

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

QSB1 – Tag 4

7. Analyse des Zusammenhangs quantitativer Merkmale

Dr. rer. med. habil. Marita Ziepert /
Dr. rer. nat. Meinhard Mende,
ZKS & IMISE, Universität Leipzig

imise.

Lernziele

Am Ende der Vorlesung

- können Sie zwei metrische Merkmale im Scatterplot darstellen,
- die Richtung und die Stärke von Zusammenhängen beurteilen,
- können Sie den Korrelationskoeffizienten nach Pearson interpretieren und kennen das Szenarium, wann man diesen anwenden darf,
- sind Sie vor möglichen Fehlschlüssen bei Korrelation auf der Hut,
- kennen Sie den Rangkorrelationskoeffizienten von Spearman,
- die Regressionskoeffizienten aus einer vorgegebenen Regressionsgeraden bestimmen und interpretieren,
- eine Regressionsgerade zeichnen,
- das Bestimmtheitsmaß interpretieren.

Übersicht: Zusammenhang mehrerer Merkmale

Vorlesung 7: Zusammenhang quantitativer Merkmale

7.1 – 7.4 Korrelation

- Der Scatterplot
- Der Korrelationskoeffizient von Pearson
- Fehlschlüsse bei der Korrelation
- Der partielle Korrelationskoeffizient
- Der Rangkorrelationskoeffizient von Spearman

7.5 Die lineare Regression

- Die Güte einer linearen Regression
- Die multiple lineare Regression

Vorlesung 8: Zusammenhang qualitativer Merkmale

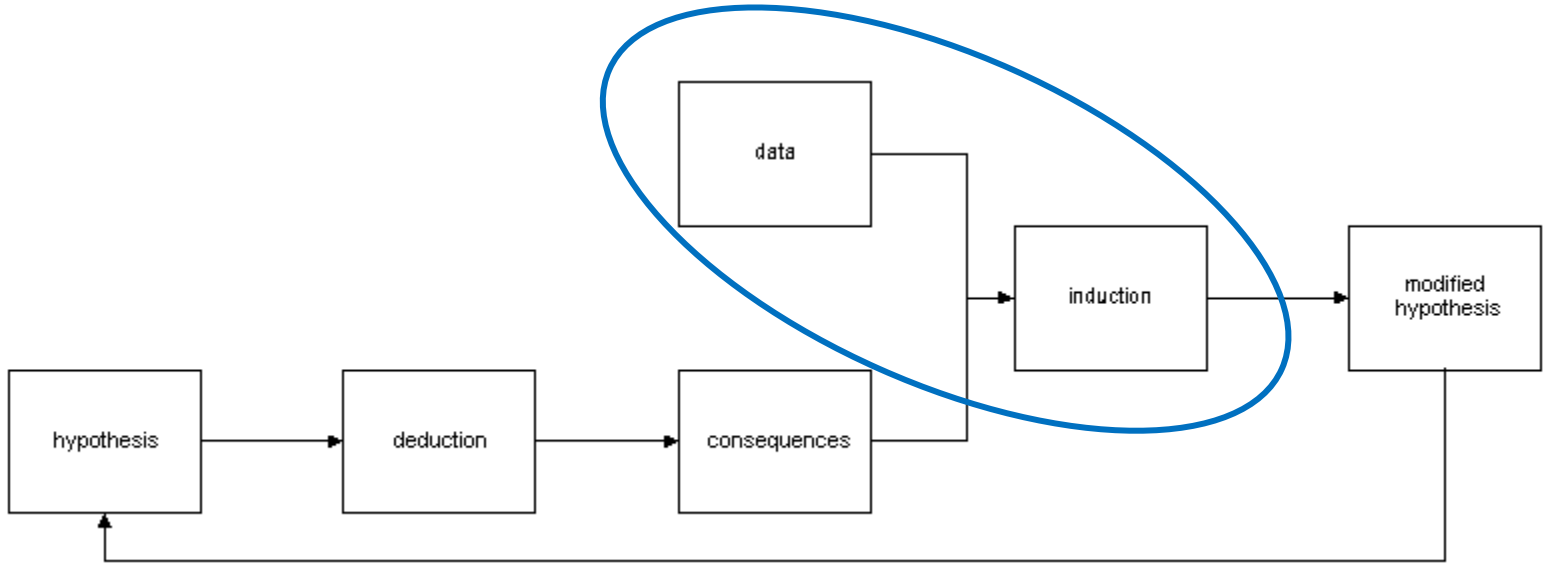
V 7 Zusammenhang I

7.1 Einführung

Empirische Forschung

Remember that all models are wrong;
the practical question is how wrong do they have to be to not be useful.

Box GEP, Draper NR. Empirical model-building and response surfaces. John Wiley & Sons 1987



Grafik aus Chance BL, Journal of Statistics Education 2002;10(3).jse.amstat.org/v10n3/chance.html

Scatterplot I

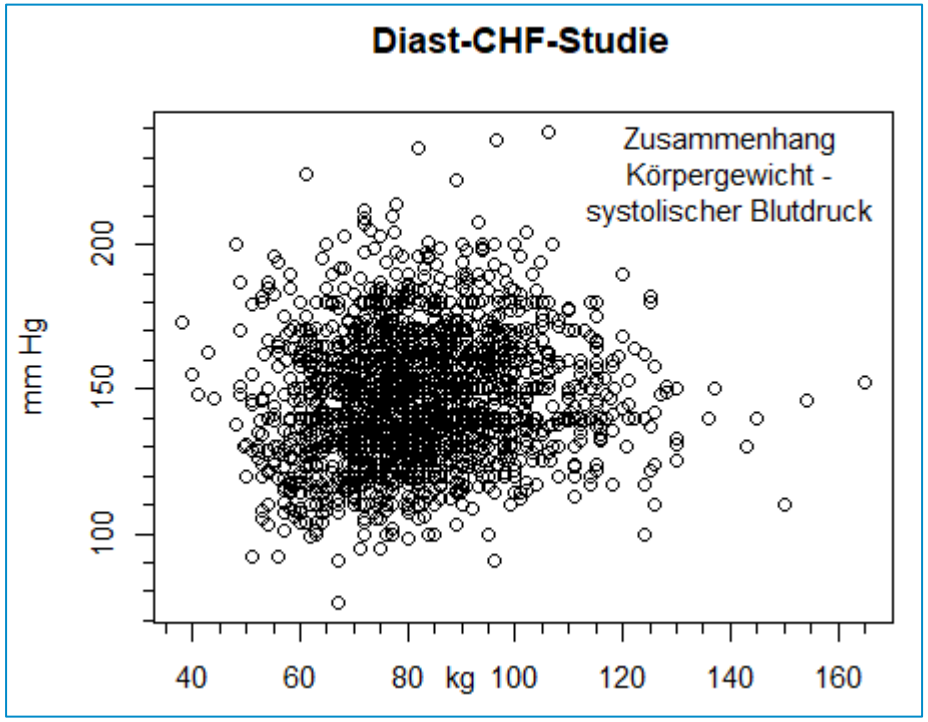
Bei den meisten statistischen Analysen geht es darum, Zusammenhänge zwischen Merkmalen aufzudecken und zu quantifizieren:

- Erster Schritt: grafische Darstellung zweier quantitativer Merkmale im Scatterplot (syn.: Punkt- / Streudiagramm)
- Der lineare Zusammenhang ist umso größer, je enger sich die Punktwolke um eine gedachte Gerade schmiegt.
- Eine steigende Tendenz der Punktwolke nennt man positive Korrelation, eine abfallende Tendenz negative Korrelation.

Frage: Wie quantifiziert man diese Zusammenhänge ?

Beispiel 7.1: Diast-CHF-Studie

- Zusammenhang Körpergewicht – systolischer Blutdruck

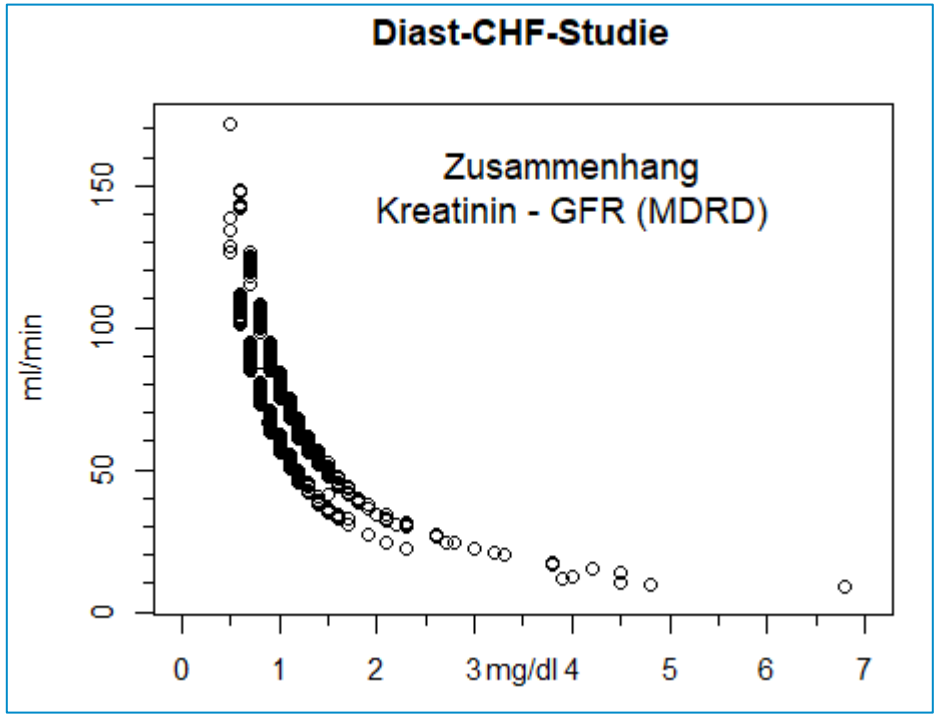


Richtung: positiv,

Stärke: kein bzw. sehr schwacher Zusammenhang

Beispiel 7.2: Diast-CHF-Studie

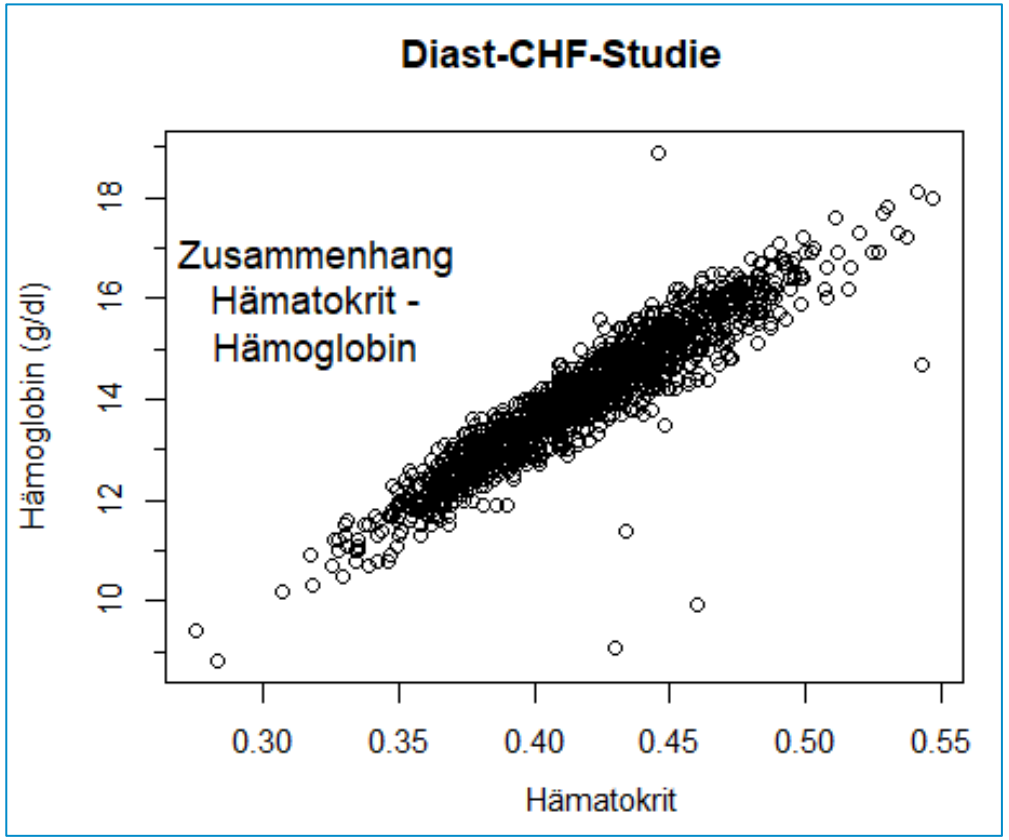
- Zusammenhang Kreatinin - GFR



Anmerkung:
Dieser Zusammenhang entsteht durch die MDRD-Formel.

