Wie informativ ist der Korrelationskoeffizient?

Michael Spielmann

1 Zusammenfassung

Das folgende Beispiel zeigt, dass die rechnerischen Kenngrößen der Regressionsanalyse nur unzureichende Informationen liefern.
Man kann daher auf eine graphische Darstellung nicht verzichten.
Umgekehrt vermittelt die oberflächliche Betrachtung der graphischen Veranschaulichung ein falsches Bild von der Qualität der Regressionsbeziehung.

2 Kann man von den rechnerischen Kenngrößen der Regressionsbeziehung auf die Form der zweidimensionalen Verteilung schließen?

Die Verteilungen der Tabelle 1 stimmen in den wesentlichen Kenngrößen überein: arithmetisches Mittel, Standardabweichung und Covarianz sind bis zur zweiten Nachkommastalle gleich. Die Verteilungen haben dieselbe Regressionsgerade mit demselben Korrelationskoeffizienten. Sie wurden nach der Methode der kleinsten Quadrate bestimmt. Die unabhängige Variable ist für die drei Verteilungen je gleich, die abhängige Variable ist jeweils y1, y2, und y3; yt sind die aus der Regressionsgeraden berechneten Werte.

X	Y1	Y2	Y3	Yt
5	11,722	12,879	13,500	9,75
6	7,610	8,476	6,500	9
7	9,830	7,16	6,500	8,25
8	4,280	4,255	6,500	7,5
9	5,070	5,46	6,500	6,75
10	7,810	6,52	6,500	6
11	6,178	7,75	6,500	5,25

Tabelle 1

Die Kenngrößen sind:

$\bar{y}$	Mittelwert der yi	7,5
s_y	Streuung der yi	2,5
C_{xy}	Covarianz	3,0
Gerade		- 0,75 x + 13,5
r_{xy}	Korrelationskoeffizient	0,6

Tabelle 2

Wir stellen die Verteilungen als Punktwolke dar. Beim Betrachten der Graphik fallen die Unterschiede sofort ins Auge.

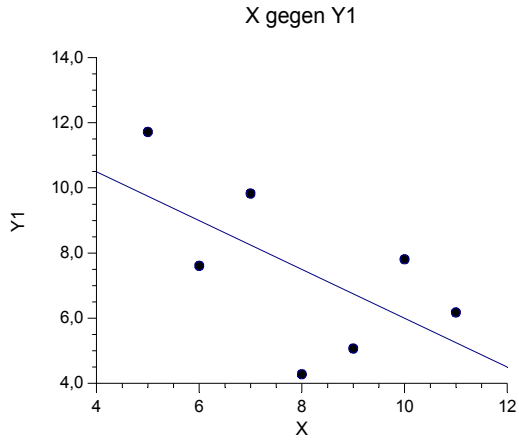

Abb. 1

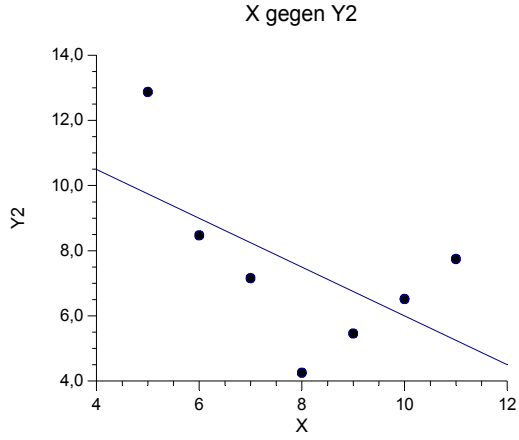

Abb. 2

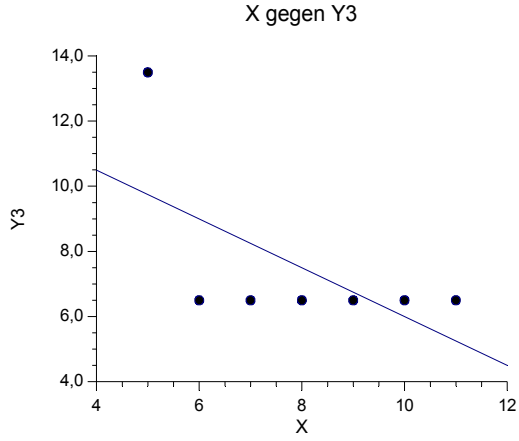

Abb. 3

Eine Anregung zu dieser Übersicht verdanke ich F.J.Anscombe [1], der vier verschiedene Verteilungen entworfen hat. Es ist reizvoll, durch Versuch und Irrtum langsam aber sicher an entsprechende eigene Daten heran zu kommen. Mit dem Tabellenkalkulationsprogramm EXCEL ist es möglich, statistische Tabellen zu veranschaulichen, gleichzeitig aber auch die Kennzahlen sofort berechnen zu lassen. Mit der Maus lassen sich durch Verschieben der Datenpunkte in einer Grafik die Eingaben so verändern, dass die Kennzahlen immer wieder auf gleiche Werte korrigiert werden.
Dieses Probierverfahren ist sicher für Schüler interessant und ansprechend. Lehrreich ist es allemal: man erhält einen guten Eindruck, wieweit die Punkte in unterschiedlicher Lage vom Schwerpunkt der Wolke die Kenngrößen beeinflussen.

