Der Umgang mit fehlenden Werten in epidemiologischen und Versorgungssforschungsstudien

Oliver Kuß

Institut für Medizinische Epidemiologie, Biometrie und Informatik, Medizinische Fakultät, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale)

Inhalt

- Einleitung
- Mechanismen für fehlende Werte
- Umgang mit fehlenden Werten
- Fazit

Was ist ein fehlender Wert?

Ein fehlender Wert ist **kein …**

- … gruppierter, aggregierter oder zensierter Wert
- … Wert, der aus inhaltlichen Gründen nicht vorkommen kann (Alter bei erster Schwangerschaft bei einem Mann, Antworten auf Skip-Question in Fragebögen)
- … Werte, bei dem auch ein Nicht-Vorliegen eine Information enthält ("Weiß ich nicht", "Trifft für mich nicht zu")

Ein fehlender Wert ist ein Wert, der eigentlich vorhanden sein sollte …

Ursachen für fehlende Werte

- Untersucher/Befrager (schlecht instruiert, Überlastung, …)
- Instrumente (unklare Fragen, unpassende Antworten, …)
- Proband/Patient (mangelnde Compliance, Scham, ..)
- Dateneingabe
- Auswertung (Division durch Null, …)
- Sonstiges (Datenverlust durch EDV, Fehler beim Postversand, …)

Mechanismen für fehlende Werte

Der zugrundeliegende Mechanismus, der zu fehlenden Werten geführt hat,

… bestimmt den Grad der Verzerrung der Ergebnisse

… bestimmt die korrekte Auswahl von Methoden zum Umgang mit fehlenden Werten

3 Mechanismen:

1. MCAR (Missing completely at random)
2. MAR (Missing at random)
3. MNAR (Missing not at random)

Mechanismen für fehlende Werte

Beispiel: Blutdruckmessung an zwei Zeitpunkten
Zum ersten Zeitpunkt (X) wird dieser bei N=30 Patienten gemessen, zum zweiten (Y) bei N=10 Patienten, für 20 Patienten haben wir also einen fehlenden Wert für Y.

Mechanismen für fehlende Werte: MCAR

Das Auftreten eines fehlenden Wertes in der Variable Y ist **unabhängig** vom

… tatsächlichen Wert von Y UND

… von allen anderen Variablen im Datensatz

Die Patienten mit fehlenden Werten sind eine zufällige Stichprobe aus allen Patienten.

Beispiel:
Die Patienten bei der zweiten Blutdruckmessung wurden zufällig ausgewählt.

Mechanismen für fehlende Werte: MAR

Das Auftreten eines fehlenden Wertes in der Variable Y kann **vollständig** durch die Ausprägungen aller anderen Variablen im Datensatz erklärt werden

Beispiel:
Zur zweiten Blutdruckmessung wurden nur die 10 Patienten einbestellt, die bei der ersten Messung einen Blutdruck >140 (X>140) hatten.

Unglückliche Bezeichnung:
Besser wäre „conditionally missing at random“ o.ä.

Mechanismen für fehlende Werte: MNAR

Das Auftreten eines fehlenden Wertes in der Variable Y ist **abhängig** vom

… tatsächlichen Wert von Y UND

… kann nicht durch andere Variablen im Datensatz erklärt werden

Beispiel:

Bei der zweiten Blutdruckmessung wurde der Blutdruck nur bei den 10 Patienten registriert, die einen Wert >140 (Y>140) hatten.

Mechanismen für fehlende Werte: Andere Bezeichnungen

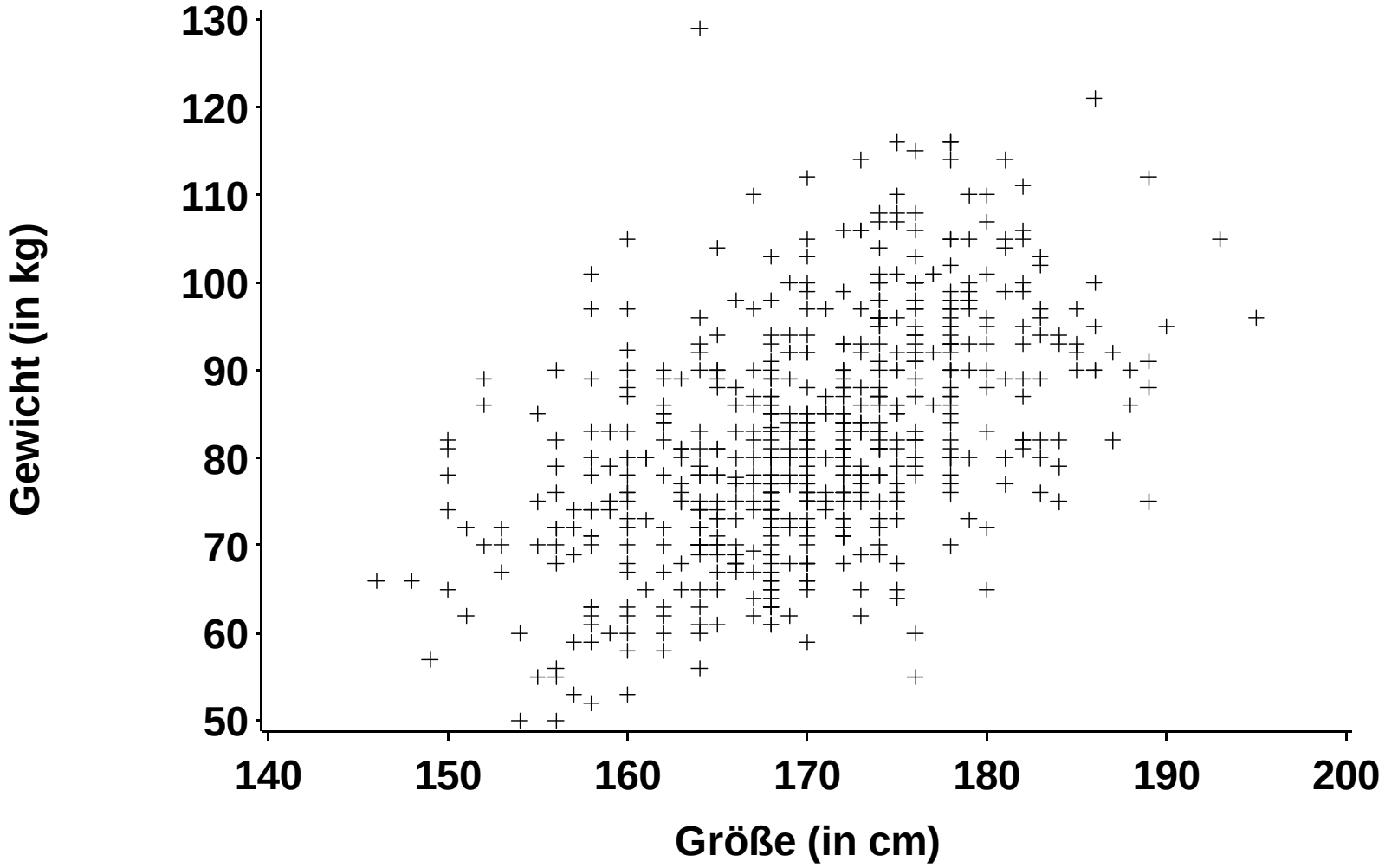
MCAR	Ignorable	Non-informative
MAR		
MNAR	Non-ignorable	Informative

