Lineare Regression und Korrelation

U. Mortensen
FB Psychologie und Sportwissenschaften,
Westfälische Wilhelms Universität

Korrigierte Version vom 4. 3. 2008

Inhaltsverzeichnis

1 Einleitung

In vielen Untersuchungen ist die Beziehung zwischen gemessenen Größen $X_1, X_2, \ldots, X_k$ von Interesse; man versucht dann, einen funktionalen Zusammenhang ϕ etwa zwischen X_1 und den restlichen Variablen herzustellen:

$$X_1 = \phi(X_2, \ldots, X_k).$$

Galton (1889) folgend spricht man von *Regression*; die Werte von X_1 werden auf die der übrigen zurückgeführt. Für $k > 2$ spricht man von *multipler Regression*, für $k = 2$ von *einfacher Regression*. Es wird zunächst die einfache Regression betrachtet.

Generell bedeutet die Interdependenz von Variablen noch nicht ihre kausale Abhängigkeit. So fand Yule (1926) (zitiert nach Kendall und Stuart, 1973, p. 291) eine ausgeprägte Beziehung zwischen (i) der Neigung zum Selbstmord und (ii) der Zugehörigkeit zur Church of England. Es wäre vermutlich verfehlt, die Zugehörigkeit zur Kirche als Ursache für die Selbstmordneigung anzusehen, oder umgekehrt die Selbstmordneigung als Ursache für die Zugehörigkeit zur Kirche. Mayo, White und Eysenck (1977) fanden eine Beziehung zwischen der Variablen "Geburt in einem bestimmten Sternzeichen" und bestimmten Persönlichkeitseigenschaften; auch hier läßt sich zeigen, dass von einer kausalen Beziehung zwischen Sternbild und Charakter nicht die Rede sein kann; im Kapitel über Partialkorrelationen wird darauf noch näher eingegangen. Die Frage, was denn Kausalität ist, ist sehr diffizil. Glücklicherweise muß sie hier nicht beantwortet werden, es genügt, festzuhalten, dass die Existenz einer Beziehung zwischen Variablen noch nicht die Existenz einer kausalen Beziehung impliziert.

Diskutiert man den Zusammenhang zwischen zwei Variablen $X_1 = X$ und $X_2 = Y$, so bezieht man sich jeweils auf eine *bestimmte* Art von Zusammenhang. Von besonderem Interesse ist dabei die lineare Beziehung

$$y = \varphi(x) = ax + b + e \tag{1}$$

wobei e einen Meßfehler ("error") repräsentiert. X heißt auch *Prädiktorvariable* und Y wird auch die *Kriteriumsvariable* oder *Zielgröße* genannt. Natürlich können auch nichtlineare Beziehungen betrachtet werden, etwa

$$y = ax^b + e \tag{2}$$

oder

$$y = ax + bx^2 + cx^3 + d + e, \tag{3}$$

wobei a, b, c und d reelle Konstanten sind und e wieder einen "Meßfehler" repräsentiert. Es wird zunächst auf die lineare Beziehung (1) diskutiert.

Es gibt zwei Fälle, die oft als verschiedene *Modelle* der Regression behandelt werden, obwohl der eine Fall sich aus dem anderen als Spezialfall herleiten läßt. Die Modelle sind wie folgt charakterisiert:

> **Modell 1:** Es wird die lineare Beziehung (1) angenommen und man hat für jeden X-Wert $x_1, ..., x_k$ jeweils $n_1, ..., n_k$ Meßwerte y_{ij}, $i = 1, ..., k$, $j = 1, ..., n_i$.
>
> **Modell 2:** In vielen Untersuchungen erhebt man eine Stichprobe von UEen und bestimmt bei jeder UE zwei Meßwerte, d.h. das Paar (x, y). Man erhält die Menge $\{(x_i, y_i)|i = 1, ..., n\}$, die sich im (X, Y)-Koordinatensystem als "Wolke" von Punkten darstellen läßt. Zu jedem x-Wert gibt es also im Unterschied zu Modell 1 nur einen y-Wert.

Beide Modelle sind *lineare Modelle*. Gegeben sind also stets nur die Meßwerte x_i und y_{ij}, und die Aufgabe besteht darin, die unbekannten *Parameter* a und b aus diesen Meßwerten zu bestimmen. Man kann nun die Beziehung $y = ax + b + e$ als eine *Regel* auffassen, die es uns gestattet, aufgrund eines gegebenen X-Wertes einen Y-Wert vorherzusagen. Natürlich soll die Vorhersage so gut wie möglich sein, d.h. der Fehler bei der Vorhersage soll so gering wie möglich sein. Je größer die Fehler e sind, desto schlechter wird uns eine Vorhersage gelingen. In jedem Fall benötigt man ein Verfahren, die Parameter a und b für einen gegebenen Datensatz $\{(x_i, y_{ij})\}$ so zu bestimmen, dass der Vorhersagefehler so klein wie möglich wird. Der gängige Ansatz besteht in der Anwendung der *Methode der Kleinsten Quadrate*.

2 Schätzung der Parameter

2.1 Modell 1

Bei der j-ten Messung für den X-Wert x_i ergibt sich y_{ij}, und für diesen Wert soll gelten

$$y_{ij} = ax_i + b + e_{ij}, \quad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \tag{4}$$

a und b sind die zu schätzenden *Regressionsparameter*. Es sei

$$\hat{y}_i = ax_i + b. \tag{5}$$

Dann kann y_{ij} auch in der Form

$$y_{ij} = \hat{y}_i + e_{ij} \tag{6}$$

geschrieben werden. Mittelt man andererseits für festes i die y_{ij}, so erhält man

$$\bar{y}_{i\cdot} = ax_i + b + \bar{e}_i \tag{7}$$

wobei $\bar{e}_i$ der mittlere Fehler bei den Meßwerten für x_i ist. Die vorhergesagten Werte $\hat{y}_i$ und die mittleren Werte $\bar{y}_i$ unterscheiden sich also durch den mittleren Fehler $\bar{e}_i$.

Gemäß der Methode der Kleinsten Quadrate (MKQ) ist die Funktion

$$Q(a, b) = \sum_{ij} e_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - ax_i - b)^2 \tag{8}$$

als Funktion von a und b zu minimieren. Dazu werden die partiellen Ableitungen vonn Q nach a bzw. b gebildet und jeweils gleich Null gesetzt; man erhält zwei Gleichungen in zwei Unbekannten $\hat{a}$ und $\hat{b}$, die die gesuchten Schätzungen für a und b sind. Man wird zu dem folgenden Satz geführt:

Satz 2.1 *Es werde angenommen, dass die Variablen X und Y Intervallskalenniveau haben und dass die Beziehung $Y = aX + b + e$ gilt, wobei e Meßfehler sowie den Effekt nicht kontrollierter Variablen repräsentiere. Die Kleinste-Quadrate-Schätzungen $\hat{a}$ und $\hat{b}$ für a und b auf der Basis der Meßwerte x_i, y_{ij}, $i = 1, \cdots k$, $j = 1, \cdots, n_i$ ergeben sich als Lösungen der beiden Gleichungen*

$$\begin{aligned} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_i y_{ij} &= \hat{a} \sum_{i=1}^{k} n_i x_i^2 + \hat{b} \sum_{i=1}^{k} n_i x_i \\ \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} &= \hat{a} \sum_{j=1}^{k} n_i x_i + n\hat{b} \end{aligned} \tag{9}$$

mit $n = n_1 + n_2 + \cdots + n_k$. Die Lösungen $\hat{a}$ und $\hat{b}$ sind durch

$$\hat{a} = \frac{\sum_i n_i x_i \bar{y}_i - n\,\bar{x}\,\bar{y}}{\sum_i n_i x_i^2 - n\,\bar{x}^2} \tag{10}$$

$$\hat{b} = \bar{y} - \hat{a}\,\bar{x} \tag{11}$$

gegeben.

Beweis: Aus (4) folgt $e_{ij} = y_{ij} - ax_i - b$. Dann ist

$$Q(a,b) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - a\,x_i - b)^2$$

Für die partiellen Ableitungen nach a und b an der Stelle $(\hat{a}, \hat{b})$ gilt dann

$$\frac{\partial Q(a,b)}{\partial a}|_{a=\hat{a}} = -2\sum_{i,j}(y_{ij} - \hat{a}x_i - \hat{b})x_i = \sum_{i,j} e_{ij}x_i = 0 \tag{12}$$

$$\frac{\partial Q(a,b)}{\partial b}|_{b=\hat{b}} = -2\sum_{i,j}(y_{ij} - \hat{a}x_i - \hat{b}) = \sum_{i,j} e_{ij} = 0 \tag{13}$$

Die erste Gleichung ist äquivalent $\sum_{ij} x_i y_{ij} - \hat{a}\sum_{ij} x_i^2 - \hat{b}\sum_{ij} x_i = 0$; dabei ist $\sum_{ij} x_i^2 = \sum_i n_i x_i^2$ und $\sum_{ij} x_i = \sum_i n_i x_i = n\overline{x}$, und $n\overline{y} = \sum_{ij} y_{ij}$, denn es wird ja jeweils n_i-mal über die Konstante x_i^2 bzw. x_i summiert. Analoge Betrachtungen gelten für die zweite Gleichung. Dann folgen sofort die Normalgleichungen (9), deren Auflösung die Gleichungen (10) und (11) liefern. □

Anmerkungen:

1. **Normalgleichungen:** Die Gleichungen (9) heißen auch *Normalgleichungen.*

2. **Kovarianz:** Der Ausdruck $\sum_i n_i x_i \bar{y}_i - n\,\bar{x}\,\bar{y}$ im Zähler für $\hat{a}$ ist von spezieller Bedeutung und rechtfertigt eine eigene Definition:

   Definition 2.1 *Es seien X und Y zwei mindestens intervallskalierte Variablen mit den arithmetischen Mitteln $\bar{x}$ und $\bar{y}$. Dann heißt*

   $$s_{xy} = Kov(x,y) = \frac{1}{n-k}\sum_{ij}(x_i - \bar{x})(y_{ij} - \bar{y}) = \frac{1}{n}\sum_{ij} n_i x_i \bar{y}_i - \bar{x}\bar{y} \tag{14}$$

   *die (*Stichproben-) Kovarianz *für X und Y.*

   Der Ausdruck "Kovarianz" soll andeuten, dass es sich bei $\mathrm{Kov}(y,x)$ bzw. $\mathrm{Kov}(x,y)$ um ein Maß für die Kovariation der Merkmale, d.h. für das Ausmaß, in dem die Variablen in jeweils gleicher Richtung von den entsprechenden Mittelwerten $\bar{x}$ und $\bar{y}$ abweichen.

3. Der Nenner $\sum_i n_i x_i^2 - n\bar{x}^2$ in (10) entspricht dem Schätzer

   $$\hat{s}_x^2 = \frac{1}{n-k}\sum_i n_i x_i^2 - n\bar{x}^2 \tag{15}$$

   für die Varianz der x-Werte; mithin kann

   $$\hat{a} = \frac{\mathrm{Kov}(x,y)}{s_x^2} \tag{16}$$

   geschrieben werden.

4. **Interpretation von $\hat{a}$:** Die Variablen x und y haben nicht notwendig die gleichen Maßeinheiten. Man kann etwa die Regression des Körpergewichts (y) auf die Körperlänge (x) betrachten; x und y haben dann verschiedene Maßeinheiten. Bezeichnet man mit $[x]$ und $[y]$ die Maßeinheiten von x und y, so findet man, dass $\mathrm{Kov}(x, y)$ durch $[x][y]$ charakterisiert ist und demnach $\hat{a}$ durch $[x][y]([x]^2 = [y]/[x]$. Dies erschwert die Interpretation des numerischen Wertes von $\hat{a}$; es ergibt sich die Frage, was ein "großer" und was ein "kleiner" Wert von $\hat{a}$ ist. Eine Möglichkeit zur Umgehung dieser Frage besteht darin, zu standardisierten Variablen $z_x = (x-\bar{x})/s_x$ und $z_y = (y-\bar{y})/s_y$ überzugehen. Der Regressionskoeffizient für standardisierte Variablen hat die Form

$$\hat{a} \to r_{xy} = \frac{\mathrm{Kov}(x, y)}{s_x s_y} \tag{17}$$

und ist unabhängig von den speziellen Maßeinheiten. r_{xy} heißt *Produkt-Moment-Korrelationskoeffizient* und wird in Abschnitt 5 ausführlich besprochen.

Beispiel 2.1 Eine Gruppe von Gedächtnispsychologen hat einen Test für die Merkfähigkeit entwickelt, bei dem die Testwerte (*Scores*) als Meßwerte auf einer Intervallskala

Tabelle 1: Scores für Merkfähigkeit

	Alter			
x_i	35	45	55	65
n_i	5	5	4	6
y_i	120	85	120	55
	90	95	65	70
	110	90	80	110
	75	73	70	45
	115	120		50
				52
$\bar{y}_i$	102	92.6	83.75	63.67
$\hat{y}_i$	104.08	91.45	78.82	66.18

Abbildung 1: Merkfähigkeit und Alter

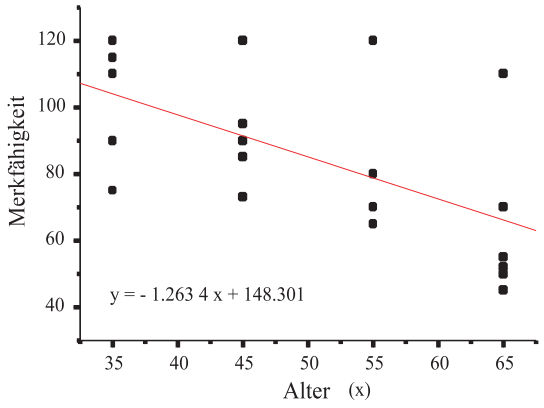
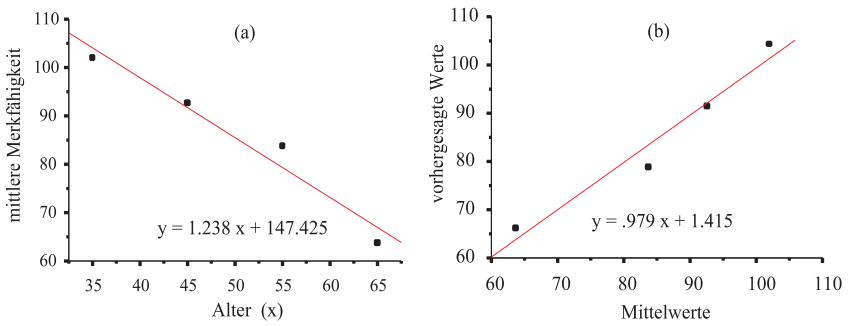


Abbildung 2: Merkfähigkeit und Alter



aufgefaßt werden können. Die Psychologen wollen nun die Merkfähigkeit in Abhängigkeit vom Alter untersuchen. Dazu bilden sie vier Stichproben von zufällig ausgewählten 35-, 45-, 55- und 65-jährigen Personen. Die Werte in der Tabelle (2.1) enthalten die Scores der Vpn.

Es ist $n = n_1 + \cdots + n_4 = 5+5+4+6 = 20$, $\bar{x} = (5\times 35+5\times 45+4\times 55+6\times 65)/20 = 50.50$, $\bar{y} = (120+90+\cdots+52)/20 = 84.50$. Weiter findet man $\sum_i n_i x_i \bar{y}_i = 8194.13$ und $\sum_i n_i x_i^2 = 53700$. Dann ergeben sich die Schätzungen $\hat{a} = -1.2634$, $\hat{b} = 148.301$. Dann lassen sich die vorhergesagten Werte $\hat{y}_i$ gemäß (5) bestimmen und mit den $\overline{y}_i$ vergleichen: s. Tabelle 2.1. Man kann auch die $\hat{y}_i$ gegen die $\overline{y}_i$ auftragen und für diese Werte eine Regressionsgerade bestimmen. Die $\hat{y}_i$-Werte sollen so gut wie möglich den $\overline{y}_i$-Werten entsprechen; demnach sollte $\hat{a} \approx 1$ und $\hat{b} \approx 0$ gelten; man findet $\hat{a} = .975, \hat{b} = .176$. Man sieht, dass der lineare Ansatz den Daten gut entspricht.

Offenbar ist die Linearität eine sinnvolle Annahme für die ausgewählten x_i-Werte. Gleichwohl ist das Ergebnis mit Vorsicht zu interpretieren. Denn zunächst muß bedacht werden, dass es sich hier um eine *Querschnittsuntersuchung* handelt, so dass über individuelle Verläufe der Merkfähigkeit mit zunehmendem Alter nichts oder nur wenig ausgesagt wird. Für die betrachteten Altersgruppen können Vorhersagen über die Merkfähigkeit einer bestimmten Person nur gemäß des folgenden Ansatzes gemacht werden: hat man eine Person in einer der Altersgruppen und hat man keine Vorinformation über sie, dann ist der beste Vorhersagewert für die Merkfähigkeit der Person der $\hat{y}_i$-Wert ihrer Altersgruppe. Liegen etwa Informationen über berufliche Tätigkeit etc. der Person vor, so können sie zur Verbesserung der Vorhersage dienen und der hier berechnete Wert $\hat{y}_i$ kann suboptimal sein.

Man kann sicher nicht *generell* eine lineare Beziehung zwischen dem Alter und der Merkfähigkeit postulieren. So könnte die Merkfähigkeit z.B. im Alter zwischen 20 und 35 Jahren konstant bleiben und zwischen 35 und 65 Jahren geringer werden, wobei die Beziehung zwischen Alter und Merkfähigkeit in erster Näherung linear sein mag, und für Personen, die älter als 65 Jahre alt sind, könnte die Merkfähigkeit beschleunigt abnehmen. □

2.2 Modell 2

Während es im Modell 1 für festgelegte x_i-Werte jeweils $n_i > 1$ Meßwerte y_{ij} gibt, wird beim Modell 2 für jede UE ein Paar (x_i, y_i) von Meßwerten bestimmt. Sicherlich werden

hierbei die x_i-Werte im allgemeinen nicht mehr gleichabständig sein, sie werden vielmehr in dem Maße zufällig variieren, in dem die UEen zufällig gewählt werden. Statt (4) wird jetzt jedenfalls

$$y_i = a\,x_i + b + e_i, \quad i = 1, \ldots, n \tag{18}$$

angesetzt. Die Methode der Kleinsten Quadrate liefert die Schätzungen

$$\hat{a} = \frac{\sum_i x_i y_i - n\,\overline{xy}}{\sum_i x_i^2 - n\,\overline{x}^2} \tag{19}$$

$$\hat{b} = \overline{y} - \hat{a}\overline{x} \tag{20}$$

Die Gleichungen folgen aus denen für das Modell I, indem einfach $n_i = 1$ gesetzt wird und die Summation von $i = 1$ bis n läuft. Teilt man in (19) auf der rechten Seite Zähler und Nenner durch $n - 1$, so erhält man für $\hat{a}$

$$\hat{a} = \frac{\mathrm{Kov}(x, y)}{s_x^2} \tag{21}$$

mit

$$\mathrm{Kov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy} \tag{22}$$

Dieser Ausdruck entspricht dem in (14) angegebenen Ausdruck für die Kovarianz für den Spezialfall $n_i = 1$ für alle i; und

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \tag{23}$$

Im Modell 1 sind die x_i-Werte vorgegeben und die Y-Werte sollen vorausgesagt werden. In Modell 2 werden Paare (x_i, y_i) von Meßwerten betrachtet, und ebenso, wie man die y_i-Werte nun aufgrund der x_i-Werte voraussagen kann, können auch die x_i-Werte aufgrund der y_i-Werte vorausgesagt werden. Um diese beiden Fälle zu unterscheiden, ist es sinnvoll, die Regressionsparameter entsprechend zu indizieren. Man schreibt

$$\hat{y}_i = \hat{a}_{yx} x_i + \hat{b}_{yx} \tag{24}$$

und analog

$$\hat{x}_i = \hat{a}_{xy} y_i + \hat{b}_{xy} \tag{25}$$

Statt (19) und (20) hat man dann

$$\hat{a}_{yx} = \frac{\mathrm{Kov}(x, y)}{s_x^2} \tag{26}$$

$$\hat{b}_{yx} = \overline{y} - \hat{a}_{yx}\overline{x} \tag{27}$$

Für (25) erhält man

$$\hat{a}_{xy} = \frac{\mathrm{Kov}(x, y)}{s_y^2} \tag{28}$$

$$\hat{b}_{yx} = \overline{x} - \hat{a}_{xy}\,\overline{y} \tag{29}$$

Die Parameter a_{yx} und a_{xy} geben die Steigungen der Regressionsgeraden an. Der Wert dieser Parameter wird einerseits durch Kov(x,y) bestimmt, andererseits durch s_x^2 bzw.

s_y^2. Gilt also $s_x^2 \neq s_y^2$, so wird auch $a_{yx} \neq a_{xy}$ gelten. Aus den Gleichungen (26) und (28) folgt sofort die Beziehung

$$\hat{a}_{yx} s_x^2 = \hat{a}_{xy} s_y^2. \tag{30}$$

Die Steigmaße $\hat{a}_{yx}$ und $\hat{a}_{xy}$ sind Maße für die Veränderung von $\hat{y}$ in Abhängigkeit von X bzw. von $\hat{x}$ in Abhängigkeit von Y; $\hat{y}$ etwa verändert sich um $\hat{a}_{yx}$ Einheiten, wenn sich X um eine Einheit verändert. Ist insbesondere $a_{yx} = 0$, so findet keine Veränderung der Y-Werte mit den X-Werten statt, X und Y sind dann unabhängig voneinander. Dann ist $\mathrm{Kov}(x, y) = 0$, und natürlich ist dann auch $a_{xy} = 0$.

