16.07.2015

Regression:

Lineare Regression

Nachteile der Interpolation

(*) Mit wachsender Anzahl der Stützstellen wächst auch die Ordnung des Interpolationspolynoms.

(**) Durch Interpolation - bedingt

$P(x_k) = y_k \quad \forall k = 0, 1, \ldots, n$

werde mehrere Werte pro Stützstelle (z.B. wiederholte Messungen!) nicht möglich.

$\Downarrow$

Regression

$\Downarrow$

Lineare Regression

(LGS)

(x_i, y_i) seien die gegebenen Datenpaare.

Ab jetzt sind mehrere Werte pro Stützstelle erlaubt.

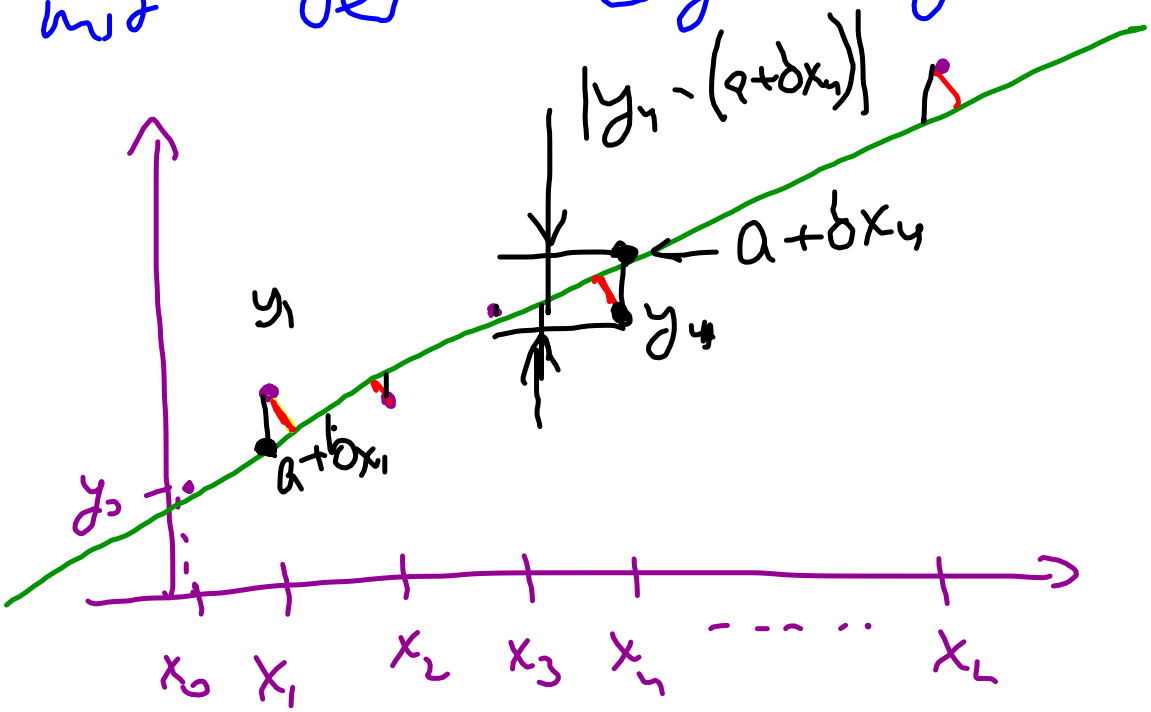
Z.B.: x – Brüder
y – Schwester

$(x_i; y_i)$ seien der Körpergrößenwerte.

Wir suchen eine Trendgerade (Regressionsgerade)

$y = a + bx$

mit der Eigenschaft:



$$\sum_{i=0}^{n} \left(y_i - (a + x b_i) \right)^2 \longrightarrow \min$$

$$F(a, b) \longrightarrow \min$$

Quadratische Abstände sollen minimiert werde!

(1) $(|\alpha|)^2 = \alpha^2 \quad \forall \alpha \in \mathbb{R}$

(2) $|\beta| < 1 \Rightarrow \beta^2 < \beta$

$|\gamma| > 1 \Rightarrow \gamma^2 > \gamma$

(3) Um Min / Max zu finden, nutzt man aus, dass, wenn die Funktion

$f(\xi)$ an der Stelle $\xi = \Theta$ min oder max hat, dann gilt:

$$f'(\Theta) = 0$$

(4) $$|x|' = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ \text{existiert nicht} & \text{für } x = 0 \end{cases}$$

$\Downarrow$

$|x - \alpha|$ ist nicht überall differenzierbar; und somit auch die Summe der Beträge ist nicht überall diff'bar.

