Fehlende Werte und Multiple Imputation

Oliver Kuß
Deutsches Diabetes-Zentrum (DDZ),
Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf,
Institut für Biometrie und Epidemiologie
oliver.kuss@ddz.uni-duesseldorf.de

Inhalt

Deutsches Diabetes-Zentrum

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

Was ist ein fehlender Wert?

DDZ Deutsches Diabetes-Zentrum

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)
- … Wert, bei dem auch ein Nicht-Vorliegen eine Information enthält ("Weiß ich nicht", "Trifft für mich nicht zu" usw.)

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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 Beobachtungen mit fehlenden Werten sind eine zufällige Stichprobe aus allen Beobachtungen.

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 andere 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.

Mechanismen für fehlende Werte: MNAR

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

MCAR	Ignorable	Non-informative
MAR		
MNAR	Non-ignorable	Informative

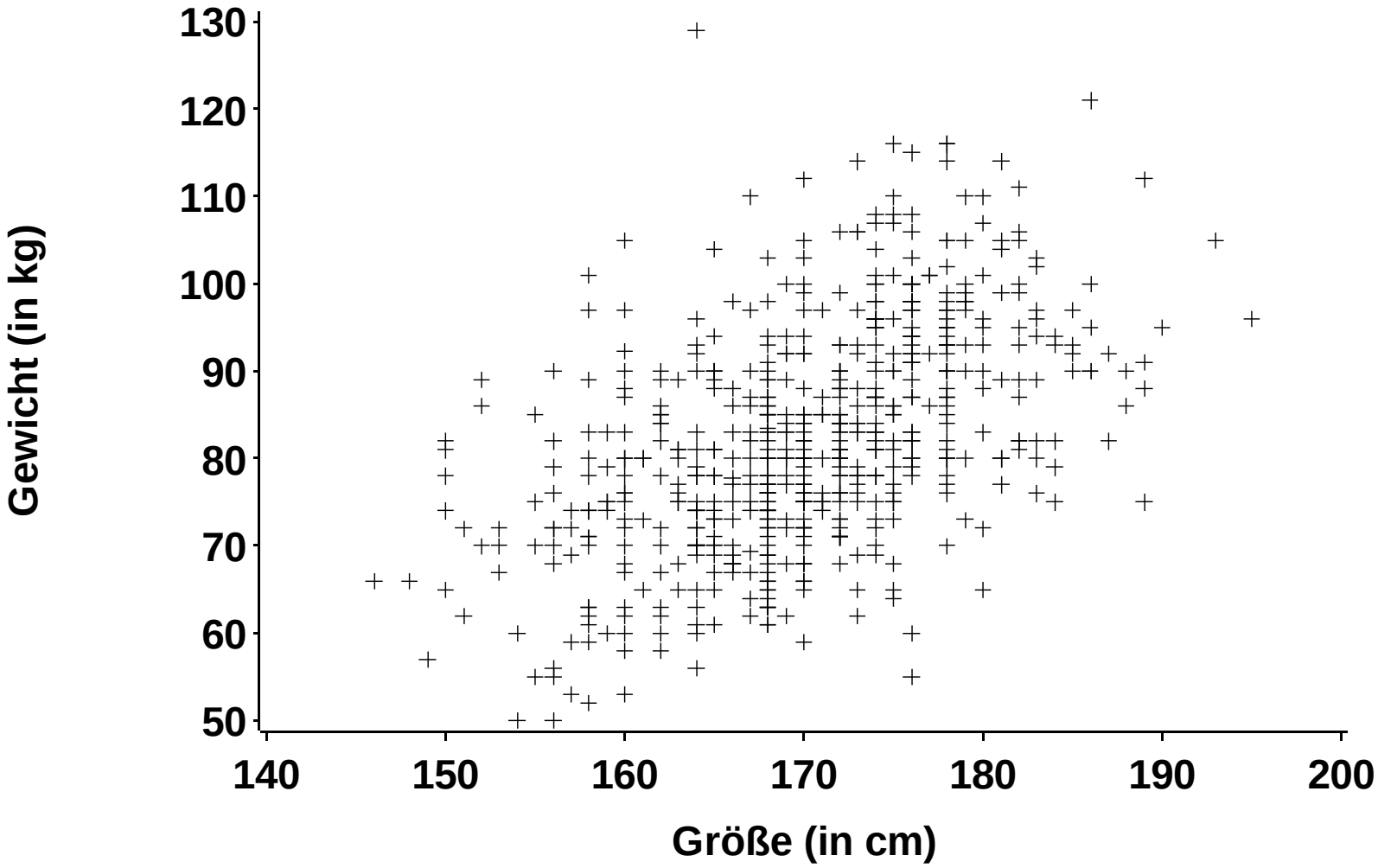
Umgang mit fehlenden Werten

- 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.

