Korrelation unabhängig von Regression? Wie man die Korrelation lieber doch nicht einführen sollte

RAPHAEL DIEPGEN, BOCHUM

Zusammenfassung: *Der Autor kritisiert – in exemplarischer Auseinandersetzung mit dem Beitrag von Engel und Sedlmeier (2010) – die übliche Einführung der Korrelation unabhängig von der Regression, nämlich als standardisierte Kovarianz, und skizziert als Alternative ihre Einführung über das viel besser interpretierbare Bestimmtheitsmaß. Skizziert wird auch eine weiter reichende Alternative zur Behandlung des Regressionseffektes.*

1 Einleitung

Es ist schon seltsam: Da wollen Engel und Sedlmeier (2010) den üblichen Fehlvorstellungen zur Korrelation vorbeugen, bleiben aber in ihrer Unterrichtsskizze just bei der doch gerade so missverständniträchtigen traditionellen Einführung der Korrelation losgelöst von der Regression. Sie führen – wie man das leider nach wie vor in den meisten Lehr- und Schulbüchern findet – den Korrelationskoeffizienten als standardisierte Kovarianz ein, also als durchschnittliches Produkt von z-standardisierten x- und y-Werten – wobei sie selbst noch den Blick auf diese Durchschnittlichkeit erschweren, weil sie – obwohl dies weit und breit kein inferenzstatistischer Kontext motivieren könnte – immer den „komischen" Nenner $n-1$ statt n nutzen (und damit auch noch den falschen Eindruck erwecken könnten, die Verwendung des Nenners $n-1$ bringe Erwartungstreue nicht nur bei der Schätzung von Varianz und Kovarianz, sondern auch bei der Schätzung von Standardabweichung und Korrelation). So eingeführt lässt sich dann der Korrelationskoeffizient „geometrisch" als „Summe von gerichteten Flächen" interpretieren (eigentlich müsste es heißen: als durchschnittliche gerichtete Fläche). Nur: Was hat diese Interpretation des Korrelationskoeffizienten als durchschnittliches Produkt bzw. als durchschnittliche gerichtete Rechtecksfläche mit dem *statistischen* Konzept der Korrelation als „Maß für die Stärke des Zusammenhangs zwischen den beiden Variablen" zu tun, von dem Engel und Sedlmeier (2010, S. 14) so beredt beklagen: „Er bemisst aber nur den ‚linearen' Zusammenhang, was leicht übersehen wird und zu Fehlinterpretationen führt", denen man dann durch von Engel und Sedlmeier (2010, S. 15 ff.) skizzierte Demonstrationen im Unterricht vorbeugen solle. Wie um Gottes willen soll ein Schüler denn erkennen und beherzigen, dass der im Unterricht durch den Lehrer „abbilddidaktisch" schlicht mitgeteilte – allenfalls ex post als durchschnittliches Produkt oder als durchschnittliche gerichtete Rechtecksfläche interpretierte – Ausdruck

$$\frac{\frac{1}{n}\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}} =: r$$

ein Maß ist für die „Stärke" – was auch immer das konkret sein soll – eines *linearen* (!) „Zusammenhangs" – was auch immer das konkret sein soll?[1] Wer die Korrelation so einführt – insbesondere also ohne Bezug zur Regression –, braucht sich nicht zu wundern, dass er durch viele nachträgliche Reparaturarbeiten die schon in der Begriffseinführung angelegten Missverständnisse, insbesondere aber Überinterpretationen, mühsam beseitigen muss.

2 Das Bestimmtheitsmaß r^2 als relativer Verlustrückgang

Wenn man überhaupt den – trotz seiner Beliebtheit recht informationsarmen – Korrelationskoeffizienten

r im schulischen Unterricht thematisieren will (warum eigentlich?), dann sollte man ihn allenfalls als Wurzel aus dem Bestimmtheitsmaß (oder Determinationskoeffizienten) r^2, also der Güte eines regressionsanalytischen Modells, einführen (vgl. auch Diepgen 2000). Etwa so:

