

Statistik1

Punktmodelle 2: Zusammenhänge

Sebastian Sauer

Einstieg

Lernziele

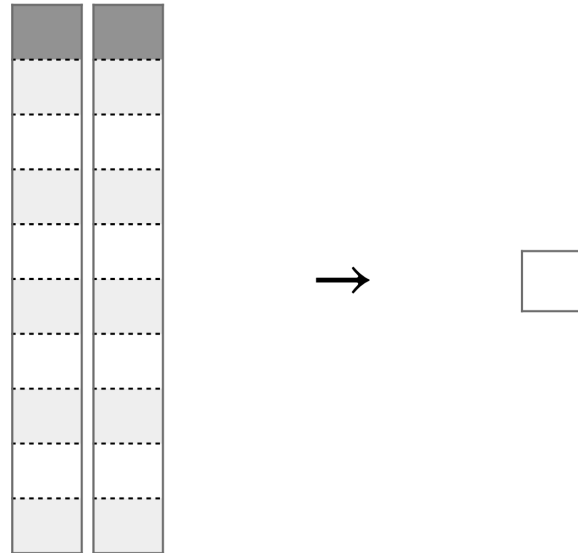
- Sie können die Begriffe Kovarianz und Korrelation definieren und ihren Zusammenhang erläutern.
- Sie können die Stärke einer Korrelation einschätzen.

Zusammenfassen zum Zusammenhang

Von einer zu zwei Spalten

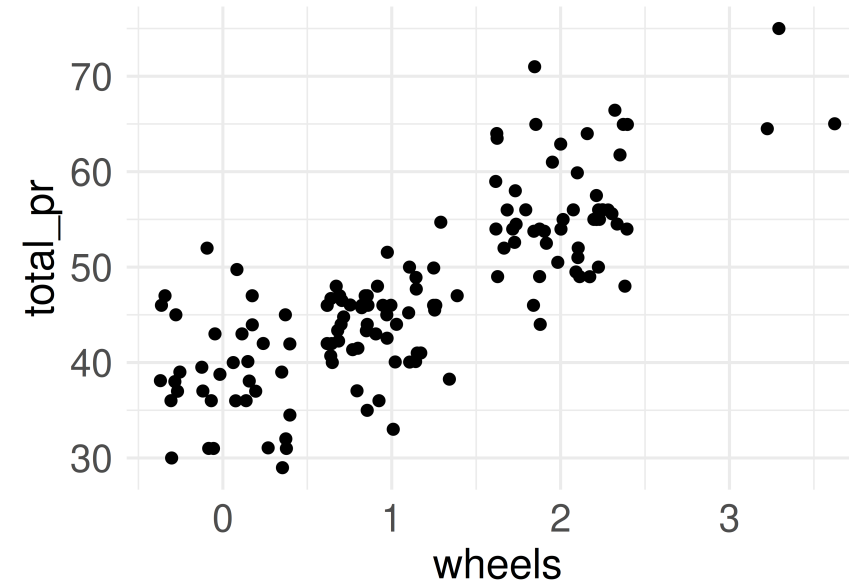
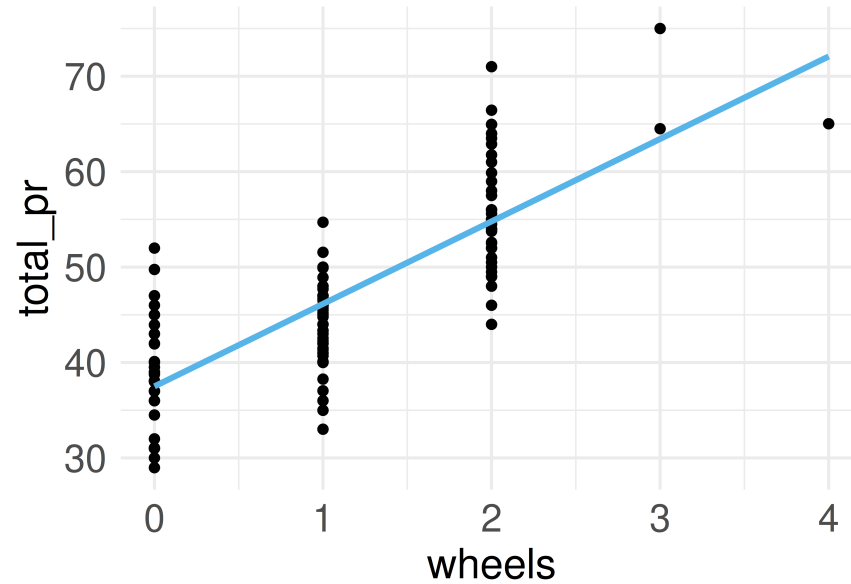
Bisher haben wir *eine* Spalte (einen Vektor) zu einer einzelnen Zahl zusammengefasst — einem “Punkt”.

