Lineaarse regressiooni eeldused

- Kuidas näha probleeme ja vigu?
- Mida teha, kui midagi on korrast ära?

Eelmises loengus ja praktikumis

- Hindasime lineaarse regressioonmudeli
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
parameetreid β_0, β_1;
- kirjeldasime usaldusvahemiku abil parameetrite hinnangute täpsust;
- testisime, kas Y väärtuste prognoosimisel on kasu tunnuse X väärtuste teadmisest;
- kirjeldasite Y ja X vahelise seose tugevust (determinatsioonikordaja R^2 või korrelatsioonikordaja r-i abil) ja tegime veel palju muudki.

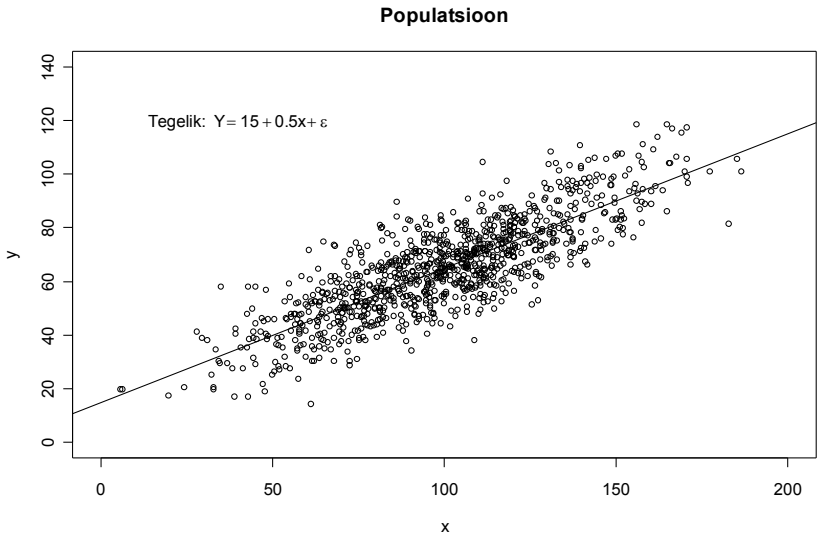
Paraku on saadud tulemused õiged ja kasutuskõlblikud vaid siis, kui on täidetud lineaarse regressioonanalüüsi eeldused.

Eeldused

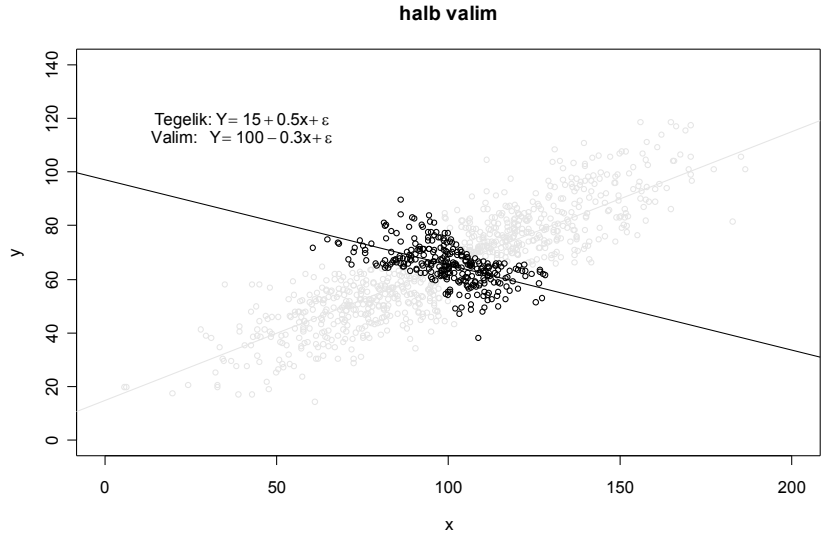
- Esindav valim (lihtne juhuslik valim);
- mudeli kuju on õige;
- uuritava tunnuse (*Y*) hajuvus ei muutu seletava ehk sõltumatu tunnuse (*X*) väärtuste muutudes;
- mudeli jäägid on normaaljaotusega.

Lisaks: vajadus kirja panna teadaolevat lisainformatsiooni…

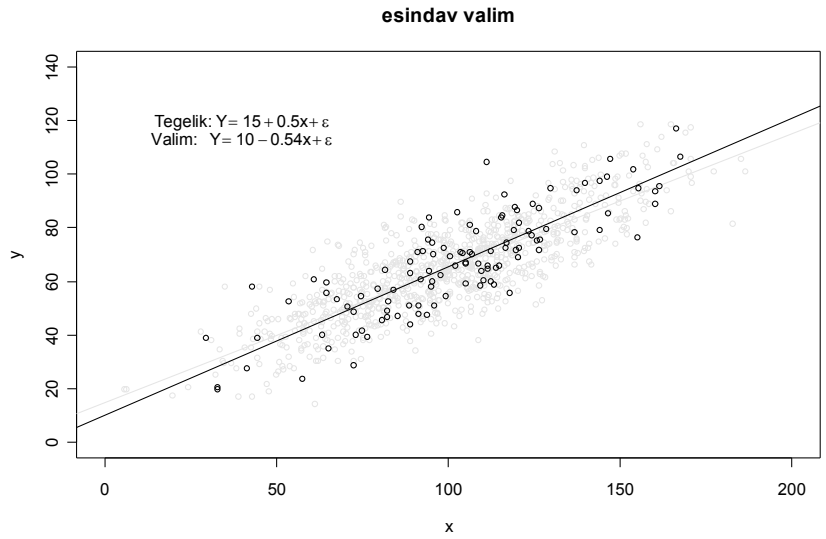
Esindav valim



Esindav valim



Esindav valim



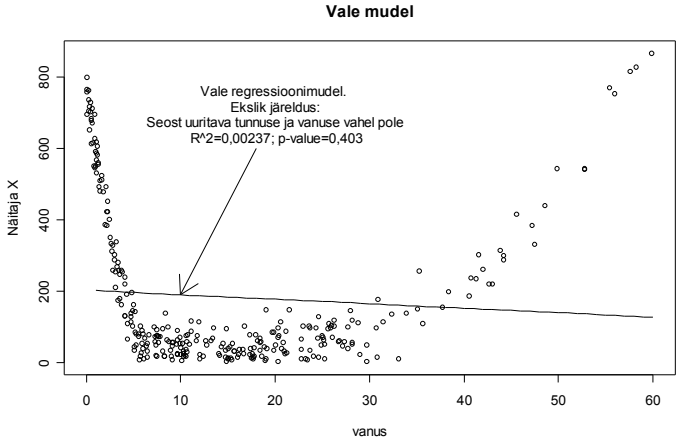
Esindav valim

Kasutatav valim peab kirjeldama populatsiooni õiglaselt ja objektiivselt, olema esindav (representative sample). Eeldatakse, et igal objektil, mis kuulub populatsiooni, on võrdne võimalus sattuda valimisse ja objekti valimisse sattumise tõenäosus ei sõltu sellest millised objektid on juba valituks osutunud.

