Projekt 9.8 Multilineær regression

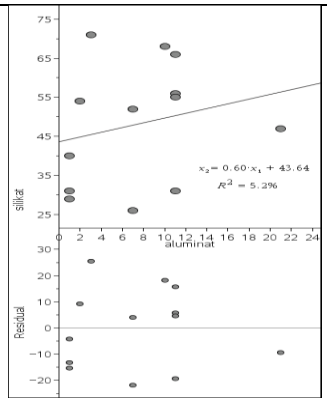
Vi undersøger i dette projekt den problemstilling, at en afhængig variabel *y*, kaldet *responsvariablen,* kan tænkes at afhænge af to eller flere uafhængige variable x_1, x_2,… , kaldet de *forklarende variable* (explanatory variables). Som et konkret eksempel ser vi på hærdning af cement. Under hærdningen udvikles varme, og varmeudviklingen afhænger af cementens sammensætning. Ved fremstilling af klinker er man især interesseret i indholdet af tricalcium-aluminat, $3CaO \cdot Al_2O_3$, og tricalcium-silicat, $3CaO \cdot SiO_2$. Ved måling af vægtprocenterne og varmeudviklingen (i kalorier pr. gram cement) fandt man følgende data:

x_1: Vægtprocent (aluminat)	7	1	11	11	7	11	3	1	2	21	1	11	10
x_2: Vægtprocent (silikat)	26	29	56	31	52	55	71	31	54	47	40	66	68
y: Varme (kalorier pr gram)	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

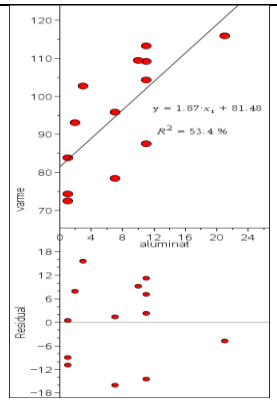
Øvelse 1

Konstruer et punktplot for *y* som funktion af x_1 henholdsvis *y* som funktion af x_2,

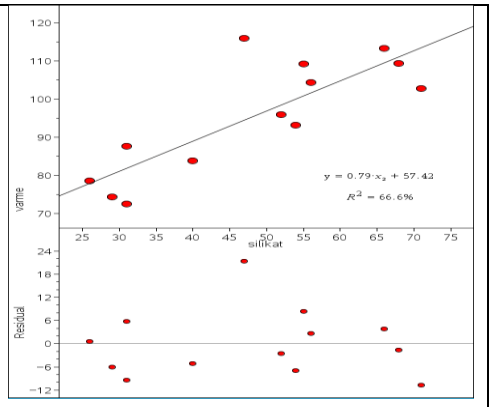
Du skal få grafer som de følgende (de følgende er beregnet og tegnet i TI Nspire):



Aluminat versus silikat



Aluminat versus varme

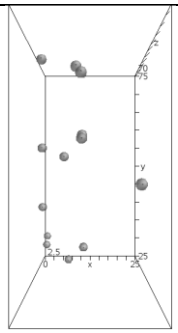


Silikat versus varme

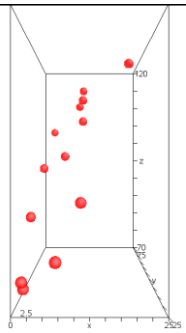
Ingen af dem ser specielt overbevisende ud som lineære sammenhænge, men vi ser dog tydeligt voksende sammenhænge.

Residualerne ligger typisk fra -10 til 10 (eller højere), så de lineære modeller er ret grumsede. Heldigvis ser der ikke ud til at være nogen sammenhæng mellem de to uafhængige variable, aluminat- og silikatindholdet. Her er forklaringsgraden helt nede på 5.2%, og de spreder sig godt ud over det todimensionale område i x_1x_2-planen. De to uafhængige variable kan altså godt antages at være lineært uafhængige.

