



**Universität
Zürich^{UZH}**

Schätzgenauigkeit von Standardfehlern und deren
Einfluss auf die Power bei der Analyse von
hierarchischen Daten

Ein Vergleich zwischen linearen und hierarchischen linearen Modellen

Masterarbeit von

Noah Bosshart

Mat-Nr.: 13-747-141

Betreut durch

Prof. Dr. Carolin Strobl

15. April 2020

Inhaltsverzeichnis

Abbildungsverzeichnis	III
Tabellenverzeichnis	IV
1 Einleitung	1
2 Konzept und Anwendung von Multilevel Analyse	4
2.1 Beispiel zur Theorie	4
2.2 Intraklassen Korrelation	5
2.3 Lineare Modelle	8
2.3.1 Aggregation	9
2.3.2 Disaggregation	11
2.4 Hierarchische Linearen Modelle	13
2.4.1 <i>Random Intercept</i> Modell	14
2.4.2 <i>Random Intercept and Slope</i> Modell	18
2.4.3 <i>Intercept</i> und <i>Slope</i> Variabilität	21
2.5 Anwendung von Multilevel Analyse in R	23
2.5.1 Informationen und Syntax von <code>lme4</code>	23
2.5.2 Interpretation eines Outputs	24
2.5.3 Berechnung der Intraklassen Korrelation in R	27
2.5.4 Aufbau und Vergleich von Modellen	29
2.5.5 Auswertung von Modellen	33
3 Simulationsstudie zur Multilevel Analyse	37
3.1 Herleitung der Forschungsfrage	37
3.2 Studiendesign	39
3.3 Simulationsdesign	40
3.4 Studie 1: Genauigkeit von Schätzparametern	42
3.4.1 Ergebnisse Studie 1	43

3.4.2	Diskussion Studie 1	48
3.5	Studie 2: Statistische Power von HLM	50
3.5.1	Ergebnisse Studie 2	51
3.5.2	Diskussion Studie 2	54
3.6	Abschliessende Diskussion und Shiny App	55
4	Literaturverzeichnis	58
5	Anhang	60
A	Shiny App Screenshots	60
B	R Code	60

Abbildungsverzeichnis

1	Zusammenhang zwischen der durchschnittlich gelösten Anzahl an Übungsaufgaben und der durchschnittlich erreichten Punktzahl pro Klasse	10
2	Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und erreichte Punktzahl mittels Disaggregation und Anwendung dieses Zusammenhangs auf jede der fünf Klassen	12
3	Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl mittels HLM im Vergleich zu einem LM (Disaggregation). Links: Geraden von LM und HLM im gesamten Datensatz. Rechts: Geraden von LM und HLM auf jede Klasse aufgeteilt.	16
4	Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl unter Berücksichtigung der Klassenzugehörigkeit und deren Interaktion mit der Anzahl gelöster Übungsaufgaben	19
5	Darstellung einer negativen und nicht vorhandenen Korrelationen zwischen Achsenabschnitt und Steigung	20
6	Residuenplot des <i>Random Intercept and Slope</i> Modells	21
7	Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im ersten Simulationsdesign.	46
8	Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im zweiten Simulationsdesign.	47
9	Statistische Power von ML und HLM in den verschiedenen IKK Bedingungen bei einer Intervention auf Level-1.	52
10	Statistische Power von ML und HLM in den verschiedenen IKK Bedingungen bei einer Intervention auf Level-2.	53

Tabellenverzeichnis

1	Beispiele für Level-1 und Level-2 Einheiten	1
2	Ausschnitt des simulierten Datensatzes	5
3	Mittlere Anzahl gelöster Übungsaufgaben und erreichte Punktzahl	9
4	Mögliche Syntax für <code>lmer()</code> nach Bates et al. (2015)	25
5	SE Genauigkeit beider Regressionskoeffizienten in beiden Simulationsdesigns und für jede Analysemethode in allen IKK Bedingungen.	45
6	Statistische Power in beiden Simulationsdesigns für jede Analysemethode in allen IKK Bedingungen.	51

Abstract

1 Einleitung

Hierarchische 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 Schulkindern innerhalb verschiedener Schulklassen oder von Mitarbeitenden aus mehreren Teams erhoben werden. Aber auch Daten aus Langzeitstudien werden als gruppiert bezeichnet, da mehrere Messzeitpunkte innerhalb einer Person gruppiert sind. Hierarchische Daten werden in Levels unterteilt, wobei Daten aus der niedrigsten Stufe als Level-1 Einheiten bezeichnet werden (Snijders & Bosker, 2012). Ein Beispiel für Level-1 Einheiten sind Schulkinder. Diese Schulkinder befinden sich wiederum in Klassen, die in der Hierarchiestufe höher sind und folglich als Level-2 Einheiten bezeichnet werden. Würde man nun in einer Studie nicht nur Schulkinder in Schulklassen, sondern auch die Schulen selbst berücksichtigen, würden die Schulen als Level-3 Einheit bezeichnet werden. Die Anzahl der Levels könnte man theoretisch beliebig hoch wählen, solange es das Studiendesign erlaubt und es aus der Perspektive der Forschungsfrage sinnvoll ist. Der Einfachheit halber beschränken wir uns im Laufe dieser Arbeit aber auf hierarchische Daten mit zwei Levels. In Tabelle 1 werden einige Beispiele für Level-1 und Level-2 Einheiten aufgeführt. Dabei ist zu beachten, dass sich das Level derselben Einheit je nach Untersuchungsgegenstand ändern kann. Wie man in der Tabelle 1 erkennen kann, sind Familien einmal als Level-1 und einmal als Level-2 Einheit aufgeführt. Daher ist es wichtig die Level

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

Level-1	Level-2
Schulkinder	Klasse
Studierende	Studienrichtungen
Kinder	Familien
Familien	Nachbarschaften
Mitarbeitende	Teams
Teams	Unternehmen
Behandelte Personen	Therapierende
Therapierende	Kliniken
Mehrere Messzeitpunkte	Person

Bezeichnung nicht als starr zu betrachten. Vielmehr sollte man sich grundsätzlich an den niedrigsten Einheiten im Datensatz orientieren. Diesen Einheiten 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 und Bosker (2012) *Multistage Sampling*. Unter *Multistage Sampling* versteht man, dass die Forschenden in der Datenerhebung auf in der Population vorhandene Gruppen zugreifen. So wäre es beispielsweise kostengünstiger eine Intervention mit 10 Schülkindern aus jeweils 100 Schulklassen durchzuführen, als mit einem Schülkind aus jeweils 1000 Schulklassen. Dadurch erhalten die Forschenden eine vermeintlich gleichgrosse Stichprobe von 1000 Beobachtungen, müssen die Intervention aber nur an 100 Schulklassen und nicht an 1000 durchführen. Dieses Auswahlverfahren führt aber dazu, dass die erhobenen Daten nicht mehr voneinander unabhängig sind. Werden nun aus jeder Schulkasse 10 Schülkinder für eine Studie ausgewählt, ist es sehr wahrscheinlich, dass Schülkinder aus derselben Klasse zueinander ähnlichere Leistungen erzielen werden. Dieser Zusammenhang kann auf unterschiedliche Ursache zurückzuführen sein. Beispielsweise könnte die didaktischen Fähigkeiten der Lehrpersonen oder die Lichtverhältnisse im Klassenzimmer einen Einfluss auf die Leistungen der Kinder aus derselben Klasse haben.

Nach Snijders und Bosker (2012) gibt es unterschiedliche Formen, wie diese Einheiten zueinander in Beziehung stehen können. Ein Beispiel für einen Zusammenhang auf Level-1 wäre, dass die Lernmotivation eines Schülkindes sich auf seine schulische Leistung auswirkt. Aber auch Level-2 Einheiten können sich gegenseitig beeinflussen. Das Klima der Schulkasse könnte sich beispielsweise auf das Stressempfinden der Lehrperson auswirken. Hier wird von einem Zusammenhang innerhalb des Levels gesprochen, weil die unabhängige Variable (z.B. Lernmotivation, Klima der Schulkasse) auf dem gleichen Level wie die abhängige Variable (z.B. schulische Leistung, Stressempfinden) ist. Häufig ist es allerdings der Fall, dass es levelübergreifende Zusammenhänge zwischen den Einheiten gibt. So können beispielsweise die didaktischen Fähigkeiten einer Lehrperson (Level-2) und

die Lernmotivation der Schulkinder (Level-1) die individuelle Leistung (Level-1) beeinflussen. Dieser Zusammenhang muss nicht zwingend direkt sein. Es kann auch vorkommen, dass die didaktischen Fähigkeiten den Zusammenhang zwischen Lernmotivation und individueller Leistung moderiert. In diesem Fall wird gemäss Snijders und Bosker (2012) von einer *Cross-Level* Interaktion gesprochen.

Werden diese Abhängigkeiten in der Analyse nicht berücksichtigt, kann dies unter anderem zu einer erhöhten Fehler Typ 1 Rate führen (Dorman, 2008; McNeish, 2014). Das heisst, dass Forschende vermehrt zu Fehlschlüssen bezüglich des Einflusses ihrer Unabhängigen Variablen gelangen und irrtümlich annehmen, einen Effekt eines Verfahrens gefunden zu haben, obwohl es diesen Effekt gar nicht gibt. Missachten dieser Abhängigkeiten kann aber nicht nur zu einer erhöhten Fehler Typ 1 Rate führen, sondern auch die Fehler Typ 2 Rate erhöhen (Moerbeek et al., 2003). Eine erhöhte Fehler Typ 2 Rate resultiert in einer niedrigeren Power einen Effekt zu entdecken, wenn dieser auch effektiv vorhanden ist. Das Vorhandensein von hierarchischen Daten ist allerdings kein unlösbares Problem. Mit Analyseansätzen, die diese hierarchische Struktur der Daten berücksichtigen, lassen sich solche erhöhten Fehler Typ 1 und 2 Raten vermeiden. Einer dieser Ansätze ist die Multilevel Analyse, die im Fokus dieser Arbeit steht.

Diese Arbeit ist in zwei Teile unterteilt. Im ersten Teil wird das Konzept und die Theorie der Multilevel Analyse behandelt. Dabei wird kurz auf die verschiedenen Methoden eingegangen, wie man Daten auf ihre hierarchische Struktur überprüfen kann. Anschliessend wird das zugrundeliegende statistische Modell der Multilevel Analyse vorgestellt und wie genau solche Modelle aufgebaut sind. Darauf folgend wird die Anwendung dieser Methoden in der Statistikumgebung R besprochen (R Core Team, 2019). Im zweiten Abschnitt dieser Arbeit wird eine Simulationsstudie durchgeführt, deren Ziel es ist, bereits vorhandene Ergebnisse in der Literatur zu replizieren und zu illustrieren, dass eine Multilevel Analyse bei hierarchischen Daten verlässliche Resultate liefert. Begleitend zu dieser Studie wird eine Shiny App programmiert (Chang et al., 2019), die zum einen das Konzept der Multilevel Analyse visualisiert und die Ergebnisse der hier durchgeführten Simulationsstudie interaktiv abbildet.

