



Technische Universität München



PROTECT
Behavioral Health Promotion &
Technology Lab

Evaluation Randomisiert-Kontrollierter Studien und Experimente mit R

Missing Data & Multiple Imputation

Prof. Dr. David Ebert & Mathias Harrer

Graduiertenseminar TUM-FGZ

Psychology & Digital Mental Health Care, Technische Universität München

Missing Data-Mechanismen

“Obviously the best way to treat missing data is not to have them.”

Orchard & Woodbury (1972)

Fehlende Werte sind aber, insbesondere in der medizinischen und psychologischen Forschung, häufig **unvermeidbar!**

Taxonomie: Gründe für fehlende Werte in klinischen Studien

- **Instrumente:** Response Burden, zu langes Assessment, ...
- **Teilnehmende:** Überforderung, Privatsphäre, Motivationsprobleme, ...
- **Center:** Inadäquate Umsetzung des Studienmethodik, Personalmangel, ...
- **Personal:** Falsche Datenerfassung, Datenverlust, falsche Dateneingabe, ...
- **Studie:** Technische Fehler, Zeitverzögerungen bei Follow-Ups, ...

Palmer et al. (2018)

Der Umgang mit fehlenden Werten in der (medizinisch-psychologischen) Forschung ist oftmals mangelhaft: (Akl et al., 2015; Bell et al., 2014; Van Buuren, 2018, Kapitel 1.1.2; Wood et al., 2004)

- Verteilung fehlender Werte nicht transparent berichtet
- Unpassendes missing data handling (z.B. **listwise deletion**)
- Adäquate Imputationsmethoden (z.B. MI, FIML) häufig nicht benutzt, inadäquat angewendet, oder unzureichend berichtet

R macht es nicht (automatisch) “richtig!”

```
y <- 1:10
x <- c(1, NA, NA, NA, 3, 5, 8, 10, -1, 10)
summary(lm(y ~ x))
```



```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9403      1.7738   2.785   0.0387 *
## x             0.3172      0.2709   1.171   0.2945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.904 on 5 degrees of freedom
## (3 observations deleted due to missingness)
```

Was dagegen tun?

- Umsetzung von **Reporting-Standards** z.B. STROBE (Von Elm et al., 2007) & CONSORT (Schulz et al., 2010), s.a. Sterne et al. (2009).
- Adäquates **Missing Data Handling**
 - Basierend auf plausiblen Annahmen, warum fehlende Werte entstanden sind (“Missing Data Mechanism”)
 - ggf. Schätzung (Imputation) fehlender Werte unter Einbezug von deren Unsicherheit

MCAR, MAR & MNAR



Donald B. Rubin

Grundannahme: Das Fehlen oder Vorhandensein von Daten ist Resultat eines probabilistischen Prozesses.

Diesem Prozess versucht man sich durch ein Modell (***missing data model***) anzunähern.

(Rubin, 1976)



(Harrer et al., 2021)

Nach Rubin (1976) können Missing Data-Mechanismen in 3 Untertypen klassifiziert werden:

MCAR

Missing Completely At Random: rein zufällig fehlende Werte einer Variable.

MAR

Missing At Random: das Fehlen von Werten einer Variable ist abhängig von anderen (observierten) Variablen.

MNAR

Missing Not At Random / "Nonignorable Missing Data": das Fehlen von Werten einer Variable ist (u.A.) abhängig von den Werten der Variable selbst.

→ Für jede Annahme ergeben sich unterschiedliche Auswirkungen bei der Datenauswertung!

Notation (Van Buuren, 2018, Kapitel 2.2.3 & 2.2.4)

- \mathbf{Y} : $n \times p$ Matrix mit teils fehlenden Werten (n Personen, p Variablen).
- \mathbf{X} : Matrix mit (vollständig observierten) Kovariaten.
- $\mathcal{D} = (\mathbf{Y}, \mathbf{X})$: Gesamter Datensatz.
- \mathbf{R} : $n \times p$ Matrix mit 0 (Datenpunkt fehlt) und 1 (Datenpunkt observiert; “response indicator”).
- \mathbf{Y}_{obs} , \mathbf{Y}_{mis} : Observierte Daten, fehlende Daten.
- ψ : Parameter des Missing Data-Modells (typischerweise nicht für die wiss. Fragestellung selbst relevant).

