Missing Data

Kai Arzheimer | Vorlesung Forschungsmethoden

Übersicht

- Missing Data: Typen
- Strategien
  - Traditionelle Ansätze
  - (Full Information) Maximum Likelihood
  - Multiple Imputation
    - Vor- und Nachteile
    - Multiple Imputation: Ansatz
    - Implementation
- Fazit

Missing Data: Typen

Literaturempfehlung für heute

Paul Allison: Missing Data. Sage, 2001.

Was meinen wir mit „missing“?

Was meinen wir mit „missing“?

unit nonresponse: Ausgewählte Personen nehmen nicht an Umfrage teil

Was meinen wir mit „missing“?

unit nonresponse: Ausgewählte Personen nehmen nicht an Umfrage teil

item nonresponse: Antworten fehlen für einzelne Fragen

Was meinen wir mit „missing“?

unit nonresponse: Ausgewählte Personen nehmen nicht an Umfrage teil

item nonresponse: Antworten fehlen für einzelne Fragen

missing by design: Fragen werden nur einer zufällig ausgewählten Teil-Stichprobe gestellt

Was meinen wir mit „missing“?

unit nonresponse: Ausgewählte Personen nehmen nicht an Umfrage teil

item nonresponse: Antworten fehlen für einzelne Fragen

missing by design: Fragen werden nur einer zufällig ausgewählten Teil-Stichprobe gestellt

- Heute: item nonresponse

Was tun bei unit nonresponse?

- Typischer Fall: Zu wenige Niedriggebildete in der Stichprobe
  - Standardprozedur: Repräsentativgewichtung, Anpassung der Randverteilung bekannter Merkmale
  - Hoffnung: Gewichtung paßt auch Verteilung unbekannter Merkmale an (Voraussetzung?)
- In der Praxis oft kaum Unterschiede, vor allem, wenn Gewichtungsvariablen als unabhängige Variablen fungieren
- Extreme Gewichtungsfaktoren problematisch
- Effizienz vs. bias

Welche Arten von item nonresponse gibt es?

Bezeichnung	Bedeutung	Beispiel
„Missing Completely At Random“ (MCAR)	Ausfall von y ist unabhängig vom Wert y und vom Wert anderer Variablen ($x_1 \ldots$)	Übertragungsfehler beim Eingeben der Fragebögen / missing by design

Welche Arten von item nonresponse gibt es?

Bezeichnung	Bedeutung	Beispiel
„Missing Completely At Random“ (MCAR)	Ausfall von y ist unabhängig vom Wert y und vom Wert anderer Variablen ($x_1 \ldots$)	Übertragungsfehler beim Eingeben der Fragebögen / missing by design
„Missing At Random“ (MAR)	Ausfall von y ist unabhängig vom Wert y, wird aber vom Wert anderer Variablen ($x_1 \ldots$) beeinflußt	Niedriges Politikinteresse führt zu Ausfällen bei Fragen, die sich auf Politik beziehen

Welche Arten von item nonresponse gibt es?

Bezeichnung	Bedeutung	Beispiel
„Missing Completely At Random“ (MCAR)	Ausfall von y ist unabhängig vom Wert y und vom Wert anderer Variablen ($x_1 \ldots$)	Übertragungsfehler beim Eingeben der Fragebögen / missing by design
„Missing At Random“ (MAR)	Ausfall von y ist unabhängig vom Wert y, wird aber vom Wert anderer Variablen ($x_1 \ldots$) beeinflußt	Niedriges Politikinteresse führt zu Ausfällen bei Fragen, die sich auf Politik beziehen
„Non-Ignorable“ (NI)/ „Missing Not At Random“ (MNAR)	Ausfall von y wird vom Wert von y und/oder nicht beobachteten Variablen beeinflußt	Antwortausfall bei heiklen Fragen

Welche Arten von item nonresponse gibt es?

- MCAR in der Regel völlig unrealistisch
- MAR $\approx$ „ignorable“: Ausfallmechanismus muß nicht modelliert werden
- NI: Ausfallmechanismus muß Bestandteil des substantiellen Modells sein (geringe praktische Relevanz)
- In der Praxis Kontinuum zwischen MAR und NI (Einkommen)
- Alle Techniken sind nur Krücken, missingness vermeiden

Warum ist missingness ein Problem?

- Bias wenn Ausfälle nicht zufällig
- Stichprobenumfang reduziert $\rightarrow$ ineffizient
- Standardfehler zu optimistisch (“listwise deletion is evil” – or is it?)

Was kann man tun?

1. Listwise Deletion
2. Pairwise Deletion
3. Dummy Variable Adjustment
4. (Konventionelle) Imputation

Was kann man tun?

