Regression und Korrelation: Alles klar, oder voller Tücken?

JOACHIM ENGEL, LUDWIGSBURG UND PETER SEDLMEIER, CHEMNITZ

Zusammenfassung: *Der Aufsatz zeigt einige Fehlvorstellungen zu den Begriffen Korrelation und Regression auf und weist auf Wege zu ihrer Überwindung hin*

1 Einleitung

Regression und Korrelation gehören zu den zentralen Methoden der angewandten Statistik. Aufsätze zum Lehren und Lernen dieser Methoden sind auch in *Stochastik in der Schule* nicht neu. Im Jahre 1988 waren zwei Hefte der Zeitschrift ausschließlich dieser Thematik gewidmet.

Die formale mathematische Herleitung der Formeln für Steigung und y-Achsenabschnitt der Regressionsgerade und für den Korrelationskoeffizienten ist mit guten Grundkenntnissen der Schulmathematik wenig problematisch. Dennoch ist dieser Themenbereich mit einer Reihe inhaltlicher Schwierigkeiten versehen. Wir knüpfen an die Aufsätze von Borovcnik (1988) und Sandrock (1989) an, stellen einige zentrale Fehlvorstellungen vor und weisen auf Wege zu ihrer Überwindung hin.

2 Formal alles klar …

Für gegebene bivariate, metrisch skalierte Daten (x_1, y_1), …, (x_n, y_n) ist bei der linearen Regression eine Gerade

$$y = mx + b$$

gesucht, so dass die Summe der quadrierten Abstände

$$S(m, b) = \sum_{i=1}^{n} [y_i - (mx_i + b)]^2$$

minimal wird.

Die Formeln für die Steigung und den Achsenabschnitt einer Regressionsgeraden können dank der Zerlegung der Summe der Quadrate

$$S(m, b) = \sum_{i=1}^{n} [y_i - (mx_i + b)]^2$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y} - m(x_i - \bar{x})) - (b - \bar{y} + m\bar{x})]^2$$

$$= \sum_{i=1}^{n} [y_i - \bar{y} - m(x_i - \bar{x})]^2 + n(b - \bar{y} + m\bar{x})^2$$

in ein eindimensionales Minimierungsproblem überführt werden, da zu jeder Wahl der Steigung m der zweite Summand durch geeignete Wahl von b zu Null gemacht werden kann. Daher ist m so zu wählen, dass der erste nur von m abhängende Summand minimal wird. Die optimale Wahl des y-Achsenabschnitts folgt dann unmittelbar. Wie in der Statistik üblich, bezeichnet hierbei ein Querstrich über einer Variablen das arithmetische Mittel der mit diesem Buchstaben bezeichneten Daten, d. h. $\bar{x} = \frac{1}{n}\sum x_i$. Die Bestimmung der Steigung besteht somit in der Aufgabe, den ersten Summanden zu minimieren. Dies kann nun mit Hilfe von Schulanalysis erfolgen oder auch – vorausgesetzt man erschrickt nicht vor umfangreicheren algebraischen Ausdrücken – mit Hilfe quadratischer Ergänzung gelöst werden.

Es ergeben sich sofort die bekannten Formeln

$$m = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2},\ b = \bar{y} - m\bar{x}.$$

Vertauscht man Regressor und Regressand, d. h. geht man von Daten $(y_1, x_1), \ldots, (y_n, x_n)$ aus, so erhält man für die Steigung der dazugehörigen Regressionsgeraden natürlich einen ganz analogen Ausdruck, bei dem im Nenner lediglich alle x's durch y's zu ersetzen sind. Ein in den beiden Variablen symmetrischer Ausdruck führt formal zum Korrelationskoeffizienten

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2}\ \sqrt{\sum (x_i - \bar{x})^2}}.$$

Inhaltlich zu interpretieren und geometrisch zu deuten ist der Korrelationskoeffizient, wenn man ihn als Summe der Produkte von standardisierten Daten darstellt. Diese erhält man, indem man von den Daten jeweils das arithmetische Mittel abzieht und durch die Standardabweichung dividiert

$$x_i^* = \frac{x_i - \bar{x}}{s_x},\ y_i^* = \frac{y_i - \bar{y}}{s_y} \text{ mit}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2},\ s_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2}.$$

Man beachte, dass die standardisierten Werte das arithmetische Mittel 0 und die Varianz 1 haben. Den Korrelationskoeffizienten erhält man, indem man die standardisierten Werte multipliziert, dann aufaddiert und durch $n - 1$ teilt.