Jetzt fassen wir *zwei* Spalten zusammen, wieder zu *einer* Zahl.



Es geht um die Frage: Wie hängen zwei (metrische) Variablen zusammen?

Beispiel: `wheels` und `total_pr`



Abweichungsrechtecke

Beispiel: Noten und Lernzeit

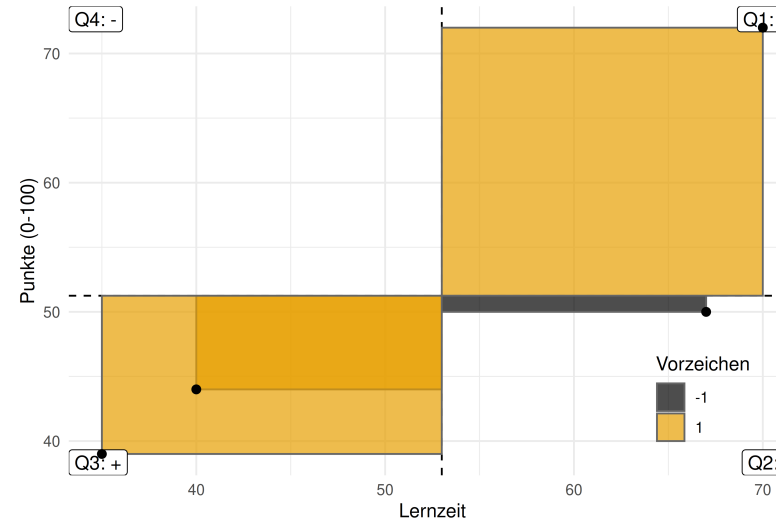
Anton, Bert, Carl und Daniel haben ihre Statistikklausur zurückbekommen. Die Lernzeit X scheint mit der erreichten Punktzahl Y (0–100) zusammenzuhängen.

id	y (Punkte)	x (Lernzeit)
1	72	70
2	44	40
3	39	35
4	50	67

Definition: Abweichungsrechteck

i Definition: Abweichungsrechteck

Im zweidimensionalen Fall spannt sich ein Abweichungsrechteck vom Mittelwert $\bar{x}$ bis zum Messwert x_i und genauso für Y . Mit $dx_i = x_i - \bar{x}$ (analog dy_i) ist die Fläche des Abweichungsrechtecks das Produkt der Abweichungen: $dx_i \cdot dy_i$.

