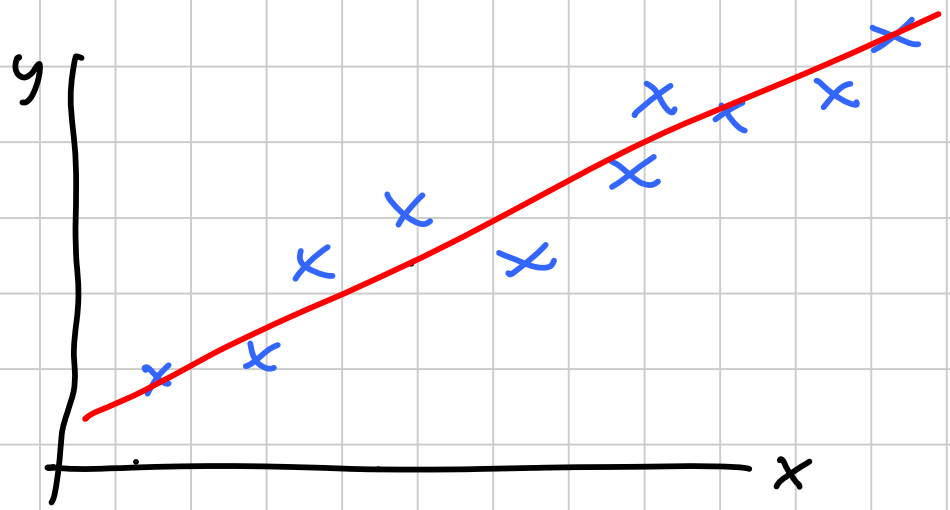
§5 Lineare Regression



Reale Daten liegen (nach geeigneter Skalierung, z.B. log-log-Plot) bestenfalls näherungsweise auf einer Geraden.

Gründe
- Messungenauigkeiten
- Änderung der äußeren Bedingungen
- Unregelmäßigkeiten der Natur
- Meßgrößen voneinander unabhängig

Ziel: finde "bestapproximierende" Gerade

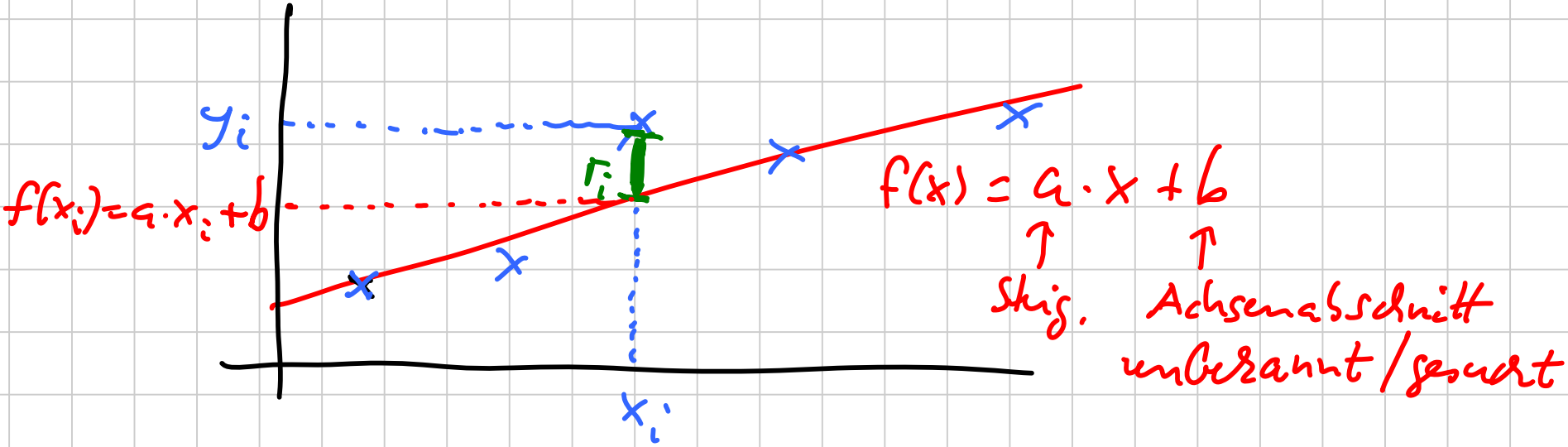
Genauer: Finde zu gegebenen Daten

x	x_1	x_2	..	x_n
y	y_1	y_2	...	y_n

eine Gerade $f(x) = a \cdot x + b$, sodaß die "durchschnittliche Abweichung" zwischen realen Daten y_i und theoretischen Werten $f(x_i) = a \cdot x_i + b$ Minimal

5.1 Bestimmung der Regressionsgeraden

(C. F. Gauss 1777 – ...)