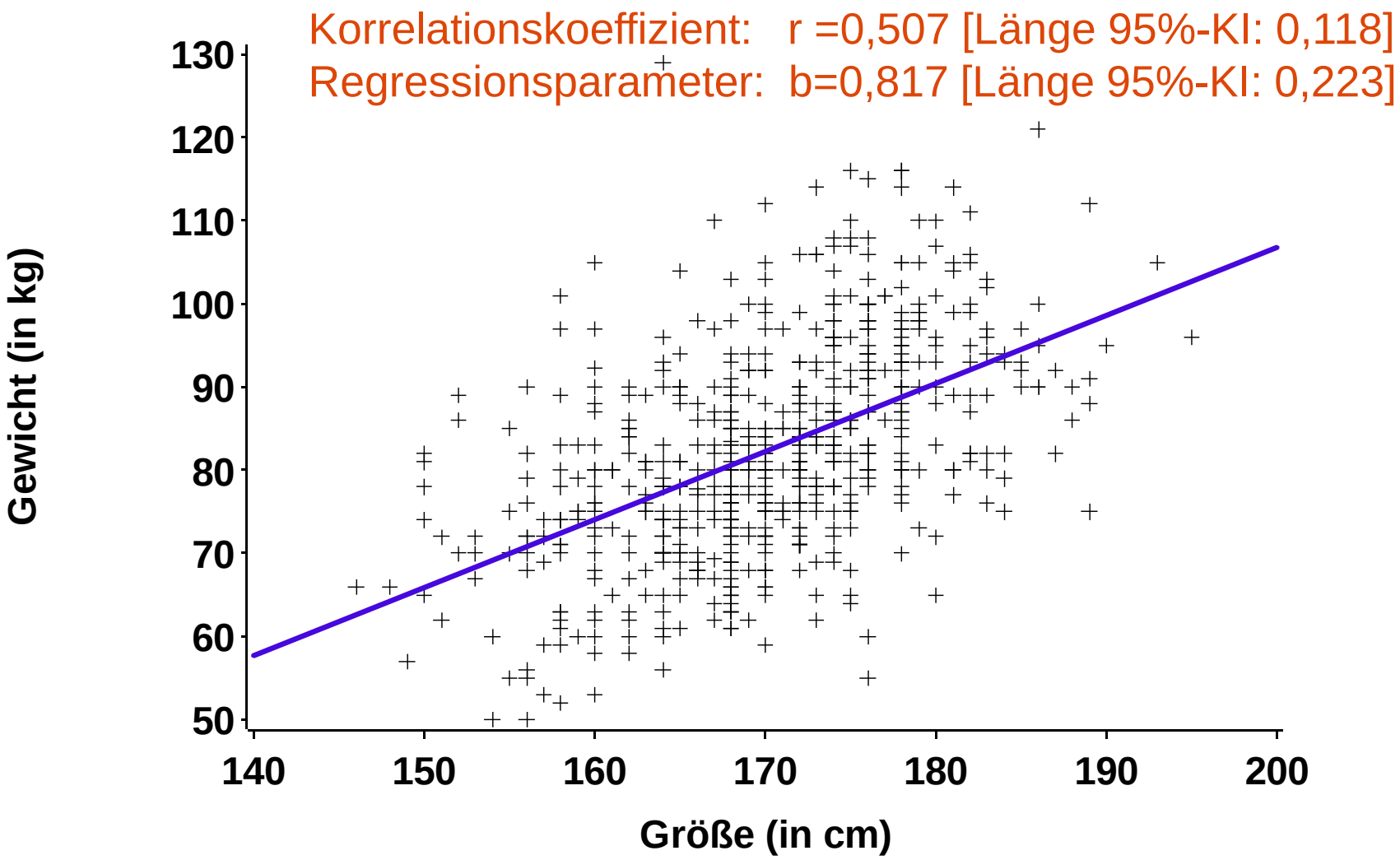
Umgang mit fehlenden Werten: Beispiel

DDZ
Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Beispiel

DDZ Deutsches Diabetes-Zentrum

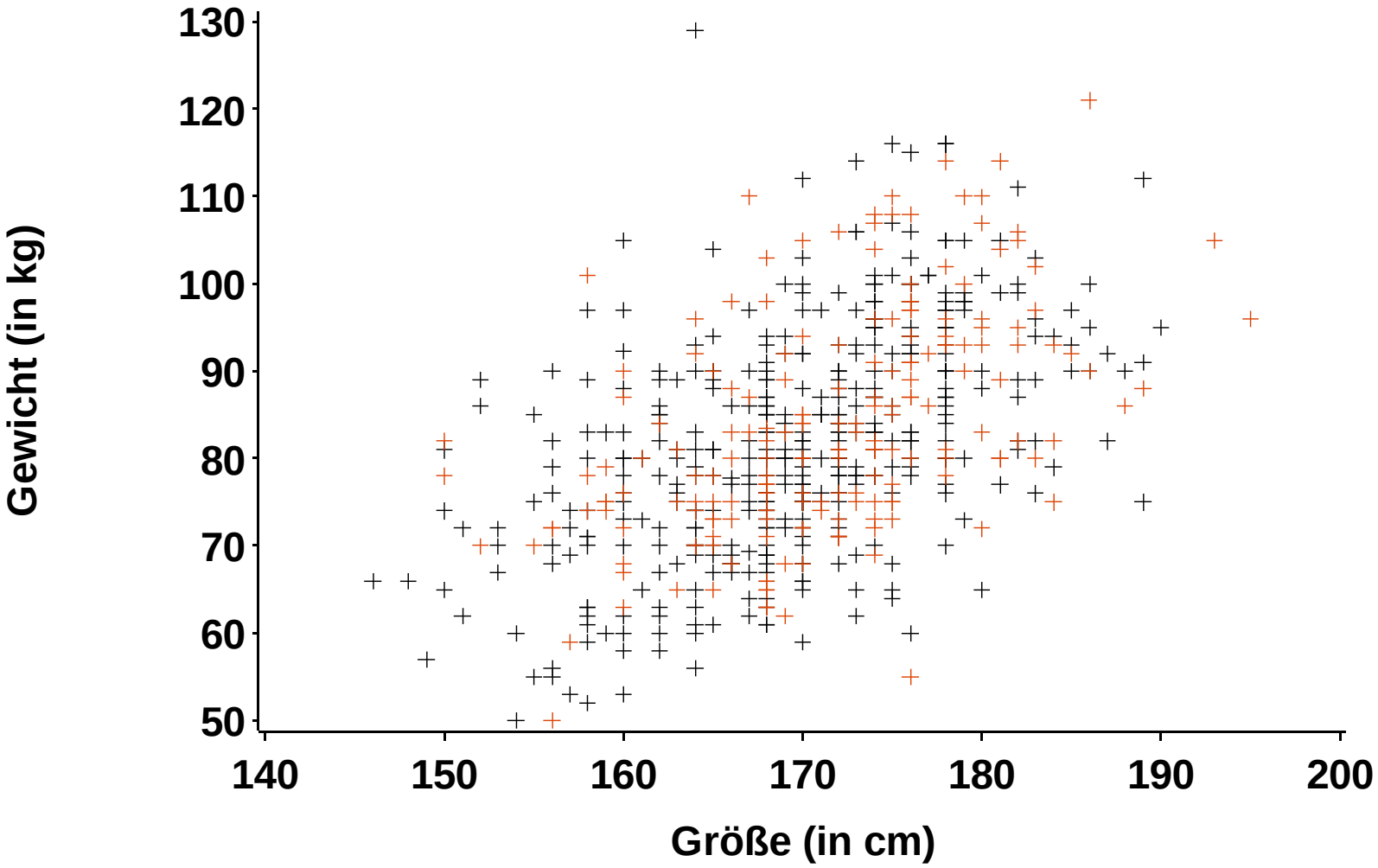


