ZUGANG ZUR LINEAREN REGRESSION MIT MICROCOMPUTERN ÜBER VERTEILUNGSTAFELN

von Jorj Kowszun, Haberdashers' Aske's School, Elstree, England
Originaltitel in "Teaching Statistics" Vol.10 (1988) Nr.3: A Microcomputer Approach to Teaching Linear Regression Using Spreadsheets
Übertragung: Bernd Wollring, Universität Münster

In den meisten höheren mathematischen Lehrbüchern wird die lineare Regression primär unter numerischem Aspekt gesehen. Nach dem Versuch die Methode der kleinsten Quadrate zu rechtfertigen geht der Lehrende zügig zur Minimierung des quadratischen Fehlers (ggf. über Gradienten) über und versucht, ein lineares Modell auf den vorliegenden Datensatz anzuwenden. In der Regel erfolgt dies mit Hilfe partieller Differentiation - nicht sehr praktisch im Grundkurs in der Schule - oder durch eine schreckliche quadratische Ergänzung in zwei Variablen. Dieses traditionelle Vorgehen erscheint mir nicht sehr zufriedenstellend, denn es scheint, daß der eigentliche Zweck dieser Art der Einführung darin besteht, die Lernenden mit Algorithmen für die erforderlichen Rechnungen auszuüsten, so daß sie technische, aber letztendlich nur subsidiäre Fragen, beantworten können.

Das Kriterium der kleinsten Quadrate

Ich finde inzwischen Untersuchungen von Verteilungstafeln (spreadsheets) bei der Einführung des Konzeptes der kleinsten Quadrate nützlich. In einer idealen Unterrichtssituation, in der die Lernenden an ihren Verteilungstafeln allein oder zu zweit arbeiten können, kann der Lehrende sinnvoll eingreifen. Obgleich bei diesen Untersuchungen einige ganz bestimmte Ergebnisse angestrebt sind, ist es von Bedeutung, daß der Lehrer interessante Beobachtungen und Vermutungen mitbekommt, die zu fruchtbarer

Stochastik in der Schule 8 (1988), Heft 2

weiterer Arbeit führen können. Wir wollen einige Übungen mit Verteilungstafeln beschreiben.

1. Setze einige Daten in der ersten Spalte der Verteilungstafel ein, gerade die ersten Zahlen, die einem so einfallen. Setze eine Zahl in einen externen Speicherplatz latz außerhalb dieser Spalte, hier sollen später weitere passende Werte erprobt werden. Die Zahl heiße "Bezugszahl" ("odd value"). Nun gebe man in der zweiten Spalte die Differenz der Bezugszahl und des jeweiligen Wertes in der ersten Spalte an. Am Ende dieser Spalte halte die Summe der Werte fest. Probiere verschiedene Werte als Bezugszahl, bis eine "interessante" Summe in der zweiten Spalte auftritt. Welcher Wert der Bezugszahl erzeugt diese Summe ?

	1	2	3	4	5	6	7
1	5	-5					
2	7	-3					
3	3	-7			10 ("odd value")		
4	9	-1					
5	12	2					
6	1	-9					
7	5	-5					
8	8	-2					
9		-					
10							
11		-30					

Hier ist der entscheidende Punkt, daß die Summe in der zweiten Spalte Null wird, wenn die Bezugszahl der Mittelwert der Beträge in der ersten Spalte ist. Ich fand, daß dieses für die Lernenden keineswegs eine offensichtliche Tatsache war. Es ist zudem oft erforderlich, darauf hinzuweisen, was ein "interessanter" Wert sein könnte.

2. Setze in der dritten Spalte die Absolutwerte der Zahlen in der zweiten Spalte ein, abgesehen von der Summe. Bilde am Ende der Spalte die Summe der Einträge. Wann ist diese Summe der dritten Spalte minimal?

