Regression und Korrelation

Nach einem Skritp von Boris Girnat
Torsten Linnemann, April 2019

1.4 Lineare Regression

Die bisherigen statistischen Methoden haben stets nur *eine* Merkmalsausprägung für sich ins Auge gefasst. Solche Analysen nennt man *eindimensional* oder *univariat*. Werden *Zusammenhänge* zwischen *mehreren* Merkmalsausprägungen untersucht, so spricht man von *mehrdimensionalen* oder *multivariaten* Analysemethoden. In der Abbildung 1.12 wurde ein erster Schritt in diese Richtung gegangen: Man hat ein kategoriales Merkmal zur Klassierung des Datensatzes benutzt, dann die univariaten Methoden für jede Klasse getrennt angewandt und die Ergebnisse anschliessend miteinander verglichen – sei es grafisch oder sei es anhand der arithmetischen Werte der Lage- und Streuparameter. Dieses Verfahren nennt sich *Clustern* und eignet sich nur, wenn man Abhängigkeiten eines Merkmals von einem *kategorialen* Merkmal untersuchen möchte.

In diesem Abschnitt werden Methoden vorgestellt, mit denen man Zusammenhänge zwischen *metrischen* Merkmalen analysieren kann. Wir beschränken uns dabei auf zwei Merkmale, d. h. auf den *zweidimensionalen* oder *bivariaten* Fall. Ausserdem behandeln wir hier nur die Möglichkeit eines *linearen Zusammenhangs*. Dieser Spezialfall ist einerseits als realitätsbezogenen Gründen wichtig, weil solche Zusammenhänge verhältnismässig oft auftreten; andererseits ist ein linearer Zusammenhang der Fall, für den gute und trotzdem verhältnismässig einfach zu handhabende mathematische Methoden zur Verfügung stehen, die sich teilweise auch auf nicht-lineare Fälle übertragen lassen.

1.4.1 Punktwolken

Abhängigkeiten zwischen Merkmalen lassen sich oft aus dem Sachkontext heraus vermuten: Das Gewicht einer Person hängt vermutlich (auch) von ihrer Grösse ab; wer mehr verdient, hat eher

eine grössere Wohnung als einer, der wenig verdient; je stärker man ein Gewebe radioaktiv bestrahlt, desto stärker ist der Schaden. In all diesen Fällen scheint ein Zusammenhang zu bestehen. Ob der aber tatsächlich besteht und – wenn ja – ob er dann auch linear ist, soll nun mit statistischen Methoden untersucht werden. Ein erster Schritt besteht darin, sich einen grafischen Überblick über Daten zu verschaffen, die zu zwei Merkmalen gehören.

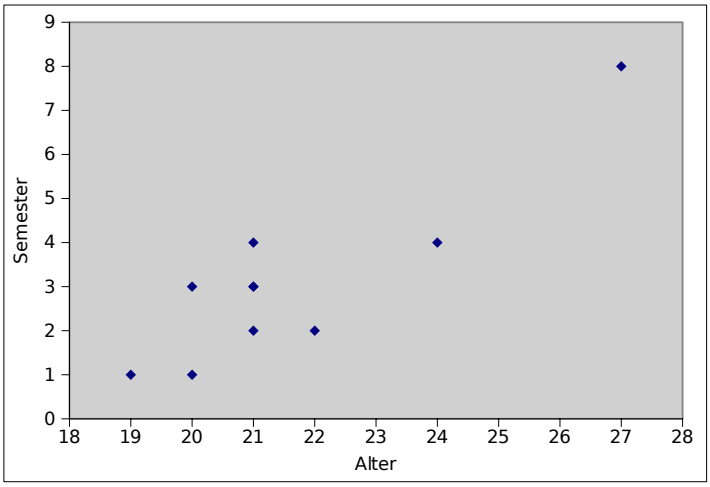
Beispiel 1.13

Wir beginnen mit einem kleinen Datensatz, in dem das Alter und die Semesterzahl einer Studentengruppe erhoben ist. Man vermutet: Je älter ein Student ist, desto höher ist sein Semester. Auch dass dieser Zusammenhang linear sein könnte, ist nicht unplausibel: Sieht man von Studienwechsel und Auslands-, Urlaubs- und Freisemestern ab, so erhöhen sich Alter und Semesterzahl proportional zueinander, sogar mit dem Proportionalitätsfaktor 1.