Die Abhängigkeit zwischen X und Y impliziert also, dass $a_{yx} \neq 0, a_{xy} \neq 0$. Wie schon im Zusammenhang mit dem Modell I angemerkt, sind aber die Werte dieser Regressionskoeffizienten schwer zu interpretieren, da sie von den Maßeinheiten von X einerseits und Y andererseits abhängen. So folgt aus der Definition (14) von $\mathrm{Kov}(x, y)$, dass diese Größe die Dimensionen [X][Y] hat, wobei [X] und [Y] für die Einheiten von X und Y stehen. s_x^2 hat sicherlich die Dimension $[X]^2$. Also muß a_{yx} in Termen von $[X][Y]/[X]^2 = [Y]/[X]$ interpretiert werden. Umgekehrt wird a_{xy} in Termen von $[X][Y]/[Y]^2 = [X]/[Y]$ interpretiert. Die Komplexitäten einer solchen Interpretation lassen sich vermeiden, wenn man zum Begriff des Korrelationskoeffizienten übergeht, vergl. Abschnitt 5

3 Vorsichtsmaßnahmen: Pseudoregression

Korrelation durch Extremgruppen bei gleichzeitiger Nullkorrelation

4 Die Varianzzerlegung

Bisher ist angenommen worden, dass der lineare Ansatz $y = ax + b + e$ korrekt ist. Die Frage ist aber, ob diese Annahme zu rechtfertigen ist. Es zeigt sich, dass die Varianz s_y^2 der Y-Werte stets in additive Komponenten zerlegt werden kann, die zur Varianz der $\overline{y}_j$-Werte, der der $\hat{y}_i$-Werte und schließlich der Varianz der Fehler e_{ij} korrespondieren, und das Verhältnis der Größe dieser Komponenten zueinander erlaubt dann eine Bewertung der Angemessenheit des linearen Modells. Es wird zunächst die Zerlegung der Varianz der y-Werte vorgestellt; sie erweist sich auch in anderen Zusammenhängen als nützlich.

4.1 Die allgemeine Zerlegung der Quadratsummen

Es gebe k Gruppen von jeweils n_i y-Werten, $i = 1, \cdots, k$, vergl. Tabelle 2: Es sei noch

Tabelle 2: Gruppen von y-werten

G_1	G_2	$\cdots$	G_k
y_{11}	y_{12}	$\cdots$	y_{1k}
y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$	$\cdots$	$y_{n_k k}$
$\overline{y}_1$	$\overline{y}_2$	$\cdots$	$\overline{y}_k$
s_{1y}^2	s_{2y}^2	$\cdots$	s_{ky}^2

darauf hingewiesen, dass an dieser Stelle noch nicht angenommen wird, dass die y_{ij}-Werte in Zusammenhang mit einerr Regressionsanalyse betrachtet werden, die folgenden Betrachtungen gelten also ganz allgemein.

Die Gesamtvarianz der y-Werte kann nun in eine Varianz "innerhalb" und eine Varianz "zwischen" zerlegt werden. Die Varianz "innerhalb" entspricht einer Mittelung der in Tabelle 2 aufgeführten Varianzen s_{jy}^2, $j = 1, \cdots, k$ der y-Werte in den einzelnen Gruppen. Die Varianz "zwischen" entspricht der Varianz der Mittelwerte $\overline{y}_j$. Die Gesamtvarianz der y_{ij} ist dann durch

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \overline{y})^2 = \frac{QS_{ges}}{n}, \quad \text{mit } \overline{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} \tag{31}$$

gegeben. Die Zerlegung der Varianz s_y^2 entspricht eine Zerlegung der QS_{ges}, es gilt insbesondere der Satz

Satz 4.1 *(Varianzzerlegung) Es sei*

$$QS_{ges} = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \overline{y})^2, \quad QS_{inn} = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \overline{y}_j)^2, \quad QS_{zw} = \sum_{j} n_j(\overline{y}_j - \overline{y})^2 \tag{32}$$

Dann ist QS_{ges} gemäß

$$QS_{ges} = QS_{inn} + QS_{zw} \tag{33}$$

zerlegbar.

Beweis: Es ist

$$\begin{aligned} QS_{ges} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 &= \sum_{j=1}^{n_j}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j + \overline{y}_j - \overline{y})^2 \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + \sum_{j=1}^{k} n_j(\overline{y}_j - \overline{y})^2 \\ &\quad +2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})(y_{ij} - \overline{y}_j) \end{aligned} \tag{34}$$

Sicherlich ist

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})(y_{ij} - \overline{y}_j) = \sum_{j=1}^{k}(\overline{y}_j - \overline{y})\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j) = 0, \tag{35}$$

da stets $\sum_i(y_{ij} - \overline{y}_j) = 0$; also folgt (33). □

Es sei darauf hingewiesen, dass (35) gerade der Kovarianz zwischen den $(\overline{y}_j - \overline{y})$ und den $(y_{ij} - \overline{y}_j)$ entspricht:

> *Die Kovarianz zwischen den Mittelwerten $\overline{y}_j$ einerseits und den Meßwerten y_{ij} innerhalb der Gruppen andererseits ist stets gleich Null!*

4.2 Quadratsummenzerlegung bei linearer Regression

Bei den folgenden Betrachtungen wird das Modell 1 angenommen. Die Resultate können aber auf das Modell 2 übertragen werden, wenn $n_j = 1$ für alle j gesetzt wird; im Abschnitt über den Korrelationskoeffizienten wird hierauf explizit eingegangen.

Macht man den Ansatz $y_{ij} = ax_j + b + e_{ij}$, so kann man für die Parameter a und b Kleinste-Quadrate-Schätzungen (KQ-Schätzungen) $\hat{a}$ und $\hat{b}$ bestimmen, | es muß dann aber noch untersucht werden, ob der Ansatz, d.h. das lineare Modell, auch tatsächlich angemessen ist. Allgemein gilt für die k Gruppen, die durch die Werte x_j der unabhängigen Variablen definiert werden, der Satz 4.1. Darüber hinaus kann der folgende Satz bewiesen werden:

Satz 4.2 *Für den X-Wert x_j mögen die die Meßwerte $y_{ij}, j = 1, \cdots, n_j$ vorliegen, es sei $n = n_1 + \cdots + n_k$, und es gelte*

$$y_{ij} = ax_j + b + e_{ij}$$

Dann folgt

$$\overline{y} = \overline{\hat{y}}, \tag{36}$$

d.h. der Mittelwert der $\hat{y}_j$ ist gleich dem Gesamtmittelwert $\overline{y}$. Weiter gilt

$$QS_{Abw.lin.Reg} = \sum_{j=1}^{k} n_i(\overline{y}_j - \hat{y}_j)^2, \quad QS_{lin.Reg} = \sum_{j=1}^{k} n_j(\hat{y}_j - \overline{y})^2 \tag{37}$$

die Quadratsummen, die die Abweichungen der Mittelwerte $\overline{y}_j$ (d.h. der Mittelwerte von Y für x_j) von den durch die lineare Regression vorhergesagten Werten $\hat{y}_j$ bzw. der Abweichungen der $\hat{y}_j$-Werte von ihrem Gesamtmittelwert repräsentieren. Es gilt

$$QS_{zw} = QS_{Abw.lin.Reg} + QS_{lin.Reg} \tag{38}$$

Schließlich sei $s^2_{\hat{y}} = \sum_j (\hat{y}_j - \overline{y})^2/n$ und $s^2_e = \sum_j \sum_i (y_{ij} - \hat{y}_j)^2 = \sum_j \sum_i e^2_{ij}/n$. Dann gilt

$$s^2_y = s^2_{\hat{y}} + s^2_e, \tag{39}$$

Beweis: Im Folgenden wird zur Abkürzung $\sum_{j,i}$ statt $\sum_{j=1}^{k} \sum_{i=1}^{n_j}$ geschrieben.

Zu (36): Es ist $y_{ij} = \hat{y}_j + e_{ij}$. Also ist $\sum_{j,i} y_{ij} = \sum_j n_j \hat{y}_j + \sum_{j,i} e_{ij}$. Aber nach (13) ist $\sum_{j,i} e_{ij} = 0$, und nach Division durch n folgt (36).

Zu (38): Es ist sicherlich

$$QS_{zw} = \sum_j n_j(\overline{y}_j - \overline{y})^2 = \sum_j n_j(\overline{y}_j - \hat{y}_j + \hat{y}_j - \overline{y})^2,$$

denn in der Summe auf der rechten Seite wird ja $\hat{y}_j$ ja nur subtrahiert, um gleich wieder addiert zu werden. Ausquadriert ergibt sich

$$QS_{zw} = \sum_j n_j(\overline{y}_j - \hat{y}_j)^2 + \sum_j n_j(\hat{y}_j - \overline{y})^2 + 2\sum_j n_j(\overline{y}_j - \hat{y}_j)(\hat{y}_j - \overline{y})$$

Die dritte Summe auf der rechten Seite entspricht wieder der Kovarianz zwischen den $(\overline{y}_j - \hat{y}_j)$ und den $(\hat{y}_j - \overline{y})$.

Nun ist einerseits $y_{ij} - \hat{y}_j = e_{ij}$ und folglich $\overline{y}_j - \hat{y}_j = e_{i.} = \overline{e}_j$, andererseits ist $\hat{y}_j - \overline{y} = \hat{a}(x_j - \overline{x})$. Dementsprechend ist

$$\begin{aligned}\sum_j n_j(\overline{y}_j - \hat{y}_j)(\hat{y}_j - \overline{y}) &= \hat{a}\sum_j n_j\overline{e}_j(x_j - \overline{x}) \\ &= \hat{a}\left(\sum_j n_j\overline{e}_j x_j - \overline{x}\sum_j n_j\overline{e}_j\right)\end{aligned}$$

Aber aus (12) und (13) ist bekannt, dass $\sum_{j,i} e_{ij}x_j = 0$ und $\sum_{j,i} e_{ij} = 0$. Es ist aber

$$\sum_{j,i} e_{ij}x_j = \sum_j x_j \sum_i e_{ij} = \sum_j x_j n_j \overline{e}_j,$$

denn $\sum_j e_{ij} = n_j\overline{e}_j$. Dann ist aber $\sum_j n_j\overline{e}_j x_j = 0$ und $\sum_j n_j\overline{e}_j = 0$, so dass das Kreuzprodukt in der ersten Gleichung verschwindet, damit ist die Behauptung (38) nachgewiesen.

Es bleibt, (39) zu zeigen. Aus $y_{ij} = \hat{y}_j + e_{ij}$ folgt $y_{ij} - \overline{y} = \hat{y}_j - \overline{y} + e_{ij}$, so dass

$$QS_{ges} = \sum_{j,i}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{k} n_j(\hat{y}_j - \overline{y})^2 + \sum_{j,i} e_{ij}^2 + 2\sum_{j,i}(\hat{y}_j - \overline{y})e_{ij}$$

Hier ist $\sum_{j,i}(\hat{y}_j - \overline{y})e_{ij}$ die Kovarianz zwischen den $(\hat{y}_j - \overline{y})$ und den e_{ij}. Es ist

$$\sum_{j,i}(\hat{y}_j - \overline{y})e_{ij} = \sum_{j,i}\hat{y}_j e_{ij} - \overline{y}\sum_{j,i} e_{ij}$$

Aber $\sum_{j,i}\hat{y}_j e_{ij} = \hat{a}\sum_{j,i} x_j e_{ij} = 0$, nach (12), und $\sum_{j,i} e_{ij} = 0$ nach (12), und damit ist die Aussage bewiesen. $\square$.

Anmerkungen: (33) und (38) entsprechend läßt sich also die Quadratsumme QS_{ges} und damit die Gesamtvarianz s_y^2 gemäß

$$QS_{ges} = QS_{inn} + QS_{Abw.lin.Reg} + QS_{lin.Reg} \tag{40}$$

in additive Komponenten zerlegen. Es ist

$$QS_{inn} = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2\right)$$

Die Summe $\sum_i(y_{ij} - \overline{y}_j)^2$ entspricht der Quadratsumme, die man berechnen würde, wollte man die Varianz in der j-ten Gruppe berechnen. Die weitere Summation über j, d.h. über die Gruppen, entspricht einer Mittelung dieser Varianzschätzungen. Es ist demnach plausibel, dass QS_{inn} eine Schätzung für s_y^2 darstellt, falls sich die Mittelwerte $\overline{y}_j$ *nur zufällig* unterscheiden. Dies wird u.a. dann der Fall sein, wenn es keinen An- bzw. Abstieg der y_{ij}-Werte mit steigenden oder fallenden x_j-Werten gibt, wenn also der Regressionsparameter a gleich Null ist, d.h. wenn $\hat{a}$ nur zufällig von Null verschieden ist. In diesem Fall ist $\hat{y}_j = \hat{a}x_j + \hat{b} \approx \hat{b}$ und die Abweichungen $(\overline{y}_j - \hat{y}_j)$ und $(\hat{y}_j - \overline{y})$, die die Werte von $QS_{Abw.lin.Reg.}$ und $QS_{lin.Reg.}$ definieren, sollten nicht systematisch von Null abweichen. Demnach tragen dann die Quadratsummen $QS_{Abw.lin.Reg.}$ und $QS_{lin.Reg.}$ nur durch zufällige Werte zum Gesamtwert QS_{ges} bei. Dieser Fall kann auch dann eintreten, wenn die Beziehung zwischen Y und X nichtlinear ist. QS_{inn} *kann* dann trotz systematisch

variierender $\overline{y}_j$-Werte eine gute Schätzung für QS_{ges} sein; im Kapitel über nichtlineare Regression wird hierauf näher eingegangen.

Die Zerlegung (38) reflektiert insbesondere den Einfluß der linearen Regression, ohne dass eine Abschätzung der Varianz s_e^2 auftritt. Die Variabilität der e_{ij} ist in den Komponenten QS_{inn}, $QS_{Abw.lin.Reg.}$ und $QS_{lin.Reg.}$ enthalten. (39) ist eine Zerlegung von s_y^2 in die Komponente s_e^2 und in eine Komponente $s^2(\hat{y})$, die die Unterschiede zwischen den $\hat{y}_j$ reflektiert. Ist $a = 0$, so wird $s^2(\hat{y}) \approx 0$ gelten und mithin $QS_{ges} \approx s_e^2$. Die Beziehung (39) ist insbesondere beim Modell 2 von Interesse.

Abschließend sei darauf hingewiesen, dass die *Additivität* der Zerlegungen von QS_{ges} einschließlich der von QS_{inn} eine Implikation der Methode der Kleinsten Quadrate ist; dieser Sachverhalt wird beim Beweis des Satzes deutlich werden, weil dann auf die Gleichungen (12) und (13) zurückgegriffen wird. Die Additivität bedeutet, dass die Kovarianzen der Abweichungen, die die Quadratsummen QS_{inn}, $QS_{Abw.lin.Reg.}$ und $QS_{lin.Reg.}$ definieren, sämtlich gleich Null sind. Da diese Quadratsummen Varianzkomponenten repräsentieren, heißt dies, dass sich s_y^2 aus *voneinander unabhängig variierenden Komponenten* zusammensetzt.

Insbesondere impliziert (39), dass die Kovarianz $\mathrm{Kov}(\hat{Y}, e)$ gleich Null ist, d.h. die Fehler e_{ij} variieren unabhängig von den $\hat{y}_j$. Für Anwendungen der Theorie der linearen Regression, etwa im Rahmen der Klassischen Theorie psychometrischer Tests, ist dies von Interesse: die Unabhängigkeit der Meßfehler von den geschätzten "wahren" Werten ist keine psychologische Annahme, sondern eine Implikation der Methode.

4.3 Signifikanztests und Konfidenzintervalle

Will man die Nullhypothese, dass es keinen linearen Zusammenhang zwischen der Prädiktorvariablen X und der Kriteriumsvariablen, d.h. der abhängigen Variablen gibt, so kann man auf die Varianzzerlegung (40), d.h. auf

$$QS_{ges} = QS_{inn} + QS_{Abw.lin.Reg} + QS_{lin.Reg}$$

zurückgreifen. Wie in der Varianzanalyse gezeigt wurde, kann man unter der Nullhypothese die Quadratsummen als Schätzungen für die Fehlervarianz σ^2 auffassen, wenn man voraussetzen kann, dass die Fehler normalverteilt, d.h. $N(0, \sigma^2)$, sind. Bekanntlich gilt

$$\frac{QS_{ges}}{\sigma^2} = \chi^2_{n-1}, \quad \frac{QS_{inn}}{\sigma^2} = \chi^2_{n-k}, \quad \frac{QS_{zw}}{\sigma^2} = \chi_{k-1}.$$

Wegen

$$QS_{zw} = QS_{Abw.lin.Reg} + QS_{lin.Reg}$$

ergibt sich die Möglichkeit, auch die Quadratsummen $QS_{Abw.lin.Reg}$ und $QS_{lin.Reg}$ auf ihre Verträglichkeit bezüglich der Nullhypothese zu überprüfen. Mit aus der Varianzanalyse bekannten Techniken zeigt man, dass insbesondere

$$\frac{QS_{lin.Reg}}{\sigma^2} = \chi^2_1 \tag{41}$$

gilt; daraus folgt sofort, dass

$$\frac{QS_{Abw.lin.Reg}}{\sigma^2} = \chi^2_{k-2} \tag{42}$$

gelten muß. Damit ergibt sich der Test der Hypothese $H_0 : a = 0$ als F-Test

$$F = \frac{QS_{inn}/(n-k)}{QS_{Abw.lin.Reg/(k-2)}}, \quad df = n-k, k-2 \tag{43}$$

5 Der Produkt-Moment-Korrelationskoeffizient

5.1 Herleitung

Zu maßeinheitsfreien Skalen gelangt man durch Standardisierung der Variablen X und Y, d.h.

$$z_i(x) = \frac{x_i - \bar{x}}{s_x}, \quad z_i(y) = \frac{y_i - \bar{y}}{s_y}$$

Die lineare Beziehung zwischen den $z_i(x)$ und den $z_i(y)$ ist durch den in (17) bereits eingeführten Korrelationskoeffizienten, genauer gesagt: durch den Pearson-Bravaisschen Produkt-Moment-Korrelationskoeffizienten gegeben. Der Ausdruck wird zunächst hergeleitet. Nach Definition des Korrelationskoeffizienten in (17) gilt

$$r_{xy} = \frac{\mathrm{Kov}(x,y)}{s_x s_y}, \tag{44}$$

während für den Regressionskoeffizienten

$$\hat{a} = \frac{\mathrm{Kov}(x,y)}{s_x^2} \tag{45}$$

gilt. Offenbar resultiert der Ausdruck für r_{xy}, wenn man Zähler und Nenner des Ausdrucks für $\hat{a}$ mit dem Quotienten s_x/s_y multipliziert:

$$r_{xy} = \hat{a}\frac{s_x}{s_y} = \frac{\mathrm{Kov}(x,y)}{s_x^2}\frac{s_x}{s_y} = \frac{\mathrm{Kov}(x,y)}{s_x s_y}. \tag{46}$$

Von der Beziehung $r_{xy} = \hat{a}_{xy} s_x/s_y$ bzw. $\hat{a}_{xy} = r_{xy} s_y/s_x$ wird öfter Gebrauch gemacht.

