Regression mit dem MMS der TI-Nspire-Software erklären und durchführen

Wolfgang Häfner

Der Beitrag soll auch für Schülerinnen und Schüler verständlich sein, ist also nicht theoretisch überladen, geht aber doch auf einige Grundlagen ein. Damit soll das Verständnis dafür, wie Regression funktioniert, in den Blick genommen werden.

Lineare Regression

Gesucht ist hier eine lineare Funktion $y = g(x) = a \cdot x + b$, die den n gegebenen Wertepaaren (x, y) optimal angenähert ist. Das ist bestimmt der Fall, wenn die Abweichung der y-Werte der Wertepaare von den entsprechenden Funktionswerten möglichst gering ist.

Wir betrachten deshalb die Summe der Quadrate dieser Abweichungen:

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

Nun ermittelt man a und b, so dass $f(a,b)$ minimal ist.
Das ist nun deshalb problematisch, da f eine Funktion mit zwei unabhängigen Variablen ist und die Analysis hierzu in der Schule nicht behandelt wird.
Sind wir also hier nicht so anspruchsvoll und beschränken uns auf die notwendige Bedingung.
Angenommen b_{min} wäre der gesuchte Wert der Variablen b, dann müsste a_{min} die Gleichung
$\frac{d}{da}\big(f(a\,, b_{min})\big) = 0$ erfüllen.
Analog gilt dies dann für b_{min} und die Gleichung $\frac{d}{db}\big(f(a_{min}\,, b)\big) = 0$.

Verwendet man für $\frac{d}{da}\big(f(a\,, b)\big)$ und $\frac{d}{db}\big(f(a\,, b)\big)$ die Schreibweise $f_a(a,b)$ bzw. $f_b(a,b)$, dann sucht man also die Lösung (a,b) des folgenden Gleichungssystems:

$I)\ \ f_a(a,b) = 0$
$II)\ f_b(a,b) = 0$

Weiterhin gilt:

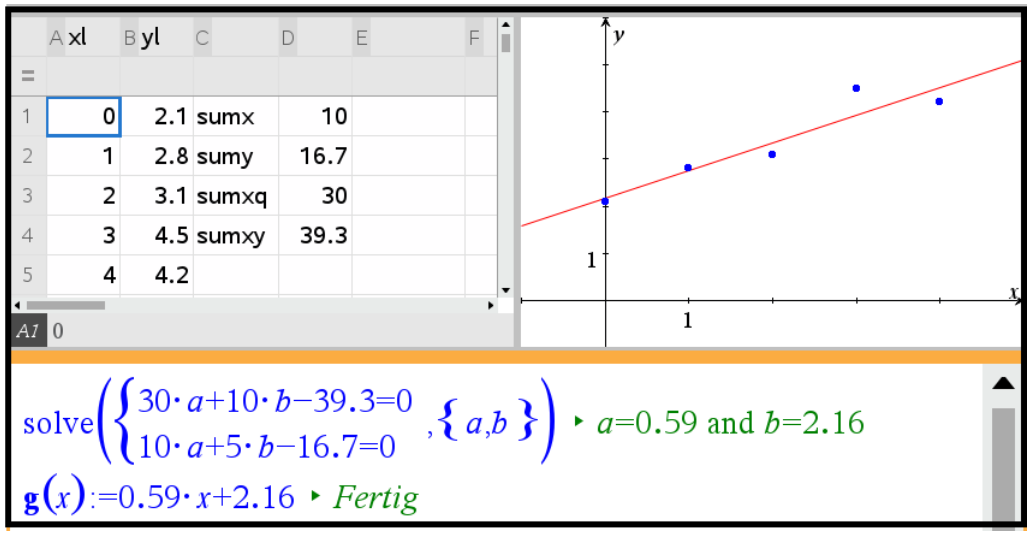
$$f_a(a,b) = \sum_{i=1}^{n} 2 \cdot (a \cdot x_i + b - y_i) \cdot x_i = 2 \cdot \sum_{i=1}^{n} \left(a \cdot x_i^2 + b \cdot x_i - x_i \cdot y_i\right)$$
$$= 2 \cdot \left(a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \cdot y_i \right)$$
$$f_b(a,b) = \sum_{i=1}^{n} 2 \cdot (a \cdot x_i + b - y_i) \cdot 1 = 2 \cdot \sum_{i=1}^{n} (a \cdot x_i + b - y_i)$$
$$= 2 \cdot \left(a \cdot \sum_{i=1}^{n} x_i + n \cdot b - \sum_{i=1}^{n} y_i \right)$$

