BEZIEHUNGSHALTIGE MATHEMATIK IN REGRESSION UND KORRELATION

von Helmut Wirths, Oldenburg

Kurzfassung: Im Analysisunterricht lernen die Schüler, den Graphen zu einer gegebenen Zuordnungsvorschrift zu zeichnen. In diesem Beitrag wird die Umkehrung betrachtet: Zu einer gegebenen Punktmenge soll eine dazu passende Funktionsgleichung bestimmt werden. Die Behandlung von mehrdimensionalen Zufallsgrößen ist sinnvoll, da mit den Begriffen Regression und Korrelation bedeutende Mathematisierungsmuster zur Verfügung stehen (vgl. [10], S. 284). Hier können Schüler vor allem beim Sammeln, Darstellen und Auswerten des Datenmaterials, aber auch im Entdecken und Interpretieren von Zusammenhängen vielfältige Aktivitäten entwickeln.

1. Einleitung

Beim Auswerten von Meßreihen stellt sich folgendes Problem: Gegeben sind n Meßwerte (x|y), gesucht ist die Gleichung einer Funktion f: x → f(x), mit der die Punktmenge "am besten" beschrieben werden kann. Dies kann auf verschiedene Weisen in eine mathematische Optimierungsaufgabe umformuliert werden. Das Problem läßt sich schulgemäß vor allem auf drei Wegen behandeln:

- mit elementaren Mitteln aus der Algebra und Stochastik (vgl. z.B. [2], [3], [4], [5], [11]),
- mit Mitteln aus der analytischen Geometrie (vgl. z.B. [8], [12], [13]) oder
- mit elementaren Mitteln aus Analysis und Stochastik (vgl. z.B. [1], [6], [7]).

Vor allem in den Heften 1 und 2 des Jahrganges 1988 der Zeitschrift "Stochastik in der Schule" findet man Motivationshilfen und eine Einführung in grundlegende Konzepte.

2. Das Lösungsmodell

Wir machen folgende **Modellannahme**:

Die Funktionsgleichung sei $\hat{y}(x) = f(x) = m \cdot x + b$. Wir nehmen also zunächst einmal an, daß der Zusammenhang zwischen den Koordinaten der Meßwerte durch eine lineare Funktion beschrieben werden kann. Es sollen hier kurz die Ergebnisse zusammengestellt werden. Eine schülergerechte Herleitung kann z.B. der in 1. angegebenen Literatur entnommen werden.

Folgende Abkürzungen lassen uns die einschlägigen Gleichungen besser strukturieren. Der Term nach dem letzten Gleichheitszeichen ist für die direkte numerische Auswertung leichter zu handhaben, sofern man nicht Taschenrechner oder Computer mit Statistik-Software benutzt.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n \cdot (\bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n \cdot \bar{x}\,\bar{y}$$

$$S_{\hat{y}\hat{y}} = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2$$ Streuung der Ausgleichspunkte

$$S_{y\hat{y}} = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2$$ Reststreuung (Residuenstreuung).

Man kann für die Parameter m und b in der Gleichung der Regressiongeraden folgende Ergebnisse herleiten (siehe Literatur):

(1) Der Punkt $(\bar{x}|\bar{y})$, der Schwerpunkt der Punktwolke, liegt auf der Regressionsgeraden, d.h. es gilt: $\bar{y} = m \cdot \bar{x} + b$. Bei bekannter Steigung m kann man daraus den Achsenabschnitt berechnen.

(2) $m = \dfrac{S_{xy}}{S_{xx}}$, falls $S_{xx} \neq 0$.

Die Ausgleichsgerade f, deren Gleichung man mit Hilfe der obigen Ergebnisse berechnen kann, läßt bestmögliche Schätzwerte $\hat{y}$ (im Sinne der Optimierung aus Abschnitt 1) für y ausrechnen, wenn man den Wert von x kennt:

1. Ausgleichsgerade f $\quad \hat{y} = \hat{y}(x) = \dfrac{S_{xy}}{S_{xx}} \cdot x + \left(\bar{y} - \dfrac{S_{xy}}{S_{xx}} \cdot \bar{x}\right)$.

Oft ergibt sich aus der Problemstellung, daß man die Zuordnung umkehren möchte, d.h. man kennt die Werte von y und möchte diejenigen von x voraussagen. Dann hat man ein anderes Optimierungsproblem und eine andere Lösung:

2. Ausgleichsgerade f_2 $\quad \hat{x} = \hat{x}(y) = \dfrac{S_{xy}}{S_{yy}} \cdot y + \left(\bar{x} - \dfrac{S_{xy}}{S_{yy}} \cdot \bar{y}\right)$.

Auch ohne explizite Herleitung kann man das Zustandekommen der zweiten Geradengleichung einsehen, wenn man in der Gleichung von f die Rollen von x und y

vertauscht. In beiden Fällen ist x die 1. Koordinate und y die 2. Koordinate. Die Gleichung von f_2 schreiben wir nicht in der üblichen Form y = f(x). In der vorgelegten Darstellung von f_2 soll deutlich werden, daß x die statistisch abhängige und y die statistisch unabhängige Variable ist.

Für die Sonderfälle gilt:

$S_{xy} = 0$: Der Graph von f ist eine Parallele zur x-Achse, der von f_2 eine Parallele zur y-Achse. Hat man S_{xy} als ein (intuitiv) zugängliches Maß des Kovariierens der beiden Variablen eingeführt (vgl. z.B. [2]), dann muß man diesen Fall so interpretieren, daß gar keine Zusammenhänge zwischen x und y bestehen.

In diesem Fall gilt: $\hat{y}(x) = \bar{y}$. Die in dieser Gleichung enthaltene Prognose $\hat{y}(x)$ für die Variable y ist unabhängig davon, ob wir Kenntnis von x haben oder nicht. Sie läßt sich auch bei noch so guter Kenntnis von x nicht mehr verbessern. Der Mittelwert von y wird zum besten Schätzwert. Analog ist für die Voraussage von x aus y der Mittelwert von x der beste Schätzwert.

$S_{xx} = 0$: Eine Summe von Quadraten ist genau dann Null, wenn alle Quadrate Null sind. In Bezug auf S_{xx} heißt das, daß alle x-Koordinaten gleich ihrem Mittelwert sind und daher alle übereinstimmen. Alle Punkte liegen auf einer Parallelen zur y-Achse, so daß eine Regressionsrechnung in diesem Fall nicht erforderlich ist.