→ Missing Data-Modelle treffen Aussagen darüber, **in welcher Beziehung** \mathbf{Y}_{obs} , \mathbf{Y}_{mis} und \mathbf{R} miteinander stehen.

MCAR

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow P(\mathbf{R} = 0 | \psi)$$

Werte in \mathbf{Y} fehlen “zufällig” (unabhängig von Werten von \mathbf{Y}). Das Fehlen von Werten ist nur bestimmt durch die allgemeine Wahrscheinlichkeit, dass Werte fehlen (im Datensatz gab es eher viele oder wenige Missings).

Beispiel

Das Fehlen der Variable “Alter” ist weder von der Variable “Neurotizismus”, noch vom Alter der Person selbst abhängig.

MAR

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \psi)$$

Werte in \mathbf{Y} fehlen abhängig von ψ **und** **observierter** Information \mathbf{Y}_{obs} .

Beispiel

Personen mit höheren Neurotizismus-Werten geben ihr Alter seltener an als Personen mit niedrigem Neurotizismus, unabhängig ihres Alters.

MNAR

$$P(\mathbf{R} = 0 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}, \psi) \Rightarrow ?$$

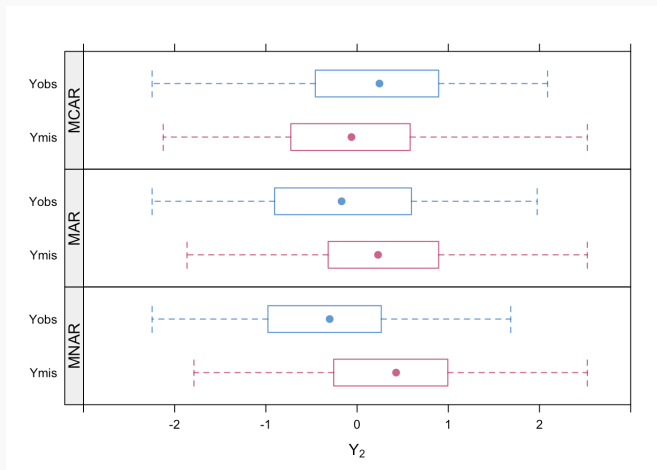
Die Formel lässt sich nicht “vereinfachen”!

Werte in \mathbf{Y} fehlen abhängig von ψ , observierter Information \mathbf{Y}_{obs} , **und** der **unobservierten Information** \mathbf{Y}_{mis} selbst.

Beispiel

Das Fehlen der Alters-Variable hängt (u.A.) vom Alter der Person selbst ab. Eventuell geben ältere Personen mit hohem Neurotizismus ihr Alter seltener an. **Aber das wissen wir nicht**, da die Daten fehlen!

Observierte und unobservierte Werte unter drei Missing Data-Annahmen



aus Van Buuren (2018), Kapitel 2.2.4

Implikationen

- **MCAR:** Da die Werte zufällig fehlen, ist beim Einsatz passender statistischer Verfahren keine Verzerrung der Ergebnisse zu erwarten. Es besteht kein systematischer **Bias**; nur ein Verlust der statistischen Power durch den Datenverlust.
- **MAR:** Bei der Schätzung von Parametern ergeben sich evtl. Verzerrungen (Bias), wenn die abhängigen Variablen nicht im Modell berücksichtigt werden. Werden die abhängigen Variablen adäquat mit einbezogen, können Verzerrungen vermieden werden.

- **MNAR:** Bei der Schätzung von Parametern ergeben sich Verzerrungen. Da das Fehlen von Werten von den fehlenden Werten selbst abhängig ist, können nur “best guesses” hinsichtlich des **zugrundeliegenden Dropout-Mechanismus** getroffen werden. Ob diese korrekt sind, kann nicht empirisch überprüft werden.

