Regressionsgerade und Korrelationskoeffizient

Für *n* Merkmalsträger seien die Beobachtungswerte i = 1…n der Merkmale *x* und *y* festgestellt worden. Gegeben sind also *n* Wertepaare der Merkmalsausprägungen x_i und y_i

$$\left\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots,(x_n,y_n)\right\}$$

Die durchschnittliche Ausprägung der Merkmale ist

(1) $$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i$$

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

Unter der Voraussetzung, dass alle Merkmalsausprägungen der Grundgesamtheit erhoben wurden, ist die Varianz der Merkmale

(3) $$s_x^{\,2} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

(4) $$s_y^{\,2} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

Die Kovarianz ist

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

Der Korrelationskoeffizient ist

(6) $$r = \frac{s_{xy}}{s_x\cdot s_y} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$$

Würden alle Beobachtungspunkte in einem Koordinatensystem mit der Abszisse *x* und der Ordinate *y* auf einer Geraden liegen, bestünde ein funktionaler linearer Zusammenhang zwischen *x* und *y*. Das heißt, mit den Parametern ***a*** für die Höhe der Geraden bei x = 0 und *b* für ihre Steigung, besteht der Zusammenhang

(7) $$y = a + b\cdot x$$

Wenn man in diese lineare Funktion einen bestimmten Wert von *x* einsetzt, also ein beliebiges x_i, erhält man aus der Funktion genau den zugehörigen Wert von y_i.

Tatsächlich kann man aber nicht davon ausgehen, dass alle Beobachtungspunkte auf einer Geraden liegen. Es könnte auch ein anderer, nicht-linearer funktionaler Zusammenhang bestehen, ein funktionaler Zusammenhang könnte durch unbekannte Einflüsse überlagert sein, und es könnte auch überhaupt kein Zusammenhang bestehen. Die Beobachtungspunkte stellen dann keine Gerade dar, sondern eine mehr oder weniger geordnete Ansammlung von Punkten, sie bilden eine Punktwolke, die in einem Streudiagramm dargestellt wird:

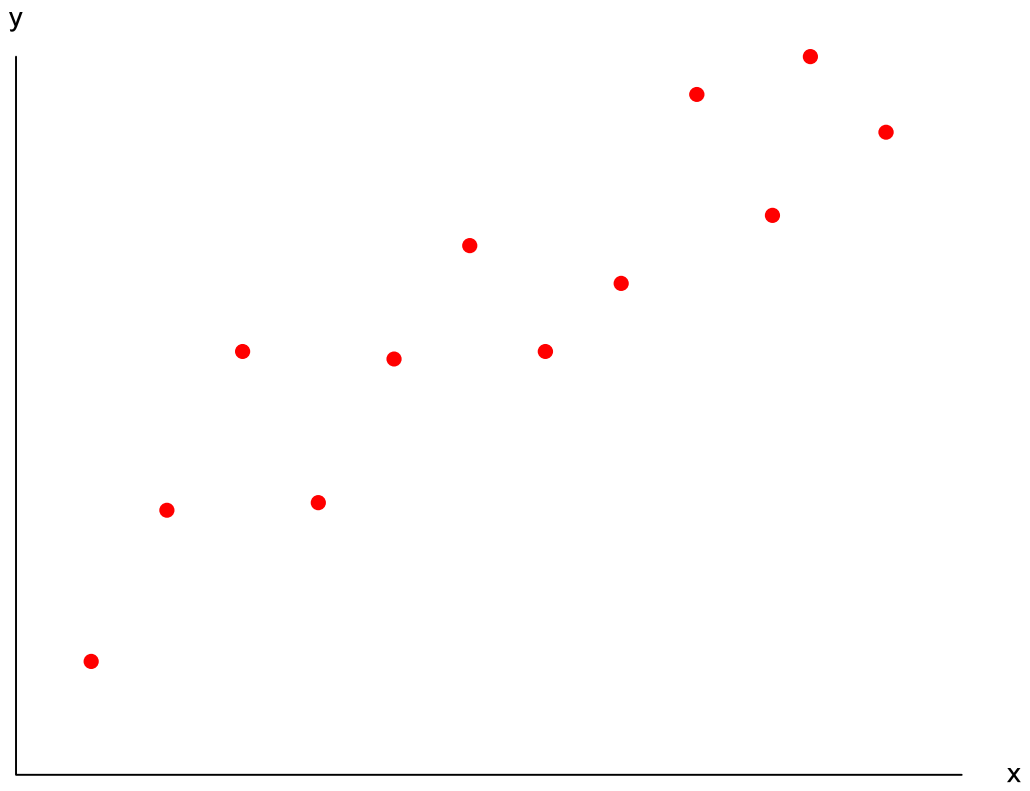

Abbildung 1: Streudiagramm

Man sieht, dass diese Wolke eine gewisse Richtung aufweist, sie weist nach oben. Je größer die Werte von *x* sind, desto größer wird auch *y*, aber mit kleineren Ausnahmen. Ein Zusammenhang zwischen *x* und *y* ist offensichtlich. Es liegt nahe, diesen Zusammenhang durch eine mathematische Funktion darzustellen und die Stärke des Zusammenhangs zu messen.

