

Lineaarse regressiooni eeldused

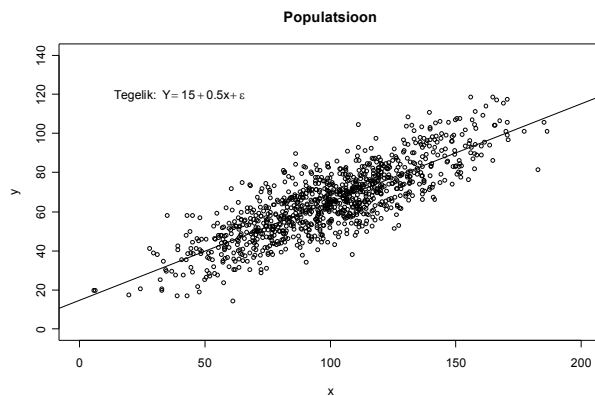
- Kuidas näha probleeme ja vigu?
- Mida teha, kui midagi on korrast ära?

Eelmises loengus ja praktikumis

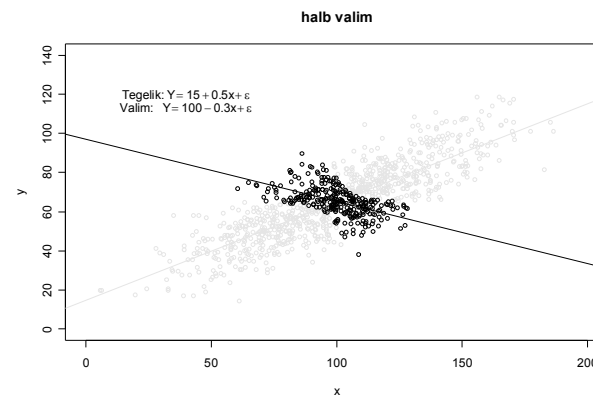
- Hindasime lineaarse regressioonimodeli
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
parameetreid β_0, β_1 ;
- kirjeldasime usaldusvahemiku abil parameetrite hinnangute täpsust;
- testisime, kas Y väärtuste prognoosimisel on kasu tunnuse X väärtuste teadmisest;
- kirjeldasite Y ja X vahelise seose tugevust (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 regressioonianalüüsi eeldused.

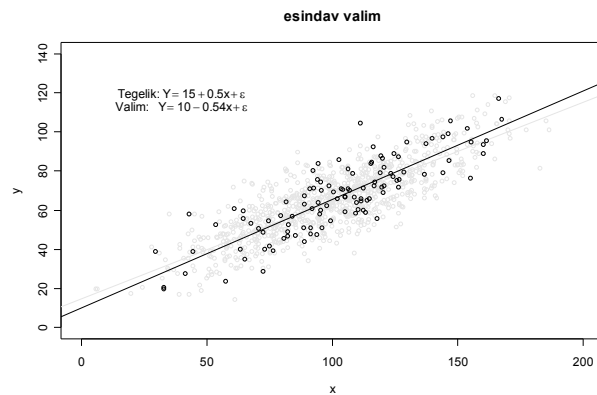
Esindav valim



Esindav valim



Esindav valim



Esindav valim

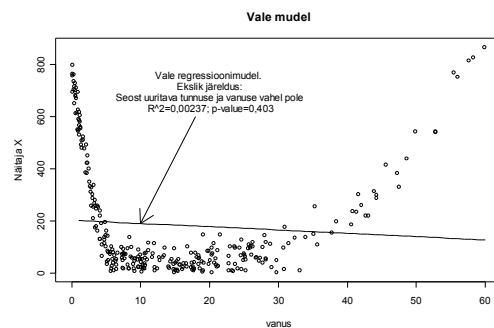
Kasutatav valim peab kirjeldama populatsiooni õiglaselt ja objektiivselt, olema esindav (representative sample). Eeldatakse, et igal objektil, mis kuulub populatsiooni, on võrdne võimalus sattuda valimisse ja objekti valimisse sattumise tõenäosus ei sõltu sellest millised objektid on juba valituks osutunud.

