3. Lineare Regression.

3.1. Summen und Mittelwerte.

Sind $x_1, \ldots, x_n$ reelle Zahlen, so bezeichnen wir mit

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

die **Summe** dieser Zahlen. Die abkürzende Schreibweise mit dem "Summenzeichen"

$$\sum_{i=1}^{n} \qquad \text{oder auch} \qquad \sum\nolimits_{i=1}^{n}$$

ist sehr praktisch und wir werden sie oft verwenden; unter dem griechischen Buchstaben Groß-Sigma $\sum$ (oder an seiner rechten unteren Ecke) steht der "Lauf-Index" (hier i) zusammen mit seinem Beginn (hier $i = 1$), über dem Zeichen $\sum$ (oder an seiner rechten oberen Ecke) Seite steht, bis zu welchem Index die Summenbildung fortzusetzen ist (hier $i = n$), man sagt in diesem Fall, dass "über i summiert wird, von 1 bis n". Analog ist $\sum_{i=2}^{4} x_i = x_2 + x_3 + x_4$ (hier wird über i summiert, und zwar von 2 bis 4). Der Lauf-Index braucht nicht i zu heißen, wir hätten ebenso $\sum_{t=1}^{n} x_t$ schreiben können, das Ergebnis wäre ebenfalls $x_1 + x_2 + \cdots + x_n$ (es ist also i oder t nichts anderes als ein "Platzhalter"). Sind die Zahlen $x_1, \ldots, x_n$ gegeben, und schreibt man einfach $\sum x_i$, so soll dies nichts anderes als $\sum_{i=1}^{n} x_i$ bedeuten (man geht also stillschweigend davon aus, dass i der Lauf-Index ist und dass von 1 bis n summiert wird). Der Index i kann in den Summanden mehrfach vorkommen, so ist $\sum_{i=1}^{3} (x_i y_i)^{2i}$ nichts anderes als $(x_1 y_1)^2 + (x_2 y_2)^4 + (x_3 y_3)^6$; genauso gut kann es passieren, dass i gar nicht vorkommt: es ist $\sum_{i=1}^{3} 2 = 2 + 2 + 2$; hier sind also drei Summanden zu addieren, und alle sind gleich 2; entsprechend ist $\sum_{i=1}^{n} a = n \cdot a$ für jede Zahl a. Hier eine wichtige (aber offensichtliche) Rechenregel:

$$a \cdot \sum\nolimits_{i=1}^{n} x_i = \sum\nolimits_{i=1}^{n} a \cdot x_i,$$

dies ist gerade das Distributivgesetz (ausgeschrieben: $a(x_1 + x_2 + \cdots + x_n) = ax_1 + ax_2 + \cdots + ax_n$). Entsprechend übertragen sich die weiteren Rechengesetze der Addition.

Sind $x_1, \ldots, x_n$ reelle Zahlen (mit $n \geq 1$), so bezeichnet man mit

$$\overline{x} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i$$

den **Mittelwert** der Zahlen x_i.

Der Mittelwert ist durch die folgende Eigenschaft charakterisiert:

Satz. *Seien $x_1, \ldots, x_n \in \mathbb{R}$ (und $n \geq 1$). Die Funktion*

$$h(t) = \sum_{i=1}^{n} (t - x_i)^2$$

nimmt für $t = \overline{x}$ ihr Minimum an.

Man kann dies auch folgendermaßen formulieren: *Für den Mittelwert $\overline{x}$ der Zahlen $x_1, \ldots, x_n$ ist die die Summe der Quadrate der Abweichungen $\overline{x} - x_i$ minimal.*

Man betrachtet hier also die "Summe der quadratischen Abweichungen" und minimiert sie! Diese Methode, mit der wir uns hier beschäftigen, heißt die **Methode der kleinsten Quadrate**.

Beweis des Satzes. Wir wollen zeigen, dass die Funktion $h(t)$ für $t = \overline{x}$ ihr Minimum annimmt. Verwenden wir die zweite binomische Formel, so können wir $h(a)$ folgendermaßen umscheiben:

$$h(t) = \sum (t - x_i)^2 = \sum (t^2 - 2x_i t + x_i^2) = nt^2 - \big(2 \sum x_i\big)t + \sum x_i^2,$$

dies ist aber (**als Funktion in** t) nichts anderes als eine quadratische Funktion: konstanter Koeffizient ist $\sum x_i^2$, der Koeffizient von t ist $-2 \sum x_i$, der von t^2 ist n; man beachte, dass dies wirklich Konstanten sind: wir gehen davon aus, dass die Zahlen $x_1, \ldots, x_n$ fest gegebene Zahlen sind, an denen nicht gewackelt wird). Da der Koeffizient n von t^2 positiv ist, wissen wir, dass der Graph von h eine nach oben geöffnete Parabel ist.

> Verwiesen sei auf Teil 4, in dem wir uns umfassend mit quadratischen Funktionen beschäftigen werden.

Um das Minimum zu finden kann man nun entweder (wie im SI-Unterricht) mit Hilfe der quadratischen Ergänzung die Scheitelpunktsform herstellen, oder man kann die Funktion einfach ableiten und die Nullstelle der Ableitung suchen: an dieser Stelle muß das Minimum liegen. Die Ableitung (wohlgemerkt, nach t) ist

$$h'(t) = 2nt - 2 \sum x_i$$

also gilt $h'(t) = 0$ genau dann, wenn $t = \frac{1}{n} \sum x_i$ ist.

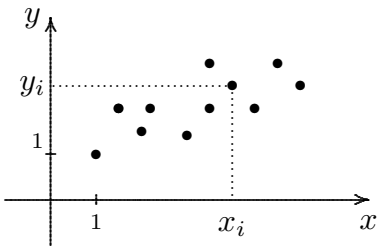
3.2. Einführung.