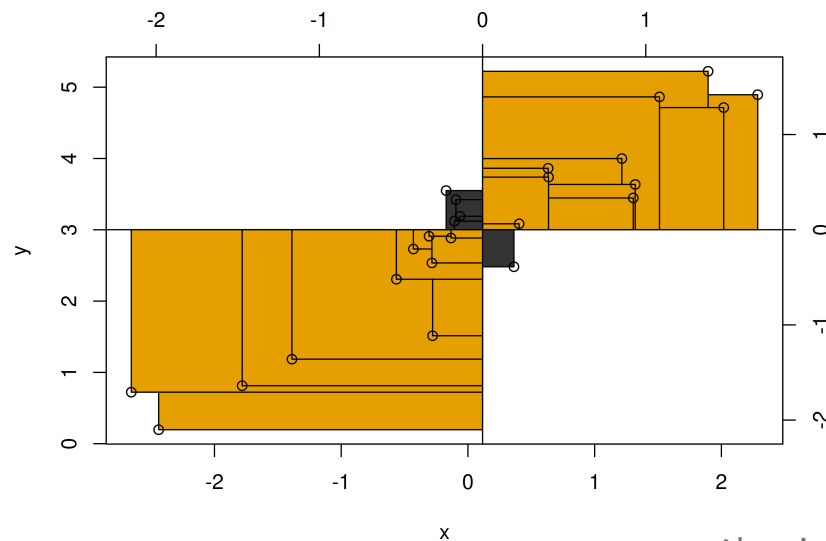


Abweichungsrechtecke

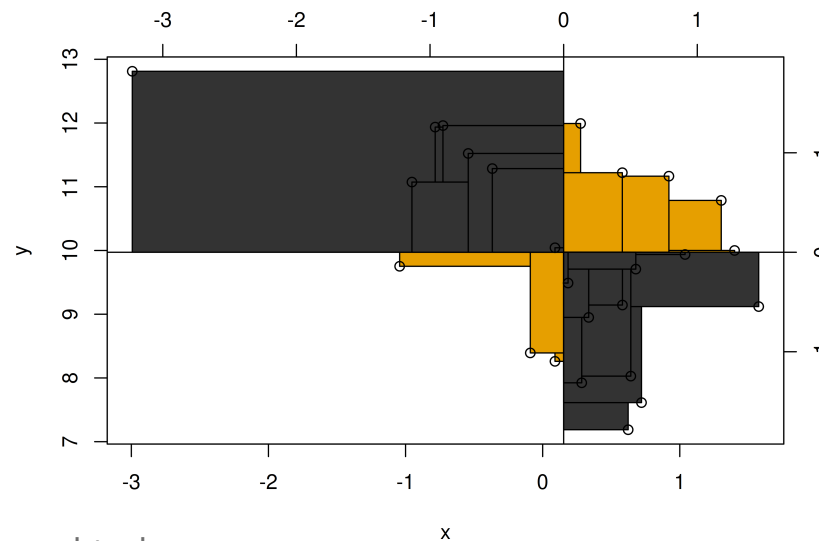
Vom Rechteck zur Summe

Legt man alle Abweichungsrechtecke zusammen, ergibt sich ein “Summenrechteck”. Je größer seine Fläche, desto stärker der (lineare) Zusammenhang.

Das *Vorzeichen* der Summe zeigt an, ob der Zusammenhang positiv (gleichsinnig) oder negativ (gegensinnig) ist.



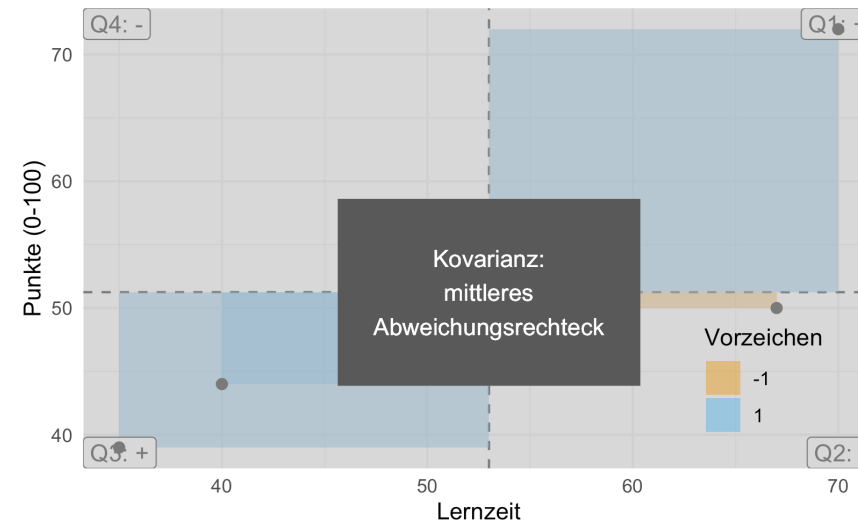
Abweichungsrechtecke



Definition: Kovarianz

Definition: Kovarianz

Die Kovarianz ist definiert als die Fläche des *mittleren* Abweichungsrechtecks. Sie ist ein Maß für Stärke und Richtung des linearen Zusammenhangs zweier metrischer Variablen.



Für die Noten und Lernzeit der vier Studierenden beträgt die Kovarianz **162**.

Formel der Kovarianz

$$\text{cov}(x, y) = s_{xy} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n dx_i \cdot dy_i$$

In Worten:

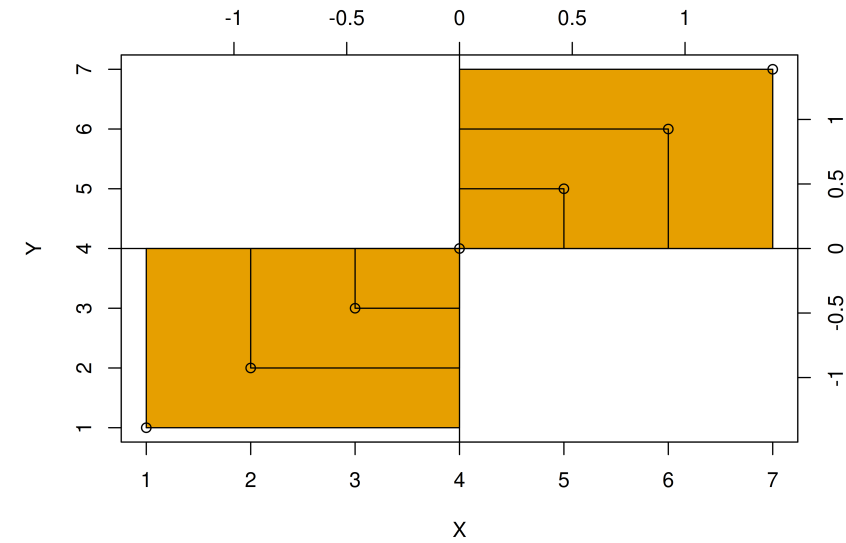
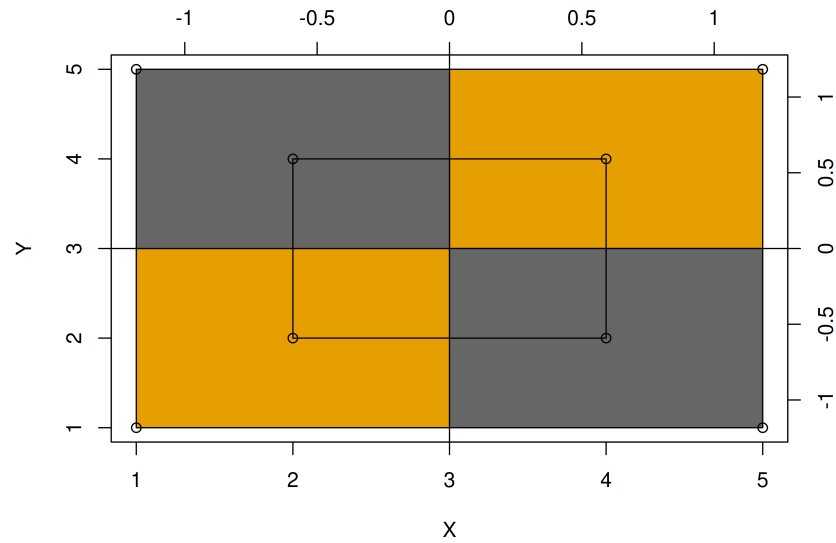
1. Abweichung jedes x_i vom Mittelwert $\bar{x}$ berechnen: dx_i .
2. Abweichung jedes y_i vom Mittelwert $\bar{y}$ berechnen: dy_i .
3. Für alle i die Abweichungsrechtecke $dx_i \cdot dy_i$ bilden.
4. Flächen der Abweichungsrechtecke addieren.
5. Durch die Anzahl der Beobachtungen n teilen.

Positive und negative Kovarianz

Positive Kovarianz (Beispiele): Größe & Gewicht, Lernzeit & Klausurerfolg, Distanz zum Ziel & Reisezeit, Temperatur & Eisverkauf.

Negative Kovarianz (Beispiele): Lernzeit & Freizeit, Alter & Restlebenszeit, Temperatur & Schneemenge, Lebenszufriedenheit & Depressivität.

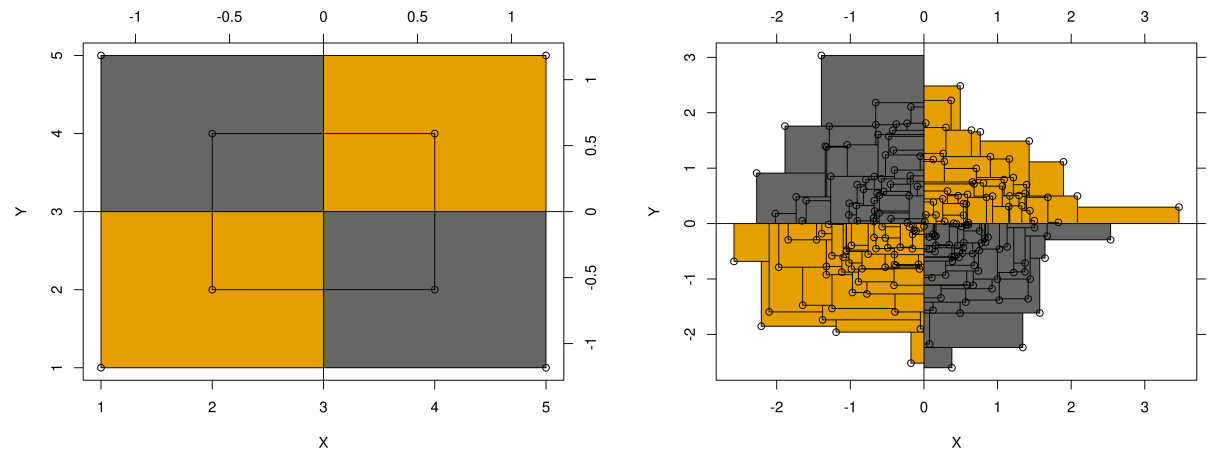
Extremwerte der Kovarianz



Wann ist die Kovarianz Null?

