

Kovarianz und Korrelation

Verstehen, wie zwei Variablen gemeinsam variieren

Ratiomera Statistics

Inhaltsverzeichnis

Zusammenfassung der Kursseite	2
Von gepaarten Abweichungen zur Korrelation	2
Was du vor der Interpretation von r prüfst	2
Die Brücke zur Regression	3
Erweiterte Referenz	4
Zweck und Grundlagen	4
Zentrale Ideen	4
Formelleitfaden	5
Die erklärende Abbildung lesen	7
Checkliste zur Interpretation	7
Verbindung zu anderen Themen	8

Zusammenfassung der Kursseite

Kovarianz und Korrelation beschreiben, wie zwei quantitative Variablen über korrekt gepaarte Fälle gemeinsam variieren. Betrachte immer zuerst das Streudiagramm. Ein Koeffizient fasst dieses Diagramm zusammen, ersetzt es aber nicht.

Von gepaarten Abweichungen zur Korrelation

Für gepaarte Beobachtungen (x_i, y_i) lautet die Stichprobenkovarianz

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

Produkte sind positiv, wenn beide Abweichungen in dieselbe Richtung zeigen, und negativ, wenn sie in entgegengesetzte Richtungen zeigen. Das Vorzeichen beschreibt deshalb die Richtung. Die Grösse hängt von den Messeinheiten ab, wodurch Kovarianzen über Skalen hinweg schwer vergleichbar sind.

Die Pearson-Korrelation standardisiert die Kovarianz:

$$r = \frac{s_{XY}}{s_X s_Y}.$$

Sie liegt zwischen -1 und $+1$. Das Vorzeichen gibt die Richtung des linearen Musters an. $|r|$ beschreibt, wie eng die Punkte einer Geraden folgen. Ein Wert nahe null bedeutet wenig linearen Zusammenhang. Ein stark gekrümmtes Muster kann trotzdem eine kleine Pearson-Korrelation haben.

Was du vor der Interpretation von r prüfst

Tabelle 1: Lies Datenmuster und Studiendesign, bevor du den Koeffizienten liest.

Prüfung	Weshalb sie wichtig ist
Gepaarte Zeilen	Jeder x_i -Wert muss mit seinem eigenen y_i -Wert verbunden bleiben
Form im Streudiagramm	Pearson- r fasst linearen Zusammenhang zusammen, nicht jede Abhängigkeitsform
Ungewöhnliche und einflussreiche Punkte	Ein einzelner Punkt kann den Koeffizienten stark verändern
Beobachteter Wertebereich	Eine Einschränkung des Bereichs kann den beobachteten Koeffizienten abschwächen
Untergruppen	Ein Gesamtmuster kann unterschiedliche Muster innerhalb von Gruppen verdecken
Messung und Stichprobe	Messfehler, Selektion und ungenügende Abdeckung können den Zusammenhang verzerren

Die Spearman-Rangkorrelation wendet die Pearson-Idee auf Ränge an. Sie beschreibt ein monotones Muster, bei dem sich eine Variable tendenziell durchgehend nach oben oder unten bewegt, wenn sich die andere verändert. Eine Gerade ist nicht erforderlich. Gebundene Werte erhalten passende gemeinsame Ränge.

Für die Inferenz über eine Populationskorrelation nach Pearson gilt unter dem hier verwendeten Modell

$$t = r \sqrt{\frac{n-2}{1-r^2}}, \quad df = n - 2.$$

Der Test fragt, ob die lineare Populationskorrelation unter den Modellannahmen null sein könnte. Er macht aus einem beobachtenden Zusammenhang keinen kausalen Effekt.

Die Brücke zur Regression

Korrelation ist symmetrisch und einheitenfrei. Die einfache Regression behält dasselbe gepaarte Muster bei, gibt den Variablen aber verschiedene Rollen: Eine wird Prädiktor, die andere Ergebnisvariable. Thema 5 macht aus der Punktwolke eine angepasste Gerade, stellt die ursprünglichen Einheiten wieder her und untersucht die Residuen, die zeigen, was die Gerade nicht erfasst.