Dann ist

$$r = \frac{1}{n-1}\sum x_i^* \cdot y_i^*.$$

Somit kann der Korrelationskoeffizient als Summe von gerichteten Flächen gedeutet werden. Dies ist in Abbildung 1 dargestellt. Aus illustrativen Gründen beschränken wir uns auf fünf Datenpunkte (linke Darstellung). Der Korrelationskoeffizient ergibt sich durch Aufsummieren der Flächeninhalte der vom Ursprung und den standardisierten Datenpunkten bestimmten Quadrate dividiert durch $n - 1$, wobei der Flächeninhalt von Rechtecken im 2. und 4. Quadranten negativ zu verrechnen sind.

Der Korrelationskoeffizient ist ein Maß für die Stärke des Zusammenhangs zwischen den beiden Variablen. Er bemisst aber nur den „linearen" Zusammenhang, was leicht übersehen wird und zu Fehlinterpretationen führt. Es gilt stets $|r| \leq 1$.

Abschließend sei angemerkt, dass die Regressionsgerade bezüglich der standardisierten Werte (x_1^*, y_1^*), …, (x_n^*, y_n^*) eine Ursprungsgerade mit dem Korrelationskoeffizienten r als Steigung ist, d. h. es gilt

$$y_i^* = \frac{y_i - \bar{y}}{s_y} \approx \frac{r \cdot x_i - \bar{x}}{s_x} = r \cdot x_i^*.$$

3 … aber in der Interpretation voller Tücken?

Auch wenn es Schülern mit solidem Wissen in Algebra (bzw. Grundkenntnissen der Analysis) nicht schwer fällt, die Formeln für Steigung und Achsenabschnitt der Regressionsgerade und für den Korrelationskoeffizienten herzuleiten, so sind die Bedeutung dieser Formeln und ihre Beziehungen zueinander dennoch nicht leicht zu verstehen. Bei der Interpretation von Regressionsgeraden und Korrelationskoeffizienten zeigen sich oft Missverständnisse und Fehlvorstellungen.

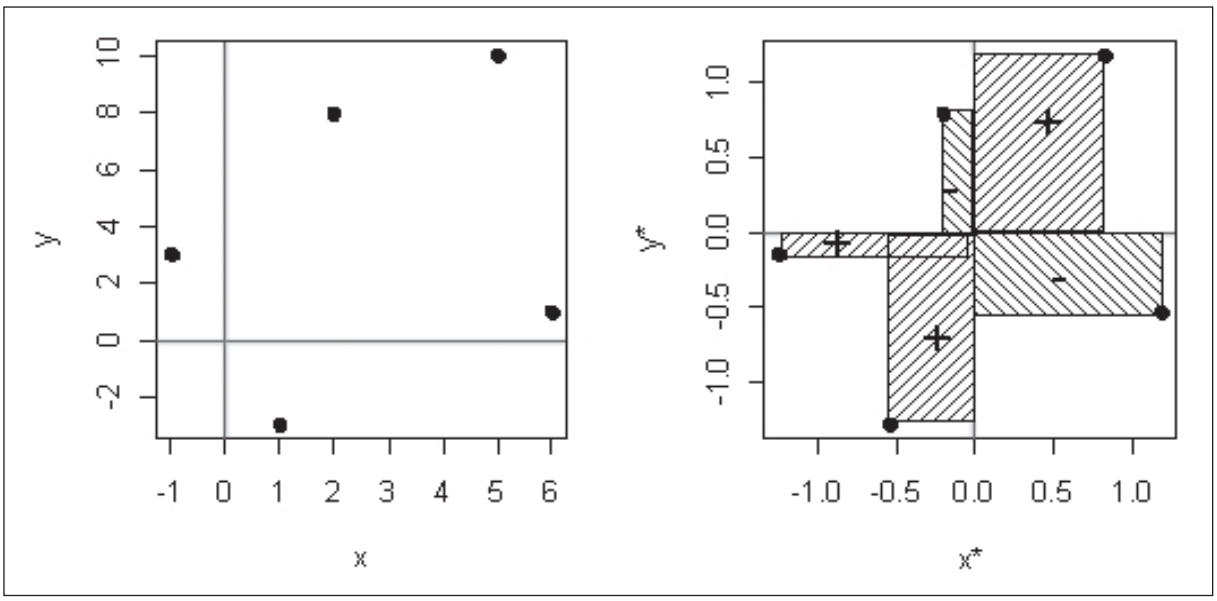

Abb. 1: Berechnung des Korrelationskoeffizienten: Originaldaten (links), normierte Daten (rechts). *r* ist die Summe der orientierten Flächeninhalte geteilt durch 4 (= *n* – 1)

3.1 Simpsons Paradox

Das Handelsblatt titelte in einem Beitrag zur Studiendauer und dem Gehalt von Akademikern mit „Methusalems machen Kasse“. Zu diesem überraschenden Ergebnis komme eine Studie über die Einstiegsgehälter von Berufsanfängern, für die die Deutsche Gesellschaft für Personalführung 44 Firmen befragt habe. Die empirischen Daten bewiesen eindeutig, dass ein Akademiker umso mehr verdient, je länger er studiert.

