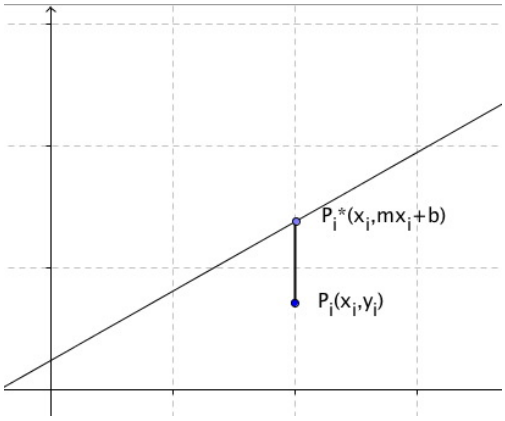
8. Lineare Regression

8.1. Die Methode der kleinsten Quadrate

Regressionsgeraden bzw. Ausgleichsgeraden sind eine Auswertung von statistischen Messdaten. Dabei sind *n* Datenpunkte $(x_1,y_1),(x_2,y_2),\ldots(x_n,y_n)$ gegeben. Es soll nun eine Gerade gefunden werden, die am besten durch die Datenpunkte verläuft. Ziel dieser Analyse ist es, Beziehungen zwischen den beiden Merkmalen festzustellen. Zusammenhänge können dann quantitativ beschrieben und prognostiziert werden. Theoretisch sind verschiedene Methoden denkbar, eine Ausgleichsgerade zu einer Menge von Messpunkten zu definieren. Die *Methode der kleinsten Quadrate*, seltener auch Kriterium der kleinsten Quadrate, wurde von Gauß entwickelt und erfolgreich angewendet. Sie hat sich als wesentliches Verfahren durchgesetzt.

Für die Ausgleichsgerade wird die Funktionsgleichung $y = mx + b$ gesucht, also die beiden Parameter *m* und *b*. Zu jedem Datenpunkt $P(x_i,y_i)$ können wir mit dem x-Wert den Punkt P* bestimmen, der auf der gesuchten Geraden liegt. Er hat die y-Koordinate $y_i^* = mx_i + b$. Somit erhalten wir zu jedem Datenpunkt den Fehler $y_i - y_i^*$.



$y_1-(mx_1+b)=$ Fehler 1

$y_2-(mx_2+b)=$ Fehler 2

…

$$F(m,b)=\sum_{i=1}^{n}\left(y_i-(mx_i+b)\right)^2$$

Die Summe der quadratischen Fehler ist dann:

$$F(m,b)=\sum_{i=1}^{n}\left(y_i-(mx_i+b)\right)^2$$ und dieser soll minimiert werden!

Wir lösen das Quadrat in der Summe mit der binomischen Formel auf.

$$F(m,b)=\sum_{i=1}^{n}\left(y_i^2-2y_i(mx_i+b)+(mx_i+b)^2\right)$$

$$=\sum_{i=1}^{n}\left(y_i^2-2mx_iy_i-2by_i+m^2x_i^2+2mbx_i+b^2\right)$$

Wir arbeiten durch Umformungen die beiden Parameter b und m heraus.

$$F(m,b)=\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}2mx_iy_i-\sum_{i=1}^{n}2by_i+\sum_{i=1}^{n}m^2x_i^2+\sum_{i=1}^{n}2mbx_i+\sum_{i=1}^{n}b^2$$

$$=\sum_{i=1}^{n}y_i^2-2m\sum_{i=1}^{n}x_iy_i-2b\sum_{i=1}^{n}y_i+m^2\sum_{i=1}^{n}x_i^2+2mb\sum_{i=1}^{n}x_i+b^2\sum_{i=1}^{n}1$$

Dieser Term hängt von zwei Variablen, m und b ab und wir wollen dazu das Minimum bestimmen. Dazu bildet man die partiellen Ableitungen, bei denen nach einer Variablen abgeleitet wird, die andere wird dabei als Konstante betrachtet.
Partielle Ableitung von $F(m,b)$ nach *m*:

$\frac{\partial}{\partial m}F(m,b) = 0 - 2\sum_{i=1}^{n} x_i y_i - 0 + 2m\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i + 0$ und nach b

$\frac{\partial}{\partial b}F(m,b) = 0 - 0 - 2\sum_{i=1}^{n} y_i + 0 + 2m\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} 1$

Im Minimum ist die erste Ableitung Null. Daher bestimmen wir die Nullstellen bezüglich m und b dieser beiden Ableitungen.