Erweiterte Referenz

Zweck und Grundlagen

Kovarianz und Korrelation beschreiben, wie zwei numerische Variablen bei denselben Fällen gemeinsam variieren. Jeder Fall muss ein zusammengehöriges Wertepaar (x_i, y_i) beitragen. Ein Streudiagramm ist der unverzichtbare Ausgangspunkt: Lege eine Variable auf die horizontale und die andere auf die vertikale Achse und stelle jeden Fall als einen Punkt dar. Die entstehende Punktwolke zeigt Richtung, Form, Stärke, Häufungen und ungewöhnliche Beobachtungen auf eine Weise, die ein einzelner Koeffizient nicht leisten kann.

Ein positiver Zusammenhang bedeutet, dass Fälle mit grösseren x -Werten tendenziell auch grössere y -Werte besitzen. Ein negativer Zusammenhang bedeutet, dass grössere x -Werte tendenziell mit kleineren y -Werten einhergehen. Ein Wert nahe null bei einem linearen Koeffizienten bedeutet wenig linearen Zusammenhang, aber nicht zwingend überhaupt keinen Zusammenhang. Gekrümmte Muster, getrennte Untergruppen oder ein eingeschränkter Wertebereich können einen Koeffizienten unvollständig oder irreführend machen.

Die Kovarianz beginnt mit den Abweichungen von den beiden Mittelwerten. Ein Fall trägt positiv bei, wenn beide Werte über ihren Mittelwerten oder beide darunter liegen. Er trägt negativ bei, wenn ein Wert über und der andere unter seinem Mittelwert liegt. Der Mittelwert dieser Kreuzprodukte ergibt die Stichprobenkovarianz. Ihr Vorzeichen ist aufschlussreich, ihr Betrag hängt jedoch von den Masseinheiten ab. Werden Stunden beispielsweise in Minuten gemessen, verändert sich die Kovarianz, obwohl die zugrunde liegende Paarung der Fälle unverändert bleibt.

Merkmal	Kovarianz	Pearson-Korrelation
Richtung	Das Vorzeichen zeigt positive oder negative gemeinsame Variation	Das Vorzeichen zeigt einen positiven oder negativen linearen Zusammenhang
Skala	Hängt von den Einheiten beider Variablen ab	Einheitenlos, weil beide Variablen standardisiert werden
Wertebereich	Nicht auf ein festes Intervall beschränkt	Liegt immer zwischen -1 und 1
Hauptaufgabe	Baustein für Zusammenhang und Regression	Vergleichbarer Kennwert für lineare Richtung und Stärke

Zentrale Ideen

Die Pearson-Korrelation r standardisiert die Kovarianz, indem sie durch die beiden Stichprobenstandardabweichungen geteilt wird. Ein Wert nahe 1 zeigt, dass die Punkte einem starken positiven geradlinigen Muster folgen. Ein Wert nahe -1 zeigt ein starkes negatives geradliniges Muster. Ein Wert nahe null zeigt wenig geradliniges Muster. Der Koeffizient beschreibt die Stichprobe. Die Populationskorrelation wird gewöhnlich als ρ geschrieben. Wenn das Ziel eine Aussage über die Grundgesamtheit ist, wird Inferenz benötigt.

Die Spearman-Rangkorrelation ersetzt die beobachteten Werte durch ihre Ränge und beurteilt, ob die Variablen einer **monotonen** Beziehung folgen. Monoton bedeutet, dass sich die Tendenz durchgehend in eine Richtung bewegt: Wenn eine Variable zunimmt, nimmt die andere im Allgemeinen ebenfalls zu oder im Allgemeinen ab. Das Muster darf dabei gekrümmt sein, solange diese Reihenfolge erhalten bleibt. Die Spearman-Korrelation kann deshalb bei einem gekrümmten monotonen Zusammenhang hoch bleiben, den die Pearson-Korrelation weniger vollständig zusammenfasst. Keiner der beiden Koeffizienten stellt eine U-förmige Beziehung gut dar, weil sich die Richtung über den Wertebereich umkehrt.

