METHODE DER KLEINSTEN QUADRATE

Manfred Borovcnik, Klagenfurt

Kurzfassung: In [1] habe ich Ideen und Konzepte zu Regression und Korrelation möglichst ohne bzw. mit elementarster Mathematik entwickelt. In dieser Note sollen *mathematische* Ergänzungen dazu in einfacher Form dargestellt werden.

1. Einleitung

Daten von zwei Zufallsvariablen können durch Punktwolken dargestellt werden. Die Problemgeschichte, wann und warum man an eine Punktwolke eine Regressionsgerade anpassen möchte, habe ich in [1] ausführlich geschildert. Nicht näher erläutert wurde in [1] jedoch, wie man zur speziellen Gestalt der Regressionsgeraden kommt. Darauf soll im folgenden eingegangen werden. Dabei werden gleichzeitig wichtige mathematische Konzepte der Regression und Korrelation miterschlossen. Die Regressionsgerade:

$$x \mapsto \hat{y}(x) = \hat{a} + \hat{b}x$$

wurde in [1] in der folgenden speziellen Gestalt verwendet:

$$\frac{\hat{y} - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}.$$

Löst man dies nach $\hat{y}$ auf so erhält man die Standardform der Geraden mit Achsenabschnitt und Steigung

$$\hat{y} = \bar{y} + r \cdot \frac{s_y}{s_x} \cdot (x - \bar{x}) = \bar{y} - r \cdot \frac{s_y}{s_x} \cdot \bar{x} + r \cdot \frac{s_y}{s_x} \cdot x.$$

Die Regressionsgerade $\hat{y}(x)$ bietet einen Schätzwert für den Mittelwert für die abhängige Variable, wenn man sich nur auf "Objekte" bezieht, deren Wert für die unabhängige Variable mit x bekannt ist. Dieser Schätzwert $\hat{y}$ unterscheidet sich vom gewöhnlichen Mittelwert umso mehr, je weiter x von $\bar{x}$ entfernt ist, je größer das Verhältnis der Standardabweichungen s_y/s_x ist und je größer der Korrelationskoeffizient r ist. Die Berechnungsmethode für den Korrelationskoeffizienten in [1] kann man in einer Formel so zusammenfassen:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{s_x}\cdot\frac{y_i-\bar{y}}{s_y} = \frac{\frac{1}{n-1}\Sigma\,(x_i-\bar{x})\,(y_i-\bar{y})}{s_x\,s_y} = \frac{\mathrm{cov}(x,y)}{s_x\,s_y}$$

Bemerkung: In [1] wurde der Einfachheit halber 1/n als Faktor gewählt, nunmehr wird 1/(n-1) genommen. Auf die innermathematischen Gründe kann hier nicht eingegangen werden.

Für die numerische Berechnung kann man folgende Formel verwenden (Die Daten brauchen dabei nur einmal in den Taschenrechner eingegeben werden):

$$r = \frac{\Sigma x_i y_i - \frac{1}{n}\,\Sigma x_i\,\Sigma y_i}{\sqrt{[\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2]\;[\Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2]}}$$

Dabei kann man die x- bzw. y-Daten noch durch eine lineare Transformation vereinfachen, z.B.

$$x \mapsto x' = \frac{x-v}{c}\,,$$

ohne daß dadurch der Wert von r verändert würde. Wie aber kommt man zu den Formeln für r sowie zur speziellen Gleichung der Regressionsgeraden? Gerade das soll im folgenden ausgeführt werden.

2. Optimierungsproblem zur Festlegung der Regressionsgeraden

Eine beliebige Gerade ist durch Wahl der Parameter a (Achsenabschnitt) und b (Steigung) festgelegt: $\hat{y} = a + bx$. Durch Wahl der Parameter a und b kann man eine Gerade der Punktwolke möglichst gut anpassen. Gut anpassen soll dabei folgendes bedeuten: Hat die unabhängige Variable den Wert x_i, so schätzt man den Wert der abhängigen Variablen durch den Punkt $(x_i, \hat{y}_i)$ auf der Regressionsgeraden, der Schätzfehler dabei ist $d_i = (y_i - \hat{y}_i)$.

