Übersicht zu fehlenden Daten

Seminar: Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten

Alexander Pokatilo

Betreuerin: Eva Endres

Seminarleiter: Prof. Dr. Thomas Augustin

Institut für Statistik, LMU

21. November 2014

Übersicht

Definition

Als existierend angenommene und als Reaktion auf einen Reiz hervorrufbare Werte, deren Beobachtung intendiert ist, die nicht beobachtet werden und nicht ohne Unsicherheit aus anderen Informationen ableitbar sind. (M.Spieß, 2008)

Mögliche Ursachen

- fehlerhaftes Untersuchungsdesign (missverständliche Fragen)
- interviewerbedingte (Skip-Fehler)
- befragtenbedingte (Antwortverweigerung)
- Unvolständigkeit von Sekundärdaten
- Datenaufbereitungsfehler (Codierung-, Übertragungsfehler)

- fehlerhaftes Untersuchungsdesign (missverständliche Fragen)
- interviewerbedingte (Skip-Fehler)
- befragtenbedingte (Antwortverweigerung)
- Unvolständigkeit von Sekundärdaten
- Datenaufbereitungsfehler (Codierung-, Übertragungsfehler)

Fehlende Werte können auf allen Stadien der Datenerhebung und Datenaufbereitung entstehen. Wichtig ist die Frage ob dieser Ausfall zufälliger Natur oder systematisch ist. Wenn der Ausfallmechanismus bekannt ist, dann ist auch die angemessnen Bechandlung von fehlenden Werten möglich.

1. Bsp. für fehlerhaftes Untersuchungsdesign - nicht Ausschöpfende Antwortmöglichkeiten
2. Bei Skip-Fehler vergisst der Interviewer eine Frage zu stellen
3. Besonders bei sensiblen Fragen oder auch wegen Konzentrationverlust, mangelnder Motivation
4. z.B. Verlorene Patientenakte

Konsequenzen

- technische Probleme bei der Auswertung
- verzerrte Schätzer
- reduzierte Aussagekraft der Analyse

- technische Probleme bei der Auswertung
- verzerrte Schätzer
- reduzierte Aussagekraft der Analyse

1. Statistische Methoden verlangen vollständige Datenmatrizen, sonst Fehler
2. Korrekte empirische Schätzer lassen sich nur auf Basis vollständiger Daten berechnen
3. Fehlende Werte bringen zusätzliche Unsicherheit in die Ergebnisse und schränken die Aussagekraft der Studie ein

Übersicht

Missing Data Pattern

Missing Data Pattern beschreiben die Struktur der fehlenden Werten im Datensatz.
Man kann drei Grundmuster der fehlenden Daten definieren:

- **Univariater Ausfall** (Item-Nonresponse)
- **Multivariater Ausfall** (Unit-Nonresponse)
- **Monoton fehlende Werte** (Panelattrition)
- **Allgemeiner Ausfallmuster**

Ausfallmuster (Uni- und Multivariater Ausfall)

Item-Nonresponse

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
x_{31}	x_{32}	x_{33}	x_{34}	.
x_{41}	x_{42}	x_{43}	x_{44}	.
x_{51}	x_{52}	x_{53}	x_{54}	.

- Einzelne Werte bei einer Variable fehlen
- z.B. Einkommen

Unit-Nonresponse

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
x_{31}	x_{32}	.	.	.
x_{41}	x_{42}	.	.	.
x_{51}	x_{52}	.	.	.

- Extremfall von Item-Nonresponse
- Ganze Einheiten fehlen
- z.B. jede Auskunft verweigert

Ausfallmuster (Monotoner und Allgemeiner Muster)

Monotoner Ausfall

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_{21}	x_{22}	x_{23}	x_{24}	.
x_{31}	x_{32}	x_{33}	.	.
x_{41}	x_{42}	.	.	.
x_{51}	.	.	.	.

