Universität Zürich[UZH]

Analyse von hierarchischen Daten in R mittels Multilevel Analyse

Masterarbeit von

Noah Bosshart

Betreut durch

Prof. Dr. Carolin Strobl

9. Januar 2020

Inhaltsverzeichnis

1 Abstract

2 Einleitung

Hierarchische oder gruppierte Daten treten häufig in den Sozialwissenschaften auf, unter anderem auch in der Psychologie (Snijders & Bosker, 2012). Von hierarchischen Daten wird gesprochen, wenn beispielsweise Daten von Schülern innerhalb verschiedener Klassen oder von Mitarbeitern aus mehreren Teams erhoben werden. Aber auch Daten aus Langzeitstudien werden als gruppiert bezeichnet, da mehrere Messzeitpunkte innerhalb einer Person genested sind.

In diesen Datenstrukturen bestehen gewissen Abhängigkeiten zwischen den einzelnen Messeinheiten.

Das bedeutet beispielsweise, dass Schüler aus der selben Klasse zueinander ähnlicher sind als zu Schülern aus anderen Klassen. Diese Gegebenheit kann auf viele verschiedene Ursachen zurückzuführen sein, wie zum Beispiel die didaktischen Fähigkeiten der Lehrperson oder die Qualität der Lehrmaterialien.

Würden Daten mit diesen Strukturen mit einem linearen Regressionsmodell analysiert werden, könnte das zu fehlerhaften Ergebnissen führen, da diese Form der Anal Diese Arbeit befasst sich nun damit, wie solche Datenstrukturen mittels Multilevel Analyse berücksichtigt werden können, um Fehlschlüsse zu vermeiden.

Bevor wir uns aber mit den theoretischen Grundlagen der Multilevel Analyse befassen können, muss geklärt werden, wie genau solche genesteten Datenstrukturen aufgebaut sind. Dazu werden im folgenden Abschnitt genestete Datenstrukturen genauer beschrieben und es wird beschrieben, wie man Daten mit solchen Strukturen in der Statistiksoftware R simmulieren kann (R Core Team, 2019).

2.1 Was ist eine Multilevel Analyse?

Multilevel Analyse ist eine Methode zur Analyse von Datenstrukturen

2.2 Warum braucht es Multilevel Analyse?

Im folgenden Abschnitt wird die Frage geklärt, warum es in manchen Situationen notwendig ist eine Multilevel Analyse durchzuführen. Dazu werden als erstes hierarchische Datenstrukturen besprochen. Es wird darauf eingegangen, was für Eigenschaften diese hierarchischen Daten mit sich bringen und in welchen Situationen, diese Datenstrukturen vorkommen. Anschliessend werden Probleme besprochen, die man anhand der Theorie und der Annahmen erwarten würde, wenn man wie gewohnt eine multiple lineare Regression auf solche Datenstrukturen anwendet.

2.2.1 Genestete Datenstrukturen

2.2.2 Problematik von linearen Modellen

Was passiert wenn genestete Strukturen ignoriert (aggregiert) werden (Snijders & Bosker, 2012).

Stichproben sollten immer zufällig gezogen werden, dies ist häufig aber nicht der Fall, da es aus Kostengründen einfacher ist bereits vorhandene Gruppen (Cluster) zu ziehen. Beispielsweise sind das Klassen, Teams, Nachbarschaften, etc. Sobald aber solche Cluster gezogen werden, bestehen Abhängigkeiten zwischen den einzelnen Datenpunkte innerhalb der Cluster. Folglich ist die Annahme der Unabhängigkeit der Varianzen von linearen Modellen verletzt.

Bei steigender Intraklassenkorrelation nimmt ebenfalls der α-Fehler zu (Dorman, 2008).

2.3 Wann wird Multilevel Analyse eingesetzt?

3 Theorie zur Multilevel Analyse

Im folgenden Abschnitt wird nun die Theorie der Multilevel Analyse genau besprochen. Zuerst wird auf das zugrundeliegende statistische Modell der Multilevel Analyse eingegangen. Dabei wird auch die erste Notation eines hierarchischen linearen Modells vorgestellt.

