

Aalto Yliopisto

SCI-C0200 - Fysiikan ja matematiikan menetelmien studio

# Tietokoneharjoitus 3: Regressiomallit

Elli Kiiski

## Sisällys

<b>1</b>	<b>Tehtävä A: Lineaarinen energiankulutusmall</b>	<b>3</b>
1.1	Lineaarinen malli . . . . .	3
1.2	Luottamusväli . . . . .	3
<b>2</b>	<b>Tehtävä B: Epälineaarinen kasvumalli</b>	<b>4</b>
2.1	Malli ruisjyvien painolle . . . . .	5
<b>3</b>	<b>Kotitehtävä: Planckin vakion estimointi</b>	<b>6</b>
3.1	Painotetun neliösumman minimointi . . . . .	7
3.2	Luottamusväli . . . . .	9
3.3	Planckin vakion laskeminen . . . . .	9

## 1 Tehtävä A: Lineaarinen energiankulutusmalli

Halutaan muodostaa taulukon 1 datasta lineaarinen regressiomalli kotitalouden energiankulutuksen riippuvuudesta sen varallisuudesta.

varallisuus	energiankulutus
20.0	1.8
30.5	3.0
40.0	4.8
55.1	5.0
60.3	6.5
74.9	7.0
88.4	9.0
95.2	9.1

Taulukko 1: Dataa kotitalouksien varallisuudesta ja energiankulutuksesta

### 1.1 Lineaarinen malli

Muodostetaan ensin lineaarinen malli ja plotataan se samaan kuvaan (kuva 1) alkuperäisten datapisteiden kanssa. Se hoituu seuraavalla koodilla.

---

```
varallisuus = [20.0, 30.5, 40.0, 55.1, 60.3, 74.9, 88.4, 95.2]';
energia = [1.8, 3.0, 4.8, 5.0, 6.5, 7.0, 9.0, 9.1]';
% luodaan lineaarinen malli
model = fitlm(varallisuus, energia, 'linear');
% plotataan datapisteet
figure
plot(varallisuus, energia, 'x', 'color', 'red');
% plotataan mallin antama lineaarinen ennuste
...
t = [10:120]';
plot(t, predict(model, t), 'color', 'blue');
...
```

---

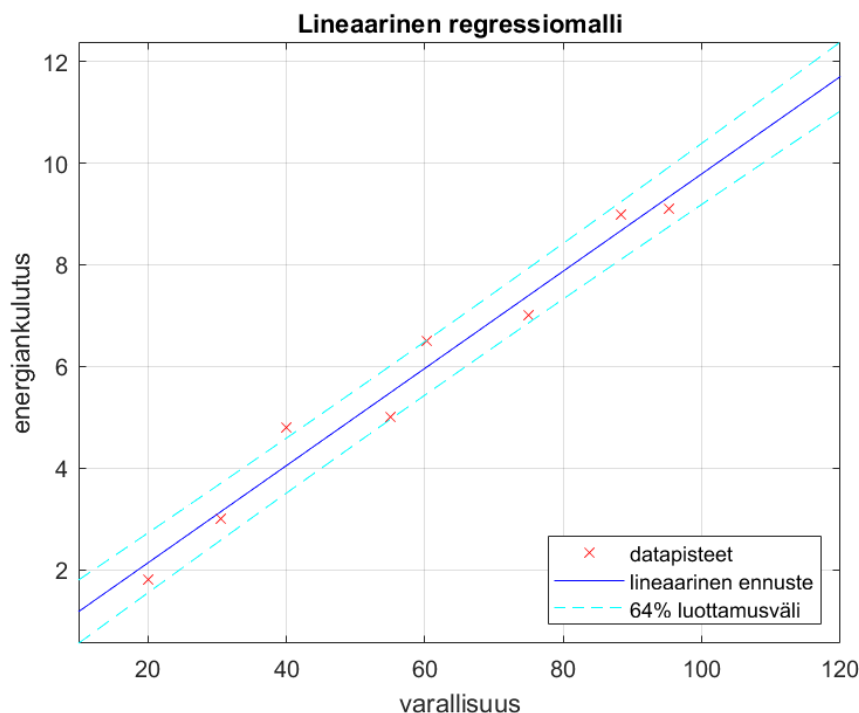
### 1.2 Luottamusväli

Lasketaan seuraavaksi ennusteelle 64% luottamusvälit ja plotataan ne samaan kuvaan (kuva 1) seuraavalla koodilla.