$S_{yy} = 0$: Hier gilt analog, daß eine Regressionsrechnung nicht erforderlich ist, weil alle y-Koordinaten übereinstimmen, die Punkte bereits auf einer Parallelen zur x-Achse liegen.

$S_{xx}=S_{yy}=0$: Dieser Fall kann nur eintreten, wenn alle x-Koordinaten gleich sind und alle y-Koordinaten übereinstimmen. Liegen mehrere Meßwerte vor, stellen sie nur einen einzigen Punkt dar. Es ist auch hier keine Regressionsrechnung erforderlich.

3. Der Korrelationskoeffizient

Das Berechnen der Gleichung der Regressionsgeraden sollte nicht alleiniger Inhalt bei der Behandlung dieses Themas sein. Man benötigt eine Größe, genannt Korrelationskoeffizient r, die als Maß für die Güte der Beschreibung der Daten im Modell dienen soll. Schülergerechte Überlegungen zur Einführung dieser Größe findet man in der in 1. genannten Literatur. Hier sollen wieder nur Ergebnisse zitiert werden.

1. Möglichkeit zur Definition von r (vgl. z.B. [2],[5]):

Der Korrelationskoeffizient r kann als normiertes gemischtes Moment zweiter Ordnung eingeführt werden:

$$(3) \qquad r = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sqrt{S_x}} \cdot \frac{y_i - \bar{y}}{\sqrt{S_y}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} .$$

Durch die Standardisierung $(x_i - \bar{x})/\sqrt{S_x}$ bzw. $(y_i - \bar{y})/\sqrt{S_y}$ erhält man jeweils Größen mit dem Mittelwert 0 und der Standardabweichung 1. Motivationshilfen für eine an den mathematischen Beziehungen orientierte Einführung findet man in [2].

2. Möglichkeit zur Definition von r (vgl. z.B. [1], [3], [11]):

Mit $S_{\hat{y}\hat{y}}$ wird die Streuung der Ausgleichspunkte gemessen. Diese Streuung liegt im Modell, kann also im Modell erklärt werden. Die Streuung der Punkte um die Ausgleichsgerade mißt $S_{y\hat{y}}$. Diese Streuung wird nicht im Modell erklärt. Das Streuen der y-Koordinaten um ihren Mittelwert mißt S_{yy}. Es gilt: $S_{yy} = S_{\hat{y}\hat{y}} + S_{y\hat{y}}$.

Der Quotient $S_{\hat{y}\hat{y}}/S_{yy}$ kann als wichtige Kenngröße eingeführt werden. Liegen alle Punkte auf einer Geraden, ist dieser Quotient 1; sonst ist er kleiner als 1, aber größer oder gleich 0. Erweitert man diesen Bruch mit 1/n, dann vergleicht man den Teil der Varianz , der im Modell erklärt werden kann, mit der gesamten Varianz. Wenn $r^2 = 1$ ist, wird die gesamte Varianz durch das Modell erklärt und für $r^2 = 0$ ist nichts im Modell erklärbar. Wieviel Prozent der Varianz im Modell erklärt wird, kann man also r^2 entnehmen. Man kann daher auf diesem Weg definieren

$$r^2 := \frac{S_{\hat{y}\hat{y}}}{S_{yy}} .$$

Den Korrelationkoeffizienten definiert man dann durch:

$$(4) \qquad r := \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = m \cdot \sqrt{\frac{S_{xx}}{S_{yy}}}$$

r ist somit eine Zahl zwischen -1 und 1, die das gleiche Vorzeichen wie m hat.

3. Möglichkeit zur Definition von r (vgl. z.B. [12]):

Wenn alle Meßwerte exakt auf einer Geraden liegen, sind die Regressionsgeraden f und f_2 identisch. Im allgemeinen Fall schließen sie einen Winkel miteinander ein. Die Größe des Winkels kann man als Maß für die Güte der Darstellung der Punktwolke durch die Regressionsgerade auffassen, wobei 0° und 180° auf einen exakt linearen Zusammenhang

deuten und der Winkel 90° darauf hinweist, daß gar kein linearer Zusammenhang zwischen den beiden Größen besteht.

Bei diesem Weg definiert man r als Kosinus des Winkels, den die beiden Ausgleichsgeraden einschließen. Diese Definition führt auf eine der Darstellungen (3) oder (4).

Für r = 0 können die folgenden Überlegungen mit zwei Beispielen weiteren Einblick liefern: Wenn r = 0 ist, muß $S_{xy} = 0$ sein, falls $S_{xx} \cdot S_{yy} \neq 0$ ist. Dann ist die Steigung der einen Regressiongeraden Null, sie verläuft also parallel zur x-Achse. Die zweite Regressionsgerade ist dann eine Parallele zur y-Achse.

Beispiel 1:

x	-5	-4	-4	-3	-3	0	0	3	3	4	4	5
y	0	3	-3	4	-4	5	-5	4	-4	3	-3	0

Alle diese Punkte liegen auf einem Kreis um den Ursprung mit einem Radius von 5 Einheiten. Es liegt also kein linearer (auch kein funktionaler!) Zusammenhang vor. Aus der Annahme, daß eine lineare Funktion den Zusammenhang beschreibt, errechnet man: r = 0, x = 0 bzw. y = 0 als Gleichungen der Regressionsgeraden, (0|0) ist Schwerpunkt der Punktwolke.

Beispiel 2:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	25	16	9	4	1	0	1	4	9	16	25

Alle diese Punkte liegen auf der Normalparabel. Es liegt also auch hier kein linearer Zusammenhang vor. Aus der Annahme, daß eine lineare Funktion den Zusammenhang beschreibt, errechnet man:
r = 0, x = 0 bzw. y = 10 sind Gleichungen der Regressionsgeraden, (0|10) ist Schwerpunkt der Punktewolke.

Weitere Wertetafeln, die zu r = 0 führen, lassen sich leicht nach dem Muster der beiden Beispiele konstruieren.

Der Korrelationskoeffizient erfaßt sehr indirekt, wie gut eine lineare Beziehung an die Daten anpaßt. Liegen nicht lineare Beziehungen vor, kann das, wie die Beispiele zeigen, durchaus r = 0 zur Folge haben.

