Bestimmung von Regressionsgeraden ohne Differentialrechnung*

von C.W.Puritz
(übersetzt von M.Nuske)

Die Anpassung einer Modellgeraden $y = a+bx$ an einen Datensatz $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ erfordert die Festlegung der beiden Zahlen a und b. Die Herleitung entsprechender Formeln nach der Methode der Kleinsten Quadrate basiert auf partiellen Ableitungen. Um diese Differentialtechnik zu vermeiden, verwenden einige Lehrbücher eine algebraische Methode, die mir immer ziemlich kompliziert erschien und einen entsprechend abstoßenden Effekt auf den Durchschnittsschüler der Oberstufe hatte. Ich habe deshalb stets partielle Differentiation verwendet (ohne die zugehörige spezielle Notation) und hatte keine Schwierigkeiten damit. Der algebraische Ansatz hat jedoch seine Vorzüge. Insbesondere garantiert er ein globales Minimum anstelle eines lediglich lokalen Extremums, welches mit Hilfe der Differentialrechnung bestimmt wird.

In diesem Jahr habe ich eine einfachere algebraische Methode entdeckt und angewendet, die mir für Schüler meiner Statistikkurse der Oberstufe mit ihren sehr unterschiedlichen Begabungen und Fähigkeiten hinreichend einfach und klar erschien. Wir begannen mit der Durchrechnung eines numerischen Beispiels und verwendeten dabei die Differentialrechnung. Auf diese Weise lernen die Schüler beide Wege kennen. Im allgemeinen bin ich wie folgt vorgegangen:

Die Gleichung

$$y = a + bx \qquad (1)$$

soll für Schätzungen von y bei vorgegebenem x-Wert verwendet werden. Wenn man sie auf die x-Daten anwendet, erzeugt sie Schätzwerte $\hat{y}_i = a+bx_i$, $i = 1,\ldots,n$, die im allgemeinen von den vorliegenden y-Daten, $y_1,\ldots,y_n$, abweichen. Sei $e_i = y_i - \hat{y}_i$ der Schätzfehler für den i-ten Datenpunkt (x_i,y_i). Wir suchen nun diejenige Größe, für die die Summe der quadrierten Abweichungen der geschätzten $\hat{y}$-Werte von den beobachteten y-Werten minimal wird. Die Größe $\overline{e^2} = (e_1^2+e_2^2+\ldots+e_n^2)/n$, der mittlere quadratische Schätzfehler also, soll also durch Wahl geeigneter Werte für a und b minimiert werden.

---

* Originaltitel in 'TEACHING STATISTICS'(1981) Heft 3, Band 3
'Deriving Regression Lines without Calculus'

Aus der bekannten Formel für die (empirische) Varianz eines Datensatzes ergibt sich

$$\overline{e^2} = \mathrm{Var}(e) + \bar{e}^2 \qquad (2)$$

Wir werden sehen, daß a und b so gewählt werden können, daß jeder der beiden Ausdrücke auf der rechten Seite von (2) sein absolutes Minimum annimmt.

Zunächst ist

$$\mathrm{Var}(e) = \mathrm{Var}(y-a-bx) = \mathrm{Var}(y-bx)$$

da die Addition der Konstanten a den Wert der Varianz nicht verändert.

Weiter ergibt sich

$$\begin{aligned}\mathrm{Var}(y-bx) &= \overline{[y-bx]^2} - \overline{[y-bx]}^2 \\ &= \overline{y^2}-2b\overline{xy}+b^2\overline{x^2}-\bar{y}^2+2b\bar{x}\bar{y}-b^2\bar{x}^2 \\ &= b^2(\overline{x^2}-\bar{x}^2)-2b(\overline{xy}-\bar{x}\bar{y})+\overline{y^2}-\bar{y}^2 \\ &= b^2s_x^2 - 2bs_{xy} + s_y^2 \qquad (3)\end{aligned}$$

wobei s_x^2, s_y^2 und s_{xy} die Varianz der x-Daten, die Varianz der y-Daten beziehungsweise die Kovarianz* der (x,y)-Datenpaare bedeuten.

