LINEARE REGRESSION UND KORRELATION IN EINEM EINFÜHRUNGSKURS ÜBER EMPIRISCHE METHODEN

von Horst Dieter Vohmann, Bielefeld

Kurzfassung: Ein Arbeitstext zum Thema Regression und Korrelation wird vorgestellt. Die Besonderheiten daran sind: Der Themenzusammenhang wird handlungs- und anwendungsbezogen erschlossen, die mathematische Darstellung kommt nicht zu kurz, kommt aber ohne partielle Ableitungen aus der Differentialrechnung aus. Die Kurssequenz wurde am Bielefelder Oberstufen-Kolleg erprobt, eine Übertragung auf das reguläre Gymnasium wird zur Diskussion gestellt.

1. Vorbemerkungen zur differenzierten Mathematikausbildung am Bielefelder Oberstufen-Kolleg

Das Oberstufen-Kolleg des Landes Nordrhein-Westfalen an der Universität Bielefeld verbindet in seinem Ausbildungssystem eine frühzeitige Spezialisierung in zwei Studienfächern mit einer interdisziplinären Allgemeinbildung, die in Kurssequenzen und Einzelkursen wissenschaftspropädeutisch ("Ergänzungsunterricht") und in Projekten handlungs- und anwendungsbezogen ("Gesamtunterricht") ausgestaltet ist (zur Allgemeinbildung siehe Hoffmann, 1986). Mit dieser Struktur einer vierjährigen Ausbildungszeit, die die gymnasiale Oberstufe und das Eingangssemester der Hochschule umfaßt, erprobt das Oberstufen-Kolleg eine Alternative zum traditionellen deutschen Ausbildungssystem (siehe Hentig, 1980).

Das Fach Mathematik tritt am Oberstufen-Kolleg zunächst in zwei etwas unterschiedlichen Fachausbildungsgängen in Kombination mit Physik sowie in freier Kombination mit anderen Wahlfächern auf. Außerdem gibt es spezielle Mathematik-Kurse in allen naturwissenschaftlichen Ausbildungsgängen (insbesondere Analysis) und in einigen sozialwissenschaftlichen Ausbildungsgängen (Statistik). Schließlich arbeiten Mathematik-Lehrende gemeinsam mit Fachlehrenden anderer Fächer ein Angebot von Kurssequenzen und Einzelkursen aus, die im Rahmen der interdisziplinären Allgemeinbildung die Funktion von Mathematik, ihre Verwendung und ihre Bedeutung innerhalb der Wissenschaften und der gesellschaftlichen Praxis vorstellen sollen. Diese Anforderung ist durch die Ausbildungs- und Prüfungsordnung vorgegeben, die vorsieht, daß alle

Schüler und Schülerinnen, die Mathematik nicht als Wahlfach oder Hilfswissenschaft kennen lernen, drei "mathematikhaltige" Kurse im Ergänzungsunterricht besuchen müssen (siehe auch Effe-Stumpf u.a., 1988).

Bisher sind am Oberstufen-Kolleg eine Reihe von derartigen Kurssequenzen unter Oberthemen wie "Mathematische Modelle in Alltag und Politik", "Logik und Sprache", "Frauen und Mathematik" oder "Informationsverarbeitung/Informatik" entwickelt und erprobt worden. Mathematik tritt in derartigen Sequenzen sehr verschiedenartig auf; gemeinsam ist allen Ansätzen, daß die Auswahl der mathematischen Inhalte unter der Leitidee exemplarischen Lernens steht. Auswahlgesichtspunkte sind beispielsweise, eine typische mathematische Vorgehens- und Denkweise an einem Theorieabschnitt vorzustellen, den Charakter mathematischer Modellbildung zu thematisieren, die Nützlichkeit und die Begrenztheit der Verwendung mathematischer Methoden zu erläutern. Die Auswahl der mathematischen Inhalte, ihre Einordnung in den jeweiligen Kursverlauf und ihre Vermittlung unterliegen nicht den curricularen und didaktischen Anforderungen der Rahmenrichtlinien für das Fach Mathematik an der Gymnasialen Oberstufe. Dennoch werden am Oberstufen-Kolleg in Einzelfällen im Ergänzungsunterricht Kursabschnitte entwickelt und erprobt, deren Übertragbarkeit in Grund- oder Leistungskurse denkbar ist und diskussionsanregend sein kann. Dies soll mit dem folgenden Beispiel ausgeführt werden.