2 Konzept und Anwendung von Multilevel Analyse

Wie in der Einleitung erläutert wurde, gibt es viele Situationen in denen hierarchische Daten vorhanden sind und man zu Fehlschlüssen gelangen kann, wenn man diese Strukturen nicht berücksichtigt. In diesem Abschnitt wird nun etwas genauer auf das Konzept und die dahintersteckende Theorie der Multilevel Analyse eingegangen. Dazu wird zuerst ein simulierter Beispieldatensatz vorgestellt, anhand dessen die besprochenen Modelle erklärt werden. Als erstes wird auf die Probleme eingegangen, die durch die Verwendung von einfachen linearen Modellen (LM) entstehen. Anschliessend wird das hierarchische lineare Modell (HLM) als das zugrundeliegende statistische Modell der Multilevel Analyse eingeführt. Das HLM gilt als eine Erweiterung der einfachen LM (Snijders & Bosker, 2012). Dabei werden bei HLM in *Random Intercept* und *Random Intercept and Slope* Modelle unterschieden. Es werden beide Modellformen besprochen und dabei wird erläutert wie die beiden Faktoren Achsenabschnitt (engl. *Intercept*) und Steigung (engl. *Slope*) zusammenhängen. Nachdem die verschiedenen Formen von HLM besprochen worden sind, wird in einem etwas praktischeren Teil die Anwendung von Multilevel Analyse in R anhand von Beispielen etwas näher gebracht.

2.1 Beispiel zur Theorie

In den folgenden Abschnitten wird die Theorie zur Analyse von hierarchischen Daten anhand eines Beispieldatensatzes erläutert. Bei dem Beispiel handelt es sich um insgesamt 150 Schulkindern aus 5 Schulklassen, die eine Mathematikprüfung geschrieben haben. Neben der erreichten Punktzahl wurde für jedes Kind zufällig ein Geschlecht, die Anzahl an gelöster Übungen, einen Wert für sozioökonomische Status und einen Intelligenzquotienten simuliert. Auf Stufe der Klasse wurden ausserdem noch die Anzahl Fenster im Klassenzimmer simuliert. Da dieser Datensatz selbst generiert wurde und aus keiner Studie entstammt, sollten Ergebnisse, die aus diesen Berechnungen entstehen nicht weiter interpretiert werden. In Tabelle 2 sind zur Veranschaulichung dieser Daten eine Auswahl von 10 Schulkindern aufgeführt.

Tabelle 2: Ausschnitt des simulierten Datensatzes

Schulkind Nr.	Klasse	Übungen	Punktzahl	Geschlecht	Anz. Fenster	SES	IQ
101	4	17	21	m	3	16	104
75	3	7	29	m	8	27	112
126	5	23	26	w	4	14	110
14	1	10	29	m	4	21	84
137	5	16	18	w	4	17	109
100	4	7	16	w	3	20	98
78	3	28	44	w	8	23	105
121	5	25	33	w	4	21	99
16	1	7	24	w	4	30	77
116	4	14	29	m	3	19	90

Betrachtet man die Variablen des Datensatzes, kann man erkennen, dass es sich um einen hierarchischen Datensatz mit zwei Levels handelt. Zu den Level-1 Variablen gehören alle Variablen die sich auf der Stufe der tiefsten Einheit (Schulkinder) befinden. Dazu zählen die Anzahl gelösten Übungen, die erreichte Punktzahl, das Geschlecht, der sozioökonomische Status und der IQ. Die beiden anderen Variablen Klasse und die Anzahl Fenster im Klassenzimmer gehören zur Level-2 Ebene. Durch ein kurzes Betrachten der Daten, kann man also relativ schnell herausfinden, ob es sich um einen hierarchischen Datensatz handelt. Allerdings gibt uns diese Betrachtung der Daten keine Auskunft darüber, ob und wie stark die Gruppenzugehörigkeit (Klasse) einen Einfluss auf die erreichte Punktzahl in der Mathematikprüfung hat. Die Intraklassen Korrelation kann diesbezüglich mehr Klarheit schaffen und wird im nächsten Abschnitt vorgestellt.

2.2 Intraklassen Korrelation

Der Einfluss einer hierarchischen Struktur auf eine abhängige Variable kann durch die Intraklassen Korrelation (IKK) beschrieben werden. Die IKK beschreibt den Grad der Ähnlichkeit von Level-1 Einheiten innerhalb einer Level-2 Einheit und kann als Verhältnis der Varianz zwischen den Level-2 Einheiten und der Gesamtvarianz beschrieben werden (Field et al., 2013; Snijders & Bosker, 2012; Twisk, 2006). Diese Varianzen ergeben sich

gemäss Snijders und Bosker (2012) aus dem *Random Effects ANOVA* Modell, das bei der Modellierung von Multilevel Modellen oft auch als leeres Modell bezeichnet wird:

$$Y_{ij} = \mu + U_j + R_{ij} \quad (1)$$

Die abhängige Variable Y_{ij} beschreibt in unserem Beispiel die erreichte Punktzahl des Schulkindes i aus der Klasse j . Der Gesamtmittelwert aller Schulkinder wird mit μ bezeichnet, wobei U_j die zufällige Abweichung einer Klasse j und R_{ij} die zufällige Abweichung eines Schulkindes i der Klasse j von diesem Gesamtmittelwert beschreiben. Dabei ist zu beachten, dass der Erwartungswert beider Zufallsvariablen U_j und R_{ij} als 0 angenommen wird. Die Varianz von U_j wird als *between-group variance* τ_0^2 und von R_{ij} als *within-group variance* σ^2 bezeichnet.

Die IKK beschreibt, wie viel Varianz in der abhängigen Variabel durch die Gruppenzugehörigkeit erklärt wird. Bezogen auf unser Beispiel gibt die IKK an, wie stark sich Schulkinder aus derselben Klasse bezüglich ihrer erreichten Punktzahl ähneln. Ist die Korrelation zwischen den Schulkindern hoch, kann man davon ausgehen, dass die Klasse als Level-2 Einheit einen bedeutenden Anteil an der Gesamtvarianz erklärt. Ist die Korrelation niedrig hat die Klassenzugehörigkeit eher einen kleineren Einfluss auf die Prüfungsleistung. Dieser Zusammenhang wird etwas klarer, wenn man ihn anhand der Formel zur Berechnung des IKK Koeffizienten ρ_I erklärt:

$$\rho_I = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \quad (2)$$

In unserem Beispiel ist die Varianz der erreichten Punktzahl zwischen den verschiedenen Klassen die *between-group variance* τ_0^2 . Die Gesamtvarianz des Datensatzes setzt sich aus der *between-group variance* und der *within-group variance* zusammen. Dabei ist die Varianz innerhalb der Klassen, die bereits erwähnte *within-group variance* und wird mit σ^2 bezeichnet. Besteht nun innerhalb der Klassen eine kleine Varianz zwischen den Ergebnissen der Schulkinder, ergibt sich eine grössere Intraklassen Korrelation. Steigt die Varianz innerhalb der Klassen an, wird der Nenner der Formel grösser. Mit einem wachsenden

Nenner, verringert sich schlussendlich die IKK.

Um nun zu überprüfen, ob in unserem Datensatz überhaupt abhängige hierarchische Strukturen vorhanden sind, können wir die IKK für unser Datensatz berechnen. Da die Populationswerte oft nicht bekannt sind, gibt es viele statistische Verfahren, um Schätzer für die nötigen Varianzen zu berechnen. Da diese Verfahren den Umfang dieser Arbeit sprengen würden und es viele Statistikprogramme gibt, die diese Berechnungen mit präziseren Methoden durchführen können, werden in dieser Arbeit nur die computerbasierten Verfahren behandelt. Die restlichen Verfahren können aber in der gängigen Literatur zur Multilevel Analyse nachgeschlagen werden (z.B. Snijders & Bosker, 2012). Mit Hilfe des Statistikprogramms R wurden nun alle nötigen Varianzen unseres generierten Datensatzes geschätzt und in die Formel (2) eingesetzt¹:

$$\rho_I = \frac{9.57}{9.57 + 44} = 0.18 \quad (3)$$

Die daraus resultierende IKK von $\rho_I = 0.18$ weist darauf hin, dass 18% der Varianz in der erreichten Punktzahl in der Mathematikprüfung durch die Klassenzugehörigkeit erklärt wird. Eine IKK von $\rho_I > 0$ bedeutet aber noch nicht, dass die Varianz, die durch die Klassenzugehörigkeit entsteht, auch zu signifikanten Unterschieden zwischen den Klassen führt. Unter der Annahme, dass die zufällige Abweichungen der Schulkinder R_{ij} normalverteilt sind, kann gemäss Snijders und Bosker (2012) eine Varianzanalyse durchgeführt werden, um zu untersuchen, ob Gruppenunterschiede vorhanden sind. In unserem Fall führte die Varianzanalyse zu einem hoch signifikantem Ergebnis ($p < .001$) und es bestehen folglich Unterschiede zwischen den Klassen. Wir wissen nun nicht nur, wie viel Varianz durch die Klassen erklärt wird sondern auch, dass diese sich signifikant unterscheiden.

Mit Hilfe der IKK kann man also den Einfluss der Gruppenzugehörigkeit quantifizieren und man erhält einen Einblick darin, wie stark dieser Einfluss ist. Im folgenden Abschnitt wird nun besprochen, was für Probleme entstehen wenn lineare Modelle zur Analyse verwendet werden und man den Einfluss von Gruppen missachtet.

¹Die Berechnung dieser Schätzer in R werden in Abschnitt 2.5 erläutert.

2.3 Lineare Modelle

Bevor wir uns mit den hierarchischen linearen Modellen beschäftigen, werden die Grundlagen der linearen Modellen (LM) kurz erläutert und aufgezeigt zu welchen Problemen es führen kann, wenn die hierarchische Datenstruktur ignoriert wird. Gemäss Gelman und Hill (2007) ist die lineare Regression eine Methode, die Veränderungen von Durchschnittswerten einer abhängigen Variablen durch eine lineare Funktion von Prädiktoren beschreibt. Einfacher formuliert, versucht die lineare Regression durch die Kombination von unabhängigen Variablen die mittlere Ausprägung einer abhängigen Variable zu beschreiben. Ein lineares Regressionsmodell kann wie folgt formuliert werden:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{ik} x_{ik} + \epsilon_{ij}, \text{ für } i = 1, \dots, n \text{ und } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (4)$$

Dabei ist y_i die abhängige Variable von der Person i . In unserem Beispiel wäre das die erreichte Punktzahl des Schulkindes i . β_0 beschreibt den Achsenabschnittes (*Intercept*) und ist die durchschnittlich erreichte Punktzahl in der Mathematikprüfung, wenn alle weiteren Prädiktoren 0 sind. Die Steigung (*Slope*) des linearen Regressionsmodells wird durch die weiteren Regressionskoeffizienten β_1 bis β_k beschrieben. Für jede unabhängige Variable x_{i1} bis x_{ik} geben diese Regressionskoeffizienten an, wie stark y_i des i -ten Schulkindes bei einer Zunahme des entsprechenden Prädiktors um eine Einheit ansteigt. Möchten wir in unserem Beispiel die erreichte Punktzahl durch die Anzahl gelöster Übungsaufgaben beschreiben, wäre x_{i1} die Anzahl gelöster Übungsaufgaben des i -ten Schulkindes und der dazugehörige Regressionskoeffizient β_1 gibt die Zunahme der Punktzahl in der Mathematikprüfung an, wenn die Anzahl der gelösten Übungsaufgaben um 1 steigt. Der letzte Parameter des Regressionsmodells ist ϵ_{ij} und wird als zufälliger Fehler oder Residuum bezeichnet. Das Residuum ist die normal verteilte zufällige Abweichung jedes i -ten Schulkindes, mit einem Erwartungswert von 0 und Varianz von σ^2 . Das bedeutet, dass es zwischen den Kindern zufällige Unterschiede in ihrer Prüfungsleistung gibt, die nicht durch das Regressionsmodell erklärt werden. Diese Unterschiede sind im Mittel aber 0.