Alter	20	21	24	21	22	21	27	19	20	21
Semester	3	3	4	2	2	3	8	1	1	4

Wenn ein linearer Zusammenhang zwischen diesen beiden Merkmalen besteht, so könnte man die zehn Paare von Messwerten in ein Koordinatensystem einzeichnen, und die Datenpunkte müssten dann «ungefähr» auf einer Geraden liegen. Den Graphen, der aus den Datenpaaren besteht, nennt man *Punktwolke*.

Punktwolken kann man wie jede Menge reeller Zahlenpaare grafisch in einem Koordinatensystem darstellen. In der Abbildung ist die Punktwolke der beiden Merkmale aus der letzten Tabelle in einem kartesischen Koordinatensystem dargestellt.



1.4.2 Lineare Regression nach Augenmass

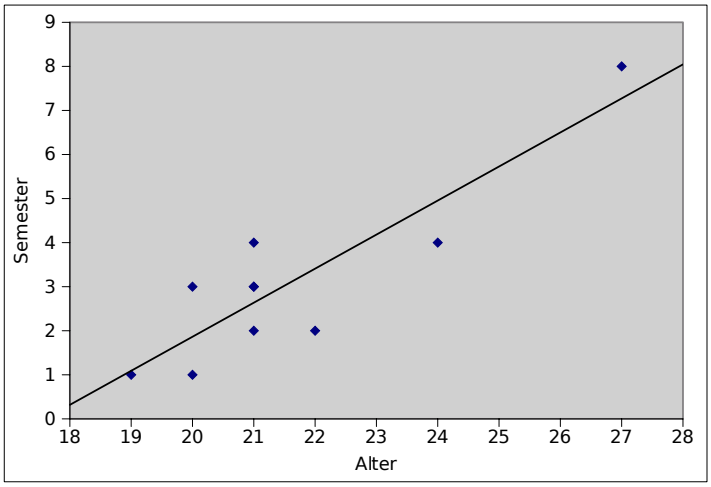
Die Punktwolke in der obigen Abbildung scheint linear anzusteigen. Eine Gerade, auf der alle Punkte liegen, wird man nicht finden – allein schon, weil dem Alter 23 drei verschiedene Semesterzahlen zugeordnet sind. Aufgabe der linearen Regression ist es, dennoch eine Gerade zu finden, welche die Punktwolke «möglichst gut» annähert, d. h. die «lineare Grundtendenz» der Punktwolke möglichst gut darzustellen, so dass die Abweichung von dieser «Tendenz» möglichst gering sind. Das Ziel ist also eine lineare Funktion y_{fit} zu finden, die zu jedem x-Wert x_i einen y-Wert $y_{\text{fit}}(x_i)$ liefert, sodass die Abweichung vom tatsächlich gemessenen y-Wert y_i für

alle Messwerte x_i möglichst gering ist. Diese Abweichung $r_i = y_i - y_{\text{fit}}(x_i)$ bezeichnet man als *Residuum*.