Man sammelt bei den Schülern n (x_i, y_i)-Datenpaare – etwa zu den notorischen Variablen Körpergröße X und Gewicht Y – und notiert sie für jeden Schüler auf ein Zettelchen; diese Zettelchen sammelt man dann in einem Gefäß, aus dem blind gezogen wird[2]. Vor der Ziehung muss auf den y-Wert gewettet werden, der auf dem gezogenen Zettelchen steht. Trifft man nicht, bezahlt man die quadratische[3] Differenz zwischen Tipp und Ziehung in Eurocent an den Spielleiter[4]. Was wäre der beste Tipp, wenn man auf lange Sicht im Schnitt möglichst wenig verlieren will? Einfache und schultaugliche Kleinste-Quadrate-Überlegungen liefern hierfür den Mittelwert $\bar{y}$, und die Varianz s_Y^2 ergibt sich hier als durchschnittlicher finanzieller Verlust bei Verwendung dieses optimalen Tipps $\bar{y}$.

Wie ließe sich die Wette verbessern, der mittlere Wettverlust also verringern, wenn man vor dem Tipp auf y den zugehörigen x-Wert kennte? Dann könnte man sicher einen besseren Tipp aufgrund dieses bekannten x-Wertes abgeben, etwa aufgrund einer einfachen Berechnung – denn offensichtlich sind größere Menschen „tendenziell" auch schwerer, warum auch immer. Lässt man dafür „der Einfachheit halber" nur die einfachstmögliche Berechnung zu – nämlich nur eine lineare Funktion mit zwei Parametern –, dann ergibt sich diese optimale lineare Vorhersagefunktion als

$$\hat{y} = bx + c$$

mit den üblichen Parametern der L^2-Regression, also

$$b = \frac{\frac{1}{n}\sum(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2} \text{ und } c = \bar{y} - b\bar{x},$$

und zwar – wie von Engel und Sedlmeier (2010, S. 14) zu Recht erwähnt – notfalls sogar mit bescheidenen Mitteln der Mittelstufenmathematik (Stichwort: quadratische Ergänzung)[5]. s_{XY} ist hier – zunächst – nur eine bequeme Abkürzung für das durchschnittliche Produkt der Abweichungswerte und nicht etwa eine den „Zusammenhang" von X und Y charakterisierende Kennziffer; das durchschnittliche Produkt der Abweichungswerte taucht hier halt by the way beim Minimierungsprozess auf – und zwar als „Mittelglied" einer binomischen Formel[6].

Die Frage an die Schüler nun, was sie denn für die Nennung des zugehörigen x-Wertes vor Abgabe ihres – davon dann linear abhängigen – Tipps auf y maximal zu zahlen bereit wären, was ihnen also das Wissen um x für die (lineare) Vorhersage von y bringe, führt dann sehr schnell zu der Frage, um welchen relativen Anteil sich der mittlere Wettverlust verringere, wenn man von der „blinden" Wette $\bar{y}$ zur „intelligenten" Wette in Kenntnis von x, also gemäß der berechneten Regressionsgleichung übergehe. Für diesen relativen Verlustrückgang

$$\frac{s_Y^2 - \frac{1}{n}\sum\left(\left(\frac{s_{XY}}{s_x^2}x_i + \left(\bar{y} - \frac{s_{XY}}{s_X^2}\bar{x}\right)\right) - y_i\right)^2}{s_Y^2}$$

ermittelt man dann – durch schulübliche algebraische Umformungen (insbesondere auch die binomische Formel) – den Ausdruck

$$\frac{{s_{XY}}^2}{s_X^2 s_Y^2} =: r^2.$$

Diese – missverständlich Determinationskoeffizient oder Bestimmtheitsmaß heißende – Größe r^2 ergibt sich hier ganz natürlich aus der problemorientierten Frage, um relativ wie viel der kostenträchtige mittlere quadratische Vorhersagefehler beim Übergang von der „blinden" Vorhersage $\bar{y}$ zur regressionsanalytischen Vorhersage in linearer Abhängigkeit vom Prädiktor X zurückgehe.