2. Statistik im Rahmen von Kursen über empirische Methoden

Unter dem Oberthema "Wissenschaft und Gesellschaft" wurde für den Ergänzungsunterricht des Oberstufen-Kollegs eine Kurssequenz "Empirische Methoden in den Sozialwissenschaften" entwickelt und mehrfach erprobt (Effe-Stumpf u.a., 1988). Im Verlauf der Sequenz sollen beispielhaft Anlässe und Möglichkeiten entstehen, den in den Naturwissenschaften und für die empirischen Sozialwissenschaften grundlegenden Prozeß des Messens und damit die Idee des quantitativen Vorgehens verstehbar und diskutierbar zu machen. Als Einstiegskurs wurden verschiedenartige Kurse zum

Meßprozeß erprobt, die jeweils zum Ziel hatten, sowohl Begriffsklärungen für "Messen" in Alltag und Wissenschaften zu leisten als auch die mathematische Seite des Berechnens von und Rechnens mit Meßergebnissen vorzustellen (vgl. 2.1 weiter unten). Hierzu wurde auch ein Stoffabschnitt zur beschreibenden Statistik und linearen Regressionsrechnung entwickelt, der in Abschnitt 3 hier vorgestellt wird. Als *Folgekurs* wurde ein Kurs "Intelligenztests auf dem Prüfstand" angeboten, der sich am Beispiel der Intelligenzmesseung mit der Entstehung und Anwendung wissenschaftlicher Theorien und Verfahren befaßt. In einem Kursabschnitt wird sehr elementar und überblicksartig in das Konzept "Normalverteilung" eingeführt; im Verlauf des Kurses entstehen vielfach Anlässe zu Wiederholung und Vertiefung der Grundbegriffe aus der beschreibenden Statistik und des Verfahrens der Korrelationsrechnung. - Der Kursaufbau insgesamt und damit seine sozialwissenschaftlichen und test-theoretischen Teile können und sollen hier nicht dargestellt werden.

Innerhalb der Kurssequenz wurde ein *weiterer Folgekurs* angeboten, der auch unabhängig vom Kurs über Intelligenztests besucht werden kann; es werden methodologische Fragen (z.B. über Schätzen und Testen von Hypothesen) und wissenschaftstheoretische Begründungen (als Beispiel sei der Ansatz des "Kritischen Rationalismus" nach K. Popper genannt) behandelt, die für das Verständnis des Vorgehens der empirischen Wissenschaften von besonderer Bedeutung sind.

2.1 Messen und Statistik

Wenn zu Beginn eines Kurses zum Meßprozeß den Lernenden Anstöße und Gelegenheiten gegeben werden, ihre Vorstellungen von und Erfahrungen mit Meßvorgängen zu erinnern und einzubringen, entsteht ein reicher Schatz an Ausgangsmaterial für Diskussionen und Erläuterungen im Unterricht. Ein solcher Einstieg macht kritisch aufnahmebereit dafür, entsprechende Definitionen und Beschreibungen aus Lexika, Handbüchern und methodisch einführender Fachliteratur herauszusuchen und zu vergleichen. Dabei sind die Fähigkeit, verschiedene Meßniveaus (Meßskalen) zu unterscheiden,

und ein Verständnis für die eingrenzende Kennzeichnung der quantitativen Messung eines Merkmals (Messung auf Intervall- oder Verhältnisniveau) besonders herauszubilden.

Mit Messen verbinden sich zunächst und meist Vorstellungen naturwissenschaftlicher Art; Naturgesetze werden bei Meßvorgängen und Meßgeräten benutzt, Unregelhaftes oder Gesetzmäßiges soll durch Messen entdeckt, aufgedeckt werden. In der Regel werden verschiedene Merkmale der Untersuchungsobjekte gemessen, und diese Meßergebnisse sollen miteinander verglichen oder untereinander in Beziehung gesetzt werden; dabei wird häufig ein linearer oder allgemeiner ein funktionaler Zusammenhang zwischen den Meß-Variablen vermutet und gesucht. Dies legt nahe, im Statistik-Unterricht überhaupt und frühzeitig Darstellungs- und Beurteilungsmethoden der Statistik mehrerer Variablen (zumindest in sehr einfacher Form) zu behandeln. Solche Methoden stehen mit der linearen Regressions- und Korrelationsrechnung zur Verfügung; in Teil 3 wird ein elementarer Zugang zu diesem Gebiet vorgestellt.

2.2 Anmerkungen zur beschreibenden Statistik

Grundbegriffe der beschreibenden Statistik wie Häufigkeitsverteilungen und Darstellungsformen, Mittelwerte und Streuungsmaße werden manchmal schon im Unterricht der Sekundarstufe I behandelt. Bei einer sehr heterogenen Schülerpopulation erscheint es dennoch auch für Statistik-Kurse in der Sekundarstufe II sinnvoll, zunächst unterschiedliche Beispiele von Merkmalsmessungen konkret vorzunehmen und auszuwerten. Im folgenden wird eine Liste mit Meßergebnissen aus einem derartigen Anfängerkurs vorgelegt; nach dem unter 2.1 angedeuteten Einstieg über das Messen wurden Beispiele für Messungen in der Kursgruppe abgesprochen und ausgeführt (siehe die Tabelle auf der folgenden Seite).

