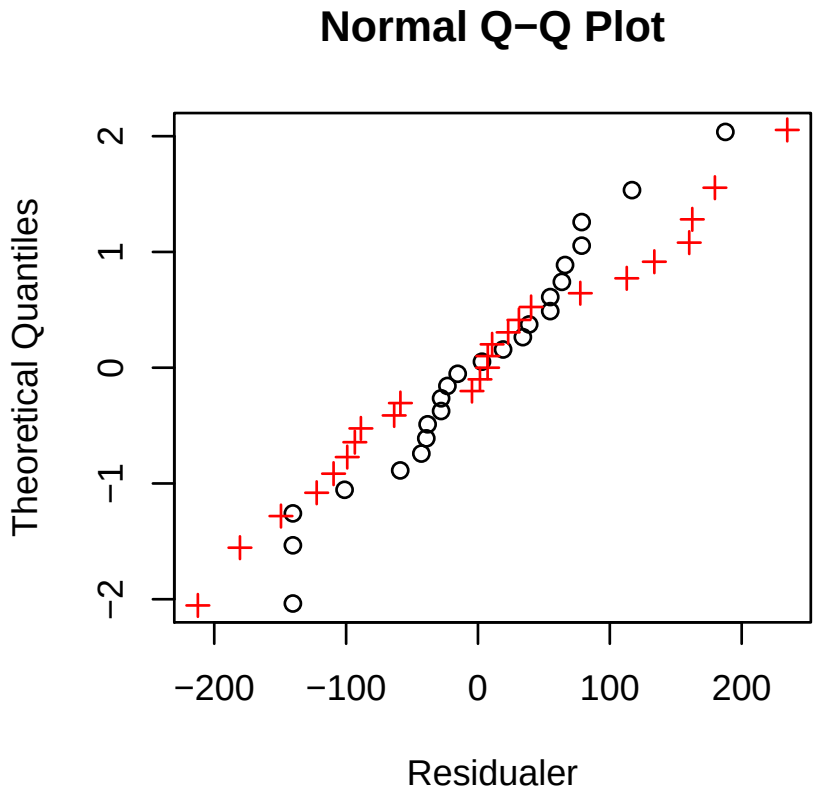
Rød: cirkeldesign, sort: rektangeldesign

Data

Eksperiment	Afstand	Bredde	Tid
Rektangel	370	50	625
Rektangel	370	50	625
Rektangel	370	50	547
Rektangel	370	50	531
Rektangel	370	50	469
Rektangel	370	50	609
Rektangel	370	50	532
Rektangel	370	50	687
Rektangel	240	10	703
Rektangel	240	10	922
⋮			
Cirkel	278	11	860
Cirkel	283	37	718
Cirkel	40	32	531
Cirkel	233	10	1047
Cirkel	191	50	625
Cirkel	179	18	734

Plot af data





1) Teste lineær sammenhæng for Rektangel-data (sort) (sidste uge)

2) Teste samme varians omkring de to linjer

3) Teste samme hældning i de to linjer

TidRekt, IdRekt: målinger fra Rektangeldata

TidCirc, IdCirc: målinger fra Cirkeldata

Tid = (TidRekt,TidCirc)

Id = (IdRekt,IdCirc)

Design=factor(rep(c("Rekt","Circ"),c(24,25)),levels=c("Rekt","Circ"))

levels=c("Rekt","Circ") for at undgå lexikografisk ordning

Teste samme varians

Model:

$$\text{TidRekt}_i \sim N(\alpha_1 + \beta_1 \cdot \text{IdRekt}_i, \sigma_1^2)$$

$$\text{TidCirc}_i \sim N(\alpha_2 + \beta_2 \cdot \text{IdCirc}_i, \sigma_2^2)$$

$(\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ kan variere frit

Teste $\sigma_1^2 = \sigma_2^2$

lmUD1=lm(TidRekt∼IdRekt); lmUD2=lm(TidCirc∼IdCirc)
$s_{\text{Rekt}} = 85, \quad s_{\text{Circ}} = 122$

var.test(lmUD1,lmUD2) (bartlett.test(list(lmUD1,lmUD2))

Resultat: p-værdi = 0.10
Konklusion: data strider ikke mod samme varians i de to eksperimenter

Prøv selv i R: mtcars

Undersøg om der er samme varians omkring linje for 4, 6 eller 8 cylindre
Model: mpg $\sim N(\alpha_{D_i} + \beta_{D_i}\text{vaegt}_i, \sigma^2_{D_i})$, hypotese: $\sigma^2_4 = \sigma^2_6 = \sigma^2_8$

```
mpg=mtcars[,1]
vaegt=mtcars[,6]
Design=factor(mtcars[,2])

lmUD1=lm(mpg[Design=="4"]~vaegt[Design=="4"])
lmUD2=lm(mpg[Design=="6"]~vaegt[Design=="6"])
lmUD3=lm(mpg[Design=="8"]~vaegt[Design=="8"])
```

Modelkontrol: plot(vaegt[Design=="4"],lmUD1$residuals) og qqnorm(lmUD1$residuals,datax=TRUE)

```
bartlett.test(list(lmUD1,lmUD2,lmUD3))
```

Konklusion: Data strider ikke mod samme varians omkring linjen da ...