Hier geht es einzig und allein um die – lineare – Vorhersagbarkeit von Y durch X: *Worin* diese ihren Grund haben mag – etwa in einem kausalen Einfluss von X auf Y oder von Y auf X oder gegenseitig von X auf Y und von Y auf X oder in einem kausalen Einfluss einer oder mehrerer Drittvariablen sowohl auf X als auch auf Y – all dies ist hier zunächst irrelevant[7]: Es geht nur um Vorhersagbarkeit, nicht etwa um Beeinflussbarkeit, nicht etwa um kausale Effekte. (Konkret: Ob die Körpergröße das Gewicht beeinflusst, oder ob das Gewicht die Körpergröße, ob Körpergröße und Gewicht gemeinsam von Drittvariablen wie Genetik, Ernährung, …, beeinflusst sind, oder ob der Zusammenhang zwischen Körpergröße und Gewicht gar nur sozial konstruiert ist – etwa durch die soziale oder gesundheitspädagogische Norm, Norm- oder Idealgewicht zu halten – all dies interessiert für die Wette nicht. Und keine dieser konkurrierenden theoretischen Möglichkeiten würde durch den Wetterfolg belegt.) Ebenso irrelevant bleibt, ob der die Vorhersagbarkeit konstituierende Zusammenhang zwischen X und Y „wirklich" linear ist, wie über-

haupt die „Form“ der (x_i, y_i)-Punktwolke – und ggf. ihre „Approximierbarkeit“ durch eine Regressionsgerade – keine wesentliche Rolle spielt: Wenn die „ansteigende“ Gesamtpunktwolke in „abfallende“ Teilpunktwolken zerfällt, ist dies in diesem Zusammenhang kein Paradox, sondern schlicht irrelevant: Zum (Simpson’schen) Paradox gerät dies nur dem, der – ohne jeden Grund – die bescheidene Vorhersagbarkeit als Beeinflussbarkeit (oder vice versa Erklärbarkeit) überhöht, wer – ohne jeden Grund – von der Gesamtpunktwolke auf ihre Teilpunktwolken schließt, wer – ohne jeden Grund – unterstellt, die für den ganzen Datentopf erfolgreiche Wettstrategie mittels eines bestimmten Regressionsterms müsse auch für jedes Teildatentöpfchen erfolgreich sein.

Solchen grundlosen Überinterpretationen wird leider durch gängige Redeweisen – vor allem in den Humanwissenschaften – Vorschub geleistet. Wer – wie leider auch Engel und Sedlmeier (2010, S. 17) – r^2 als „erklärte“ Varianz bezeichnet, dürfte den Eindruck erwecken, dass hier – aus Ursachen – „erklärt“ wird, wo doch tatsächlich nur vorhergesagt wird.

3 Und die Korrelation *r*?

Es gibt von der Sache her eigentlich keinen Grund, neben dem wohl interpretierten Determinationskoeffizienten r^2 auch noch dessen – informationsidentische – Wurzel, also den Korrelationskoeffizienten r einzuführen – außer der seltsamen Beliebtheit dieses Korrelationskoeffizienten r in der mathematikfernen statistischen Praxis. (Ironie dabei: Zumeist wird dort die Mitteilung eines Korrelationskoeffizienten – etwa 0,7 – sofort mit dem Zusatz erläutert, also sei die Varianz in Y durch X zu $0{,}7^2$ sprich etwa 50 % „erklärt“.) Und es dürfte auf der Schule auch nicht ganz einfach sein, den Schülern gegenüber überzeugend zu begründen, warum man aus dem wohl definierten Konzept eines relativen Verlustrückgangs, also aus dem Prozentsatz r^2, die Wurzel ziehen sollte. Wenn man doch eine Begründung versuchen will, könnte man – gleichsam in zweidimensionaler Extension des Übergangs von der Varianz zur Standardabweichung – auf den Nutzen von r für die Standardisierung einer Punktwolke oder auf die Rolle von r in der zweidimensionalen Tschebyscheff’schen Ungleichung hinweisen – zugegebener Weise alles eher dunkel.