3 Kann man vom Korrelationskoeffizienten auf die Streuung der Punkte um die Regressionsgerade schließen?

Ein hoher Korrelationskoeffizient legt die Vermutung nahe, die Punktwolke schmiege sich eng um die Regressionsgerade. Wir wollen den Zusammenhang zwischen Korrelationskoeffizient und Steigung der Geraden untersuchen. (Siehe auch [2].)

Der Korrelationskoeffizient sei r_{xy}, die Covarianz C_{xy}, die Steigung der Regressionsgeraden sei b, die Streuungen s_x und s_y.

Es gilt dann

$$r_{xy} = \frac{c_{xy}}{s_x \cdot s_y} \quad und \quad b = \frac{c_{xy}}{s_x^2} \quad also \quad b = r_{xy} \cdot \frac{s_y}{s_x}$$

Die letzte Gleichung gibt die Abhängigkeit zwischen b und r_{xy} sehr gut wieder.

Wenn s_x und r_{xy} konstant sind, dann muss bei kleinerem s_y auch b kleiner werden. Andersherum: Wenn s_x und b konstant sind und s_y kleiner wird, dann muss r_{xy} entsprechend größer werden.

Damit ist die Abhängigkeit rechnerisch beschrieben und eigentlich mathematisch klar. Aber ist sie auch anschaulich klar? Der Vergleich zweier Grafiken erzeugt wohl doch noch einige Verwunderung.

X	Y1	Y2
1	4	7,1239
2	5	7,1983
2	3	7,0494
3	7	7,3471
3	9	7,4959
3	5	7,1983
4	6	7,2727
4	4	7,1239
5	7	7,3471
5	8	7,4215
5	6	7,2727
6	9	7,4959
6	11	7,6447
6	8	7,4215
7	12	7,7192
7	14	7,8680

Tabelle 3

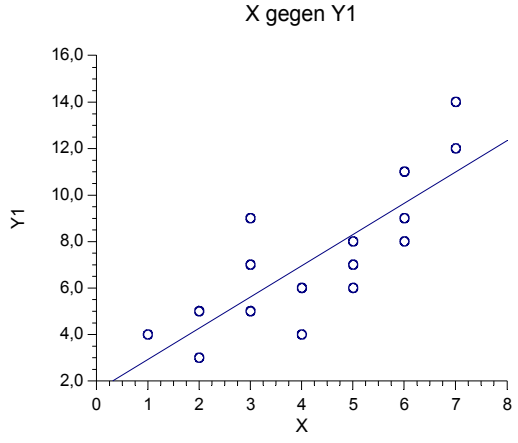

Abb. 4

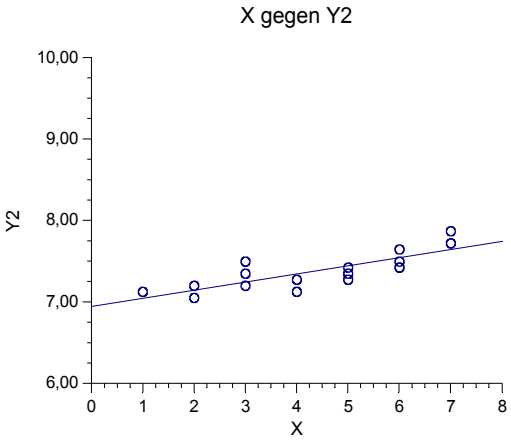

Abb. 5

Abb. 4 vermittelt den Eindruck eines eher mäßig guten Zusammenhangs, eine Schätzung wird den Korrelationskoeffizient bei etwa 0,5 ansiedeln. Dagegen erscheint der Zusammenhang in der Wolke von Abb. 5 hervorragend; die Schüler schätzen den Korrelationskoeffizienten in der Größe von etwa 0,9. Die Korrelationskoeffizienten sind jedoch in beiden Fällen gleich 0,66!

4 Wertung

An den beiden hier präsentierten Beispielen wird deutlich, dass Berechnung von Kenngrößen und graphische Veranschaulichung erst in gegenseitiger Ergänzung ein zutreffendes Bild von der Stichprobe vermitteln. John Tukey [3] betonte schon 1977, dass eine statistische Analyse der guten graphischen Unterstützung bedarf. Die Beurteilung anhand der Punktwolke reicht allein nicht aus, sie ist aber trotz rechnerischer Analyse unerlässlich.

5 Literatur

[1] **Anscombe, Francis J.:** *Graphs in Statistical Analysis*, American Statistician, 27, 17-21,

[2] **Iversen, Gudmund R.:** *Statistics : the conceptual approach***,** New York [u.a.] : Springer, 1997

[3] **Tukey, John W**.*: Exploratory data* Reading, Mass. [u.a.] : Addison-Wesley, 1977.

Anschrift des Verfassers:
StD Michael Spielmann, Wolfgangstr. 14, 42655 Solingen.
email: spielmann@wtal.de