Korrelation und lineare Regression

Lektion B11 – BSc Wahrscheinlichkeit und Statistik

Digital Finance

Gliederung

1. Streudiagramm
2. Kovarianz
3. Pearson-Korrelation
4. Spearman-Rangkorrelation
5. Lineare Regression (OLS)
6. Regressionsdiagnostik
7. R-Code
8. Zusammenfassung

Lernziele

Am Ende dieser Lektion werden Sie in der Lage sein:

1. Ein Streudiagramm zu erstellen und Zusammenhangsrichtung/-stärke abzulesen
2. Die Kovarianz $\mathrm{Cov}(X, Y)$ zu berechnen und ihre Grenzen zu benennen
3. Den Pearson-Korrelationskoeffizienten r zu berechnen und zu interpretieren
4. Den Spearman-Rangkorrelationskoeffizienten r_s anzuwenden und von Pearson abzugrenzen
5. Eine einfache lineare Regression (OLS) durchzuführen und a, b, R^2 zu interpretieren
6. Residuenplots und Q-Q-Plots für die Regressionsdiagnostik zu nutzen
7. Die Aussage "Korrelation $\neq$ Kausalität" zu erklären und zu begründen

Diese Ziele leiten, was Sie aus dieser Lektion beherrschen sollten.

Streudiagramm (Scatter Plot)

Idee: Wir tragen für jedes Beobachtungspaar (x_i, y_i) einen Punkt in ein Koordinatensystem ein.

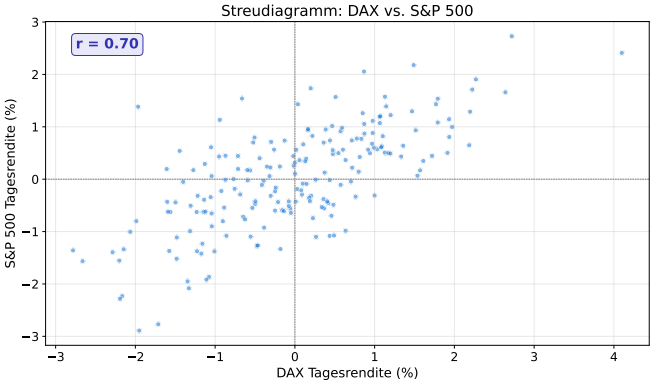
Was kann man ablesen?

- **Richtung:** positiv (steigend), negativ (fallend), kein Trend
- **Stärke:** eng beieinander = starker Zusammenhang
- **Form:** linear oder nicht-linear (Kurve, Cluster)
- **Ausreißer:** Punkte, die stark abweichen

Finanzbeispiel: DAX-Tagesrendite vs. S&P 500-Tagesrendite – erwartet man einen positiven Zusammenhang?

Das Streudiagramm ist immer der erste Schritt vor jeder Korrelations- oder Regressionsanalyse.

Beispiel: DAX vs. S&P 500 Tagesrenditen



Beobachtungen:

- Die Punkte bilden eine **Ellipse** mit positiver Steigung
- Je stärker die Korrelation, desto schmaler die Ellipse
- Einzelne Ausreißer (z. B. extreme Handelstage) sind sichtbar

Globale Aktienmärkte bewegen sich oft in die gleiche Richtung – positive Korrelation.

Kovarianz $\mathrm{Cov}(X, Y)$

Definition: Die Kovarianz misst die gemeinsame Variabilität zweier Zufallsvariablen:

$$\boxed{\mathrm{Cov}(X, Y) = E\big[(X - \mu_X)(Y - \mu_Y)\big] = E[XY] - E[X] \cdot E[Y]}$$

Interpretation des Vorzeichens:

- $\mathrm{Cov}(X, Y) > 0$: X und Y tendieren in die **gleiche** Richtung
- $\mathrm{Cov}(X, Y) < 0$: X und Y tendieren in **entgegengesetzte** Richtungen
- $\mathrm{Cov}(X, Y) = 0$: kein *linearer* Zusammenhang

Problem: Die Kovarianz hängt von den **Maßeinheiten** ab!