- Fehler im i^{ten} Datenpunkt: $r_i = y_i - f(x_i)$
  $= y_i - (a x_i + b)$
- quadratischer Fehler im i^{ten} Datenpunkt: $r_i^2 = (y_i - (a \cdot x_i + b))^2$

r_i^2 immer ≥ 0 (gut)

Auch andere Größen denkbar, z.B. $|r_i| = \begin{cases} r_i, & r_i \geq 0 \\ -r_i, & r_i < 0 \end{cases}$

aber dann Theorie komplizierter.

- mittlerer quadratischer Fehler ($\frac{\text{Summe}}{\text{Anzahl}}$ der r_i^2):

$$R(a,b) = \frac{1}{n}\left(r_1^2 + r_2^2 + \dots + r_n^2\right)$$

R: Funktion, die von den Parametern a und b der Gerade abhängt

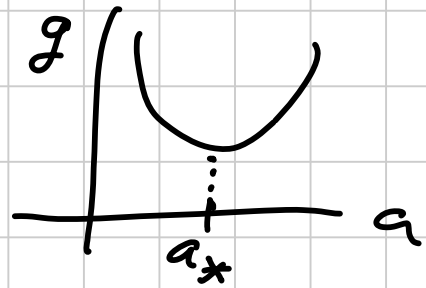
$$= \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

$\sum$: Summenzeichen

Regressionsgerade: Wähle a und b so, daß der mittlere quadratische Fehler $R(a,b)$ minimal wird.

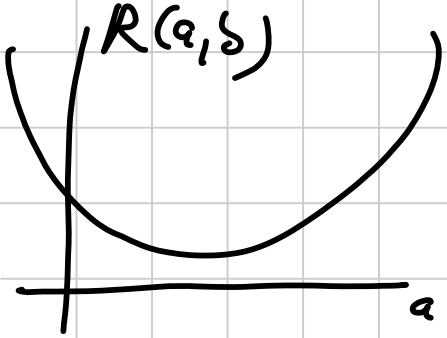
Methode zur Minimierung von $g(a)$:

An der Minimumsstelle a_* gilt $g'(a_*) = 0$



Also 1) Ableitung bestimmen
2) Nullstelle der Ableitung bestimmen

Hier: 1) $R(a,b)$ nach a bzw. b ableiten
2) gleich Null setzen $\rightsquigarrow$ 2 Gleichungen
3) a und b aus den 2 Gl'en bestimmen



1),2) $R(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2$

Ableitung nach a: $\frac{1}{n}\cdot\sum_{i=1}^{n} \underbrace{2(y_i - ax_i - b)}_{\text{äußere Abl.}} \cdot \underbrace{(-x_i)}_{\text{innere Abl.}} \overset{!}{=} 0$ (1)

(Kettenregel: $h(g(a))' = h'(g(a)) \cdot g'(a)$

$h(g) = g^2 \rightsquigarrow h'(g) = 2g$

$g(a) = y_i - ax_i - b \rightsquigarrow g'(a) = -x_i$)

Ableitung nach b: $\frac{1}{n}\sum_{i=1}^{n} 2(y_i - ax_i - b)\cdot(-1) \overset{!}{=} 0$ (2)

