14 Lineare Regression

Problemstellung: Gegeben n Punkte $(x_i, y_i) \in \mathbb{R}^2$, $i \in \{1, \ldots, n\}$, $n \geq 2$, die näherungsweise auf einer Geraden liegen. Man bestimme die Gerade, die am besten durch diese Punkte hindurchführt. Diese Gerade heißt *Ausgleichsgerade* oder *Regressionsgerade*. Die Bestimmung der Regressionsgeraden heißt *lineare Regression*.

Mathematische Präzisierung: Wir geben uns eine Vorschrift, wie die *Abweichung* D einer Geraden g von den gegebenen Punkten quantitativ zu bewerten ist, und suchen dann g so, dass die Abweichung minimiert wird. Dabei ist g der Graph der Funktion

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

(Wir bezeichnen sowohl die Funktion als auch ihren Graphen mit g.) Die Gerade wird auf diese Weise durch 2 Zahlen, 2 *Parameter*, bestimmt: $m, b \in \mathbb{R}$. Daher ist D eine Funktion von m und b. Wie ist D sinnvoll zu wählen?

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

D soll messen, wie gut der Vektor $v \in \mathbb{R}^n$, mit den Einträgen $v_i = g(x_i)$, zusammen passt mit dem Vektor $y \in \mathbb{R}^n$, mit den Einträgen y_i, den gegebenen Werten. Daher betrachtet man den Abstand der beiden in $\mathbb{R}^n$,

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

Das liefert uns das präzise Problem: finde $\tilde{m}, \tilde{b}$, die D minimieren!

Ermittlung der Regressionsgeraden. Jetzt wissen Sie bereits, wie vorzugehen ist: Sie berechnen den Gradienten ∇D und setzen ihn Null. Um die Rechnung zu erleichtern, sei Ihnen aber ein Trick verraten. Eine rechnerische Vereinfachung besteht darin, statt $D(m, b)$ die Funktion $f(m, b) = D(m, b)^2$ zu minimieren. Beide Funktionen D, f haben dasselbe Minimum, weil $f = h \circ D$ und $h(x) = x^2$ eine streng wachsende Funktion auf $[0, \infty)$ ist: wenn $D(m', b') < D(m, b)$, dann $f(m', b') < f(m, b)$. Also minimieren wir

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

Daher heißt diese Methode zur Bestimmung einer Ausgleichsgeraden auch die *Methode der kleinsten Quadrate*; sie geht auf C. F. Gauß (1777-1855) zurück. Um das Minimum zu ermitteln, berechnen wir den Gradienten von f,

$$\nabla f = \Big(\frac{\partial f}{\partial m}, \frac{\partial f}{\partial b} \Big).$$

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

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

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

$$\Big(2\sum_i x_i^2\Big)m + \Big(2\sum_i x_i\Big)b = \sum_i x_i y_i \tag{1}$$

$$\Big(2\sum_i x_i\Big)m + 2nb = \sum_i y_i \tag{2}$$

Dies ist ein lineares Gleichungssystem mit 2 Gleichungen und 2 Unbekannten m, b.

Satz. Sei $n \geq 2$ und seien nicht alle x_i gleich. Dann besitzt das LGS (1), (2) eine eindeutige Lösung (d.h. eine und nur eine Lösung).

Bemerkung. Wenn doch alle x_i gleich sind, liegen offenbar alle gegebenen Punkte auf einer senkrechten Geraden.

Beweis des Satzes. Zunächst ist $\sum_i x_i^2 > 0$, denn Null sein könnte es nur, wenn alle $x_i = 0$ sind, aber nach Voraussetzung sind nicht alle x_i gleich. Also können wir die erste Gleichung durch $2\sum_i x_i^2$ dividieren, und erhalten als Koeffizienten:

$$\begin{array}{cc|c} 1 & \frac{\sum_i x_i}{\sum_i x_i^2} & \frac{\sum_i x_i y_i}{2\sum_i x_i^2} \\ 2\sum_i x_i & 2n & \sum_i y_i \end{array}$$

Dem Gauß-Verfahren folgend eliminieren wir m aus der zweiten Gleichung: $II \to II - (2\sum x_i)I$

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

Wir werden gleich als Lemma (= Hilfssatz) beweisen, dass der Koeffizient vor b, $2n - 2(\sum x_i)^2/\sum x_i^2$, nicht verschwindet (d.h. $\neq 0$ ist). Daraus folgt, dass sich die untere Gleichung nach b auflösen lässt; die obere lässt sich (bei gegebenem b) nach m auflösen.□