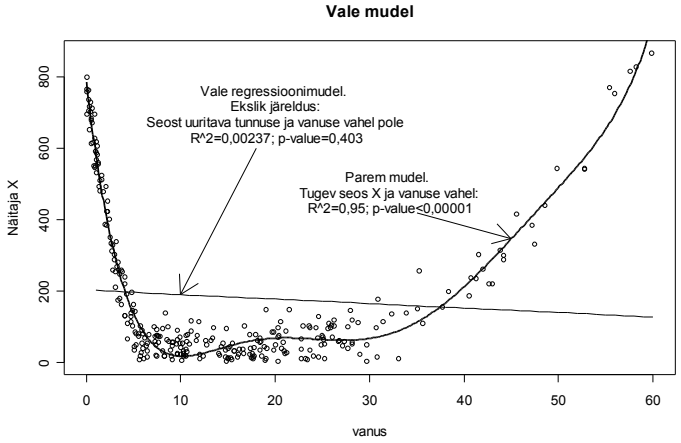
Näiteid võimalikest probleemidest:
- vastama nõustuvad inimesed ei pruugi olla esindavad kõigi inimeste jaoks;
- katsepõllud ei pruugi olla „tüüpilised“ põllud, katsepõldude eest hoolitsemine ei pruugi olla „tüüpiline“;
- pikast küsimustikust tüdinud inimesed vastavad küsimustele juhuslikult;
- uuringusse võivad hooletu katseplaani korral sagedamini sattuda need objektid, mida on mugavam kätte saada (proovid põllu või metsa servast, mitte keskelt; kohtadest kuhu saab mugavalt autoga ligi aga mitte metsataludest või saartelt jne).

Seda, kas antud valim on esindav, ei saa vaid oma valimit kasutades kontrollida.

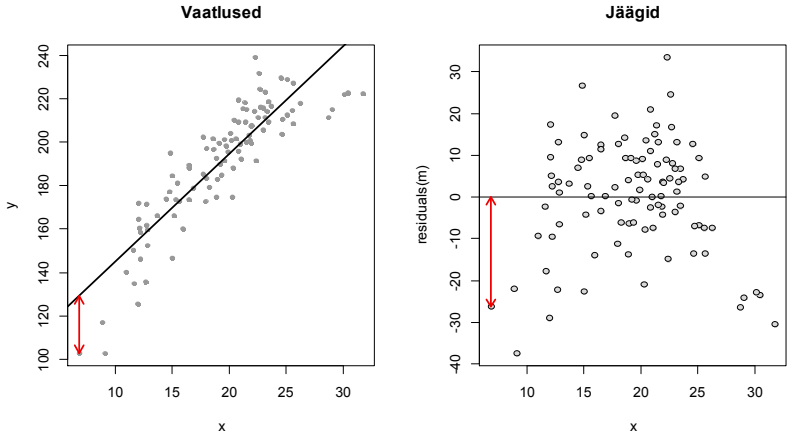
Eeldus II. Seos tunnuste vahel peab olema lineaarne (või selline, nagu mudel ütleb).



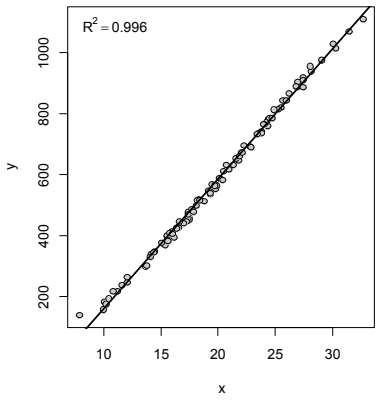
Eeldus II. Seos tunnuste vahel peab olema lineaarne (või selline, nagu mudel ütleb).



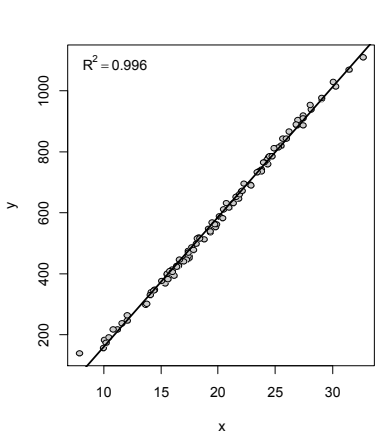
Kuidas märgata?

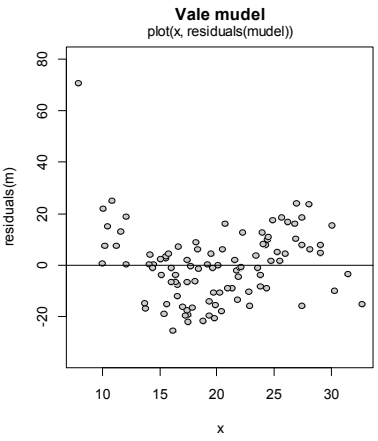


Miks uurida jääke?

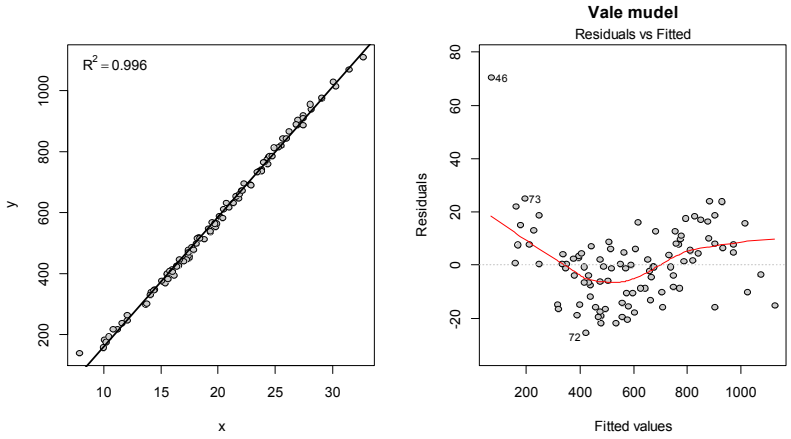


Miks uurida jääke?

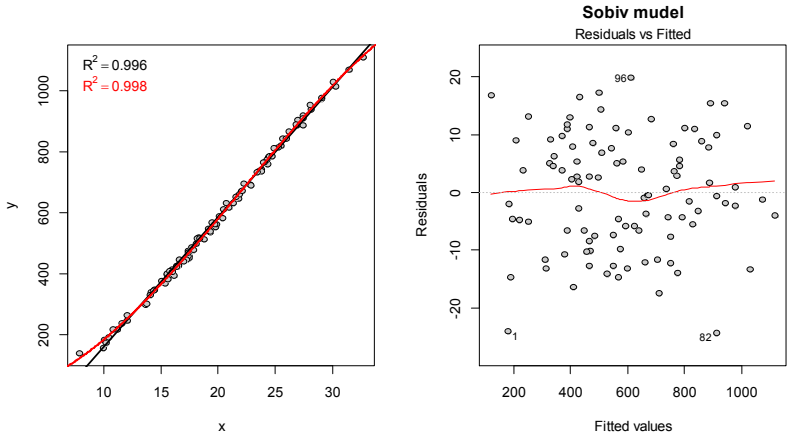




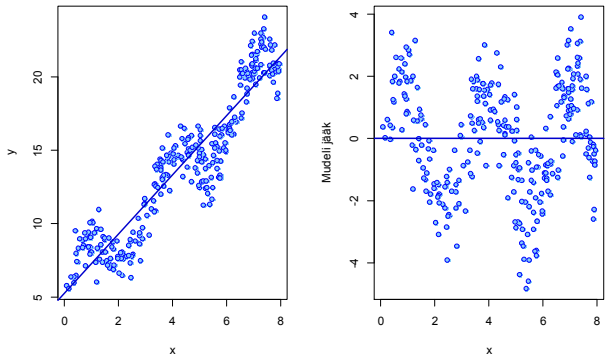
Miks uurida jääke?