Beispiel: $\mathrm{Cov}(\text{DAX}, \text{S\&P 500}) = 0{,}0003$ – ist das viel oder wenig?
$\Rightarrow$ Wir brauchen eine **normierte** Größe: den Korrelationskoeffizienten.

Die Kovarianz zeigt die Richtung, aber nicht die Stärke des Zusammenhangs.

Pearson-Korrelationskoeffizient r

Definition (Bravais-Pearson):

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

Eigenschaften:

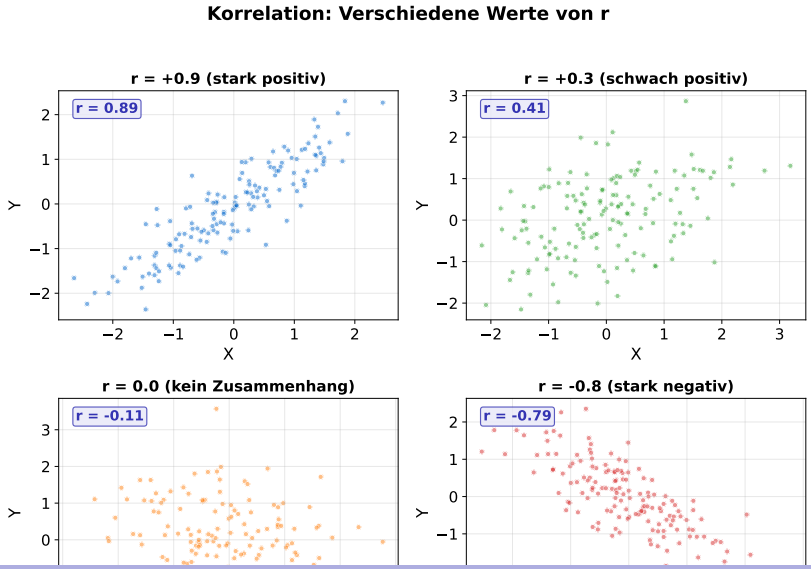
1. $-1 \leq r \leq +1$ (dimensionslos, normiert)
2. $r = +1$: perfekter **positiver** linearer Zusammenhang
3. $r = -1$: perfekter **negativer** linearer Zusammenhang
4. $r = 0$: kein linearer Zusammenhang

Achtung: r misst nur den *linearen* Zusammenhang! Ein starker nicht-linearer Zusammenhang kann $r \approx 0$ ergeben.

---

r ist die "normierte Kovarianz" – unabhängig von den Masseinheiten.

Korrelation: Verschiedene Werte von r



Finanzbeispiel: Gold vs. USD/CHF

Historische Beobachtung: Gold und der US-Dollar (gemessen in CHF) zeigen oft eine **negative Korrelation**.

Warum?

- Gold wird in USD gehandelt – steigt der USD, wird Gold für andere Währungen teurer
- In Krisenzeiten: Flucht in Gold *und* CHF (beides "sichere Häfen")
- $r \approx -0{,}4$ bis $-0{,}6$ (variiert je nach Zeitraum)

Portfolio-Konsequenz:
Negative Korrelation ermöglicht **Diversifikation**:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

Bei $\mathrm{Cov}(X, Y) < 0$ wird das Portfolio-Risiko **reduziert**!

Diversifikation funktioniert, weil verschiedene Anlagen nicht perfekt korreliert sind.

Spearman-Rangkorrelation r_s

Idee: Statt der Originalwerte verwende die Ränge (Rangnummern).