Andererseits kann man von der Regression zwischen standardisierten Variablen ausgehen. Allgemein gilt für standardisierte Variable

$$\hat{a} = \frac{\frac{1}{n-1}\sum_i z_{xi} z_{yi} - \bar{z}_x \bar{z}_y}{s_{z_x}^2}.$$

Aber für standardisierte Variable gilt $\bar{z}_x = \bar{z}_y = 0$ und $s_{z_x}^2 = 1$, so dass sich

$$\hat{a} = \frac{1}{n-1}\sum_{i=1}^{n} z_{xi} z_{yi} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y} = r_{xy}$$

ergibt.

Der Korrelationskoeffizient ist symmetrisch:

$$r_{yx} = r_{xy} = r \tag{47}$$

gilt im Unterschied zu den Regressionsparametern a_{yx} und a_{xy}, für die im allgemeinen $a_{yx} \neq a_{xy}$ gilt. Für standardisierte Variable ist also die Richtung der Regression irrelevant.

Anmerkung: Ein etwas ausführlicherer Name für r ist *Pearson-Bravaisscher Produkt-Moment-Korrelationskoeffizient*, nach den Statistikern Pearson und Bravais; r heißt Produkt-Moment-Korrelationskoeffizient, weil $\mathrm{Kov}(x,y)$ ein arithmetisches Mittel (erstes zentrales Moment) der Produkte $(x_i - \bar{x})(y_i - \bar{y})$ ist. r bzw. $\hat{r}$ hat offenbar die Dimensionen [X][Y]/[X][Y], d.h. r und damit $\hat{r}$ ist frei von Dimensionen.

Der Korrelationskoeffizient hat die folgenden Eigenschaften:

Satz 5.1 *Es gilt*

$$-1 \leq \quad r \quad \leq 1, \tag{48}$$

$$r \quad = \quad \sqrt{\hat{a}_{yx}\hat{a}_{xy}}. \tag{49}$$

Beweis: Man sieht leicht, dass für $x_i = \pm y_i$ für alle i die Gleichung $r_{xy} = \pm 1$ resultiert; man muß aber noch zeigen, dass für $x_i \neq y_i$ zumindest für einige i nicht $r_{xy} > 1$ bzs. $r_{xy} < -1$ resultieren kann. Deswegen macht man zum Beweis von (48) Gebrauch von der Schwarzschen Ungleichung (auch Cauchy'sche oder Cauchy-Bunjakowskische Ungleichung genannt), derzufolge für irgendwelche reellen Zahlen a_i, b_i, $i = 1, \cdots, n$

$$\left| \sum_{i=1}^{n} a_i b_i \right|^2 \leq \sum_{i=1}^{n} |a_i|^2 \sum_{i=1}^{n} |b_i|^2 \tag{50}$$

gilt; das Gleichheitszeichen gilt genau dann, wenn $b_i = \alpha a_i$ für alle i, $\alpha \neq 0$. Die Herleitung von (50) wird im letzten Abschnitt dieses Kapitels gegeben.

Es ist $r = \mathrm{Kov}(x, y)/s_x s_y$ und $\mathrm{Kov}(x, y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$; nach (50) ist mit $a_i = x_i - \bar{x}$, $b_i = y_i - \bar{y}$

$$\left| \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right|^2 \leq \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|^2 \, \frac{1}{n} \sum_{i=1}^{n} |y_i - \bar{y}|^2 = s_x^2 s_y^2$$

und das Gleichheitszeichen gilt für $y_i - \overline{y} = a(x_i - \overline{x})$ für alle i, $|a| \neq 0$. Mithin folgt

$$\frac{|\sum_i (x_i - \bar{x})(y_i - \bar{y})|^2/n}{s_x^2 s_y^2} \leq 1,$$

womit (48) bewiesen ist. Für $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$ folgt $r < 0$.

Die Gleichung (49) folgt sofort aus der Tatsache, dass einerseits $r = \hat{a}_{yx} s_x / s_y$ und andererseits $r = \hat{a}_{xy} s_y / s_x$; multipliziert man die linken und die rechten Seiten, so erhält man $r^2 = \hat{a}_{yx}\hat{a}_{xy}$. Zieht man die Wurzel, so erhält man die Gleichung (49).

□

Die Bedingung $y_i - \overline{y} = a(x_i - \overline{x})$ für alle i mit $|a| \neq 0$ bedeutet, dass $e_i = 0$ für alle i, dass der Zusammenhang zwischen X und Y fehlerfrei ist. Der numerische Wert von a ist dagegen für den Wert von r nicht von Bedeutung, es sei denn, es gilt $a = 0$; in diesem Fall ist $\mathrm{Kov}(x, y) = 0$ und es folgt $r = 0$. Für $a \neq 0$ wird der Wert von r also durch die e_i bestimmt; dieser Sachverhalt legt nahe, über den Korrelationskoeffizienten ein Maß für die Güte der Vorhersage der Y-Werte anhand der X-Werte herzuleiten, wobei sich der Begriff der "Güte" auf das Ausmaß, in dem Fehler in die Vorhersage eingehen, bezieht. Der Effekt der Fehler läßt sich nun wie folgt fassen. Es ist ja

$$\hat{y}_i = \hat{a}_{yx} x_i + \hat{b}.$$

$\hat{a}_{yx}$ definiert gewissermaßen die Kopplung zwischen den X- und Y-Werten. In dem Maße, in dem X variiert, wird auch $\hat{y}$ variieren, insbesondere gilt

$$s_{\hat{y}}^2 = \hat{a}_{yx}^2 s_x^2 \tag{51}$$

Der folgende Begriff liefert ein Maß für die Güte der Vorhersage der Y- aufgrund der X-Werte:

Definition 5.1 *Für die Variablen X und Y werde die lineare Beziehung $\hat{y}_i = \hat{a}_{yx}x_i + \hat{b}_{yx} + e_i$ angenommen. Dann heißt der Quotient*

$$D := \frac{s^2_{\hat{y}}}{s^2_y} \tag{52}$$

Determinationskoeffizient.

Es gilt

Satz 5.2 *Für die Variablen X und Y werde die Beziehung $Y = a_{yx}X + b_{yx} + e$ angenommen. Die Vorhersagen der Y-Werte aufgrund der X-Werte seien durch $\hat{y}_i = \hat{a}_{yx}x_i + \hat{b}_{yx}$ gegeben, wobei die Parameter $\hat{a}_{yx}$ und $\hat{b}_{yx}$ die Kleinste-Quadrate-Schätzungen für a_{yx} und b_{yx} seien. Weiter sei $s^2_e = \sum_i e_i^2/n$. Dann gelten die Aussagen*

$$\begin{aligned}
\overline{y} &= \overline{\hat{y}} && (53)\\
QS_{ges} &= QS_{Abw.lin.Reg} + QS_{lin.Reg}, && (54)\\
s^2_y &= s^2_{\hat{y}} + s^2_e, && (55)\\
D &= \hat{a}_{yx}\frac{s^2_y}{s^2_x} = \hat{r}^2 && (56)\\
D &= 1 - \frac{s^2_e}{s^2_y} = \hat{r}^2, && (57)\\
0 \leq\ & D \leq 1 && (58)
\end{aligned}$$

Beweis: Die Aussagen (53) und (55) entsprechen den Aussagen (36) und (39) des Satzes 4.1 und sind hier nur der Vollständigkeit halber noch einmal aufgeführt worden. Da die Bedingung $n_i > 1$ beim Beweis dieser Ausagen des Satzes 4.1 keine notwendige Voraussetzung war, genügt es, darauf hinzuweisen, dass für den Fall $n_i = 1$ für alle i die Ausssagen (53) und (55) ebenfalls resultieren.

Die Aussage (54) folgt aus (40) für den Fall $n_i = 1$ für alle i. Denn $QS_{inn} = \sum_{ij}(y_{ij} - \overline{y}_i)^2$, und im Falle $n_i = 1$ für alle i ist $y_{ij} = y_i$ und $\overline{y}_i = y_i$, so dass $QS_{inn} = 0$.

Da einerseits $\hat{r}^2 = \hat{a}^2_{yx}s^2_x/s^2_y$ und andererseits $s^2_{\hat{y}} = \hat{a}^2_{yx}s^2_x$, folgt $\hat{r}^2 s^2_y = s^2_{\hat{y}}$ und somit (56).

Teilt man (55) durch s^2_y, so erhält man sofort $1 = s^2_{\hat{y}}/s^2_y + s^2_e/s^2_y$, und wegen (52) folgt (57).

(58) ist eine unmittelbare Konsequenz von (55): als Quotient von Quadraten kann D nicht kleiner als Null werden, und der Wert $D = 1$ kann nur für $s^2_e = 0$ angenommen werden. □

Anmerkungen:

1. Aus (40) und der Definition von $s^2_{\hat{y}}$ und s^2_e folgt, wenn $n_i = 1$ für alle i gilt, dass

   $$s^2_{\hat{y}} = QS_{lin.Reg.}/k, \quad s^2_e = QS_{Abw.lin.Reg.}/k \tag{59}$$

   Im Modell 2 sind also die Aussagen (54) und (55) äquivalent.

2. D ist der Anteil der "vorhergesagten" Varianz der Y-Werte an der Gesamtvarianz der Y-Werte. Denn wegen $\hat{y}_i = a\,x_i + b$ ist ja $s^2_{\hat{y}} = a^2 s^2_x$, d.h. die Varianz $s^2_{\hat{y}}$ geht

allein auf die Unterschiede zwischen den x_i, die ja durch s_x^2 repräsentiert werden, zurück. Nach (55) ist s_y^2 um den Betrag von s_e^2 größer als $s_{\hat{y}}^2$; s_e^2 "erklärt" also die nicht auf die Unterschiede zwischen den x_i-Werten zurückführbare Variation der Y-Werte.

3. Nach (57) gilt $\hat{r}^2 = 1 - s_e^2/s_y^2$. Diese Beziehung verdeutlicht noch einmal die Beziehung zwischen s_e^2 und $\hat{r}$ (bzw. r; der Unterschied zwischen r und $\hat{r}$ ist bei diesen Betrachtungen nicht wesentlich). Für $s_e^2 = s_y^2$ folgt $\hat{r}^2 = \hat{r} = 0$ und damit $a = 0$; für $s_e^2 = 0$ folgt $\hat{r}^2 = 1$ und damit $|\hat{r}| = 1$.

Man kann (57) nach s_e auflösen:$s_e^2 = s_y^2(1 - \hat{r}^2)$, so dass

$$s_e = s_y\sqrt{1 - \hat{r}^2}. \tag{60}$$

Der hier gegebene Ausdruck für die Streuung der Fehler hat einen speziellen Namen, der dementsprechend besonders eingeführt werden soll:

Definition 5.2 *Die in (60) angegebene Streuung s_e der "Fehler" heißt* Standardfehler der Schätzung *oder* Standardschätzfehler.

Natürlich kann man s_e^2 und damit s_e auch direkt aus (55) bestimmen. Der Reiz von (60) liegt darin, dass (i) s_e zum (Stichproben-) Regressionskoeffizienten $\hat{a}$ in Beziehung gesetzt und (ii) die Kenntnis von $s_{\hat{y}}^2$ nicht vorausgesetzt wird.

Beispiel 5.1 Zur Illustration sollen noch einmal die Daten aus Beispiel (2.1) herangezogen werden. Der Determinationskoeffizient für die Beziehung zwischen Alter und Merkfähigkeit ergibt ist

$$D = s_{\hat{y}}^2/s_y^2 = \hat{a}_{yx}^2 s_x^2/s_y^2 = (1.2634)^2 \times 134.75/595.15 = .361.$$

Dies ist der Anteil der *Varianz der vorhergesagten Y-Werte* an der *Varianz der Y-Werte*. Hieraus folgt sofort $s_{\hat{y}}^2 = D\, s_y^2 = .361 \times 595.15 = 215.085$, und $s_e^2 = s_y^2 - s_{\hat{y}}^2 = 595.15 - 215.085 = 380.065$; die Wurzel daraus ist der Standardschätzfehler, $s_e = 19.495$.

Inhaltlich wird der Wert von D oft durch die Aussage interpretiert, dass ein Anteil von .361 (oder ≈ 36 Prozent) der Unterschiede in der Merkfähigkeit auf Unterschiede im Alter zurückzuführen seien, also 64 Prozent der Variation der Merkfähigkeit auf andere Faktoren zurückgehen. Es wird dann weitergefolgert, dass das Nachlassen der Merkfähigkeit zu ca. 36 % auf Alterungsprozesse zurückführbar sei. Dieser Schluß ist sehr salopp und stellt eine unzulässige Vereinfachung dar, denn die Varianz ist ein spezielles Maß für Unterschiedlichkeit und der Anteil D bezieht sich eben auf dieses spezielle Maß. Zu welchem Anteil das Altern ein Nachlassen der Merkfähigkeit nach sich zieht, kann nur durch zusätzliche Betrachtungen bzw. Datenerhebungen bestimmt werden. □

Zusammenfassend läßt sich sagen:

1. Während $\hat{a}_{yx}$ die durchschnittliche Veränderung von Y mit X repräsentiert (entsprechendes gilt für $\hat{a}_{xy}$), reflektiert der Regressionsparameter r der standardisierten Variablen das Ausmaß des statistischen Zusammenhanges zwischen X und Y.

2. Man kann nicht auf die Unabhängigkeit schließen, wenn man $r \approx 0$ findet. Denn die Variablen können in nachgerade deterministischer Abhängigkeit zueinander stehen und dennoch eine Korrelation von Null aufweisen:

Beispiel 5.2 (Feller, 1968, *p.* 236) Die Variable X nehme die Werte -2, -1, 1, 2 an, und es sei $Y = X^2$; dann ist Y durch X eindeutig bestimmt. Die möglichen Meßwertpaare (-2, 4), (-1, 1), (1, 1) und (2, 4) seien gleichhäufig aufgetreten. Dann ist $\bar{x} = 0$, $\bar{y} = 2.5$, $\text{Kov}(x, y) = (\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y})/n$, d.h. $\text{Kov}(x, y) = -8 - 1 + 1 + 4 - 0 \cdot 2.5 = 0$, also ist auch $r = 0$. Man sieht, dass der genaue Wert von n keine Rolle spielt, so lange nur die Paare alle gleich häufig auftreten. □

Das Beispiel mag ein wenig gekünstelt erscheinen, aber es weist auf die Möglichkeit hin, dass sogar bei einer deterministischen funktionalen Abhängigkeit Werte von $r \approx 0$ auftreten können. Dieser Fall kann u.a. dann eintreten, wenn die Beziehung zwischen X und Y nichtlinear ist.

Nach (49) ist der Korrelationskoeffizient r gerade gleich dem geometrischen Mittel der Regressionskoeffizienten $\hat{a}_{yx}$ und $\hat{a}_{xy}$; die Bildung des geometrischen Mittels ist tatsächlich immer möglich, da für $\hat{a}_{yx} < 0$ auch $\hat{a}_{xy} < 0$ und für $\hat{a}_{yx} > 0$ auch $\hat{a}_{xy} > 0$, so dass das Produkt der beiden Regressionskoeffizienten stets positiv ist. Ist einer der Regressionskoeffizienten gleich Null, so ist auch $r = 0$. Aus (49) ergibt sich eine Beziehung zwischen den Regressionskoeffizienten $\hat{a}_{yx}$ und $\hat{a}_{xy}$, denn es folgt

$$\hat{a}_{yx} = r^2/\hat{a}_{xy}, \quad \hat{a}_{xy} = r^2/\hat{a}_{yx}. \tag{61}$$

Die Regressionskoeffizienten sind also reziprok zueinander mit dem Proportionalitätsfaktor r^2. Nur für $r = 1$ folgt die direkte Reziprozität $\hat{a}_{yx} = 1/\hat{a}_{xy}$, wie man sie für die fehlerfreien Beziehungen $y_i = \hat{a}_{yx} x_i + \hat{b}_{yx}$, $x_i = \hat{a}_{xy} y_i + \hat{b}_{xy}$ kennt.

5.2 Korrelation und zweidimensionale Normalverteilung

Die zufälligen Veränderlichen X und Y seien gemeinsam normalverteilt, d.h. die gemeinsame Dichte sei durch

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\phi(x, y)\right] \tag{62}$$

mit

$$\phi(x, y) = \frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right] \tag{63}$$

Hierin sind $\mu_x = E(X)$ und $\mu_y = E(Y)$ die Erwartungswerte von X bzw. Y, σ_x^2 und σ_y^2 sind die entsprechenden Varianzen und ρ ist die Korrelation zwischen X und Y. Für

$$\phi(x, y) = C, \tag{64}$$

C eine Konstante, definiert $\phi(x, y)$ eine Ellipse, d.h. die Punkte (x, y), die der Bedingung (64) genügen, liegen auf einer Ellipse. In Bezug auf das Regressionsproblem ist die Bedingte Verteilung von Y, gegeben ein x-Wert, von Bedeutung. Allgemein sind bedingte Dichten durch

$$f_{Y|x}(y|x) = \frac{f(x, y)}{f_X(x)} \tag{65}$$

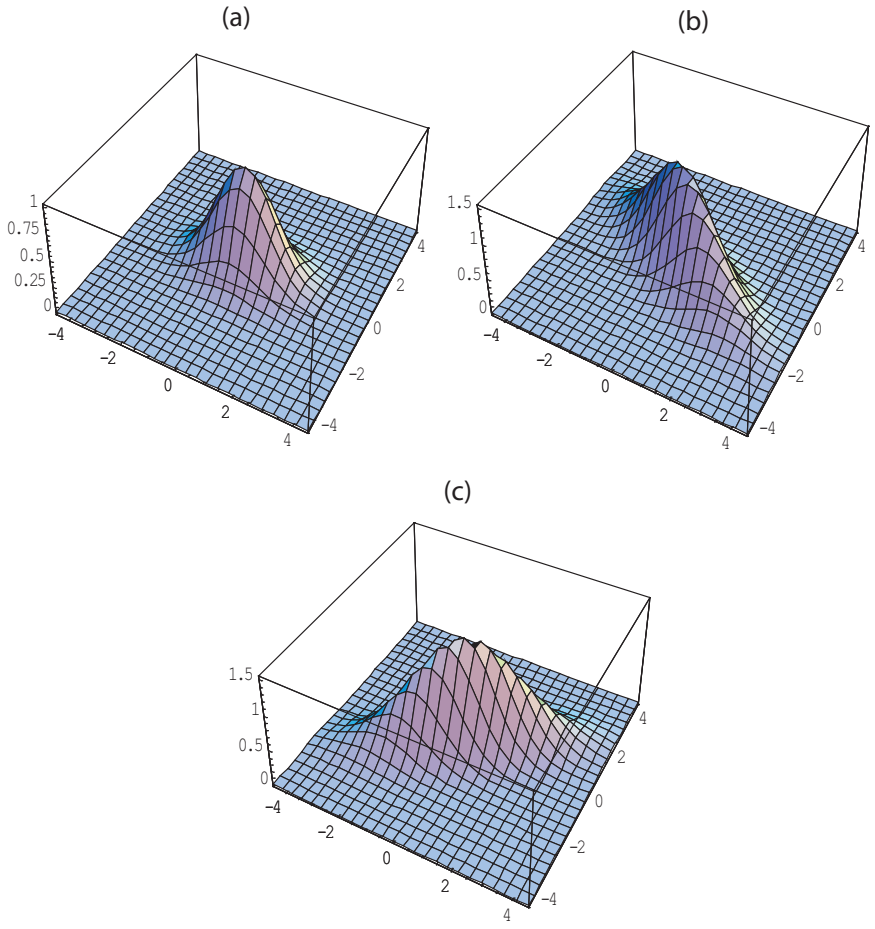
gegeben, wobei

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Für die Normalverteilung ergibt sich

$$f_Y(y|x) = \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2\sigma_x^2(1-\rho^2)} \left(y - \mu_y - \frac{\rho\sigma_y^2}{\sigma_x^2}(x-\mu_x)^2 \right) \right] \tag{66}$$

Abbildung 3: Zweidimensionale Normaldichten; (a) $r = 0$, (b) $r = .75$, (c) $r = -.75$



Der zugehörige bedingte Erwartungswert von Y ist

$$E(Y|x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x), \quad \sigma_{y|x} = \sigma_x\sqrt{1-\rho^2}. \tag{67}$$

Aus $\hat{y}_i = \hat{b}_{yx} + \hat{a}_{yx}x_i$, $\bar{\hat{y}} = \hat{b}_{yx} + \hat{a}_{yx}\bar{x}$ und wegen (53), also $\bar{\hat{y}} = \bar{y}$, folgt

$$\hat{y}_i - \bar{y} = \hat{a}_{yx}(x_i - \bar{x}),$$

d.h.

$$\hat{y}_i = \bar{y} + \hat{a}_{yx}(x_i - \bar{x}) \tag{68}$$

Diese Beziehung korrespondiert zu (67); die Regressionsgerade entspricht der Geraden, die die bedingten Erwartungswerte für Y unter der Bedingung eines x-Wertes angibt.