Im Kurs hatte sich eine Phase mit Gruppenarbeit in drei Kleingruppen mit verschiedenen Arbeitsaufträgen angeschlossen; die Ergebnisse wurden anschließend im Plenum zusammengetragen, vertieft und zum Teil weitergeführt. Eine Kleingruppe hat gelernt und anschließend den übrigen erläutert, wie mit kleinen Computer-Pro-

grammen die Meßdaten eingelesen und in Listenform übersichtlich ausgedruckt werden können.

Tabelle: Messungen in der Kursgruppe

Nr.	Name	Alter Mon.	Schul-abschl.	Fächer	Größe cm	Gew. kg	Schuh-größe	Motivation
1	Martin	193	LS/QV	Tec/Mat	195.0	82.5	46.5	interessant
2	Lars	203	LS	Spo/Gew	166.0	50.0	41	brauchbar
3	Doris	242	RS/QV+LE	Öko/Geg	174.0	55.0	38	interessant
4	Carsten	200	HS/QV	Öko/Mat	184.5	68.0	43.5	wissenswert
5	Guido	233	HS/QV	Phy/Mat	176.5	66.0	41	brauchbar
6	Jörg	257	GY/QV+LE	Phy/Mat	188.5	95.0	45	interessant
7	Sabine	281	RS+LE	Geg/Spa	179.0	60.0	41	interessant
8	Thomas	229	RS	Che/Mat	183.0	75.0	43	wissenswert
9	Christian	199	HS	Phy/Mat	178.0	69.0	42	interessant
10	Richard	258	ARS/QV	Tec/Mat	186.5	77.0	44	interessant
11	Michael	216	RS	Che/Mat	196.0	84.5	44	brauchbar
12	Holger	208	RS	Bio/Mat	167.0	75.5	41	wissenswert
13	Sven	197	HS/QV	Phy/Mat	190.0	84.5	44	interessant
14	Karsten	220	RS	Tec/Mat	178.5	51.0	42.5	interessant
15	Thorsten	198	LS/QV	Tec/Mat	178.5	76.0	42	interessant
16	David	202	HS	Phy/Mat	186.0	76.0	46	wissenswert

'Schulabschluß': GY Sek.I am Gymnasium; ARS, RS (Abend-)Realschule; HS Hauptschule; LS Laborschule,eine besondere Bielefelder Gesamtschule; QV Qualifikationsvermerk; LE abgeschlossene Lehre.

'Fächer': (hier vorläufig noch) Wahlfächer.

'Motivation': Skala: sehr interessant - interessant - wissenswert - brauchbar - uninteressant.

Eine weitere Kursgruppe hat sich in einfachen Schul- und Lehrbüchern über diskrete Häufigkeitsverteilungen und Möglichkeiten ihrer Darstellung informiert; dabei hat der Lehrende die Textauswahl (in der Bibliothek) beraten und zeitweise Verständnisfragen beantwortet. Die Gruppe hat sich auch mit Mittelwerten und Streuungsmaßen befaßt; sie hat der Gesamtgruppe die wichtigsten Grundbegriffe vorgestellt und ihre Verwendung an einigen der gemessenen Merkmale erläutert. Die meisten Schüler und Schülerinnen waren von den verschiedenen Möglichkeiten, Mittelwerte als Modalwert, Median oder als arithmetisches Mittel zu definieren, überrascht und zugleich angeregt, weitere Parameter zu suchen und auszuprobieren, die Verteilungen charakterisieren können. Es wurde dabei ausführlicher diskutiert, welche Hinweise das jeweilige Meßniveau eines Merkmals auf die Wahl eines Mittelwerts oder eines anderen Parameters gibt, und, welche Aussagemöglichkeiten sich damit ergeben.

Der Arbeitsbereich der dritten Kleingruppe wird hier nur angedeutet: Sie ist der Verwendung des Begriffs Messen und seiner Bedeutung in einer Reihe verschiedener, selbst gewählter Fächer (Jura, Ökonomie, Sport, Musik, Biologie) in Fachbüchern und Gesprächen mit Lehrenden nachgegangen und hat versucht, die Ergebnisse in den Kontext des Kurseinstiegs einzuordnen.