Wenn man etwas genauer hinschaut, so wurde die Studiendauer über *alle* Fächer betrachtet. Das Anfangsgehalt hängt aber von (mindestens) zwei Einflussgrößen ab, dem Studienfach und der Studiendauer. Betrachtet man diese Daten von Fach zu Fach, so ergibt sich sehr wohl, dass eine lange Studiendauer Gehalt und Berufsaussichten reduziert.

Anstelle der umfangreichen Originaldaten benutzen wir für die Demonstration dieser Fehlvorstellung in Abbildung 2 konstruierte Daten. Simpsons Paradox tritt auf, wenn eine verborgene dritte Variable (ein so genannter Confounder) beim Studium des Zusammenhangs zwischen zwei Größen übersehen wird, die das Vorzeichen der Korrelation umkehrt. Im Beispiel ist diese verborgene Variable das Fach, in dem der akademische Abschluss erzielt wurde. Statistisch gesehen ist die Studiendauer bis zum Diplom in Chemie deutlich länger als ein Fachhochschulabschluss im Fach Betriebswirtschaft. Innerhalb jedes einzelnen Faches jedoch verdienen diejenigen Absolventen am meisten, die zügig ihren Studienabschluss erreicht haben. Die Regressionsgerade der Daten über alle Fächer hinweg hat eine positive Steigung. Errechnet man jedoch die Regressionsgeraden separat für jedes einzelne Studienfach, so ist jedes Mal die Steigung negativ, d. h. auf jedes Fach bezogen gilt genau der umgekehrte Effekt, den man erhält, wenn man alle Daten auf einmal betrachtet ohne Berücksichtigung des „Confounders“. Entsprechend gilt für den Korrelationskoeffizienten: Studiendauer und Gehalt im ersten Berufsjahr sind positiv korreliert, wenn man alle Daten betrachtet. Differenziert man aber nach Studienfach, so ist die Korrelation bei Betrachtung der Daten jedes einzelnen Studienfachs negativ.

Wie kann man auf die Möglichkeit eines solchen Phänomens deutlich genug hinweisen? Sicher ist es

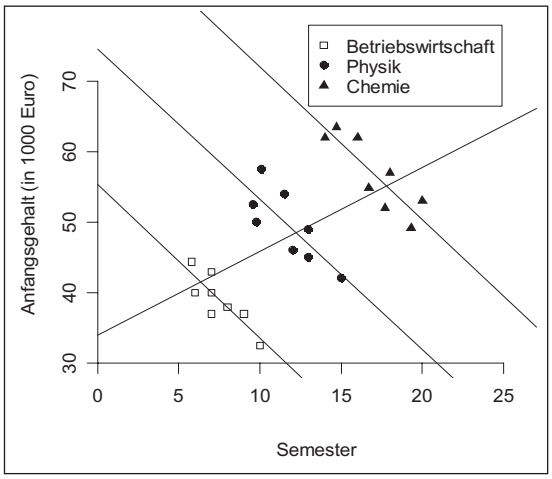

Abb. 2: Studiendauer in Semestern und Gehalt im ersten Jahr der Berufstätigkeit (in €). Die Regressionsgerade für alle Daten hat eine positive Steigung, während bei Differenzierung nach Studienfach *alle* Steigungen negativ sind

zunächst wichtig, um die Existenz von Simpsons Paradox zu wissen, und auch bei der Analyse von Zahlenbeispielen (vielleicht unter Bezugnahme auf reale Daten aus aktuellen Zeitungsberichten) Simpsons Paradox konkret erfahren zu haben. Eine wichtige Rolle spielt hier auch das Wissen um die Art der Datenerhebung: Besonders bei Beobachtungsstudien, bei denen man keinerlei Einfluss auf die Datenerzeugung hat, ist die Gefahr groß, dass eine explanative Drittvariable übersehen wurde. Hingegen ist die Gefahr eines Fehlschlusses bei kontrollierten Experimenten mit randomisierter Zuweisung zu Behandlungs- und Kontrollgruppen weitaus geringer, da dann tendenziell beide Gruppen bezüglich *aller* Variabler (außer der Art der Behandlung) ausgeglichen sind, und zwar unabhängig davon, ob sich der Forscher eines Einflusses dieser Variabler bewusst ist oder nicht. Dies ist ein wichtiger Grund, warum z. B. bei klinischen Studien die Zuweisung zu Kontroll- und Behandlungsgruppe per Zufallsentscheid durchgeführt wird.

