Korrelation und lineare Regression – Ein Zugang über das Verfahren der Standardisierung

Hans-Ulrich Lampe

Vorbemerkung

Im Zentrum der Betrachtungen stehen die Analyse von Zusammenhängen zwischen zwei Merkmalen und die Formulierung einer Trendaussage zu den Daten. Hierbei wird die oft genutzte black-box Regressionsmodul des Taschenrechners zu einer white-box.

Wichtige Eingangsvoraussetzung ist eine klare Vorstellung von der Standardabweichung. Dabei plädiere ich dafür, zunächst die Standardabweichung unabhängig von dem Aspekt Stichprobe oder Grundgesamtheit nicht in $\mathbf{s_{n-1}}$ oder $\mathbf{s_n}$ zu unterscheiden, sondern einheitlich mit „geteilt durch n" zu rechnen. Dieser Zugang ist für die Schülerinnen und Schüler natürlicher (z.B. über die Berechnung in einer Tabelle im Data-Matrix-Editor) und kann später in einer nachfolgenden Betrachtung präzisiert werden (vgl. PINKERNELL 2006). Berechnet man die Standardabweichung mit dem Taschenrechner (`0neVar` oder `TwoVar`), so muss man - jedenfalls für kleine Werte für n - aufpassen. Bei der Datenanalyse werden sowohl S_x (als s_{n-1}) als auch σ_x (als s_n) ausgegeben, wenn beim Voyage™ 200 das Betriebssystem 3.10 installiert ist. Bei älteren Betriebssystemen lässt sich σ_x mit einem Trick entlocken: im Home-Editor [2nd]GS eingeben, es erscheint σ, dann die betrachtete Größe X oder Y anfügen, also z.B. σx.

Aufgabenbeispiel

Das Verfahren der Standardisierung soll anhand des folgenden Beispiels verdeutlicht werden. In einer Datenerhebung sind Körpergröße und Körpergewicht von jeweils 10 Personen ermittelt worden.

Größe [cm]	154	166	157	169	167	176	158	173	161	175
Gewicht [kg]	51	56	52	58	68	69	49	60	52	63

Daran soll die Frage untersucht werden, ob die Faustformel „Körpergewicht gleich Körpergröße minus 100 abzüglich 15 Prozent" zutrifft (nach JAHNKE/WUTTKE 2005, S. 322).

Berechnung des Korrelationskoeffizienten

Zunächst ist die Frage zu klären, ob denn überhaupt ein Zusammenhang zwischen den Daten vorliegt, sie also korrelieren. Hierzu müssten sie einem erkennbaren Trend folgen („das Gewicht nimmt mit wachsender Körpergröße zu").

Zur besseren Beurteilung werden die Daten grafisch als sog. Punktwolken dargestellt (die Trend- oder Regressionsgerade ist hier nur zur Verdeutlichung eines linearen Zusammenhangs eingezeichnet worden, sie wird erst später thematisiert):

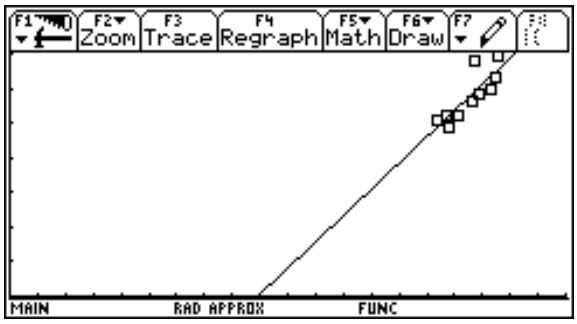

Abb. 1: $0 < x < 180$ und $0 < y < 70$

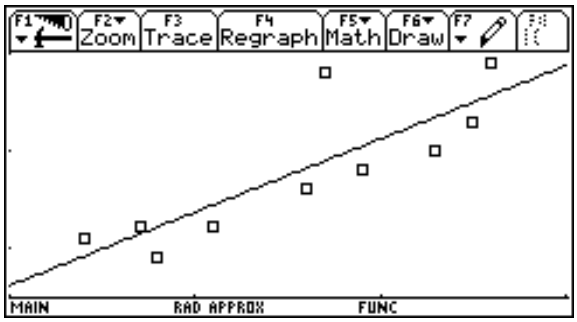

Abb. 2: $150 < x < 180$ und $45 < y < 70$

Die beiden Abbildungen kommen im Schulalltag durchaus vor, da „man bei Abb. 1 die Achsen sieht". Deutlich wird aber, dass die verschiedenen Skalierungen der Achsen eine Aussage erschweren oder verschleiern. Abhilfe schafft hier a) ein einheitliches Koordinatensystem oder b) die Standardisierung der Daten. Die Standardisierung ist ein universeller Ansatz und soll hier weiter verfolgt werden. Neben dem Vorteil einer Vereinheitlichung ist besonders das einfache, durchschaubare Rechenverfahren hervorzuheben. Außerdem kann dieses Verfahren bei der Approximation der Binomialverteilung durch die Normalverteilung wieder aufgegriffen werden.

