Korrelation und Regression als Lineare Algebra?

RAPHAEL DIEPGEN, BOCHUM

Zusammenfassung: *Der Autor kritisiert den Vorschlag von Mantyk (2005), Korrelation und Regression über die Lagebeziehung zweier Vektoren im n-dimensionalen Raum einzuführen.*

1 Vorbemerkung

Als ehemaliger Stochastikschulbuch- und SiS-Autor und vor Jahren aus dem Schuldienst entlassener Mathematiklehrer schaue ich immer wieder einmal mit nostalgischem Blick in diese Zeitschrift, zuletzt in Heft 3 des Jahres 2005. Meine leise Hoffnung auf didaktische Anregungen für meine eigene Statistiklehre für Psychologiestudenten an der Universität wird freilich zumeist enttäuscht. In genanntem Heft erregte aber trotz seines statistikfernen Titels der augenscheinlich ernsthaft didaktisch motivierte Beitrag von Mantyk (2005) meine Aufmerksamkeit.

2 Statistik als Lineare Algebra

Sein Vorschlag: Es bestehe die "realisierbare Möglichkeit", die Themen Regression und Korrelation aus der in der Schule von Zeitknappheit und übriger Stofffülle bedrohten Beschreibenden Statistik im Rahmen der Linearen Algebra einzuführen. Und zwar in etwa so:

Die an n Beobachtungseinheiten erhobenen und jeweils ohne Beschränkung der Allgemeinheit auf den Mittelwert 0 zentrierten Merkmalsvariablen ("Listen") X und Y lassen sich im durch die n Fälle aufgespannten n-dimensionalen (euklidischen) Raum als Ursprungsvektoren auffassen. Dann lässt sich der Regressionskoeffizient von Y auf X als "Modifikationsfaktor" m einführen, definiert durch das Lot des Vektors $\vec{y}$ auf den Vektor $\vec{x}$, also als Lösung der Aufgabe "suche m so, dass $|\vec{y} - m\vec{x}|$ minimal ist", ebenso wie der Korrelationskoeffizient von X und Y als "Proportionalitätsmaß" r, nämlich als Cosinus des Winkels zwischen $\vec{x}$ und $\vec{y}$. Diese zunächst nur über die Lagebeziehungen ("Abstand") zwischen den n-dimensionalen Vektoren $\vec{x}$ und $\vec{y}$ motivierten Konzepte lassen sich dann – übertragen auf den durch die beiden Variablen X und Y aufgespannten zweidimensionalen Raum mit den n Beobachtungseinheiten als Punktwolke – im üblichen Sinne reinterpretieren.

Regression und Korrelation werden bei dieser "Einführung" "komplett auf die Lagebeziehungen der beiden n-dimensionalen Vektoren … reduziert. Insbesondere ist der explizite Zugriff auf die Standardabweichung nicht notwendig. Auch die mitunter schwer nachvollziehbare Bevorzugung der vertikalen Abstandsquadrate im zweidimensionalen Kontext steht auf der n-dimensionalen Schiene an keiner Stelle zur Diskussion" (Mantyk, 2005, S. 17)