3.2 Regressionseffekt

In einem Projekt zur Förderung rechenschwacher Kinder wurden alle Schüler einer Grundschule getestet. Die Kinder mit einem niedrigen Testwert wurden in einem speziellen Programm gefördert. Als sie nach einem Jahr wieder getestet wurden, stellte die Untersuchungsleiterin mit Befriedigung fest, dass die speziell geförderten Kinder gegenüber ihren Schulkameraden aufgeholt hatten.

In einem Schulbezirk wurden alle Zweitklässler getestet, um Kinder für ein Hochbegabten-Förderprogramm zu wählen. Als zwei Jahre später wieder alle Kinder getestet wurden, stellte man mit Erstaunen fest, dass sich die Gruppe der Hochbegabten im Vergleich zu den restlichen Schülern verschlechtert hatte.

Lässt sich aus den Testergebnissen auf einen Erfolg der Förderung rechenschwacher Kinder und das Versagen des Programms für die Hochbegabten schließen?

Es ist eine oft beobachtete Tatsache bei vielen Studien, in denen Personen zweimal (z. B. vor und nach einer Intervention) getestet werden: diejenigen, die beim ersten Test sehr gut abgeschnitten haben, sind zwar beim Wiederholungstest oft immer noch gut, aber nicht mehr ganz so gut wie beim ersten Mal. Bei den im ersten Test schlecht abschneidenden Personen ist der Trend genau umgekehrt. Sie sind beim zweiten Test oft immer noch unterdurchschnittlich, aber verbessert gegenüber dem ersten Ergebnis. Dieses Phänomen kann jedoch auftreten, ohne dass ein Wirkungseffekt vorhanden wäre, und ist dann als *Regressionseffekt* (oder *Regression zur Mitte* hin) bekannt. Es ist Ursache für viele Missverständnisse und Fehlinterpretationen bei Interventionsstudien. Sowohl Laien wie auch Forscher haben manchmal ein wenig ausgeprägtes Verständnis des Regressionseffektes und neigen dazu, ihn als einen realen Effekt anzusehen (Jennings et al. 1982; Stelzl 1982). Der Regressionsbegriff geht auf Sir Francis Galton (1822–1911) zurück, der den Zusammenhang der Körpergrößen zwischen Vätern und Söhnen untersuchte. Galton beobachtete, dass große Väter zwar immer noch überdurchschnittlich große Söhne haben, die aber in der Tendenz nicht mehr ganz so groß wie ihre Väter sind. Entsprechendes gilt für kleine Väter, deren Söhne zwar immer noch eher klein, aber größer als ihre Väter sind. Was Galton hier beobachtete, ist ein statistisches Artefakt: der Regressions-Effekt. Dieser Effekt kann niemals vermieden werden außer im Fall, wenn die beiden Variablen perfekt korrelieren.

Eine heuristische Erklärung ist wie folgt: Stellen wir uns vor, die Fähigkeit einer Person wird zweimal gemessen, ohne dass in der Zwischenzeit die Kompetenz sich geändert hätte. Ein Testergebnis hängt aber nicht allein nur vom Können der Person ab, sondern ist durch eine Reihe anderer Einflüsse (Tagesform, Auswahl der Testfragen, etc.) geprägt, die wir als zufällig ansehen wollen.

$$\textit{Beobachteter Testwert} = \textit{wahres Können} + \textit{zufällige Einflüsse}.$$

Dieses Modell zeigt seine Alltagsnähe durch gebräuchliche Redewendungen wie „Einen schlechten Tag haben“, „das Quäntchen Glück“, „das Glück des Tüchtigen“. Dies sind Umschreibungen und Interpretationen des zweiten Summanden, der zusammen mit dem wahren Können das tatsächliche Abschneiden in Testsituationen bestimmt. Bei einer Wiederholung des Tests wird das Können (nach Voraussetzung) exakt das gleiche wie beim ersten Test sein, aber die zufälligen Einflüsse werden sich nicht exakt wiederholen. Wer beim ersten Mal sehr gut war, kann entweder tatsächlich sehr gut sein oder einfach nur viel Glück gehabt haben. Wenn letzteres zutrifft, wird die Person beim Wiederholungstest deutlich schwächer abschneiden. Entsprechendes gilt in der Tendenz für Schüler, die im ersten Test sehr schlecht abgeschnitten haben. Beim Wiederholungstest werden die Extreme tendenziell zur Mitte hin gezogen, ohne dass sich irgendetwas an den wirklichen Fähigkeiten geändert haben muss.