Damit erhält das Gleichungssystem folgende Form:

$$I)\ \ a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \cdot y_i = 0$$
$$II)\qquad a \cdot \sum_{i=1}^{n} x_i + n \cdot b - \sum_{i=1}^{n} y_i = 0$$

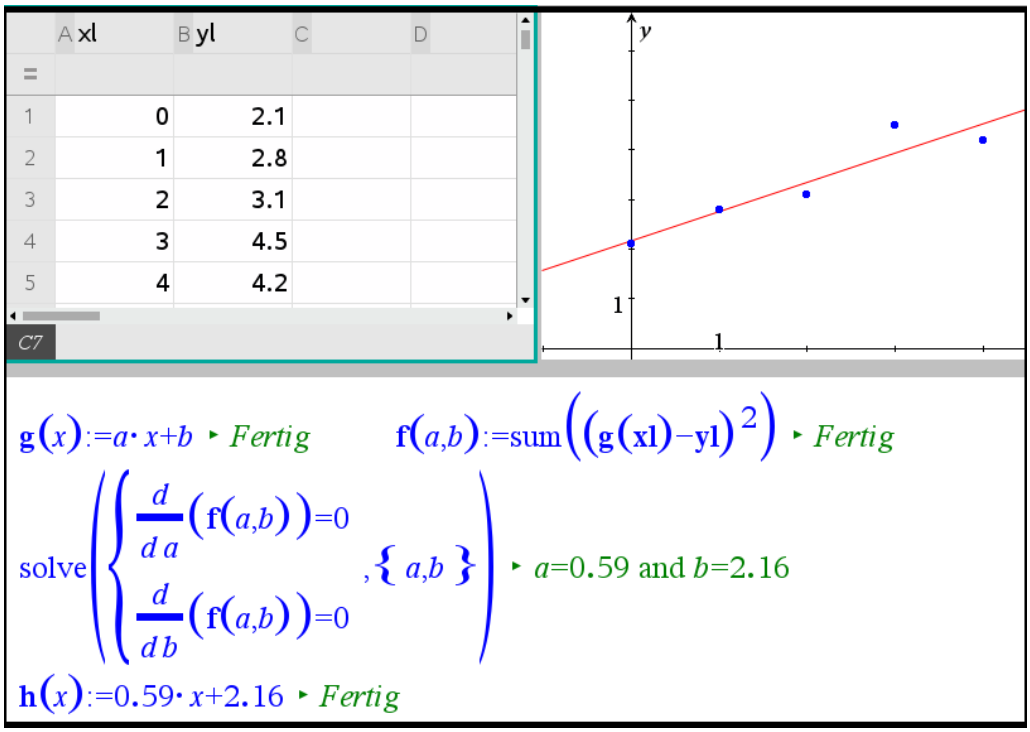
Mit etwas Mühe beim Berechnen der Summen ließe sich dieses Gleichungssystem auch ohne ein MMS lösen.
Im nachfolgenden Beispiel verwenden wir es trotzdem.
Wir werden dieses Beispiel danach noch einmal betrachten und feststellen, dass der bisherige Lösungsaufwand eigentlich nicht nötig ist, wenn man die Möglichkeiten des MMS besser nutzt. Die ausführliche Vorgehensweise soll vor allem dazu dienen, das eigentliche Verfahren transparenter zu machen. Der effizientere Lösungsweg kann dieses Anliegen nicht unterstützen.