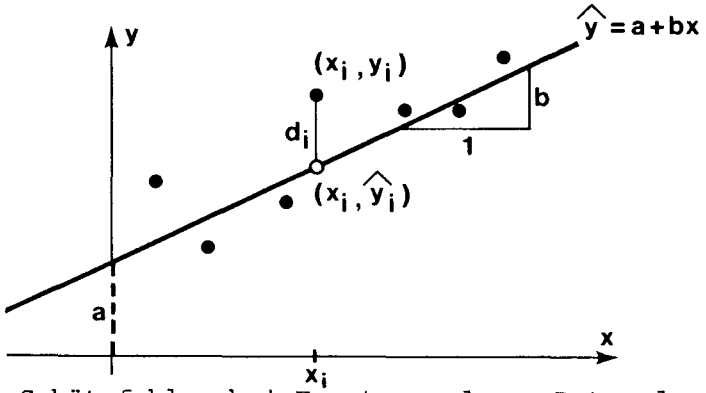

Fig.1: Schätzfehler bei Ersetzung der y-Daten durch entsprechende Punkte auf der Regressionsgeraden.

Zwei Forderungen an die Schätzfehler d_i erscheinen "vernünftig":

1) *Summe der Schätzfehler gleich Null:*

$$\Sigma d_i = 0$$

2) *Summe der Quadrate der Schätzfehler minimal:*

$$\Sigma d_i^2 = \text{minimal}$$

Der Ausdruck Σd_i^2 ist abhängig von a und b:

$$Q(a,b) = \Sigma d_i^2 = \Sigma (y_i - \hat{y}_i)^2 = \Sigma (y_i - a - bx_i)^2 ,$$

und soll durch Wahl von a und b minimiert werden.
Die Forderung 1) ergibt folgende Nebenbedingung:

$$\Sigma d_i = 0 \text{ , d.h. } \Sigma (y_i - \hat{y}_i) = 0, \text{ also:}$$

$$\Sigma (y_i - a - bx_i) = 0.$$

Daraus ergibt sich:

$$\Sigma y_i - n \cdot a - b \cdot \Sigma x_i = 0$$

Aufgelöst nach a:

$$a = \bar{y} - b\bar{x}.$$

Das bedeutet gleichzeitig, daß der Punkt $(\bar{x},\bar{y})$ auf der Regressionsgeraden liegen soll. Die aufgelöste Nebenbedingung in die Zielfunktion Q eingesetzt, ergibt:

$$Q(b) = \Sigma [y_i - \bar{y} - b(x_i - \bar{x})]^2$$

Ausquadrieren ergibt:

$$Q(b) = \Sigma (y_i - \bar{y})^2 - 2b\Sigma (x_i - \bar{x})(y_i - \bar{y}) + b^2 \Sigma (x_i - \bar{x})^2$$

Die Minimalstelle der Parabel Q(b) erhält man am einfachsten durch Nullsetzen der Ableitung:

$$Q'(b) = -2\Sigma (x_i - \bar{x})(y_i - \bar{y}) + 2b\Sigma (x_i - \bar{x})^2 = 0$$

Lösung $\hat{b}$:

$$\hat{b} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} = \frac{\text{cov}(x,y)}{s_x^2} = \frac{s_y}{s_x} \cdot \frac{\text{cov}(x,y)}{s_x s_y} = \frac{s_y}{s_x} \cdot r$$

$\hat{b}$ ist Minimum, wie man sich anhand der zweiten Ableitung von Q überzeugen kann. Für $\hat{a}$ ergibt sich:

$$\hat{a} = \bar{y} - \frac{s_y}{s_x} \cdot r \cdot \bar{x}$$

3. Restfehler und Korrelationskoeffizient

Die y-Daten werden durch die Punkte $(x_i, \hat{y}_i)$ auf der Regressionsgeraden geschätzt. Der minimale quadratische Fehler bei der bestmöglichen Wahl der Regressionsgeraden ist dabei:

$$\begin{aligned} Q(\hat{b}) &= \Sigma(y_i - \hat{y}_i)^2 = \\ &= \Sigma[y_i - \bar{y} - \frac{s_y}{s_x} \cdot r \cdot (x_i - \bar{x})]^2 = \\ &= \underbrace{\Sigma(y_i-\bar{y})^2}_{(n-1)s_y^2} + \frac{s_y^2}{s_x^2} \cdot r^2 \cdot \underbrace{\Sigma(x_i-\bar{x})^2}_{(n-1)s_x^2} - 2\frac{s_y}{s_x} \cdot r \cdot \underbrace{\Sigma(x_i-\bar{x})(y_i-\bar{y})}_{(n-1)rs_xs_y} \\ &= (n-1)s_y^2 + (n-1)s_y^2r^2 - 2(n-1)s_y^2r^2 = \\ &= (n-1)s_y^2[1-r^2] \end{aligned}$$

Messen wir "Varianz" vorübergehend durch die Summe quadratischer Abweichungen und nicht wie bisher üblich durch die "mittlere" quadratische Abweichung, also durch:

$$\Sigma(y_i - \bar{y})^2 \qquad \text{statt} \qquad \frac{1}{n-1}\Sigma(y_i - \bar{y})^2,$$

so können wir die letzte Formel plakativ wie folgt aussprechen:

$$(1)\quad Q(\hat{b}) = \Sigma(y_i - \bar{y})^2 \cdot (1 - r^2)$$

Restvarianz "Streuen um die Regressionsgerade" = ursprüngliche Varianz "Streuen um den Mittelwert $\bar{y}$" × 1 minus Quadrat des Korrelationskoeffizienten

Naiv kann man die Differenz "Ursprüngliche Varianz" minus "Restvarianz"

$$\Sigma(y_i - \bar{y})^2 - \Sigma(y_i - \hat{y}_i)^2$$

als *Varianzreduktion* ansprechen und erhält durch entsprechende Umformung von (1) die Beziehung:

Varianzreduktion durch Bezug auf Regressionsbeziehung zwischen x,y

$$\boxed{\text{Ursprüngliche Varianz}} \quad \text{minus} \quad \text{Restvarianz}$$

$$\boxed{\text{Ursprüngliche Varianz} \cdot r^2} = \Sigma(y_i - \bar{y})^2 \cdot r^2$$

Das Quadrat des Korrelationskoeffizienten ist daher einer ganz einfachen Deutung zugänglich: Es gibt an, um welchen Anteil die ursprüngliche Varianz der y-Daten um den Mittelwert $\bar{y}$ durch Bezugnahme auf die (unterstellte) Regressionsbeziehung verringert wird.

4. Zerlegung der Streuung

Für die mathematische Theorie der Regressionsrechnung ist es wichtig, daß man diesen naiv formulierten Sachverhalt abstützen kann: Die individuellen Abweichungen $y_i - \bar{y}$ kann man so zerlegen:

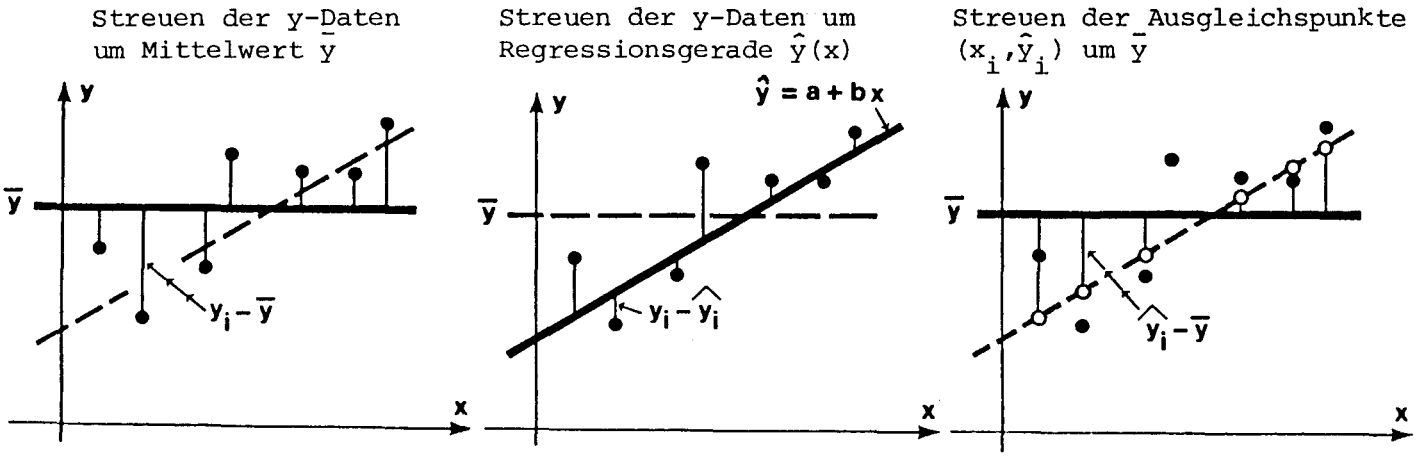

Fig.2: Zerlegung der Streuung in Komponenten, die verschiedenen "Ursachen" zugeordnet werden.

Die Verringerung der Abweichung

von	$y_i - \bar{y}$	Abweichung vom Mittelwert
auf nunmehr	$y_i - \hat{y}_i$	Restabweichung
hat die Größenordnung	$\hat{y}_i - \bar{y}$	Abweichung, *erklärt* durch Regressionsbeziehung

Würden die y-Daten *exakt* dem Verlauf der Regressionsgeraden folgen, so wäre ihre Streuung (quadratische Abweichung) durch $\Sigma(\hat{y}_i - \bar{y})^2$ gegeben. Diese Streuung ist der Regression von y und x zuzuschreiben, sie ist "durch die Regressionsbeziehung erklärt".
Darüberhinaus jedoch weichen die y-Daten durch weitere "Fehlerquellen" von der Regressionsgeraden ab, und zwar in der Größenordnung $\Sigma(y_i - \hat{y}_i)^2$. Dies ist die in den y-Daten verbleibende Streuung, *nachdem* man schon Bezug auf die Regressionsbeziehung genom-

men hat, diese *Reststreuung* ist *nicht* durch die Regressionsbeziehung erklärt.

Nicht nur das Streuverhalten einzelner Daten läßt sich in die angesprochenen zwei Komponenten zerlegen, sondern auch die Summen der Abweichungsquadrate. Behauptung:

(2) $\Sigma(y_i - \bar{y})^2$ (Ursprüngliche Varianz) $= \Sigma(y_i - \hat{y}_i)^2$ (Restvarianz) $+ \Sigma(\hat{y}_i - \bar{y})^2$ (Varianz, erklärt durch Regressionsbeziehung)

Bemerkung: Diese Streuungszerlegung hat u.a. dazu geführt, daß sich in der Entwicklung der Statistik die Varianz und nicht die mittlere lineare Abweichung als Kennziffer für das Streuverhalten von Daten durchgesetzt hat.

Der Beweis ist leider nur formal, ohne begleitende tiefere und davon unabhängige Einsicht zu führen. Durch Ausquadrieren der individuellen Abweichungszerlegung und Umordnen ergibt sich:

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Der Beweis ist geführt, wenn die gemischte Summe verschwindet, d.h. falls

$$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

Es gilt (siehe Fig.2)

$$y_i - \hat{y}_i = y_i - \bar{y} + \bar{y} - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y}).$$

Weil $(x_i, \hat{y}_i)$ und $(\bar{x}, \bar{y})$ auf der Regressionsgeraden liegen, gilt ferner:

$$\hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x}).$$

Dies kann man zur Umformung beider Faktoren verwenden also gilt:

$$\begin{aligned}\Sigma(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \Sigma[y_i-\bar{y}-\hat{b}(x_i-\bar{x})]\,\hat{b}(x_i-\bar{x}) \\ &= \hat{b}\Sigma(y_i-\bar{y})(x_i-\bar{x}) - (\hat{b})^2\,\Sigma(x_i-\bar{x})^2\end{aligned}$$

Jetzt kann man die spezielle Gestalt der Lösung $\hat{b}$ des Regressionsproblems ausnützen:

$$\hat{b} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}, \text{ woraus sich ergibt:}$$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = \hat{b} \cdot \Sigma(x_i-\bar{x})^2$$

Setzt man diesen Ausdruck oben ein, so sieht man, daß die gemischte Summe wirklich verschwindet.