Technologiegestützte Simulationen können zum Verständnis wesentlich beitragen und den Regressionseffekt konkret erfahrbar machen. Dazu gehen wir von z. B. 50 Zahlenwerten t_i, $i = 1, \ldots, 50$, die im Intervall [0, 1] gleichverteilt sind, aus und die die „wahren“ (entsprechend standardisierten) Fähigkeiten von 50 Schülern repräsentieren. Zusätzlich generieren wir 50 (standard-normalverteilte) Zufallswerte $e_i^{(1)}$, $i = 1\ldots, n$. Nun konstruieren wir 50 „beobachtete“ Testwerte gemäß

$$x_i = t_i + ke_i^{(1)},\ i = 1, \ldots, 50.$$

Die reelle Zahl k ist ein noch zu wählender Skalierungsfaktor, der es uns erlaubt, die Intensität des Zufallseinflusses zu variieren. Um Simulationsdaten für den Wiederholungstest zu erhalten, erzeugen wir weitere 50 normalverteilte Zufallsvariable $e_i^{(2)}$ und definieren

$$y_i = t_i + ke_i^{(2)},\ i = 1, \ldots, 50.$$

Man beachte, dass hierbei dieselben Werte t_i verwendet werden, da sich – nach Voraussetzung – beim Wiederholungstest das „wahre Können“ nicht geändert hat, dass nur andere Zufallseinflüsse wirksam werden. Schließlich berechnen wir im Streudiagramm der Punkte (x_i, y_i), $i = 1, \ldots, n$ die Steigung der Regressionsgeraden. Sie wird (tendenziell) kleiner 1 sein, und dieser Effekt ist umso stärker, je größer der Skalierungsfaktor k gewählt wurde. Dieses Phänomen lässt sich durch eine verallgemeinernde Betrachtung plausibel machen. Gehen wir zu standardisierten Werten (x_i^*, y_i^*), i = 1, …, n, über, so ist die Steigung der Regressionsgeraden der standardisierten Werte gerade der Regressionskoeffizient r. Da aber $|r| \leq 1$, bedeutet dies, dass Abweichungen von 1, 2 … Standardabweichungen vom Mittelwert der x-Variable eine um den Faktor r geringere Abweichung (gemessen in Standardabweichungen der y-Variable) zur Folge haben.

Abbildung 3 zeigt in der oberen Darstellung eine Realisierung simulierter Daten mit einem Skalierungsfaktor $k = 0{,}2$ mitsamt der Regressionsgerade $y = 0{,}6095x + 0{,}2148$.

Die mittlere Darstellung zeigt ein Histogramm der Steigungen von 100 simulierten Regressiongeraden. Die untere Darstellung zeigt Regressionsgeraden von 20 Simulationen.

Weitere Beispiele, schülernahe handlungsorientierte Aktivitäten sowie eine theoriebegründete Herleitung des Regressionseffekts finden sich bei Engel und Vogel (2001).

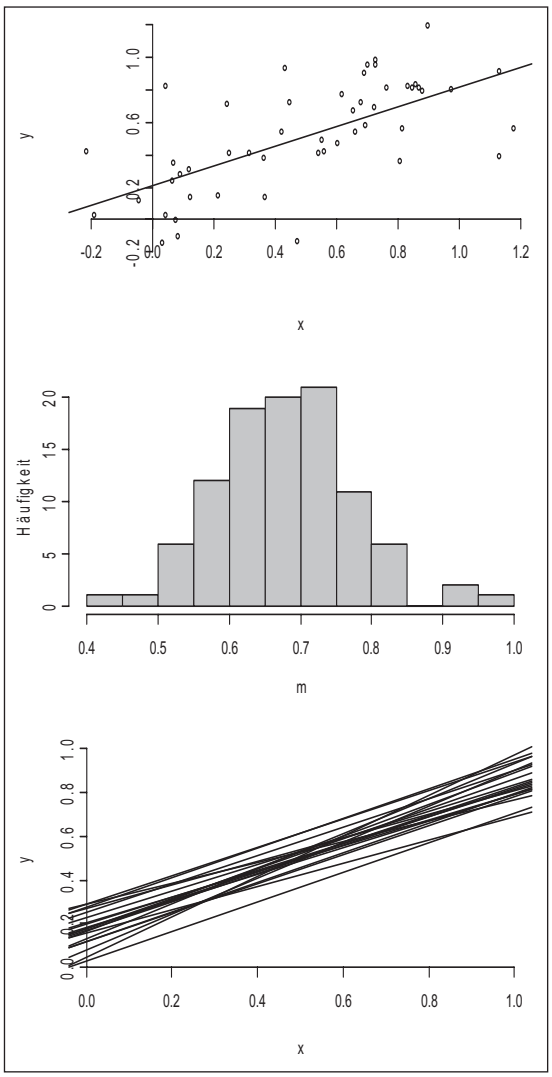

Abb. 3: Simulierte Daten zum Regressionseffekt (oben); Histogramm von 100 per Simulation erhaltenen Steigungen von Regressionsgeraden (Mitte) und Regressionsgeraden zu 20 Simulationen (unten)