Gruppespecifik regressionskoefficient

Tid: alle tidsmålinger
Id: alle Id-værdier
Design: factor med værdier Rekt og Circ

Model: $\text{Tid}_i \sim N(\alpha_{\text{Design}_i} + \beta_{\text{Design}_i} \cdot \text{Id}_i, \sigma^2)$

Model hvor hvert eksperiment har sin egen lineære sammenhæng

Modelformel, R: Tid ~ Design*Id samme som Tid ~ Design+Design*Id

Design giver α_{Rekt} og α_{Circ}

Design*Id giver $\beta_{\text{Rekt}} \cdot \text{Id}_i$ og $\beta_{\text{Circ}} \cdot \text{Id}_i$

Vektoren af middelværdier kan skrives som $\boldsymbol{\xi} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \beta_1 \mathbf{t}_1 + \beta_2 \mathbf{t}_2$

$$\mathbf{e}_1 = \begin{cases} 1 & \text{på gruppe 1} \\ 0 & \text{på gruppe 2} \end{cases} \qquad \mathbf{e}_2 = \begin{cases} 0 & \text{på gruppe 1} \\ 1 & \text{på gruppe 2} \end{cases}$$

$$\mathbf{t}_1 = \begin{cases} t_i & \text{på gruppe 1} \\ 0 & \text{på gruppe 2} \end{cases} \qquad \mathbf{t}_2 = \begin{cases} 0 & \text{på gruppe 1} \\ t_i & \text{på gruppe 2} \end{cases}$$

Generelt: underrum L udspændes af søjlerne i $\mathbf{H}$, $\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}\mathbf{x}$$

Antag $\sum_{\mathrm{I}} t_i = \sum_{\mathrm{II}} t_i = 0$, hvor I og II refererer til gruppe 1 og gruppe 2
Lad $\mathbf{H} = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{t}_1, \mathbf{t}_2)$

$$\mathbf{H}^{\mathsf{T}}\mathbf{H} = \begin{pmatrix} n_1 & 0 & 0 & 0 \\ 0 & n_2 & 0 & 0 \\ 0 & 0 & \mathrm{SS}_1 & 0 \\ 0 & 0 & 0 & \mathrm{SS}_2 \end{pmatrix}, \quad \mathrm{SS}_1 = \sum_{\mathrm{I}} t_i^2,\ \mathrm{SS}_2 = \sum_{\mathrm{II}} t_i^2$$

$$\mathbf{H}^{\mathsf{T}}\mathbf{x} = \left(\sum_{\mathrm{I}} x_i, \sum_{\mathrm{II}} x_i, \sum_{\mathrm{I}} t_i x_i, \sum_{\mathrm{II}} t_i x_i\right)^{\mathsf{T}}$$

Dette giver estimaterne

$$\hat{\alpha}_1 = \frac{\sum_{\mathrm{I}} x_i}{n_1},\ \hat{\alpha}_2 = \frac{\sum_{\mathrm{II}} x_i}{n_2},\ \hat{\beta}_1 = \frac{\sum_{\mathrm{I}} t_i x_i}{\mathrm{SS}_1},\ \hat{\beta}_2 = \frac{\sum_{\mathrm{II}} t_i x_i}{\mathrm{SS}_2}$$

Parametrisering i R af modellen: Design+Design*Id

$$\text{Intercept} = \alpha_{\text{Rekt}}, \quad \text{DesignCirc} = \alpha_{\text{Circ}} - \alpha_{\text{Rekt}}$$

$$\text{Id} = \beta_{\text{Rekt}}, \quad \text{DesignCirc:Id} = \beta_{\text{Circ}} - \beta_{\text{Rekt}}$$

```
Design=factor(rep(c("Rekt","Circ"),c(24,25)),levels=c("Rekt","Circ"))

summary(lm(Tid~F+F*Id))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     37.10      63.32   0.586  0.56089
DesignCirc     315.89      91.46   3.454  0.00122
Id             173.62      18.71   9.280 5.12e-12
DesignCirc:Id  -27.82      27.03  -1.029  0.30888
Residual standard error: 105.3 on 45 degrees of freedom
```

Bemærk *levels=...* for at undgå lexikografisk ordning

Teste samme hældning

Stiger *Tid* med *index of diffculty* på samme måde i de to eksperimenter?