Umgang mit fehlenden Werten

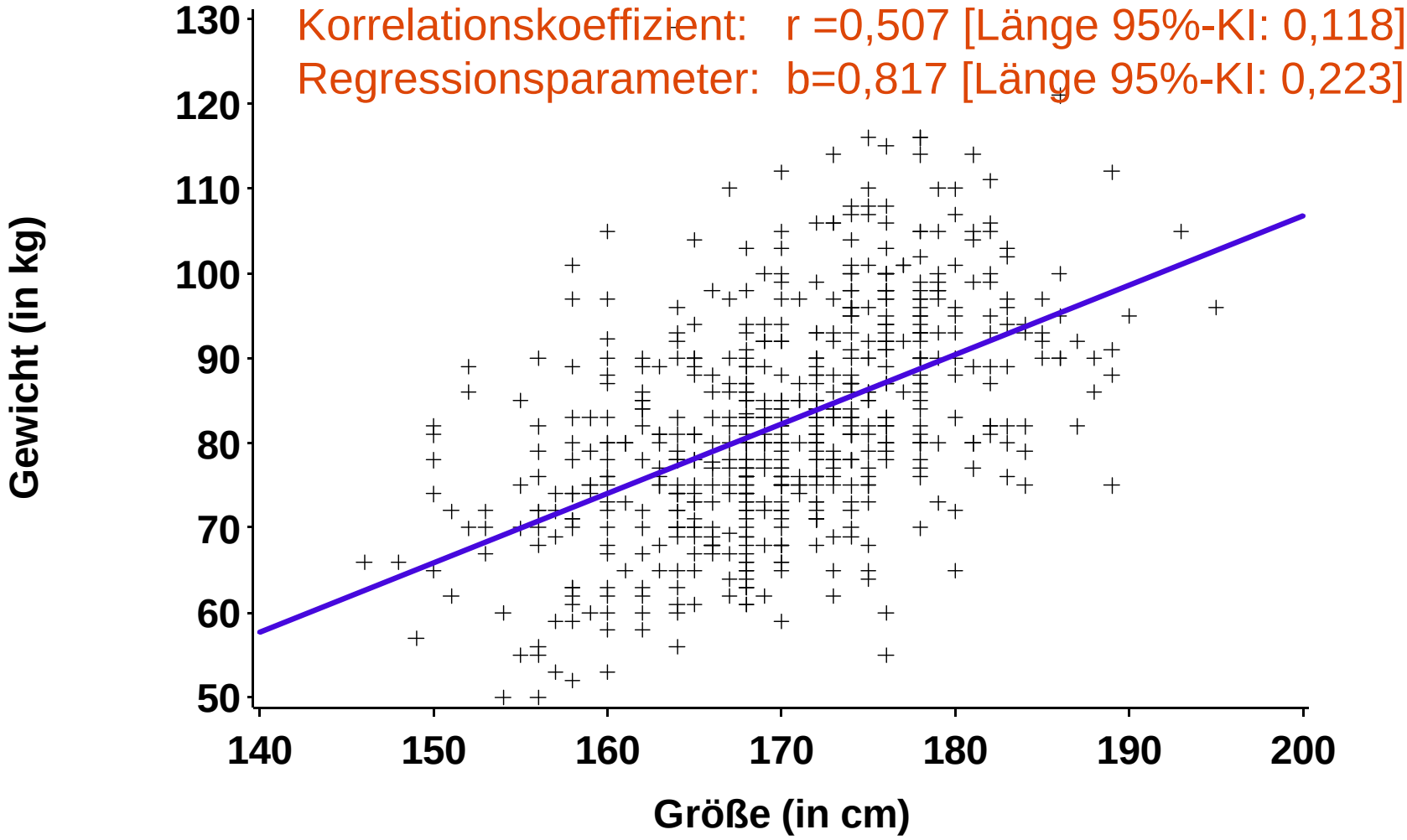
- Complete case analysis
- Mean imputation
- Regression imputation
- Multiple imputation

Beispiel: Zusammenhang zwischen Größe (X) und Gewicht (Y) bei N=604 Patienten aus einer randomisierten Studie zum Vergleich dreier OP-Techniken in der Bypass-Chirurgie.

