Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Zweidimensionale Häufigkeitsverteilungen

Im Folgenden betrachten wir zu jeder Untersuchungseinheit $i = 1, \ldots, n$ zwei Beobachtungswerte x_i und y_i

Die Urliste lautet in diesem Fall dann

$$(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$$

Beispiele

- (IQ, Hirngewicht)
- (Ausbildungsgrad, Einkommen)
- (Mathematiknote, Deutschnote)
- (Zigarettenkonsum, Lebenserwartung)
- (Körpergröße des Vaters, Körpergröße des Sohnes)
- (Alter des Ehemanns, Alter der Ehefrau)

Zweidimensionale Häufigkeitsverteilungen

Frage

- Was ist für jedes der Beispiele die jeweilige Untersuchungseinheit?

Die Darstellung der entsprechenden Verteilungen hängt wiederum von der Art der betrachteten Merkmale ab

- Quantitative versus qualitative Merkmale
- Bei quantitativen Merkmalen
  - Diskrete versus stetige/quasi-stetige Merkmale

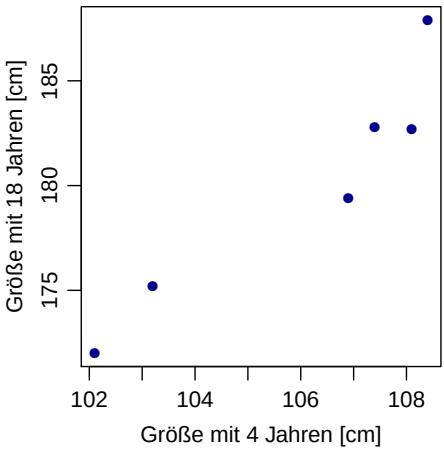
Fall 1: Beide Merkmale sind stetig oder quasi-stetig

Beispiel

X Körpergröße mit 4 Jahren
Y Körpergröße mit 18 Jahren

i	x_i (cm)	y_i (cm)
1	108.1	182.7
2	102.1	172.0
3	107.4	182.8
4	106.9	179.4
5	103.2	175.2
6	108.4	187.9

Streudiagramm (scatter plot)




```
dat <- read.table(header = TRUE, text = "
  vp     x     y
   1 108.1 182.7
   2 102.1 172.0
   3 107.4 182.8
   4 106.9 179.4
   5 103.2 175.2
   6 108.4 187.9
")

plot(
  y ~ x, dat, pch = 16,
  xlab = "Groesse mit 4 Jahren [cm]",
  ylab = "Groesse mit 18 Jahren [cm]"
)
```

Fall 2: Beide Merkmale sind diskret

x Merkmal A

$a_1, \ldots, a_j, \ldots, a_m$

y Merkmal B

$b_1, \ldots, b_k, \ldots, b_l$

Wir betrachten also $m \cdot l$ Merkmalskombinationen

$$(a_j, b_k) \text{ für alle } j = 1, \ldots, m \text{ und } k = 1, \ldots, l$$

und die entsprechenden absoluten Häufigkeiten

$$h_{jk} = h(a_j, b_k) \text{ für alle } j = 1, \ldots, m \text{ und } k = 1, \ldots, l$$

Kontingenztafel der absoluten Häufigkeiten

		B						
		b_1	b_2	$\cdots$	b_k	$\cdots$	b_l	$\sum$
	a_1	h_{11}	h_{12}	$\cdots$	h_{1k}	$\cdots$	h_{1l}	$h_{1.}$
	a_2	h_{21}	h_{22}	$\cdots$	h_{2k}	$\cdots$	h_{2l}	$h_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A	a_j	h_{j1}	h_{j2}	$\cdots$	h_{jk}	$\cdots$	h_{jl}	$h_{j.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a_m	h_{m1}	h_{m2}	$\cdots$	h_{mk}	$\cdots$	h_{ml}	$h_{m.}$
	$\sum$	$h_{.1}$	$h_{.2}$	$\cdots$	$h_{.k}$	$\cdots$	$h_{.l}$	$h_{..} = n$

