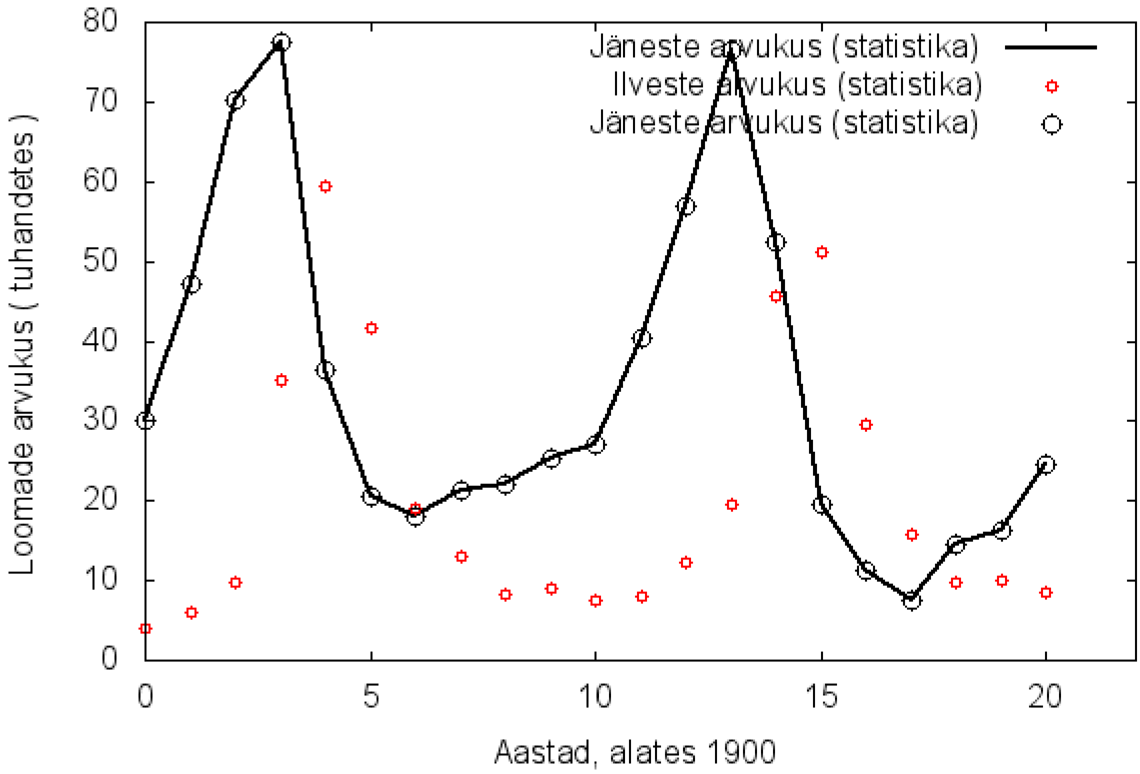
Funktsioonide lähendamine, interpoleerimine

Käesolevas peatükis kasutame kohati loengus [1] toodud materjale.

Katseandmete töötlemise juures on sageli vaja kasutada interpoleerimist. Probleemi olemus seisneb järgnevas. Olgu meil teostatud mingi lõplik arv katseid ja iga katse korral teostatud mingi mõõtmiste kompleks. Näiteks võib mingil ajavahemikul mõõta temperatuuri.

Joonisel on aastatel 1900 kuni 1920 igal aastal üle loetud Kanada valgejäneste ja ilveste ligikaudne arv.



Kui eeldada, et kahe mõõdetud suuruse vahel on funktsionaalne sõltuvus, siis funktsiooni väärtuse arvutamiseks nendes punktides, kus katseid pole läbi viidud, tuleks konstrueerida mingi katseandmeid lähendav (interpoleeriv) funktsioon. Interpoleerivate funktsioonide konstrueerimist järgnevalt vaatlemegi.

Joonisel on jäneste arvukus ühendatud sirgjoontega. Ilveste arvukus on kantud joonisele ainult punktidena (mõõtetulemustena). Kui meid huvitab arvukus nendes punktides, kus mõõtetulemus puudub, siis tuleb luua funktsioon, mis läbib antud punkte täpselt või siis võimalikult lähedalt.

Lineaarne interpoleerimine

Käsk **linterp(x-vektor , y-vektor , x)** väljastab väärtuse sirge peal punktis x, kus sirge ühendab vektoris "y-vektor" olevaid väärtusi.

