Reinhold Hoffmann

Lineare Regression - von der Black-Box zur White-Box

Anwendungen von Modulen werden unterrichtlich oft nur dann akzeptiert, wenn die zugrunde liegenden Algorithmen verstanden sind, und die zugehörigen elementaren Operationen (jederzeit) ohne Hilfsmittel realisiert werden können. Ein in diesem Sinne verständiger Umgang mit den verfügbaren Regressionen ist aber schulisch kaum zu vermitteln. Dennoch sollte auf Regressionen als „Black-Boxes“ nicht verzichtet werden. Sie gewähren Schülerinnen und Schülern Einsichten in funktionale Zusammenhänge zwischen Größenpaaren, die auf herkömmliche Weise nur mit erhöhtem Zeit- und Arbeitsaufwand zu bewirken sind. So führen Weg-Zeit-Experimente, Abkühlungsversuche, Optimierungsaufgaben, Wachstumsuntersuchungen, Zufallsversuche u.a. zu endlich vielen Messpunkten $P(x_1;y_1)$,$P(x_2;y_2)$, ... , $P(x_n;y_n)$. Sie werden im Data-Menü des TI-89/92 Plus/Voyage 200 tabellarisch erfasst und durch Regressionen ausgewertet, ohne dass das Verfahren, wohl aber die Wahl der speziellen Regression kritisch zu hinterfragen ist.

Eine lineare Regression

Eine Messwertreihe mit 16 Messpunkten:

DATA	i	xi	yi				
	c1	c2	c3	c4	c5	c6	c7
1	1.00	19	12				
2	2.00	17	12				
3	3.00	15	11				
4	4.00	18	14				
5	5.00	16	14				
6	6.00	21	16				
7	7.00	24	15				
8	8.00	13	12				
9	9.00	12	10				
10	10.0	22	17				
11	11.0	11	12				
12	12.0	20	14				
13	13.0	22	14				
14	14.0	9	9				
15	15.0	10	11				
16	16.0	11	9				
17							

r11c1=11.

MAIN DEG EXACT PAR

Abb. 1

(Die Betitelung der Listen c1, c2, ... mit den Namen i, xi, yi, ... soll die Übersichtlichkeit verbessern.)

Graphische und arithmetische Auswertung der 16 Messpunkte $P(x_1;y_1)$, $P(x_2;y_2)$,... , $P(x_{16};y_{16})$ mit den Möglichkeiten, die der Data/Matrix Editor bietet:
Die Lineare Regression wird im Menue Calc aktiviert. c2 und c3 sind Listennamen. Die Regressionsgleichung wird unter y1(x) gespeichert.

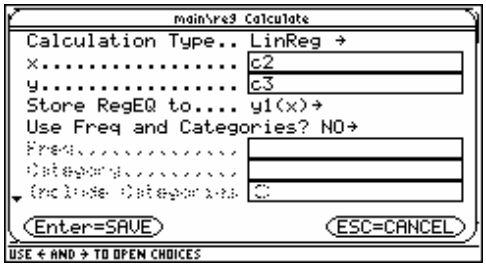


Abb. 2

Die Auswertung liefert:

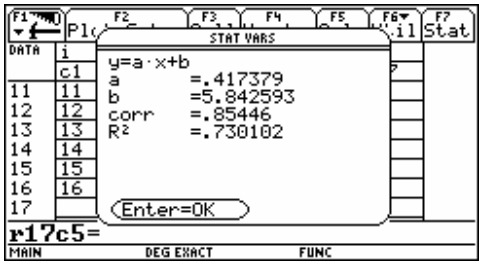


Abb. 3

Die graphische Darstellung der Punktwolke mit der zugehörigen linearen Funktion.

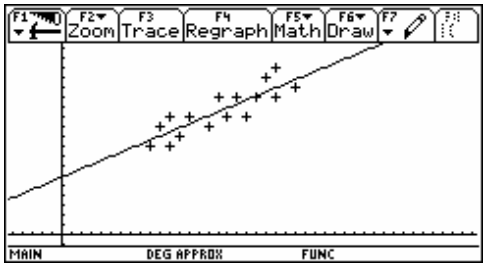


Abb. 4

Einordnung

Die Ausleuchtung der durchgeführten Regression ist Anlass, lineare Regressionsverfahren als innermathematische Modellierung aufzubereiten.

