Mathematik I für Biologen, Geowissenschaftler und Geoökologen

Lineare Regression

Stefan Keppeler

23. Januar 2008

Eberhard Karls
Universität
Tübingen

Problemstellung

Extrema mehrdimensional

Lineare Regression

- Maß für Abweichung
- Trick
- Berechnung
- Minimum?
- Regression anderer Zusammenhänge

Problemstellung:

- Gegeben seien 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 nächsten an diesen Punkte liegt.
- Diese Gerade heißt Ausgleichsgerade oder Regressionsgerade. Die Bestimmung der Regressionsgeraden heißt lineare Regression.

Satz: Hat die differenzierbare Funktion $f : \mathbb{R}^d \to \mathbb{R}$ in $x \in \mathbb{R}^d$ ein (lokales) Minimum oder Maximum, so ist $\nabla f(x) = 0$. Ist umgekehrt

- $\nabla f(x) = 0$ und die Hesse-Matrix $H(x)$ positiv definit, so hat f in x ein lokales Minimum;
- $\nabla f(x) = 0$ und die Hesse-Matrix $H(x)$ negativ definit, so hat f in x ein lokales Maximum.

Definition: Eine Matrix $A \in \mathcal{M}(n, n)$ heißt

- positiv definit, wenn für alle $u \in \mathbb{R}^n$ mit $u \neq 0$ gilt $u^T Au > 0$.
- negativ definit, wenn für alle $u \in \mathbb{R}^n$ mit $u \neq 0$ gilt $u^T Au < 0$.

- 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 Zahlen, 2 Parameter, bestimmt: $m, b \in \mathbb{R}$, d.h. D ist eine Funktion von m und b. 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 $v \in \mathbb{R}^n$, mit Einträgen $v_i = g(x_i)$,
beim Vektor $y \in \mathbb{R}^n$, mit Einträgen y_i, liegt.

$\rightsquigarrow$ Betrachte den Abstand der beiden in $\mathbb{R}^n$,

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

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

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

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

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

Wir minimieren also

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

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

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

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

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

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

$$\left(2\sum_i x_i^2\right)m + \left(2\sum_i x_i\right)b = \sum_i x_i y_i \quad (1)$$

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

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

Satz: Sei $n \geq 2$ und seien nicht alle x_i gleich. Dann besitzt das LGS eine eindeutige Lösung.

Beweis: Zunächst ist $\sum_i x_i^2 > 0$, da nicht alle x_i gleich.

$$\left(\begin{array}{cc|c} 2\sum_i x_i^2 & 2\sum_i x_i & \sum_i x_i y_i \\ 2\sum_i x_i & 2n & \sum_i y_i \end{array}\right) \quad \left| \cdot 1/\left(2\sum_i x_i^2\right)\right.$$

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

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

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

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

D.h. es gibt immer genau ein Paar (m, b) für das

$$\nabla f = \left(\frac{\partial f}{\partial m}, \frac{\partial f}{\partial b} \right) = 0\,.$$

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

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

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

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

$$\frac{\partial^2 f}{\partial b\,\partial m} = \frac{\partial^2 f}{\partial m\,\partial b} = \sum_{i=1}^{n} 2x_i$$

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

$$H = 2\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}$$ hängt nicht von m und b ab.

Lemma: Sei $n \geq 2$ und seien nicht alle x_i gleich.
Dann ist H positiv definit.

Beweis: Betrachte $u^T H u$ für beliebiges $u = (\alpha, \beta)^T \in \mathbb{R}^2$:

$$u^T H u = (\alpha, \beta)\ 2\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

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

$$= 2\sum_i (\alpha x_i + \beta)^2 \quad \geq 0$$

"$= 0$" nur dann, wenn für jedes i gilt $\alpha x_i + \beta = 0$. Widerspruch zu "nicht alle x_i gleich". Also gilt für $u \neq 0$, dass $u^T H u > 0$.

Eberhard Karls Universität Tübingen

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 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 $c = \mathrm{e}^b$, $\lambda = m$.

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

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

und kann wieder lineare Regression anwenden.