Undervisning 8

Regression

Regression

Hvis man foretager en række målinger i forbindelse med et eksperiment, er man ofte interesseret i at opstille en matematisk model over eksperimentet. Dette betyder, at man ønsker at afgøre, om målingerne følger en bestemt sammenhæng.

Ved et eksperiment kan man selvfølgelig ikke regne med, at målepunkterne præcis ligger på en ret linje eller præcis på en anden kurve. I reglen er man derfor tilfreds med at fastslå, at punkterne med god tilnærmelse følger en bestemt kurve. Man siger, at man fortager en *regression* på de målte data.

Regression(Lineær)

Man har en gruppe af datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, som kan man teste ved hjælp af programmer som Maple og Geogebra om disse værdier kan beskrives ud fra modellen

$$y = a \cdot x + b$$

Det kaldes at bestemme bedste rette linje.

Regression(Lineær).

Bemærk her at modellens type beskrives i opgaven og f(x) anvendes i stedet for y. (Dette kan variere fra opgave til opgave).

Tit vil den ønsket anvendte model kunne fremstå beskrevet med tekst og ikke som formel.

Dvs. f.eks. ”Anvend lineær regression til at bestemme en model for udviklingen”.

Eksempel 1(Lineær regression)

Simone driver fabrik med tilhørende butik. Der produceres på fabrikken kun det antal enheder af en bestemt vare, som matcher efterspørgslen i butikken.

Produktionen følges over en tiårig periode fra 2005 til 2015.

Nedenstående tabel viser antal producerede enheder i den ovennævnte periode.

år	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Antal producerede enheder (målt i tusinder)	171	321	471	589	712	811	929	1043	1192	900	1081

Udviklingen i produktionen af enheder på Simones fabrik kan beskrive ud fra funktionen

$$f(x) = a \cdot x + b$$

Hvor $f(x)$ er antal solgte enheder til tiden x målt i år efter 2005.

a) Bestem forskriften for sammenhængen $f(x)$ ved hjælp af lineær regression i Maple. Samt beskriv betydningen af konstanten a og b i $f(x)$.

Regression(Lineær)

Vigtig!
Bemærk her, at indholdet af tabellen fra opgaven indsættes i lister. Elementerne i listerne adskilles af alm. komma og ikke punktummer.

Samt husk at navngive listerne det som de hedder i opgaverne, og ikke x eller y. Idet det kan give problemer.

Løsning : **Eksempel 1**

I Maple laves 2 lister som vist nedenfor, hvori værdien fra tabellen i opgaven indsættes.

Først indsættes værdierne fra tabellen i 2 lister med navnene *år* og *enheder*.

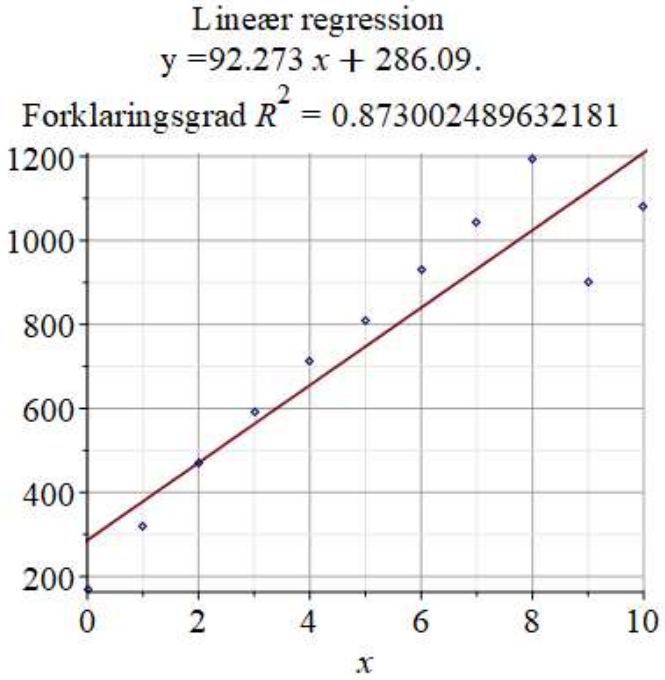
$$år := [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]:$$
$$enheder := [171, 321, 471, 591, 712, 811, 929, 1043, 1192, 900, 1081]:$$

(Husk vi arbejder med antal år efter begyndelsesåret, som vores x-værdier.)

