

Lineaarsed mudelid  
3. praktikum  
**Suurima tõepära hinnang ja  
ühe suure kuriteo lugu**

Tekitame esmalt miniandmestiku oma edaspidiste katsetuste jaoks:

```
set.seed(1)
n=100
x=runif(n)
y=2+3*x+rnorm(n, sd=1.2)
```

Suurima tõepära hinnangu leidmine algab enamasti tõepärafunktsiooni kirjapanemisega. Teeme seda meiegi (täida ise lüngake saamaks korrektselt vormistatud funktsiooni):

```
log_toepara=function(parameetrid,y,X){
  sigma=parameetrid[1]
  beeta=parameetrid[-1]
  n=length(y)
  l=(-n/2)*log(2*pi*sigma**2)-0.5*t(y- .....
  return(l)
}
```

Funktsiooni maksimiseerimiseks/minimiseerimiseks saab R'is kasutada funktsiooni optim:

```
X=cbind(1, x)
```

```
tul=optim(c(1,1,1), log_toepara, y=y, X=X,
          control=list(fnscale=-1))
```

```
tul
```

Parametrite  
algväärtused

Maksimiseeritav  
funktsioon

Maksimiseeritava funktsiooni parameetrid,  
mida maksimiseerimise käigus ei muudeta

Vaikimisi minimiseerib  
optim-käsk funktsiooni  
väärtuse. Maksimumi  
leidmiseks tuleb lisada  
*fnscale*-lisaparameter.

Võrdle tulemust tavameetodil saadud hinnanguga:

```
b_hat = solve( t(X) %*% ....
P_X = X %*% ....
sigma_ML = t(y) %*% (diag(n) - .....
```

```
b_hat
sigma_ML
```

Kas tulevad samad tulemused?

Kui kõik on tehtud hästi, peaksid mõlemal viisil hinnatud parameetrid tulema lähedased. Väikesed erinevused on võimalikud, sest optim-käsk maksimeerib funktsiooni numbrilisel meetodil ning ei pruugi saavutada piisavalt suurt täpsust – kaugemad komakohad ei pruugi olla õiged. Veendume, et meie poolt arvutatud parameetrite väärtustega saame suurema tõepärafunktsiooni väärtuse kui optim-käsu abil saadud parameetritega:

```
log_toepara(tul$par, y, X)
log_toepara(c(sqrt(sigma2_ML), b_hat), y, X)
log_toepara(c(sqrt(sigma2_ML), b_hat), y, X) -
log_toepara(tul$par, y, X)
```

Võrdle saadud tulemusi ka R'i lm –käsu poolt raporteeritud väärtustega:

```
summary(lm(y~x))
```

Millised tulemused langevad lm-käsu poolt raporteeritute kokku, millised mitte?

Mida peaksid olema õppinud sellest esimesest näitest:

- kui oskad kirja panna vajalikku (log-)tõepära, siis on suurima tõepära hinnangu leidmine numbrilisi meetodeid kasutades üsnagi lihtne – lähenemine on kasutatav ka siis, kui peaksid kokku puutama keerukamate, näiteks mittelineaarsete mudelitega;
- kui oskad ise avaldada lahendi valemi, siis saadud tulemus on täpsem kui numbriliste meetodite abil leitud tulemus (lisandub märkimisväärne ajavõit ja kaob ära mure valelahendite pärast, mida numbrilised meetodid võivad vahel produtseerida).

## Osa II. Toimunud on röövimine.

Loe sisse andmestik roov.txt, kättesaadav aadressilt [www.ms.ut.ee/mart/linmud2018/roov.txt](http://www.ms.ut.ee/mart/linmud2018/roov.txt). Sisselugemist saab teha näiteks järgmise käsu abil:

```
and=read.table(url("http://www.ms.ut.ee/mart/linmud2018/roov.txt"),
header=T)
attach(and)
```

Andmestiku taustast:

Linnas Z (muuseas, Z=New York) võitis avalikul konkursil firma X õiguse korraldada parkimist. Mõnda aega hiljem tekkis valvsatel linnaametnikel kahtlus, et sissetulekud parkimisest pole need, mis nad olema peaksid. Kontrollimise käigus õnnestus varguselt tabada 5 firma X töötajat, kõik kuritöölt tabatud mõisteti ka kohtu poolt süüdi. Seepeale otsustas linn firmalt X nõuda sisse saamata jäänud tulu (firma X kohustus on ka töötajate järelevalve ja töö selline organiseerimine, et vargused oleksid võimatud). Kohus leidis linna nõude põhjendatud olevat ning käskis linnal esitada kalkulatsioon kahju võimaliku suuruse kohta.