Vorgehen:

1. Ersetze x_i durch Rang $\mathrm{rg}(x_i)$ und y_i durch Rang $\mathrm{rg}(y_i)$
2. Berechne Pearson-Korrelation der Ränge

Alternative Formel (bei keinen Bindungen):

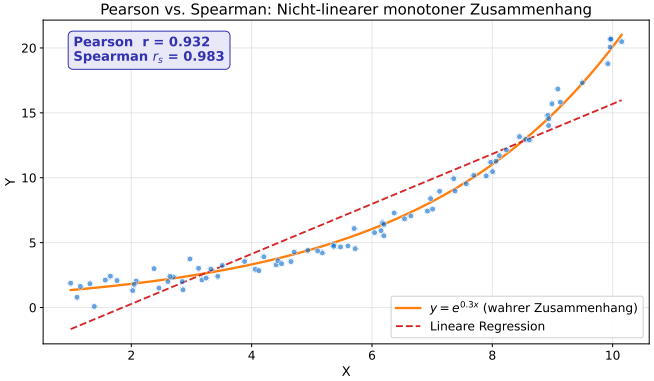
$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad d_i = \mathrm{rg}(x_i) - \mathrm{rg}(y_i)$$

Wann Spearman statt Pearson?

- Daten auf **Ordinalniveau** (Schätznoten, Rankings)
- **Nicht-linearer**, aber *monotoner* Zusammenhang (z. B. exponentiell)
- **Ausreißer** vorhanden (Ränge sind robuster)

Spearman misst monotonen Zusammenhang – Pearson nur linearen.

Pearson vs. Spearman: Wann unterscheiden sie sich?



Beobachtung: Bei einem **nicht-linearen monotonen** Zusammenhang (hier: exponentiell) gilt:

- Pearson r ist deutlich kleiner als 1 (misst nur Linearität)
- Spearman $r_s \approx 1$ (erkennt die monotone Beziehung)

Wenn die Beziehung monoton, aber nicht linear ist, ist Spearman die bessere Wahl.

Einfache lineare Regression

Modell: Wir suchen die "beste" Gerade durch die Punktwolke:

$$\boxed{Y = a + bX + \varepsilon}$$

- $a =$ **Achsenabschnitt** (Intercept): Wert von Y, wenn $X = 0$
- $b =$ **Steigung** (Slope): Änderung von Y pro Einheit X
- $\varepsilon =$ **Fehlerterm** (Residuum): nicht erklärte Abweichung

Methode der kleinsten Quadrate (OLS):
Minimiere die Summe der quadrierten Abweichungen:

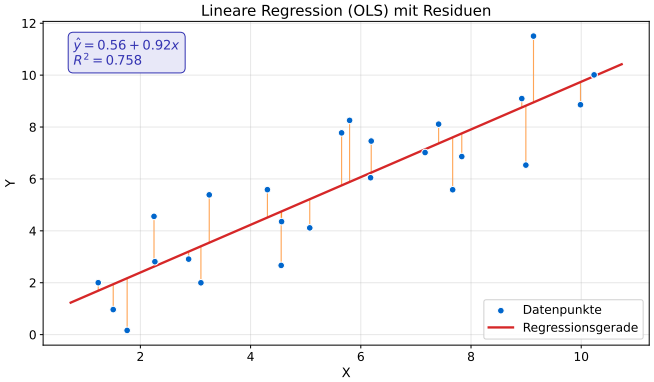
$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Lösung:

$$\boxed{b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}}, \qquad a = \bar{y} - b\bar{x}$$

OLS = "Ordinary Least Squares" – die Standardmethode zur Schätzung linearer Modelle.

Regressionsgerade und Residuen



Interpretation:

- Die rote Linie ist die geschätzte Regressionsgerade $\hat{y} = a + bx$
- Die vertikalen Linien zeigen die Residuen $e_i = y_i - \hat{y}_i$
- OLS minimiert die **Summe der quadrierten** Residuen

Die Regressionsgerade geht immer durch den Punkt $(\bar{x}, \bar{y})$.

Bestimmtheitsmass R^2

Frage: Wie gut erklärt die Regression die Daten?

Zerlegung der Gesamtvariation:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{SST (Total)}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression)}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE (Residuen)}}$$

Definition:

$$\boxed{R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}}$$

Interpretation:

- $R^2 \in [0, 1]$: Anteil der erklärten Varianz
- $R^2 = 0{,}85$: 85 % der Variation in Y wird durch X erklärt
- Bei einfacher Regression: $R^2 = r^2$ (Quadrat des Korrelationskoeffizienten!)