Eine Funktion, die den Zusammenhang am besten darstellt, nennt man eine Regressionsfunktion, weil der Zusammenhang zwischen den Werten auf eine mathematische Funktion zurückgeführt wird. Welche Funktion infrage kommt, kann man mit dem Augenschein abzuschätzen versuchen. Wenn der Zusammenhang linear zu sein scheint, verwendet man eine lineare Funktion, die Regressionsgerade. Wenn sich die Steigung der Wertepaare abschwächt, kann man eine logarithmische Funktion verwenden, bei einer zunehmenden Steigung eine Exponentialfunktion; bei einer negativen Steigung ist wiederum eine Gerade geeignet, wenn die Steigung konstant zu sein scheint, sonst kommt auch eine quadratische Funktion infrage. Wie gut die Annäherung an die tatsächlichen Werte durch die Regressionsfunktion ist, sollte man durch eine Kennzahl messen können.

Das notwendige Instrumentarium wird im Folgenden anhand eines linearen Zusammenhangs entwickelt, der durch eine Regressionsgerade dargestellt wird. Dieser lineare Zusammenhang scheint hier auch vorzuliegen, wie man erkennen kann, wenn man mit der freien Hand versucht, eine Regressionsgerade einzuzeichnen, die den beobachteten Werten möglichst nahe kommt:

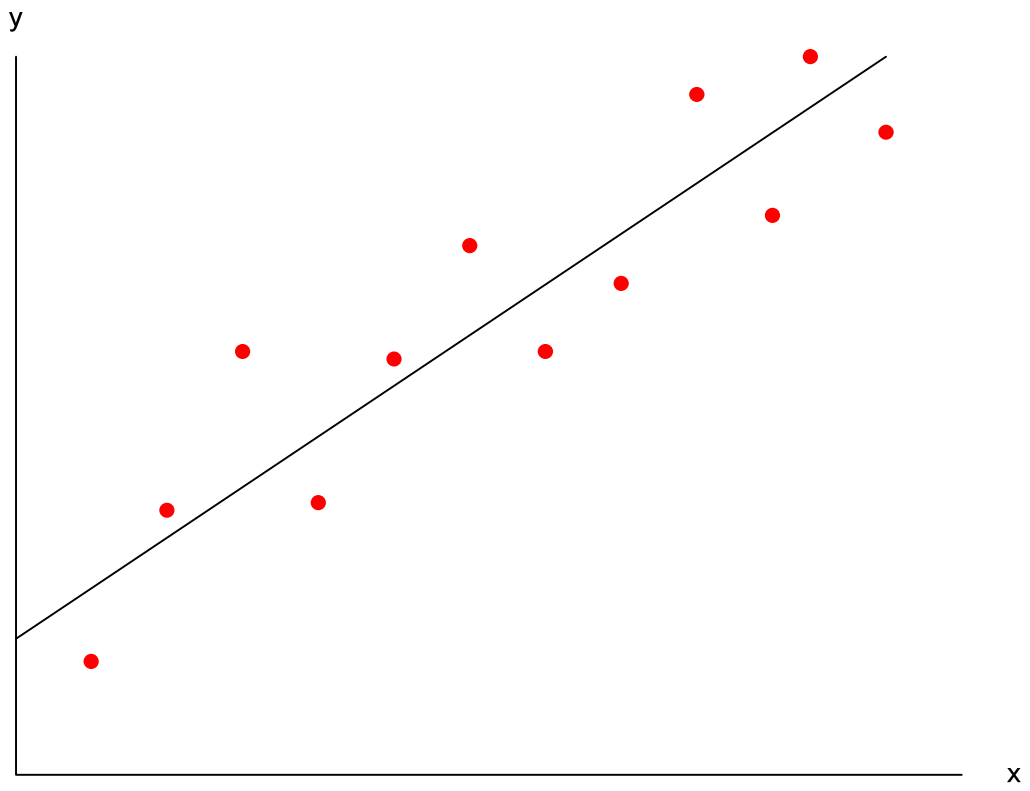

Abbildung 2: Freihand-Regressionsgerade

Ob man nun mit der freien Hand die beste Anpassung der Geraden an die Punktewolke getroffen hat, kann man nicht sagen, solange nicht definiert ist, wie die Güte der Anpassung gemessen wird. Kein einziger der Beobachtungspunkte liegt ja hier auf der Regressionsgeraden, sondern jeder Punkt liegt entweder darüber oder darunter. Da *x* die unabhängige Variable ist, deren Werte im Koordinatensystem vorgegeben sind und hiernach aufgrund der Funktion die Werte der abhängigen Variablen *y* berechnet werden, bestehen Abweichungen zwischen den tatsächlichen y-Werten, also jeweils y_i, und dem jeweiligen Wert von *y*, der sich aufgrund der Regressionsgeraden bei $x = x_i$ ergibt. Wird dieser Wert als $\hat{y}_i$ bezeichnet und die Abweichung als e_i, so gilt:

(8) $$e_i = y_i - \hat{y}_i$$

An einem ausgewählten Punkt (x_i,y_i) der Punktwolke dargestellt:

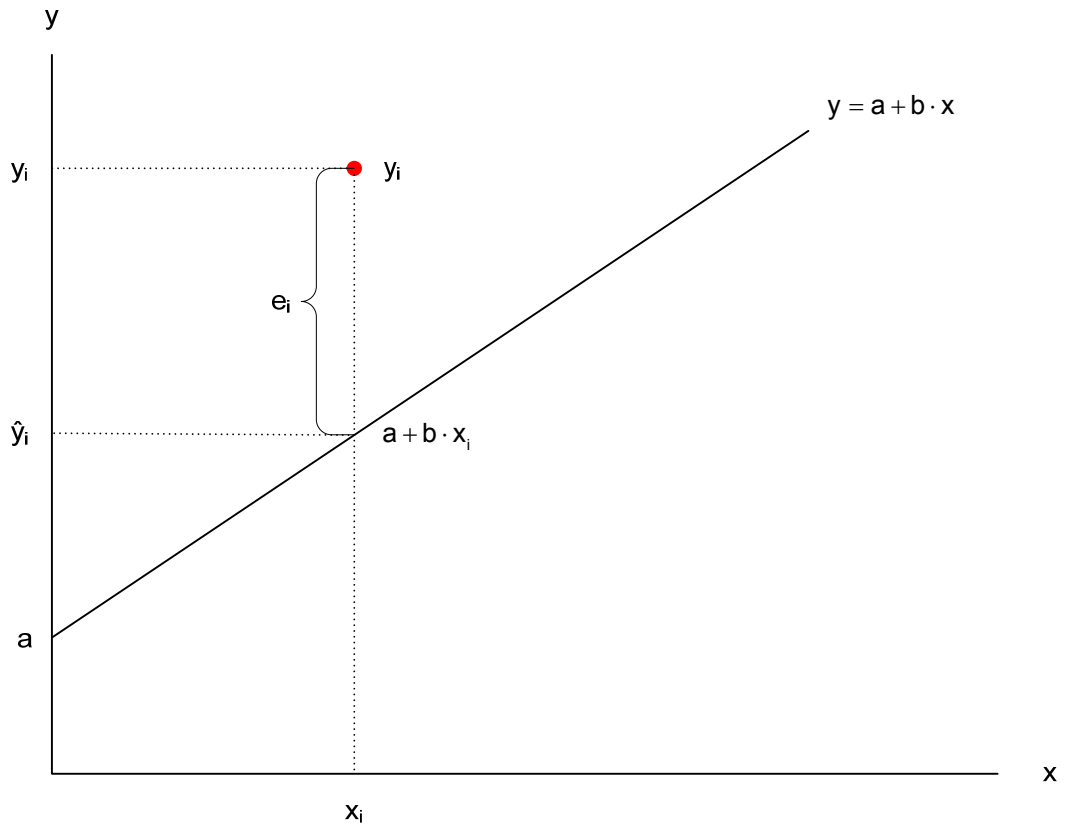

Abbildung 3: Abweichung von der Regressionsgeraden

Wird $\hat{y}_i$ als abhängige Variable der Regressionsfunktion betrachtet, so gilt

(9) $$\hat{y}_i = a + b \cdot x_i$$

Dies in die Definition der Abweichung gemäß Gleichung (8) eingesetzt:

(10) $$e_i = y_i - (a + b \cdot x_i) = y_i - a - b \cdot x_i$$

Da man versuchen wird, die Regressionsgerade mitten durch die Punktwolke zu ziehen, um die Abweichungen zu minimieren, werden die Beobachtungswerte teils über und teils unter der Regressionsgeraden liegen, sodass es positive und negative Abweichungen geben wird. Nun ist aber sowohl eine positive wie eine negative Abweichung eben eine Abweichung, und man müsste die absoluten Abweichungen addieren, um eine Summe der Abweichungen zu erhalten, die man minimieren könnte. Quadriert man indessen die Abweichungen und sucht das Minimum der Quadratsumme, so ist das Quadrat immer positiv, und man braucht auf das Vorzeichen keine Rücksicht zu nehmen. Dies ist die von Gauß eingeführte Methode der kleinsten Quadrate, die seitdem verwendet wird.

Die Parameter *a* und *b* der Regressionsgeraden sind also so zu bestimmen, dass die Regressionsgerade die Summe der Abweichungsquadrate minimiert. Mit *SQA* für die Summe der Abweichungsquadrate und der Abweichung gemäß (10) lautet das mathematische Problem:

(11) $$SQA(a,b) = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2 \rightarrow \min.$$

Wenn diese Funktion ein Minimum in Bezug auf *a* und *b* hat, dann liegt es dort, wo die Steigung gleich null ist. Es sind also die partiellen Ableitungen nach *a* und *b* zu bilden und diese gleich null zu setzen. Da alle Summanden die gleiche Struktur haben, genügt es, zunächst nur einen der Summanden zu differenzieren:

$$\frac{\delta}{\delta a}\left(y_i - a - b \cdot x_i\right)^2$$

$$\frac{\delta}{\delta b}\left(y_i - a - b \cdot x_i\right)^2$$

Man sieht, dass in der abzuleitenden Funktion die Variablen *a* und *b* eine Funktion bilden, nämlich die Funktion $y_i - a - b \cdot x_i$, und dass hierauf eine weitere Funktion anzuwenden ist, nämlich dass sie quadriert werden muss. Man könnte die Quadratklammer ausmultiplizieren und dann erst ableiten, einfacher ist aber die Anwendung der Kettenregel. Diese gilt für derartige verschachtelte Funktionen und lautet in allgemeiner Form

$$y = u\left[v(x)\right]$$

$$\frac{dy}{dx} = \frac{du}{dv} \cdot \frac{dv}{dx}$$

Hierbei ist *u* die Funktion, die auf die Funktion *v* anzuwenden ist, und *x* die Variable, nach der zu differenzieren ist. Diese ist hier *a* bzw. *b*, sodass zunächst für *a* gilt:

$$y = u\left[v(a)\right]$$

$$\frac{dy}{da} = \frac{du}{dv} \cdot \frac{dv}{da}$$

Die Funktion v(a), auch innere Funktion genannt, ist

$$v(a) = y_i - a - b \cdot x_i$$

Die Funktion *u*, auch äußere Funktion genannt, ist diejenige, die auf die innere Funktion anzuwenden ist. Da die innere Funktion quadriert werden soll, gilt hier:

$$u = v^2$$

Die Ableitung ist dann

$$\frac{\delta}{\delta a}\left(y_i - a - b \cdot x_i\right)^2 = \frac{du}{dv} \cdot \frac{dv}{da} = 2 \cdot v \cdot (-1) = 2 \cdot \left(y_i - a - b \cdot x_i\right) \cdot (-1)$$