"x-vektor" on x-teljel olevate (reaalarvuliste) väärtuste vektor (ehk mõõtmishetked, mõõtmispunktid). **Andmed peavad olema järjestatud kasvavas järjestuses.**

"y-vektor" on vastavad (reaalarvulised) väärtused y-teljel (ehk siis mõõtmistulemused), kusjuures nende kahe vektori dimensioonid peavad olema samad.

$$\text{aastad} := (1920 \;\; 1915 \;\; 1910 \;\; 1905 \;\; 1900)^T$$

Mõõtmishetked

$$\text{jänku} := (24.7 \;\; 19.5 \;\; 27.1 \;\; 20.6 \;\; 30)^T$$

Jäneste arv tuhandetes nendel aastatel

$$\text{andmed} := \text{augment}(\text{aastad}, \text{jänku}) = \begin{pmatrix} 1920 & 24.7 \\ 1915 & 19.5 \\ 1910 & 27.1 \\ 1905 & 20.6 \\ 1900 & 30 \end{pmatrix}$$

Moodustame andmete maatriksi, kus esimeses veerus on mõõtmishetked ja teises mõõtmistulemused.

$$\text{linterp}\left(\text{andmed}^{\langle 0\rangle}, \text{andmed}^{\langle 1\rangle}, 1908\right) = \blacksquare$$

Kohe ei saa funktsiooni linterp kasutada, kuna esimene veerg ei ole kasvavas järjekorras.

$$\text{andmed} := \text{csort}(\text{andmed}, 0) = \begin{pmatrix} 1900 & 30 \\ 1905 & 20.6 \\ 1910 & 27.1 \\ 1915 & 19.5 \\ 1920 & 24.7 \end{pmatrix}$$

Sorteerime oma maatriksi kasvavas järjekorras esimese veeru järgi (käsuga "csort"). Andmete segamini mineku vältimiseks ühendasime enne oma mõõtmihsetkete ja mõõtmistulemuste vektorid. Nüüd võib julgelt maatriksit sorteerida, sest ühel real hoitakse alati õigeid väärtusi.

$$\text{linterp}\left(\text{andmed}^{\langle 0\rangle}, \text{andmed}^{\langle 1\rangle}, 1908\right) = 24.5$$

Väljastatakse jäneste arvukus aastal 1908.

$$\text{linterp}\left(\text{andmed}^{\langle 0\rangle}, \text{andmed}^{\langle 1\rangle}, 1930\right) = 35.1$$

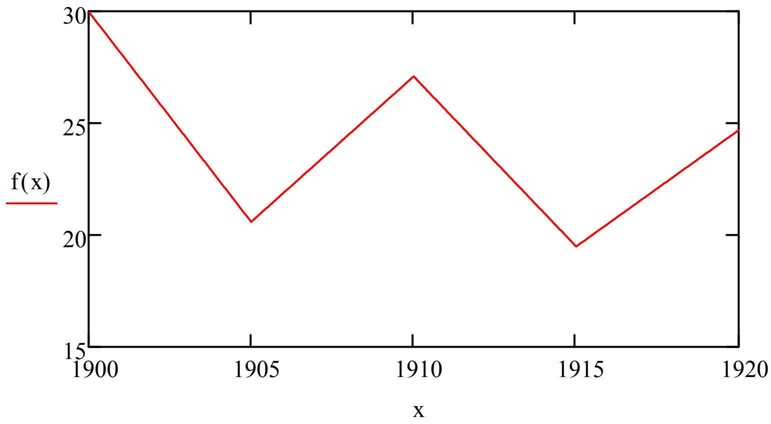
Sedasi võib ka ennustada, näiteks 1930 oleks jäneseid 35 tuhat, kuid ilmselgelt leitakse väärtused väga ebatäpselt, kui üritame neid leida liialt kaugetes punktides võrreldes katsetulemustega (mõõtmishetketega)

$$f(x) := \text{linterp}\left(\text{andmed}^{\langle 0\rangle}, \text{andmed}^{\langle 1\rangle}, x\right)$$

Edaspidiseks on kasulik teha funktsioon f, mis annab meile jäneste arvukuse suvalises punktis x (kuigi teda väljaspool lõiku [1900,1920] ei ole tark kasutada).