Umgang mit fehlenden Werten: Beispiel



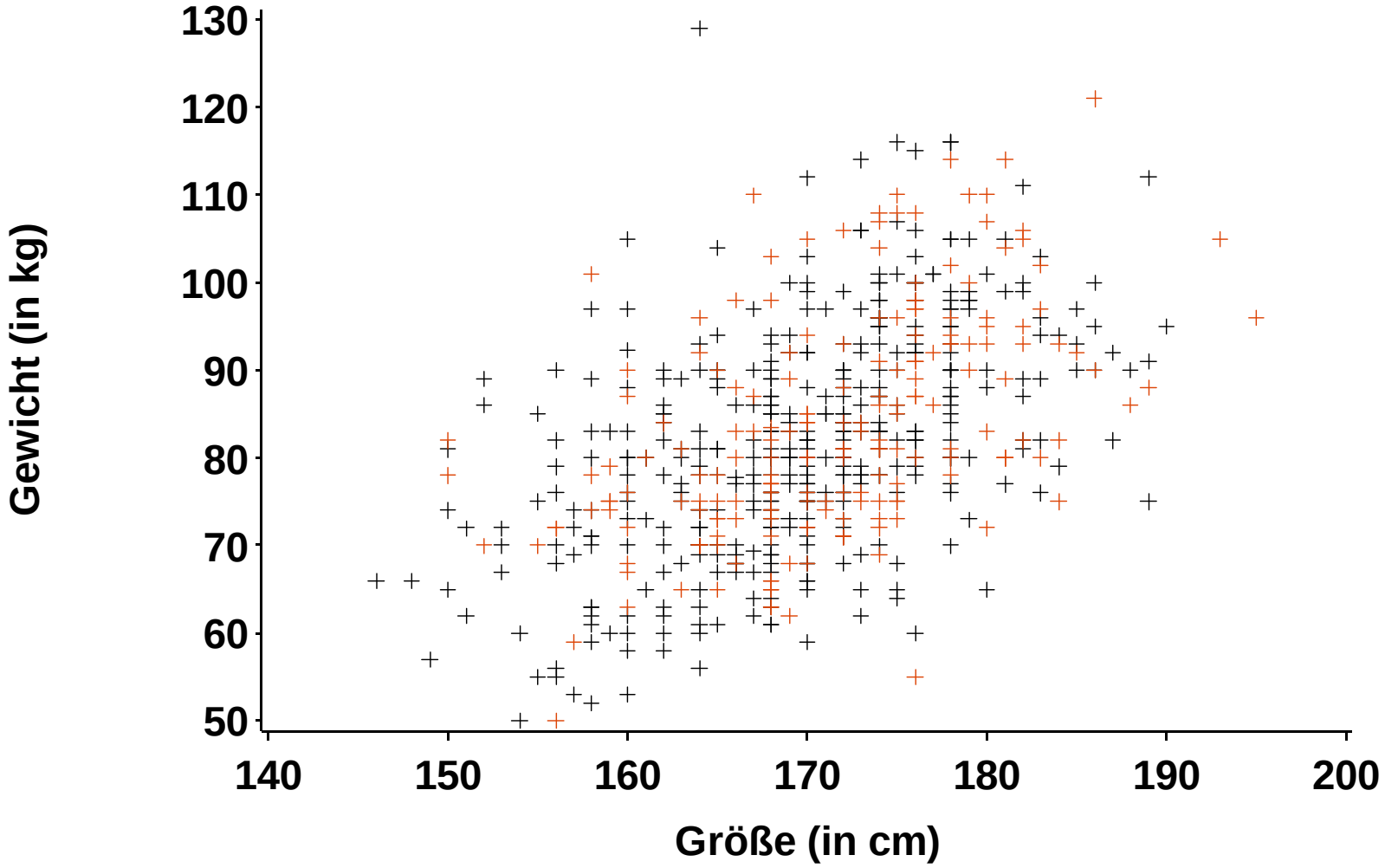
Umgang mit fehlenden Werten: Beispiel



Umgang mit fehlenden Werten: Beispiel

- Lösche zufällig ein Drittel der Werte auf der x-Achse (z.B. Teile des Studienpersonals waren nicht unterrichtet, dass eine Größenmessung statt finden soll)
- Neuer Stichprobenumfang N=410
- Daten sind MCAR.