Beim Vorliegen von MNAR spricht man von **"nonignorable missing data"** (Little & Rubin, 2019). Dies bedeutet, dass nicht von der gemeinsamen (Posterior-) **Verteilung der observierten Daten auf die der fehlenden Daten geschlossen werden kann**:

$$P(\mathbf{Y}|\mathbf{Y}_{\text{obs}}, \mathbf{R} = 1) \neq P(\mathbf{Y}|\mathbf{Y}_{\text{obs}}, \mathbf{R} = 0)$$

Das impliziert, dass eine Schätzung (Imputation) auf Basis der vorliegenden Werte nicht ohne weiteres möglich ist. Es müssen Annahmen getroffen werden, die **"über die Daten hinaus gehen"**.

Was bedeutet das für die Analyse von RCT-Daten?

MCAR

- Complete Case-Analysen und andere Ad Hoc-Verfahren führen nicht zu einer systematischen Verzerrung der Ergebnisse (aber durchaus zu einem Verlust statistischer Power/“Effizienz”).
- Die MCAR-Annahme ist typischerweise für RCT-Daten **nicht sehr plausibel**. (Bell et al., 2014; Mallinckrodt et al., 2004)
- Selbst wenn die MCAR-Annahme zutrifft, können Verfahren wie Multiple Imputation genutzt werden, z.B. um Konfidenzintervalle korrekter zu schätzen. (vgl. Pedersen et al., 2017)

Was bedeutet das für die Analyse von RCT-Daten?

MAR

- Complete Case-Analysen führen zu einer Verzerrung der Ergebnisse.
- Verfahren wie Multiple Imputation, Full Information Maximum Likelihood (FIML) oder Mixed-Effect Models (eingeschränkt) können aber **genutzt werden, um die MAR-Annahme abzubilden**.
- Werden diese Modelle korrekt angewandt, vermeidet dies gebiaste Ergebnisse und führt zu einem korrekten Miteinbezug der Unsicherheit durch fehlende Werte (→ passende Konfidenzintervalle).

Was bedeutet das für die Analyse von RCT-Daten?

MNAR

- Auch Verfahren wie Multiple Imputation (denen nur die observierten Daten zugrunde liegen) können zu **Verzerrungen in den Ergebnissen führen**.
- Methoden wie **Pattern-Mixture/Selektionmodelle** oder **referenzbasierte Imputation** können genutzt werden, um die Ergebnisse unter Annahme bestimmter Dropout-Mechanismen zu analysieren. (Little & Rubin, 2019, Kapitel 15; Carpenter et al., 2013; Heckman, 1976)
- Diese Annahmen sind jedoch nicht direkt empirisch nachweisbar.

*“MNAR models are [...] typically **highly dependent on untestable and often implicit assumptions** regarding the distribution of the unobserved measurements given the observed measurements.”*

—Molenberghs et al. (2004), S. 447

Was soll ich für meinen Trial annehmen?

“[W]e recommend that in trials [...], all data should be used in an analysis that makes a plausible assumption about missing data. Usually this will be a MAR assumption.”

—Bell et al. (2014)

“The assumption of ignorability is often sensible in practice, and generally provides a natural starting point.”

—Van Buuren (2018), Kapitel 2.2.6

- Dropout & Fehlende Daten sind bei RCTs **kaum zu vermeiden**.
- Das Fehlen von Daten kann man sich als **“Produkt” eines wahrscheinlichkeitsbasierten Prozesses** vorstellen, der von einem (realen oder angenommenen) **Missing Data-Mechanismus** gesteuert wird.
- Derartige Mechanismen lassen sich in drei “Archetypen” zusammenfassen: **MCAR, MAR** und **MNAR**.
- Wird der zugrundeliegende Missing Data-Mechanismus nicht berücksichtigt, kann dies bei der Analyse zu **Verzerrungen** und **falschen Schlußfolgerungen** führen.



- Ob Daten MAR oder MNAR sind, kann **nicht anhand der Daten selbst bestimmt werden**; beide Annahmen sind immer nur mehr oder weniger plausibel.
- Bei RCTs ist die **Annahme von MAR häufig ein guter Startpunkt**; die Ergebnisse im Fall von nonignorable missing data (MNAR) können dann z.B. durch Sensitivitätsanalysen geprüft werden (Bell et al., 2014).
- Insbesondere **multiple Imputationsverfahren** sind ein gutes Mittel, Verzerrungen der Ergebnisse durch Dropout unter Annahme von MAR vorzubeugen.