Gegeben seien Zahlenpaare $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$. Wir stellen die Frage, ob sich die Zahlen y_i als Werte einer linearen Funktion $x_i \mapsto y_i$ beschreiben lassen; genauer: wie sieht eine lineare Funktion $f(x) = a + bx$ aus, so dass $y_i \approx f(x_i)$ gilt. Anders formuliert: Tragen wir die Paare (x_i, y_i) in ein x-y-Koordinatensystem ein (wir erhalten damit eine "Punktwolke", man sagt auch "Scatter-Plot" oder "Streu-Diagramm"), so suchen wir eine Gerade, auf der (im schönsten Fall) alle diese Punkte liegen, zumindest sollen alle Punkte so nah wie möglich an dieser Gerade liegen; man nennt eine solche

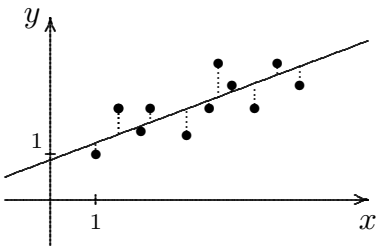
Gerade eine *Ausgleichsgerade* oder eine *Regressionsgerade*. Was soll dabei $\approx$ bedeuten, was wollen wir unter "so nah wie möglich" verstehen? Wir betrachten die Abweichungen, also die Differenzen $f(x_i) - y_i$, diese Abweichungen sollen so klein wie möglich sein: es hat sich herausgestellt, dass es günstig ist, die Quadrate dieser Abweichungen zu betrachten, deren Summe zu bilden, und zu verlangen, dass diese Zahl

$$\sum (f(x_i) - y_i)^2$$

so klein wie möglich ist: man will also *die Summe der quadratischen Abweichungen minimieren.*



Die gegebene Punktwolke



Die Regressionsgerade

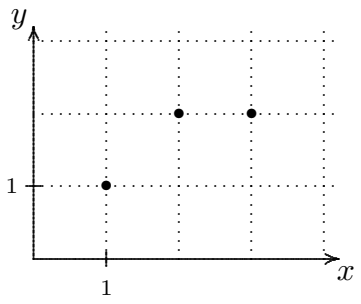
Noch einmal: Was wir hier betrachten, sind die Abweichungen

$$f(x_i) - y_i,$$

die so klein wie möglich sein sollen (man nennt diese Abweichungen auch **Residuen.**) Warum ist es die **Summe der quadratischen Abweichungen,** die von Bedeutung ist? Wenn wir Zahlen $x_1, x_2, \ldots, x_n$ gegeben haben, so haben wir gesehen, dass für den Mittelwert $\overline{x}$ der Zahlen $x_1, \ldots, x_n$ gilt: die Summe der Quadrate der Abweichungen $\overline{x} - x_i$ ist minimal. Nun wollen wir nicht einen Mittelwert für eine Menge von Zahlen bilden, sondern wir haben eine Menge von Zahlen**paaren** gegeben, und suchen eine lineare Funktion $f(x)$, so dass nun die Summe der Quadrate der Abweichungen $\overline{f}(x_i) - y_i$ minimal ist.

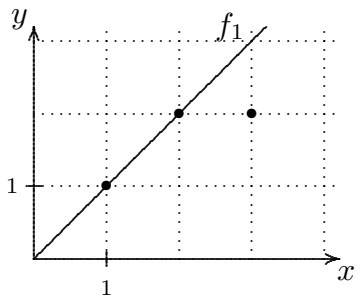
Beispiel. Wir betrachten ein ganz einfaches Beispiel. Gegeben seien drei Zahlenpaare (x_i, y_i), etwa

$$(x_1, y_1) = (1 \mid 1)$$
$$(x_2, y_2) = (2 \mid 2)$$
$$(x_3, y_3) = (3 \mid 2)$$



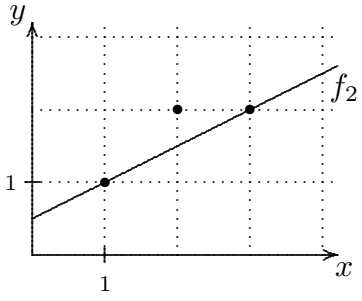
Wenn nur zwei Zahlenpaare (x_1, y_1) und $(x_2.y_2)$ gegeben wären (und nicht gerade $x_1 = x_2$ gilt!) so ist die eindeutig bestimmte Gerade durch diese beiden Punkte der Graph einer linearen Funktion f und natürlich sind dann die beiden Residuen gleich Null. Der erste interessante Fall ist daher der Fall, dass **drei** Zahlenpaare gegeben sind!

Wir betrachten also den Fall der drei Zahlenpaare (x_i, y_i). Legen wir durch zwei dieser Zahlenpaare jeweils eine Gerade, so erhalten wir drei verschiedene Geraden, nämlich



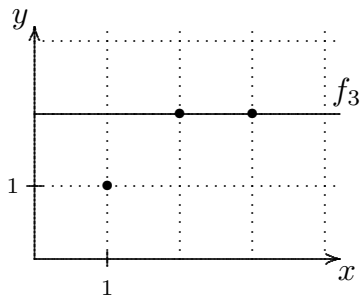
$f_1(x) = x$

also $a = 0$ und $b = 1$



$f_2(x) = \frac{1}{2}x + \frac{1}{2}$

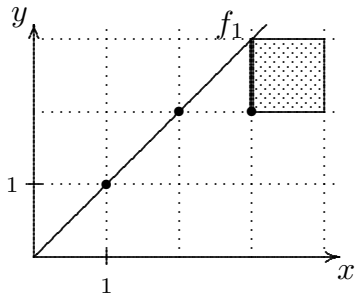
also $a = \frac{1}{2}$ und $b = \frac{1}{2}$