Regression(Lineær)

Løsning fortsat: **Eksempel 1:** Efterfølgende, for at bestemme bedste rette linje anvendes kommandoen LinReg i Maple, som vist nedenfor.

$$LinReg(år, enheder)$$



Regression(Lineær)

Løsning fortsat: **Eksempel 1**

Alternativt kan følgende kommando anvendes i Maple til at bestemme funktionen, som beskriver bedste rette linje. (Dog tegnes grafen for f ikke).

$$f := x \rightarrow LinReg(år, enheder, x) : f(x)$$

$$92.2727272727273\, x + 286.090909090909$$

Fordelen ved denne er, hvis man skal bestemme en funktionsværdi til et bestemt tidspunkt. f.eks. $f(11) = 1301.09090909091$

Regression(Lineær)

Løsning fortsat: **Eksempel 1**

Konstanterne a og b fra funktionen f kan skrives ud vha. Maple ved at bruge følgende kommando.

$$reg\_koeff = \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 286.090909090909 \\ 92.2727272727273 \end{bmatrix}$$

Regression(Lineær)

Løsning fortsat: **Eksempel 1**

Hvad fortæller konstanterne a og b ?

a = 92.272 fortæller, at produktionen af enheder vokser med cirka 92 enheder om året efter 2005 (iht modellen).

b = 286.09 fortæller, at produktionen af enheder i 2005 (iht. modellen var cirka 286 enheder.

Regression(Lineær)

Bemærk at kvaliteten af en lineær model ofte kan variere. Hvor god en model er afgøres af dens forklaringsgrad, benævnt R^2.

Kvaliteten af lineær sammenhæng

Forklaringsgraden R^2 et tal, som ligger mellem 0 og 1. Den afgøre kvaliteten af den lineær sammenhæng.

R-værdi	Kvalitet af den lineær sammenhæng
1	Perfekt lineær sammehæng.
0	Ingen lineær sammenhæng
0.9	Stærk lineær sammenhæng.
0.5	Moderat lineær sammenhæng.
0.2	Svag lineær sammenhæng.

Regression(Lineær)

I vores eksempel fra tidligere, så vi en forklaringsgrad $R^2 = 0.87$.

Det betyder altså at sammenhængen mellem antal producerede enheder og antal år efter 2005 er en næsten stærk lineær sammenhæng.

Regression(Lineær)

Teorien om hvordan bedste rette linje forekommer stammer fra den så-kaldte ”mindste kvadrats metode”.