Möchte man mit einem linearen Regressionsmodell die Daten unseres Beispiels untersuchen gibt es zwei Möglichkeiten. Die erste Möglichkeit ist die Aggregation, die häufig in den Sozialwissenschaften angewandt wird (Snijders & Bosker, 2012). Bei dieser Methode werden Mittelwerte für jede Klasse berechnet und anhand dieser wird dann ein lineares Modell erstellt. Die zweite Möglichkeit ist die Disaggregation, bei der die Klassenstruktur aufgelöst wird und alle 150 Schulkinder als unabhängige Werte in die Analyse einfließen.

2.3.1 Aggregation

Wie bereits erwähnt, werden bei der Aggregation für jede Level-2 Einheit Mittelwerte berechnet, die später in das Regressionsmodell einfließen. Ausgehend von unserem Beispiel könnte man sich nun für den Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl in der Mathematikprüfung interessieren. In Tabelle 3 sind die relevanten Mittelwerte für jede der fünf Schulklassen aufgelistet.

Wird nun anhand dieser aggregierter Werte überprüft, wie genau die erreichte Punktzahl eines Schulkindes mit der Anzahl an gelösten Übungsaufgaben zusammenhängt, entstehen mehrere Probleme, die zu Verzerrungen und Fehlschlüssen führen können. Zum einen verändert sich die Forschungsfrage, da sich durch die Aggregation der Daten der Fokus von der Level-1 Ebene auf die Level-2 Ebene verschiebt (Snijders & Bosker, 2012; Woltman et al., 2012). Die abhängige Variable ist nun nicht mehr die erreichte Punktzahl jedes einzelnen Schulkindes, sondern die durchschnittlich erreichte Punktzahl einer Schulklasse. Ein weiteres Problem ist der Verlust von Variabilität, die durch individuelle

Tabelle 3: Mittlere Anzahl gelöster Übungsaufgaben und erreichte Punktzahl

Klasse	Übungen	Punktzahl
1	13.1	21.5
2	12.8	29.3
3	13.5	30.7
4	15.7	25.6
5	17.5	24.7

Unterschiede zwischen den Schulkindern entsteht. Dieser Verlust an Variabilität beträgt nach Raudenbush und Bryk 80-90% und kann zu Fehlschlüssen über den Zusammenhang der Variablen führen (Raudenbush & Bryk, 2002).

Betrachtet man die Regressionsgerade in Abbildung 1, sieht man, dass ein höhere Anzahl an gelöster Übungsaufgaben mit einer tieferen durchschnittlich erreichten Punktzahl zusammenhängt. Folglich könnte man daraus schliessen, dass dies auch auf Ebene der Schüler zutrifft und eine Erhöhte Anzahl an gelösten Übungsaufgaben mit einer tieferen Punktzahl in der Prüfung einhergeht. Diese Schlussfolgerung ist allerdings unzulässig, da man nicht von einer Korrelation zweier Level-2 Variablen auf den Zusammenhang von Level-1 Variablen schliessen darf (Snijders & Bosker, 2012). Diese fehlerhafte Schlussfolgerung wird auch als ökologischer Fehlschluss bezeichnet (Robinson, 2009).

Die Analyse mittels Aggregation führt folglich nicht zu einem zufriedenstellenden Ergebnis und ist aufgrund der besprochenen Einschränkungen nicht geeignet, um Zusammenhänge auf Level-1 zu untersuchen.

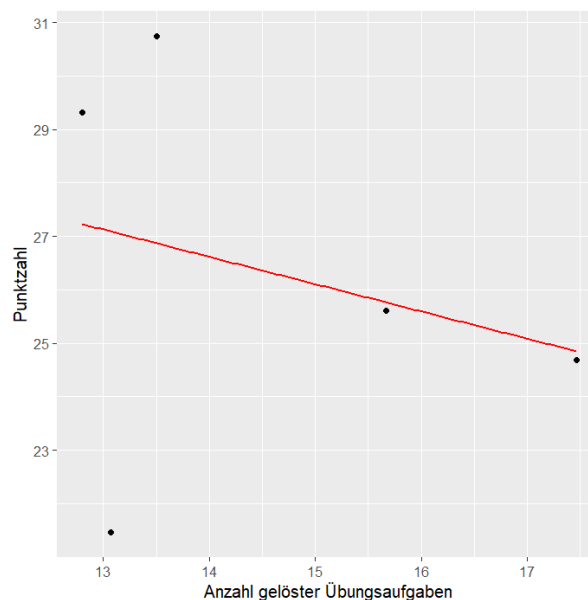


Abbildung 1: Zusammenhang zwischen der durchschnittlich gelösten Anzahl an Übungsaufgaben und der durchschnittlich erreichten Punktzahl pro Klasse

2.3.2 Disaggregation

Die zweite Möglichkeit um hierarchische Daten mit einem linearen Regressionsmodell zu untersuchen ist die Disaggregation. Wie bereits angedeutet werden bei der Disaggregation alle Level-2 Variablen auf Level-1 Einheiten verteilt.

In unserem Beispiel werden also alle Schulkinder als von einander unabhängige Datenpunkte in die Analyse miteinbezogen. Dazu werden jedem Schulkind aus derselben Klasse die gleichen Werte der Level-2 Variablen zugeschrieben. In Tabelle 2 aus Abschnitt 2.1 kann man dieses Vorgehen bei der Level-2 Variable *Fenster* beobachten. Durch diese Disaggregation von Level-2 Variablen auf Level-1 Einheiten werden Datensätze künstlich vergrößert und mögliche Variabilität, die zwischen den Level-2 Variablen besteht, wird ignoriert (Snijders & Bosker, 2012; Woltman et al., 2012). Folglich wird die geteilte Varianz zwischen Level-1 Einheiten einer Gruppe nicht berücksichtigt und die Annahme, dass Fehler voneinander unabhängig sind, ist verletzt. Das führt dazu, dass die Effekte von Level-1 und Level-2 Variablen auf die abhängige Variable nicht voneinander getrennt werden können (Woltman et al., 2012). In unserem Beispiel würde das bedeuten, dass man den Einfluss der Anzahl gelöster Übungsaufgaben nicht vom Einfluss der Klasse trennen kann. Ein grundlegendes Problem der Disaggregation entsteht durch die Annahme von linearen Regressionsmodellen, dass einzelne Beobachtungen voneinander unabhängig sind (Woltman et al., 2012). Da bei hierarchischen Daten einzelne Beobachtungen voneinander Abhängig sind ist diese Annahme verletzt und Analysen basierend auf linearen Regressionsmodellen (z.B. Disaggregation) führen folglich zu ungenauen Ergebnissen (Gelman & Hill, 2007; Snijders & Bosker, 2012; Woltman et al., 2012).

Auf der linken Seite der Abbildung 2 befindet sich die Regressionsgerade, die durch ein lineares Regressionsmodell entsteht, wenn man mit einem disaggregierten Datensatz arbeitet. Anhand dieser Regressionsgerade besteht ein positiver Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl in der Mathematikprüfung, so dass die erreichte Punktzahl mit steigender Anzahl an gelöster Übungsaufgaben zunimmt. Wie vorhin bereits erwähnt, wird in dieser Analyse aber nicht berücksichtigt, dass



Abbildung 2: Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und erreichte Punktzahl mittels Disaggregation und Anwendung dieses Zusammenhangs auf jede der fünf Klassen

die Schulklasse selbst einen Effekt auf die erreichte Punktzahl haben kann. Dieser Effekt wird klar, wenn man die rechte Seite der Abbildung 2 betrachtet. Für jede der fünf Klassen wurde dieselbe Regressionsgerade, die aus dem disaggregierten Datensatz entsteht, über die Daten gelegt. Man kann relativ einfach erkennen, dass es gewisse Klassen gibt, bei denen mehr Schulkinder über oder unter der Regressionsgerade liegen. Des weiteren kann man erkennen, dass es nicht optimal ist, wenn für alle Klassen dieselbe Steigung der Regressionsgerade verwendet wird. Betrachten wir beispielsweise die zweite Klasse, kann man erkennen, dass diese Schulkinder einen viel stärkeren Zusammenhang zwischen gelösten Übungsaufgaben und erreichter Punktzahl verzeichnen als die erste Klasse. Man könnte nun mit Hilfe einer Dummy-Kodierung den Einfluss von Klassen berücksichtigen, dazu müsste aber für jede Klasse einen zusätzlichen Parameter in das Modell aufgenommen werden. Bei unserem Datensatz mit nur fünf Klassen wäre das kein grosses Problem. Bei Datensätzen mit sehr vielen Klassen würde das bedeuten, dass ganz viele Parameter in das Modell aufgenommen werden müssten. Da es grundsätzlich erstrebenswert ist, möglichst sparsame Modelle zu bilden, ist auch dies keine optimale Lösung.

Im letzten Abschnitt hat sich gezeigt, dass weder die Aggregation noch die Disaggregation der Daten zu zufriedenstellenden Ergebnissen führen, da sie massiven Einschränkungen unterliegen. Es erfordert folglich eine Erweiterung von LMs, die Zusammenhänge innerhalb und zwischen Level-2 Einheiten abbilden kann, ohne sich dabei auf eine Analyseeinheit festzulegen.

2.4 Hierarchische Linearen Modelle

In den letzten Abschnitten wurde angenommen, dass sich die Regressionskoeffizienten β_0 und β_1 feste Werte sind, die sich zwischen den Klassen nicht unterscheiden. In Abbildung 2 aus dem vorherigen Abschnitt konnte man aber erkennen, dass diese Annahme nicht in allen Fällen zu einem erwünschten Ergebnis führt. In unserem Beispiel gibt es offensichtlich Klassen, die eine über- oder unterdurchschnittliche Punktzahl erreichen. In HLMS können diese Unterschiede in den Regressionskoeffizienten zwischen den Klassen mittels zufälligen Effekten berücksichtigt werden. Unter einem zufälligen Effekt versteht man in diesem Kontext, dass die Koeffizienten der einzelnen Klassen zufällig voneinander abweichen. Die Varianz dieser zufälligen Abweichung wird von HLMS geschätzt und kann später dazu genutzt werden die einzelnen Regressionskoeffizienten der Klassen zu berechnen.

