Datenanpassung

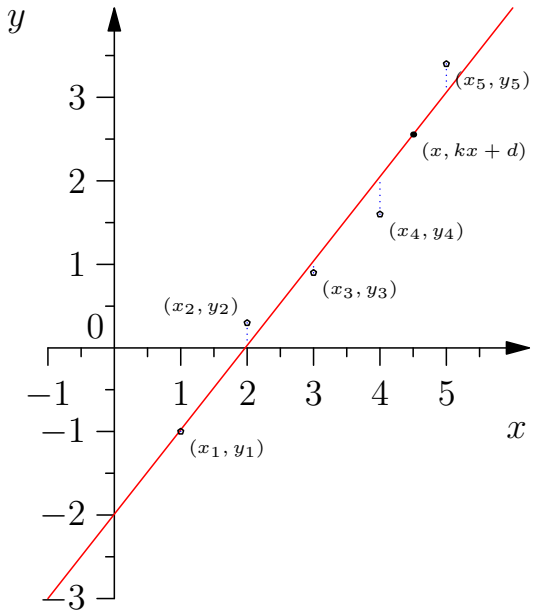
1 Lineare Regression, Datenanpassung mit Hilfe affiner Funktionen $f(x) = kx + d$

Aufgabe: Es seien n Datenpaare (x_1, y_1), (x_2, y_2),..., (x_n, y_n) mit $x_1, y_1, x_2, y_2, \ldots, x_n, y_n \in \mathbb{R}$ gegeben. Gesucht ist dann eine Funktion f der Form $f(x) = kx + d$, eine sogenannte *affine* Funktion, die „bestmöglich“ zu den Daten paßt.

Bestmöglich soll dabei heißen: Man versuche, die reellen Zahlen k und d so zu bestimmen, daß die Summe der quadratischen Abweichungen

$$(y_1 - (kx_1 + d))^2 + (y_2 - (kx_2 + d))^2 + \ldots + (y_n - (kx_n + d))^2$$

kleinstmöglich wird.



Wenn die Daten gut zu zeichnen sind, wenn also alle Daten bei geeigneter Skalierung auf die Zeichenfläche passen, kann man versuchen eine Gerade mit Augenmaß und Gefühl einzupassen. Nimmt man dann zwei Punkte (u_1, v_1) und (u_2, v_2) dieser Geraden, bei denen u_1 und u_2 weit auseinander liegen, so findet man für k den Wert $k = \frac{v_2 - v_1}{u_2 - u_1}$ und dann (z. B.) $d = v_1 - ku_1$.

Es gibt aber auch Formeln (lineare Regression, Gaußsche Methode der kleinsten Quadrate). In diesen Formeln ist es zweckmäßig, mit dem Summenzeichen ($\sum$) zu hantieren.

Wenn $a_1, a_2, a_3, \ldots, a_n, a_{n+1}, \ldots$ reelle Zahlen sind, ist der Ausdruck $\sum_{i=1}^{n} a_i$ eine Kurzschreibweise für die Summe $a_1 + a_2 + \ldots + a_n$. Der *Summationsindex* i ist

frei wählbar. Es ist also $\sum_{i=1}^{n} a_i = \sum_{k=1}^{k} a_k$. Aus der Deutung der Formel ergibt sich $\sum_{i=1}^{1} a_i = a_1$, $\sum_{i=1}^{n+1} a_i = (\sum_{i=1}^{n} a_i) + a_{n+1}$.

Zurück zur Berechnung von k und d. Es seien $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ die *Mittelwerte* der x- bzw. y-Daten. Ferner seien $\overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ und $\overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2$ die Mittelwerte der Produkte $x_i y_i$ bzw. der Quadrate x_i^2. Dann berechnen sich k und d wie folgt:

$$k = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \quad d = \overline{y} - k\overline{x}.$$

Beispiel Das obige Bild paßt zu den Daten

$n = 5$ und

i	1	2	3	4	5
x_i	1	2	3	4	5
y_i	-1	0.3	0.9	1.6	3.4

Zur Berechnung der Parameter kann man eine Tabelle der folgenden Art verwenden.