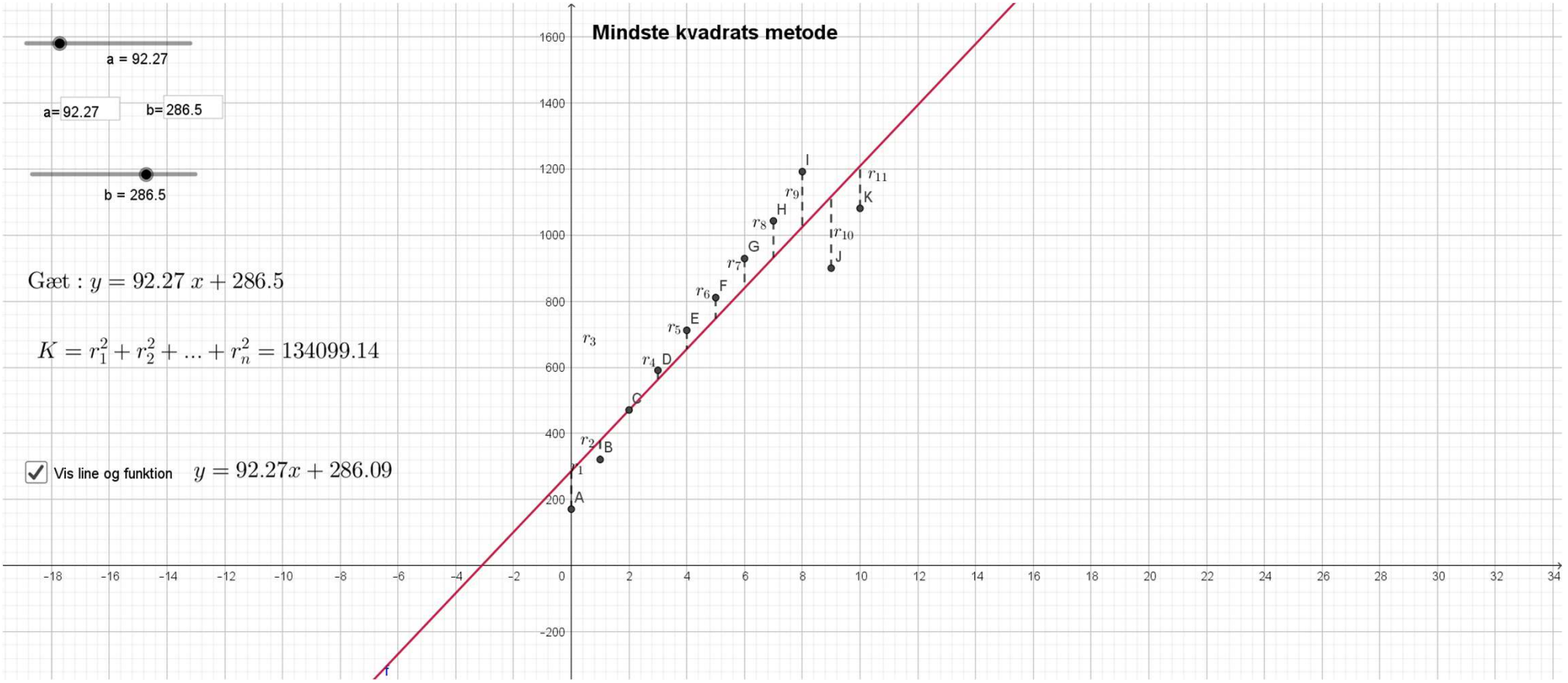
Regression(Lineær)

Kort om mindste kvadrats metode.

Man plotter sine punkter i et koordinatsystem, og bestemmer derefter den rette linje hvor der er kortest afstand mellem linje og hvert punkt fra målinger. Desuden summen af kvadratet på hvert af afstandene K skal være mindst mulig.

Formel : $K = r_1^2 + r_2^2 + \cdots .. + r_n^2$

Regression(Lineær)



Regression(Lineær)

Residualer, hvad er det?

Tallene $r_1, r_2, \ldots r_n$ er forskellen mellem tabelværdi og modelværdi. Disse kaldes residualer.

Et residual bestemmes vha følgende formel: $r_n = y_n - f(x_n)$

Regression(Lineær)

Lidt mere om residualer.

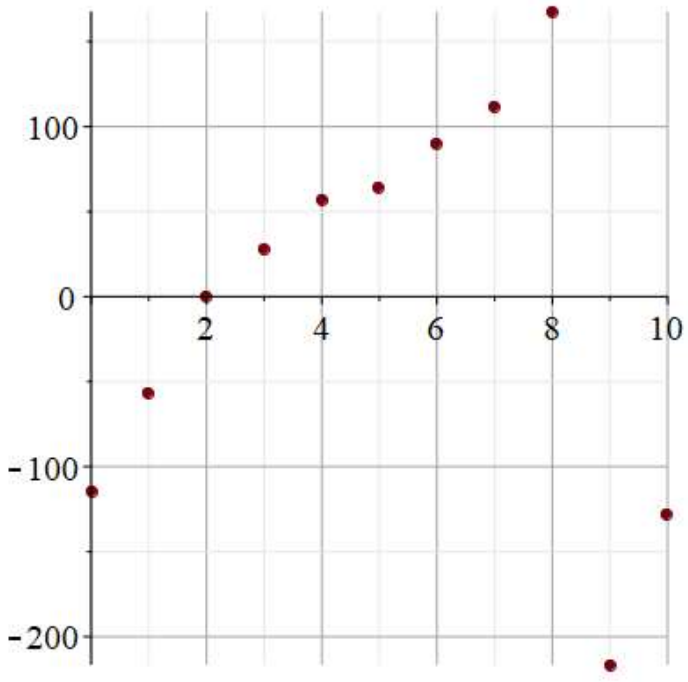
Et andet værktøj til at bestemme kvaliteten af en model er et plot over residualerne, et såkaldt residualplot.