5.3 Signifikanztests

In Abschnitt 4.3 wurde ein F-Test zur Überprüfung der Nullhypothese $a = 0$ hergeleitet, wobei die Quadratsumme QS_{inn} als Schätzung des Fehlers benützt wurde. Da es aber im

Modell 2 nur einen Wert für jede Gruppe (d.h. nur einen y_i-Wert für einen x_i-Wert) gibt, ist $QS_{inn} = 0$, kann also nicht für einen F-Test verwendet werden. Andererseits konnte die Quadratsumme QS_{zw} in einen Anteil $QS_{Abw.lin.Reg.}$ und in einen Anteil $QS_{lin.Reg.}$ zerlegt werden, wobei $QS_{Abw.lin.Reg.}/\sigma^2 = \chi_{k-2}$ und $QS_{lin.Reg.}/\sigma^2 = \chi_1^2$ ist. Also kann man den F-Test

$$F = \frac{QS_{lin.Reg.}}{QS_{Abw.lin.Reg.}}, \quad df = 1, k-2 \tag{69}$$

bilden, der bei Gültigkeit der Nullhypothese $\rho_{xy} = 0$ wie F mit $k-2$ und 1 Freiheitsgraden verteilt ist; ρ ist der Populationskorrelationskoeffizient. Darüber hinaus ist bekannt, dass F mit $n-2$ und 1 Freiheitsgraden wie t^2 mit $n-2$ Freiheitsgraden verteilt ist; man kann also den Test

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}, \quad df = n-2 \tag{70}$$

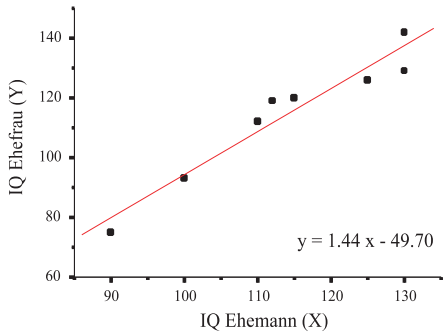
bilden (er ergibt sich aus (69) durch Umformung.

Beispiel 5.3 8 Ehepaare sind gebeten worden, ihre Intelligenzquotienten messen zu lassen. Es ergaben sich die folgenden Werte: Man errechnet $\bar{x} = 113.38$, $\bar{y} = 114.5$,

Tabelle 3: Intelligenzquotienten von Ehepartnern

	Intelligenzquotienten							
IQ-Ehemann	90	110	100	115	117	120	125	130
IQ-Ehefrau	75	112	93	120	119	129	126	142

Abbildung 4: Ehe und Intelligenz



$s_x^2 = \sum_i (x_i - \bar{x})^2/n = 151.044$, ($s_x = 12.29$) bzw. $\hat{s}_x^2 = \sum_i (x_i - \bar{x})^2/(n-1) = 172.66$ bzw. ($\hat{s}_x = 13.14$), $s_y^2 = 397.205$ oder $s_y = 19.93$, bzw. $\hat{s}_y^2 = 451.116$, ($\hat{s}_y = 21.31$), $\mathrm{Kov}(x, y) = (\sum_i (x_i - \bar{x})(y_i - \bar{y}))/n = 241.44$, bzw. $(\sum_i (x_i y_i - (\sum_{i=1}^n x_i \sum_{i=1}^n y_i)/n)/(n-1) = 275.93$. Dann ergibt sich ein Korrelationskoeffizient $r_{xy} = .9858$ bzw. $\hat{r}_{xy} = .9857$ (man sieht, dass hier die Korrektur des Effekts der Unterschätzung von Stichprobenvarianzen und -kovarianzen kaum von Bedeutung ist). Wir schreiben $r_{xy} \approx .99$. Dann ist der Determinationskoeffizient $r_{xy}^2 = .972$, d.h. $\approx 97\%$ der Varianz der Y-Werte (IQ Ehefrau) wird durch die Varianz der X-Werte (IQ Ehemann) erklärt. Aus r_{xy}, s_x und s_y lassen sich die Regressionskoeffizienten $\hat{a}_{yx}$ und $\hat{a}_{xy}$, sowie die entsprechenden Standardschätzfehler

berechnen: aus (26) folgt $\hat{a}_{yx} = .99(19.93/12.29) = 1.61$, $\hat{a}_{xy} = .99(12.29/19.93) = .61$. Wegen (48) ist gleichzeitig $\hat{a}_{yx} = r^2/\hat{a}_{xy}$; in der Tat findet man $1.61 \approx .98/.61$. Sollen die IQen der Ehefrauen aufgrund der IQen der Ehemänner vorhergesagt werden, so hat man einen Standardschätzfehler $s_y(e) = s_x\sqrt{1-r^2} = 19.93 \times .141 = 2.81$; sagt man umgekehrt die Intelligenz der Männer aufgrund des IQen ihrer Frauen voraus, so hat man einen Standardschätzfehler $s_x(e) = s_x\sqrt{1-r^2} = 12.29 \times .141 = 1.73$. Die Intelligenz der Männer läßt sich genauer auf der Basis der Intelligenz ihrer Ehefrauen voraussagen als die Intelligenz der Frauen auf der Basis der Intelligenz ihrer Männer. Man überlege sich die psychologische Bedeutung des Befundes $a_{yx} > 1$ und $a_{xy} < 1$! $\square$

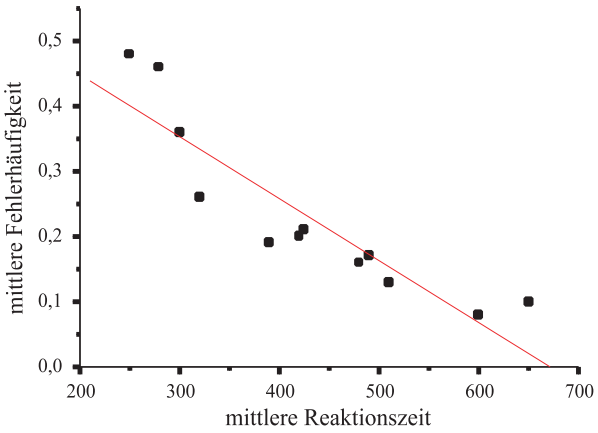
Beispiel 5.4 (Speed-Accuracy-Trade-Off) Im Rahmen einer Untersuchung über die Leistungsfähigkeit von Operateuren, die die Arbeit eines Kernreaktors kontrollieren, soll die Beziehung zwischen Präzision und Reaktionszeit bestimmt werden. Dazu müssen die

Tabelle 4: Reaktionszeiten und Fehlerhäufigkeiten

	Reaktionszeiten											
O	1	2	3	4	5	6	7	8	9	10	11	12
$\bar{t}$	250	279	300	320	390	420	425	480	490	510	600	650
$\hat{p}$	.48	.46	.36	.26	.19	.20	.21	.16	.17	.13	.08	.10

Operateure auf ein zu einer zufälligen Zeit am Bildschirm erscheinendes Zeichen so schnell

Abbildung 5: Reaktionszeiten und Fehler (Speed-accuracy trade-off)



es geht einen dem Zeichen entsprechenden Schalter umlegen; dabei können sie einen Fehler machen, indem sie einen falschen Schalter wählen. Für 12 Operateure wurde ihre mittlere Reaktionszeit (in ms) sowie ihre relative Häufigkeit von Fehlern in der folgenden Tabelle zusammengefaßt: Der die Untersuchung durchführende Psychologe entscheidet nach einem Blick auf die Daten, dass vermutlich eine lineare Beziehung zwischen den mittleren Zeiten $\bar{t}_i$ der $i = 1, ..., 14$ Operateure und ihren relativen Fehlerhäufigkeiten besteht. Mit $x_i = \bar{t}_i$, $y_i = \bar{p}_i$ ergibt sich $\bar{x} = 426.17$, $\bar{y} = .23$, $s_x = 126.33$, $s_y = .13$, $\text{Kov}(x, y) = -15.18$. Mithin ist $r_{xy} = -.91$. Diesen Werten entsprechen die Regressionsgeraden $y_i = \hat{a}_{yx}x_i + \hat{b}_{yx} + e_i$, $x_i = \hat{a}_{xy}y_i + \hat{b}_{xy} + \tilde{e}_i$ mit $\hat{a}_{yx} = rs_y/s_x = -.000936$,

$\hat{b}_{yx} = \bar{y} - \hat{a}_{yx}\bar{x} = -.64$. Der Vorhersage der der $\bar{p}_i$ durch die $\bar{t}_i$ entspricht ein Standardschätzfehler $s_y(e) = s_y\sqrt{1-r^2} = .0539$. Der Determinationskoeffizient ist $r^2 = .83$, d.h. 83 % der Varianz der $\bar{p}$-Werte wird durch die Varianz der $\bar{t}$-Werte erklärt. Zwar scheint der Schluß, dass Operateure mit größeren mittleren Reaktionszeiten weniger Fehler machen gerechtfertigt zu sein, denn der Wert von r^2 ist einigermaßen hoch, doch läßt die Inspektion der Daten vermuten, dass dennoch keine lineare Beziehung vorliegt. □

6 Regressionsdiagnostik und Anscombe's Quartett

Der Verfügbarkeit von Programmpaketen zur Statistik verleitet dazu, die Daten einzugeben, sich die Schätzungen der Regressionskoeffizienten und die entsprechenden Resultate von Signifikanztest ausgeben zu lassen und dann direkt zur Interpretation überzugehen. Anscombe (1973)[1] hat zu Recht darauf hingewiesen, dass die Rolle der kritischen Betrachtung der Daten in Lehrbüchern, Programmen und wohl auch in Skripten zu wenig berücksichtigt wird. Deswegen soll an dieser Stelle auf die Notwendigkeit einer Inspektion der Daten hingewiesen werden. In Bezug auf Regressionsanalysen können die folgenden Anmerkungen gemacht werden (vergl. Abb. 6)

1. Die Punkte (x, y) liegen fast einer geraden Linie (der Regressionsgeraden),
2. Die Punkte (x, y) liegen fast auf einer Linie, die aber keine Gerade ist,
3. Die y-Werte streuen ohne Bezug auf die x-Werte,
4. Es liegt eine Mischung der drei vorangegangenen Punkte vor,
5. Die meisten Puntke (x, y) liegen auf einer Geraden oder einer anderen Linie, aber einige liegen weit von dieser Kurve entfernt ("Ausreißer").

Ausreißer können auf einfache Tippfehler zurückgehen, oder aber eine systematische Komponente haben, die näher untersucht werden muß. Sie entgehen der Aufmerksamkeit, wenn nur "mechanisch", d.h. ohne Dateninsepktion, gerechnet wird, und können natürlich die Schätzungen verfälschen.

Plot der Residuen Ist sei die Regressionsgerade

$$\hat{y}_i = a_{yx}x_i + b_{by}$$

angepasst worden; man kann

$$y_i = \hat{y}_i + e_i$$

schreiben. Die Differenzen

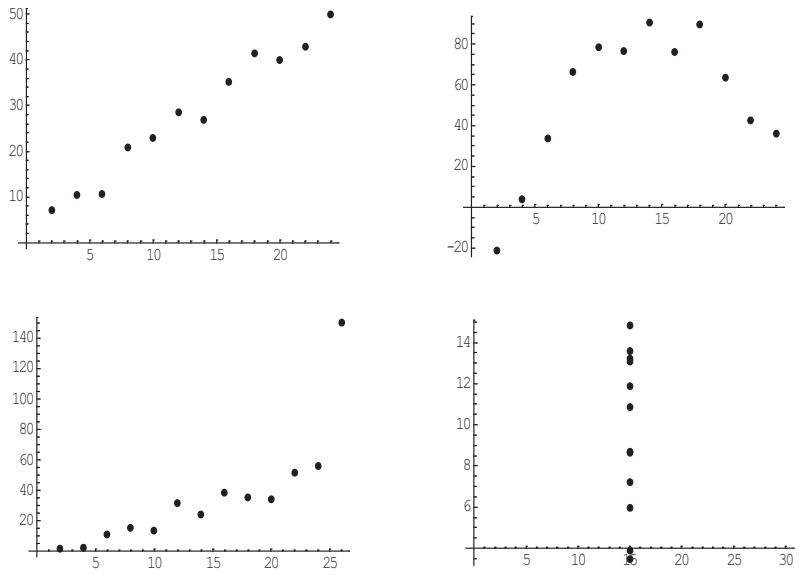
$$e_i = y_i - \hat{y}_i \tag{71}$$

sind dann die *Residuen*. Eine Forderung an die Daten war die nach der Homoskedaszidität: die "Fehler" sollten für alle x_i die gleichen Verteilung haben ($N(0, \sigma^2)$). Eine erste Überprüfung dieser Annahme ist dann ein Plot der e_i, wie sie in (71) definiert wurden, gegen die x_i-Werte. Alternativ dazu können die e_i gegen die $\hat{y}_i$ aufgetragen werden, wobei die e_i und die $\hat{y}_i$ durch die gleiche Skala repräsentiert werden. Idealerweise sollten dann die e_i als normalverteilt mit gleicher Varianz erscheinen. Bei diesem Plot sind vier Aspekte zu überprüfen:

1. Einige wenige Residuen sind viel größer als die übrigen. Dies ist ein Hinweis auf Ausreißer in den Daten.

[1] Anscombe, F. J. (1973) Graphs in Statistical Analysis. *The American Statistician*, 27, No. 1, 17-21

Abbildung 6: Anscombe's Quartet



2. Die Regression der Residuen auf die Werte $\hat{y}_i$ erscheint als nichtlinear, – das Modell $y_i + a_{yx}x_i + b_{yx} + e_i$ ist nicht adäquat.

3. Die e_i werden größer mit den $\hat{y}_i$, – es gibt eine systematische Abweichung von der Varianzhomogenität.

4. Die Residuen erscheinen als nicht normalverteilt, z. B. ist die Verteilung ist schief.

Im Falle eines Hinweises auf eine Nichtlinearität der Beziehung zwischen den x- und y-Werten kann man sich nähern, indem man ein Polynom anpasst:

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_k x_i^k + e_i, \tag{72}$$

wobei man zunächst nur das quadratische Glied berücksichtigt. Erst wenn dies nicht genügt, wird man den kubischen Term berücksichtigen, etc. Ein solches Vorgehen ist zunächst rein deskriptiv; das Nachdenken über die Ursachen der Nichtlinearität und ein geeignetes Modell für die Daten kann dadurch nicht ersetzt werden. Die Bestimmung der Regressionsparameter $a_0, a_1, a_2, \ldots$ kann mit den Techniken der multiplen Regression erfolgen.

7 Der Vierfelder-Korrelationskoeffizient

Bisher ist vorausgesetzt worden, dass die beiden Variablen X und Y auf dem Niveau von Intervallskalen gemessen werden können. In diesem Abschnitt soll der Fall betrachtet werden, bei dem sowohl X als auch Y dichotom sind, d.h. beide Variable nehmen entweder den Wert 0 oder den Wert 1 an. Man hat dann das in Tab. 5 gegebene Schema. Dabei ist $N = a + b + c + d$.

Man zeigt nun leicht, dass der folgende Satz gilt:

Tabelle 5: 4-Felder-Schema für ϕ

		Y		
		1	0	Σ
X	1	a	b	a + b
	0	c	d	c + d
	Σ	a + c	b + d	N

Satz 7.1 *Die Variablen X und Y seien beide dichotom, d.h. sie können nur die Werte 0 oder 1 annehmen. Der Produktmoment-Korrelationskoeffizient hat die Form*

$$\phi_{xy} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = r_{xy} \tag{73}$$

Anmerkung: Die Bezeichnung dieses Korrrelationskoeffizienten mit ϕ_{xy} dient lediglich als Hinweis, dass der Koeffizient für dichotome Variablen berechnet wurde.

Beweis: Es ist

$$r_{xy} = \frac{\sum_i x_i y_i/N - \overline{xy}}{s_x s_y}$$

Nun ist aber das Produkt $x_i y_i \neq 0$ nur dann, wenn $x_i = 1$ und $y_i = 1$; also ist

$$\frac{1}{N}\sum_i x_i y_i = \frac{a}{N}$$

Weiter ist

$$\overline{x} = \frac{a+b}{N}, \quad \overline{y} = \frac{a+c}{N}$$

Es ist ja

$$s_x^2 = \frac{1}{N}\sum_i x_i^2 - \overline{x}^2.$$

Nun ist aber $\sum_i x_i^2/N = (a+b)\,1^2/N = (a+b)/N$ und $\overline{x}^2 = (a+b)^2/N^2$, mithin

$$s_x^2 = \frac{a+b}{N} - \frac{(a+b)^2}{N^2} = \frac{a+b}{N}(1 - \frac{a+b}{N}) \tag{74}$$

Auf analoge Weise findet man

$$s_y^2 = \frac{a+c}{N} - \frac{(a+c)^2}{N^2} = \frac{a+c}{N}(1 - \frac{a+c}{N}) \tag{75}$$

Setzt man diese Größen in die Formel für r_{xy} ein, so erhält man

$$r_{xy} = \frac{Na - (a+b)(a+c)}{N^2\sqrt{((a+b)/N)(1-(a+b)/N))((a+c)/N)(1-(a+c)/N)}}$$

Berücksichtigt man, dass $N = a+b+c+d$, so vereinfacht sich dieser Ausdruck zu (73). □

In Kapitel 5 ist das χ^2-Maß für Kontigenztabellen eingeführt worden. Es ist ein Maß für den Zusammenhang zwischen den Zeilen- und Spaltenkategorien. Da der Korrelationskoeffizient ein Maß für die statistische Abhängigkeit zweier Variablen ist, liegt es nahe, nach der Beziehung zwischen dem ϕ-Koeffizienten und dem χ^2-Maß zu fragen.

Zur Erinnerung sei das χ^2-Maß noch einmal aufgeführt:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N} \tag{76}$$

Das 4-Felder-Schema ist ein Spezialfall einer Kontingenztabelle; es ist $n_{11} = a, n_{12} = b, n_{21} = c$ und $n_{22} = d$. Weiter ist $n_{1.} = a + b, n_{2.} = c + d$, etc. Wendet man also die Bezeichnungen aus Tab. 5 auf (76) an, so erhält man für den ersten Summanden

$$\frac{(a - (a+b)(a+c)/N)^2}{(a+b)(a+c)/N} = \frac{(Na - (a+b)(a+c))^2}{N(a+b)(a+c)}$$

Berücksichtigt man wieder, dass $N = a + b + c + d$, multipliziert die Klammern aus und faßt entsprechende Terme zusammen, so erhält man

$$\frac{(a - (a+b)(a+c)/N)^2}{(a+b)(a+c)/N} = \frac{(ad - bc)^2}{N(a+b)(a+c)}$$

Auf analoge Weise verfährt man mit den restlichen drei Summanden; das Resultat ist

$$\begin{aligned} \chi^2 &= \frac{(ad-bc)^2}{N(a+b)(a+c)} + \frac{(ad-bc)^2}{N(a+b)(b+d)} + \\ &+ \frac{(ad-bc)^2}{N(a+c)(c+d)} + \frac{(ad-bc)^2}{N(b+d)(c+d)} \end{aligned}$$

Bringt man nun noch die Summanden auf den gleichen Nenner, so erhält man

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{77}$$

Es gilt also, wie der Vergleich mit (73) sofort zeigt, der

Satz 7.2 *Zwischen dem Korrelationskoeffizienten ϕ_{xy} und dem χ^2-Maß besteht die Beziehung*

$$|\phi_{xy}| = \sqrt{\frac{\chi^2}{N}} \tag{78}$$

Da die Wahrscheinlichkeitsverteilung von χ^2 unter der Hypothese der Unabhängigkeit von X und Y bekannt ist, liefert (78) sofort eine Möglichkeit, zu testen, ob ϕ_{xy} nur zufällig von Null abweicht oder nicht.

