Imputationsverfahren

Minh Ngoc Nguyen

Betreuer: Eva Endres

München, 09.01.2015

Imputation

- Prinzip: fehlende Werte sollen durch möglichst passenden Werte ersetzt werden
- Vorteil
  - Erzeugen den vollständigen Datensatz
  - Anwenden Standardmethode für den vervollständigten Datensatz
- Es gibt eine Vielzahl von Verfahren
- Man unterscheidet zwischen Single-Imputation und Multiple-Imputation

Imputation

Single-Imputation: jeder fehlende Wert wird durch einen Wert ersetzt
Aber

- Berücksichtigen die Unsicherheit nicht
- Unterschätzen die Varianz
- Führen fälschlicherweise signifikante Ergebnisse möglich

Multiple-Imputation: jeder fehlende Wert wird durch mehrere Werte ersetzt

Notation

- $Y = (y_{ij})$: vollständige $(n \times p)$ Datenmatrix mit n Beobachtungen von p Variablen
- Y_{beob}: beobachtete Teil u. Y_{fehl}: fehlende Teil, d.h. $Y = (Y_{beob}, Y_{beob})$
- $R = (r_{ij})$: Indikatormatrix mit $r_{ij} = 0$ falls y_{ij} fehlt oder $r_{ij} = 1$ falls y_{ij} beobachtet
- ξ: unbekannter Parameter steuert den Fehlendmechanismus R
- θ: unbekannter Parameter steuert die Daten Y

Fehlendmechanismen

- MCAR: Fehlende Werte sind missing completely at random, wenn gilt

$$g(R|Y,\xi) = g(R|\xi)\ \forall Y,\xi$$

$\rightarrow$ Unproblematischster, aber unwahrscheinlicher Fall

- MAR: Fehlende Werte sind missing at random, wenn gilt

$$g(R|Y,\xi) = g(R|Y_{beob},\xi)\ \forall Y_{fehl},\xi$$

$\rightarrow$ Mäßig problematisch und wahrscheinlicher Fall

- MNAR: Fehlende Werte sind missing not at random, wenn gilt

$$g(R|Y,\xi) = g(R|Y_{beob}, Y_{fehl},\xi)\ \forall Y,\xi$$

$\rightarrow$ Sehr problematisch und nicht unwahrscheinlicher Fall

Ignorierbarkeit

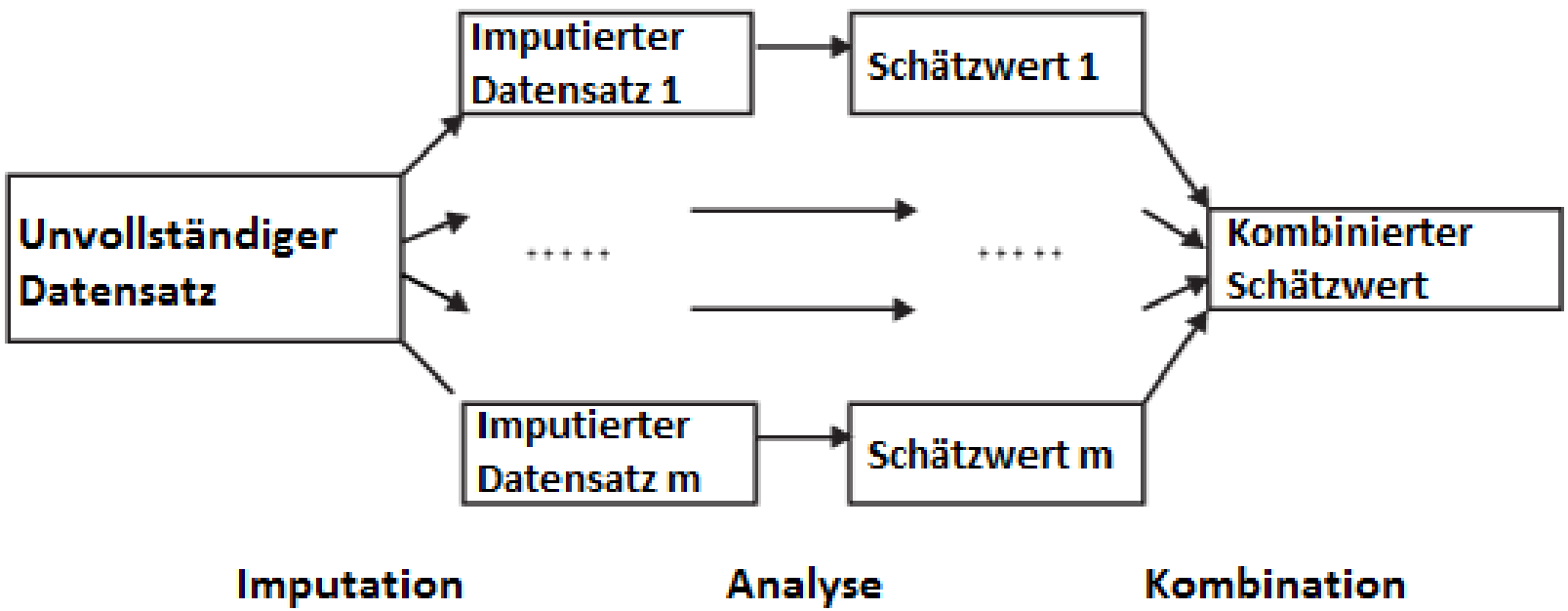
Ignorieren des Fehlendmechanismus möglich, wenn