Mit der Behandlung des Korrelationskoeffizienten kann man eine Unterrichtsreihe "Regression und Korrelation" beschließen. Empfehlenswert ist es jedoch, die Theorie auf

solche Fälle auszuweiten, bei denen man aus der Lage der Punktmenge einen nicht linearen Zusammenhang vermutet. Man erschließt sich damit eine Fülle von interessanten Anwendungsaufgaben.

4. Linearisierung der Meßwerte

Die Überlegungen bezogen sich bisher nur auf den Fall, daß der Zusammenhang zwischen den beiden Größen x und y durch eine lineare Funktion f: $x \to y$ mit $y = m \cdot x + b$ beschrieben werden kann und als Graph der Punktwolke eine Gerade in Frage kommt. Dies ist das **Modell M_1**: Die lineare Funktion f: $x \to y$ mit $y = m \cdot x + b$.

In den Fällen, in denen als Graph eine Gerade unpassend erscheint, muß man versuchen die Meßwerte so zu transformieren, daß der Graph der transformierten Daten eine Gerade ist. Ist eine solche Linearisierung gelungen, schließt sich eine Regressionsrechnung mit den transformierten Daten an. Der Korrelationkoeffizient informiert dann, wie gut eine Gerade die transformierten Daten approximiert. Mit s bzw. t seien im folgenden die 1. bzw. die 2. Koordinate der gegebenen Daten bezeichnet, mit x bzw. y die zugehörige Koordinaten nach der Transformation und mit f(t) die 2. Koordinate der Modellfunktion. Folgende einfache Modelle bieten sich im Mathematikunterricht vor allem an:

Modell M_2:

Die Potenzfunktion f: $t \to s$ mit $s = a \cdot t^n$.

$s = a \cdot t^n$
$\Leftrightarrow \ln s = \ln(a \cdot t^n)$ (falls alle Koordinaten positiv sind!)
$\Leftrightarrow \ln s = n \cdot \ln t + \ln a$

Setzt man $y = \ln s$ und $x = \ln t$, erhält man: $y = n \cdot x + \ln a$. Dabei ist n die Steigung m und (ln a) der Achsenabschnitt b der Regressionsgeraden. Also lautet die Gleichung der Ausgleichsfunktion $f(t) = e^b \cdot t^m$.

Modell M_3:

Die Exponentialfunktion f: $t \to s$ mit $s = a \cdot e^{k \cdot t}$.

$s = a \cdot e^{k \cdot t}$
$\Leftrightarrow \ln s = \ln(a \cdot e^{k \cdot t})$ (falls alle 2. Koordinaten positiv!)
$\Leftrightarrow \ln s = k \cdot t + \ln a$

Setzt man $y = \ln s$ und $x = t$, erhält man: $y = k \cdot x + \ln a$. Dabei ist k die Steigung m und (ln a)

der Achsenabschnitt b der Regressionsgeraden. Also lautet die Gleichung der Ausgleichsfunktion $f(t) = e^b \cdot e^{m \cdot t}$.

Modell M_4:

Die quadratische Funktion f: $t \to s$ mit $s = a \cdot t^2 + b$. Setzt man $y = s$ und $x = t^2$, so erhält man $y = a \cdot t^2 + b$. Also ist a die Steigung m und b der Achsenabschnitt der Regressiongeraden. Die Gleichung der Ausgleichsfunktion lautet $f(t) = m \cdot t^2 + b$.

In den folgenden Abschnitten werden einige Beispiele zur Arbeit mit diesen Modellen gegeben.

5. Prognosewerte und Beurteilungen von Residuen

Der Korrelationskoeffizient ist eine sehr schwierig zu interpretierende Größe, auch wenn man bei der Herleitung Wert auf einige Deutungen legt. Es werden daher zwei weitere Methoden vorgestellt, die über den Korrelationskoeffizienten hinausgehend eine Prüfung der Güte des an die Daten angepaßten Modells ermöglichen sollen.

Methode 1 (Vergleich Modell - Realität):

Man setzt in jedem Modell, das zur Beschreibung des Zusammenhangs zwischen den Meßgrößen geeignet erscheint, bei allen Meßwerten die statistisch unabhängige Variable in die Gleichung der Regressionsfunktion ein, errechnet einen Prognosewert für die andere Variable und vergleicht die theoretischen Daten mit den gegebenen.

Methode 2 (Vorhersage):

Man benutzt den funktionalen Zusammenhang zwischen den Größen, um eine Vorhersage für die zweite Koordinate bei anderen als den gegebenen Daten zu machen.

An der folgenden **Aufgabe 1** sollen Ausführungen zur Wahl des Modells sowie zu den beiden Methoden gemacht werden:

Von deutschen Wetterstationen wurden folgende Jahresmittelwerte für den Luftdruck veröffentlicht (Daten aus [14]):

Station	List/Sylt	Freudenstadt	Feldberg	Wendelstein	Zugspitze
h in km	0.006	0.797	1.486	1.832	2.960
s in mbar	1009.3	924.2	848.5	814.2	705.2

h: Höhe über NN s: mittlerer Luftdruck

Bemerkungen:

Setzen wir voraus, daß die Höhen der Wetterstationen fehlerfrei, die mittleren Werte für den Luftdruck fehlerbehaftet sind. Durch die Regressionsrechnung sollen diese Fehler eliminiert werden. Wir berechnen also Prognosewerte $\hat{s}(h)$.

Schüler, die die barometrische Höhenformel kennen, werden Modell M_3 einsetzen. In der Regel wird jedoch zunächst in Modell M_1 gearbeitet, zumal die Schüler dem Graphen meist keine überzeugenden Gründe für einen nicht-linearen Zusammenhang entnehmen. Modell M_2 scheidet aus, weil es einen Graphen durch den Koordinatenurspung voraussetzt, was hier wohl nicht der Fall ist. Die Ergebnisse der Regressionsrechnung für die Modelle M_1, M_2 und M_4 werden in der folgenden Tabelle dargestellt:

Modell	r	r^2	Gleichungen
M_1 lin. Fkt.	-0.99950	0.9990	$\hat{s}_1 = -103.19 \cdot h + 1006.6$
M_3 Exp-Fkt.	-0.99956	0.9991	$\hat{s}_3 = 1014.9 \cdot e^{-0.1217 \cdot h}$
M_4 Parabel	-0.93864	0.8810	$\hat{s}_4 = -30.89 \cdot h^2 + 952.9$

Der wesentlich schlechtere Korrelationskoeffizient macht M_4 uninteressant.