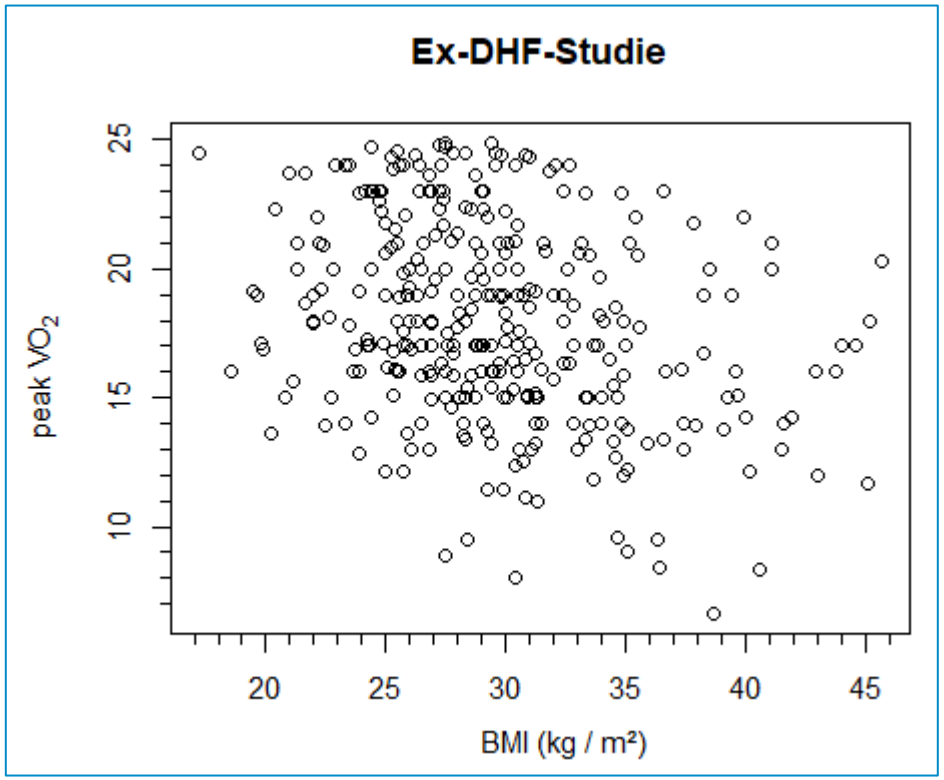
Beispiel 7.3: Diast-CHF-Studie

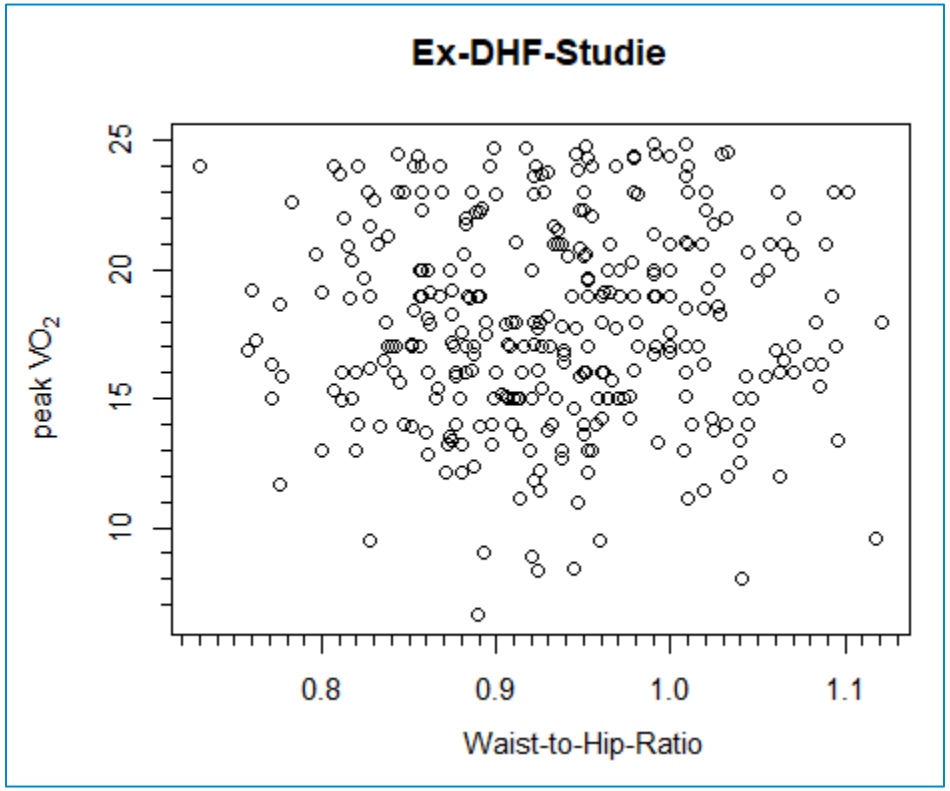
- Zusammenhang Hämatokrit - Hämoglobin



Anmerkung:
Beide Merkmale hängen zusammen über die Erythrozyten.

Beispiel 7.4: Ex-DHF-Studie





Korrelation - Einführung I

- Die Korrelationsanalyse ist ein Werkzeug der Statistik, den ungerichteten Zusammenhang zweier Merkmale zu untersuchen,

Korrelationsmaße

- quantifizieren die Stärke des Zusammenhangs und
- geben an, ob die Korrelation gleichläufig, d.h. positiv oder gegenläufig, d.h. negativ ist

Korrelation - Einführung II

– Geschätzte Streuungen (empirische Varianzen) in X und Y:

$$SQ(X) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \Rightarrow {s_x}^2 = Sxx = \frac{SQ(X)}{n-1}$$

$$\widehat{Var(X)} = \frac{SQ(X)}{n-1}$$

$$SQ(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \Rightarrow {s_y}^2 = Syy = \frac{SQ(Y)}{n-1}$$

$$\widehat{Var(Y)} = \frac{SQ(Y)}{n-1}$$

Korrelation - Einführung III

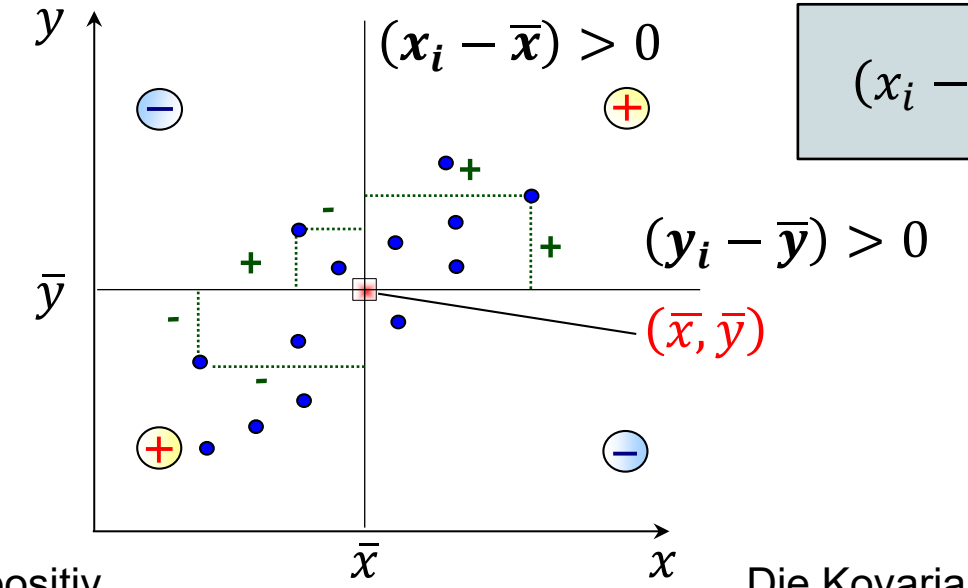
- Summe der Produkte von X und Y:
- Empirische Kovarianz von X und Y

$$SP(X,Y) = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \qquad \Rightarrow S_{xy} = \frac{SP(X,Y)}{n-1}$$

$$\widehat{Cov}(X,Y) = \frac{SP(X,Y)}{n-1}$$

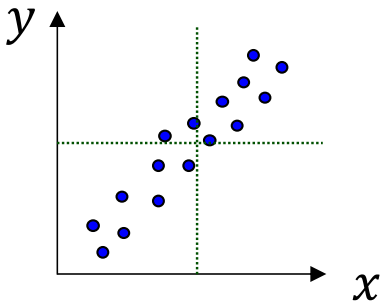
Korrelation – Einführung IV

Die Kovarianz misst, ob X und Y im Trend gleich- oder gegenläufig sind.

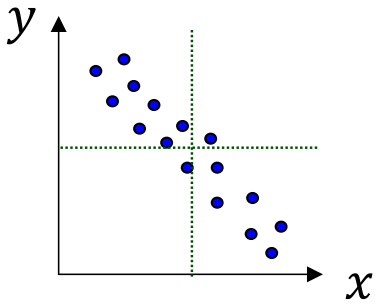


$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Die Kovarianz ist positiv, wenn die Daten so liegen



Die Kovarianz ist negativ, wenn die Daten so liegen



7.2 Der lineare Korrelationskoeffizient (Pearson)

Linearer Korrelationskoeffizient I

auch: Produkt-Moment-Korrelation (Galton 1880, Pearson 1896)

Modell-Ebene:

- Voraussetzung: Normalverteilung der Merkmale X, Y
- Standardisiertes Maß für den linearen Zusammenhang
-

$$Korrelation(X,Y) = \frac{Kovarianz\ (X,Y)}{\sqrt{Varianz\ (X) \cdot Varianz\ (Y)}}$$

$$r = Corr\ (X,Y) = \frac{Cov\ (X,Y)}{\sqrt{Var\ (X)} \cdot \sqrt{Var\ (Y)}}$$

Linearer Korrelationskoeffizient II

Empirische Ebene (aus den Daten errechnet):

$$\hat{r}_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = \frac{SP\,(X,Y)}{\sqrt{SQ(X) \cdot SQ(Y)}}$$

- misst den linearen Zusammenhang metrischer Merkmale X und Y,
- hat Werte in [-1, 1]
- je größer der Absolutbetrag, desto stärker der Zusammenhang
- einheitenlos durch die Normierung mit den Standardabweichungen

Kovarianz – Korrelation

- Formelzeichen: ρ
- misst den linearen Zusammenhang metrischer Merkmale
- Aussagen über gleichsinnige Variation
- abhängig von der Maßeinheit
- Wertebereich $(-\infty, +\infty)$

- Formelzeichen r
- misst den linearen Zusammenhang metrischer Merkmale
- Aussagen über gleichsinnige Variation
- standardisiert, d. h. einheitenlos
- Wertebereich $[-1, +1]$

Pearson‘s Korrelationskoeffizient: Interpretation

- Der Korrelationskoeffizient r nimmt Werte an im Intervall $-1 \leq r \leq +1$

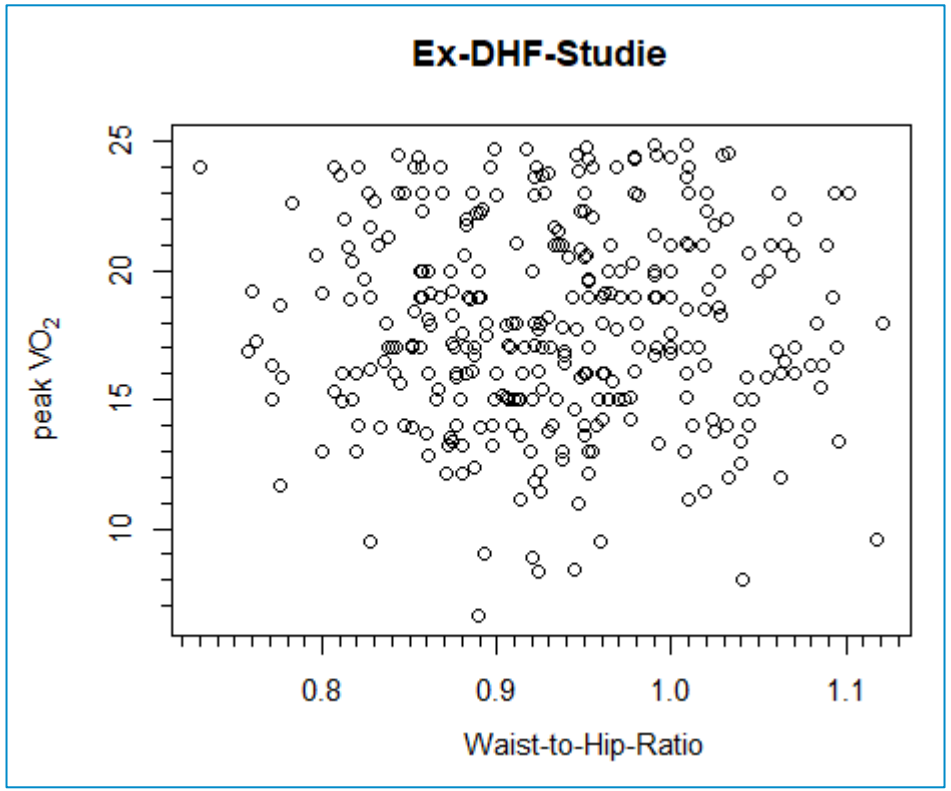
Grobe Orientierung zur Beurteilung der Stärke des Zusammenhangs:

- $|r| < 0{,}4$ schwach
- $0{,}4 \leq |r| < 0{,}6$ mittel
- $0{,}6 \leq |r| < 0{,}8$ stark
- $0{,}8 \leq |r| < 1{,}0$ sehr stark
- $|r| = 1{,}0$ funktionaler Zusammenhang

https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/regression-correlation (2026-10-06).