1. Listwise Deletion
2. Pairwise Deletion
3. Dummy Variable Adjustment
4. (Konventionelle) Imputation
5. Full Information Likelihood (FIML)
6. Multiple Imputation

Strategien

Was ist listwise deletion?

- Nur vollständiger Fälle (Voreinstellung)
- Ist listwise deletion „evil“?
  - **Ja…**
  - Korrekte Schätzungen/Standardfehler wenn MCAR (Stichprobe aus Stichprobe)
  - *Bei MAR Verzerrungen möglich*

Was ist listwise deletion?

- Nur vollständiger Fälle (Voreinstellung)
- Ist listwise deletion „evil“?
  - **Ja...**
  - Korrekte Schätzungen/Standardfehler wenn MCAR (Stichprobe aus Stichprobe)
  - *Bei MAR Verzerrungen möglich*
- Jedoch...
  - Relativ unproblematisch für missingness der unabhängigen Variablen (wenn nicht vom Wert der abhängigen Variablen beeinflußt)
  - Für logistische Regression sogar missingness der abhängigen Variablen unproblematisch, wenn missingness nicht von unabhängigen Variablen beeinflußt

Was ist listwise deletion?

- Nur vollständiger Fälle (Voreinstellung)
- Ist listwise deletion „evil“?
  - **Ja...**
  - Korrekte Schätzungen/Standardfehler wenn MCAR (Stichprobe aus Stichprobe)
  - *Bei MAR Verzerrungen möglich*
- Jedoch...
  - Relativ unproblematisch für missingness der unabhängigen Variablen (wenn nicht vom Wert der abhängigen Variablen beeinflußt)
  - Für logistische Regression sogar missingness der abhängigen Variablen unproblematisch, wenn missingness nicht von unabhängigen Variablen beeinflußt
- Aber: Bei komplexeren Modellen dramatische Reduktion der Stichprobengröße
- *Beispiel fünf Prozent missingness, 20 Variablen:* $0{,}95^{20} \approx 0{,}36$

Was ist pairwise deletion?

- Zur Schätzung vieler linearer Modelle genügt statt Rohdaten Kovarianzmatrix
- Pairwise Deletion: Für jede Kovarianz alle verfügbaren Fälle verwenden $\rightarrow$ mehr Fälle als bei listwise deletion $\rightarrow$ unterschiedliche Fallzahlen
- Ambiguitäten bei Tests etc.; Standardfehler nicht korrekt
- *Bei ernsthafter missingness Inkonsistenzen möglich/wahrscheinlich; Modell nicht schätzbar*
- (Spielt eigentlich nur in SPSS eine Rolle)

Was ist Dummy Variable Adjustment?

- Fehlende Werte für unabhängige Variable x
  - Dummy d für missingness
  - ergänzte Variable x^* mit konstantem Wert, der fehlende Werte ersetzt
- Regression von y auf d und x^*
- Alle Fälle werden genutzt, einfache Interpretation
- Früher in Standardwerken empfohlen
- *Leider selbst bei MCAR sehr stark verzerrte Parameterschätzungen möglich*

Was ist Imputation?

- Fehlende Werte werden ersetzt
- Z. B. durch Mittelwert (ganz schlecht) oder durch Regression auf beobachtete Werte (funktioniert bei MCAR)
- Wird kompliziert, wenn mehrere Variablen betroffen sind
- *Generell sind Standardfehler zu optimistisch, weil Imputation wie reale Daten betrachtet werden*

Was ist das Zwischenfazit?

- Konventionelle Methoden machen die Sache oft noch schlimmer

Was ist das Zwischenfazit?

- Konventionelle Methoden machen die Sache oft noch schlimmer
  - bias
  - Falsche (optimistische) Standardfehler
  - Sehr anfällig wenn Daten NI

Was ist das Zwischenfazit?

- Konventionelle Methoden machen die Sache oft noch schlimmer
  - bias
  - Falsche (optimistische) Standardfehler
  - Sehr anfällig wenn Daten NI
- *Unter konventionellen Ansätzen listwise deletion oft die am wenigsten schlechte Alternative*

Was ist das Zwischenfazit?

- Konventionelle Methoden machen die Sache oft noch schlimmer
  - bias
  - Falsche (optimistische) Standardfehler
  - Sehr anfällig wenn Daten NI
- *Unter konventionellen Ansätzen listwise deletion oft die am wenigsten schlechte Alternative*
- Es geht auch besser (theoretisch)

Wie kann ML hier helfen?