3 Was als Lineare Algebra?

Was ist das, was hier die Schüler als angewandte Lineare Algebra lernen? Statistik? Wohl kaum: Es geht nämlich überhaupt nicht um statistische Probleme, etwa das der linearen Regression und Korrelation: Wie kann man in Kenntnis der Variablen X möglichst einfach (nämlich linear) und möglichst gut (nämlich: mit im Schnitt möglichst geringen - zur Not: quadrierten – Fehlern) Vorhersagen oder Wetten für die Variable Y machen? Und was bringt das im Vergleich zur blinden Wette auf Y ohne Kenntnis von X? Keine Rede von solchen echten statistischen Problemen! Stattdessen Ersatzprobleme der Form: Welchen "Abstand" haben zwei unanschauliche Vektoren im n-dimensionalen Raum? Die formal-abstrakte Strukturgleichheit von Problem und Ersatzproblem ist kein vernünftiger Grund, den Unterricht statt am Problem am Ersatzproblem zu orientieren. "Statistik" hat hier soviel Problemcharakter wie die früheren "eingekleideten" Aufgaben, in denen der Mathematiklehrer seine Schüler die von ihm versteckte Strukturgleichheit von im Unterricht gelehrtem mathematischem Modell und abstrahiertem Realitätsausschnitt nachentdecken ließ. Es gibt aber keinen guten Unterricht ohne echte Problemorientierung, auch keinen guten Statistikunterricht. Und es vermittelt dem Schüler ein falsches Bild, wenn eine Disziplin unabhängig von den Problemen dargestellt wird, deretwegen sie sich überhaupt entwickelt hat. Unterricht wird so "sinnlos".

Nimmt man den Vorschlag von Mantyk ernst, so könnte man auch das fundamentale statistische Konzept des Mittel- oder Durchschnittswertes $\overline{x}$ von n-Daten in einer Variablen X einführen über die Projektion des n-dimensionalen Vektors $\vec{x}$ auf den n-dimensionalen winkelhalbierenden Einheitsvektor $\vec{1}$: Kaum ein Schüler würde so ein angemessenes Verständnis dieser grundlegenden statistischen Begrifflichkeit entwickeln.

4 Allenfalls umgekehrt!

Könnte die von Mantyk skizzierte vektorielle Interpretation von Regression und Korrelation (in Jahrgangsstufe 12) umgekehrt eine sinnvolle Ergänzung für einen "konventionellen", nämlich an der Idee einer optimalen linearen Prognose von Y durch X – nicht etwa an der abwegigen Idee der Approximation einer Punktwolke durch eine Gerade – orientierten Unterricht über Regression und Korrelation (in Jahrgangsstufe 11) sein? Vielleicht – und zwar aus folgenden Gründen:

a. Legitimation des Quadratischen

Wer sich im konventionellen Statistikunterricht am grundlegenden Konzept der Vorhersage oder Wette orientiert, dem wird es zum Problem, warum die zu minimierenden durchschnittlichen Vorhersagefehler in der üblichen Statistik nicht betragsmäßig, sondern quadratisch zählen. (Die "vertikalen Abstandsquadrate" werden dem prognoseorientierten Statistikunterricht – anders als im obigen Zitat von Mantyk unterstellt – nicht zum Problem, weil sie "vertikal", sondern nur, weil sie "Quadrate" sind.) Warum um Gottes willen soll denn ein Prognosefehler vom Ausmaß 2 soviel zählen wie vier Prognosefehler des Ausmaßes 1? Begründet ist die Orientierung an den quadratischen Vorhersagefehlern in der traditionellen "Kleinst-Quadrate-Statistik" nicht von der typischen Problemsituation her, sondern rein innermathematisch: Das hier gestellte Fehlerminimierungsproblem lässt sich mit den traditionellen mathematischen Mitteln elegant halt nur für quadratische Fehler lösen: Denn nur hier geht es um die einfache Scheitelpunktsuche einer Parabel, nur hier um die einfache Minimierung einer quadratischen Funktion. Ist der Lehrer ehrlich, so konfrontiert er an dieser Stelle den Schüler mit der mangelnden Fähigkeit der Mathematik, das Prognoseproblem für die ursprüngliche Fragestellung mit betragsmäßigen Fehlern lösen zu können. Kurz: Traditionelle Statistik befasst sich mit quadratischen statt betragsmäßigen Fehlern schlicht deshalb, weil sie die betragsmäßigen nicht in den Griff bekommt. Eine geradezu klassische Problemverschiebung: vom exakt nicht lösbaren Ursprungsproblem hin zu einem leicht veränderten, aber lösbaren Problem! Bescheidene Einsicht in die Begrenztheiten einer Disziplin, nichtresignativer Abschied von Omnipotenzphantasien gehören zur Bildung. (Dass diese Begrenztheiten in der professionellen außerschulischen Statistik kaum mehr eine Rolle spielen, weil es längst praktikable numerische Verfahren auch für die approximative Minimierung des durchschnittlichen betragsmäßigen Prognosefehlers gibt, mag dann später Trost sein – ebenso wie die ex-post-Legitimation quadratischer Statistik durch die quadratische Struktur der Normalverteilung.)