Für die Standardisierung des Datensatzes werden im Data-Matrix-Editor weitere Spalten nach der folgenden Vorschrift angefügt, wobei die Tabellenspalte c5 zunächst außer Acht gelassen wird.

c1	c2	c3	c4	c5
x_i	y_i	$\tilde{x}_i = \frac{(x_i - \bar{x})}{s_x}$	$\tilde{y}_i = \frac{(y_i - \bar{y})}{s_y}$	$\tilde{x}_i \cdot \tilde{y}_i$

Mittelwerte und Standardabweichungen werden vorher über die Datenanalyse (`TwoVar`) berechnet:

$$\bar{x} = 165{,}6\ ;\ \bar{y} = 57{,}8\ ;\ s_x = 7{,}43\ ;\ s_y = 6{,}75$$

DATA	xi	yi	sxi	syi	sxi*syi	
	c1	c2	c3	c4	c5	c6
1	154.	51.	-1.5612	-1.0074	1.5728	8.33857
2	166.	56.	.053836	-.26667	-.01436	
3	157.	52.	-1.1575	-.85926	.994567	
4	169.	58.	.457604	.02963	.013559	
5	167.	68.	.188425	1.51111	.284732	
6	176.	69.	1.39973	1.65926	2.32252	
7	158.	49.	-1.0229	-1.3037	1.33353	
8	173.	60.	.995962	.325926	.32461	
9	161.	52.	-.61911	-.85926	.531977	
10	175.	63.	1.26514	.77037	.974627	

c3=(c1-165.6)/(7.43)

Abb. 3: Datenblatt (hier montiert)

Durch die (affin-lineare) Transformation haben die standardisierten Datensätze jeweils den arithmetischen Mittelwert 0 und jeweils die Standardabweichung 1. Die grafische Darstellung der standardisierten Daten ergibt ein einheitlicheres Bild. Man sieht auch deutlich, dass die relative Lage der Punkte in der Punktwolke erhalten geblieben ist.

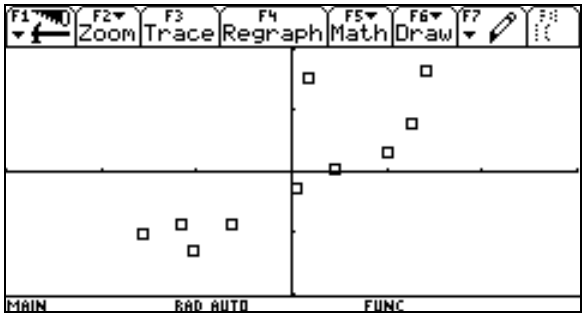

Abb. 4: Punktwolke der standardisierten Daten
$(-3 < \tilde{x}_i < 3$ und $-2 < \tilde{y}_i < 2)$

Die Daten werden in der Tabelle oder Graphik nun danach beurteilt, wie sie zu dem jeweiligen Durchschnitt als Bezugsgröße liegen: Im Vergleich zum Durchschnitt größer oder kleiner bzw. schwerer oder leichter. Hiermit wird die Abweichung vom Mittelwert betont. Weiterhin wird durch die Berücksichtigung der Standardabweichung die absolute Abweichung vom Mittelwert gewichtet. Durch die Transformation entsteht eine neue Skalierung in der Einheit *Standardabweichung*. Die erste Person ist um 1,5612 Standardabweichungen kleiner und um 1,0074 Standardabweichungen leichter als der Durchschnitt. Hiermit zeigt sich der Vorteil der Standardisierung deutlich: die Formulierung ist präziser geworden. Vergleicht man die Datensätze verschiedener Bezugsgruppen (Männer/Frauen, Japaner/Nordeuropäer u.ä.) wird der Vorteil der Standardisierung noch deutlicher.