4 Zum Regressionseffekt

Ob man auf der Schule – wie von Engel und Sedlmeier (2010, S. 16) vorgeschlagen – den doch sehr speziellen Regressionseffekt, also die Regression zur Mitte bei Messwiederholung, thematisieren sollte, sei dahingestellt: Die mögliche Missinterpretation der statistisch zwingenden Mittelwertsveränderungen von Extremgruppen (hin zur Mitte) bei wiederholter messfehlerbehafteter Messung scheint viel weniger bildungsrelevant[8] als etwa die grundsätzliche Problematik der kausalen Überinterpretation von Korrelationen (Stichwort „Scheinkorrelation“). Wenn man daher auf der Schule den speziellen Re-gressionseffekt behandelt, sollte man dies möglichst so tun, dass dabei auch noch allgemeine Einsichten für die Regression abfallen – mehr jedenfalls als bloß die abstrakte Erkenntnis, dass im Falle gleicher Varianzen von X und Y die Steigung der Regressionsgerade dem Korrelationskoeffizienten r entspricht und daher bei nichtperfekter Korrelation betragsmäßig kleiner 1 sein muss. Etwa so:

Man erzeugt – dem Vorschlag von Engel und Sedlmeier (2010, S. 17) folgend – durch Computersimulation eine Punktwolke (nicht nur aus fünfzig, sondern aus vielen hundert oder tausend) Datenpunkten (x_i, y_i) zu den Variablen

$$X = T + E_X \text{ und } Y = T + E_Y$$

mit einer einheitlichen, $N(\mu, \sigma^2)$-verteilten „wahren“ Variablen T und zwei (untereinander und von T) unabhängigen $N(0, \sigma_E^2)$-verteilten Messfehlervariablen E_X und E_Y.[9] Man sieht dann, dass die resultierende Punktwolke ellipsenförmig ist[10]. (Genauer: Dass der „Sandhaufen“, als der diese Punktwolke interpretiert werden kann, durch ellipsenförmige Höhenlinien beschrieben ist[11].) Die Längsachse – also die lange Symmetrieachse – dieser Ellipse(n) ist selbstverständlich aufgrund der einheitlichen „wahren“ Komponente T und der nichtkorrelierenden unsystematischen Zufallsfehlerkomponenten (mit gleicher Verteilung) die 45°-Winkelhalbierende (vgl. Abb. 1).[12]

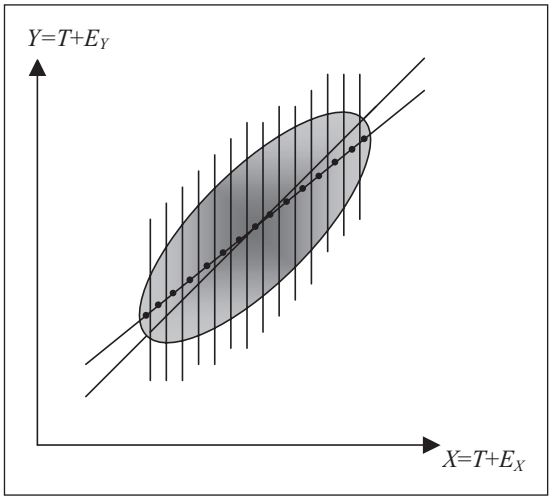

Abb. 1: „Ellipsenförmige Punktwolke“ mit Symmetrieachse und – flacherer – Regressionsgerade

Schneidet man dann diese Punktwolke längs der *Y*-Achse in feine Scheibchen und fragt sich, welches die nach dem Kleinste-Quadrate-Kriterium beste Prognose jeweils für das entsprechende kleine *X*-Intervall ist, dann ergibt sich jeweils der scheibchenspezifische Mittelwert in *Y*, in Abb. 1 jeweils als dicker Punkt markiert. Offensichtlich[13] liegen alle diese Punkte auf einer Geraden, die dann – da sie sich an demselben Kleinste-Quadrate-Kriterium orientiert – gerade die L^2-Regressionsgerade sein muss. Diese hat aber offensichtlich gegenüber der Symmetrieachse eine geringere Steigung, woraus der spezielle Regressionseffekt folgt[14]. Die zusätzliche allgemeine Einsicht zur Regression ist hier aber: Bei „symmetrischen" Punktwolken stimmt die Regressionsgerade eben nicht – wie von vielen immer spontan vermutet – mit der Symmetrieachse überein: Diese nämlich minimiert – ganz anders als eine Regressionsgerade – allenfalls die orthogonalen Abstände. Und diese Erkenntnis könnte man dann nutzen, um die – auch und gerade in Schulbüchern verbreitete – verkürzte Vorstellung zu korrigieren, bei der Regression gehe es zuvörderst um die möglichst gute („enge") Approximation einer Punktwolke durch eine Gerade.