3.3 Hohe Korrelation impliziert nicht die Gültigkeit des Modells

Beim Anpassen linearer Modelle trifft man immer wieder die Vorstellung an, dass ein Wert von r nahe + 1 oder – 1 ein Beweis dafür ist, dass das zugrunde liegende Modell gültig ist. Dies ist jedoch keineswegs der Fall. Der quadrierte Korrelationskoeffizient r^2 ist lediglich ein Maß dafür, wieviel der Varianz gemessen als Summe der Quadrate der y-Variablen mit dem mathematischen Modell erklärt werden kann. Der Koeffizient allein ist jedoch ungeeignet, um die Linearität zu evaluieren. Ein berühmtes Beispiel dazu stammt von Anscombe (1973), das in Abbildung 4 in leicht abgeänderter Form dargestellt wird (Daten in Anhang 1).

Alle vier Datensätze haben den gleichen Korrelationskoeffizienten $r = 0{,}99$ und die Regressionsgerade

$y = 0{,}5x$. Während Datensatz *A* ein ideales Muster mit zufällig um die Regressionsgerade gestreuten Residuen darstellt, sind die Daten in *B* unverrauscht, allerdings mit Krümmung. Die Datensätze in *C* und *D* illustrieren die Auswirkungen eines Ausreißers. Diese vier Beispiele zeigen eindrucksvoll, wie wichtig es ist, graphische Darstellungen der Daten zu betrachten und sich nicht ausschließlich auf errechnete Parameter zu verlassen. Das Beispiel zeigt, wie wichtig es ist, auch Regressionsanalysen und Kurvenanpassungen für Daten mit nichtlinearer Struktur zu behandeln.

3.4 Interpretation des Korrelationskoeffizienten

In einführenden Statistikkursen wird oft behauptet, dass die Korrelation zwischen zwei Variablen *X* und *Y* positiv ist, falls mit ansteigendem *X* auch *Y* wächst. Analog dazu wird gesagt, die Korrelation sei negativ, falls große Werte von *X* mit kleinen Werten von *Y* korrespondieren. Kharsikar und Kunte (2002) geben ein einfaches Beispiel dafür, dass diese Aussage nicht immer korrekt ist. Betrachten wir das Werfen von Münzen, bis zweimal hintereinander Zahl oder zweimal hintereinander Kopf erscheint. *X* bezeichne die Anzahl der beobachteten Ergebnisse Kopf und *Y* die Anzahl von Zahl. Die zwei Variablen *X* und *Y* sind dann negativ korreliert mit $r = -0{,}2$ (siehe Anhang 2). Welchen Wert *X* auch annimmt, *Y* wird nur einen der Werte $X-2$, $X-1$, $X+1$ oder $X+1$ annehmen können. Dies bedeutet, dass größere *X* dazu tendieren, zusammen mit größeren Werten für *Y* aufzutreten. Eine exakte Formulierung zur Charakterisierung einer positiven Korrelation lautet hingegen:

Überdurchschnittlich große X-Werte kommen tendenziell mit überdurchschnittlich großen Y-Werten,