Die Korrelation lässt sich jedoch nur qualitativ angeben. Es stellt sich die Frage nach einer quantitativen Angabe. Die Berechnung eines Korrelationskoeffizienten geht auf GALTON zurück, ist aber später als der BRAVAIS-PEARSON-Korrelationskoeffizienten bekannt geworden. Für die Berechnung des Korrelationskoeffizienten wird die Tabellenspalte c5 benötigt, also die paarweise multiplizierten standardisierten Daten. Plausibel wird dieses Vorgehen durch eine Überlegung, die anhand der Abb. 5 hergeleitet wird.

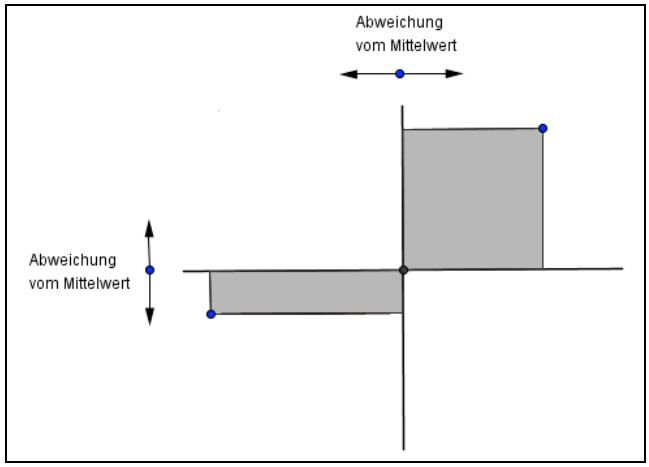

Abb. 5: Veranschaulichung zum Korrelationskoeffizienten (nach Krämer 2008, S. 189)

Zunächst betrachtet man die Verteilung der Punkte. Ein positiver Zusammenhang entsteht, wenn die Punkte überwiegend im I. und III. Quadranten liegen, ein negativer, wenn sie im II. und IV. liegen. Trägt man, wie in der Abb. 5 gezeigt, Rechtecke an den Punkten ab, so unterstützen die großen Flächen den positiven Zusammenhang stärker als kleine Flächen. Zusammengefasst ist der Korrelationskoeffizient die mittlere Fläche, welche die Punkte mit den Mittelwert-Achsen bilden. Die Flächen werden entsprechend ihres Quadranten als positiv oder negativ aufgeführt. Als Formel ergibt sich

$$r_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{x}_i \cdot \tilde{y}_i \,.$$

Für das Beispiel erhält man den Korrelationskoeffizienten $r_{xy} = 0{,}83386$ (in c6 berechnet über 1/10*sum(c5)). Diesen Wert wird auch der Taschenrechner mit dem Regressionsmodul bestätigen.

Durch eine Betrachtung wird man unmittelbar einsehen, dass für den Korrelationskoeffizienten gilt $|r_{xy}| \le 1$. Je näher der Wert an 1 oder -1 liegt, desto größer ist der positive bzw. negative lineare Zusammenhang, für $r_{xy} = 0$ oder in der Nähe davon liegt kein linearer Zusammenhang vor. (Weitere Hintergründe zur Standardisierung findet man bei KRÄMER 2008, BÜCHTER/HENN 2005 und BOROVCNIK 1994.)

Die obige Formel ist konform zu denen, die meist in den Schulbüchern stehen. Die nachfolgenden Umformungen sollen dies verdeutlichen.

$$r_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{x}_i \cdot \tilde{y}_i = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}$$

$$= \frac{s_{xy}}{s_x \cdot s_y} \qquad \text{mit } s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Der im letzten Schritt definierte Zählerterm s_{xy} heißt auch Kovarianz der Datenreihen.

Ein Vorteil der Berechnungsformel für die standardisierten Daten ist, dass man unmittelbar sehen kann, dass der Korre-

lationskoeffizient von der Zuordnungsrichtung unabhängig ist, hier also eine Symmetrie vorliegt.

Als vertiefende Übung bietet sich die Untersuchung von Körperproportionen an: Körpergröße – Ellenbogenlänge (Fingerspitze bis Ellenbogen), Schuhgröße – Handspannenweite usw.
Verdeutlicht werden sollte aber auch, dass ein erfolgreicher Nachweis einer Korrelation nichts über die Kausalität aussagt (siehe die berühmte und in jedem Übungsteil zu findende Storchenaufgabe). Kann man bei empirisch ermittelten Datenreihen x_i kontrollieren, dann hat man bei der Beurteilung eine sichere Position.

