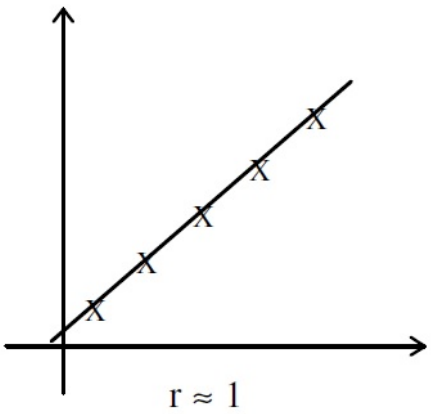
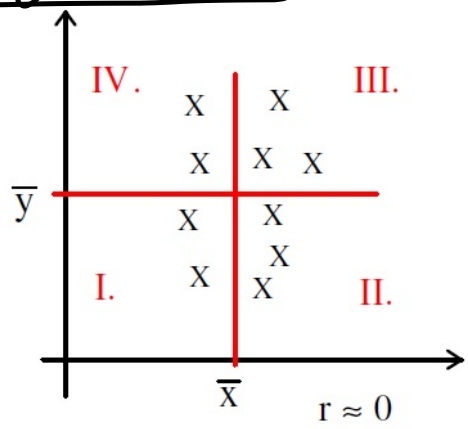
Regressionsanalyse

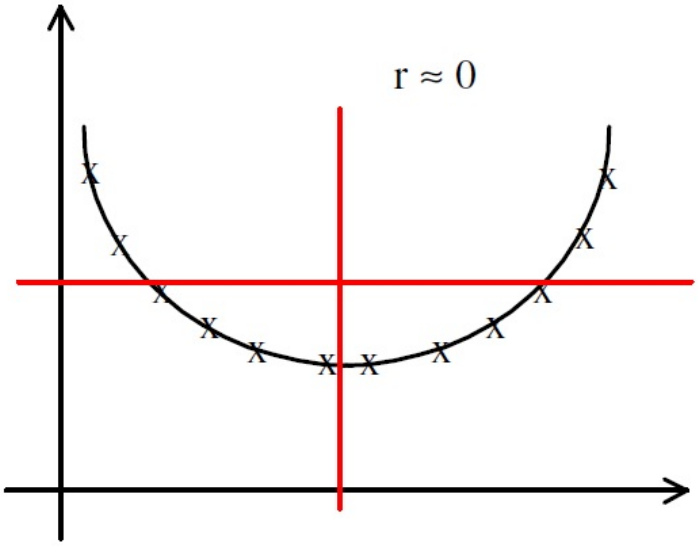




r misst nur den linearen Zusammenhang:

$y = f(x)$, und sei f keine Gerade, sei z. B.

$y = ax^2 + bx + c \qquad (x_i, y_i), \quad i = 1,\ldots,n$



⇒ Ich erkenne mit r keinen Zusammenhang, obwohl es einen gibt (weil f nicht linear ist)

Aufgabenstellung

f soll abgeschätzt werden,

(f beschreibt Zusammenhang zwischen x und y):

Modellbildung:

$$y = f(x) + \varepsilon$$

Ziel: y vorhersagen

können anhand der Werte von x

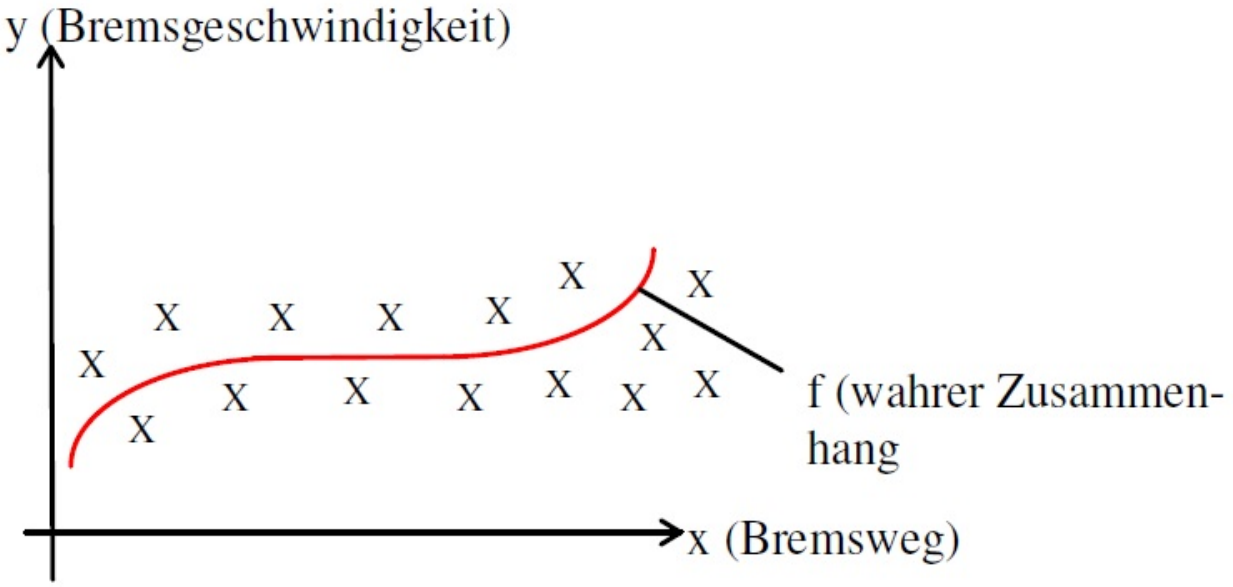
$\longrightarrow f(x)$ soll durch eine Funktion $\hat{f}(x)$ abgeschätzt werde

