

Mathematik I für Biologen, Geowissenschaftler und Geoökologen
Lineare Regression

Stefan Keppeler

21. Januar 2013



Problemstellung

Beispiel

Lineare Regression

Maß für Abweichung

Trick

Berechnung

Minimum?

"Kochrezept"

Anhang: Regression anderer Zusammenhänge

Exponentialfunktionen

Potenzfunktionen



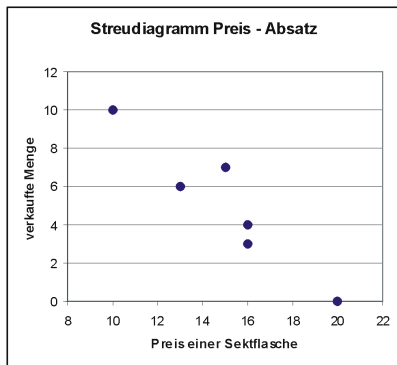
Problemstellung:

- ▶ Gegeben seien n Punkte $(x_i, y_i) \in \mathbb{R}^2$, $i \in \{1, \dots, n\}$, $n \geq 2$, die näherungsweise auf einer Geraden liegen.
- ▶ Man bestimme die Gerade, die am nächsten an diesen Punkten liegt.
- ▶ Diese Gerade heißt **Ausgleichsgerade** oder **Regressionsgerade**. Die Bestimmung dieser Geraden heißt **lineare Regression**.



Quelle: de.wikipedia.org/wiki/Regressionsanalyse

Eine renommierte Sektkellerei möchte einen hochwertigen Rieslingsekt auf den Markt bringen. Für die Festlegung des Abgabepreises soll zunächst eine Preis-Absatz-Funktion ermittelt werden. Dazu wurde in $n = 6$ Geschäften ein Testverkauf durchgeführt. Man erhielt sechs Wertepaare mit dem Ladenpreis x (in Euro) einer Flasche und die verkaufte Menge y an Flaschen:



Laden i	1	2	3	4	5	6
Preis einer Flasche x_i	20	16	15	16	13	10
verkaufte Menge y_i	0	3	7	4	6	10

- ▶ Wir benötigen: Maß für die **Abweichung** D einer Geraden g von den gegebenen Punkten.
- ▶ Suche dann g so, dass die Abweichung minimiert wird. Dabei ist g der Graph der Funktion

$$g(x) = mx + b.$$

- ▶ Die Gerade wird durch **2 Parameter**, bestimmt: $m, b \in \mathbb{R}$, d.h. D ist eine Funktion von b und m .

Wie ist D sinnvoll zu wählen?



Wie ist D sinnvoll zu wählen?

x_1	x_2	\dots	x_n
y_1	y_2	\dots	y_n
$g(x_1)$	$g(x_2)$	\dots	$g(x_n)$

- ▶ D soll messen, wie nahe
der Vektor $\vec{v} \in \mathbb{R}^n$, mit Einträgen $v_i = g(x_i)$,
beim Vektor $\vec{y} \in \mathbb{R}^n$, mit Einträgen y_i , liegt.
- ▶ Betrachte den Abstand der beiden im \mathbb{R}^n ,

$$D(b, m) = d(\vec{v}, \vec{y}) = \|\vec{v} - \vec{y}\| = \sqrt{\sum_{i=1}^n (v_i - y_i)^2}.$$



Aufgabe: Finde b und m , die D minimieren!
Fordere also $\nabla D = 0$.



Trick: Statt $D(b, m)$ minimieren wir $f(b, m) = D(b, m)^2$, denn

$$f \text{ minimal} \quad \Leftrightarrow \quad D \text{ minimal},$$

da $f = h \circ D$ mit $h(x) = x^2$ streng monoton wachsend auf $[0, \infty)$ – d.h. wenn $D(b', m') < D(b, m)$, dann auch $f(b', m') < f(b, m)$.

Wir minimieren also

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



Daher auch Bezeichnung
Methode der kleinsten (Fehler-)Quadrate;
geht auf C.F. Gauß (1777-1855) zurück.



Gradient von $f = \sum_{i=1}^n (mx_i + b - y_i)^2$: $\nabla f = \left(\frac{\partial f}{\partial b}, \frac{\partial f}{\partial m} \right)$

$$\frac{\partial f}{\partial b} = \sum_{i=1}^n 2(mx_i + b - y_i)$$

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

Am Minimum ist $\nabla f = 0$, also

$$n b + \left(\sum_i x_i \right) m = \sum_i y_i \quad (1)$$

$$\left(\sum_i x_i \right) b + \left(\sum_i x_i^2 \right) m = \sum_i x_i y_i \quad (2)$$

Lineares Gleichungssystem mit 2 Gleichungen
 für die 2 Unbekannten b und m .