Näiteid võimalikest probleemidest:

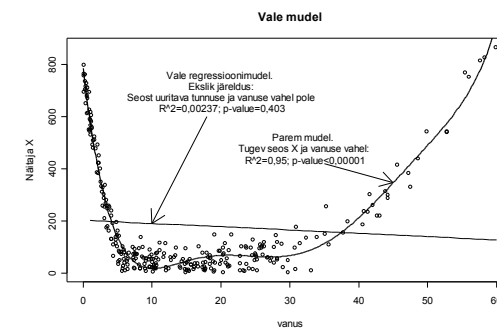
- vastama nõustuvad inimesed ei pruugi olla esindavad kõigi inimeste jaoks;
- katsepõllud ei pruugi olla „tüüpilised“ põllud, katsepõldude eest hoolitsemine ei pruugi olla „tüüpiline“;
- pikast küsimustikust tüdinud inimesed vastavad küsimustele juhuslikult;
- uuringusse võivad hooletu katseplaani korral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

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

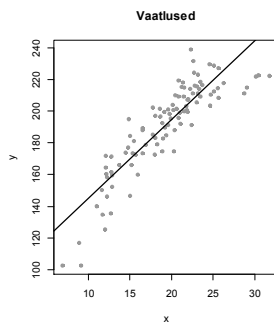
Eeldus II. Seos tunnuste vahel peab olema lineaarne (või selline, nagu mudel ütleb).



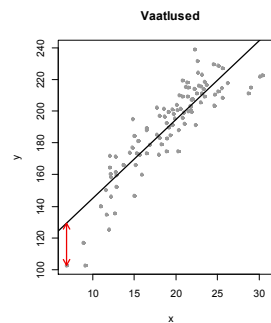
Eeldus II. Seos tunnuste vahel peab olema lineaarne (või selline, nagu mudel ütleb).



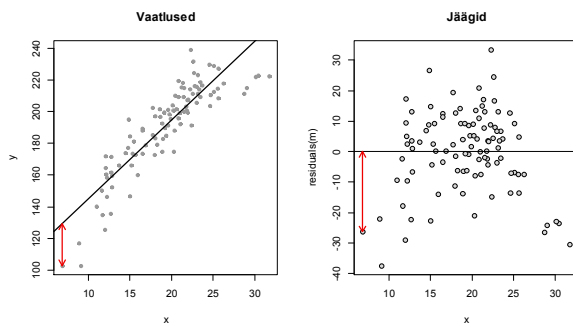
Kuidas märgata?



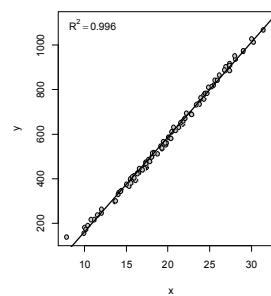
Kuidas märgata?



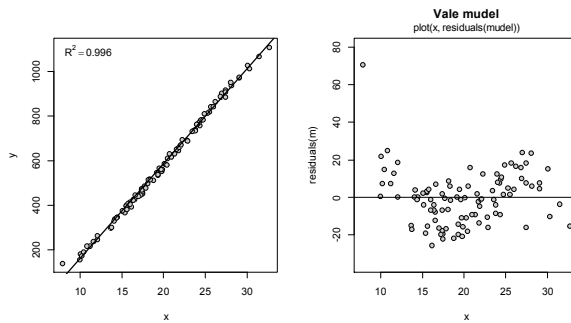
Kuidas märgata?



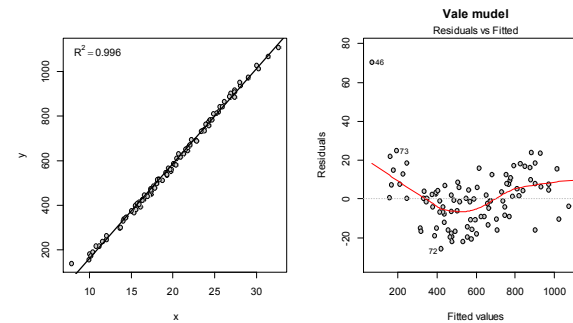
Miks uurida jääke?