Umgang mit fehlenden Werten: Beispiel

- 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

DDZ
Deutsches Diabetes-Zentrum



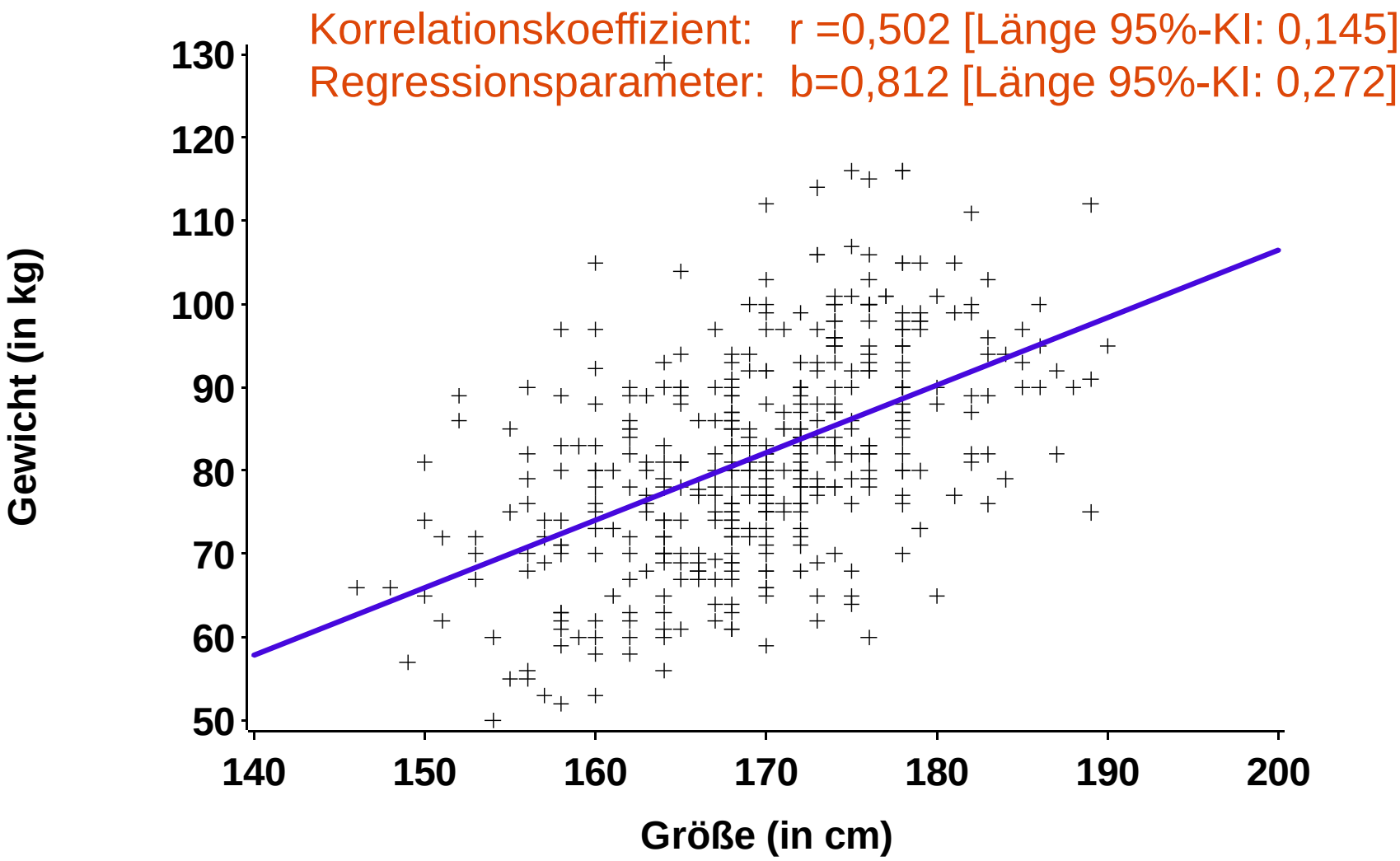
Umgang mit fehlenden Werten: Complete case analysis

