

Lineare Regression

Roland Heynkes

18. April 2006, Aachen

Es kommt in der Natur relativ oft vor, daß zwei Größen statistisch miteinander verbunden sind. Wenn sich diese Verbundenheit mathematisch durch eine Funktion beschreiben lässt, dann lässt sich von der Art der Funktion auf die Art des Zusammenhangs schließen. Diese Funktionen können logarithmisch, exponentiell, quadratisch oder auch linear sein. Um letztere soll es in diesem Text gehen.

Inhaltsverzeichnis

1	Lineare Regression	1
1.1	Methode der kleinsten Quadrate nach Carl Friedrich Gauß	1
1.2	Vereinfachung nach dem Verschiebungssatz der Statistik	3
2	Regressionsgerade bezüglich x	4
3	Korrelation	4
	Quellenverzeichnis	5

1 Lineare Regression

Man kann die Ausprägungen zweier Merkmale entweder bei einer Reihe von Merkmalsträgern oder mehrfach beim selben Merkmalsträger erfassen. Beispiele wären die Merkmalpaare Körpergröße und Gewicht oder die Länge einer Stahlfeder und das an ihr hängende Gewicht. Dann kann man ein zweidimensionales $X|Y$ -Koordinatensystem zeichnen, in dem die beiden Achsen für zwei verschiedene stetige oder zumindest quasistetige quantitative Merkmale stehen. Für jedes Merkmalspaar wird ein Punkt in das Koordinatensystem eingezeichnet. Die Ausprägungen der beiden Merkmale werden dazu als Koordinaten verwendet.

Oft sind dann die Punkte nicht zufällig verteilt, sondern wie um eine unsichtbare Gerade herum verteilt. In solchen Fällen besteht ein linearer Zusammenhang zwischen beiden Merkmalen. Die Methode der linearen Regression dient dem Zweck, eine möglichst gut an die Punkte angepasste Gerade zu finden. Gesucht wird also die Gleichung einer Geraden, von welcher die Punkte insgesamt möglichst wenig weit entfernt liegen. [2, 3]

$$y = g(x) = m \cdot x + b \quad (1)$$

Die durch diese Geradengleichung mit der Steigung¹ a und dem y -Achsenabschnitt b charakterisierte Gerade nennt man Regressionsgerade [3, S.51]. Mit den Parametern einer solchen Geraden lassen sich noch nicht gemessene Wertepaare im Rahmen der Meßgenauigkeit im Größenbereich der gemessenen Werte vorhersagen [2, S.132]. Über den von den Daten abgedeckten Bereich hinaus sind in der beschreibenden Statistik keine Aussagen erlaubt [3, S.53]. Zu beachten ist auch, daß bei der linearen Regression davon ausgegangen wird, daß tatsächlich eine Abhängigkeit der einen von der anderen Größe besteht und nicht etwa umgekehrt [3, S.50]. Das wird schon dadurch ausgedrückt, daß man die (vermeintlich) unabhängige Größe der x -Achse zuordnet, die mutmaßlich abhängige Größe hingegen der y -Achse. Außerdem entscheidet man sich deshalb, die Abstände der gemessenen y -Koordinaten und nicht etwa die Abstände der x -Koordinaten zur Geraden zu minimieren. [2, 3].

1.1 Methode der kleinsten Quadrate nach Carl Friedrich Gauß

Die Standardmethode für die Ermittlung der günstigsten Parameter zur Anpassung einer Geradengleichung an die vorhandenen Daten ist die sogenannte Methode der kleinsten Quadrate nach Carl Friedrich Gauß. Dabei werden im Prinzip die Steigung m und der y -Achsenabschnitt b so gewählt, daß die Summe aller Quadrate der einzelnen Abweichungen zwischen den gemessenen y_i und den errechneten $y(x_i)$ bzw. $g(x_i) = m \cdot x_i + b$ minimiert wird. [3, S.50]

$$\text{Minimum} = [g(x_1) - y_1]^2 + [g(x_2) - y_2]^2 + \dots + [g(x_n) - y_n]^2 = \sum_{i=1}^n [g(x_i) - y_i]^2 \quad (2)$$