Allgemeiner Muster

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	.	x_{15}
.	x_{22}	x_{23}	.	x_{25}
x_{31}	x_{32}	.	x_{34}	.
x_{41}	.	x_{43}	x_{44}	.
x_{51}	.	x_{53}	x_{54}	x_{55}

- Panelattrition
- Einheiten scheiden im Verlauf der Studie aus und kehren nicht zurück (Umzug, Tot)

Übersicht zu fehlenden Daten

Ausfallmuster (Monotoner und Allgemeiner Muster)

Monotoner Ausfall

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_{21}	x_{22}	x_{23}	x_{24}	.
x_{31}	x_{32}	x_{33}	.	.
x_{41}	x_{42}	.	.	.
x_{51}	.	.	.	.

Allgemeiner Muster

Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	x_{12}	x_{13}	.	x_{15}
.	x_{22}	x_{23}	.	x_{25}
x_{31}	x_{32}	.	x_{34}	.
x_{41}	.	x_{43}	x_{44}	.
x_{51}	.	x_{53}	x_{54}	x_{55}

- Panelattrition
- Einheiten scheiden im Verlauf der Studie aus und kehren nicht zurück (Umzug, Tot)

Panalattrition (Panalmortalität) bezeichnet eine Abnutzung der Stichprobe über die Zeit, also die Situation, bei der Untersuchungseinheiten, die am Anfang der Studie (Längsschnittstudie oder Panel) dabei waren, im weiteren Verlauf der Studie aus verschidensten Gründen ausscheiden und nicht zurückkehren

Übersicht

Missingmechanismen

Rubin (1976) und Little und Rubin (2002) entwarfen ein Konzept zur Beschreibung von Mechanismen, die zu fehlenden Werten führen. Diese Messingmechanismen sind:

- **Missing Completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Not Missing at Random (NMAR)**

Falsche Annahmen bzgl. Mechanismus können zu verzerrten Ergebnissen führen.

Missing Completely at Random (MCAR)

Der Antwortausfall ist rein zufällig und steht weder mit der Ausprägung des Merkmals noch mit anderen Variablen in Verbindung.

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|\phi) \quad \text{für alle } Y_{obs}, Y_{mis}, \phi$$

- Idealfall
- beobachtete Werte als Zufallsstichprobe von der ursprünglich volständigen Stichprobe
- Test auf MCAR von Little für stetige Variablen

Der Antwortausfall ist rein zufällig und steht weder mit der Ausprägung des Merkmals noch mit anderen Variablen in Verbindung.

$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|\phi)$ für alle Y_{obs}, Y_{mis}, ϕ

- Idealfall
- beobachtete Werte als Zufallsstichprobe von der ursprünglich volständigen Stichprobe
- Test auf MCAR von Little für stetige Variablen

