Regressionsanalyse

Die **Regressionsanalyse** ist eine Sammlung von statistischen Analyseverfahren. Ziel bei den am häufigsten eingesetzten Analyseverfahren ist es, Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen. Sie wird insbesondere verwendet, wenn Zusammenhänge quantitativ zu beschreiben oder Werte der abhängigen Variablen zu prognostizieren sind.

Die früheste Form der Regression war die Methode der kleinsten Quadrate (frz.: *méthode des moindres carrés*), 1805 von Legendre und 1809 von Gauß veröffentlicht. Beide verwendeten die Methode, um die Umlaufbahnen der Planeten um die Sonne anhand von astronomischen Beobachtungen zu bestimmen. Gauß veröffentlichte eine Weiterentwicklung der Theorie der kleinsten Quadrate im Jahr 1821, die eine Version des Satzes von Gauß-Markow enthielt.

Mathematisch kann die Beziehung dargestellt werden als

$y = f(x) + e$, im eindimensionalen Fall und
$y = f(x_1, x_2, \cdots, x_n) + e$ im *n*-dimensionalen Fall,

wobei *y* die abhängige Variable und *x* eine oder mehrere unabhängige Variablen bezeichnen. *f* ist die gesuchte oder angenommene Funktion und *e* bezeichnet den Fehler bzw. das Residuum des Modells.

Regressionsverfahren haben viele praktische Anwendungen. Die meisten Anwendungen fallen in eine der folgenden beiden Kategorien:

- Wenn das Ziel die Prognose oder Vorhersage ist, dann kann der durch das Regressionsverfahren ermittelte funktionale Zusammenhang verwendet werden, um ein Vorhersagemodell zu erstellen. Wenn nun zusätzliche Werte *x* ohne zugehörigen Wert *y* vorliegen, dann kann das angepasste Modell zur Vorhersage des Wertes von *y* verwendet werden.

- Wenn eine Variable *y* und eine Anzahl von Variablen x_1, ..., x_p vorliegen, die mit *y* in Verbindung gebracht werden können, dann können Regressionsverfahren angewandt werden, um die Stärke des Zusammenhangs zu quantifizieren. So können diejenigen x_j ermitteln werden, die gar keinen Zusammenhang mit *y* haben; oder diejenigen Teilmengen x_i, ..., x_j, die redundante Information über *y* enthalten.

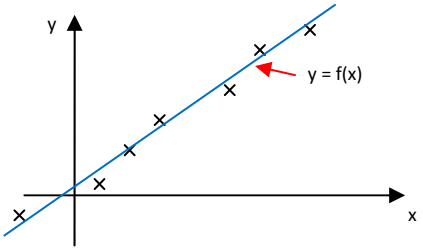
Nachfolgend sollen drei Formen der Regressionsanalyse dargestellt werden, und zwar
- die lineare,
- die quadratische (polynomische mit Grad 2) und
- die exponentielle Regression,

allerdings nur für den eindimensionalen Fall mit der Methode der kleinsten Quadrate.

1. Lineare Regression

Gegeben sind Datenpaare, z.B. in Form einer Tabelle, die grafisch in einem x-y-Diagramm dargestellt werden können.

x	x_1	x_2	.....	x_n
y	y_1	y_2	.....	y_n



Gesucht wird nun die Funktion $y = f(x) + e$ mit minimalem e, also einer optimalen Anpassung.

Im Falle der linearen Regression wird die Funktion

$$\hat{y} = f(\hat{x}) = m \cdot \hat{x} + n \tag{1}$$

unter der Bedingung gesucht, dass die Summe der Quadrate der Abstände der tatsächlichen *y*-Werte von den berechneten $\hat{y}$ -Werten ein Minimum hat (Methode der kleinsten Quadrate). Die zu minimierende Größe sei allgemein

$$V = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \,. \tag{2}$$

Zur Bestimmung der Konstanten *m* und *n* in Gleichung (1) wird Gleichung (2) modifiziert und die partiellen Ableitungen $\frac{\partial V}{\partial m}$ und $\frac{\partial V}{\partial n}$ gleich null gesetzt, um jeweils das Minimum zu erhalten:

$$V(m,n) = \sum_{i=1}^{k} (y_i - (m \cdot x_i + n))^2 = \sum_{i=1}^{k} (y_i - m \cdot x_i - n)^2$$