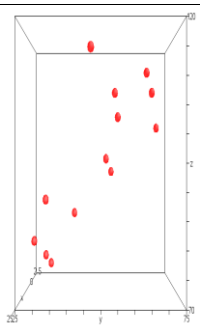
Vi vil derfor i stedet prøve os frem med en samlet model for de tre variable. Vi afsætter da datasættet som et tredimensionalt punktplot med x_1 og x_2 ud af første- og andenaksen, og *y* op ad tredjeaksen. Kigger vi lige ind langs x_2-aksen eller lige ind langs x_1-aksen fås de samme billeder som før med en noget mudret lineær sammenhæng.



Et lodret kig ind på x_1x_2-planen.



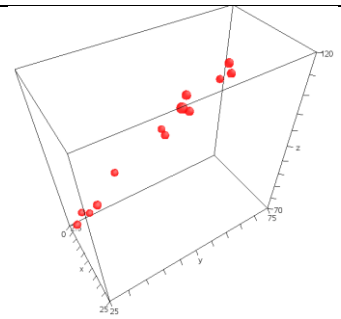
Et vandret kig ind på x_1y-planen.



Et vandret kig ind på x_2y-planen.

Øvelse 2.

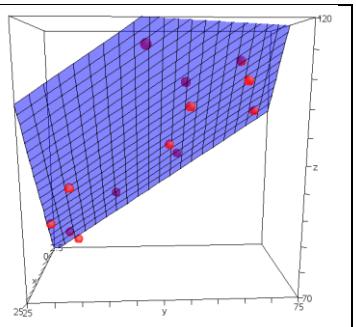
Prøv nu at dreje koordinatsystemet med punkterne afsat, så du ser skråt ind på datasættet.
Ved at prøve dig frem, vil du se at datapunkterne faktisk klumper sammen, bare på en anden måde, end vi først eftersøgte.
Du skal få noget der ligner dette billede:
(Et skråt kig ind i x_1x_2y -rummet:
Datapunkterne klumper sig sammen.)



Det viser, at datapunkterne med god tilnærmelse ligger samlet i en plan, som vi kan betragte som grafen for en lineær funktion i to variable:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 .$$

For at finde den bedste lineære funktion, der går gennem datasværmen vil vi som sædvanligt forsøge at minimere summen af de kvadratiske afvigelser. Vi kunne gå frem ligesom ved den lineære regression og bestemme parametrene én for én, men løsningsformlerne bliver stadigt mere komplicerede, så vi nøjes med at vise den konkrete *numeriske* løsning af ligningerne. De generelle løsningsformler er gemt inde i værktøjsprogrammet, så det er trivielt at finde ligningen for den multilineære regression ved hjælp af et værktøjsprogram, også i de tilfælde, hvor der er mere end to uafhængige variable.



Den bedste rette plan gennem datapunkterne. Data punkterne følger planen tæt, nogen over andre under.

Øvelse 3. Bedste lineære sammenhæng

Bestem i dit værktøjsprogram den bedste lineære sammenhæng.

I Maple ser det således ud:

Vi lægger datafilen ind som en matrix og anvender kommandoen MultiLinReg i Gympakken:

MultiLinReg(M): $y = 52.577 + 1.468 \cdot x_1(\text{aluminat}) + 0.662 \cdot x_2(\text{silikat})$

Samtidig får vi en række oplysninger om kvaliteten af regressionen, der bla. angiver hvad usikkerheden er på bestemmelsen af koefficienterne:

Koefficient	Estimat	Standardfejl	t-stat	p-værdi	Nedre 95 %	Øvre 95%
a_0	52.577	2.286	22.998	0.	47.483	57.671
a_1	1.468	0.121	12.105	0.	1.198	1.739
a_2	0.662	0.046	14.442	0.	0.560	0.764
	R^2	R^2 justeret	Observationer			
	0.979	0.974	13			
Model	Frihedsgrader	SSM	MSM	F	Signifikans F	
	2	2657.9	1328.9	229.50	$4.41\ 10^{-9}$	
Residualer	Frihedsgrader	SSE	MSE	Standardafvigelse		
	10	57.904	5.7904	2.406		