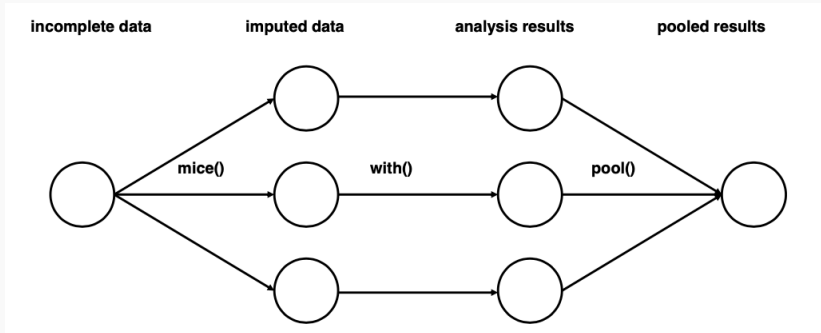
Kurze Anmerkung zu Little's MCAR Test

- Der χ^2 MCAR Test von Little findet immer noch häufig bei der Analyse von RCT Verwendung.
- Die Nullhypothese ist hierbei, dass Missings in den Daten zufällig auftreten (\rightarrow MCAR).
- Der **praktische Nutzen dieses Tests ist häufig begrenzt**: die Fähigkeit des Tests, die Nullhypothese (MCAR) zu verwerfen, hängt von der Größe des Datensets, und damit der statistischen Power ab. Bei großen Datensets können schon kleine Abweichungen zu einem Wert $p < 0.05$ führen.
- Ein signifikanter MCAR-Test sagt nichts darüber, ob die Daten MAR oder MNAR sind!
- **Tipp**: stattdessen MAR als plausiblen "first start" annehmen, MNAR-Auswirkungen ggf. durch Sensitivitätsanalysen explorieren.

siehe Little (1988).

Multiple Imputation

- Multiple Imputation (**MI**, Rubin, 1987) ist eine der flexibelsten und **gängigsten Methoden** zum Umgang mit fehlenden Werten.
- **Ziel von MI:** auf Basis der Verteilung der beobachteten Daten plausible Werte für die fehlenden Werte schätzen (“imputieren”).
- Um die Unsicherheit bei der Schätzung fehlender Werte abzubilden, werden **mehrere (“multiple”) Imputationen** für jeden fehlenden Datenpunkt erzeugt.
- Die so generierten vollständigen Datensätze werden dann **simultan analysiert** (z.B. Berechnung des Stichprobenmittelwerts) und Ergebnisse abschließend **gepoolt**.



aus Van Buuren & Groothuis-Oudshoorn (2011).

- MI kann unter **Annahme von MCAR und MAR** (sowie unter bestimmten Voraussetzungen auch MNAR) verwendet werden.
- Wenn korrekt angewandt, führt MI zur **verzerrungsfreien Schätzung von Populationsparametern** (Regressionsgewichte, Populationsmittelwerte, Korrelationen, etc.) sowie deren Varianz (“asymptotically unbiased”) – trotz des Vorliegens fehlender Werte.
(White et al., 2011)

Es können 2 MI-Ansätze unterschieden werden:

- **Joint Modeling (JM)**: Spezifikation einer multivariaten Verteilung (“joint model”) für die fehlenden Daten, auf deren Basis Imputationen mit Markov Chain Monte Carlo (MCMC) gesampelt werden (Schafer, 1997).
- **Fully Conditional Specification (FCS)**: unvollständige Variablen werden in einem sequentiell-iterativen Prozess imputiert, eine Variable nach der Anderen.

Für FCS ist es im Gegensatz zu JM nicht notwendig, eine multivariate Verteilung der fehlenden Daten zu finden. FCS ist in dieser Hinsicht “atheoretisch” (vgl. White et al., 2011).

Der MICE Algorithmus

Der **'MICE'** (*Multiple Imputation by Chained Equations*) Algorithmus ist eine der am häufigsten verwendeten und am besten erprobten FCS-Ansätze.

