Lineare Regression und Korrelation

– Artikel Nr. 5 der Statistik-Serie in der DMW –

Linear regression and correlation

Autoren S. Lange [1] R. Bender [1]

Institut [1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

Neben der univariaten, das heißt auf ein einzelnes Merkmal bezogenen Analyse von Daten aus einer klinischen Studie, ist man häufig daran interessiert, den Zusammenhang zwischen zwei (bivariat) oder mehreren (multivariat) Variablen zu betrachten. Bei Betrachtung von zwei quantitativen Merkmalen bietet sich als anschauliche, grafische Darstellungsweise die Punktwolke an, bei der die Wertepaare durch einen Punkt in einem Koordinatensystem abgebildet werden (**Abb. 1**). Damit wird sofort visuell erfassbar, ob überhaupt ein Zusammenhang besteht, und wenn ja, wie stark er ist. **Tab. 1** enthält die Werte für den systolischen Blutdruck und das Körpergewicht von 24 zufällig ausgewählten Patienten einer dermatologischen Ambulanz. **Abb. 1** zeigt die dazugehörige Punktwolke, die einen recht deutlichen Zusammenhang zwischen den beiden Merkmalen erkennen lässt.

Eine Möglichkeit, den Zusammenhang zwischen Merkmalen statistisch zu beschreiben, bietet die Regressionsanalyse. Bei der einfachen, linearen Regression, erfolgt anhand einer Geradengleichung die Vorhersage von Werten einer abhängigen Variablen aus den Werten einer als unabhängig angesehenen Variablen; es wird also ein Modell verwendet. Modelle können die Realität meist nur unvollkommen beschreiben, aber das lineare Modell hat sich für viele medizinische Anwendungen als sinnvoll und hilfreich erwiesen. Die Angemessenheit lässt sich häufig bereits bei der visuellen Betrachtung der Punktwolke beurteilen.

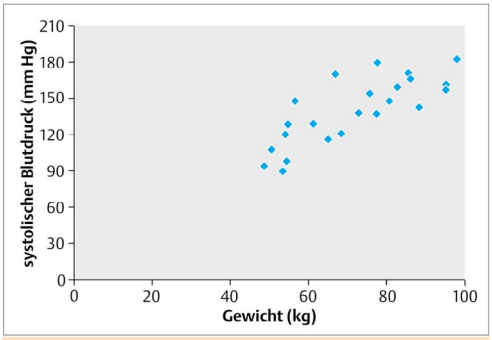

Abb. 1 Punktwolke für den Zusammenhang zwischen Körpergewicht (kg) und systolischen Blutdruck (mm Hg) von 24 zufällig ausgewählten Patienten einer dermatologischen Ambulanz.

Tab. 1 Körpergewicht (kg) und systolische Blutdruckwerte (mm Hg) von 24 zufällig ausgewählten Patienten einer dermatologischen Ambulanz.

Patientennummer	Körpergewicht (kg)	systolischer Blutdruck (mm Hg)
1	54,5	128
2	77,0	154
3	78,5	180
4	48,0	96
5	90,0	142
6	86,5	170
7	54,6	122
8	61,0	130
9	66,0	118
10	54,0	98
11	85,0	172
12	80,0	149
13	80,5	150
14	96,7	181
15	68,0	170
16	50,0	109
17	71,5	140
18	55,0	150
19	78,5	139
20	94,5	157
21	68,7	121
22	97,2	160
23	53,0	91
24	84,0	161

Schlüsselwörter
- Lineare Regression
- Korrelation
- Methode der kleinsten Quadrate
- Bestimmtheitsmaß (R^2)

Key words
- Linear regression
- Correlation
- Ordinary least squares
- Coefficient of determination (R^2)

Bibliografie
DOI 10.1055/s-2007-959028
Dtsch Med Wochenschr 2007; 132: e9–e11 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. Stefan Lange
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail stefan.lange@iqwig.de

