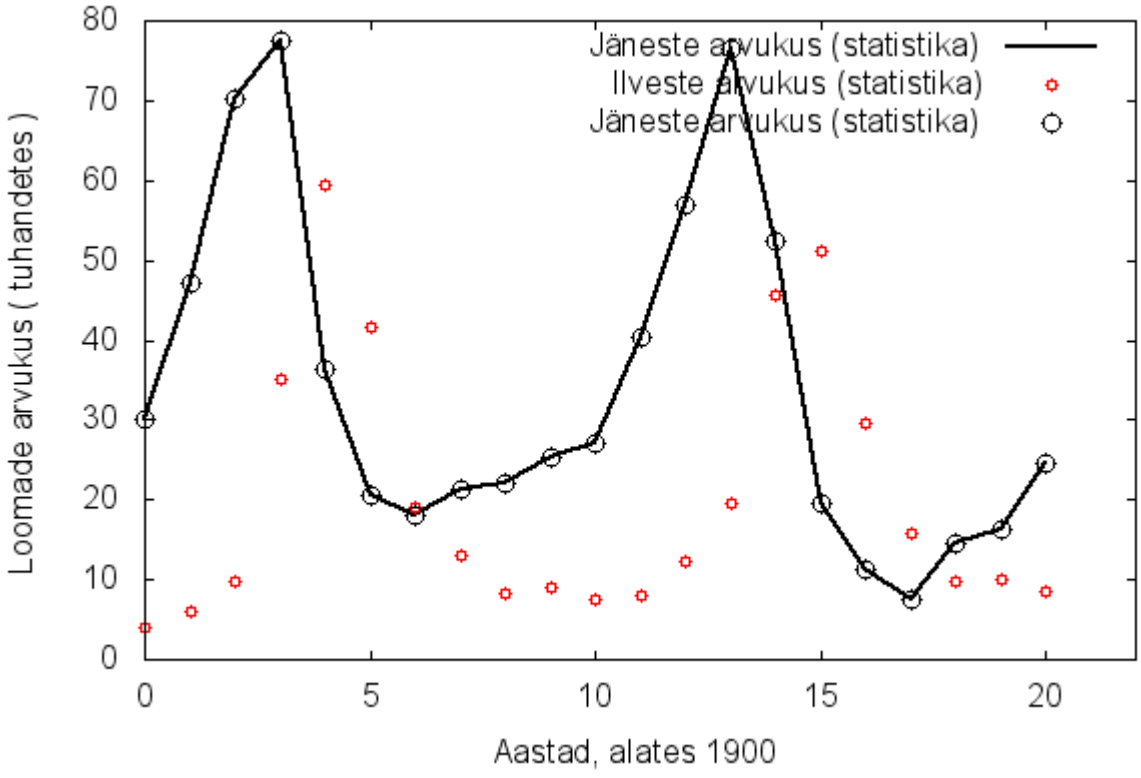
Funktsioonide lähendamine, interpoleerimine

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 mõõta mingil ajavahemikul 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 näiteks 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 punktis x sirge peal, mis ühendab vektoris "y-vektor" olevaid väärtusi (kusjuures ühenduslõigud läbivad täpselt kõiki antud 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.

$$aastad := (1920 \quad 1915 \quad 1910 \quad 1905 \quad 1900)^T$$

Mõõtmishetked

$$jänku := (24.7 \quad 19.5 \quad 27.1 \quad 20.6 \quad 30)^T$$

Jäneste arv tuhandetes nendel aastatel

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

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

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

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

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

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.

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

Väljastatakse jäneste arvukus aastal 1908.