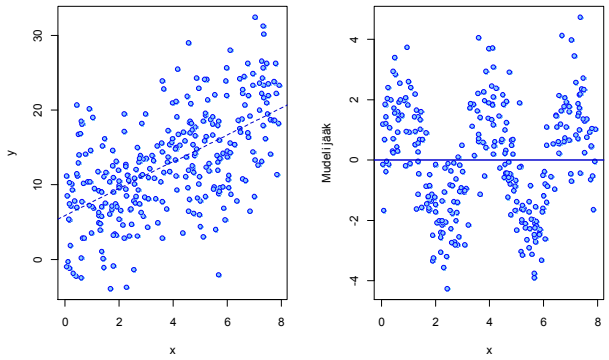
Miks uurida jääke?



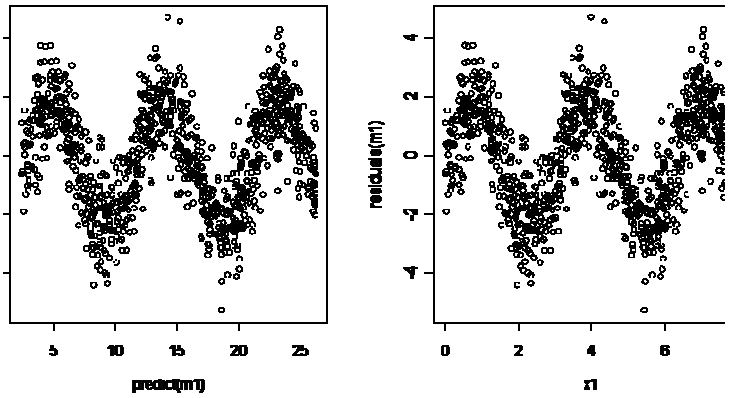
Üks sõltumatu tunnus



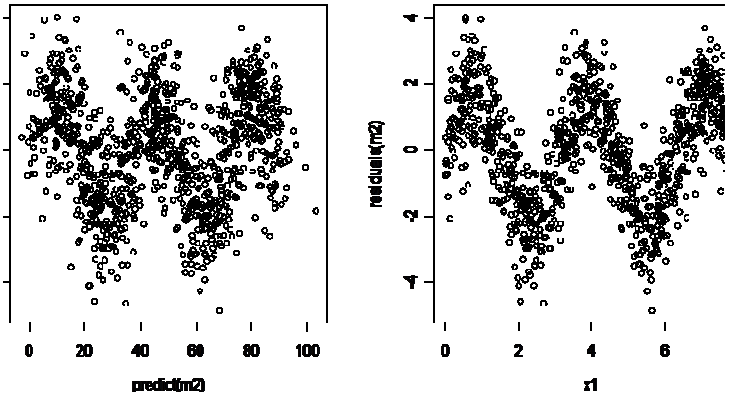
Kaks sõltumatut tunnust



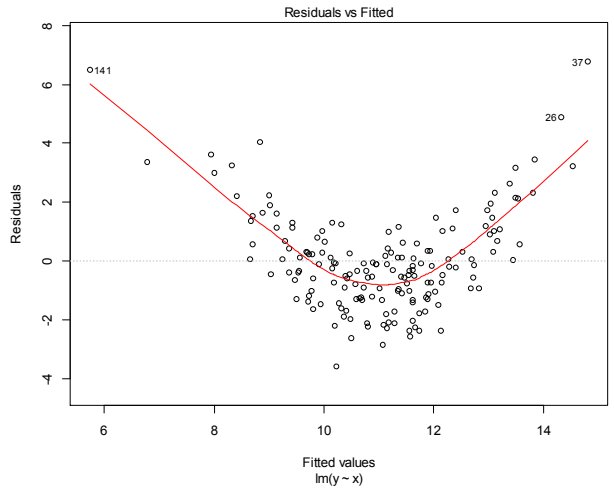
Kas siis, kui mudelis on 100 tunnust, peame tegema 100 graafikut või on olemas ka lihtsam pääsetee???
Mudelis 1 sõltumatu tunnus...



Kas siis, kui mudelis on 100 tunnust, peame tegema 100 graafikut või on olemas ka lihtsam pääsetee???
Mudelis 2 sõltumatut tunnust...



plot(m1)



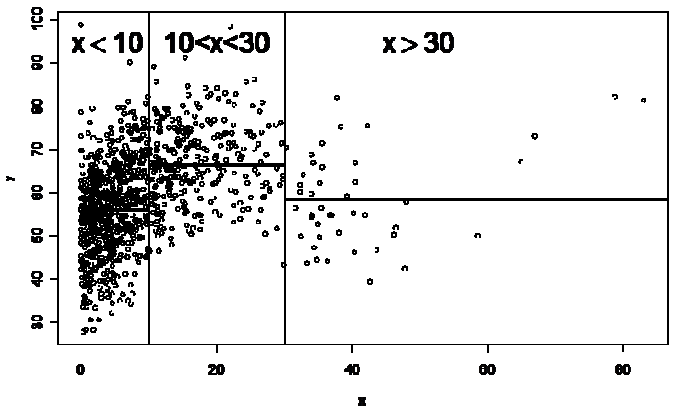
Mida teha siis, kui seos pole lineaarne (sirgega kirjeldatav)?

- Jagada pidev tunnus vahemikeks…
- Lähendada tundmatut seost polünoomi abil (Weierstrass'i teoreem 1885):

$$y = c_0+c_1x+c_2x^2+c_3x^3+\ldots+\varepsilon$$

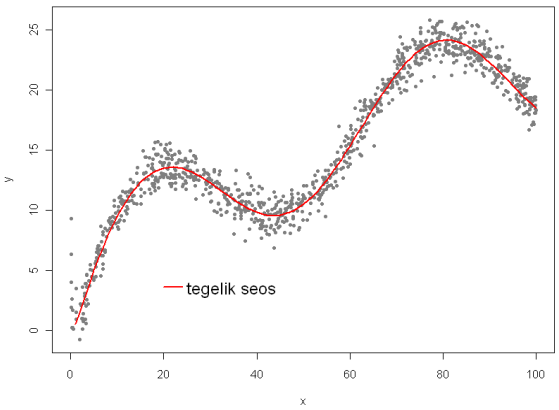
- Lähendada tundmatut seost murdjoone (b-splaini) abil
- Kasutada lokaalset regressiooni
- …

Üks võimalus...