Nach Untersuchung des linearen Zusammenhangs besteht für Schüler ohne Kenntnis der barometrischen Höhenformel (Regelfall vor allem in Grundkursen) meist kein Anlaß mehr, einen nicht-linearen Zusammenhang zu vermuten. Ein Korrelationskoeffizient, der fast 1 ist, läßt die Möglichkeit einer besseren Beschreibung des Zusammenhangs unwahrscheinlich erscheinen. Wenn man die physikalischen Zusammenhänge nicht kennt, besteht überhaupt kein Anlaß, das einfache Modell des linearen Zusammenhanges zugunsten eines komplizierteren aufzugeben. So denken auch die Schüler.

Wie kann man nach der Untersuchung des linearen Zusammenhangs an M_1 Zweifel wecken und auf M_3 aufmerksam machen, ohne als Lehrer die entscheidenden Argumente einzubringen? Dies soll nun mit Hilfe der beiden oben vorgestellten Methoden geschehen, wobei sofort neben die Ergebnisse in M_1 die entsprechenden in M_3 mit angegeben werden.

Anwendung von Methode 1 auf Aufgabe 1:

Station	List/Sylt	Freudenstadt	Feldberg	Wendelstein	Zugspitze
h in km	0.006	0.797	1.486	1.832	2.960
s in mbar	1009.3	924.2	848.5	814.2	705.2

In Modell M_1 gilt:

$\hat{s}_1$ in mbar	1006.0	924.4	853.3	817.6	701.2
Δ_{1s} in mbar	3.3	0.2	-3.8	-3.4	4.0
$\lvert\Delta_{1s}\rvert$: s*100	0.33	0.02	0.45	0.42	0.57

$\hat{s}_1$: Prognosewert $\hat{s}_1(h)$ im Modell M_1

$\Delta_1 s$: absoluter Fehler (Residuum): = s - $\hat{s}_3$

$|\Delta_1 s|$: s * 100: relativer Fehler in %

In Modell M_3 gilt:

$\hat{s}_3$ in mbar	1014.2	921.2	847.1	812.2	708.0
$\Delta_3 s$ in mbar	-4.7	3.1	2.4	2.1	-2.8
$\lvert\Delta_s s\rvert$: s*100	0.47	0.34	0.28	0.26	0.40

$\hat{s}_3$: Prognosewert $\hat{s}_3(h)$ im Modell M_3

$\Delta_3 s$: absoluter Fehler (Residuum):= s - $\hat{s}_3$

$|\Delta_3 s|$. s * 100: relativer Fehler in %.

Die Höhen wurden in die Gleichung der Regressionsfunktion eingesetzt und die zugehörigen Prognosewerte $\hat{s}_1$ bzw. $\hat{s}_3$ errechnet.

Mit Methode 1 läßt sich die Prognosefähigkeit des jeweiligen Modells innerhalb des durch die Wertetabelle gegebenen Argumentbereichs der Regressionsfunktion gut studieren.

Es gilt: Residuen = Daten – Modell. Sowohl die Residuen als auch die prozentualen Fehler werden von den Schülern akzeptiert. Wenn man Zweifel an der Brauchbarkeit von Modell M_1 wecken will, muß man andere Argumente finden.

Versuchen wir für fünf Wetterstationen der Schweiz (Daten aus einem Wetterblatt der Schweizerischen Meteorologischen Anstalt Zürich) Prognosewerte s(h) zu berechnen, wobei eine Wetterstation (Jungfraujoch) wesentlich höher als die auf der Zugspitze liegt, und wir vergleichen sowohl die Residuen als auch die relativen Fehler.

Anwendung von Mehode 2 auf Aufgabe 1:

Station	Basel	Gstaad	Arosa	Weißfluhjoch	Jungfraujoch
h in km	0.316	1.099	1.821	2.690	3.580
s in mbar	979.5	892.5	812.5	733.0	653.0

In Modell M_1 gilt:

$\hat{s}_1$ in mbar	974.0	893.2	818.7	729.0	637.2
$\Delta_1 s$ in mbar	5.5	-0.7	-6.2	4.0	15.8
$\lvert\Delta_1 s\rvert$: s*1000	0.56	0.08	0.76	0.55	2.42

In Modell M_3 gilt:

$\hat{s}_3$ in mbar	976.7	887.9	813.3	831.7	656.6
$\Delta_3 s$ in mbar	2.8	4.6	-0.8	1.3	-3.6
$\lvert\Delta_3 s\rvert$: s*1000	0.29	0.52	0.10	0.16	0.55

Auffallend ist, daß in Modell M_1 der absolute Fehler beim Jungfraujoch wesentlich über dem der anderen Stationen liegt, während bei den anderen Stationen die an den deutschen Wetterstationen ermittelten Zusammenhänge auch auf die neue Tabelle übertragen werden können. Folgende Tatsachen lassen Zweifel an M_1 aufkommen:

- Bei größeren Höhen muß man offensichtlich mit erheblich größeren Fehlern bei den Prognosewerten rechnen.
- Die Nulltestelle der Ausgleichsgeraden liegt bei ca. 9.76 km. Nur bis zu diesem Wert ist die Gleichung brauchbar, denn negative Werte für den Luftdruck sind physikalisch nicht sinnvoll. Aus dem gleichen Grund kann die Regressionsfunktion von M_4 nur bis ca. 5.5 km benutzt werden.
- Bei großen Höhen liegen die gemessenen Werte offenbar immer über den Prognosewerten im Modell M_1. Läßt man Schüler skizzieren, wie eine bessere Anpassung vor allem für größere Höhen graphisch aussehen könnte, stellen sie eine (leicht) linksgekrümmte Kurve, die die s-Achse bei ca. 1010 mbar schneidet, dar. Solch eine Kurve für negatives exponentielles Wachstum sollte ihnen aus dem Unterricht der 10. Klasse und aus der Analysis bekannt sein. Damit ist der Weg zu Modell M_3 angezeigt.

Erst eine Regressionsrechnung in M_3 (vgl. die obigen Tabellen zu Methode 1 und 2) zeigt, daß in diesem Modell überhaupt eine bessere Approximation der Daten als in M_1 erreicht wird. Wenn in einem Modell bessere Vorhersagen als in anderen gemacht werden können, ziehen wir dieses Modell den anderen vor. Auch Schüler handeln so. In diesem Sinn erweist sich das Modell M_3 erst nach Einbeziehung von Meßwerten aus großer Höhe dem Modell M_1 überlegen.