Anmerkungen

1 Falls man an das Vorwissen anknüpfen kann, dass ein Rechteck gegebenen Umfangs dann maximalen Flächeninhalt hat, wenn es ein Quadrat ist, könnte man im Unterricht mit einer Fülle von Zusatzüberlegungen zeigen, dass die durchschnittliche gerichtete Rechtecksfläche betragsmäßig maximal und gleich der beiderseitigen Varianz – kraft Standardisierung also 1 – ist, sofern die gemittelten Rechtecke Quadrate sind, sofern also die Datenpunkte alle auf der 45°-Winkelhalbierenden liegen, insbesondere also falls mit den standarisierten *x*-Werten immer betragsmäßig gleich große *y*-Werte verbunden sind bzw. die unstandardisierten *y*-Werte eine lineare Funktion der zugehörigen unstandardisierten *x*-Werte sind. Man hätte mit diesen durchaus nicht trivialen Überlegungen aber nur gezeigt, dass die Korrelation betragsmäßig 1 ist, falls alle Datenpunkte auf einer Geraden liegen; man hätte damit nicht gezeigt, dass und in welchem konkret interpretierbaren Sinne die Korrelation ein sinnvolles Maß für den linearen Zusammenhang ist, sofern dieser eben nicht perfekt ist. Elegant ist das alles nicht.

2 Natürlich muss man dieses Wettspiel bei – üblicher – Zeitknappheit im Unterricht nicht tatsächlich durchführen; es geht nur darum, eine paradigmatische konkrete Situation für die – letztlich theoretische – Diskussion zu präsentieren, in der die Frage nach möglichst guter Vorhersage praktische Relevanz hat.

3 Ob für die statistikbasiertes Handeln sicherlich recht gut repräsentierenden Wettspiele quadratische Verlustfunktionen angemessen sind – oder doch eher lineare – sei dahingestellt. Diese Frage ließe sich aber ex post durch den Vergleich von algorithmisch bequemer L^2-Regression, die den mittleren quadratischen Vorhersagefehler minimiert, und algorithmisch aufwändiger L^1-Regression, die den mittleren absoluten Vorhersagefehler minimiert, diskutieren – und für den Fall „symmetrischer" Datenwolken als irrelevant erkennen. Erleben ließe sich dabei, dass es bei der stochastischen Modellierung nicht nur auf Realitätsangemessenheit, sondern auch auf – nicht zuletzt algorithmische – Einfachheit ankommt.

   Zu den vielleicht nicht jedermann bekannten statistiküblichen Bezeichnungen L^1 bzw. L^2: Diese beziehen sich darauf, dass das Regressionsproblem geometrisch im durch die Personen (also nicht Variablen) aufgespannten Raum interpretiert werden kann als Fällen des Lotes von einem Punkt auf eine Ebene; L^1 bzw. L^2 markiert dann die Metrik dieses Raumes, also was man jeweils unter dem Abstand *d* zweier (mehrdimensionaler) Punkte bzw. Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ versteht (also $d = \sum|x_i - y_i|$ in L^1 und $d = \sqrt{\sum(x_i - y_i)^2}$ in L^2). Die L^2-Regression ist also die „normale" Regression mit den üblichen Kleinst-Quadrate-Schätzern.

4 Zur besseren Motivation ließe sich natürlich so modifizieren: Man erhält einen festen Eurobetrag minus der quadratischen Differenz zwischen Tipp und Ziehung ☺.

5 Sicherlich ist diese mittelstufenmathematische Bestimmung der Kleinst-Quadrate-Schätzer *b* und *c* weder sonderlich elegant noch sonderlich bildungswirksam, weil sie halt nur für die einfache Regression mit einem Prädiktor funktioniert. Bei mindestens zwei Prädiktoren führt dann kein Weg mehr an der eleganteren und allgemeineren Bestimmung über die partiellen Ableitungen der Oberstufenmathematik vorbei.