In den folgenden Abschnitten werden nun zwei Formen von HLMS besprochen, mit denen es möglich ist die Varianzen dieser Koeffizienten zu schätzen. Als erstes wird das *Random Intercept* Modell vorgestellt. Dieses Modell geht davon aus, dass nur die Achsenabschnitte (*Intercept*) der verschiedenen Klassen zufällig voneinander abweichen und die Steigung (*Slope*) über alle Klassen konstant bleibt. Das zweite besprochene Modell ist das *Random Intercept and Slope* Modell. Bei diesem Modell wird angenommen, dass sich nicht nur die Achsenabschnitte der verschiedenen Klassen unterscheiden, sondern dass auch die Steigung zwischen den Klassen zufällig variiert.

2.4.1 *Random Intercept* Modell

Die einfachste Form eines *Random Intercept* Modells ist ein Modell, das nur den Koeffizienten für den Achsenabschnitt β_{0j} und das Residuum ϵ_i enthält. Dieses Modell wird wie folgt beschrieben:

$$\begin{aligned}\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j}\end{aligned}\tag{5}$$

Bei dieser Darstellung handelt es sich um die hierarchische Notation der Gleichung, da die einzelnen Gleichungen gleich dem dazugehörigen Level zugeordnet werden. Dies wird klarer, wenn man es in Bezug zu unserem Beispiel betrachtet. Auf Level-1 befindet sich die Regressionsgleichung für die erreichte Punktzahl jedes einzelnen Schulkindes i aus der Klasse j . Dabei kann man erkennen, dass der Regressionskoeffizient β_{0j} von der Klasse j abhängt und folglich für jede Klasse einen anderen Wert annimmt. Da die Klasse eine Level-2 Variable ist, befindet sich die Gleichung für β_{0j} auf Level-2. Dabei ist γ_{00} der Gesamtmittelwert und U_{0j} die zufällige Abweichung der Klasse j vom Gesamtmittelwert. Substituiert man die Gleichung von Level-2 in die Gleichung von Level-1, erhält man folgende Darstellung des *Random Intercept* Modells:

$$\begin{aligned}y_{ji} &= \beta_{0j} + \epsilon_{ij} \\ &= \gamma_{00} + U_{0j} + \epsilon_{ij}\end{aligned}\tag{6}$$

Diese Gleichung entspricht dem leeren Modell aus Abschnitt 2.2, anhand dessen man die IKK berechnet. Ähnlich wie bei der linearen Regression können diesem Modell weitere Variablen hinzugefügt werden, um die Varianz der erreichten Punktzahl des Schulkindes i aus der Klasse j zu erklären. In unserem Beispiel ergänzen wir das Modell mit nur einer weiteren Variable x_{ij} , die der Anzahl gelöster Übungsaufgaben entspricht:

$$\begin{aligned}
\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\
\text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j} \\
\beta_1 &= \gamma_{10}
\end{aligned} \tag{7}$$

Da es sich hier um ein *Random Intercept* Modell handelt, bleibt die Steigung für alle Klassen gleich. Dies kann man an der Gleichung des Koeffizienten β_1 erkennen, da es keine zufällige Abweichung in Abhängigkeit der Klasse j von der Gesamtsteigung γ_{10} gibt. Werden nun aus (7) die beiden Gleichungen auf Level-2 in die Gleichung auf Level-1 eingesetzt, können wir das *Random Intercept* Modell wieder mit nur einer Gleichung beschreiben:

$$\begin{aligned}
y_{ji} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\
&= \gamma_{00} + U_{0j} + \gamma_{10} x_{ij} + \epsilon_{ij} \\
&= \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + \epsilon_{ij}
\end{aligned} \tag{8}$$

In Abbildung 3 kann man nun die Geraden für jede einzelne Klasse erkennen. Die rote Gerade entspricht der linearen Regressionsgerade, die durch die Disaggregation entsteht und wird hier als Vergleichswert verwendet. Die schwarzen Geraden entsprechen den Geraden, die durch das *Random Intercept* Modell geschätzt werden. Auf der linken Seite der Abbildung erkennt man relativ schnell, dass es bedeutende Unterschiede zwischen den Klassen gibt. Werden die Geraden für jede einzelne Klasse betrachtet, erhält man einen Überblick über das Leistungsniveau der verschiedenen Klassen. Beispielsweise kann man erkennen, dass die Klassen 1 und 5 eher tiefere und die Klassen 2 und 3 eher höhere Punktzahlen erreichen. Diese Unterschiede kommen durch die unterschiedliche Ausprägungen der zufälligen Abweichungen U_{0j} zustande.

Da in HLMS nicht die zufällige Abweichung jeder einzelnen Klasse, sondern nur ihre Varianz geschätzt wird, müssen die zufällige Abweichung U_{0j} und die daraus resultierenden Regressionskoeffizienten β_{0j} zuerst berechnet werden. Diese Berechnung folgt einem in der psychologischen Testtheorie bekannten Prinzip, dem sogenannten *shrinkage to the mean*

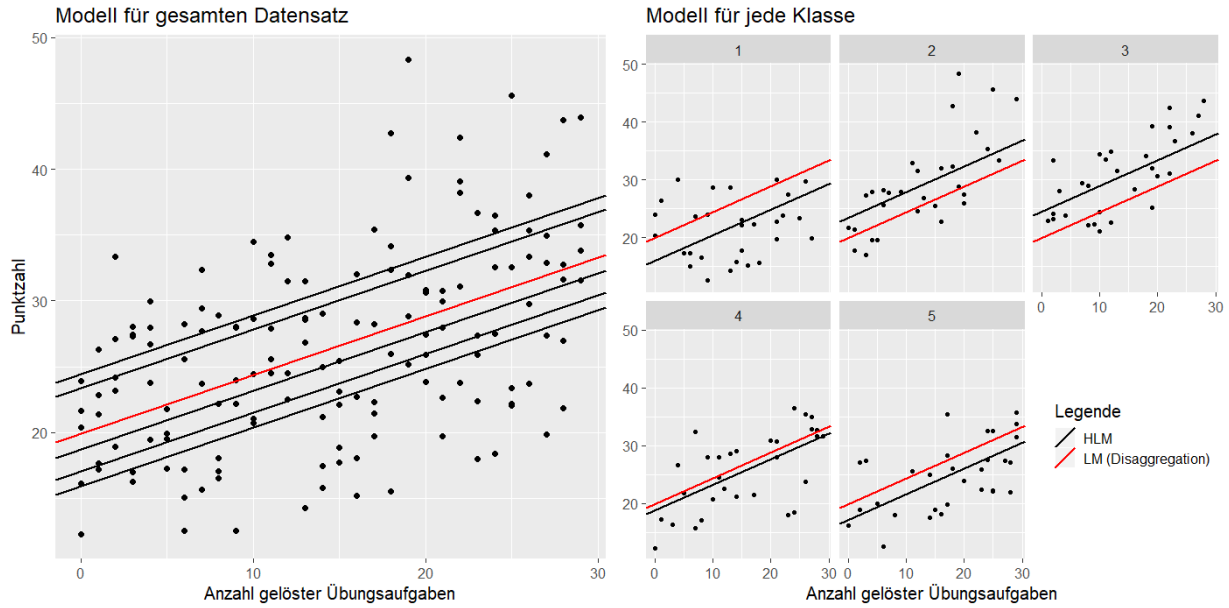


Abbildung 3: Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl mittels HLM im Vergleich zu einem LM (Disaggregation). Links: Geraden von LM und HLM im gesamten Datensatz. Rechts: Geraden von LM und HLM auf jede Klasse aufgeteilt.

(Snijders & Bosker, 2012). Dabei wird angenommen, dass sich Gruppenmittelwerte leicht in Richtung des Gesamtmittelwerts verschieben. Diese Annahme wird damit begründet, dass bei normalverteilten Daten Werte nahe am Gesamtmittelwert häufiger sind als extreme Werte. Da man bei HLMs davon aus geht, dass die zufällige Abweichung normal verteilt und im Mittel 0 ist, kann also das selbe Prinzip auch hier angewendet werden. In unserem Beispiel würde das bedeuten, dass der Mittelwert der erreichten Punktzahl der Klasse j sich leicht in Richtung des Gesamtmittelwertes aller Beobachtungen γ_{00} verschiebt. Wie stark aber dieser Mittelwert sich in Richtung des Gesamtmittelwertes verschiebt, wird durch eine Gewichtung dieser beider Mittelwerte festgelegt. Diese Gewichtung wird über die geschätzte Varianz der zufälligen Abweichung U_{0j} berechnet. Ist die geschätzte Varianz der zufälligen Abweichung gross, dann werden Gruppenmittelwerte stärker gewichtet, ist die geschätzte Varianz klein, wird der Gesamtmittelwert stärker gewichtet. Die genauen Formeln und eine ausführlichere Beschreibung dieser Berechnung können in Snijders und Bosker (2012, Abschnitt 4.8) nachgeschlagen werden.

Für die Klasse 1 ergibt sich nun beispielsweise aus unserem *Random Intercept* Modell

eine berechnete gewichtete zufällige Abweichung von $U_{01} = -4.01$ und einen Schätzer für den Gesamtmittelwert $\gamma_{00} = 19.9$. Setzt man diese beiden Werte in die Gleichung aus (7) erhält man den klassenspezifischen Achsenabschnitt β_{01} :

$$\beta_{01} = 19.9 - 4.01 = 15.89 \quad (9)$$

Schulkinder der Klasse 1 erreichen bei 0 gelösten Übungsaufgaben im Mittel eine Punktzahl von 15.89. Für Klasse 2 lässt sich ihr Achsenabschnitt β_{02} genau gleich bestimmen. Aus der klassenspezifischen Abweichung $U_{02} = 3.48$ und dem Gesamtmittelwert von $\gamma_{00} = 19.9$ ergibt sich ein Achsenabschnitt von $\beta_{02} = 23.38$. Schulkinder aus Klasse 2 erreichen bei 0 gelösten Übungsaufgaben im Mittel also eine höhere Punktzahl als Schulkinder aus Klasse 1. Diese Aussage stimmt ebenfalls mit den Geraden aus Abbildung 3 überein.

Ebenfalls kann man beobachten, dass die Geraden des hierarchischen linearen Modells besser zu den Daten passen. Betrachtet man in Abbildung 3 auf der rechten Seite die Geraden der einzelnen Klassen kann man erkennen, dass die schwarzen Geraden des HLMs die Beobachtungen besser beschreiben, als die rote Gerade, die durch das LM geschätzt wird. Allerdings gibt es immer noch Schulkinder, die noch nicht optimal durch die Gerade des HLMs beschrieben werden. Bei der zweiten und dritten Klasse erreichten beispielsweise Schulkinder, die viele Übungen gelöst haben eine noch viel höhere Punktzahl als vom Modell angenommen wird.

Anscheinend gibt es in unserem Datensatz Klassen, bei denen die Schulkinder einen stärkeren oder schwächeren Anstieg der erreichten Punktzahl bei steigender Anzahl gelöster Übungsaufgaben verzeichnen. Daraus könnte man nun folgern, dass sich nicht nur der Achsenabschnitt zwischen den Klassen unterscheidet, sondern auch die Steigung. Im folgenden Abschnitt wird nun das *Random Intercept and Slope* Modell vorgestellt und überprüft, ob diese Form von HLM besser zu unseren Daten passt als ein *Random Intercept* Modell.