Es folgen nun noch Anmerkungen zur Besprechung der Ergebnisse der zweiten Arbeitsgruppe und zur Weiterführung durch den Lehrenden im Plenum: Die Summen- und Indexschreibweise waren zumeist unvertraut; sie lassen sich allerdings in diesem Bereich dann auch vielfältig üben. Beispielsweise wurden für (quantitative) Meßwerte $x_1, x_2, .. x_n$ und einen vorgegebenen Wert x verschiedene Abweichungsmaße erläutert und diskutiert; sodann wurden mathematische Eigenschaften (teilweise in Verbindung mit geometrischen Überlegungen) algebraisch formuliert und zumeist vom Lehrenden im Tafelvortrag bewiesen; Schüler und Schülerinnen haben Beweisteile danach wiederholend für die anderen erläutert:

- Es gilt: $\Sigma\ (x_i - x) = 0$ genau dann, wenn $x = \bar{x}$ (arithmetisches Mittel).
- $\Sigma\ |x_i - x|$ hat für $x = \tilde{x}$ (Median) einen minimalen Wert.
- $\Sigma\ (x_i - x)^2$ hat als in x quadratischer Funktionsterm für $x = \bar{x}$ seinen minimalen Wert.

Die Varianz wurde mit

$$s_x^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

eingeführt, die Standardabweichung als positive Wurzel daraus. Als weiteres Beispiel wurde die für Berechnungen meist günstigere Darstellung der Formel als

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

bewiesen. Die Übung mit den Schreibweisen bereitet auch schon auf die späteren Rechnungen zur linearen Regression und Korrelation vor und erleichtert die dortigen Überlegungen.

Es sei noch angemerkt, daß durchgehend bei der Streuungsmessung, linearen Regressions- und Korrelationsrechnung bewußt auf das Stichproben-Maß (mit Faktor $\frac{1}{n-1}$ statt $\frac{1}{n}$) zugunsten der Einfachheit und leichteren Erklärbarkeit verzichtet wurde. Eine

ausführliche Darstellung dieser verschiedenen Maße mit Begründungen für ihre Verwendung findet sich in den meisten Lehrbüchern zur Statistik (etwa Bosch, 1976).

3. Lineare Regression und Korrelationskoeffizient

Zunächst wird in 3.1 in einem konkreten Beispiel vorgestellt, wie Motivation und Vorverständnis für den Ansatz der linearen Regression angeregt werden können. In 3.2 und 3.3 wird in Form eines Arbeitstextes ein Zugang zu den Begriffen Regressionsgerade und (Produkt-Moment-)Korrelationskoeffizient vorgestellt, der keine (partiellen) Ableitungen aus der Differentialrechnung verwendet. Ein ähnlicher Text wurde als Grundlage einer ausführlichen Bearbeitung im Unterricht erprobt.

3.1 Ein einführendes Beispiel

Aus der Urliste zu den Messungen in der Kursgruppe (vgl. 2.2) wurden die quantitativen Merkmale Körpergröße (x) und Körpergewicht(y) herausgegriffen.

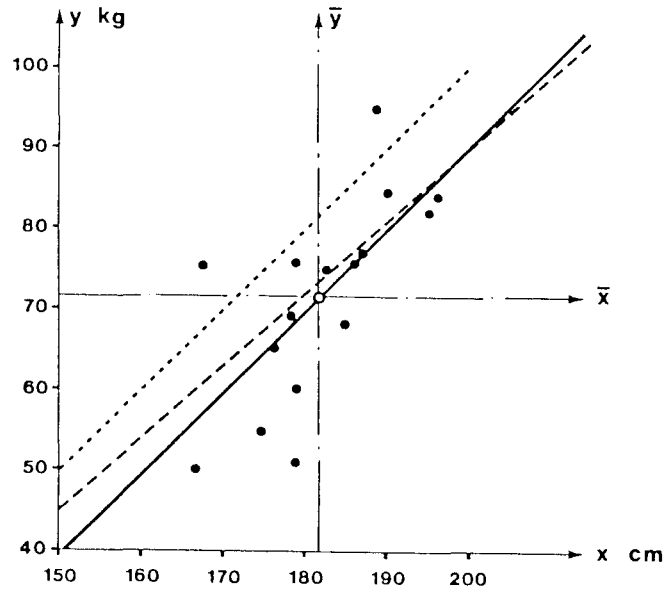
i	1	2	3	4	5	6	7	8	9	10
x_i	195,0	166,0	174,0	184,5	176,5	188,5	179,0	183,0	178,0	186,5
y_i	82,5	50,0	55,0	68,0	66,0	95,0	60,0	75,0	69,0	77,0

11	12	13	14	15	16		
196,0	167,0	190,0	178,5	178,5	186,0	(cm)	$\bar{x} = 181{,}6875$
84,5	75,5	84,5	51,0	76,0	76,0	(kg)	$\bar{y} = 71{,}5625$

Die Meßwert-Paare (x_i, y_i) wurden als Punktwolke dargestellt (vgl. die Graphik auf der folgenden Seite). Bei Versuchen, die Gestalt der Punktwolke zu beschreiben, wurden zunächst die jeweiligen arithmetischen Mittelwerte $\bar{x}$, $\bar{y}$ hinzugezogen und ein Hilfs-Achsenkreuz durch den 'Schwerpunkt' $(\bar{x}, \bar{y})$ gelegt; dann lassen sich Merkmalspaare durch Kombination der Kennzeichnungen unter- und überdurchschnittlich groß bzw. schwer unterscheiden. Sodann wurden weitergehende Zusammenhänge zwischen den Variablen x und y diskutiert. Dabei spielten Vorkenntnisse über Normal- und Ideal-

Gewichtstabellen und entsprechende Formeln eine anleitende Rolle.