Beispiel I

– B 7.4 Quotient Taille- zu Hüft-Umfang – peak VO_2

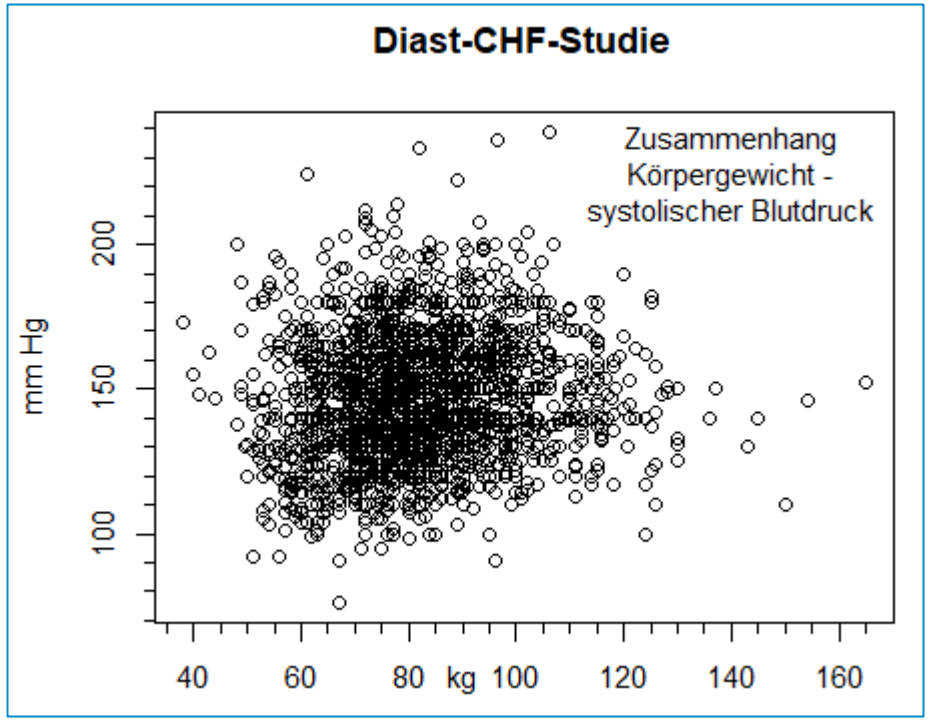


$r = -0{,}002$

keine Korrelation

Beispiel II

– B 7.1 Diast-CHF: Körpergewicht – systolischer Blutdruck



$r = 0{,}11$

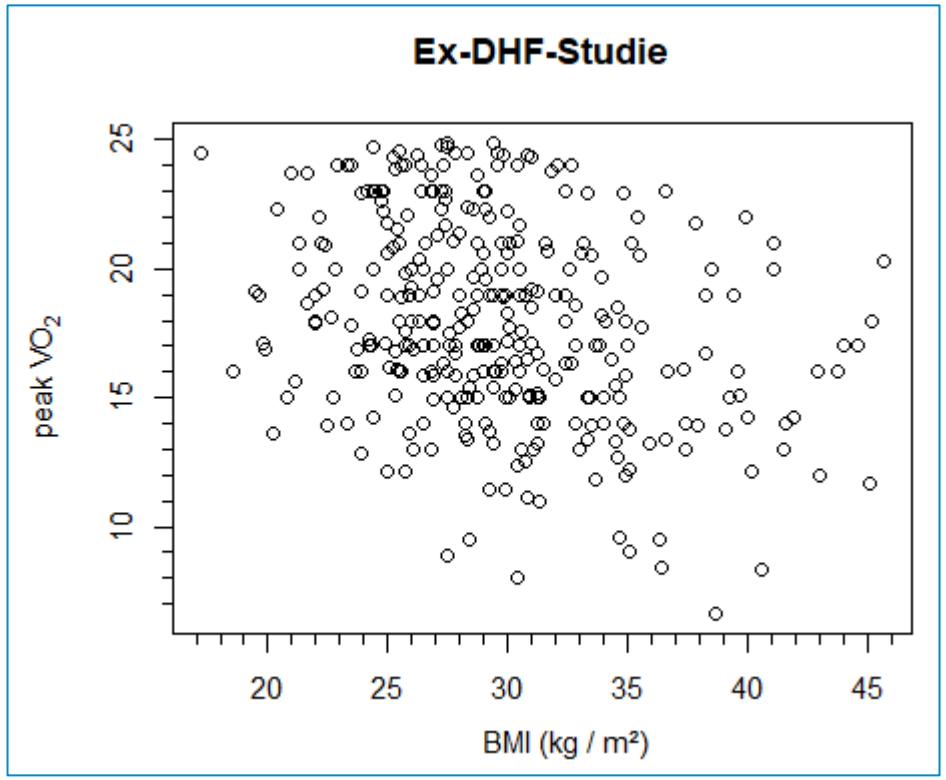
(sehr) schwache Korrelation

CAVE!
Verzerrungen ggf. durch
Geschlechtsunterschiede möglich



Beispiel III

– Ex-DHF-Studie: Probanden mit BMI > 20 kg/m²: BMI – peak VO_2



schwacher negativer Zusammenhang

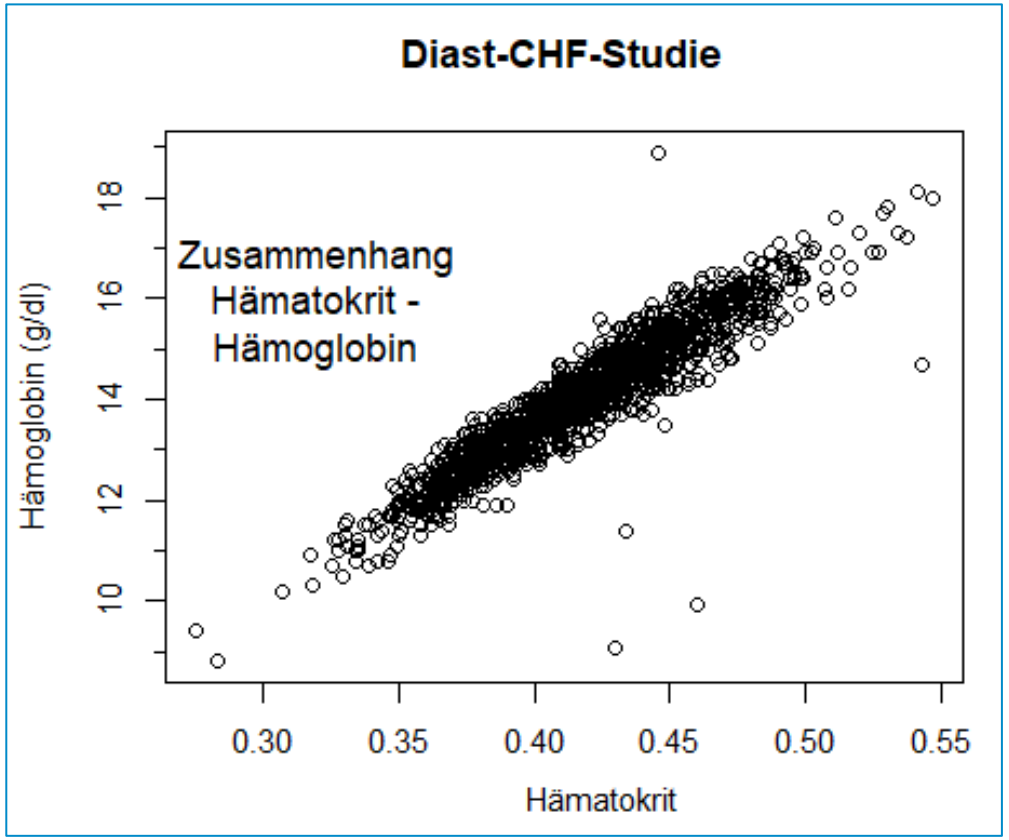
$r = -0{,}30$



Gonick & Smith: Statistik in Cartoons

Beispiel IV

- Diast-CHF-Studie: Hämatokrit – Hämoglobin

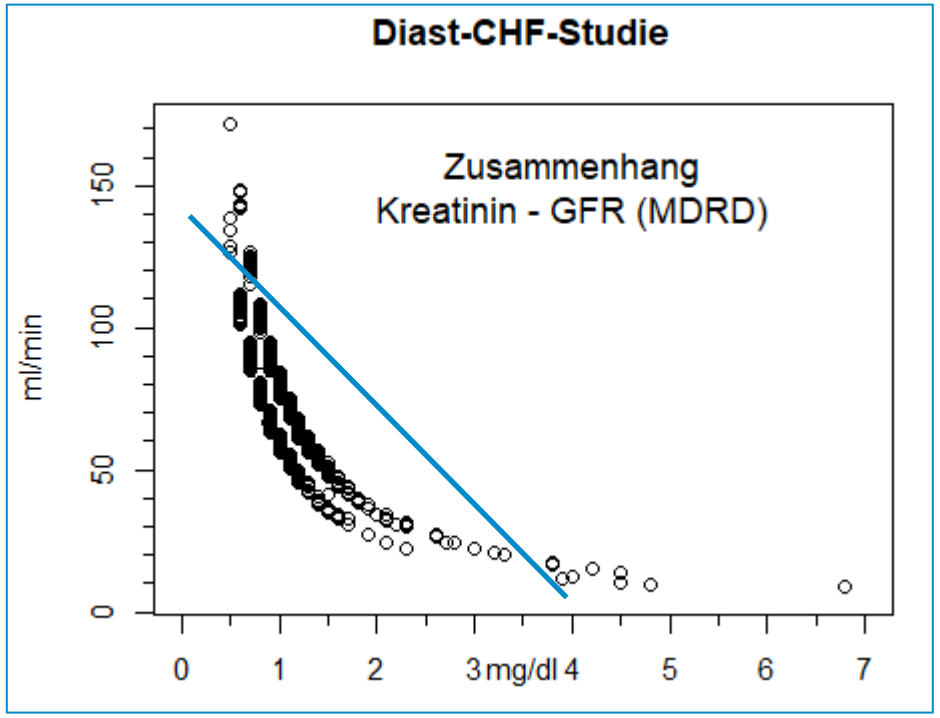


sehr starke positive Korrelation

r = 0,93

Beispiel V

- Diast-CHF-Studie: Kreatinin – berechnete GFR (MDRD)



starker negativer Zusammenhang

r = -0,65



CAVE:

r misst nur den linearen Zusammenhang;

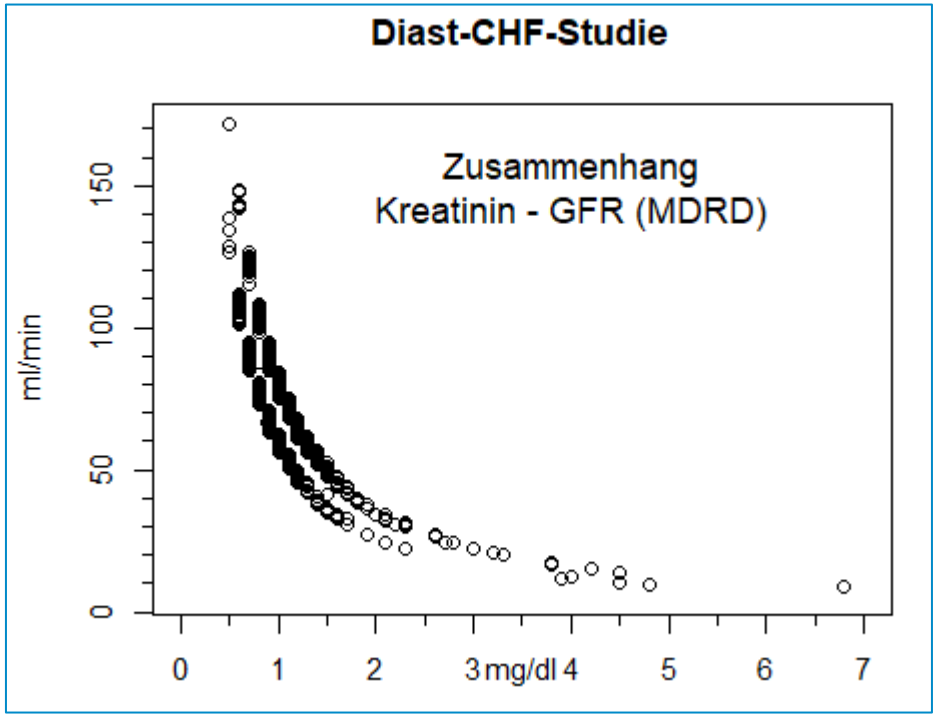
Pearson‘s R ist hier nicht adäquat!

Beispiel V b

- Diast-CHF-Studie: Kreatinin – berechnete GFR (MDRD)

- offensichtlich nichtlinearer Zusammenhang

➔ Rangkorrelationskoeffizient (Spearman) ist adäquat: R = 0,76



Übungsaufgabe 1

Ex-DHF-Studie: 322 HFpEF-Patienten (heart failure with preserved ejection fraction) randomisiert:

- UC (Usual Care): Standardbehandlung für HFpEF-Patienten
- ExT (Exercise Training): angeleitetes Training zusätzlich zu UC

Der primäre Endpunkt nach 12 Monaten war ordinal (*worsened, unchanged, improved*).

- Die peak VO_2 im spiroergometrischen Test ist ein Marker für die Leistungsfähigkeit.

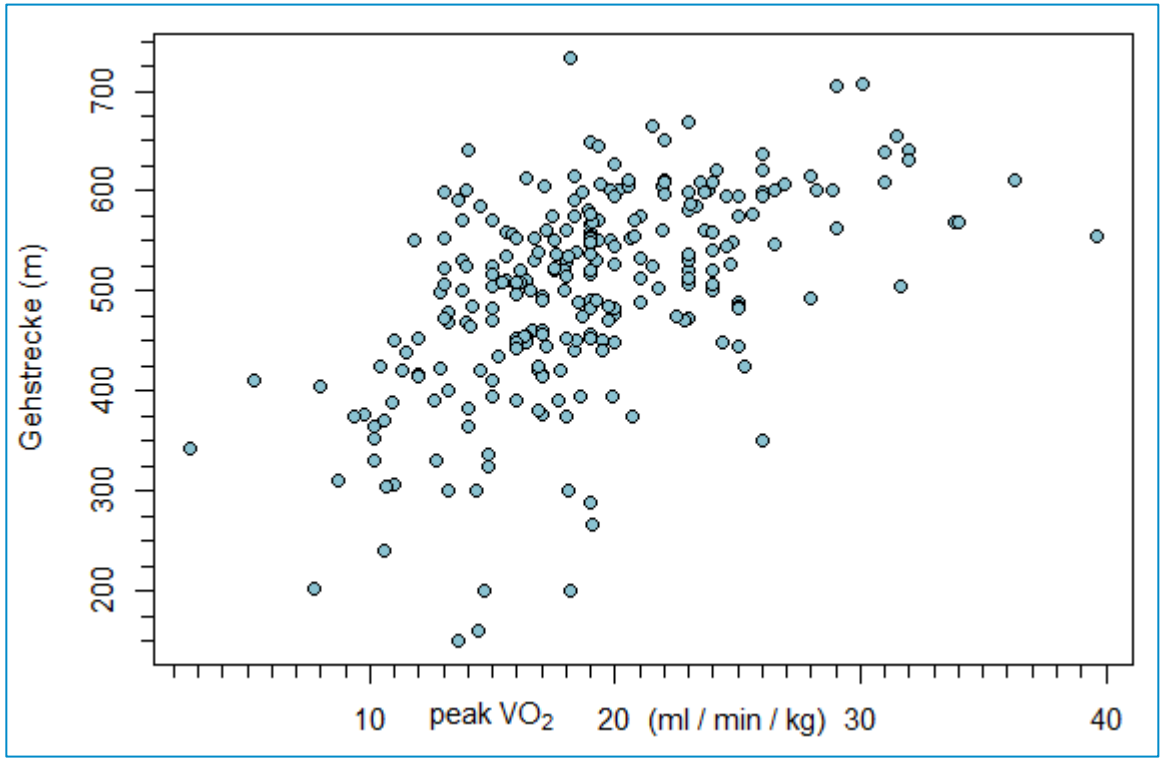
Edelmann F et al. Nature Medicine 2025;31:306-314

Fragestellung:

- Beurteilen Sie den Zusammenhang zwischen den peak-VO_2-Werten und der 6-Minuten-Gehstrecke nach 12 Monaten (Richtung und Stärke, s. nächste Folie)!

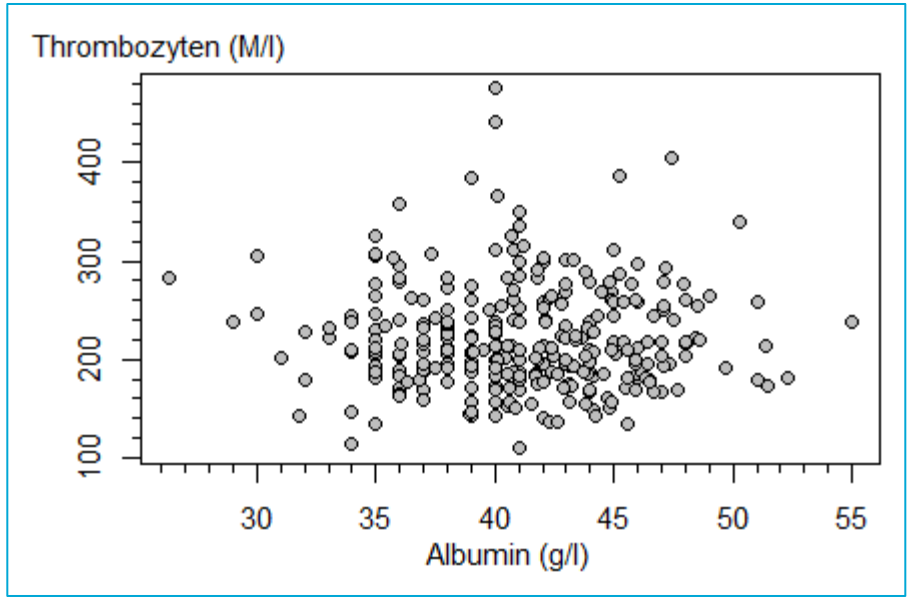
Übungsaufgabe 1 (Fortsetzung)

- Wie groß vermuten Sie R (Pearson)!
  -
  -
- Wie interpretieren Sie das?
  -



Übungsaufgabe 2

- In derselben Studie wurden baseline Laborwerte gemessen.
- Uns interessiert der Zusammenhang Albumin – Thrombozyten:
  Schätzen Sie R (Pearson)
  und beurteilen Sie Stärke
  und Richtung des
  Zusammenhangs!



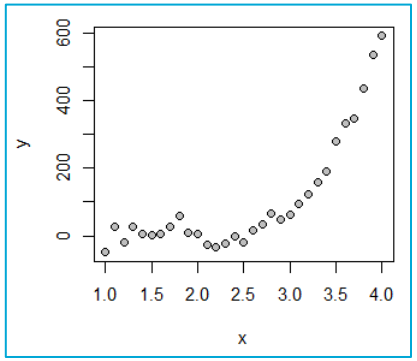
V 7

7.3 Rangkorrelation (Spearman)

Rangkorrelation I

Szenarien:

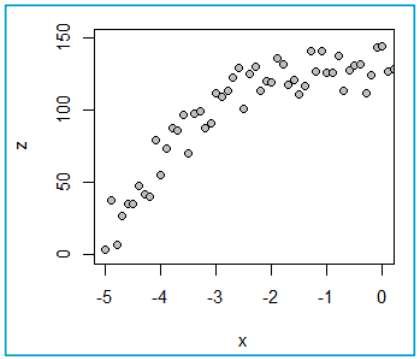
- monotoner Zusammenhang zwischen metrischen Variablen
- oder Ausreißer / extreme Werte



Dann ist Pearsons R nicht geeignet und man muss alternativ die Rangkorrelation berechnen.

