Regression:

Lineare Regression

Nachteile der Interpolation:

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

(*) Durch die Interpolationsbedingung $P(x_k) = y_k \quad \forall k = 0, 1, \ldots, n$ werden mehrere Stützwerte pro Stützstelle (z.B. wiederholte

Messungen) ausgeschlossen.

⇓

Regression

⇓

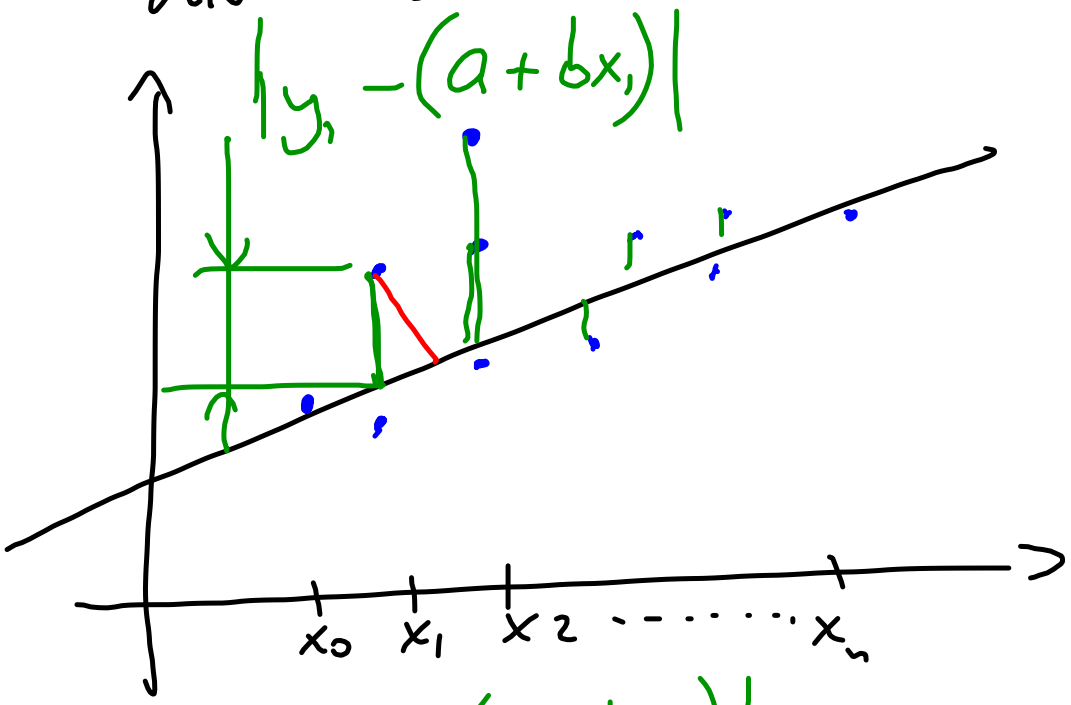
Lineare Regression (LGS)

$(x_i; y_i)$ seien die Ortskoordinaten, mehrere Schätzstelle Werte pro nun erlaubt!

Wir suchen eine Trendgerade (Regressionsgerade)

$$y = a + bx$$

mit der Eigenschaft:



$$|y_k - (a + bx_k)|$$

$$\sum_{k=0}^{n} \left(y_k - (a + bx_k) \right)^2 \longrightarrow \min$$

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

Quadratische Abstände sollen minimiert werden!

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

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

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

(3) Um min/max zu finden, nutzt man aus, dass, wenn die Funktion $f(\xi)$ an der Stelle $\xi = \delta$ max oder min hat und diff'bar ist,

dann gilt:

$$f'(\xi) = 0.$$

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

existiert nicht für $x = 0$

$\Downarrow$

$|x-a|$ ist nicht überall differenzierbar;

damit ist auch die Summe der Beträge nicht überall diff'bar!

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

$\Downarrow$

$\Rightarrow (x-a)^2$ ist überall diffbar!

(5) $\sum_{k=0}^{n} \left(y_k - (q + b x_k)\right)^2 =$

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

$F(q,b)$ ist bzgl. Var. q eine nach oben geöffnete Parabel $\Rightarrow$ genau ein Extremum $\Rightarrow$ Minimum !!