(12) $$\frac{\delta}{\delta a}\left(y_i - a - b \cdot x_i\right)^2 = -2 \cdot \left(y_i - a - b \cdot x_i\right)$$

Für die Ableitung nach *b* gilt entsprechend:

$$\frac{\delta}{\delta b}\left(y_i - a - b \cdot x_i\right)^2 = \frac{du}{dv} \cdot \frac{dv}{db} = 2 \cdot v \cdot \left(-x_i\right) = 2 \cdot \left(y_i - a - b \cdot x_i\right) \cdot \left(-x_i\right)$$

(13) $$\frac{\delta}{\delta b}\left(y_i - a - b \cdot x_i\right)^2 = -2 \cdot \left(y_i - a - b \cdot x_i\right) \cdot x_i$$

Die Ableitungen der einzelnen Summanden können nun zu den partiellen Ableitungen der Quadratsumme zusammengefügt werden, um die Nullstellen zu bestimmen:

(14) $$\frac{\delta SQA(a,b)}{\delta a} = -2 \cdot \sum_{i=1}^{n}\left(y_i - a - b \cdot x_i\right) = 0$$

(15) $$\frac{\delta SQA(a,b)}{\delta b} = -2 \cdot \sum_{i=1}^{n} (y_i - a - b \cdot x_i) \cdot x_i = 0$$

Aus Gleichung (14) folgt:

$$-2 \cdot \sum_{i=1}^{n} (y_i - a - b \cdot x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - a - b \cdot x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} b \cdot x_i = 0$$

Hierin bedeutet $\sum_{i=1}^{n} a$, dass für alle *i* der Wert *a* zu addieren ist, insgesamt also *n* Mal. Dies lässt sich auch durch das Produkt $n \cdot a$ darstellen, sodass:

$$\sum_{i=1}^{n} y_i - n \cdot a - \sum_{i=1}^{n} b \cdot x_i = 0 \quad |: n$$

$$\frac{1}{n} \cdot \sum_{i=1}^{n} y_i - a - b \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 0$$

Hierin Gleichung (1) und Gleichung (2) eingesetzt:

(16) $$\bar{y} - a - b \cdot \bar{x} = 0$$

Die entsprechenden Umformungen an Gleichung (15) vorgenommen:

$$-2 \cdot \sum_{i=1}^{n} (y_i - a - b \cdot x_i) \cdot x_i = 0$$

$$\sum_{i=1}^{n} (y_i - a - b \cdot x_i) \cdot x_i = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} a \cdot x_i - \sum_{i=1}^{n} b \cdot x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a \cdot \sum_{i=1}^{n} x_i - b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a \cdot n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i - b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

(17) $$\sum_{i=1}^{n} x_i \cdot y_i - a \cdot n \cdot \bar{x} - b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

Die Gleichungen (16) und (17) müssen nach den gesuchten Parametern *a* und *b* aufgelöst werden. Hierzu wird aus beiden Gleichungen ein Lösungsblock gebildet:

$$\bar{y} - a - b \cdot \bar{x} = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a \cdot n \cdot \bar{x} - b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

Die erste Gleichung des Lösungsblocks wird mit $n \cdot \bar{x}$ multipliziert:

$$n \cdot \bar{x} \cdot \bar{y} - a \cdot n \cdot \bar{x} - b \cdot n \cdot \bar{x}^2 = 0$$

$$\sum_{i=1}^{n} x_i \cdot y_i - a \cdot n \cdot \bar{x} - b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

Es wird die Differenz zwischen den Gleichungen gebildet:

$$n \cdot \bar{x} \cdot \bar{y} - a \cdot n \cdot \bar{x} - b \cdot n \cdot \bar{x}^2 - \sum_{i=1}^{n} x_i \cdot y_i + a \cdot n \cdot \bar{x} + b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

$$n \cdot \bar{x} \cdot \bar{y} - b \cdot n \cdot \bar{x}^2 - \sum_{i=1}^{n} x_i \cdot y_i + b \cdot \sum_{i=1}^{n} x_i^2 = 0$$

Die Gleichung wird nach *b* aufgelöst:

$$b \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2 \right) = \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}$$

$$b = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}$$

Zähler und Nenner des Bruches werden durch *n* geteilt:

(18)
$$b = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

Man kann nun zeigen, dass im Zähler dieses Bruches nichts anderes steht als die Kovarianz s_{xy} und im Nenner die Varianz s_x^2. Hierzu werden zunächst in Gleichung (5) die Klammern ausmultipliziert:

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

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

$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot \bar{y} - \frac{1}{n} \cdot \sum_{i=1}^{n} \bar{x} \cdot y_i + \frac{1}{n} \cdot \sum_{i=1}^{n} \bar{x} \cdot \bar{y}$$

$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{y} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i - \bar{x} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} y_i + \frac{1}{n} \cdot \sum_{i=1}^{n} \bar{x} \cdot \bar{y}$$

Hierin bedeutet $\sum_{i=1}^{n} \bar{x} \cdot \bar{y}$, dass der Summand $\bar{x} \cdot \bar{y}$ für alle *i* addiert werden muss, insgesamt also *n* Mal, was sich auch durch das Produkt $n \cdot \bar{x} \cdot \bar{y}$ darstellen lässt. Neben diesem Ausdruck werden auch Gleichung (1) und Gleichung (2) in die obige Gleichung eingesetzt, sodass:

$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{y} \cdot \bar{x} - \bar{x} \cdot \bar{y} + \bar{x} \cdot \bar{y}$$

(19)
$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

Die entsprechenden Umformungen an Gleichung (3) vorgenommen:

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

$$s_x^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2 \right)$$

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

Hierin Gleichung (1) und $\sum_{i=1}^{n} \bar{x}^2 = n \cdot \bar{x}^2$ eingesetzt:

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

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

Die Gleichungen (19) und (20) können nun in (18) eingesetzt werden:

(21) $$b = \frac{s_{xy}}{s_x^2}$$

Den Zusammenhang von *b* mit dem Korrelationskoeffizienten erkennt man, wenn *b* mit s_y erweitert wird:

$$b = \frac{s_{xy}}{s_x \cdot s_x} \cdot \frac{s_y}{s_y} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x}$$

Hierin Gleichung (6) eingesetzt:

(22) $$b = r \cdot \frac{s_y}{s_x}$$

Um *a* zu ermitteln, wird Gleichung (21) in (16) eingesetzt:

$$\bar{y} - a - \frac{s_{xy}}{s_x^2} \cdot \bar{x} = 0$$

(23) $$a = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x}$$

Die Parameter *a* und *b* in Gleichung (7) eingesetzt ergibt die Regressionsgerade:

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

(25) $$y - \bar{y} = \frac{s_{xy}}{s_x^2} \cdot \left(x - \bar{x} \right)$$

Diese Regressionsgerade soll die Summe der Abweichungsquadrate gemäß Gleichung (11) minimieren. Hierzu wurden die Nullstellen der Funktion *SQA(a,b)* ermittelt. Durch die Gleichsetzung der partiellen Ableitungen mit null findet man aber nur den Punkt, bei dem die Steigung der Funktion gleich null ist. An dieser Stelle kann ein Minimum, ein Maximum oder ein Sattelpunkt liegen. Welcher

Art dieser Extremwert ist, erkennt man an der zweiten Ableitung. Diese wird für einen der Summanden von *SQA(a,b)* ermittelt und ergibt sich aus der Ableitung der Gleichungen (12) und (13):

$$\frac{\partial^2}{\partial a^2}\left(y_i - a - b \cdot x_i\right)^2 = 2 > 0$$

$$\frac{\partial^2}{\partial b^2}\left(y_i - a - b \cdot x_i\right)^2 = 2x_i^2 > 0$$

Die zweiten Ableitungen sind also positiv. Da die Ableitung einer Funktion die Steigung angibt, also die Veränderung der Funktion bei größer werdender unabhängiger Variable, bedeutet eine positive zweite Ableitung, dass die erste Ableitung eine positive Steigung hat, also immer größer wird, wenn die unabhängige Variable größer wird. Wenn die erste Ableitung nun bei einem bestimmten Wert der unabhängigen Variablen gleich null ist, muss sie bei einem ständigen Ansteigen vor der Nullstelle negativ gewesen sein und wird nach der Nullstelle positiv. Die ursprünglich abgeleitete Funktion weist also vor der Nullstelle einen fallenden Verlauf auf und danach einen steigenden Verlauf. Das aber ist das Kennzeichen für ein Minimum. Die Regressionsgerade minimiert also wie gewünscht die Quadrate der Abweichungen.

Die Eigenschaften der Regressionsgeraden können nun weiter untersucht werden. Setzt man zunächst in Gleichung (25) $x = \bar{x}$, so wird $y = \bar{y}$. Die Regressionsgerade läuft also durch den Punkt $\left(\bar{x}, \bar{y}\right)$, auch Schwerpunkt der Beobachtungsreihe genannt.

Aus Gleichung (22) für die Steigung der Regressionsgeraden ergibt sich, dass die Richtung der Steigung nur vom Korrelationskoeffizienten *r* abhängt, denn die Standardabweichungen s_x und s_y sind definitionsgemäß stets positiv. Ist auch *r* positiv, dann ist b > 0, die Regressionsgerade hat eine positive Steigung. Eine durch ein positives *r* angezeigte Korrelation zwischen x_i und y_i spiegelt sich also in einer ansteigenden Regressionsgeraden wider. Je größer *r* ist, je stärker also der Zusammenhang zwischen den Beobachtungswerten, desto steiler steigt die Regressionsgerade an. Bei $r = 1$ wird die Steigung nur noch vom Verhältnis der Standardabweichungen s_y und s_x bestimmt. Wenn $r = 0$ ist, ist auch $b = 0$, und die Regressionsgerade verläuft horizontal. Dies zeigt an, dass es keinen Zusammenhang zwischen x_i und y_i gibt.

Ein negativer Wert von *r*, also eine negative Korrelation mit einer gegenläufigen Entwicklung von x_i und y_i , führt zu einer negativen Steigung der Regressionsgeraden, die für $r = -1$ bei gleichbleibendem s_y und s_x am stärksten ist. Je schwächer die negative Korrelation, desto geringer ist die Steigung der Regressionsgeraden; bei $r = 0$ verläuft diese wiederum horizontal.