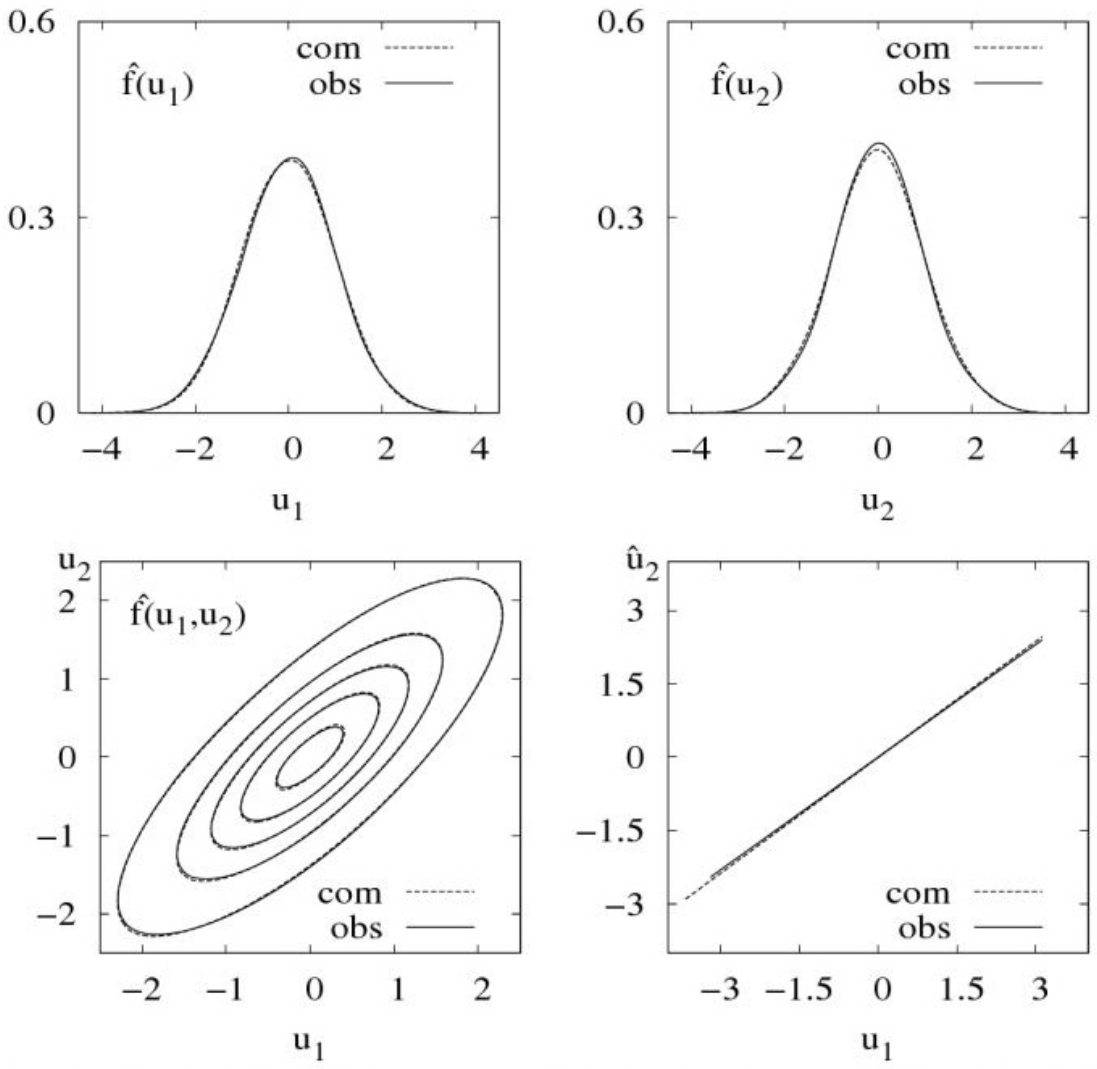
$Y = (y_{ij}) = Y_{obs} + Y_{mis}$
wobei Y alle, Y_{obs} beobachtete und Y_{mis} fehlende Werte bezeichnen
$M = (m_{ij})$ Indikatormatrix der fehlenden Werte

$$m_{ij} = \begin{cases} 1 & \text{, falls } y_{ij} \text{ fehlt} \\ 0 & \text{, sonst} \end{cases}$$

$f(M|Y, \phi)$ bedingte Verteilung von M, wobei ϕ unbekannte Parameter definiert.

Test von Little, dass in SPSS(Modul *Missing Values* nötig) vorhanden ist, erlaubt globale beurteilung von MCAR-Bedingung

Missing Completely at Random (MCAR)

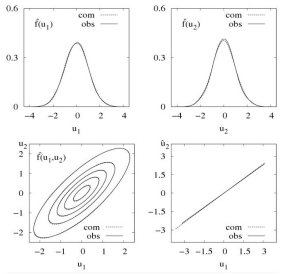


- 2000 Wertepaare (u_1, u_2)
- ca. 50% der Werte von u_1 nach MCAR gelöscht
- (fast) keine Verzerrungen sichtbar

Abbildung: aus Spieß(2008)[S.6]

Übersicht zu fehlenden Daten

Missing Completely at Random (MCAR)



- 2000 Wertepaare (u_1, u_2)
- ca. 50% der Werte von u_1 nach MCAR gelöscht
- (fast) keine Verzerrungen sichtbar

Abbildung: aus Spieß(2008)[S.6]

Oben links und rechts die geglätteten Verteilungen der Variablen u_1 und u_2 vor (mit com für complete" bezeichnet) und nach dem Löschen von Daten (mit obs für observed). Unten links die Höhenlinien der bivariaten Dichte. Unten rechts die Stichprobenregressionsgeraden für Regression von u_2 auf u_1. Dieselbe Struktur bleibt auch für folgene Abbildungen von MAR und NMAR-Mechanismen erhalten.