Partielle Ableitung nach *m*:

$$\rightarrow \frac{\partial V(m,n)}{\partial m} = 2 \cdot \sum_{i=1}^{k} (y_i - m \cdot x_i - n) \cdot (-x_i) = 0$$

$$\rightarrow \sum_{i=1}^{k} (x_i y_i) = m \cdot \sum_{i=1}^{k} x_i^2 + n \cdot \sum_{i=1}^{k} x_i \tag{3}$$

Partielle Ableitung nach *n*:

$$\rightarrow \frac{\partial V(m,n)}{\partial n} = 2 \cdot \sum_{i=1}^{k} (y_i - m \cdot x_i - n) \cdot (-1) = 0$$

$$\rightarrow \sum_{i=1}^{k} y_i = m \cdot \sum_{i=1}^{k} x_i + n \cdot k \tag{4}$$

Werden die Gleichungen (3) und (4) jeweils durch die Anzahl *k* der Datenpaare dividiert, so werden die Summenzeichen (z.B. $\sum_{i=1}^{k} x_i$) durch die Mittelwerte (z.B. $\frac{1}{k} \cdot \sum_{i=1}^{k} x_i = \overline{x_i}$) ersetzt. Es entsteht ein lineares Gleichungssystem in zwei Variablen *m* und *n*, dessen Lösung die optimalen Konstanten liefert:

$$\left.\begin{aligned} \overline{x_i y_i} &= m \cdot \overline{x_i^2} + n \cdot \overline{x_i} \\ \overline{y_i} &= m \cdot \overline{x_i} + n \end{aligned}\right\} \rightarrow \quad m = \frac{\overline{x_i y_i} - \overline{y_i} \cdot \overline{x_i}}{\overline{x_i^2} - \left(\overline{x_i}\right)^2} \tag{5}$$

$$n = \overline{y_i} - m \cdot \overline{x_i} = \overline{y_i} - \frac{\overline{x_i y_i} - \overline{y_i} \cdot \overline{x_i}}{\overline{x_i^2} - \left(\overline{x_i}\right)^2} \cdot \overline{x_i} \tag{6}$$

Beispiel:

Messtabelle **Mittelwerte**

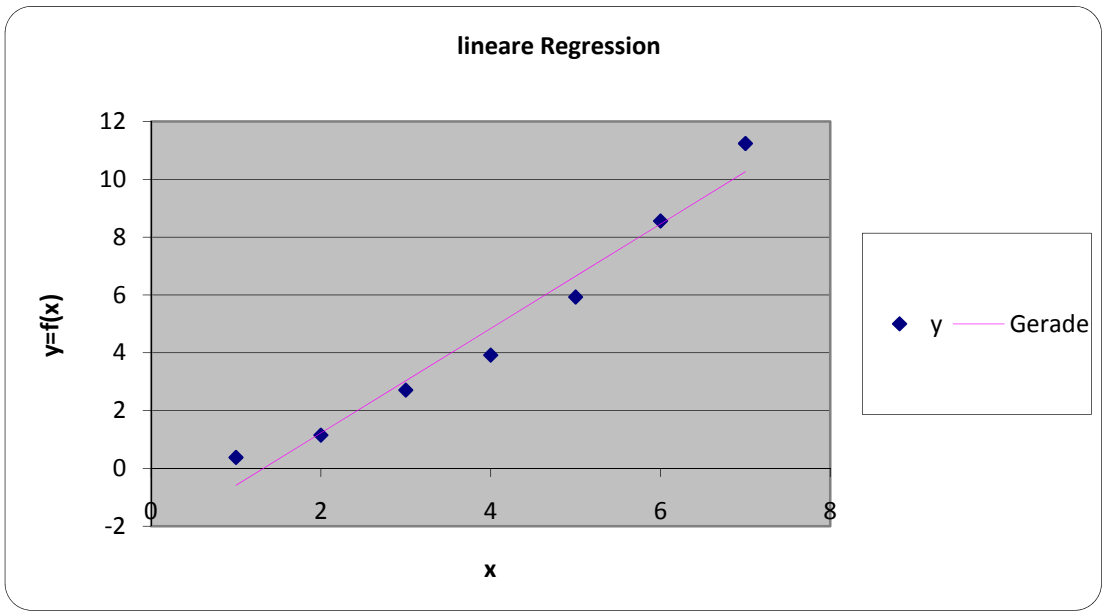
x	1	2	3	4	5	6	7	4
y	0,38	1,15	2,71	3,92	5,93	8,56	11,24	4,84142857
xy	0,38	2,3	8,13	15,68	29,65	51,36	78,68	26,5971429
x^2	1	4	9	16	25	36	49	20