Diagnosefrage	Was du untersuchst	Weshalb es die Interpretation verändert
Ist die Form annähernd linear?	Streudiagramm und mögliches geglättetes Muster	Pearson- r fasst eine geradlinige Tendenz zusammen
Sind ungewöhnliche Punkte einflussreich?	Beschriftetes Streudiagramm und Empfindlichkeitsvergleich	Ein weit entfernter Punkt kann Richtung oder Betrag verändern
Werden Gruppen vermischt?	Farben oder getrennte Felder für inhaltlich bedeutsame Gruppen	Ein zusammengefasster Zusammenhang kann sich von den Zusammenhängen innerhalb der Gruppen unterscheiden
Ist der beobachtete Bereich eingeschränkt?	Wertebereiche der Variablen und Stichprobenverfahren	Begrenzte Variation kann einen beobachteten Koeffizienten abschwächen
Ist die Paarung gültig?	Fallkennungen und Messzeitpunkte	Bei einer Korrelation müssen sich beide Werte auf denselben Fall beziehen

Korrelation belegt keine Kausalität. Ein beobachteter Zusammenhang kann einen direkten Einfluss, die umgekehrte Einflussrichtung, eine mit beiden Variablen verbundene Drittvariable, die Auswahl in die Stichprobe, Messartefakte oder Zufall widerspiegeln. Zeitliche Reihenfolge und ein glaubwürdiges Forschungsdesign liefern Informationen, die ein Koeffizient allein nicht geben kann. Auch wenn keine kausale Aussage beabsichtigt ist, entscheidet der inhaltliche Kontext darüber, ob die gepaarten Variablen und ihre Interpretation sinnvoll sind.

Wenn eine Aussage über die Grundgesamtheit beabsichtigt ist, kann der Stichprobenkoeffizient gegen $H_0 : \rho = 0$ getestet werden. Halte diesen Test von Betrag und praktischer Bedeutung getrennt. Ein kleiner p-Wert betrifft unter dem Modell die Vereinbarkeit mit einer Populationskorrelation von null. Er macht den Zusammenhang nicht automatisch gross, wichtig oder kausal.

Formelleitfaden

Die Stichprobenkovarianz bildet den Mittelwert gepaarter Kreuzprodukte von Abweichungen und verwendet $n - 1$ im Nenner:

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

Das Vorzeichen jedes Produkts zeigt, ob die beiden Abweichungen in dieselbe oder in entgegengesetzte Richtungen weisen. Grosse Abweichungen erhalten mehr Gewicht, weil ihr Produkt einen grösseren Betrag besitzt.

Die Pearson-Stichprobenkorrelation teilt die Kovarianz durch das Produkt der beiden Stichprobenstandardabweichungen:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Wenn korrigierte Summen vorliegen, kann derselbe Koeffizient direkt berechnet werden als

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Beide Darstellungen benötigen in den beiden Variablen dieselben gepaarten Fälle. Wird eine Spalte unabhängig von ihrer Partnerin umgeordnet, werden die Paare zerstört und die Fragestellung verändert.

Dieselbe Berechnung kann als Summe der Produkte standardisierter Werte geschrieben werden. Dadurch wird sichtbar, weshalb der Koeffizient einheitenlos ist:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n z_{xi} z_{yi}$$

In einer Stichprobe ohne verbundene Ränge kann die Spearman-Rangkorrelation aus der Differenz d_i zwischen den beiden Rängen jedes Falles berechnet werden:

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

Bei verbundenen Rängen berechnest du stattdessen die Pearson-Korrelation der zugewiesenen Ränge. In jeder Form muss der Koeffizient zusammen mit dem Streudiagramm oder Rangmuster gelesen werden, aus dem er entstanden ist.

Um die Pearson-Korrelation der Grundgesamtheit gegen $H_0 : \rho = 0$ zu testen, wird

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad df = n - 2.$$

Die Alternativhypothese bestimmt, ob die Referenzfläche einseitig oder zweiseitig ist. Die Berechnung setzt unabhängige gepaarte Fälle, einen Zusammenhang, für den eine lineare Pearson-Zusammenfassung vertretbar ist, sowie das Fehlen von Designproblemen oder einflussreichen Punkten voraus, welche die Interpretation ungültig machen würden.