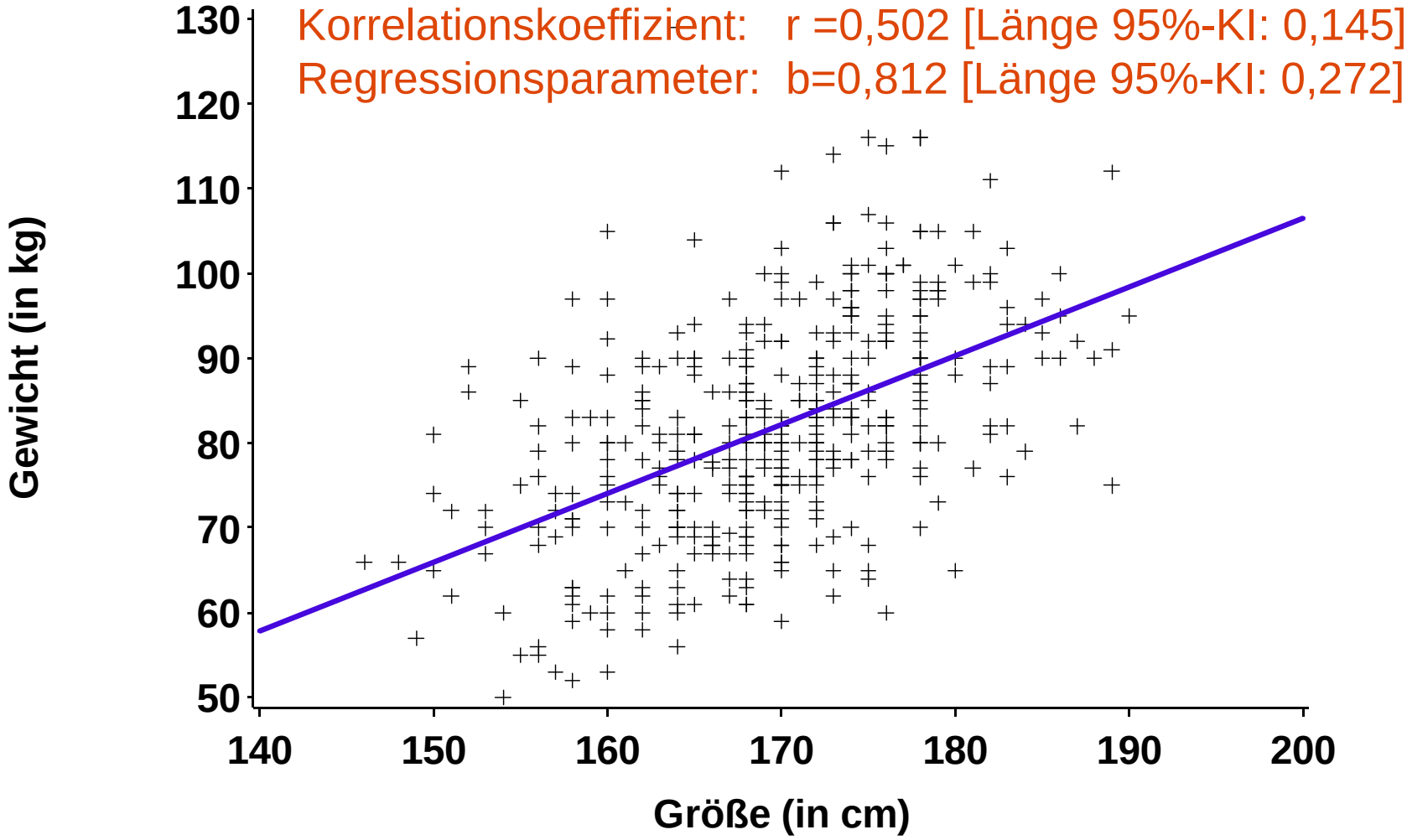
Umgang mit fehlenden Werten: Beispiel



Umgang mit fehlenden Werten: Complete case analysis

- Andere Bezeichnungen:
  case deletion, listwise deletion
- Verwende nur die vorhandenen Daten für die Analyse

Umgang mit fehlenden Werten: Complete case analysis



Umgang mit fehlenden Werten
Complete case analysis

	r [Länge 95%-KI]	b [Länge 95%-KI]
Vollständiger Datensatz (n=604)	0,507 [0,118]	0,817 [0,223]
Complete case analysis (n=410)	0,502 [0,145]	0,812 [0,272]

Umgang mit fehlenden Werten: Complete case analysis

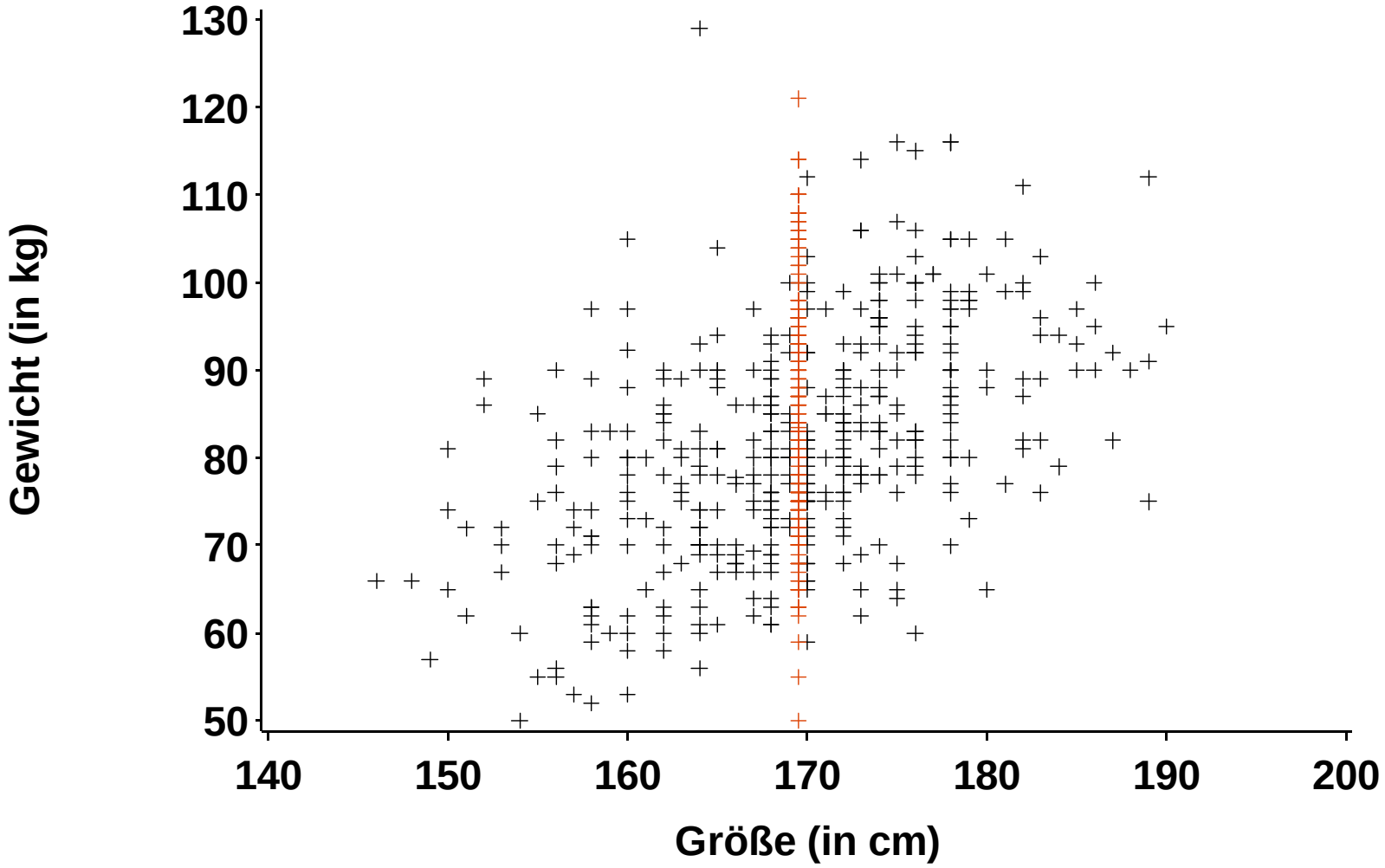
- Andere Bezeichnungen:
  case deletion, listwise deletion
- Verwende nur die vorhandenen Daten für die Analyse