Mit den berechneten Mittelwerten werden mit Hilfe der Formeln (5) und (6) die Konstanten berechnet:

m = 1,80785714

n = -2,39

Die grafische Darstellung der Datenpunkte und der Regressionsgeraden ergibt:



2. Quadratische (polynomische) Regression

Im Falle der polynomischen Regression vom Grad 2 (quadratisch) wird die Funktion

$$\hat{y} = f(\hat{x}) = a \cdot \hat{x}^2 + b \cdot \hat{x} + c \tag{7}$$

unter der Bedingung gesucht, dass die Funktion

$$V(a,b,c) = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{k} (y_i - a \cdot x_i^2 - b \cdot x_i - c)^2 \tag{8}$$

der Summe der Quadrate der Abstände der tatsächlichen *y*-Werte von den berechneten $\hat{y}$ -Werten ein Minimum hat.

Zur Bestimmung der Konstanten *a*, *b* und *c* in Gleichung (8) werden die partiellen Ableitungen $\frac{\partial V}{\partial a}$, $\frac{\partial V}{\partial b}$ und $\frac{\partial V}{\partial c}$ gleich null gesetzt, um jeweils das Minimum zu erhalten:

Partielle Ableitung nach *a*:

$$\rightarrow \frac{\partial V(a,b,c)}{\partial a} = 2 \cdot \sum_{i=1}^{k} (y_i - a \cdot x_i^2 - b \cdot x_i - c) \cdot (-x_i^2) = 0$$

dividiert durch *k* ergibt: $\overline{y_i x_i^2} = a \cdot \overline{x_i^4} + b \cdot \overline{x_i^3} + c \cdot \overline{x_i^2}$ (9)

Partielle Ableitung nach *b*:

$$\rightarrow \frac{\partial V(a,b,c)}{\partial b} = 2 \cdot \sum_{i=1}^{k} (y_i - a \cdot x_i^2 - b \cdot x_i - c) \cdot (-x_i) = 0$$

dividiert durch *k* ergibt: $\overline{y_i x_i} = a \cdot \overline{x_i^3} + b \cdot \overline{x_i^2} + c \cdot \overline{x_i}$ (10)

Partielle Ableitung nach *c*:

$$\rightarrow \frac{\partial V(a,b,c)}{\partial c} = 2 \cdot \sum_{i=1}^{k} (y_i - a \cdot x_i^2 - b \cdot x_i - c) \cdot (-1) = 0$$

dividiert durch *k* ergibt: $\overline{y_i} = a \cdot \overline{x_i^2} + b \cdot \overline{x_i} + c$ (11)

Es entsteht wieder ein lineares Gleichungssystem in drei Variablen *a, b* und *c*, dessen Lösung die optimalen Konstanten liefert:

$$a = \frac{\left(\overline{y_i x_i^2} - \overline{y_i} \cdot \overline{x_i^2}\right) \cdot \left(\overline{x_i^2} - \left(\overline{x_i}\right)^2\right) - \left(\overline{y_i x_i} - \overline{y_i} \cdot \overline{x_i}\right) \cdot \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)}{\left(\overline{x_i^4} - \left(\overline{x_i^2}\right)^2\right) \cdot \left(\overline{x_i^2} - \left(\overline{x_i}\right)^2\right) - \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)^2} \qquad (12)$$

$$b = \frac{\overline{y_i x_i} - \overline{y_i} \cdot \overline{x_i} - a \cdot \left(\overline{x_i^3} - \overline{x_i} \cdot \overline{x_i^2}\right)}{\overline{x_i^2} - \left(\overline{x_i}\right)^2} \qquad (13)$$

$$c = \overline{y_i} - a \cdot \overline{x_i^2} - b \cdot \overline{x_i} \qquad (14)$$