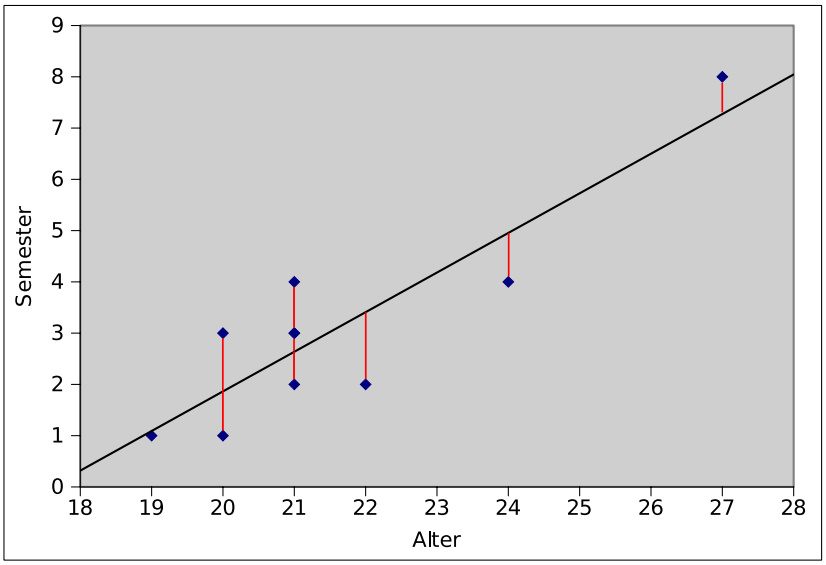
Definition 1.11

Sind $M_X = (x_1, x_2, \ldots, x_n)$ und $M_Y = (y_1, y_2, \ldots, y_n)$ die n-Tupel der Merkmalsausprägungen der Merkmale X und Y einer Stichprobe Ω und ist y_{fit} eine reelle Funktion, die X als Definitionsbereich einschliesst, so ist $r_i = y_i - y_{\text{fit}}(x_i)$ das Residuum von y_{fit} an der Stelle x_i.

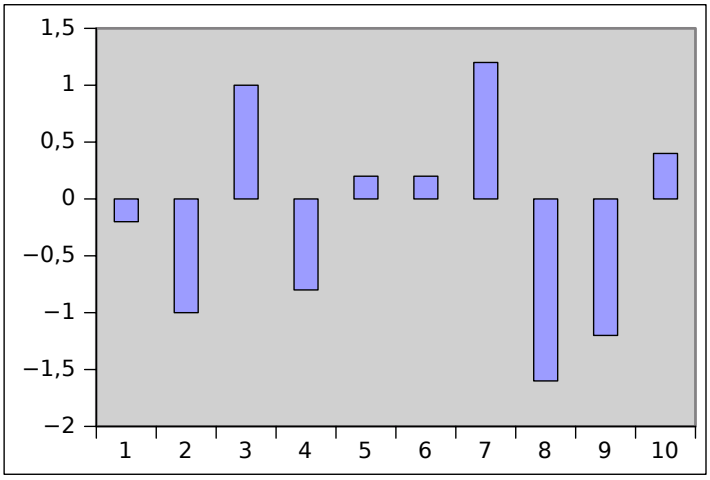
Eine einfache und für Schulzwecke oft ausreichende Methode, eine passende Regressionsfunktion y_{fit} zu finden – insbesondere wenn es sich (vermutlich) um einen linearen Zusammenhang handelt und die Regressionsfunktion eine Gerade ist – besteht darin, den Graphen von y_{fit} nach Augenmass möglichst «passend» in die Punktwolke einzuzeichnen und aus der grafischen Darstellung die Funktionsgleichung von y_{fit} abzulesen. In der Abbildung **??** ist per Augenmass eine Regressionsgerade eingezeichnet worden, so dass die Gerade möglichst dicht an den Punkten verläuft und möglichst gleichmässig Punkte ober- und unterhalb der Geraden liegen. Als Funktionsgleichung der Regressionsgeraden kann mit den üblichen Verfahren näherungsweise die Funktionsgleichung $y_{\text{fit}} = 0.8x - 14$ aus dem Schaubild ermitteln.



In der nächsten Grafik sind die Residuen $r_i = y_i - y_{\text{fit}}(x_i)$ rot eingezeichnet. Eine Möglichkeit, die Einpassung der Geraden per Augenmass zu verbessern, ist es, die Residuen als Säulendiagramm darzustellen und nach «systematischen Fehlern» zu suchen.