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

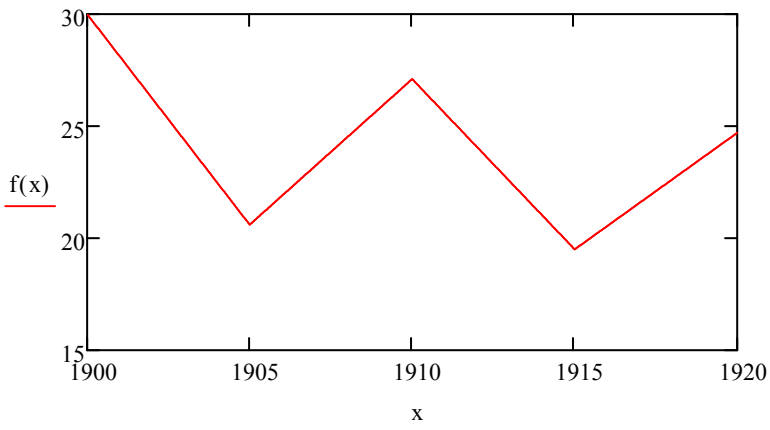
Sedasi võib ka ennustada, näiteks 1930 oleks jäneseid 35 tuhat, kuid ilmselgelt leitakse väärtused väga ebatäpselt, kui üritame neid leida liialt kaugetes punktides võrreldes katsetulemustega (mõõtmishetketega)

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

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.

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

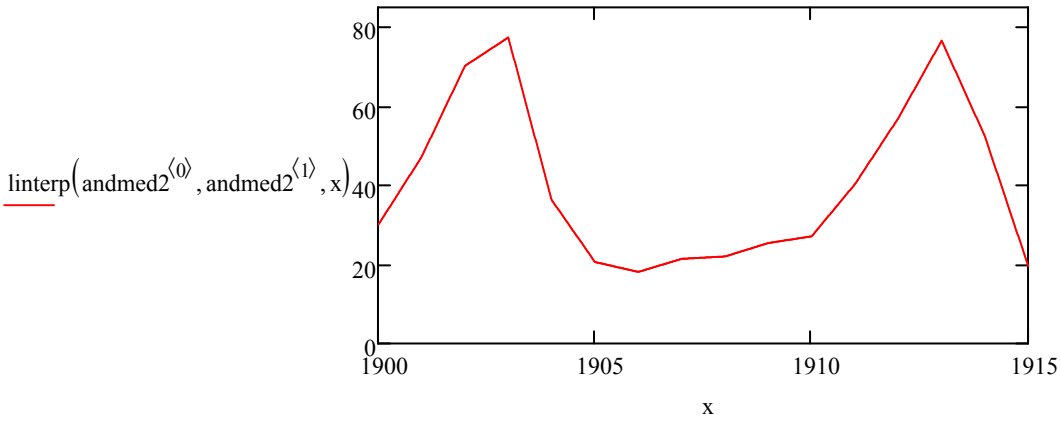


Vaatleme mõõtmistulemusi rohkemate andmetega. Sel juhul on tulemus parem.

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

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

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



Käsk **predict(y-vektor , m , N)** väljastab N järgmist uut väärtust andmete "y-vektor" põhjal, kus "y-vektor" on andmete vektor. Seejuures kasutatakse m viimast väärtust andmete vektoris.

$$\text{linterp}\left(\text{andmed2}^{\langle 0 \rangle}, \text{andmed2}^{\langle 1 \rangle}, 1920\right) = -144.5$$

"Ennustamine" tavalisel juhul oleks üsna tänamatu töö ja lõpeks katastroofiga.

$$\text{predict}\left(\text{andmed2}^{\langle 1\rangle}, 3, 5\right) = \begin{pmatrix} 5.942 \\ 8.883 \\ 15.04 \\ 16.683 \\ 14.13 \end{pmatrix}$$

Kasutades 3 viimast väärtust, saaksime 1920 jäneste arvuks u. 14 tuhat.

$$\text{predict}\left(\text{andmed2}^{\langle 1\rangle}, 15, 5\right) = \begin{pmatrix} 12.909 \\ 19.818 \\ 18.071 \\ 26.762 \\ 22.514 \end{pmatrix}$$

Kasutades kõiki andmeid, saaksime 22.5 tuhat. Mõõdetud tulemus on tegelikult 24.7 ja vastav diferentsiaalvõrrandi mudel annab 22.6. Seega käsk predict töötab siin päris hästi.

Siiski tuleb ilmselt väga suure kahtlusega suhtuda nn. pikkadesse prognoosidesse.