Voraussetzungen:

- ordinale / metrische Skala (ggf. Rangdaten),
- Zusammenhang ist monoton



⇨ Rangkorrelationskoeffizient ρ_S (nach Spearman)

Rangkorrelation II

Idee:

- Von beiden sortierten Wertereihen werden Ränge gebildet.
- Danach berechnet man den Korrelationskoeffizienten anhand der Ränge anstatt mit den Werten:

$$\rho_s = \frac{Kovarianz\ (R(X), R(Y))}{\sqrt{Var\ \big(R(X)\big) \cdot Var\ (R(Y))}}$$

$$\hat{\rho}_s = r_s = \frac{SP(R(X), R(Y))}{\sqrt{SQ\ \big(R(X)\big) \cdot SQ\ (R(Y))}}$$

$\rho = rho$

Rangkorrelation III

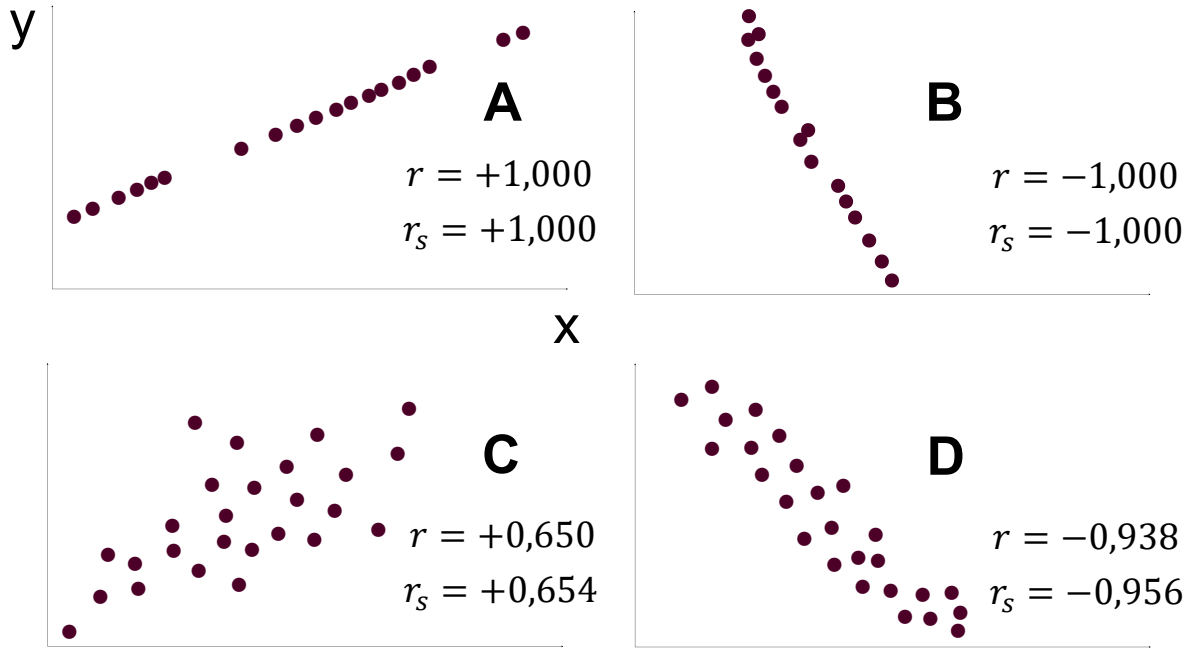
- Spearman-Rang-Korrelationskoeffizient - vereinfachte Rechnung:

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

- Dabei sind die
- $D_i = R(X_i) - R(Y_i)$ Differenzen der Ränge

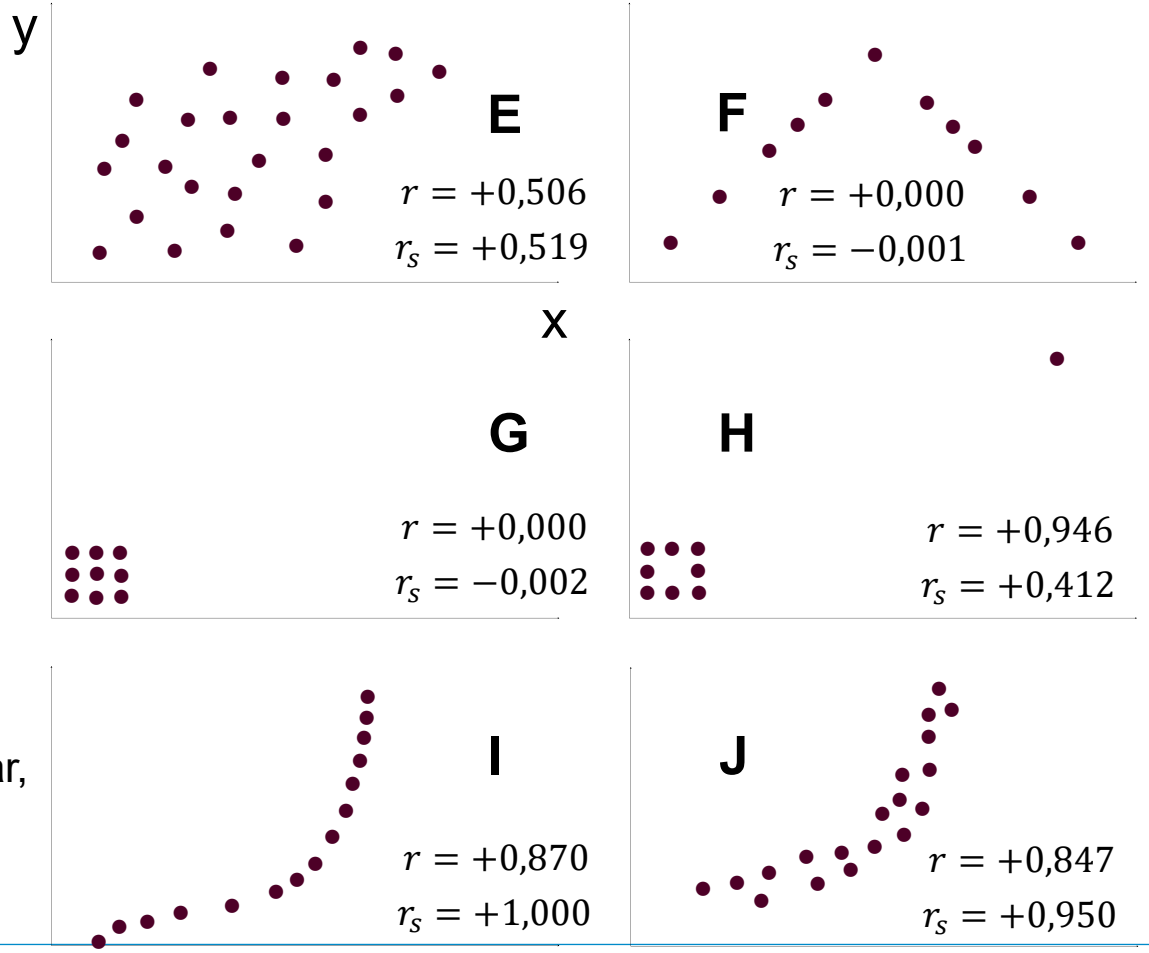
Rangkorrelation IV

– Vergleich der Korrelationskoeffizienten von Pearson und Spearman:



Rangkorrelation V

– Vergleich der Korrelationskoeffizienten von Pearson und Spearman:



Nichtlinear, aber monoton

Spearman:
⇨ geeignet für nichtlineare Zusammenhänge

V 7

7.4 Fehlschlüsse zu Korrelationen vermeiden!

Korrelation - Fehlschlüsse

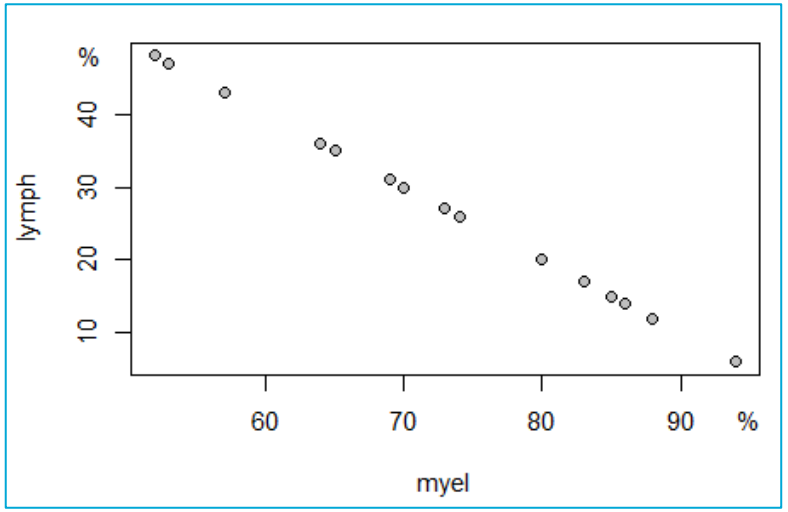
Korrelationen werden oft fehlinterpretiert, z. B.

1. formale (mathematische) Korrelation
2. durch Ausreißer bedingte Korrelation
3. durch Inhomogenitäten bedingte Korrelation
4. Gemeinschaftskorrelation
5. vorschnell kausale Schlussfolgerungen aus statistischer Assoziation

1. Formale Korrelation

– mathematische Korrelation: Beispiel: Differentialblutbild

i	Myeloische Blutzellen (%) x	Lymphatische Blutzellen (%) y
1	73	27
2	85	15
3	69	31
4	94	6
5	70	30
6	52	48
7	73	27
…	…	…
22	64	36
23	88	12
24	65	35
25	83	17
26	80	20
27	57	43
28	74	26
29	53	47
30	86	14



$$r = -1$$

Formale Korrelation entsteht hier durch Differenzbildung zu 100%.

1. Formale Korrelation

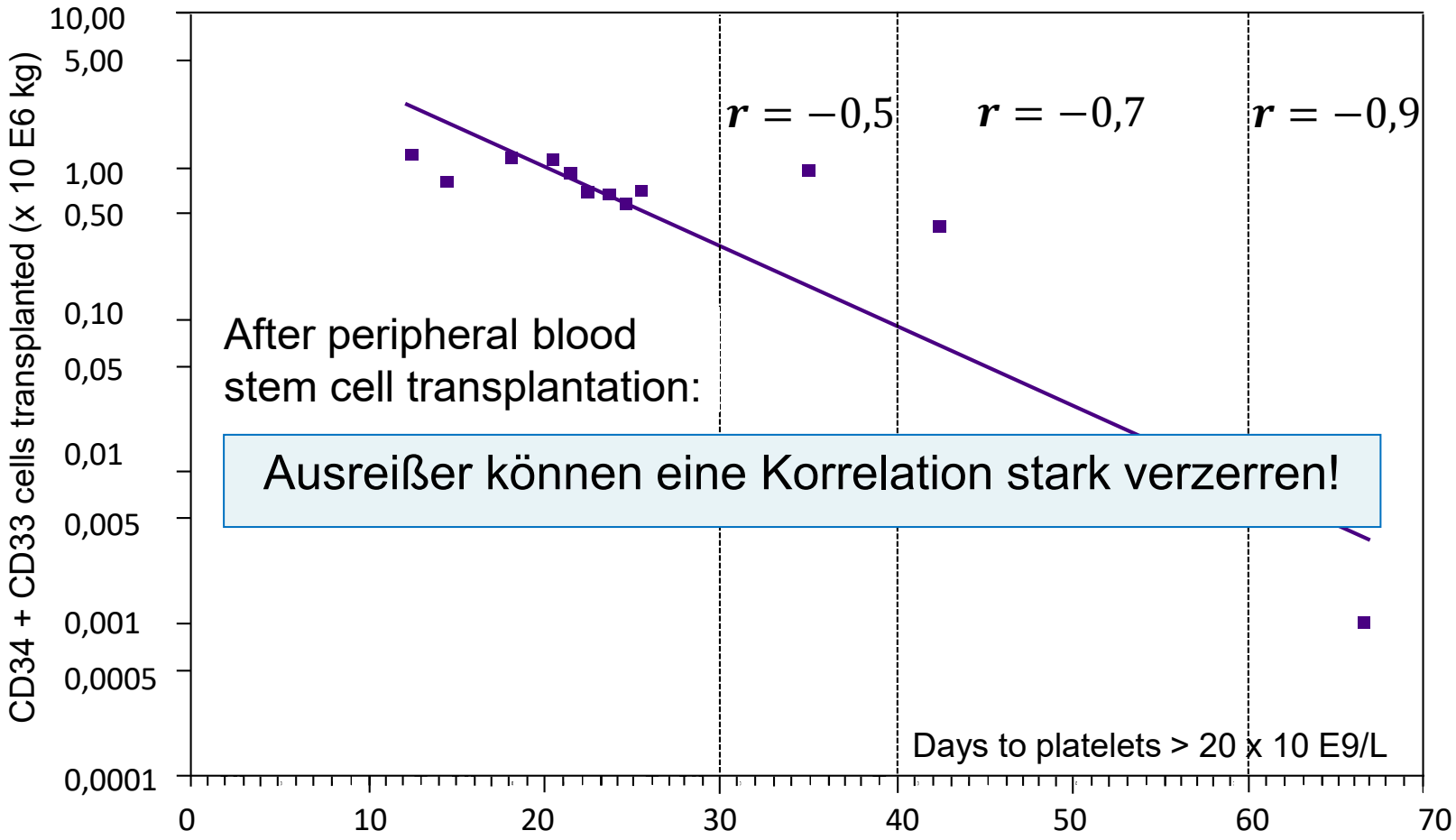
Mathematische Korrelation:

Szenario:

- Zwei und mehr Merkmale sind voneinander rechnerisch abhängig, wenn die Information des einen in die Berechnung der anderen eingeht.
- zum Beispiel: MDRD-Fomel
- $eGFR = 175 \cdot (Creatinine_{i.S.})^{-1.154} \cdot Age^{-0.203} \cdot (0.742 \cdot [female])$
  https://flexikon.doccheck.com/de/MDRD-Formel#cite_note-1 (2026-10-10)

⇨ Die Korrelation zwischen eGFR und Kreatinin ist deshalb kaum informativ.