Kontingenztafel der absoluten Häufigkeiten

Beispiel: X Note in Statistik 1, Y Note in Forschungsmethoden

		Y					
		1	2	3	4	5	$\sum$
	1	2	3	2	1	0	8
	2	6	9	5	3	0	23
X	3	2	12	14	8	2	38
	4	0	6	8	9	3	26
	5	2	1	1	1	0	5
	$\sum$	12	31	30	22	5	100

Kontingenztafel der relativen Häufigkeiten

In der Kontingenztafel der absoluten Häufigkeiten werden alle Einträge durch n geteilt

Damit erhält man

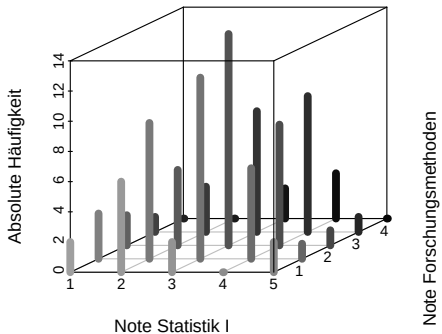
$$f_{jk} = \frac{h_{jk}}{n}$$

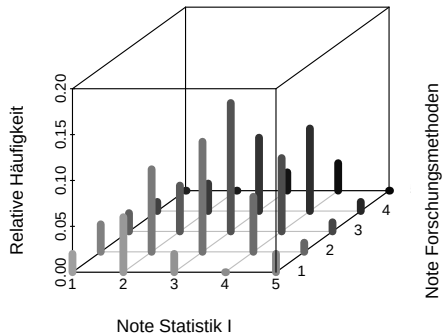
Beispiel: X Note in Statistik 1, Y Note in Forschungsmethoden

		Y					
		1	2	3	4	5	$\sum$
	1	0.02	0.03	0.02	0.01	0.00	0.08
	2	0.06	0.09	0.05	0.03	0.00	0.23
X	3	0.02	0.12	0.14	0.08	0.02	0.38
	4	0.00	0.06	0.08	0.09	0.03	0.26
	5	0.02	0.01	0.01	0.01	0.00	0.05
	$\sum$	0.12	0.31	0.30	0.22	0.05	1.00

Dreidimensionales Stabdiagramm

Beispiel: Grafische Darstellung von Kontingenztafeln absoluter bzw. relativer Häufigkeiten als 3D-Stabdiagramm





Kontingenztafeln und 3D-Stabdiagramm in R

```
library(scatterplot3d)  # Zusatzpaket laden

h <- c(2,3,2,1,0,6,9,5,3,0,2,12,14,8,2,0,6,8,9,3,2,1,1,1,0)

tab <- matrix(h, nrow = 5, ncol = 5, byrow = TRUE,        # abs Hfkt
              dimnames = list(stat = 1:5, forsch = 1:5))
tab <- tab/sum(h)                                         # rel Hfkt

addmargins(tab)                          # mit Randsummen

dat <- as.data.frame.table(tab)          # Tabelle zu Dataframe
tab <- xtabs(Freq ~ stat + forsch, dat) # Dataframe zu Tabelle

scatterplot3d(
  dat, type = "h", lwd = 7, pch = "", color = gray(25:1/40),
  xlab = "Note Statistik I", ylab = "Note Forschungsmethoden",
  zlab = "Absolute/Relative Haeufigkeit"
)
```

Vierfeldertafel

Ein häufiger Spezialfall resultiert, wenn beide Merkmale jeweils nur zwei Merkmalsausprägungen haben

Es gibt dann genau vier gemeinsame Häufigkeiten

Die resultierende Kontingenztafel heißt in diesem Fall auch *Vierfeldertafel*

Vierfeldertafel der absoluten Häufigkeiten

Beispiel: Von $n = 1000$ Personen ließen sich 190 gegen Grippe impfen. Von den geimpften Personen erkranken 16 und von den nicht geimpften Personen 121 an Grippe