In den Gleichungen (13) und (14) wird auf das Einsetzen der kompletten Formel der zuvor berechneten Konstanten aus Gründen der Übersichtlichkeit verzichtet.

Beispiel:

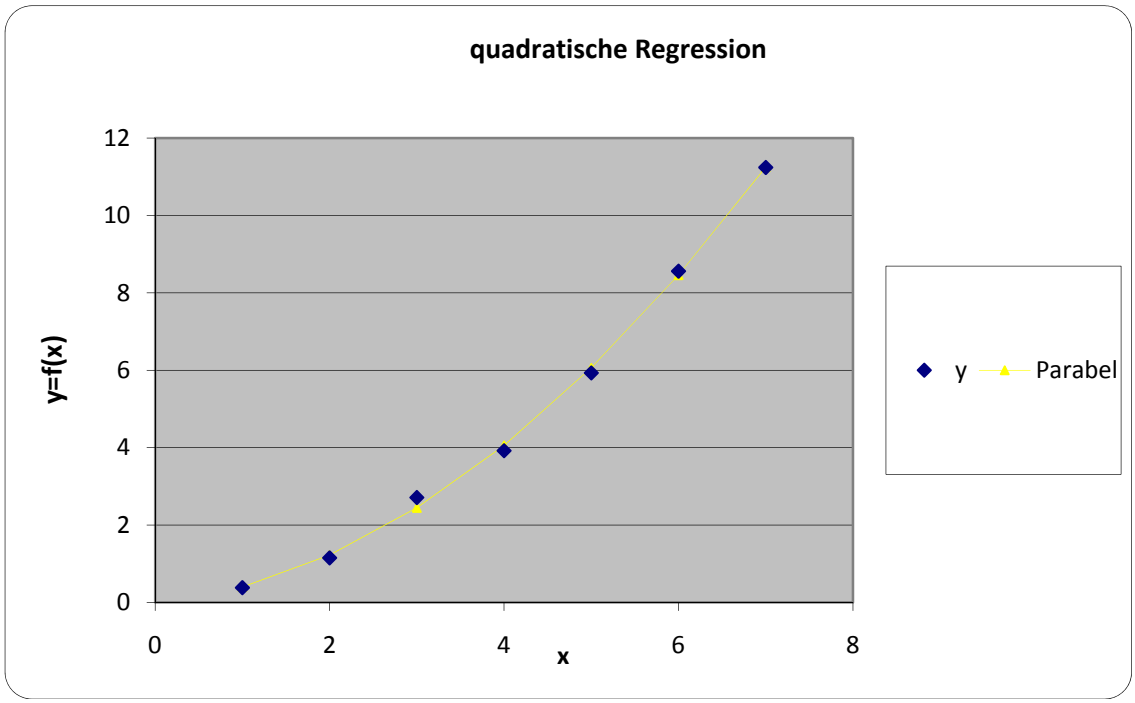
Messtabelle **Mittelwerte**

x	1	2	3	4	5	6	7	4
y	0,38	1,15	2,71	3,92	5,93	8,56	11,24	4,84142857
xy	0,38	2,3	8,13	15,68	29,65	51,36	78,68	26,5971429
x^2	1	4	9	16	25	36	49	20
x^3	1	8	27	64	125	216	343	112
x^4	1	16	81	256	625	1296	2401	668
yx^2	0,38	4,6	24,4	62,72	148,25	308,2	550,76	157,037143

Die mit den Formeln (12) bis (14) berechneten Konstanten sind:

a =	0,19642857
b =	0,23642857
c =	-0,03285714

Die grafische Darstellung der Datenpunkte und der Regressionsparabel ergibt:



3. Exponentielle Regression

Bei der exponentiellen Regression wird die Funktion

$$\hat{y} = f(\hat{x}) = d \cdot e^{k \cdot \hat{x}} \tag{15}$$

wieder unter der Bedingung gesucht, dass die Summe der Quadrate der Abstände der tatsächlichen y-Werte von den berechneten $\hat{y}$-Werten ein Minimum hat. Anders als bei der polynomischen Regression wird die Funktion (15) zunächst in eine lineare Funktion umgewandelt, indem die Gleichung logarithmiert wird. Dann erhält sie die Form

$$\ln \hat{y} = \ln\left(d \cdot e^{k \cdot \hat{x}}\right) = k \cdot \hat{x} + \ln d \,. \tag{16}$$

Ein Vergleich mit der linearen Regression unter Punkt 1 liefert damit folgende zu minimierende Funktion:

$$V(m,n) = \sum_{i=1}^{k} (\ln y_i - m \cdot x_i - n)^2 \rightarrow V(k, \ln d) = \sum_{i=1}^{k} (\ln y_i - k \cdot x_i - \ln d)^2 \tag{17}$$

Da das Verfahren mit den partiellen Ableitungen identisch zur linearen Regression ist, wird hier auf die ausführliche Berechnung verzichtet. Es ergeben sich folgende Konstanten:

$$\left.\begin{aligned} \overline{x_i \ln y_i} &= k \cdot \overline{x_i^2} + \ln d \cdot \overline{x_i} \\ \overline{\ln y_i} &= k \cdot \overline{x_i} + \ln d \end{aligned}\right\} \rightarrow \quad k = \frac{\overline{x_i \ln y_i} - \overline{\ln y_i} \cdot \overline{x_i}}{\overline{x_i^2} - \left(\overline{x_i}\right)^2} \tag{18}$$

$$\ln d = \overline{\ln y_i} - k \cdot \overline{x_i} \rightarrow d = e^{\ln d} \tag{19}$$

Beispiel:

Messtabelle **Mittelwerte**

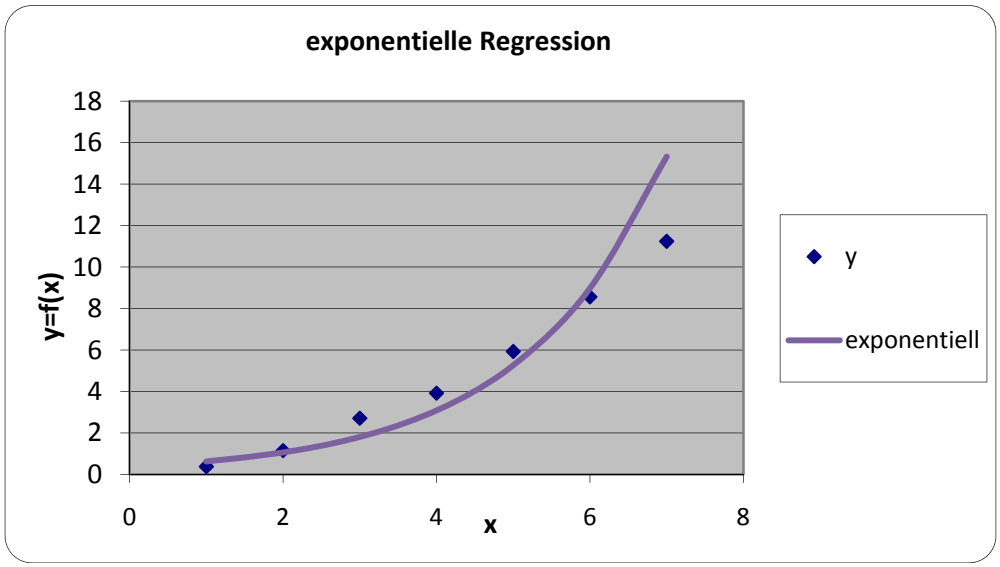
x	1	2	3	4	5	6	7	4
y	0,38	1,15	2,71	3,92	5,93	8,56	11,24	4,84142857
ln y	-0,97	0,14	1	1,366	1,78	2,147	2,4195	1,12597449
xy	0,38	2,3	8,13	15,68	29,65	51,36	78,68	26,5971429
x^2	1	4	9	16	25	36	49	20
x lny	-0,97	0,28	2,99	5,464	8,9001	12,88	16,936	6,6408895