Zuerst werden wir uns auf das *Random Intercept* Modell konzentrieren. Dazu werden auch noch weitere wichtige Kennwerte eingeführt, die bei einer Multilevel Analyse zu beachten sind. Am Ende dieses Kapitels werden die *Random Intercept and Slope* Modelle vorgestellt.

Besprechen von ICC und Design Effekt (Vlg. Dazu Guide ML Analysis von J. Peugh 2009)

3.1 Aufbau von Hierarchischen Linearen Modellen

Das zugrundeliegende statistische Modell, das zur Multilevel Analyse verwendet wird ist das Hierarchische lineare Modell (auch HLM). Dieses Modell ist eine Erweiterung der multiplen linearen Regression, das zusätzlich genestete zufällige Koeffizienten beinhaltet (Snijders & Bosker, 2012).

Aufbau erklären. Was ist das richtige Vorgehen um ein Multilevel Modell zu erstellen. Nullmodell bis hin zu Cross-Level Modellen etc. An Guides zu Multi Level Modellen Orienteieren! (Snijders & Bosker, 2012) (Weitere Guides / Tutorials zu MLM finden)

Die meisten Modelle erlauben nicht mehr als 2-3 Random Slopes und konvergieren nicht (Snijders & Bosker, 2012)

3.1.1 Das Null Modell

3.1.2 Das Level-1 Modell

3.1.3 Das Level-2 Modell

3.1.4 Cross-Level Interaktion

3.2 Vergleich von Hierarchischen Linearen Modellen

Modelle welche sich nur in fixen Effekten unterscheiden sollten mit ML und Modelle welche sich in zufälligen Effekten unterscheiden mit REML verglichen werden (Snijders & Bosker, 2012)

Tests für feste Effekte Wald-Test (Snijders & Bosker, 2012) Inkl. Dummy-Test

Deviance Tests ebenfalls verwendbar für feste Effekte. Bei Random Intercept an chi-square verteilung mit df = anz. veränderte variable teile (wichtig fixed effect müssen gleich bleiben, wenn mit REML, sonst ML)

Da Varianzen nicht negativ werden können, wird oft einseitig getestet. Konservativere Möglichkeit druch halbierung des testwertes (SZweiseitiges Testen").

Deviance Tests für Random Slope etwas aufwändiger, df = m1 - m0 = p + 1 (anz. covarianzen p, von denen sich das m0 zu m1 unterscheiden + 1 varianz) Prüfwert wird für df = p und für df = p+1 in einer chi-quadrat verteilung bestimmt. danach mittelwert davon ergibt den eigentlichen prüfwert.

Konfidenzintervall am besten durch profile likelihood (via lme4 Paket). Profile likelihood verhindert, dass Konfidenzintervalle den Wert 0 Unterschreiten, da Varianzen nicht negativ sein können.

Wenn diese Methode nicht vorhanden ist können andere Methoden gewählt werden, die allerdings nicht so genau/reliabel sind.

Proportionale Reduktion der Varianz und Pseude R Squared (Zitation nötig!)

3.3 Kennwerte von Hierarchischen Linearen Modellen

3.4 R Pakete für die Multilevel Analyse

Beschreibung von lme4 und grund warum in dieser Arbeit nur mit diesem Paket gearbeitet wird. (Buch und Studie von D. Bates)

4 Literatur und Forschungsfrage

5 Design der Simulationsstudie

6 Ergebnisse der Simulationsstudie

7 Anwendung und Beschreibung der Shiny App

8 Diskussion

9 Literaturverzeichnis

Dorman, J. P. (2008). The effect of clustering on statistical tests: an illustration using classroom environment data. *Educational Psychology*, *28* (5), 583–595.

R Core Team. (2019). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `https://www.R-project.org/`

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis : an introduction to basic and advanced multilevel modeling* (2. Aufl.). Los Angeles: SAGE.

10 Abbildungsverzeichnis

11 Anhang