Kahju suuruse leidmiseks ongi kasutada andmefail roov.txt, mis sisaldab järgmiseid tunnuseid:

AEG           – aeg kuudes: 1=Mai 1977, 47 = Märts 1981  
RAHA         – linnale antud parkimistasu summa  
LINN         – raekoja esiselt parkimisplatsilt usaldusväärsete linnaametnike poolt  
              kogutud parkimistasu suurus  
FIRMAX       – tunnus, mis näitab, kas antud kuul kogus parkimistasusid FIRMAX või mitte  
              (0-kogus keegi teine; 1- parkimistasusid kogus firma X).

Lühipilt andmetele:

```
algus=ISOdate(1977, 5,1)
kuup=seq(algus, len=length(RAHA), by="1 month")
plot(kuup, RAHA, type="l", xlab="Aeg")
abline(v=range(kuup[FIRMAX==1]), lty=2)

points(kuup, RAHA, col=c("red","blue")[2-FIRMAX], pch=21,
       bg=c("yellow","skyblue")[2-FIRMAX], cex=1.2)
text(ISOdate(1979, 5,1), 900000, "röövimiste \n ajajärk" )
arrows(min(kuup[FIRMAX==1]), 800000, max(kuup[FIRMAX==1]), 800000,
       code=3, length=0.15)
```

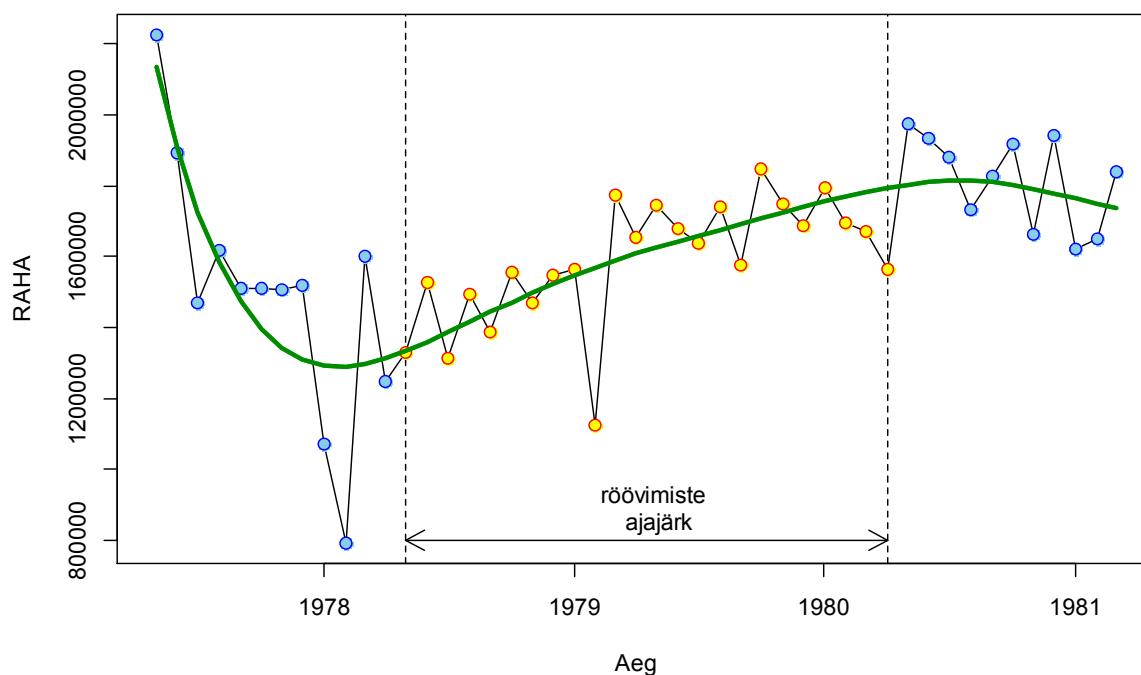
## Tänane ülesanne

Leia hinnang röövitud parkimistasude kogusummale!

