Usaldusriba (Confidence Band)

Olgu meil väike andmestik,

```
set.seed(4)
n=25
x=runif(n, 0,10)
y=2+0.6*x-0.1*x*x+rnorm(n,sd=1)
plot(x,y)
```

ja soovime seda andmestikku kasutades hinnata regressioonanalüüsi mudelit:

```
m1=lm(y~x+I(x**2))

abi=seq(-1,11, length=200)
yhat=predict(m1, data.frame(x=abi))
lines(abi,yhat, lwd=2)
```

Loomulikult tahame ka iseloomustada seda, kui täpselt me ikka seda mudelit teame. Üheks võimaluseks on lisada joonisele 95%-usaldusintervallid:

```
abi=seq(-1,11, length=200)
yhat=predict(m1, data.frame(x=abi), interval="confidence")

lines(abi,yhat[,2], lty=2)
lines(abi,yhat[,3], lty=2)
```

Aga nende usaldusintervallidega on üks häda: iga *x*-i väärtuse korral võime küll väita 95%-kindlusega, et tegelik regressioonjoon asub usalduspiiride sees, aga meil on joonistatud praegu väga palju erinevaid usalduspiire (arvutasime praegu tegelikult kakssada usalduspiiri....). Seega on täiesti võimalik, et joonisel kuskil (näiteks 5% ulatuses) on tegelik regressioonjoon väljaspool usalduspiire...

Kuna antud juhul on tegemist meie poolt genereeritud andmetega, siis me teame tegelikku seost. Kanname ka tegeliku x-tunnuse ja y-tunnuse vahelise seose joonisele:

```
# Tegelik seos
Ey=2+0.6*abi-0.1*abi^2
lines(abi,Ey, lwd=2, col="red")
```

Näemegi – nagu kahtlustasime – et tegelikku seost iseloomustav regressioonjoon sattub korraks väljapoole leitud 95%-usalduspiire (näiteks x=3,5 korral).

Tee 20 simulatsiooni (iga kord uute andmetega, ilma set.seed-käsku R'le andmata). Loe kokku mitme genereeritud andmestiku korral sattus tegelik regressioonijoon joonisel väljapoole 95%-usaldusvahemikku (silma järgi hinnates):

Tegeliku parameetri väärtuse (näiteks keskväärtust) äraarvamisel peaks 95%-usaldusintervall eksima ootuspäraselt ühel korral 20 valimi kohta. Teadmata küll teie simulatsioonide tulemusi arvan siiski, et tegelik regressioonjoon sattus joonistel korraks väljapoole usaldusintervalle märksa enam kui üks kord.

Kuidas saada regressioonjoonele selliseid usalduspiire, et 95% kindlusega saaksime väita – tegelik regressioonjoon asub täies ulatuses joonistatud piiride vahel – joonistatud usaldusriba sees (95% valimitest on sellised, et regressioonjoon jääb täies ulatuses joonisele kantud usaldusriba sisse)?

Appi tuleb Scheffé meetod, mis võimaldab leida samaaegseid usalduspiire (võime nõuda, et kõik leitud usalduspiirid sisaldaksid parameetri õiget väärtust tõenäosusega 1-α ehk maksimaalselt tõenäosusega α võib leitud usalduspiiride seas leiduda mõni ekslik usaldusvahemik mis tegelikku keskväärtust ei kata):

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}-\sqrt{(p-p_0)\mathrm{D}\widehat{\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}}\ \leq\right.\right.$$
$$\left.\left.\leq\ \boldsymbol{\lambda}^T\boldsymbol{\beta}\leq\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}+\sqrt{(p-p_0)\mathrm{D}\widehat{\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}}\right\}\right)=1-\alpha$$