- Unverzerrte Schätzung unter MCAR
- Verlust an statistischer Power, größere Konfidenzintervalle

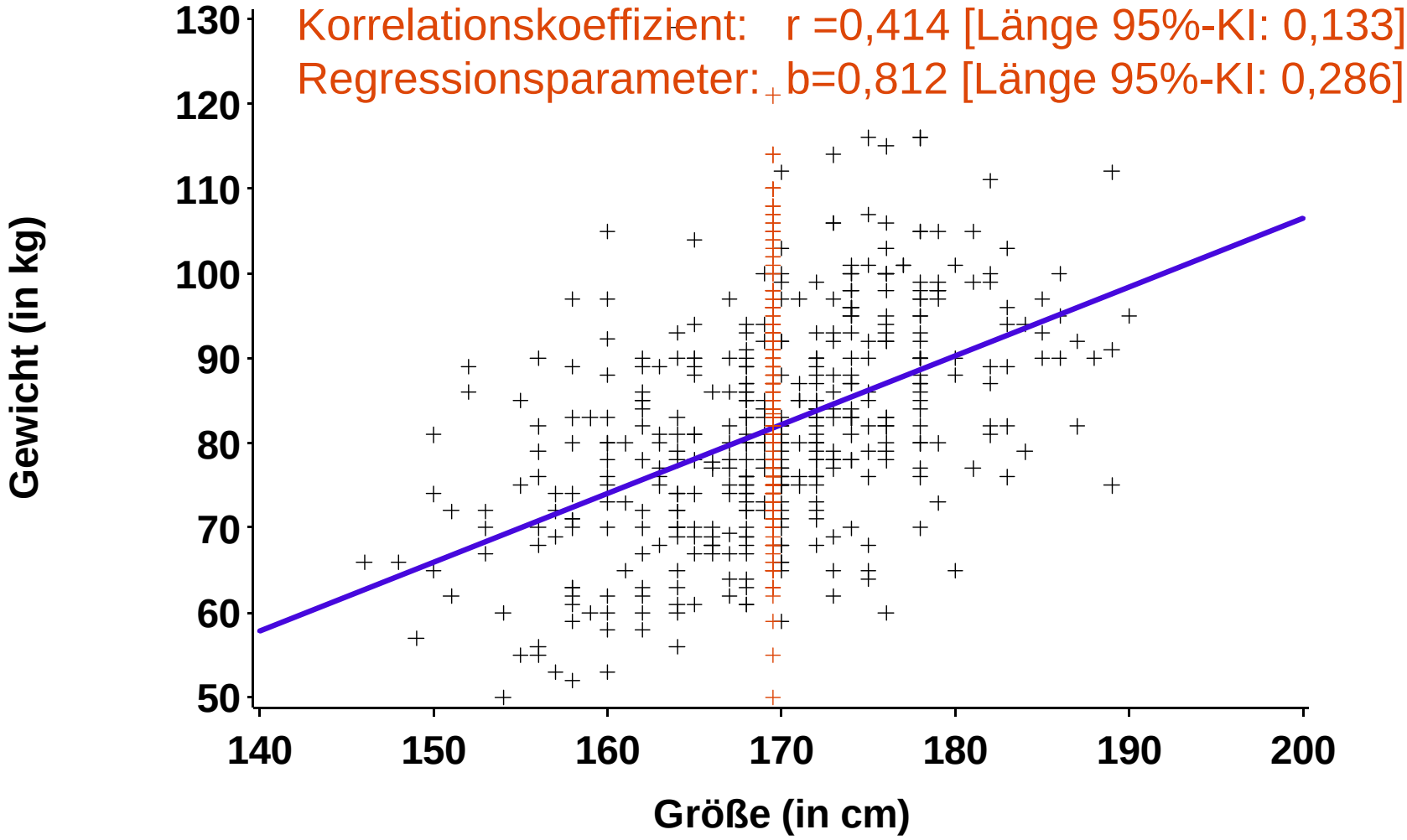
Umgang mit fehlenden Werten: Mean imputation

- Ersetze alle fehlenden Werte durch den Mittelwert der beobachteten Werte (hier: Mittlere Größe= 169,5 cm)
- Vorteil: Der Mittelwert der Daten bleibt erhalten.