DDZ
Deutsches Diabetes-Zentrum

- Verwende nur die vorhandenen Daten für die Analyse

Umgang mit fehlenden Werten: Complete case analysis

DDZ Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Complete case analysis

DDZ Deutsches Diabetes-Zentrum

	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

DDZ Deutsches Diabetes-Zentrum

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

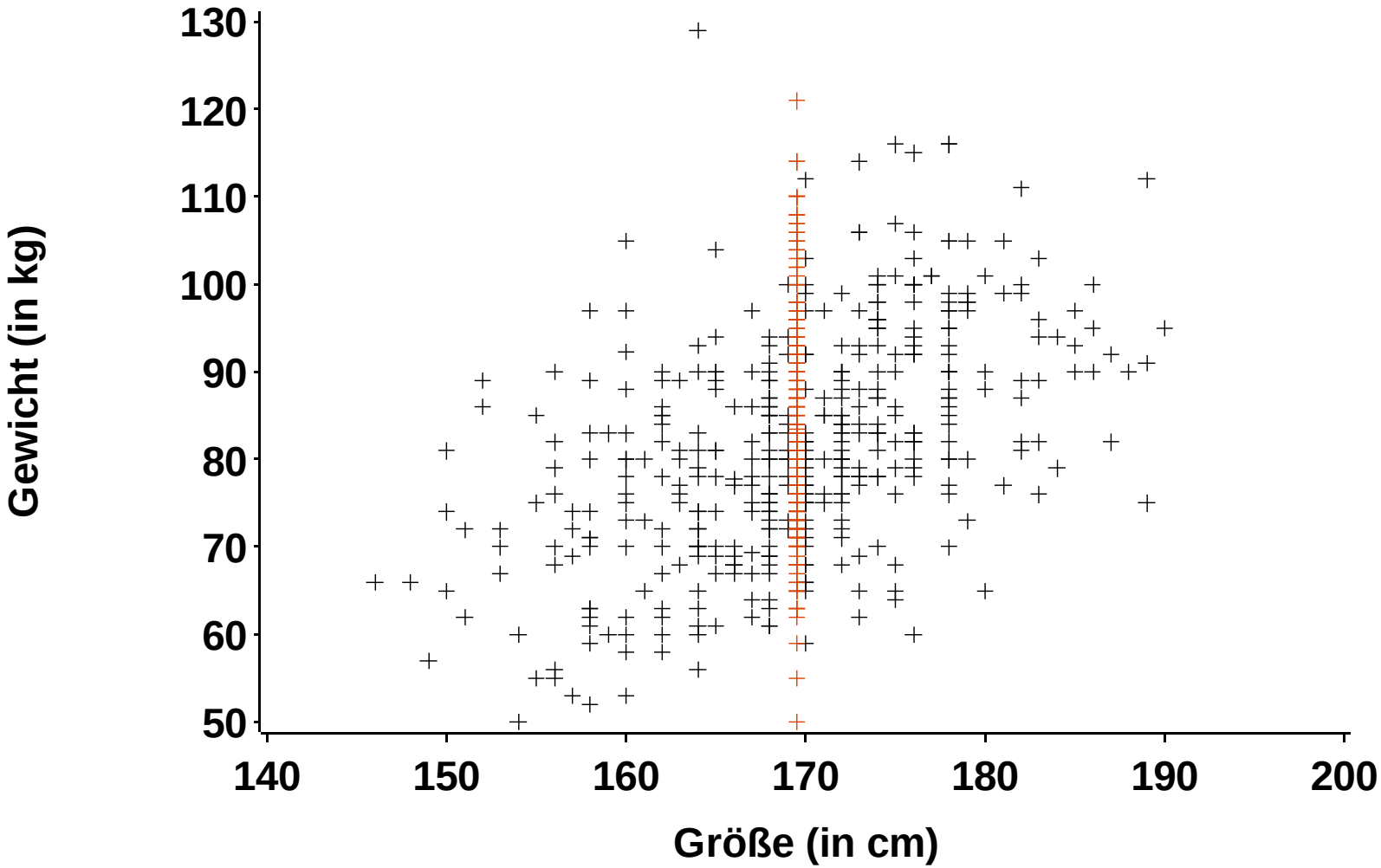
Umgang mit fehlenden Werten: Mean imputation

