Lineare Regression

Was ist der Hintergrund bei der linearen Regression? Betrachten wir ein einfaches Beispiel:

Gegeben sind die Datenpunkte (2 | 1), (3 | 6) und (7 | 2). Gesucht ist eine Gerade (die „Ausgleichs"-

oder „Regressionsgerade" mit der Gleichung $y = m \cdot x + n$), die „gut" zu den Datenpunkten passt.

Was soll „gut" heißen?
Die Datenpunkte stammen häufig aus Messungen, bei denen die x-Werte gut messbar sind und die y-Werte messfehlerbehaftet sein können (Beispiel: Eine Größe y wird zu festen Zeitpunkten x erhoben). Man kann also die x-Werte nicht einfach mit den y-Werten vertauschen.

Zu jedem x-Wert gibt es einerseits den zugehörigen gemessenen y-Wert, andererseits aber auch den durch $y = m \cdot x + n$ berechneten y-Wert:

x-Wert	gemessener y-Wert	berechneter y-Wert	Differenz
2	1	$m \cdot 2 + n$	$m \cdot 2 + n - 1$
3	6	$m \cdot 3 + n$	$m \cdot 3 + n - 6$
7	2	$m \cdot 7 + n$	$m \cdot 7 + n - 2$

Die Werte für m und n sind unbekannt. Beide Werte passen „gut" zu den Datenpunkten, wenn die Summe der Differenzen zwischen den gemessenen und berechneten y-Werten null ergibt:

$$12 \cdot m + 3 \cdot n - 9 = 0$$

bzw.

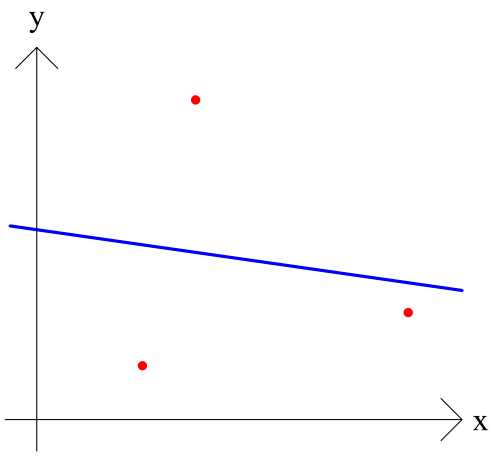
$$n = 3 - 4 \cdot m .$$

Die Regressionsgerade hat also die Gleichung $y = m \cdot (x - 4) + 3$; man sieht, dass sie durch den Schwerpunkt (4 | 3) der Datenpunkte geht.

Nun muss man nur noch m bestimmen. Man bekommt die klassische Regressionsgerade, wenn man fordert, dass die Summe der quadrierten Differenzen möglichst klein ist, dass also

$$\begin{aligned} &(2 \cdot m + 3 - 4 \cdot m - 1)^2 + (3 \cdot m + 3 - 4 \cdot m - 6)^2 + (7 \cdot m + 3 - 4 \cdot m - 2)^2 \\ &= (-2 \cdot m + 2)^2 + (-m - 3)^2 + (3 \cdot m + 1)^2 \\ &= 14 \cdot m^2 + 4 \cdot m + 14 \end{aligned}$$

minimal wird; dies ist für $m = -\frac{1}{7}$ der Fall.

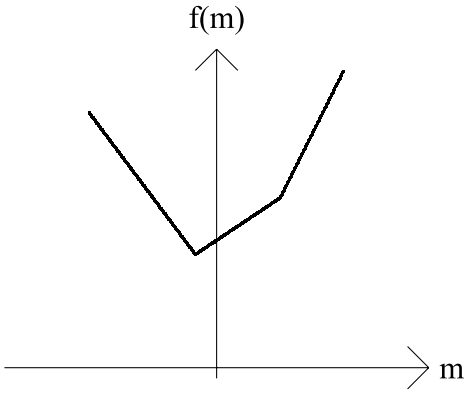


Ersetzt man die Forderung „Summe der quadrierten Differenzen möglichst klein“ durch „Summe der absoluten Differenzen möglichst klein“, bekommt man eine andere Gerade; die Minimierung von

$$f(m) = |2 \cdot m + 3 - 4 \cdot m - 1| + |3 \cdot m + 3 - 4 \cdot m - 6| + |7 \cdot m + 3 - 4 \cdot m - 2|$$
$$= |-2 \cdot m + 2| + |-m - 3| + |3 \cdot m + 1|$$

liefert nämlich

$$m = -\frac{1}{3}.$$



Die Minimierung von $f(m)$ hat zu tun mit dem Median: Schreibt man

$f(m) = 2 \cdot |m - 2| + |m + 3| + 3 \cdot \left|m + \frac{1}{3}\right|$ und erinnert sich daran, dass der Median m von

$x_1,\ x_2,\ ...,\ x_n$ derjenige Wert ist, der $\sum_{i=1}^{n} |x_i - m|$ minimiert, so wird $f(m)$ vom Median von

$2;\ 2;\ -3,\ -\frac{1}{3};\ -\frac{1}{3};\ -\frac{1}{3}$ minimiert.

Verallgemeinerungen, Bezüge zur Vektorgeometrie und einen Einblik in nichtlineare Regression findet man bei

J. Meyer: Vernetzungen zwischen Vektorgeometrie und Beschreibender Statistik. In: Stochastik in der Schule **24** (2); S. 24 - 29 (2004).