Bei einer Kovarianz von (ungefähr) Null ist die Summe der Abweichungsrechtecke pro Quadrant (ungefähr) gleich groß.

$$\begin{aligned}\sum (dX \cdot dY) &= 0 \\ \Leftrightarrow \text{MW} (dX \cdot dY) &= 0 \\ \Leftrightarrow \text{cov}(X, Y) &= 0\end{aligned}$$



Abweichungsrechtecke

Die Kovarianz ist schwer zu interpretieren

- Sie hängt von der Skalierung ab: Rechnet man z. B. Einkommen von Euro in Cent um, steigt die Kovarianz um den Faktor 100.
- Der Zusammenhang zwischen z. B. Einkommen und Lebenszufriedenheit sollte aber unabhängig von der Maßeinheit sein.
- Sie hat keinen Maximalwert, der einen perfekten Zusammenhang anzeigt.

Fazit: Die Kovarianz ist schwer zu interpretieren und wird in der Praxis wenig verwendet.

Korrelation

Korrelation als mittleres z-Produkt

Der Korrelationskoeffizient r nach Karl Pearson (1896) löst das Problem der schweren Interpretierbarkeit der Kovarianz.

Wertebereich: -1 (perfekte negative Korrelation) bis $+1$ (perfekte positive Korrelation).
 $r = 0$ bedeutet *kein linearer* Zusammenhang.

So wird r berechnet:

1. Alle x_i durch ihre Standardabweichung s_x teilen.
2. Alle y_i durch ihre Standardabweichung s_y teilen.
3. Mit diesen (z-transformierten) Werten die Kovarianz berechnen.

Definition: Korrelationskoeffizient r

Definition: Korrelationskoeffizient r

Der Korrelationskoeffizient r (nach Pearson) ist definiert als das mittlere Produkt der z-Wert-Paare (vgl. [Cohen et al., 2003](#)). Er misst den linearen Zusammenhang zweier metrischer Variablen; Wertebereich $[-1; 1]$, wobei 0 keinen linearen Zusammenhang anzeigt und $|r| = 1$ perfekten linearen Zusammenhang.

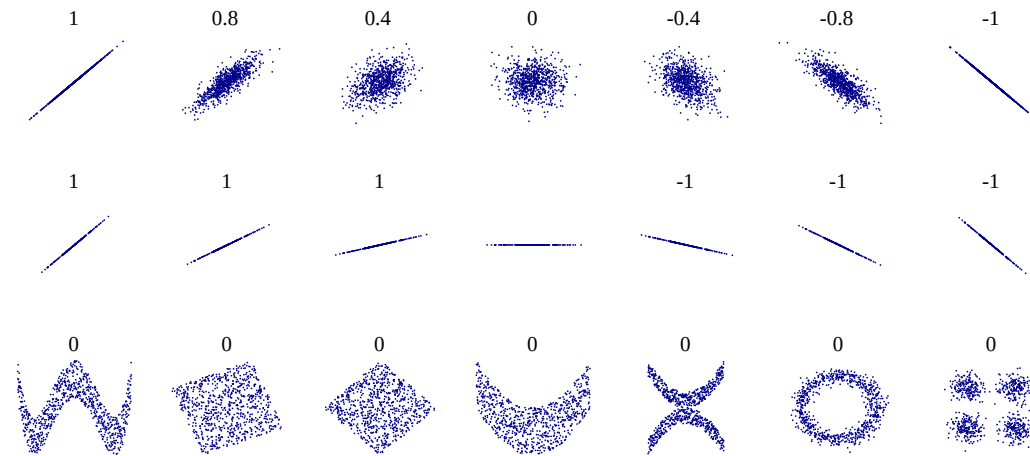
$$r_{xy} = \frac{1}{n} \sum_{i=1}^n z_{x_i} z_{y_i}$$

Was sagt r aus?

1. **Vorzeichen:** positiv $\rightarrow$ gleichsinniger Zusammenhang;
negativ $\rightarrow$ gegensinniger Zusammenhang.
2. **Absolutwert:** gibt die Stärke des linearen Zusammenhangs an — je näher an 1, desto stärker.

Korrelation ist (wie die Kovarianz) nur für metrische Variablen definiert.

Streudiagramme und Korrelationskoeffizient



Verschiedene Streudiagramme und ihr Korrelationskoeffizient ([DenisBoigelot, 2011](#))

Untere Zeile: nicht-lineare Zusammenhänge — hier ist $r = 0$, obwohl ein starker *nicht*-linearer Zusammenhang besteht.