Et residualplot fremkommer ved hjælp af kommandoen plotResidualer i Maple.

$$plotResidualer(år, enheder, LinReg)$$

Regression(Lineær)

Nedenfor ses en repræsentation af et residualplot i Maple.



Regression(Lineær)

Kort om residualplot.

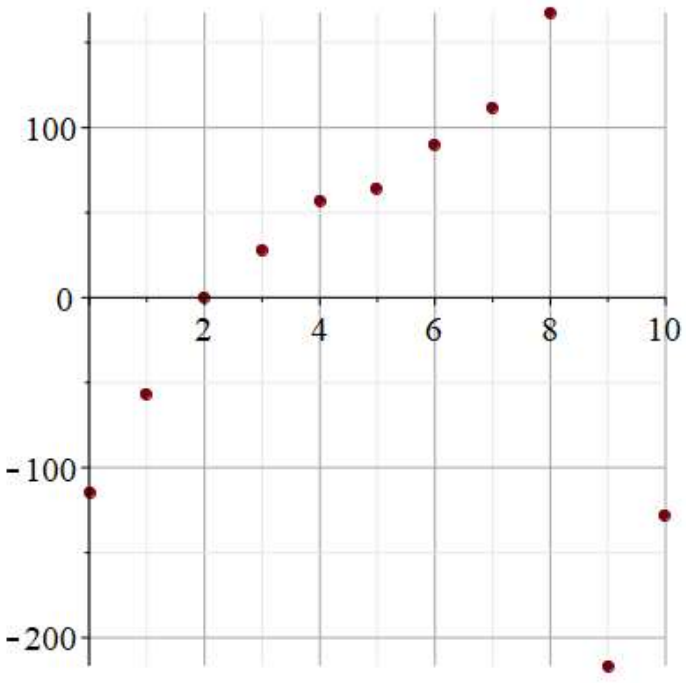
Residualplottet af et plot af residualerne. Dvs. et plot af forskellen mellem model og virkelighed.

Hvis der et fast mønster af residualerne, så siges det at modellen ikke er en god lineær sammenhæng. Men hvis der er kaos i punkternes placering, så er det en god lineær sammenhæng.

Regression(Lineær)

Hvis man ser på punkternes placering i residualplottet fra tidligere, så ses en delvis sammenhæng (dog med enkelte afvigelser). Så derfor er der muligvis en model som passer bedre på punkterne end den lineær.

En anden måde at analyser residualerne på er deres placering om x-aksen.



Regression(lineær)

Residualernes variation om *x*-aksen beskrives ved hjælp af den såkaldte *residualspredning*, der defineres ved

$$s = \sqrt{\frac{r_1^2 + r_2^2 + \cdots + r_n^2}{n-2}}$$

Tallet *s* kan fortolkes som datapunkternes gennemsnitlige afstand til regressionslinjen. Hvis der for en given *x*-værdi findes flere datapunkter lodret over hinanden, viser *s* hvor meget *y*-værdierne forventes at variere.

Regression(lineær)

Generelt om residualspredning.

- En god lineær model af et datasæt kendetegnes ved, at følgende to betingelser er opfyldt:
- *s* er lille i forhold til *y*-værdierne
- de fleste residualer har en numerisk værdi, der omtrent er lig med *s*.