DDZ Deutsches Diabetes-Zentrum

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

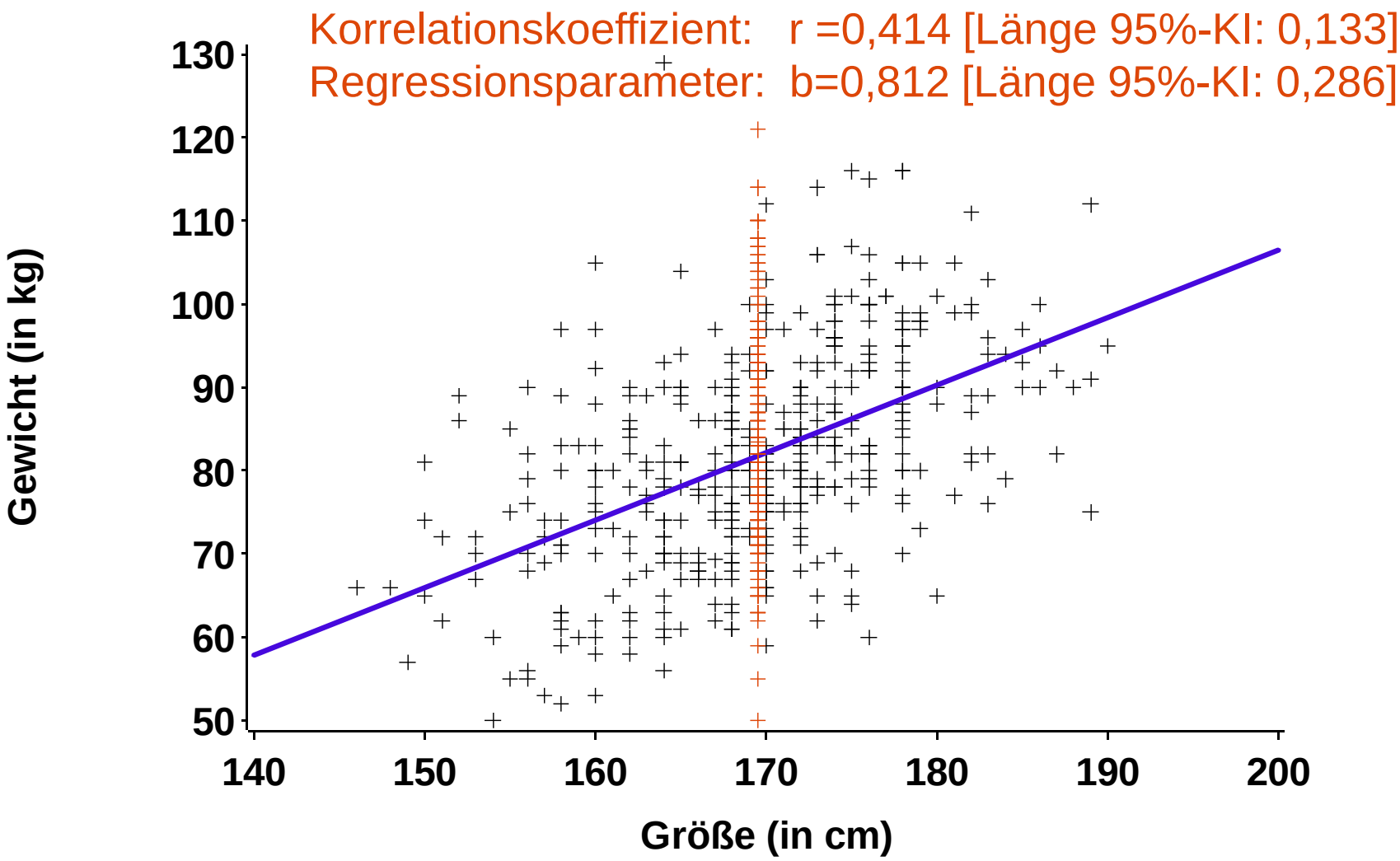
Umgang mit fehlenden Werten: Mean imputation

DDZ Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Mean imputation

DDZ Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Mean imputation

DDZ Deutsches Diabetes-Zentrum

	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]

