Entwicklung und Evaluation eines Emotionswissenstrainings für Kindergartenfachkräfte
Datenimputation

Eric Stemmler

23.04.2020

1 Vorüberlegungen zur Datenimputation

Eine Imputation ermöglichst die Auswertung aller *vorhandenen* Daten in einem Datensatz. Im Normalfall ist es so, dass Software zur statistischen Analyse von Daten den gesamten einzelnen Dateneintrag (d.h. z.B. die Daten ID, Kontrollgruppe/ Experimentalgruppe, Alter, abhängige Variable) verwirft oder igrnoriert, wenn einer der Werte fehlt (z.B. Alter). Eine Imputation des Datenwertes für die Variable Alter würde hierbei entsprechend helfen, die *vorhandenen* Informationen, d.h. Kontroll/ Experimentalgruppe und abhängigen Variable für das statistische Verfahren zu erhalten.

Es gibt im Prinzip zwei Aspekte, die über die Sinnhaftigkeit einer Datenimputation entscheiden. (1) Ist die Imputation für sich genommen überhaupt valide? und (2) Lassen sich durch Auswertung des imputierten Datensatzes statistisch fundierte Schlussfolgerungen ziehen?

1.1 Validität einer Imputation

Es gibt verschienen Verfahren Daten zu imputieren. Wie man imputatiert hängt von den Antworten auf folgende Fragen ab: (a) Warum treten die Lücken im Datensatz auf? (b) Wie treten die Lücken auf? (c) Wo treten die Lücken auf?

Man unterscheidet, ob Lücken in einer (univariat) oder mehrerer (multivariat) Variablen auftauchen.

1.1.1 Missing completely at random

Treten die Lücken rein zufällig auf (*missing completely at random*), dann wäre das gleichbedeutend damit, dass die Teilnehmer jedesmal um die Beantwortung einer Frage gewürfelt hätten.

1.1.2 Missing at random

Treten die Lücken nicht rein zufällig auf, so wie es meistens der Fall ist, spricht man von zufälligen Lücken (*missingness at random*). Dieser Fall lässt sich erkennen, wenn die Auslassungsraten über andere Variablen hinweg unterschiedlich groß sind. Beispiel: Unterschiedlich viele Angaben bei der Frage nach dem Einkommen für ehemalige West- und Ostdeutsche. Eine Implikation dieser Annahme ist, dass die Wahrscheinlich für das Fehlen eines Wertes alleinig von vorhandenen Daten abhängt (Gelman and Hill 2006). Bei ausreichend vorhandenen Daten bedeutet das, dass man die fehlenden Dateneinträge ignorieren kann, solange man die entsprechend konfundierden Variablen in das statistische Modell mit einschließt. Das würde die Verzerrung aufgrund der fehlenden Werte reduzieren. Falls nicht genügend viele Daten vorhandenen sind um noch eine sinnvolle Auswertung durchzuführen ist eine Imputation zu empfehlen.

Tabelle 1: Lücken-Modellierung: Ergebnisse der logistische Regression für alle Variablen mit fehlenden Werten. Gezeigt sind p-Werte < .8

predictor	Std. Error	p	dependent variable
year3-5 Jahre	2.2320109	0.7924073	train
age	0.3632144	0.5947806	
fexpweniger als 11 Jahre	5.0990754	0.1198445	
age	0.2154595	0.0697236	
year11-20 Jahre	2.5929530	0.2201406	emo
year3-5 Jahre	2.0198049	0.5356446	
trainDipl. Sozialpädagogin oder B.A. Sozialpädagogin	2.5987167	0.5256632	

1.1.3 Missing not at random/ Zensur

Demgegenüber gibt es den Fall, dass die Wahrscheinlichkeit einer Lücke von nicht vorhandenen Daten abhängt. Man hat dann die Möglichkeit das Fehlen explizit zu modellieren oder es bleibt nichts anderes übrig, als eine Verzerrung in Kauf zu nehmen. Ein Spezialfall von letzteren ist, wenn die Wahrscheinlichkeit des Fehlens von der betroffenen Variable selbstabhängt, d.h. Zensur.

1.1.4 Fazit

Die Fälle *missing completly at random* (MCAR) und *missing at random* (MAR) sind jene, bei denen eine valide Imputation, d.h. keine Verzerrung wahrscheinlicher und vor allem leichter zu erreichen ist. Die Güte der Imputation hängt dann nur vom Imputationsmodell ab. Handelt es sich um *missing not at random* (MNAR) ist dies schwieriger, jedoch verringert eine Imputation sicher die Verzerrung die man sich durch schlichtes irgnorieren der Datenlücken zulassen würde.