Dank einer sogenannten mehrdimensionalen Differentialrechnung soll der Nachweis möglich sein, daß dieses Minimum und damit eine optimale Anpassung der Geradengleichung genau dann erreicht wird, wenn die Parameter-Schätzwerte m und b die beiden folgenden Bedingungen erfüllen: [3, S.51]

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(\sum_{i=1}^n x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{(\sum_{i=1}^n x_i^2) - n \cdot \bar{x}^2} \quad \text{und} \quad b = \bar{y} - m \cdot \bar{x} \quad (3)$$

¹Regressionskoeffizient [1]

Die kleine zweite Gleichung 3 auf der vorherigen Seite lässt erkennen, daß die Regressionsgerade stets durch den Schwerpunkt (\bar{x}, \bar{y}) bzw. das arithmetische Mittel der Punktwolke verläuft [3, S.51]. Man erhält die Koordinaten des Schwerpunkts der Datenwolke durch getrennte Berechnung der arithmetischen Mittel der Ausprägungen beider Merkmale x und y [2, S.133].

Das Schroedel-Mathematikbuch setzt neben der Forderung nach einer Minimierung der Summe der Differenzenquadrate als zweite Bedingung für die richtige Lage der Regressionsgeraden einfach voraus, daß sie durch den Schwerpunkt $M(\bar{x}|\bar{y})$ verläuft. Daraus folgt, daß die auch Ausgleichsgerade oder Trendlinie genannte Regressionsgerade $g(x) = m \cdot x + b$ natürlich auch im Schwerpunkt gilt, wo man die Gleichung nach b auflösen kann. [2, S.134]

$$\bar{y} = m \cdot \bar{x} + b \iff b = \bar{y} - m \cdot \bar{x} \quad (4)$$

So erhält man einen Term, den man für das b in die Ausgleichsgerade 1 auf der vorherigen Seite einsetzen kann: [2, S.134]

$$g(x) = m \cdot x + b \iff g(x) = m \cdot x + \bar{y} - m \cdot \bar{x} \iff g(x) = m(x - \bar{x}) + \bar{y} \quad (5)$$

Die rechte Seite der rechten Gleichung in 5 lässt sich nun in den Summenterm rechts in 2 auf der vorherigen Seite einsetzen, der anschließend umgeformt wird [2, S.134].

$$\sum_{i=1}^n [g(x_i) - y_i]^2 = \sum_{i=1}^n [m(x_i - \bar{x}) + \bar{y} - y_i]^2 = \sum_{i=1}^n [m(x_i - \bar{x}) - (y_i - \bar{y})]^2 \quad (6)$$

Wenn dieser Summenterm den kleinstmöglichen Wert annimmt, hat man die optimale Regressionsgerade gefunden. Auf seine ganz rechts in 6 stehende Form lässt sich die zweite binomische Formel anwenden [2, S.134].

$$\sum_{i=1}^n m^2(x_i - \bar{x})^2 - 2m(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2 \quad (7)$$

Den Summenterm 7 kann man in drei Summenterme zerlegen.

$$\sum_{i=1}^n m^2(x_i - \bar{x})^2 - \sum_{i=1}^n 2m(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2 \quad (8)$$

Die Steigung m sowie deren Quadrat lassen sich vor die Summenterme ziehen, da sie für jedes i in 7 und 8 gleich bleiben.

$$m^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2m \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2 \quad (9)$$

Machen wir aus dem Term eine Funktion, dann erhalten wir eine Parabel.

$$f(x) = \sum_{i=1}^n (x_i - \bar{x})^2 \cdot m^2 - 2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \cdot m + \sum_{i=1}^n (y_i - \bar{y})^2 \quad (10)$$

Die Parabel ist nach oben geöffnet und an ihrem Scheitelpunkt finden wir als x-Koordinate die Steigung m, deren Einsetzen in den Summenterm das gesuchte kleinstmögliche y liefert, welches dem in 2 auf der vorherigen Seite gesuchten Minimum des Summenterms entspricht [2, S.134].