Analog für b

$\Downarrow$

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

F hängt von 2 Var. ab

$\Longrightarrow$ partielle Ableitungen!

$$\frac{\partial F(a,b)}{\partial a} = \frac{\partial}{\partial a}\left(\sum_{k=1}^{n}(y_k - a - bx_k)^2\right)$$

Summenregel $= \sum_{k=1}^{n} \frac{\partial}{\partial a}(y_k - a - bx_k)^2 =$

$\frac{df}{dx}$	$\frac{\partial F(x,y)}{\partial x}$

Kettenregel

$$= \sum_{k=0}^{n} 2\left(y_k - a - bx_k\right)(-1) =$$

$$= -2 \sum_{k=0}^{n} \left(y_k - a - bx_k\right);$$

$$\frac{\partial F}{\partial b} = \frac{\partial}{\partial b}\left(\sum \left(y_k - a - bx_k\right)^2\right) =$$

$$= \sum \frac{\partial}{\partial b}\left(y_k - a - bx_k\right)^2 =$$

$$= \sum 2\left(y_k - a - bx_k\right)\left(-x_k\right)$$

An einer (wegen der Parabel — anders m.i.-Stelle)

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

$$\Downarrow$$

$$\begin{cases} \sum_{k=0}^{n} (y_k - a - bx_k) = 0 \\ \sum_{k=0}^{n} (y_k - a - bx_k)x_k = 0 \end{cases}$$

Ein lineares GS mit 2 Gl und 2 Var: a und b

$$\begin{cases} (n+1)\,a + b \sum_{k=0}^{n} x_k = \sum_{k=0}^{n} y_k \\ a \sum_{k=0}^{n} x_k + b \sum_{k=0}^{n} x_k^2 = \sum_{k=0}^{n} y_k x_k \end{cases}$$

Normalgleichung (en).

Umformulierung durch Umnummerierung (Indexverschieben):

wir fangen bei 1 an!:

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

$$\begin{cases} n \cdot a + \left(\sum_{k=1}^{n} x_k\right) b = \sum_{k=1}^{n} y_k \\ \left(\sum_{k=1}^{n} x_k\right) a + \left(\sum_{k=1}^{n} x_k^2\right) b = \sum_{k=1}^{n} x_k y_k \end{cases}$$

Matrixform:

$$\underbrace{\begin{pmatrix} n & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\vec{c}} = \underbrace{\begin{pmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \end{pmatrix}}_{\vec{d}}$$

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

↗ gesucht

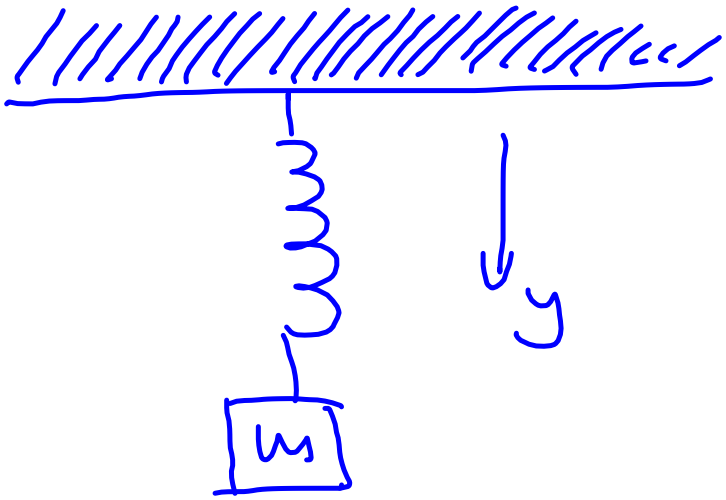
Cramer'sche Regel:

$$\det A = n \cdot \sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2;$$

$$a = \frac{\left(\sum y_k\right)\left(\sum x_k^2\right) - \left(\sum x_k\right)\left(\sum x_k y_k\right)}{n \cdot \sum_{k=1}^{n} x_k^2 - \left(\sum x_k\right)^2};$$

$$b = \frac{n \sum_{k=1}^{n} (x_k y_k) - \left(\sum x_k\right)\left(\sum y_k\right)}{n\left(\sum x_k^2\right) - \left(\sum x_k\right)^2};$$

Bsp:



Hook'sches Gesetz:

"Die Auslenkung ist proportional zur Masse":

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

m_k	5	10	15	20	25	30
y_k	34	52	66	79	97	110

$n = 6$

m_k	y_k	m_k^2	$m_k y_k$
5	34	25	170
10	52	100	520
15	66	225	990
20	79	400	1580
25	97	625	2425
30	110	900	3300
$\sum m_k =$ $= 105$	$\sum y_k =$ $= 438$	$\sum m_k^2 =$ $= 2275$	$\sum m_k y_k =$ $= 8885$

$$a = \frac{\left(\sum y_k\right)\left(\sum x_k^2\right) - \left(\sum x_k\right)\left(\sum x_k y_k\right)}{n \cdot \sum_{k=1}^{n} x_k^2 - \left(\sum x_k\right)^2};$$

$$b = \frac{n \sum_{k=1}^{n} (x_k y_k) - \left(\sum x_k\right)\left(\sum y_k\right)}{n\left(\sum x_k^2\right) - \left(\sum x_k\right)^2};$$

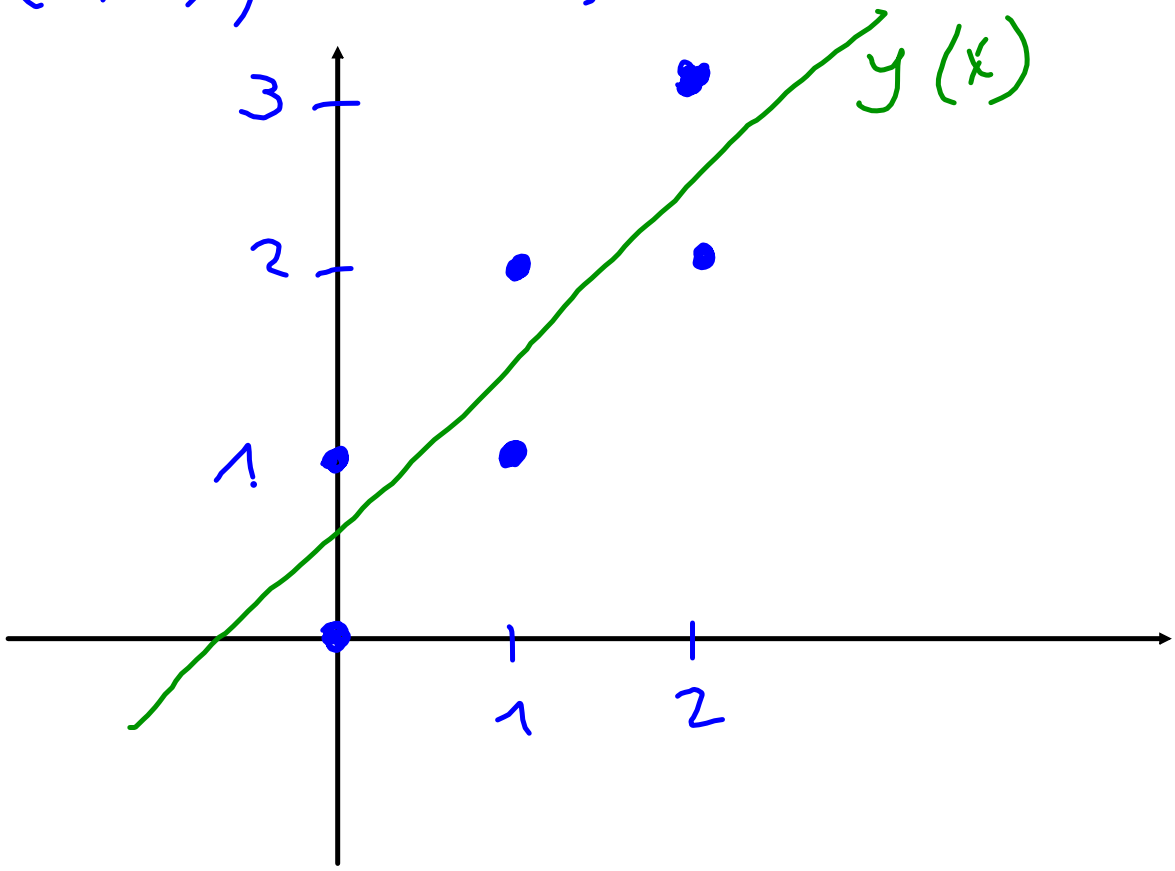
$$a = \frac{438 \cdot 2275 - 105 \cdot 8385}{6 \cdot 2275 - (105)^2} = 20{,}2$$

$$b = \frac{6 \cdot 8385 - 105 \cdot 438}{6 \cdot 2275 - (105)^2} \approx 3{,}02$$

$y = 3{,}02\,m + 20{,}2$

"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

die Regressionsgerade:

$$y = a + bx$$

$$y_k \approx a + b x_k$$

↑

nicht exakt, aber ungefähr.