2. Korrelation durch Ausreißer

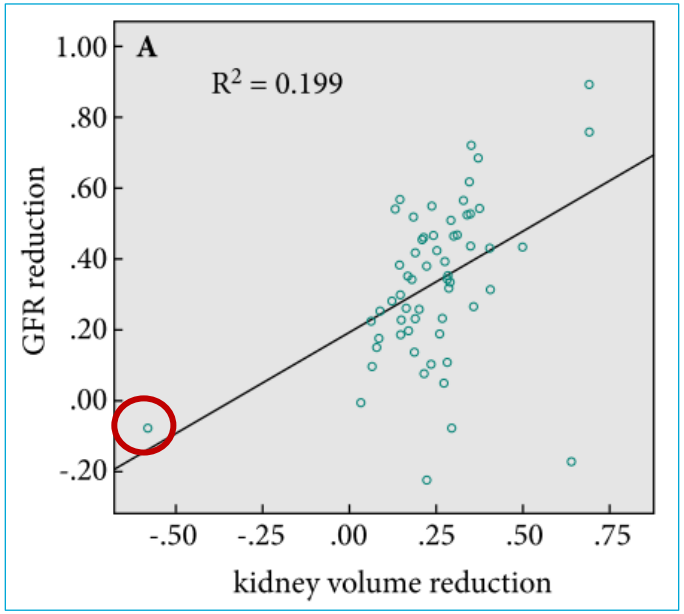


Literaturbeispiel 7.5

Ein krasses Beispiel falsch dargestellter und interpretierter Korrelation:

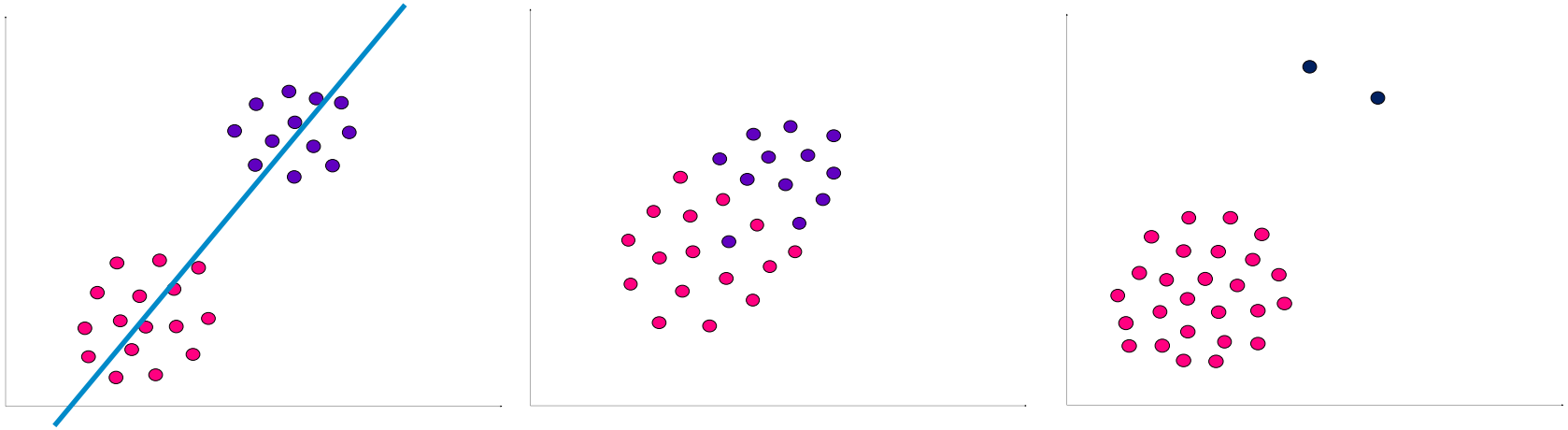
Li et al. BJU Int 2016;118: 610–617

- Artikel zur Teilresektion der Nieren
- Dargestellt wird im Scatterplot
  - das entfernte Nierenvolumen (X)
  - die Verringerung der GFR (Y)
- Was fällt Ihnen auf?
  -
  -



3. Korrelation durch Inhomogenitäten

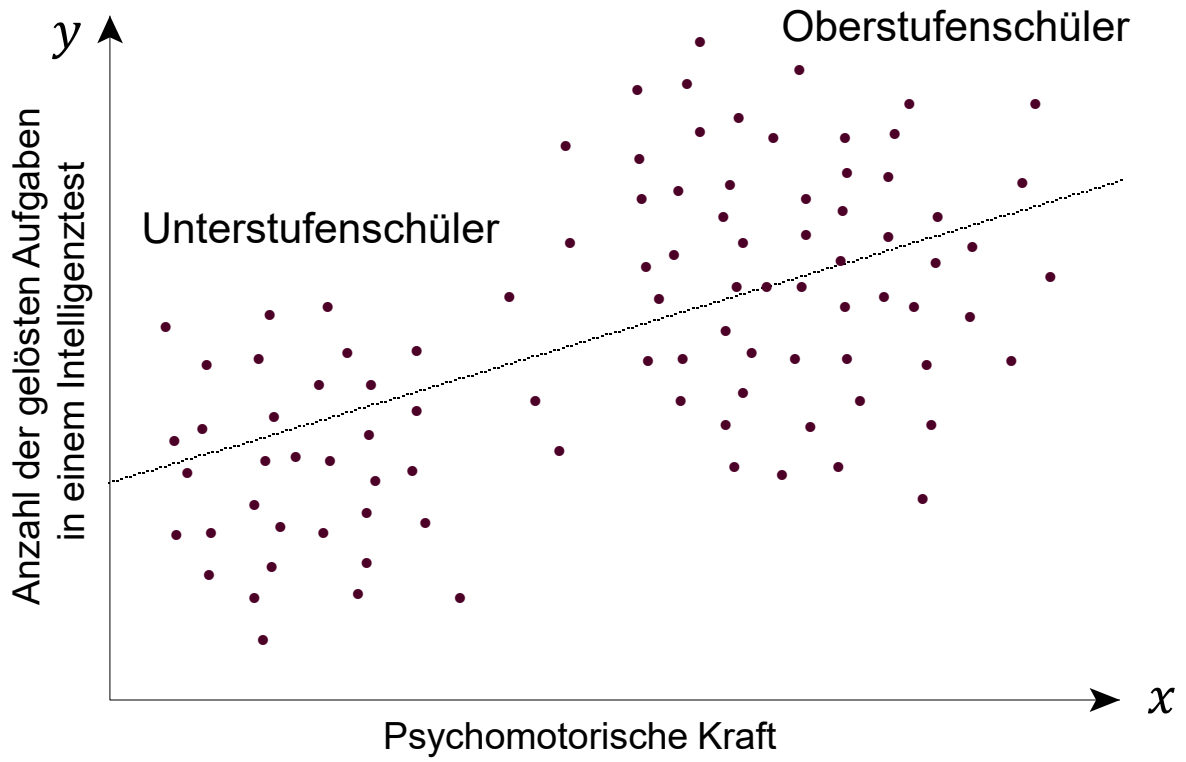
Szenario: Wertebereich ist in Subgruppen der Kohorte verschieden:



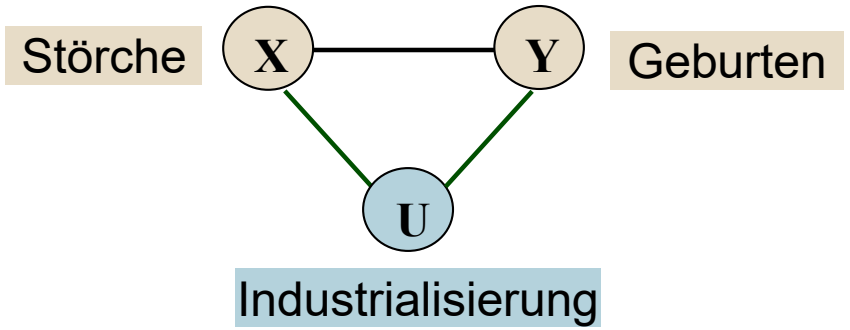
– Zusammenfassung der Subgruppen schafft eine Korrelation, die ggf. völlig anders ist als die Korrelation innerhalb der Subgruppe

3. Korrelation durch Inhomogenitäten

– Beispiel: Intelligenztest und Psychomotorische Kraft in zwei Schülergruppen



4. Gemeinschaftskorrelation



Szenario:

- Zwei Merkmale, zwischen denen keine kausale Beziehung besteht,
- werden entscheidend durch eine dritte Größe beeinflusst

Z. B.:

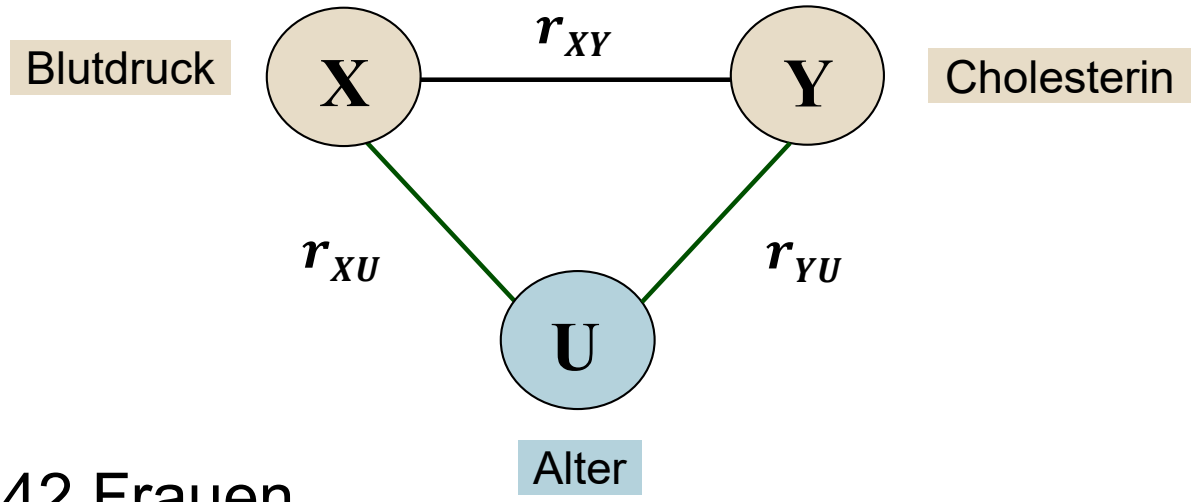
Anzahl der Störche (Merkmal X) korreliert mit dem Rückgang der Geburtenzahl (Merkmal Y) in Ostpreußen.

→ Beide Merkmale werden von der Industrialisierung (Merkmal U) beeinflusst.

4 b Partielle Korrelation I

Szenario:

– Die Korrelation (Assoziation) zwischen zwei metrischen Variablen wird durch eine dritte hervorgerufen.



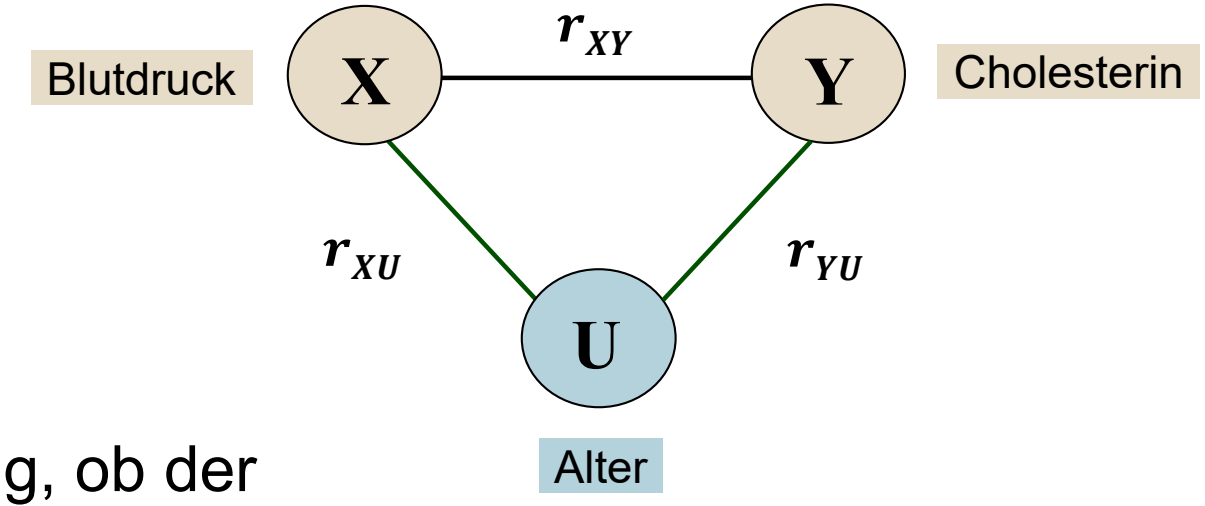
Beispiel 7.6:

– Daten von n = 142 Frauen

Partieller Korrelationskoeffizient

$$r_{XY/U} = \frac{r_{XY} - r_{XU} \cdot r_{YU}}{\sqrt{(1 - r_{XU}^2) \cdot (1 - r_{YU}^2)}}$$

Adjustierung



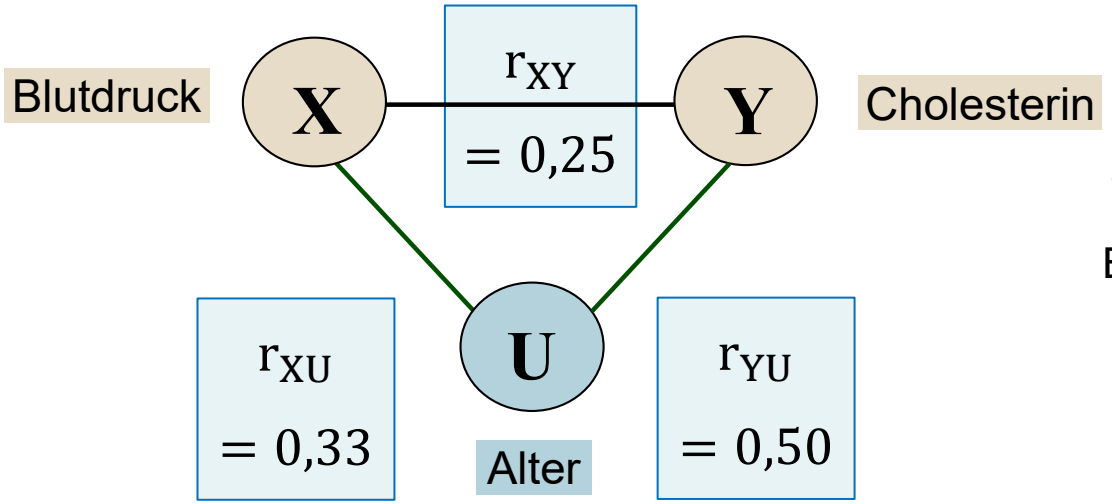
These adjustments proceed by reporting not what was observed in the real world, but instead modeling what would have been observed in an imaginary world in which specific nuisances and improprieties are absent.

Kaufmann JS. Amer J Law Med 2017;43: 193-206.