Unterrichtliche Umsetzung

Die Suche nach einer sinnvollen Ausgleichsgeraden führt unterrichtlich nach Schülerintentionen meist zur der Berechnung der zugehörigen Gleichung aus zwei „mittleren Punkten“ bzw. den Start- und Endwerten. Auch die Idee der Minimierung der Abstände wird gelegentlich im Zusammenhang mit der graphischen Darstellung des Problems genannt. Die auf Legendre und Gauß zurückgehende Methode, aus gegebenen Messwerten Werte zu berechnen, für welche die

Summe der Fehlerquadrate $\sum_{i=1}^{n}(f(x_i)-y_i)^2$ minimal

ist, muss als Modellierungsansatz vorgegeben und von anderen Möglichkeiten abgegrenzt werden.
Im speziellen Fall ist zur Beschreibung der Ausgleichsgeraden eine Gleichung f(x) = mx + b zu

finden, die der Setzung genügt, dass

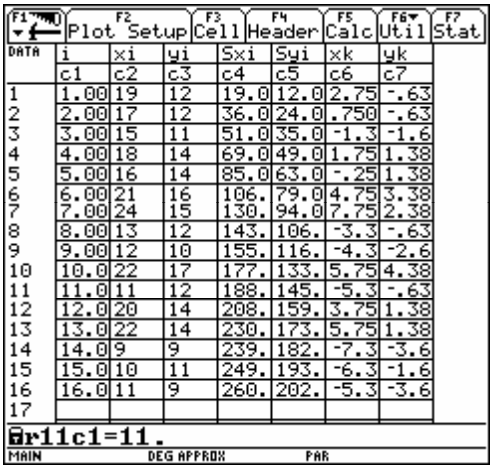
$\sum_{i=1}^{16}(m \cdot x_i + b - y_i)^2$ minimal ist.

Alle Bemühungen, den Summenterm zur Bestimmung von m und b brauchbar umzuwandeln, führen zu komplexen Ausdrücken, die eine elementare Lösung nicht zulassen.

Mit $x_q = \frac{1}{16}\sum_{i=1}^{16} x_i$ und $y_q = \frac{1}{16}\sum_{i=1}^{16} y_i$ wird stattdessen der sogenannte Regressionsschwerpunkt $P(x_q;y_q)$ als ein Punkt der Ausgleichsgeraden festgelegt.
Zur Bestimmung der Geradensteigung wird eine Koordinatentransformation so durchgeführt, dass der Regressionsschwerpunkt $P(x_q;y_q)$ Ursprung des neuen Koordinatensystems ist. Die Richtung der Koordinatenachsen und der Maßstab bleiben unverändert.
Es gelten die Transformationsgleichungen:

$x_i - x_q = x_k$ und $y_i - y_q = y_k$.

Mit c4 = cumSum(c2) und c5 = cumSum(c3) lassen sich x_q = 260/16 und y_q = 202/16 einfach bestimmen und damit die transformierten Koordinaten x_k und y_k einfach in c6 und c7 berechnen.



	i	xi	yi	Sxi	Syi	xk	yk
	c1	c2	c3	c4	c5	c6	c7
1	1.00	19	12	19.0	12.0	2.75	-.63
2	2.00	17	12	36.0	24.0	.750	-.63
3	3.00	15	11	51.0	35.0	-1.3	-1.6
4	4.00	18	14	69.0	49.0	1.75	1.38
5	5.00	16	14	85.0	63.0	-.25	1.38
6	6.00	21	16	106.	79.0	4.75	3.38
7	7.00	24	15	130.	94.0	7.75	2.38
8	8.00	13	12	143.	106.	-3.3	-.63
9	9.00	12	10	155.	116.	-4.3	-2.6
10	10.0	22	17	177.	133.	5.75	4.38
11	11.0	11	12	188.	145.	-5.3	-.63
12	12.0	20	14	208.	159.	3.75	1.38
13	13.0	22	14	230.	173.	5.75	1.38
14	14.0	9	9	239.	182.	-7.3	-3.6
15	15.0	10	11	249.	193.	-6.3	-1.6
16	16.0	11	9	260.	202.	-5.3	-3.6
17							

r11c1=11.

MAIN DEG APPROX PAR

Abb. 5

Bei einer Koordinatentransformation werden alle Punkte der Ebene in zwei verschiedenen Koordinatensystemen dargestellt, c6 und c7 sind die Listen der Koordinaten im transformierten System. Die zum neuen Listenpaar (c6,c7) zugehörige Ausgleichsgerade verläuft durch den Ursprung und genügt der Gleichung g(x) = mx. Mit dem TI-92/Voyage 200 lassen sich die zum Listenpaar (c6,c7) zugehörigen Punkte nun im selben Koordinatensystem darstellen. Es handelt sich um eine Abbildung der Ebene (Punkte).