- die Daten MAR oder MCAR sind
- θ und ξ voneinander unabhängig sind

Das Vorliegen eines ignorierbaren Fehlendmechanismus ist Voraussetzung für die meisten Imputationsverfahren

Grundkonzepte der Multiple-Imputation

Jeder fehlende Werte wird durch m Werte ersetzt (mit $m > 1$)
$\rightarrow$ m vervollständigte Datensätze erhalten gleichen beobachteten Werte aber unterschiedlichen imputierten Werten



Kombinationsregeln (Little und Rubin, 2002)

- $\hat{Q}$: Schätzung interessierender Parameter (z.B: Mittelwert, Regressionskoeffizienten)
- $\hat{V}$: Schätzung der Varianz
- Für $i = 1, \cdots, m$
  - $\hat{Q}_i$: Schätzung aus i-ten Datensatz
  - $\hat{V}_i$: zugehörige geschätzte Varianz aus i-ten Datensatz

Kombinationsregeln (Little und Rubin, 2002)

Schätzwert

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

Gesamtvarianz

$$T = (1 + \frac{1}{m})B + \bar{V}$$

mit

- Within-Varianz $\bar{V} = \frac{1}{m} \sum_{i=1}^{m} \hat{V}_i$
- Between-Varianz $B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2$

Kombinationsregeln (Little und Rubin, 2002)

Standardabweichung

$$\sqrt{T}$$

Konfidenzintervall

$$KI = \bar{Q} \pm t_{1-\frac{1}{\alpha}} \sqrt{T}$$

mit

- $t_{1-\frac{1}{\alpha}}$: Quantil der t-Verteilung
- $df = (m-1)(1+\frac{m\bar{V}}{(m+1)B})^2$

Multipe-Imputation: Vorteil

- Unsicherheit durch Variabilität der m Schätzwert wird berücksichtigt
- Verwendung von sämtlicher zur Verfügung stehender Information (beobachtete Daten)
- Analyse der vollständigen Daten mit beliebiger Standardmethode
- Schätzer haben hohen Effizienz bei bereits niedrigem m

Multipe-Imputation: Nachteil

- Aufwendige Vorgehen zur Erzeugung der vervollständigten Datensätze
- Benötigen größere Rechen- und Speicherkapazität
- Steigern die Analysezeit bei separaten Betrachtung jeder einzelnen imputierten Datensatz erheblich
- Ist der Anteil der fehlenden Werte groß, so muss auch m gesteigert werden und der Aufwand steigt ebenfalls

Maximum-Likelihood Theorie

Falls die fehlenden Daten ignorierbar sind, kann die Likelihood der beachteten Daten wie folgt

$$L(\theta|Y_{beob}) = \int f(Y_{beob}, Y_{fehl}|\theta)dY_{fehl}$$

- bei normaler Maximum-Likelihood gilt es hier das Maximum dieser Funktion zu finden
- bei Fälle von fehlenden Daten sind kompliziert