Zur Bestimmung von Korrelationskoeffizienten sollte man die Schülerinnen und Schüler nicht nur Berechnungen durchführen lassen, sondern ihnen Abbildungen von (standardisierten) Punktwolken zur Beurteilung vorlegen (z.B. BÜCHTER/HENN 2005, S. 107). An dieser Stelle ist das Angebot der Vokabeln „*stark*, *mittel*, *schwach* korreliert“ hilfreich.

Bestimmung der Regressionsgeraden

Ist der lineare Zusammenhang von Datensätzen mit dem Korrelationskoeffizienten nachgewiesen, so liegt die Bestimmung einer Trendgeraden (Regressionsgerade) nahe. Mit dieser Trendgeraden ist dann im begrenzten Rahmen eine Vorhersage möglich.

Jetzt besteht die Möglichkeit, quasi von vorn zu beginnen und mit den Originaldatenreihen zu arbeiten. Hier bieten die Schulbücher Wege in sehr unterschiedlicher Komplexität an. Nach den Mühen der Standardisierung wäre es aber angebracht, in dem Konzept zu verbleiben. Dieser Weg soll im Folgenden gezeigt werden.

In einem ersten Schritt zeichnen die Schülerinnen und Schüler „nach Gefühl“ eine Gerade durch die (standardisierte) Punktwolke. Die unterschiedlichen Variationen führen zu der Frage nach einer optimalen Gerade. An dieser Stelle sollte herausgearbeitet werden, dass der Punkt (0|0) sinnvollerweise zu der Geraden gehört muss, da er ja den Mittelwert der jeweiligen Datenreihen repräsentiert und somit eine Art Schwerpunkt bildet. Durch diese Forderung reduziert sich die Bestimmung auf eine Ursprungsgerade $y = a \cdot x$, wobei a gemäß der Konvention den Steigungswert bezeichnet und x und y (idealisierte) standardisierte Werte sind.

Zu jeder Körpergröße $\tilde{x}_i$ gehört ein standardisiertes Trendgewicht y, dass aber von dem tatsächlichen Gewicht $\tilde{y}_i$ mehr oder weniger abweichen wird. Den dabei entstehenden Fehler gilt es klein zu halten. Dies führt auf die Idee der *kleinsten Fehlerquadrate*. Hier wiederholen sich bei entsprechender Vorbereitung Überlegungen, die schon bei Lage- und Streumaßen angestellt wurden. Für die Fehlerbetrachtung wird der Ansatz aufgestellt

$$f(a) = \sum_{i=1}^{n} (\tilde{y}_i - y)^2 \quad \text{mit} \quad y = a \cdot \tilde{x}_i \text{, also gilt:}$$

$$f(a) = \sum_{i=1}^{n} (\tilde{y}_i - a \cdot \tilde{x}_i)^2 \text{ .}$$

Diese Summe gilt es zu minimieren. Dieses kann man geometrisch sehr schön darstellen und mit *Cabri Geometry* dynamisieren. Dazu schaltet man mit `F8:9Format...` die Koordinatenachsen ein und „zieht“ mit ⎋ an der x-Achse die gewünschte Skalierung (ähnlich der in Abb. 4). Dann setzt man Punkte auf das Zeichenblatt (Abb. 6) und ordnet ihnen mit `F6:5Equation & Coordinates` Koordinaten zu und verschiebt die Punkte an die passende Position der Datenpunkte. Aufgrund der Pixeldarstellung ist hier leider keine Genauigkeit von 4 Nachkommastellen zu verwirklichen. Ebenfalls mit `F5:5` lässt sich eine Ursprungsgerade einzeichnen (Abb. 7). Beim Ziehen an einem Geradenpunkt ⎋ kann man die Gerade um den Ursprung drehen. Über Senkrechten zu der x-Achse durch die Datenpunkte erhält man die Schnittpunkte mit der Ursprungsgeraden. Die Datenpunkte werden mit den Punkten der Ursprungsgeraden über Strecken verbunden, alles andere wird mit `F7:1Hyde / Show` verborgen (Abb. 8). Mit `F6:1Distance & Length` misst man die Streckenlängen und mit `F6:6 Calculate` addiert man die Quadrate dieser Streckenlängen (Abb. 9). Zieht man nun wieder an dem Geradenpunkt ⎋, so kann man die Veränderung der Quadratsumme beobachten und einen annähernd minimalen Wert finden. Wer mag, konstruiert über den Strecken mit einem Makro die zugehörigen Quadrate. Für dieses Beispiel wäre jedoch die Übersichtlichkeit verloren gegangen.