Die Interpretation des ϕ-Koeffizienten als deskriptives Maß ist nicht immer einfach. Generell läßt sich sicherlich sagen, dass

$$-1 \leq \phi_{xy} \leq 1$$

denn $\phi = \phi_{xy}$ ist ja aus dem Korrelationskoeffizienten r_{xy} abgeleitet worden, d.h. er ist ein Spezialfall des Produkt-Moment-Korrelationskoeffizienten. Die Interpretation des ϕ-Koeffizienten ist allerdings an Randbedingungen gebunden. So werde angenommen, dass zwischen X und Y ein perfekter Zusammenhang besteht. Dieser Fall ist sicherlich dann gegeben, wenn $b = c = 0$. Andererseits folgt aus (73) für $\phi = 1$

$$(ad - bc)^2 = (a+b)(a+c)(b+d)(c+d)$$

Diese Gleichung ist offenbar genau dann erfüllt, wenn $b = c = 0$; dies sieht man, wenn man beide Seiten ausmultipliziert. Also ist der Fall $b = c = 0$ notwendig und hinreichend für $\phi = 1$. Für $a = d = 0$ erhält man analog $\phi = -1$.

Gilt $b = c = 0$, so folgt für die Randsummen $a + b = a + c = a$; d.h. aber $p_x = p_y$, wobei p_x der Anteil der X-Werte mit $X = 1$ ist, und p_y ist der Anteil der Y-Werte mit $Y = 1$. Dann ist natürlich auch $1 - p_x = 1 - p_y$.

Die Bedingung $p_x = p_y$ ist notwendig für $|\phi| = 1$, aber nicht hinreichend. Denn $p_x = p_y$ bedeutet ja nur, dass $a + b = a + c$ ist. Deswegen folgt nur $b = c$, aber nicht $b = c = 0$. Für $b = c \neq 0$ folgt $|\phi| < 1$; insbesondere ist auch $\phi = 0$ möglich. Damit folgt, dass ϕ nur dann alle Werte zwischen -1 und 1 annehmen kann, wenn $p_x = p_y$ gilt; gilt aber $p_x \neq p_y$, so folgt

$$\phi_{min} \leq \phi \leq \phi_{max}, \tag{79}$$

wobei

$$\phi_{max} = \sqrt{\frac{p_x(1-p_y)}{(1-p_x)p_y}}, \quad \phi_{min} = -\phi_{max}. \tag{80}$$

Die Herleitung von (80) findet man in Carroll (1961).

Die Ungleichung (79) bedeutet, dass der Wert des ϕ-Koeffizienten, falls er $\neq \pm 1$ ist, etwas schwierig zu interpretieren ist, falls $p_x \neq p_y$ gilt, denn die Begrenzung des Wertebereichs von ϕ reflektiert nicht einen mangelnden statistischen Zusammenhang, sondern nur den Sachverhalt, dass die Randhäufigkeiten nicht gleich sind. Eine Möglichkeit, den Wert von ϕ in diesem Fall mit anderen Korrelationskoeffizienten, die diesen Randbedingungen nicht unterliegen, vergleichbar zu machen, besteht darin, ihn durch ϕ_{max} zu teilen, also

$$\tilde{\phi} = \frac{\phi}{\phi_{max}} \tag{81}$$

zu betrachten. Für eine weitere Diskussion vergl. Carroll (1961).

Beispiel 7.1 In Beispiel **??** wurde das Wahlverhalten von Frauen und Männern betrachtet; es ergab sich die folgende Tabelle Man kann nun über ϕ die Korrelation zwischen

Tabelle 6: Wahlverhalten und Geschlecht

	Wahl		
Geschlecht	ja (X = 1)	nein (X = 0)	Σ
w (Y = 1)	40	5	45
m (Y = 0)	25	30	55
Σ	65	35	100

dem Geschlecht einerseits und der Entscheidung andererseits bestimmen:

$$\phi = \frac{40 \times 30 - 5 \times 25}{\sqrt{45 \times 65 \times 55 \times 35}} = .453$$

Diesem Wert von ϕ entspricht nach (78) der Wert

$$\chi^2 = N\,\phi^2 = 100 \times .205 = 25.52$$

Inspektion der Zeilen- und Spaltensummen von Tab. 6 zeigt aber, dass $p_x \neq p_y$. Nach (80) erhält man für diese Daten

$$\phi_{max} = \sqrt{\frac{.45 \times .35}{.65 \times .55}} = .664$$

Nach (81) ergibt sich dann der korrigierte Korrelationskoeffizient

$$\tilde{\phi} = \frac{.453}{.664} = .682$$

□

8 Die punkt-biseriale Korrelation

Gelegentlich ist man daran interessiert, den statistischen Zusammenhang zwischen zwei Variablen X und Y zu erfassen, wobei X auf einer Intervallskala gemessen wurde, Y aber nur die Werte 1 oder 0 annimmt. Diese Situation kann etwa dann eintreten, wenn man sich hinsichtlich des Meßniveaus einer Variablen nicht sicher ist und man deshalb ihren Wertebereich dichotomisiert, so dass $Y = 1$ für Werte größer als der Median und $Y = 0$ für Werte kleiner als der Median kodiert werden.

Wie beim ϕ-Koeffizienten kann nun die Formel für die Produkt-Moment-Korrelationen anwenden. Es gilt der

Satz 8.1 *Es sei X eine kontinuierlich variierende Größe und Y nehme entweder den Wert 1 oder 0 an. Für n_1 Meßwertpaare gelte $Y = 1$, und X habe für diese Paare den Mittelwert $\overline{x}_1$. Für $n_2 = n - n_1$ Paare gelte $Y = 0$, und der Mittelwert der X-Werte sei $\overline{x}_2$. Die Produkt-Moment-Korrelation zwischen X und Y hat dann die Form*

$$r = r_{pbis} = \frac{Kov(x,y)}{s_x s_y} = \frac{\overline{x}_1 - \overline{x}_2}{s_x}\sqrt{\frac{n_1 n_2}{n^2}} \tag{82}$$

Beweis: Nur für diejenigen Paare (x_i, y_i), für die $Y = 1$ ist, ist das Produkt $x_i y_i \neq 0$. Es ist $n = n_1 + n_2$ der Umfang der gesamten Stichprobe. Die Paare können stets so durchnummeriert werden, dass für $x_1, \cdots, x_{n_1}$ $Y = 1$ gilt. Der Mittelwert $\overline{x}$ setzt sich dann aus den Mittelwerten in den Teilstichproben wie folgt zusammen:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_2} x_i\right) \\ &= \frac{1}{n}(n_1\overline{x}_1 + n_1\overline{x}_2) \end{aligned}$$

Weiter ist $\overline{y} = n_1/n$ und

$$\begin{aligned} s_y &= \sqrt{\frac{n_1}{n} - \left(\frac{n_1}{n}\right)^2} = \sqrt{\frac{n\,n_1 - n_1^2}{n^2}} \\ &= \sqrt{\frac{n_1(n-n_1)}{n^2}} = \sqrt{\frac{n_1 n_2}{n^2}} \end{aligned}$$

Dann ist

$$\begin{aligned} r = \frac{\mathrm{Kov}(x,y)}{s_x s_y} &= \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n s_x s_y} \\ &= \frac{n_1\overline{x}_1 - (n_1\overline{x}_1 + n_2\overline{x}_2)n_1/n}{n s_x s_y} \\ &= \frac{n\,n_1\overline{x}_1 - n_1^2\overline{x}_1 - n_1 n_2\overline{x}_2}{n^2 s_x s_y} \end{aligned}$$

$$
\begin{aligned}
&= \frac{(n-n_1)n_1\overline{x}_1 - n_1n_2\overline{x}_2}{ns_xs_y} \\
&= \frac{n_1n_2(\overline{x}_1-\overline{x}_2)}{s_xn^2\sqrt{n_1n_2/n^2}} \\
&= \frac{\overline{x}_1-\overline{x}_2}{s_x}\sqrt{\frac{n_1n_2}{n^2}}
\end{aligned}
$$

und dies ist (82). □

Beispiel 8.1 In Beispiel 5.3 wurden die Intelligenzquotienten von Eheleuten betrachtet. Im Vergleich zu dem dort berechneten Korrelationskoeffizienten soll hier der punktbiseriale Koeffizient bestimmt werden. Dazu werde angenommen, dass die IQen der Ehemänner bezüglich des Medians dichotomisiert seien; es ergibt sich die Tabelle 7, d.h. die

Tabelle 7: Intelligenzquotienten von Ehepartnern

	Intelligenzquotienten							
IQ-Ehemann	0	0	0	0	1	1	1	1
IQ-Ehefrau	75	112	93	120	119	129	126	142

X-Werte sind hier die IQen der Ehefrauen, die dichotomisierten IQen der Ehemänner seien die Y-Werte. Der Ausdruck (82) liefert dann

$$r_{pbis} = \frac{100-129}{19.93}\sqrt{\frac{4\times 4}{8^2}} = .73$$

Dieser Wert ist geringer als der in Beispiel 5.3 gefundene von .986; der Unterschied resultiert aus der Tatsache, dass man beim punkt-biserialen Korrelationskoeffizienten mit weniger Information auskommen muß. □

9 Rangkorrelationen

Die Berechnung des Produkt-Moment-Korrelationskoeffizienten

$$r_{xy} = \frac{\mathrm{Kov}(X,Y)}{s_xs_y}$$

zwischen zwei Variablen X und Y setzt Intervallskalenqualität der Messungen X und Y voraus. In diesem Abschnitt sollen zwei Korrelationsmaße vorgestellt werden, die berechnet werden können, wenn die Messungen X und Y zumindest Ordinalskalenqualität haben.

Gegeben seien also die Objekte $\omega_i, i = 1, ..., n$, und für jedes ω_i sei das Paar (x_i, y_i) von Meßwerten erhoben worden, wobei x_i und y_i mindestens Ordinalskalenqualität haben.

Es müssen zwei Fälle betrachtet werden:

1. Für mindestens ein Paar (i, j), $i \neq j$ gelte $x_i = x_j$. Dann enthält die Rangordnung der x_i eine *Rangbindung*.
2. Für alle $i \neq j$ gelte entweder $x_i < x_j$ oder $x_i > x_j$. Dann existiert für die x_i eine *Rangordnung ohne Rangbindung*.

Für die y_i sind diese Charakterisierungen natürlich analog.

Für Rangordnungen ohne Rangbindungen gelte

$$x_{i_1} < x_{i_2} < \cdots < x_{i_n}. \tag{83}$$

$(i_1, i_2, \cdots, i_n)$ ist eine Permutation von $(1, 2, \cdots, n)$. Es sei r_i der Rangplatz von x_i; es ist

$$r_i = k \quad \text{genau dann, wenn} \quad x_i = x_{i_k}$$

Analog gelte

$$y_{j_1} < y_{j_2} < \cdots < y_{j_n} \tag{84}$$

$(j_1, j_2, \cdots, j_n)$ ist ebenfalls eine Permutation von $(1, 2, \cdots, n)$. Es sei s_i der Rangplatz von y_i; es ist

$$s_i = k' \quad \text{genau dann, wenn} \quad y_i = y_{j_{k'}}$$

Statt der "<" - Beziehung kann natürlich ebensogut die ">"- Beziehung zur Definition der Rangplätze verwendet werden.

Es werde noch eine Vereinbarung für den Fall von Rangbindungen getroffen. Angenommen, es gelte

$$x_{i_1} = x_{i_2} = \cdots = x_{i_j} \tag{85}$$

und x_{i_1-1} sei der größte Wert, der kleiner als x_{i_1} ist. x_{i_1-1} habe den Rang $k-1$. Würde nun

$$x_{i_1} < \cdots < x_{i_j}$$

gelten, so erhielten sie die Rangplätze $k, k+1, \cdots, k+j$. Da nun aber (85) gilt, kann man so verfahren, dass diese Ränge *aufgeteilt* werden. Dazu wird das arithmetische Mittel

$$\overline{k} = \frac{1}{j} \sum_{m=k}^{j} m \tag{86}$$

berechnet und die $x_{i_1}, \cdots, x_{i_j}$ bekommen alle den Rangplatz $\overline{k}$ zugewiesen.

9.1 Spearmans ρ

Spearman (1904, 1906) schlug vor, als Maß des Zusammenhanges zwischen X und Y einfach die gewöhnliche Produkt-Moment-Korrelation zwischen den Rangreihen der x_i- und y_i-Werte zu bilden.

R und S seien im folgenden Variablen, die die Werte r_i bzw. s_i annehmen können.

Definition 9.1 *Es seien R und S die Rangplätze der Meßwerte x_i und y_i, die mindestens Ordinalskalenqualität haben mögen. Dann heißt*

$$\rho = \frac{Kov(R, S)}{s_R s_S} \tag{87}$$

der Spearmansche Rang-Korrelationskoeffizient (Spearmans ρ).

Es gilt der

Satz 9.1 *Die Meßwerte X und Y mögen den Bedingungen der Definition 9.1 genügen, und die Meßwerte x_i und y_i mögen ohne Rangbindungen in eine Rangreihe gebracht*

werden können. Weiter sei $d_i = r_i - r_i$, $i = 1, \cdots, n$ die Differenz der Ränge r_i und s_i für das Objekt ω_i. Dann gilt für den in (87) definierten Korrelationkoeffizienten

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{88}$$

Beweis: Bekanntlich gelten die Formeln

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \tag{89}$$

Weiter sei $\overline{r} = \sum_i r_i/n$, $\overline{s} = \sum_i s_i/n$. Aber die Ränge r_i und s_i durchlaufen jeweils die ersten natürlichen Zahlen $1, ..., n$, so dass sicherlich $\sum_i r_i = \sum_i s_i$, denn beim Summieren kommt es auf die Reihenfolge der Zahlen nicht an. Also folgt

$$\overline{r} = \overline{s} = a = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

Weiter gilt

$$\begin{aligned} n\, s_R^2 &= \sum_{i=1}^{n} (r_i - a)^2 = \sum_{i=1}^{n} r_i^2 - na^2 \\ n\, s_S^2 &= \sum_{i=1}^{n} (s_i - a)^2 = \sum_{i=1}^{n} s_i^2 - na^2 \end{aligned}$$

und natürlich

$$s_R^2 = s_S^2 = s^2$$

da $\sum_i r_i^2 = \sum s_i^2$, so dass $s_R s_S = s^2$. Wegen (89) folgt überdies

$$n\, s^2 = \frac{2n(n+1)(2n+1) - 3n(n+1)^2}{12} \tag{90}$$

Es sei

$$d_i = r_i - s_i = (r_i - a) - (r_i - a)$$

Dann ist

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (r_i - a)^2 + \sum_{i=1}^{n} (s_i - a)^2 - 2\sum_{i=1}^{n} (r_i - a)(s_i - a)$$

so dass

$$\begin{aligned} n\mathrm{Kov}(R,S) &= \sum_{i=1}^{n} (r_i - a)(s_i - a) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n} (r_i - a)^2 + \sum_{i=1}^{n} (s_i - a)^2 - \sum_{i=1}^{n} d_i^2\Big) = ns^2 - \sum_{i=1}^{n} d_i^2/2 \end{aligned} \tag{91}$$

Mithin ist

$$\rho = \frac{\mathrm{Kov}(R,S)}{s_R s_S} = \frac{s^2 - \sum_i d_i^2/2n}{s^2} = 1 - \frac{\sum_i d_i^2}{2ns^2}$$

und zusammen mit (90) folgt die Behauptung. □

Es soll noch der Fall von Rangbindungen betachtet werden:

Satz 9.2 *Es mögen die Bedingungen von Satz 9.1 gelten, allerdings sollen Rangbindungen erlaubt sein. Es sei t_i die Anzahl der Meßwerte X in der i-ten Gruppe mit gleichen Meßwerten, und u_j sei die Anzahl der Meßwerte in der j-ten Gruppe mit gleichen Y-Werten. Dann kann ρ gemäß*

$$\rho = \frac{2A - T - U - \sum_{k=1}^{n} d_k^2}{2\sqrt{(A-T)(A-U)}} \tag{92}$$

mit

$$A = \frac{n(n^2-1)}{12}, \quad T = \sum_{i=1}^{n_x}(t_i^3 - t_i)/12, \quad U = \sum_{j=1}^{n_y}(u_j^3 - u_j)/12,$$

berechnet werden, wobei n_x die Anzahl von Gruppen gleicher X-Werte und n_y die Anzahl von Gruppen gleicher Y-Werte ist.

Beweis: Horn(1942)

Beispiel 9.1 (Fortsetzung Beispiel 5.3) Zur Illustration soll das Spearmansche ρ für die Intelligenzquotienten von Ehepartnern bestimmt werden. Die Daten seien hier noch einmal vorgestellt: Die Paare können nun nach Maßgabe aufsteigender x_i-Werte angeordnet

Tabelle 8: Intelligenzquotienten von Ehepartnern

	Intelligenzquotienten							
IQ-Ehemann	90	110	100	115	117	120	125	130
IQ-Ehefrau	75	112	93	120	119	129	126	142

werden: Die Rangordnungen werden in Tab. 9 gegeben. Es ist

Tabelle 9: Rangordnung nach Maßgabe der x_i-Werte

	Intelligenzquotienten							
IQ-Ehemann	x_1	x_3	x_2	x_4	x_5	x_6	x_7	x_8
IQ-Ehefrau	y_1	y_3	y_2	y_5	y_4	y_6	y_7	y_8
d_i	0	0	0	-1	1	0	0	0

$$\rho = 1 - \frac{2}{8(64-1)} = .996$$

Der Produkt-Moment-Korrelationskoeffizient war in Beispiel 5.3 mit $r_{xy} = .986$ angegeben worden. Erwartungsgemäß ist $\rho > r_{xy}$, denn bei der Berechnung von ρ werden ja die Abweichungen von der perfekten Regression nicht berücksichtigt, so lange die übereinstimmung der Rangreihen nicht verletzt wird.