Lineaarme regressioonikõver

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 (seega erinevus interpoleerimisega on selles, et regressioonikõver ei pruugi mõõdetud punkte täpselt läbida).

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.

Käsk **stderr(x-vektor , y-vektor)** väljastab lineaarse regressiooniga seotud vea.

Näide

$N := 50$ Mõõdetavate punktide arv.

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

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

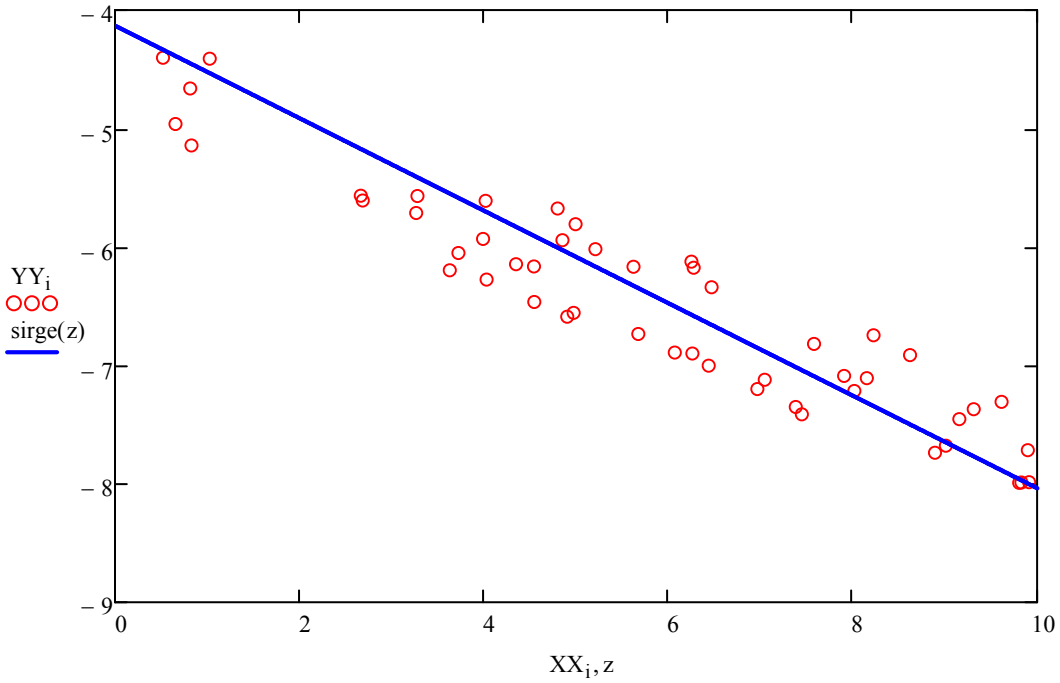
$YY_i := -\frac{1}{3} \cdot XX_i + rnd(1) - 5$ Oletame, et YY väärtused on leitud sellisel moel.

$sirge := line(XX, YY) = \begin{pmatrix} -4.13 \\ -0.391 \end{pmatrix}$ Leiame regressioonisirge nihke ja tõusu.

$slope(XX, YY) = -0.391$ Võrdluseks sirge tõus slope käsuga

$intercept(XX, YY) = -4.13$ Võrdluseks sirge nihe intercept käsuga

$sirge(x) := sirge_1 \cdot x + sirge_0$ Defineerime tõusu ja nihke abil sirge funktsioonina.



$stderr(XX, YY) = 0.694$ Regressiooniga seotud viga.

Polünomiaalne regressioonikõver

Lineaarne regressioon töötab juhtudel, kus andmed asuvad lähedaselt sirgjoonele, kuid läbi kõverjoone ei saa seda edukalt kasutada. Mathcadis on olemas veel kaks abivahendit polünoomide jaoks, s.t. regressioonikõverat üritatakse lähendada polünoomidega.

Käsk **regress(x-vektor , y-vektor , m)** väljastab vektorina kordajad m-järku polünoomi jaoks

(leitud vähimruutude meetodiga). Väljund on mõeldud kasutamiseks koos interp() funktsiooniga. Tüüpiliselt ei ole praktikas soovitatav kasutada kõrgemat kui 4-järku polünoome.