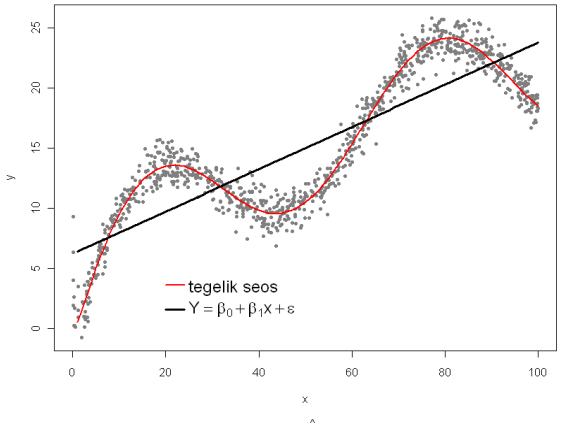
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil)



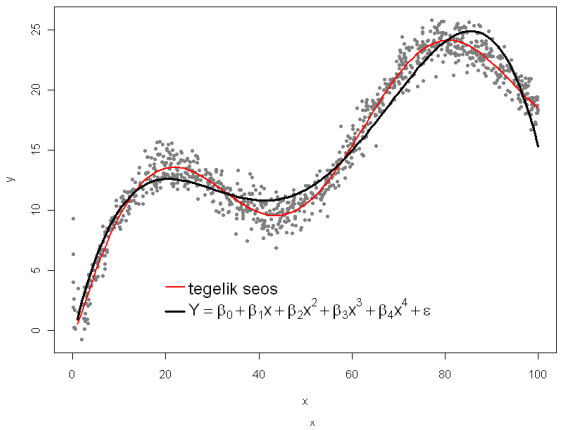
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil)



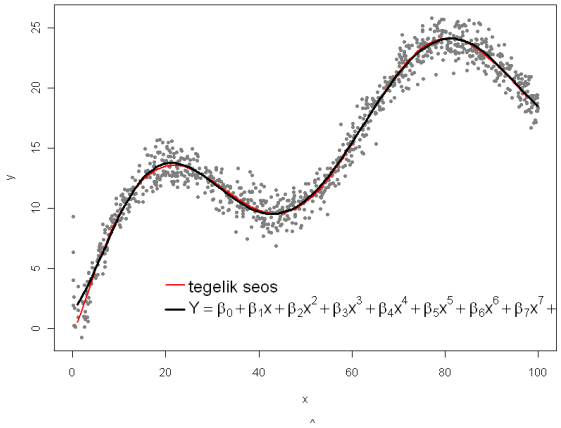
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil)



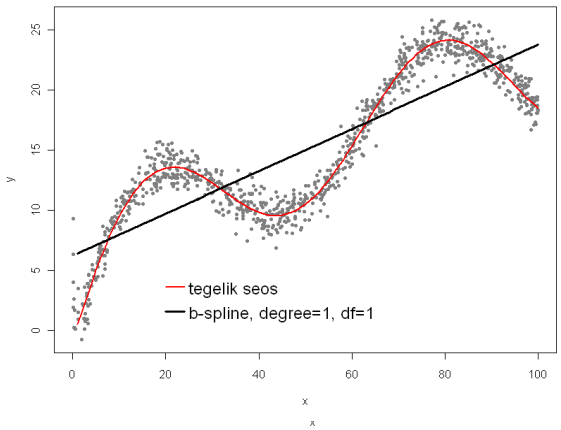
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil)



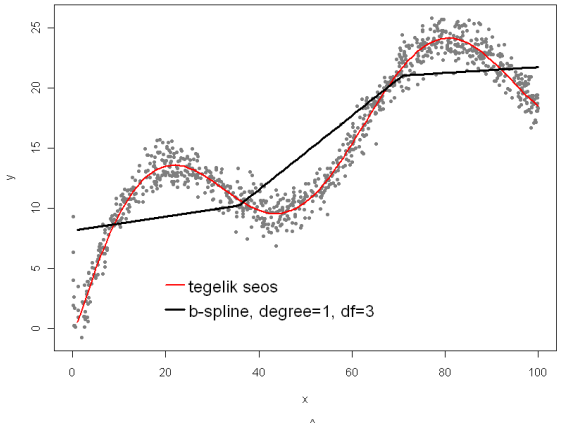
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil, b-spline kasutades)



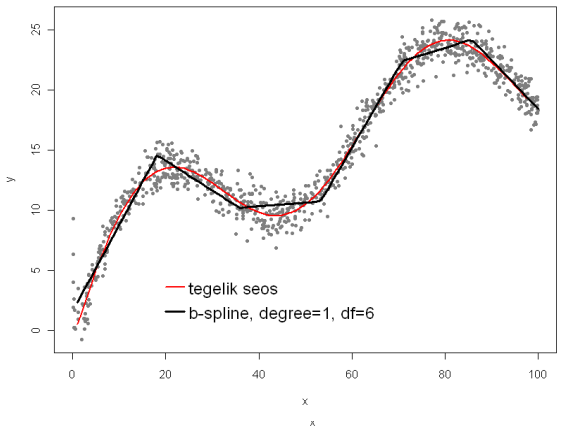
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil, b-spline kasutades)



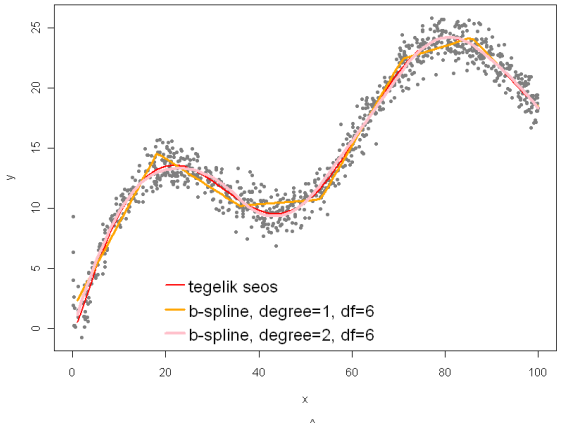
Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil, b-spline kasutades)



Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil, b-spline kasutades)



Mittelineaarne seos – lineaarne mudel

- Kõigi kirjeldatud võimaluste korral (modeleerimine polünoomida abil, b-splainid) on ikka tegemist lineaarse mudeliga:
- Esialgsest tunnusest *x* tehakse mitu uut:

  $x \rightarrow x_1, x_2, x_3, \ldots$ (näiteks $x_1=x$, $x_2=x^2$, $x_3=x^3$, ...)
- Hinnatakse lineaarne mudel (vähimruutude meetodil – minimiseeritakse teadaolevate vaatluste "prognoosimisel" tehtav viga):

  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon$
- Tegemist on lineaarse (regressioni)mudeliga, sest uuritav tunnus on tundmatute parameetrite ($\beta_0, \beta_1, \beta_2, \ldots$) lineaarne funktsioon.

Lisamärkus: sessoonsuse modelleerimine

Olgu tunnus X näiteks päev aastas (1..365). Tahame iseloomustada mingi näitaja Y sõltuvust aastaajast (sessoonset – aastaaegadest tingitud - muutumist, mitte pikaajalist trendi). Tahame, et kohal X=366 oleks graafik sealsamas kus päeval X=1 (sest mõlemad näitavad sama kohta ajas). Seega X+365=X.