---

```
% tallennetaan 64% luottamusvalin ylä- ja alarajat matriisiin eci
[epred, eci] = predict(model, t, 'alpha', 0.36 ,
    'Prediction', 'observation');
% plotataan luottamusvalin ylä- ja alarajat
plot(t, eci(:,1), '--', 'color', 'cyan');
plot(t, eci(:,2), '--', 'color', 'cyan');
```

---



Kuva 1: Lineaarinen regressiomalli kotilaouksien varallisuuden vaikutuksesta energiankulutukseen

Esimerkiksi varallisuustason 65 energiankulutus osuu 64% luottamusvälin perusteella välille [5.91, 6.97].

## 2 Tehtävä B: Epälineaarinen kasvumalli

Tiedetään, että jyvien painon  $w$  keshitystä ajan  $t$  suhteen voidaan mallintaa yhtälöllä

$$w = f(t, w_{max}, t_{max}, t_v) = w_{max} \left(1 + \frac{t_{max} - t}{t_{max} - t_v}\right) \left(\frac{t}{t_{max}}\right)^{\frac{t_{max}}{t_{max} - t_v}} \quad (1)$$

kun jyvä ei ole vielä saavuttanut maksimipainoaan, eli kun  $0 \leq t \leq t_{max}$ . Kun maksimipaino on saavutettu eli kun  $t \geq t_{max}$ , paino pysyy vakiona eli  $f(t, w_{max}, t_{max}, t_v) = w_{max}$ .

Mallin parametrit tarkoittavat seuraavaa:

- $w_{max}$  on jyvän saavuttama maksimipaino
- $t_{max}$  on ajanhetki, jona maksimipaino saavutetaan
- $t_v$  on ajanhetki, jona jyvän kasvunopeus on suurimmillaan

## 2.1 Malli ruisjyvien painolle

Halutaan selvittää millä parametrien  $w_{max}$ ,  $t_{max}$  ja  $t_v$  arvoilla malli sopii taulukon 2 mittausdataan mahdollisimman hyvin.

jyvän paino (0.1g)	päiviä 1. kukinnosta
2	2
8	10
19	18
32	26
43	31
45	35

Taulukko 2: Jyvän painon mittaustuloksia

Käytetään tehtävän ratkaisuun PNS-menetelmää. Muodostetaan ensin funktio, joka ottaa parametreikseen vektorin  $\beta = (w_{max}, t_{max}, t_v)$  sekä vektorin ajanhetkiä, joiden ajalle malli sovitetaan. Funktio palauttaa vektorin  $w$ , joka sisältää mallilla lasketut jyvän painot kullakin vektorin  $t$  ajanhetkellä.

---

```
function w = kasvumalli(beta, t)
% tallennetaan parametrien arvot kuvaaviin muuttujiin
wmax = beta(1); tmax = beta(2); tv = beta(3);
% lasketaan jyvien painot vektorin t ajanhetkellä
w = wmax.*(1+(tmax-t)/(tmax-tv)).*(t./tmax).^(tmax/(tmax-tv));
% huolehditaan, ettei jyvän paino lähde laskuun
indmax = find(w==max(w));
w(indmax:end)=max(w);
```

---

Etsitään seuraavaksi PNS-menetelmällä vektori  $\beta$  eli parametrit  $w_{max}$ ,  $t_{max}$  ja  $t_v$ , joilla malli sopii taulukon 2 mittausdataan parhaiten. Käytetään alkuarvausta  $\beta = (40, 40, 20)$ .