In diesem Zusammenhang dürfte es dann für die Schüler eine verblüffende Entdeckung im auf Statistik rückblickenden Unterricht über Lineare Algebra sein, dass das Quadratische der traditionellen Regressions- und Korrelationsstatistik letztlich Pendant der quadratischen Struktur unseres euklidischen Raumes ist – Pythagoras lässt grüßen.

b. Ein knackiges Aha-Erlebnis

Das aber sollte man die Schüler dann tunlichst selbst entdecken lassen, indem man die vektorielle Darstellung von auf dem Mittelwert 0 zentrierten Merkmalsvariablen im n-dimensionalen Raum untersuchen und dort nach geometrischen Interpretationen der aus dem prognoseorientierten Kleinst-Quadrate-Statistikunterricht bekannten Konzepte suchen lässt, dies dann allerdings viel ausführlicher als von Mantyk skizziert – insbesondere für die die mittleren quadratischen Vorhersagefehler charakterisierenden oder vergleichenden Konzepte: Varianz s^2 einer Variablen gleich Quadrat über die entsprechende Vektorlänge; Determinationskoeffizient r^2 (relativer Rückgang des durchschnittlichen quadratischen Vorhersagefehlers beim Übergang von der blinden Wette $\bar{y}$ ohne Kenntnis von X zur Wette mX in Kenntnis und linearer Abhängigkeit von X) gleich dem Verhältnis des Quadrates über dem Vektor X (Ankathete) zum Quadrat über dem Vektor Y (Hypothenuse), sprich den quadrierten Kosinus des Winkels zwischen den beiden Vektoren $\vec{x}$ und $\vec{y}$. Und dann mag man die im traditionellen Statistikunterricht ohnehin nicht ganz einfach – nämlich über Standardisierungsüberlegungen, die Ungleichung von Tschebyscheff oder die "binomische" Gleichung $s^2_{X+Y} = s^2_X + 2s_{XY} + s^2_Y$ – zu legitimierenden "radizierten" Konzepte interpretieren: Standardabweichung $s = \sqrt{s^2}$ als Länge eines Vektors, Korrelationskoeffizient $r = \sqrt{r^2}$ als Kosinus des Winkels zweier Vektoren und zugleich als standardisierte Kovarianz s_{XY} bzw. standardisiertes Skalarprodukt. Es macht hier freilich wenig Sinn, wie von Mantyk vorgeschlagen nur den Korrelationskoeffizienten r zu interpretieren, nicht aber die Standardabweichung s: Beide Konzepte resultieren schließlich aus den gleichen Standardisierungsüberlegungen; Mantyks "Proportionalitätsmaß" r ist schließlich nichts anders als sein "Modifikationsfaktor" m bei auf die einheitliche Länge 1 standardisierten Vektoren.

Die Entdeckung, dass die Kleinst-Quadrate-Schätzung einer Regressionsgerade geometrisch als Lotfällen im euklidischen Raum interpretiert werden kann – nicht oft gibt es in der Schulmathematik die Gelegenheit zu einer derartigen Entdeckung der Strukturgleichheit verschiedener Modelle -, dürfte die Schüler verblüffen. Damit sie weiteren Bildungswert bekommt, sollte man schließlich auch nach dem geometrischen Äquivalent zur Minimierung absoluter Abweichungen suchen, also zur L^1- statt zur L^2-Statistik – zumindest am Beispiel des üblicherweise einzigen L^1-Konzeptes im Statistikunterricht, nämlich dem Median. Dann sieht der Schüler, dass es für diesen Median, anders als für seinen Kleinst-Quadrate-Konkurrenten Mittelwert, keine geometrische Entsprechung gibt, so lange man nicht die quadratische euklidische Metrik aufgibt. Hier wird's dann spannend, denn dann eröffnet sich der Blick darauf, dass es neben der natürlichen Abstandsnorm unseres euklidischen Anschauungsraumes noch andere Abstandsnormen geben könnte – etwa die "nichtquadratische" oder "nichtpythagoreische" City-Block-Metrik.