Die x-Koordinate m des Scheitelpunktes einer Parabel $f(x) = ax^2 - bx + c$ oder $f(x) = am^2 - bm + c$ berechnet man nach der Formel $m = b/2a$. Angewandt auf 10 ergibt sich 11, was der nun aus der zweiten Bedingung abgeleiteten ersten Bedingung in 4 entspricht.

$$m = -b/2a = \frac{2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (11)$$

1.2 Vereinfachung nach dem Verschiebungssatz der Statistik

Der Nenner des rechts in 11 auf der vorherigen Seite stehenden Bruches läßt sich nach dem Verschiebungssatz vereinfachen, den ich hier durch schrittweise Umformung des linken Terms in 12 zum rechten Term in 15 beweise. Zunächst wird die zweite binomische Formel angewendet. Danach kann man aus dem einen Summenterm drei machen.

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \left(\sum_{i=1}^n x_i^2 \right) - 2\bar{x} \left(\sum_{i=1}^n x_i \right) + \left(\sum_{i=1}^n \bar{x}^2 \right) \quad (12)$$

Den dritten Summenterm kann man besser als einfache Multiplikation schreiben.

$$\left(\sum_{i=1}^n \bar{x}^2 \right) = \bar{x} \left(\sum_{i=1}^n \bar{x} \right) = \bar{x} \cdot n\bar{x} = n\bar{x}^2 \quad (13)$$

Beim zweiten Summanden auf der rechten Seite der Gleichung 13 sieht man, daß der Unterschied zwischen dem \bar{x} und dem daneben stehenden Summenterm nur darin besteht, daß beim arithmetischen Mittel \bar{x} die Summe aller x_i noch durch n dividiert wird. Deshalb kann man den Summenterm auch ersetzen durch $n \cdot \bar{x}$.

$$\left(\sum_{i=1}^n x_i^2 \right) - 2\bar{x} \left(\sum_{i=1}^n x_i \right) + n\bar{x}^2 = \left(\sum_{i=1}^n x_i^2 \right) - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2 \quad (14)$$

Fasst man die Faktoren im zweiten Summanden zusammen, dann wird leichter erkennbar, daß sich der zweite und der dritte Summand zusammenfassen lassen.

$$\left(\sum_{i=1}^n x_i^2 \right) - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2 = \left(\sum_{i=1}^n x_i^2 \right) - 2n\bar{x}^2 + n\bar{x}^2 = \left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2 \quad (15)$$

Nach dem Verschiebungssatz der Statistik läßt sich auch der Zähler des rechts in 11 auf der vorherigen Seite stehenden Bruches vereinfachen. Die in 16 bis 19 folgende Herleitung ist nur etwas komplizierter als beim Nenner. Dazu wird zunächst ausmultipliziert.

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i \cdot y_i - x_i \cdot \bar{y} - \bar{x} \cdot y_i + \bar{x} \cdot \bar{y}) \quad (16)$$

Man kann nun diesen Summenterm aufteilen.

$$\sum_{i=1}^n (x_i \cdot y_i - x_i \cdot \bar{y} - \bar{x} \cdot y_i + \bar{x} \cdot \bar{y}) = \sum_{i=1}^n (x_i y_i) - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n\bar{x}\bar{y} \quad (17)$$

Danach lohnt sich wieder die Umformulierung zweier Summenterme zu mit n multiplizierten arithmetischen Mittelwerten.