Im Folgenden sei noch gezeigt, dass der Korrelationskoeffizient nur Werte zwischen +1 und -1 annehmen kann und dass bei $|r| = 1$ alle Beobachtungswerte auf der Regressionsgeraden liegen.[1]

Ausgangspunkt ist die Funktion der Regressionsgeraden nach Gleichung (25):

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} \cdot \left(x - \bar{x}\right)$$

Wenn alle Beobachtungswerte auf der Regressionsgeraden liegen, folgen die Beobachtungswerte x_i und y_i eben dieser Funktion, es gilt also

(26)
$$y_i - \bar{y} = \frac{s_{xy}}{s_x^2} \cdot \left(x_i - \bar{x}\right)$$

Hieraus folgt

$$\left(y_i - \bar{y}\right) - \frac{s_{xy}}{s_x^2} \cdot \left(x_i - \bar{x}\right) = 0$$

[1] vgl. hierzu K. Bosch, Großes Lehrbuch der Statistik, München / Wien 1996, S. 78 f.

Wenn dieser Ausdruck quadriert wird, ist das Quadrat positiv oder gleich null:

$$\left[\left(y_i-\bar{y}\right)-\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)\right]^2\geq 0$$

Das Quadrat wird für alle *i* summiert:

$$\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)-\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)\right]^2\geq 0$$

Um den durchschnittlichen Wert der Summanden zu bestimmen, wird durch die Anzahl der Beobachtungswerte *n* geteilt, falls alle Merkmalsausprägungen der Grundgesamtheit erhoben wurden, bei einer Stichprobe durch n – 1. Für die Beweisführung kommt es darauf nicht an. Hier wird durch *n* geteilt:

(27)
$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)-\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)\right]^2\geq 0$$

Die eckige Klammer wird ausmultipliziert:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)^2-2\cdot\left(y_i-\bar{y}\right)\cdot\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)+\frac{s_{xy}^{\,2}}{s_x^{\,4}}\cdot\left(x_i-\bar{x}\right)^2\right]\geq 0$$

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2-2\cdot\frac{s_{xy}}{s_x^{\,2}}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)+\frac{s_{xy}^{\,2}}{s_x^{\,4}}\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\geq 0$$

Hierin die Gleichungen (3), (4) und (5) eingesetzt:

$$s_y^{\,2}-2\cdot\frac{s_{xy}\cdot s_{xy}}{s_x^{\,2}}+\frac{s_{xy}^{\,2}\cdot s_x^{\,2}}{s_x^{\,4}}=s_y^{\,2}-2\cdot\frac{s_{xy}^{\,2}}{s_x^{\,2}}+\frac{s_{xy}^{\,2}}{s_x^{\,2}}\geq 0$$

$$s_y^{\,2}-\frac{s_{xy}^{\,2}}{s_x^{\,2}}\geq 0$$

$s_y^{\,2}$ wird ausgeklammert:

$$s_y^{\,2}\cdot\left(1-\frac{s_{xy}^{\,2}}{s_x^{\,2}\cdot s_y^{\,2}}\right)\geq 0$$

Hierin Gleichung (5) eingesetzt:

(28)
$$s_y^{\,2}\cdot\left(1-r^2\right)\geq 0$$

Aus dieser Bedingung für eine Regressionsgerade können die möglichen Werte von *r* abgeleitet werden. Da nämlich $s_y^{\,2}$ stets positiv ist, wird das Vorzeichen von $s_y^{\,2}\cdot\left(1-r^2\right)$ durch das Vorzeichen von $\left(1-r^2\right)$ bestimmt. Dieser Ausdruck darf nach (28) null sein, was dann der Fall ist, wenn $r^2=1$. $s_y^{\,2}\cdot\left(1-r^2\right)$ wäre negativ und damit die Bedingung verletzt, wenn $r^2 > 1$. Das heißt, *r* darf nicht größer also 1 sein und nicht kleiner als -1. Dies sind also die Grenzen, in denen sich *r* bewegen kann. Ein Wert von $r > 1$ oder $r < -1$ wäre kein zulässiger Parameter für eine Regressionsgerade. Der Korrelationskoeffizient kann nicht größer als 1 und nicht kleiner als -1 sein.

Der mögliche Fall $r^2=1$ sei weiter untersucht. Der Korrelationskoeffizient ist in diesem Fall entweder +1 oder -1. Da Bedingung (28) aus (27) abgeleitet wurde, gilt bei $s_y^{\,2}\cdot\left(1-r^2\right)=0$ hiernach auch:

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)-\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)\right]^2=0$$

Dann ist auch

$$\left(y_i-\bar{y}\right)-\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)=0$$

$$y_i-\bar{y}=\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)$$

Das aber ist die Funktion der Regressionsgeraden nach Gleichung (26), auf der alle Beobachtungspunkte liegen. Mit anderen Worten: Wenn $r^2 = 1$, also der Korrelationskoeffizient 1 oder -1 beträgt, liegen alle Beobachtungspunkte auf der Regressionsgeraden. Für jedes x_i ergibt sich das zugehörige y_i, indem x_i in die Regressionsfunktion eingesetzt wird. Die Abweichung e_i [s. Gleichung (8)] zwischen dem Beobachtungswert y_i und dem sich bei x_i auf der Regressionsgeraden ergebenden Wert $\hat{y}_i$ ist gleich null. Es gilt dann

$$e_i=y_i-\hat{y}_i=0$$
$$y_i=\hat{y}_i$$

Das heißt, bei perfekter Korrelation werden alle Beobachtungswerte y_i durch die Regressionsfunktion erklärt.