$(x-a)^2$ oder schon!

$\Downarrow$

$\sum (x_i - a_k)^2$ ist überall

diff'bar! bekannt bekannt

$$(S) \quad \sum_{i=0}^{n} \left(y_k - (a + b x_k)\right)^2 =$$

Variablen

$$= F(a,b) \longrightarrow \min$$

$F(a,b)$ ist bzgl. der Variabl. a eine nach oben geöffnete Parabel $\Rightarrow$

gesucht ist Extremum,
Minimum!

Analog für b

⇓

$F(a,b)$ hat gesucht
ein Minimum!

F hängt von 2 Var. ab
⇒ Partielle Ableitungen!

$$\frac{\partial F(a,b)}{\partial a} = \frac{\partial}{\partial a}\left(\sum (y_i - a - bx_i)^2\right)$$

Summenregel

$$= \sum_{i=0}^{n} \frac{\partial}{\partial a}\left(y_i - bx_i - a\right)^2 =$$

Kettenregel

$$= \sum_{i=0}^{n} 2\left(y_i - bx_i - a\right)(-1) =$$

$$= -2 \sum_{i=0}^{n} \left(y_i - bx_i - a\right);$$

$$\frac{\partial F}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=0}^{n} (y_i - a - bx_i)^2\right) =$$

$$= \sum_{i=0}^{n} \frac{\partial}{\partial b}(y_i - a - bx_i)^2 =$$

$$= \sum_{i=0}^{n} 2(y_i - a - bx_i)(-x_i).$$

An eler (wege der Parabel – an der)

Min - Stelle:

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0.$$

$$\begin{cases} \sum_{i=0}^{n} (y_i - a - bx_i) \overset{!}{=} 0 \\ \sum_{i=0}^{n} (y_i - a - bx_i) x_i \overset{!}{=} 0 \end{cases}$$

ein lineares GS und

2 Gl und 2 Var: a und b

$$\begin{cases} (n+1)a + b \sum_{i=0}^{n} x_i = \sum_{i=0}^{n} y_i \\ a \sum_{i=0}^{n} x_i + b \sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} x_i y_i \end{cases}$$

Normalgleichung

Umformulierung durch Umnumerierung! Wir fangen bei 1 an!

Hat man n Datenpaare (x_k, y_k), so erhält man die Koeffizienten a und b der Ausgleichsgeraden $y = a + bx$ die Normalgleichungen:

$$\begin{cases} n a + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i\right)^2 b = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Matrix form:

$$\underbrace{\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\bar{c}} = \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}}_{\bar{d}}$$

$$A\bar{c} = \bar{d}$$

↗ gesucht

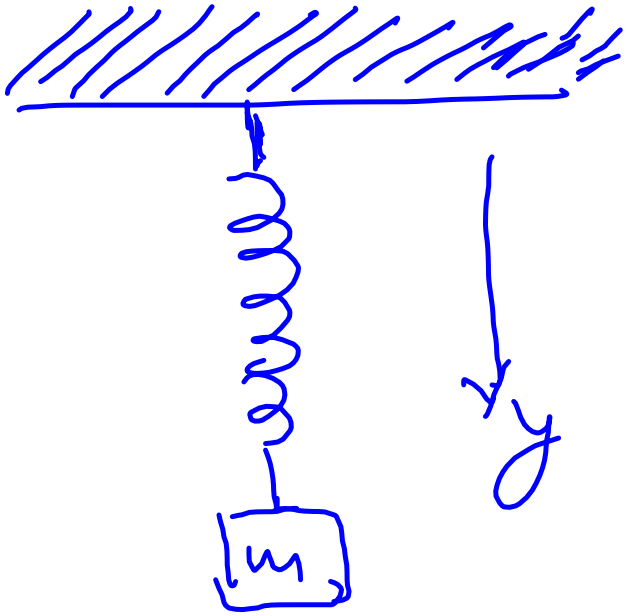
Cramer'sche Regel:

$$\det A = n \cdot \sum (x_i^2) - \left(\sum x_i\right)^2;$$

$$a = \frac{(\Sigma y_i)(\Sigma (x_i)^2) - (\Sigma x_i)(\Sigma x_i y_i)}{n \cdot \Sigma (x_i)^2 - (\Sigma x_i)^2};$$

$$b = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{n \cdot \Sigma (x_i)^2 - (\Sigma x_i)^2}$$