Hypotese: $\beta_{\text{Circ}} = \beta_{\text{Rekt}}$

Teste reduktion fra model $\text{Tid}_i \sim N(\alpha_{\text{Design}_i} + \beta_{\text{Design}_i} \cdot \text{Id}_i, \sigma^2)$
til $\text{Tid}_i \sim N(\alpha_{\text{Design}_i} + \beta \cdot \text{Id}_i, \sigma^2)$

fra model Tid~Design*Id til model Tid~Design+Id

Fortolkning af model:

Additivitet: uanset eksperiment er der samme forskel i middelværdi mellem to værdier af *Id*

uanset værdien af *Id* er der samme forskel i middelværdi mellem de to eksperimenter

Teste samme hældning

Generelle F-test:

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

```
anova(lm(Tid~Design+Id),lm(Tid~Design*Id))

Model 1: Tid ~ Design + Id
Model 2: Tid ~ Design * Id
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1     46 510747.7
2     45 499001.3  1   11746.4 1.0593  0.3089
```

Konklusion: data strider ikke mod hypotesen om samme hældning $\beta_{\text{Rekt}} = \beta_{\text{Circ}}$ (p-værdi = 0.31)

Gå tilbage til parametertabel og find p-værdi der

Teste samme skæring

Vi beskriver data med modellen $\text{Tid}_i \sim N(\alpha_{\text{Design}_i} + \beta \cdot \text{Id}_i, \sigma^2)$

teste reduktion til model $N(\alpha + \beta \cdot \text{Id}_i, \sigma^2)$

fortolkning: samme lineære sammenhæng i de to eksperimenter

```
anova(lm(Tid~Id),lm(Tid~Design+Id))

Model 1: Tid ~ Id
Model 2: Tid ~ Design + Id
  Res.Df       RSS Df Sum of Sq       F  Pr(>F)
1     47 1141624.8
2     46  510747.7  1  630877.1 56.8193  0.0000 ***
```

Konklusion: data strider kraftigt mod hypotesen om samme skæring $\alpha_{\text{Rekt}} = \alpha_{\text{Circ}}$

(Teste samme skæring i fulde model: p-værdi $= 0.0012$)

Parametertabel

```
summary(lm(Tid~Design+Id))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     79.54      48.09   1.654    0.105
DesignCirc     227.00      30.11   7.538 1.43e-09
Id             160.29      13.51  11.864 1.35e-15
Residual standard error: 105.4 on 46 degrees of freedom
```

F-test for at fjerne 1 parameter = t-test kvadreret

Fulde model: Residual standard error: 105.3 on 45 degrees of freedom

Konfidensintervaller i model $\text{Tid}_i \sim N(\alpha_{\text{Design}_i} + \beta \cdot \text{Id}_i, \sigma^2)$

```
confint(lm(Tid~Design+Id))
```

```
                2.5 %   97.5 %
(Intercept) -17.26535 176.3364
DesignCirc  166.37933 287.6124
Id          133.09496 187.4859
```

I cirkel-eksperiment er tidsforbruget cirka 200 (ms) over rektangel-eksperiment

Fordobling af sværhedsgrad: stigning på cirka 160 ms

Konfidensinterval for spredning omkring linjen:

Webbog afsnit 2.6 med $s^2 = 105.4^2$ og $df = 46$: $[88,\ 132]$

Prøv selv i R: mtcars

Undersøg om der er samme hældning for de tre designs

Model: ..., Hypotese: ...

```
mpg=mtcars[,1]
vaegt=mtcars[,6]
Design=factor(mtcars[,2])
```

Vi bruger det generelle F-test fra afsnit 4.7

```
anova(lm(mpg~Design+vaegt),lm(mpg~Design*vaegt))
```

F=... som vurderes i en $F(..,..)$-fordeling, og p-værdien for testet er ... Da denne er under 0.05, siger vi, at ...

Vores model er nu ...