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

Vierfeldertafel der relativen Häufigkeiten

Gemeinsame relative Häufigkeiten

		Erkrankung		
		ja	nein	$\sum$
Impfung	ja	0.016	0.174	0.190
	nein	0.121	0.689	0.810
	$\sum$	0.137	0.863	1.000

Bedingte relative Häufigkeiten

		Erkrankung		
		ja	nein	$\sum$
Impfung	ja	0.084	0.916	1
	nein	0.149	0.851	1
	$\sum$			2

Korrelationsrechnung

Mittels Korrelationsrechnung soll der Zusammenhang zwischen zwei Merkmalen untersucht werden

Vorsicht:

- Die hieraus ableitbaren Aussagen sind **nicht** als Kausalzusammenhang (d. h. Ursache-Wirkungs-Beziehung) zu interpretieren

Beispiel

- In einer Untersuchung an Grundschulkindern wird ein positiver Zusammenhang zwischen Fingerfertigkeit und Körpergewicht festgestellt (d. h. die Fingerfertigkeit wächst tendenziell mit dem Körpergewicht)
- Dieser Zusammenhang ist durch die mit den beiden untersuchten Merkmalen kovariierende Variable Alter bedingt

Kovarianz

Um die Kovariation zweier Merkmale zu erfassen wird das Konzept der Varianz in naheliegender Weise verallgemeinert

Die (empirische) Kovarianz zweier Merkmale X und Y ist definiert durch

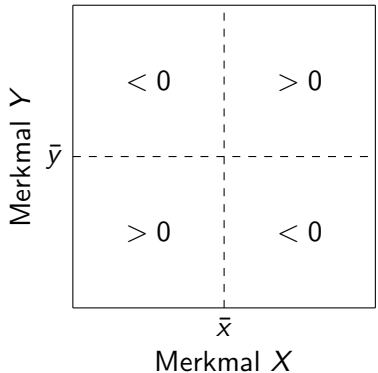
$$\tilde{s}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Beachte:

- Im Gegensatz zur Varianz kann die Kovarianz auch negative Werte annehmen

Kovarianz

Vorzeichenbetrachtung für das Produkt $(x_i - \bar{x}) \cdot (y_i - \bar{y})$



Kovarianz

Beispiel

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	2	3	-0.4	-0.2	0.08
2	1	2	-1.4	-1.2	1.68
3	3	4	0.6	0.8	0.48
4	2	3	-0.4	-0.2	0.08
5	4	4	1.6	0.8	1.28
$\sum$	12	16	0.0	0.0	3.60

mit

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

Kovarianz

Beispiel

- Für die Kovarianz errechnet man dann

$$\tilde{s}_{xy} = \frac{3.60}{5} = 0.72$$

Zur Klärung der Frage, ob dieser Wert vergleichsweise klein oder groß ist, betrachten wir den Wertebereich der Kovarianz s_{xy}

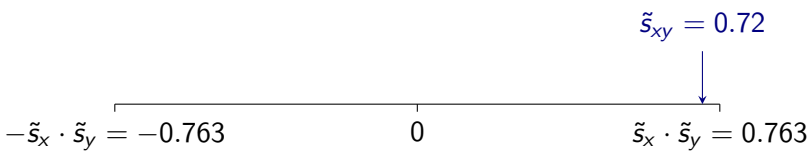
Man kann zeigen, dass allgemein gilt

$$|\tilde{s}_{xy}| \leq \tilde{s}_x \cdot \tilde{s}_y$$

(Der Beweis dieser Tatsache basiert auf der Cauchy-Schwarzschen Ungleichung)

Kovarianz

Beispiel



Zur besseren Vergleichbarkeit der Werte für den Zusammenhang zweier Merkmale wird die Kovarianz normiert, und zwar auf den maximal möglichen Wert von $\tilde{s}_x \cdot \tilde{s}_y$