Siis teeme abitunnused järgmise eeskirja alusel:

$X_1=\cos(2\cdot\pi/365 \cdot X)$; $X_2=\sin(2\cdot\pi/365 \cdot X)$

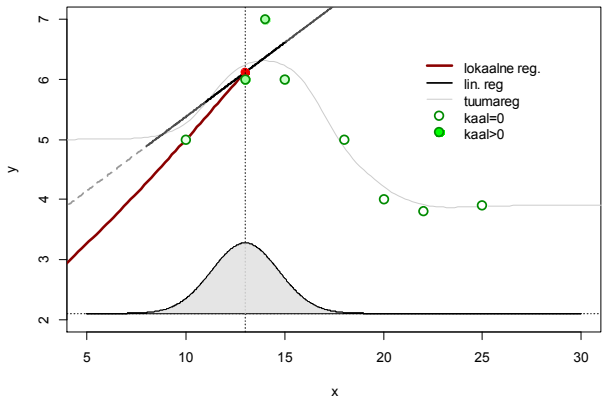
Vajadusel lisa veel täiendavaid tunnuseid:

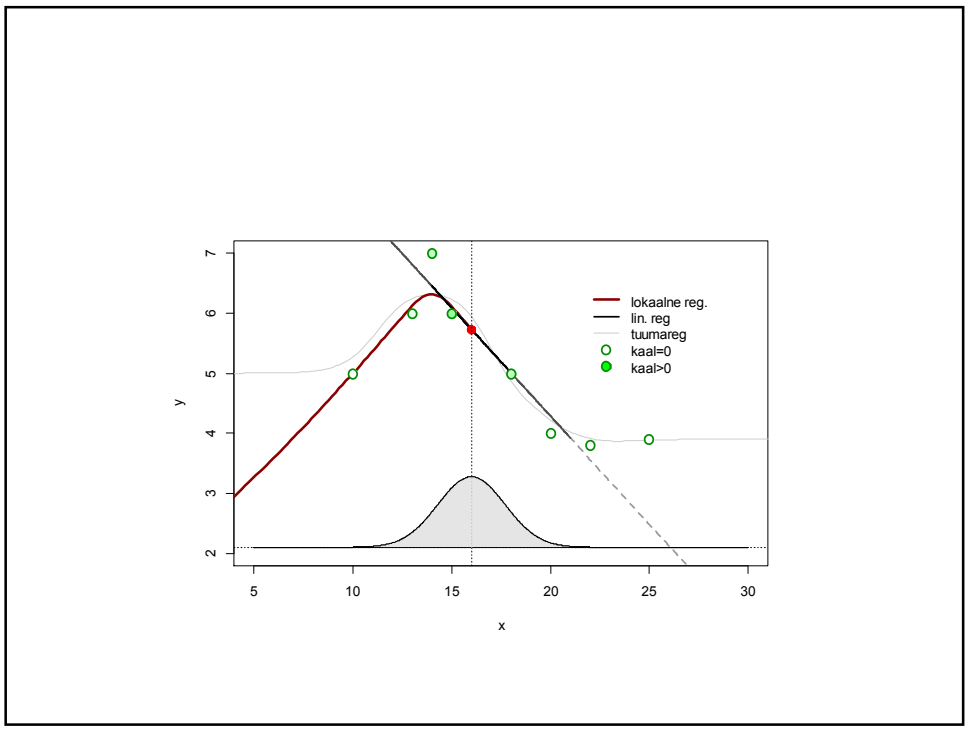
$X_3=\cos(2\cdot 2\cdot\pi/365 \cdot X)$; $X_4=\sin(2\cdot 2\cdot\pi/365 \cdot X)$

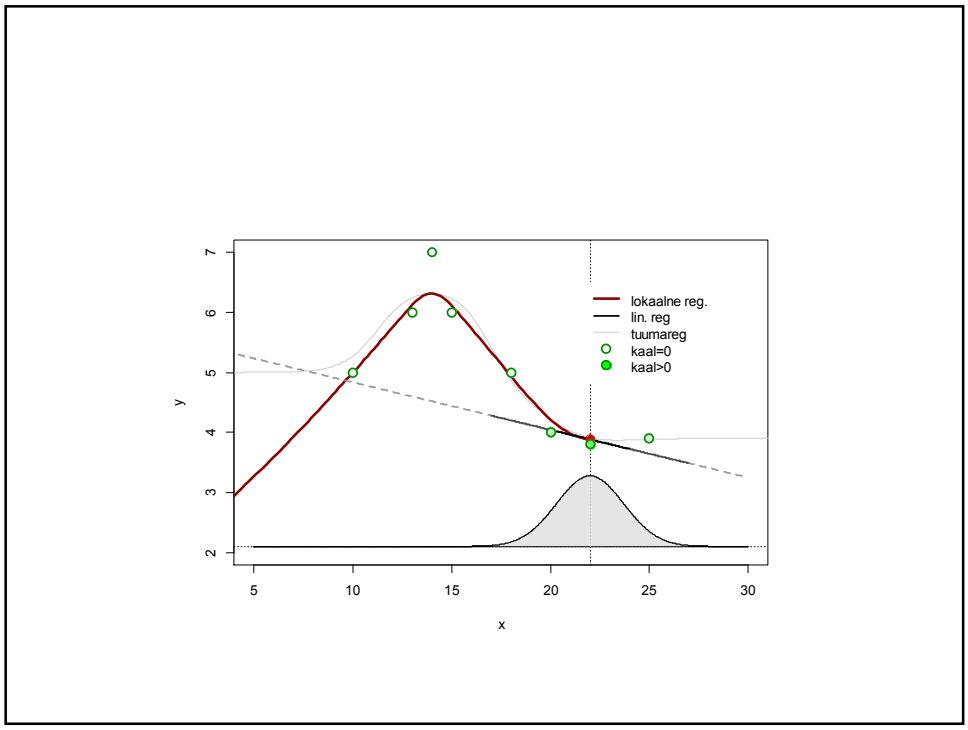
$X_5=\cos(3\cdot 2\cdot\pi/365 \cdot X)$; $X_6=\sin(3\cdot 2\cdot\pi/365 \cdot X)$

.....

Lokaalne regressioon varasemal andmestikul...




lokaalne reg.
lin. reg
tuumareg
kaal=0
kaal>0
y
x
2
3
4
5
6
7
5
10
15
20
25
30


lokaalne reg.
lin. reg
tuumareg
kaal=0
kaal>0
y
x
2
3
4
5
6
7
5
10
15
20
25
30

Tulemuste edastamismugavus

Edastamismugavus

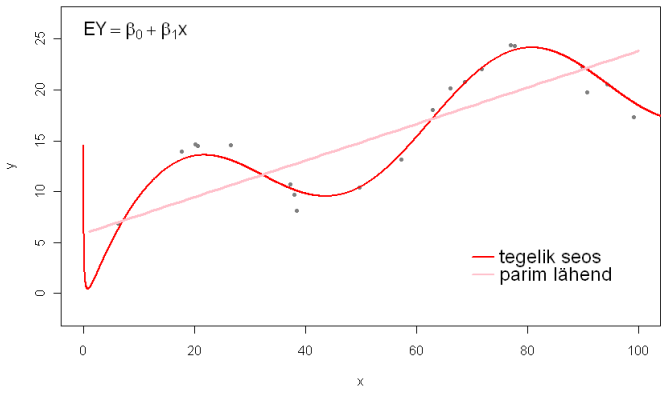
Vahemikud
kaal<50kg
50kg≤kaal<60kg
…

Polünoomid
$y = c_0+c_1 kaal+c_2 kaal^2+\ldots+\varepsilon$

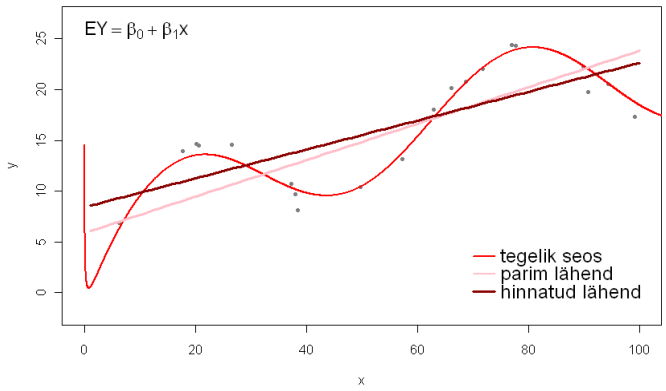
Splainid
(b-splain, tp-splain)