Skift til webbog afsnit 5.1

Samme model kan parametriseres på flere måder, for eksempel

$$\alpha_1,\ \alpha_2-\alpha_1,\ \alpha_3-\alpha_1,\ldots,\ \alpha_k-\alpha_1$$

eller: $\alpha_1,\ \alpha_2,\ \alpha_3,\ldots,\ \alpha_k$

Opskrevet med vektorer svarer dette til:

$$\alpha_1\mathbf{e}_1+\alpha_2\mathbf{e}_2+\cdots+\alpha_k\mathbf{e}_k$$

$$=\alpha_1(\mathbf{e}_1+\cdots+\mathbf{e}_k)+(\alpha_2-\alpha_1)\mathbf{e}_2+\cdots+(\alpha_k-\alpha_1)\mathbf{e}_k$$

Model med gruppebestemt regressionkoefficient er omtalt

Næste: Forstå fordelingsudsagn og uafhængighedsudsagn

Først: vektor af normalfordelte stokastiske variable

Fitts law for thumbs

Middelværdi og varians af vektor

Vektor af stokastiske variable: $\mathbf{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$

$$\mathrm{E}(\mathbf{Z}) = (E(Z_1), \ldots, E(Z_n))^\mathsf{T}$$

$$\mathrm{Var}(\mathbf{Z}) = \begin{pmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_n) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(Z_n, Z_1) & \mathrm{Cov}(Z_n, Z_2) & \cdots & \mathrm{Var}(Z_n) \end{pmatrix}$$

$$\mathrm{Var}(Z) = E(Z - E(Z))^2, \quad \mathrm{Cov}(Z_1, Z_2) = E(Z_1 - E(Z_1))(Z_2 - E(Z_2))$$

Kendte regneregler:

$$E\left(\sum_i a_i Z_i\right) = \sum_i a_i E(Z_i)$$

$$\mathrm{Var}\left(\sum_i a_i Z_i\right) = \sum_{i,j} a_i a_j \mathrm{Cov}(Z_i, Z_j)$$

$$\mathrm{Cov}\left(\sum_i a_i Z_i, \sum_i b_i Z_i\right) = \sum_{i,j} a_i b_j \mathrm{Cov}(Z_i, Z_j)$$

Dette giver $\mathrm{E}(\mathbf{BZ}) = \mathbf{B}\mathrm{E}(\mathbf{Z})$ og $\mathrm{Var}(\mathbf{BZ}) = \mathbf{B}\mathrm{Var}(\mathbf{Z})\mathbf{B}^\mathsf{T}, \quad \mathbf{B}: \ k \times n$

Bevis:

$$(\mathbf{BZ})_i = \sum_j B_{ij} Z_j \rightarrow E(\mathbf{BZ})_i = \sum_j B_{ij} E(Z_j) = (\mathbf{B}\mathrm{E}(\mathbf{Z}))_i$$

$$\mathrm{Cov}\left(\sum_i B_{ui} Z_i, \sum_j B_{vj} Z_j \right) = \sum_{i,j} B_{ui} B_{vj} \mathrm{Cov}(Z_i, Z_j)$$

$$= \sum_{i,j} B_{ui} \mathrm{Var}(\mathbf{Z})_{ij} \mathbf{B}^\mathsf{T}_{jv}$$

$$= (\mathbf{B}\mathrm{Var}(\mathbf{Z})\mathbf{B}^\mathsf{T})_{uv}$$

$Z_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, uafhængige, skrives som

$$\mathbf{Z} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \quad \mathbf{I} \text{ diagonal med 1 i diagonalen}$$

Da $\mathrm{E}(\mathbf{BZ}) = \mathbf{B}\mathrm{E}(\mathbf{Z})$ og $\mathrm{Var}(\mathbf{BZ}) = \mathbf{B}\mathrm{Var}(\mathbf{Z})\mathbf{B}^\mathsf{T}$ skriver vi

$$\mathbf{W} = \mathbf{BZ} \sim N_n\big(\mathbf{B}\boldsymbol{\mu}, \sigma^2 \mathbf{B}\mathbf{B}^\mathsf{T}\big),$$

Vi skriver nu $\mathbf{W} \sim N_n\big(\boldsymbol{\xi}, \boldsymbol{\Sigma}\big)$, $\boldsymbol{\xi} = \mathbf{B}\boldsymbol{\mu}$, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{B}\mathbf{B}^\mathsf{T}$ og får

$$\mathbf{AW} \sim N_n\big(\mathbf{A}\boldsymbol{\xi}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}\big)$$

idet $\mathbf{AW} = (\mathbf{AB})\mathbf{Z} \sim N_n\big((\mathbf{AB})\boldsymbol{\mu}, \sigma^2(\mathbf{AB})(\mathbf{AB})^\mathsf{T}\big)$

$$= N_n\big(\mathbf{A}\boldsymbol{\xi}, \mathbf{A}(\sigma^2 \mathbf{B}\mathbf{B}^\mathsf{T})\mathbf{A}^\mathsf{T}\big)$$