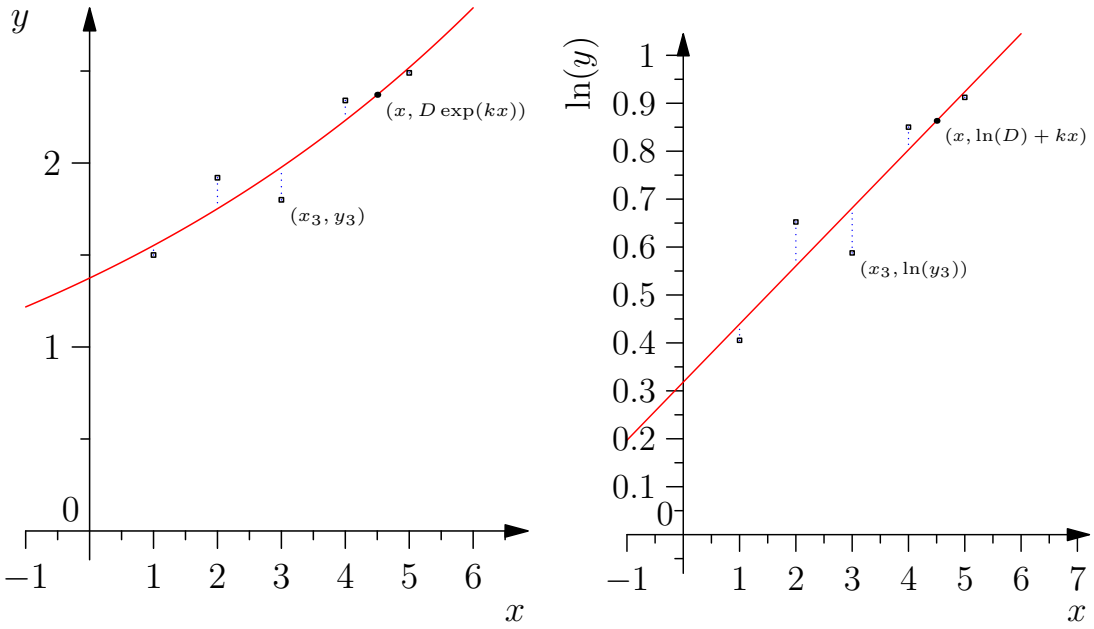
i	x_i	y_i	$x_i y_i$	x_i^2
1	1	-1	-1	1
2	2	0.3	0.6	4
3	3	0.9	2.7	9
4	4	1.6	6.4	16
5	5	3.4	17	25
$\sum$	15	5.2	25.7	55
$\sum/n$	3	1.04	5.14	11

Es ist also $k = \frac{5.14 - 3 \cdot 1.04}{11 - 3^2} = 1.01$ und $d = 1.04 - 1.01 \cdot 3 = -1.99$.

2 Datenanpassung mit Exponentialfunktionen

Nicht immer sind gegebene Daten so beschaffen, daß man sinnvoll eine Gerade zum Anpassen verwenden kann (vgl. die markierten Punkte im linken Bild unten). Es könnte aber sein, daß eine *Exponentialfunktion* zu den Daten paßt.

Aufgabe: Es seien n Datenpaare (x_1, y_1), (x_2, y_2),..., (x_n, y_n) mit $x_1, y_1, x_2, y_2, \ldots, x_n, y_n \in \mathbb{R}$ gegeben. Gesucht ist dann eine Funktion f der Form $f(x) = D\exp(kx)$, die „bestmöglich“ zu den Daten paßt.



Betrachtet man die Funktion $g = \ln \circ f$, so ist also $g(x) = \ln(D\exp(kx)) = \ln(D) + kx$. g ist folglich eine affine Funktion, die zu den Datenpunkten $(x_i, \ln(y_i))$ passen soll.

Deshalb bekommt man D und k so: Bestimme die Parameter k und d der Regressionsgeraden für die Datenpunkte $(x_1, \ln(y_1))$, $(x_2, \ln(y_2))$,..., $(x_n, \ln(y_n))$. Die gesuchte Funktion ist dann $f(x) = D\exp(kx)$ mit dem zuvor berechneten k und mit $D = \exp(d)$. (Man vgl. mit obigem Bild, rechts hat man eine sogenannte *semilogarithmische* Darstellung der Daten.)

Natürlich geht das nur für solche Daten, bei denen alle y_i-Werte positiv sind.