$2\sum_{i=1}^{n} x_i y_i + 2m\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i = 0$ und $2\sum_{i=1}^{n} y_i + 2m\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} 1 = 0$

Daraus folgt:

$$\begin{cases} m\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ m\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Wir erhalten 2 Gleichungen mit 2 Unbekannten *m* und *b*! Um diese zu lösen benutzen wir folgendes Bezeichnungssystem:

$\sum x_i^2 = A \quad \sum x_i = B \quad \sum y_i = C \quad \sum x_i y_i = D$

$$\Rightarrow \begin{cases} m \cdot A + b \cdot B = D \\ m \cdot B + n \cdot b = C \end{cases}$$

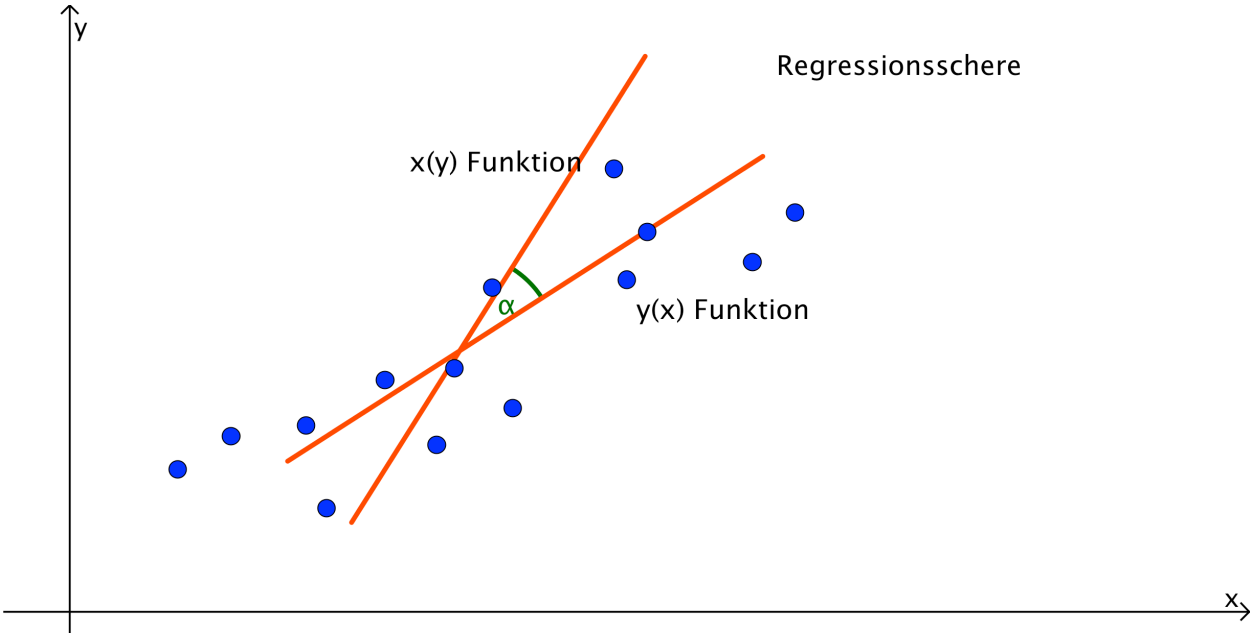
Ergebnisse:

$\Rightarrow m = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$ und $b = \frac{1}{n}\left(\sum y_i - m\sum x_i\right)$