Interpretation der Stärke

Grobe (!) Richtlinien nach Cohen ([1988](#)):

$ r $	Grobe Interpretation
0.01 – 0.09	sehr schwach
0.10 – 0.29	schwach
0.30 – 0.49	mittel
≥ 0.50	stark

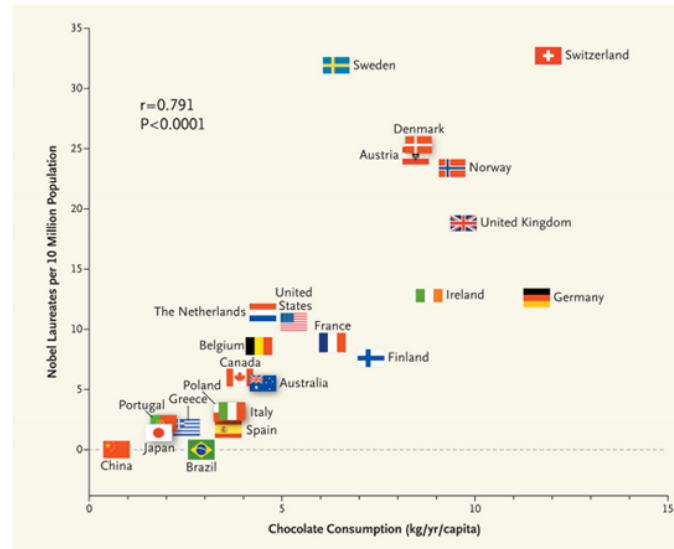
Korrelation mit R berechnen

```
1 mariokart |>  
2   select(total_pr, duration, n_bids) |>  
3   correlation() |> # aus `easystats`  
4   summary()
```

Parameter	n_bids	duration
total_pr	0.13	-0.04
duration	-0.12	

Korrelation ist nicht Kausation

Eine Studie fand eine starke Korrelation zwischen Schokoladenkonsum und Anzahl der Nobelpreise eines Landes ([Messerli, 2012](#)).



Schoki futtern macht schlau? ([Messerli, 2012](#))

Korrelation $\neq$ Kausation! Liegt Korrelation ohne Kausation vor, spricht man von *Scheinkorrelation*.

Scheinkorrelation: Störche und Babys

Ein Mythos besagt: Die Anzahl der Störche pro Landkreis korreliert mit der Anzahl der Babys (vgl. [Matthews, 2000](#)).

Mögliche Erklärung: Die “Naturbelassenheit” des Landkreises ist gemeinsame Ursache von Storchenzahl *und* Geburtenzahl — eine Scheinkorrelation ohne direkte Kausalität.

Glatze macht Corona?

Männer mit Glatze bekommen häufiger einen schweren Corona-Verlauf ([Goren et al., 2020](#)).

Alternative Erklärung: *Alter* wirkt sowohl auf Glatze (ältere Männer häufiger kahl) als auch auf die Schwere des Corona-Verlaufs (ältere Menschen häufiger schwerer betroffen).

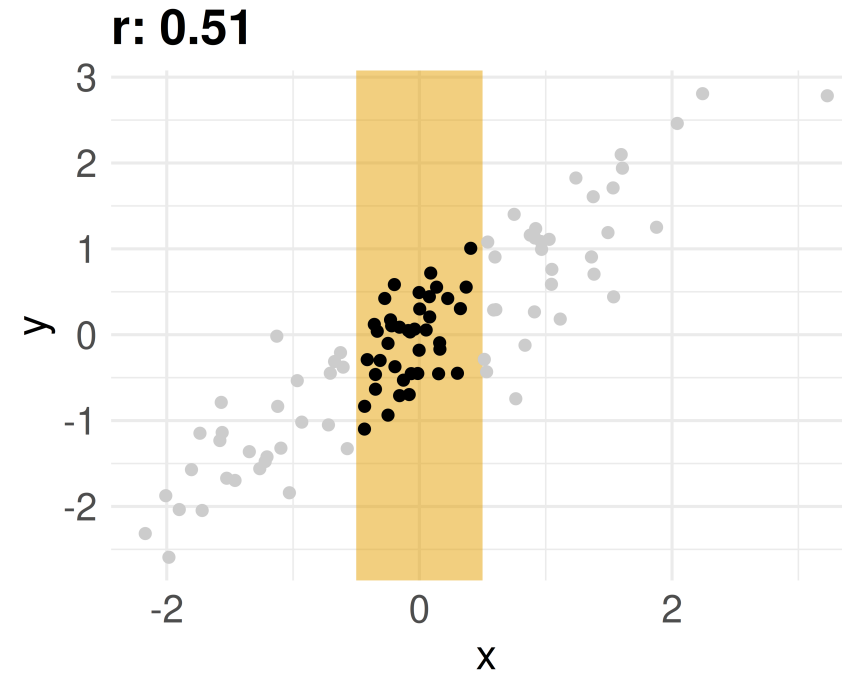
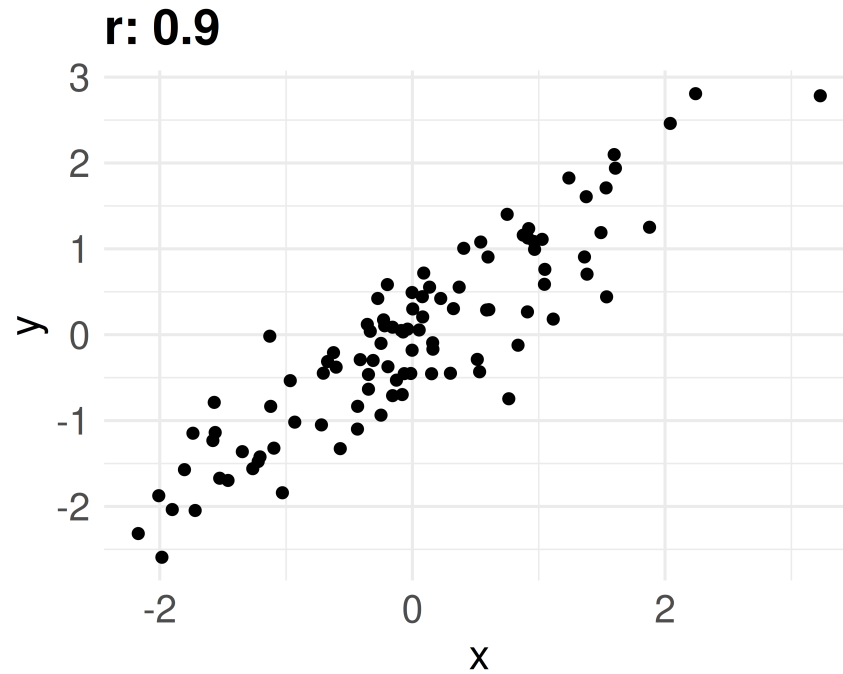
Wie man mit Statistik lügt