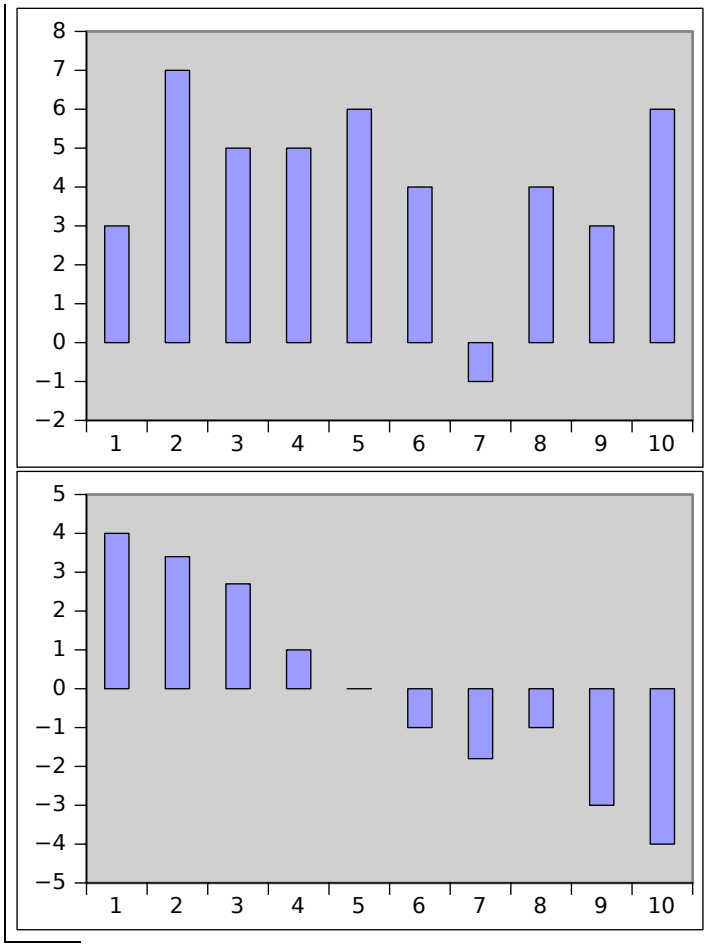


In der nächsten Abbildung sind die Residuen zur Regressionsgeraden aus der Grafik 1.4.2 aufgetragen. Man kann keinen systematischen Fehler erkennen: Die Residuen sind nicht allzu gross und verteilen sich einigermassen gleichmässig im positiven wie im negativen Bereich.



Auftrag 1.5

In den Abbildungen 1.5 und 1.5 sehen Sie Residuenplots, die auf einen systematischen Fehler beim Einzeichnen der Regressionsgeraden hindeuten. a) Beschreiben Sie, worin der Fehler besteht; b) erläutern Sie, wie sich dieser Fehler auf die Lage der Regressionsgeraden bezüglich der Punktwolke auswirkt, und c) geben Sie begründet an, wie man die Regressiongeraden verändern sollte, um den Fehler zu vermeiden, und wie sich diese Änderung in der Funktionsgleichung der Regressionsgeraden bemerkbar macht.



1.4.3 Lineare Regression mit der Methode der kleinsten Quadrate

Bei der Einpassung einer guten Regressionsgerade geht es darum, die Residuen möglichst klein zu halten. Das heisst nicht anderes, als die Summe $\sum_{i=1}^{n} |r_i|$ zu minimieren. Wenn man sich nicht auf das Augenmass verlassen kann oder will, sondern wenn man einer rechnerische Lösung sucht, die man insbesondere auch einem Computer anvertrauen kann, dann ist die Forderung, die Summe $\sum_{i=1}^{n} |r_i|$ zu minimieren, problematisch. Minimierungsprobleme lassen sich oft mit Methoden der Analysis lösen. Dass in der Summe Beträge auftreten, verhindert eine Anwendung des Ableitungskalküls. Auch andere brauchbare Lösungen ohne analytische Methoden hat man nicht gefunden. Aus diesem Grunde hat man sich dafür entschieden, statt der Summe $\sum_{i=1}^{n} |r_i|$ die Summe $\sum_{i=1}^{n} r_i^2$ zu minimieren, also nicht die Summe der absoluten Abstände der Datenpunkte zur Geraden, sondern die Summe der Quadrate der Abstände zur Geraden, die man auch schon für die Minimalitätseigenschaft des arithmetischen Mittels in Abschnitt 1.3.3 betrachtet hat. Die Minimierung dieser Summe ist über den Ableitungskalkül möglich und führt sogar für jeden Datensatz zu einer eindeutigen Lösung.

Satz 1.4

Ist $P = \{(x_i, y_i) \in \mathbb{R}^2 \mid 1 \leq i \leq n\}$ *eine Punktwolke, so ist*

$$y = a \cdot x + b$$

genau dann eine Regressionsgerade, welche die Summe

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

minimiert, wenn

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sigma_X^2} = \frac{Cov(X,Y)}{\sigma_X^2}$$

und

$$b = \overline{y} - a\overline{x}$$

gilt.