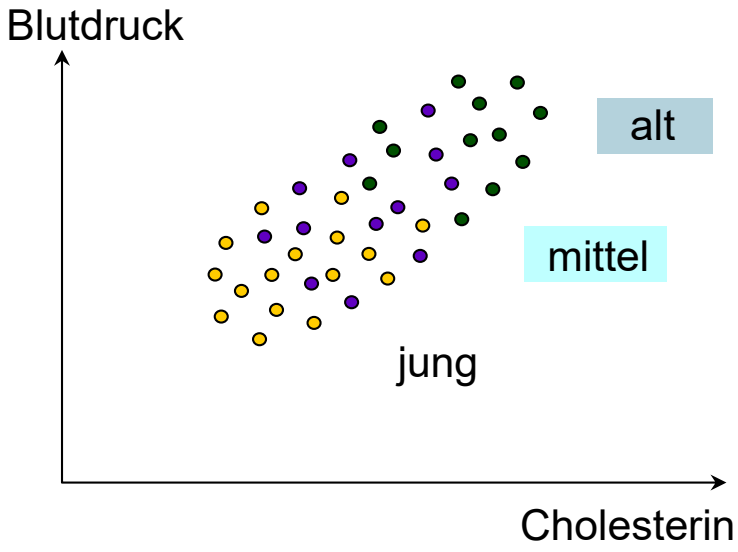
Idee:

- Prüfung, ob der Zusammenhang X – Y bedeutungslos wird, wenn auf die dritte metrische Größe U adjustiert wird (→ Lineare Modelle)

4 b Partielle Korrelation II



Skizze der Zusammenhänge:



$$r_{XY/U} = \frac{0{,}25 - 0{,}33 \cdot 0{,}50}{\sqrt{(1 - 0{,}33^2) \cdot (1 - 0{,}50^2)}} = \mathbf{0,1}$$

(aus Hartung, S. 562)

5. Korrelation → Kausalität??



Jotquadrat: Grafitto in Bonn. commons. wikipedia.org

- Patienten mit hohem NT-proBNP-Wert haben häufig Vorhofflimmern (atrial fibrillation – AF): starke Assoziation!
  (z. B. Werhahn et al. ESC Heart Failure 2022; 9: 100 – 109)
- Darf man deshalb sagen: Hohes NT-proBNP verursacht Vorhofflimmern?

Aus statistischer Assoziation darf höchstens dann eine kausale Wirkung abgeleitet werden, wenn alle anderen Ursachen ausgeschlossen werden können. (vgl. V 1 / V 14: Confounder)

→ V 1: randomisierte, kontrollierte klinische Studien

CAVE! Keine Zusammenhänge konstruieren



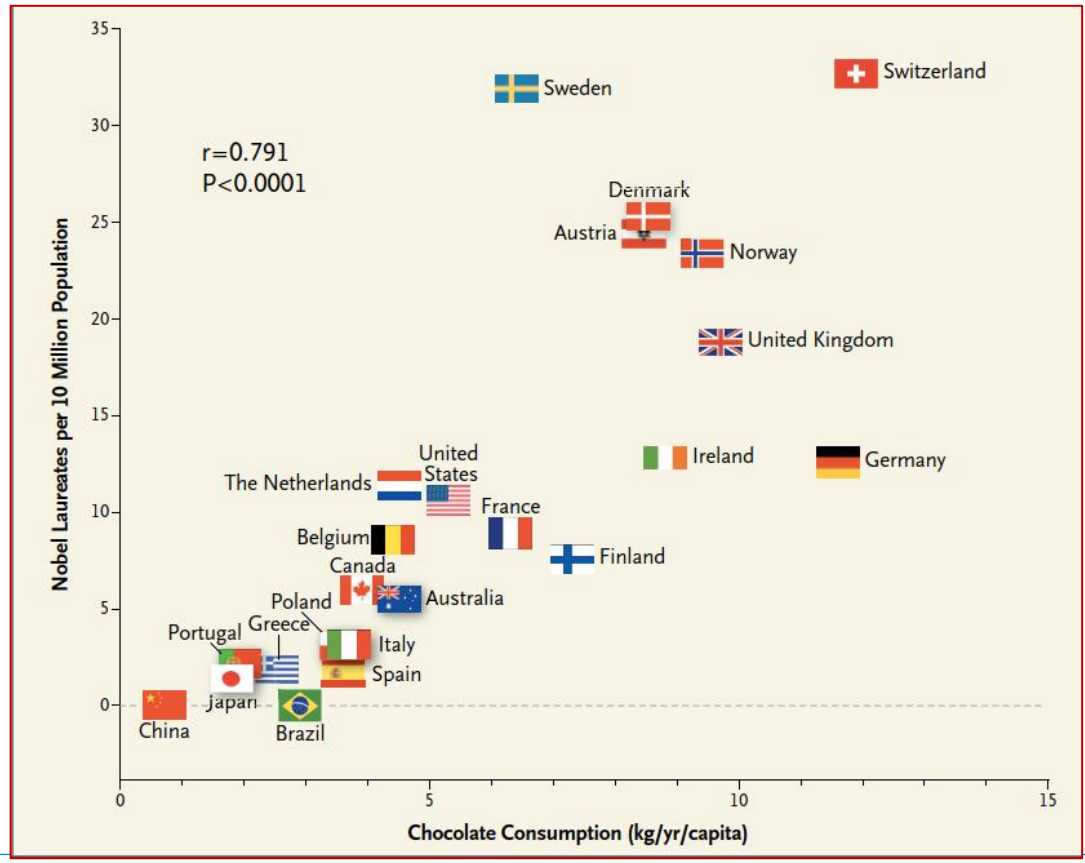
Mensch neigt dazu, Beobachtungen kausal zu interpretieren!

Die einfache Antwort auf diese Frage lautet, dass man, wenn man sich auf seine Intuition verlässt, häufig irrt, indem man ein Zufallsereignis als ein systematisches fehlinterpretiert.

Daniel Kahnemann. Schnelles Denken, langsames Denken.
Dt. Ausgabe: Siedler Verlag München 2011, S. 149

Messerli FH, NEJM 2012;367:1562-4

(Erwiderung: Maurage P et al. J Nutr. 2013;143(6):931-3)



V 7

7.5 Lineare Regression

Lineare Regression I

Lineare Korrelation

- Korrelationskoeffizienten messen die Stärke des Zusammenhangs zwischen 2 metrischen Variablen

Lineare Regression

- Verfahren, mit dem man die funktionale Form (Gleichung) eines linearen Zusammenhangs bestimmt.

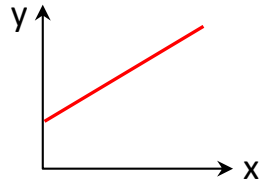
Lineare Regression II

Oft stellt sich die Frage nach der funktionalen Beziehung zwischen Zufallsvariablen: ⇨ Um wieviel ändert sich Y, wenn sich $X_1 \rightarrow X_2$ verändert?

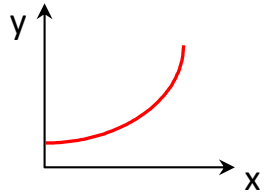
– $Y = f(X)$?

zum Beispiel:

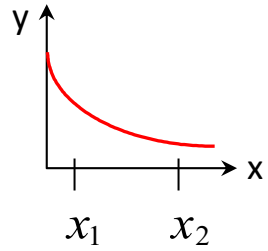
$y = a + b \cdot x$ linear



$y = a + b \cdot x + c \cdot x^2$ polynomial



$y = ae^{-bx}$ exponentiell

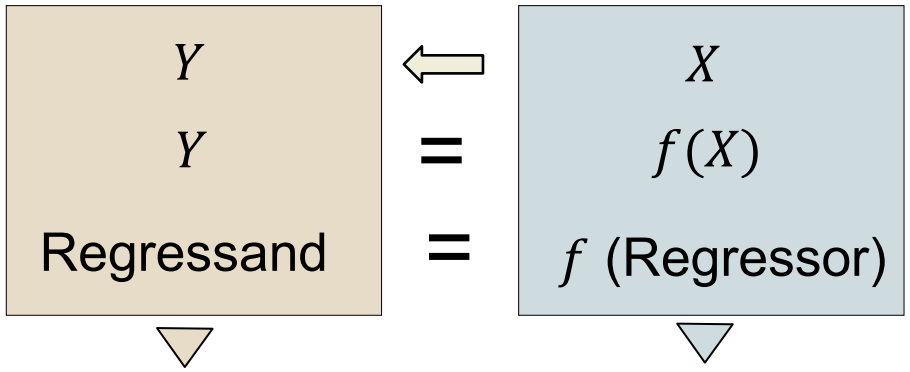


– Dabei heißen a, b, c Parameter / Koeffizienten

Lineare Regression III

Modellszenario:

- metrische Variablen X, Y von denselben Probanden
- gerichteter Zusammenhang von X → Y



- abhängig
- erklärte Größe
- Zielgröße
- Responsevariable

- unabhängig
- explanatorische Größe
- Prädiktor
- Kovariable

Ziele der (linearen) Regression

1. Prüfen eines funktionalen Zusammenhangs

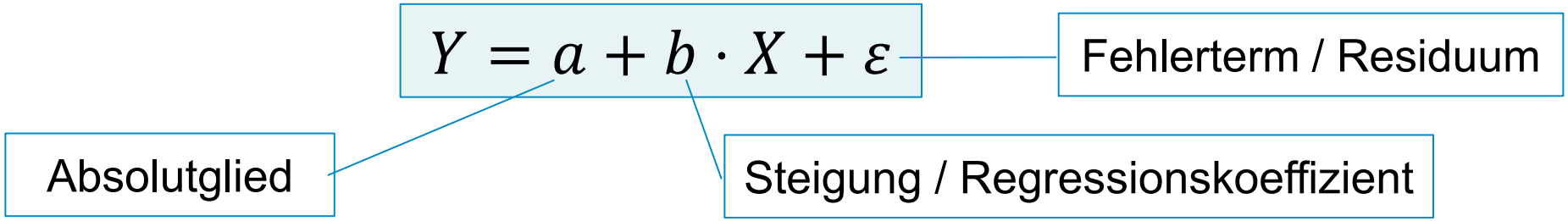
2. Quantifizieren des Zusammenhangs
   (= Schätzen von Parametern der funktionalen Beziehung)

3. Interpolation fehlender Werte

4. Vorsichtige Vorhersage (Prädiktion / Prognose)

Modellannahmen der linearen Regression

1. Die metrische Variable Y hängt von der metrischen Variablen X in einer linearen Beziehung ab:



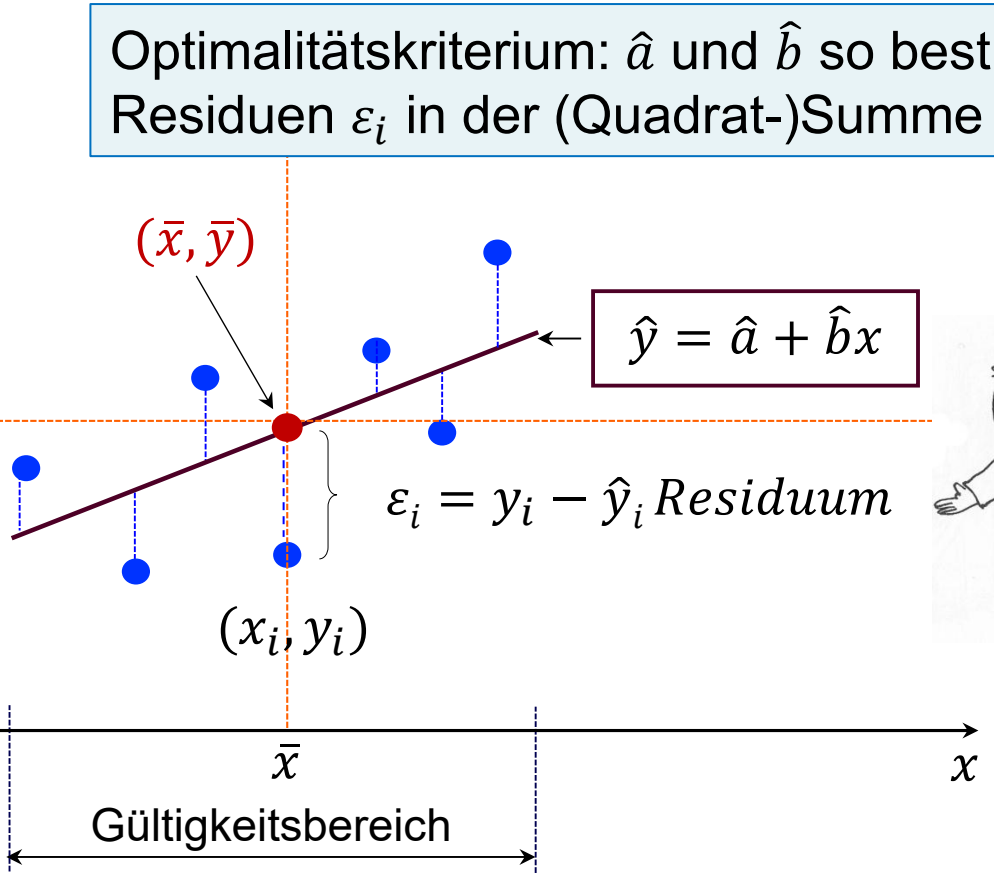
2. Die Parameter a und b haben feste (aber unbekannte) Werte.
3. Die Residuen sind normalverteilt mit Erwartungswert 0: $\varepsilon \sim N(0, \sigma^2)$
4. Die n Realisierungen von (x_i, y_i) für $i = 1, \ldots, n$ sind voneinander unabhängig.

Methode der Kleinsten Quadrate (MKQ) 1

Optimalitätskriterium: $\hat{a}$ und $\hat{b}$ so bestimmen, dass die Residuen ε_i in der (Quadrat-)Summe minimal werden.



Gonick & Smith:
Statistik in Cartoons





Gonick & Smith:
Statistik in Cartoons

Die Regressionsgerade verläuft immer genau durch den Schwerpunkt $(\bar{x}, \bar{y})$.