Sammen med udregningen kan vi få residualerne. Vi kan i tabellen se, at spredningen på residualerne er 2,4.
I Maple ligger der forklaringer på, hvad hver enkelt af disse felter står for – vi vil ikke her gå længere ind i dette. Men vi konstaterer, at vi får en ligning for den bedste lineære model i 2 dimensioner, dvs den plan, der tilnærmer punkterne bedst.

I TI Nspire kan det se således ud:

$$\textbf{\textit{kvad}} = \text{sum}\left((\textbf{\textit{varme}} - b_0 - b_1 \cdot \textbf{\textit{aluminat}} - b_2 \cdot \textbf{\textit{silikat}})^2\right)$$

$$= 13b_0^2 + 1139b_1^2 + 33050b_2^2 + 194b_0b_1 + 1252b_0b_2 + 9844b_1b_2 - 2481b_0 - 20064b_1 - 124056b_2 + 121088$$

$$\text{fMin}(\textbf{\textit{kvad}}, b_0) \blacktriangleright b_0 = \tfrac{1}{13} \cdot (-97b_1 - 626b_2 + 1240.5)$$

$$\text{fMin}(\textbf{\textit{kvad}}, b_1) \blacktriangleright b_1 = \tfrac{1}{1139} \cdot (-97b_0 - 4922b_2 + 10032)$$

$$\text{fMin}(\textbf{\textit{kvad}}, b_2) \blacktriangleright b_2 = \tfrac{1}{16525} \cdot (-313b_0 - 2461b_1 + 31014)$$

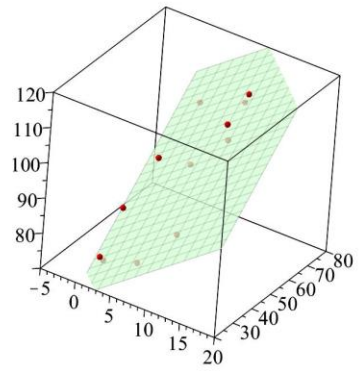
$$\text{solve}\left(\begin{cases}\text{fMin}(\textbf{\textit{kvad}}, b_0)\\ \text{fMin}(\textbf{\textit{kvad}}, b_1)\\ \text{fMin}(\textbf{\textit{kvad}}, b_2)\end{cases}, b_0, b_1, b_2\right) \blacktriangleright b_0 = 52.5773 \text{ and } b_1 = 1.46831 \text{ and } b_2 = 0.66225)$$

Konklusion: $y = 52.577 + 1.468 \cdot x_1(\text{aluminat}) + 0.662 \cdot x_2(\text{silikat})$

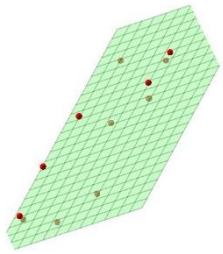
Altså præcis samme ligning, når vi bestemmer koefficienterne skridt for skridt.

Her står konstantleddet 52.57 for varmeudviklingen i fravær af aluminat og silikat (skæringen med *y*-aksen). Tilsvarende står koefficienterne 1.47 og 0.66 for hældningerne langs x_1- og x_2-akserne, dvs. for hver gang aluminatindholdet stiger med 1 vægtprocent vokser varmeudviklingen med 1.47 kalorier pr. gram og for hver gang silikatindholdet stiger med 1 vægtprocent vokser varmeudviklingen med 0.66 kalorier pr. gram.