Beispiel:
Variante 1:



Die Spalten A und B der Applikation Lists- & Spreadsheet enthalten die x- und y-Werte,
Die benötigten Summen befinden sich in den Zellen D1 bis D4.
$D1:=sum(xl)$ $D2:=sum(yl)$ $D3:=sum(xl^2)$ $D4:=sum(xl\cdot yl)$

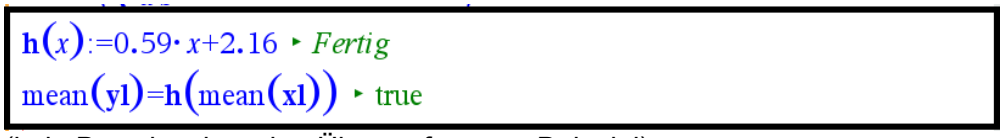
Variante 2:



Das MMS ist doch sehr leistungsfähig. Die aufwendigen Vorüberlegungen entfallen, wenn man dem Rechner die Ermittlung der Ableitungen und die Lösung des Gleichungssystems überlässt.

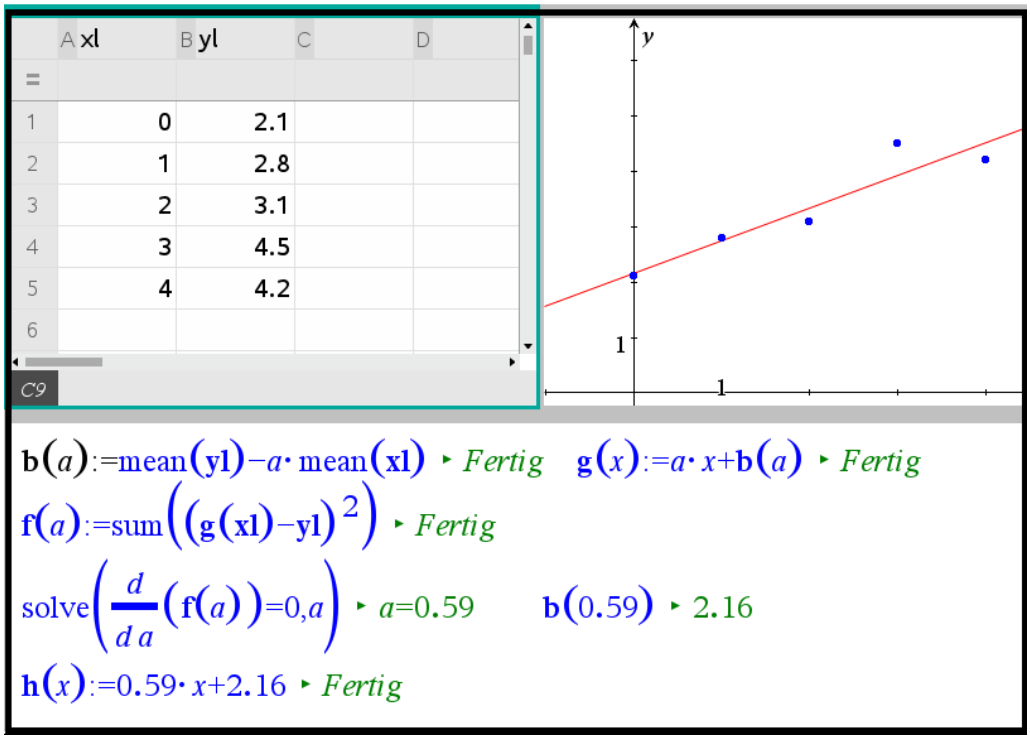
Weitere Zusammenhänge

Wenn $\bar{x}$ der Mittelwert der Liste der gegebenen x-Werte und $\bar{y}$ der Mittelwert der Liste der gegebenen y-Werte und $y = h(x) = a \cdot x + b$ die Gleichung der Regressionsgerade ist, so gilt $\bar{y} = h(\bar{x}) = a \cdot \bar{x} + b$.



(kein Beweis, aber eine Überprüfung am Beispiel)

Damit kann man dann b in der Funktionsgleichung $f(a, b)$ durch $\bar{y} - a \cdot \bar{x}$ substituieren, wodurch nur noch eine unabhängige Variable verbleibt.



Korrelationskoeffizient und Bestimmtheitsmaß

In den folgenden Beispielen wird die lineare Regression mit der statistischen Berechnung „Lineare Regression (mx+b)“, die vom CAS zur Verfügung gestellt wird, durchgeführt. Dabei wird auch der Korrelationskoeffizient r und das Bestimmtheitsmaß r^2 berechnet und ausgegeben. Die Bedeutung dieser Maße kann auch im Unterricht besprochen werden. Ihre Definition über die Varianz der x- und y-Werte und die Kovarianz von x und y würde im Unterricht zu weit führen.