Zu berichtendes Ergebnis	Beantwortete Frage	Was es allein nicht belegen kann
Form des Streudiagramms	Welches Muster, welche Gruppen, welcher Wertebereich und welche ungewöhnlichen Punkte sind sichtbar?	Eine Aussage über die Grundgesamtheit
r oder r_s	Welche Richtung und lineare oder monotone Stärke zeigt sich in der Stichprobe?	Kausalität
t , df und p-Wert	Wie vereinbar ist der Pearson-Koeffizient der Stichprobe unter dem Modell mit $\rho = 0$?	Praktische Bedeutung oder ein grosser Zusammenhang

Die erklärende Abbildung lesen

Ein Streudiagramm zeigt, was ein einzelner Koeffizient übersehen kann

Spearman folgt der monotonen Rangfolge;
eine U-Form wird von keinem der beiden Koeffizienten gut zusammengefasst

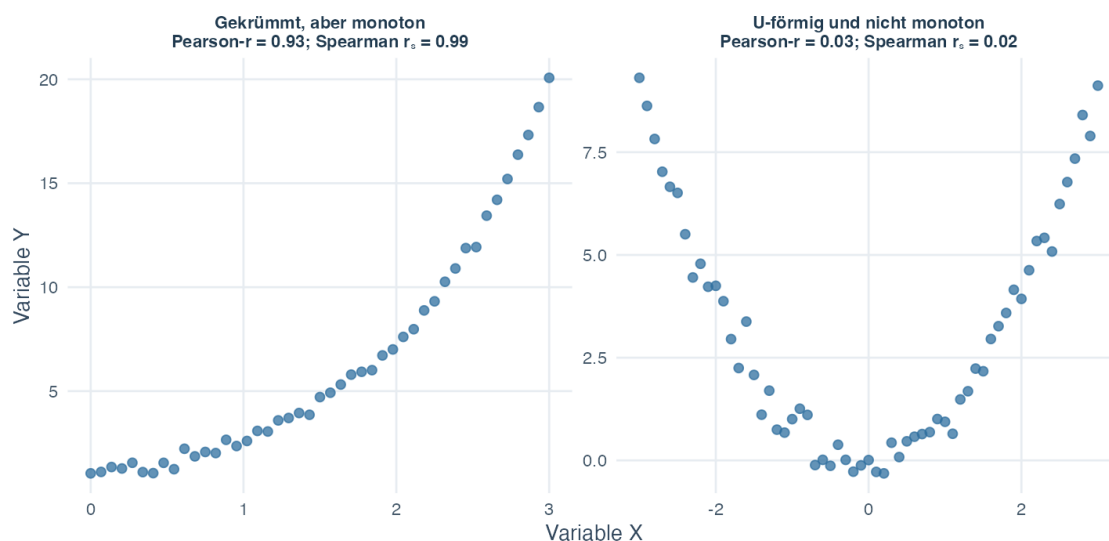


Abbildung 1: Zwei Streudiagramme vergleichen ein gekrümmtes monotonen Muster mit einem U-förmigen Muster und geben über jedem Feld die Pearson- und Spearman-Korrelation an.

Das linke Feld steigt über den gesamten beobachteten Bereich an. Der Anstieg ist gekrümmt und nicht geradlinig, doch die Reihenfolge ist sehr beständig: Grössere x -Werte gehen fast immer mit grösseren y -Werten einher. Die Spearman-Korrelation liegt deshalb nahe bei eins, weil die Ränge diese aufsteigende Reihenfolge bewahren. Auch die Pearson-Korrelation ist stark positiv, konzentriert sich aber weiterhin auf den gesamten geradlinigen Anteil des Musters. Die sichtbare Krümmung zeigt dir, dass eine geradlinige Zusammenfassung nicht jedes Merkmal erfasst.