Bsp.:



Hook'sches Gesetz:

"Auslenkung proportional zur Masse"

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

m_k	5	10	15	20	25	30
y_k	34	52	66	78	97	110

$n = 6$

x_k	y_k	x_k^2	$x_k \cdot y_k$
5	34	25	170
10	52	100	520
15	66	225	990
20	78	400	1580
25	87	625	2425
30	110	900	3300
$\sum x_k = 105$	$\sum y_k = 438$	$\sum x_k^2 = 2275$	$\sum x_k \cdot y_k = 8985$

$$a = \frac{(\Sigma y_i)(\Sigma (x_i)^2) - (\Sigma x_i)(\Sigma x_i y_i)}{n \cdot \Sigma (x_i)^2 - (\Sigma x_i)^2};$$

$$b = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{n \cdot \Sigma (x_i)^2 - (\Sigma x_i)^2}$$

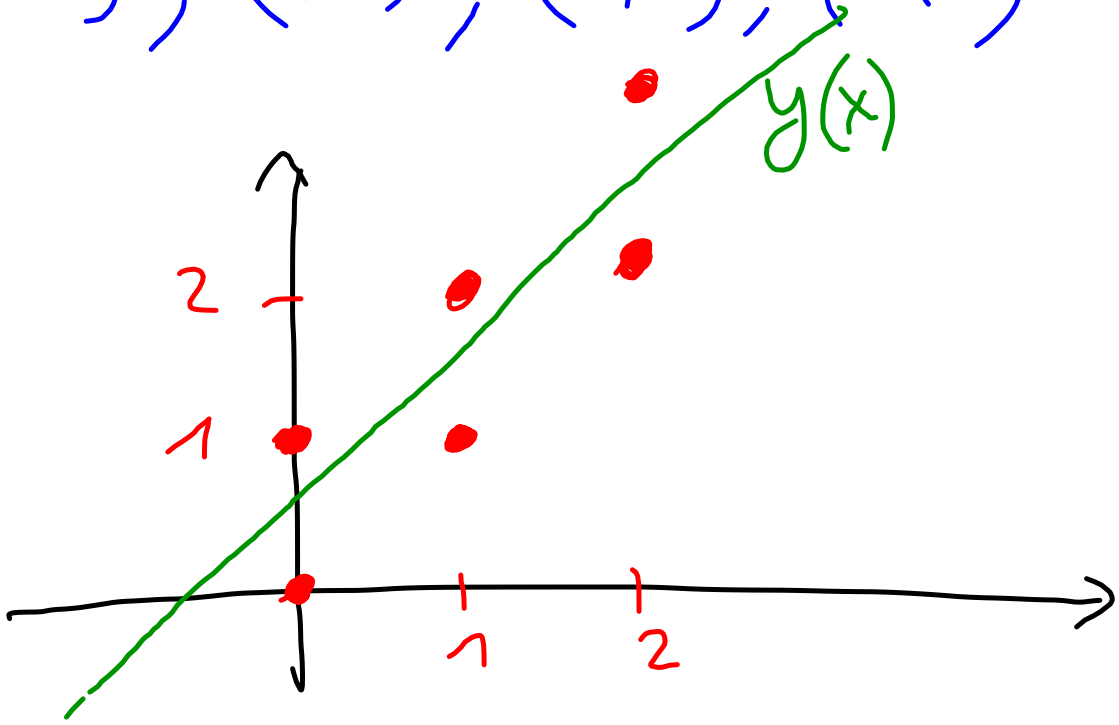
$$a = \frac{438 \cdot 2275 - 105 \cdot 8585}{6 \cdot 2275 - (105)^2} \approx 20,2$$

$$b = \frac{6 \cdot 8585 - 105 \cdot 438}{6 \cdot 2275 - (105)^2} = 3,017 \approx 3$$

$y = 3 \cdot m + 29{,}2$

"Optimale" Gerade

Bsp 2: (0, 0); (0, 1); (1, 1); (1, 2); (2, 2); (2, 3)



Hypothese:

$$y(x) = x + \frac{1}{2}$$

Nun berechnen wir:

$$y = a + bx$$

$$y_k \approx a + bx_k$$

nicht exakt, aber ungefähr.