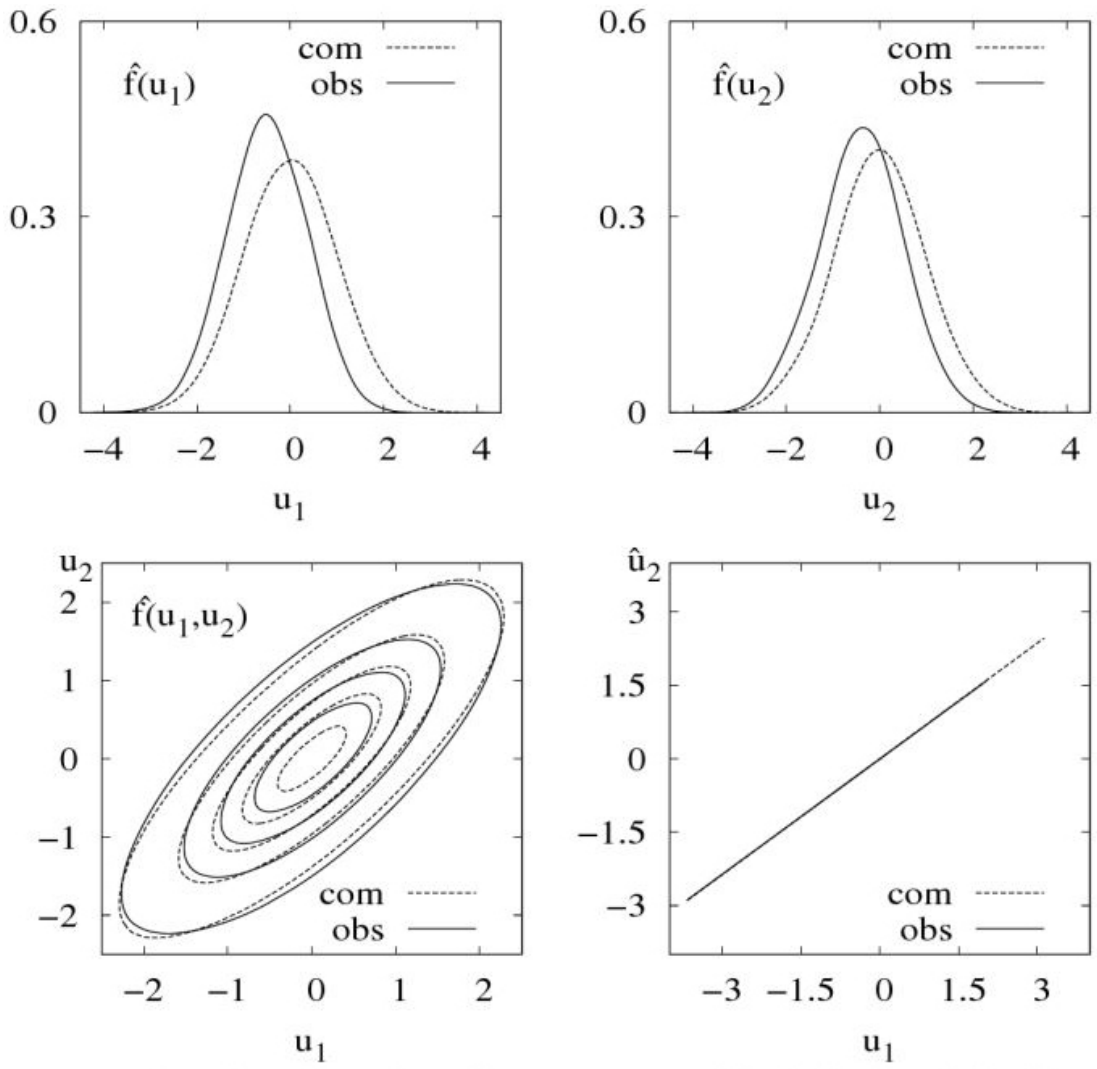
Missing at Random (MAR)