R^2 beantwortet: "Wie viel Prozent der Streuung erklärt das Modell?"

Finanzbeispiel: CAPM-β

Capital Asset Pricing Model (CAPM):

$$R_{\text{Aktie}} - R_f = \alpha + \beta \cdot (R_{\text{Markt}} - R_f) + \varepsilon$$

Interpretation:

- $\beta =$ **Steigung** der Regression: Sensitivität der Aktie zum Markt
- $\beta > 1$: Aktie schwankt stärker als der Markt (aggressiv)
- $\beta < 1$: Aktie schwankt weniger als der Markt (defensiv)
- $\beta = 1$: Aktie bewegt sich wie der Markt
- α: Überrendite ("Jensen's Alpha")

Beispiel:
$\beta_{\text{Novartis}} \approx 0{,}7$ (defensiv), $\beta_{\text{UBS}} \approx 1{,}3$ (aggressiv)
$R^2 \approx 0{,}35$: Der Markt erklärt 35 % der Renditevarianz.

Das CAPM-β ist die bekannteste Anwendung der linearen Regression in der Finanzwelt.

Die vier Annahmen der linearen Regression

Damit OLS zuverlässige Ergebnisse liefert, müssen gelten:

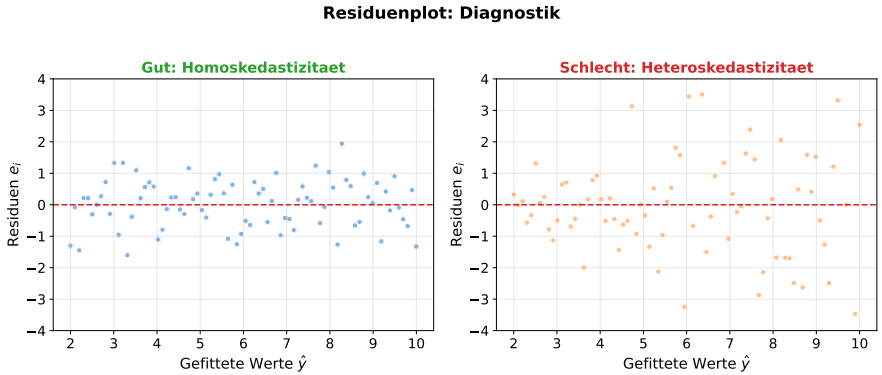
1. **Linearität:** $E[Y|X] = a + bX$ (der wahre Zusammenhang ist linear)
2. **Homoskedastizität:** $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (konstante Fehlervarianz)
3. **Normalverteilung:** $\varepsilon_i \sim N(0, \sigma^2)$ (für Konfidenzintervalle/Tests)
4. **Unabhängigkeit:** ε_i sind untereinander unabhängig (keine Autokorrelation)

Wie prüft man das?

- Annahmen 1 & 2: **Residuenplot** (Residuen vs. gefittete Werte)
- Annahme 3: **Q-Q-Plot** (Quantile der Residuen vs. Normalverteilung)
- Annahme 4: Durbin-Watson-Test, Autokorrelationsplot (nicht in dieser Lektion)

Vor jeder Interpretation der Regression: erst die Annahmen prüfen!