2.4.2 *Random Intercept and Slope* Modell

Im letzten Abschnitt wurde das *Random Intercept* Modell besprochen und aufgezeigt, dass man durch die Hinzunahme von variierenden Achsenabschnitten eine bessere Passung zwischen dem Modell und den Daten erreicht. Um eine noch bessere Passung zu erreichen und genauere Vorhersagen zu treffen kann man nun nicht nur den Achsenabschnitt, sondern auch die Steigung variieren lassen. Dies führt zum *Random Intercept and Slope* Modell, das die Interaktion zwischen Klassenzugehörigkeit und der Anzahl gelöster Übungsaufgabe berücksichtigt. Der Regressionskoeffizient β_1 aus dem *Random Intercept* Modell (7) ist nun von der Klasse j abhängig und wird durch die zufällige Abweichung U_{1j} erweitert. Dies führt zum folgenden Modell in der hierarchischen Notation:

$$\begin{aligned}\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j} \\ \beta_{1j} &= \gamma_{10} + U_{1j}\end{aligned}\tag{10}$$

Durch Einsetzen und Umformen erhalten wir wieder die flache Notation unseres Modells:

$$\begin{aligned}y_{ji} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{10}x_{ij} + U_{0j} + U_{1j}x_{ij} + \epsilon_{ij}\end{aligned}\tag{11}$$

Dabei wurde die Gleichung so umgeformt, dass der erste Teil $\gamma_{00} + \gamma_{10}x_{ij}$ die jeweiligen Gesamtmittelwerte enthält. Diese Werte sind unveränderbar und werden folglich als fester Teil des Modells bezeichnet. Der zweite Teil der Gleichung mit $U_{0j} + U_{1j}x_{ij} + \epsilon_{ij}$ wird als zufälliger Teil bezeichnet, weil er alle veränderbaren Werte enthält. Der Term $U_{1j}x_{ij}$ beschreibt die zufällige Interaktion zwischen Gruppenzugehörigkeit und der Variable x_{ij} . Bezogen auf unser Beispiel bedeutet dieser Term, wie stark und in welche Richtung sich die erreichte Punktzahl verändert in Abhängigkeit der Klassenzugehörigkeit und der Anzahl gelöster Übungen. In Abbildung 4 ist diese zufällige Interaktion einfach zu erkennen. Für

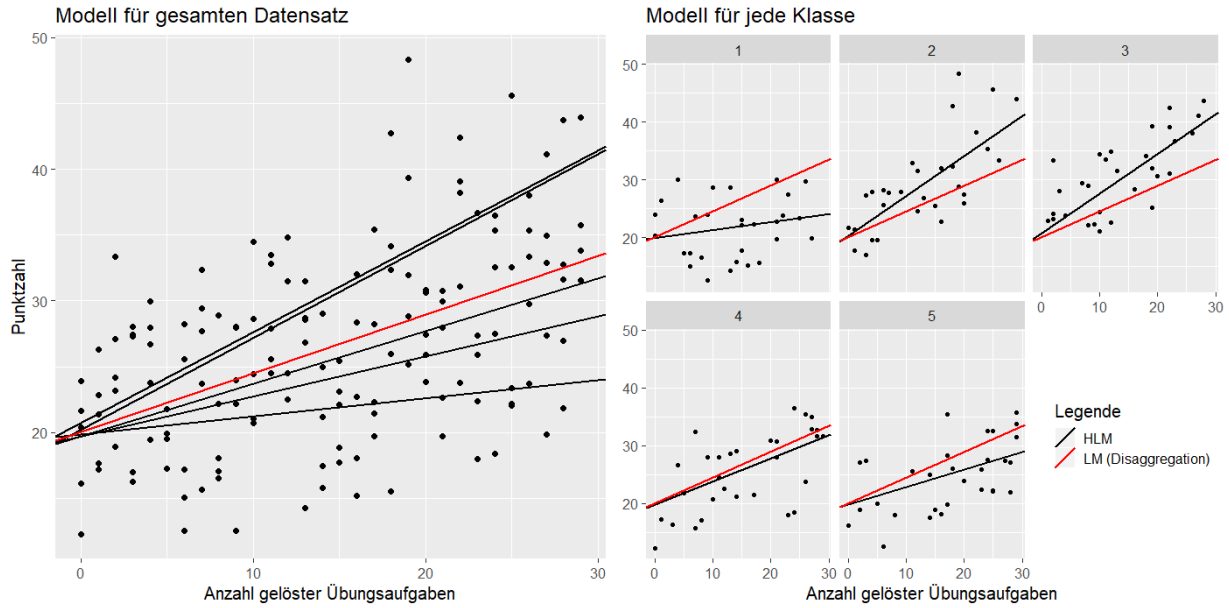


Abbildung 4: Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl unter Berücksichtigung der Klassenzugehörigkeit und deren Interaktion mit der Anzahl gelöster Übungsaufgaben

gewisse Klassen nimmt die Zufallsvariable U_{1j} einen positiven und für andere einen negativen Wert an. Dies spiegelt sich wiederum in klassenspezifischen Steigungen β_{1j} , die höher oder tiefer als die mittlere Steigung γ_{10} sind. Betrachten wir die Klasse 3 unseres Beispiels. Das *Random Intercept and Slope* Modell schätzt für diese Klasse eine zufällige Abweichung $U_{13} = 0.22$ von der mittleren Steigung $\gamma_{10} = 0.45$. Setzt man nun diese beiden Werte in die Gleichung für β_{1j} aus (10) ein, erhält man die klassenspezifische Steigung β_{13} .

$$\beta_{13} = 0.45 + 0.22 = 0.67 \quad (12)$$

Für jede weitere gelöste Übungsaufgabe eines Schulkindes j aus der Klasse 3 steigt also die erwartete Punktzahl im Mittel um 0.67 Punkte an. Betrachten wir nun die Klasse 1 aus unserem Beispiel, schätzt unser Modell eine negative zufällige Abweichung $U_{11} = -0.24$ von der mittleren Steigung. Wird dieser Wert wieder in die Gleichung für β_{1j} aus (10) eingesetzt, erhalten wir eine klassenspezifische Steigung von $\beta_{11} = 0.2$. Diese Steigung ist nun kleiner als die mittlere Steigung aller Klassen $\gamma_{10} = 0.45$. Folglich nimmt die erreichte

Punktzahl bei einer weiteren gelösten Übungsaufgabe eines Schulkindes j aus der Klasse 1 im Mittel nur um 0.2 zu.

In unserem Beispiel ist es also so, dass je höher der klassenspezifische Achsenabschnitt ist, desto höher ist auch die klassenspezifische Steigung. Man spricht hier auch von einer positiven Korrelation zwischen Achsenabschnitt und Steigung. Diese beiden Koeffizienten müssen aber nicht zwingend positiv miteinander korreliert sein. Es gibt auch die Möglichkeit, dass diese Koeffizienten negativ korreliert oder sogar unkorreliert sind. In Abbildung 5 kann man betrachten, wie die weiteren Korrelationen von Achsenabschnitt und Steigung sich auswirken. Bei einer negativen Korrelation nimmt die Steigung mit der Höhe des Achsenabschnittes ab. Sind die beiden Koeffizienten unkorreliert, bildet sich kein offensichtliches Muster zwischen der Höhe des Achsenabschnittes und der Steigung.

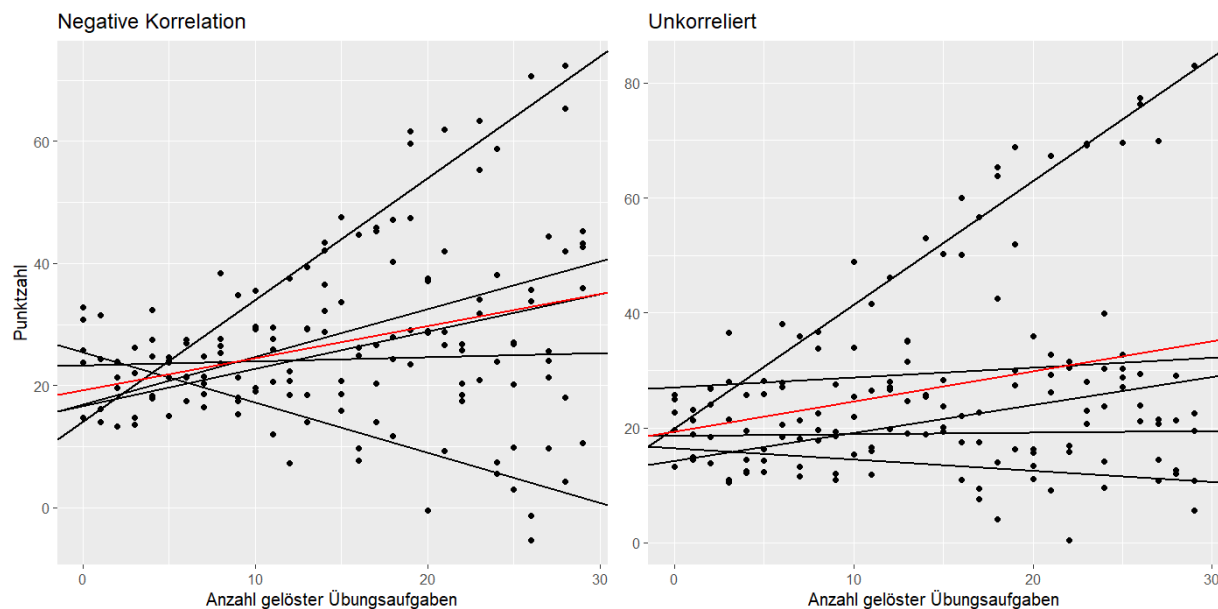


Abbildung 5: Darstellung einer negativen und nicht vorhandenen Korrelationen zwischen Achsenabschnitt und Steigung

In Abschnitt 2.4.1 wurden die Residuenplots eines linearen Modells und eines *Random Intercept* Modells verglichen. Dabei ist aufgefallen, dass das *Random Intercept* Modell zwar kleinere Residuen aufwies, diese aber an den Endpunkten nicht um den Nullpunkt normalverteilt waren. Vergleicht man nun den Residuenplot aus Abbildung 6 mit den

anderen aus der Abbildung ??, hat sich die Lage bezüglich der Verteilung der Residuen an den Endpunkten deutlich verbessert im *Random Intercept and Slope* Modell.