$$\sum_{i=1}^n (x_i y_i) - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n\bar{x}\bar{y} = \sum_{i=1}^n (x_i y_i) - \bar{y} \cdot n \cdot \bar{x} - \bar{x} \cdot n \cdot \bar{y} + n\bar{x}\bar{y} \quad (18)$$

Die letzten drei Summanden sind trotz unterschiedlicher Schreibweisen identisch und können zusammengefasst werden.

$$\sum_{i=1}^n (x_i y_i) - \bar{y} \cdot n \cdot \bar{x} - \bar{x} \cdot n \cdot \bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y} \quad (19)$$

Insgesamt konnte mit diesen beiden Herleitungen gezeigt werden, daß sich die rechte Seite der Gleichung 11 auf der vorherigen Seite so umformen und die Berechnung der Steigung m vereinfachen läßt, wie dies ohne Erläuterung schon bei der Berechnung von m in 3 auf Seite 1 geschah.

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(\sum_{i=1}^n x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{(\sum_{i=1}^n x_i^2) - n \cdot \bar{x}^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad (20)$$

2 Regressionsgerade bezüglich x

Man kann bei der linearen Regression entweder die Abweichungen der y-Werte oder die Abweichungen der x-Werte von den Mittelwerten minimieren. Bisher habe ich nur beschrieben, wie man die sogenannte Regressionsgerade bezüglich y berechnet, indem man die Summe der quadratischen Abweichungen in vertikaler Richtung minimiert. Deshalb konnte ich zugunsten größerer Klarheit auf eine Unterscheidung verzichten, die nun aber notwendig wird, um zwischen den nicht identischen Regressionsgeraden bezüglich x und bezüglich y zu unterscheiden. [2, S.136]

	Regressionsgerade bezüglich y	Regressionsgerade bezüglich x
Regressionsgerade	$y = g(x) = m_x \cdot x + b_x$	$x = g(y) = m_y \cdot y + b_y$
Abweichungs- optimierung	vertikal	horizontal
Summe der Fehler- quadrate	$\sum_{i=1}^n [g(x_i) - y_i]^2$	$\sum_{i=1}^n [g(y_i) - x_i]^2$
Steigung	$m_x = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$	$m_y = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}$
y-Achsenabschnitt	$b_x = \bar{y} - m_x \cdot \bar{x}$	$b_y = \bar{x} - m_y \cdot \bar{y}$

Tabelle 1: Unterscheidung der Regressionsgeraden bezüglich x oder y

Bevor man die Regressionsgerade bezüglich x in ein normales x/y-Koordinatensystem einzeichnen kann, muß man ihre Geradengleichung nach y auflösen [2, S.136].

$$x = g(y) = m_y \cdot y + b_y \iff x - b_y = m_y \cdot y \iff y = \frac{1}{m_y}x - \frac{b_y}{m_y} \quad (21)$$

3 Korrelation

Wenn alle Punkte einer Punktwolke auf einer perfekten Geraden liegen, dann sind die Regressionsgeraden bezüglich x und y identisch. Je größer aber die zufälligen Meßfehler sind, umso größer wird auch die Streuung der x- und y-Koordinaten der gemessenen Punkte um die Regressionsgerade. Daraus folgt auch, daß sich die entweder hinsichtlich der vertikalen oder hinsichtlich der horizontalen Abweichungen optimierten Regressionsgeraden mit zunehmender Streuung zunehmend von einander unterscheiden. Da beide möglichen Regressionsgeraden durch den Datenschwerpunkt verlaufen, nimmt also der von beiden gebildete Winkel mit der Streuung der Daten zu. Aus den beiden Steigungen m_x und m_y kann man daher eine Maßzahl für den linearen Zusammenhang zwischen den betrachteten Merkmalen ableiten, in die allerdings auch die Meßgenauigkeit mit einfließt. Nach Bravais-Pearson berechnet man den Korrelationskoeffizienten r als geometrischen Mittelwert der Steigungen m_x und m_y der beiden Regressionsgeraden bezüglich y und x. [2, S.139]

$$r = \sqrt{m_x \cdot m_y} \iff \sqrt{\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \cdot \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}} \quad (22)$$