Lokaalne regressioon

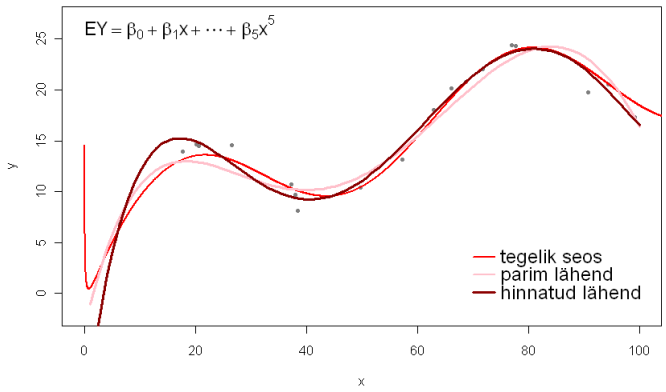
Automaatse modelleerimise piirid I



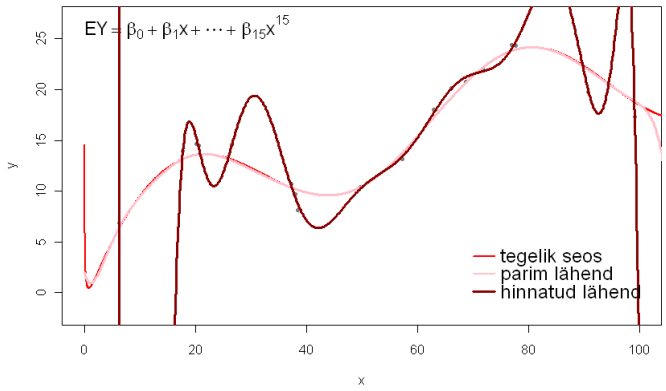
Automaatse modelleerimise piirid I



Automaatse modelleerimise piirid I



Automaatse modelleerimise piirid I



Järeldus I

Väga rikas – kõikelubav – mudel võib praktikas hindamisvigade pärast halvasti töötada.

Mida lihtsam on mudel, seda lähemal on ta üldjuhul “parimale” mudeli poolt lubatule – seda väiksem on hindamisviga (*variance*). Aga lubatutest parim võib olla jälle väga kaugel tegelikust (nn hinnangu nihe – *bias* - võib olla suur)

Vajalik on mõistlik kompromiss.

Vajalik on mõistlik kompromiss….

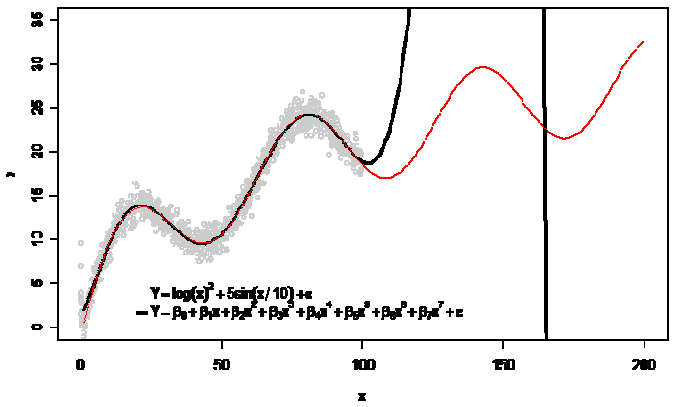
Mudel 1: $y = c_0+c_1x +\varepsilon$
Mudel 2: $y = c_0+c_1x+c_2x^2 +\varepsilon$
Mudel 3: $y = c_0+c_1x+c_2x^2+c_3x^3 +\varepsilon$
Mudel 4: $y = c_0+c_1x+c_2x^2+c_3x^3+c_4x^4 +\varepsilon$
Mudel 5: $y = c_0+c_1x+c_2x^2+c_3x^3+c_4x^4+c_5x^5 +\varepsilon$

mudel	AIC	R2_adj	BIC	p.väärtus
1	657.7827	0.4429994	667.1901	0.0000
2	506.3223	0.7728005	518.8655	0.0000
3	491.5500	0.7929058	**507.2290**	**0.0001**
4	**490.8187**	**0.7949714**	509.6335	0.1040
5	492.8181	0.7937218	514.7687	0.9820

Väikseim parim — Suurim parim — Väikseim parim

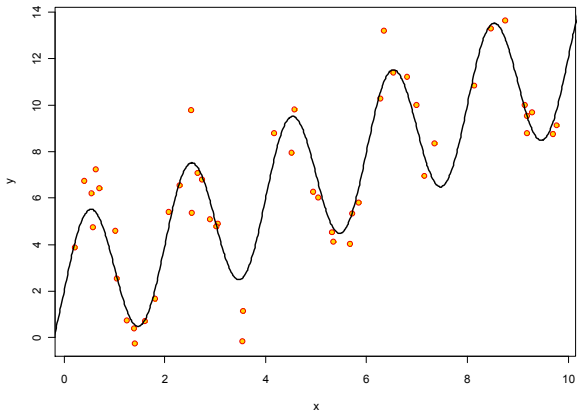
Ohud polünoomide kasutamisel

(sarnaseid probleeme esineb ka teistel lähendusmeetoditel)



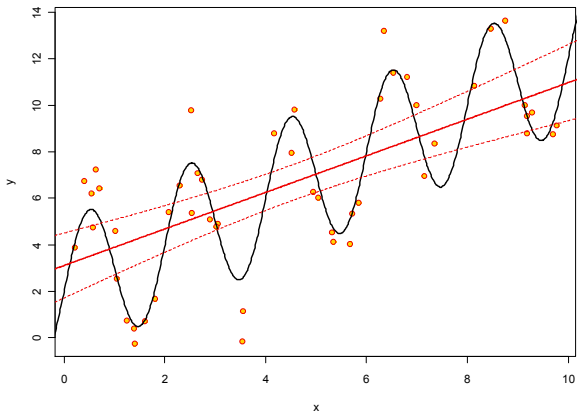
Usaldusintervall ja vale mudel...