Abb.: Punktwolke der Körpergröße - Körpergewicht - Daten mit drei Schätzgeraden:
..... Normalgewichts-Gerade
--- Idealgewichts-Gerade
—— Regressionsgerade (von y aus x)



Für Männer wurden folgende Formeln erinnert (vgl. Meyer, 1980, S. 194ff; die Funktion und Problematik solcher Formeln wird dort auch angesprochen):

(1) Normalgewicht (in kg) = Körpergröße (in cm) - 100

$y = x - 100$

(2) Idealgewicht (in kg) = 90% (für Frauen 85%) vom Normalgewicht (in kg)

$y = 0{,}9(x-100)$

Die Abweichung zwischen den Punkten im Diagramm und den jeweiligen Geraden sollten beurteilt werden. Die Lage der Normalgewichtsgeraden in Bezug auf die Punktwolke erschien offensichtlich als zu hoch (2 Punkte liegen darüber, 14 darunter). Für die Idealgewichtsgerade wurden Möglichkeiten einer genaueren Abweichungsmessung erörtert und (in Erinnerung an das Streumaß) wieder die Quadratsumme aller Abweichungen vorgeschlagen. Die (willkürliche) Entscheidung für die Messung vertikaler Abweichungen (d.h. in y-

Richtung war hier naheliegend, weil die Formeln eine Schätzung des Gewichts, ausgehend von der Größe, vornehmen. Für die mittlere quadratische Abweichung (in y-Richtung) zwischen den Meßwert-Paaren (x_i,y_i) und den zugehörigen Punkten $(x_i,\hat{y}_i)$ auf einer Schätzgeraden wurde daher folgendes Maß gewählt:

$$(3) \qquad s^2_{y\hat{y}} := \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Zur Übung wurden mit Hilfe einer Tabelle, die neben den Meßwerten die gemäß (1) und (2) berechneten Schätzwerte enthielt, folgende Werte berechnet: $s^2_{y\hat{y}} \approx 180{,}72$ bzw. $s^2_{y\hat{y}} \approx 83{,}10$ für die Normal- bzw. die Idealgewichtsgerade. Schließlich wurde noch der Vorschlag diskutiert, die Idealgewichtsgerade parallel so zu verschieben, daß sie durch den Schwerpunkt $(\bar{x},\bar{y})$ verläuft, also die Schätzgerade mit folgender Gleichung zu betrachten:

$$(4) \qquad \hat{y} = 0{,}9(x-\bar{x}) + \bar{y} \quad \text{mit } \bar{x} = 181{,}6875;\ \bar{y} = 71{,}5625.$$

(Diese Gerade wurde nicht in das Diagramm eingetragen.) Als Begründung für die Parallel-Verschiebung wurde u.a. vorgebracht, daß diese Gerade noch "näher" zu den Meßpunkten liege (eine Kontrollrechnung ergab tatsächlich $s^2_{y\hat{y}} \approx 79{,}27$) und daß der Mittelwert der Schätzwerte $\hat{y}_i$ ebenfalls $\bar{y}$ sein soll. Die Diskussion dieser Behauptung führte zu der (zutreffenden) verallgemeinernden Vermutung:

(5) Werden an den Stellen $x_1,x_2,..,x_n$ Schätzwerte $\hat{y}_1,\hat{y}_2,..\hat{y}_n$ durch die Schätzgerade mit Gleichung $\hat{y} = m(x-\bar{x})+d$ berechnet, so gilt für ihren Mittelwert: $\frac{1}{n}\sum \hat{y}_i = d$.

Der Beweis dieser Vermutung folgt durch einfaches Nachrechnen, wobei $\sum (x_i-\bar{x}) = 0$ benutzt wird (vgl. 2.2).

Bei diesem Verlauf der Behandlung des einführenden Beispiels bot sich nun die Frage an, ob und wie sich unter den Schätzgeraden, die durch den Schwerpunkt verlaufen, eine solche mit minimalem mittleren quadratischen Abstand finden läßt. Es wurden versuchsweise einige $s^2_{y\hat{y}}$-Werte für verschiedene Steigungszahlen m berechnet:

m	0,8	0,9	1,0	1,1	1,2	1,02
$s^2_{y\hat{y}}$	81,76	79,27	78,20	78,56	80,34	78,16

Abschließend wurde vermutet, daß knapp über m=1 ein günstigster Wert liegen müsse; einige Schüler haben die Werte mit einem Computer-Programm nachgerechnet und einen günstigsten Wert mit $m \approx 1{,}025$ "eingeschachtelt".