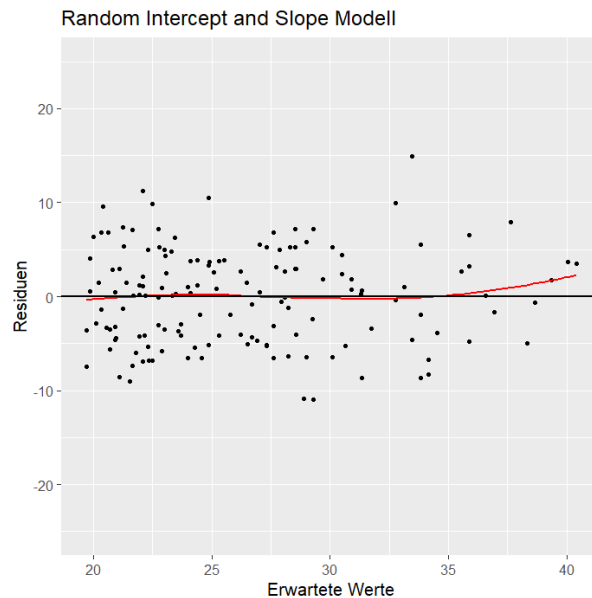


Abbildung 6: Residuenplot des *Random Intercept and Slope* Modells

2.4.3 *Intercept* und *Slope* Variabilität

Wie bei linearen Regressionsmodellen wird auch bei hierarchischen linearen Regressionsmodellen versucht die Variabilität einer bestimmten Abhängigen Variablen zu erklären. Diese unerklärte Variabilität hängt in hierarchischen linearen Modellen nicht nur von der Varianz des Residuums ϵ_{ij} ab, sondern auch von der Varianz der zufälligen Abweichung des Achsenabschnittes U_{0j} und der Steigung U_{1j} (Snijders & Bosker, 2012). Um nun unerklärte Variabilität in hierarchischen Modellen zu beschreiben, können alle dieser Komponenten angegangen werden. Um die Varianz des Residuums zu verringern können, wie bei der normalen linearen Regression, weitere Level-1 Variablen in das Modell aufgenommen werden. Um die Varianz der beide zufälligen Abweichungen zu verringern wird es etwas anspruchsvoller, da diese beiden Varianzen nicht durch Unterschieden innerhalb der Gruppen sondern zwischen den Gruppen entstehen. Folglich können diese Varianzen nicht durch das Hin-

zufügen von Level-1 Variablen reduziert werden, sondern erfordern das Hinzufügen von Level-2 Variablen. Snijders und Bosker (2012) erweitern hierfür die beiden Gleichungen für die Regressionskoeffizienten β_{0j} und β_{1j} :

$$\begin{aligned}
\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij} \\
\text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + \gamma_{01}z_j + U_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}z_j + U_{1j}
\end{aligned} \tag{13}$$

Dabei ist z_j eine Level-2 Variable, die sich zwischen den Gruppen unterscheidet. Auf unser Beispiel bezogen, könnte die Variable z_j die Anzahl Fenster im Klassenzimmer sein. Durch das Hinzufügen dieser Level-2 Variablen werden die Regressionskoeffizienten selbst zu einer Abhängigen Variablen eines Regressionmodells. Setzt man nun die beiden Koeffizienten in die Level-1 Gleichung ein erhält man wieder die flache Notation des Modells:

$$\begin{aligned}
y_{ji} &= \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij} \\
&= \gamma_{00} + \gamma_{01}z_j + U_{0j} + (\gamma_{10} + \gamma_{11}z_j + U_{1j})x_{ij} + \epsilon_{ij} \\
&= \gamma_{00} + \gamma_{01}z_j + U_{0j} + \gamma_{10}x_{ij} + \gamma_{11}z_jx_{ij} + U_{1j}x_{ij} + \epsilon_{ij} \\
&= \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}z_jx_{ij} + U_{0j} + U_{1j}x_{ij} + \epsilon_{ij}
\end{aligned} \tag{14}$$

Auch wenn es in der hierarchischen Notation einfacher ist zu erkennen, welche Varianz genau durch die Hinzunahme dieser Level-2 Variable verringert wird, erkennt man in der flachen Notation einen weiteren wichtigen Zusammenhang. Der Term $\gamma_{11}z_jx_{ij}$ beschreibt eine besondere Interaktion zwischen einer Level-1 und einer Level-2 Variable und wird, wie bereits in der Einleitung kurz erwähnt, als *Cross-Level* Interaktion bezeichnet. Da diese *Cross-Level* Interaktion durch das Hinzufügen einer Level-2 Variable als Prädiktor in der Gleichung des Steigungskoeffizienten entsteht, ist diese Interaktion vor allem wichtig, um unerklärte Varianz in der Steigung zu erklären. In unserem Beispiel würde diese *Cross-Level* Interaktion also durch die Interaktion zwischen der Anzahl gelösten Übungsaufgaben und der Anzahl Fenster im Klassenzimmer beschreiben werden.

In den letzten Abschnitten wurden zwei verschiedene hierarchische lineare Modelle besprochen und das Prinzip ihrer Anwendung etwas näher gebracht. Es ist dabei zu berücksichtigen, dass hier nur die Grundlagen zu den hierarchischen linearen Modellen behandelt wurden. Für eine weitere Vertiefung dieses Themas wird auf die gängige Literatur zur Multilevel Analyse verwiesen (Gelman & Hill, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Twisk, 2006). Im anschliessenden Abschnitt geht es nun darum, wie man solche Modelle in R berechnet und Multilevel Analysen durchführt.

2.5 Anwendung von Multilevel Analyse in R

Das Konzept von hierarchischen linearen Modellen wurde in den letzten Abschnitten ausführlich besprochen. In den nächsten Abschnitten wird vor allem die Anwendung dieser Modelle behandelt. Dabei wird der Fokus auf die Programmiersprache R gelegt und dessen Zusatzpaket `lme4` (Bates et al., 2015), das neben weiteren Paketen die Analyse mittels HLMs ermöglicht. Dabei wird davon ausgegangen, dass die Grundlagen dieser Programmiersprache verstanden wurden. Zuerst werden allgemeine Informationen und die Syntax von `lme4` besprochen. Anschliessend wird anhand unseres Beispiels ein erstes Modell geschätzt und dessen `summary()`-Output interpretiert. Dabei werden noch weitere Möglichkeiten besprochen, wie man Ergebnisse eines Geschätzten Modells präsentieren kann. Nachdem die Syntax und Interpretation des Modells besprochen wurden, wird Schritt für Schritt eine Multilevel Analyse unseres Beispiels durchgeführt. Hierbei wird aufgezeigt, wie HLMs aufgebaut und miteinander verglichen werden, um das geeignetste Modell zu identifizieren. Am Ende dieses Abschnittes werden Effektsterkemasse besprochen, die bei der Auswertung und Berichterstattung von HLMs wichtig sind.

2.5.1 Informationen und Syntax von `lme4`

Mit dem Paket `lme4` lassen sich verschiedenste Formen von hierarchischen Modellen schätzen und analysieren. Dabei werden wir uns hier hauptsächlich auf die Funktion `lmer()` beschränken. Diese Funktion wird verwendet, um hierarchische lineare Modelle zu schätzen

und dessen Syntax ist relativ ähnlich mit der Syntax des Befehls für die Berechnung normaler linearer Modelle `lm()`. Sie ist wie folgt aufgebaut:

```
lmer(formula, data, REML)
```

Dabei wird in **formula** die gewünschte Formel des Modells eingegeben, bei **data** wird der Datensatz festgelegt anhand das Modell geschätzt werden soll und bei **REML** wird durch einen logischen Operator (*TRUE* oder *FALSE*) eingestellt, ob das Modell mit *Restricted Maximum Likelihood* (REML) oder mit *Maximum Likelihood* (ML) geschätzt werden soll. In unserem Fall ist diese Einstellung vor allem beim Vergleich von Modellen wichtig. Bei Modellen die mit REML geschätzt wurden, können nur die zufälligen Effekte miteinander verglichen werden. Wenn man also feste als auch zufällige Effekte vergleichen möchte, sollte man Modelle mit ML schätzen (Peugh, 2010).

```
lmer(Abhängige Variable ~ Feste Effekte + (Zufällige Effekte | Gruppe), ... )
```

Bis zum Term innerhalb der Klammer ist die Syntax von `lmer()` genau die gleiche wie bei `lm()`. Dabei wird auf zuerst die zu erklärende Variable aufgeführt und anschliessend alle Variablen, die als feste Effekte in das Modell aufgenommen werden sollen. Innerhalb der Klammern können nun die zufälligen Effekte und die variierende Gruppe festgelegt werden. Dabei kann innerhalb der Klammern eingestellt werden, ob der Achsenabschnitt und die Steigung korrelieren sollten oder nicht. In Tabelle 4 findet man einen Überblick über die möglichen Interaktionen, die man innerhalb der Klammer festlegen kann und die für uns relevant sind².

2.5.2 Interpretation eines Outputs

Wir haben nun den groben Aufbau von `lmer()` besprochen und schätzen nun ein erstes Modell. Dafür muss zuerst das Paket `lme4` und unser Beispieldatensatz geladen werden.

²Weitere Funktionen und eine ausführliche Beschreibung des Paketes `lme4` können in Bates et al. (2015) nachgeschlagen werden.

Tabelle 4: Mögliche Syntax für `lmer()` nach Bates et al. (2015)

Formel	Bedeutung
$(1 \mid \text{Gruppe})$	Zufälliger Achsenabschnitt
$(x \mid \text{Gruppe})$	Korrelierter Achsenabschnitt und Steigung
$(x \parallel \text{Gruppe})$	Unkorrelierter Achsenabschnitt und Steigung

```
library(lme4)
beispiel_data <- readRDS(file = "dataset_theory")
```