Tegelik seos ja vaatlused



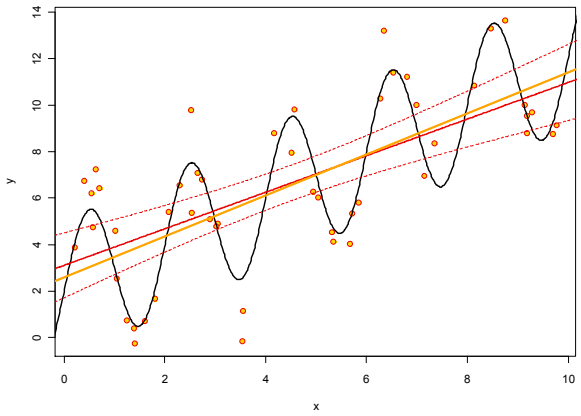
Usaldusintervall ja vale mudel...

Lisame hinnatud regressioonisirge ja usalduspiirid



Usaldusintervall ja vale mudel...

+ parim võimalik regressioonisirge...



Kuidas lähendada polünoomiga:

```
m1=lm(kaal~pikkus); summary(m1)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -118.00340    6.39372  -18.46   <2e-16 ***
pikkus         1.05852    0.03731   28.37   <2e-16 ***

Residual standard error: 7.973 on 641 degrees of freedom
Multiple R-Squared: 0.5567,     Adjusted R-squared: 0.556
F-statistic: 805.1 on 1 and 641 DF,  p-value: < 2.2e-16


m3=lm(kaal~pikkus+I(pikkus^2)+I(pikkus^3)); summary(m3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.622e+03  1.250e+03   2.897  0.00389 **
pikkus      -6.165e+01  2.153e+01  -2.863  0.00434 **
I(pikkus^2)  3.493e-01  1.234e-01   2.829  0.00481 **
I(pikkus^3) -6.461e-04  2.354e-04  -2.745  0.00623 **

Residual standard error: 7.868 on 639 degrees of freedom
Multiple R-Squared: 0.5696,     Adjusted R-squared: 0.5676
F-statistic: 281.9 on 3 and 639 DF,  p-value: < 2.2e-16
```

Märkimisväärset kirjeldatuse tõusu pole, seega kaalu ja pikkuse seos on mõistlikult hästi kirjeldatav ka lineaarse seose abil (kuigi kuuppolünoom on parem).

Testi, kas keerukam mudel on tõestatavalt parem lihtsamast:

```
> m1=lm(y~x)
> m2=lm(y~x+I(x^2)+I(x^3))
> anova(m1,m2)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2) + I(x^3)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      3 0.30054
2      1 0.18873  2   0.11181 0.2962 0.7924
```

Antud näites olulisustõenäosus = 0,7924 > 0,05 ja seega pole keerukama mudeli tõestatavalt parem lihtsast mudelist (mis ei pruugi veel tähendada, et me erinevatel põhjustel ei võiks keerukamat mudelit kasutada).

AIC ja BIC

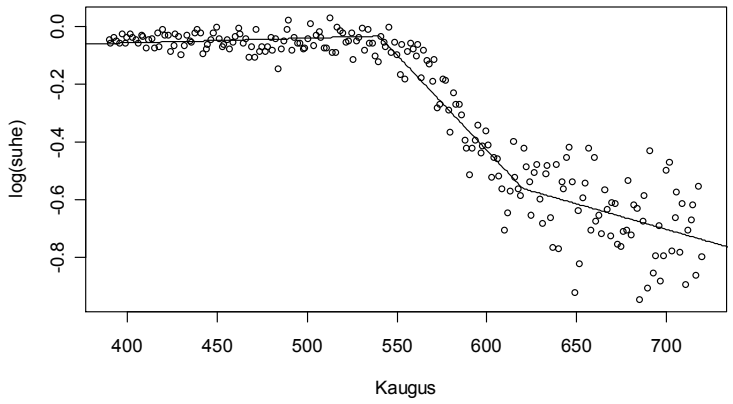
Akaike Information Criterion

Beysian Information Criterion

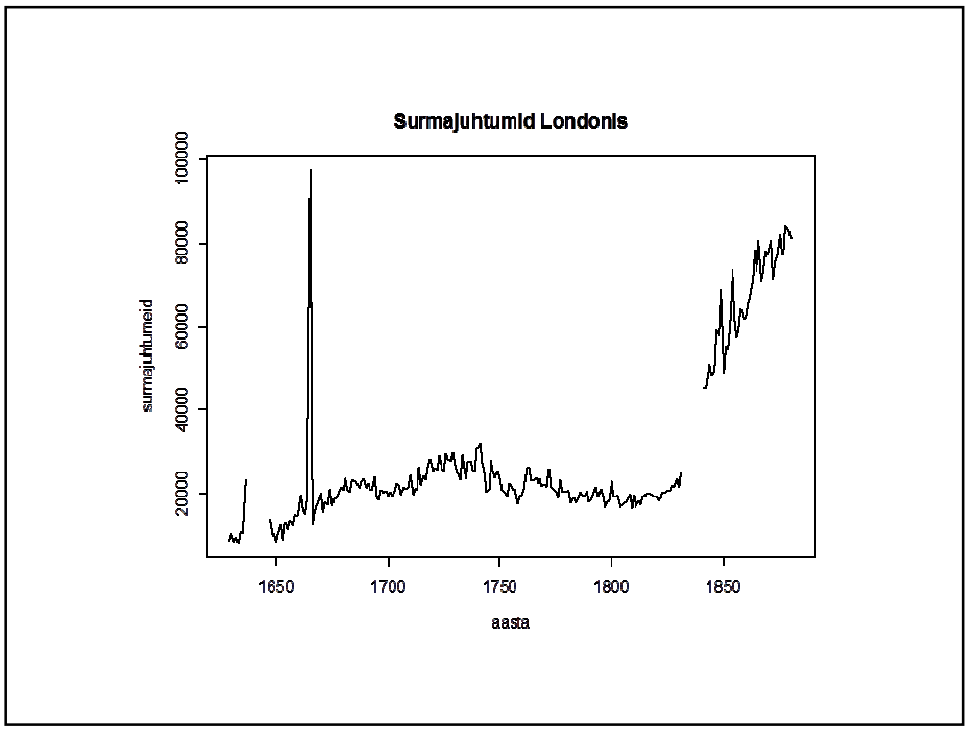
```
> m3=lm(y~x+I(x^2)+I(x^3))
> m4=lm(y~x+I(x^2)+I(x^3)+I(x^4))
> m5=lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5))
> AIC(m3)
[1] 491.55
> AIC(m4)
[1] 490.8187
> AIC(m5)
[1] 492.8181
> BIC(m3)
[1] 507.229
> BIC(m4)
[1] 509.6335
```