“The MICE algorithm possesses a touch of magic.”

—Van Buuren & Groothuis-Oudshoorn (2011)

Die Idee der “Chained Equations”

Annahme: unsere Daten $\mathbf{Y} = Y_1, Y_2, \dots, Y_p$ (teils vorliegend, teils fehlend) sind das “Produkt” von θ , einem (unbekannten) Vektor von Populationsparametern. Der multivariaten Verteilung von θ wollen wir uns annähern, um korrekte Imputationen zu erzeugen.

Der MICE Algorithmus tut dies implizit, indem er iterativ für jede einzelne unvollständige Variable Y aus deren bedingter Verteilung sampelt, in folgender Form (Van Buuren & Groothuis-Oudshoorn, 2011):

$$P(Y_1 | \mathbf{Y}_{\setminus 1}, \theta_1)$$

$$\vdots$$

$$P(Y_p | \mathbf{Y}_{\setminus p}, \theta_p)$$

Dies wird durch sogenanntes **Gibbs-Sampling** umgesetzt. Basierend auf Ausgangswerten wird dabei iterativ für jede Variable j jeweils zuerst θ_j geschätzt, was wiederum direkt genutzt wird, um imputierte Werte Y_j^* zu erzeugen. Diese Werte bilden dann die Basis für das weitere Sampling.

Für eine beliebige Iteration t ergibt sich so:

$$\begin{aligned}\theta_1^{*t} &\sim P(\theta_1 | Y_1^{\text{obs}}, Y_2^{t-1}, \dots, Y_p^{t-1}) \\ Y_1^{*t} &\sim P(Y_1 | Y_1^{\text{obs}}, Y_2^{t-1}, \dots, Y_p^{t-1}, \theta_1^{*t}) \\ &\vdots \\ \theta_p^{*t} &\sim P(\theta_p | Y_p^{\text{obs}}, Y_1^t, \dots, Y_{p-1}^t) \\ Y_p^{*t} &\sim P(Y_p | Y_p^{\text{obs}}, Y_1^t, \dots, Y_{p-1}^t, \theta_p^{*t})\end{aligned}$$

Im Normalfall **konvergiert** dieser Prozess nach einer bestimmten Anzahl Iterationen und erreicht Stationarität (“pendelt sich ein”).

Da in MICE die vorherige Imputation Y_j^{t-1} zur Imputation von Y_j nicht direkt einfließt, wird dies **relativ schnell erreicht** (oft schon nach 5-10 Iterationen).

Dieser iterative Prozess wird **parallel mehrfach durchgeführt**, um die m Imputationssets zu erzeugen.

Die genaue Technik, mit der die Imputationen erzeugt werden, kann bei MICE **flexibel für jede Variable festgelegt werden**:

Technik	Implementierung	Skala
Predictive Mean Matching	pmm	sämtliche
Bayesian Linear Regression	norm	kontinuierlich
Uncodition Mean Imputation	mean	kontinuierlich
Bayesian Logistic Regression	logreg	binär
Bayesian Polytomous Regression	polyreg	faktoriell

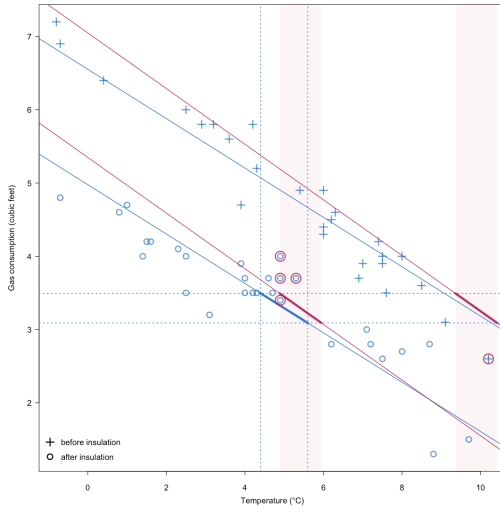
Typische Techniken zur Imputation mit MICE, nach Skalentyp.