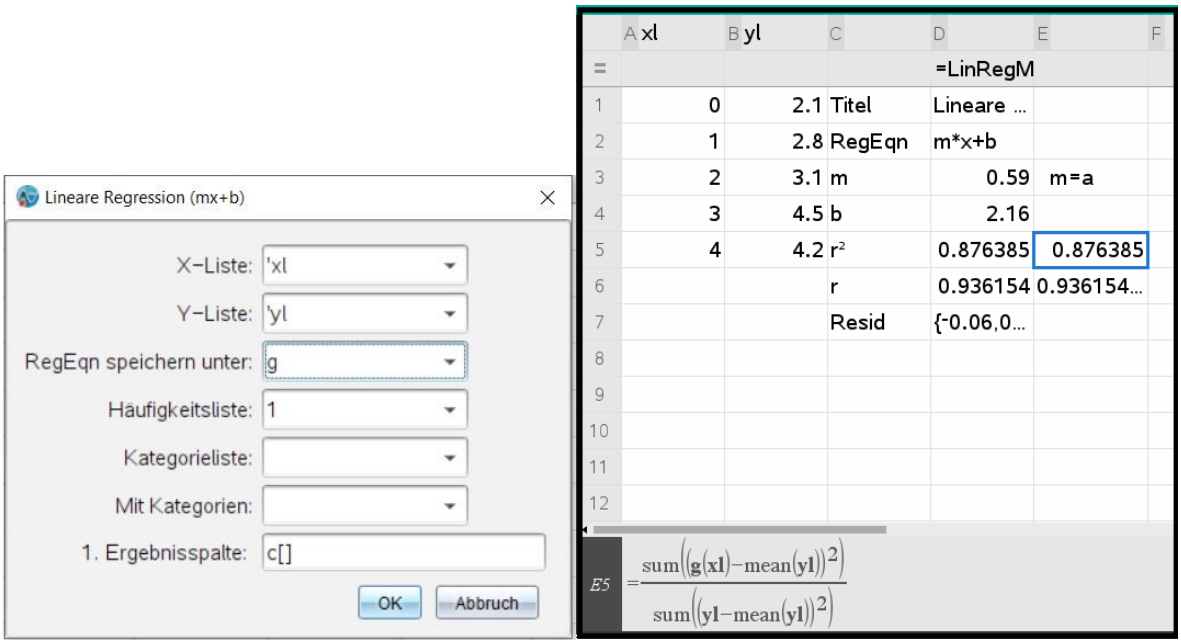
Wenn überhaupt, so kann man die vom Rechner ausgegeben Werte durch die Berechnung folgender Terme bestätigen.

$$r^2 = \frac{\sum_{i=1}^{n}(g(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad r = sign(m) \cdot \sqrt{r^2}$$

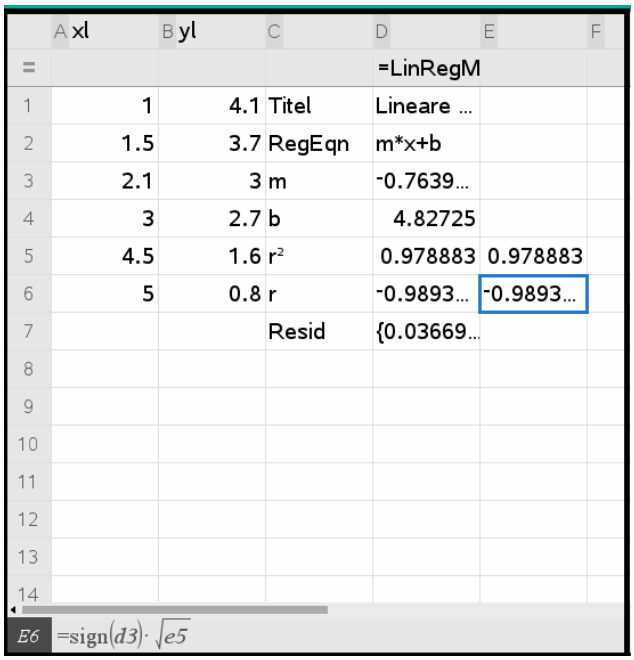
Rechnereingabe:

$$E5: \quad = \frac{sum\left(\left(g(xl) - mean(yl)\right)^2\right)}{sum((yl - mean(yl)^2)} \qquad E6: = sign(D3) \cdot \sqrt{E5}$$

Das erste Beispiel ist das bereits behandelte (siehe oben).