Führt man für *x* und *y* die Mittelwerte ein $\bar{x} = \frac{1}{n}\sum x_i$ und $\bar{y} = \frac{1}{n}\sum y_i$, so kann man in der Formel für *m* die Summen über x und y ersetzen: $\sum x_i = n\bar{x}$ und $\sum y_i = n\bar{y}$. Damit erhält man $m = \frac{n\sum x_i y_i - n\bar{x} \cdot n\bar{y}}{n\sum x_i^2 - (n\bar{x})^2} = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2}$. Die Formel für b wird dann $b = \bar{y} - m\bar{x}$. Führt man noch die Abkürzungen $S_{xy} = \sum x_i y_i - n\overline{xy}$ und $S_{xx} = \sum x_i^2 - n\bar{x}^2$ ein, so erhält man:

$$\boxed{m = \frac{S_{xy}}{S_{xx}} \text{ und } b = \bar{y} - m\bar{x}}$$

8.2. Korrelationskoeffizienten



(Abb. 1)

In den bisherigen Betrachtungen wurde von einer Punktwolke ausgegangen, durch die man die Regressionsgeraden legen kann. Dabei hat die Regressionsgerade **bezüglich x** $(y(x))$ die **Steigung a_1** und die Regressionsgerade **bezüglich y** $(x(y))$ die Steigung **a_2.**
Man sieht (Abb. 1), dass die Größe der Steigungen a_1 und a_2 ein Maßstab für die Stärke des Zusammenhangs zwischen den beiden Variablen x und y darstellt (die Steigung könnte auch weiterhin mit m betitelt werden, ich habe einfachheitshalber darauf verzichtet).

Wäre der Zusammenhang streng linear, wie dies z.B. für die beiden angegebenen Funktionen y(x) und x(y) der Fall ist – beide Funktionen haben den gleichen Graphen, sie sind identisch – so ist das Produkt der Steigungen a_1 und a_2 gleich eins (**Beispiel 1**).

Beispiel 1:

$y(x) = 0,5x + n \Rightarrow x(y) = 2y - 2n$

$a_1 = 2; \quad a_2 = 0,5; \quad a_1 \cdot a_2 = 1$

Beispiel 2:

$a_1 = -2,65; \quad a_2 = -0,36;$

$a_1 \cdot a_2 = -2,65 \cdot (-0,36) = 0,954 \approx 95\%$

Je stärker der Zusammenhang zwischen den Merkmalen, desto enger rücken die Punkte der Punktwolke zusammen und desto kleiner wird der Winkel α zwischen den beiden Regressionsfunktionen.
Ein wichtiges Maß für die Stärke des Zusammenhangs ist das Produkt $a_1 \cdot a_2$.
Dieses Maß wird mit r^2 bezeichnet und heißt **Bestimmtheitsmaß** $\left[r^2 = a_1 \cdot a_2 \right]$**.**
Dieses Maß gibt an, wie viel Prozent der Veränderung der y-Werte auf Einflüsse der x-Werte zurückzuführen sind. Das sind im obigen ca. 95% (**Beispiel 2**).
Wichtiger als das Bestimmtheitsmaß ist der Korrelationskoeffizient **r**. Er ist die Wurzel aus dem Bestimmtheitsmaß $\left[r = \sqrt{a_1 \cdot a_2} \right]$.

Da sich die Steigungen a_1 und a_2 jeweils durch andere Terme (die Summen aus vorigem Kapitel) ersetzen lassen, ergibt sich folgender Satz:
Sind *n* Paare $\left((x_1,y_1),(x_2,y_2),\ldots(x_n,y_n)\right)$ von Merkmalswerten gegeben, dann berechnet sich der lineare Korrelationskoeffizient $r=\sqrt{a_1\cdot a_2}$ nach $r=\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$.

Mit den oben eingeführten Abkürzungen
$S_{xy}=\sum x_i y_i-n\bar{x}\bar{y}$ und $S_{xx}=\sum x_i^2-n\bar{x}^2$ und der noch fehlenden $S_{yy}=\sum y_i^2-n\bar{y}^2$ lässt sich der Korrelationskoeffizient berechnen durch $r=\dfrac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}$.