6 Denn die Vorhersagefehler $(y_i - (bx_i + c))^2$ enthalten die Produkte $x_i y_i$, wie der Schüler beim Ausmultiplizieren sieht. An dieser Stelle kann man klar machen, dass in diesen Produkten $x_i y_i$ – letztlich also in der Kovarianz – alle Information über den linearen „Zusammenhang" von *X* und *Y* stecken muss, denn alle anderen Bestandteile der quadratischen Vorhersagefehler hängen ersichtlich jeweils nur von einer dieser beiden Variablen ab, also nicht von deren „Kovariieren".

7 Deshalb sollte man das Wettspiel im Unterricht tunlichst auch in umgekehrter Richtung – Wette auf *X* in Kenntnis von *Y* – untersuchen.

8 Für die Allgemeinbildung relevant ist wohl eher das theoretische Durchdenken der psychologischen oder soziologischen Wirkungen latenter Regressionseffekte im Alltag. Beispiel: Warum glauben so viele Lehrer (Eltern, Chefs, …), dass „ihrer Erfahrung nach" der Tadel schlechter Leistungen nutze, das Lob guter Leistungen aber nicht? Unterstellt man plausibler Weise, dass die wahrgenommenen („gemessenen") Leistungen mit einem zufälligen „Fehler" behaftet sind, muss sich eine

gezeigte schlechte Leistung beim nächsten Mal tendenziell verbessern, eine gute tendenziell verschlechtern, obwohl die „wahre“ Leistung gleich geblieben ist, und zwar völlig unabhängig von einem zwischenzeitlichen Tadel oder Lob des Lehrers. Dieser aber „erlebt“ und „erfährt“ immer wieder, dass sein Tadel die Leistung zumeist verbessert, sein Lob aber zumeist verschlechtert. Das Aufdecken solcher „erfahrungsgetränkter“ Fehlattributionen von Regressionseffekten auf eigenes Bewirken dürfte durchaus zur Allgemeinbildung gehören. Und das Demaskieren von Lehrerhandeln als „abergläubig“ dürfte Schülern sicherlich auch Spaß machen.

9 Tunlichst sollte die Fehlervarianz σ_E^2 viel kleiner sein als die „wahre“ Varianz σ^2. Wer will, kann dann auch noch das Verhältnis von wahrer Varianz σ^2 zur gesamten Varianz $\sigma^2 + \sigma_E^2$ mit dem psychometrischen Fachbegriff „Reliabilität“ benennen. Übrigens könnte der fleißig SiS lesende Lehrer zur Illustration der bivariaten Normalverteilung von (E_X, E_Y) das Logo des Vereins zur Förderung des schulischen Stochastikunterrichtes nutzen ☺.

10 Diese Ellipsenförmigkeit hatte Galton übrigens bei seiner „Entdeckung“ des Regressionseffektes bei Vater-Sohnes-Größen gleich mitentdeckt; sie wurde dann im Konzept der bivariaten Normalverteilung formalisiert, die sich ja gerade dadurch auszeichnet, dass die Kurven gleicher Dichte Ellipsen sind – was mathematisch übrigens nicht ganz leicht zu zeigen ist.

11 Es kommt hier im Unterricht überhaupt nicht auf – ohnehin nicht erreichbare – begriffliche Exaktheit an, also darauf, dass es tatsächlich im mathematischen Sinne exakte Ellipsen sind – von denen die meisten Schüler nur eine umgangssprachliche Vorstellung haben dürften. Die nur auf den visuellen Eindruck gestützte unterrichtliche Argumentation greift auch, wenn es sich nur „ungefähr“ um so etwas wie Ellipsen handelt.

12 Man kann dies übrigens auch ohne Computer bewerkstelligen: Man bastelt sich auf Folie ein „kreisförmiges“ (E_X, E_Y)-Punktwölkchen (mit sich um den Ursprungspunkt häufenden Punkten), markiert sich auf der 45°-Winkelhalbierenden im XY-Koordinatensystem halbwegs normalverteilt einige Punkte für die wahre Variable T, legt auf diese „wahren“ Punkte jeweils mit seinem Mittelpunkt das Fehlerpunktwölkchen als Schablone und erzeugt so die gewünschte „ellipsenförmige“ (X, Y)-Punktwolke.