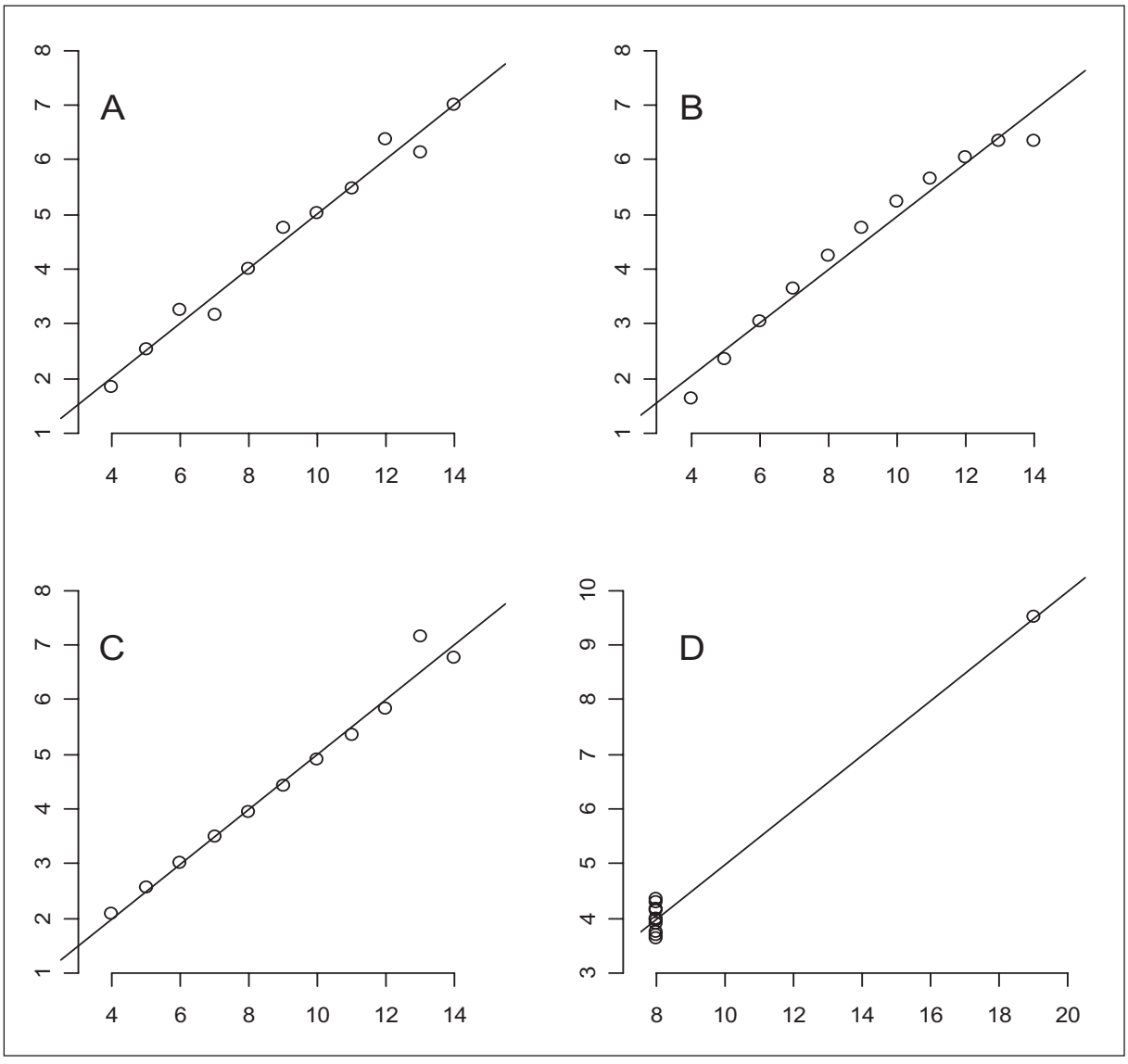

Abb. 4: Dasselbe lineare Modell angepasst an vier unterschiedliche Datensätze: Regressionsgerade und Korrelationskoeffizient sind in allen Fällen identisch $y = 0{,}5x$, $r = 0{,}99$

unterdurchschnittliche X kommen tendenziell mit unterdurchschnittlichen Y vor.

Zur Illustration dieser Aussage verweisen wir noch einmal auf die rechte Darstellung in Abbildung 1: eine positive Korrelation bedeutet in dieser Darstellung möglichst viel Fläche in den Rechtecken im 1. und 3. Quadranten. Da durch die Standardisierung der Koordinatenursprung genau dem Punkt $(\bar{x}, \bar{y})$ entspricht, entspricht dies aber gerade der oben gegebenen Interpretation: überdurchschnittliche X-Werte korrespondieren mit überdurchschnittlichen Y-Werten, entsprechendes für unterdurchschnittliche Werte. In diesem Beispiel haben 87,5 % der erwarteten Daten einen der beiden Werte oberhalb und den anderen unterhalb ihres Erwartungswertes von 1,5, liegen somit im 2. bzw. 4 Quadranten und tragen zum negativen Vorzeichen des Korrelationskoeffizienten bei. Abbildung 5 zeigt das Ergebnis einer 100-fachen Simulation von Münzwürfen, bis zweimal hintereinander Kopf oder zweimal hintereinander Wappen erscheint. Da einige wenige Beobachtungen, z. B. 2-mal Zahl, 0-mal Wappen, sehr oft vorkommen, stellen wir die Daten mit Hilfe eines sogenannten Sunflowerplots dar, das die Häufigkeiten von Beobachtungen im Streudiagramm durch die Anzahl der radial abgehenden Striche darstellt. In Abbildung 5 sehen wir, dass sehr viele Daten die Koordinaten (2/0) bzw. (0/2) haben, d. h. zweimal Wappen und keinmal Zahl bzw. kein Wappen und zweimal Zahl: Etliche weitere Daten haben die Koordinaten (1/2) und (2/1) sowie (1/3) und (3/1). Im Streudiagramm der standardisierten Daten liegen diese Punkte alle im 2. bzw. 4. Quadranten, während nur sehr wenige Daten im 1. Quadranten liegen. Daher ist der negative Korrelationskoeffizient (in der Simulation von $r = -0{,}223$) nicht mehr überraschend.