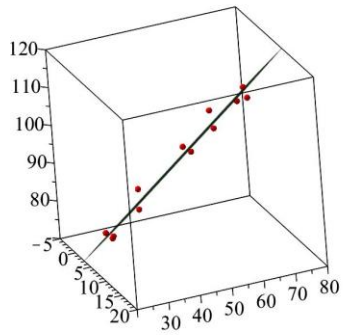
Den grafiske fremstilling af regressionsplanen sammen med punkterne (denne gang i Maple):



Planen er gjort transparent, så vi svagt ser punkterne på den anden side



Vi har fjernet koordinatsystemets ramme, så vi bedre ser plan og punkter



Vi ser her ind langs kanten af regressionsplanen

Øvelse 4. Residualerne

Beregn residualerne ud fra definitionen, dvs som:

Forventede værdier (ud fra modellen) – Empiriske værdier (datasættet)

Du skal finde, at residualerne ligger mellem -3 og 5 og at langt de fleste ligger mellem -2 og 2.

Konklusionen på det visuelle i øvelse 3, og beregningerne i øvelse 4 er altså, at den multilineære regressionsmodel giver en bedre og mere sammenhængende beskrivelse af varmeudviklingens afhængighed af cements sammensætning end de enkelte lineære regressionsmodeller gør det.

Anvendelser af multilineær regression

Potensmodeller med to eller flere uafhængige variable

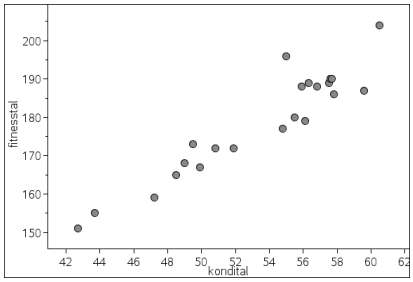
De *generelle potensmodeller* er meget udbredte i naturvidenskab samt i økonomi og samfundsvidenskab:

$$y = b_0 \cdot x^{b_1} \cdot x_2^{b_2} \cdot \ldots \cdot x_n^{b_n}.$$

Ved at transformere både responsvariablen *y* og de forklarende variable $x_1, x_2, \ldots, x_n$ med en logaritmisk transformation, omformes denne model netop til en multilineær model i $\left(\ln(x_1), \ldots, \ln(x_n), \ln\left(y\right)\right)$:

$$\ln(y) = \ln(b_0) + b_1 \cdot \ln(x_1) + b_2 \cdot \ln(x_2) + \ldots + b_n \cdot \ln(x_n) \qquad (*)$$

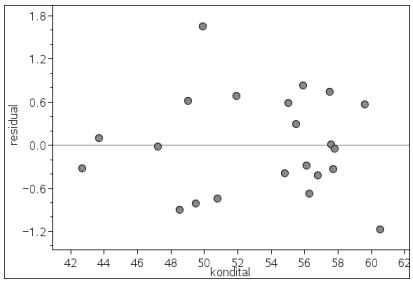
Lad os som et simpelt eksempel se på fitnesstallet fra U18-datasættet. Du kan hente datasættet ***her***. Afsættes fitnesstallet som funktion af konditallet, er det klart, at der ikke er tale om en simpel variabelsammenhæng. Vi inddrager derfor også vægten, for at se om fitnesstallet med rimelighed kan beskrives som en potensfunktion af konditallet og vægten. Vi udfører derfor multilineær regression på de transformerede data med ln(fitnesstal) som funktion af ln(vægt) og ln(kondital).



Øvelse 5.

Gennemfør selv dette. Du skal få potensmodellen

$$\text{fitnesstal} = 1.35 \cdot \text{kondital}^{0.961} \cdot \text{vægt}^{0.238}$$



Resdiualplottet viser da også en rimelig tilfældig fordeling af fejlene med en typisk fejl mellem -1 og 1, hvilket synes rimeligt nok, da fitnesstallet er udregnet som et helt tal. Vi har altså en udmærket deskriptiv model. Men vi har ingen begrundelse for modellen, og det ville være mærkeligt, hvis det var lige netop den, Idrætsforskeren Lars Michalsik havde brugt til at omsætte vægt og kondital til et fitnesstal

