REKURSIVE KLEINSTE QUADRATE

von Ian. H.W.M. Grant

Originaltitel: Recursive least squares

Bearbeitung: G. Ihorst, Dortmund.

ZDM-Klassifikation: K 84

Einleitung: Die schädliche Wirkung von Computern auf die algebraischen Fähigkeiten von Schülern und Studenten wird in letzter Zeit häufig diskutiert. Es steht außer Frage, daß Computer eine sehr wichtige Rolle in einem Fach wie Statistik spielen. Wie von *Searle* (1983) betont wird, darf man dabei jedoch nicht vergessen, daß vernünftige algebraische Vorüberlegungen oft die Genauigkeit von Schätzprozeduren verbessern können, von denen viele mit Hilfe des Computers durchgeführt werden.

Es gibt eine Vielzahl von Fragestellungen, bei denen Schülern schnell bewußt wird, daß das Modell regelmäßig dadurch auf den neuesten Stand gebracht wird, daß neue Daten verfügbar werden. Wenn diese neuen Daten in regelmäßigen Zeitabständen eintreffen und schnelle Neuberechnungen erforderlich sind, dann sind rekursive Prozeduren ausgesprochen nützlich. Viele rekursive Algorithmen erfordern einige entscheidende algebraische Umformungen, damit sie einfach auf dem Computer zu implementieren sind. Wenn dies jedoch einmal geleistet worden ist, können die Neuberechnungen schnell und billig durchgeführt werden.

Anpassung einer Geraden

Wir betrachten das Problem, eine Gerade an die Punkte (x_1,y_1), (x_2,y_2),..., (x_n,y_n) mit Hilfe des Kleinst-Quadrate-Prinzips anzupassen. Die Gleichung der angepaßten Geraden für diese n (≥ 2) Punkte lautet:

$$y_i = A_n + B_n(x_i - M_n) \quad , \quad i = 1,...,n \quad ,$$

wobei

$$M_n = (x_1 + x_2 + ... + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad . \tag{1}$$

Die Kleinst-Quadrate-Gleichungen zur Schätzung von A_n und B_n sind bekanntlich

$$nA_n = \sum_{i=1}^{n} y_i \tag{2}$$

und

$$\sum_{i=1}^{n} (x_i - M_n)^2 B_n = \sum_{i=1}^{n} y_i(x_i - M_n) \quad . \tag{3}$$

Mit Searle's Notation

$$S_n^2 = \sum_{i=1}^{n} (x_i - M_n)^2$$

läßt sich (3) schreiben als

$$S_n^2 B_n = \sum_{i=1}^{n} y_i(x_i - M_n) \quad . \tag{4}$$

Zusätzliche Daten

Wir nehmen an, ein weiterer Datenpunkt (x_{n+1},y_{n+1}) kommt hinzu, und wir möchten die Schätzungen für den Achsenabschnitt (A) und die Steigung (B) der Geraden auf der Basis aller n+1 Datenpunkte berechnen. Zur Bestimmung von A_{n+1} erhält man mit Gleichung (2):

$$(n+1)A_{n+1} = \sum_{i=1}^{n+1} y_i = \sum_{i=1}^{n} y_i + y_{n+1}$$

oder

$$(n+1)A_{n+1} = nA_n + y_{n+1} ,$$

so daß

$$A_{n+1} = A_n + (y_{n+1} - A_n)/(n+1) . \qquad (5)$$

Um B_n neu zu berechnen, benutzen wir Gleichung (4) und schreiben

$$S^2_{n+1}B_{n+1} = \sum_{i=1}^{n+1} y_i(x_i - M_{n+1})$$

$$= \sum_{i=1}^{n} y_i(x_i - M_{n+1}) + y_{n+1}(x_{n+1} - M_{n+1}) . \qquad (6)$$

Es gilt

$$\sum_{i=1}^{n} y_i(x_i - M_{n+1}) = \sum_{i=1}^{n} y_i(x_i - M_n + M_n - M_{n+1})$$

$$= \sum_{i=1}^{n} y_i(x_i - M_n) + (M_n - M_{n+1})\sum_{i=1}^{n} y_i$$

$$= S^2_n B_n + n(M_n - M_{n+1})A_n$$

unter Ausnutzung von (2) und (4).