5. Prüfgrößen als Anteil erklärter Streuung

Eine Regressionsgerade paßt anschaulich dann gut, wenn die Varianzreduktion in Formel (1) groß ist, d.h. falls r^2 groß ist. Eine andere Möglichkeit, die Güte der Anpassung der Geraden an die Punktwolke zu prüfen, besteht darin, die durch die Regressionsbeziehung erklärte "Varianz" $\Sigma(\hat{y}_i-\bar{y})^2$ in Formel (2) anteilmäßig

a) auf die ursprüngliche Varianz in den y-Daten zu beziehen:

$$(3)\quad \frac{\Sigma(\hat{y}_i-\bar{y})^2}{\Sigma(y_i-\bar{y})^2} = r^2$$ (leichte Umformung von (1)!)

b) auf die "Restvarianz" zu beziehen:

$$(4)\quad \frac{\Sigma(\hat{y}_i-\bar{y})^2}{\Sigma(y_i-\hat{y}_i)^2}$$

Sind die Werte von (3) und (4) groß, dann paßt die Regressionsgerade gut, dann ist die Regressionsbeziehung als Quelle, die Variabilität in den y-Daten erzeugt, groß im Vergleich zu allen Quellen, die Variabilität in den y-Daten erzeugen (Nenner in a)) bzw. groß im Vergleich zu in der Regressionsbeziehung nicht erfaßten Quellen, die "verursachen", daß die y-Daten streuen (Nenner in b)). Statt den Quadratsummen in (3) und (4) verwendet man in der statistischen Literatur mittlere Quadrate, (3) bleibt davon unberührt, (4) wird zu (4') (siehe dazu [4]).

$$(4')\quad \frac{\Sigma(\hat{y}_i-\bar{y})^2}{\frac{1}{n-2}\Sigma(y_i-\hat{y}_i)^2}$$

6. Abschließende Bemerkungen

Die Methode, die Regressionsgerade durch Minimierung der quadratischen Fehler festzulegen, heißt Methode der kleinsten Quadrate. Die voranstehenden Überlegungen sollten zeigen: Das der Regressionsrechnung zugrunde liegende Optimierungsproblem ist eigentlich leicht zu fassen. Die Lösung ist durch die einsichtige Zusatzforderung 1) (Summe der Schätzfehler ist Null) direkt und überschaubar - siehe jedoch [2].

Der Korrelationskoeffizient hat in der Form $1-r^2$ als Varianzreduktion (Formel (1)) und in der Form von (3) als Anteil erklärter Streu-

ung an der Gesamtstreuung eine sehr direkte Interpretation. Die Zerlegung der Streuung von y in Komponenten, die dann verschiedenen "Ursachen" zugeordnet werden sollen, Formel (2), erschließt eine wichtige inhaltliche Beziehung, ist jedoch exakt nur sehr umständlich zu beweisen. Die in der Statistik üblichen Prüfgrößen, (3) und (4'), dafür, daß die Regression gut paßt, sind eigentlich durch die Deutung als "relative, durch Regression erklärte Varianz" auch sehr plausibel.

Neben den inhaltlichen Überlegungen in [1] sollten auch die hier erläuterten mathematischen Beziehungen untermauern, daß Regression und Korrelation "gar nicht so schwierig" sind und daß die Konzepte einen tieferen, durchaus verständlichen Sinn haben.

Literatur:

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

(2) Goode, S.M. und Gold, E.M.: Lineare Regression und Korrelation - Ein elementarer Zugang.
In: Stochastik in der Schule 8(1988), S 33-35.

(3) Koßwig, F.W.: Auswertung von Meßdaten im naturwissenschaftlichen Unterricht. Ein elementarer Zugang zur Regressionsrechnung. In: Beiträge zum Mathematikunterricht 1983, S 180 - 183.

(4) Riedwyl, H.: Regressionsgerade und Verwandtes.
UTB Taschenbücher. Haupt: Bern und Stuttgart, 1980.