- ML findet gute Parameterschätzungen, indem (Log-)Likelihood-Funktion maximiert wird
- (Log-)Likelihood ist eine (modellspezifische) Funktion der Daten und der Vermutungen über den Wert der Parameter
- Fallweise Berechnung, Multiplikation (Fälle unabhängig)
- Wenn MAR gilt
  - kann für fehlende Werte die Summe (diskrete Variable)
  - beziehungsweise das Integral (kontinuierliche Variable)
  - der Likelihood-Funktion über die möglichen Werte eingesetzt werden
- $\rightarrow$ Full Information Maximum Likelihood = FIML

Probleme/Komplikationen?

- Besondere Algorithmen, Vorkehrungen für Standardfehler
- Konkrete Vorgehensweise hängt vom Modell und vom Muster der Ausfälle ab
- Problem: *Modell der gemeinsamen Verteilung aller Variablen mit fehlenden Werten* erforderlich
- In der Regel: Annahme *multivariater Normalverteilung*
  - Völlig unrealistisch
  - Erstaunlicherweise trotzdem recht robust
- *Methode muss für jedes statistische Modell gesondert implementiert werden*
- In Stata: (G)SEM; Standard in MPlus

Probleme/Komplikationen?

- Besondere Algorithmen, Vorkehrungen für Standardfehler
- Konkrete Vorgehensweise hängt vom Modell und vom Muster der Ausfälle ab
- Problem: *Modell der gemeinsamen Verteilung aller Variablen mit fehlenden Werten* erforderlich
- In der Regel: Annahme *multivariater Normalverteilung*
  - Völlig unrealistisch
  - Erstaunlicherweise trotzdem recht robust
- *Methode muss für jedes statistische Modell gesondert implementiert werden*
- In Stata: (G)SEM; Standard in MPlus
- Allgemeinere Lösung? $\rightarrow$ Multiple Imputation

Was sind die Vorteile der multiplen Imputation (MI)?

- Hat dieselben optimalen Eigenschaften wie ML
- Kann mit (praktisch) jedem statistischen Modell kombiniert werden
- Extrem flexibel solange MAR
- Anwendung mit Standardsoftware (STATA) relativ leicht möglich (Zusatzmodule oder `mi`-Befehle ab Stata 11)

Wo ist der Haken?

- (Vor der Analyse mit Standardsoftware Einsatz spezieller Programme notwendig)
- (Etwas) mühsam, zeit- und rechenaufwendig
- Organisationsaufwand, Fehleranfälligkeit, wenn keine speziellen Erweiterungen für Standardsoftware genutzt werden
- Auch hier: Vorüberlegungen, Zweifelsfälle
- Kommunkation der Ergebnisse nicht unproblematisch

Wie funktioniert die einfache random imputation?

Beispiel aus Allison:

- x und y sind bivariat standard-normalverteilt mit einer Korrelation von 0,3
- Die Hälfte der y wird zufällig gelöscht (MCAR)
- Fehlende Werte von y durch Regression auf x ersetzen $\rightarrow$ Analyse der komplettierten Daten $\rightarrow$ Korrelation von 0,42

Wie funktioniert die einfache random imputation?

Beispiel aus Allison:

- x und y sind bivariat standard-normalverteilt mit einer Korrelation von 0,3
- Die Hälfte der y wird zufällig gelöscht (MCAR)
- Fehlende Werte von y durch Regression auf x ersetzen $\rightarrow$ Analyse der komplettierten Daten $\rightarrow$ Korrelation von 0,42
- Warum? $\rightarrow$ Vorhergesagte Werte berücksichtigen nur systematischen Teil (deterministischer Zusammenhang)
- bias, da zuwenig Streuung für imputierte Werte

Wie funktioniert die einfache random imputation?

Beispiel aus Allison:

- x und y sind bivariat standard-normalverteilt mit einer Korrelation von 0,3
- Die Hälfte der y wird zufällig gelöscht (MCAR)
- Fehlende Werte von y durch Regression auf x ersetzen $\rightarrow$ Analyse der komplettierten Daten $\rightarrow$ Korrelation von 0,42
- Warum? $\rightarrow$ Vorhergesagte Werte berücksichtigen nur systematischen Teil (deterministischer Zusammenhang)
- bias, da zuwenig Streuung für imputierte Werte
- Zufällig Werte aus der Verteilung der Residuen (von y auf x) ziehen und zu imputierten Werten addieren
- Bias verschwindet fast vollständig, da nun Verteilung der imputierten Werte mit Verteilung der beobachteten Werte identisch

Wie funktioniert die multiple random imputation?