Kanname tulemused ka joonisele, kuigi võime märgata, et eriti lähedast tulemust me ei saa, sest meil puuduvad vahepealsed väga olulised mõõtmistulemused.

$$\text{andmed2} := \begin{pmatrix} 1900 & 30 \\ 1901 & 47.2 \\ 1902 & 70.2 \\ 1903 & 77.4 \\ 1904 & 36.3 \\ 1905 & 20.6 \\ 1906 & 18.1 \\ 1907 & 21.4 \\ 1908 & 22.0 \\ 1909 & 25.4 \\ 1910 & 27.1 \\ 1911 & 40.3 \\ 1912 & 57 \\ 1913 & 76.6 \\ 1914 & 52.3 \\ 1915 & 19.5 \end{pmatrix}$$

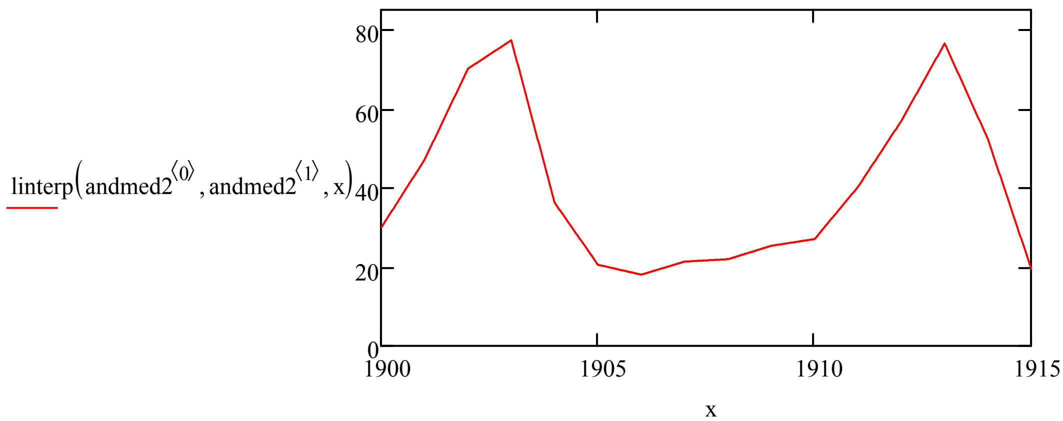


Vaatleme mõõtmistulemusi rohkemate andmetega. Sel juhul on tulemus parem.

$$\text{linterp}\left(\text{andmed2}^{\langle 0\rangle}, \text{andmed2}^{\langle 1\rangle}, 1912.5\right) = 66.8$$

$$\text{linterp}\left(\text{andmed2}^{\langle 0\rangle}, \text{andmed2}^{\langle 1\rangle}, 1913.5\right) = 64.45$$

$$\text{linterp}\left(\text{andmed2}^{\langle 0\rangle}, \text{andmed2}^{\langle 1\rangle}, 1914.5\right) = 35.9$$



Regressioonisirge

Kui on teada, et lähteandmed sisaldavad mingit (mõõtmis)viga, ei ole väga suure hulga andmeid otstarbekas lähendada intepoleerides (interpoleeriv funktsioon läbib **kõiki** katsepunkte), vaid koostades regressioonikõvera. Regressioonikõver koostatakse nii, et ta graafik oleks võimalikult lähedal katsepunktidele, kuid ta ei pea neid tingmata läbima.

Lihtsaimaks viisiks on lähendada katseandmeid sirgega, koostades sirge võrrandi nii, et punktide kauguste ruutude summa sirgest oleks minimaalne. Kuna sirge võrrandiks on **y = ax+b**, siis tuleb meil määrata kaks parameetrit - a ja b. Nende parameetrite määramiseks on Mathcadis funksioonid ***slope()*** (ingl. k. tõus, kalle) ja ***intercept()*** (ingl. k. nihe).

Käsk **slope(x-vektor , y-vektor)** väljastab sirge tõusu, mis on leitud vähimruutude meetodil andmete "x-vektor" ja "y-vektor" põhjal.

Käsk **intercept(x-vektor , y-vektor)** väljastab vastava sirge nihke, mis on leitud vähimruutude meetodil andmete "x-vektor" ja "y-vektor" põhjal.

Käsk **line(x-vektor , y-vektor)** väljastab vektorina sirge nihke ja tõusu, mis on leitud vähimruutude meetodil andmete "x-vektor" ja "y-vektor" põhjal.