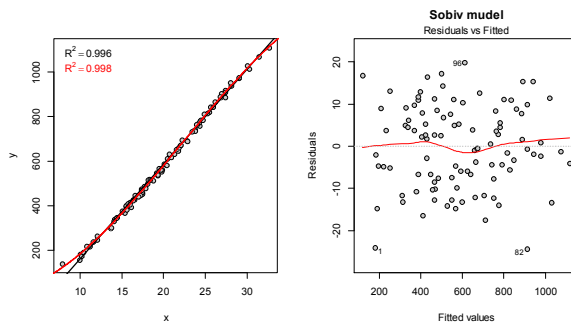
Miks uurida jääke?



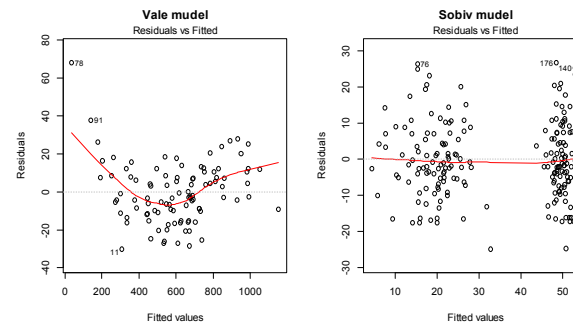
Miks uurida jääke?



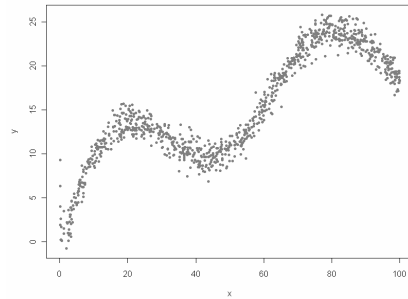
Miks uurida jääke?



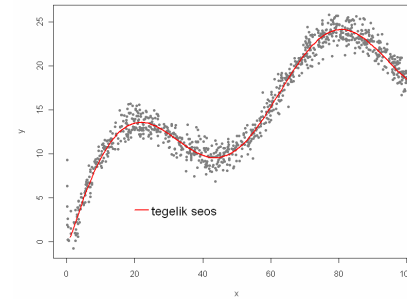
Miks uurida jääke?



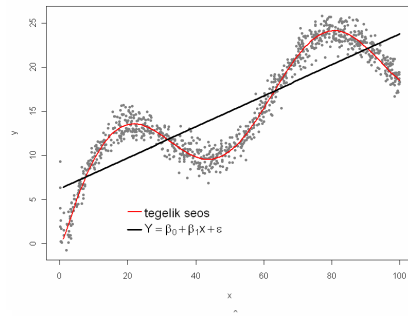
Mittelineaarse seose modelleerimisest (lineaarse mudeli abil)



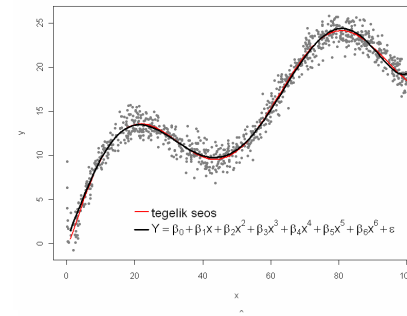
Mittelineaarse seose modelleerimisest (lineaarse mudeli abil)



Mittelineaarse seose modelleerimisest (lineaarse mudeli abil)

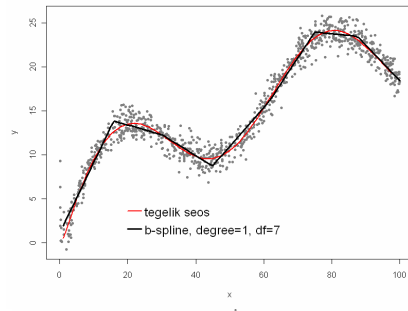


Mittelineaarse seose modelleerimisest (lineaarse mudeli abil)