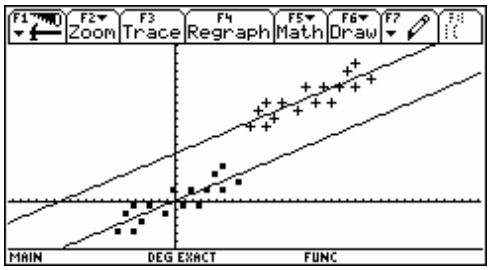


Abb. 6

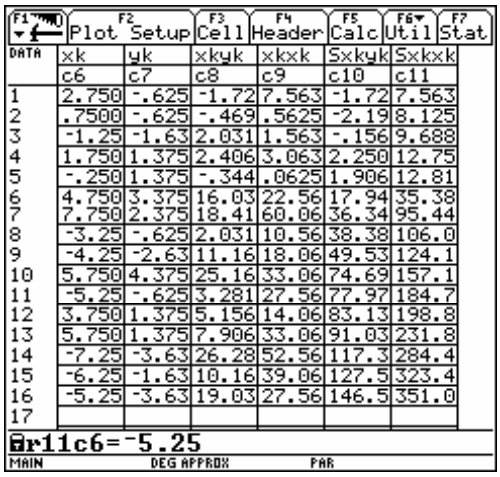
Somit ist die Summe der Fehlerquadrate

$\sum_{k=1}^{16}(g(x_k) - y_k)^2 = \sum(mx_k - y_k)^2$ zu minimieren.

Diese Aufgabe ist einfacher, denn

$\sum(mx_k - y_k)^2 = m^2 \cdot \sum {x_k}^2 - 2m \cdot \sum x_k \cdot y_k + \sum {y_k}^2$

ist ein quadratischer Term in m und sein Graph entspricht einer nach oben geöffneten Parabel. Die erste Koordinate des Scheitelpunktes bestimmt also die Steigung m so, dass diese Summe minimal ist. Damit ergibt sich der Regressionskoeffizient der Linearen Regression als $m_{min} = \frac{\sum x_k \cdot y_k}{\sum {x_k}^2}$.



	xk	yk	xkyk	xkxk	Sxkyk	Sxkxk
	c6	c7	c8	c9	c10	c11
1	2.750	-.625	-1.72	7.563	-1.72	7.563
2	.7500	-.625	-.469	.5625	-2.19	8.125
3	-1.25	-1.63	2.031	1.563	-.156	9.688
4	1.750	1.375	2.406	3.063	2.250	12.75
5	-.250	1.375	-.344	.0625	1.906	12.81
6	4.750	3.375	16.03	22.56	17.94	35.38
7	7.750	2.375	18.41	60.06	36.34	95.44
8	-3.25	-.625	2.031	10.56	38.38	106.0
9	-4.25	-2.63	11.16	18.06	49.53	124.1
10	5.750	4.375	25.16	33.06	74.69	157.1
11	-5.25	-.625	3.281	27.56	77.97	184.7
12	3.750	1.375	5.156	14.06	83.13	198.8
13	5.750	1.375	7.906	33.06	91.03	231.8
14	-7.25	-3.63	26.28	52.56	117.3	284.4
15	-6.25	-1.63	10.16	39.06	127.5	323.4
16	-5.25	-3.63	19.03	27.56	146.5	351.0
17						

r11c6=-5.25

MAIN DEG APPROX PAR

Abb. 7

Im Data/Matrix Editor wird in Spalte c8 die Folge der $x_k y_k$ und in Spalte c9 die der ${x_k}^2$ eingetragen. Die entsprechenden kumulativen Summen stehen in Spalten c10 und c11. Damit berechnet sich der Regressionskoeffizient als m = 146,5/351 = 0,417379.
Zur Bestimmung des fehlenden Achsenabschnittes wird entsprechend der Transformation die Gleichung $f(x) = m(x - x_q) + y_q$ herangezogen.
Dies ergibt f(x) = 0,417379(x - 16,25) + 12,625 = 0,417379 x + 5,84259, in Übereinstimmung mit der vom TI-92/Voyage 200 ermittelten Regressionsgleichung (vgl. Abb. 3).