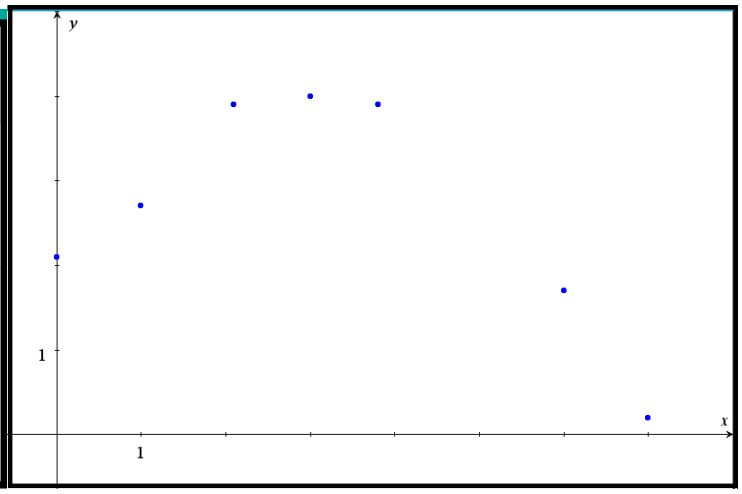


2. Beispiel



Quadratische Regression

	A xl	B yl	C
=			
1	0	2.1	
2	1	2.7	
3	2.1	3.9	
4	3	4	
5	3.8	3.9	
6	6	1.7	
7	7	0.2	



Dass hier kein linearer Zusammenhang vorliegt ist am Streudiagramm sofort zu sehen.
Vielleicht ist hier eine Parabel die optimale Näherung der Punkte.

Das Ziel ist also, die Gleichung einer quadratischen Funktion zu finden, deren Graph eine optimale Annährung an die Punkte des Streudiagramms ergibt.
Dem Verfahren liegen praktisch die gleichen Überlegungen zu Grunde, die schon bei der linearen Regression zum Ziel führten.

Daher ist die nachfolgend ersichtliche Vorgehensweise prinzipiell bekannt.
Neu ist, dass es drei zu ermittelnde Parameter und damit auch drei Ableitungen und ein Gleichungssystem mit drei Gleichungen gibt.

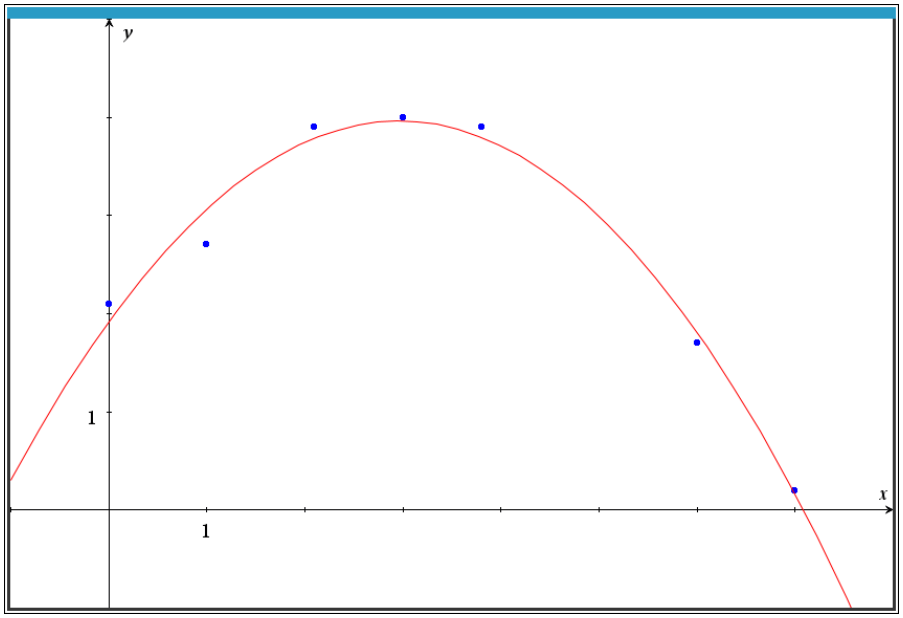
$$g(x):=a\cdot x^2+b\cdot x+c \quad \blacktriangleright \textit{Fertig}$$