Einschränkung der Spannweite

Schränkt man den Wertebereich (die Spannweite) einer oder beider Variablen (nicht-randomisiert) ein, sinkt die Stärke (der Absolutwert) einer Korrelation (vgl. [Cohen et al., 2003](#)).

```
1 n <- 1e2
2 d <- tibble(x = rnorm(n = n, mean = 0, sd = 1),
3             e = rnorm(n = n, mean = 0, sd = .5),
4             y = x + e)
5
6 d_filtered <- d |> filter(between(x, -0.5, 0.5))
```

Effekt der Spannweiten-Einschränkung



Fallbeispiel

Verkaufspreis im Mariokart-Datensatz

In Ihrer Arbeit beim Online-Auktionshaus analysieren Sie, welche Variablen mit dem Verkaufspreis von Computerspielen zusammenhängen.

```
1 mariokart %>%  
2   dplyr::select(duration, n_bids, start_pr, ship_pr,  
3                 total_pr, seller_rate, wheels) %>%  
4   cor() %>%  
5   round(2)
```

Namensverwechslung: `select`

Sind mehrere Pakete geladen, die je eine Funktion `select` enthalten, verwendet R die Funktion aus dem zuletzt geladenen Paket — das kann zu verwirrenden Fehlermeldungen führen.

Abhilfe: Paket explizit angeben, z. B. `dplyr::select(...)`.

Korrelationstabelle (tidy)

```
1 mariokart %>%
2   dplyr::select(duration, n_bids, start_pr, ship_pr,
3                 total_pr, seller_rate, wheels) |>
4   correlation() |>
5   summary()
```

Parameter	wheels	seller_rate	total_pr	ship_pr	start_pr	n_bids
duration	-0.30**	-0.15	-0.04	0.27*	0.13	-0.12
n_bids	-0.08	-0.11	0.13	0.03	-0.63***	
total_pr	0.33**	0.01				

Sternchen zeigen die statistische Signifikanz der Korrelation an — ein Thema, das wir hier ignorieren.