Definiere

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

und schreibe LGS in Kurzform

$$\left(\begin{array}{cc|c} n & n\bar{x} & n\bar{y} \\ n\bar{x} & \sum_i x_i^2 & \sum_i x_i y_i \end{array} \right) \begin{array}{l} \left[\begin{array}{l} -\bar{x} \\ + \end{array} \right] \cdot 1/n \end{array}$$

$$\left(\begin{array}{cc|c} 1 & \bar{x} & \bar{y} \\ 0 & \sum_i x_i^2 - n\bar{x}^2 & \sum_i x_i y_i - n\bar{x}\bar{y} \end{array} \right)$$

$$\begin{aligned} \sum_i x_i y_i - n\bar{x}\bar{y} &= \text{pencil} \\ &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

...und mit $y \rightarrow x$ auch

$$\sum_i x_i^2 - n\bar{x}^2 = \sum_i (x_i - \bar{x})^2$$




Damit lesen wir die **eindeutige Lösung** ab,

$$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2},$$

$$b = \bar{y} - m\bar{x}.$$

Bemerkungen:

- ▶ Nenner von m ungleich 0? 
- ▶ Mit $g(x) = mx + b$ bedeutet die 2. Gleichung: $g(\bar{x}) = \bar{y}$.

Noch zu klären:

- ▶ Liegt an dieser Stelle ein **Minimum** von f ?
 (also weder Maximum noch Sattel)
- ▶ Untersuche Hesse-Matrix $H = f''$.



Hesse-Matrix:


$$f(m, b) = \sum_{i=1}^n (mx_i + b - y_i)^2, \quad H = f'' = \begin{pmatrix} \frac{\partial^2 f}{\partial b^2} & \frac{\partial^2 f}{\partial b \partial m} \\ \frac{\partial^2 f}{\partial m \partial b} & \frac{\partial^2 f}{\partial m^2} \end{pmatrix},$$

wobei

$$\frac{\partial^2 f}{\partial b^2} = \frac{\partial}{\partial b} \sum_i 2(mx_i + b - y_i) = 2n,$$

$$\frac{\partial^2 f}{\partial m \partial b} = \frac{\partial}{\partial m} \sum_i 2(mx_i + b - y_i) = \sum_i 2x_i = 2n\bar{x} = \frac{\partial^2 f}{\partial b \partial m},$$

$$\frac{\partial^2 f}{\partial m^2} = \frac{\partial}{\partial m} \sum_i 2(mx_i + b - y_i)x_i = 2 \sum_i x_i^2,$$

d.h. $H = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2 \sum_i x_i^2 \end{pmatrix}$. Positiv definit? 



- ▶ Datenpunkte $(x_i, y_i) \in \mathbb{R}^2$, $i \in \{1, \dots, n\}$
- ▶ Berechne die **Mittelwerte**


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

- ▶ Bestimme die **Steigung** der Regressionsgeraden,

$$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

- ▶ Bestimme den **Achsenabschnitt** der Regressionsgeraden

$$b = \bar{y} - m\bar{x}.$$

Beispiel Sektpreise:  oder MATLAB



Regression von Exponentialfunktionen:

Vermutet wird der Zusammenhang

$$z = ce^{\lambda x}$$

zwischen den Größen x und z , und aus Messwerten für (x_i, z_i) sollen die Konstanten $c, \lambda \in \mathbb{R}$ geschätzt werden.

Führe durch **Logarithmieren**¹ zurück auf lineare Regression:

$$y = \log z = \log c + \lambda x$$

Lineare Regression mit Daten $(x_i, y_i) = (x_i, \log z_i)$

liefert Ausgleichsgerade $y = mx + b$.

Daraus erhalten wir $c = e^b$, $\lambda = m$.

¹vgl. log-Plot, Vorlesung 5



Regression von Potenzfunktionen:

Vermutet wird der Zusammenhang

$$p = \alpha q^\beta$$

zwischen den Größen q und p , und aus Messwerten für q und p sollen die Konstanten $\alpha, \beta \in \mathbb{R}$ geschätzt werden.

Führe durch **Logarithmieren**² zurück auf lineare Regression:

$$y = \log p = \log \alpha + \beta \log q = \log \alpha + \beta x$$

Lineare Regression mit Daten $(x_i, y_i) = (\log q_i, \log p_i)$

liefert Ausgleichsgerade $y = mx + b$.

Daraus erhalten wir $\alpha = e^b$, $\beta = m$.

²vgl. log-log-Plot, Vorlesung 5

