Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Rangkorrelation nach Spearman

Produkt-Moment-Korrelation

- setzt Intervallskalenniveau voraus
- unterschätzt den Zusammenhang von x und y, wenn dieser nichtlinear ist
- reagiert empfindlich auf Ausreißerwerte

Rangkorrelation r_s

- eignet sich für ordinalskalierte (quasi-)stetige Merkmale
- geeignet bei monotonem, aber nichtlinearem Zusammenhang von x und y
- geeignet bei Verdacht auf Ausreißerwerte

Rangkorrelation nach Spearman

r_s ist definiert als die Korrelation der Rangplätze der Messwerte

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}$$

Dabei ist $R_i = rg(x_i)$ der Rang von x_i (bzgl. aller erhaltenen Werte für das Merkmal X) und entsprechend $S_i = rg(y_i)$ der Rang von y_i (bzgl. aller erhaltenen Werte für das Merkmal Y)

$\bar{R}$ und $\bar{S}$ sind die Mittelwerte der Ränge R_i und S_i mit $i = 1, \ldots, n$

Rangkorrelation nach Spearman

Wegen der Gleichung

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

gilt für die Mittelwerte der Ränge

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{n+1}{2}$$

und

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{n+1}{2}$$

Rangkorrelation nach Spearman

Benutzt man diese Beziehungen, so erhält man nach einigen elementaren Rechenschritten

$$r_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}$$

$$\vdots$$

$$= 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n\,(n^2 - 1)}$$

Rangplätze

r_s ist definiert als die Korrelation der Rangplätze der Messwerte

Beispiel

Messwerte	x_i	33	15	17	11	40	22
	y_i	47	32	29	28	56	38
Rangplätze	$R_i = rg(x_i)$	5	2	3	1	6	4
	$S_i = rg(y_i)$	5	3	2	1	6	4

$$r_s = 1 - \frac{6 \cdot [0^2 + (-1)^2 + 1^2 + 0^2 + 0^2 + 0^2]}{6 \cdot (36 - 1)} = 0.94 \quad (r = 0.98)$$

Rangplätze

Sowohl bei den x- bzw. y-Werten als auch den Rängen können identische Werte auftreten

- In diesem Fall spricht man von *Bindungen* ("ties")
- Die Problematik der Rangvergabe wird mit Durchschnittsbildung über die in Frage kommenden Ränge gelöst, d. h. es wird jedem der entsprechenden Werte der durchschnittliche Rang zugeordnet
- Beispiel

Messwerte	x_i	1.09	2.17	2.17	2.17	3.02	4.5
Rangplätze	$R_i = rg(x_i)$	1	3	3	3	5	6

Rangkorrelation: Berechnung in R

Mit dem Argument `method = "spearman"` in der Funktion `cor()` wird die Spearman-Rangkorrelation berechnet

```
x <- c(33, 15, 17, 11, 40, 22)
y <- c(47, 32, 29, 28, 56, 38)
cor(x, y, method = "spearman")     # 0.943
```

Berechnung der Ränge

```
rank(x)
[1] 5 2 3 1 6 4
rank(y)
[1] 5 3 2 1 6 4
```

Durchschnittsbildung bei Ties

```
c(1.09, 2.17, 2.17, 2.17, 3.02, 4.5) |>
  rank(ties.method = "average")
[1] 1 3 3 3 5 6
```

φ-Koeffizient

Allgemeiner Fall

	x_1	x_2	$\sum$
y_1	a	b	$A = a + b$
y_2	c	d	$B = c + d$
$\sum$	$C = a + c$	$D = b + d$	n

Der Zusammenhang zwischen den Merkmalen X und Y lässt sich in diesem Fall durch die Vierfelder-Korrelation φ beschreiben

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{A \cdot B \cdot C \cdot D}}$$

φ-Koeffizient

Beispiel

		Erkrankung ja	nein	$\sum$
Impfung	ja	16	174	190
	nein	121	689	810
	$\sum$	137	863	1000

$$\varphi = \frac{16 \cdot 689 - 121 \cdot 174}{\sqrt{190 \cdot 810 \cdot 863 \cdot 137}} = -0.074$$