$\rightarrow$ Verwendung des Expectation-Maximization Algorithmus

Expectation-Maximization Algorithmus

- EM Algorithmus nach Dempster (1977)
- Iteratives Verfahren zur Bestimmung von Maximum Likelihood Schätzwerten
- Grundidee
  - Ersetze fehlende Werte durch plausible Startwerte
  - Schätze die Parameter
  - Ersetze die fehlende Werte unter Verwendung des jeweilig zuvor geschätzten Parameter aktualisiert
  - Die fehlende Werte und die Parameter werden so lange neu geschätzt bis es zur Konvergenz kommt

EM Algorthimus

Verwende Startwert $\theta^{(0)}$, z.B. Mittelwerte und Kovarianzen

Es besteht aus 2 Schritten
E-Schritt: Berechne

$$Q(\theta|\theta^{(t)}) = E(l(Y|\theta)|Y_{beob}, \theta^{(t)})$$

M-Schritt: Wähle $\theta^{(t+1)}$ so, dass

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

Dieser iterative Prozess wird solange fortgeführt, bis es konvergiert

EM Algorithmus

Vorteil

- In jedem Iteration-Schritt wird Likelihood der beobachteten Daten erhöht
  $\rightarrow$ EM Algorithmus einfach zu konstruieren
  $\rightarrow$ jeder Schritt sich leicht interpretieren lässt
- Parameter-Restriktionen sind meist automatisch erfüllt

Nachteil

- Langsame Rate der Konvergenz (hängt von dem Anteil der fehlenden Daten ab)
- Schwierig die Varianz zu bestimmen

Bayes Ansatz

Erzeuge m unabhängige Zufallsziehungen für die fehlenden Daten aus ihren a-posteriori prädiktive Verteilung. Unter Annahmen des ignorierbaren Fehlendmechanismus

$$f(Y_{fehl}|Y_{beob}) = \int f(Y_{fehl}|Y_{beob},\theta) f(\theta|Y_{beob}) d\theta$$

mit
$f(Y_{fehl}|Y_{beob},\theta)$: bedingte prädiktive Verteilung
$f(\theta|Y_{beob})$: a-posteriori Verteilung

Bayes Ansatz

Realisation direkt durch

1. Zufallszüge für θ aus den a-posteriori Verteilung $f(\theta|Y_{beob})$
2. Zufallszüge für Y_{fehl} aus ihren bedingten prädiktiven Verteilung $f(Y_{fehl}|Y_{beob}, \theta)$ für aktuelle θ

Anschließend werden zwei Algorithmus zur Beschaffung solcher Zufallsziehungen vorgestellt:

- Data-Augmentation
- Multivariate Imputation by Chained Equations

Data-Augmentation Algorithmus

- "Datenmehrung" nach Tanner und Wong (1987)
- Iteratives Verfahren zur Simulation der a-posteriori Verteilung von θ
- Bayesianische Variante des bereits EM Algorithmus
- Häufig für Imputation verwendete Markov Chain Monte Carlo (MCMC) Methode
- Eine gemeinsame Verteilung aller Variable wird vorgenommen (z.B. multivariate Normalverteilung)

DA Algorithmus

Verwende Startwert $\theta^{(0)}$, z.B ML durch EM-Algorithmus

Es besteht aus 2 Schritten:

I-Schritt: Ziehe

$$Y_{fehl}^{(t+1)} \sim f(Y_{fehl}|Y_{beob}, \theta^{(t)})$$

P-Schritt: Ziehe

$$\theta^{(t+1)} \sim f(\theta|Y_{beob}, Y_{fehl}^{(t+1)})$$

Die Ergebnis des Verfahren ist eine Markov Kette $\left\{(\theta^{(t)}, Y_{fehl}^{(t)}), t = 1, 2, \cdots\right\}$ mit der stationären Verteilung $f(\theta, Y_{fehl}|Y_{beob})$

DA Algorithmus

Konvergenz

- Benötigen viele Iterationen bis zur Konvergenz
- Bestimmung der Konvergenz mit Hilfe grafischer Darstellung und einfacher Kennwerte wie Autokorrelationen oder Varianzen

Konvergenzgeschwindigkeit hängt von

- dem Anteil der fehlenden Werten
- Startwerte

ab.