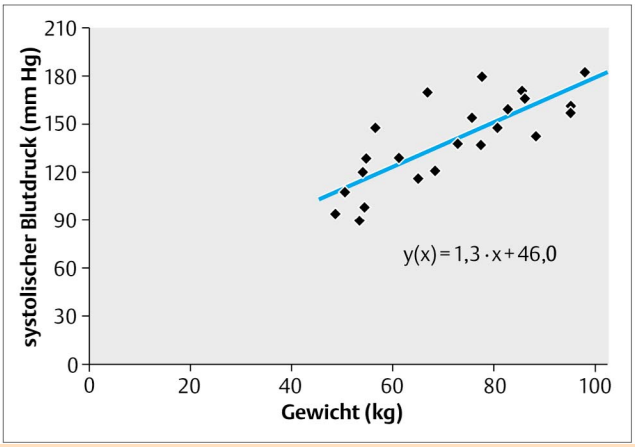

Abb. 2 Punktwolke mit Regressionsgerade und Regressionsgleichung für den Zusammenhang zwischen Körpergewicht (kg) und systolischen Blutruck (mm Hg). Das Körpergewicht ist die unabhängige (Prädiktor), der systolische Blutdruck die abhängige Variable (Outcome).

Ähnlich wie der Mittelwert im univariaten Fall einen typischen Wert der Stichprobe für das betrachtete Merkmal repräsentiert [2], liefert die Regressionsgerade einen typischen Wert der abhängigen Variablen bei gegebenem Wert der unabhängigen. Das Stichwort der „Vorhersage" macht deutlich, dass bei der Regression die Richtung des Zusammenhangs üblicherweise vorgegeben wird, das heißt es können schon a priori sinnvoll eine abhängige Variable (Outcome), deren Werte vorhergesagt werden sollen, und eine unabhängige Variable (Prädiktor) definiert werden. Für die Punktwolke wird als Konvention die abhängige Variable zumeist auf der Ordinate (y-Achse) und die unabhängige Variable auf der Abszisse (x-Achse) abgebildet.

Eine Geradengleichung benötigt zwei Parameter: Zum einen die Steigung der Geraden, die angibt, um wie viel die Werte der abhängigen Variable steigen oder fallen, wenn sich die unabhängige Variable um eine Einheit verändert, und zum zweiten der Achsenabschnitt, der das Basisniveau der abhängigen Variable angibt, wenn also die unabhängige Variable den Wert Null annimmt. Die Steigung der Geraden wird als Regressionskoeffizient bezeichnet.

In **Abb. 2** ist die Regressionsgerade mit der entsprechenden Regressionsgleichung für die Daten aus **Tab. 1** dargestellt. Es erscheint plausibel, den Blutdruck in Abhängigkeit vom Gewicht und nicht umgekehrt zu betrachten. Die Geradengleichung zeigt an, dass der Wert des systolischen Blutdrucks im Mittel um ca. 1,3 mm Hg ansteigt, wenn der Wert des Körpergewichts um 1 kg zunimmt. Bei einer 70 kg schweren Person ist mit einem Blutdruck von $70 \times 1{,}3 + 46{,}0 \approx 137$ mm Hg zu rechnen. Die am besten „passende" Regressionsgerade wird durch ein besonderes statistisches Schätzverfahren – die Kleinste-Quadrate-Methode – gefunden, und zwar ist es diejenige Gerade, bei der die Summe der quadrierten (vertikalen) Abstände zwischen den einzelnen Punkten und der Geraden minimal wird.

kurzgefasst

Mit der Regression lässt sich der Zusammenhang zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen darstellen. Die Regressionsgleichung liefert den Wert der abhängigen Variable, wenn die unabhängige bekannt ist.

Für eine weitere Quantifizierung des beobachteten Zusammenhangs zwischen den Merkmalen ist das Bestimmtheitsmaß (R^2) ein sehr anschaulicher Parameter. Hierfür muss man sich zunächst vergegenwärtigen, dass die Werte der abhängigen Variable – im Beispiel die Blutdruckwerte – bei univariater Betrachtung um ihren Mittelwert „streuen"; diese Streuung wird als Summe der quadratischen Abweichungen (der Einzelwerte von ihrem Mittelwert) ausgedrückt [3]. Die Blutdruckwerte streuen auch um die Regressionsgerade, aber in einem geringeren Ausmaß als um ihren Mittelwert. Das Bestimmtheitsmaß bezeichnet nun den Anteil, um den die Variabilität der abhängigen Variable durch die Regression, also durch die zusätzliche Betrachtung der unabhängigen Variable, vermindert wird. Als Maß für die Streuung um die Regressionsgerade wird wieder die Summe von Abweichungsquadraten (der Einzelwerte von der Regressionsgerade) verwendet. Im Beispiel ergibt sich ein Bestimmtheitsmaß von 0,62, also 62 % der „rohen" Variabilität der Blutdruckwerte aus der Stichprobe kann durch das Körpergewicht der Patienten „erklärt" werden (unter Annahme des linearen Modells).