Odds Ratio

Als *Odds* (Wettchance) bezeichnet man einen Quotienten der Art

$$\frac{\frac{a}{A}}{\frac{b}{A}} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}$$

Für das Beispiel gibt diese Größe das Verhältnis der relativen Häufigkeit der Erkrankung vs. Nicht-Erkrankung für den Fall an, dass man geimpft ist

Den Quotienten

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

bezeichnet man als *Odds Ratio* (auch relative Chance, Chancen- oder Kreuzproduktverhältnis)

Odds Ratio

Für das Beispiel berechnet man

$$\frac{a \cdot d}{b \cdot c} = \frac{16 \cdot 689}{174 \cdot 121} = 0.52$$

Interpretation

- Wenn man geimpft ist, dann ist das Risiko (die Chance) zu erkranken nur halb so groß

Den Begriffen Odds und Odds Ratio werden wir im Kontext wahrscheinlichkeitstheoretischer Betrachtungen in der Psychologie noch häufiger begegnen

Künstliche Dichotomie

Merkmale mit zwei Ausprägungen können entweder natürliche Dichotomien beschreiben (z. B. Geschlecht) oder aus der dichotomen Kategorisierung (quasi-)stetiger Merkmale resultieren

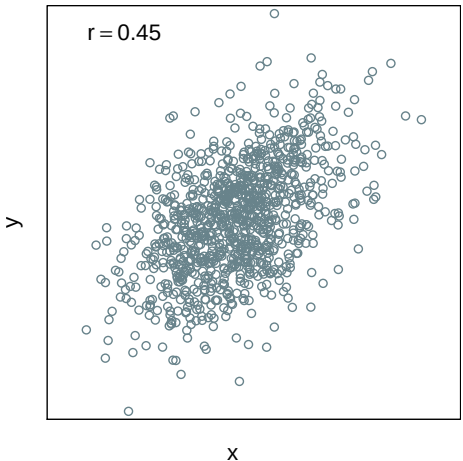
Im letzteren Fall spricht man auch von künstlichen Dichotomien

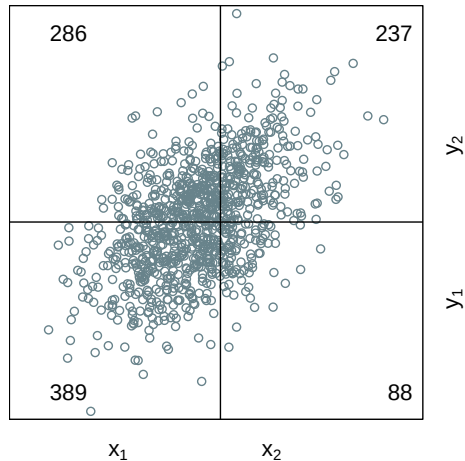
Beispiele

- Dichotomisierung des Alters in Kategorien "jung" vs. "alt"
- Dichotomisierung des Einkommens in Kategorien "arm" vs. "wohlhabend"

Berechnung der Vierfelder-Korrelation φ (manifeste Korrelation) führt im Allgemeinen zu einer Unterschätzung der tatsächlich vorliegenden Produkt-Moment-Korrelation (latente Korrelation)

Künstliche Dichotomie





Künstliche Dichotomie

	x_1	x_2	$\sum$
y_1	389	88	477
y_2	286	237	523
$\sum$	675	325	1000

$$\varphi = \frac{389 \cdot 237 - 88 \cdot 286}{\sqrt{477 \cdot 523 \cdot 675 \cdot 325}} = 0.29$$

Tetrachorische Korrelation

Um diesen Mangel auszugleichen, wurden verschiedene Näherungsformeln zur Schätzung der latenten Produkt-Moment-Korrelation aus der Vierfeldertafel vorgeschlagen, die als tetrachorische Korrelation bezeichnet werden