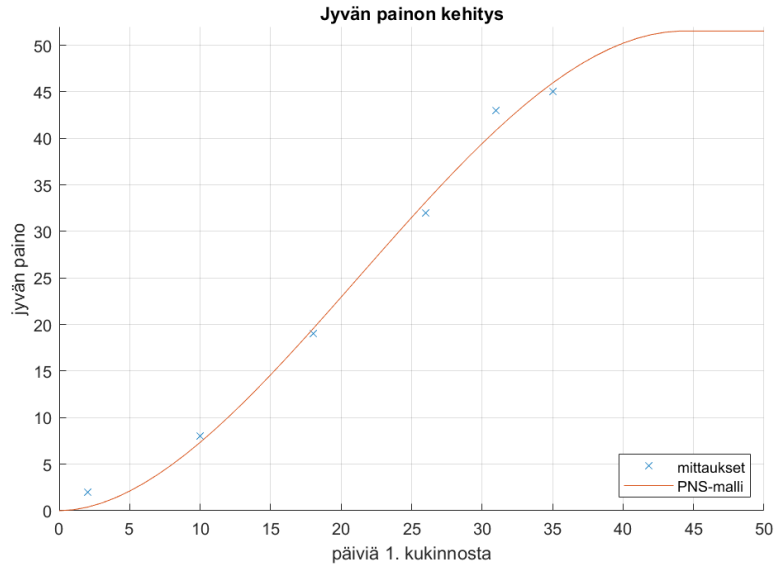
---

```
% alustetaan mittausdata
days = [2, 10, 18, 26, 31, 35];
weight = [2, 8, 19, 32, 43, 45];
% estimoidaan parhaat parametrien arvot
t = [0:50];
est = lsqcurvefit(@(b,t) kasvumalli(b,t), [40,40,20], days, weight);
```

---

Kuvassa 2 näkyvät sekä alkuperäiset datapisteet että mallin tuottama ennustekäyrä. Huomataan, että mallin mukaan jyvän paino kasvaa aluksi maltillisesti kunnes vauhti kiihtyy ja lopuksi hyytyy taas. Kun maksimipaino on saavutettu, ei paino luonnollisestikaan lähde siitä enää laskemaan. Itse asiassa käyrä muistuttaa hieman Poisson-jakaumaa.

Voidaan tietenkin vielä määrittää ennustettu maksipaino ja päivä, jolloin se



Kuva 2: Jyvän painon kehitys ajan suhteen mallinnettuna PNS-menetelmällä mittausdatan pohjalta.

saavutetaan. Kuvaajan mukaan maksimipaino 5.15g saavutetaan 45. päivänä ensimmäisen kukinnon jälkeen. Siispä esimerkiksi 35. päivänä tiedetään, että sato kannattaa korjata kymmenen päivän kuluttua.

### 3 Kotitehtävä: Planckin vakion estimointi

Tavoitteena on selvittää kokeellisen mittausdatan perusteella arvio Planckin vakiolle. Hyödynnetään tehtävässä yhtälöä

$$V_0 = -\frac{h}{e}f + C, \quad (2)$$

jossa  $V_0$  on pysäytysjännite,  $f$  valon tajuus,  $h$  Planckin vakio  $e$  elektronin varaus ja  $C$  vakiotermi. Tämän avulla voidaan taulukon 3 mittaustulosten avulla selvittää niihin sovitetun suoran kulmakerroin, joka vastaa termiä  $-\frac{h}{e}$ , josta Planckin vakio  $h$  voidaan helposti laskea.

Mittausdatan tarkkuuden vaihtelun vuoksi ei ole järkevää käyttää tavallista `fitlm`-mallia, joten käytetään hyödyksi menetelmää, jossa minimoidaan painotettua neliösummaa

$$\sum_{i=1}^n \left( \frac{y_i - (b_1 x_i + b_0)}{\Delta y_i} \right)^2, \quad (3)$$

missä  $\Delta y_i$  on havainnon  $i$  virhetermi.

taajuus $f$ ( $10^{12}$ Hz)	pysäytysjännite $V_0$ (V)
519	$1.0 \pm 0.15$
549	$1.2 \pm 0.40$
688	$1.9 \pm 0.20$
740	$2.4 \pm 0.30$
821	$2.3 \pm 0.10$

Taulukko 3: Valon taajuuden mittaustietoon virheetöntä, mutta pysäytysjännitteelle on arvioitu virherajat.

Analyttisesti voidaan ratkaista neliösumman minimoivat parametrien arvot:

$$b_1 = \frac{1}{D} \left( \left( \sum_{i=1}^n \frac{1}{(\Delta y_i)^2} \right) \left( \sum_{i=1}^n \frac{x_i y_i}{(\Delta y_i)^2} \right) - \left( \sum_{i=1}^n \frac{x_i}{(\Delta y_i)^2} \right) \left( \sum_{i=1}^n \frac{y_i}{(\Delta y_i)^2} \right) \right) \quad (4)$$

ja

$$b_0 = \frac{1}{D} \left( \left( \sum_{i=1}^n \frac{x_i^2}{(\Delta y_i)^2} \right) \left( \sum_{i=1}^n \frac{y_i}{(\Delta y_i)^2} \right) - \left( \sum_{i=1}^n \frac{x_i}{(\Delta y_i)^2} \right) \left( \sum_{i=1}^n \frac{x_i y_i}{(\Delta y_i)^2} \right) \right) \quad (5)$$

missä

$$D = \left( \sum_{i=1}^n \frac{1}{(\Delta y_i)^2} \right) \left( \sum_{i=1}^n \frac{x_i^2}{(\Delta y_i)^2} \right) - \left( \sum_{i=1}^n \frac{x_i}{(\Delta y_i)^2} \right)^2. \quad (6)$$