Mittelineaarse seose modelleerimisest

(lineaarse mudeli abil, b-spline kasutades)



Mittelineaarne seos – lineaarne mudel

- Kõigi kirjeldatud võimaluste korral (modelleerimine polünoomida abil, b-splainid) on ikka tegemist lineaarse mudeliga:
- Esialgselt tunnusest x tehakse mitu uut:
 $x \rightarrow x_1, x_2, x_3, \dots$ (näiteks $x_1=x, x_2=x^2, x_3=x^3, \dots$)
- Hinnatakse lineaarne mudel (vähimruutude meetodil – minimiseeritakse teadaolevate vaatluste “proгноosisel” tehtav viga):

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

Lisamärkus: sessoonsuse modelleerimine

Olgu tunnus X näiteks päev aastas (1..365). Tahame iseloomustada mingi näitaja Y sõltuvust aastaajast (sessoonset – aastaegadest 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:

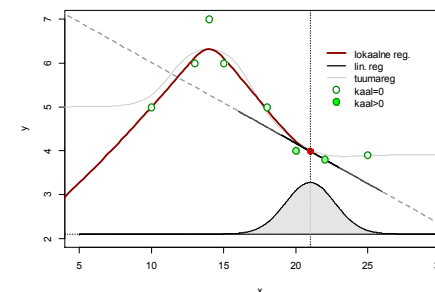
$$X_1 = \cos(2 \cdot \pi / 365 \cdot X); \quad X_2 = \sin(2 \cdot \pi / 365 \cdot X)$$

Vajadusel lisa veel täiendavaid tunnuseid:

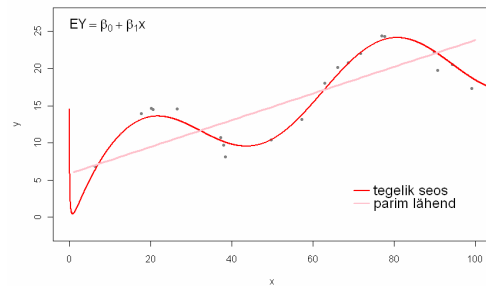
$$X_3 = \cos(2 \cdot 2 \cdot \pi / 365 \cdot X); \quad X_4 = \sin(2 \cdot 2 \cdot \pi / 365 \cdot X)$$

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

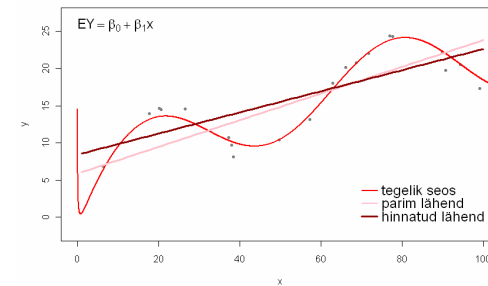
.....



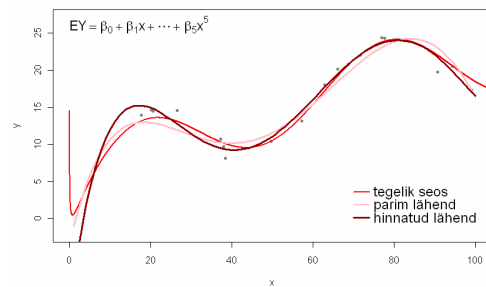
Automaatse modelleerimise piirid I



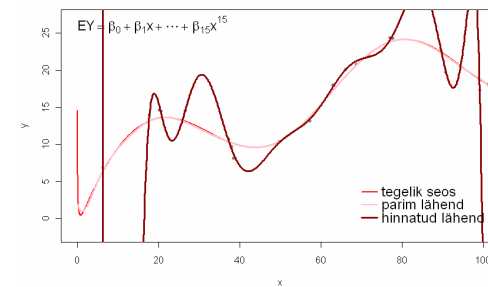
Automaatse modelleerimise piirid I



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.