Umgang mit fehlenden Werten: Mean imputation



Umgang mit fehlenden Werten: Mean imputation



Umgang mit fehlenden Werten
Mean imputation

	r [Länge 95%-KI]	b [Länge 95%-KI]
Vollständiger Datensatz (n=604)	0,507 [0,118]	0,817 [0,223]
Complete case analysis (n=410)	0,502 [0,145]	0,812 [0,272]
Mean imputation	0,414 [0,133]	0,812 [0,286]

Umgang mit fehlenden Werten: Mean imputation

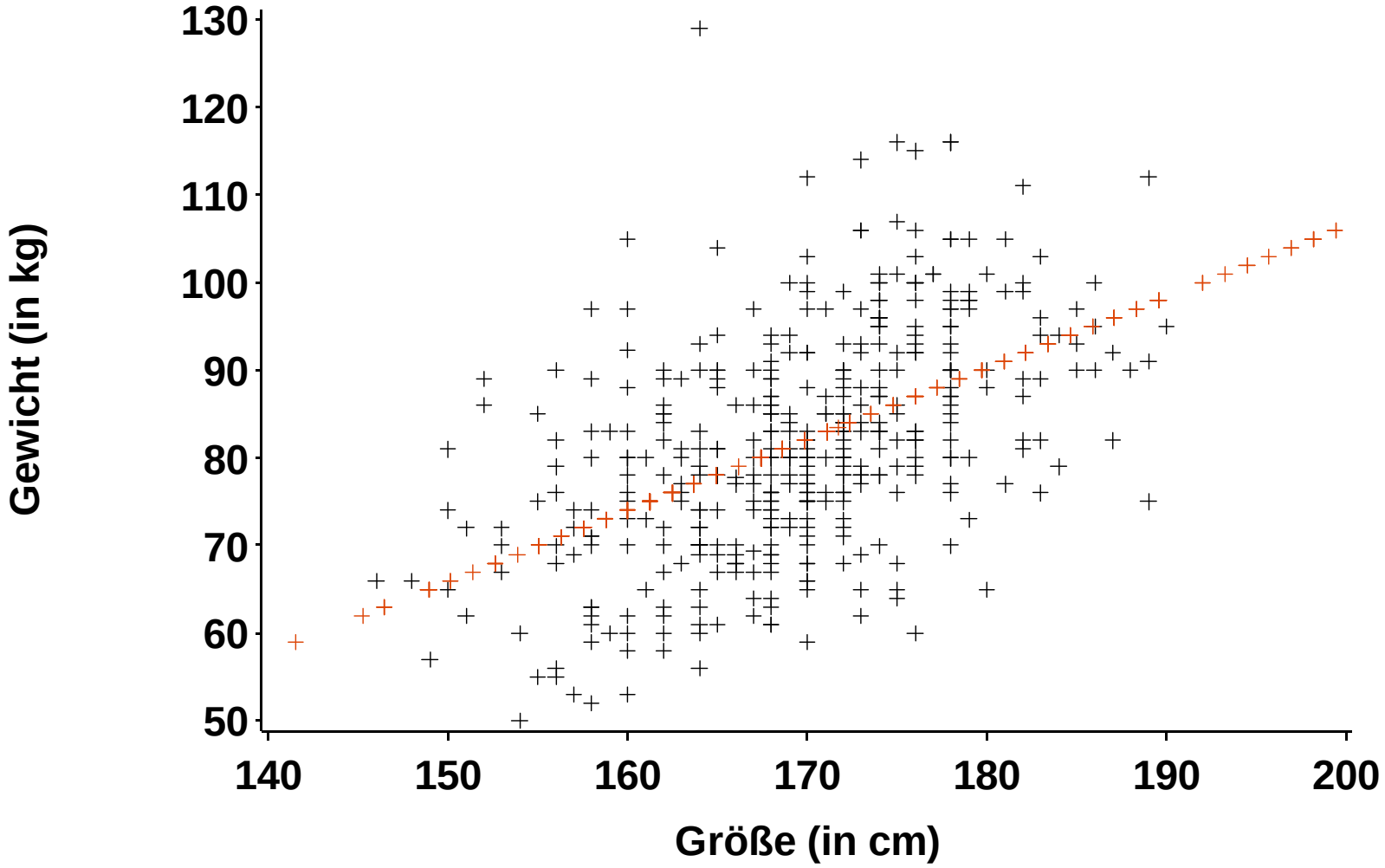
- Ersetze alle fehlenden Werte durch den Mittelwert der beobachteten Werte (hier: Mittlere Größe= 169,5 cm)
- Vorteil: Der Mittelwert der Daten bleibt erhalten.

- Nachteil für KIs dagegen gleich doppelt: Variabilität der Daten wird niedriger, Fallzahl wird größer.

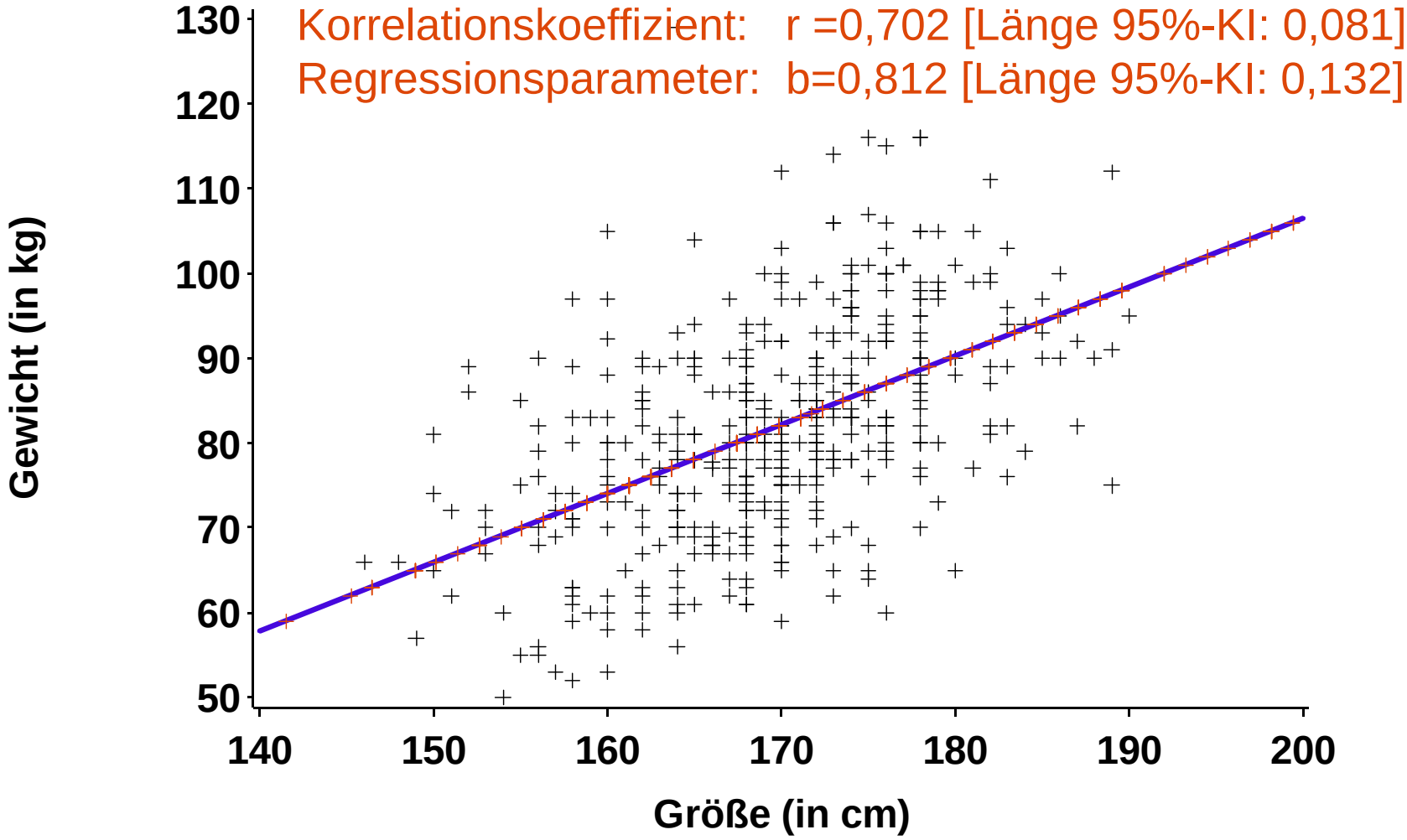
Umgang mit fehlenden Werten: Regression imputation