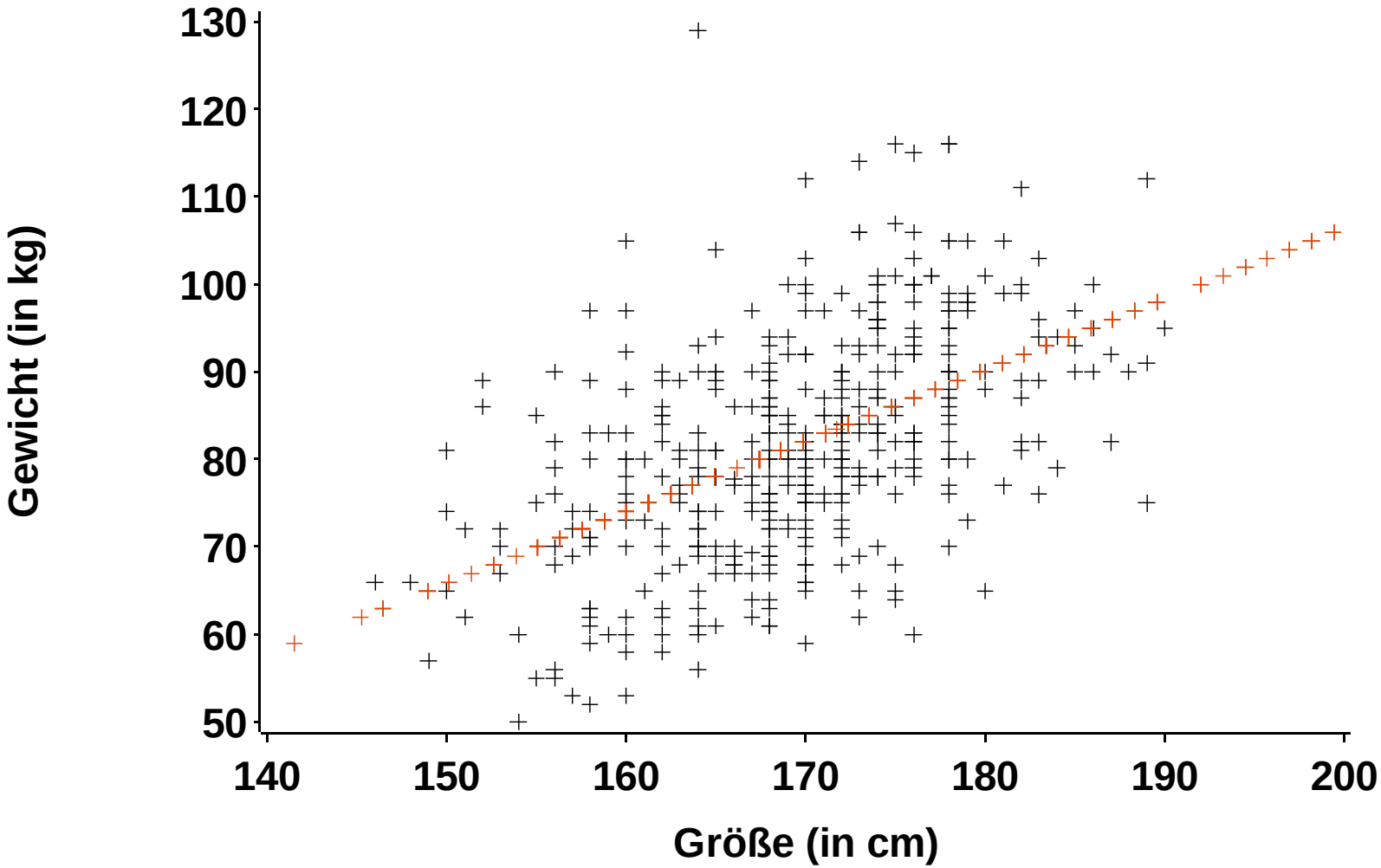
Umgang mit fehlenden Werten: Regression imputation