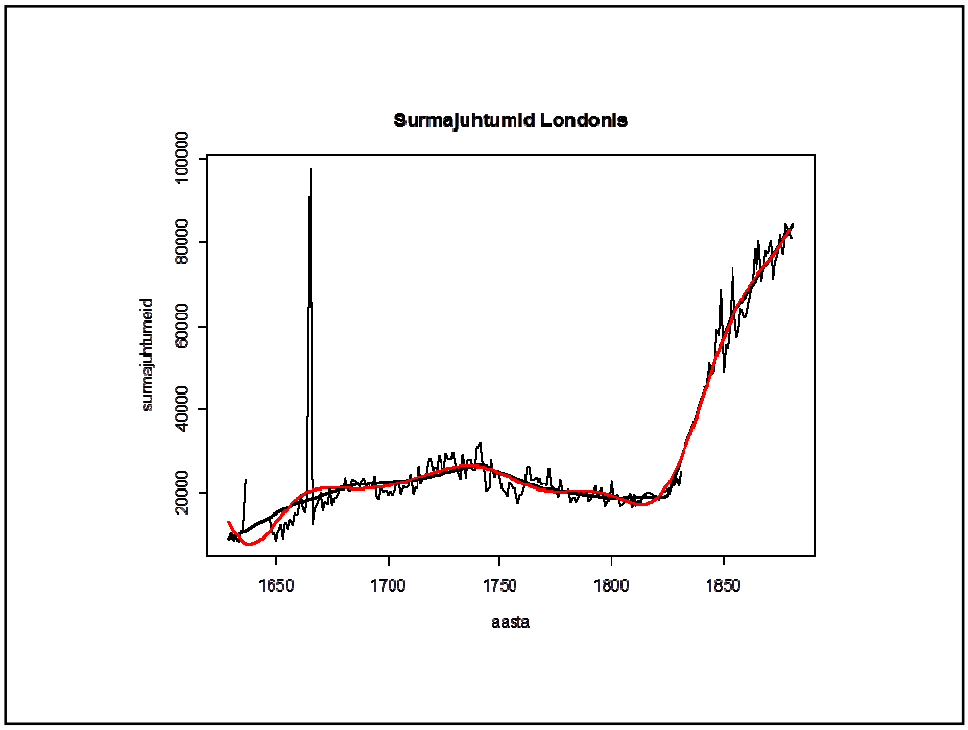
```
mudel=step(m5)
```

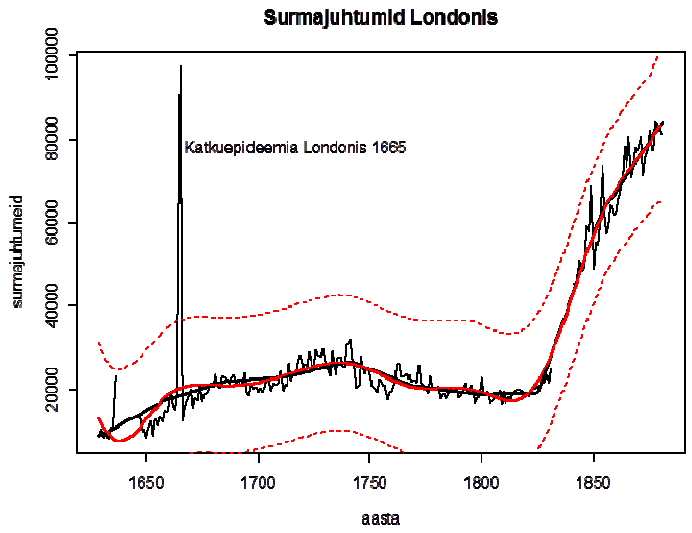
Splainid R-s




```
library(splines)

# Variant 1 – anname ise murdepunktid ette
lm(log(suhe)~bs(kaugus, knots=c(540, 620), degree=1))

# Variant 2 – murdepunktid automaatselt (6 murdepunkti)
mudel=lm(log(suhe)~bs(kaugus,
       knots=quantile(kaugus, (1:6)/7))

# Prognoosimine:
predict(mudel, data.frame(kaugus=500))

# Variant 3 – murdepunktid automaatselt (määrame lihtsalt
#      hinnatavate parameetrite arvu)

m=lm(kaal~bs(pikkus, df=10, degree=3))
summary(m)
```


Surmajuhtumid Londonis
surmajuhtumeid
20000
40000
60000
80000
100000
1650
1700
1750
1800
1850
aasta

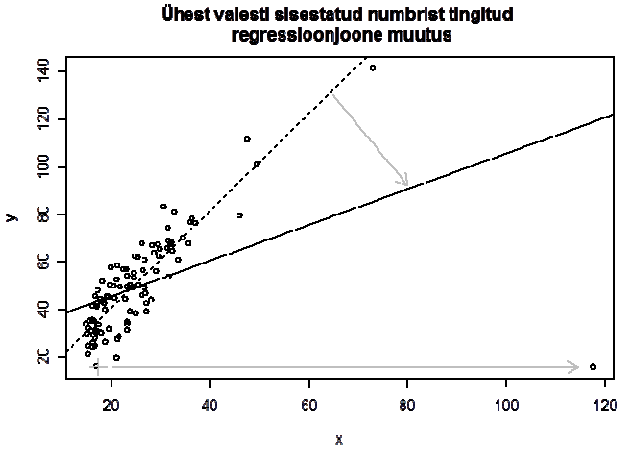

Surmajuhtumid Londonis
surmajuhtumeid
20000
40000
60000
80000
100000
1650
1700
1750
1800
1850
aasta



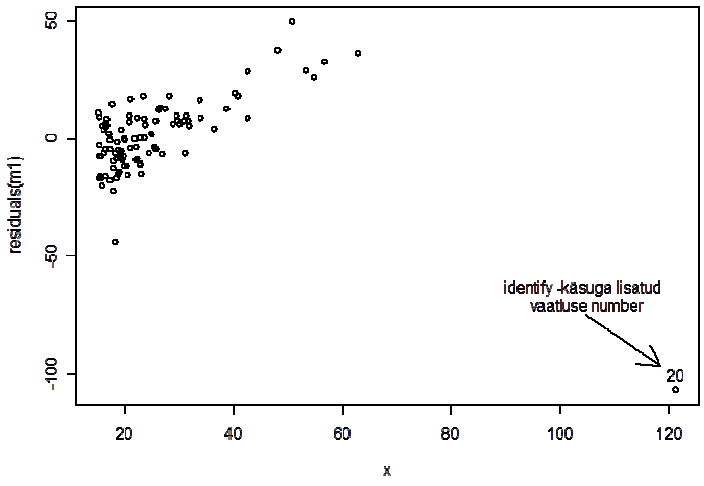
Eelmise joonise tegemiseks kasutatud käsud

```
plot(aeg, surmi, type="l", main="Surmajuhtumid Londonis",
    xlab="aasta", ylab="surmajuhtumeid")


m1=lm(surmi~bs(aeg, knots=seq(min(aeg), max(aeg), length=10),
    degree=1))
m1a=lm(surmi~bs(aeg, knots=seq(min(aeg), max(aeg), length=10),
    degree=3))


x=seq(min(aeg), max(aeg), length=200)
y=predict(m1, data.frame(aeg=x))
ya=predict(m1a, data.frame(aeg=x))

lines(x,y, lwd=2)
lines(x,yb, col=2, lwd=2)

ya=predict(m1b, data.frame(aeg=x), interval="prediction",
        level=0.99)
lines(x,ya[,2], col=2, lty=2)
lines(x,ya[,3], col=2, lty=2)
text(1667, 80000, "Katkuepideemia Londonis 1665", adj=c(0,1))
```

Jämedad vead ja erindid




```
plot(x, residuals(m1))
identify(x, residuals(m1))
```



Standardiseeritud jäägid, mõjukus ja Cook'i kaugus

```
plot(mudel)
```