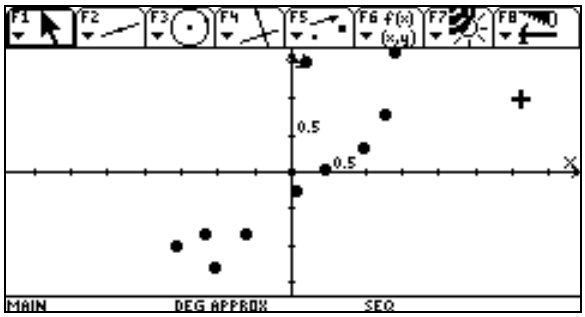

Abb. 6

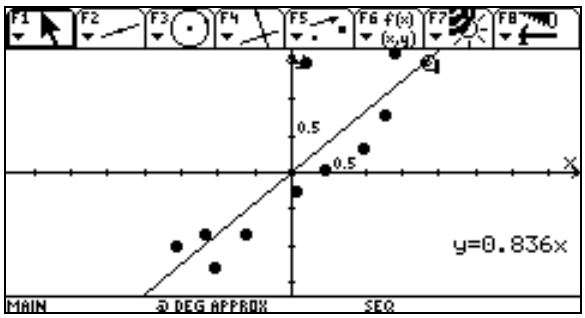

Abb. 7

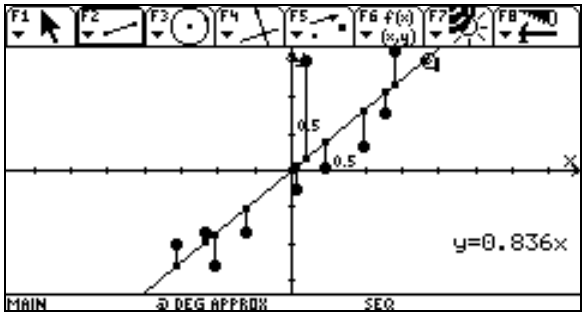

Abb. 8

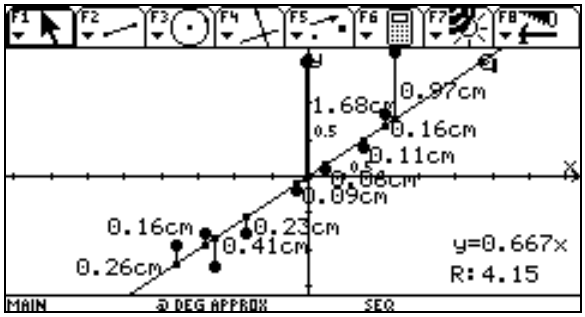

Abb. 9

Wieder stellt sich die Frage nach einem rechnerischem Weg. Da nun eine Reihe von gezielten Termumformungen im Sinne von Rückwärtsarbeiten notwendig sind, gibt man den Schülerinnen und Schülern ein Arbeitsblatt, auf dem sie die Umformungen identifizieren und kommentieren können. Selbst ein CAS ist bei den strukturierenden Termumformungen eher hinderlich. Dabei ist auch ein differenzierendes Arbeiten möglich, indem man entweder alle Umformungsschritte aufführt oder passende Lücken setzt.

$$
\begin{aligned}
f(a) &= \sum_{i=1}^{n} (\tilde{y}_i - a \cdot \tilde{x}_i)^2 \\
&= \sum_{i=1}^{n} (\tilde{y}_i^2 - 2a \cdot \tilde{x}_i \cdot \tilde{y}_i + a^2 \cdot \tilde{x}_i^2) \\
&= \sum_{i=1}^{n} \tilde{y}_i^2 - 2a \sum_{i=1}^{n} \tilde{x}_i \cdot \tilde{y}_i + a^2 \sum_{i=1}^{n} \tilde{x}_i^2 \\
&= n \cdot s_y^2 - 2a \sum_{i=1}^{n} \tilde{x}_i \cdot \tilde{y}_i + a^2 \cdot n \cdot s_x^2 \\
&= n \cdot (s_y^2 - 2a \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \tilde{x}_i \cdot \tilde{y}_i + a^2 \cdot s_x^2) \\
&= n \cdot (1 - 2a \cdot r_{xy} + a^2)
\end{aligned}
$$

$$
\begin{aligned}
f'(a) &= n \cdot (-2 \cdot r_{xy} + 2a) \\
f'(a) &= 0 \\
n \cdot (-2 \cdot r_{xy} + 2a) &= 0 \quad \Leftrightarrow \quad a = r_{xy}
\end{aligned}
$$