$$f(a,b,c):=\text{sum}\left((g(xl)-yl)^2\right) \quad \blacktriangleright \textit{Fertig}$$

$$\text{solve}\left(\begin{cases}\frac{d}{da}\left(f(a,b,c)=0\right)\\ \frac{d}{db}\left(f(a,b,c)=0\right)\\ \frac{d}{dc}\left(f(a,b,c)=0\right)\end{cases},\{a,b,c\}\right)$$

$$\blacktriangleright\ a=-0.233108 \text{ and } b=1.38056 \text{ and } c=1.91777$$

$$h(x):=-0.233108\cdot x^2+1.38056\cdot x+1.91777 \quad \blacktriangleright \textit{Fertig}$$

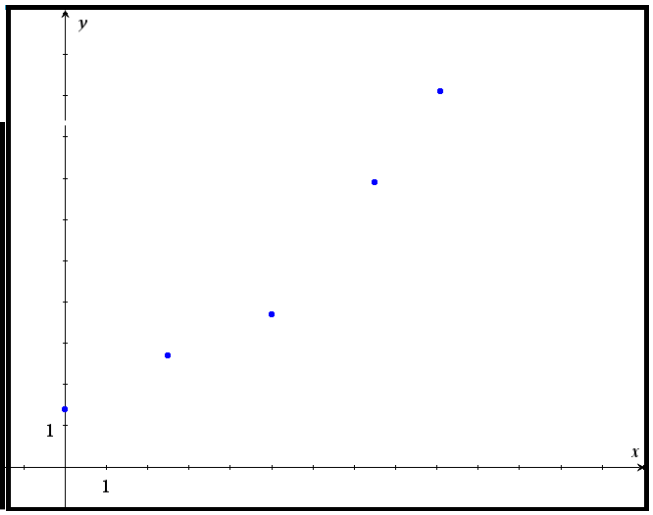


Mit der statistischen Berechnung „Quadratische Regression“ erhält man diese Ergebnis selbstverständlich einfacher.

	A xl	B yl	C	D	E
=				=QuadReg('xl,'yl,1): Copy'	
1	0	2.1	Titel	Quadratische Regression...	
2	1	2.7	RegEqn	a*x^2+b*x+c	
3	2.1	3.9	a	-0.233108	
4	3	4	b	1.38056	
5	3.8	3.9	c	1.91777	
6	6	1.7	R^2	0.983184	
7	7	0.2	Resid	{0.1822260392222,-0.36...	

Exponentielle Regression

	A xl	B yl	C
=			
1	0	1.4	
2	2.5	2.7	
3	5	3.7	
4	7.5	6.9	
5	9.1	9.1	



Wenn es auch der betrachtete Sachverhalt zulässt, könnte hier ein exponentieller Zuwachs vermutet werden.

Das Ziel ist also, die Gleichung einer Exponentialfunktion zu finden, deren Graph eine optimale Annährung an die Punkte des Streudiagramms ergibt.
Hierbei kann man die exponentielle Regression in eine lineare überführen.
Dazu wird die Ansatzgleichung logarithmiert.

$y = a \cdot e^{k \cdot x}$

$\ln y = k \cdot x + \ln a$

$g(x) = \ln y = k \cdot x + \ln a$

Damit ergibt sich folgende Vorgehensweise:

$g(x) := k \cdot x + \ln(a)$ ▸ *Fertig*

$f(k,a) := \text{sum}\left(\left(g(xl) - \ln(yl)\right)^2\right)$ ▸ *Fertig*

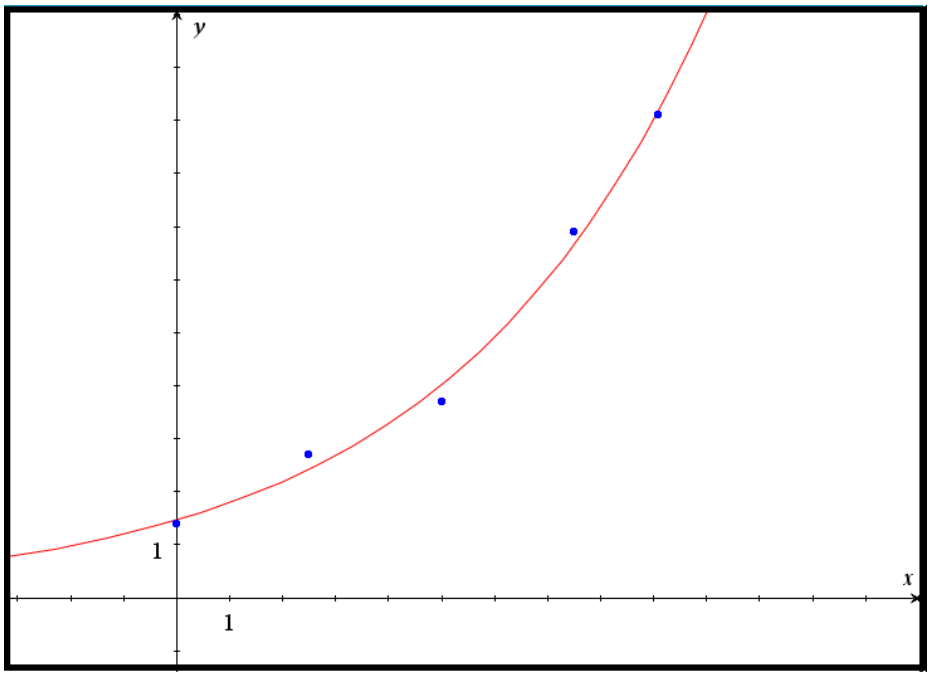
$\text{solve}\left(\begin{cases} \frac{d}{da}\left(f(k,a)\right) = 0 \\ \frac{d}{dk}\left(f(k,a)\right) = 0 \end{cases}, \{k,a\}\right)$ ▸ $a = 1.46519$ and $k = 0.201988$

$g1(x) := 1.46519 \cdot e^{0.201988 \cdot x}$ ▸ *Fertig*

oder

$e^{0.201988}$ ▸ 1.22383

$g2(x) := 1.46519 \cdot (1.22383)^x$ ▸ *Fertig*



Mit der statistischen Berechnung „Exponentielle Regression“ erhält man ebenfalls dieses Ergebnis.

	A xl	B yl	C	D
=				=ExpReg('xl,'yl,1): CopyV
1	0	1.4	Titel	Exponentielle Regression..
2	2.5	2.7	RegEqn	a*b^x
3	5	3.7	a	1.46519
4	7.5	6.9	b	1.22383
5	9.1	9.1	r^2	0.990275
6			r	0.995125
7			Resid	{-0.065194359732212,0...
8			ResidTra...	{-0.045515665807287,0...

Probieren Sie einmal aus, was man erhält, wenn man die hier vorgestellte Methode direkt auf den exponentiellen Ansatz anwendet.

Quellen:

1. B. Berchthold, „Lineare Regression und Korrelation“ (Arbeitsblätter),
   https://www.mathematik.ch/anwendungenmath/Korrelation_Regression/Regression_Korrelation.pdf

2. Evelyn Fiedler, Dr. Kinga Szücs, Dr. Wilfried Zappe,
   Lehrerhandreichung zur Kompetenzentwicklung entsprechend des Lehrplans „Der Schüler kann Zusammenhänge zwischen zwei Datenreihen durch eine geeignete Funktion näherungsweise beschreiben.“ – TI-NspireCX CAS
   Mathematik – Klassenstufe 7 – 12
   Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien Bad Berka,
   Juli 2014
   https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=5037&tspt=%3A%3B%3AbackUrl%3A%3D%3A%2Fmedia%2Fsearch%3Ftspt%3Dnosearch&csthl=Lehrerhandreichung

   Hinweis:
   Der in der Lehrerhandreichung angegebene Link
   http://www.uni-magdeburg.de/exph/mathe_gl/regression.pdf
   ist nicht mehr möglich.