Wenn die Korrelation nicht so perfekt ist, dass alle Beobachtungswerte y_i durch die Regressionsfunktion erklärt werden können, zerlegt man die gesamte Varianz der Beobachtungswerte in einen Teil, der durch die Regressionsgerade erklärt wird und in einen restlichen Teil, der dadurch nicht erklärt wird.[2] Hierzu wird $y_i-\bar{y}$ folgendermaßen ausgedrückt:

$$y_i-\bar{y}=\left(\hat{y}_i-\bar{y}\right)+\left(y_i-\hat{y}_i\right)$$

Der Ausdruck wird quadriert:

$$\left(y_i-\bar{y}\right)^2=\left[\left(\hat{y}_i-\bar{y}\right)+\left(y_i-\hat{y}_i\right)\right]^2$$

$$\left(y_i-\bar{y}\right)^2=\left(\hat{y}_i-\bar{y}\right)^2+2\cdot\left(\hat{y}_i-\bar{y}\right)\cdot\left(y_i-\hat{y}_i\right)+\left(y_i-\hat{y}_i\right)^2$$

(29)
$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+2\cdot\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\cdot\left(y_i-\hat{y}_i\right)+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

Der in dieser Gleichung enthaltene Ausdruck $\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\cdot\left(y_i-\hat{y}_i\right)$ wird gesondert betrachtet. Hierzu ist es zunächst erforderlich, die Funktion der Regressionsgeraden für den Fall zu formulieren, dass nicht alle Beobachtungswerte auf der Regressionsgeraden liegen. Die unabhängige Variable ist dann $\hat{y}_i$. Dies statt y_i in Gleichung (26) eingesetzt:

$$\hat{y}_i-\bar{y}=\frac{s_{xy}}{s_x^{\,2}}\cdot\left(x_i-\bar{x}\right)$$

Zur Erhöhung der Übersichtlichkeit wird hierin Gleichung (21) eingesetzt, sodass:

[2] Zur Beweisführung vgl. K. Bosch, Das große Lehrbuch der Statistik, München / Wien 1996, S. 93 ff. und J. Schira, Statistische Methoden der VWL und BWL – Theorie und Praxis –, 2. Aufl. München 2005, S. 111 ff.

(30) $$\hat{y}_i - \bar{y} = b \cdot \left(x_i - \bar{x} \right)$$

Für $y_i - \hat{y}_i$ lässt sich auch schreiben

$$y_i - \hat{y}_i = \left(y_i - \bar{y} \right) - \left(\hat{y}_i - \bar{y} \right)$$

Hierin (30) eingesetzt:

(31) $$y_i - \hat{y}_i = \left(y_i - \bar{y} \right) - b \cdot \left(x_i - \bar{x} \right)$$

(30) und (31) in den zu untersuchenden Ausdruck $\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right) \cdot \left(y_i - \hat{y}_i \right)$ aus Gleichung (29) eingesetzt:

$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right) \cdot \left(y_i - \hat{y}_i \right) = \sum_{i=1}^{n} b \cdot \left(x_i - \bar{x} \right) \cdot \left[\left(y_i - \bar{y} \right) - b \cdot \left(x_i - \bar{x} \right) \right]$$

$$= \sum_{i=1}^{n} b \cdot \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) - \sum_{i=1}^{n} b^2 \cdot \left(x_i - \bar{x} \right)^2$$

$$= b \cdot n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) - b^2 \cdot n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Hierin (3) und (5) eingesetzt:

$$= b \cdot n \cdot s_{xy} - b^2 \cdot n \cdot {s_x}^2$$

Hierin (21) eingesetzt:

$$= \frac{s_{xy}}{{s_x}^2} \cdot n \cdot s_{xy} - \frac{{s_{xy}}^2}{{s_x}^4} \cdot n \cdot {s_x}^2$$

$$= \frac{{s_{xy}}^2}{{s_x}^2} \cdot n - \frac{{s_{xy}}^2}{{s_x}^2} \cdot n = 0$$

Damit wird der Ausdruck $2 \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right) \cdot \left(y_i - \hat{y}_i \right)$ in Gleichung (29) zu null, sodass

$$\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \quad | : n$$

(32) $$\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 + \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Auf der linken Seite von Gleichung (32) steht die Varianz von y_i. Der Ausdruck $\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$ wäre die Varianz von $\hat{y}_i$, wenn $\bar{y} = \bar{\hat{y}}$. Dass dies tatsächlich der Fall ist, sei im Folgenden gezeigt.

Nach Gleichung (9) gilt für die auf der Regressionsgeraden liegenden Werte $\hat{y}_i$:

$$\hat{y}_i = a + b \cdot x_i$$

Dies summiert über alle *i*:

$$\sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n}\left(a + b\cdot x_i\right) \quad |: n$$

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(a + b\cdot x_i\right)$$

Der Ausdruck auf der linken Seite ist definitionsgemäß das arithmetische Mittel von $\hat{y}_i$:

$$\bar{\hat{y}} = \frac{1}{n}\cdot\sum_{i=1}^{n}a + b\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$$

Mit $\sum_{i=1}^{n}a = n\cdot a$ und $\frac{1}{n}\cdot\sum_{i=1}^{n}x_i = \bar{x}$ ergibt sich

$$\bar{\hat{y}} = \frac{1}{n}\cdot n\cdot a + b\cdot\bar{x}$$

$$\bar{\hat{y}} = a + b\cdot\bar{x}$$

Aus (16) folgt $a + b\cdot\bar{x} = \bar{y}$. Dies eingesetzt:

(33) $$\bar{\hat{y}} = \bar{y}$$

Diese Identität in Gleichung (32) eingesetzt:

(34) $$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 + \frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Die linke Seite dieser Gleichung stellt die Varianz der Beobachtungswerte y_i dar. Der erste Summand auf der rechten Seite ist die Varianz der entsprechenden y-Werte auf der Regressionsgeraden, während der zweite Summand einfach die Summe der quadrierten Abweichungen $y_i - \hat{y}_i = e_i$ darstellt. Es liegt zwar der Gedanke nahe, dass diese Quadratsumme die Varianz der Abweichungen e_i bilden könnte, aber der Ausdruck

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}e_i^2$$

entspricht nicht der Definition einer Varianz. Hierfür wäre es erforderlich, die Differenzen von e_i zu ihrem Mittelwert zu bilden und diese zu quadrieren. Die Varianz von e_i ist dementsprechend definiert als

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2$$

Nur für $\bar{e} = 0$ wäre $\frac{1}{n}\cdot\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}e_i^2$ und damit $\frac{1}{n}\cdot\sum_{i=1}^{n}e_i^2$ auch die Varianz von e_i.