Anschliessend wird ein Modell mit der Anzahl gelösten Übungen als fester Effekt und mit variierenden Achsenabschnitten und Steigungen geschätzt und mittels `summary()` übersichtlich ausgegeben.

```
beispiel_model <- lmer(punktzahl ~ uebung + (uebung | klasse),
  data = beispiel_data, REML = FALSE)
summary(beispiel_model)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: punktzahl ~ uebung + (uebung | klasse)
## Data: beispiel_data
##
##      AIC      BIC   logLik deviance df.resid
##      944      963    -466     932      144
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.1080 -0.8358  0.0364  0.7288  2.9172
##
## Random effects:
##  Groups   Name                Variance Std.Dev. Corr
##  klasse  (Intercept)    0.9191   0.959
##          uebung         0.0349   0.187   1.00
## Residual                26.5996   5.157
## Number of obs: 150, groups:  klasse, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  19.9971    0.9360   21.36
```

```
## uebung      0.4481      0.0973      4.61
##
## Correlation of Fixed Effects:
##      (Intr)
## uebung 0.000
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

Dabei kann man im Output direkt die geschätzten Werte für die zufälligen und festen Effekte ablesen. Bei den zufälligen Effekten werden für den Achsenabschnitt als auch für die Steigung die geschätzte Varianzen, die Standardabweichungen, als auch die Korrelation dieser beiden angegeben. Ebenfalls werden direkt unterhalb der zufälligen Effekte aufgeführt wie gross die verwendete Stichprobe war und wie viele Gruppen für die Schätzung verwendet wurden. In unserem Fall waren das 150 Schulkinder, verteilt auf 5 Klassen. Der Abschnitt zu den festen Effekten ähnelt stark dem Output der Funktion `lm()` und lässt sich auch dementsprechend interpretieren. Es werden hier aber keine p -Werte angezeigt, da man sich in der Forschung noch uneinig darüber ist, wie genau die Anzahl der Freiheitsgrade geschätzt werden soll (Peugh, 2010; Snijders & Bosker, 2012)³. Eine Interpretation des Achsenabschnittes unter Berücksichtigung der zufälligen Effekte würde dann wie folgt lauten:

Die erwartete erreichte Punktzahl in der Mathematikprüfung ist $20 \pm 2 \cdot 0.96$ Punkte, wenn keine einzige Übungsaufgabe gelöst wurde.

Und für die Steigung:

Die erwartete Zunahme der Punktzahl beträgt $0.45 \pm 2 \cdot 0.19$ pro gelöste Übungsaufgabe.

Dabei ist 0.96 die Standardabweichung des Achsenabschnittes und 0.19 die Standardabweichung der Steigung. Bei der Interpretation ist es von Vorteil die Standardabweichung zu verwenden, da diese nicht wie die Varianz quadriert ist und somit der Masseinheit (Punktzahl) entspricht.

³Möchte man p -Werte anzeigen lassen, kann man das mit dem Paket `lmerTEST` (Kuznetsova et al., 2017).

2.5.3 Berechnung der Intraklassen Korrelation in R

Wir gelangen nun zur Analyse unseres Beispieldatensatzes. Bevor wir mit der Analyse starten können ist es wichtig, dass wir das Level unserer Forschungsfrage klar definieren. In unserem Fall möchten wir herausfinden, wie genau sich die Prüfungsleistung von Schulkindern mit der Anzahl gelöster Übungsaufgaben unter Berücksichtigung der Klassenzugehörigkeit verändert. Das bedeutet, dass sich unsere abhängige Variable auf Level-1 befindet. Ebenfalls sollte berücksichtigt werden, dass es zu Interaktionen zwischen Level-1 und Level-2 variablen kommen kann. Da wir sehr wahrscheinlich nicht nur feste Effekte sondern auch zufällige Effekte miteinander Vergleichen möchten, müssen wir zudem die Modelle mit ML schätzen (Peugh, 2010).

Als Erstes sollte immer geprüft werden, ob eine Multilevel Analyse überhaupt nötig ist. Dies geschieht zum einen durch die Berechnung der Intraklassen Korrelation und der Überprüfung, ob überhaupt mittlere Unterschiede zwischen den Klassen bestehen. Wie bereits im Abschnitt 2.2 besprochen, lässt sich die IKK aus den Varianzen des leeren Modells berechnen, das hier noch einmal kurz aufgeführt ist:

$$\begin{aligned}\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j}\end{aligned}\tag{15}$$

Dieses Modell wird wie folgt in R mittels der Funktion `lmer()` geschätzt:

```
m_leer <- lmer(punktzahl ~ (1 | klasse), data = beispiel_data,
               REML = FALSE)

summary(m_leer)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: punktzahl ~ (1 | klasse)
## Data: beispiel_data
##
##      AIC      BIC   logLik deviance df.resid
##    1009     1018    -502     1003      147
##
```

```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.0308 -0.7713 -0.0074  0.6433  2.9282
##
## Random effects:
##   Groups   Name                Variance Std.Dev.
##   klasse   (Intercept)         9.57      3.09
##   Residual                    44.00      6.63
## Number of obs: 150, groups:  klasse, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    26.36      1.49    17.7
```

Aus dem Output können nun die Varianzen für den Achsenabschnitt und das Residuum abgelesen und in die Formel (2) eingesetzt werden. Daraus ergibt sich die bereits berechnete intraklassen Korrelation von $\rho_I = 0.18$. Folglich werden 18% der Variabilität in der erreichten Punktzahl durch die Klassenzugehörigkeit erklärt. Um nun noch herauszufinden, ob diese Klassenunterschiede auch signifikant sind, kann eine Varianzanalyse durchgeführt werden (Snijders & Bosker, 2012). Dies geschieht indem man ein normales lineares Modell schätzt, das nur anhand der Klasse Leistungsunterschiede erklären möchte. Wie ebenfalls aus Abschnitt 2.2 bekannt, ergibt sich aus der Varianzanalyse einen hoch signifikanten p -Wert. Es bestehen folglich mittlere Klassenunterschiede in der erreichten Punktzahl, die in der Analyse zu berücksichtigen sind.

```
m_linear <- lm(punktzahl ~ klasse, data = beispiel_data)

anova(m_linear)

## Analysis of Variance Table
##
## Response: punktzahl
##              Df Sum Sq Mean Sq F value  Pr(>F)
##   klasse       4   1656     414    9.41 8.7e-07 ***
## Residuals   145    6381      44
## ---
```

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.5.4 Aufbau und Vergleich von Modellen

Da nun die Frage geklärt ist, ob unser Datensatz eine Multilevel Analyse verlangt, können wir damit anfangen unser leeres Modell mit weitere Prädiktoren aufzubauen. In unserem Beispiel fügen wir die Anzahl gelöste Übungsaufgaben als festen Effekt dem Modell hinzu. Natürlich könnten hier noch weitere feste Effekte hinzugefügt werden, damit das Beispiel aber übersichtlich bleibt, ist das Modell hier auf einen festen Effekt beschränkt:

$$\begin{aligned} \text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_1 \cdot \text{uebungen}_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j} \\ \beta_1 &= \gamma_{10} \end{aligned} \tag{16}$$

In der oberen Gleichung kann man erkennen, dass es sich hier um ein einfaches *Random Intercept* Modell handelt, da bei β_1 keine zufällige Abweichung hinzugefügt wurde. Um nun ein *Random Intercept* Modell in R zu schätzen wird der Funktion die Variable wie folgt hinzugefügt:

```
m_intercept <- lmer(punktzahl ~ uebung + (1 | klasse),
  data = beispiel_data, REML = FALSE)
```

Dabei ist zu beachten, dass der Term innerhalb der Klammern der Vorgabe aus Tabelle 4 entspricht, um ein *Random Intercept* Modell zu schätzen. Möchte man nun zusätzlich ein *Random Intercept and Slope* Modell schätzen wird entsprechend des Abschnittes 2.4.2 der Gleichung von β_{1j} eine zufällige Abweichung von der mittleren Steigung hinzugefügt:

$$\begin{aligned} \text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_1 \cdot \text{uebungen}_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j} \\ \beta_{1j} &= \gamma_{10} + U_{1j} \end{aligned} \tag{17}$$

Dieser zufällige Effekt wird nun auch in der Formel in R gemäss Tabelle 4 hinzugefügt:

```
m_uncorr <- lmer(punktzahl ~ uebung + (uebung || klasse),
  data = beispiel_data, REML = FALSE)

m_corr <- lmer(punktzahl ~ uebung + (uebung | klasse),
  data = beispiel_data, REML = FALSE)
```

Wir haben hier gleich zwei Varianten eines *Random Intercept and Slope* Modell geschätzt. Zum einen eines bei dem der Achsenabschnitt und die Steigung unkorreliert sind und zum anderen ein Modell, bei dem beide Koeffizienten miteinander korrelieren. Um das Modell zu identifizieren, das am meisten erklärte Varianz aufweist, kann man wie beim Vergleich von normalen linearen Modellen mit dem Befehl `anova()` die Modelle vergleichen. Dabei wird hier nicht wie für normale lineare Modelle üblich ein *F* Test sondern ein *Likelihood Ration Test* verwendet (Peugh, 2010; Snijders & Bosker, 2012).

Wenn Modelle mit einer *Maximum Likelihood* Methode geschätzt werden, wird immer eine *Likelihood* des Modells angegeben, die wiederum in einen sogenannten *Deviance* Wert umgewandelt werden kann. Betrachtet man den Output aus Abschnitt 2.5.2, findet man den *Deviance* Wert in der oberen Hälfte des Outputs. Dieser *Deviance* Wert gibt an, wie genau das Modell zu den Daten passt und kann folglich verwendet werden, um Modelle zu vergleichen (Snijders & Bosker, 2012). Die Differenz des *Deviance* Werts zweier Modelle kann als Testwert einer χ^2 -Verteilung verwendet werden.

Betrachtet man nun den folgenden Outputs wird klar, dass ein *Random Intercept and Slope* Modell mit unkorrelierten Achsenabschnitten und Steigungen am besten zu unseren Daten passt.

```
anova(m_intercept, m_corr, m_uncorr, method = "LRT")

## Data: beispiel_data
## Models:
## m_intercept: punktzahl ~ uebung + (1 | klasse)
## m_uncorr: punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
```

```
## m_corr: punktzahl ~ uebung + (uebung | klasse)
##           Df AIC BIC logLik deviance Chisq Chi Df
## m_intercept 4 952 964   -472     944
## m_uncorr     5 944 959   -467     934 10.46      1
## m_corr       6 945 963   -466     933  1.31      1
##           Pr(>Chisq)
## m_intercept
## m_uncorr      0.0012 **
## m_corr        0.2529
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bis jetzt haben wir uns nur mit Level-1 Variablen beschäftigt. Möchten wir dem Modell die Anzahl Fenster im Klassenzimmer als Level-2 Variable hinzufügen, kann man diese Variable mit *Cross-Level* Interaktion oder ohne hinzufügen. Ebenfalls ist zu beachten, dass Level-2 Variablen in einem hierarchischen linearen Modell mit zwei Level nur mit festen Effekten hinzugefügt werden können. Da es in einem zwei Level Modell nicht noch ein drittes höheres Level gibt, von dem eine Level-2 Variable abhängen könnte. Mathematisch würde das Hinzufügen einer Level-2 Variable ohne *Cross-Level* Interaktion wie folgt aussehen:

$$\begin{aligned}
\text{Level 1: } y_{ji} &= \beta_{0j} + \beta_1 \cdot \text{uebungen}_{ij} + \epsilon_{ij} \\
\text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01} \cdot \text{fenster}_j + U_{0j} \\
\beta_{j1} &= \gamma_{1j} + U_{1j}
\end{aligned} \tag{18}$$

In Abschnitt 2.4.3 wurde allerdings mathematisch gezeigt, dass *Cross-Level* Interaktionen wichtig für die Erklärung von Steigungsvarianz sind. In der folgenden Gleichung wurde die Variable dementsprechend mit einer *Cross-Level* Interaktion hinzugefügt:

$$\begin{aligned}
\text{Level 1: } y_{ji} &= \beta_{0j} + \beta_1 \cdot \text{uebungen}_{ij} + \epsilon_{ij} \\
\text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01} \cdot \text{fenster}_j + U_{0j} \\
\beta_{j1} &= \gamma_{1j} + \gamma_{11} \cdot \text{fenster}_j + U_{1j}
\end{aligned} \tag{19}$$

In R ist das Hinzufügen von Level-2 Variablen relativ einfach, da eine Level-2 Variable

in unserem Fall nur als fester Effekt hinzugefügt werden kann, wird sie einfach in die Gleichung vor dem Term in der Klammer eingefügt. Folgend wurden die beiden Modelle geschätzt, eines ohne und eines mit *Cross-Level* Interaktion.

```
m_lvl2 <- lmer(punktzahl ~ uebung + fenster + (uebung || klasse),
              data = beispiel_data, REML = FALSE)

m_cross <- lmer(punktzahl ~ uebung * fenster + (uebung || klasse),
              data = beispiel_data, REML = FALSE)
```