Also 6 nicht exakte (zunächst) Gleichungen:

$$0 \approx a + b \cdot 0$$

$$1 \approx a + b \cdot 0$$

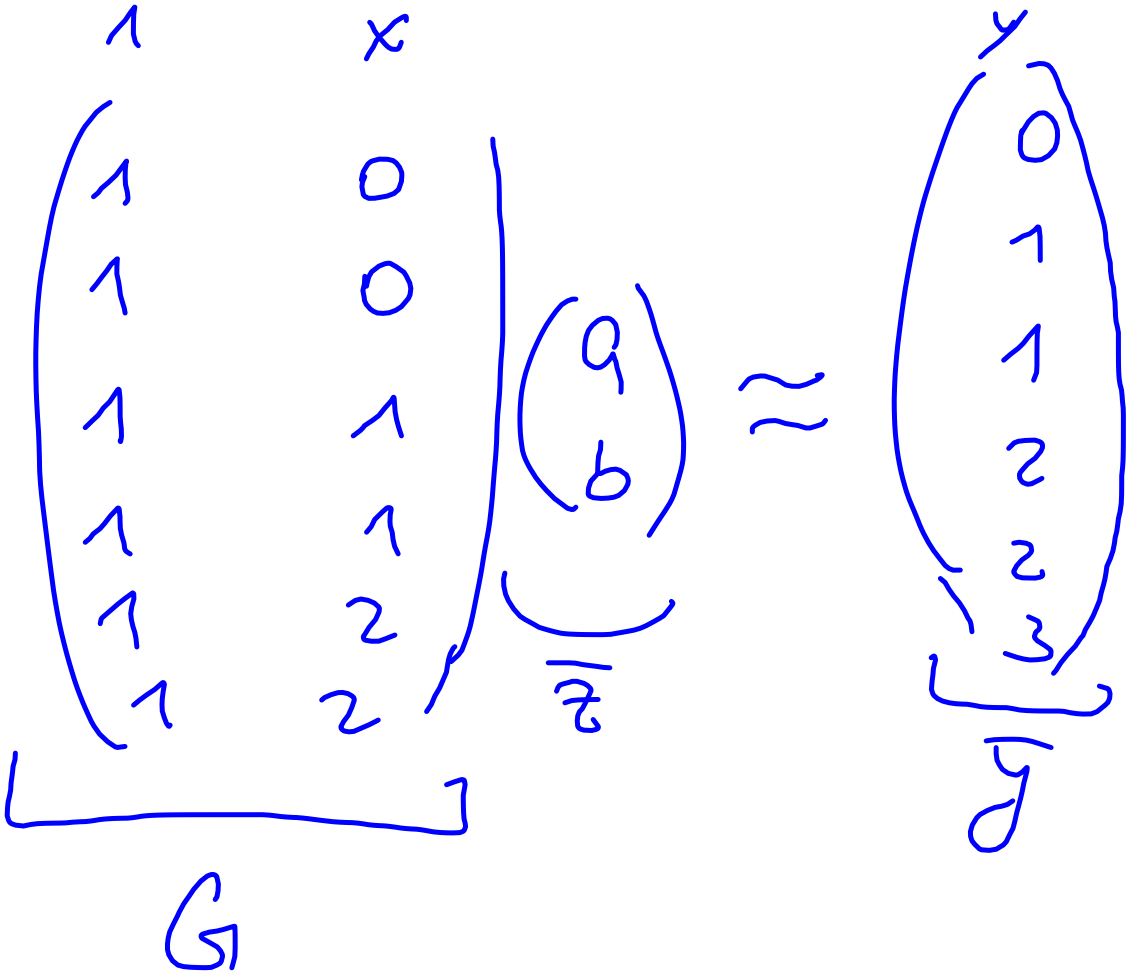
$$1 \approx a + b \cdot 1$$

$$2 \approx a + b \cdot 1$$

$$2 \approx a + b \cdot 2$$

$$3 \approx a + b \cdot 2$$

LGS



$$G\bar{z} \simeq \bar{y}$$

mult

$$G = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}; \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\bar{z} = \begin{pmatrix} a \\ b \end{pmatrix};$$