3) (2) nach b auflösen $(2)\cdot\left(-\frac{1}{2}\right)$

$$\Rightarrow \quad 0 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{=\bar{y}\ \text{Mittelwert der } y_i} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} a x_i}_{=a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i = a\cdot\bar{x}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} b}_{=\underbrace{b+\ldots+b}_{n\ \text{mal}}}$$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$: Mittelwert der x_i

$$= \bar{y} - a\cdot\bar{x} - b \qquad \Big| +b$$

$$\Rightarrow \boxed{b = \bar{y} - a\cdot\bar{x}}$$

Einsetzen in (1), nach a auflösen: $(1)\cdot\left(-\frac{n}{2}\right)$

$$0 = \sum_{i=1}^{n} (y_i - a x_i - b)\cdot x_i$$

b einsetzen $\Rightarrow$ $\sum_{i=1}^{n} (y_i - a x_i - \bar{y} + a\bar{x})\cdot x_i$

$$= \sum_{i=1}^{n} \left((y_i - \bar{y}) - a (x_i - \bar{x}) \right) \cdot x_i$$

$$= \sum_{i=1}^{n} (y_i - \bar{y}) \cdot x_i - a \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot x_i \qquad (*)$$

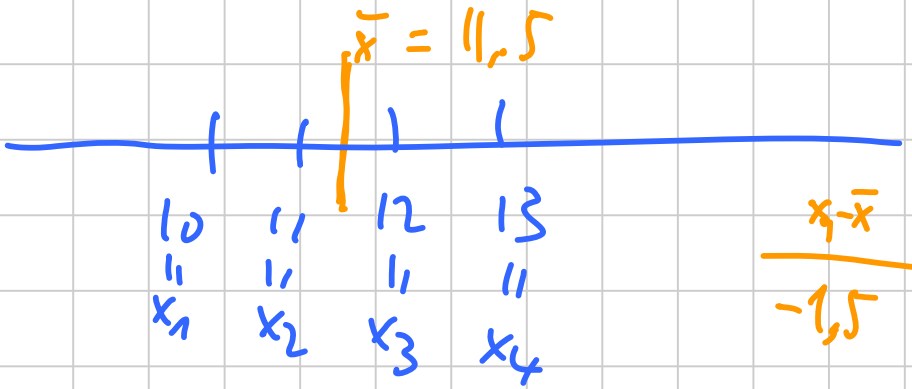
⚠ könnte 0 sein?

Nebenrechnung: Seien $x_1, \ldots, x_n$ beliebige Zahlen

Sei $\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ der Mittelwert.

Dann gilt

$$(x_1 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0$$



$x_1 - \bar{x}$	$x_2 - \bar{x}$	$x_3 - \bar{x}$	$x_4 - \bar{x}$
$-1{,}5$	$-0{,}5$	$0{,}5$	$1{,}5$

Beweis: $(x_1 - \bar{x}) + \ldots + (x_n - \bar{x})$

$= (\underbrace{x_1 + \ldots + x_n}_{n\bar{x}}) - n \cdot \bar{x} = 0$

Also:

$$(*) = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - a \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})$$

$$\Rightarrow 0 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - a \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$ schon wenn $n \geq 2$ (mindestens 2 Datenpunkte)

$$\Leftrightarrow \boxed{a = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Satz 5.1 Die "bestapproximierende" Gerade eines Datensatzes

x	x_1	...	x_n
y	y_1	...	x_n

, d.h. diejenige mit dem kleinsten mittleren quadratischen Fehler, ist eindeutig bestimmt. Steigung und Achsenabschnitt sind gegeben durch

$$a = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$b = \bar{y} - a\bar{x},$$

wobei $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ (Mittelwert der x_i),

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ (—"— y_i).

Diese Gerade heißt <u>Regressionsgerade</u>.