Im Allgemeinen ist es leider unmöglich zu beweisen/ zweifelsfrei zu zeigen, dass MAR vorliegt (Gelman and Hill 2006).

1.2 Betrachtung des vorliegenden Datensatzes

Abbildung 1 zeigt das Lückenmuster im vorliegenden Datensatz. Die Variablen *htrain* und *train*, sowie *educ01*, *educ* und *edu* fehlen demnach immer im Verbund. Diese Variablen lassen sich einfach imputieren, wenn jeweils die Variablen *train* und *edu* imputiert wurde.

In Abbildung 2 ist ein sogenannter *margin plot* dargestellt. Der Plot zeigt boxplots (d.h. Mittelwert, 1. und 2. Quartil und doppelte Standardabweichung) an den Rändern unten und links (engl. *margins*) für jeweils zwei Variablen. In diesem Fall sind es Alter (*age*) und Vorerfahrung mit EmotionTalk (*emo*). Der rote Boxplot zeigt alle Alterswerte für welche der entsprechende Wert für Vorerfahrung mit EmotionTalk fehlt der untere blaue für alle bei denen er vorhanden ist. Da sich roter und blauer Boxplot voneinander unterscheiden, kann man davon ausgehen, dass *emo* nicht rein zufällig fehlt (MAR). Zum Großteil überlappen sich die beiden Boxplots, was dafür spricht, dass die Varianz der Variable *age* für entsprechend fehlende Einträge bei *emo* innerhalb der Varianz der vollständigen Einträge liegt und deshalb eine Imputation auf Basis des Alters zu einer Reduktion der ansonsten vorliegenden Verzerrung führen kann.

Um zu statitisch fundiert zu ermitteln, welche Variablen einen Einfluss auf das fehlen von Werten in einer anderer Variablen haben, kann man das fehlen durch eine logistische Regression modellieren. Hierfür lässt sich jeweils eine neue binomiale Variable definieren mit welche die Werte 1 (vorhanden) oder 0 (fehlt) annehmen kann.

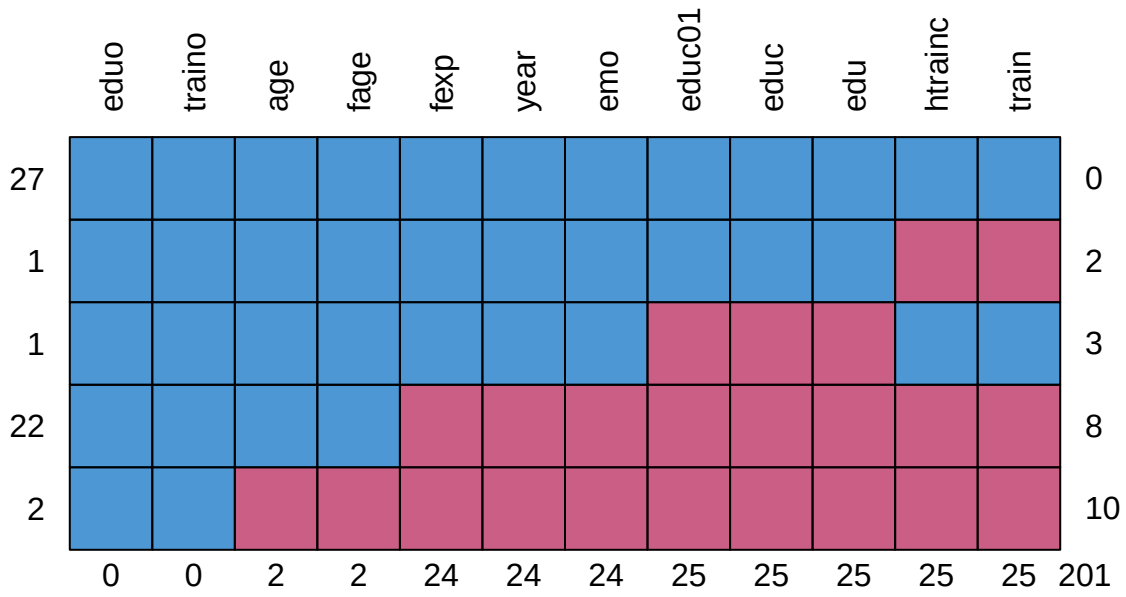

Abbildung 1: Lückenmuster (rot = fehlt, blau = vorhanden) der demographischen Daten (v.l.n.r.): Anzahl der Berufsjahre (kategorial), Anzahl Berufsjahre als Kitafachkraft, Alter, Alter (kategorial), Erfahrung EmotionTalk, Schulabschluss zusammengefasst (kategorial: 0, 1), Schulabschluss zusammengefasst, Schulabschluss, Schulab (sonstige), hoechste Ausbildung (zusammengefasst), Ausbildung, Ausbildung (sonstige). Die Zahlen am linken Rand geben an wieviele Zeilen im Datensatz mit einem Lückenmuster vorkommen. Die Zahlen am rechten Rand geben an, wieviele Variablen in dieser Konstellation Lücken haben. Die Zahlen am unteren Rand geben an, wieiviele Werte in der Variable fehlen.