Käsk **loess(x-vektor , y-vektor , ümbrus)** väljastab vektorina kordajad ruutpolünoomide hulga jaoks, mis kõige paremini lähendab etteantud väärtusi. Siin ümbrus>0 ja tähistab punktide maksimaalset ümbruse raadiust, kus kohast lähend peaks läbi minema. Liiga väikese väärtuse korral ei pruugi lahendit leiduda ja liiga suure väärtuse korral ei ole soovitud lähend eriti hea. Väljund on mõeldud kasutamiseks koos interp() funktsiooniga.

Käsk **interp(kordajate-vektor , x-vektor , y-vektor , x)** väljastab lähisväärtuse punktis x, kasutades mõõtmishetki (kohti) "x-vektor", mõõtmisandmeid "y-vektor" ja näiteks käskudega regress() loess() leitud polünoomide kordajate vektorit "kordajate-vektor". Kordajate vektoriks võib olla ka kuupsplaini käskudega leitud vektor (seda vaatleme järgmises loengus).

Näide. Kasutame eespool olevaid andmeid jäneste arvukuse koha.

andmed2 =

	0	1
0	1900	30
1	1901	47.2
2	1902	70.2
3	1903	77.4
4	1904	36.3
5	1905	20.6
6	1906	18.1
7	1907	21.4
8	1908	22
9	1909	25.4
10	1910	27.1
11	1911	40.3
12	1912	57
13	1913	76.6
14	1914	52.3
15	1915	19.5

Funktsiooni loess() väljund:

$abi2 := loess\left(andmed2^{\langle 0 \rangle}, andmed2^{\langle 1 \rangle}, 0.4\right) =$

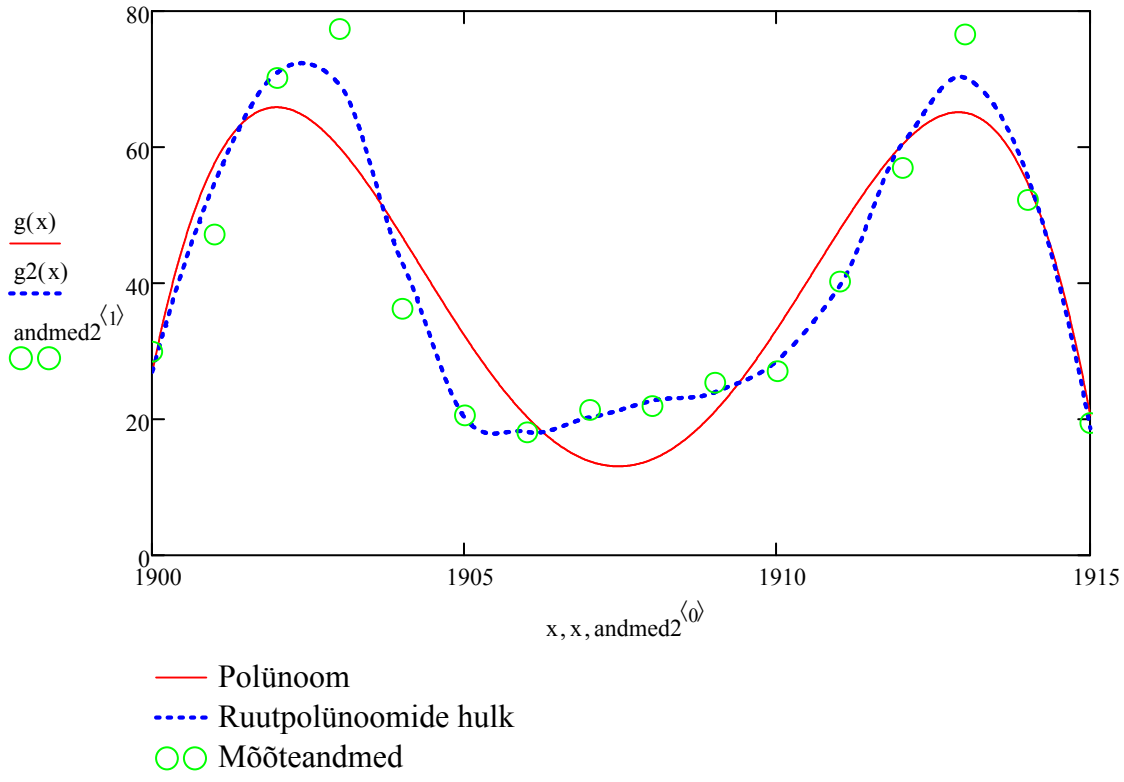
	0
0	1
1	160
2	27.003
3	54.886
4	71.043
5	69.186
6	42.944
7	20.229
8	18.185
9	20.317
10	22.773
11	24.031
12	28.74
13	39.728
14	60.573
15	...