Multivariate Imputation durch Chained Equations

- Bekannt auch unter anderen Namen wie Full Conditional Specification oder Chained Equations
- Beim MICE
  - Vermeiden die gemeinsame Verteilung aller Variablen
  - Spezifizieren separat für jede Variable eine bedingte Verteilung

MICE Algorithmus

- Multivariate Imputationen stellen eine Kette von univariaten Imputationen dar
- Für jede Variablen lassen sich verschiedene Modelle spezifizieren (je nach Skalierung)
- Iterative Imputationsverfahren
- Mit Hilfe der iterativen Ziehungen aus den bedingten Verteilungen kann die gemeinsame Verteilung modelliert werden
  $\rightarrow$ Idee des Gibbs-Sampler

MICE Algorithmus

$$\theta_1^{(t)} \sim f(\theta_1 | Y_{beob,1}, Y_2^{(t-1)}, \cdots, Y_p^{(t-1)})$$

$$Y_{fehl,1}^{(t)} \sim f(Y_{fehl,1} | Y_{beob,1}, Y_2^{(t-1)}, \cdots, Y_p^{(t-1)}, \theta_1^{(t)})$$

$$\vdots$$

$$\theta_p^{(t)} \sim f(\theta_p | Y_{beob,p}, Y_1^{(t)}, \cdots, Y_{p-1}^{(t)})$$

$$Y_{fehl,p}^{(t)} \sim f(Y_{fehl,p} | Y_{beob,p}, Y_1^{(t)}, \cdots, Y_{p-1}^{(t)}, \theta_p^{(t)})$$

Die wird bis zu einer vorgegebenen Anzahl der Iteration $10 - 20$ weiter durchgeführt.

MICE Algorithmus

MICE-Algorithmus bietet den Vorteil, dass komplexe Datenstrukturen berücksichtigt werden können

- Zählvariable: Poisson Regression
- Stetige Variable: normale lineare Regressionsmodell
- Kategoriale Variable: logistisches oder verallgemeinertes logistisches Modell

DA- vs. MICE- Algorithmus

Gemeinsamkeit

- iterative Algorithmus als MCMC-Methode
- Zufallszug von Parameter aus der a-posteriori Verteilung

Unterschied

- DA modelliert eine gemeinsame Verteilung während MICE Zug um Zug bedingte Verteilungen konstruieren
- Bei dem MICE-Algorithmus kommt Konvergenz meist schnell zustande, aber bei dem DA-Algorithmus nicht

Simulationsdesign

Insgesamt werden die drei Variablen X_1, X_2 und Y mit jeweils 10000 Beobachtungen generiert.
Es sei

$$(X_1, X_2) \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix})$$

$$Y = \beta_1 X_1 + \beta_2 X_2$$

mit $\beta_1 = \beta_2$ und $\epsilon \sim N(0, 1)$

Simulationsdesign

Datensatz mit fehlenden Werten

- MCAR: Beobachtungen in X_2 fehlen mit einer Wahrscheinlichkeit von γ unabhängig von Y und X_1
- MAR abhängig von X_1: Beobachtungen in X_2 fehlen wenn X_1 kleiner als das γ Quantil von X_1 ist
- MNAR: Beobachtungen in X_2 fehlen wenn X_2 kleiner als das γ Quantil von X_2 ist.