Die Güte der Angleichung

Wie aber sind *corr* und R^2 zu interpretieren? Je besser unterrichtlich ermittelte Regressionsgraphen die Messpunkte treffen, desto mehr nähern sich *corr*

bzw. R^2 dem Wert 1. Sie geben offensichtlich Auskunft über die Güte der Angleichung (vgl. Abb. 3). Hier ist zu thematisieren, dass die Wurzel aus dem Produkt der Steigungen der Ausgleichsgeraden zu den jeweiligen Listenpaaren (x;y) bzw. (y;x) als Maß für die die Abhängigkeit der Größen x und y vereinbart wird und Korrelationskoeffizient heißt. Es gilt:

$$R^2 = \frac{\sum x_k y_k \cdot \sum x_k y_k}{\sum x_k^{\,2} \cdot \sum y_k^{\,2}}$$

Im Falle strenger Linearität genügen die Funktionen zu den Listenpaaren (x;y) und (y;x) der Relation „ist Umkehrfunktion von“, so dass die Wurzel den Wert 1 liefert.

Die Funktionen der Regressionsgeraden zu den Listenpaaren (c2;c3) und (c3;c2) der Ausgangsmesswerte (Abb. 1) genügen nicht der Relation „ist Umkehrfunktion von“, d.h. das Produkt ihrer Steigungskoeffizienten ist ungleich 1.

Darstellung der Ausgleichsgeraden zu den Listenpaaren (c2;c3) und (c3;c2):

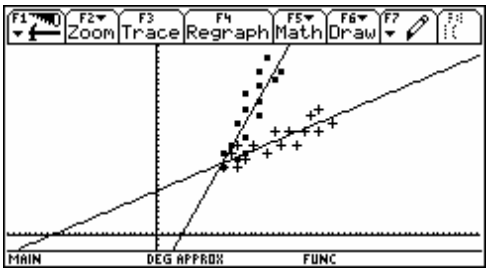


Abb. 8

In der Tabelle wird noch in Spalte c12 die Folge der $y_k^{\,2}$ und in Spalte c13 ihre kumulative Summe berechnet.

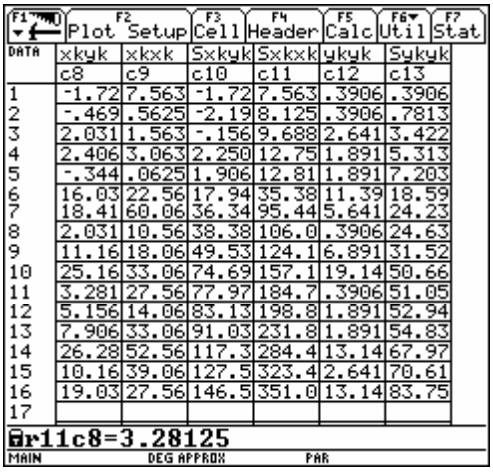


Abb. 9

Im Hauptbildschirm erhält man dann die in Abb. 3 berechneten Korrelationskoeffizienten.

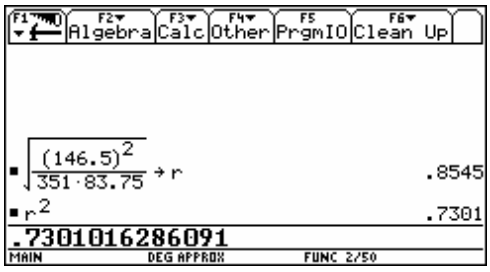


Abb. 10

Anregungen und Anwendungsbeispiele unter
www.studienseminar-leer.de
Fachleiter – Mathematik.

Der Autor:
Reinhold Hoffmann
Glansdorfer Str.47
26817 Rhauderfehn
email: K.R.Hoffmann@t-online.de

Anmerkung der Redaktion:
Nach Angaben des Autors wurde diese Unterrichtssequenz in der 9. Jahrgangsstufe, also in der Sek I behandelt.
Für die Sek II kann man vom gleichen Datenmaterial ausgehen, aber die CAS-Fähigkeiten des Voyage 200 ausnutzen. Man speichert die Koordinaten der Messpunkte in 2 Listen xi (für die Abszisse) und yi (für die Ordinate). Dann deklariert man einen Term m*x+b als Funktion g(x). Die Summe der quadratischen Abweichung $(g(xi) - yi)^2$ ist ein biquadratischer Ausdruck in m und b. Ihn weist man einer Funktion, z.B. d(m, b) zu. Setzt man nun die Ableitungen von d nach m bzw. nach b gleich Null, so erhält man ein System linearer Gleichungen in m und b. Die Lösung dieses Gleichungssystems liefert genau die vom CAS berechnete Regressionsgerade.