Predictive Mean Matching (PMM)

- PMM (Morris et al., 2014) ist eine sogenannte “Hot Deck”-Methode.
- Dabei werden fehlende Werte anhand der eines **“Spenders”** imputiert.
- Der Spender wird dabei zufällig aus einer Anzahl von Kandidaten ausgewählt, die der Person mit fehlenden Werten **statistisch “ähnlich”** sind (d.h. ähnliche vorhergesagte Werte hat):

$$\arg \min_i \left| \hat{y}_i^{(\text{obs})} - \hat{y}_j^{(\text{mis})} \right|$$

- Eine Anzahl von $d = 3-10$ Spenderkandidaten ist dabei häufig passend (Morris et al., 2014).



(Van Buuren, 2018, Kapitel 3.4.1)

Predictive Mean Matching (PMM)

- PMM ist eine breit einsetzbare und typischerweise **robuste Technik** zur Imputation (Kleinke, 2017).
- Da alle Werte von beobachteten Spendern stammen, ist das Imputieren **"unmöglicher" Werte** (z.B. Alter von -2) **ausgeschlossen**.


Multiple Imputation: Praktische Fragen

- **Anzahl der Imputationssets?** $m = 20$ Sets sind häufig ausreichend bei moderater Menge an Missings. Mehr (z.B. $m \approx 50-100$) sind notwendig, um schwer zu schätzende Parameter (z.B. Standardfehler, Varianzkomponenten) genau zu schätzen; oder wenn zahlreiche Missings vorliegen (van Buuren, 2018, Kapitel 2.8).
- **Anzahl der Iterationen?** 20-30 Iterationen häufig bereits ausreichend; dies kann erhöht werden, wenn Konvergenz in Frage steht.
- **Breite des Imputationsmodells?** "So breit wie möglich, so schmal wie nötig". Eine breite Auswahl von (sinnvollen) "Hilfsvariablen" macht die MAR-Annahme plausibler, kann aber zu Softwareproblemen führen (\rightarrow zu hohe Komplexität). Deshalb: "goldene Mitte" finden, Modellkomplexität reduzieren, ohne die Plausibilität des Modells zu sehr zu vermindern.

*“In general, one should try to **simplify the imputation structure without damaging it**; for example, omit variables that seem on exploratory investigation unlikely to be required in ‘reasonable’ analysis models, but avoid omitting variables that are in the analysis model or variables that clearly contribute towards satisfying the MAR assumption.”*

— White et al. (2011)

Praxis-Teil



```
128 }
129
130 }
131
132 .mail{
133     background: url(../img/mailico.png) no-repeat center;
134     display: inline-block;
135     width: 120px;
136     height: 140px;
137     float: left;
138     margin: 2px 7px 0 0;
139 }
140 .phone{
141     background: url(../img/phoneico.png) no-repeat center;
142     display: inline-block;
143     width: 280px;
144     height: 180px;
145     float: left;
146     margin: 2px 7px 0 0;
147 }
```

Der **Influx-** & **Outflux-Koeffizient** kann genutzt werden, um potenziell (un-)relevante Prädiktoren für das Imputationsmodell zu identifizieren.

Bei gleicher Anzahl von Missings zeigen diese Koeffizienten an, welche Variablen besser mit dem Rest der (beobachteten) Daten “verbunden” sind.

Bei RCTs ist der heuristische Werte der Koeffizienten oft begrenzt, da Daten Zeitpunkt-abhängig fehlen (z.B. fehlen für Person i durch Dropout sämtliche Daten zum Postzeitpunkt).

(Van Buuren, 2018, Kapitel 4.1.3)

Der **Influx-Koeffizient** I_j einer Variable j gibt an, wie gut *fehlende* Werte in j durch *beobachtete* Werte in anderen Variablen abgedeckt werden.

Es sei R_{ij} der Responseindikator (beobachtet = 1; fehlend = 0) für Person i bei Variable j . R_{ik} sei der Indikator für selbige Person bei einer beliebigen anderen Variable k .