Parametrien arvioille voidaan laskea myös virheet

$$\Delta b_1 = \sqrt{\frac{1}{D} \sum_{i=1}^n \frac{1}{(\Delta y_i)^2}} \quad (7)$$

ja

$$\Delta b_0 = \sqrt{\frac{1}{D} \sum_{i=1}^n \frac{x_i^2}{(\Delta y_i)^2}}, \quad (8)$$

joiden avulla arvioille saadaan 64% luottamusväli  $[b_i - \Delta b_i, b_i + \Delta b_i]$ .

### 3.1 Painotetun neliösumman minimointi

Luodaan ensin MATLABilla funktio, joka ottaa parametreikseen taajuus-, jännite- ja virherajavektorit. Funktio palauttaa ensimmäisenä arvonaan vektorin  $b = (b_0, b_1)$ , joka sisältää arviot termeille  $b_0$  ja  $b_1$ , jotka ovat siis sovittavan suoran  $V = b_0 + b_1 f$  vakiotermi ja kulmakerroin. Toisena arvona palautetaan matriisina ala- ja ylärajat termien  $b_0$  ja  $b_1$  virheille.<sup>1</sup>

<sup>1</sup> Aiemmassa versiossa, jossa funktio tuotti vääriä tuloksia oli tapahtunut klassinen murtovirhe: funktion parametreina nykyisten `xdata` ja `ydata` tilalla oli `x` ja `y`, jolloin niiden arvoja muutettiin myöhemmillä riveillä, vaikka tarkoitus oli luoda uudet muuttujat.

---

```

function [b, bci] = sovittaja(xdata, ydata, dy)
% LASKETAAN YKSITTAISET SUMMALAUSEKKEET
% 1 / dy^2
e = sum(1./(dy.^2));
% x / dy^2
x = sum(xdata./(dy.^2));
% y / dy^2
y = sum(ydata./(dy.^2));
% x * y / dy^2
xy = sum((xdata.*ydata)./(dy.^2));
% x^2 / dy^2
x2 = sum((xdata.^2)./(dy.^2));
% kerroin 1/D
d = 1./((e*x2)-(x^2));
% LASKETAAN TERMIT b1 JA b0
b1 = d*((e*xy)-(x*y));
b0 = d*((x2*y)-(x*xy));
% LASKETAAN VIRHEET TERMEILLE b1 JA b0
eb1 = sqrt(d*e);
eb0 = sqrt(d*x2);
% TALLENNETAAN ARVOT PALAUTETTAVIIN MUUTTUJIIN
b = [b0 b1];
bci = [(b0 - eb0) (b1 - eb1); (b0 + eb0) (b1 + eb1)];

```

---

Arvioidaan parametrit  $b_0$  ja  $b_1$  käyttäen yllä määriteltyä funktiota `sovittaja` ja muodostetaan saatujen arvojen avulla lineaarinen regressiosuora. Plotaan samaan kuvaajaan (kuva 3) sekä mittausdata virherajoiheen että kyseinen suora taajuusvälillä [400, 900].