E.: nichtquadratisches LGS.

$$G \in \mathbb{R}^{n \times 2}$$

$$G^T \in \mathbb{R}^{2 \times n}$$

$$G^T G \in \mathbb{R}^{2 \times 2}$$

$(2 \times n) \cdot (n \times 2) = (2 \times 2)$

$=$

↓ existiert

$\bar{y} \in \mathbb{R}^n \Rightarrow (G^T \bar{y}) \in \mathbb{R}^2$

$$G\,\vec{z} = \vec{y} \quad | \; G^T \cdot \ldots$$

$$\left(G^T G\right)\vec{z} = G^T y$$

Normalgleichungsgerade

Wieder!

$$G^T G = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} =$$

$$= \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} (x_i^2) \end{pmatrix}$$

$$G^T \vec{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$\Rightarrow$ dasselbe LGS wie früher!

In unserem Bsp:

$$G = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix}$$

$$G^T G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & 6 \\ 6 & 10 \end{pmatrix}$$

$$G^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix} =$$

$$= \begin{pmatrix} 9 \\ 13 \end{pmatrix}$$

$\Rightarrow$ Normalsystem:

$$\begin{pmatrix} 6 & 6 \\ 6 & 10 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 9 \\ 13 \end{pmatrix}$$

$$\left(\begin{array}{cc|c} 6 & 6 & 9 \\ 6 & 10 & 13 \end{array}\right)$$

GA $\rightsquigarrow$

$z_2 \leftarrow z_2 - z_1$

$$\left(\begin{array}{cc|c} 6 & 6 & 9 \\ 0 & 4 & 4 \end{array}\right)$$

$\Downarrow$

$b = 1 \Rightarrow a = \frac{1}{2}$

$y = b \cdot x + a = x + \frac{1}{2}$

Unsere Hypothese stimmt!

Regressionsgerade oder Regressionsparabel?

x_k	0	0	1	1	2	2
y_k	0	1	1	2	4	5

(1) Regr. - Gerade:

$$G = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 4 \\ 5 \end{pmatrix}$$

1 x

$$G^T G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} =$$

(wie in Bsp 2!)

$$= \begin{pmatrix} 6 & 6 \\ 6 & 10 \end{pmatrix} .$$

$$G^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 4 \\ 5 \end{pmatrix} =$$

$$= \begin{pmatrix} 13 \\ 21 \end{pmatrix} .$$

$$\left(\begin{array}{cc|c} 6 & 6 & 13 \\ 6 & 10 & 21 \end{array}\right) \qquad z_2 \leftarrow z_2 - z_1$$

$$\left(\begin{array}{cc|c} 6 & 6 & 13 \\ 0 & 4 & 8 \end{array}\right)$$

$$\Downarrow$$

$$b = 2$$

$$a = (13 - 6 \cdot 2)/6 = \frac{1}{6}$$

$$\Rightarrow \quad y_g = a + x b = \frac{1}{6} + 2x \,;$$

(2) Regressionsfunktion:

$$y_i = a + bx + cx^2$$

3 Unbekannte

$$\begin{cases} y_1 \approx a + bx_1 + cx_1^2 \\ y_2 \approx a + bx_2 + cx_2^2 \\ \dots \\ y_n \approx a + bx_n + cx_n^2 \end{cases}$$

n Gl.; 3 Unbekannte

$$G = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}; \quad \vec{z} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$\bar{y} = (y_1, y_2, y_3 \dots, y_n)^T$$

Normalgleichung:

$$(G^T G)\,\bar{z} = G^T \bar{y}$$

$$G^T G = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_n^2 \end{pmatrix} \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} (x_i)^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} (x_i)^2 & \sum_{i=1}^{n} (x_i)^3 \\ \sum_{i=1}^{n} (x_i)^2 & \sum_{i=1}^{n} (x_i)^3 & \sum_{i=1}^{n} (x_i)^4 \end{pmatrix}$$

$$G^T \vec{y} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x_i y_i) \\ \sum_{i=1}^{n} (x_i^2 y_i) \end{pmatrix}$$