Die folgende Formel beruht auf dem Odds Ratio OR

$$r_{tet} \approx \frac{OR^{0.74} - 1}{OR^{0.74} + 1}$$

Beispiel

$$OR = \frac{389 \cdot 237}{88 \cdot 286} = 3.66, \quad r_{tet} \approx \frac{3.66^{0.74} - 1}{3.66^{0.74} + 1} = 0.45$$

Natürliche und künstliche Dichotomie

Repräsentiert das Merkmal X der Vierfeldertafel eine natürliche Dichotomie und das Merkmal Y eine künstliche Dichotomie, so ist zur Schätzung der latenten Korrelation folgende Formel vorgeschlagen worden

$$\nu = \frac{\log(OR)}{\sqrt{[\log(OR)]^2 + \frac{2.89}{\frac{C}{n} \cdot \frac{D}{n}}}}$$

Beachte: Die Randsummen C und D beziehen sich auf die durch das Merkmal X repräsentierte natürliche Dichotomie

Regressionsrechnung

Bei der Korrelationsrechnung steht der (ungerichtete) Zusammenhang zwischen zwei Merkmalen im Vordergrund

Mit der Regressionsrechnung soll, auf der Grundlage eines funktionalen Zusammenhangs, die Vorhersage eines Merkmals durch ein anderes ermöglicht werden

Im Allgemeinen wird ein Kriterium (Regressand, Zielgröße), nachfolgend durch Y bezeichnet, durch einen Prädiktor (Regressor, Einflussgröße) vorhergesagt, nachfolgend durch X bezeichnet

Beispiele

Empirische Überprüfung der Abhängigkeit der Längung einer Federwaage durch eine Masse

- Idealerweise ist dieser Zusammenhang linear (Proportionalität, Hooksches Gesetz)

Empirische Überprüfung der Abhängigkeit des Bremswegs eines Autos von der Geschwindigkeit

- Idealerweise kann dieser Zusammenhang durch eine Parabel beschrieben werden

In der konkreten empirischen Situation werden aber mehr oder weniger große Abweichungen davon auftreten (Messungenauigkeiten, Zustand der Bremsen oder des Straßenbelags, . . .)

Kausalität?

Da man in diesen physikalischen Beispielen von der Existenz eines deterministischen funktionalen Zusammenhangs ausgehen kann (jedenfalls idealerweise nach Ausschluss möglicher Fehlerquellen), scheint mit der Regression ein Zugriff auf Kausalbeziehungen möglich

In der Psychologie ist dies aber sicher nicht der Regelfall

- Häufig werden nicht alle relevanten Einflussgrößen als Prädiktoren erfasst sein und der statistische Zusammenhang der Größen wird deutlich schwächer ausfallen als in den genannten physikalischen Anwendungen

Kausalität ist also nicht mit einem bestimmten statistischen Verfahren verbunden, sondern resultiert aus dem unterliegenden Mechanismus; ihr empirischer Nachweis hängt davon ab, wie gut es gelingt, die untersuchten Größen von Fehlerquellen zu isolieren

Problemstellung

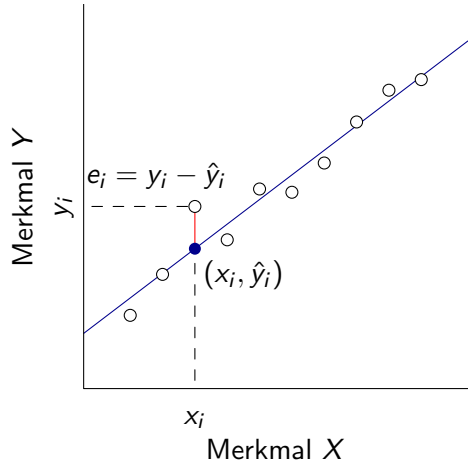
An die zweidimensionale Punktewolke

$$(x_1, y_1), \ldots, (x_n, y_n)$$

soll möglichst gut eine Gerade

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

angepasst werden



Dabei bezeichnen der Parameter a den Ordinatenabschnitt (Schnittpunkt mit der y-Achse) und der Parameter b die Steigung der Geraden