Der Beweis benötigt Mittel aus der Differentialrechnung und findet sich beispielsweise in Kütting (2011).

Im Nenner des Terms, der die Steigung einer optimalen Regressionsgeraden nach der Methode der kleinsten Quadrate ausdrückt, tritt die empirische Varianz des Merkmals X auf. Im Zähler steht ein Ausdruck, der eine hohe Ähnlichkeit mit der Varianz hat, nur dass beide Merkmale X und Y darin «verarbeitet» werden. Diesen Ausdruck nennt man *Kovarianz* von X und Y.

Definition 1.12

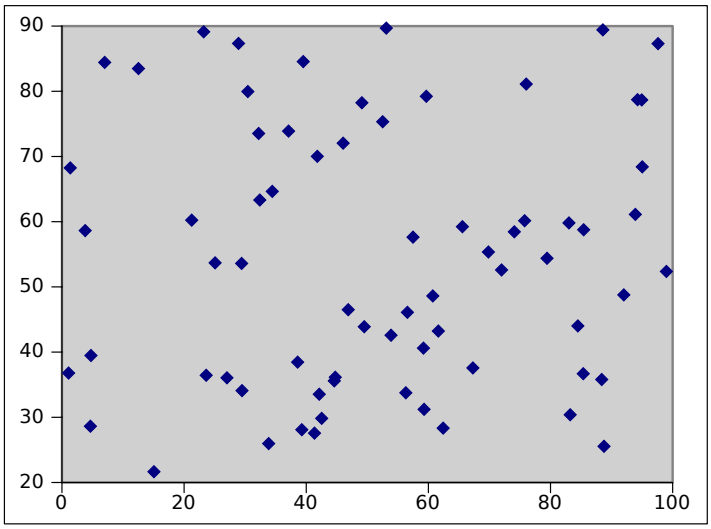
Es seien $X = (x_1, x_2, \ldots, x_n)$ *und* $Y = (y_1, y_2, \ldots, y_n)$ *metrische Merkmale, dann heisst*

$$Cov(X,Y) = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

die Kovarianz von X *und* Y.

1.4.4 Korrelationskoeffizienten

Mit der Methode der kleinsten Quadrate hat man ein Verfahren, das für jeden Datensatz mit zwei numerischen Merkmalen eine optimale Regressionsgerade angibt. So schön es ist, eine algorithmische Lösung für dieses Problem zu haben, so fraglich ist es, ob diese Lösung in jedem Fall brauchbar ist. In der Abbildung 1.4.4 ist eine Punktwolke eingezeichnet, in der kein linearer Zusammenhang (und auch nicht irgendein anderer) zwischen den x- und y-Werten erkennbar ist. Auch für diese Punktwolke würde die Methode der kleinsten Quadrate eine Regressionsgerade ermitteln. Die Summe der Residuenquadrate wäre zwar minimal, aber trotzdem noch so hoch, dass die Abweichung von der Geraden im allgemeinen sehr gross ist. *Jede* Gerade wäre ungeeignet, die Punktwolke anzunähern, auch die Regressionsgerade nach der Methode der kleinsten Quadrate.



Erforderlich wäre also eine Entscheidung darüber, ob zwischen zwei Merkmalen überhaupt ein linearer Zusammenhang besteht. Hier wäre eine Ja-Nein-Entscheidung allerdings ungeeignet, da selbst «sehr schön lineare» Punktwolken wie jene in der Abbildung 1.4.1 nicht vollkommen auf einer Geraden liegen. Sinnvoller wäre ein *Mass*, das *Grade der Linearität* angibt. Ein solches Mass nennt man *Korrelationskoeffizienten*. Für die Interpretation dieses Masses wird folgende Konvention getroffen:

Definition 1.13

Es seien $X = (x_1, x_2, \ldots, x_n)$ und $Y = (y_1, y_2, \ldots, y_n)$ metrische Merkmale. Eine Funktion $r(X, Y)$ heisst Korrelationskoeffizient, wenn r die folgenden Eigenschaften besitzt:

1) $r(X, Y) = r(Y, X)$,
2) $r(X, X) = 1$,
3) $r(X, -X) = -1$.

