Universität Zürich[UZH]

Analyse von hierarchischen Daten in R mittels Multilevel Analyse

Masterarbeit von

Noah Bosshart

Betreut durch

Prof. Dr. Carolin Strobl

14. Januar 2020

Inhaltsverzeichnis

1 Abstract **4**

2 Einleitung **5**

3 Theorie zur Multilevel Analyse **7**

3.1 Methoden zum Umgang mit hierarchischen Daten 7

3.1.1 Aggregation und Disaggregation 7

3.1.2 Intraklassen Korrelation und Design Effect 8

3.2 Hierarchische Linearen Modelle 8

3.2.1 *Random Intercept* Modell 8

3.2.2 *Random Intercept and Slope* Modell 8

3.3 Vergleich von Hierarchischen Linearen Modellen 8

3.4 R Pakete für die Multilevel Analyse 9

4 Literatur und Forschungsfrage **10**

4.1 Stand der Literatur zur HLM 10

4.2 Herleitung der Forschungsfrage 10

5 Design der Simulationsstudie **10**

5.1 Manipulierte Faktoren 10

5.2 Konstante Faktoren 10

5.3 Untersuchte Faktoren 10

6 Ergebnisse der Simulationsstudie **10**

7 Anwendung und Beschreibung der Shiny App **10**

8 Diskussion **10**

9 Literaturverzeichnis **11**

10 Abbildungsverzeichnis **11**

11 Anhang **11**

1 Abstract

2 Einleitung

Hierarchische Daten treten häufig in den Sozialwissenschaften auf, unter anderem auch in der Psychologie (Snijders & Bosker, 2012). Von hierarchischen Strukturen wird gesprochen, wenn beispielsweise Daten von Schulkindern innerhalb verschiedener Schulklassen 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. Hierarchische Daten werden in unterschiedliche Levels unterteilt, wobei Level-1 die niedrigste Stufe der Daten beschreibt. Dabei erhöht sich mit steigender Stufe in der Hierarchie auch das Level (Snijders & Bosker, 2012). Würde man in einer Studie nicht nur Schulkinder in Schulklassen, sondern auch noch die Schulen selbst berücksichtigen, werden Schulen als Level-3 Variable bezeichnet. Der Einfachheit halber beschränken wir uns im Laufe dieser Arbeit auf hierarchische Daten mit zwei Levels. In Tabelle 1 werden einige Beispiele für Level-1 und Level-2 Variablen aufgeführt.

Level-1	Level-2
Schulkinder	Klasse
Studierende	Studienrichtungen
Kinder	Familien
Familien	Nachbarschaften
Mitarbeiter	Teams
Patienten	Therapeuten
Therapeuten	Kliniken
Mehrere Messzeitpunkte	Person

Tabelle 1: Beispiele für Level-1 und Level-2

Dabei ist zu beachten, dass sich das Level der selben Variable je nach Untersuchungsgegenstand ändern kann. Wie man in der Tabelle 1 erkennen kann, sind Therapeuten einmal als Level-1 und einmal als Level-2 Variable aufgeführt. Daher ist es wichtig die Level Bezeichnung als starr betrachtet werden, vielmehr sollte man sich grundsätzlich an der niedrigsten Einheit orientieren. Dieser Einheit wird dann das Level-1 zugeschrieben.

In der Forschung ist es aus Kostengründen oder aus Gründen des Studiendesigns oft nicht möglich, solche gruppierte Datenstrukturen zu vermeiden (Snijders & Bosker, 2012; Woltman et al., 2012). Als eine von vielen Ursachen, die zur Entstehung solcher Datenstrukturen führt, nennen Snijders & Bosker (2012) *multistage sampling*. Unter *multistage sampling* wird verstanden, dass die Forschenden auf in der Population vorhandene Gruppen zugreifen in der Datenerhebung. Beispielsweise ist es Kostengünstiger zufällig 100 Schulkassen und von diesen Schulklassen wieder jeweils 10 Kinder auszuwählen als von 1000 Schulklassen jeweils nur einen Schulkind auszuwählen, da man sonst in 1000 verschiedenen Schulklassen eine Studie durchführen müsste, um die gleiche Stichprobengrösse zu erreichen.

Dieses Auswahlverfahren führt dazu, dass die erhobenen Daten nicht mehr unabhängig voneinander sind. Werden nun also aus jeder Schulklasse 10 Schulkinder für die Studie ausgewählt, ist es sehr wahrscheinlich, dass diese 10 Schulkinder bezüglich ihrer Daten zueinander ähnlicher sind als zu Schulkindern aus anderen Schulklassen. Dieser Zusammenhang kann alleine dadurch entstehen, weil die Schulkinder unterschiedliche Lehrpersonen haben oder in einem anderen Klassenzimmer unterrichtet werden. Das heisst, dass Einflussfaktoren aus unterschiedlichen Levels sich gegenseitig beeinflussen können. Bleiben wir beim Beispiel mit den Schulkindern aus Schulklassen

Wird diese Abhängigkeit der Daten ignoriert und in der Analyse nicht berücksichtigt, kann dies zu einer erhöhten Fehler Typ-1 Rate führen (Dorman, 2008; McNeish, 2014). Das heisst, dass Forschende vermehrt zu Fehlschlüssen bezüglich des Einflusses ihrer Abhängigen Variablen gelangen und irrtümlich annehmen, einen Effekt eines Verfahren gefunden zu haben, obwohl es diesen Effekt gar nicht gibt.

Das Vorhandensein von hierarchischen Daten ist allerdings kein Problem, wenn die Struktur bei der Analyse dieser Daten korrekt berücksichtigt wird.

Was versteht man unter Levels (Stufen) und Einheiten. und wie können diese verschiedenen Levels sich gegenseitig beeinflussen. Beispiele Bringen. Wird diese Beeinflussung nicht berücksichtigt kann es zu Fehlschlüssen kommen.

Wie in dieser Einleutung kurz erläutert wurde, gibt es viele Situationen in denen hierarchische Daten vorhanden sind und wenn diese Strukturen nicht berücksichtigt werden, kann man zu Fehlschlüssen gelangen. Im nächsten Abschnitt wird nun etwas genauer auf die Theorie zur Multilevel Analyse eingegangen. Dabei werden zuerst mögliche Methoden besprochen, wie man hierarchische Datenstrukturen in der Analyse berücksichtigen kann und warum auch diese Methoden nicht immer völlig unproblematisch sein können.

3 Theorie zur Multilevel Analyse

Nach dieser kurzen Einleitung zum Thema wird in diesem Abschnitt nun etwas genauer auf die Theorie der Multilevel Analyse eingegangen. Als erstes werden

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.

3.1 Methoden zum Umgang mit hierarchischen Daten

3.1.1 Aggregation und Disaggregation

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 (Fehler Typ-1) zu Dorman (2008).

3.1.2 Intraklassen Korrelation und Design Effect

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

3.2 Hierarchische Linearen Modelle

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)

3.2.1 *Random Intercept* Modell

3.2.2 *Random Intercept and Slope* Modell

3.3 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!)

3.4 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

4.1 Stand der Literatur zur HLM

4.2 Herleitung der Forschungsfrage

5 Design der Simulationsstudie

5.1 Manipulierte Faktoren

5.2 Konstante Faktoren

5.3 Untersuchte Faktoren

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.

McNeish, D. M. (2014). Analyzing clustered data with ols regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, *40* (1), 11–16.

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

Woltman, H., Feldstain, A., MacKay, J. C. & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, *8* (1), 52–69.

10 Abbildungsverzeichnis

11 Anhang