---

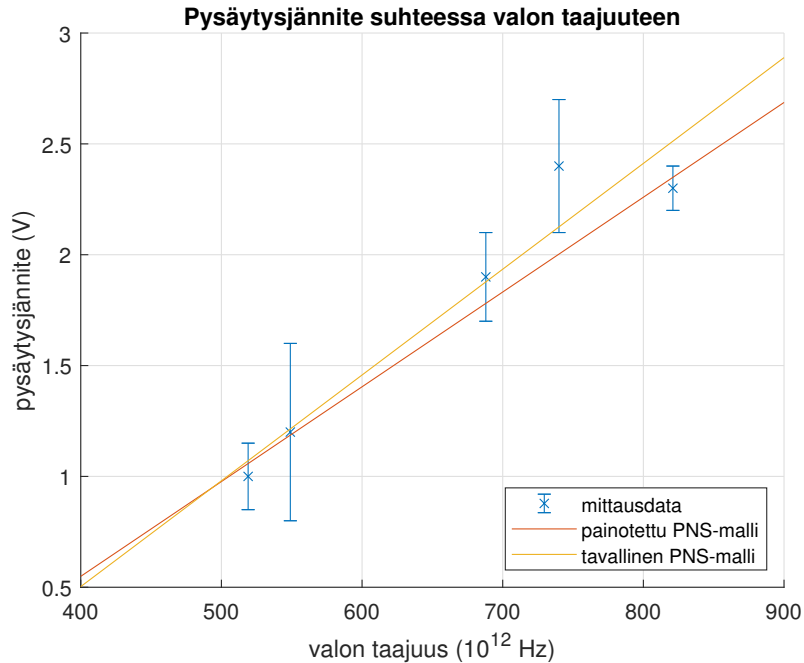
```

% alustetaan mittausdata vaektoreihin
f = [519, 549, 688, 740, 821];
v = [1.0, 1.2, 1.9, 2.4, 2.3];
e = [0.15, 0.40, 0.20, 0.30, 0.10];
% estimoidaan b0 ja b1 sovittaja-funktiolla
[b, bci] = sovittaja(f, v, e);
b0 = b(1);
b1 = b(2);
% määritetään regressiosuora välillä [400, 900]
x = [400:900]';
y = b0 + b1 .* x;
% sovitetaan mittausdataan myös tavallinen lineaarinen malli
model = fitlm(f, v, 'linear');
% plotataan kaikki samaan kuvaan
...
errorbar(f, v, e, 'x');
plot(x, y);
plot(x, predict(model, x));
...

```

---





Kuva 3: Mittaustulokset virheväleineen sekä niihin sovitetut painotettu ja tavallinen lineaarinen regressiosuora.

### 3.2 Luottamusväli

Mallin luottamusvälien päätepisteet saadaan niin ikään **sovittaja**-funktion palautusarvona, ja voidaan laskea seuraavasti

---

```
% parametrien b0 ja b1 luottamusvalit
eb0 = [bci(1,1) bci(2,1)]
eb1 = [bci(1,2) bci(2,2)]
```

---

Luottamusväleiksi saadaan  $eb0 = [-1.5796, -0.7421]$  ja  $eb1 = [0.0037, 0.0048]$ , mikä tarkoittaa erityisesti sitä, että välille  $[0.0037, 0.0048]$  pitäisi kuulua se kulmakertoimen arvo, joka on kirjallisuudesta löytyvän Planckin vakion mukainen.

### 3.3 Planckin vakion laskeminen

Lasketaan vielä saatujen tulosten pohjalta arvio Planckin vakiolle. Kuten aiemmin todettu, sovitetun regressiosuoran kulmakerroin kuvaa termiä  $-\frac{h}{e}$ , jossa  $e = -1.6 \cdot 10^{-19} C$  on elektronin varaus ja  $h$  selvitettävänä oleva Planckin vakio. Toisaalta suoran kulmakertoimen  $b_1$  on myös mitattu kuuluvan todennäköisesti välille  $[0.0037 \cdot 10^{-12} Vs, 0.0048 \cdot 10^{-12} Vs] = [b_{1min}, b_{1max}]$ .

Voidaan ratkaista yhtälöt

$$h_{min} = -e \cdot b_{1_{min}} \quad (9)$$

$$= -(-1.6 \cdot 10^{-19} \text{ C}) \cdot 0.0037 \cdot 10^{-12} \text{ V s} \quad (10)$$

$$= 5.92 \cdot 10^{-34} \text{ J s}. \quad (11)$$

ja

$$h_{max} = -e \cdot b_{1_{max}} \quad (12)$$

$$= -(-1.6 \cdot 10^{-19} \text{ C}) \cdot 0.0048 \cdot 10^{-12} \text{ V s} \quad (13)$$

$$= 7.68 \cdot 10^{-34} \text{ J s}. \quad (14)$$

Huomamantaan, että oikealle Planckin vakiolle pätee  $h = 6.63 \cdot 10^{-34} \in [h_{min}, h_{max}]$ . Näin ollen voidaan uskoa mallinnuksen ja laskelmien onnistuneen.