$$abi := regress\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 4\right) = \begin{pmatrix} 3 \\ 3 \\ 4 \\ -785226230365.524 \\ 1646670310.114 \\ -1294936.87 \\ 452.592 \\ -0.059 \end{pmatrix}$$

Neljandat järku polünoome kasutades näeb funktsiooni regress() väljund välja selline.

$g(x) := interp\left(abi, andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$ Polünoomiga leitud kõver.

$g2(x) := interp\left(abi2, andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$ Ruutpolünoomide hulgaga leitud lahend.



Toome võrdluseks veel mõned erinevad kõverad erinevate parameetrite jaoks.

$$rg1(x) := interp\left(regress\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 3\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$

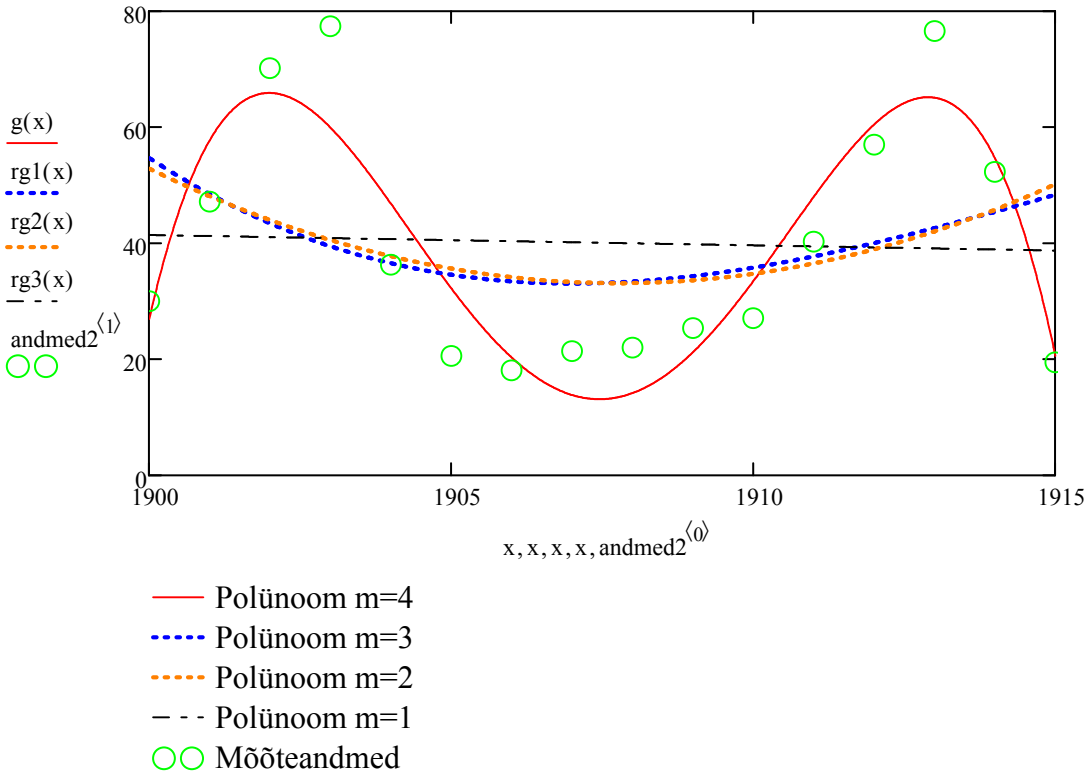
$$rg2(x) := interp\left(regress\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 2\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$