Näide

$N := 50$ Mõõdetavate punktide arv.

$i := 1, 2 .. N$ Indeks 1..N

$XX_i := \mathrm{rnd}(10)$ Käsk rnd, (r n d), leiab siin juhuslikud reaalarvud 0..10.

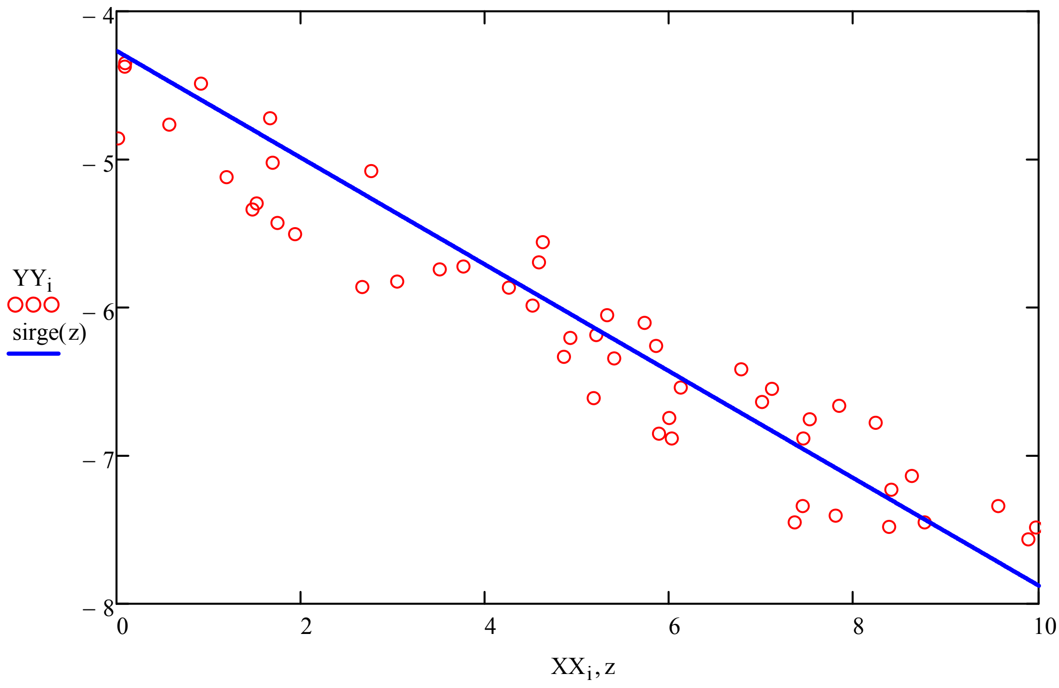
$YY_i := -\frac{1}{3} \cdot XX_i + \mathrm{rnd}(1) - 5$ Oletame, et YY väärtused on leitud sellisel moel.

$\mathrm{sirge} := \mathrm{line}(XX, YY) = \begin{pmatrix} -4.266 \\ -0.361 \end{pmatrix}$ Leiame regressioonisirge nihke ja tõusu.

$\mathrm{slope}(XX, YY) = -0.361$ Võrdluseks sirge tõus slope käsuga

$\mathrm{intercept}(XX, YY) = -4.266$ Võrdluseks sirge nihe intercept käsuga

$\mathrm{sirge}(x) := \mathrm{sirge}_1 \cdot x + \mathrm{sirge}_0$ Defineerime tõusu ja nihke abil sirge funktsioonina.



Üldine regressioonikõver

Sirgega lähendamine ei ole piisavalt paindlik. Anname skeemi üldiste funktsioonide jaoks.

Käsk **linfit(x-vektor , y-vektor , F-vektor)** väljastab vektorina kordajad n-järku lineaarsele kombinatsioonile $Y(x) = c_1 \cdot f_1(x) + c_2 \cdot f_2(x) + \ldots + c_n \cdot f_n(x)$, kus lahend Y käitub eeldatavasti mingite antud funktsioonide lineaarse kombinatsioonina.

Näide, [1]. Lineaarse kombinatsiooni näide, linfit().

$i := 0, 1 .. 50$ $\qquad x_i := \mathrm{rnd}(10)$

$$y_i := \left(x_i\right)^2 \cdot 0.5 + \frac{0.2}{x_i + 1} + 0.2 + \mathrm{rnd}(5)$$

Olgu arvutatud väärtused sellise funktsiooni järgi, kus siis argument x saab juhuslikke suurusi ja samuti y väärtustele lisanduvad juhuslikud suurused liikme rnd(5) näol.