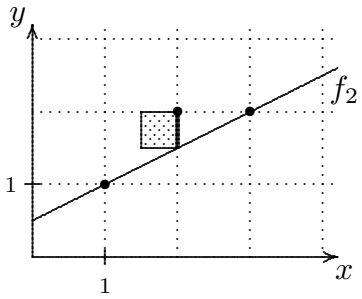


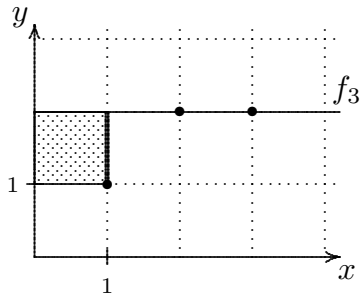
$f_3(x) = 2$

also $a = 2$ und $b = 0$

In allen drei Fällen erhalten wir jeweils ein einziges von Null verschiedenes Residuum. Die folgenden Bilder zeigen die Quadrate dieser Residuen (die Residuen sind fett gezeichnet, links oder rechts sieht man die zugehörigen Quadrate):

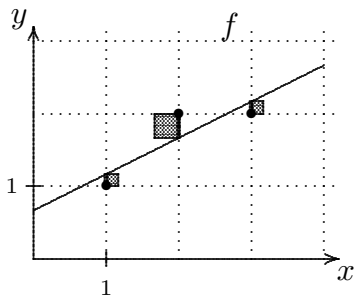






Wenn man also nur diese drei Geraden vergleicht, so ist bezüglich der Minimierungsfrage die zweite Gerade viel besser als die anderen beiden Geraden: bei der zweiten Geraden ist die Summe der quadratischen Abweichungen gleich $\frac{1}{4}$, während im ersten und in dritten Fall die Summe der quadratischen Abweichungen gleich 1 ist. Aber auch die zweite Gerade ist nicht optimal!

Hier ist die bestmögliche Gerade, sie ist durch $f(x) = \frac{2}{3} + \frac{1}{2}x$ gegeben:



Die Summe der Quadrate der Residuen ist hier $(\frac{1}{6})^2 + (\frac{2}{6})^2 + (\frac{1}{6})^2 = \frac{1+4+1}{36} = \frac{1}{6}$. Dass dies der optimale Wert ist, ist nicht offensichtlich! Im nächsten Abschnitt wird gezeigt, dass es immer eine eindeutig bestimmte "Regressionsgerade" $f(x) = a + bx$ gibt und wie man die Koeffizienten a und b berechnet. Die Formeln, die im nächsten Abschnitt hergeleitet werden, besagen:

$$b = \frac{\sum x_i y_i \; - n \cdot \overline{x} \cdot \overline{y}}{\sum x_i^2 \; - n \cdot \overline{x}^2} \qquad \text{und} \qquad a = \overline{y} - b\overline{x}.$$

Beim Rechnen mit der Hand empfiehlt es sich, mit folgendem Rechenschema zu arbeiten:

x_i	1	2	3	$\sum x_i$	$=$ ...	$\overline{x}$	$=$ ...
y_i	1	2	2	$\sum y_i$	$=$ ...	$\overline{y}$	$=$ ...
$x_i y_i$	...	...	...	$\sum x_i y_i$	$=$ ...		
x_i^2	...	...	...	$\sum x_i^2$	$=$ ...		

Führt man diese Rechnungen durch, so erhält man:

x_i	1	2	3	$\sum x_i$	$= 6$	$\overline{x}$	$= 2$
y_i	1	2	2	$\sum y_i$	$= 5$	$\overline{y}$	$= \frac{5}{3}$
$x_i y_i$	1	4	6	$\sum x_i y_i$	$= 11$		
x_i^2	1	4	9	$\sum x_i^2$	$= 14$		

also

$$b = \frac{11 - 3 \cdot 2 \cdot \frac{5}{3}}{14 - 3 \cdot 2 \cdot 2} = \frac{11 - 10}{14 - 12} = \frac{1}{2},$$

$$a = \frac{5}{3} - \frac{1}{2} \cdot 2 = \frac{5}{3} - 1 = \frac{2}{3}.$$

3.3. Die Regressionsgerade.

Gegeben seien also Zahlenpaare (x_i, y_i) mit $1 \leq i \leq n$. Wir suchen eine lineare Funktion $f(x)$, für die

$$\sum (f(x_i) - y_i)^2$$

minimal wird. Eine lineare Funktion hat die Form $f(x) = a + bx$, gesucht sind also reelle Zahlen a und b, so dass

$$\sum (a + bx_i - y_i)^2$$

minimal ist.

Wir werden voraussetzen, dass die Zahlen $x_1, \ldots, x_n$ nicht alle gleich sind. Gilt nämlich $x_i = c$ für alle i, so liegen alle Punkten $(x_i, y_i) = (c, y_i)$ auf der Geraden $x = c$ und es macht wenig Sinn, nach einer Funktion f mit $f(x_i) \approx y_i$ zu suchen.

Satz. *Seien n Zahlenpaare (x_i, y_i) gegeben. Wir setzen voraus, dass die Zahlen x_i nicht alle gleich sind. Dann gibt es Zahlen a, b in $\mathbb{R}$, sodass*

$$\sum (a + bx_i - y_i)^2$$

minimal ist, und diese Zahlen a, b sind eindeutig bestimmt. Die folgenden Formeln liefern zuerst b und dann a:

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\sum x_i y_i \ - n \cdot \overline{x} \cdot \overline{y}}{\sum x_i^2 \ - n \cdot \overline{x}^2} \qquad \text{und} \qquad a = \overline{y} - b\overline{x}$$

.