Das Auftreten von fehlenden Werten steht zumindest teilweise im Zusammenhang mit beobachteten Werten anderer Variablen.

$$f(M|Y_{obs}, Y_{mis}, \phi) = f(M|Y_{obs}, \phi) \quad \text{für alle } Y_{mis},\ \phi$$

- abhängig fehlende Werte
- nach Kontrolle der Abhängigkeiten von beobachteten Variablen hängen M_j nicht mehr von Y_j
- weniger strenge Bedingung als MCAR
- Bsp.: fehlende Einkommen-Angaben hängen vom Alter, nicht aber von der Hohe des Einkommens

Missing at Random (MAR)



- deutliche Verzerrungen uni- und bivariaten Verteilungen
- Regression unverzerrt

Abbildung: aus Spieß(2008)[S.8]

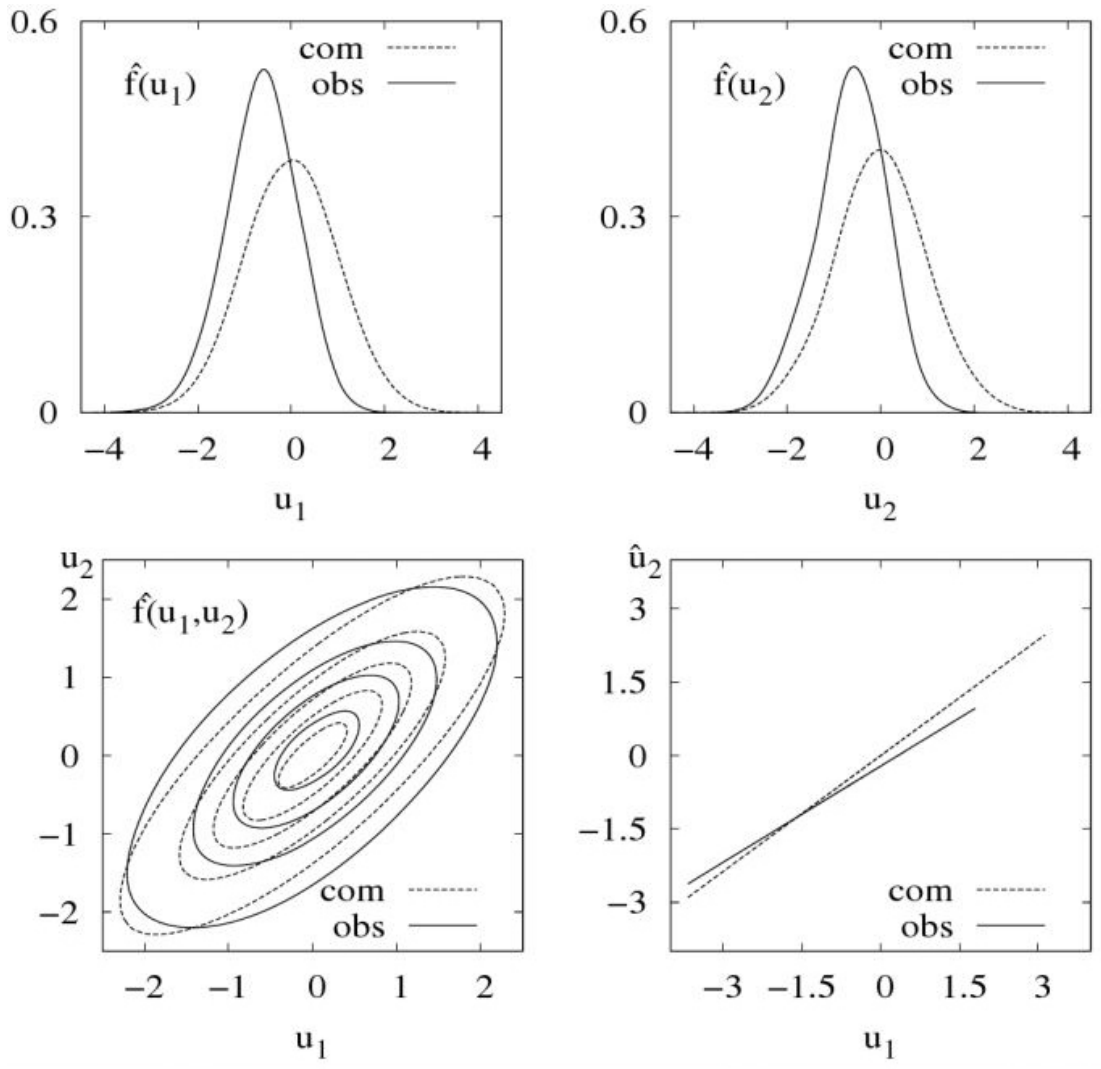
Not Missing at Random (NMAR)