Zur Illustration werde weiter der Fall von Rangbindungen betrachtet. Es seien die folgenden Daten gegeben. Die Rangordnungen sind dann in Tab. 11 gegeben. Nach Tabelle 10 ist aber $x_2 = x_3 = x_4 = x_5$; diesen vier Werten wird demnach der Rang $(2 + 3 + 4 + 5)/4 = 14/4 = 3.5$ zugeordnet. Weiter ist $x_6 = x_7$, so dass diesen Werten der Rang $(6 + 7)/2 = 6.5$ zugeordnet wird. Es gibt zwei Gruppen mit gleichen Werten; dabei ist

Tabelle 10: Intelligenzquotienten von Ehepartnern

	Intelligenzquotienten							
IQ-Ehemann	90	110	100	100	100	115	115	130
IQ-Ehefrau	75	110	110	95	115	115	120	120

Tabelle 11: Rangordnungen nach Maßgabe der X-Werte

	Intelligenzquotienten							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Rangordnung (Ehemann)	1	2	3	4	5	6	7	8
Rangordnung (Ehefrau)	1	3	4	2	5	6	7	8
Rangordn. (Bind.,E'mann)	1	3.5	3.5	3.5	3.5	6.5	6.5	8
Rangordn. (Bind.,E'frau)	1	3.5	3.5	2	5.5	5.5	7.5	7.5
d_i	0	0	0	1.5	-2	1	-1	.5

$t_1 = 4$, $t_2 = 2$. Für die Ehefrauen ist $y_2 = y_3$ mit den Rängen 3 und 4, so dass hier der Rang 3.5 zugeordnet wird. Weiter ist $y_5 = y_6$, was den Rang 5.5 für diese beiden Werte ergibt, und schließlich erhalten die Werte $y_7 = y_8$ den Rang 7.5. Es gibt 3 Gruppen mit jeweils 2 gleichen Werten, so dass $u_1 = u_2 = u_3 = 2$. Es ist $A = 8(64-1)/12 = 42$, $T = (4^3 - 4 + 2^3 - 2)/12 = 5.5$ und $U = (2^3 - 2 + 2^3 - 2)/12 = (2^3 - 2)/6 = 1$. Weiter ist $\sum_i d_i^2 = 8.5$. Demnach ist

$$\rho = \frac{84 - 5.5 - 1 - 8.5}{2\sqrt{(42 - 5.5)(42 - 1)}} = .892$$

□

9.2 Kendalls τ

Es wird angenommen, dass keine Rangbindungen vorliegen. Den x_i-Werten seien die Ränge $r_1 = 1, \cdots, r_n = n$ zugeordnet worden. Die Folge $r_1, \cdots, r_n$ wird nun als *Ankerreihe* gewählt, in bezug auf die die Ränge s_i der y_i-Werte diskutiert werden. Gilt $r_i = s_i$ für alle i, so ist der Zusammenhang zwischen den beiden Rangreihen perfekt. Der Zusammenhang wird um so geringer sein, je häufiger der Fall $s_i \neq i$ auftritt. Insbesondere heiße der Fall $s_i > s_j$ für $i < j$ eine *Inversion*; dementsprechend heiße der Fall $s_i < s_j$ für $i < j$ eine *Proversion*. Der Zusammenhang zwischen den Rangreihen wird also um so kleiner sein, je größer die Anzahl der Inversionen ist. Die Anzahl I der Inversionen läßt sich bestimmen, indem man alle Paare (i, j) mit $i < j$ betrachtet. Es sei

$$k_{ij} = 1, \quad \text{wenn} \quad s_i > s_j, \quad i < j$$
$$k_{ij} = 0, \quad \text{wenn} \quad s_i < s_j, \quad i < j$$

Dann ist

$$I = \sum_{i<j} k_{ij} \tag{93}$$

Insgesamt gibt es $\binom{n}{2} = n(n-1)/2$ verschiedene Paare von Rängen. Ist P die Anzahl der Proversionen, so muß

$$P + I = \binom{n}{2} = \frac{n(n-1)}{2}$$

gelten. Gilt $P = n(n-1)/2$, so muß $I = 0$ sein, und der Zusammenhang zwischen den Rängen ist perfekt; dieser Fall sollte einem Korrelationsmaß von 1 entsprechen. Gilt umgekehrt $I = n(n-1)/2$, so muß $P = 0$ sein. In diesem Fall ist die Rangreihe der y_i-Werte exakt die Umkehrung der Rangreihe der x_i-Werte. Diesem Fall sollte ein Korrelationsmaß von -1 entsprechen. Ein Korrelationsmaß kann dann wie folgt eingeführt werden.

Definition 9.2 *Es sei*

$$W = P - I \tag{94}$$

Dann heißt W Kendall-Summe.

Offenbar ist $W = n(n-1)/2$ für $I = 0$ und $W = -n(n-1)/2$ für $P = 0$; W ist dann ein allerdings noch unnormiertes Maß für den Zusammenhang der Rangreihen. Da $|W|$ maximal den Wert $n(n-1)/2$ annehmen kann, wird die Normierung durch Division von W durch $n(n-1)/2$ erreicht. Dementsprechend hat man

$$\tau = \frac{W}{n(n-1)/2} = \frac{2W}{n(n-1)} \tag{95}$$

Diese Größe wurde zuerst von Kendall eingeführt; man spricht dementsprechend von *Kendalls* τ.

Beispiel 9.2 Zur Illustration werde Kendalls τ für die Daten aus Tabelle (8), Beispiel 9.1 berechnet. Es werden zunächst die beiden Rangreihen angegeben: In der folgenden

Tabelle 12: Ränge der Intelligenzquotienten

	Ränge							
IQ-Ehemann	1	3	2	4	5	6	7	8
IQ-Ehefrau	1	3	2	5	4	7	6	8

Tabelle sind die Ränge für die IQen der Ehemänner in eine aufsteigende Ordnung gebracht worden, die Ränge der Ehefrauen sind darunter zu finden: Für die Ränge der IQen der

Tabelle 13: Geordnete Ränge der Intelligenzquotienten

	Ränge							
IQ-Ehemann	1	2	3	4	5	6	7	8
IQ-Ehefrau	1	2	3	5	4	7	6	8

Ehefrauen wird jetzt die Anzahl der Inversionen berechnet; offenbar gibt es nur zwei: (5,4) und (7,6), also $I = 2$. Die Gesamtzahl von Rangpaaren ist 8(8 - 1)/2 = 28; wegen $P + I = 28$ folgt $P = 26$ und die Kendall-Summe ist $W = P - I = 26 - 2 = 24$. Dementsprechend ist Kendalls τ

$$\tau = \frac{24}{28} = .857$$

Kendalls τ ist also deutlich kleiner als Spearmans ρ. □

Der Fall von Rangbindungen ist bei Kendalls τ komplizierter als bei Spearmans ρ, man vergl. dazu etwa Bortz (1990), p. 427.

9.3 Abschließende Bemerkungen

Für gegebene Rangreihen von Meßwerten ergeben sich für Spearmans ρ und Kendalls τ im allgemeinen verschiedene Werte. Da bei Spearmans ρ die Formel für den Produkt-Moment-Korrelationskoeffizienten auf die Ränge angewandt wird, wird im Grunde unterstellt, dass den Abständen zwischen den Rangplätzen auch entsprechende Abstände zwischen den ursprünglichen Meßwerten entsprechen, was natürlich nicht der Fall zu sein braucht. Dementsprechend wird ρ im allgemeinen größer ausfallen, als es dem tatsächlichen Zusammenhang entspricht. Bei inhaltlichen Interpretationen von ρ ist also eine gewisse Zurückhaltung geboten.

Bei Kendalls τ tritt dieses Problem nicht auf, denn τ hängt ja nur von der Anzahl der Proversionen und der der Inversionen ab; möglicherweise wird der Zusammenhang sogar unterschätzt.

10 Der Regressionseffekt

Galton (1885) bemerkte bei Untersuchungen zur Vererbung, dass sich Söhne hinsichtlich ihrer Körpergröße *im Durchschnitt* weniger vom Durchschnitt der Menge der Söhne unterscheiden als sich ihre Väter vom Durchschnitt der Menge der Väter unterscheiden. Für andere Merkmale kann man das gleiche Phänomen beobachten. Man spricht deshalb von einer Regression zur Mitte. Man kann sofort fragen, ob sich ein analoger Effekt beobachten läßt, wenn man die die Körpergrösse der Väter aufgrund der Körpergröße der Söhne vorhersagen läßt. Es zeigt sich, dass tatsächlich bei jeder Vorhersage von Messwerten auf der Basis anderer Messwerte eine Regression zur Mitte findet.

Tatsächlich ist die Regression zur Mitte in erster Linie ein stochastisches Phänomen. Nachtigall und Suhl (2002)[2] haben vorgeschlagen, den Regressionseffekt über den Begriff der bedingten Erwartung zu definieren.

Definition 10.1 *Es sei $E(Y|X = x)$ der Erwartungswert von Y unter der Bedingung, dass X den Wert x hat, $E(X)$ sei der Erwartungswert von X, und σ_x und σ_y seien die Streuungen (Standardabweichungen) von X bzw. Y. Dann existiert ein Regressionseffekt genau dann, wenn*

$$\frac{|E(Y|X = x) - E(Y)|}{\sigma_y} < \frac{|x - E(X)|}{\sigma_y} \tag{96}$$

In das Beispiel der Größe von Söhnen (Y) und Vätern $/X$) übersetzt bedeutet diese Definition, dass die *standardisierte* durchschnittliche Größe der Söhne, gegeben eine bestimmte Vatergröße $X = x$, vom allgemeinen Durchschnitt der Größe der Söhne (also gemittelt über alle Vätergrößen) kleiner ist als der korrespondierende, *standardisierte* Absolutbetrag der Differenz zwischen x und dem Erwartungswert, also dem Durchscnitt, der Vätergrößen. Es sei angemerkt, dass $EY|X = x)$ eine lineare Regression implizieren kann (vergl. (67), Seite 19), aber nicht muß. Weiter wird aus der Definition 96 ersichtlich, dass man sich beim Regressionseffekt auf standardisierte Differenzen bezieht. Dies reflektiert den Sachverhalt, dass der Begriff einer "großen" oder "kleinen" Abweichung nur Sinn macht, wenn man ihn in Bezug zur Standardabeichung der Messungen setzt; bekanntlich kommen Werte, die mehr als 3 Standardabweichungen vom Erwartungswert abweichen, kaum noch vor und sind in diesem Sinne "sehr sehr groß". Weicht ein Messwert mehr als

[2] Nachtigall,C. Suhl, U.: Der Regressionseffekt. Mythos und Wirklichkeit. *Methevalreprot* , 4, (2), 2002. S.a. Internet: http://www.uni-jena.de/swv/metheval/report/

eine Standardabweichung vom Erwartungswert ab, kann man ihn "groß" nennen, weicht er zwischen zwei und drei Standardabweichungen ab, könnte man ihn "sehr groß" nennen.

Beispiel 10.1 Gegeben sei ein Würfel, und X die Augenzahl, die bei einem Wurf des Würfels "oben" liegt. Alle sechs Seiten des Würfels seien gleichwahrscheinlich, so dass $P(X = i) = p_i = 1/6$ für $i = 1, \ldots, 6$. Der Erwartungswert von X ist $E(X) = \mu = \sum_i ip_i = \sum_i i/6 = 6(6+1)/12 = 3.5$. Es werde nun eine Folge von Würfen ausgeführt; natürlich sind die Ergebnisse der einzelnen Würfe unabhängig voneinander, d.h. das Ergebnis des einen Wurfes hat keinen Einfluß auf das Ergebnis eines anderen Wurfes. Die Wahrscheinlichkeit, dass X einen Wert größer oder gleich 2 und kleiner oder gleich 5 annimmt, ist $P(2 \leq X \leq 5) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 4/6 = 2/3$, und dementsprechend ist $P(X \notin [2, 5]) = P(X = 1) + P(X = 6) = 2/6 = 1/3$. Für einen gegebenen Wurf resultiere nun eine Augenzahl, die nicht in dem Intervall $[2, 5]$ liegt, es trete also ein Ereignis mit der Wahrscheinlichkeit 1/3 ein. Die Wahrscheinlichkeit, dass beim nächsten Wurf wieder eine 1 oder eine 6 auftritt, hat wieder nur die Wahrscheinlichkeit 1/3, während die Wahrscheinlichkeit, dass die Augenzahl gleich 2, 3, 4 oder 5 ist, gleich 2/3 ist. Die Wahrscheinlichkeit, dass nach einem "extremen" Ergebnis (die 1 oder die 6 liegt "oben") eine Augenzahl gewürfelt wird, die näher am Erwartungswert 3.5 liegt, ist 2/3, also zweimal so hoch wie die Wahrscheinlichkeit, noch einmal ein extremes Resultat zu erhalten. □

Diese Betrachtung gilt natürlich nicht nur für die Gleichverteilung, sondern für beliebige Verteilungen: es sei X eine zufällige Veränderliche mit der Verteilungsfunktion $F(x) = P(X \leq x)$ und dem Erwartungswert $E(X) = \mu$. Es sei $0 < p < 1/2$. Man kann nun Zahlen a_1 und a_2 finden derart, dass $P(X \leq \mu - a_1) = p/2$ und $P(X > \mu + a_2) = p/2$. Dann ist $P(X \leq \mu - a_1) + P(X > \mu + a_2) = p$ und $P(\mu - a_1 \leq X \leq \mu + a_2) = 1 - p$, und da nach Voraussetzung $p < 1/2$, muß $1 - p > 1/2$, also größer als 1/2 sein. Beobachtet man voneinander unabhängige von Werte von X und erhält man insbesondere einen Wert aus den "extremen" Bereichen $X \leq \mu - a_1$ und $X > \mu + a_2$, so beträgt die Wahrscheinlichkeit, bei der nächsten Beobachtung wieder einen Wert aus diesen Bereichen zu finden, nur p, während die Wahrscheinlichkeit, einen weniger von μ abweichenden Wert aus dem Intervall $\mu - a_1 \leq X \leq \mu + a_2$ zu finden den größeren Wert $1 - p$ hat.

Beispiel 10.2 Es sei X exponentialverteilt, d.h. $F(x) = 1 - \exp(-\lambda x)$. Der Einfachheit halber werde $\lambda = 1$ gesetzt. Der Erwartungswert ist dann $E(X) = 1/\lambda = 1$. Es sei $p = .45$, so dass $1 - p = .55$, $p/2 = .275$. Aus $P(X \leq x_1) = 1 - \exp(-x_1) = .275$ folgt $x_1 = -\log(1 - .275) = .1397$, und $P(X > x_2) = .275 = \exp(-x_2)$, d.h. $x_2 = -\log .275 = .561$. Also ist $a_1 = \mu - x_1 = .86$, $a_2 = x_2 + \mu = .439$, und sicherlich gilt $P(X \in [x_1, x_2]) > P(X \notin [x_1, x_2])$. □

Das folgende Beispiel aus dem Bereich der Pädagogik illustriert den Regressionseffekt weniger formal:

Beispiel 10.3 Es sei X die musikalische Qualität[3], mit der ein Studierender der Musik sein Instrument spielt. Der Lehrer des Studierenden vertritt die Theorie, dass ein Lob für Aufführungen mit einer Qualität, die größer als ein bestimmte kritische Qualität x_2 sind keinen Effekt habe, da der Student mit großer Wahrscheinlichkeit bei der nächsten Aufführung unterhalb dieser Qualtiätsmarke bleibe. Tadele man ihn hingegen für Aufführungen unterhalb einer bestimmten Qualtiätsmarke x_1 so habe dies sehr wohl einen Effekt, weil er dann nämlich mit großer Wahrscheinlichkeit beim nächsten Male besser

[3]Es muss hier nicht vertieft werden, wie musikalische Qualität definiert ist!

spiele. Der Lehrer behauptet dementsprechend, seiner Erfahrung nach habe Tadel einen größeren psychologischen Effekt als Lob.

Die Qualität werde (vereinfachend) durch die zufällige Veränderliche X repräsentiert. X sei normalverteilt mit *konstantem* Erwartungswert μ und Varianz σ^2. Die Wahrscheinlichkeit, dass sein Spiel bei einer Aufführung eine Qualität X zwischen der unteren Qualitätsmarke $x_1 = \mu - a_1$ und der oberen $x_2 = \mu + a_1$ bekommt, sei gleich .8; der Einfachheit halber wird hier angenommen, dass der Lehrer die untere und die obere Qualtiätsmarke symmetrisch zu μ wählt.

Geht man von der (ebenfalls vereinfachenden) Annahme aus, dass die musikalischen Leistungen Studierenden von einer Aufführung zur anderen stochastisch unabhängig[4] sind, so sieht man, dass die Schlußfolgerung des Lehrers, seine Erfahrung lehre, dass Tadel stärker wirke als Lob, keineswegs notwendig richtig ist: mit der Wahrscheinlichkeit $p = .9$ ist der Schüler bei der jeweilig nächsten Aufführung besser als die untere Marke x_1 und schlechter als x_2, ganz unabhängig von Lob und Tadel. □

Das vorangegangene Beispiel ist insofern extrem, als angenommen wird, dass der Schüler gar nichts mehr lernt und seine Leistungen um einen in der Zeit konstanten Mittelwert zufällig schwanken. Man kann nun fragen, ob und wie sich der Regressionseffekt auswirkt, wenn der Studierende tatsächlich lernt, μ also eine nicht konstante Funktion der Zeit ist. Im einfachsten Fall ist μ eine lineare Funktion der Zeit (für einen gewissen Zeitabschnitt während des Studiums wird dieseine gute Näherung sein), so dass man $X(t) = a\,t + b + \varepsilon$ schreiben kann. ε repräsentiert die zufälligen, d.h. nicht kontrollierten Einflüsse auf die Leistung. Es zeigt sich, dass die Existenz dieser zufälligen Einflüsse auch in diesem Fall einen Regressionseffekt erzeugt.

Es werde also ganz allgemein angenommen, dass zwischen zwei Variablen x und y die Beziehung

$$y = bx + a + \varepsilon = b\hat{y} + a, \tag{97}$$

mit

$$\hat{y} = bx + a, \quad b = \varrho\frac{\sigma_y}{\sigma_x}, \tag{98}$$

besteht; ϱ ist die Korrelation zwischen x und y. Es sei $\mu_x = E(x)$, $\mu_y = E(y)$, $Var(x) = \sigma_x$ und $Var(y) = \sigma_y$. Wegen $E(\varepsilon) = 0$ gilt

$$\mu_y = b\mu_x + a \tag{99}$$

und

$$\hat{y} - \mu_y = b(x - \mu_x). \tag{100}$$

Die Abweichungen $\hat{y} - \mu_y$ und $x - \mu_x$ lassen sich stets in σ_y- bzw. σ_y-Einheiten ausdrücken, d.h. es existieren Zahlen α und β derart, dass

$$x - \mu_x = \alpha\sigma_x, \quad \hat{y} - \mu_y = \beta\sigma_y \tag{101}$$

gilt. Die Gleichungen (98), (100) und (101) implizieren dann

$$\beta\sigma_y = b\alpha\sigma_x = \varrho\frac{\sigma_y}{\sigma_x}\alpha\sigma_x, \tag{102}$$

d.h.

$$\beta = \rho\alpha. \tag{103}$$

[4]Dieser Begriff reflektiert den Sachverhalt, dass die Leistungen zufällig schwanken, in dem Sinne, dass von einer Leistung nicht auf die Nächste geschlossen werden kann.

Nun ist sicher $|\beta| = |\rho||\alpha|$, so dass

$$|\beta| \leq |\alpha| \quad \text{für} \quad |\rho| \leq 1. \tag{104}$$

Dies bedeutet, dass die *Vorhersage* $\hat{y}$ eines y-Wertes, gemessen in σ_y-Einheiten, weniger von μ_y abweicht wie der korrespondierende x-Wert in σ_x-Einheiten von μ_x. Diese Aussage defniert den Regressionseffekt.

Auf den ersten Blick mag diese Aussage als verwickelt und schwer zu durchschauen erscheinen. Man muß sich aber in Erinnerung rufen, dass der Begriff eines "grossen" oder "kleinen" x- bzw. y-Wertes nur Sinn in bezug auf die Varianz oder die Streuung σ_x bzw. σ_y macht. Eine numerisch grosse Abweichung eines y-Wertes von μ_y reflektiert ja nur dann eine "grosse" Abweichung, wenn sie groß ist relativ zur durchschnittlichen Abweichung der Werte, also relativ zu σ_y. Eine analoge Aussage gilt für die x-Abweichungen. Die Aussage (104) bedeutet demnach, dass für eine Korrelation, deren Betrag kleiner als 1 ist (wenn also x und y nicht perfekt positiv oder negativ korrelieren), die vorhergesagten Werte $\hat{y}$ weniger σ_y-Einheiten von μ_y abweicht als die korrespondierenden x-Werte σ_x-Einheiten von μ_x abweichen. Ein Teil der tatsächlichen Messwerte y können aber durchaus größer (oder noch kleiner) sein, denn in sie gehen ja noch die zufällig variierenden ε-Werte ein.

Es versteht sich, dass all diese Aussagen gelten, wenn man statt der Populationsparameter a, b, μ_x, μ_y etc. die entsprechenden Schätzungen $\hat{a}$, $\hat{b}$, $\bar{x}$, $\bar{y}$ etc. einsetzt.

Es soll nun gezeigt werden, dass der Regressionseffekt genau dann auftritt, wenn die Fehlervarianz σ_ε^2 grösser als Null ist; genau dann gilt ja für die Korrelation $|\varrho| < 1$.