Weiterhin erhalten wir aus Searle's Beitrag

$$x_{n+1} = (n+1)M_{n+1} - nM_n .$$

Der Term

$$y_{n+1}(x_{n+1} - M_{n+1})$$

in (6) läßt sich damit schreiben als

$$y_{n+1}((n+1)M_n - nM_n - M_{n+1}) = n(M_{n+1} - M_n)y_{n+1} .$$

Mit diesen Ergebnissen läßt sich Gleichung (6) umformulieren zu

$$S^2_{n+1}B_{n+1} = S^2_n B_n + n(M_n - M_{n+1})A_n + n(M_{n+1} - M_n)y_{n+1}$$

$$= S^2_n B_n + n(M_n - M_{n+1})(A_n - y_{n+1}) . \qquad (7)$$

Die rekursiven Formeln (5) und (7) können zunächst für n = 2 bewertet werden, indem eine Gerade an die Punkte (x_1,y_1) und (x_2,y_2) exakt angepaßt wird. Die darauffolgenden Neuberechnungen lassen sich leicht ausführen, wie das folgende Beispiel zeigt.

Ein Beispiel:

i	x_i	y_i
1	1	3
2	5	11
3	6	12
.	.	.
.	.	.

(i) Die Gerade $y_i = 1 + 2x_i = 7 + 2(x_i - 3)$ verbindet die ersten beiden Punkte (1,3) und (5,11), somit ist

$$A_2 = 7 \quad \text{und} \quad B_2 = 2 .$$

Zusätzlich benötigen wir

$$M_2 = (1+5)/2 = 3$$

und

$$S^2_2 = (1-3)^2 + (5-3)^2 = 8 .$$

(ii) Neuberechnung von M, S, A, und B unter Verwendung des zusätzlichen Datenpunktes (6,12). Gemäß Searle's Artikel gilt

$$M_{n+1} = M_n + (x_{n+1} - M_n)/(n+1) , \qquad (8)$$

also

$$M_3 = 3 + (6 - 3)/3 = 4 ;$$

außerdem

$$S^2_{n+1} = S^2_n + n(n+1)(M_{n+1} - M_n)^2 , \qquad (9)$$

also hier

$$S^2_3 = 8 + 6(4 - 3)^2 = 14 .$$

Aus (5) ergibt sich

$$A_3 = (2A_2 + y_3)/3 = (14 + 12)/3 = 26/3 ,$$

und aus (7) erhält man

$$S^2_3B_3 = S^2_2B_2 + 2(M_2 - M_3)(A_2 - y_3) ,$$

$$14B_3 = 16 + 2(3 - 4)(7 - 12) ,$$

$$B_3 = 13/7 .$$

Die neue Gleichung für die Gerade lautet

$$y_i = 26/3 + 13(x_i - 4)/7 = 26/3 + 13x_i/7 ,$$

was sich leicht durch direkte Berechnung verifizieren läßt. Wie umfangreich die ursprünglichen Daten auch sein mögen, für die nächste Berechnung müssen nur die Größen n, M_n, S^2_n, A_n und B_n, d.h. 5 Werte, abgespeichert werden. Für

den soeben demonstrierten Algorithmus läßt sich leicht ein Computer-Programm schreiben.

Zusammenfassung

In der einfachen linearen Regression kann die Gleichung der Regressionsgeraden rekursiv mit Hilfe der Formeln (5), (7), (8) und (9) neu berechnet werden, wenn zusätzliche Beobachtungen verfügbar werden. Diese Methode ist übertragbar auf multiple lineare Regression; weitergehende Texte wie z.B. Harvey diskutieren die Vorteile einer solchen Vorgehensweise. Das wachsende Interesse an rekursiven Verfahren auf Seiten von Ingenieuren, Ökonomen und anderen Wissenschaftlern legt es nahe, diese Methoden schon früh im Statistikunterricht zu behandeln.

Literatur

HARVEY, A.C. (1981): "*The ecometric analysis of time series*", Philip Allan, Oxford.

SEARLE, S.R. (1983): "The recurrence formulae for means and variances", *Teaching Statistics*, Vol. 5, No. 1, 7-10.