Datenergänzung mit multipler Imputation und Predictive Mean Matching – nichts Neues, nur eine Erinnerung

Axel Albrecht, Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg

Meistens verwendet man bei der Auswertung nur Datensätze mit vollständigen Beobachtungen, also „Zeilen“ mit gültigen Werten für jede „Spalte“. Diese Reduktion des Datensatzes – auch als complete case analysis bekannt – kann aber nicht nur zu lückenhafter sondern sogar fehlerhafter und verzerrter Inferenz führen. Gründe sind hier hoher Bias und verfälschte Korrelationen. Schlimm ist auch der Ansatz, den Mittelwert einer Variablen in die Lücken einzutragen. Denn hierdurch wird die Varianz der Variablen reduziert, was sich in einer Regression reduzierend auf den Standardfehler der Koeffizientenschätzung auswirkt. Und das wiederum führt zur überoptimistischen Signifikanzschätzung. Stattdessen sollten lückenhafte Datensätze mit multipler Imputation ergänzt werden. Dabei werden meist datenbasiert Fehlwerte wiederholt ergänzt. Ein mögliches Verfahren für diese Ergänzung ist das schon lange bekannte predictive mean matching. In dessen Zuge werden zunächst für die complete cases multiple lineare Regressionen zwischen komplett beobachteten und teilweise lückigen Variablen errechnet. Die durch x-fache Wiederholung geschätzten Koeffizienten werden anschließend für die x-fache Schätzung der Zielvariablen für alle Beobachtungen, also solche ohne und mit Beobachtungen, verwendet. Eingesetzt wird schließlich durch Auswahl von Beobachtungen mit möglichst ähnlichen geschätzten Werten für Beobachtungen mit und ohne Fehlwert. Die so erzeugten mehrfachen vervollständigten Datensätze können dann für die eigentliche Analyse, z.B. eine Regression, verwendet werden. Die Unsicherheit des Ergänzungsverfahrens wird hierbei durch wiederholte Regression für die mehrfachen Datensätze berücksichtigt, da die Unterschiede zwischen den Datensätzen die Variabilität widerspiegeln. Anschließend besteht die Möglichkeit, die mehrfachen Regressionsergebnisse durch das Poolen zu einer konsolidierten Endschätzung zusammenzuführen. Hierbei werden nur die mehrfach geschätzten Koeffizienten zusammengeführt, nicht die Schätzwerte. Anhand eines Beispieldatensatzes wird die Anwendung in R im Paket mice demonstriert. Die Auswirkungen der complete case analysis, single mean imputation und multiple imputation mit predictive mean matching auf die Variablenauswahl und die Koeffizientenschätzung wird bildhaft dargestellt.

Primär-Literatur:
- Buuren, S. v. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R." Journal of Statistical Software 45(3): 1-67.