D.h. wir haben 6

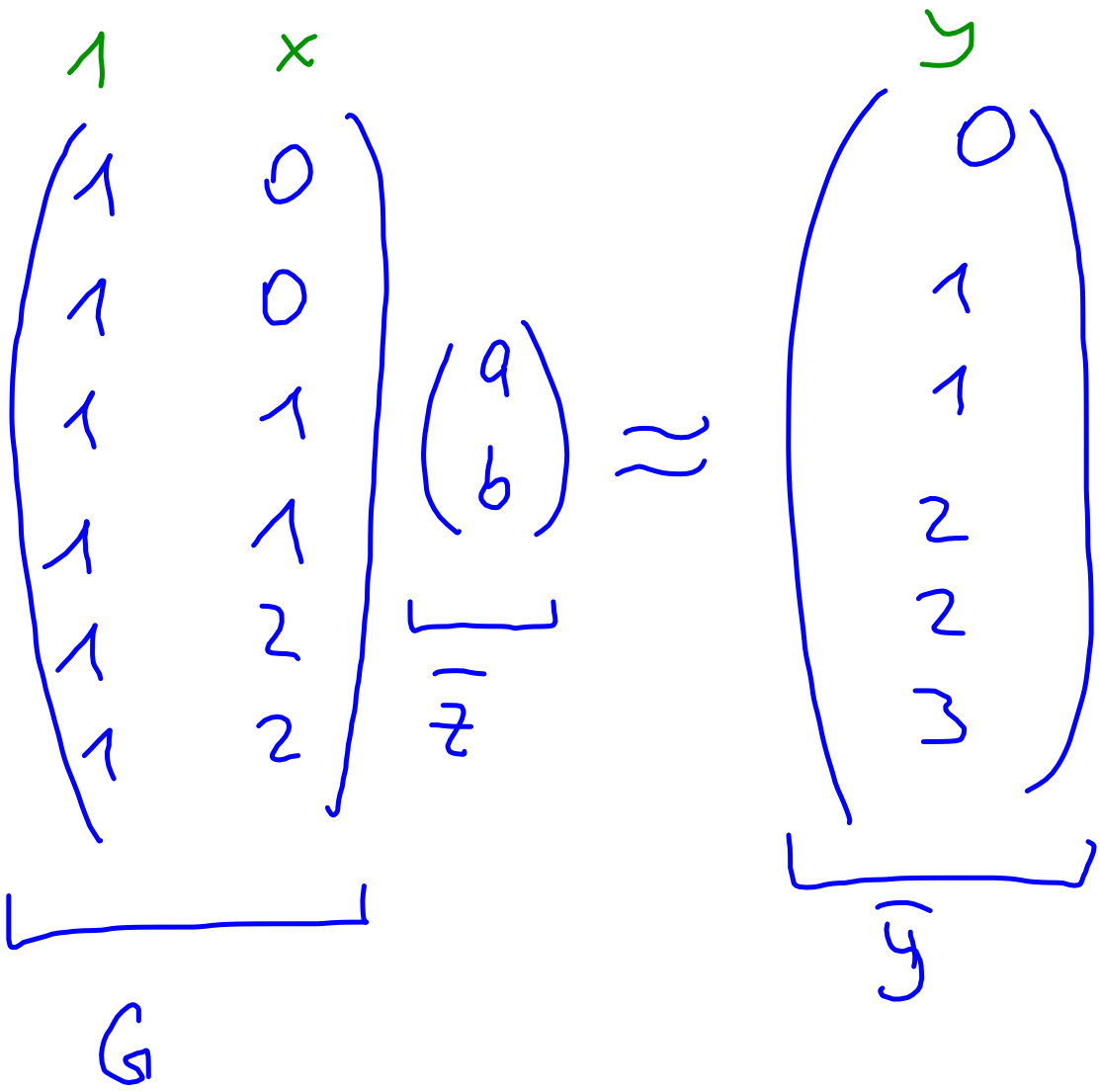
(zunächst nicht exakte)