3.2 Lineare Regression

Der folgende Optimierungssatz läßt sich elementar, d.h. ohne Differentialrechnung beweisen:

(6) Gegeben seien n Meßwert-Paare $(x_1,y_1), (x_2,y_2),..,(x_n,y_n)$ mit Standardabweichung $s_x \neq 0$ (d.h. nicht alle x_i sind untereinander gleich). Dann gibt es genau eine Gerade mit der Gleichung $\hat{y} = m(x-\bar{x})+d$, für die die mittlere quadratische Abweichung in y-Richtung $s^2_{y\hat{y}} = \frac{1}{n} \Sigma (y_i-\hat{y}_i)^2$ minimal ist.
Diese beste Schätzgerade (bei Schätzung von y durch x) wird Regressionsgerade genannt; sie verläuft durch den Schwerpunkt $(\bar{x},\bar{y})$, d.h. $d=\bar{y}$, und für ihre Steigung gilt: $m = s_{xy}/s_x^2$, wobei $s_{xy} = \frac{1}{n} \Sigma (x_i-\bar{x})(y_i-\bar{y})$ die Kovarianz bezeichnet. Die Regressionsgerade hat also die Gleichung:

$$\hat{y} = \frac{s_{xy}}{s_x^2}(x-\bar{x}) + \bar{y}$$

Beweis: Die Summation erstreckt sich, wie schon in der Formulierung des Satzes jeweils über i=1,..,n. Es wird die Geradengleichung in der Form $\hat{y} = m(x-\bar{x}) + (\bar{y}+c)$, d.h. mit $c := d-\bar{y}$, benutzt. Es gilt:

$$s^2_{y\hat{y}} = \frac{1}{n} \Sigma \{y_i - m(x_i-\bar{x}) - (\bar{y}+c)\}^2 = \frac{1}{n} \Sigma \{(y_i-\bar{y}) - m(x_i-\bar{x}) - c\}^2 =$$

$$= \frac{1}{n} \Sigma \{(y_i-\bar{y})^2 + m^2(x_i-\bar{x})^2 + c^2 - 2m(x_i-\bar{x})(y_i-\bar{y}) - 2c(y_i-\bar{y}) + 2mc(x_i-\bar{x})\}$$

$$= \frac{1}{n}\Sigma (y_i-\bar{y})^2 + \frac{m^2}{n}\Sigma(x_i-\bar{x})^2 + c^2 - \frac{2m}{n}\Sigma (x_i-\bar{x})(y_i-\bar{y}) - \frac{2c}{n}\Sigma (y_i-\bar{y}) + \frac{2mc}{n}\Sigma (x_i-\bar{x})$$

$$= s_y^2 + m^2 s_x^2 + c^2 - 2m s_{xy}.$$

Offensichtlich ist $c^2=0$, also $d=\bar{y}$ eine notwendige Bedingung dafür, daß $s^2_{y\hat{y}}$ einen minimalen Wert annimmt; damit ist bereits gezeigt: Wenn es eine optimale Schätzgerade gibt, dann muß diese durch den Schwerpunkt verlaufen. Für solche Schätzgeraden mit der Gleichung $\hat{y} = m(x-\bar{x})+\bar{y}$ hängt aber die Abweichung $s^2_{y\hat{y}}$ quadratisch von m ab:

$$\begin{aligned} s^2_{y\hat{y}} &= m^2 s_x^2 + s_y^2 - 2m\, s_{xy} \\ &= s_x^2(m^2 - 2m\, s_{xy}/s_x^2) + s_y^2 \\ &= s_x^2(m - s_{xy}/s_x^2)^2 + s_y^2 - s_{xy}^2/s_x^2 . \end{aligned}$$

Wegen $s_x^2 > 0$ läßt sich aus dieser Scheitelpunktgleichung der in m quadratischen Funktion nunmehr unmittelbar ablesen:

(7) $s^2_{y\hat{y}}$ hat für $m = s_{xy}/s_x^2$ ein (einziges) Minimum und dort gilt:

$$s^2_{y\hat{y}} = s_y^2 - \frac{s_{xy}^2}{s_x^2} .$$

Fortführung des Beispiels in 3.1:

Mit den Werten des Beispiels in 3.1 gilt: $s_x^2 = 71{,}12109$; $s_y^2 = 152{,}87109$; $s_{xy} = 72{,}89453$; also $m = 1{,}02494$. Somit hat die Regressionsgerade die Gleichung $\hat{y} = 1{,}025(x-181{,}69) + 71{,}56$.
Im Unterricht sollten nun eine Reihe weiterer Anwendungsbeispiele anschließen, die mit der Methode und den Anwendungsmöglichkeiten des Modells der linearen Regression vertraut machen; damit kann zugleich Motivation für die Frage nach der Güte der linearen Annäherung (vgl. 3.3) geweckt werden.