Nach (104) ergibt sich der Regressionseffekt, wenn $|r| < 1$ ist. Dieser Sachverhalt soll noch etwas vertiefender behandelt werden. Es ist

$$\rho^2 = \sigma_y^2/\sigma_x^2 = 1 - \sigma_e^2/\sigma_y^2, \quad \text{d.h. } \varrho = \pm\sqrt{1 - \sigma_e^2/\sigma_y^2}. \tag{105}$$

Es ist also $|\rho| < 1$ wenn $\sigma_e^2 > 0$, d.h. der Regressionseffekt wird durch die *zufälligen*, d.h. hier von x unabhängigen Effekte erzeugt, die in den Wert von y eingehen und für die σ_e^2 ein Maß ist. Der Regressionseffekt hängt also nicht vom Wert der Varianz σ_x^2 ab; von σ_y^2 hängt er nur dann ab, wenn $\sigma_\varepsilon^2 \neq 0$ in die Varianz von σ_y^2 eingeht (es ist ja $\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_\varepsilon^2$).

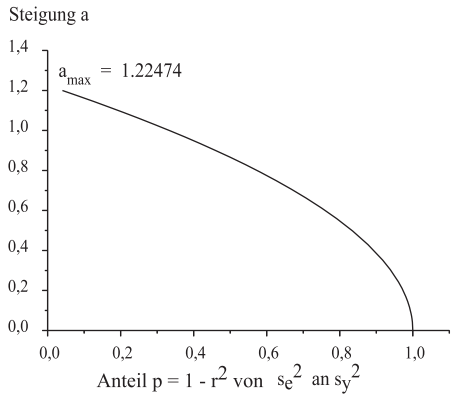
Damit kann die Beziehung zu den zu Beginn dieses Abschnitts angestellten und in den Beispielen 10.2 und 10.3 illustrierten Betrachtungen hergestellt werden. Es wird zunächst die Beziehung zwischen dem Regressionsparameter a und der Fehlervarianz σ_ε^2 explizit gemacht; da $a = r\sigma_y/\sigma_x$ folgt aus (105)

$$a = \frac{\sigma_y}{\sigma_x}\sqrt{1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2}}. \tag{106}$$

In Abbildung 10.2 wird die Beziehung zwischen dem Wert von $a = a_{xy}$ und σ_ε^2 gezeigt, wobei $\sigma_\varepsilon^2 = p\sigma_y^2$ mit $0 \leq p \leq 1$ gesetzt wurde; p ist der Anteil von σ_ε^2 and σ_y^2. Da $p = \sigma_\varepsilon^2/\sigma_y^2 = 1 - \rho^2$ wir auf diese Weise die Beziehung zwischen dem Parameter a_{xy} und der Korrelation $r = r_{xy}$ deutlich. Für $p = 0$ ist $\sigma_\varepsilon^2 = 0$, d.h. man hat den Fall fehlerfreier Vorhersage der y-Werte aufgrund der y-Werte und dementsprechend $a = a_{max}$, und für $p = 1$ ist $\sigma_\varepsilon^2 = \sigma_y^2$, d.h. $a = 0$; in diesem Fall kann y nicht anhand von x vorhergesagt werden. Es ist $\sigma_x^2 = 100$ und $\sigma_y^2 = 150$ angenommen worden, so dass $a_{max} = \sigma_y/\sigma_x = 1.22474$.

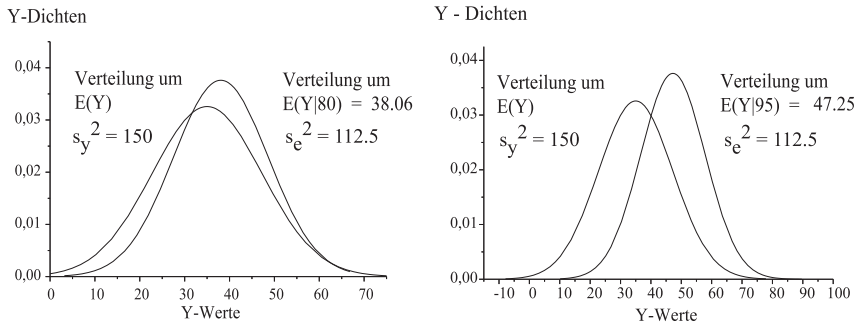
Generell folgt $a \to 0$ für $\sigma_\varepsilon \to \sigma_y$; in der Tat bedeutet $\sigma_\varepsilon \to \sigma_y$ ja, dass $\hat{s}_{\hat{y}} \to 0$, d.h. dass die Variation der Y-Werte unabhängig von der der X-Werte ist und nicht auf der X-Werte zurückgeführt werden kann. Je größer also σ_ε^2 im Vergleich zu σ_y^2 ist, desto kleiner

Abbildung 7: Die Steigung a_{xy} der Regressionsgeraden als Funktion des Anteils p von σ_ε^2 an σ_y^2



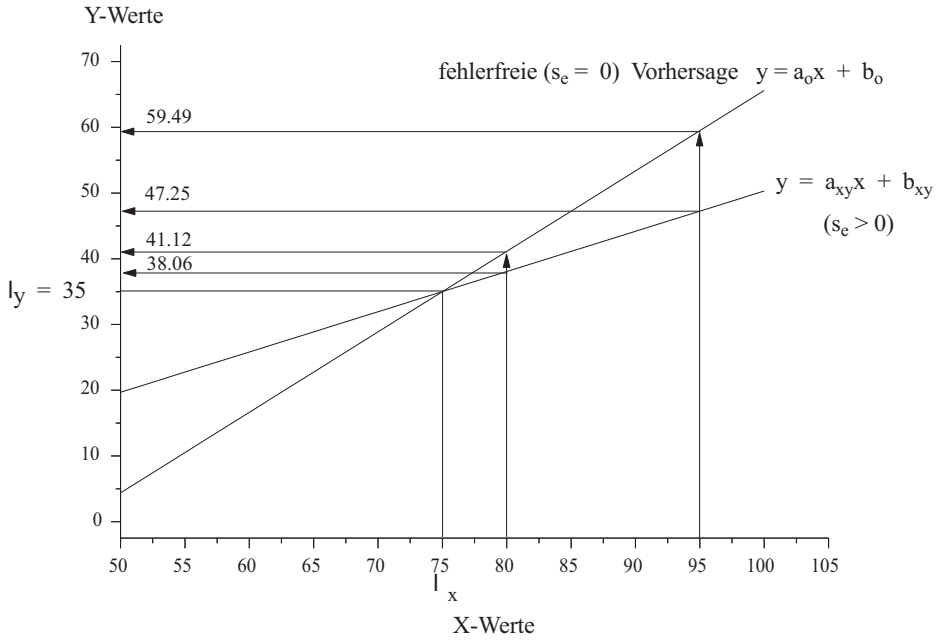
ist der Wert von a, und dies bedeutet, dass die $\hat{y}_i$-Werte langsamer mit x_i wachsen als im Falle eines größeren a-Werts, also eines kleinerem σ_ε^2-Werts. Dieses wiederum bedeutet, dass für kleineren a-Wert bzw. größerem σ_ε^2-Wert *die Verteilung der Werte um* $\hat{y}_i$, d.h. der Y-Werte unter der Bedingung $X = x_i$, näher an den Mittelwert $\bar{y}$ der Y-Werte heranrückt. Dieses wiederum bedeutet, dass sich der $\hat{y}_i$-Wert weniger von $\bar{y}$ (in Einheiten von σ_y) unterscheidet als x_i von $\bar{x}$ (in Einheiten von σ_x); die Abweichung $y_i - \bar{y}$ gilt ja als "groß", wenn es nur wenige Werte gibt, die ebenfalls so sehr wie y_i von $\bar{y}$ abweichen; der Bezug auf die Anzahl ebensosehr abweichender Werte wird durch den Wert von s hergestellt. Unter der Bedingung x_i treten mit größerer Wahrscheinlichkeit Y-Werte auf, die sich in σ_y-Einheiten weniger von $\bar{y}$ unterscheiden als sich x_i von $\bar{x}$ in σ_x-Einheiten unterscheidet.

Abbildung 8: Unbedingte und bedingte Verteilungen der Y-Werte



In Abbildung 9 wird der Regressionseffekt illustriert. Betrachtet werden zwei normalverteilte Variablen X und Y mit $E(X) = \mu_x = 35$, $Var(X) = \sigma_x^2 = 100$, $E(Y) = \mu_y = 35$, $Var(Y) = \sigma_y^2 = 150$. Für $\sigma_\varepsilon = 0$ ist $a_{xy} = a_{max} = a_0 = \sigma_y/\sigma_x = \sqrt{150/100} = \sqrt{1.5} = 1.2247$. Für $x = 80$ und $x = 95$ erhält man die Vorhersagen $\hat{y} = 41.12$ und $\hat{y} = 59.49$. Ist dagegen $\sigma_\varepsilon^2 \neq 0$, etwa $\sigma_\varepsilon^2 = .75 \times \sigma_y^2 = 112.5$, so erhält man die Regressionsgerade mit der Steigung $a_{xy} = .612$; dem entspricht ein Korrelationskoeffizient von $r = .5$

Abbildung 9: Zum Regressionseffekt



und damit ein Determinationskoeffizient $\rho^2 = .25$, und dieser Wert entspricht dem Anteil der Varianz der vorhergesagten Y-Werte an der Gesamtvarianz der Y-Werte, korrespondierend zu $\sigma_\varepsilon^2/\sigma_y^2 = .75$. Für $x = 80$ bzw. $x = 95$ erhält man die Vorhersagen $\hat{y} = 38.06$ bzw. $\hat{y} = 47.25$. Diese Vorhersagen sind kleiner als die entsprechenden Vorhersagen für den fehlerfreien Fall. Für $x = 80$ beträgt die Abweichung vom Mittelwert $\mu_x = 75$ $x - \mu_x = 5 = \alpha\sigma_x = \alpha 10$, d.h. die Abweichung entspricht $\alpha = 5/10 = .5$ σ_x-Einheiten. Die Abweichung des vorhergesagten y-Wertes $\hat{y} = 38.06$ von μ_y beträgt $38.06 - 35 = 3.06 = \beta\sigma_y = \beta 12.247$, d.h. $\hat{y} = 38.06$ weicht nur $\beta = .25$-σ_y-Einheiten von μ_y ab, ist also, in σ_y-Einheiten, nur halb so groß wie die Abweichung $x - \mu_x$, wenn diese in σ_x-Einheiten ausgedrückt wird. Für $x = 95$ ergibt eine analoge Rechnung, dass die Abweichung $95 - \mu_x = 20$ gerade $\alpha = 2$ σ_x-Einheiten beträgt, die entsprechende Abweichung der Vorhersage $\hat{y} = 47.25$ von $\mu_y = 35$ aber nur $\beta = 1$ σ_y-Einheiten.

Beispiel 10.3 ist ein Spezialfall des Regressionseffekts: hier ist $a_{xy} = 0$ und damit $r_{xy} = 0$. Es können ebenfalls die x_i Werte auf der Basis der y_i-Werte vorausgesagt werden. Dann gilt $\hat{x}_i = \hat{a}_{xy} y_i + \hat{b}_{xy}$. Für die Abweichungen vom Mittelwert gilt dann, ausgedrückt in Einheiten der Streuungen σ_x und σ_y, $\hat{x}_i - \bar{x} = \xi_i \sigma_x = \hat{a}_{xy} \eta_i \sigma_y = \hat{a}_{xy}(y_i - \bar{y})$. Da $\hat{a}_{xy} = \varrho\sigma_x/\sigma_y$ folgt $\xi_i \sigma_x = \varrho(\sigma_x/\sigma_y)\eta_i\sigma_y$ und damit

$$\xi_i = \varrho\eta_i. \tag{107}$$

Diese Gleichung besagt, dass die Abweichung der $\hat{x}_i$-Werte von $\bar{x}$ für $\varrho < 1$ stets kleiner ausfällt als die korrespondierende Abweichung des y_i-Wertes von $\bar{y}$. Wird wieder die Größe der Väter durch x repräsentiert und die Größe der Söhne durch y, so besagt (**??**), dass die Söhne jeweils weniger vom Mittelwert abweichen als ihre Väter. Die Gleichung (107) sagt, dass gleichzeitig die Väter weniger vom Mittelwert abweichen als die Söhne. Dies macht deutlich, dass der Regressionseffekt primär ein Effekt von zufälligen Variationen bei der Vorhersage einer Variablen aufgrund einer anderen ist.

11 Zusammenfassende Betrachtungen

Die Ergebnisse dieses Kapitels sollen noch einmal kurz zusammengefaßt werden, insbesondere sollen die Zielsetzungen bei der Berechnung von Regressionskoeffizienten im Vergleich zu Korrelationskoeffizienten noch einmal aufgeführt werden. Die Bemerkungen über die Regressionskoeffizienten a und b beziehen sich auf die Modelle 1 und 2; explizit werden die Betrachtungen anhand des Modells 1 durchgeführt.

Unter der Annahme, dass die Beziehung zwischen X und Y linear ist, dass also eine Beziehung der Form

$$y_i = a\,x_i + b + e_i$$

zwischen den Variablen existiert, können die Parameter a und b mit der Methode der Kleinsten Quadrate geschätzt werden. Generell ist a ein Maß für die Veränderung, die Y erfährt, wenn X variiert wird. Für $a = 0$ folgt, dass X und Y unabhängig voneinander sind. Für $a \neq 0$ kann man davon ausgehen, dass zwischen X und Y eine Beziehung besteht. Der numerische Wert von $a \neq 0$ hängt aber von den Maßeinheiten von X und Y ab; auch ein sehr kleiner Wert von a kann eine ausgeprägte Abhängigkeit der beiden Variablen bedeuten. Wie gut die Vorhersage der Y-Werte aufgrund der X-Werte ist, kann dem Wert von $a \neq 0$ nicht entnommen werden. Die Güte der Vorhersage hängt von den e_i ab. Ist s_e^2 groß, so bedeutet auch ein numerisch großer Wert von a noch nicht, dass eine genaue Vorhersage möglich ist.

Dafür können sowohl mit a als auch mit b inhaltliche Betrachtungen verknüpft werden.

Beispiel 11.1 Die in Beispiel 2.1 vorgestellte Untersuchung über den Zusammenhang zwischen Alter und Merkfähigkeit soll nach Geschlechtern getrennt analysiert werden. Der *Vergleich* der Parameterwerte kann hier sinnvoll sein. Findet man dann etwa $|\hat{a}(\text{weiblich})| < |\hat{a}(\text{männlich})|$, so heißt dies, dass sich die Merkfähigkeit der Frauen langsamer verändert, d.h. sie werden nicht so schnell senil wie die Männer. Gilt ebenfalls $|\hat{b}(\text{weiblich})| < |\hat{b}(\text{männlich})|$, so heißt dies, dass die Männer in jüngeren Jahren eine bessere Merkfähigkeit haben als die Frauen. Gilt $|\hat{b}(\text{weiblich})| > |\hat{b}(\text{männlich})|$ und $|\hat{a}(\text{weiblich})| \approx |\hat{a}(\text{männlich})|$, so heißt dies, dass die Frauen den Männern hinsichtlich der Merkfähigkeit überlegen sind und bleiben, sich der Abbau der Merkfähigkeit aber mit der gleichen Geschwindigkeit wie bei den Männern vollzieht. □

Ist man an einem Maß für die Genauigkeit der Vorhersage von Y aufgrund von X interessiert, so kann man den Korrelationskoeffizienten

$$r_{xy} = r_{yx} = r = \frac{\text{Kov}(x,y)}{s_x s_y}$$

berechnen. r ist unabhängig von den Maßeinheiten von X und Y. Die Güte der Vorhersage läßt sich dann durch den Determinationskoeffizienten

$$D = r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

berechnen. r^2 entspricht also dem Anteil der "vorhergesagten" Varianz der Y-Werte an der Gesamtvarianz der Y-Werte, denn es gilt ja

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

Einer Korrelation von $r = .5$ entspricht also ein $D = r^2 = .25$, d.h. nur 25% der Varianz der Y-Werte lassen sich durch die Unterschiede in den X-Werten erklären, - 75%

der Unterschiede zwischen den Y-Werten gehen auf andere Ursachen als die durch X ausgedrückten zurück. So gesehen ist $r = .5$ also noch keineswegs ein "hoher" Wert.

12 Die Schwarzsche Ungleichung

Es seien $a_i \epsilon \mathbb{R}$, $b_i \epsilon \mathbb{R}$ und es sei

$$|a| = \sqrt{\sum_{i=1}^{n} a_i^2}, \quad |b| = \sqrt{\sum_{i=1}^{n} b_i^2}$$

Weiter sei

$$x_i = \frac{a_i}{|a|}, \, y_i = \frac{b_i}{|b|}. \tag{108}$$

Es ist

$$\sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} x_i y_i \geq 0$$

und mithin

$$\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 \geq \sum_{i=1}^{n} {x_i}^2 + 2 \sum_{i=1}^{n} x_i y_i$$

Aber es ist $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = 1$, mithin gilt $2 \geq 2 \sum_{i=1}^{n} x_i y_i$ oder $\sum_{i=1}^{n} x_i y_i \leq 1$. Setzt man (108) für x_i und y_i ein, so erhält man

$$\sum_{i=1}^{n} \frac{a_i}{|a|} \frac{b_i}{|b|} = \frac{1}{|a| \, |b|} \sum_{i=1}^{n} a_i b_i \leq 1,$$

und mithin

$$\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}$$

bzw.

$$\left| \sum_{i=1}^{n} a_i b_i \right|^2 \leq \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2$$

□

13 Partielle Korrelationen

13.1 Die Fragestellung

Der praktische Wert von Korrelationen besteht u.a. darin, dass sie ein Maß für die Güte der Vorhersagbarkeit der Werte einer Variablen aufgrund der Werte einer anderen Variablen repräsentieren. Hat man aber keine inhaltliche Theorie über die Beziehung zwischen den betrachteten Variablen, so kann der Versuch, eine solche Theorie aus empirisch bestimmten Korrelationen zu destillieren, ein schwieriges Unternehmen sein. Denn eine noch so hohe Korrelation zwischen zwei Variablen muß keineswegs auch eine direkte inhaltliche Beziehung zwischen den Variablen reflektieren. So sei z.B. X_1 der Alkoholkonsum in einem Jahr in einem bestimmten Land; X_1 habe Intervallskalenniveau. X_2

sei die Anzahl der Männer in diesem Land, die sich entscheiden, Priester zu werden. Bestimmt man die Werte für X_1 und X_2 für eine Anzahl von Jahren und berechnet dann die Korrelation zwischen X_1 und X_2, so findet man eine positive Korrelation. Kausale Interpretationen wie "Wer trinkt, wird auch Priester" oder "Wer Priester wird, fängt früher oder später an zu trinken" sind aber vermutlich wenig realistisch. Der Schluß, dass nun deutlich werde, dass die Korrelationsrechnung eben Ausdruck der fehlgeleiteten Neigung dröger Formelfetischisten zu Kalauern sei, denen das Spielen mit Zahlen Ersatz für den abhanden gekommenen Bezug zur realen Welt sei, ist aber voreilig. Denn die Korrelation zwischen Alkoholkonsum und der Häufigkeit, mit der der Priesterberuf ergriffen wird, legt die Wirkung einer dritten, nicht explizit in die Betrachtung aufgenommenen Variablen nahe, die man global als "Wirtschaftliche Situation" bezeichnen könnte. Ist die nämlich schlecht und steigt also die Arbeitslosigkeit, so kann diese Situation für die einen erhöhten Alkoholkonsum bedeuten, während sie für die anderen eine Hinwendung zu den spirituellen Aspekten des Lebens nach sich zieht.

Natürlich kann mehr als nur eine Variable auf die Beziehung zwischen zwei bestimmten Variablen einwirken, d.h. die Korrelation r_{12} zwischen X_1 und X_2 kann durch gleichzeitige Einwirkung der Variablen $X_3, \cdots, X_k$ zustande kommen. Zunächst wird nur der Einfluß einer Variablen, X_3, betrachtet.