Bemerkung: Da wir voraussetzen, dass die x_i nicht alle gleich sind, ist der Nenner von b von Null verschieden (ansonsten würde die Formel für b keinen Sinn machen).

Der Satz behauptet **dreierlei:**

- Erstens: das Minimierungproblem besitzt eine Lösung.
- Zweitens: es gibt **nur eine** Lösung.
- Drittens: Es gibt eine Formel, die die Lösung liefert.

Man muss sich hier klar machen, dass **keine** dieser Aussagen offensichtlich ist. Erstens: Man kann viele Minimierungsprobleme formulieren, die gar keine Lösung besitzen (Beispiel: Finde die kleinste reelle Zahl $r > 0$. Oder: Finde die kleinste ganze Zahl...). Zweitens: Es gibt viele Minimierungsprobleme, für die es mehrere Lösungen gibt (Beispiele später). Drittens: Es gibt Minimierungsprobleme, für die man zeigen kann, dass es eine einzige Lösung gibt, wo es

aber schwierig oder sogar unmöglich ist, eine Lösung explizit anzugeben.

Beweis, dass die beiden Formeln für b das gleiche liefern: Wir zeigen als erstes, dass die Zähler gleich sind:

$$\begin{aligned}\sum (x_i - \overline{x})(y_i - \overline{y}) &= \sum (x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{x} \cdot \overline{y}) \\ &= \sum x_i y_i - \sum x_i \overline{y} - \sum \overline{x} y_i + \sum \overline{x} \cdot \overline{y} \\ &= \sum x_i y_i - \left(\sum x_i\right) \cdot \overline{y} - \overline{x} \left(\sum y_i\right) + \sum \overline{x} \cdot \overline{y} \\ &= \sum x_i y_i - n \cdot \overline{x} \cdot \overline{y} - \overline{x} \cdot n \cdot \overline{y} + n \cdot \overline{x} \cdot \overline{y} \\ &= \sum x_i y_i - n \cdot \overline{x} \cdot \overline{y}\end{aligned}$$

dabei haben wir einerseits verwendet, dass $\sum x_i = n \cdot \overline{x}$ und $\sum y_i = n \cdot \overline{y}$ gilt, andererseits muß man beachten, dass $\sum \overline{x} \cdot \overline{y}$ die Summe von n gleichen Termen der Form $\overline{x} \cdot \overline{y}$ ist. Dies zeigt, dass die beiden Zähler gleich sind. Eine entsprechende Rechnung kann für die Nenner durchgeführt werden, aber man kann auch unmittelbar sehen, dass die Gleichheit der Nenner ein Spezialfall der gerade bewiesenen Gleichheit der Zähler ist.

Beweis der Formeln für a und b. Gesucht sind Zahlen a, b, so dass der folgende Wert

$$H(a, b) = \sum (a + b x_i - y_i)^2$$

(der von a und b abhängt) minimal wird (dabei sind die Zahlen x_i, y_i fest vorgegebene Zahlen). Wir werden folgendermaßen vorgehen: Wir wählen als erstes ein willkürliches b und zeigen, dass $H(a, b)$ nur dann minimal sein kann, wenn a die angegebene Form hat. Wir betrachten also jetzt die Funktion

$$h(a) = \sum (a + b x_i - y_i)^2 ,$$

die nur noch von a abhängt (alle x_i, y_i und auch b sind feste Zahlen). Dies ist eine quadratische Funktion in der Variablen a, die nur nicht-negative Werte annimmt, sie wird also durch eine (nach oben geöffnete) Parabel beschrieben. Eine derartige Funktion hat ein eindeutig bestimmtes Minimum, das wir durch das Nullsetzen der Ableitung (wir differenzieren nach a) berechnen können: Die Ableitung ist

$$h'(a) = \sum 2 (a + b x_i - y_i)$$

Ist $h'(a) = 0$, so ist $\sum (a + b x_i - y_i) = 0$, also

$$\begin{aligned}0 = \sum (a + b x_i - y_i) &= \sum a + b \sum x_i - \sum y_i \\ &= n a + b n \overline{x} - n \overline{y},\end{aligned}$$

also $a = \overline{y} - b \overline{x}$.

Nun betrachten wir entsprechend die Abhängigkeit der Funktion $H(a,b)$ von b, also

$$g(b) = H(a,b) = \sum (a + bx_i - y_i)^2 ;$$

wie wir wissen, gilt bei einem optimalen Paar (a,b) die Beziehung $a = \overline{y} - b\overline{x}$, also

$$g(b) = \sum (a + bx_i - y_i)^2 = \sum (\overline{y} - b\overline{x} + bx_i - y_i)^2 .$$

Wir schreiben die Klammer in der Form $(x_i - \overline{x})b - (y_i - \overline{y})$ und erhalten durch Quadrieren:

$$\begin{aligned} g(b) &= \sum \left((x_i - \overline{x})^2 b^2 - 2(x_i - \overline{x})b(y_i - \overline{y}) + (y_i - \overline{y})^2 \right) \\ &= \sum (x_i - \overline{x})^2 \cdot b^2 - 2 \sum (x_i - \overline{x})(y_i - \overline{y}) \cdot b + \sum (y_i - \overline{y})^2 \end{aligned}$$

auch dies ist (**als Funktion in** b) eine quadratische Funktion, deren höchster Koeffizient $\sum (x_i - \overline{x})^2$ eine Summe von Quadratzahlen, also positiv ist. Also sehen wir: die Funktion $g(b)$ wird wieder durch eine nach oben geöffnete Parabel beschrieben; sie hat wieder ein eindeutig bestimmtes Minimum, das wir durch das Nullsetzen der Ableitung (jetzt differenzieren wir nach b) berechnen können: Die Ableitung ist

$$g'(b) = 2 \sum (x_i - \overline{x})^2 b - 2 \sum (x_i - \overline{x})(y_i - \overline{y})$$