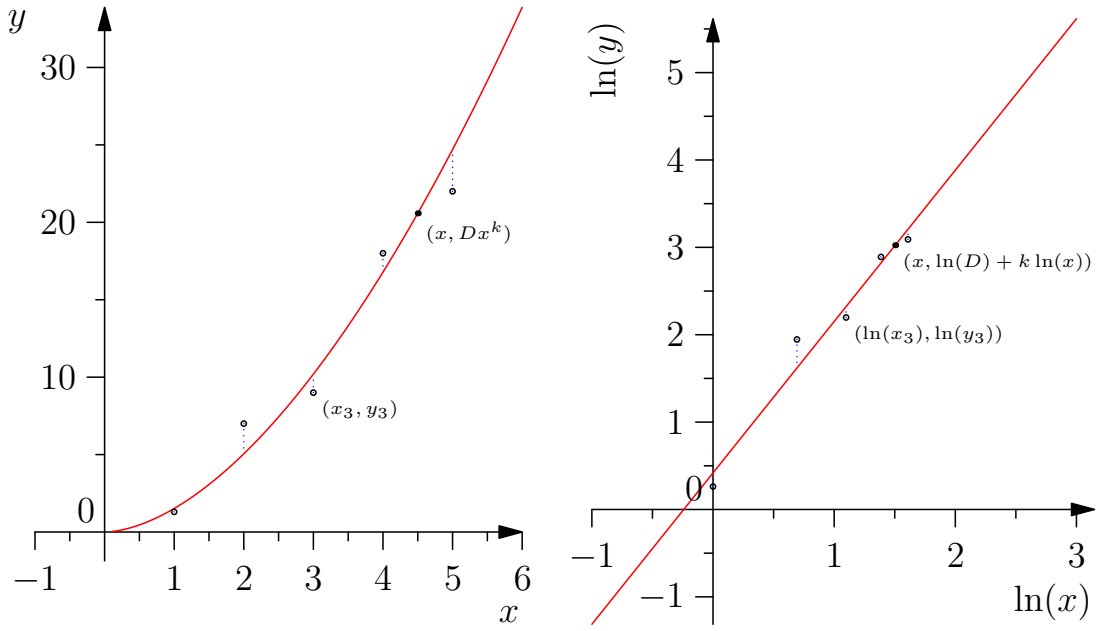
Praktische Durchführung: Sind die Datenpunkte (x_i, y_i), $1 \leq i \leq n$, gegeben, so bilde man daraus neue Datenpunkte (x_i, z_i) mit $z_i = \ln(y_i)$. Aus diesen neuen Daten bestimme man die Parameter k und d mit Hilfe der *linearen Regression*. Die gesuchte Funktion f ist dann von der Form $f(x) = \exp(d)\exp(kx)$.

Die bestimmten Parameter sind nur im transformierten Modell bestmöglich. Wollte man $\sum_{i=1}^{n}(y_i - D\exp(kx_i))^2$ selbst minimieren, so bekäme man zwar bessere Parameter, die Rechnung wäre aber viel komplizierter und nur näherungsweise durchzuführen.

3 Datenanpassung mit Potenzfunktionen

Nicht immer sind gegebene Daten so beschaffen, daß man sinnvoll eine Gerade oder eine Exponentialfunktion zum Anpassen verwenden kann (vgl. die markierten Punkte im linken Bild unten). Es könnte aber sein, daß eine *Potenzfunktion* zu den Daten paßt.

Aufgabe: Es seien n Datenpaare (x_1, y_1), (x_2, y_2),..., (x_n, y_n) mit $x_1, y_1, x_2, y_2, \ldots, x_n, y_n \in \mathbb{R}$ gegeben. Gesucht ist dann eine Funktion f der Form $f(x) = Dx^k$, die „bestmöglich" zu den Daten paßt.



Betrachtet man die Funktion $g = \ln \circ f$, so ist also $g(x) = \ln(D\exp(kx)) = \ln(D) + k\ln(x)$. Die Funktion g ist folglich eine affine Funktion von $\ln(x)$, die zu den Datenpunkten $(\ln(x_i), \ln(y_i))$ passen soll.

Deshalb bekommt man D und k so: Bestimme die Parameter k und d der Regressionsgeraden für die Datenpunkte $(\ln(x_1), \ln(y_1))$, $(\ln(x_2), \ln(y_2))$,..., $(\ln(x_n), \ln(y_n))$. Die gesuchte Funktion ist dann $f(x) = Dx^k$ mit dem zuvor berechneten k und mit $D = \exp(d)$. (Man vgl. mit obigem Bild, rechts hat man eine sogenannte *doppeltlogarithmische* Darstellung der Daten.) Natürlich geht das nur für solche Daten, bei denen alle y_i-Werte und alle x_i-Werte positiv sind.