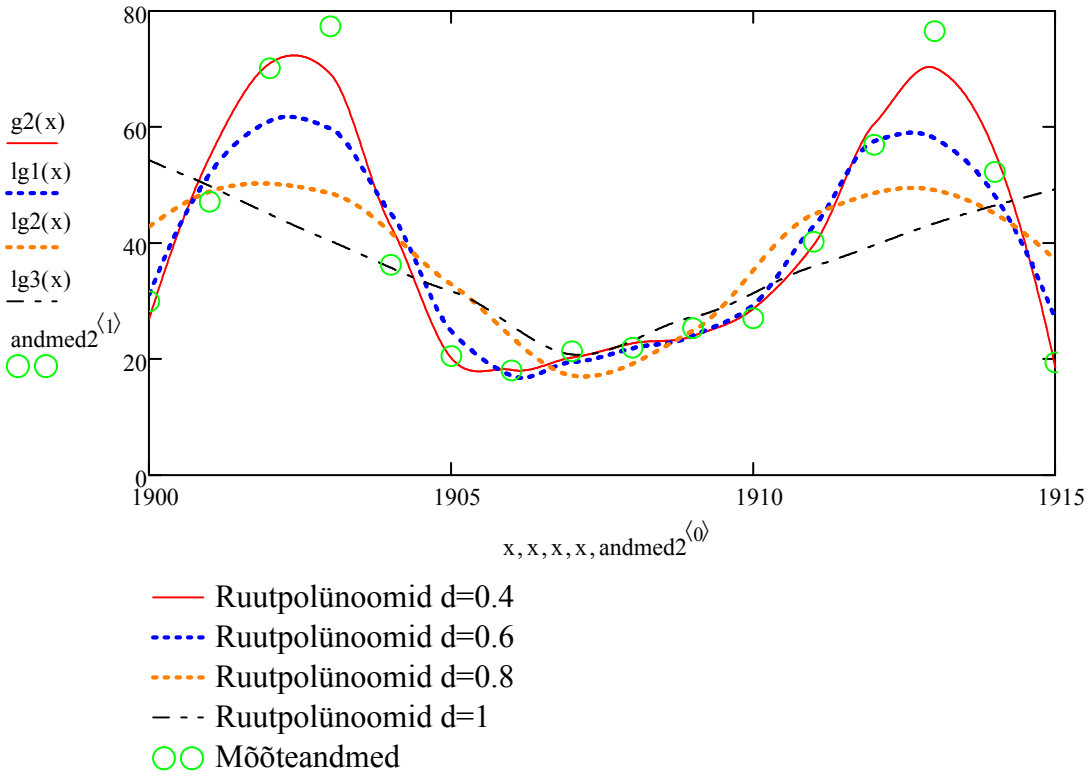
$$rg3(x) := interp\left(regress\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 1\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$

$$lg1(x) := interp\left(loess\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 0.6\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$

$$lg2(x) := interp\left(loess\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 0.8\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$

$$lg3(x) := interp\left(loess\left(andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, 1\right), andmed2^{\langle 0\rangle}, andmed2^{\langle 1\rangle}, x\right)$$





Üldine regressioonikõver

Tihti võib ka polünomiaalne lähenemine olla liiga jäik. Sel juhul võib kasutada üldisi regressioonikõveraid, kuid enamasti need eeldavad seda, et teada on andmete ligikaudne käitumine.

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.

Käsk **genfit(x-vektor , y-vektor , alglähendi-vektor , F-vektor)** väljastab vektorina kordajad mittelineaarsele kombinatsioonile, näiteks Y käitub kui funktsioon $2 \cdot \sin(a_1 \cdot x) + 3 \cdot \tanh(a_2 \cdot x)$. Siin vektor F sisaldab antud funktsiooni ja osatuletisi iga parameetri järgi.