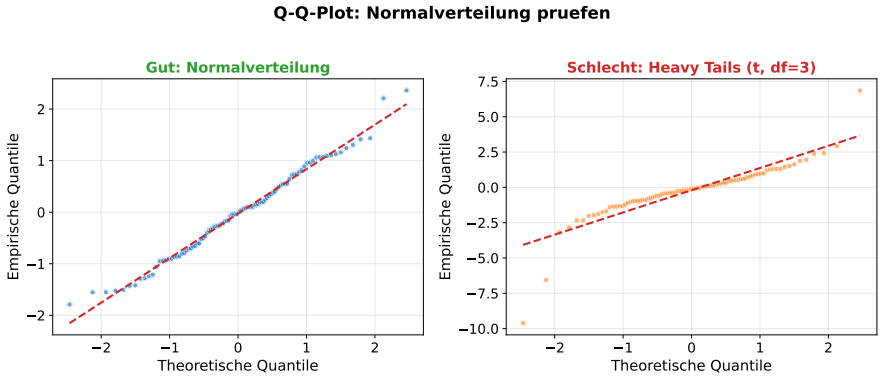
Residuenplot: Gut vs. Schlecht



Links (gut): Residuen zufällig um 0 verteilt, keine Muster $\Rightarrow$ Annahmen erfüllt.
Rechts (schlecht): Trichterform (Heteroskedastizität) – die Streuung wächst mit $\hat{y}$.

Ein guter Residuenplot zeigt "zufälliges Rauschen" – kein Muster, kein Trichter.

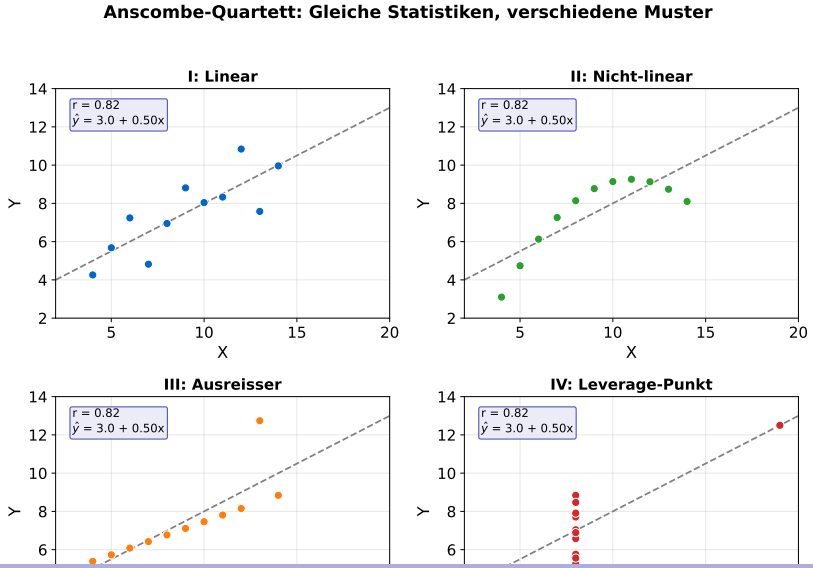
Q-Q-Plot: Normalverteilung prüfen



Links (gut): Punkte liegen auf der Diagonalen ⇒ Residuen sind normalverteilt.
Rechts (schlecht): Abweichungen an den Enden ⇒ schwere Schwanztails (heavy tails), typisch für Finanzdaten.

Der Q-Q-Plot vergleicht die empirischen Quantile mit den theoretischen Normalquantilen.

Anscombe-Quartett: Warum Visualisierung wichtig ist



R-Code: Korrelation berechnen

```r
# Daten einlesen
x <- c(1.2, 2.3, 3.1, 4.0, 5.2, 6.1, 7.3, 8.0)
y <- c(2.1, 3.5, 4.2, 5.8, 6.3, 7.9, 8.5, 10.1)

# Pearson-Korrelation
cor(x, y)                      # r = 0.993

# Spearman-Rangkorrelation
cor(x, y, method = "spearman")  # r_s = 1.0

# Korrelationstest (H0: rho = 0)
cor.test(x, y)
# t = 25.43, df = 6, p-value = 2.1e-07
# 95%-KI: [0.965, 0.999]
```

Interpretation:

- `cor()` berechnet r oder r_s (je nach `method`)
- `cor.test()` liefert p-Wert und Konfidenzintervall für ρ
- $p < 0{,}001$: Die Korrelation ist statistisch hochsignifikant

---

`cor.test()` **testet** $H_0: \rho = 0$ **gegen** $H_1: \rho \neq 0$.