13 Dies ist übrigens eine „altehrwürdige“ Erkenntnis der Kegelschnittgeometrie, üblicherweise zugeschrieben Apollonius von Perge (ca. 260–190 v. Chr.): Betrachtet man zu einem beliebigen Ellipsendurchmesser (also einer Ellipsensehne durch den Ellipsenmittelpunkt) alle parallelen Sehnen, so liegen deren Mittelpunkte ebenfalls auf einem Ellipsendurchmesser, dem sogenannten „konjugierten“ Durchmesser.

Handelt es sich übrigens bei der Punktwolke nur „ungefähr“ um eine Ellipse, dann gilt die Argumentation ersichtlich „im Wesentlichen“ immer noch. Dann liegen halt die Scheibchenmittelwertspunkte nur ungefähr auf einer Geraden, wie ja auch die Symmetrieachse dann nur ein ungefähres – damit aber nicht sinnloses – Konzept ist.

14 Für den, der es genau wissen will: Unter den im Text genannten Modellannahmen – es sind abgesehen von den Normalverteilungsannahmen die der sogenannten „klassischen Testtheorie“ der Psychometrie – ist die zweidimensionale Zufallsvariable (X, Y) bivariat normalverteilt, also $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \varrho)$-verteilt, und zwar mit den Parametern $\mu_X = \mu_Y = \mu$, $\sigma_X^2 = \sigma_Y^2 = \sigma^2 + \sigma_E^2$ und $\varrho = \sigma^2/(\sigma^2 + \sigma_E^2)$. Die Regressiongerade hat den Funktionsterm $\varrho X + (1 - \varrho)\mu$; sie beschreibt – dies wurde für diesen „scheibchenweisen“ Unterrichtsvorschlag gerade ausgenutzt – auch den bedingten Erwartungswert von Y gegeben X; ihre – der Reliabilität entsprechende – Steigung $\varrho = \sigma^2/(\sigma^2 + \sigma_E^2)$ ist kleiner 1, entsprechend 45°, solange die Messung nicht perfekt, also $\sigma_E^2 > 0$ ist. Insofern ist der Regressionseffekt hier Folge nichtperfekter Messung.

Dass für die Korrelation ϱ von X und Y $\varrho = \sigma^2/(\sigma^2 + \sigma_E^2)$ gilt, ergibt sich aus den Modellannahmen durch Anwendung der üblichen Rechenregeln für Kovarianzen („Ausmultiplizieren“), wobei sich insbesondere auch $\sigma_{XY} = \sigma^2$ zeigt, also dass die Kovarianz der „Tests“ X und Y gleich der „wahren“ Varianz von T ist. Die Identität der Korrelation ϱ mit der Reliabilität $\sigma^2/(\sigma^2 + \sigma_E 2)$ wird in der Psychometrie zur empirischen Schätzung der Reliabilität genutzt: Man schätzt die Reliabilität eines Tests X, indem man diesen (in einer Stichprobe) mit einem „parallelen“ – also (vermutlich) dieselbe „wahre“ Variable T mit derselben Genauigkeit messenden – Test Y korreliert, wobei Y häufig schlicht eine Wiederholung des Testes X zu einem späteren Zeitpunkt ist („Retestreliabilität“).

Selbstverständlich ließe sich der Regressionseffekt im Sinne einer gegenüber der Symmetrieachse einer symmetrischen Punktwolke flacheren Regressionsgerade auch unabhängig von dieser praktischen Problematik der Wiederholung fehlerbehafteter Messungen als theoretische Eigenschaft einer jeden bivariaten Normalverteilung darstellen; allein, dies wäre wohl viel weniger bildungsrelevant.

Literatur

Diepgen, R. (2000): Kovarianz oder Bestimmtheitsmaß? Didaktische Überlegungen zur Basis der Korrelation. *Stochastik in der Schule* 20 (3), 29–32.

Engel, J. & Sedlmeier, P. (2010): Regression und Korrelation: Alles klar, oder voller Tücken? *Stochastik in der Schule* 30 (2), 13–20.

Anschrift des Verfassers

Dr. Raphael Diepgen
Ruhr-Universität Bochum
Fakultät für Psychologie
44780 Bochum
raphael.diepgen@rub.de