Regression lineær

Eksempel fortsat

Vi bestemmer først residualerne for tabelværdierne vha. kommandoen residualer i Maple.

Regresion(lineær)

Kort om residualer i Maple.

Residualerne for eksemplet bestemt vha. Maple.

Bemærk, at de residualerne opstilles automatisk af Maple i en tabel som den til højre.

$$residualer(år, enheder, LinReg)$$

$$\begin{bmatrix} 0 & -115.090909090909 \\ 1 & -57.3636363636361 \\ 2 & 0.363636363636601 \\ 3 & 28.0909090909092 \\ 4 & 56.8181818181820 \\ 5 & 63.5454545454547 \\ 6 & 89.2727272727275 \\ 7 & 111.000000000000 \\ 8 & 167.727272727273 \\ 9 & -216.545454545454 \\ 10 & -127.818181818182 \end{bmatrix}$$

Regression(Lineær)

Eksempel fortsat

Vi bestemmer dernæst residualspredning vha. Maple og kommandoen Residualspredning.

$$residualspredning(år, enheder, LinReg)$$

123.037729275169

Regression(Lineær)

Eksempel fortsat

s viser at, den gns. variation mellen antal producerede enheder i begyndelsen af perioden er forholdvis stor i begyndelsen af perioden. Men i mindskes i slutningen af perioden.

s er forholdvis lille i forhold til y-værdierne, men y-værdierne ligger langt fra s. Derfor er modellen ikke en god lineær model.

Opgaver(Residualspredning)

Opgave 1019-1024, side 153-154, Mat2-bogen.

Opgaver(Regression)

Øvelse 41, side 33 – Mat1-bogen.

Opgave 236, side 42 – Mat1-bogen.

Opgave 237, side 42 – Mat1-bogen.

Opgave 1023, side 154-Mat2-bogen (spg a-c).

Regression(procent-afvigelse)

Procentvis afvigelse er når man ønsker at bestemme afvigelsen mellem en tabelværdi og modelværdi.

$$\text{Procent afvigelse} = \frac{modelværdi - tabelværdi}{tabelværdi} \cdot 100\ \%$$

Regression

Eksempel(procentvis afvigelse)

En tabelværdi for en sammenhæng er aflæst til 117 og modelværdien 121. Bestem den procentvise afvigelse:

$$\frac{121 - 117}{117} = 3.418803419$$

Dvs. der en afvigelse på cirka 3% mellem model og tabelværdi.

Regression(Eksponentiel)

Antag du at har mængde af en serie punkterne (x_1, y_1), … … . (x_n, y_n) som du ønsker at teste om disse kan beskrives ud fra modellen

$$y = b \cdot a^x$$

Således man finder bedste eksponentiel funktion. Det kaldes også eksponentiel regression .

Regression(Eksponentiel).

Eksempel.

Tabellen viser antallet af robotter, der blev benyttet i dansk industri, i årene 2001-2008.

År	2001	2002	2003	2004	2005	2006	2007	2008
Antal robotter	2093	2342	2630	2926	3258	3626	4115	4622

Det oplyses, at antallet af industrirobotter med god tilnærmelse er vokset eksponentielt i denne periode.

a) Benyt alle tabellens oplysninger til at opstille en model for antallet af industrirobotter som funktion af antal år efter 2001.

Regression(Eksponentiel).

Løsning: Eksempel

Vi indsætter år og antal i hver deres liste her i Maple (Ligesom ved lineær regression).