Diese Aufgabe weist auch auf eine Besonderheit hin, der man bei der Auswertung von physikalischen Tabellen häufig begegnet. Das Modell, das durch eine Einbettung in eine

physikalische Theorie (wie M_3 bei Aufgabe 1) ausgezeichnet ist, hat keinen Korrelationskoeffizienten, der die der anderen Modelle "haushoch" überragt. Um solch ein Modell auch dem Schüler ohne physikalische Erfahrungen gegenüber M_1 in das Blickfeld zu rücken, muß man versuchen, überzeugende Argumente zu finden. Für solche Überlegungen, die über den Korrelationskoeffizienten hinausgehen, können die beiden Methoden hilfreich sein.

An diesem Beispiel wird schon deutlich, daß der Computer mit geeigneter Software (Rechenblatt oder Programm) bei der Vorbereitung des Unterrichts ein wichtiges, vielleicht sogar unverzichtbares, Hilfsmittel darstellt. Nach der Bearbeitung einiger Aufgaben nach diesen Methoden erfahren die Schüler auch, daß bei Daten um den Schwerpunkt der Punktewolke wesentlich geringere Abweichungen zwischen im Modell errechnetem und dem gemessenen Wert vorkommen als dies bei Daten am Rand oder außerhalb des Meßbereichs der gegebenen Daten der Fall ist.

6. Korrelation und Kausalität

Ein weiteres Problem der Interpretation von Zusammenhängen, die mit Hilfe der Korrelationsrechnung beschrieben werden, ist das Problem der Kausalität. Wie z.B. in [2] ausgeführt wird, existieren verschiedene Zusammenhänge. Ein hoher Korrelationskoeffizient bedingt keineswegs, daß es sich um eine kausale Beziehung handelt. Immer wenn zwei monotone Folgen miteinander korrelieren, ergibt sich ein hoher Korrelationskoeffizient. Besonders deutlich wird das in folgendem, schon legendären Beispiel von **Aufgabe 2**:

Jahr	1930	1931	1932	1933	1934	1935	1936
t	132	142	166	188	240	250	252
s	55.4	55.4	65	67.7	69.8	72.3	76

t: Storchenpaare in Oldenburg

s: Einwohner in Oldenburg (in 1000) Zahlen aus [4].

Bemerkungen:

Hier ist zunächst einmal nicht klar, ob es überhaupt eine angemessene Darstellung des Zusammenhangs gibt. Die Ergebnisse in den vier Modellen sind:

Modell	M_1	M_2	M_3	M_4
r	0.945	0.954	0.938	0.928

In Modell M_2 wird der Zusammenhang zwischen s und t offenbar am besten dargestellt. Die Gleichung der Regressionsfunktion für Prognosewerte $\hat{s}(t)$ lautet $\hat{s} = \hat{s}(t) = 6638.79 \cdot t^{0.4364}$.

Wenigstens eine solche "Nonsens"-Korrelation sollte behandelt werden, um auf die Problematik Korrelation-Kausalität aufmerksam zu machen. Ein hoher Korrelationskoeffizient beweist nichts. Er weist auf einen möglichen Zusammenhang hin. Ob es sich dabei um einen Kausalzusammenhang handelt, muß mit anderen Mitteln untersucht werden (vgl. auch [2] oder das Beispiel aus [4] auf S. 197).

7. Physikalische Zusammenhänge und Meßfehler

Physikalische Beziehungen gelten recht genau, falls man über eine theoretische Fundierung und genaue Meßgeräte verfügt. Die Abweichung von der Beziehung, die durch eine mathematische Funktion formuliert wird, haben also den Charakter von restlichen Störgrößen, die um vieles kleiner als die Meßgrößen sind. Dies gilt nicht nur für das in der Physik formulierte Gesetz, sondern macht sich z.B. auch im linearen Modell bemerkbar. Diese Phänomen soll nun erläutert werden.

In den folgenden Abschnitten wird jeweils nur noch Modell M_1 dem mit dem größten Korrelationskoeffizienten gegenübergestellt.

Aufgabe 3:

Bei den Jupitermonden gilt folgender Zusammenhang zwischen der Umlaufdauer T und der halben großen Achse ihrer Bahn x:

Name	Jo	Eurapa	Ganymed	Callisto
T in Tage	1.769	3.551	7.1555	16.689
s in 1000 km	412	671	1070	1880

Anwendung von Methode 1 auf Aufgabe 3:

In Modell M_1 gilt mit $r_1 = 0.995$:

$\hat{s}_1$ in 1000 km	479	650	995	1910
$\Delta_1 s$ in 1000 km	-67	21	75	-30
$\lvert\Delta_1 s\rvert$: s•100	16 .3	3.1	7.0	1.6

In Modell M_2 gilt mit $r_2 = 0.99993$:

$\hat{s}_2$ in 1000 km	415	665	1067	1889
Δ_1 s in 1000 km	-3	6	3	-9
$\lvert\Delta_2$ s $\rvert$: s•100	0.7	0.9	0.3	0.5

Bei der Rechnung wurde vorausgesetzt, daß die Zeiten T ohne Fehler gemessen worden sind, die Werte für die Entfernung jedoch fehlerbehaftet sind (bzw. der Fehler in der Zeitmessung vernachlässigbar klein gegenüber dem in der Entfernungsmessung ist). Daher soll ein Schätzwert $\hat{s}(T)$ für die Entfernung s berechnet werden.

Man sieht, wie gut in M_2 die gegebenen Daten approximiert werden. In M_2 erreicht der Korrelationskoeffizient den größten Wert von allen vier Modellen.

In $\hat{s} = \hat{s}(T) = 282.5 \cdot T^{0.675}$, der Gleichung der Regressionsfunktion, ist das 3. Kepplersche Gesetz (T^2 proportional zu s^3) recht gut zu erkennen.