Dieses Ergebnis ist zunächst verblüffend, aber plausibel zu machen, da ja mit Standardabweichungseinheiten gearbeitet wurde. Der erhaltene Steigungswert gilt jedoch nur für die standardisierten Werte, deswegen wird festgehalten:

$$a_{stand} = r_{xy} .$$

Um den Steigungswert für die Originaldaten zu bekommen, muss die Standardisierung rückgängig gemacht werden. Zwei Punkte der (standardisierten) Regressionsgeraden seien $P(\tilde{x}_1 \mid \tilde{y}_1)$ und $Q(\tilde{x}_2 \mid \tilde{y}_2)$, dann gilt:

$$a_{stand} = \frac{\tilde{y}_2 - \tilde{y}_1}{\tilde{x}_2 - \tilde{x}_1} = \frac{(\frac{y_2 - \bar{y}}{s_y}) - (\frac{y_1 - \bar{y}}{s_y})}{(\frac{x_2 - \bar{x}}{s_x}) - (\frac{x_1 - \bar{x}}{s_x})} = \frac{y_2 - y_1}{x_2 - x_1} \cdot \frac{s_x}{s_y} = a \cdot \frac{s_x}{s_y}$$

$$\Rightarrow \quad a = a_{stand} \cdot \frac{s_y}{s_x} .$$

Auch diese Formel ist konform zu den Formeln der Schulbücher:

$$a = a_{stand} \cdot \frac{s_y}{s_x} = r_{xy} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} .$$

Da nach der Transformation in der Regel keine Ursprungsgerade mehr vorliegt, muss für die Gerade $y = a \cdot x + b$ noch der Wert für b ausgerechnet werden. Dies ist aber ganz einfach, da ja nach obiger Forderung auch die beiden Mittelwerte der Original-Datenreihen $\bar{x}$, $\bar{y}$ die Koordinaten eines Punktes der Geraden sein müssen. Aus

$$\bar{y} = a \cdot \bar{x} + b \quad \Rightarrow \quad b = \bar{y} - a \cdot \bar{x}$$

und damit auch die Schulbuchformel:

$$b = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} .$$

Für den Beispielsdatensatz kann nun die Regressionsgerade bestimmt werden. Ein erneuter Vergleich mit der durch das Regressionsmodul bestimmten Geradengleichung wird die Übereinstimmung zeigen. Die Nachkommastellen können entsprechend der Rundung bei der Standardabweichung geringfügig abweichen. Nun kann der Vergleich mit der Faustregel erfolgen.

Datensatz $y = 0{,}76 \cdot x - 67{,}57$

Faustregel: $y = (x - 100) - (x - 100) \cdot 0{,}15$
$= 0{,}85 \cdot x - 85$

Zurückblickend erweist sich der Weg der Standardisierung als ein vorteilhafter Weg, da man in einem Konzept verbleibt. Die Berechnungsformeln lassen sich plausibel herleiten und sind auch von ihrer Struktur her eingängiger. Die Rückführung auf die Schulbuchformeln ist nur für den Leser geschehen und müsste im Unterricht nicht thematisiert werden.

Literatur

[1] Borovcnik, M.: *Korrelation und Regression – Ein inhaltlicher Zugang zu den grundlegenden mathematischen Konzepten*; Stochastik in der Schule, Jhg. 8 (1988), Heft 1 (www.mathematik.uni-kassel.de/stochastik.schule/sisonline/08-01.htm)

[2] Büchter, A.; Henn, H.-W.: *Elementare Stochastik*, Springer Verlag, 2005 (1. Aufl.)

[3] Pinkernell, G.: S≠σ - *oder: Standardabweichung ist nicht gleich Standardabweichung*, TI-Nachrichten, 1/2006, S. 22 (www.ti-unterrichtsmaterialien.de)

[4] Jahnke, T. / Wuttke, H.: *Mathematik Stochastik* (blauer Band); Cornelsen Verlag, 2005

[5] Krämer, W.: *Statistik verstehen*; Piper, 2008

Autor:

Hans-Ulrich Lampe, Stadthagen (D)
Studienseminar Stadthagen für das Lehramt an Gymnasien
UlrichLampe@t-online.de