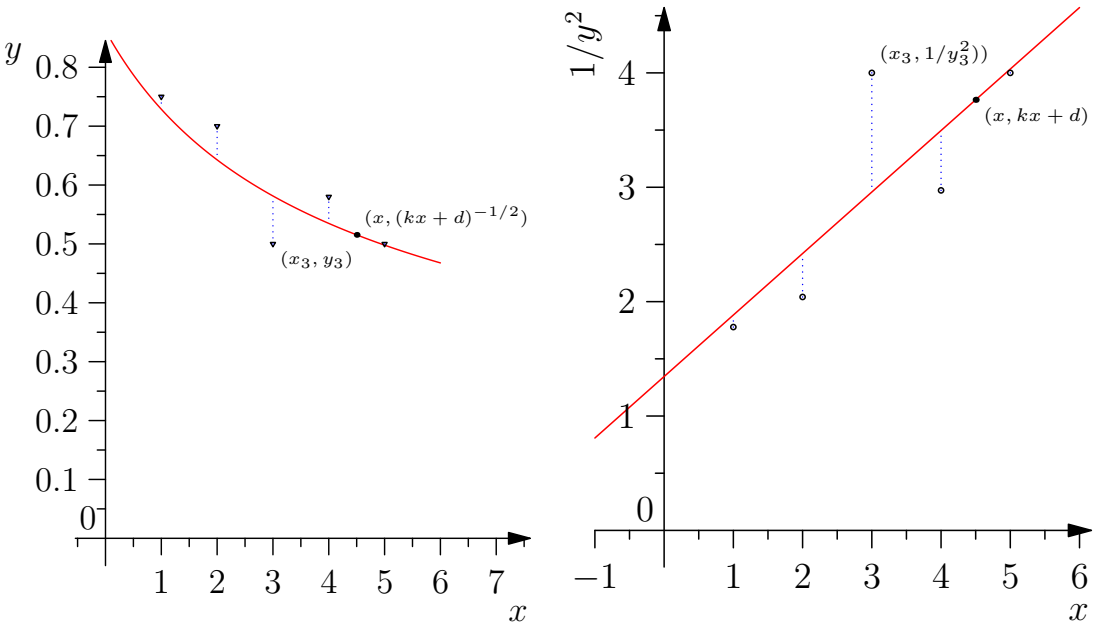
Praktische Durchführung: Sind die Datenpunkte (x_i, y_i), $1 \le i \le n$, gegeben, so bilde man daraus neue Datenpunkte (w_i, z_i) mit $z_i = \ln(y_i)$ und $w_i = \ln(x_i)$. Aus diese neuen Daten bestimme man die Parameter k und d mit Hilfe der *linearen Regression*. Die gesuchte Funktion f ist dann von der Form $f(x) = \exp(d)x^k$.

Die bestimmten Parameter sind nur im transformierten Modell bestmöglich. Wollte man $\sum_{i=1}^{n}(y_i - Dx_i^k)^2$ selbst minimieren, so bekäme man zwar bessere Parameter, die Rechnung wäre aber viel komplizierter und nur näherungsweise durchzuführen.

4 Datenanpassung mit anderen Funktionen

Die Konzentration $c(t)$ eines Stoffes bei einer chemischen Reaktion zur Zeit t hat die Form $c(t) = B\exp(-at)$, wenn jeweils ein ($m = 1$) Molekül des Stoffes reagiert. Benötigt die Reaktion $m \geq 2$ Moleküle, so hat c die Form $c(t) = (at + b)^{-1/(m-1)}$. Für $m = 1$ führt das Logarithmieren auf die Beziehung $\ln(c(t)) = \ln(B) - at$. Die Bestimmung der Parameter erfolgt somit wie im Abschnitt über Exponentialfunktionen, wobei B durch D und a durch $-k$ ersetzt wird. Im Fall $m \geq 2$ bedeutet $c(t) = (at + b)^{-1/(m-1)}$, daß $c(t)^{-(m-1)} = at + b$

Aufgabe: Es seien n Datenpaare (x_1, y_1), (x_2, y_2),..., (x_n, y_n) mit $x_1, y_1, x_2, y_2, \ldots, x_n, y_n \in \mathbb{R}$ gegeben. Ferner sei $m \geq 2$. Gesucht ist dann eine Funktion f der Form $f(x) = (kx + d)^{-1/(m-1)}$, die „bestmöglich“ zu den Daten paßt.



Praktische Durchführung: Sind die Datenpunkte (x_i, y_i), $1 \leq i \leq n$, gegeben, und ist $m \geq 2$, so bilde man daraus neue Datenpunkte (x_i, z_i) mit $z_i = y_i^{-(m-1)}$. Aus diesen neuen Daten bestimme man die Parameter k und d mit Hilfe der *linearen Regression*. Die gesuchte Funktion f ist dann von der Form $f(x) = (kx + d)^{-1/(m-1)}$.

Die bestimmten Parameter sind nur im transformierten Modell bestmöglich. Wollte man $\sum_{i=1}^{n} \left(y_i - (kx + d)^{-1/(m-1)}\right)^2$ selbst minimieren, so bekäme man zwar bessere Parameter, die Rechnung wäre aber viel komplizierter und nur näherungsweise durchzuführen.