Literatur

Anscombe, F. J. (1973): Graphs in statistical analysis. *The American Statistician* 27, 17–21.

Borovcnik, M. (1988): Korrelation und Regression – Ein inhaltlicher Zugang zu den grundlegenden mathematischen Konzepten. *Stochastik in der Schule* 8 (1), 5–32.

Engel, J. und Vogel, M. (2001): Versinken wir alle im Mittelmaß? Zum Verstehen des Regressionseffektes. *MNU Der Mathematisch Naturwissenschaftliche Unterricht* 55 (7), 397–402.

Jennings, D. L., Amabile, T. M. und Ross, L. (1982): Informal covariation assessment: Databased versus theorybased judgments. In: D. Kahneman, P. Slovic & A. Tversky (Eds.): Ludgment under uncertainty: Heuristics and biases (S. 211–230). New York: Cambridge University Press..

Kharsikar, A. V. und Kunte, S. (2002): Understanding correlation. *Teaching Statistics* 24 (2), 66–67.

Sandrock, K. (1989): Ein häufiger Patzer bei einfacher linearer Regressionsanalyse. *Stochastik in der Schule* 9 (2), 40–47.

Stelzl, I. (1982). Fehler und Fallen der Statistik. Bern: Huber.

Anhang 1:

Die Daten in Abbildung 4 sind wie folgt:

x_1	y_1	y_2	y_3	x_2	y_4
10	5,01	5,23	4,89	8	3,92
8	3,99	4,23	3,95	8	3,75
13	6,12	6,35	7,15	8	4,14
9	4,76	4,75	4,42	8	4,37
11	5,47	5,65	5,36	8	4,29
14	6,99	6,62	6,77	8	4,01
6	3,25	3,03	3,02	8	3,65
4	1,85	1,62	2,08	19	9,5
12	6,37	6,03	5,83	8	3,71
7	3,16	3,65	3,48	8	4,18
5	2,54	2,35	2,55	8	3,98

Daten zu Abb. 4: (x_1, y_1), (x_1, y_2), (x_1, y_3) bzw. (x_2, y_4). Für alle gilt $r = 0{,}99$ mit Regressionsgleichung $y = 0{,}5x$

Anhang 2:

Eine Münze werde geworfen, bis zweimal hintereinander Zahl oder zweimal hintereinander Kopf erscheint. X bezeichne die Anzahl von Kopf, Y die Anzahl von Zahl. Ist $X = 0$, dann ist $Y = 2$. Wenn $X = 1$, dann ist $Y = 2$ (KZZ) oder $Y = 3$ (ZKZZ). Allgemein gilt für jeden Wert $X > 1$, dass für Y nur die Werte X-2, $X - 1$, $X + 1$ oder $X + 2$ in Frage kommen. Der genaue Wert hängt davon ab, ob die beobachtete Folge mit K oder Z beginnt und mit KK oder ZZ endet. Wenn $X = x$, so ergeben sich folgende Möglichkeiten:

1. $Y = x - 2$, falls die beobachtete Folge mit K beginnt und mit KK endet
2. $Y = x - 1$, falls die beobachtete Folge mit Z beginnt und mit KK endet
3. $Y = x + 1$, falls die beobachtete Folge mit K beginnt und mit ZZ endet
4. $Y = x + 2$, falls die beobachtete Folge mit K beginnt und mit KK endet.

Damit errechnet sich

$$P(X = 0,\ Y = 2) = \frac{1}{4},\ P(X = 1,\ Y = 2) = \frac{1}{8},$$

$$P(X = 1,\ Y = 3) = \frac{1}{16},$$

$$= \frac{7}{16} + \frac{27}{4} \cdot \frac{53}{108} - 9 \cdot \frac{7}{36} = 2.$$

Institut für Psychologie
Technische Universität Chemnitz
09107 Chemnitz
peter.sedlmeier@phil.tu-chemnitz.de