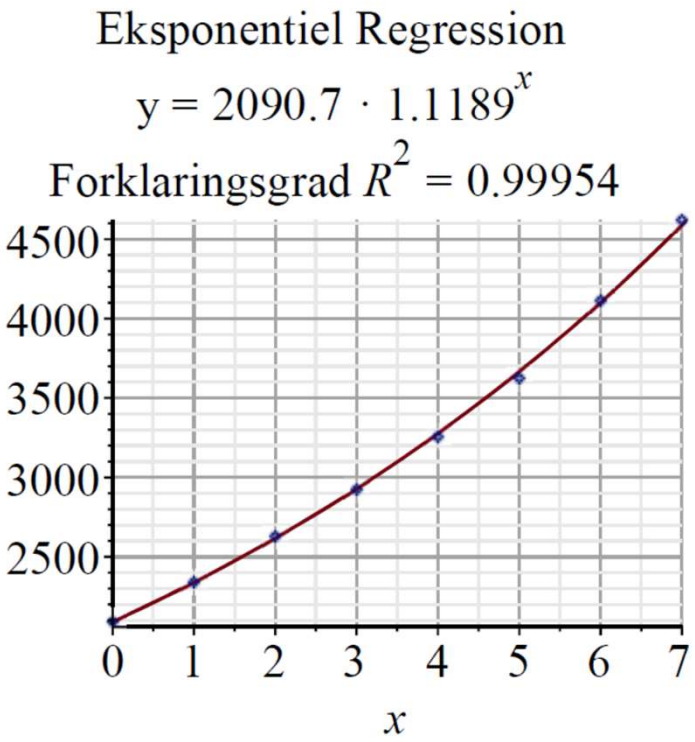
$År := [0, 1, 2, 3, 4, 5, 6, 7]:$

$antal := [2093, 2342, 2630, 2926, 3258, 3626, 4115, 4622]:$

Herefter anvendes kommandoen *ExpReg* fra Gym-pakken i Maple på listerne.

Regression(Eksponentiel).

ExpReg(*År*, *antal*)



Herved opnås den eksponentielle funktion, som viser sammenhængen mellem år og antal i eksemplet.

Regression(Eksponentiel).

Løsning fortsat: Eksempel

Den ses desuden, at forklaringsgraden $R^2 = 0.99954$ dvs. meget tæt på 1. Derfor kan man sige er her er tale om næsten perfekt eksponentiel sammehæng.

Den eksponentielle sammehæng kan bestemmes uden graf ved hjælp af følgende Maple kommando.

$$f := x \rightarrow ExpReg(\mathit{År}, \mathit{antal}, x) : f(x) = 2090.71216331838\ 1.11886827573393^{x}$$

Sidste er især praktisk, hvis man skal bruges funktionen $f(x)$ f.eks. til at bestemme en bestemt y-værd hvis man kender x-værdien.

Eller en bestemt x-værdi, hvis man kender y-værdien.

Regression(Eksponentiel)

Forklaringsgrad og residualer.

Bemærk, at begrebet forklaringsgrad R^2 også bruges i forbindelse med eksponentiel sammenhæng. Dvs. desto tættere forklaringsgraden er på 1, desto bedre er sammenhængen.

Ligeledes gælder det som ved lineær sammenhæng, at hvis der er kaos i residualplottet vise hvor god den eksponentielle sammenhæng er.

Opgaver

Opgave 730, side 155, mat1-bogen

Opgave 731, side 155, mat1-bogen

Opgave 732, side 156, mat1-bogen

(Plus supplerende opgaver.)

Regression(Eksponentiel)

Husk at man eksponentiel regression kan bestemme fordoblings eller halveringstid for sammenhængen.

Formel for fordoblingstid: $T_2 = \frac{\log_{10}(2)}{\log_{10}(a)}$

Formel for halveringsstid: $T_{0.5} = \frac{\log_{10}(2)}{\log_{10}(0.5)}$

Regression(Potens).

Hvad er potensregression?

Ved potensregression menes, at man bestemmer om en gruppe af punkter $(x_1, y_1), (x_2, y_2), \ldots\ldots (y_n, y_n)$ kan udtrykkes ud fra en potensfunktion på formen $y = b \cdot x^a$

Det kaldes bedste potensfunktion.

Regression(Potens).

Eksempel(Potensregression) Side 176, Mat1-bogen.