X	Y	ε
Bremsweg	Bremsgeschwindigkeit	Straßenverhältnisse, Zustand der Bremsen, Reaktionsverhalten
Geschwindigkeit	Benzinverbrauch	Luftdruck; Wetter; Lastung; Straßenverhältnisse;
Kohlenstoffgehalt einer Stahlsorte	Reißfestigkeit	Gleichmäßigkeit; Bearbeitung; Legierung; Temperatur;
Durchmesser eines Flaschenhalses	Berstdruck der Flasche	z.B. Temperatur; Flüssigkeit;
Körpergröße	Intelligenz	
Risthöhe beim Rind	Gewicht	

Gesucht wird eine Funktion mit mehreren Einflußgrößen:

$$y = f(x_1, x_2, \ldots, x_n) + \varepsilon$$

Methode der Schätzung von f für f:



(1) Bestimme die Beobachtungspaare (x_i, y_i), $i = 1, \ldots, n$ und trage sie in ein Koordinatensystem ein.

(2) Anhand der Punktwolke wird Funktionstyp von f festgelegt.

Hier z.B. Polynom vom Grad 2 oder 3:

$$\hat{f}_2(x) = a_2 x^2 + a_1 x + a_0$$

$$\hat{f}_3(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0$$

Unbekannt sind die Koeffizienten $a_0, a_1, \ldots$

Allgemeine Form der Ansätze in der linearen Regressionsanalyse

$$\hat{f}(x) = \sum_{k=0}^{m} a_k \varphi_k(x),$$

$\varphi_k(x)$ – bekannte Funktionen,

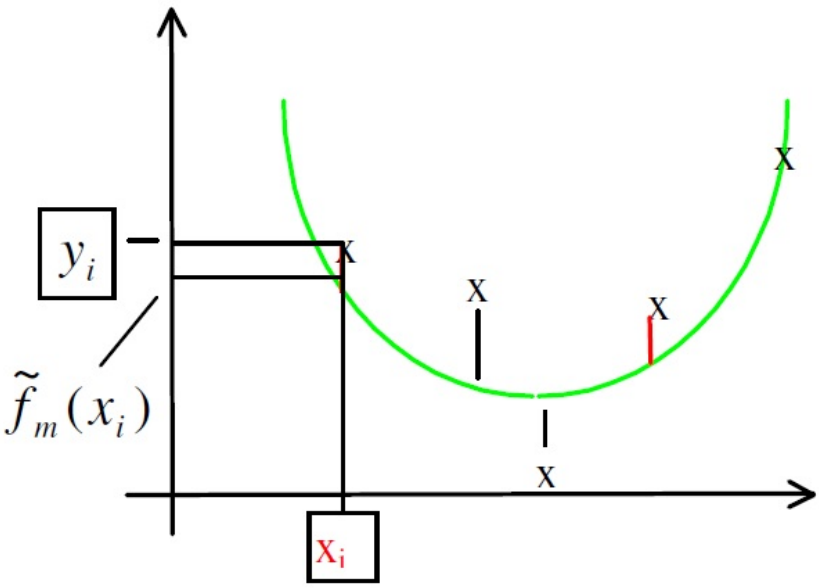
z.B. x^k

$$(3)\quad \mathcal{K}_m = \left\{ \tilde{f}_m \;\middle|\; \hat{f}_m(x) = \sum_{k=0}^{m} a_k \varphi_k(x) \right\},$$

$a_k \in \mathbb{R}$

ist die Menge der Kandidaten.