3 Größen sind zu bestimmen (zur Geradenbestimmung) zu berechnen!

3x3 Lineares GS

Zurück zum Bsp.:

							Σ
x_k	0	0	1	1	2	2	6
y_k	0	1	1	2	4	5	13
x_k^2	0	0	1	1	4	4	10
x_k^3	0	0	1	1	8	8	18
x_k^4	0	0	1	1	16	16	34
$x_k y_k$	0	0	1	2	8	10	21
$x_k^2 y_k$	0	0	1	2	16	20	39

Nor-LGS:

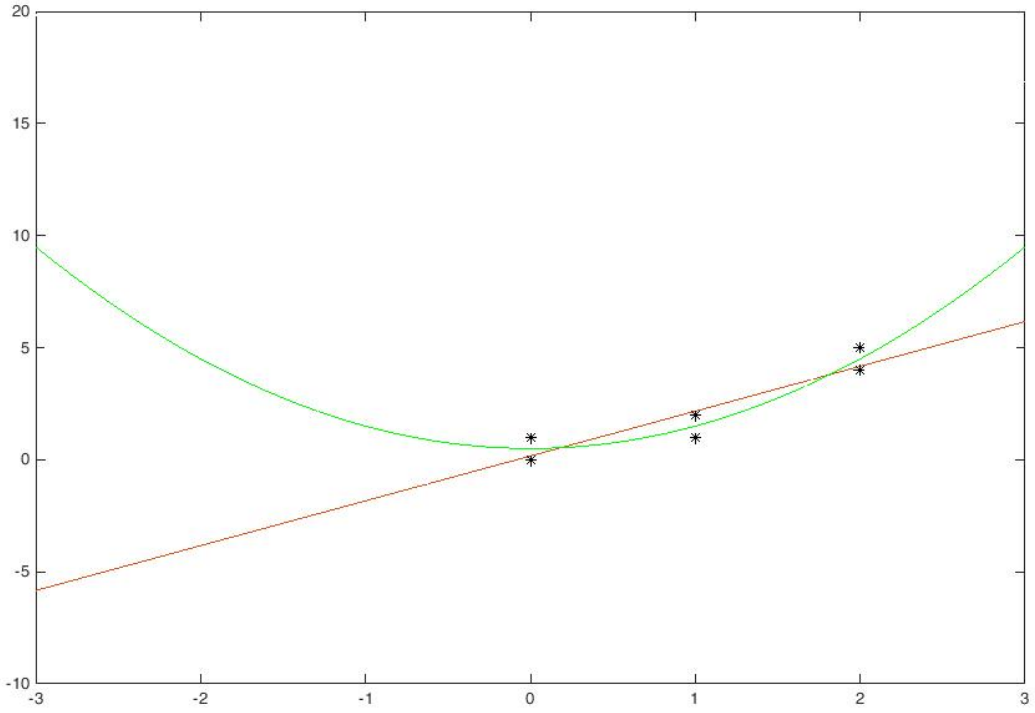
$$\begin{pmatrix} 6 & 6 & 13 \\ 6 & 13 & 18 \\ 13 & 18 & 34 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 13 \\ 21 \\ 39 \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -2,3 \\ 0,5 \\ 2,3 \end{pmatrix}$$

$y_p = -2,3 + 0,5x + 2,3x^2$;

Rechenfehler

$y_p = \frac{1}{2} + x^2$;



Was ist hier besser, Parabel oder Gerade?

Maß:

$\sum (y_k - g(x_k))^2$ wird

mit

$$\sum \left(y_k - p(x_k)\right)^2$$

Vergleichen.

Die Regressionsfunktionen mit kleinerer quadratischer Abweichung ist "besser".

$$\sum_{k=1}^{6} \left(y_k - \left(\frac{1}{6} + 2 x_k\right)\right)^2 = 2{,}83$$

$$\sum_{k=1}^{6} \left(y_k - \left(\frac{1}{2} + x_k^2\right)\right)^2 = 1{,}5$$

$\Rightarrow$ Die Regressionsgerade

β_1 in diesem Fall

besser, weil Fehler

kleiner β_1!