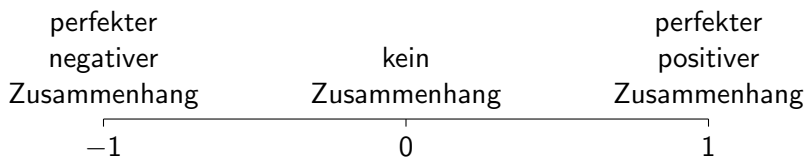
Korrelationskoeffizient (Produkt-Moment-Korrelation)

Definition der Produkt-Moment-Korrelation zweier Merkmale X und Y

$$r = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y}$$

Damit gilt

$$-1 \leq r \leq 1$$



Korrelationskoeffizient (Produkt-Moment-Korrelation)

Beispiel

$$r = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y}$$

$$= \frac{0.72}{0.763}$$

$$= 0.94$$

Korrelationskoeffizient (Produkt-Moment-Korrelation)

Beispiel: Berechnung in R

```
x  <- c(2, 1, 3, 2, 4)
y  <- c(3, 2, 4, 3, 4)
n  <- length(x)

xq <- mean(x)
yq <- mean(y)

sx <- sqrt(mean(x^2) - xq^2)      # Standardabweichung
sy <- sqrt(mean(y^2) - yq^2)

sxy <- sum((x - xq)*(y - yq))/n  # Kovarianz
r   <- sxy / (sx*sy)             # Korrelation
[1] 0.943
```

Korrelationskoeffizient (Produkt-Moment-Korrelation)

Beispiel: Berechnung in R

- Die Funktion `cor()` berechnet die (Pearsonsche) Produkt-Moment-Korrelation

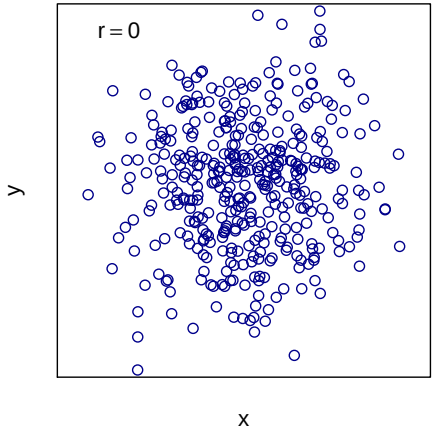
```
x <- c(2, 1, 3, 2, 4)
y <- c(3, 2, 4, 3, 4)
cor(x, y, method = "pearson")  # default: pearson
[1] 0.943
```

- Die Funktion `cov()` gibt die Kovarianz an, wobei die Produktsumme durch $n - 1$ geteilt wird anstelle von n

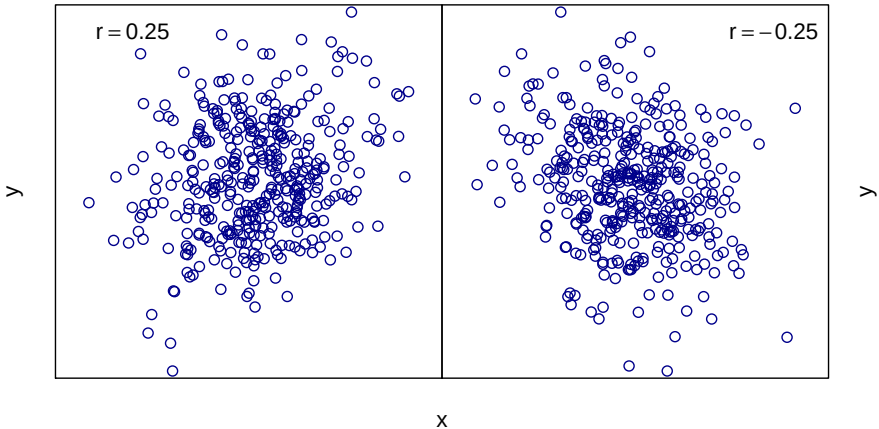
```
cov(x, y)
[1] 0.9
```

Wie drückt sich die Korrelation zwischen zwei Merkmalen im Streudiagramm aus?