- Ersetze alle fehlenden Werte durch den prognostizierten Wert aus einer Regressionsanalyse mit den beobachteten Werten

Umgang mit fehlenden Werten: Regression imputation



Umgang mit fehlenden Werten: Regression imputation



Umgang mit fehlenden Werten: Regression imputation

	r [Länge 95%-KI]	b [Länge 95%-KI]
Vollständiger Datensatz (n=604)	0,507 [0,118]	0,817 [0,223]
Complete case analysis (n=410)	0,502 [0,145]	0,812 [0,272]
Mean imputation	0,414 [0,133]	0,812 [0,286]
Regression imputation	0,702 [0,081]	0,812 [0,132]

Umgang mit fehlenden Werten: Zwischenfazit

- Alle Verfahren, die einen festen Wert ersetzen, führen zu verzerrten Ergebnissen!
- **Lösung:** Multiple Imputation

Umgang mit fehlenden Werten: Multiple Imputation

3 Schritte:

- Generiere mehrere (m) imputierte Datensätze mit „plausiblen“ Ersetzungen (konkret: ziehe Werte aus der „posterior predictive distribution“)
- Analysiere alle m imputierten Datensätze
- Kombiniere m Parameterschätzer nach „Rubin‘s Regeln“

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
        NIMPUTE=10 SEED=6900123;
    MCMC;
    VAR groesse gew;
RUN;
```

Aufruf der Prozedur

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
        NIMPUTE=10 SEED=6900123;
    MCMC;
    VAR groesse gew;
RUN;
```

Datensatz (mit fehlenden Werten)

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
      NIMPUTE=10 SEED=6900123;
   MCMC;
   VAR groesse gew;
RUN;
```

Datensatz (mit imputierten Werten)

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
        NIMPUTE=10 SEED=6900123;
    MCMC;
    VAR groesse gew;
RUN;
```

Anzahl der Imputationen (m)

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
      NIMPUTE=10 SEED=6900123;
   MCMC;
   VAR groesse gew;
RUN;
```

Startwert des Zufallszahlengenerators

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
        NIMPUTE=10 SEED=6900123;
    MCMC;
    VAR groesse gew;
RUN;
```

Imputationstechnik

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 1: Generierung der imputierten Datensätze

```
PROC MI DATA=einsmiss OUT=einsmissMI
       NIMPUTE=10 SEED=6900123;
    MCMC;
    VAR groesse gew;
RUN;
```

Angabe der Variablen, die für die Imputation beutzt werden

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 2: Analyse

```
PROC MIXED DATA=einsmissMI;
     MODEL gew=groesse / s covb;
     BY _Imputation_;
     ODS OUTPUT SolutionF=regparms;
RUN;
```

Für jeden Wert der Variable _Imputation_ wird eine neue Analyse gemacht

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 2: Analyse

```
PROC MIXED DATA=einsmissMI;
     MODEL gew=groesse / s  covb;
     BY _Imputation_;
     ODS OUTPUT SolutionF=regparms;
RUN;
```

Herausschreiben von Schätzern und Varianzen in die Datei regparms

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 3: Zusammenfassen der Schätzer

```
PROC MIANALYZE PARMS=regparms;
    MODELEFFECTS Intercept groesse;
RUN;
```

Aufruf der Prozedur

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 3: Zusammenfassen der Schätzer

```
PROC MIANALYZE PARMS=regparms;
    MODELEFFECTS Intercept groesse;
RUN;
```

Datensatz der Schätzer

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Schritt 3: Zusammenfassen der Schätzer

```
PROC MIANALYZE PARMS=regparms;
    MODELEFFECTS Intercept groesse;
RUN;
```

Angabe der zu zusammenfassenden Parameter

Umgang mit fehlenden Werten: Multiple Imputation in SAS

	r [Länge 95%-KI]	b [Länge 95%-KI]
Vollständiger Datensatz (n=604)	0,507 [0,118]	0,817 [0,223]
Complete case analysis (n=410)	0,502 [0,145]	0,812 [0,272]
Mean imputation	0,414 [0,133]	0,812 [0,286]
Regression imputation	0,702 [0,081]	0,812 [0,132]
Multiple imputation (m=10)	0,501 [0,149]	0,810 [0,239]
Multiple imputation (m=1000)	0,503 [0,141]	0,812 [0,253]