Gleichungen:

$$0 \approx a + b \cdot 0$$
$$1 \approx a + b \cdot 0$$
$$1 \approx a + b \cdot 1$$
$$2 \approx a + b \cdot 1$$
$$2 \approx a + b \cdot 2$$
$$3 \approx a + b \cdot 2$$

Approx. LGS.

$\Longrightarrow$ Matrixform:



$$G\bar{z} \approx \tilde{y}$$

$$\therefore A. \quad G = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

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

E.g. nichtquadratisches LGS

$$G \in \mathbb{R}^{4\times 2}$$

$$G^T \in \mathbb{R}^{2\times 4}$$

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

$$(2\times 4)\cdot(4\times 2) = (2\times 2)$$

$$\bar{y} \in \mathbb{R}^4 \Rightarrow G^T\bar{y} \in \mathbb{R}^2$$

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

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

(2×2)

Wieder Normalgleichungs-System!

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

$$= \begin{pmatrix} n & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} (x_k^2) \end{pmatrix};$$

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

$$= \begin{pmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \end{pmatrix};$$

$\Rightarrow$ Dasselbe LGS wie

früher!

In unserem Bsp:

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

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

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

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

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

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

$\Rightarrow$ Normalsystem:

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

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

$$z_2 \leftarrow z_2 - z_1$$

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

$z_1 \cdot \frac{1}{3}$, $z_2 \cdot \frac{1}{4}$

$$\left(\begin{array}{cc|c} 2 & 2 & 3 \\ 0 & 1 & 1 \end{array}\right)$$

$\Downarrow$

$b = 1$

$\Downarrow$

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

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

$= \frac{1}{2} + x$

Also: 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:

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

1 x

$$y = a + bx \quad , \quad \vec{z} = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$G^T G = \begin{pmatrix} 6 & 6 \\ 6 & 10 \end{pmatrix}$$

(gleich wie im Bsp. 2)

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

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

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

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

$$\Downarrow$$

$$b = 2$$

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

$$\Rightarrow \quad y_{\text{obs}} = a + bx =$$

$$= \frac{1}{6} + 2x \; ;$$

(2) Regressionsparabel:

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

Unbekannt

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

n Gl.; 3 Unbekannte

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

$$\bar{z} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Normalgleichung:

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

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

$$= \begin{pmatrix} n & \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 & \sum_{k=1}^{n} x_k^3 \\ \sum_{k=1}^{n} x_k^2 & \sum_{k=1}^{n} x_k^3 & \sum_{k=1}^{n} x_k^4 \end{pmatrix};$$

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

$$= \begin{pmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \\ \sum_{k=1}^{n} x_k^2 y_k \end{pmatrix}$$

3 Größen sind zusätzlich (zur Gerade bestimmung) 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

$$\begin{pmatrix} n & \sum x_k & \sum x_k^2 \\ \sum x_k & \sum x_k^2 & \sum x_k^3 \\ \sum x_k^2 & \sum x_k^3 & \sum x_k^4 \end{pmatrix} = G^T G =$$

$$= \begin{pmatrix} 6 & 6 & 10 \\ 6 & 10 & 18 \\ 10 & 18 & 34 \end{pmatrix}; \quad \bar{z} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$G^T \bar{y} = \begin{pmatrix} \sum y_k \\ \sum x_k y_k \\ \sum x_k^2 y_k \end{pmatrix} = \begin{pmatrix} 13 \\ 21 \\ 39 \end{pmatrix}$$