Wir vervollständigen nun das Quadrat mittels der sogen. quadratischen Ergänzung (oder könnten auch die Analysis ohne partielle Differentiation anwenden, falls gewünscht):

$$\begin{aligned}\mathrm{Var}(e) &= s_x^2[b^2-2bs_{xy}/s_x^2 + s_{xy}^2/s_x^4] + s_y^2 - s_{xy}^2/s_x^2 \\ &= s_x^2[b-s_{xy}/s_x^2]^2 + [s_x^2s_y^2 - s_{xy}^2]/s_x^2 \qquad (4)\end{aligned}$$

Offenbar nimmt dieser Ausdruck sein absolutes Minimum dann an, wenn gilt

$$b = s_{xy}/s_x^2 \qquad (5)$$

Nun betrachten wir den zweiten Ausdruck in (2):

$$\bar{e}^2 = [\overline{y-(a+bx)}]^2 = [\bar{y}-(a+b\bar{x})]^2$$

---

* Die Definition der Kovarianz einer Datenpaarmenge lautet:

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Das absolute Minimum dieses Ausdrucks ist 0 und wird genau dann angenommen, wenn gilt

$$\bar{y} = a + b\bar{x} \qquad (6)$$

Das heißt $\overline{e^2}$ nimmt seinen minimalen Wert an, wenn b gemäß (5) gewählt wird - das gibt uns die Steigung der Regressionsgeraden - und wenn a die Bedingung (6) erfüllt - daraus ersehen wir, daß die Regressionsgerade durch das Datenzentrum $(\bar{x},\bar{y})$ verläuft. Die Bedingungen (5) und (6) ergeben also die Lösung des Problems.

Zusätzlich läßt sich aus (4) der Minimalwert von $\overline{e^2}$ bestimmen:

$$[s_x^2s_y^2-s_{xy}^2]/s_x^2 = [s_x^2s_y^2-(rs_xs_y)^2]/s_x^2 = s_y^2 - r^2s_y^2 \qquad (7)$$

Dabei ist $r = s_{xy}/s_xs_y$ der Korrelationskoeffizient der Datenpaare (x_i,y_i). Hieraus ergibt sich unmittelbar, daß r^2 stets kleiner oder gleich 1 ist und den Wert 1 genau dann annimmt, wenn $\overline{e^2}$ gleich 0 ist, das heißt, wenn die Datenpunkte alle exakt auf einer Geraden liegen.

Darüberhinaus können wir zeigen, daß $r^2s_y^2$ die Varianz der geschätzten y-Werte, $\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_n$, ist:

$$\begin{aligned}\mathrm{Var}(\hat{y}) &= \mathrm{Var}(a+bx) = b^2\mathrm{Var}(x) \\ &= (s_{xy}^2/s_x^4)s_x^2 = (s_{xy}/s_x)^2 = r^2s_y^2\end{aligned}$$

Aus (7) ergibt sich also:

$$\mathrm{Var}(e) = \mathrm{Var}(y) - \mathrm{Var}(\hat{y})$$

oder

$$\mathrm{Var}(y) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(e)$$

Das heißt die Gesamtvarianz der y-Daten ist die Summe aus der erklärten Varianz (die durch die Beziehung zwischen y und x zustande kommt) und der nichterklärten Varianz (die durch die Abweichungen der Datenpunkte von der Geraden entsteht). Und in dieser Summe ist das Verhältnis von erklärter zu nichterklärter Varianz gerade r^2.

Dieser letzte Teil (nach (7)) war für die Schüler schwierig, erforderte aber nicht viel Zeit und erzeugte vielleicht das auf dieser Stufe bestmögliche Verständnis dafür, wie der Korrelationskoeffizient tatsächlich den Grad mißt, in welchem einem empirischen Datensatz eine theoretische Modellgerade angepaßt werden kann.