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

$i := 0, 1 .. 50$

$x_i := rnd(10)$

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

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

$$M := \mathrm{csort}(\mathrm{augment}(x, y), 1)$$

Moodustame maatriksi ja järjestame:

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 kuuluvate funktsioonide vektori:

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

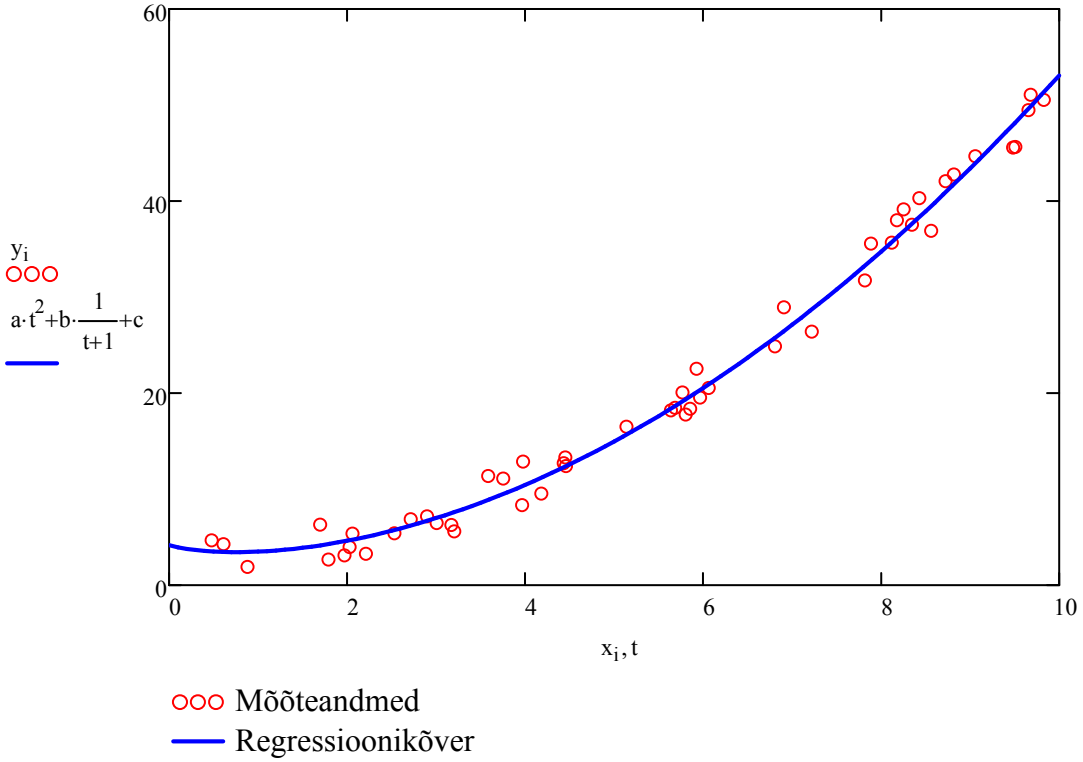
$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} := \mathrm{linfit}\left(M^{\langle 0 \rangle}, M^{\langle 1 \rangle}, F\right)$$

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

$a = 0.511 \qquad b = 2.354 \qquad c = 1.833$

Kujutame regressioonikõvera koos katseandmetega ka graafikul:

$$t := 0, 0.1 .. 10$$



Näide, [1]. Mittelineaarse kombinatsiooni näide, genfit().

Otsime suuruste x, ja y vahelist sõltuvust kujul

$$f(x) = c_1 \cdot e^{c_2 \cdot x}$$

Siin parameetrid c_1 ja c_2 on vaja määrata.

$$x := \begin{pmatrix} -1.4 \\ 0.2 \\ 0.5 \\ 0.7 \\ 1.0 \\ 1.5 \\ 2.4 \end{pmatrix} \qquad z := \begin{pmatrix} 0.442 \\ 0.207 \\ 0.726 \\ 0.309 \\ 0.931 \\ 0.945 \\ 1.389 \end{pmatrix}$$

Mõõtekohad ja mõõtetulemused.