Umgang mit fehlenden Werten: Beispiel mit MAR-Daten

Jetzt: „Schwierigere“ Situation

Generiere MAR-Daten:
Alle Beobachtungen mit X>175cm fehlen
(zu kurzes Maßband …)

Umgang mit fehlenden Werten: Beispiel mit MAR-Daten

	r [Länge 95%-KI]	b [Länge 95%-KI]
Vollständiger Datensatz (n=604)	0,507 [0,118]	0,817 [0,223]
Complete case analysis (n=453)	0,370 [0,160]	0,720 [0,334]
Mean imputation	0,294 [0,146]	0,720 [0,374]
Regression imputation	0,675 [0,088]	0,720 [0,126]
Multiple imputation (m=10)	0,398 [0,145]	0,831 [0,331]
Multiple imputation (m=1000)	0,398 [0,166]	0,832 [0,358]

Umgang mit fehlenden Werten: Multiple Imputation

Vorteile:

- Verlangt nur MAR, nicht MCAR.
- Man kann für die imputierten Datensätze die altbekannte Software verwenden.
- Die Berechnung der adjustierten Schätzer ist simpel.
- Man kann imputierte Datensätze für mehrere Analysen verwenden.
- Es sind gar nicht soo viele Imputationen nötig.
- Man kann für Imputation und Schätzung verschiedene Modelle benutzen. Für Imputation kann (muss!) man also alle verfügbaren Merkmale benutzen.

Umgang mit fehlenden Werten: Software

- Norm 2.03
  Public Domain, entwickelt von Joseph Schafer, nur Export nach Imputation möglich
  http://www.stat.psu.edu/~jls/misoftwa.html
- SPSS (PASW)
  Modul „Missing Values“
  http://www.spss.com/media/collateral/statistics/missing-values.pdf
- R
  hat eine Reihe von Paketen zur Imputation
- STATA
  ab Version 11: **mi** command

Umgang mit fehlenden Werten: Was tun, wenn MNAR vorliegt?

- Reweighting
- Selection models
- Pattern mixture models

Hier muss der Missing Data-Mechanismus explizit mitmodelliert werden.

Graham/Schafer: Nur notwendig in klinischen Studien mit Längsschnitts-Messungen, wenn man sich relativ sicher ist dass eine Ausscheiden aus der Studie mit den fehlenden Werten zu tun hat. In anderen Bereichen reichen in der Regel die bisher besprochenen Methoden.

Fazit I

- Single Imputationsmethoden sollten nicht verwendet werden.
- Eine ausführliche Non-Responderanalyse ist Pflicht.
- Vor der Studie darauf achten, dass Prädiktoren für Fehlwerte erhoben werden.
- Bei der Multiplen Imputation so viel Merkmale wie möglich (auch die Zielgröße!) zur Imputation verwenden. Nicht-normalverteilte Merkmale notfalls transformieren.

Fazit II

Aus Sterne JA et al. BMJ 2009:

- „The cost of multiple imputation analysis is small compared with the cost of collecting the data.“
- „It is no longer excussable for missing values and the reasons they arose to be swept under the carpet, nor for potentially misleading and inefficient analyses of complete cases to be considered adequate.“

Guideline zum Umgang mit fehlenden Werten

Box 2 | Guidelines for reporting any analysis potentially affected by missing data

- Report the number of missing values for each variable of interest, or the number of cases with complete data for each important component of the analysis. Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study. If possible, describe reasons for missing data in terms of other variables (rather than just reporting a universal reason such as treatment failure)
- Clarify whether there are important differences between individuals with complete and incomplete data—for example, by providing a table comparing the distributions of key exposure and outcome variables in these different groups
- Describe the type of analysis used to account for missing data (eg, multiple imputation), and the assumptions that were made (eg, missing at random)

For analyses based on multiple imputation

- Provide details of the imputation modelling:
  - Report details of the software used and of key settings for the imputation modelling
  - Report the number of imputed datasets that were created (Although five imputed datasets have been suggested to be sufficient on theoretical grounds,[10 11] a larger number (at least 20) may be preferable to reduce sampling variability from the imputation process[29])
  - What variables were included in the imputation procedure?
  - How were non-normally distributed and binary/categorical variables dealt with?
  - If statistical interactions were included in the final analyses, were they also included in imputation models?
- If a large fraction of the data is imputed, compare observed and imputed values
- Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations, bearing in mind that analyses of complete cases may suffer more chance variation, and that under the missing at random assumption multiple imputation should correct biases that may arise in complete cases analyses
- Discuss whether the variables included in the imputation model make the missing at random assumption plausible
- It is also desirable to investigate the robustness of key inferences to possible departures from the missing at random assumption, by assuming a range of missing not at random mechanisms in sensitivity analyses. This is an area of ongoing research[30 31]

Sterne JA et al. BMJ 2009.

Literatur

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002 Jun;7(2):147-77.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393.

Wirtz M. Über das Problem fehlender Werte: Wie der Einfluss fehlender Information auf Analyseergebnisse entdeckt und reduziert werden kann. Rehabilitation 2004 Apr;43(2):109-15.

Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. Obes Rev. 2003 Aug;4(3):175-84.

Donders AR, van der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006 Oct;59(10):1087-91.

Haukoos JS, Newgard CD. Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework. Acad Emerg Med. 2007 Jul;14(7):662-8.

Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research--part 2: multiple imputation. Acad Emerg Med. 2007 Jul;14(7):669-78.

Pigott TD. A Review of Methods for Missing Data. Educational Research and Evaluation 2001, Vol. 7, No. 4, pp. 353-383