Nüüd on meil vektoritena olemas mõõtehetked x ja mõõteandmed y.

Otsime regressioonikõverat kujul $y = a \cdot x^2 + \frac{b}{x + 1} + c$, kus a, b ja c on otsitavad parameetrid.

Moodustame a, b ja c juurde kuuluvatest funktsioonidest vektorfunktsiooni:

$$F(x) := \begin{pmatrix} x^2 \\ \frac{1}{x+1} \\ 1 \end{pmatrix}$$

See on vektorfunktsioon, mille argumentideks on arvud.

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} := \mathrm{linfit}(x, y, F)$$

Leiame parameetrid a, b ja c kasutades käsku linfit().

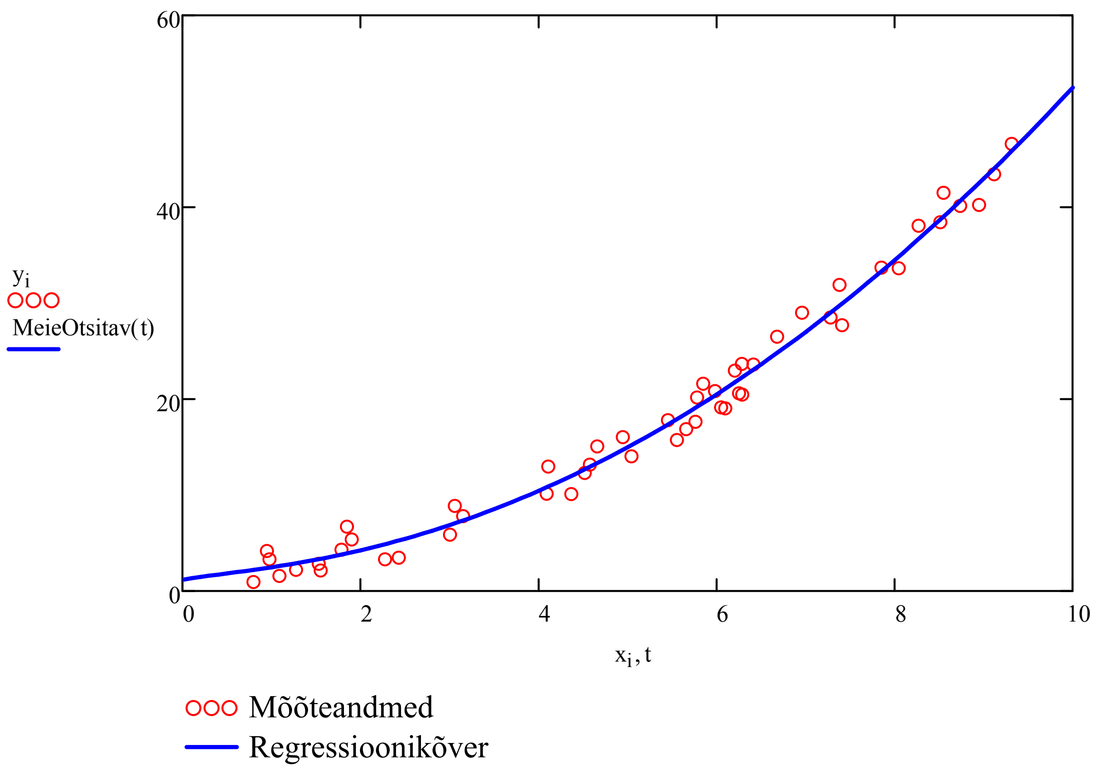
$a = 0.498 \qquad b = -1.64 \qquad c = 2.836$

$$\mathrm{MeieOtsitav}(x) := a \cdot x^2 + \frac{b}{x+1} + c$$

Defineerime otsitud (regressiooni)funktsiooni.

Kujutame regressioonikõvera koos katseandmetega ka graafikul:

$t := 0, 0.1 .. 10$



Kasutatud kirjandus

[1] U. Hämarik. "MTMM.00.216 Arvutiõpetus: Mathcad, MS Office. Mathcad: mõõtühikud". Tartu Ülikool. http://math.ut.ee/~uno_h/arvutiopf.html

[2] "Mathcad 2000. User's Guide." USA, 1999.