c. Vereinfachung durch Modellwechsel

So schön solche tiefgehenden theoretischen Entdeckungen sind, noch schöner ist es, wenn sie nicht nur theoretisches Staunen erzeugen, sondern auch Vereinfachung. Schön wäre also im Unterricht auch die Demonstration, was man denn von der geometrischen Interpretation der Regression hat – außer einem knackigen Aha-Erlebnis. Kann man in dem geometrischen Modell bestimmte statistische Sachverhalte etwas einfacher "sehen" als in der Statistik selbst? Einige Ideen:

Erstens: Dass die Varianz einer Summe von unkorrelierten Variablen gleich der Summe der Varianzen der einzelnen Summanden ist – nun, diesen für viele "varianzanalytische" Anwendungen fundamentalen Sachverhalt kann man zwar auch "zu Fuß" zeigen, aber in der geometrischen Interpretation ist er als der vertraute Satz von Pythagoras sofort präsent.

Zweitens: Kann der Determinationskoeffizient r^2 ("erklärter Varianzanteil") in der multiplen linearen Regression mit zwei Prädiktorvariablen größer sein als die Summe der Determinationskoeffizienten aus den beiden einfachen linearen Regressionen mit jeweils nur einem Prädiktor? Können also zwei Prädiktoren in der linearen Regression gemeinsam mehr "erklären" als die Summe dessen, was sie jeweils allein "erklären" können? Die Antwort ("Nein!") lässt sich im geometrischen Modell viel einfacher begründen oder "sehen" als im statistischen Modell. Versuchen Sie's.

Drittens: In der multiplen Regressionsanalyse gibt es, wenn die Prädiktorvariablen untereinander hoch korrelieren, das berühmt-berüchtigte und praktisch durchaus relevante Problem der Multikollinearität: Die zufallsstichprobengestützte Schätzung der Regressionskoeffizienten in der Population wird dann notorisch unzuverlässig, schwankt also von Stichprobe zu Stichprobe sehr. Auch dieses Phänomen – analytisch innerhalb der Inferenzstatistik nur kompliziert ableitbar – lässt sich in dem geometrischen Modell einfach demonstrieren: Wenn die hintereinander montierten Prädiktorvektoren alle fast in dieselbe Richtung zeigen, dann wirken sich leichte Zufallsschwankungen – insbesondere im Zielvektor – offensichtlich massiv darauf aus, wie weit jeweils die einzelnen Prädiktorvektoren verkürzt oder verlängert werden müssen, um zusammen dem schwankenden Zielpunkt jeweils möglichst nahe zu kommen.

Kurzum: Die geometrische Interpretation der Kleinst-Quadrate-Regression sollte exemplarisch demonstrieren, dass der Wechsel von der statistischen zur geometrischen Betrachtungsweise die Erkenntnis komplexer Sachverhalte vereinfachen kann. Vereinfachung durch Modellwechsel – dies ist aber zentrales Momentum in der Mathematik.

5 Fazit

Ganz klar: Die Anregung von Mantyk, Statistik (auch) geometrisch zu betreiben, ist berechtigt. Man muss sie nur vom Kopf auf die Füße stellen.

Literatur

Mantyk, R. (2005): Proportionalität ist "relativ". Stochastik in der Schule 25, Heft 3, 9-17.

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