3.3 Der Korrelationskoeffizient

Um zu beurteilen, wei gut n Meßwert-Paare (x_1,y_1), (x_2,y_2),.., (x_n,y_n) durch die Regressionsgerade angenähert werden können, wird der Korrelationskoeffizient r_{xy} bestimmt. Er stellt ein Beurteilungsmaß dafür dar, inwieweit die Varianz der y_i-Werte durch die Annahme eines linearen Zusammenhangs der y_i-Werte mit den x_i-Werten über die Varianz der x_i-Werte erklärt werden kann (dabei wird nur der nicht-triviale Fall $s_x^2 \neq 0$ und $s_y^2 \neq 0$ behandelt). Zu-

nächst gilt stets für die Varianz der mit einer Schätzgeraden $\hat{y} = m(x-\bar{x}) + d$ aus den x_i-Werten berechneten $\hat{y}_i$-Werte:

$$s^2_{\hat{y}} = m^2 s^2_x \ .$$

Dies folgt daraus, daß die $\hat{y}_i$-Werte durch eine affine Transformation aus den x_i-Werten entstehen. Es kann auch leicht mit der Varianz-Formel nachgerechnet werden, wobei wegen (5) der Mittelwert der $\hat{y}_i$-Werte gleich d ist.
Speziell für die Regressionsgerade mit $m = s_{xy}/s^2_x$ (vgl.(6)) gilt daher:

(8) $$s^2_{\hat{y}} = \left(\frac{s_{xy}}{s^2_x}\right)^2 \cdot s^2_x = \frac{s^2_{xy}}{s^2_x} \ .$$

Deshalb gilt für die Regressionsgerade die Gleichung (vgl.(7)):

(9) $$s^2_{y\hat{y}} = s^2_y - s^2_{\hat{y}} \quad \text{bzw.} \quad s^2_y = s^2_{\hat{y}} + s^2_{y\hat{y}} \ ,$$

d.h. die mittlere quadratische Abweichung zwischen den Meßwert-Paaren und der Regressionsgeraden (bei Schätzungen von y durch x) ergibt sich als Differenz zwischen der Gesamt-Varianz der y_i-Werte und der (nie größeren) Varianz der Schätzwerte $\hat{y}_i$.
Die mittlere quadratische Abweichung $s^2_{y\hat{y}}$ läßt sich für Schätzgeraden durch den Schwerpunkt (ebenfalls) als Varianz auffassen. Für die Differenz-Variable $y-\hat{y}$ gilt nämlich:

$$s^2_{y-\hat{y}} = \frac{1}{n} \sum ((y_i-\hat{y}_i) - \overline{(y-\hat{y})})^2 \quad \text{und} \quad \overline{y-\hat{y}} = 0,$$

da die y_i- und die $\hat{y}_i$-Werte den gleichen Mittelwert $\bar{y}$ haben.
Der Anteil $s^2_{\hat{y}}/s^2_y$ der durch die Regressionsgerade "erklärbaren" Varianz $s^2_{\hat{y}}$ an der Gesamtvarianz s^2_y kann als Maß für die Güte der linearen Regression (bei Schätzung von y durch x) verwendet werden; wegen (8) gilt:

$$\frac{s^2_{\hat{y}}}{s^2_y} = \frac{s^2_{xy}}{s^2_x s^2_y}$$

In der Statistik wird als Korrelationskoeffizient r_{xy} definiert:

(10) $$r_{xy} := \frac{s_{xy}}{s_x s_y}$$

Für den Korrelationskoeffizienten gilt:

(11) $r^2_{xy} \leq 1$; $r^2_{xy} = 1$ gilt genau dann, wenn alle Punkte auf einer Geraden liegen.

Beweis: (8) in (9) eingesetzt ergibt:

$$s^2_{y\hat{y}} = s^2_y - s^2_{\hat{y}} = s^2_y - \frac{s^2_{xy}}{s^2_x} .$$

Da $s^2_{y\hat{y}}$ als Summe von Quadraten nicht-negativ ist, gilt daher:

$$s^2_y \geq \frac{s^2_{xy}}{s^2_x} \quad \text{bzw.} \quad \frac{s^2_{xy}}{s^2_x \cdot s^2_y} \leq 1.$$

Es ist klar, daß $r^2_{xy} = 1$ gleichwertig zu $s^2_{\hat{y}} = s^2_y$, also gleichwertig zu $s^2_{y\hat{y}} = 0$ ist; genau dann, wenn $s^2_{y\hat{y}} = 0$ gilt, liegen alle Punkte auf einer Geraden.