Dass dies tatsächlich der Fall ist, sei im Folgenden gezeigt. Nach Gleichung (10) gilt

$$e_i = y_i - a - b\cdot x_i$$

Es wird über alle *i* summiert:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - a - b \cdot x_i \right)$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b \cdot \sum_{i=1}^{n} x_i$$

Mit $\sum_{i=1}^{n} a = n \cdot a$ wird hieraus

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - n \cdot a - b \cdot \sum_{i=1}^{n} x_i \quad | : n$$

$$\frac{1}{n} \cdot \sum_{i=1}^{n} e_i = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - a - b \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Durch $\frac{1}{n} \cdot \sum_{i=1}^{n} e_i$, $\frac{1}{n} \cdot \sum_{i=1}^{n} y_i$ und $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ sind aber die arithmetischen Mittel $\bar{e}$, $\bar{y}$ und $\bar{x}$ definiert, sodass:

$$\bar{e} = \bar{y} - a - b \cdot \bar{x}$$

Nach Gleichung (16) gilt $\bar{y} - a - b \cdot \bar{x} = 0$ und damit in der Tat $\bar{e} = 0$. Die Varianz von e_i ist also

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \left(e_i - \bar{e} \right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Damit kann Gleichung (34) folgendermaßen interpretiert werden: Der Ausdruck auf der linken Seite ist die Gesamtvarianz der Beobachtungswerte. Diese kann zerlegt werden in die Varianz der Werte auf der Regressionsgeraden und in die Varianz der Abweichungen von der Regressionsgeraden. Setzt man für die Varianz der Beobachtungswerte wieder s_y^2, für die Varianz der Regressionswerte $s_{\hat{y}}^2$ und für die Varianz der Abweichungen s_e^2, dann besteht der Zusammenhang

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

Da $s_{\hat{y}}^2$ aus den Werten der Regressionsgeraden berechnet und damit erklärt wird, nennt man diese Größe auch erklärte Varianz. Die Varianz der Abweichungen von den Werten auf der Regressionsgeraden s_e^2 ist die nicht erklärte Varianz. Beide zusammen bilden die Gesamtvarianz s_y^2.

Teilt man Gleichung (35) durch s_y^2, wird daraus

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

$$1 = \frac{s_{\hat{y}}^2}{s_y^2} + \frac{s_e^2}{s_y^2}$$

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

Das Verhältnis der erklärten Varianz zur Gesamtvarianz ist das Bestimmtheitsmaß *B*:

(36) $$B = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

B nimmt den Wert 1 an, wenn die nicht erklärte Varianz s_e^2 null ist. Alle Beobachtungswerte werden durch die Regressionsfunktion erklärt. Dann liegen auch alle Beobachtungswerte auf der Regressionsgeraden, denn $s_e^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i^2$ kann nur dann gleich null sein, wenn alle e_i gleich null sind. Da auch alle Beobachtungswerte auf der Regressionsgeraden liegen, wenn der Korrelationskoeffizient gleich 1 ist, muss es einen Zusammenhang zwischen dem Bestimmtheitsmaß und dem Korrelationskoeffizienten geben.

B nimmt den Wert 0 an, wenn $s_e^2 = s_y^2$. Dann ist die gesamte Varianz unerklärt, es gibt also keine Korrelation. Dies äußert sich auch in einem Korrelationskoeffizienten von null.

Um den Zusammenhang zwischen dem Bestimmtheitsmaß und dem Korrelationskoeffizienten aufzuklären, wird angesetzt

(37) $$B = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}} \right)^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}$$

Nach Gleichung (30) gilt

$$\hat{y}_i - \bar{y} = b \cdot \left(x_i - \bar{x} \right)$$

Hierin die Identität (33) $\bar{y} = \bar{\hat{y}}$ eingesetzt:

$$\hat{y}_i - \bar{\hat{y}} = b \cdot \left(x_i - \bar{x} \right)$$

Dies in die Definitionsgleichung (37) eingesetzt:

$$B = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \left[b \cdot \left(x_i - \bar{x} \right) \right]^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} = \frac{b^2 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}$$

Hierin Gleichung (3) und Gleichung (4) eingesetzt:

$$B = \frac{b^2 \cdot s_x^2}{s_y^2}$$

Hierin Gleichung (21) eingesetzt:

$$B = \frac{s_{xy}^2 \cdot s_x^2}{s_x^4 \cdot s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_{xy}}{s_x \cdot s_y}$$

Hierin Gleichung (6) eingesetzt:

(38) $$B = r^2$$

Das Bestimmtheitsmaß ist also nichts anderes als das Quadrat des Korrelationskoeffizienten. Wenn der Korrelationskoeffizient gleich null ist, wird auch das Bestimmtheitsmaß null. Je stärker die Beobachtungswerte korreliert sind, desto größer wird das Bestimmtheitsmaß, gleichgültig, ob die Korrelation positiv oder negativ ist. Bei vollständiger positiver oder negativer Korrelation ist das Bestimmtheitsmaß 1.