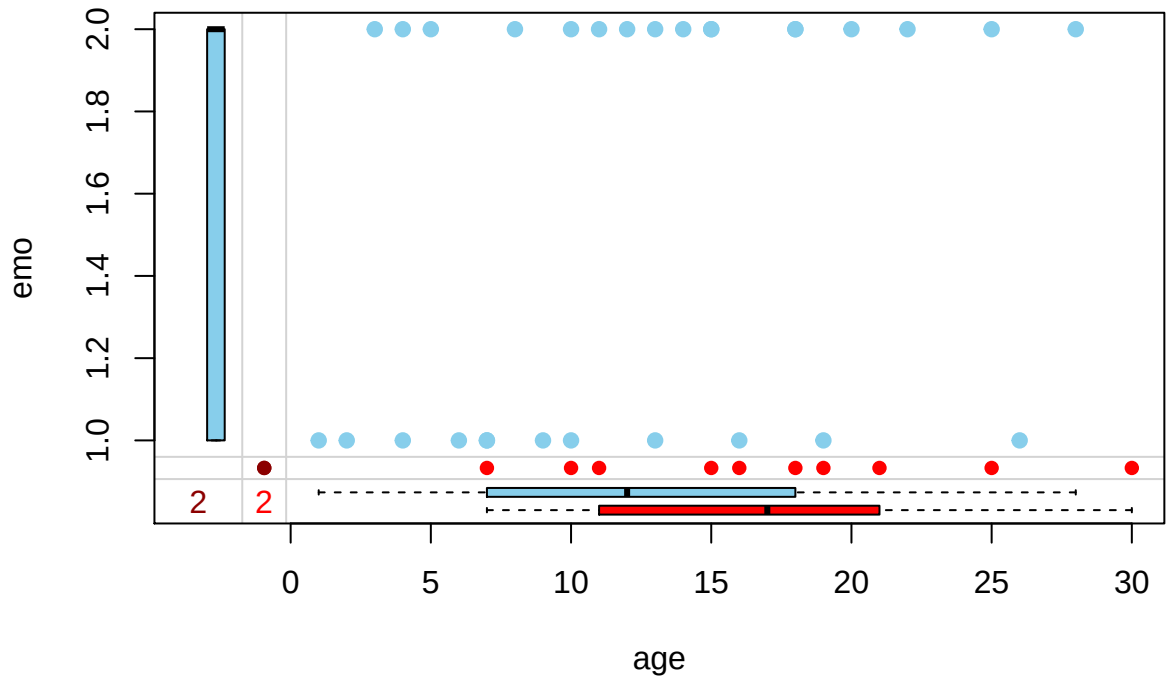

Abbildung 2: Die roten Punkte und Balken sind jene Alters-Werte, für die der entsprechende Erfahrungs-Wert fehlt. Die blauen Werte repräsentieren vorhandene Datenwerte.

In Tabelle 1 sind die Ergebnisse der logistischen Regressionen für die Variablen mit fehlenden Werten gezeigt. Der Übersichtlichkeit halber sind hier nur jene Prädiktoren aufgelistet, für welche der entsprechende p-Wert $< .8$ ergeben hat. Variablen die sich signifikant oder mit geringerem p-Wert modellieren lassen, ist die zusätzliche Streuung im imputierten Datensatz, welche durch die Imputation zwangsläufig hinzukommt geringer als für jene mit größerem p-Wert. Grundsätzlich ist es jedoch legitim, für die Modellierung von Lücken alle möglichen Prädiktoren zu verwenden, solange das Ergebnis plausible ist (Gelman and Hill 2006).

Ein weitere Analyse um zu überprüfen, ob die Imputation valide ist, ist im Falle von FCS (implementiert im R-Packet MICE) zu überprüfen ob über die Iterationen hinweg Imputation konvergieren. Konvergenz von Imputation ist ein Indikator für ungültige Imputationen (Buuren and Groothuis-Oudshoorn 2010).

2 Literatur

Buuren, S van, and Karin Groothuis-Oudshoorn. 2010. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*. University of California, Los Angeles, 1–68.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.