Verschiedene Anteile der fehlenden Werte werden betrachtet, mit $\gamma = 30, 50, 70, 90\%$

Verwendung des DA Algorithmus

- Multiple-Imputation auf Basis des Data-Augmentation Algorithmus
- $m = 5$
- mit 20 Iteration
- Verwendung von Paket norm der Software R

DA Augmentation: MCAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.0097	0.0119	-0.0046	1.0086	0.50224
50	1.0085	0.0116	-0.0115	1.0308	0.51370
70	1.0163	0.0151	0.0109	1.0178	0.51264
90	1.0403	0.0163	0.0176	1.0429	0.53444

- β_2 treffen den wahren Wert von 1 sehr gut mit geringen Standardfehler
- β_2 und SD steigen mit zunehmenden Fehlwertanteil leicht an
- $\mu_{X_2}, \sigma^2_{X_2}, \rho_{X_1X_2}$ bleiben annähernd

DA Augmentation: MAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.0057	0.0126	0.0131	0.9952	0.48987
50	1.0234	0.0119	0.0201	0.9781	0.49584
70	1.0047	0.0172	0.1127	0.9067	0.43830
90	1.0619	0.0159	0.1518	0.8998	0.43623

- β_2 sind unverzerrt
- Standardfehler sind auch niedrig
- Struktur der Daten bleibt nach der Imputation durch $\mu_{X_2}, \sigma^2_{X_2}, \rho_{X_1X_2}$ nahezu

DA Augmentation: MNAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.1628	0.0296	0.3089	0.5634	0.46311
50	1.2358	0.0217	0.5489	0.4203	0.42106
70	1.3803	0.0270	0.8547	0.3034	0.39576
90	1.4644	0.0355	1.4524	0.1860	0.33250

- β_2 sind stark verzerrt
- SD sind erhöht
- $\mu_{X_2}, \sigma^2_{X_2}$ weichen stark vom wahren Wert ab

Verwendung des MICE Algorithmus

- Multiple-Imputation auf Basis des MICE Algorithmus
- $m = 5$
- Imputationen auf der Basis eines linearen Modells
- Verwendung von Paket mice der Software R

MICE Algorithmus: MCAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.0088	0.0138	-0.0086	1.0088	0.5072
50	1.0151	0.0141	-0.0140	1.0109	0.5156
70	1.0154	0.0134	0.0087	1.0331	0.5095
90	1.0147	0.0126	-0.0124	1.0246	0.5138

- β_2 sind unverzerrt bei geringen SD
- $\mu_{X_2}, \sigma^2_{X_2}$ werden gut die Daten widerspiegelt
- $\rho_{X_1X_2}$ werden durch diesem Algorithmus zudem korrekt nachgebildet

MICE Algorithmus: MAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.0023	0.0118	0.0086	1.0088	0.4950
50	1.0176	0.0127	0.0164	0.9802	0.4952
70	1.0103	0.0143	0.0470	0.9573	0.4793
90	1.0412	0.0129	0.0764	0.9297	0.4690

- Parameterschätzer sind unverzerrt bei geringen SD
- $\mu_{X_2}, \sigma^2_{X_2}$ können nicht exakt wiedergeben, aber diese Werte liegen noch nahe den wahren Wert

MICE Algorithmus: MNAR

γ	β_2	SD	μ_{X_2}	$\sigma^2_{X_2}$	$\rho_{X_1X_2}$
30	1.1623	0.0175	0.3032	0.5746	0.4671
50	1.2534	0.0231	0.5374	0.4282	0.4340
70	1.3613	0.0316	0.8451	0.3134	0.3957
90	1.4515	0.0347	1.4295	0.1933	0.3602
	0.3543				

- Bei erhöhten SD sind Schätzer stark verzerrt
- $\mu_{X_2}, \sigma^2_{X_2}$ werden nicht korrekt wiedergegeben

Fazit

Für MCAR und MAR Fehlendmechanismus

- Unverzerrte Parameterschätzung mit geringen Standardfehler
- Mittelwert und Varianz bleiben gut mit dem wahren Wert

Für MNAR Fehlendmechanismus

- Stark verzerrt

Literatur & Software

B.Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York, USA: John Wiley & Sons.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society,Series B 39(1), 138.

Little, R. and D. Rubin (2002). Statistical Analysis with Missing Data. Hoboken, USA: Wiley & Sons.

Novo, A. A. (2013). Package norm: Analysis of multivariate normal datasets with missing values.

Tanner, M. A. and W. Wong (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 39(1), 138.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45(3), 167.

Vielen Dank für die Aufmerksamkeit!