Anmerkungen:

- Für den Korrelationskoeffizienten lassen sich folgende Fälle unterscheiden:
  1. $r>0$ *steigende* Regressionsgerade, **2.** $r<0$ *fallende* Regressionsgerade

- Für die Bewertung der Korrelation gilt folgende Tabelle:

r	0	(0 ; 0.3)	(0.3 ; 0.7)	(0.7 ; 1)	1
Korrelation	keine	schwache	mittlere	starke	volle

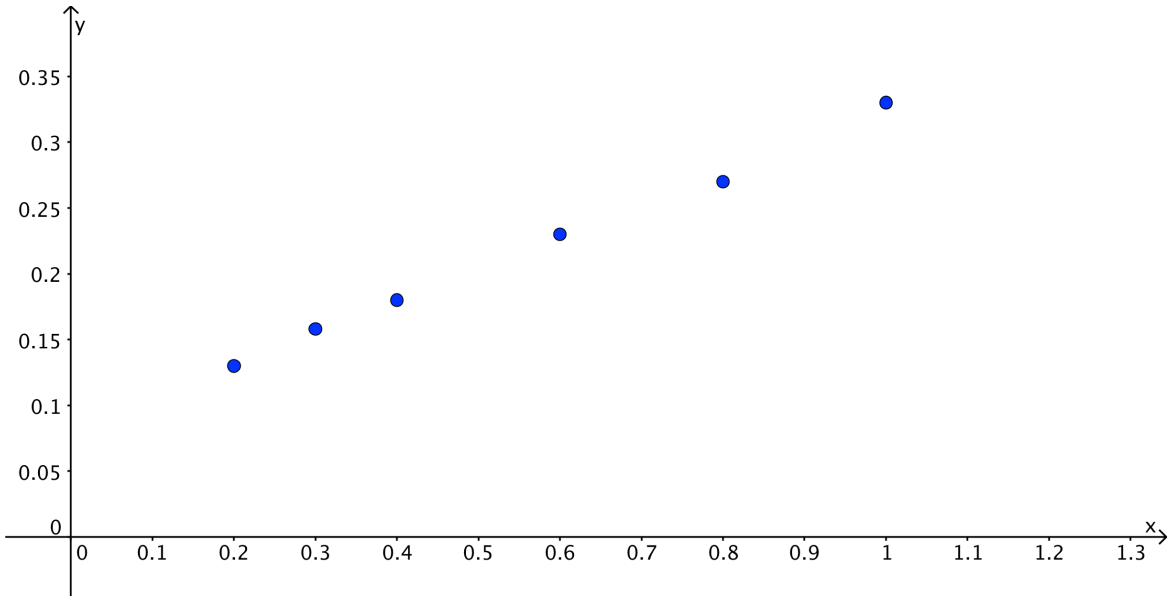
Wir schauen uns all diese Zusammenhänge an einem ausführlichen Beispiel an:
Bei einer landesweit durchgeführten Polizeikontrolle wurde die Reaktionsfähigkeit von Menschen, abhängig vom Alkoholgehalt in ihrem Blut, untersucht.

Alkoholgehalt in ‰	0,2	0,3	0,4	0,6	0,8	1,0
Reaktionszeit in s	0,13	0,158	0,18	0,23	0,27	0,33

a.) Zeichnen Sie die Wertepaare in ein Diagramm!
b.) Bestimmen Sie die lineare Korrelation!
c.) Ermitteln Sie die Regressionsgerade bezüglich *x* und zeichnen Sie sie in das Diagramm unter a.)!