Nüüd moodustame vektorfunktsiooni, mille esimeseks komponendiks on otsitav funktsioon, teiseks komponendiks selle osatuletis esimese otsitava parameetri järgi ja kolmandaks komponendiks osatuletis teise otsitava parameetri järgi:

$$F(x,c) := \begin{pmatrix} c_0 \cdot e^{c_1 \cdot x} \\ e^{c_1 \cdot x} \\ c_0 \cdot c_1 \cdot e^{c_1 \cdot x} \end{pmatrix}$$

Funktsioon genfit() kasutab parameetrite leidmiseks iteratsiooni, seetõttu on vaja ka alglähendite vektorit:

$$\text{alglahend} := \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

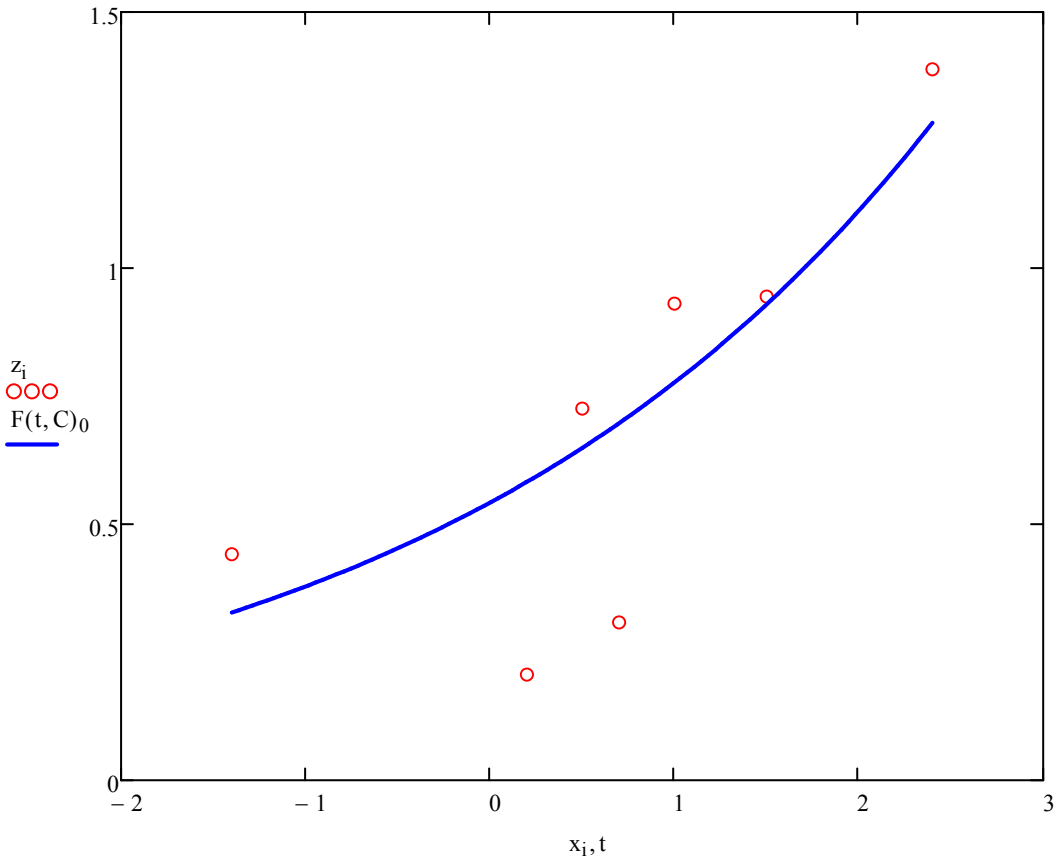
$$C := \text{genfit}(x, z, \text{alglahend}, F) = \begin{pmatrix} 0.543 \\ 0.359 \end{pmatrix}$$

Otsitavate parameetrite leidmiseks rakendame funktsiooni *genfit()* ja salvestame parameetrid vektorisse C.

Kujutame tulemuse graafikul:

$$i := 0 .. 6$$

$$t := \min(x), \min(x) + \frac{\max(x) - \min(x)}{100} .. \max(x)$$



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.