6.2 Polynomielle regressionsmodeller

De *polynomielle modeller*, der også er meget udbredte i naturvidenskaberne, ser således ud:

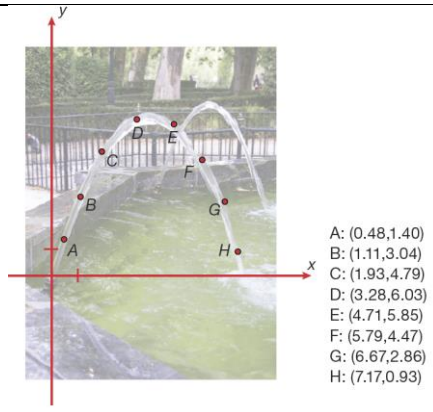
$$y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + \ldots + b_n \cdot x^n.$$

Ved at erstatte potenserne med forklarende variable omformes den til en multilineær model:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n$$

$$x_1 = x, \quad x_2 = x^2, \ldots, \quad x_n = x^n$$

De polynomielle modeller er indbygget direkte i de fleste værktøjsprogrammer. Vi illustrerer dette med en undersøgelse af de kurver, som et springvands vandstråler aftegner. Vi undersøgte eksemplet i HEM2, kapitel 2. Den gang anvendte vi en kvadratisk regression, men nu vil vi i stedet se på en multilineær regression og tilføjer derfor variablen $x_2 = x^2$ til modellen.



Øvelse 6

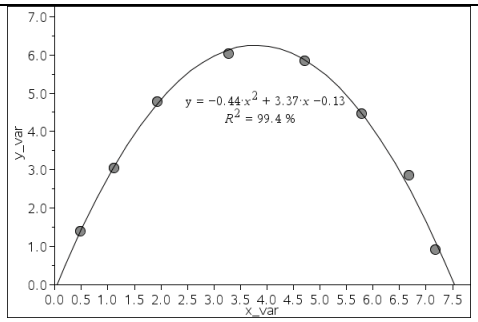
a) Indtast selv datasættet og udfør multilineær regression. Du skal få:

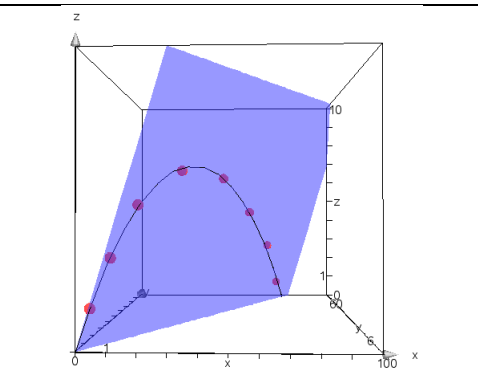
$$y = -0.13 + 3.37 \cdot x_1 - 0.44 \cdot x_2$$

b) Andengradsregression giver naturligvis "det samme":

$x_1 = t, x_2 = t^2, y = -0.13 + 3.37 \cdot t - 0.44 \cdot t^2$

Tegn den plane kurve i et 2d koordinatsysteme, og tegne regressionsplanen med punkterne i et 3d koordinatsystem. Du skal få noget der ligner disse grafer:





c) Udfører vi nu en multilineær regressionstest, ser vi at *p*-værdierne for de tre koefficienter er givet ved

b-værdier	$b_0 = -0.1267$	$b_1 = 3.368$	$b_2 = -0.4446$
p-værdier	0.529 = 52.9%	$1.32 \cdot 10^{-6}$	$1.10 \cdot 10^{-6}$

Vi ser da, at konstantleddet har en meget høj *p*-værdi, og vi kan derfor ikke afvise, at skæringen med *y*-aksen alene afspejler de stokastiske fluktuationer i modellen. Modellen kan derfor forenkles til et andengradspolynomium, der går gennem (0,0): $y = 3.37 \cdot x - 0.44 \cdot x^2$.