Lösung

a) Der näherungsweise lineare Zusammenhang ist in etwa an den Messpunkten zu erkennen.



b) Man berechne die Mittelwerte $\overline{x}$ und $\overline{y}$ und verwendet zur Bestimmung der linearen Korrelation die oben hergeleitete Formel:

$$\overline{x} = \frac{0{,}2+0{,}3+\ldots+1{,}0}{6} = 0{,}55\ ; \quad \overline{y} = \frac{0{,}13+0{,}158+\ldots+0{,}33}{6} = 0{,}216$$

Zur besseren Übersicht führt man die Rechnung am Besten mit Hilfe einer Tabelle durch:

	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	-0,35	-0,086	0,0301	0,1225	0,007396
2	-0,25	-0,058	0,0145	0,0625	0,003364
3	-0,15	-0,036	0,0054	0,0225	0,001296
4	0,05	0,014	0,0007	0,0025	0,000196
5	0,25	0,054	0,0135	0,0625	0,002916
6	0,45	0,114	0,0513	0,2025	0,012996
$\sum$	0	0,002	0,1155	0,475	0,028164

Der Korrelationskoeffizient zeigt eine **starke** Korrelation zwischen dem Alkoholgehalt im Blut und der Reaktionsfähigkeit.

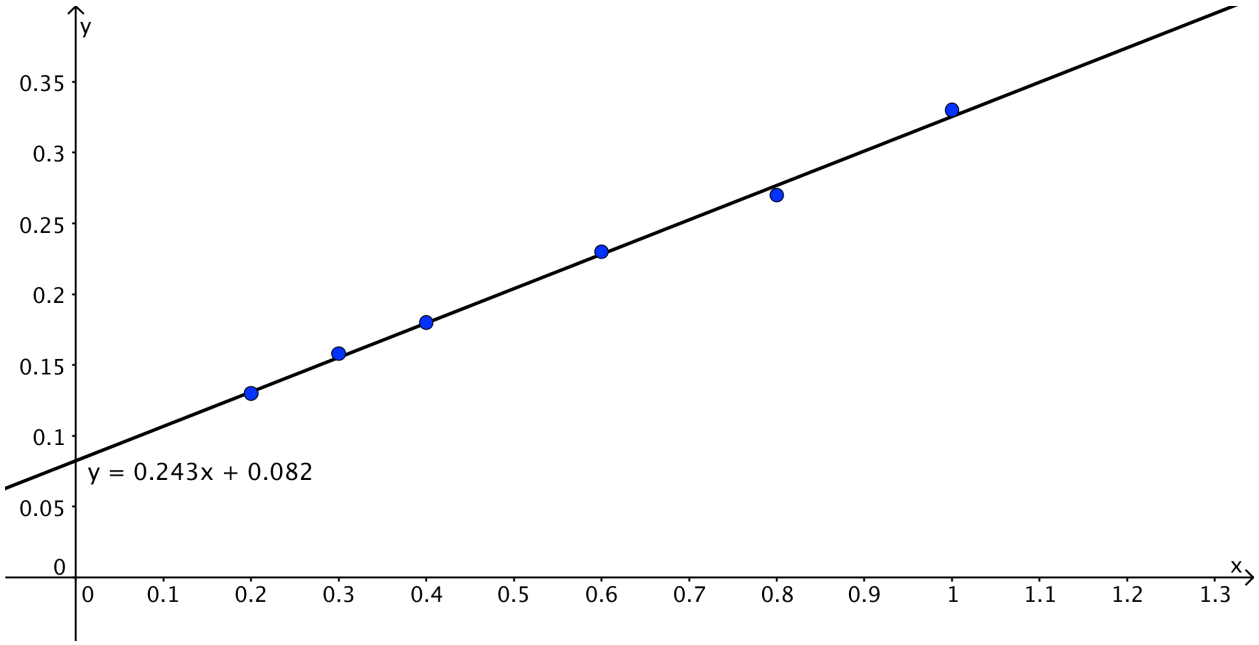
$$r = \frac{0{,}1155}{\sqrt{0{,}475 \cdot 0{,}02814}} = 0{,}999$$

c) Zur Berechnung von **m** und **b** werden die oben hergeleiteten Formeln verwendet!

$m = 0{,}243$

$b = 0{,}08235$

Die Regressionsgerade bezüglich x hat die Gleichung: $y = 0{,}243x + 0{,}08235$


y
x
y = 0.243x + 0.082
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3