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2) \cdot (\sum_{i=1}^n y_i^2 - n\bar{y}^2)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (23)$$

Man sieht, daß der Korrelationskoeffizient r im Gegensatz zur Regressionsgeraden unabhängig davon ist, welches Merkmal man als möglicherweise vom anderen abhängig betrachtet. Es ist aber nicht unbedingt leicht verständlich, wie das geometrische Mittel zweier Steigungen immer zwischen -1 und $+1$ liegen kann, wie das beim Korrelationskoeffizienten nach Bravais und Pearson der Fall ist.

Betrachtet man beispielsweise die Wertepaare $1|1, 2|4, 3|6, 4|8, 5|10$, dann sieht man eine perfekte lineare Abhängigkeit der y -Werte von den x -Werten mit der Steigung 2 ($y = g(x) = m_x \cdot x + b_x = 2 \cdot x + 0$). Da es in diesem Beispiel überhaupt keine Streuung gibt, erwartet man zwei exakt aufeinander liegende Regressionsgeraden mit identischen Steigungen, deren geometrisches Mittel dementsprechend den Wert 2 besitzen müsste. Dem ist aber nicht so, weil die Steigung m_y der Regressionsgeraden bezüglich x für ein Koordinatensystem mit vertauschten Achsen berechnet wird. Will man diese Regressionsgerade bezüglich x gemeinsam mit der Regressionsgeraden bezüglich y in ein normales Koordinatensystem einzeichnen, dann muß man ihre Geradengleichung wie schon in [21](#) auf der vorherigen Seite gezeigt nach y auflösen.

$$y = m_x \cdot x + b_x \quad \text{und} \quad y = \frac{1}{m_y} x - \frac{b_y}{m_y} \quad (24)$$

Da beide Geraden genau aufeinander liegen, muß m_y der Kehrwert von m_x sein. Nun ist offensichtlich, daß der geometrische Mittelwert von m_y und m_x 1 sein muß.

Die zum Rechnen ungünstigere aber dafür leichter lesbare rechte Formel für r in [23](#) lässt auch erkennen, warum eine Streuung um die Regressionsgeraden dazu führt, daß der Nenner kleiner als der Nenner wird. Ein Teil der Kovarianzen² im Zähler bleibt negativ, weil sie anders als die x - und y -Abweichungen im Nenner nicht quadriert werden.

Man spricht von starken Korrelationen bei Werten von $r < -0,8$ und $r > 0,8$ sowie von schwachen Korrelationen bei Werten von $r < -0,5$ und $r > 0,5$. Starke Korrelationen beweisen aber keine Kausalzusammenhänge.

Quellenverzeichnis

- [1] [1](#)
Bibliographisches Institut & F. A. Brockhaus AG: *Der Brockhaus multimedial 2004 premium*. ISBN 3-411-06673-3
- [2] [1](#), [1](#), [1.1](#), [1.1](#), [1.1](#), [1.1](#), [1.1](#), [2](#), [2](#), [3](#)
Günter Cöster ; Heinz Griesel ; Arnold Hermans ; Horst Jahner ; Andreas Meißner ; Angelika Müller ; Heinz Klaus Strick ; Frierich Suhr ; Rudolf vom Hofe ; Helmut Postel ; Lohar Profke ; Ferdinand Weber: *Elemente der Mathematik 11*. Schroedel Verlag GmbH, Hannover 1999
- [3] [1](#), [1](#), [1.1](#), [1.1](#), [1.1](#)
Dr. Sabine Lauer: *Grundlagen der Statistik*.
<http://www.vanille.de/lehre/skript.pdf>, 25.1.2006

²Produkte aus den x - und y -Abweichungen von den jeweiligen arithmetischen Mittelwerten