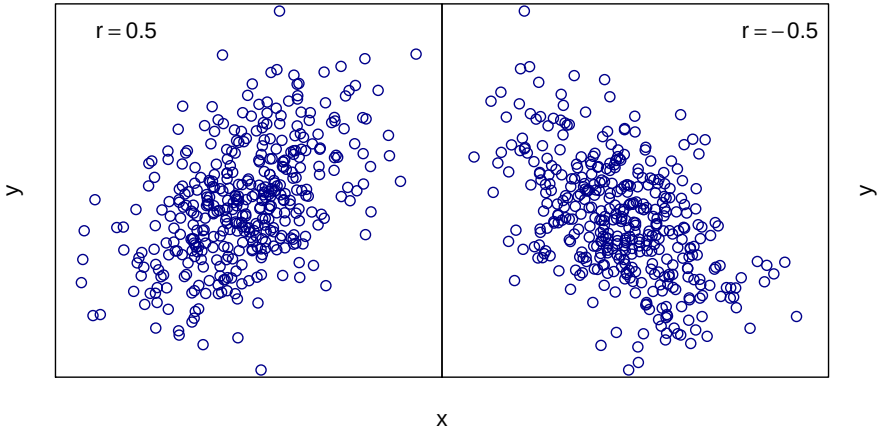
Korrelationskoeffizient



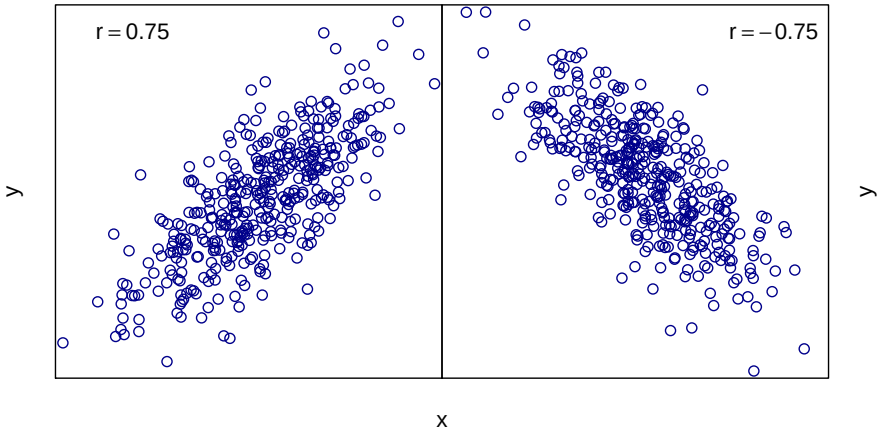
Korrelationskoeffizient



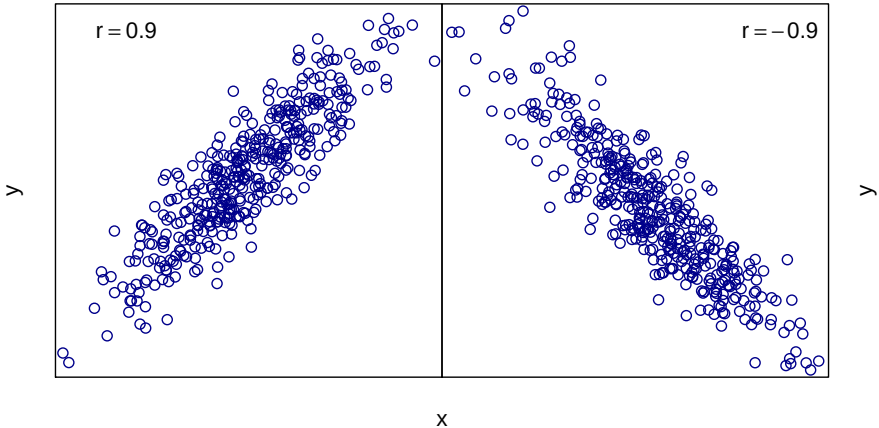
Korrelationskoeffizient



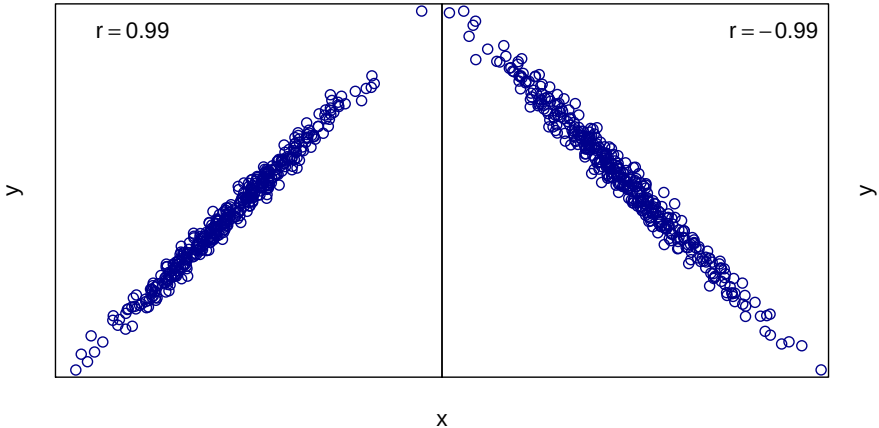
Korrelationskoeffizient



Korrelationskoeffizient



Korrelationskoeffizient



Korrelationskoeffizient (Produkt-Moment-Korrelation)

Bewertung des korrelativen Zusammenhangs (Cohen, 1988)

- $|r| \approx 0.1$ "schwacher Zusammenhang"
- $|r| \approx 0.3$ "mittlerer Zusammenhang"
- $|r| \geq 0.5$ "starker Zusammenhang"

Eigenschaften

- Symmetrie

$$r_{xy} = r_{yx}$$

- Es ist $r = 1$ (bzw. $r = -1$) genau dann, wenn die Beobachtungswerte auf einer Geraden liegen mit Steigung $b > 0$ (bzw. $b < 0$), d. h.

$$y_i = a + b \cdot x_i$$

für alle $i = 1, \ldots, n$

Korrelationskoeffizient

Beziehung zu z-transformierten Merkmalen

$$
\begin{aligned}
r &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\tilde{s}_x \cdot \tilde{s}_y} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{\tilde{s}_x} \cdot \frac{(y_i - \bar{y})}{\tilde{s}_y} \\
&= \frac{1}{n}\sum_{i=1}^{n} z(x_i) \cdot z(y_i)
\end{aligned}
$$

Korrelationskoeffizient

Vereinfachte Berechnung der Kovarianz

$$
\begin{aligned}
\tilde{s}_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i - x_i \cdot \bar{y} - \bar{x} \cdot y_i + \bar{x} \cdot \bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \cdot \bar{y} - \bar{x} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \bar{x} \cdot \bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}
\end{aligned}
$$

Korrelationskoeffizient

Invarianz unter positiv linearen Transformationen

$$x' = p \cdot x + q, \quad y' = u \cdot y + v \quad \text{mit } p, u > 0$$

Es gilt

$$\begin{aligned} r_{x'y'} &= \frac{\tilde{s}_{x'y'}}{\tilde{s}_{x'} \cdot \tilde{s}_{y'}} \\ &= \frac{p \cdot u \cdot \tilde{s}_{xy}}{(p \cdot \tilde{s}_x) \cdot (u \cdot \tilde{s}_y)} \\ &= \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} \\ &= r_{xy} \end{aligned}$$