DDZ Deutsches Diabetes-Zentrum

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

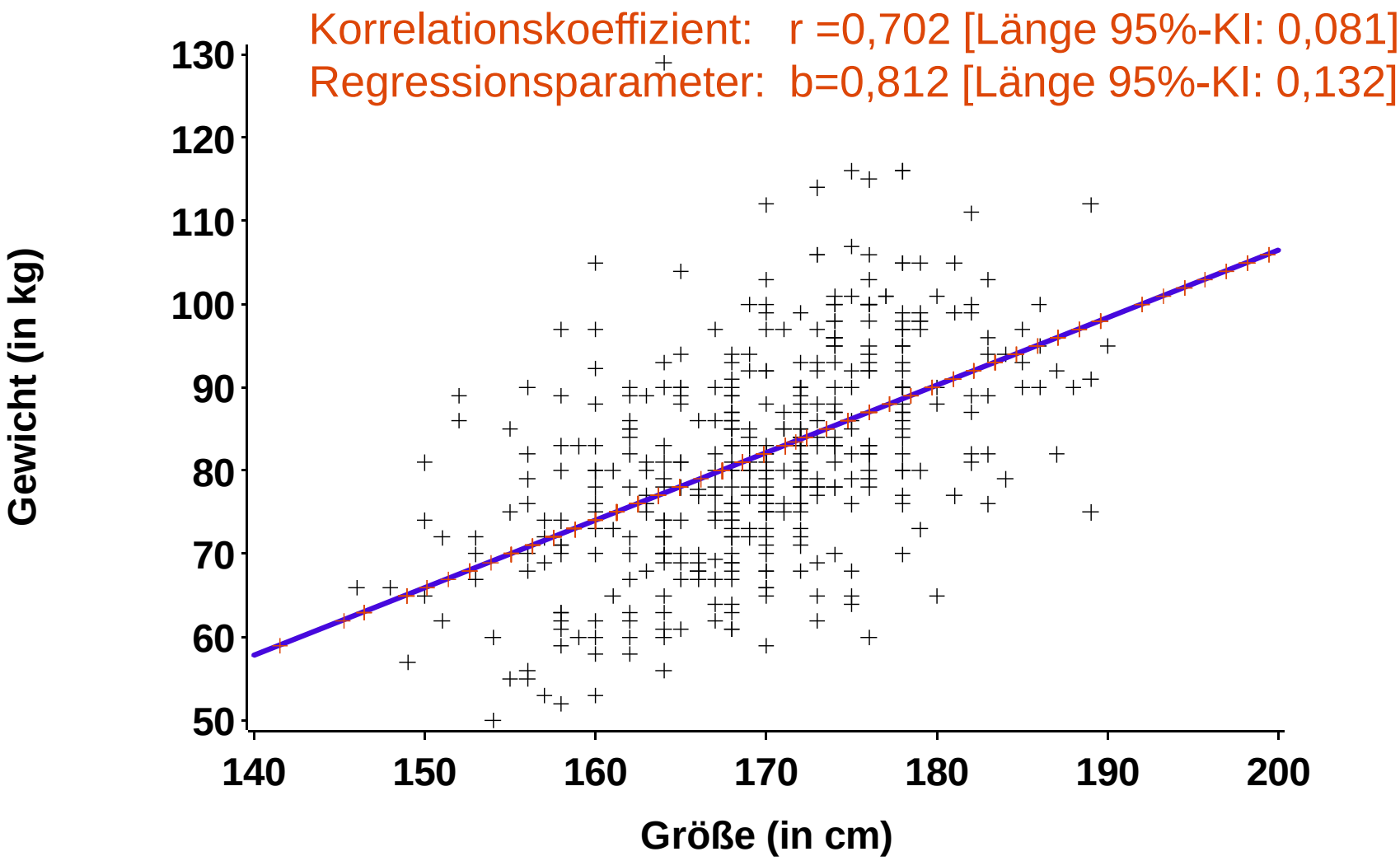
Umgang mit fehlenden Werten: Regression imputation

DDZ Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Regression imputation

DDZ Deutsches Diabetes-Zentrum



Umgang mit fehlenden Werten: Regression imputation

DDZ Deutsches Diabetes-Zentrum

	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

DDZ Deutsches Diabetes-Zentrum

- Alle Verfahren, die einen festen Wert ersetzen, führen zu verzerrten Ergebnissen oder zu „zu guten“ Ergebnissen!

- **Lösung:** Multiple Imputation

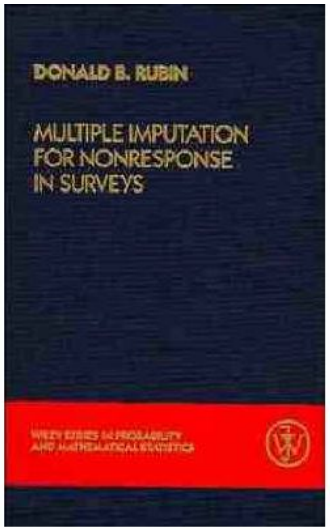
  Verlangt nur MAR (und nicht MCAR wie die Complete Case Analysis), um zu unverzerrten Schätzern zu kommen und hat dann auch korrekte Schätzvarianzen!

Umgang mit fehlenden Werten: Multiple Imputation

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

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 benutzt werden (Beachte: Auch die Zielgröße wird benutzt!)

Umgang mit fehlenden Werten: Multiple Imputation in SAS

DDZ Deutsches Diabetes-Zentrum

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 durchgeführt (insgesamt m)

Umgang mit fehlenden Werten: Multiple Imputation in SAS

Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

Schritt 3: Zusammenfassen der Schätzer

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

Aufruf der Prozedur

Umgang mit fehlenden Werten: Multiple Imputation in SAS

DDZ Deutsches Diabetes-Zentrum

Schritt 3: Zusammenfassen der Schätzer

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

Datensatz der Schätzer

Umgang mit fehlenden Werten: Multiple Imputation in SAS

DDZ Deutsches Diabetes-Zentrum

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

DDZ Deutsches Diabetes-Zentrum

	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

DDZ Deutsches Diabetes-Zentrum

Jetzt: „Schwierigere“ Situation

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

Umgang mit fehlenden Werten: Beispiel mit MAR-Daten

DDZ Deutsches Diabetes-Zentrum

	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

DDZ Deutsches Diabetes-Zentrum

Vorteile:

- Verlangt nur MAR, nicht MCAR.
- Man kann für die imputierten Datensätze die altbekannte Software verwenden.
- Die Berechnung der korrigierten 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