Das rechte Feld ist U-förmig. Von ganz links zur Mitte nimmt y ab, während x zunimmt. Von der Mitte nach ganz rechts nimmt y zu, während x zunimmt. Diese entgegengesetzten Richtungen heben sich sowohl bei der Pearson- als auch bei der Spearman-Berechnung auf und erzeugen Werte nahe null. Trotzdem besteht eine ausgeprägte Beziehung zwischen den Variablen. Die richtige Schlussfolgerung lautet nicht «kein Zusammenhang». Stattdessen fasst weder ein linearer noch ein monotoner Koeffizient diese Form gut zusammen.

Die Verbindungslinien in der Abbildung helfen, die Reihenfolge der Punkte zu erkennen. Es handelt sich nicht um angepasste Regressionsgeraden. Die über den Feldern angegebenen Koeffizienten beschreiben die dargestellten simulierten Werte. Sie sind Lehrresultate und keine Schätzungen von realen Teilnehmenden. Das Beispiel erinnert daran, zuerst die Form in der Grafik zu bestimmen und danach einen Koeffizienten jenes Merkmal zusammenfassen zu lassen, für das er entwickelt wurde.

Checkliste zur Interpretation

Bestätige, dass die Variablen numerisch sind oder dass eine rangbasierte Analyse angemessen ist. Prüfe, ob die Werte nach Fall gepaart sind. Untersuche vor der Berechnung eines Koeffizienten ein Streudiagramm. Beschreibe Richtung, Form, Stärke, Häufungen, Wertebereich und ungewöhnliche Punkte. Wähle die Pearson-Korrelation für eine inhaltlich sinnvolle lineare Zusammenfassung und die Spearman-Korrelation für eine monotone rangbasierte Zusammenfassung. Berichte Stichprobengrösse und Koeffizient. Ergänze ein Intervall oder einen Test, wenn eine Aussage über die Grundgesamtheit erforderlich ist.

Verwende keine festen allgemeingültigen Grenzwerte, um einen Zusammenhang ohne Kontext als schwach, mittel oder stark zu bezeichnen. Die praktische Bedeutung eines Koeffizienten hängt von Messzuverlässigkeit, Fachgebiet, Design und Konsequenzen ab. Prüfe, wie sich das Ergebnis verändert, wenn ein einflussreicher Punkt oder eine inhaltlich bedeutsame Untergruppe untersucht wird. Entferne jedoch keine Fälle mit dem einzigen Ziel, den Koeffizienten zu verbessern. Vermeide kausale Verben, wenn das Design sie nicht stützt.

Verbindung zu anderen Themen

Die Kovarianz bildet die Brücke von der deskriptiven Streuung zur Regression. In der einfachen linearen Regression kann die Steigung als Kovarianz zwischen Prädiktor und Zielvariable geteilt durch die Varianz des Prädiktors geschrieben werden. Die Korrelation standardisiert dieselbe gepaarte Tendenz. Die Regression behält dagegen die Einheit der Zielvariable und gibt an, wie stark sich ihr angepasster Wert bei einer Veränderung des Prädiktors um eine Einheit verändert.

Die partielle Korrelation fragt später, wie zwei Variablen nach der linearen Bereinigung um eine dritte Variable zusammenhängen. Die multiple Regression erweitert dieselbe Logik, indem sie den bedingten Zusammenhang jedes Prädiktors schätzt, während die anderen konstant gehalten werden. Auch die Varianzanalyse gehört zu dieser Familie: Sie erklärt Variation in einer Zielvariable mithilfe der Gruppenzugehörigkeit, statt mit einem numerischen Prädiktor zu beginnen. Die gemeinsame Frage lautet, wie die Variation einer Zielvariable mit den Informationen eines oder mehrerer Prädiktoren zusammenpasst.

Dokument-ID: topic-04-covariance-correlation-summary-de

Notizen 1 von 3

Kovarianz und Korrelation | Thema 4

Notizen 2 von 3

Kovarianz und Korrelation | Thema 4

Notizen 3 von 3

Kovarianz und Korrelation | Thema 4