Für physikalische Zwecke, wo man größere Genauigkeit gewöhnt ist, paßt Modell M_1 nicht so gut: relativer Fehler für Jo 16%! Einem Phänomen sollen noch einige Überlegungen gewidmet werden: In M_1 erreicht der Korrelationskoeffizient die phantastische Größe von r = 0.995, obwohl dieses Modell von der physikalischen Theorie nicht gestützt wird. In der Physik kann man unter wohldefinierten Bedingungen mit dem Auftreten fast funktionalen Beziehung rechnen, die durch Störgrößen (die auch Fehler im Erstellen der Bedingungen und Meßfehler beinhalten) etwas verschleiert werden. Läßt man zwei streng monotone Folgen korrelieren, dann ist ein hoher Korrelationskoeffizient im linearen Modell nicht ungewöhnlich.

Untersuchen wir an Aufgabe 3 einmal, wie eine unterstellte monotone Beziehung die Streuung der Residuen im Verhältnis zur Ausgangsstreuung beeinflußt, erhalten wird folgende Tabelle:

Entfernung					Standard-abweichung
s in 1000 km	412	671	1070	1880	555.2
Δ_1 s in 1000 km	-67	21	75	-30	53.5

Die Standardabweichung der Residuen beträgt 9.6% der Standardabweichung der Ausgangsdaten. Das lineare Modell reduziert dieses Maß an Unsicherheit auf ca. 10% des Ausgangsniveaus. In diesem Sinne bietet M_1 sehr viel Erklärung für den Trend in der

Punktwolke.

Kann man hier eine weitere Deutung für den Korrelationskoeffizienten entwickeln? In der Deutung als anteilige erklärte Varianz hat man $r^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}}$.

Berücksichtigt man $S_{yy} = S_{\hat{y}\hat{y}} + S_{y\hat{y}}$, erhält man

$$\frac{\sigma_{neu}}{\sigma_{alt}} = \frac{\sigma_{Residuen}}{\sigma_{Daten}} = \sqrt{\frac{S_{y\hat{y}}}{S_{yy}}} = \sqrt{\left(\frac{S_{yy}}{S_{yy}} - \frac{S_{\hat{y}\hat{y}}}{S_{yy}}\right)} = \sqrt{1 - r^2}.$$

Jetzt kann man den Einfluß der Reduktion des Maßes an Unsicherheit in Abhängigkeit von r studieren:

r	0.999	0.995	0.990	0.950	0.900	0.866	0.500
σ_n/σ_{alt}	0.045	0.100	0.141	0.312	0.436	0.500	0.866

Eine Verringerung der Standardabweichung auf die Hälfte des Ausgangsniveaus erfordert einen Korrelationskoeffizient von r = 0.866, auf ein Drittel r = 0.95, für eine Zehntelung benötigt man bereits einen so hohen Korrelationskoeffizient von 0.995. Damit haben wir eine neue Bedeutung von r kennengelernt.

Die beiden folgenden Aufgaben werden vorgestellt, weil hier eine interessante Interpretation des Achsenabschnitts möglich ist.

Aufgabe 4:

Für die mittlere Höhe h (über NN) von Satelliten und ihre Umlaufdauer T gilt:

Name	Salut I	Cosmos 184	Geos-3	Nimbus	Intel-sat-2
T in min	89.77	97	101.7	107.3	1424.4
h in km	262	619	844	1100	35558

Bemerkungen:

Trägt man die Werte in ein T-h-Koordinatensystem ein, kann man dem Graphen kaum Information entnehmen, ob ein linearer Zusammenhang vorliegt oder nicht, der letzte Meßwert liegt zu weit von den anderen entfernt. Wir legen das 3. Kepplersche Gesetz, den bei Aufgabe 3 erkannten Zusammenhang, zugrunde und führen eine lineare Regressionsrechnung mit h und $T^{2/3}$ durch. Man erhält r = 0.9999999 sowie die Gleichung $\hat{h} = \hat{h}(T) = 333.2367 \cdot T^{2/3} - 6375.4$ für die Regressionsfunktion. Im Achsenabschnitt erkennt man den mittleren Erdradius, den man zu h addieren muß, um die Abstände vom Erdmittelpunkt aus zu messen. Bei diesem hohen Korrelationskoeffizienten, diesmal bewußt

mit 7 Nachkommastellen angegeben, argwöhnen die Schüler in der Regel, daß das 3. Kepplersche Gesetz beim Aufstellen der Tabelle zugrunde lag. Das Argument, die Daten seien verschiedenen Tabellen entnommen, überzeugt sie nicht. Haben die Schüler wohl Recht ??? !!!

Aufgabe 5:

Bei einem Zentralkraftgerät wird die Abhängigkeit der Zentralkraft F_z von der Frequenz f (aus Messungen der Umlaufzeit für 20 Umdrehungen berechnet) gemessen:

f in Hz	0.84	0.71	0.60	0.49	0.39	0.29	0.19
F_z in N	1.04	0.75	0.58	0.35	0.21	0.14	0.08

Bemerkungen:

Im Modell M_4, das die Zusammenhänge zwischen f als stochastisch unabhängiger und F_z als abhängiger Variable in Übereinstimmung mit der physikalischen Theorie beschreibt, erhält man mit $r_4 = 0.9982$ den größten Korrelationskoeffizienten aller vier Modelle und die Gleichung $\hat{F}_z = \hat{F}_z(f) = 1.464 \cdot f^2 + 0.015$ für die Regressionsfunktion. Den konstanten Summanden der Funktionsgleichung kann man als systematischen Feher bei der Nullpunkteinstellung des Kraftmessers interpretieren. Von den gemessenen Kräften muß man also ca. 0.015 N abziehen, um die tatsächlich wirksame Kraft zu kennen. In M_1 erhält man wiederum einen hohen Korrelationskoeffizienten, der in diesem Fall $r_1 = 0.981$ beträgt.

Die Aufgaben 4 und 5 zeigen, daß man bei Beachtung des physikalischen Zusammenhangs der Größen auch systematische Fehler beim Erstellen der Meßtabelle (Nicht-Berücksichtigung des Erdradius, Fehler beim Messen der Kraft, . . .) bestimmen kann. Es sei auch auf Aufgabe 5 aus [12] verwiesen, wo man einer Regressionsrechnung in Modell M_4 die Masse entnehmen kann, die bei der Schwingung eines Federpendels am Schwingungsvorgang beteiligt ist und zur abgehängten Masse addiert werden muß.