R-Code: Lineare Regression

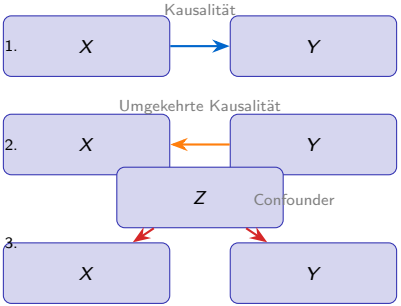
```r
# Lineares Modell schaetzen
modell <- lm(y ~ x)

# Ergebnisse anzeigen
summary(modell)
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   0.5738     0.2551   2.249   0.0656
# x             1.1464     0.0451  25.430  2.11e-07 ***
# R-squared: 0.9909

# Konfidenzintervalle fuer Koeffizienten
confint(modell, level = 0.95)

# Diagnostik-Plots
par(mfrow = c(2, 2))
plot(modell)
```

Ablesen: $\hat{y} = 0{,}57 + 1{,}15x$, $R^2 = 0{,}99$.
`plot(modell)` erzeugt automatisch: Residuen vs. Fitted, Q-Q-Plot, Scale-Location, Leverage.

`lm(y ~ x)` **ist der zentrale R-Befehl fuer lineare Regression.**

Korrelation ≠ Kausalität

Zentraler Grundsatz: Aus einer Korrelation allein kann man **keine** Ursache-Wirkungs-Beziehung ableiten!



Eisverkauf ↔ Ertrinkungsfälle
(Confounder: Temperatur)

Schuhgröße ↔ Lesekompetenz
(Confounder: Alter)

Nur ein kontrolliertes Experiment oder eine kausale Analyse kann Kausalität belegen.

Übersicht: Korrelation und Regression

	Kovarianz	Pearson r	Spearman r_s
Formel	$E[(X-\mu_X)(Y-\mu_Y)]$	$\frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$	Pearson auf Rängen
Wertebereich	$(-\infty, +\infty)$	$[-1, +1]$	$[-1, +1]$
Misst	Richtung	linearen Zusammenhang	monotonen Zusammenhang
Robust?	Nein	Nein	Ja (gegen Ausreißer)

Regression zusätzlich:

- OLS: $\hat{y} = a + bx$ mit $b = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$
- $R^2 = r^2$: Anteil der erklärten Varianz
- Diagnostik: Residuenplot + Q-Q-Plot

Korrelation beschreibt den Zusammenhang, Regression modelliert ihn quantitativ.

Lernziele – Checkliste

Prüfen Sie sich selbst:

- ✓ **Streudiagramm:** Richtung, Stärke, Form, Ausreißer ablesen
- ✓ **Kovarianz:** $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$, Vorzeichen interpretieren
- ✓ **Pearson r:** $r = \mathrm{Cov}/(\sigma_X \sigma_Y)$, $r \in [-1, +1]$, nur linearer Zusammenhang
- ✓ **Spearman r_S:** Rangkorrelation, robust, monotoner Zusammenhang
- ✓ **OLS:** $b = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$, $a = \bar{y} - b\bar{x}$
- ✓ **R^2:** Anteil erklärter Varianz, $R^2 = r^2$
- ✓ **Diagnostik:** Residuenplot (Muster?), Q-Q-Plot (Normalverteilung?)
- ✓ **Korrelation $\neq$ Kausalität:** Scheinkorrelation, Confounder, umgekehrte Kausalität

---

Wenn Sie alle Punkte verstanden haben, beherrschen Sie die Grundlagen der Regressionsanalyse!

Schlüsselformeln

Formel	Name
$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$	Kovarianz
$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$	Pearson-Korrelation
$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$	Spearman-Rangkorrelation
$b = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \quad a = \bar{y} - b\bar{x}$	OLS-Koeffizienten
$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = r^2$	Bestimmtheitsmass
$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$	Varianz einer Summe

Diese Formeln sind das Fundament der bivariaten Datenanalyse.