Ist $g'(b) = 0$, so ist

$$2 \sum (x_i - \overline{x})^2 b = 2 \sum (x_i - \overline{x})(y_i - \overline{y}).$$

Wir lösen nach b auf und erhalten:

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}.$$

Dies wollten wir zeigen! **Ende des Beweises.**

> Hinweis: Funktionen in **zwei** Variablen, wie die hier betrachtete Funktion $H(a,b)$ werden wir später noch genauer analysieren und uns dabei mit dem Differenzieren nach den beiden Variablen a und b (also in Richtung der zwei Koordinatenrichtungen) beschäftigen: man nennt dies "partielles Differenzieren".

Die Regressionsgerade geht immer durch den Punkt $(\overline{x}, \overline{y})$. Dies sieht man unmittelbar, wenn man die Formel für a umschreibt: $\overline{y} = a + b \cdot \overline{x}$.

Kennt man die Regressionsgerade $f(x)$, so verwendet man sie zum **Interpolieren** und **Extrapolieren.** Für jedes $x \in \mathbb{R}$ kann man $f(x)$ berechnen. Liegt x zwischen zwei x-Werten x_i und x_j, so geht man davon aus, dass man $f(x)$ als den Funktionswert an

der Stelle x ansehen sollte — hier handelt es sich um eine Interpolation. Man berechnet aber auch Werte $f(x)$ wobei x außerhalb der gegebenen Daten x_i liegt: hier handelt es sich um eine Extrapolation: man will auf diese Weise versuchen, Information über die weitere Entwicklung zu erhalten (über den Trend), oder aber Informationen über die Vergangenheit.

Effektives Rechnen: Die zweite Formel für b ist meist praktischer, da man weniger Rechenschritte benötigt. Insbesondere aber aus folgendem Grund: Die Mittelwerte $\overline{x}$, $\overline{y}$ werden hier erst als Letztes berechnet; fügt man ein weiteres Zahlenpaar (x_{n+1}, y_{n+1}) hinzu, so werden sich natürlich die Mittelwerte $\overline{x}$ und $\overline{y}$ ändern, man kann aber die alte Zwischensumme $\sum_{i=1}^{n} x_i y_i$ für die neue Berechnung mitverwenden.

Viele Taschenrechner und PC-Programme haben die Formeln für a und b eingebaut: man gibt also nur die Folge der Zahlenpaare ein und erhält durch Tastendruck a und b, bei graphischen Taschenrechnern und bei Programmen wie EXCEL auch die entsprechende graphische Darstellung.

Zum Arbeiten mit EXCEL gibt es ein Anleitungsblatt.

Das Arbeiten mit der Regressionsgerade bedeutet gerade, dass man (linear) interpoliert! Dies macht nur Sinn, wenn die Regressionsgerade $f(x) = a + bx$ die Abhängigkeit der y-Werte von den x-Werten hinreichend gut beschreibt, wenn also wirklich $y_i \approx f(x_i)$ für alle i gilt. Wir haben betont, dass die Regressionsgerade zu jedem x-Wert einen zugehörigen y-Wert liefert, nämlich $y = a + bx$. Auch bei diesem **Interpolieren** helfen viele Taschenrechner: Sind die Zahlenpaare (x_i, y_i) eingegeben, so reicht es oft, einen x-Wert einzugeben, um dann mit einem einzigen Tastendruck den zugehörigen Wert $a + bx$ zu erhalten.

Interpretation von Zähler und Nenner von b. Es ist

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum (x_i - \overline{x})^2},$$

hier haben wir Zähler und Nenner jeweils durch n geteilt, dabei ändert sich der Wert des Bruchs ja nicht. Wir erhalten im Nenner den Term

$$\boxed{V = V_x = \frac{1}{n}\sum (x_i - \overline{x})^2}$$

man nennt V die *Varianz* der Meßwerte $x_1, \ldots, x_n$; dabei handelt es sich also (in Worten) um die "mittlere quadratische Abweichung vom Mittelwert" (das letzte Wort "Mittelwert" bezieht sich auf $\overline{x}$; "quadratische Abweichung" meint die Bildung der Terme $(x_i - \overline{x})^2$, das Wort "mittlere" steht für das Bilden des Mittelwerts $\frac{1}{n}\sum \ldots$ dieser quadratischen Abweichungen). *Die Varianz ist ein Maß für die Abweichung der Meßwerte vom Mittelwert.* Statt der Varianz wird häufig auch die Wurzel

$$\boxed{s_x = \sqrt{V_x}}$$

betrachtet; man nennt dies die *Standard-Abweichung.* (Es ist also $V_x = s_x^2$.) Der neue Zähler

$$\boxed{s_{xy} = \frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}$$

wird entsprechend die *Kovarianz* der Zahlenpaare $(x_1, y_1), \ldots, (x_n, y_n)$ genannt. Wir können demnach unsere Formel für b auch in folgender Form schreiben:

$$\boxed{b = \frac{s_{xy}}{{s_x}^2}.}$$

Bezüglich der Definition von Varianz und Kovarianz eine **Warnung:** Der hier verwendete Faktor $\frac{1}{n}$ wird in manchen Büchern durch $\frac{1}{n-1}$ ersetzt; für große Zahlen n macht dies zwar praktisch keinen Unterschied, trotzdem muß man aufpassen, was jeweils gemeint ist. Immerhin macht es für die Berechnung von b keinen Unterschied, welchen dieser beiden Faktoren man verwendet, denn es gilt:

$$b = \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum(x_i - \overline{x})^2} = \frac{\frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum(x_i - \overline{x})^2}.$$