Vi har etableret notation for vektor med normalfordelte variable

Næste: fundamentale spaltningssætning (uden bevis)

Opsætning:

$\mathbf{Y} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}), \quad \mathbf{B}_1 : \ k_1 \times n, \quad \mathbf{B}_2 : \ k_2 \times n$

$\mathbf{Z}_1 = (Z_{11}, \ldots, Z_{1,k_1})^\mathsf{T} = \mathbf{B}_1 \mathbf{Y} \quad \text{og} \quad \mathbf{Z}_2 = (Z_{21}, \ldots, Z_{2,k_2})^\mathsf{T} = \mathbf{B}_2 \mathbf{Y}$

Hvis $\mathbf{B}_1 \mathbf{B}_2^\mathsf{T} = \mathbf{0}$ så er $\mathbf{Z}_1$ og $\mathbf{Z}_2$ stokastisk uafhængige

En $n \times n$ matrice $\mathbf{B}$ er en ortogonal projektionsmatrice hvis og kun hvis

$$\mathbf{BB} = \mathbf{B} \text{ og } \mathbf{B}^\mathsf{T} = \mathbf{B}$$

Definer $\mathbf{Z} = (Z_1, \ldots, Z_n)^\mathsf{T} = \mathbf{BY}$. Så gælder

$$\sum_{i=1}^{n} Z_i^2 = \|\mathbf{BY}\|^2 = (\mathbf{BY})^\mathsf{T}(\mathbf{BY}) \sim \sigma^2 \chi^2(k)$$

k: rang af $\mathbf{B}$ = dimensionen af rum der projektieres ned på

Spaltningssætningen er omtalt

Bruge sætningen

Et normalfordelt observationssæt

$X_i \sim N(\mu, \sigma^2),\ i = 1, \ldots, n, \quad Y_i = X_i - \mu \sim N(0, \sigma^2)$

$\bar{X} = \frac{X_1}{n} + \ldots + \frac{X_n}{n} = \bar{Y} + \mu = \mathbf{B}_1\mathbf{Y} + \mu$

$\mathbf{B}_1 = \mathbf{e}^\mathsf{T} = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$

$\text{SSD} = \sum_i (X_i - \bar{X}_i)^2 = \sum_i (Y_i - \bar{Y})^2 = \|(\mathbf{I} - \mathbf{E})\mathbf{Y}\|^2 = \|\mathbf{B}_2\mathbf{Y}\|^2$

$\mathbf{E}$, $n \times n$, alle søjler lig med $\mathbf{e}$, $\mathbf{B}_2 = \mathbf{I} - \mathbf{E}$

Betingelse fra spaltningssætningen:

$\mathbf{B}_1\mathbf{B}_2^\mathsf{T} = \mathbf{e}^\mathsf{T}(\mathbf{I} - \mathbf{E}) = (\frac{1}{n} - \frac{1}{n})(1, 1, \ldots, 1) = \mathbf{0}$

$\bar{X}$ og SSD er uafhængige og $\text{SSD} \sim \sigma^2\chi^2(n-1)$ idet

$\mathbf{B}_2\mathbf{B}_2 = (\mathbf{I} - \mathbf{E})(\mathbf{I} - \mathbf{E}) = \mathbf{I} - 2\mathbf{E} + \mathbf{E}\mathbf{E} = \mathbf{I} - \mathbf{E} = \mathbf{B}_2$

Uafhængighed mellem skøn over parametre i middelværdi og variansskøn: webbog afsnit 6.2

Model: $\boldsymbol{\xi} \in L$, L udspændt af søjlerne i $\mathbf{H}$: $\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \mathbf{B}_1\mathbf{Y} + \boldsymbol{\theta}, \quad \mathbf{B}_1 = (\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}, \quad \mathbf{Y} = \mathbf{X} - \boldsymbol{\xi}$$

Projektionsmatrix: $\mathbf{P} = \mathbf{H}(\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\mathbf{H}^\mathsf{T}$

$$\mathrm{SSD}(M) = \|\mathbf{B}_2\mathbf{Y}\|^2, \quad \mathbf{B}_2 = \mathbf{I} - \mathbf{P}$$

Betingelse i spaltningssætningen:

$$\mathbf{B}_1\mathbf{B}_2^\mathsf{T} = \mathbf{P}(\mathbf{I} - \mathbf{P}) = \mathbf{P} - \mathbf{P}\mathbf{P} = \mathbf{P} - \mathbf{P} = \mathbf{0}$$

Spaltningssætningen har givet os de resultater vi har brugt tidligere

Slut for i dag - efter en meget teoritung dag