Für den Korrelationskoeffizienten können daher folgende Eigenschaften zusammengefaßt werden:

(12) $-1 \leq r_{xy} \leq 1$;

Dabei gibt das Vorzeichen (das über die Kovarianz entsteht) die Steigung der Regressionsgeraden an und $r_{xy}=1$ gilt genau dann, wenn alle Meßpunkt-Paare auf einer Geraden liegen; $r_{xy}=0$ ist gleichbedeutend damit, daß die Regressionsgerade waagrecht durch den Schwerpunkt $(\bar{x},\bar{y})$ geht, d.h. die Varianz der y_i-Werte läßt sich durch die Annahme eines linearen Zusammenhangs mit den x_i-Werten überhaupt nicht erklären (die Variablen x und y sind unkorreliert: $s_{xy}=0$).

Abschließend sei noch angemerkt, daß die in diesem Abschnitt vorgestellte Deutung des Korrelationskoeffizienten über die Aufteilung der Varianz in der angelsächsischen Literatur üblich ist (vgl. Überla, 1971, S. 16).

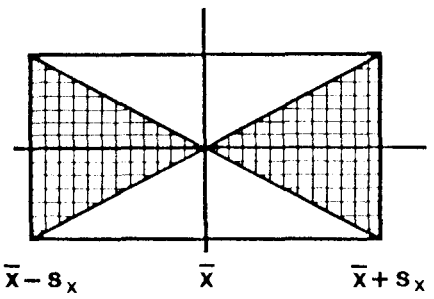
Fortführung des Beispiels in 3.1

Unter Benutzung der am Ende von 3.2 angegebenen Werte folgt: $r_{xy} \approx 0{,}699$; $r^2_{xy} \approx 0{,}489$. Etwa die Hälfte der Varianz des Körpergewichtes ließe sich durch die Annahme eines linearen Zusammenhangs aus der Varianz der Körpergröße erklären. Das Beispiel könnte unter verschiedenen Aspekten weiter bearbeitet werden. Wie wirken sich die Meßwert-Paare der beiden Frauen aus? Gibt es Untersuchungen zur Korrelation: Körpergröße/Körpergewicht in der statistischen biologischen/ medizinischen Literatur? Welche Koeffizienten werden dort angegeben, welche Erklärungen werden zur Be-

urteilung der Güte der Korrelation angeboten (z.B. verschiedene Körpertypen)?

3.4 Weitere Aspekte

1. Es ist eine nützliche Übung, die lineare Regression auch für Schätzungen von x durch y durchzuführen. Nur im Fall eines vollständigen linearen Zusammenhangs ($r_{xy}=\pm 1$) fallen beide Regressionsgeraden zusammen. Allgemein läßt sich die Lage der beiden Regressionsgeraden in einem Diagramm kennzeichnen, das sich ergibt, wenn man die in den Formeln ausgedrückte Standardisierung der Variablen x und y beachtet. Die Regressionsgerade bei Schätzung von y aus x verläuft durch den schraffierten Bereich, während die Regressionsgerade bei Schätzung von x aus y durch den nicht-schraffierten Bereich verläuft. Dies folgt aus der Hölderschen Ungleichung



$$-s_x \cdot s_y \leq s_{xy} \leq s_x \cdot s_y$$

durch Division mit s_x^2 unter Beachtung der Steigungsformel $m=s_{xy}/s_x^2$.

2. Ob und wie der hier vorgestellte Abschnitt zum Korrelationskoeffizienten in der gymnasialen Oberstufe behandelt werden kann und sollte, hängt gewiß von einer Reihe von Bedingungen ab, wie z.B. dem vorhandenen Zeitrahmen, den Lernvoraussetzungen, den Zielen des Statistik-Unterrichts insgesamt.

Literatur

Bosch, K.: Angewandte mathematische Statistik. Reinbeck bei Hamburg: Rowohlt, 1976.

Effe-Stumpf, Gerull, Kemper, Schülert, Vohmann: Anwendungsbezogene Mathematik am Oberstufenkolleg Bielefeld. Erscheint in ZDM 1988.

Hentig, H. O.: Die Krise des Abiturs und eine Alternative. Stuttgart: Klett-Cotta, 1980.

Hoffmann, B.(Hrsg.): Allgemeinbildung. AMBOS Unterricht: Wissenschaftspropädeutik, Bd. 22. Bielefeld: Oberstufen-Kolleg, 1986.

Meyer-Nachschlagewerk: Gesund und fit. Mannheim: Meyers Lexikonverlag, 1980.

Überla, K.: Faktorenanalyse. Berlin-Heidelberg-New York: Springer, 1971.