MKQ 2

- Angenommen, wir hätten optimale Parameter, $\hat{a}$ und $\hat{b}$ gefunden:

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

- Dann minimieren die Schätzwerte $\hat{a}$ und $\hat{b}$ die Summe der Abweichungsquadrate (Residuen):

$$\sum_{i=1}^{n} \left[y_i - (\hat{a} + \hat{b} \cdot x_i) \right]^2 \rightarrow Min!$$

- C. F. Gauß veröffentlichte 1809 die Lösung im Band 2 von *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (Theorie der Bewegung der Himmelskörper, welche in Kegelschnitten die Sonne umlaufen)

Schätzung der Parameter der linearen Regression

Kovarianz Summe der Quadrate

Berechne: $\bar{x}, \bar{y},$ $SP\ (X,Y),$ $SQ(X), SQ(Y)$

Berechne: $\hat{b} = \frac{SP(X,Y)}{SQ(X)};\ \hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$

Optimal angepasstes geschätztes Regressionsmodell:

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

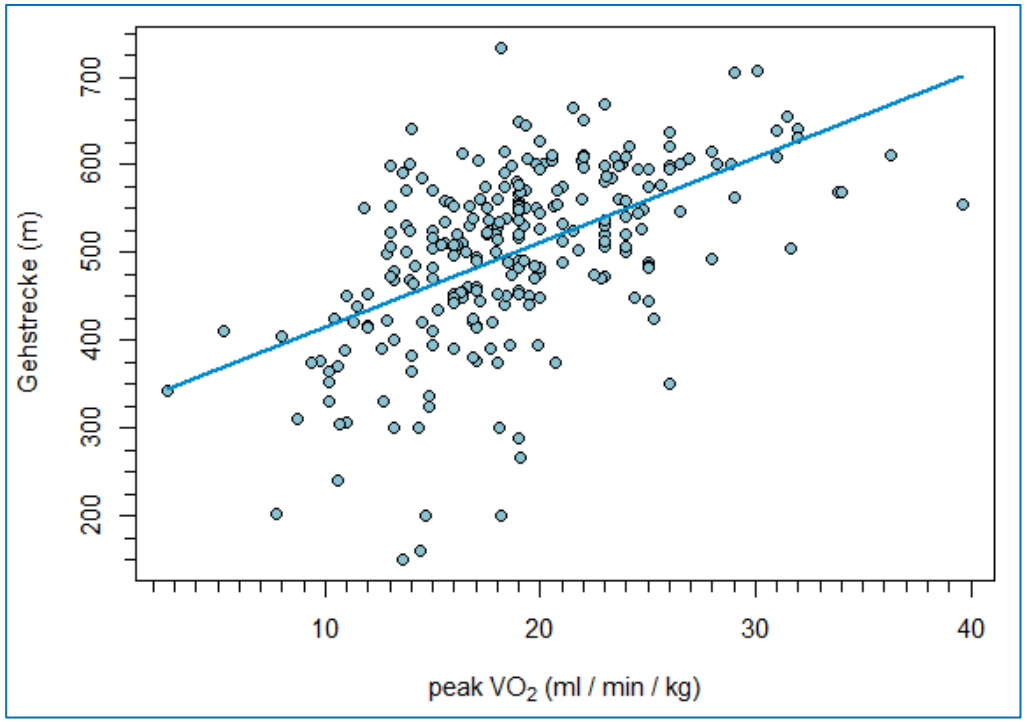
x ist gegeben / die unabhängige Variable

regression = (lat.) Zurückgehen

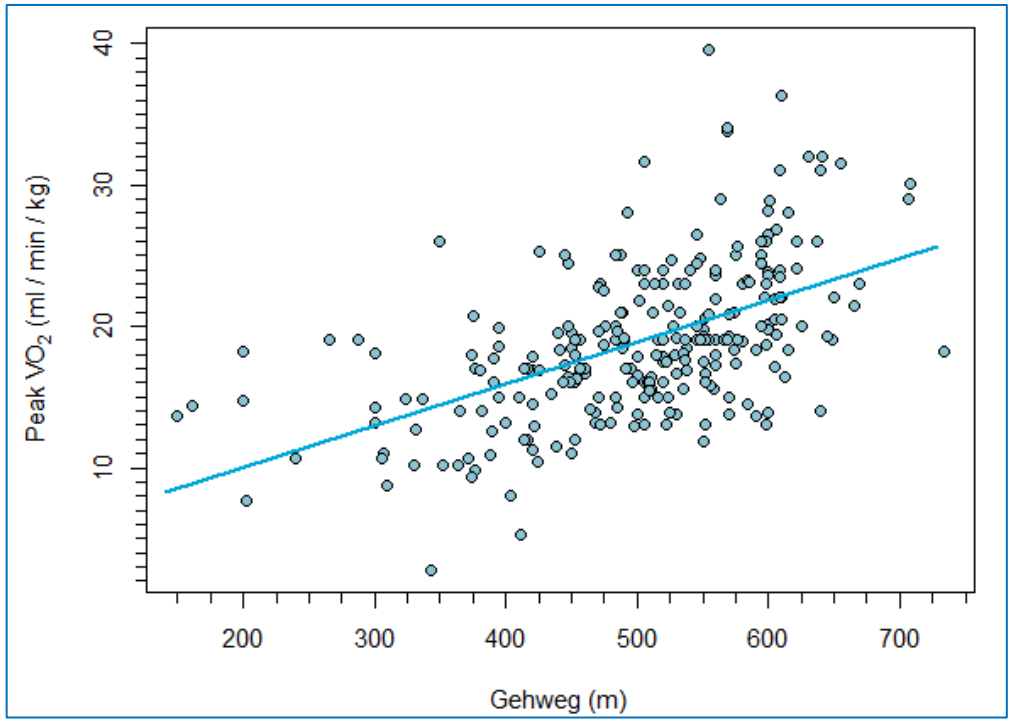
Regression von Y auf X von X auf Y

Die Rollen – Zielvariable ~ unabhängige Variable – müssen a-priori klar sein!

- Y wird auf X zurückgeführt
- $\hat{y} = \widehat{a_{yx}} + \widehat{b_{yx}} \cdot x$



- X wird auf Y zurückgeführt
- $\hat{x} = \widehat{a_{xy}} + \widehat{b_{xy}} \cdot y$



Lineare Regression: *Prädiktion* und Interpretation

Prädiktion:

– Die Regressionsgleichung: $\hat{y} = \hat{a}_{yx} + \hat{b}_{yx} \cdot x$ berechnet für vorgegebene x-Werte jeweils einen Wert $\hat{y}$.

Interpretation:

– Der Regressionskoeffizient b_{yx} gibt an, um wie viele Einheiten Y im Mittel zunimmt bzw. abnimmt, wenn X um eine Einheit zunimmt:

– Angenommen, x erhöhe sich um eine Einheit von einem Wert x_0 auf einen Wert $x_1 = x_0 + 1$, …

$\widehat{y_0} = \hat{a} + \hat{b} \cdot x_0$ $\qquad\qquad$ $\widehat{y_1} = \hat{a} + \hat{b} \cdot (x_0 + 1) = (\hat{a} + \hat{b} \cdot x_0) + \hat{b}$,

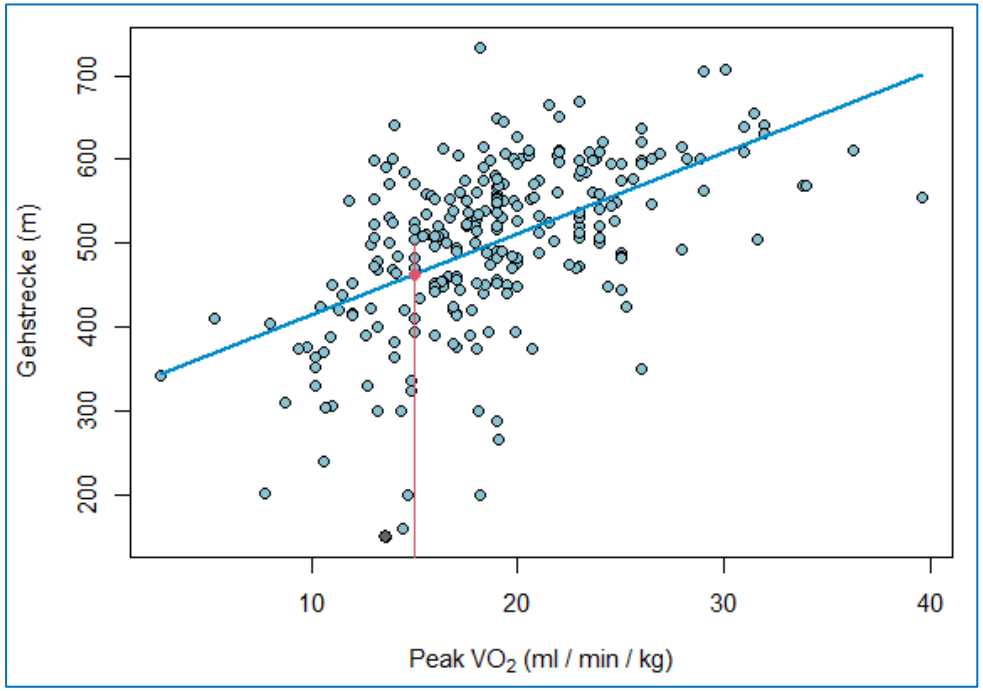
– dann erhöht sich $\widehat{y_0}$ um $\hat{b}$ Einheiten.

CAVE! *Prädiktion?*



Frau X, wenn Sie durch Training Ihre peak VO_2 auf 15 ml / min / kg steigern könnten, würde sich Ihre Gehstrecke auf 463 m verbessern!

- Die Prognose gilt nur, wenn das Modell *wahr* wäre.
- Es besteht eine erhebliche Variation: Die Gehstrecke, die für Frau X vorhergesagt wird, ist ca. 200 m geringer!

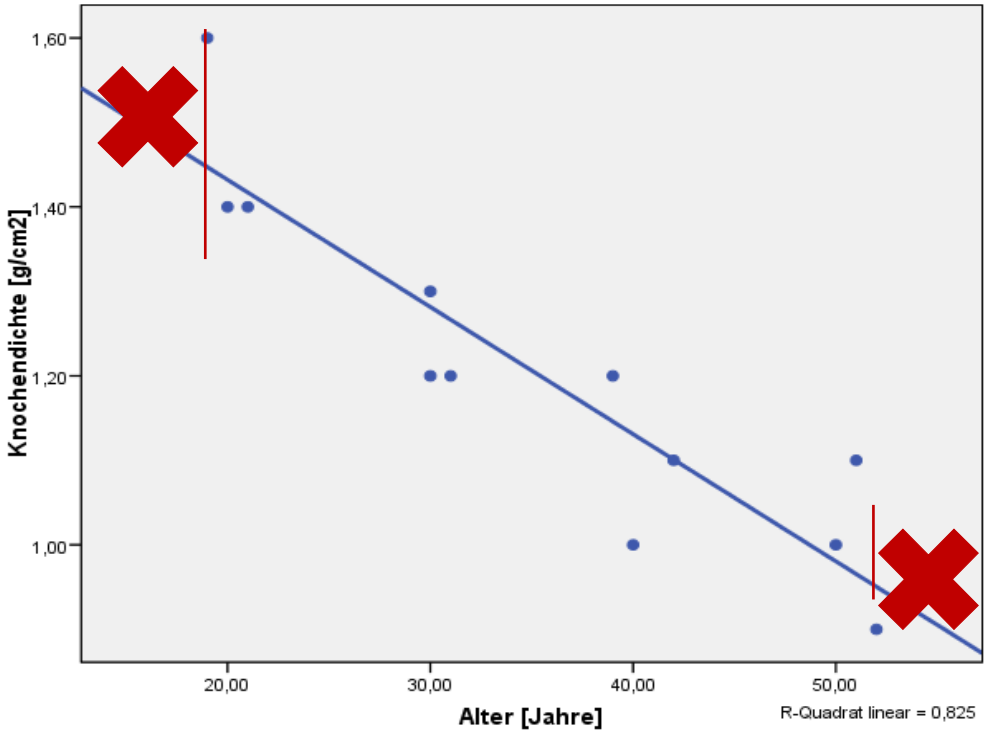


Statistik gewinnt aus vielen Daten Aussagen für Kohorten, aber nicht für Einzelpersonen!

CAVE! *Prädiktion?*



– Beispiel 7.5: Zusammenhang zwischen Alter und Knochendichte:



$$Knochendichte = 1{,}733 - 0{,}015 \cdot Alter$$

Regressionskoeffizient: $\hat{b} = -0{,}015$,
d.h. pro Altersjahr nimmt im Mittel die Knochendichte um $0{,}015\ g/cm^2$ ab.

$(r = -0{,}9;\ p = 0{,}00004)$

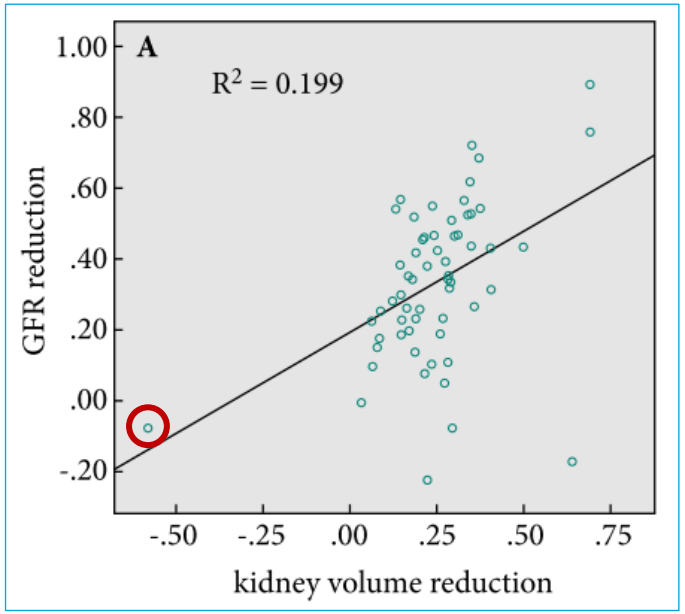
Extrapolation über den Beobachtungsbereich hinaus ist nicht erlaubt.

Empfindlich gegen Ausreißer



Beispiel 7.5 (Li et al. BJU Int 2016;118: 610–617)

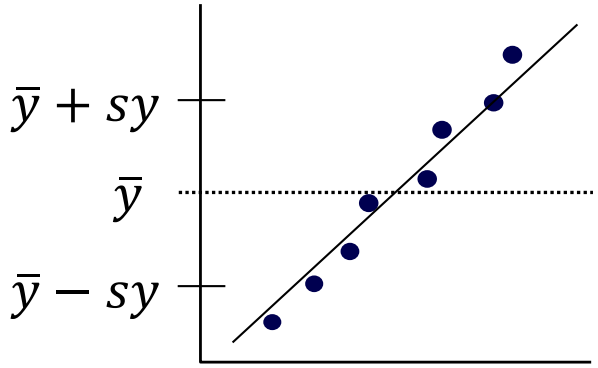
- Nierenteilresektion
- ein offensichtlich falscher Wert (negatives Resektionsvolumen) dominiert die lineare Regression
- ohne Ausreißer kaum noch eine Richtung erkennbar
- Ist hier Lineare Regression überhaupt ein valides Verfahren?



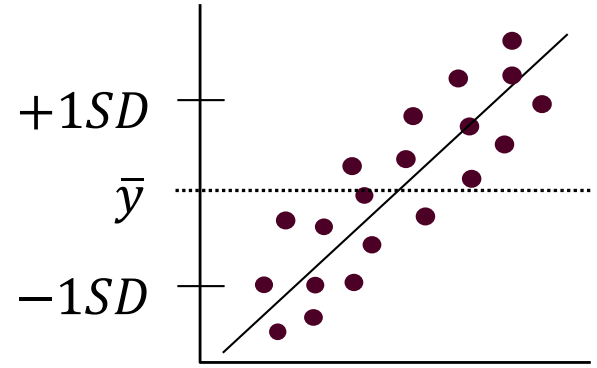
Insbesondere bei kleinen Stichproben können einzelne extreme Werte die Parameterschätzer maßgeblich verändern („Hebelwirkung“).