Man kann leicht zeigen, dass r nur Werte aus dem Intervall $[-1, 1]$ annimmt. Mit der zweiten und dritten Eigenschaft der Definition wird die Interpretation von r deutlich: X hängt von X klarerweise perfekt linear ab; und ebenso hängt auch X von $-X$ perfekt linear ab; die Regressionsgerade hätte allerdings eine negative statt einer positiven Steigung. Dies gilt allgemein: Je dichter r bei 1 oder -1 liegt, desto linearer ist der Zusammenhang; und je dichter r bei 0 liegt, desto geringer ist der lineare Zusammenhang. Im Falle $r = 0$ besteht überhaupt kein linearer Zusammenhang, so wie in den Grenzfällen 1 und -1 ein perfekter linearer Zusammenhang besteht. Zusätzlich gibt das Vorzeichen des Korrelationskoeffizienten das Vorzeichen einer optimalen Regressionsgerade an. Die Interpretation zwischen den Extremfällen $-1, 0$ und 1 unterliegt einer gewissen Willkür. Üblich, aber nicht zwingend ist es, die Grenzen folgendermassen zu ziehen:

Wertebereich von r	Interpretation
$-1 \leq r \leq -0.7$	starker negativer linearer Zusammenhang
$-0.7 < r < -0.3$	Grauzone
$-0.3 \leq r \leq 0.3$	kein linearer Zusammenhang
$0.3 < r < 0.7$	Grauzone
$0.7 \leq r \leq 1$	starker positiver linearer Zusammenhang

1.4.6 Der Korrelationskoeffizient nach Pearson

Wie Sie eben selbst nachgewiesen haben, weist der resistente Korrelationskoeffizient unter Umständen auch dann einen linearen Zusammenhang aus, wenn gar keiner vorliegt. Das ist wenig erfreulich. Aus diesem Grunde werden lieber andere Korrelationskoeffizienten benutzt, die in der Regel aber einen höheren Rechenaufwand erfordern und weniger anschaulich zu begründen sind. Am gebräuchlichsten ist der Korrelationskoeffizient von Pearson.

Definition 1.15

Es seien $X = (x_1, x_2, \ldots, x_n)$ und $Y = (y_1, y_2, \ldots, y_n)$ metrische Merkmale. Dann ist

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

der Korrelationskoeffizient nach Pearson.

Warum das eine sinnvolle Definition für einen Korrelationskoeffizient ist, lässt sich am besten aus Sicht der analytischen Geometrie verstehen. Betrachten wir die beiden Vektoren

$$\vec{x} = \begin{pmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \text{ und } \vec{y} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix}$$

Diese Vektoren beinhalten als Komponenten die Abweichungen der einzelnen Messwerte vom arithmetischen Mittel. Verwendet man das Skalarprodukt und die euklidische Norm, so sieht der Pearsonsche Korrelationskoeffizient in vektorieller Schreibweise folgendermassen aus:

$$r = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||}$$

Über diesen Term werden überlicherweise Winkel definiert. Der Winkel $\sphericalangle$ zwischen $\vec{x}$ und $\vec{y}$ ist

$$\sphericalangle(\vec{x}, \vec{y}) = \arccos\left(\frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||}\right) = \arccos(r)$$

Nun wird deutlich: Wenn r nahe bei 1 liegt, dann schliessen $\vec{x}$ und $\vec{y}$ nahezu einen Winkel von $0°$ ein, d. h. die Änderungen von X und Y «laufen» nahezu in «dieselbe Richtung», und da der Winkel nicht von der Länge von $\vec{x}$ und $\vec{y}$ abhängt, sind die beiden Änderungen sogar proportional zueinander, d. h. ein linearer Zusammenhang liegt vor. Bei r nahe bei Null liegt der Winkel zwischen $vecx$ und $\vec{y}$ nahe bei $90°$, d. h. wenn X sich in die einer Richtung ändert, ändert Y sich in eine «ganz andere» Richtung als X, d. h. es liegt kein Zusammenhang zwischen X und Y vor, erst recht kein linearer. Im Fall, dass r nahe bei -1 liegt, ist der Winkel zwischen $\vec{x}$ und $\vec{y}$ ungefähr $180°$, d. h. wenn sich X verändert, verändert sich Y genau in die entgegengesetzte Richtung, und zwar proportional zu X. Also liegt dann ein negativer linearer Zusammenhang vor.