	1	2	3	4	5	6	7
1	5	-5	5				
2	7	-3	3				
3	3	-7	7		10 ("odd value")		
4	9	-1	1				
5	12	2	2				
6	1	-9	9				
7	5	-5	5				
8	8	-2	2				
9		-	-				
10							
11		-30	34				

Hier ist der Median der Wert, der die Summe minimiert, falls eine Serie mit ungrader Anzahl in der ersten Spalte vorliegt, oder jeder Wert zwischen zwei Mittelwerten im Falle einer geraden Anzahl. Hoffentlich treten in derselben Stunde sowohl gerade als auch ungerade Anzahlen auf, so daß man eine sinnvolle Diskussion über Beliebigkeiten bei der Definition des Medians führen kann. Wenn für die verwendete Verteilungstafel eine Sortiermöglichkeit besteht, so kann man sie denen empfehlen, die steckengeblieben sind.

3. Nun füge man in der vierten Spalte die Quadrate der Werte aus der zweiten Spalte ein. Läßt sich eine Bezugszahl finden, die die Summe der vierten Spalte minimiert?

	1	2	3	4	5	6	7
1	5	-5	5	25			
2	7	-3	3	9			
3	3	-7	7	49	10 ("odd value")		
4	9	-1	1	1			
5	12	2	2	4			
6	1	-9	9	81			
7	5	-5	5	25			
8	8	-2	2	4			
9		-	-	-			
10							
11		-30	34	198			

Der minimierende Wert ist das arithmetische Mittel.

Diese Untersuchungen helfen, für das Verstehen der Varianz als Abweichungsmaß eine Grundlage zu schaffen, das ja auf kleinsten Quadraten beruht und für die parallele

Untersuchung eines Abweichungsmaßes auf der Basis kleinster Absolutbeträge. Die Nicht-Eindeutigkeit des Medians bei einer geraden Anzahl von Werten weist auf technische Probleme hin, die ein Kriterium über kleine absolute Abweichungen in komplexeren Situationen haben könnte.

Es gibt etliche Untersuchungen im Anschluß an diese, die die Lernenden selbst vorschlagen könnten oder die der Lehrer vorschlagen mag. Zum Beispiel könnte man anstatt die Summe der Spalten, die nur nichtnegative Einträge haben, deren größten Eintrag versuchen zu minimieren, was zu einer Diskussion von "Minimax"-Kriterien führt. Man kann das Muster auch auf folgende Art fortsetzen, was ganz natürlich wäre: Man untersucht die Spalte der dritten Potenzen, dann die der vierten Potenzen, usw.

"Lineare Regression von Hand"

Eine zu den oben genannten parallel verlaufende Reihe von Untersuchungen kann dazu dienen, den Lernenden ein Gefühl für die Natur und den Umfang der numerischen Aufgaben zu vermitteln, die bei linearen Regressionen auftreten. Den Ausgang bildet eine Verteilungstafel mit zwei Datenspalten, für die man einen passenden linearen Zusammenhang annimmt. An solchen Datensätzen herrscht kein Mangel, besonders wenn man in diesem Stadium nicht davor zurückschreckt, Datensätze zu erfinden. Es ist hilfreich, wenn jeder Lernende oder jedes Team einen eigenen Datensatz hat. Mit Hilfe der geläufigen Formulierung "y = mx+c" kann man die erste Spalte mit x bezeichnen, die zweite mit y und eine dritte mit den Werten mx+c anlegen, wobei man die Werte von m und c in externen Speicherplätzen ablegt. Eine vierte Spalte umfaßt die Abweichungen, d.h. die Differenz entsprechender Werte in der zweiten und in der dritten Spalte.

	1	2	3	4	5	6	7
1	x	y	mx+c	Abw.			
2	1	2	9	-7			
3	3	7	11	-3			
4	4	11	12	-1		m : 1	
5	6	11	14	-3			
6	8	18	16	2		c : 8	
7	9	19	17	1			
8	12	21	20	1			
9	13	28	21	7			
10							
11							

Ausgehend von dieser Vorlage kann man folgende Aktivitäten durchführen:

1 Variiere m und c mit dem Ziel, die Summe der Abweichungen zu Null zu machen. Man beachte, daß es zu diesem Problem keine eindeutige Lösung gibt.
2 Man lege eine Spalte mit den Absolutbeträgen der Abweichungen an. Versuche, Werte für m und c zu finden, für die die Summe dieser Spalte minimal wird.
3 Man lege eine Spalte der Quadrate der Abweichungen an und versuche, deren Summe zu minimieren.

