

Lineaarsed mudelid 7. praktikum

Usaldusellips(oid)

Näide 1

Vaatame lihtsat regressioonanalüüsi mudelit

Genereerime andmed

```
RNGkind(sample.kind = "Rejection")
set.seed(1)
x=runif(100,0,6)
y=2+1.9*x+rnorm(100)
and=data.frame(x,y)
```

Hindame mudeli

```
m=lm(y~x, data=and)
summary(m)
```

Huvitagu meid näiteks küsimus, kas nii vabaliige kui ka sirge tõus võiksid mõlemad olla võrdsed 2-ga. Võime sellele küsimusele muidugi proovida leida vastust tavaviisil:

```
> confint(m)
                2.5 %    97.5 %
(Intercept)  1.412247  2.229102
x             1.835153  2.068962
```

Kust näeme, et 2 jääb kenasti mõlema parameetri korral 95%-usaldupiiridesse. Arvestades muidugi, et vaatame kahte parameetrit korraga, peaksime kasutama Bonferroni korrektsiooni (ühe usalduspiiri jaoks lubatud vea tõenäosuse 0,05 asemel lubama vea tõenäosust 0,025):

```
> confint(m, level=0.975)
                1.25 %    98.75 %
(Intercept)  1.352172  2.289177
x             1.817957  2.086157
```

Mis teeks usalduspiire muidugi vaid laiemaks. Proovime aga antud hüpoteesipaari koos testida F-testi abil:

$$\left(\Lambda^T \hat{\beta} - \Lambda^T \beta_0\right)^T \left(\Lambda^T (\mathbf{X}^T \mathbf{X})^{-1} \Lambda \cdot \text{MSE}\right)^{-1} \left(\Lambda^T \hat{\beta} - \Lambda^T \beta_0\right) / \text{rank}(\Lambda) \stackrel{H_0}{\sim} F_{\text{df}_1=\text{rank}(\Lambda); \text{df}_2=n-\text{rank}(\mathbf{X})}.$$

Ehk R-i programmina:

```
Lambda=rbind(c(1,0), c(0,1))
stat=t(Lambda%*%coef(m)-c(2,2))%*%
      (solve(Lambda%*%vcov(m))%*%t(Lambda))/2)%*%
      (Lambda%*%coef(m)-c(2,2))
stat
1-pf(stat, df1=2, df2=98)
```

Mis annab meile tulemuseks pisikese p-väärtuse (järelikut pole võimalik, et nii vabaliige kui ka sirge tõus oleksid 2-d).

Mitut hüpoteesi saab korraga testida (kuidagimoodi) ka estimable-käsu abil. Viimane käsk küll eeldab, et teame tegelikku vaatluste hajuvust (ignoreerib seda, et me tegelikult hindame jääkide dispersiooni):

```
Lambda=rbind(c(1,0), c(0,1))

library(gmodels)
estimable(m, Lambda, beta0=c(2,2), joint.test=TRUE)
```

Märkus: sama tulemuse saaksime ise arvutusi tehes siis, kui kasutaksime valemit:

$$\left(\Lambda^T \hat{\beta} - \Lambda^T \beta_0\right)^T \left(\Lambda^T (\mathbf{X}^T \mathbf{X})^{-1} \Lambda - \sigma^2\right)^{-1} \left(\Lambda^T \hat{\beta} - \Lambda^T \beta_0\right) \stackrel{H_0}{\sim} \chi^2_{df=\text{rank}(\Lambda)}$$

ja asendaksime selles valimis jääkide dispersiooni jääkide dispersiooni hinnanguga:

```
1-pchisq(stat*2, df=2)
```

Seega antud juhul on näha, et ise arvutades on võimalik saada vahel ka täpsemaid tulemusi võrreldes R-i standardfunktsioonidega.

Sama järelduseni oleksime võinud jõuda muidugi ka parameetritele joonistatud usaldusellipsoidi vaadates:

$$\begin{aligned} \left(\Lambda^T \hat{\beta} - \Lambda^T \beta\right)^T \left(\Lambda^T (\mathbf{X}^T \mathbf{X})^{-1} \Lambda \cdot \text{MSE}\right)^{-1} \left(\Lambda^T \hat{\beta} - \Lambda^T \beta\right) / \text{rank}(\Lambda) &\leq \\ &\leq f_{1-\alpha; df_1=\text{rank}(\Lambda); df_2=n-\text{rank}(\mathbf{X})} \end{aligned}$$

ehk R-i programmi abil joonistatult:

```
omavek= eigen(vcov(m))$vectors
omavaart= eigen(vcov(m))$values

r= seq(0, 2*pi, length=1000)

k= sqrt(qf(0.95,2,98))
a= omavek%*%diag(sqrt(omavaart))%*%t(omavek)
abi= sqrt(2)*a%*%(k*rbind(sin(r), cos(r)))+coef(m)

plot(abi[1,], abi[2,], type="l")
polygon(abi[1,], abi[2,], col="gray90", border="black", lwd=3)

points(2,2, pch=20, col=2, cex=2)
```

Näeme, et punkt 2,2 ei asu 95%-usaldusellipsis, seega võime 95%-kindlusega väita, et vabaliige ja sirge tõus ei saa samaaegselt olla 2-d.

Usaldusellipsit mudeli parameetritele saab leida ka lisamooduli ellipse abil, näiteks lisa oma joonisele ellipse-käsu abil leitud usaldusellips (punasega, et näha oleks):

```
library(ellipse)
lines(ellipse(m, npoint=1000), col="red2", lwd=2)
```

Ülesanne 1

Muudame eelmises ülesandes kasutatud andmete genereerimise mehhanismi veidi. Vaata mis juhtub parameetrite usaldusellipsiga variant 1 ja variant 2 korral. Oskad sa välja mõelda, miks (mõni tingimus, millal võiksid usaldusellipsi teljed ühtida koordinaattelgedega – koos põhjendusega)?

Variant 1

```
x=runif(100,0,6)
x=x-mean(x)
y=2+1.9*x+rnorm(100)
and=data.frame(x,y)
m=lm(y~x, data=and)
```

Variant 2

```
x=runif(100,0,6)+10
y=2+1.9*x+rnorm(100)
and=data.frame(x,y)
m=lm(y~x, data=and)
```

Ülesanne 2

Genereerime teise andmestiku – dispersioonianalüüsi mudel. Joonista usaldusellips mudeli 2. ja 3. parameetri jaoks (grupp 2 mõju ja grupp 3 mõju). Selgita, kuidas usaldusellipsi pealt vaadata, kas mõlemad mõjud võivad olla ka 0-id. Anna ka oma järeldus jooniselt: kas need mõjud võivad olla samaaegselt nullid?

```
set.seed(1)
grupp=rep(1:3, c(10, 100, 1000))
y=2.7*(grupp==1)+2.6*(grupp==2)+2.4*(grupp==3)
                                     +rnorm(length(grupp))
m2=lm(y~factor(grupp))
summary(m2)
```