Wahrscheinlichkeit für Auftreten von fehlenden Werten hängt auch nach der Kontrolle aller Einflussgrösen von unbeobachteten Werten (der Variable oder anderen nicht erhobenen Merkmalen) ab.

$$f(M|Y_{obs}, Y_{mis}, \phi)$$

- *nonignorable nonresponse*
- Bsp.: Befragten mit hohem Einkommen haben dessen Eingabe häufiger verweigert

Not Missing at Random (NMAR)



- uni- und bivariaten Verteilungen und Regressionsbeziehung sind verzerrt

Abbildung: aus Spieß(2008)[S.9]

Ignorierbarkeit

Kann man bei der Analyse den unbekannten Missingmechanusmus ignorieren?
Es ist möglich, wenn günstige für die Inferenz Eigenschaften des Schätzers erhalten werden. Zwei Voraussätzungen nach Rubin(1976):

- MAR
- Distinktheit (stochastische Unabhängigkeit) der Parametern des Modells und des Missingmechanismus

$\Rightarrow$ NMAR ist nicht ignorierbar!

Anmerkungen

- Unterscheiden zwischen MAR und NMAR allein auf der Basis der beobachteten Werten unmöglich
- Missingmechanismus muss nicht für alle Einheiten identisch sein
- auch bei MCAR mögliche Probleme mit Software bei der Auswertung

- Unterscheiden zwischen MAR und NMAR allein auf der Basis der beobachteten Werten unmöglich
- Missingmechanismus muss nicht für alle Einheiten identisch sein
- auch bei MCAR mögliche Probleme mit Software bei der Auswertung

Missingmechanismus muss nicht für alle Einheiten identisch sein: Einige Personen haben die Fragae zu Einkommen einfach übersehen(MCAR), andere haben die Antwort verweigert, weil sie ihr hohes Einkommen nicht angeben wollten(NMAR)

Übersicht

1. Einführung
   - Definition
   - Ursachen
   - Kosequenzen
2. Missing Data Pattern
3. Missingmechanismen
   - Missing Completely at Random
   - Missing at Random
   - Not Missing at Random
   - Ignorierbarkeit
4. Behandlungsmethoden
   - Fallreduktion
   - Imputationsverfahren

Bechandlung von fehlenden Werten

- Fallreduktion (Complete Case, Available Case)
- Ersetzungsmethoden
  - Mittelwert-, Median-, Modusimputation
  - Regressionsmethoden
  - Hot-, Cold-Deck Techniken
  - Substitution
  - ML-Algorithmus
  - ...

Complete Case Analyse aka Listwise Deletion

Nur Fälle mit gültigen Werten für alle beteiligten Variablen werden berücksichtigt

- einfache Anwendung
- univariate Statistiken vergleichbar, da gleicher Stichprobenumfang.

- Informationsverlust
- Verzerrung bei MAR oder NMAR
- Fallzahl kann sehr gering sein (z.B. bei 20 Variablen und 10% Chance auf MD für jede Variable wird nur 13% der Fallen vollständig)

Available Case Analyse aka Pairwise Deletion

Es werden alle Fälle benutzt, die bei zu interessierenden Variablen vollständig sind

- Verwendung aller Daten
- unterschiedliche Fall-Basis für die Analyse

$\Rightarrow$ Vergleichbarkeit erschwert

In Extremfall sind Korrelationen so inkonsistent, dass eine indefinite Korrelationsmatrix resultiert $\Rightarrow$ fehlerhafte Regressionsanalyse