Vi indsætter tabelværdierne for længde og år i to Maple liste.

$$with(Gym):$$

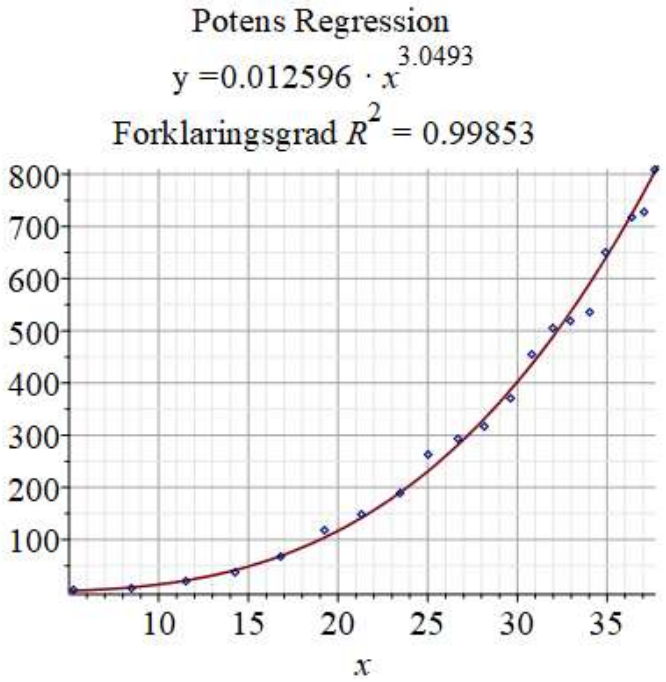
$$længde := [5.2, 8.5, 11.5, 14.3, 16.8, 19.2, 21.3, 23.5, 25, 26.7, 28.2, 29.6, 30.8, 32, 33, 34, 34.9, 36.4, 37.1, 37.7]:$$

$$vægt := [2, 8, 21, 38, 69, 117, 148, 190, 264, 293, 318, 371, 455, 504, 518, 537, 651, 719, 726, 810]:$$

Regression(Potens)

Herefter anvendes kommandoen PowReg fra Maple til at bestemme den hvis mulige bedste potensfunktion for sammenhængen.

$PowReg(længde, vægt)$



Regression(Potens).

Eksempel fortsat.

Det ses, at forklaringsgraden er $R^2 = 0.99853$. Det vil sige at sammenhæng er næsten perfekt (jf. tabellen om forklaringsgrad slide 10).

Regression(Potens).

Husk, at her kan du ligesom ved eksponentiel og lineær regression bestemme funktionen for udviklingen, uden der tegnes en graf.

Det ses desuden at residualplottet er kaos, så der er tal om en næsten perfekt potenssammenhæng.

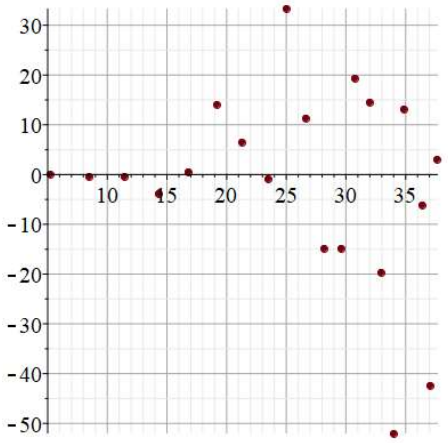
$f := x \rightarrow PowReg(længde, vægt, x) : f(x)$

$0.0125958154925056\, x^{3.04927808860249}$

Det ses at forklaringsgraden er tæt på 1, så antages det er en næsten perfekt potenssammehæng.

Vi efterprøver sammenhængen ved at plotte Residualerne og ses, at punkterne ligger sig tilfældigt om x-aksen. Derfor må det er antages at vores første antagelse er korrekt.

$plotResidualer(længde, vægt, PowReg)$



Opgaver

Øvelse 23-24-25, side. 177, Mat1-bogen.

Opgaver 935, side. 188, Mat1-bogen.

Opgave 936, side. 188, Mat1-bogen.