3.4. Lineare Korrelation.

Wir nehmen nunan, dass nicht nur die x_i nicht alle gleich sind, sondern dass auch die y_i nicht alle gleich sind. Wir setzen

$$\boxed{r_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2} \cdot \sqrt{\sum(y_i - \overline{y})^2}}}$$

und nennen dies den *(linearen) Korrelations-Koeffizienten*; mit den gerade eingeführten Bezeichungen für die Standard-Abweichung und die Kovarianz gilt offensichtlich

$$\boxed{r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}}$$

(nach Voraussetzung sind s_x und s_y beide von Null verschieden). Dieser Term ist ähnlich wie der Steigungs-Koeffizient der Regressionsgeraden gebildet ($b = \frac{s_{xy}}{s_x^2}$), im Gegensatz zur Bildung von b ist der Ausdruck r_{xy} aber symmetrisch in x und y definiert.

Die Schwarz'sche Ungleichung: *Es gilt immer:* $-1 \leq r_{xy} \leq 1$.

Zusatz.

(a) *Genau dann ist* $r_{xy} = 1$, *wenn alle Paare* (x_i, y_i) *auf einer Geraden mit positiver Steigung liegen.*

(b) *Genau dann ist* $r_{xy} = -1$, *wenn alle Paare* (x_i, y_i) *auf einer Geraden mit negativer Steigung liegen.*

Der Beweis der Schwarz'schen Ungleichung wird im Abschnitt 3.5 gegeben.

Ist also $|r_{xy}| = 1$, so liegen die gegebenen Paare (x_i, y_i) auf einer Geraden, die zu keiner Koordinatenachse parallel ist. Ist $|r_{xy}|$ nahe bei 1, also etwa $r_{xy} = 0,87$ oder $r_{xy} = -0,91$, so liegen die gegebenen Paare (x_i, y_i) in der Nähe der Regressionsgeraden, diese beschreibt also recht gut den Zusammenhang zwischen den x-Werten und den y-Werten (diese Werte sind "linear korreliert"); je näher der Korrelationskoeffizient bei 1 oder -1 liegt, um so besser wird der Zusammenhang durch die Regressionsgerade beschrieben! Ist dagegen der Korrelationskoeffizient in der Nähe von 0, so "liegt keine lineare Korrelation vor". Am Korrelationskoeffizienten kann man also ablesen, wie gut die lineare Regression die vorgegebene Situation beschreibt.

Um ein Gefühl für den Korrelations-Koeffizienten von Punktwolken zu bekommen, sei auf die interaktiven Übungen verwiesen, die unter dem Namen JUMBO (**J**ava **U**nterstützte **M**ünsteraner **B**iometrie-**O**berfläche) im Internet verfügbar sind. Dort erhält man zum Beispiel Punktwolken vorgelegt, deren Korrelations-Koeffizienten man schätzen soll. Auch kann man interaktiv nachvollziehen, in welcher Weise die Verschiebung einzelner Punkte den Korrelations-Koeffizienten ändert.

Warnung. Auch wenn die Regressionsgerade die Abhängigkeit der y-Werte von den x-Werten sehr gut beschreiben sollte, so handelt es sich hierbei zuerst einmal nur um eine **statistische** Beziehung, aus der man nicht notwendigerweise auf eine **kausale** Beziehung schließen darf!

Korrelation wird manchmal als das Vorliegen eines irgendwie gearteten Zusammenhangs zwischen zwei oder mehreren Variablen verstanden; dabei hat man die Vorstellung, dass die Werte der ersten Variablen möglicherweise einen Einfluß auf die Werte der zweiten haben sollten. Hier ist aber große Vorsicht geboten!

Beispiel 1. Es gibt eine Untersuchung, die eine starke lineare Korrelation zwischen der Anzahl von Geburten und der Anzahl der Störche in Mecklenburg in den vergangenen 100 Jahren nachweist . . . .

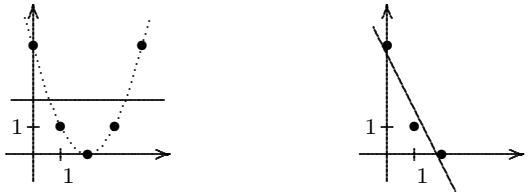
Beispiel 2. Gilt $x_1 < x_2 < \cdots < x_n$ und $y_1 < y_2 < \cdots < y_n$, so ist der Korrelations-Koeffizient auf jeden Fall positiv, oft nahe bei 1, dies besagt aber nur, dass

die beiden Zahlenreihen gleichmäßig ansteigen. Zum Beispiel: Mein Alter ($= x_i$) wächst jährlich um $+1$. Wenn nun der Benzinpreis ($= y_i$) jährlich um 10 Cent wächst, liegt eine perfekte Korrelation vor - niemand würde aber behaupten, dass die Benzinpreise steigen, weil ich älter werde, oder dass ich älter werde, weil das Benzin teurer wird . . . .