13.2 Der partielle Korrelationskoeffizient

Die Aufgabe besteht darin, ein Maß für die Korrelation zwischen X_1 und X_2 zu finden, das von der simultanen Wirkung der Variablen X_3 auf X_1 und X_2 befreit ist. Es werde dazu angenommen, dass die Beziehung zwischen X_1 und X_3 sowie die zwischen X_2 und X_3 linear sei. Dann gilt jedenfalls

$$\begin{aligned} X_{1i} &= \hat{a}_{13} X_{3i} + \hat{b}_{13} + u_i \\ X_{2i} &= \hat{a}_{23} X_{3i} + \hat{b}_{23} + v_i \end{aligned} \tag{109}$$

Dabei repräsentieren die u_i die "Meßfehler" bei der Vorhersage der X_{3i} durch die X_{1i} und v_i sind die "Meßfehler" bei der Vorhersage der X_{2i} durch die X_{3i}. Diese Meßfehler enthalten aber auch alle Aspekte der durch X_1 bzw. X_2 repräsentierten Variablen, die durch X_3 *nicht* vorhergesagt werden können. Wir können nun die Beziehung zwischen den Variablen u und v betrachten: nach dem eben Gesagten repräsentieren sie diejenigen Aspekte von X_1 und X_2, die nicht auf die durch X_3 repräsentierte Variable zurückgeführt werden können Aus (109) folgt dann

$$\begin{aligned} u_i &= X_{1i} - \hat{a}_{13} X_{3i} - \hat{b}_{13} \\ v_i &= X_{2i} - \hat{a}_{23} X_{3i} - \hat{b}_{23} \end{aligned} \tag{110}$$

Bestimmt man nun die Korrelation zwischen den Variablen u und v (die u_i, v_i sind die *Realisierungen* dieser beiden Variablen), so erhält man eine Korrelation zwischen den beiden ursprünglich interessierenden Variablen X_1 und X_2, die vom Effekt der Variablen X_3 "bereinigt" worden sind. Diese Korrelation ist durch

$$r(u,v) := r_{12.3} = \frac{\mathrm{Kov}(u,v)}{s_u s_v} \tag{111}$$

gegeben. Dabei soll die Schreibweise $r_{12.3}$ andeuten, dass es sich bei $r(u,v)$ um die Korrelation zwischen den vom Effekt der Variablen X_3 "bereinigten" Variablen X_1 und X_2 handelt.

Die Korrelation $r(u,v)$ hat einen speziellen Namen:

Definition 13.1 *Die Variablen X_1, X_2 und X_3 haben mindestens Intervallskalenniveau und genügen den Bedingungen (109). Dann heißt die in (111) eingeführte Korrelation $r(u, v)$* partielle Korrelation*; sie wird mit $r_{12.3}$ bezeichnet.*

Da nun u und v in (110) mit X_1, X_2 und X_3 verbunden werden, liegt es nahe, $r_{12.3}$ durch die Korrelationen zwischen diesen Variablen auszudrücken. Es gilt

Satz 13.1 *Es werde angenommen, dass die Variablen X_1, X_2 und X_3 den eben beschriebenen Bedingungen genügen. Dann gilt*

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{112}$$

Den Beweis des Satzes findet man im letzten Abschnitt dieses Kapitels.

Sind die drei Variablen X_1, X_2 und X_3 gegeben, so können natürlich die Korrelationen zwischen irgendzwei von ihnen bei Auspartialisierung der jeweils dritten berechnet werden. So gilt

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} \tag{113}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}. \tag{114}$$

Es sollen noch einige Eigenschaften von partiellen Korrelationen angemerkt werden; dabei beziehen wir uns auf $r_{12.3}$, die Aussagen für die anderen partiellen Korrelationen sind natürlich analog.

Es gelte $r_{12.3} = 0$. Aus (112) folgt, dass dieser Fall genau dann eintritt, wenn $r_{12} = r_{13}r_{23}$ ist. In diesem Fall kommt also die Korrelation zwischen X_1 und X_2 nur dadurch zustande, dass auf beide Variablen die Variable X_3 "einwirkt"; durch diese Redeweise soll aber noch nicht die Existenz einer *kausalen* Wirkung postuliert werden! Denn im Prinzip ist es möglich, dass die Korrelationen zwischen X_1, X_2 und X_3 wiederum nur durch die Wirkung einer weiteren Variablen X_4 zustande kommen. Das eingangs genannte Beispiel der Korrelation zwischen der Häufigkeit der Wahl des Priesterberufs und des Alkoholkonsums kann hier zur Illustration herangezogen werden.

Es sei $r_{12.3} < r_{12}$. Man kann vermuten, dass es außer der gleichzeitigen Wirkung von X_3 noch andere Beziehungen zwischen X_1 und X_2 gibt, die möglicherweise auf andere Variablen X_4. X_5, ... zurückgeführt werden können.

Nun sei $r_{12.3} > r_{12}$. Dieser Fall bedeutet, dass die Wirkung der Variablen X_3 die Korrelation zwischen X_1 und X_2 *reduziert*. Aus (112) folgt aus $r_{12.3} > r_{12}$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} > r_{12}$$

und damit

$$r_{12} - r_{13}r_{23} > r_{12}\sqrt{1 - {r_{13}}^2}\sqrt{1 - {r_{23}}^2}$$

d.h.

$$r_{12}(1 - \sqrt{1 - {r_{13}}^2}\sqrt{1 - {r_{23}}^2}) > r_{13}r_{23}.$$

Wegen $r_{13} < 1$, $r_{23} < 1$ folgt $0 < \sqrt{1 - {r_{13}}^2}\sqrt{1 - {r_{23}}^2} < 1$. Es sei zunächst $0 < r_{12}$ Man sieht nun, dass die obige Ungleichung *stets* erfüllt ist, wenn (i) r_{13} und r_{23} unterschiedliche Vorzeichen haben, denn dann ist $r_{13}r_{23} < 0$, oder wenn (ii) wenigstens eine der beiden Korrelationen r_{13}, r_{23} gleich Null ist. Betrachten wir zunächst den Fall (i); es sei $r_{13} > 0$,

$r_{23} < 0$. Große Werte von X_3 bedeuten kleine Werte von X_2 und umgekehrt, während große Werte von X_3 auch große Werte von X_1 bedeuten. Die gleichzeitige Einwirkung von X_3 auf X_1 und X_2 bewirkt also z.B., dass ein größerer Wert von X_1 einen kleineren Wert von X_2 impliziert und umgekehrt. Diese Gegenläufigkeit der Einwirkungen bewirkt eine Reduktion der Kovarianz von X_1 und X_2. Partialisiert man den Effekt von X_3 heraus, so muß sich die Kovarianz zwangsläufig erhöhen.

Im Falle (ii) wirkt X_3 nur auf eine der Variablen X_1 und X_2. Es sei $r_{23} = 0$; dann wirkt also X_3 nicht auf X_2. Aber X_3 wirkt noch auf X_1 und erzeugt damit eine Variation in den Werten von X_1, die nicht auf die von X_2 zurückgeführt werden kann, und dies reduziert wiederum die Kovariation von X_1 und X_2. Auspartialisieren von X_3 muß deshalb zu einer Erhöhung der Kovariation von X_1 und X_2 führen.

Beispiel 13.1 In Beispiel 2.1 wurde die Beziehung zwischen X_1 = Merkfähigkeit und X_2 = Alter betrachtet; eine Querschnittsuntersuchung ergab eine lineare Beziehung mit negativer Steigung zwischen den beiden Merkmalen. In einer neuen Untersuchung habe man 100 zufällig ausgewählte Personen im Alter (X_2) zwischen 20 und 70 Jahren bezüglich (i) ihrer Merkfähigkeit (X_1) und (ii) bezüglich des intellektuellen Trainings (X_3) getestet; die Problematik der Erfassung einer solchen Variablen soll hier nicht weiter diskutiert werden. Zwischen Merkfähigkeit und Alter ergab sich die Korrelation $r_{12} = -.5$. Zwischen der Merkfähigkeit und dem Training ergab sich eine Korrelation $r_{13} = .8$, und zwischen Alter und Training wurde $r_{23} = -.6$ berechnet. Um den reinen Alterseffekt auf die Merkfähigkeit besser abschätzen zu können, soll die Korrelation zwischen Alter und Merkfähigkeit nach Herauspartialisieren des Effekts des Trainings berechnet werden. Diese ist die partielle Korrelation $r_{12.3}$. Nach (112) ist

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - {r_{13}}^2}\sqrt{1 - {r_{23}}^2}} = \frac{-.5 - .8(-.6)}{\sqrt{1 - .8^2}\sqrt{1 - .6^2}} = -.04.$$

Der Determinationskoeffizient für diese "bereinigte" Korrelation ist $r^2_{12.3} = .002$, d.h. die Variation der jetzt vorhergesagten Merkfähigkeitswerte beträgt nur .2 % der Variation der tatsächlich gemessenen Merkfähigkeitswerte. Dies bedeutet, dass das Altern noch keinen Verlust der Merkfähigkeit nach sich ziehen muß, dass aber Variablen, die mit dem Altern korrelieren, wie zunehmende Bequemlichkeit, mangelnde Anregung, Dominanz von Routinetätigkeiten, einen Verlust der Merkfähigkeit nach sich ziehen können.[5] □

Beispiel 13.2 Es sei X_1 die Note, die ein(e) Student(in) der Psychologie im Diplom erhält, und X_2 sei die Punktzahl des/der gleichen Studenti(e)n in der Statistikklausur. Die Korrelation zwischen X_1 und X_2 betrage $r_{12} = .5$. Weiter sei X_3 ein Maß für das Interesse an qualitativ-phänomenologischen Betrachtungen psychologischer Variablen. Die Korrelation zwischen dem Merkmal "Diplomnote" und X_3 sei $r_{13} = .05$; dies bedeutet, dass das Interesse an der phänomenologische Psychologie weder auf ein gutes noch auf ein schlechtes Diplomzeugnis schließen läßt. Die Korrelation zwischen der Punktzahl in Statistik (X_2) und X_3 aber betrage $r_{23} = -.7$, d.h. je ausgeprägter das Interesse an der phänomenologischen Psychologie, desto schlechter fallen die Klausuren aus, und umgekehrt.

Es soll nun die Diplomnote aufgrund der Ergebnisse der Statistikklausur vorhergesagt werden. Da $r_{12} = .5$, beträgt der Anteil der Varianz der anhand der Statistikklausur *vorhergesagten* Diplomnoten an der tatsächlichen Varianz der Diplomnoten nur .25, d.h. 75 % der Varianz der Diplomnoten werden durch andere Faktoren erzeugt. Es wird nun

[5] Diese Aussagen stellen sicherlich eine starke Vereinfachung dar; hier geht es nur um das Prinzip der partiellen Korrelation.

vermutet, dass das Herauspartialisieren des Interesses für die phänomenologische Psychologie die Vorhersage aufgrund der Statistikklausur verbessern könnte. Es ergibt sich dann

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - {r_{13}}^2}\sqrt{1 - {r_{23}}^2}} = \frac{.5 - .05 \cdot -.7}{\sqrt{1 - .05^2}\sqrt{1 - .7^2}} = .75.$$

"Bereinigt" man also die Variablen "Klausurerfolg" und "Studienerfolg" (Diplomnote) vom Effekt der Variablen "Interesse an der phänomenologischen Psychologie", so läßt sich der Studienerfolg besser aufgrund des Klausurerfolgs vorhersagen; immerhin ist $r^2_{12.3} = .56$, d.h. 56 % der Varianz der Diplomnoten sind auf die Variation der Klausurpunkte zurückführbar. Vermutlich sind der Klausurerfolg einerseits und der Studienerfolg andererseits auf Variablen wie Intelligenz, Arbeitsdisziplin, Interesse an den zu erlernenden Bereichen etc. zurückführbar. Die Erhöhung der Vorhersagekraft des Klausurerfolges durch Herauspartialisieren des Interesses an der Phänomenologie ergibt sich daraus, dass einerseits diese Variable mit dem Klausurerfolg negativ korreliert, mit der Kriteriumsvariablen Diplomnote aber nicht korreliert. Etwas vergröbert ausgedrückt kann man sagen, dass sich jemand, der sich für die Phänomenologie interessiert, sich nicht für die Statistik erwärmt, und wer sich für Statistik interessiert, hat geringes Interesse an der Phänomenologie. Unterschiede hinsichtlich des Interesses an der Phänomenologie erzeugen also Unterschiede im Klausurerfolg. Da aber das Interesse an der Phänomenologie kaum mit dem Studienerfolg korreliert, ist das Phänomenologieinteresse für die Vorhersage des Studienerfolges irrelevant. Da dieses Interesse aber den Prädiktor "Klausurerfolg" beeinflußt, wirkt es wie eine Störvariable, die durch das Herauspartialisieren eliminiert wird. □

Das Auspartialisieren von Variablen kann also eine Korrelation zwischen zwei Variablen verringern oder gar auf Null bringen; die herauspartialisierte Variable hat dann einen Scheinzusammenhang zwischen den ursprünglich betrachteten Variablen erzeugt.

Variablen, deren Auspartialisierung die ursprüngliche Korrelation erhöht, heißen auch *Suppressorvariable*; sie werden bei einer an sich guten Prädiktorvariablen miterfaßt und erzeugen eine in bezug auf das interessierende Kriterium irrelevante Variation der Prädiktorvariablen. Damit *unterdrücken* sie, werden sie nicht explizit in Rechnung gestellt, die Vorhersagekraft der Prädiktorvariablen, obwohl sie selbst unabhängig vom Kriterium sind.

13.3 Der semipartielle Korrelationskoeffizient

Gelegentlich kann es von Interesse sein, eine Variable X_3 nicht aus beiden Variablen X_1 und X_2 herauszupartialisieren, sondern nur aus einer. Man spricht dann von einer *semipartiellen Korrelation*.

Beispiel 13.3 Eine Beraterfirma führt regelmäßig Seminare für dynamische Jungmanager durch. Am Ende eines Seminars läßt sie die Leiter durch die Teilnehmer auf Ratingskalen beurteilen: X_1 soll die Fähigkeit zur Einnahme einer Führungsposition abbilden, X_2 das Ausmaß, in dem die betreffende Person "sympathisch" ist, und X_3 eine Schätzung der "Intelligenz". Man kann davon ausgehen, dass alle drei Variablen positiv miteinander korrelieren. Man kann nun untersuchen, wie sehr X_1 und X_2 miteinander korrelieren, wenn X_2 von X_3 bereinigt worden ist. Es gelte

$$X_{2i} = a_{13}X_{3i} + b_{13} + u_i, \quad i = 1, \cdots, n$$

so dass

$$u_i = X_{2i} - a_{13}X_{3i} - b_{13}$$

folgt; $u = (u_1, \cdots, u_n)'$ ist die von der Intelligenz bereinigte Sympathievariable. Die Korrelation $r(X_1, u)$ ist die semipartielle Korrelation zwischen Führungsfähigkeit und Sympathie. □

Eine explizite Formel für $r(X_1, u)$ leitet man analog zu den Partialkorrelationen her. Es genügt, hier das Resultat anzugeben:

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}}, \tag{115}$$

wobei die Schreibweise $r_{1(2.3)}$ anzeigen soll, dass die Korrelation zwischen den Variablen X_1 und X_2 unter Herauspartialisierung der Variablen X_3 aus X_2 berechnet werden soll.

Das Prinzip des Herauspartialisierens kann fortgesetzt werden. So könnte man aus der "Sympathie", die man für einen (Management-)Lehrer empfindet, noch das Merkmal "hat Erfahrung" (X_4) herauspartialisieren, oder X_5: "soziale Kompetenz". Man erhält dann die *Partialkorrelationen höherer Ordnung* $r_{1.23} = r_{1(2.3)}$, $r_{1.234} = r_{1(2.34)}$, etc. Letztlich läuft dieses Vorgehen darauf hinaus, dass ein Merkmal, das zur "Vorhersage" etwa eines Merkmals X_1 herangezogen werden soll, in konstituierende Komponenten zerlegt wird bzw. in Komponenten, von denen zumindest angenommen wird, dass sie die ursprüngliche Prädiktorvariable konstituieren. Benennt man diese Komponenten mit $X_2, \cdots, X_k$, so wird man auf ein Modell der Form

$$X_{1i} = b_2 X_{2i} + b_3 X_{3i} + \cdots + b_k X_{ki} + e_i \tag{116}$$

geführt. Man spricht dann von der *multiplen Korrelation*; diese wird im folgenden Kapitel besprochen.

13.4 Anhang: Beweis zu Satz 13.1

$$\begin{aligned} \overline{u} &= x_1 - \hat{a}_{13}\, x_3 - b_{13} \\ \overline{v} &= x_2 - \hat{a}_{23}\, x_3 - b_{23} \end{aligned}$$

und folglich ist

$$\begin{aligned} \mathrm{Kov}(u, v) &= \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) = \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1 - \hat{a}_{13}(x_{3i} - \bar{x}_3))(x_{2i} - \bar{x}_2 - \hat{a}_{23}(x_{3i} - \bar{x}_3)). \end{aligned}$$

Führt man die Abkürzungen $x_{1i} = X_{1i} - \bar{x}_1$, $x_{2i} = X_{2i} - \bar{x}_2$ und $x_{3i} = X_{3i} - \bar{x}_3$ ein und berücksichtigt, dass

$$\hat{a}_{12} = r_{12}\frac{s_1}{s_2}, \quad \hat{a}_{13} = r_{13}\frac{s_1}{s_3}, \quad \hat{a}_{23} = r_{23}\frac{s_2}{s_3}$$

gilt, so erhält man, setzt man diese Ausdrücke in den für $\mathrm{Kov}(u, v)$ ein,

$$\frac{1}{n}\sum_{i=1}^{n}(x_{1i} - x_{3i}r_{12}\frac{s_1}{s_2})(x_{2i} - x_{3i}r_{23}\frac{s_2}{s_3}) =$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}(x_{1i}x_{2i} - x_{1i}x_{3i}r_{23}\frac{s_2}{s_3} - x_{2i}x_{3i}r_{12}\frac{s_1}{s_2} + x_{3i}^2 r_{12}r_{23}\frac{s_1}{s_3}) = \\
&= \frac{1}{n}\sum_{i=1}^{n} x_{1i}x_{2i} - r_{23}\frac{s_2}{s_3}\frac{1}{n}\sum_{i=1}^{n} x_{1i}x_{3i} - \\
&- r_{12}\frac{s_1}{s_2}\frac{1}{n}\sum_{i=1}^{n} x_{2i}x_{3i} + r_{12}r_{13}\frac{s_1}{s_3}\frac{1}{n}\sum_{i=1}^{n} x_{3i}^2. \qquad (117)
\end{aligned}
$$

Nun ist aber $\sum_i x_{1i}x_{2i}/n = \hat{a}_{12}s_2^2$, $\sum_i x_{1i}x_{3i}/n = \hat{a}_{13}s_3^2$ etc, und nach Substitution der in (117) gegebenen Ausdrücke für $\hat{a}_{12}$ etc hat man

$$\mathrm{Kov}(u,v) = r_{12}s_1s_2 - r_{13}r_{23}s_1s_2 - r_{12}r_{23}s_1s_3 + r_{12}r_{23}s_1s_3,$$

d.h.

$$\mathrm{Kov}(u,v) = (r_{12} - r_{13}r_{23})s_1s_2. \qquad (118)$$

Es bleibt, die Standardabweichungen s_u und s_v der Variablen u und v zu bestimmen. Dazu werde (109) betrachtet: hier entspricht z.B. X_1 der Kriteriumsvariablen Y, X_3 entspricht der Prädiktorvariablen X und u entspricht dem "Fehler" e in dem in Abschnitt 3.4 behandelten Modell. Dann entspricht s_u der Standardabweichung $s(e)$, d.h. dem Standardschätzfehler. In Kapitel 7 war dafür der Ausdruck $s(e) = s(y)\sqrt{1-r^2}$ gefunden worden. Setzt man dafür die in (109) auftretenden Größen X, Y und e ein, so erhält man

$$s_u = s(u) = s_1\sqrt{1-{r_{13}}^2}, \quad s_v = s(v) = s_2\sqrt{1-{r_{23}}^2}.$$

Damit ist dann

$$r_{12.3} = \frac{\mathrm{Kov}(u,v)}{s_u s_v} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

□