Die beiden letzten Tätigkeiten werden die Lernenden recht schwierig finden. Das Problem dabei ist, mit der Variation von zwei Werten zur selben Zeit fertig zu werden. Man könnte dieses Problem vereinfachen, indem man eine zusätzliche Bedingung einführt um die Zahl der Variablen zu reduzieren. Eine ganz natürliche Bedingung besteht darin, daß die Mittelwerte der Spalten x und y, bezeichnet als μ_x bzw. μ_y, die lineare Bedingung $\mu_y = m\mu_x + c$ erfüllen sollen. Man beachte, daß dies aquivalent dazu ist, daß die Summe der Spalte der Abweichungen Null beträgt. Diese Bedingung bewirkt, daß man den Wert für c aus dem von m berechnen kann. Damit verbleibt eine leicht lösbare Minimierungsaufgabe, bei der man die Analogie zu den Ergebnissen aus der ersten Untersuchungsserie ausnutzen kann.

Standardisierung

Nachdem einige Erfahrungen über die Natur des Problems gesammelt sind und ein gewisses Gefühl dafür besteht, wo die Schwierigkeiten im technischen Bereich liegen, halte ich es für erforderlich, einen Rahmen zu schaffen, in dem die üblichen Formeln für den kleinsten quadratischen Fehler entwickelt werden können. Ich finde zu diesem Zweck das Konzept der Standardisierung besonders nützlich. Mit dem "Standardisieren" eines Datensatzes meine ich, daß man von jedem Wert das arithmetische Mittel subtrahiert und das Ergebnis durch die Standardabweichung dividiert. So erhält man einen neuen Datensatz, der durch lineare Transformation aus dem alten entsteht und den Mittelwert Null und die Standardabweichung Eins hat.

Die Daten der x-Spalte und der y-Spalte mögen x_i und y_i sein. Dann sind die entsprechenden standardisierten Zufallsgrößen:

$$X_i = (x_i - \mu_x)/\sigma_x \qquad Y_i = (y_i - \mu_y)/\sigma_y$$

wobei μ_x und μ_y die arithmetischen Mittel und σ_x und σ_y die Standardabweichungen für die betreffenden Spalten sind. Die Verteilungstqafel kann zur Berechnung der standardisierten Daten dienen. Man muß allerdings bei der Benutzung eingebauter Funktionen etwas aufpassen, denn oft findet man in der Maschine die Stichprobenvarianz, mit dem Koeffizienten 1/(n-1) anstelle des Koeffizienten 1/n bei der Varianz für Grundgesamtheiten. Deshalb ist es zweckmäßig, die Varianzen direkt über die Summen der Spalten x^2 und y^2 zu berechnen.

Viele Taschenrechner benutzen leicht zugängliche fest verdrahtete Funktionen, die bei der Berechnung von Varianzen hilfreich sind.

Eine sehr aufschlußreiche Tätigkeit - so fand ich - besteht darin die Lernenden die Daten ihrer letzten Verteilungstafel standardisieren zu lassen und sie dann zu

bitten, eine entsprechende Punktwolke zu zeichnen. Dann bat ich sie, die beste Gerade nach "Augenmaß" zu zeichnen und ihre Steigung und ihren Achsenabschnitt zu schätzen. Es sollte dann möglich sein, diverse Paare von Steigung und Achsenabschnitt zu verschiedenen standardisierten Datensätzen zu sammeln und zu beachten, daß alle Steigungen nahe bei 1 und alle Achsenabschnitte nahe bei 0 liegen. Dies legt nahe, daß es einen allgemeinen Zusammenhang gibt, der hier zugrunde liegt, und der auf Y = X oder Y = - X hinausläuft. Die Lernenden können vielleicht eine plausible intuitive Erklärung geben, oder man kann sie bitten, eine Situationen zu betrachten, in der die ursprünglichen Datensätze exakt linear korreliert sind. In diesem Fall gilt:

$$y_i = mx_i + c$$

Daraus folgt:

$$\mu_y = m\mu_x + c \quad \text{and} \quad \sigma_y = |m|\sigma_x$$

Also gilt:

$$Y_i = (y_i - \mu_y)/\sigma_y = (mx_i + c - m\mu_x - c)/(|m|\sigma_x) = (m/|m|)(x_i - \mu_x)/\sigma_x = (m/|m|)X_i$$

Und wir erhalten $Y_i = X_i$ oder $Y_i = -X_i$ je nach Vorzeichen von m.