Noch eine Warnung. Der hier definierte Korrelations-Koeffizient beschreibt nur, ob eine **lineare** Korrelation vorliegt, also eine Abhängigkeit der jeweiligen Werte, die sich durch eine **lineare** Funktion beschreiben läßt. Betrachtet man etwa die folgenden Zahlenpaare (x_i, y_i)

$$(0,4), \quad (1,1), \quad (2,0), \quad (3,1), \quad (4,4)$$

so sieht man, dass sie alle auf der Parabel $y = (x-2)^2$ liegen (die Abhängigkeit der y-Werte von den x-Werte wird also durch eine quadratische Funktion beschrieben), dagegen ist hier $r_{xy} = 0$, die Regressionsgerade ist $y = 2$. Das Ergebnis $r_{xy} = 0$ besagt eben, dass hier **keine lineare Korrelation** vorliegt. Betrachtet man dagegen nur den linken Ast der Parabel, also etwa die drei Zahlenpaare $(0,4)$, $(1,1)$, $(2,0)$, so erhält man die Regressionsgerade $y = \frac{11}{3} - 2x$ und die Korrelation ist $r_{xy} = -0,96$ (dieser Wert liegt nah bei -1).



3.5. Beweis der Schwarz'schen Ungleichung.

3.6. Die zweite Regressionsgerade.

Seien wieder Zahlenpaare (x_i, y_i) mit $1 \leq i \leq n$ gegeben. Wir haben die Frage gestellt, ob sich die Zahlen y_i als Werte einer linearen Funktion $x_i \mapsto y_i$ schreiben lassen; genauer: wie eine lineare Funktion $f(x) = a + bx$ aussieht, so dass $y_i \approx f(x_i)$ gilt (dabei haben wir die Summe der quadratischen Abweichungen $\sum (f(x_i) - y_i))^2$ minimiert).

Oft ist allerdings gar nicht klar, ob wir die y-Werte als Funktion der x-Werte ansehen wollen, oder umgekehrt, die x-Werte als Funktion der y-Werte. Nun könnte man meinen, dass dies keinen Unterschied macht, denn die Umkehrfunktion f^{-1} einer linearen Funktion f ist wieder linear (und zum Beispiel gilt: ist b die Steigung der Funktion f, so berechnet sich die Steigung von f^{-1} als $\frac{1}{b}$). Es wird sich aber zeigen, dass man im allgemeinen eine ganz andere Gerade erhält, wenn man das lineare Regressionsproblem zur Abhängigkeit der x-Werte von den y-Werten löst.

Wie sieht diese Lösung aus? Hier noch einmal das Problem: *Zu unseren Zahlenpaaren (x_i, y_i) mit $1 \leq i \leq n$ suchen wir eine lineare Funktion $g(y) = a' + b'y$, so dass die Summe der quadratischen Abweichungen $\sum (g(y_i) - x_i)^2$ minimal ist.* Gesucht sind

also die Zahlen a' und b'. Hier werden wir nun vorausetzen, dass die Zahlen y_i nicht alle gleich sind. Die Überlegungen im Abschnitt 1.3 zeigen (man muss ja nur jeweils x durch y ersetzen):

$$b' = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(y_i - \overline{y})^2} \qquad \text{und} \qquad a' = \overline{x} - b'\overline{y}$$

Wie man am Koeffizienten a' abliest, gilt entsprechend $g(\overline{y}) = \overline{x}$. Man nennt die Gerade $g(y) = a' + b'y$ die *zweite Regressionsgerade*. Zeichnen wir sie ins x-y-Koordinatensystem (also mit waagrechter x-Achse und senkrechter y-Achse), so erhalten wir eine Gerade durch den Punkt $(\overline{x}, \overline{y})$ mit Steigung $\frac{1}{b'}$. (Warum $\frac{1}{b'}$? Die Steigung der Geraden im y-x-Koordinatensystem ist b', nun haben wir die Achsen vertauscht, daher ist das Bild, das wir vor Augen haben, das der Umkehrfunktion, und die hat die Steigung $\frac{1}{b'}$.)

Wir setzen nun voraus, dass die Zahlen x_i nicht alle gleich sind, und dass auch die Zahlen y_i nicht alle gleich sind. Dann existieren **beide** Regressionsgeraden und wir sehen: *die beiden Regressionsgeraden gehen durch den Punkt* $(\overline{x}, \overline{y})$ *und haben im* x-y*-Koordinatensystem die Steigungen* b *und* $\frac{1}{b'}$. Nun gilt aber:

$$bb' = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \cdot \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(y_i - \overline{y})^2} = r_{xy}^2.$$

Ist $r_{xy} = 0$, so ist sowohl $b = 0$, als auch $b' = 0$. In diesem Fall ist also die erste Regressionsgerade die Gerade $y = \overline{y}$, und die zweite Regressionsgerade ist die Gerade $x = \overline{x}$: es sind dies also die beiden achsenparallelen Geraden durch $(\overline{x}, \overline{y})$.

Sei nun $r_{xy} \neq 0$. Da bb' eine Quadratzahl ist, müssen b und b' das gleiche Vorzeichen haben. Auch wissen wir: es ist immer $r_{xy}^2 \leq 1$ und es ist $r_{xy}^2 = 1$ nur dann, wenn alle Zahlenpaare (x_i, y_i) auf einer Geraden liegen. Also sehen wir: $0 \leq bb' \leq 1$ und es ist $bb' = 1$ nur dann, wenn alle Zahlenpaare (x_i, y_i) auf einer Geraden liegen. Nun ist aber $bb' = 1$ gleichbedeutend mit $b = \frac{1}{b'}$. Also gilt: *nur dann fallen die beiden Regressionsgeraden zusammen, wenn alle Zahlenpaare auf einer Geraden liegen!*

Wenn nun die beiden Regressionsgeraden nicht übereinstimmen, so haben immerhin b, b' beide das gleiche Vorzeichen, die Regressionsgeraden sind also entweder beide steigend oder beide fallend. Wenn die Regressionsgeraden beide steigen (also wenn $b > 0$ und $b' > 0$ gilt), so ist $b \leq \frac{1}{b'}$; dies bedeutet: *die zweite Regressionsgerade ist steiler als die erste.* Wenn dagegen die Regressionsgeraden beide fallen (also wenn $b < 0$ und $b' < 0$ gilt), so ist $b \geq \frac{1}{b'}$; aber auch dies bedeutet (denn nun sind ja b und b' negativ): *die zweite Regressionsgerade ist steiler als die erste.*