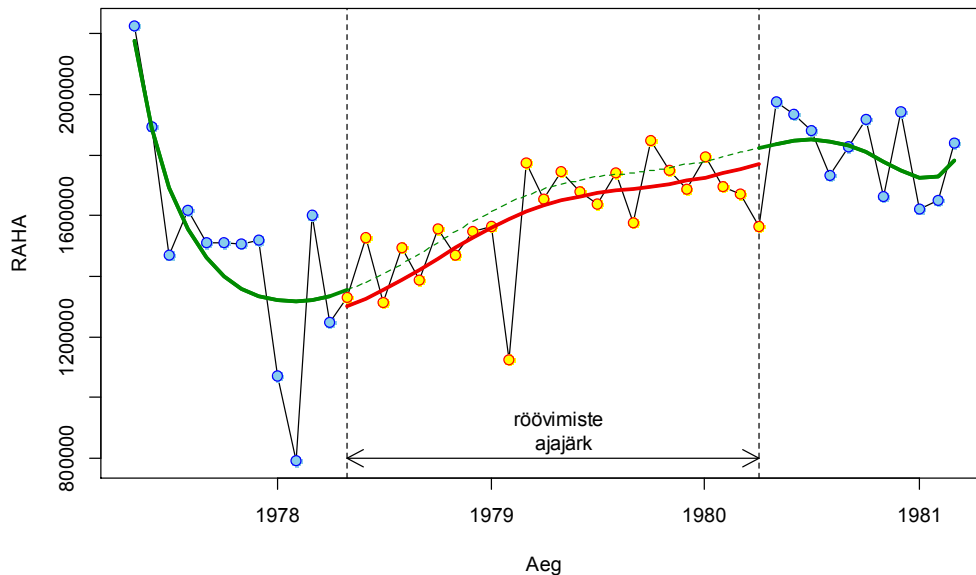
## Vihjed

NB! Hea vastuseni on võimalik liikuda ka mõnda teist teed pidi!

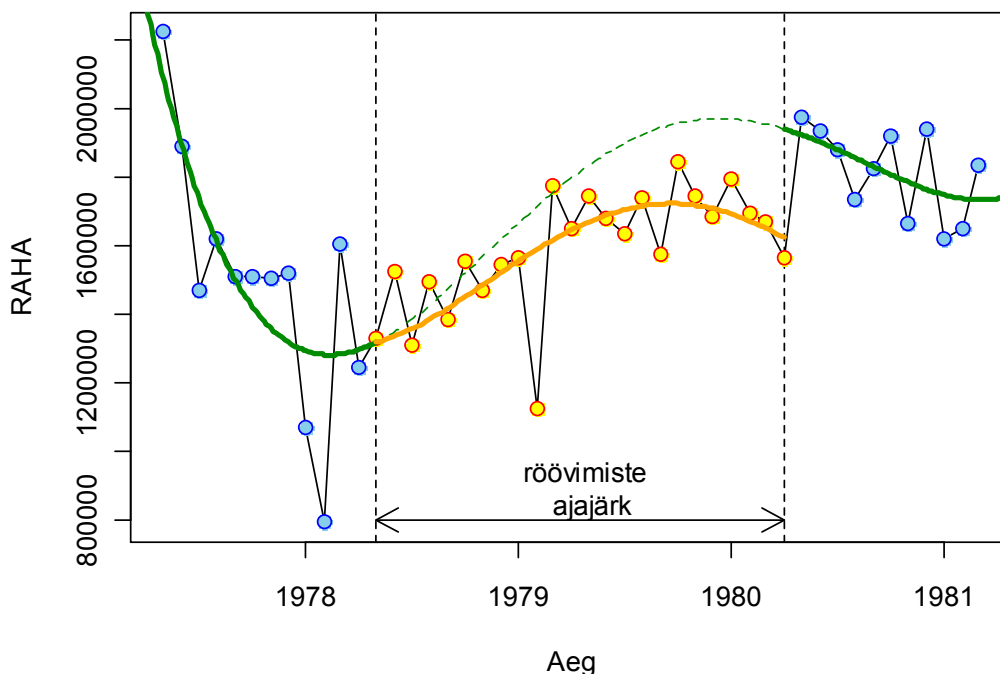
1. Hinda mudel 1, mis kirjeldab, kuidas sõltuvad laekuvad parkimistasud ajast!



2. Hinda röövimist kirjeldav mudel (mudel 2). Luba laekuvate parkimistasude kõveral olla röövimiste ajajärgul madalamal. Näiteks eeldades, et ajajärgul kui parkimistasusid korjas firma X, läks iga kuu mingi summa kõrvale (osa raha rööviti)!



3. Mõtle, kas hinnatud mudel 2 kirjeldab röövlite tegevust adekvaatselt! Järsku röövlid ei varastanud igal kuul samapalju, äkki hoopis röövlite isu ajapikku kasvas? Hinda mudel, mis lubaks röövlite saagihimul ajapikku suureneda (mudel 3), vaata ka lisatud joonist:



4. Võrdle mudeleid 2 ja 3, kumb tundub adekvaatsem?
5. Leia, millised oleksid pidanud olema laekumised siis, kui röövimisi poleks toimunud...
6. Soovi korral paranda mudeli prognoosivõimet kasutades täiendavaid tunnuseid (LINN)