Arithmetisches Verfahren bei Regression mit kleinsten Quadraten

Wir nehmen ein lineares Modell an, das die standardisierten Daten in der Form $Y_i = aX_i + b$ koppelt und betrachten die in dem Modell auftretenden Abweichungen: $E_i = Y_i - aX_i - b$. Man beachte, daß diese Abweichungen aus den entsprechenden Abweichungen bei den nicht standardisierten Daten durch eine lineare Transformation hervorgehen, so daß das Minimieren ihrer Quadratsumme dem Minimieren der Quadratsumme bei den Originaldaten äqivalent ist. Es ist algebraisch einfacher, einen quadratischen

Fehler zu minimieren. Bei n Datenpaaren hat er folgende Form:

$$\begin{aligned}\sum E_i^2/n &= \sum (Y_i - aX_i - b)^2/n \\ &= \sum (Y_i^2 + a^2X_i^2 + b^2 - 2aX_iY_i - 2bY_i + 2abX_i)/n \\ &= \sum Y_i^2/n + a^2 \sum X_i^2/n + \sum b^2/n - 2a \sum X_iY_i/n - 2b \sum Y_i/n + 2ab \sum X_i/n\end{aligned}$$

Hier kann man gut einhalten, um sich an einige Folgerungen aus der Standardisierung zu erinnern:

$$\sum X_i = 0 \qquad \sum Y_i = 0 \qquad \sum X_i^2/n = 1 \qquad \sum Y_i^2/n = 1$$

So nimmt der mittlere quadratische Fehler folgende einfachere Form an:

$$1 + a^2 + b^2 - 2a \sum X_iY_i/n$$

Zur weiteren Vereinfachung sei $R = \sum X_iY_i/n$, und wir erhalten

$$1 + a^2 + b^2 - 2aR$$

Quadratische Ergänzung ergibt:

$$(a - R)^2 + b^2 + 1 - R^2$$

Dieser Ausdruck nimmt für a = R und b = 0 als Minimum $1-R^2$ an. Man beachte, daß R der übliche Korrelationskoeffizient ist und daß aus der Tatsache, daß der mittlere quadratische Fehler nicht negativ ist und das Minimum $1-R^2$ hat, sofort die Ungleichung $|R| \leq 1$ folgt. Das Modell, das aus der Minimierung des quadratischen Fehlers für die standardisierten Daten folgt, ist Y = RX , es kann leicht in ein entsprechendes Modell für die nicht standardisierten Daten transformiert werden. Dies ist nicht ganz das anfangs vorhergesagte intiutive Modell, aber da der Korrelationskoeffizient nahe bei 1 oder -1 liegt, wenn die Daten hinreichend linear korreliert sind, paßt es zu den ersten intuitiven Vermutungen.

Man kann denselben Ansatz verwenden, wenn man mit der Methode der kleinsten Quadrate die Summe der senkrechten Abstände der Punkte von der Regressionsgeraden minimieren will. Der Abstand des Punktes $(X_i;Y_i)$ von der Geraden y = ax + b ist:

$$|Y_i - aX_i - b|/\sqrt{(a^2 + 1)}$$

Damit erhält man den mittleren quadratischen Fehler:

$$\frac{\sum (Y_i - aX_i - b)^2/n}{a^2 + 1}$$

der genau dem oben genannten "vertikalen" quadratischen Fehler bis auf die Division durch $a^2 + 1$ entspricht. Es ist eine recht einfache Übung, nachzuweisen, daß der entsprechende Ausdruck für a = 1 oder a = -1 minimal wird, was genau zu unserer intuitiven Vermutung paßt. Ich vermute, daß nach Augenmaß gezeichnete Regressionsgeraden eher denen entsprechen, bei denen die Quadratsumme der Abstände senkrecht zur Geraden als den vertikalen entsprechen.