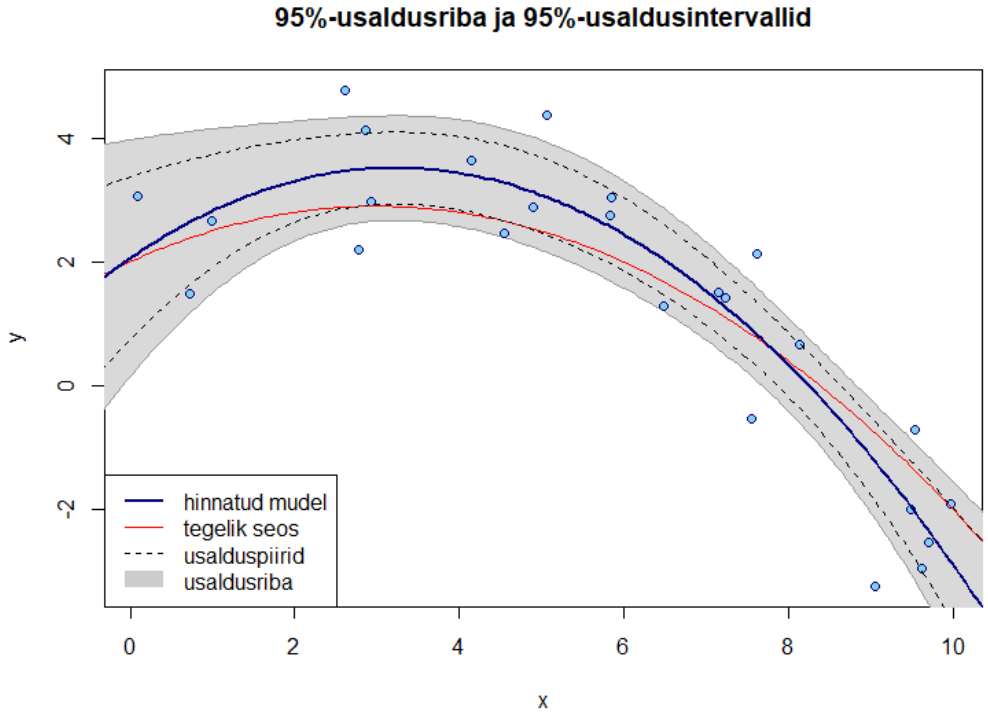
Kuna antud juhul tahame leida usalduspiire suurustele $c_0 + c_1 x + c_2 x^2$ ehk meid huvitavad parameetrite lineaarkombinatsioonid $\lambda^T\beta$ sisaldavad kõiki meie mudeli parameetreid siis $p_0=0$ ja vajalikud 95% samaaegsed usalduspiirid (95%-usaldusriba; 95%-confidence band) on leitav järgmisel viisil:

```
# 95%-usaldusriba Scheffe meetodil:
yhat=predict(m1, data.frame(x=abi), se.fit=TRUE)
ylemine= yhat$fit +sqrt(3*qf(0.95, 3, 25-3))*yhat$se.fit
alumine= yhat$fit -sqrt(3*qf(0.95, 3, 25-3))*yhat$se.fit

lines(abi, alumine, col="gray60")
lines(abi, ylemine, col="gray60")
```

Saadud usaldusriba on veidi laiem eelnevalt joonistatud punktiviisilistest usaldusintervallidest (*95%-pointwise confidence intervals*) ja peaks kogu joonise ulatuses sisaldama tegelikku seost.

Sinu saadud tulemus peaks välja nägema umbes selline:



Genereeri ka nüüd 20 andmestikku. Mitmel juhul sattus tegelikku regressioonseost kirjeldav punane sirge väljapoole usaldusriba?

……. / 20

Ootuspäraselt peaksid kahekümnest katse korral nägema ühte ekslikku usaldusriba. Väga halva õnne korral näed ehk kahte usaldusriba, kust tegelik seos korraks välja sattub.

Näide 2.

Vaatame veidi keerulisemat mudelit (taas genereeritud andmed, tõde on teada).

```
set.seed(4)
n=50
x1=runif(n, 0,10)
x2=runif(n, 0,10)
y=2+0.6*x1-0.1*x1^2+0.4*x2+0.1*x2^2+rnorm(n)

# Statistik hindab andmete põhjal mudeli
m2=lm(y~x1+I(x1**2)+x2+I(x2^2))
summary(m2)
```

Mitmese regressiooni korral soovime sageli joonisel iseloomustada ühe tunnuse mõju – milline on meie mudeli järgi näiteks seos x_1 ja y-tunnuse vahel? Kuidas x_1 väärtuse suurenemine muudaks y-tunnuse keskväärtust? Kuna mudelis pole koosmõjusid on x_1 mõju kõigi x_2 väärtuste korral samasugune ja seega saame antud mõju ühe lihtsa joonise abil iseloomustada. Teemegi x_1 mõju iseloomustava joonise:

```
abix=seq(0,11, length=1000)
M=cbind(0, abix, abix^2, 0, 0)
colnames(M)=NULL
abi=estimable(m2, cm=M)

plot(abix, abi[,1], type="l", lwd=2, xlim=c(0, 10),
xlab="x1", ylab="E(y|x1)-E(y|x1=0)",
main="Tunnuse x1 mõju keskväärtusele")

# Soovi korral lisa tegelik tunnuse x1 mõju:
# lines(abix, 0.6*abix-0.1*abix^2, col=2)
```

Märkus programmi kohta: kui soovitakse iseloomustada tunnuse x_1 mõju y-tunnuse keskväärtusele siis joonistatakse sageli graafik mis näitab y-tunnuse prognooside muutumist x_1 väärtuste muutumise korral mingi konkreetse x_2 tunnuse väärtuse korral (näiteks millised on mudeli prognoosid x_2=mean(x_2) korral). Saadud graafiku üldine kuju on täpselt samasugune kui meie poolt joonistatud graafikul (y-teljel väärtused muutuksid) kuid näiteks tulemuseks saadud usalduspiirid tuleksid sellise konkreetset x_2 väärtust kasutava lähenemise korral laiemad – mõnel juhul isegi märksa laiemad. Oskad sa arvata miks?

Antud joonisele on väga lihtne lisada 95%-punktiviisilisi usalduspiire:

```
abi=estimable(m2, cm=M, conf.int=0.95)
lines(abix, abi[,6], lty=2)
lines(abix, abi[,7], lty=2)
```

Aga soovime näha samal joonisel ka Scheffe meetodil leitud 95%-usaldusriba!

Ülesanne 1

Lisa tehtud joonisele 95%-usaldusriba (kasutades Scheffé meetodit). Sooviks näha nii programmi kui ka programmi abil leitud joonist!

Vihjeks: antud ülesande korral $p_0 \neq 0$!