Suchen wir im Rn
nach einem Kriterium der
optimalen Kandidat.



Polynom 2. Grades

Gütekriterium:

$$\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 \rightarrow \min_{f_m}$$

Def.: $u_i = y_i - \tilde{f}_m(x_i)$ heißt ***Residuum*** an der Stelle x_i.

Die Summe $\sum_{i=1}^{n}\left(y_i - \tilde{f}_m(x_i)\right)^2 = \sum_{i=1}^{n}\left(y_i - \left(\sum_{\nu=1}^{m} a_\nu g_\nu(x_i)\right)\right)^2$ heißt ***Residual Sum of Squares*** *(RSS)* von $\tilde{f}_m(x_i)$.

Die Lösung $\hat{f}_m(x)$ des Extremwertproblems $\min_{\tilde{f}_m \in \mathfrak{I}_m} \sum_{i=1}^{n}\left(y_i - \tilde{f}_m(x_i)\right)^2$ heißt ***lineare Regressionsfunktion.***

Minimierer der Summe aus der Menge K_m.

$\longrightarrow$ Extremwertproblem

Methode der Kleinsten Quadrate

Ansatz:

$$y = a_0 f_0(x) + a_1 f_1(x) + \dots + a_m f_m(x) + \varepsilon$$

$\Longrightarrow$ Für die Beobachtungen:

$$y_1 = a_0 f_0(x_1) + \dots + a_m f_m(x_1) + \varepsilon_1$$

$$y_2 = a_0 f_0(x_2) + \dots + a_m f_m(x_2) + \varepsilon_2$$

$$\dots$$

$$y_n = a_0 f_0(x_n) + \dots + a_m f_m(x_n) + \varepsilon_n$$

$$y_k - \varepsilon_k = a_0 f_0(x_k) + \dots\dots +$$

$$+ a_m f_m(x_k) =$$

$$\left\langle \begin{pmatrix} f_0(x_k) \\ f_1(x_k) \\ \dots \\ f_m(x_k) \end{pmatrix}; \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} \right\rangle =$$

$$= \underbrace{\left(f_0(x_k); f_1(x_k); \dots; f_m(x_k) \right)}_{\text{Zeile der Matrix}} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}$$

gesuchter Vektor

$\Rightarrow$

$$\begin{pmatrix} g_0(x_1) & g_1(x_1) & \cdots & g_m(x_1) \\ g_0(x_2) & g_1(x_2) & \cdots & g_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(x_n) & g_1(x_n) & \cdots & g_m(x_n) \end{pmatrix} =$$

$$=: G = \left(g_k(x_i) \right)_{k=0,\dots,m}^{i=1,\dots,n}$$

$$\in \mathbb{R}^{n \times (m+1)}$$

$$\Rightarrow \quad \bar{y} = G\bar{a} + \bar{\varepsilon}$$

$$\overline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \overline{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Satz! Sei G wie oben mit $\operatorname{rang}(G) = m+1$

Die Parameter $\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_m$ der Lösung $\tilde{f}_m = \sum_{k=0}^{m} \tilde{a}_k g_k(x)$ des Extremwertproblems

$$\sum_{i=1}^{n} \left(y_i - \tilde{f}_m(x_i)\right)^2 \longrightarrow \min_{\tilde{f}_m \in \mathcal{K}_m}$$

ergeben sich als Lösung des

Normalengleichungssystem:

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

$G^T G$

$[(m+1) \times n] \cdot [n \times (m+1)] = (m+1) \times (m+1)$

ist quadratisch und hat vollen Rang $(m+1)$

$$\vec{a} = (G^T G)^{-1} G^T \vec{y}$$

Def: Die Lösung

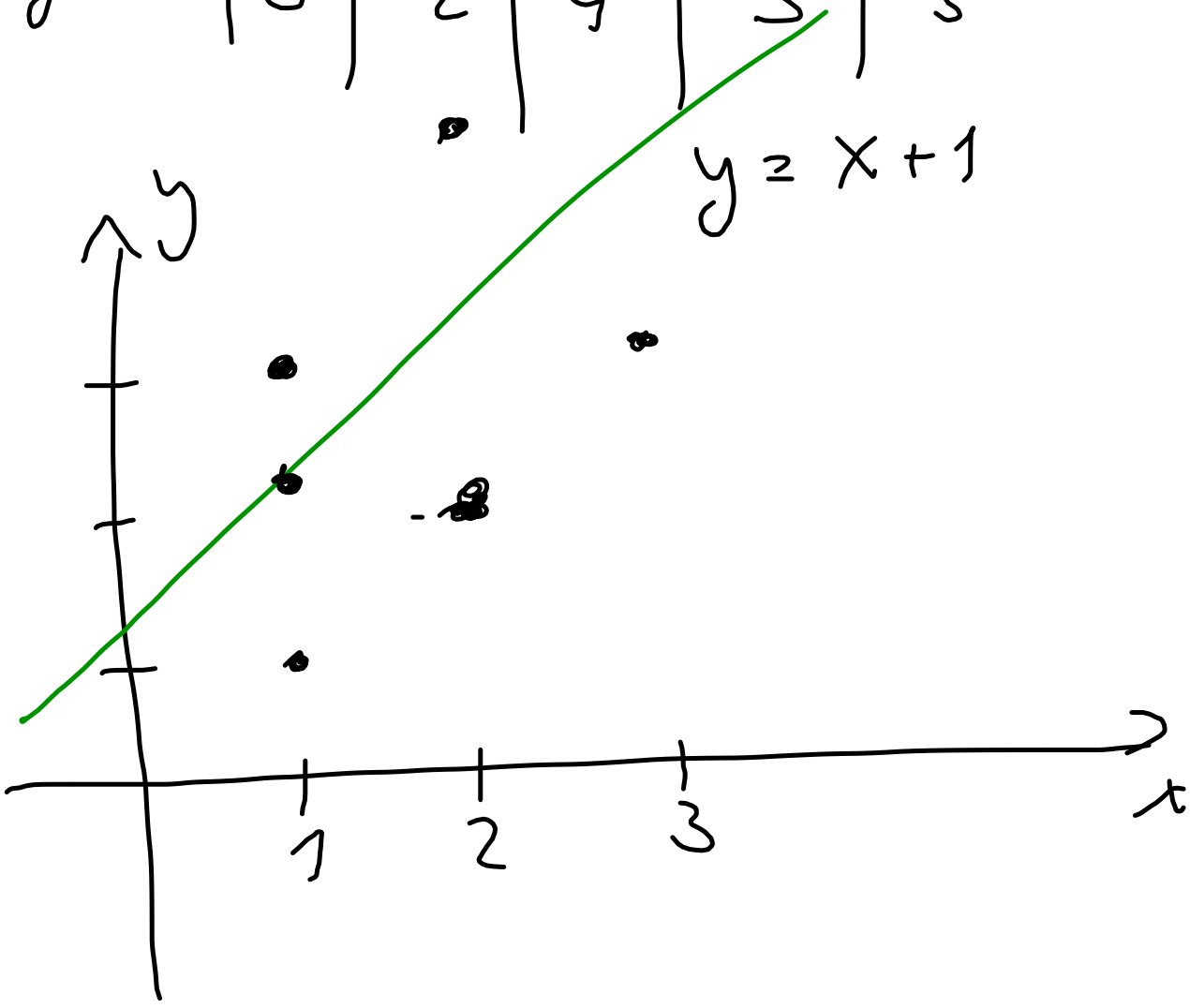
$$\tilde{a} = \begin{pmatrix} \tilde{a}_0 \\ \vdots \\ \tilde{a}_m \end{pmatrix}$$ heißt

gewöhnliche kleinste-Quadrat-

schätzung von $\vec{s}$.

Bsp.:

i	1	2	3	4	5	6
x:	1	1	2	2	3	3
y:	1	3	2	4	3	5



$$y = ax + b = a_1 x + a_0$$

$$G = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix}$$

$$\tilde{y} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix}$$

$\Rightarrow$ Normala system:

$$G^T G \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = G^T \bar{y}$$

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix}$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & \sum x_i = 12 \\ \sum x_i = 12 & \sum x_i^2 = 28 \end{pmatrix}$$

Gerade:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \end{pmatrix}$$

$$G^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 3 \\ 5 \end{pmatrix} =$$

$$= \begin{pmatrix} 18 \\ 40 \end{pmatrix}$$

$$\Rightarrow \quad a_0 = a_1 = 1$$

Weitere Veranstaltungen wurden nicht elektronisch aufgezeichnet, s. eigene Notitzen