8. Stochastische Zusammenhänge nach dem Muster der Physik

In der Physik treten unter Idealbedingungen Beziehungen zwischen Größen auf, die nur durch restliche Störfaktoren getrübt sind. Daher sind hohe Korrelationkoeffizienten und genaue Beschreibungen der Zusammenhänge die Folge. Die Beziehungen sind in einem Theoriegeflecht miteinander vernetzt. In anderen Wissenschaften, z.B. den Sozialwissenschaften, ist das anders. Auch dort versucht man, stabile Modelle und Beziehungen herauszuarbeiten, damit man einen Problemkomplex besser strukturieren kann.

Ein interessanter Punkt ist, daß ein bestimmtes Regressionsmodell für eine Reihe ähnlicher Fragestellungen eine gute Beschreibung liefert. Aus der Art der Beziehung kann man sich einen gesetzesähnlichen Zusammenhang denken und an die Interpretation der unterschiedlichen Koeffizienten herangehen.

Aufgabe 6:

Im Eisschnellauf bestanden vor Jahren folgende Weltrekorde:

x in m	500	1000	1500	5000	10000	35246
Zeit t in s	36.57	72.58	113.22	411.17	858.00	3600

Aufgabe 7:

In den leichtathletischen Laufdisziplinen bestanden vor Jahren folgende Weltrekorde:

x in m	100	400	1500	5000	10000	30000
Zeit t in s	9.93	43.86	209.46	780.4	1633.8	5490.4

Bemerkungen:

Die Strecken sollen fehlerfrei, die Zeiten hingegen "fehlerbehaftet" sein. Daher versucht man durch eine Regressionsrechnung die Fehler zu eliminieren und Prognosewerte $\hat{t}(x)$ zu berechnen.

Ein linearer Zusammenhang kommt für die Schüler nicht in Frage, weil z.B. ein 10000 m Rekordlauf nicht als 100 malige Wiederholung eines 100 m Rekordlaufs dargestellt werden kann. Läßt man Schüler darstellen, wie ein zusammenhängender Graph im Vergleich zu einer Ursprungsgeraden aussehen müßte, zeichnen sie eine Kurve durch den Ursprung (daher kommt M_3 nicht in Frage), die linksgekrümmt sich immer weiter von der Urspungsgeraden entfernt. Es erscheint daher interessant, Modell M_2 zu erproben. Das Modell M_4 scheidet aus, weil in der Gleichung der Regressionsfunktion b = 0 gesetzt werden muß und ein quadratischer Term in x als zu starke Einschränkung der Möglichkeiten erscheint.

Der Korrelationskoeffizient ist in Modell M_2 tatsächlich bei beiden Aufgaben der größte aller vier Modelle.

In M_2 erhält man für die beiden Aufgaben folgende Ergebnisse:

Aufgabe	r	r^2	Gleichungen
6	0.9998	0.9996	$t_2 = 0.04 \cdot x^{1.079}$
7	0.9999	0.9998	$t^2 = 0.06 \cdot x^{1.112}$

Mit 0.9993 bei den Werten von Aufgabe 7 hat der Korrelationskoeffizient in M_1 den zweithöchsten Wert aller vier Modelle. Die Gleichung der Regressionsgeraden lautet $\hat{t} = \hat{t}(x) = 0.184 \cdot x - 80.03$. Wie schlecht trotz dieses phantastischen Werts von r in M_1 die gegebenen Daten approximiert werden, soll mit Hilfe von Methode 1 in der folgenden Tabelle dargestellt werden:

x in m	100	400	1500	5000	10000	30000
t in s	9.93	43.86	209.46	780.4	1633.8	5490.4
$\hat{t}_1$ in s	-61.63	-6.43	195.97	840.0	1760.0	5440.0

Auch ohne Angabe der Residuen und des relativen Fehlers wird deutlich, wie wenig M_1 geeignet ist. Für ca. 435 m würde man einen Rekord mit 0 Sekunden Laufzeit erwarten. Auch die Prognosewerte für 100m und 400m sind Ergebnisse, die mit der Realität überhaupt nicht in Einklang gebracht werden können.

Dies ist wieder ein Beispiel dafür, daß auch ein extrem hoher Korrelationskoeffizient noch keine Garantie bietet, daß im zugehörigen Modell der Zusammenhang zwischen den Größen angemessen im ganzen Definitionsbereich der Meßtabelle beschrieben wird. Es ist immer hilfreich, wenigstens eine der in Aufgabe 3 vorgeführten Methoden zu weiteren Vergleichen heranzuziehen.

Die Gleichung für die Regressionsfunktion im Modell M_2 für Weltrekorde im Eisschlellauf (Aufgabe 6), in den leichtathletischen Laufwettbewerben (Aufgabe 7) und vom Freistilschwimmen (5 Meßwerte von 100 bis 1500 m) werden in der nächsten Tabelle dargestellt:

Disziplin	Regressionsfunktion
Eisschnellauf	$\hat{t}_2 = 0.043 \cdot x^{1.079}$
Laufen	$\hat{t}_2 = 0.059 \cdot x^{1.112}$
Freistilschwimmen	$\hat{t}_2 = 0.370 \cdot x^{1.069}$

Es fällt auf, daß die Exponenten k fast gleich sind und in der Nähe von 1.1 liegen. Kann man die Lauf-(Schwimm-)zeiten dieser Sportarten durch eine gemeinsame Gleichungsform $\hat{t} = a \cdot x^{1.1}$ beschreiben, wobei für a jeweils der für die betreffende Sportart (Lauf, Schwimmen, Eisschnellauf) charakteristische Wert für Eisschnellauf ca. 0.043) eingesetzt werden muß? Eine Regressionsrechnung kann zwar diese überraschende Gemeinsamkeit aufzeigen. Ob hier ein gesetzesähnlicher Zusammenhang beschrieben wird, ob die beiden

Koeffizienten interpretiert werden können und wie sie ggfs. interpretiert werden können, kann der Mathematiker nicht mehr klären.

Schüler finden es interessant, an solch eine Nahtstelle geführt zu werden und zu erleben, wie Mathematik in anderen Wissenschaften zur Strukturierung eines Problemfeldes eingesetzt werden kann.

9. Abschließende Bemerkungen

Bei der Behandlung von Regression und Korrelation lassen sich auch Probleme aus der Physik mit Gewinn einbringen. Selbst Schüler, die Physik in der Oberstufe abgewählt haben, werden durch diese Beispiele nicht abgeschreckt.