Lemma. Sei $n \geq 2$ und seien nicht alle x_i gleich. Dann gilt

$$n - \frac{(\sum x_i)^2}{\sum x_i^2} > 0.$$

Beweis. Sei $\overline{x} = (1/n)\sum_j x_j$ das (arithmetische) Mittel der x_i. Da nicht alle x_i gleich sind, ist $x_i - \overline{x} \neq 0$ für wenigstens ein i. Daher

$$0 < \sum_i (x_i - \overline{x})^2 = \sum_i \Big(x_i^2 - \frac{2x_i}{n}\sum_j x_j + \frac{1}{n^2}\big(\sum_j x_j\big)^2\Big) =$$

$$= \sum_i x_i^2 - \frac{2}{n}\Big(\sum_i x_i\Big)\Big(\sum_j x_j\Big) + \frac{n}{n^2}\Big(\sum_j x_j\Big)^2 =$$

$$= \sum_i x_i^2 - \frac{2}{n}\Big(\sum_i x_i\Big)^2 + \frac{1}{n}\Big(\sum_i x_i\Big)^2 =$$

$$= \sum_i x_i^2 - \frac{1}{n}\Big(\sum_i x_i\Big)^2.$$

Daraus folgt, indem man mit $n/\sum x_i^2$ multipliziert, die Behauptung. □

Sei $(\tilde{m}, \tilde{b})$ die Lösung von (1), (2). Um nachzuweisen, dass es sich um ein *Minimum* (statt Maximum oder Sattelpunkt) von f handelt, untersuchen wir die Hesse-Matrix von f,

$$H = f'' = \begin{pmatrix} \frac{\partial^2 f}{\partial m^2} & \frac{\partial^2 f}{\partial m\,\partial b} \\ \frac{\partial^2 f}{\partial b\,\partial m} & \frac{\partial^2 f}{\partial b^2} \end{pmatrix}.$$

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

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

$$\frac{\partial^2 f}{\partial b\,\partial m} = \frac{\partial^2 f}{\partial m\,\partial b}$$

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

Also

$$H = 2\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}. \tag{3}$$

Dies hängt nicht mehr von m und b ab, weil die Funktion f ein Polynom zweiten Grades ist und für solche Funktionen die zweite Ableitung stets konstant ist.

Lemma. Sei $n \geq 2$ und seien nicht alle x_i gleich. Dann ist die Matrix (3) positiv-definit.

Beweis. Wir betrachten $u^T H u$ für beliebiges $u \in \mathbb{R}^2$, und schreiben $u = (\alpha, \beta)^T$. Dann

$$u^T H u = 2\Big(\alpha^2 \sum_i x_i^2 + 2\alpha\beta \sum_i x_i + \beta^2 n\Big) =$$

$$= 2\sum_i (\alpha^2 x_i^2 + 2\alpha x_i \beta + \beta^2) = 2\sum_i (\alpha x_i + \beta)^2.$$

Dieser Ausdruck ist zunächst ≥ 0. Für welche α, β ist er $= 0$? Nur dann, wenn für jedes i gilt

$$\alpha x_i + \beta = 0. \tag{4}$$

Da aber nicht alle x_i gleich sind, etwa $x_1 \neq x_2$, folgt aus (4), dass $\alpha x_1 + \beta = 0 = \alpha x_2 + \beta \Rightarrow \alpha x_1 = \alpha x_2 \Rightarrow \alpha(x_2 - x_1) = 0 \Rightarrow \alpha = 0$; dann folgt aus (4), dass auch $\beta = 0$. Also gilt für $u \neq 0$, dass $u^T H u > 0$. □

Regression von Exponentialfunktionen. Vermutet wird der Zusammenhang

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

zwischen den Größen x und z, und aus Messwerten für x und z sollen die Konstanten $c, \lambda \in \mathbb{R}$ geschätzt werden. In anderen Worten, durch eine Anzahl n von Punkten $(x_i, z_i) \in \mathbb{R}^2$ soll diejenige Exponentialkurve gelegt werden, die am besten passt. Dieses Problem lässt sich auf die lineare Regression zurückführen, indem man z logarithmiert:

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

Da dies eine Geradengleichung ist, liefert die lineare Regression aus den Punkten $(x_i, \log z_i)$ hierfür eine optimale Gerade $y = mx + b$, und daraus erhalten wir $c = e^b$, $\lambda = m$.

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. Durch doppeltes Logarithmieren, $x = \log q$, $y = \log p$, erhält man die Geradengleichung

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

und kann wieder lineare Regression anwenden.