- Problem: (Zufällig) imputierte Daten werden wie reale Daten behandelt $\rightarrow$ Standardfehler zu klein
- Wie kann man Unsicherheit über fehlende Werte berücksichtigen? $\rightarrow$ multiple Imputation, z. B. acht Datensätze
- Wegen zufälliger Komponente unterscheiden sich Datensätze
  - Wenn Unsicherheit über fehlende Werte gering, sind Datensätze fast identisch
  - Je größer die Unsicherheit, desto stärker die Differenzen

Wie kommt man zu Ergebnissen?

- Analyse jedes einzelnen Datensatzes
- Parameterschätzung: Arithmetischen Mittelwert über acht Einzelschätzungen bilden
- Standardfehler: Anwendung der „Rubin-Regel“

Wie werden die Standardfehler berechnet?

$$\text{Korrigierter Standardfehler des Parameters} = \sqrt{\frac{1}{M}\sum_{j=1}^{M} s_j^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{j=1}^{M}(r_j-\bar{r})^2}$$

- M: Zahl der imputierten Datensätze
- $\frac{1}{M}\sum_{j=1}^{M} s_j^2$: Mittelwert der Schätzung des quadrierten Standardfehlers des Parameters über alle Imputationen
- $\frac{1}{M-1}\sum_{j=1}^{M}(r_j-\bar{r})^2$: Varianz der Parameterschätzungen
- $1+\frac{1}{M}$: Korrekturfaktor

Was war da mit Variation der Parameterschätzungen?

- Fehlende Werte von x werden durch Regressionsmodell $x = \beta_0 + \beta_1 y + \epsilon$ ersetzt, das zufällige Komponente berücksichtigt
- Aber: Modellparameter (β_0, β_1, σ^2_ϵ) basieren ja selbst nur auf Schätzungen
- Müssen deshalb über Schätzungen variieren
- Zufällige Ziehung der Werte aus *Verteilung möglicher Modellparameter*
- Macht bei hoher Rate von missingness einen Unterschied (größere korrigierte Standardfehler)

Was passiert in komplexeren Fällen?

- Wir brauchen ein Imputationsmodell
- Meistens: Multivariates normales Modell
  - Alle Variablen sind normalverteilt
  - Jede Variable ist als Linearkombination aus den übrigen Variablen und einem homoskedastischen Fehlerterm (ϵ) darstellbar
- In der Regel völlig unrealistisch
- Aber als Imputationsmodell relativ robust; Transformationen

Was macht der Computer konkret?

- Dummerweise ist die Verteilung der Parameter nicht bekannt
- Kann wegen der fehlenden Werte nicht mal korrekt geschätzt werden
- Iterative Algorithmen die zwischen
  - Zufälligen Ziehungen aus der Verteilung der Parameter und
  - Zufälligen Ziehungen aus der Verteilung der fehlenden Werte pendeln
- Wenn der Algorithmus konvergiert, zufällige Ziehungen aus der gemeinsamen Verteilung von Daten und Parametern
- In Abhängigkeit von beobachteten Werten (MAR)
- Verteilungen in der Regel nicht analytisch darstellbar, Zugriff über Simulationsverfahren
- Immenser numerischer Aufwand

Was ist MICE?

- Multivariates normales Modell: *Gemeinsames* Imputationsmodell für *alle* Variablen
- Multiple Imputation by Chained Equations:
  - Für jede Variable mit fehlenden Werten individuelles Imputationsmodell
  - Lineare Regression, Logit, Probit, Poisson…
  - Starke Annahme: Indivdiduelle Verteilungen sind miteinander kompatibel & brauchbare Approximation für *gemeinsame* Verteilung
- Sehr flexibel, Implementation in Stata, Behandlung von Dummies und transformierten Variablen

Fazit

Was sind die Hauptergebnisse?

- Item nonresponse als Problem wird unterschätzt und sollte vermieden werden
- Für einfache Modelle mit wenig missingness ist listwise deletion eine brauchbare Lösung
- Andere konventionelle Ansätze vermeiden
- Für komplexere Modelle Likelihood-basierte Lösung oder MI um
  - Vorhandene Daten auszuschöpfen und
  - Korrekte Parameterschätzungen/Standardfehler zu erhalten
- Möglichst viel zusätzliche Information nutzen, um MAR-Bedingung realistischer zu machen (Imputationsmodell $\neq$ Analysemodell)

Literaturempfehlung für nächste Woche (matching)

Jasjeet S. Sekhon, Opiates for the Matches: Matching Methods for Causal Inference, Annual Review of Political Science 2009, p. 487–508, DOI http://www.annualreviews.org/doi/abs/10.1146/annurev.polisci.11.060606.135444