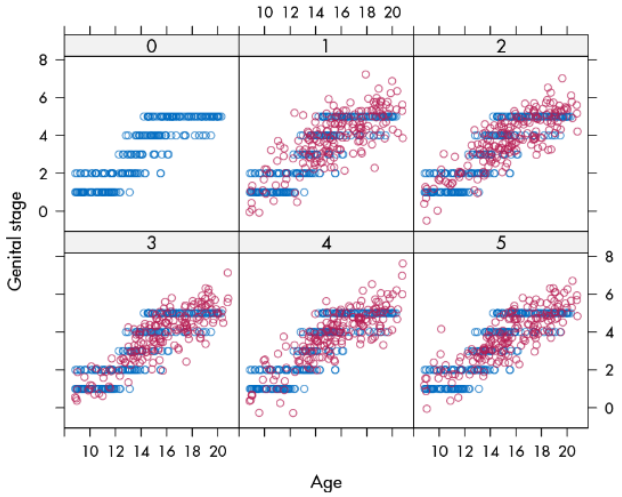
DDZ Deutsches Diabetes-Zentrum

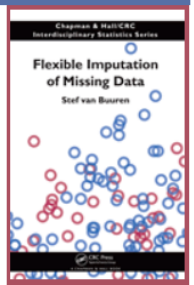
- Schöne Übersicht unter http://www.stefvanbuuren.nl/mi/Software.html

www.multiple-imputation.com

Home
MI
MICE
FIMD
Software
Course
Contact

Software





Software for multiple imputation

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

DDZ Deutsches Diabetes-Zentrum

- 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

DDZ Deutsches Diabetes-Zentrum

- Single Imputationsmethoden sollten nicht verwendet werden.
- Eine ausführliche Non-Responderanalyse sollte gemacht werden.
- Vor der Studie darauf achten, dass Prädiktoren für Fehlwerte erhoben werden.
- Bei der Multiplen Imputation so viele Variablen wie möglich (auch die Zielgröße!) zur Imputation verwenden.

Fazit II

DDZ Deutsches Diabetes-Zentrum

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.“

Auch andere Probleme in der Epidemiologie könnten mit Methoden für fehldende Werte gelöst werden, z.B. alle mit Potential outcomes (s. Westreich et al. Int J Epidemiol 2015)

Guidelines zum Umgang mit fehlenden Werten

DDZ Deutsches Diabetes-Zentrum

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.

Guidelines zum Umgang mit fehlenden Werten: STROBE statement

DDZ Deutsches Diabetes-Zentrum

Aus der „Checklist“:

Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed
Descriptive data	14*	(a) Give characteristics of study participants (eg, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest

Aus der „Explanation and Elaboration“:

Box 6 - Missing data: problems and possible solutions

A common approach to dealing with missing data is to restrict analyses to individuals with complete data on all variables required for a particular analysis. Although such 'complete-case' analyses are unbiased in many circumstances, they can be biased and are always inefficient.[108] Bias arises if individuals with missing data are not typical of the whole sample. Inefficiency arises because of the reduced sample size for analysis.

Using the last observation carried forward for repeated measures can distort trends over time if persons who experience a foreshadowing of the outcome selectively drop out.[109] Inserting a missing category indicator for a confounder may increase residual confounding.[107] Imputation, in which each missing value is replaced with an assumed or estimated value, may lead to attenuation or exaggeration of the association of interest, and without the use of sophisticated methods described below may produce standard errors that are too small.

Rubin developed a typology of missing data problems, based on a model for the probability of an observation being missing.[109, 110] Data are described as missing completely at random (MCAR) if the probability that a particular observation is missing does not depend on the value of any observable variable(s). Data are missing at random (MAR) if, given the observed data, the probability that observations are missing is independent of the actual values of the missing data. For example, suppose younger children are more prone to missing spirometry measurements, but that the probability of missing is unrelated to the true unobserved lung function, after accounting for age. Then the missing lung function measurement would be MAR in models including age. Data are missing not at random (MNAR) if the probability of missing still depends on the missing value even after taking the available data into account. When data are MNAR valid inferences require explicit assumptions about the mechanisms that led to missing data.

Methods to deal with data missing at random (MAR) fall into three broad classes:[108, 111] likelihood-based approaches,[112] weighted estimation[113] and multiple imputation.[111, 114] Of these three approaches, multiple imputation is the most commonly used and flexible, particularly when multiple variables have missing values.[115] Results using any of these approaches should be compared with those from complete case analyses, and important differences discussed. The plausibility of assumptions made in missing data analyses is generally unverifiable. In particular it is impossible to prove that data are MAR, rather than MNAR. Such analyses are therefore best viewed in the spirit of sensitivity analysis (see items 12e and 17).

Vandenbroucke et al. Epidemiology. 2007.

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.

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007 Nov;18(6):805-35.

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.

Yang Y. Multiple Imputation Using SAS Software. Journal of Statistical Software. 2011 Dec;45(6). http://www.jstatsoft.org/

Berglund P, Heeringa S. *Multiple Imputation of Missing Data Using SAS. 2014.*Cary, NC: SAS® Institute Inc.

Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, Schisterman EF. Imputation approaches for potential outcomes in causal inference. Int J Epidemiol. 2015 Oct;44(5):1731-7.