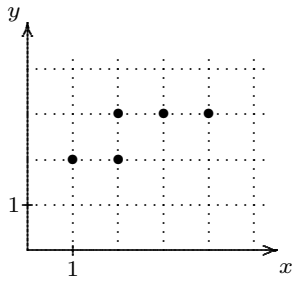
Zwei Beispiele:

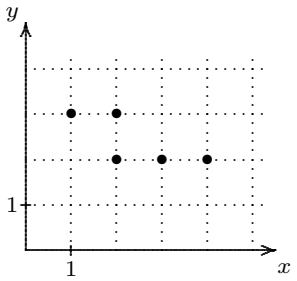
Daten

x_i	1	2	2	3	4
y_i	2	2	3	3	3

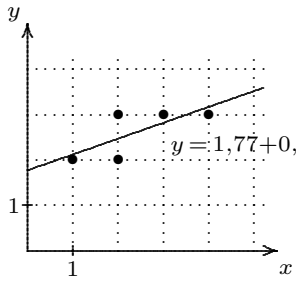
x_i	1	2	2	3	4
y_i	3	3	2	2	2

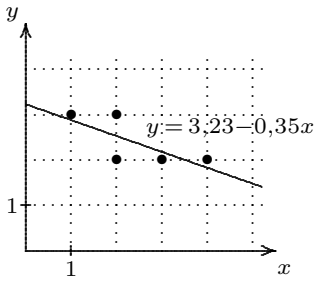
Punktwolke



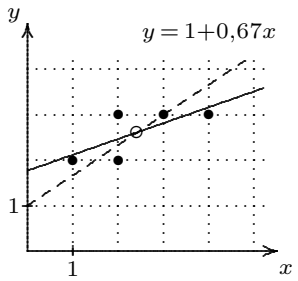


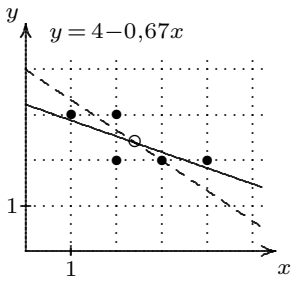
Erste Regressionsgerade





Zusätzlich, gestrichelt, die zweite Regressionsgerade (Schnittpunkt $\circ = (\overline{x}, \overline{y})$)





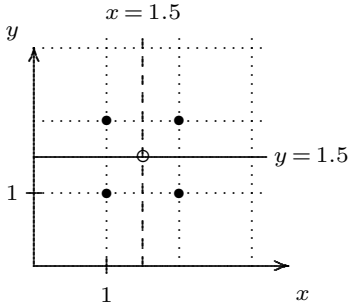
Korrelations-Koeffizient $r_{xy} \approx 0,72$ $r_{xy} \approx 0,72$

Zum Abschluss noch ein **Beispiel mit** $r_{xy} = 0$.

Daten

x_i	1	1	2	2
y_i	1	2	1	2

Regressionsgeraden



die zweite Regressionsgerade ist wieder gestrichelt gezeichnet.

Ausblick: Wir sind in diesem Abschnitt davon ausgegangen, dass n Zahlenpaare (x_i, y_i), also n Punkte in der Ebene $\mathbb{R}^2$ gegeben sind.
Oft haben wir mit den x-Werten $(x_1, x_2, \ldots, x_n)$ oder den y-Werten $(y_1, y_2, \ldots, y_n)$ gearbeitet. In der Mathematik nennt man die Menge derartiger n-Tupel

$$(x_1, x_2, \ldots, x_n)$$

den **n-dimensionalen reellen Vektorraum** $\mathbb{R}^n$. Die Fälle $n = 2$ und $n = 3$ kennt man ja aus der Schule, die Fälle $n \geq 4$ erscheinen zu Beginn vielleicht merkwürdig, es stellt sich aber heraus, dass das Arbeiten in derartigen allgemeinen "Vektorräumen" sehr praktisch ist! Beliebig lange n-Tupel treten bei vielen Messreihen auf; dass man diese n-Tupel "Vektoren" nennt, und sie damit als algebraische oder geometrische Objekte auffasst, soll niemanden stören. Worum es geht, ist folgendes: Man will mit solchen n-Tupeln algebraisch arbeiten, zum Beispiel skalare Vielfache bilden, oder Datensätze addieren oder subtrahieren: dazu ist es gut, die (algebraischen) Regeln der Vektoraddition zu kennen. Zur Interpretation solcher Datensätze ist es ebenfalls hilfreich, die geometrische Intuition, so wie man sie von der Ebene $\mathbb{R}^2$ und dem Raum $\mathbb{R}^3$ her kennt (wo man von Längen und von Winkeln spricht) auf den allgemeinen $\mathbb{R}^n$ zu übertragen. Man erhält auf diese Weise eine **geometrische Interpretation des Korrelations-Koeffizienten:** Seien reelle Zahlen $x_1, \ldots, x_n$ mit Mittelwert $\overline{x}$ und entsprechend reelle Zahlen $y_1, \ldots, y_n$ mit Mittelwert $\overline{y}$ gegeben. Der lineare Korrelations-Koeffizient r_{xy} ist nichts anderes als $\cos\phi$, wobei ϕ der "Winkel" zwischen den Vektoren

$$(x_1, \ldots, x_n) - (\overline{x}, \ldots, \overline{x}) \text{ und } (y_1, \ldots, y_n) - (\overline{y}, \ldots, \overline{y})$$

ist.