Für den Influx-Koeffizienten von j ergibt sich so:

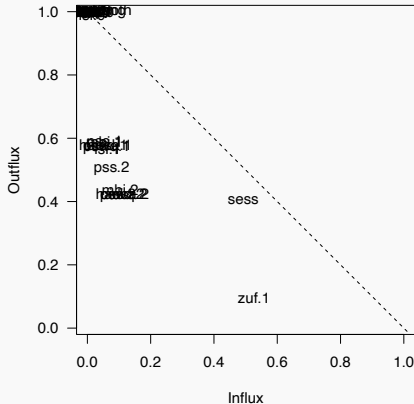
$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - R_{ij}) R_{ik}}{\sum_k^p \sum_i^n R_{ik}}$$

(Van Buuren, 2018, Kapitel 4.1.3)

Der **Outflux-Koeffizient** O_j einer Variable j gibt an, wie gut *beobachtete* Werte in j zur Abdeckung *fehlender* Werte in anderen Variablen genutzt werden können.

Für Variable j ergibt sich so:

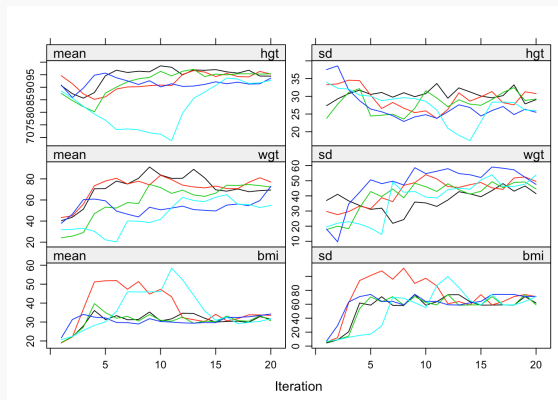
$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n R_{ij}(1 - R_{ik})}{\sum_k^p \sum_i^n 1 - R_{ij}}$$



Werte im **oberen Bereich** sowie nah an der Diagonalen sind eher hilfreich als Prädiktoren. Werte im **unteren linken Bereich** können eher entfernt werden (insofern inhaltlich nicht/wenig relevant).

CAVE: Der Flux-Koeffizient sagt nichts über die tatsächliche Vorhersagekraft der Variable aus!

Beispiel für Nonkonvergenz: Chains vermischen sich kaum & zeitlicher Trend ersichtlich!



(Van Buuren, 2018, Kapitel 6.5.2)

- Typischerweise sollte ein plausibles Imputationsmodell Werte $\dot{\mathbf{Y}}_{\text{mis}}$ erzeugen, die in ihrer Verteilung denen von \mathbf{Y}_{obs} ähneln.
- Unter MAR sind jedoch auch **systematische Unterschiede** zwischen $\dot{\mathbf{Y}}_{\text{mis}}$ und \mathbf{Y}_{obs} plausibel (z.B. Mittelwerts- oder Streuungsunterschiede)!
- Besonders **starke Divergenzen** sowie ausgesprochen unplausible/extreme Ergebnisse können jedoch auf Probleme des Imputationsmodells hinweisen.

Referenzen

- Akl, E. A., Shawwa, K., Kahale, L. A., Agoritsas, T., Brignardello-Petersen, R., Busse, J. W., Carrasco-Labra, A., Ebrahim, S., Johnston, B. C., Neumann, I. others. (2015). Reporting missing participant data in randomised trials: Systematic survey of the methodological literature and a proposed guide. *BMJ Open*, 5(12), e008431.
- Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C.-H. (2014). Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, 14(1), 1–8.
- Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6), 1352–1371.
- Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st ed.). Chapman & Hall/CRC Press.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.
- Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, 42(4), 371–404.

- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- Mallinckrodt, C. H., Watkin, J. G., Molenberghs, G., & Carroll, R. J. (2004). Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 3(3), 161–169.
- Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3), 445–464.
- Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1), 1–13.
- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Theory of statistics* (pp. 697–716). University of California Press.
- Palmer, M. J., Mercieca-Bebber, R., King, M., Calvert, M., Richardson, H., & Brundage, M. (2018). A systematic review and development of a classification framework for factors associated with missing patient-reported outcome data. *Clinical Trials*, 15(1), 95–106.

- Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9, 157.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials*, 11(1), 1–8.
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Bmj*, 338.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(1), 1–67.

- Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Bulletin of the World Health Organization*, 85, 867–872.
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399.
- Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4), 368–376.