Die mit den Formeln (18) und (19) berechneten Konstanten sind:

d = 0,36384875

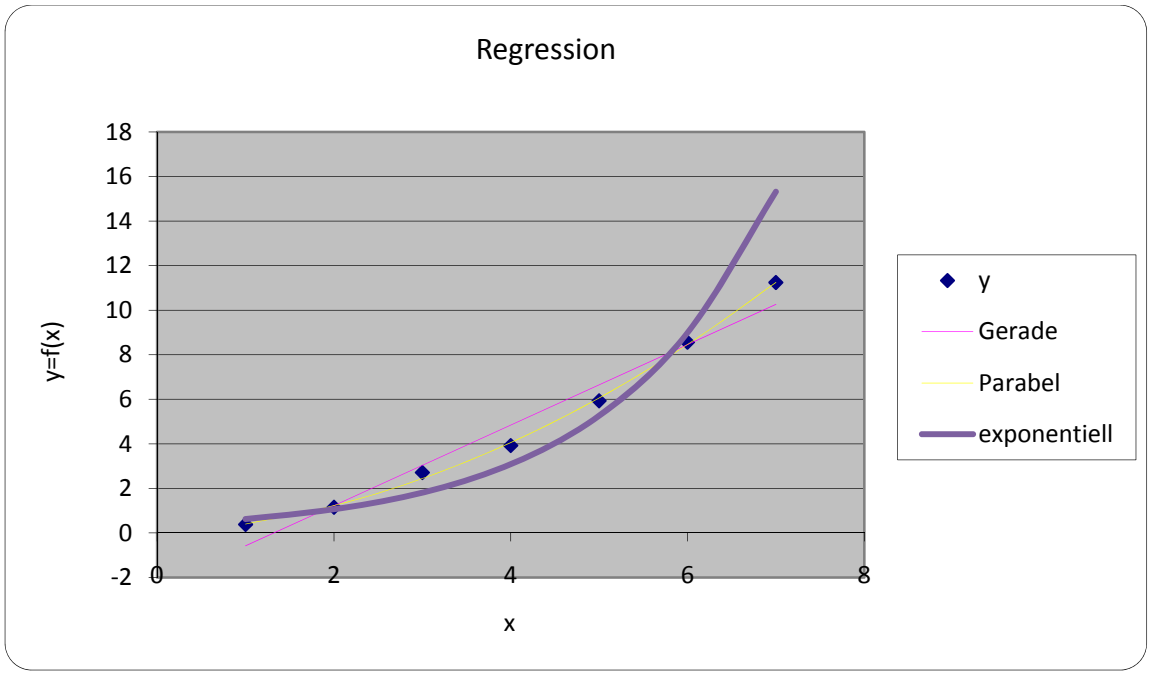
k = 0,53424788

Die grafische Darstellung der Datenpunkte und der Regressionskurve ergibt



4. Vergleich und Auswertung

In den Beispielen 1 bis 3 sind bewusst die gleichen Messwerte verwendet worden. Damit ist ein Vergleich der drei Anpassungsformen möglich. Im nachfolgenden Diagramm sind alle drei Kurven eingezeichnet.



Rein optisch ist die Parabel die Kurve mit der geringsten Abweichung von den Messpunkten.
Ein rechnerischer Nachweis ist mit einer Formel möglich, die die mittlere Abweichung σ der Messpunkte von den berechneten Punkten berechnet. Die Formel lautet:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k}} \qquad (20)$$

In den drei behandelten Fällen ergibt sich:

Regressionstyp	σ
linear	0,693
quadratisch	**0,132**
exponentiell	1,639

Die quadratische Regression liefert in unserem Fall die beste Anpassung, d.h. mit der geringsten Abweichung von den Messwerten.