Complete Case Vs. Available Case

- beide Methoden sing generell nicht zufriedenstellend
- bei mäßigen Korrelationen und MCAR - paarweise Behandlung möglich
- bei starken Korrelationen - fallweise Behandlung

- beide Methoden sing generell nicht zufriedenstellend
- bei mäßigen Korrelationen und MCAR - paarweise Behandlung möglich
- bei starken Korrelationen - fallweise Behandlung

Complete Case Analyse wird zwar häufig angewandt, ist aber generell keine gute Lösung des Problems

Mittelwertimputation

Fehlende Werte werden durch Mittelwert der Variable ersetzt

- einfache Durchführung
- aber es werden keine neue Informationen beigefügt
  $\Rightarrow$ Verzerrte Verteilung der Variable, Unterschätzte Varianz und Kovarianz
- es können unrealistische Werte auftreten

Beispiel aus Cohen et al. (2003):
Gehalt und Zitierungsniveau der Publikationen

Analysis	**N**	**r**	β	**St.Error**
Complete cases	62	.55	310.747	60.95
Mean inputation	69	.54	310.747	59.13

Keine empfehlenswerte Methode.

Regressionsimputation

Vorhersage der fehlenden Werten mit dem Regressionsmodell, anhand von Variablen mit vollständigen Beobachtungen. Voraussetzungen:

- MCAR oder MAR
- symmetrische Verteilung der Zielgrösse
- monotoner Missingmechanismus

Vor- und Nachteile:

- Berücksichtigung der Unsicherheit durch Residuum (stochastische Regressionsimputation)
- unrealistische Werte können auftreten
- Reduzierung des Standardfehlers, da grössere Stichprobe, aber keine neue Information

Hot-Deck-Methode

Ersetzen durch beobachtete Variablenwerte, die aus Einheiten mit ähnlichen Charakteristiken zufällig gezogen werden.

- Verteilung der Variable unter MCAR nicht verletzt
- Wahl der Variablen, die Ähnlichkeit beurteilen, ist subjektiv
- verschiedene Ziehungsmethoden denkbar (z.B. mit/ohne Zurücklegen, sequentielle Hot-Deck Imputation)

Cold-Deck-Methode

Donor-Einheiten für Ersetzung stammen aus einer externen Quelle derselben Population (z.B. Voruntersuchung)

- Verteilung der Variable unter MCAR nicht verletzt
- Ersetzung durch konstanten Wert oder ähnlich wie bei Hot-Deck
- Anwendbarkeit eingeschränkt, da die Bezugsquelle oft fehlt

Anmerkungen

- bei fast allen Methoden der einfachen Imputation bleibt Unsicherheit nicht berücksichtigt
- Standardfehler der Schätzer systematisch unterschätzt
- Möglichkeit scheinbar signifikanter Ergebnisse

- bei fast allen Methoden der einfachen Imputation bleibt Unsicherheit nicht berücksichtigt
- Standardfehler der Schätzer systematisch unterschätzt
- Möglichkeit scheinbar signifikanter Ergebnisse

Jede Imputation stellt einen Schätzwert für nicht beobachtete Daten dar. Mann soll diese Unsicherheit, die mit Schätzung verbunden ist, berücksichtigen, sonst kann es zu scheinbar signifikanten Ergebnissen führen

Literatur

- Baltes-Götz, B. (2013). Behandlung fehlender Werte in SPSS und Amos. Zentrum für Informations-, Medien und Kommunikationstechnologie (ZIMK), Trier
- Cohen, J. & Cohen, P., West, S. G. & Aiken, L. S. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd edition. Mahwah, N.J.: Lawrence Erlbaum
- Little, R. J. A und Rubin, D. B. (2002). Statistical Analysis with Missing Data. John Wiley & Sons, New Jersey.
- Mayer, B. (2010) Fehlende Werte in klinischen Verlaufsstudien - Der Umgang mit Studienabbrechern, Ulm
- Schnell, R. (1986) Missing-Data-Probleme in der empirischen Sozialforschung, Bochum
- Spieß, M. (2008). Missing-data-Techniken: Analyse von Daten mit fehlenden Werten. LIT Verlag, Münster

Ende

Vielen Dank für die Aufmerksamkeit!