Bsp Herzfrequenz von Fröschen

x	Temp. [°C]	2	6	10	18
y	Frequenz (Schläge/min)	5	10	22	32

$$\bar{x} = \frac{36}{4} = 9, \qquad \bar{y} = \frac{69}{4} = 17{,}25$$

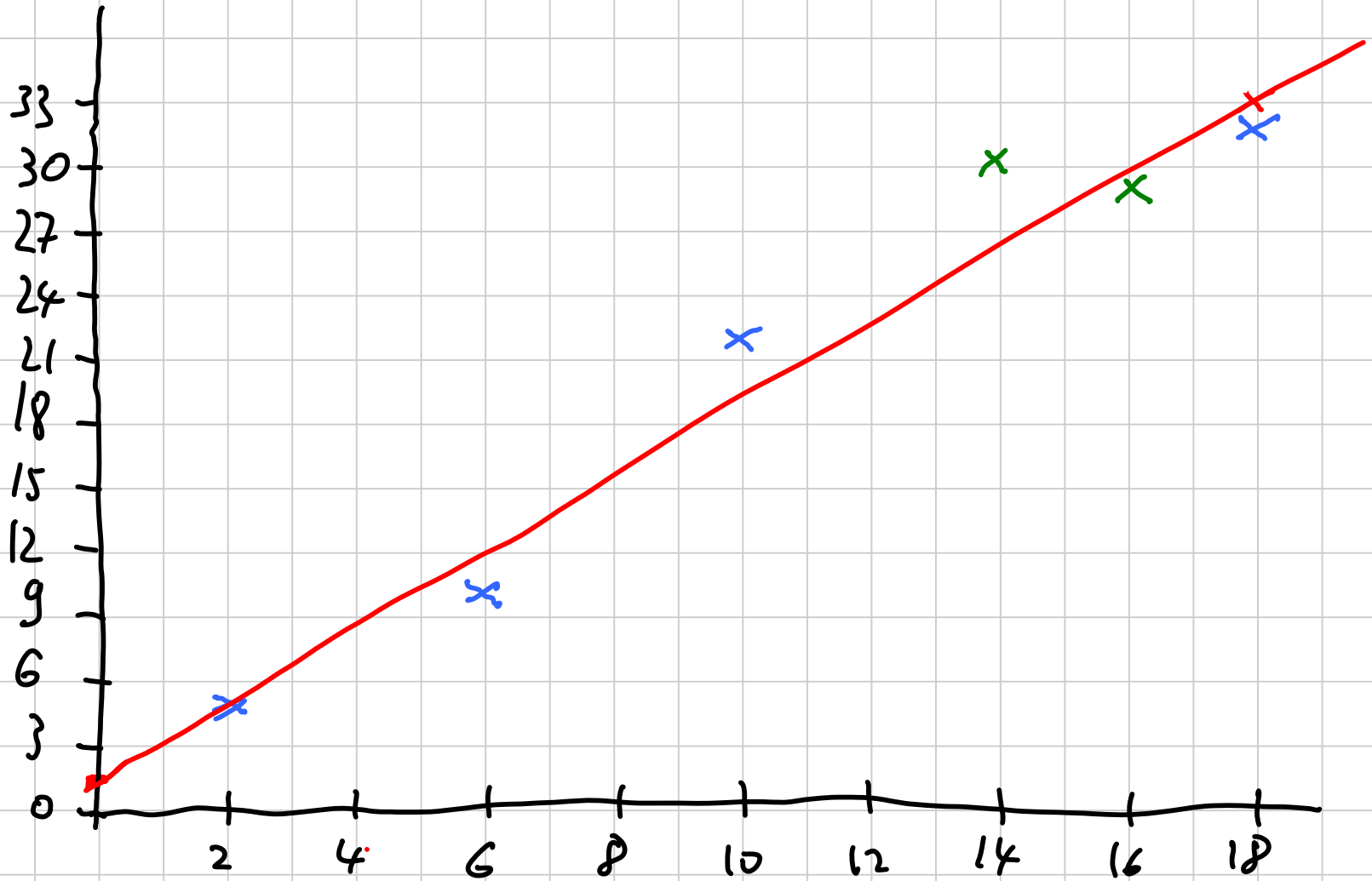
$x_i - \bar{x}$	-7	-3	1	9
$y_i - \bar{y}$	$-12{,}25$	$-7{,}25$	$4{,}75$	$14{,}75$

$$a = \frac{(-7)\cdot(-12{,}25) + (-3)\cdot(-7{,}25) + 1\cdot 4{,}75 + 9\cdot 14{,}75}{7^2 + 3^2 + 1^2 + 9^2} = 1{,}75$$

(Taschenrechner)

$$b = 17{,}25 - 1{,}75 \cdot 9 = 1{,}5$$

Regressionsgerade: $f(x) = 1{,}75x + 1{,}5$



$f(0) = 1{,}5$

$f(18) = 1{,}75 \cdot 18 + 1{,}5$

$= 33$

Vorhersage weiterer Werte

	Theorie $(a \cdot x + b)$	Messung (y_i)
$x = 14$	26	30
$x = 16$	29,5	29