Einzelne Korrelation & fehlende Werte

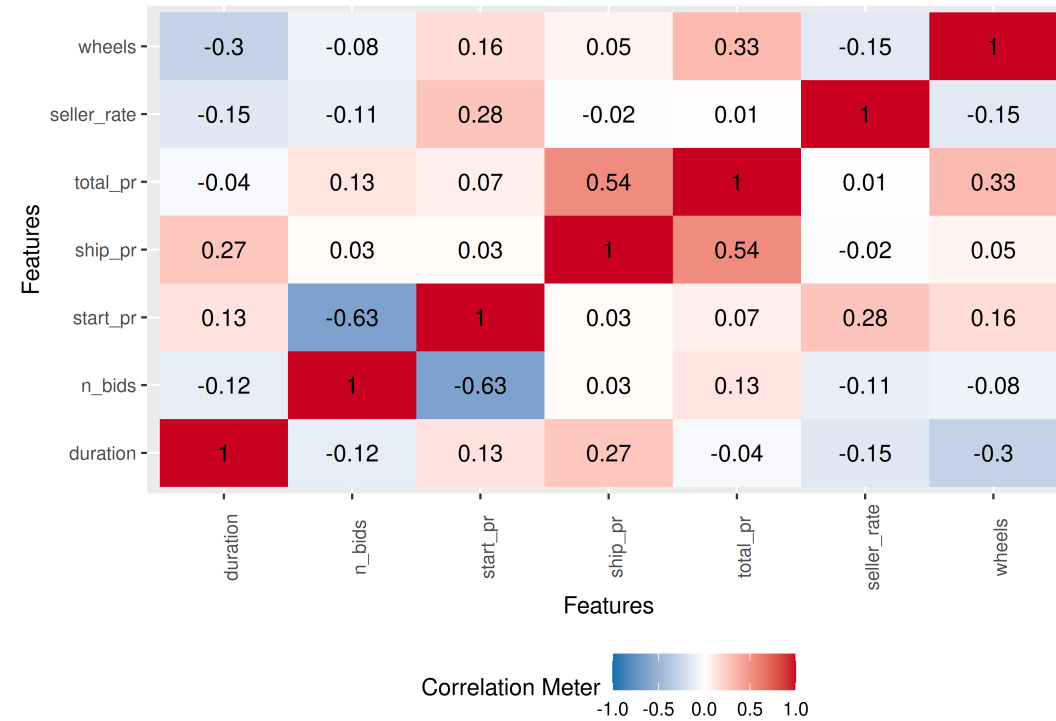
Einen einzelnen Korrelationskoeffizienten erhält man z. B. mit:

```
1 mariokart %>% summarise(korrelation = cor(total_pr, wheels))
```

Bei fehlenden Werten muss `cor` ausdrücklich dazu ermutigt werden, trotzdem zu rechnen:

```
1 mariokart %>%  
2   summarise(cor_super_wichtig = cor(total_pr, wheels,  
3                                     use = "complete.obs"))
```

Korrelationen als Heatmap



Zusammenfassung

- **Abweichungsrechteck:** Fläche $dx_i \cdot dy_i$ vom Mittelwert bis zum Messwert
- **Kovarianz:** mittleres Abweichungsrechteck — misst Stärke & Richtung des linearen Zusammenhangs, aber schwer interpretierbar (skalenabhängig)
- **Korrelation r :** standardisierte Kovarianz (mittleres z-Produkt), Wertebereich $[-1; 1]$
- Vorzeichen zeigt Richtung, Absolutwert zeigt Stärke des (linearen) Zusammenhangs
- Korrelation $\neq$ Kausation — Vorsicht vor Scheinkorrelationen
- Einschränkung der Spannweite kann Korrelationen künstlich abschwächen

Literaturhinweise

- Kaplan ([2009](#)) — unorthodoxe, geometrische Herangehensweise an Korrelation und Regression
- Poldrack ([2023](#)) — modernes, frei online verfügbares Statistikbuch
- Cohen et al. ([2003](#)) — Klassiker, etwas höheren Anspruchs
- Pearl & Mackenzie ([2018](#)) — Scheinkorrelation vs. “echte” Korrelation, mit Wissenschaftsgeschichte

Referenzen

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. <http://dx.doi.org/10.4324/9780203771587>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Ed. Lawrence Erlbaum.
- DenisBoigelot. (2011). *English: Redesign File:Correlation_examples.Png Using Vector Graphics (SVG File)* [Artwork]. <https://commons.wikimedia.org/w/index.php?curid=15165296>
- Goren, A., Vaño-Galván, S., Wambier, C. G., McCoy, J., Gomez-Zubiaur, A., Moreno-Arrones, O. M., Shapiro, J., Sinclair, R. D., Gold, M. H., Kovacevic, M., Mesinkovska, N. A., Goldust, M., & Washenik, K. (2020). A Preliminary Observation: Male Pattern Hair Loss among Hospitalized COVID-19 Patients in Spain – A Potential Clue to the Role of Androgens in COVID-19 Severity. *Journal of Cosmetic Dermatology*, 19(7), 1545–1547. <https://doi.org/10.1111/jocd.13443>
- Kaplan, D. T. (2009). *Statistical Modeling: A Fresh Approach*. CreateSpace. <https://dtkaplan.github.io/SM2-bookdown/>
- Matthews, R. (2000). Storks Deliver Babies ($P=0.008$). *Teaching Statistics*, 22(2), 36–38. <https://doi.org/10.1111/1467-9639.00013>
- Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, 367(16), 1562–1564. <https://doi.org/10.1056/NEJMon1211064>
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.
- Pearson, K. (1896). VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187, 253–318. <https://doi.org/10.1098/rsta.1896.0007>
- Poldrack, R. A. (2023). *Statistical Thinking: Analyzing Data in an Uncertain World*. Princeton University Press. <https://statstinking21.github.io/statstinking21-core-site/>