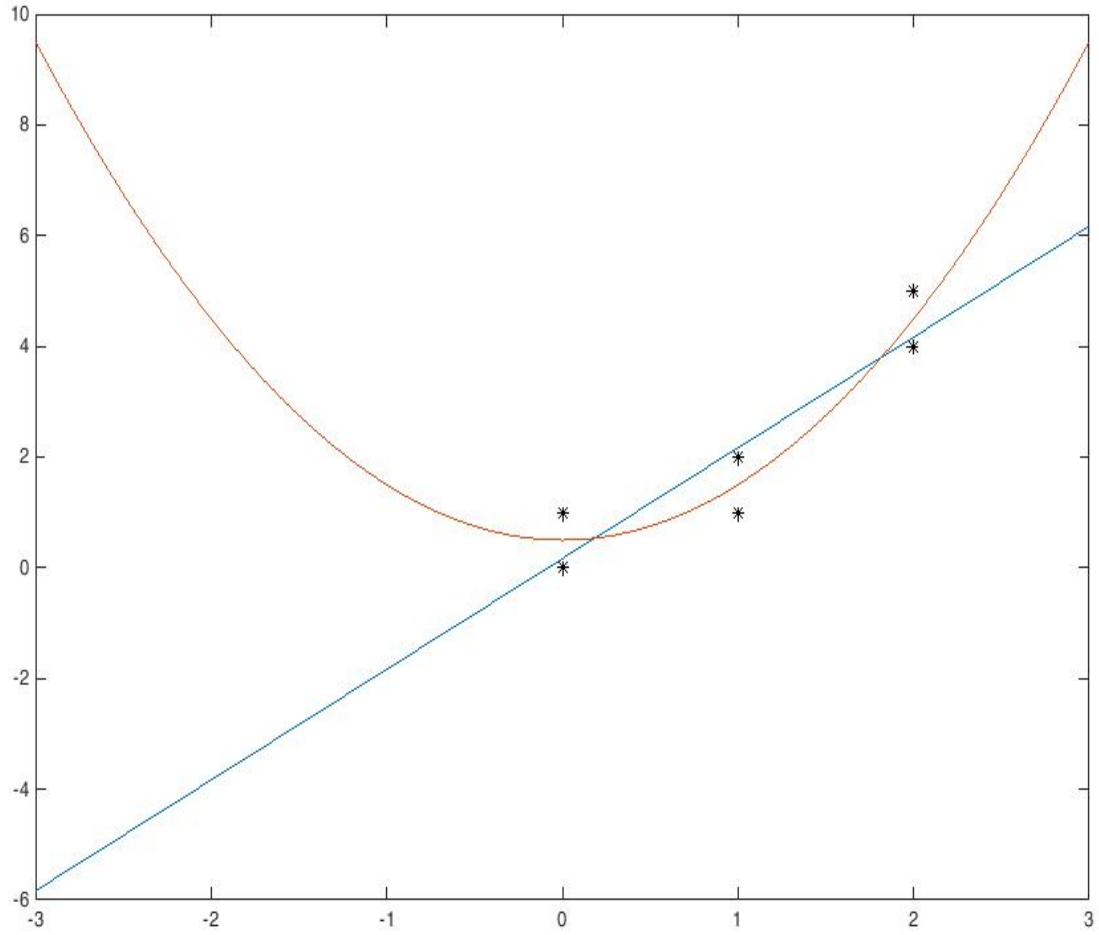
Norm. LGS:

$$\begin{pmatrix} 6 & 6 & 10 \\ 6 & 10 & 18 \\ 10 & 18 & 34 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 13 \\ 21 \\ 39 \end{pmatrix}$$

$\Rightarrow$ Lösung:

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

$$\left(a = \frac{1}{2};\ b = 0;\ c = 1\right)$$



Was ist hier besser, Parabel oder Gerade?

MaB: $\sum_{k=1}^{n} \left(y_k - f(x_k)\right)^2$

D.h.

$\sum_{k=1}^{n} \left(y_k - y_g(x_k)\right)^2$ wird

mit $\sum_{k=1}^{n} \left(y_k - y_p(x_k)\right)^2$

verglichen.

Die Regressionsfunktion

mit der kleineren

quadratischen Abweichung ist

"besser", d.h. optimaler.

Bsp:

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

$$= \left(0 - \left(\frac{1}{6} + 0 \right) \right)^2 + \left(1 - \left(\frac{1}{6} + 0 \right) \right)^2 +$$

.....

$$= \ldots\ldots = 2,8333$$

(Gerade ↗)

Parabel:

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

d.h. in unserem Bsp ist die Regressionsgerade optimaler ("besser"), weil der Fehler kleiner ist!

"Methode der kleinsten Quadrate".