

8. Lineare Regression

8.1. Die Methode der kleinsten Quadrate

Regressionsgeraden bzw. Ausgleichsgeraden sind eine Auswertung von statistischen Messdaten. Ziel dieses Verfahrens ist es, Beziehungen zwischen zwei Merkmalen festzustellen. Zusammenhänge können dann quantitativ beschrieben und prognostiziert werden.

Von den vielen Kriterien zur Bestimmung einer optimalen Ausgleichsgerade bzw. Regressionsfunktion hat sich die *Methode der kleinsten Quadrate*, seltener auch Kriterium der kleinsten Quadrate, durchgesetzt.

Dieses Kriterium baut auf folgende Gedanken auf:

- n Datenpunkte sind gegeben $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- es soll nun eine Gerade gefunden werden, die am besten durch die Datenpunkte verläuft

Zunächst nehmen wir an, dass die Gerade durch auf jeden Fall durch $(0,0)$ verläuft und somit durch $y = mx$ beschrieben wird. Für einen Punkt (x_i, y_i) , der genau auf der Geraden liegt, gilt: .

Da wir aber annehmen, dass keiner der Punkte genau auf der Geraden liegt, erhalten wir bei jedem Punkte einen Fehler $y_i - mx_i = f_i$. Nach der Summe der kleinsten

Fehlerquadrate müssen diese Fehler aufsummiert und soweit minimiert werden wie nur möglich: $\sum_{i=1}^n \underbrace{(mx_i - y_i)^2}_{\text{der Fehler wurde quadriert}} = F(m)$. Jetzt kommt die Aufgabe, dass $F(m)$ minimiert werden soll!

Finde das Minimum der Funktion $F(m) = \sum_{i=1}^n (mx_i - y_i)^2$

Wir leiten die Funktion ab und setzen die Ableitung Null: $F'(m) = 0$

$$F(m) = \sum_{i=1}^n (m^2 \cdot x_i^2 - 2mx_i y_i + y_i^2)$$

$$F'(m) = \sum_{i=1}^n (2mx_i - 2x_i y_i)$$

$$F'(m) = 0 = \sum_{i=1}^n (2mx_i - 2x_i y_i)$$

$$= \sum_{i=1}^n 2mx_i^2 - \sum_{i=1}^n 2x_i y_i$$

$$= 2m \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i y_i$$

$$\Rightarrow m = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Jetzt kann mit der Steigung weitergerechnet werden!
 $y = mx$

Bisher haben wir die Betrachtungen darauf beschränkt, dass die gesuchte Gerade durch den Ursprung (0,0) verläuft.

Allgemein gilt aber $y = mx + b$. Hier sind nun 2 Unbekannte vorzufinden, m und b .

$$y = mx + b$$

Falls der Punkt (x_1, y_1) nicht auf der Geraden liegt, gilt

$$y_1 - (mx_1 + b) = \text{Fehler 1}$$

$$y_2 - (mx_2 + b) = \text{Fehler 2}$$

...

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

Die Summe der quadratischen Fehler ist dann:

$$F(m, b) = \sum_{i=1}^n \left(\underbrace{y_i}_a - \underbrace{(mx_i + b)}_b \right)^2 \rightarrow \text{soll minimiert werden!}$$

umgeschrieben durch $(a - b)^2$:

$$\begin{aligned} F(m, b) &= \sum_{i=1}^n (y_i^2 - 2y_i(mx_i + b) + (mx_i + b)^2) \\ &= \sum_{i=1}^n (y_i^2 - 2mx_i y_i - 2by_i + m^2 x_i^2 + 2mbx_i + b^2) \end{aligned}$$

Da es wie gesagt 2 Unbekannte sind, müssen die beiden Unbekannten jeweils **einzeln** betrachtet werden und es muss die Ableitung gebildet werden.

Ableitung von $F(m, b)$ bezogen auf m :

$$F'_m(m, b) = \sum_{i=1}^n (-2x_i y_i + 2mx_i^2 + 2bx_i) = 0$$

Im Minimum ist die erste Ableitung Null.

Ableitung von $F(m, b)$ bezogen auf b :

$$F'_b(m, b) = \sum_{i=1}^n (-2y_i + 2mx_i + 2b) = 0$$

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:

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

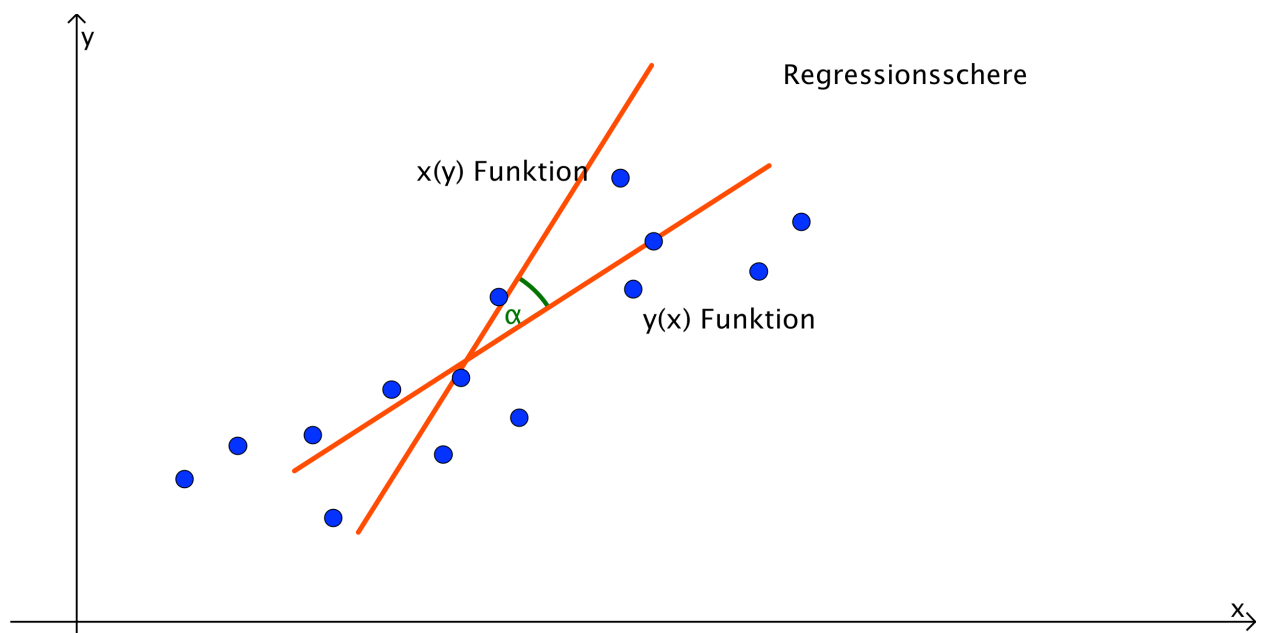
Führt man für beide Größen x und y die Mittelwerte ein:

$$\bar{x} = \frac{1}{n} \sum x_i \quad \text{und} \quad \bar{y} = \frac{1}{n} \sum y_i$$

so ergeben sich für m und b die Formeln:

$$S_{xy} = \sum x_i y_i - n \bar{x} \bar{y} \quad \text{und} \quad S_{xx} = \sum x_i^2 - n \bar{x}^2, \quad m = \frac{S_{xy}}{S_{xx}} \quad \text{und} \quad b = \bar{y} - m \bar{x}$$

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** $[r^2 = a_1 \cdot a_2]$.

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 $[r = \sqrt{a_1 \cdot a_2}]$.

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 $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ von Merkmalswerten gegeben, dann berechnet sich

der lineare Korrelationskoeffizient $r = \sqrt{a_1 \cdot a_2}$ nach
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Mit den oben eingeführten Abkürzungen

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} \quad \text{und} \quad S_{xx} = \sum x_i^2 - n\bar{x}^2 \quad \text{und der noch fehlenden} \quad S_{yy} = \sum y_i^2 - n\bar{y}^2$$
 lässt sich

der Korrelationskoeffizient berechnen durch
$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}.$$

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.

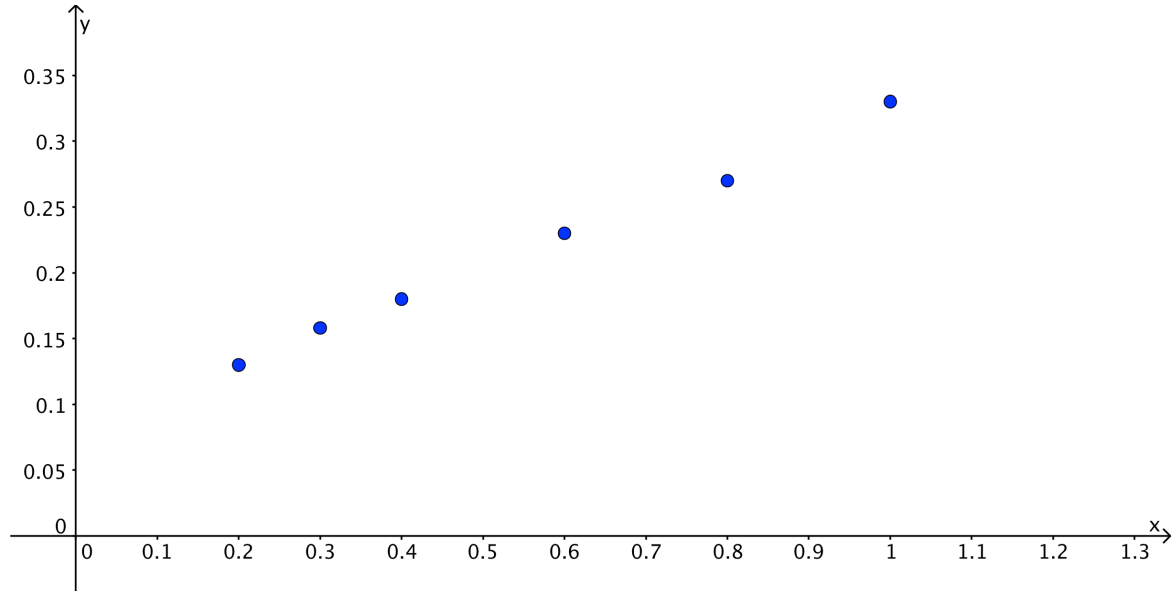
Alkoholgehalt in ‰	0,2	0,3	0,4	0,6	0,8	1,0
Reaktionszeit in s	0,13	0,158	0,18	0,23	0,27	0,33

- Zeichnen Sie die Wertepaare in ein Diagramm!
- 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 $\bar{x}$ und $\bar{y}$ und verwendet zur Bestimmung der linearen Korrelation die oben hergeleitete Formel:

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

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

	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
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
Σ	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: $y = 0,243x + 0,08235$