Güte einer linearen Regression I

Gütemaß: das Bestimmtheitsmaß B (coefficient of determination)



- wenig Abweichung von der Geraden
- geringe Restvarianz
- $\sum e_i^2$ ist klein gegenüber $SQ(y)$
- B ist groß



- größere Restvarianz
- B ist deutlich kleiner

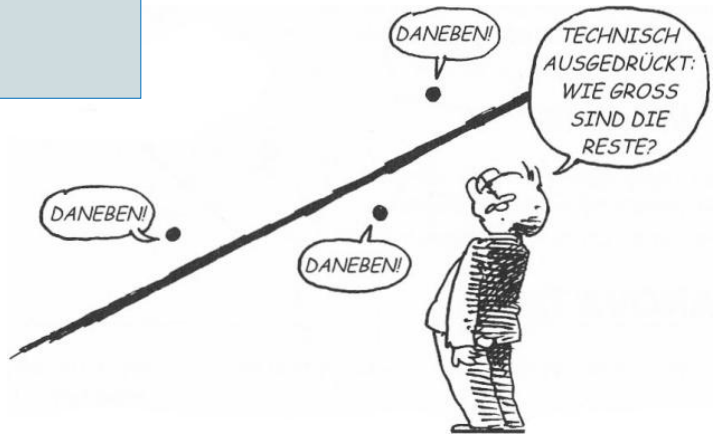
Güte einer linearen Regression II

- Das Bestimmtheitsmaß ist das Quadrat des Korrelationskoeffizienten nach Pearson:

$$B = R^2 = 1 - \frac{\text{Quadratsumme der Residuen}}{\text{Gesamtvarianz in Y}}$$

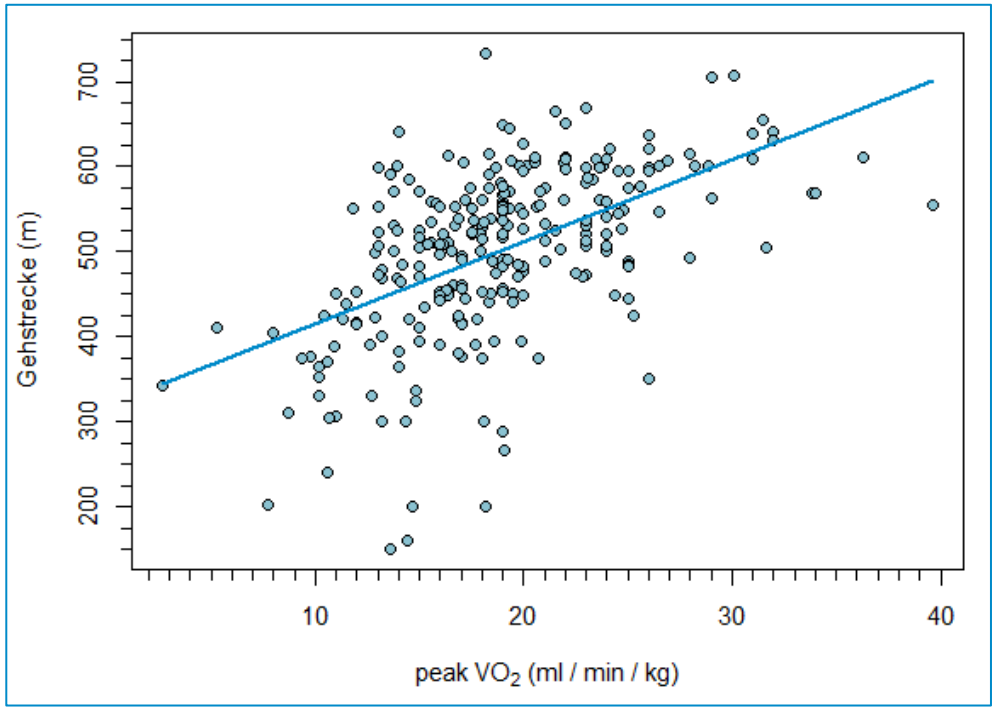
$$R^2 = 1 - \frac{\sum_{i=1}^{n} {e_i}^2}{SQ(y)}$$

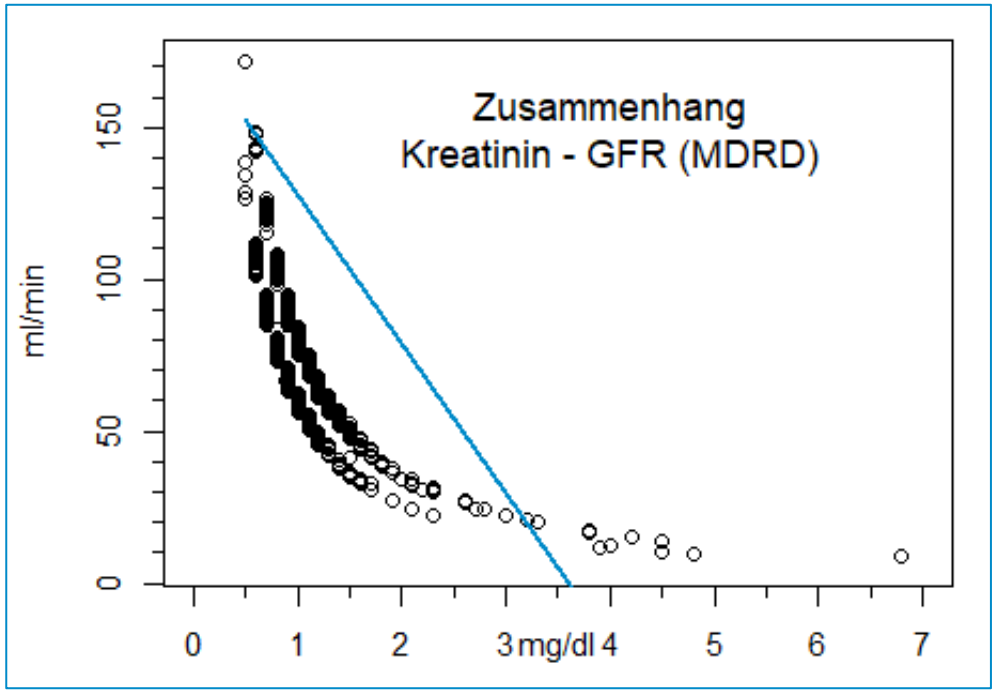
- R^2 gibt an, wieviel Streuung der Zielvariablen Y durch das Modell erklärt wird.
- mathematisches, kein fachwissenschaftliches Gütekriterium für ein Modell



Gonick & Smith: Statistik in Cartoons

Bestimmtheitsmaß – Beispiele





Ex-DHF-Studie (vgl. Folie 27)

$R^2 = 0{,}29$; relativ gut!

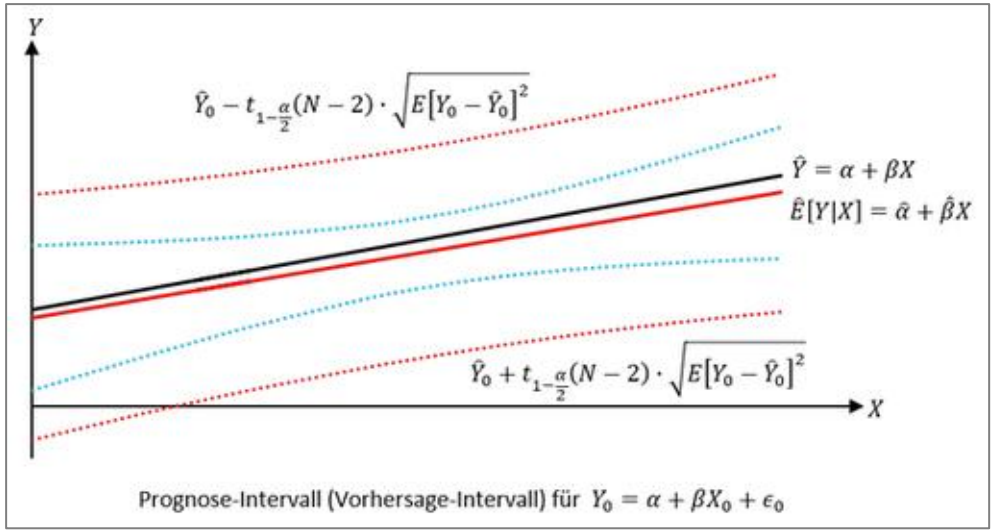
B 7.2 Diast-CHF-Studie (vgl. Folie 8)

$R^2 = 0{,}42$; eigentlich hervorragend, aber Modell hier inadäquat

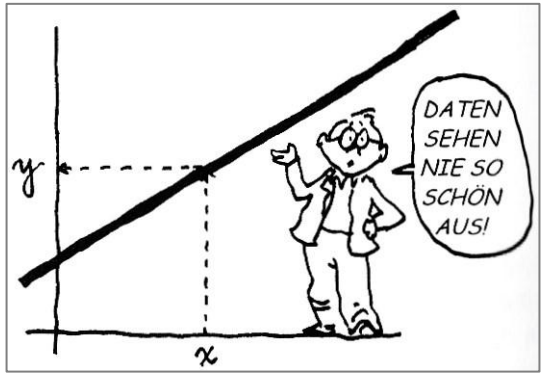
Konfidenz- und Prognoseintervall



- Das 95%-Konfidenzintervall der Regressionsgerade beschreibt, wie gut man einen mittleren Wert y für einen festen Wert x vorhersagen kann.
- Das 95% - Prognoseintervall beschreibt, wie gut man einen neuen Wert y aus einem beliebigen Wert x vorhersagen kann, ohne ihn zu messen.



https://www.fernstudium4you.de/prognoseintervall.html (2026-10-08)



Gonick & Smith: Statistik in Cartoons

Motivation für multiple lineare Regression



Häufig gibt es mehrere Merkmale, welche potenziell die Zielgröße beeinflussen. Das Szenario erfordert also, mehrere unabhängige Variablen ins lineare Modell zu einzuschließen: multiple lineare Regression, um

- z. B. auf bekannte Confounder (Störgrößen) zu adjustieren
- ein Modell zu erhalten, das besser an die Daten angepasst ist und
- die Power von Tests (→ V 9) der Koeffizienten zu erhöhen.

Beispiel: Ex-DHF, die 6-Minuten-Gehstrecke hängt u. U. auch von der NYHA-Klasse und der Trainings-Intervention ab:

6-Min.-Gehstrecke ~ peak VO_2 + NYHA + Exercise Intervention

Multiple lineare Regression II

- Die SPSS-Ausgabe zeigt die Abhängigkeit der Gehstrecke von der peak VO_2 und von der NYHA-Klasse.

Parameterschätzungen

Abhängige Variable: F12: Gehtest ja: Gehstrecke

Parameter	Regressionskoeffizient B	Std.-Fehler	T	Sig.	95% Konfidenzintervall Untergrenze	95% Konfidenzintervall Obergrenze	Partielles Eta-Quadrat
Konstanter Term	291,485	20,937	13,922	<,001	250,254	332,717	,432
peakVO2	9,008	,994	9,061	<,001	7,050	10,965	,244
Exercise Training	-5,063	10,450	-,485	,628	-25,644	15,517	,001
Usual Care	0[a]	.	.	.	.	.	.
NYHA II	47,305	16,263	2,909	,004	15,278	79,332	,032
NYHA III	0[a]	.	.	.	.	.	.

a. Dieser Parameter wird auf null gesetzt, da er redundant ist.

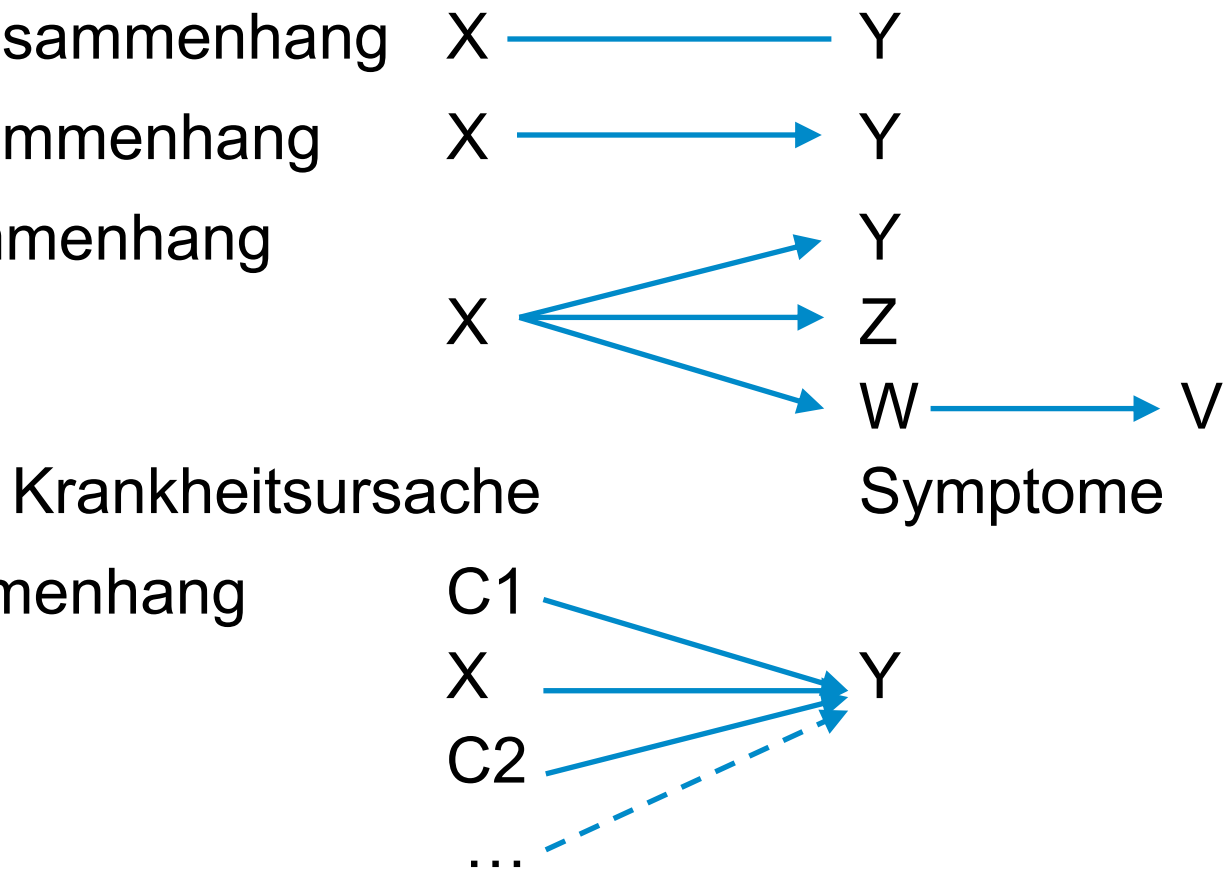
- Interpretation:
  - Konstante kaum interpretierbar
  - peak VO_2 je 1 ml / min / kg ist die Gehstrecke im Mittel um 9 m höher,
  - NYHA II im Mittel 47 m mehr als Patienten mit NYHA III
  - Ex.Training im Mittel 5 m weniger als Probanden im Kontrollarm

Herausforderungen multipler Modelle

- Medizinisch-biologische Erwägungen haben das Primat bei der Modellwahl.
- Modellannahmen (z. B. Linearität) müssen geprüft werden,
- Parameterschätzungen setzen ein gültiges Modell voraus;
  i.d.R. ist aber das beste Modell unbekannt.
- Algorithmen (automatische Variablenauswahl oder R^2) finden ein mathematisch optimales Modell, das fachwissenschaftlich unsinnig sein kann!

Zusammenfassung I

1. Ungerichteter Zusammenhang
2. Gerichteter Zusammenhang
3. Indirekter Zusammenhang
4. Multipler Zusammenhang



Zusammenfassung II

- Korrelation beschreibt Richtung und Stärke des Zusammenhangs,
- Regression den funktionalen Zusammenhang
- Rangkorrelation, falls Ausreißer bzw. bei ordinalen Merkmalen
- für lineare Regression vorher Linearität prüfen (Scatterplot!)
- Regressionskoeffizient gibt an, um wie viele Einheiten Y im Mittel zunimmt / abnimmt, wenn X um eine Einheit zunimmt
- komplexe Zusammenhänge durch multiple Regressionsmodelle beschreibbar



UNIVERSITÄT LEIPZIG

Medizinische Fakultät

VIELEN DANK!

Dr. rer. nat. Meinhard Mende
Senior-Wissenschaftler,
Institut für Medizinische Informatik, Statistik und Epidemiologie
Härtelstraße 16-18, 04107 Leipzig

Meinhard.mende@zks.uni-leipzig.de