Bestimmung der Geradengleichung

Für gegebene Stichprobenwerte

$$(x_1, y_1), \ldots, (x_n, y_n)$$

gilt es, die Parameter a und b geeignet zu wählen

Die Werte $\hat{a}$ und $\hat{b}$ werden dabei so bestimmt, dass die Quadratsumme der Residuen e_i einen minimalen Wert annimmt, dass also

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}\, x_i)^2 = \min.$$

Kleinst-Quadrat-Kriterium

Diese Vorgehensweise wird als Bestimmung der Geradengleichung nach dem *(gewöhnlichen) Kleinst-Quadrat-Kriterium* (ordinary least squares criterion) bezeichnet (Prinzip stammt von C. F. Gauss, 1777-1855)

- Maximiert die Korrelation zwischen den beobachteten Werten y und den vorhergesagten Werten $\hat{y}$

Die Quadratsumme der Residuen

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

wird als Funktion der beiden Parameter a und b aufgefasst

Bestimmung der Geradengleichung

Bei einem Minimum dieser Funktion müssen die (partiellen) Ableitungen von $Q(a, b)$ nach a bzw. b verschwinden

Daher wird die Funktion $Q(a, b)$ (partiell) nach ihren Argumenten a und b abgeleitet

$$\frac{\partial Q(a,b)}{\partial a} \qquad \frac{\partial Q(a,b)}{\partial b}$$

Die resultierenden Ausdrücke werden dann gleich Null gesetzt, wodurch sich zwei Gleichungen in zwei Unbekannten ergeben, die es zu lösen gilt

Bestimmung der Geradengleichung

Es ist

$$\frac{\partial Q(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - b\,x_i)$$

$$\frac{\partial Q(a,b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - b\,x_i)\,x_i$$

Die beiden Gleichungen sind nun gleich Null zu setzen und nach den Parametern a und b aufzulösen

Bestimmung der Geradengleichung

Für die erste Gleichung ergibt sich

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

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

$$n\,\bar{y} - n\,\hat{a} - \hat{b}\,n\,\bar{x} = 0$$

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

Bestimmung der Geradengleichung

Für die zweite Gleichung ergibt sich

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

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

$$n\,\hat{a}\,\bar{x} + \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Bestimmung der Geradengleichung

Einsetzen von

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

liefert dann die Gleichung

$$n\,\bar{x}\,(\bar{y} - \hat{b}\,\bar{x}) + \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

die nun nach $\hat{b}$ aufzulösen ist

Bestimmung der Geradengleichung

Daraus ergibt sich

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \\ &= \frac{n\,\tilde{s}_{xy}}{n\,\tilde{s}_x^2} \\ &= \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}\end{aligned}$$

Dabei bezeichnet

- $\tilde{s}_{xy}$ die empirische Kovarianz zwischen x- und y-Werten
- $\tilde{s}_x^2$ die empirische Varianz der x-Werte

Bestimmung der Geradengleichung

Für die empirische Regressionsgerade erhält man dann

$$\hat{y}_i = \bar{y} + \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}(x_i - \bar{x})$$

Für die Regressionsgerade bedeutet das

- sie verläuft durch den Punkt $(\bar{x}, \bar{y})$
- sie besitzt die Steigung

$$\hat{b} = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$$

Beispiel: Noten Statistik und Forschungsmethoden (s. o.)

Mit

$$\bar{x} = \frac{12}{5} = 2.4, \ \tilde{s}_x = 1.020$$
$$\bar{y} = \frac{16}{5} = 3.2, \ \tilde{s}_y = 0.748$$

und der Kovarianz $\tilde{s}_{xy} = 0.72$ ergibt sich

$$\hat{b} = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2} = \frac{0.72}{1.020^2} \approx 0.69$$

und

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 3.20 - 0.69 \cdot 2.40 \approx 1.54$$