Überzeugende Argumente für die Wahl eines passenden Modells sollten **vor** der Durchführung einer Regressionsrechnung gesucht werden. Unser Auge kann gut erkennen, ob gegebene Punkte (fast) auf einer Geraden liegen oder nicht, besonders wenn es von einem durchsichtigen Geodreieck unterstützt wird. Liegen andere Zusammenhänge vor, müssen die Daten transformiert werden. Erst wenn die transformierten Daten (fast) auf einer Geraden liegen, sollte im zugehörigen Modell eine Regressionsrechnung durchgeführt werden. Bei physikalischen Meßtabellen hilft dieses Vorgehen sehr. Die approximierten Modelle in den Aufgaben 2 - 6 wurden nach dieser Methode von den Schülern erfunden.

Naturgesetzmäßigkeiten werden nicht mit Hilfe eines Korrelationskoeffizienten entdeckt, der die von anderen möglichen Zusammenhängen weit überragt. Bei physikalischen Wertetabellen liegen die Werte des Korrelationskoeffizienten teilweise sehr dicht beieinander. Man muß in diesem Fall Methoden vergleichbar der Auswertung eines Zielphotos anwenden, um das Modell mit der optimalen Anpassung der Daten zu finden, aber auch um auf das Modell mit physikalischer Legitimation erst aufmerksam zu machen. Darin liegt eine gewisse Crux, aber auch der Reiz eines Unterrichts, in dem mehr als nur die Werte von r, m und b berechnet werden sollen. Hier muß der Fachlehrer in seiner Vorbereitung (wie eigentlich in jedem anderen Gebiet auch) mögliche Schülervorschläge vorweg erahnen und auf Tragfähigkeit prüfen, realisierbare Argumente für oder gegen ein Modell aufspüren oder Möglichkeiten zu Schüler-Schüler-Interpretationen anbahnen. Daß ein Computer mit entsprechender Software wertvolle Dienste leistet, sollte nach Studium dieses Beitrags einleuchten. Die Schüler müssen lernen, auf Besonderheiten in den Tabellen und der graphischen Darstellung zu achten. Starke Änderungen der Residuen oder des relativen Fehlers sind als Indiz für Grenzen des Modells ernst zu nehmen. Die beiden

vorgestellten Methoden sind, wie in den vorigen Abschnitten gezeigt wurde, gute Hilfen. Die Schüler können auch in der Bereitstellung neuen Datenmaterials für neue, aber auch für den besprochenen Problemen ähnliche, Aufgaben vielfältige Aktivitäten entwickeln.

Die zweite Ausgleichsgerade f_2 dient in erster Linie in meinem Unterricht zur Erarbeitung einer interessanten Deutung des Korrelationskoeffizienten, indem das Phänomen der "Regressionsschere" ausgenutzt wird. Das ist sehr suggestiv und wirkungsvoll. Das Zustandekommen der Gleichung ist unmittelbar einsichtig, wenn man die Gleichung von f ausführlich hergeleitet hat und ist auch nicht zeitaufwendig. Bei einer linearen Regression ist der Standpunktwechsel (Rollentausch der statistisch unabhängigen und abhängigen Variablen) sowie die Berechnung der Gleichung der zweiten Ausgleichsgeraden ohne großen Aufwand möglich und sollte durchaus auch durchgeführt werden. Vor Anwendung von Linearisierungsmethoden sollte jedoch geklärt werden, für welche der beiden Variablen Prognosewerte berechnet werden sollen. Während man in M_2 noch eine Gleichung von f_2 elementar bestimmen kann, ist in M_3 und M_4 der Aufwand nicht mehr vertretbar. Um eine Gleichung einer Regressionsfunktion f_2 z.B. in M_3 besimmen zu können, muß man die Gleichung der Regressionsfunktion $f: s \to t$ mit $t = a \cdot \ln s + c$ bestimmen. Eine weitere Linearisierung in einem neuen Modell M_3* wäre die Folge. Die Funktion f aus M_3 ist f_2 in M_3*, f_2 aus M_3 ist f aus M_3*. Daher ziehe ich es vor, aus der Fülle der möglichen Aufgaben lieber wenige ausführlich zu behandeln, Wert auf Interpretation zu legen und Raum für Entdeckungen zu geben, als viele Aufgaben lediglich rechnerisch bearbeiten zu lassen.

Literatur-Auswahl

[1] Athen/Griesel (1984):Mathematik heute: LK Stochastik, Schroedel.

[2] Borovcnik, M. (1988): Korrelation und Regression - Ein inhaltlicher Zugang zu den grundlegenden Kozepten, *Stochastik in der Schule* 1/1988, S. 5 - 32.

[3] Borovcnik, M. (1988): Methode der kleinsten Quadrate, *Stochastik in der Schule* 2/1988, S. 17 - 24.

[4] Engel, A. (1987): *Stochastik*, Klett.

[5] Goode, S. M. und E. M. Gold (1988): Lineare Regression und Korrelation - Ein elementarer Zugang, *Stochastik in der Schule* 1/1988, S. 36 - 46

[6] Kroll, W. (1980): Ausgleichsrechnung als Anwendung der elementaren Analysis in Grundkursen, *MNU* 1980/6, S. 334 ff.

[7] Kroll, W. (1985): *Analysis* Band 1, Dümmler.

[8] Martens, G. (1988): Zur Methode der kleinsten Quadrate, *DdM* 2/1988, S. 88 -93.

[9] Morgan, A. L. (1988): Korrelation zwischen den Augenzahlen von zwei Würfeln, *Stochastik in der Schule* 2/1988, S. 25 - 32.

[10] Titze, Klika, Welpers (1982): *Didaktik des Mathematikunterrichts in der Sekundarstufe II,* Vieweg.

[11] Vohman, H. D. (1988): Lineare Regression in einem Einführungskurs über empirischse Methoden, *Stochastik in der Schule* 2/1988, S. 3 - 16.

[12] Wirths, H. (1990): Regression/Korrelation, *DdM* 1/1990, S. 52 - 60.

[13] Wunderling, H. (1979): *Lineare Algebra,* bsv (1979).

[14] Wahrscheinlichkeitsrechnung und Statistik unter Einbeziehung von elektronischen Rechnern (1982), *Beschreibende Statistik DIFF* Heft SR 1, Tübingen.

Herrn Dr. M. Borovcnik sei für wertvolle Hinweise und Anregungen gedankt.