Ein weiteres Maß für die Quantifizierung des Zusammenhangs zwischen zwei (quantitativen) Merkmalen ist der Korrelationskoeffizient **„*r*"**. Der Absolutbetrag des Korrelationskoeffizienten nach Pearson ist einfach die Wurzel aus dem Bestimmtheitsmaß: $|r| = \sqrt{R^2}$. Der Korrelationskoeffizient r kann Werte zwischen -1 (negativer Zusammenhang) und + 1 (positiver Zusammenhang) annehmen. Das Vorzeichen von r ist dasselbe wie das des Regressionskoeffizienten. Ein Korrelationskoeffizient von Null bedeutet, dass kein linearer Zusammenhang besteht. Für das Beispiel ergibt sich $|r| = \sqrt{0{,}62} \approx 0{,}79$. Anstelle des Korrelationskoeffizienten nach Pearson kann auch der Rangkorrelationskoeffizient nach Spearman berechnet werden. Er basiert, wie der Name andeutet, nicht auf den Messwerten, sondern auf den Rangzahlen, die die Messwerte in der sortierten Stichprobe einnehmen. Er ist in gleicher Weise zu interpretieren wie der Korrelationskoeffizient nach Pearson und wird insbesondere bei der Betrachtung von Scores benutzt.

Der Korrelationskoeffizient ist eines der am häufigsten, leider oft auch fälschlich eingesetzten Maße in der medizinischen Statistik. Deshalb soll auf folgende, für eine adäquate Interpretation zu beachtende Punkte hingewiesen werden:

- Der Korrelationskoeffizient, genauso wie die Regressionsgerade, liefert keine Aussage über einen kausalen Zusammenhang.
- Der Wert des Korrelationskoeffizienten kann sehr stark durch Extremwerte beeinflusst werden. Das ist leicht nachzuvollziehen, da Extremwerte die Varianz eines Merkmals stark erhöhen, und dann durch die Regression sehr viel von dieser Varianz „erklärt" werden kann.
- Die gemeinsame Betrachtung von zwei sehr unterschiedlichen Gruppen kann zu einer hohen Korrelation zwischen Merkmalen führen, obwohl innerhalb jeder Gruppe nur eine geringe oder gar keine Korrelation zwischen den Merkmalen besteht (Heterogenitätskorrelation).
- Der Korrelationskoeffizient ist kein Maß für Übereinstimmung! Seine Verwendung beim Vergleich zweier Messverfahren ist daher für sich allein nicht aussagefähig und häufig nicht adäquat [1, 4, 5]. Ein Korrelationskoeffizient nahe 1 wird auch dann erreicht, wenn zum Beispiel beim Vergleich zweier Verfahren zur Blutzuckermessung das eine Verfahren doppelt so hohe Werte liefert wie das andere.

kurzgefasst

Der Korrelationskoeffizient *r* zeigt den linearen Zusammenhang zwischen 2 Variablen. Er kann Werte zwischen -1 und + 1 einnehmen. Der Korrelationskoeffizient dient NICHT der Darstellung von kausalen Zusammenhängen oder Übereinstimmungen.

Tab. 2 zeigt wieder die Übersetzungen wichtiger Begriffe für die Interpretation englischsprachiger Studien.

Tab. 2 Übersetzungen (deutsch – englisch)

Kleinste-Quadrate-Methode	least-square-method
Vorhersage	prediction
Bestimmtheitsmaß (R^2)	coefficient of determination
(Un)abhängige Variable	(in)dependent variable
Korrelationskoeffizient	correlation coefficient
Regression	regression
Regressionsgerade	regression
(-koeffizient)	line (coefficient)
Punktwolle	scatter plot

Dieser Beitrag ist eine überarbeitete Fassung aus dem Supplement Statistik aus dem Jahr 2001.

Literatur

1 *Bland JM, Altman DG.* Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307–310
2 *Lange S, Bender R.* Median oder Mittelwert? Dtsch Med Wochenschr 2007; 132: e1–e2
3 *Lange S, Bender R.* Variablitätsmaße. Dtsch Med Wochenschr 2007; 132: e5–e7
4 *Grouven U, Bender R, Ziegler A, Lange S.* Vergleich von Messmethoden. Dtsch Med Wochenschr 2007; 132: e69–e73
5 *Richter K, Lange S.* Methoden der Diagnoseevaluierung. Internist 1997; 38: 325–336