Diese Modelle werden nun wieder mit unserem unkorrelierten *Random Intercept and Slope* Modell verglichen, um herauszufinden, ob das Hinzufügen einer Level-2 Variable das Modell verbessert hat:

```
anova(m_uncorr, m_lvl2, m_cross, method = "LRT")

## Data: beispiel_data
## Models:
## m_uncorr: punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
## m_lvl2: punktzahl ~ uebung + fenster + ((1 | klasse) + (0 + uebung |
## m_lvl2:      klasse))
## m_cross: punktzahl ~ uebung * fenster + ((1 | klasse) + (0 + uebung |
## m_cross:      klasse))
##           Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m_uncorr   5 944 959   -467      934
## m_lvl2     6 942 960   -465      930  3.39     1    0.066
## m_cross    7 944 965   -465      930  0.10     1    0.754
##
## m_uncorr
## m_lvl2    .
## m_cross
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wie man erkennen kann, ist keines der neuen Modelle besser als das unkorrelierte *Random Intercept and Slope* Modell. Folglich wird für die weiteren Abschnitte dieses Modell verwendet, da es anscheinend am meisten Varianz aufklärt.

2.5.5 Auswertung von Modellen

Bis jetzt wurden nur Modelle miteinander verglichen und herausgefunden, dass ein unkorreliertes *Random Intercept and Slope* Modell am meisten Varianz aufklärt. In der normalen linearen Regression gibt es viele verschiedene Effektstärkemasse, die genutzt werden können, um Aussagen über Modelle zu treffen. Beispielsweise gibt es das Bestimmtheitsmass R^2 oder das Cohen's d , die den meisten ein Begriff sind. In den hierarchischen linearen Modell ist das Berechnen von Effektstärkemasse etwas komplizierter und es herrscht aktuell keinen Konsens darüber, was für Effektstärkemasse genau verwendet werden sollen (Peugh, 2010; Snijders & Bosker, 2012).

Dabei werden Effektstärkemasse von hierarchischen linearen Modellen in globale und lokale Effektstärkemasse getrennt (Peugh, 2010). Bei den globalen Effektstärkemasse gibt es mehrere Formen von Bestimmtheitsmassen R^2 , von denen hier nur dasjenige von Snijders und Bosker (2012) behandelt wird. Die beiden Autoren sprechen davon, dass bei hierarchischen linearen Modellen auf mehreren Levels die proportional erklärte Varianz berechnet werden kann. Ausgehend von unserem Beispiel mit zwei Level kann man proportional erklärte Varianz auf Stufe des Individuums und auf Stufe der Gruppe berechnen. Snijders und Bosker (2012) geben an, dass vor allem die proportionale erklärte Varianz des Individuums von praktischer Relevanz ist und wird durch R_1^2 gekennzeichnet. Die Berechnung von R_1^2 erfolgt dann mit:

$$\begin{aligned}
 R_1^2 &= 1 - \frac{\text{Gesamtvarianz des Modells mit Prädiktoren}}{\text{Gesamtvarianz des leeren Modells}} \\
 &= 1 - \frac{\sigma^2 + \tau_0^2}{\sigma_{leer}^2 + \tau_{0leer}^2}
 \end{aligned}
 \tag{20}$$

Dabei ist σ^2 die jeweilige Varianz des Residuums und τ_0^2 die Varianz des Achsenabschnittes des jeweiligen Modelles. Da in der Berechnung von R_1^2 die Varianz der Steigung nicht berücksichtigt wird, kann dieses Effektstärkemasse nur für *Random Intercept* Modelle berechnet werden. Snijders und Bosker (2012) empfehlen für *Random Intercept and Slope*

Modelle, dass diese mit den selben festen Effekten als *Random Intercept* Modell geschätzt werden, um anschliessend das R_1^2 zu berechnen. Diese Methode ist viel einfacher als andere gängige Methoden zur Berechnung von R_1^2 für *Random Intercept and Slope* Modelle und sollte normalerweise zu R_1^2 Werte führen, die sehr nahe an den Werten für ein *Random Intercept and Slope* Modell liegen.

Möchten wir nun R_1^2 für unser Beispiel berechnen müssen wir zuerst unser Modell als *Random Intercept* Modell schätzen.

```
m_i <- lmer(punktzahl ~ uebung + (1 | klasse),
            data = beispiel_data, REML = FALSE)

summary(m_i)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: punktzahl ~ uebung + (1 | klasse)
## Data: beispiel_data
##
##      AIC      BIC   logLik deviance df.resid
##      952      964     -472     944      146
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.0508 -0.8023 -0.0412  0.7507  3.0677
##
## Random effects:
##  Groups   Name                Variance Std.Dev.
##  klasse   (Intercept) 12.2         3.49
##  Residual                    29.1         5.39
## Number of obs: 150, groups:  klasse, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 19.9079     1.7874   11.14
## uebung       0.4453     0.0519    8.58
##
## Correlation of Fixed Effects:
##      (Intr)
## uebung -0.421
```


Im Output aus Abschnitt 2.5.3 können die Varianzen des leeren Modells abgelesen werden und mit den Varianzen aus dem Output für das *Random Intercept* Modells in die Formel eingesetzt werden:

$$R_1^2 = 1 - \frac{29.09 + 12.17}{44 + 9.57} = 0.23 \quad (21)$$

Folglich kann man nun die Aussage treffen, dass unser Modell 23% der Varianz in der erreichten Punktzahl von Schulkindern erklärt.

Es gibt aber auch Situationen in denen man genau wissen möchte, wie viel Varianz durch die Hinzunahme eines bestimmten Prädiktors im Modell erklärt wird. Dies kann man mit der proportionalen Reduktion der Varianz (PRV) erklären, die als eines der lokale Effektstärkemaße von hierarchischen linearen Modellen gilt (Peugh, 2010; Woltman et al., 2012). Die Berechnung der proportionalen Reduktion der Varianz wird für jede einzelne Varianz des Modells durchgeführt. Das bedeutet, dass die PRV für die Varianz des Residuums, des Achsenabschnittes und für die der Steigung berechnet werden kann. Die Formel zur Berechnung der PRV bleibt aber für alle Fälle die gleiche:

$$PRV = \frac{Var_0 - Var_1}{Var_0} \quad (22)$$

Dabei ist Var_0 die Varianz des Modells, das den gewünschten Prädiktor nicht enthält und Var_1 des Modells, das den Prädiktor enthält. Gehen wir nun ganz an den Anfang unserer Analyse zurück und überprüfen, wie viel Varianz im Vergleich zum leeren Modell durch die Hinzunahme der Anzahl gelösten Übungsaufgaben als Prädiktor reduziert wird. Dazu wurden folgend die Varianzkomponenten des leeren Modells und des ersten *Random Intercept* Modells ausgegeben.

```
m_leer <- lmer(punktzahl ~ (1 | klasse), data = beispiel_data,
  REML = FALSE)
print(VarCorr(m_leer), comp = "Variance")
## Groups      Name      Variance
```

```
##  klasse  (Intercept)  9.57
##  Residual                44.00

m_uebung <- lmer(punktzahl ~ uebung + (1 | klasse), data = beispiel_data,
  REML = FALSE)
print(VarCorr(m_uebung), comp = "Variance")

##  Groups   Name      Variance
##  klasse  (Intercept) 12.2
##  Residual                29.1
```

Die Varianz der Residuen des leeren Modells beträgt $\sigma_0^2 = 44$ und die Varianz der Residuen des *Random Intercept* Modells $\sigma_1^2 = 29.09$. Fügen wir diese beiden Werte in die Formel (22) ein erhalten wir die proportionale Reduktion der Varianz des Residuums:

$$PRV_R = \frac{44 - 29.09}{44} = 0.34 \quad (23)$$

Daraus kann nun geschlossen werden, dass durch die Hinzunahme der Anzahl gelösten Übungsaufgaben als Prädiktor für die erreichte Punktzahl in der Mathematikprüfung eine Reduktion der Residualvarianz von 34% erreicht wird. Da wir in unserem Modell nur eine Level-1 Variable als festen Effekt hinzugefügt haben, kann keine weitere PRV für den Achsenabschnitt und der Steigung berechnet werden. Möchte man die Varianzen dieser beiden Koeffiziente reduzieren, müssen wie bereits in Abschnitt 2.4.3 besprochen Level-2 Variablen in das Modell aufgenommen werden. Erst dann kann auch die proportionale Reduktion der Varianz mit der oberen Formel berechnet werden.

Im letzten Abschnitt wurde anhand eines einfachen Beispiels aufgezeigt, wie man eine exemplarische Multilevel Analyse in R durchführt. Das Ziel dieses Abschnittes war es in erster Linie zu zeigen, wie mit dem Paket **lme4** arbeitet und welche Möglichkeiten es gibt, um hierarchische Daten in R zu analysieren.

3 Simulationstudie zur Multilevel Analyse

Der Einfluss von hierarchischen Datenstrukturen auf die Analyse wurden konzeptionell in den letzten Abschnitten diskutiert und vorgestellt. Im folgenden Abschnitt geht es nun, um die wissenschaftliche Replikation und Überprüfung dieses Einflusses. Dabei wird vor allem der Fokus auf die Unterschiede zwischen der Analyse mittels normaler linearer Regression und hierarchischer linearer Regression legen. Um diese beide Methoden zu vergleichen, wird eine Simulationsstudie durchgeführt.

Anschliessend an die Simulationsstudie wird eine Shiny App (Chang et al., 2019) vorgestellt, die im Laufe dieser Arbeit programmiert wurde und mit der es Nutzern möglich sein wird, zum einen das Konzept der Multilevel Analyse zu verstehen und zum anderen die Simulationsstudie aus dieser Arbeit selbst durchzuführen.

3.1 Herleitung der Forschungsfrage

In der Einleitung wurde bereits erwähnt, dass in der psychologischen Forschung hierarchische Datenstrukturen keine Seltenheit sind und es wurden einige Beispiele für solche hierarchischen Daten genannt (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Woltman et al., 2012). Allerdings ist es oft so, dass sich Forschende dieser Datenstruktur oder den Möglichkeiten von hierarchischen linearen Modellen nicht bewusst sind (McNeish, 2014). Dies kann dazu führen, dass diese hierarchischen Daten mittels normalen linearen Modellen anstelle von hierarchischen linearen Modellen analysiert wird. Das muss aber nicht zwingend ein Problem darstellen, da in einigen Studien gezeigt werden konnte, dass die Schätzung der Regressionskoeffizienten beider Analysemethoden auch bei hohem Einfluss der Klassenzugehörigkeit relativ genau ist (McNeish, 2014; Mundfrom & Schults, 2002). Das bedeutet, dass die Schätzung des Effekts einer Intervention oder des Achsenabschnitts bei beiden Methoden relativ nahe am Populationsmittelwert sind.

Allerdings ist eine genaue Schätzung der Regressionskoeffizienten nicht ausreichend, um zu bestimmen, ob der Effekt einer Intervention auch signifikant ist. Um das zu überprüfen wird üblicherweise ein t Test durchgeführt (Snijders & Bosker, 2012). Die Prüfgrösse des t

Tests wird über das Verhältnis zwischen des geschätzten Regressionskoeffizienten und dessen Standardfehler bestimmt. Wird beispielsweise ein Standardfehler zu klein geschätzt, steigt die Prüfgrösse an und die Rate in der die Nullhypothese abgelehnt wird, nimmt zu. Wird der Standardfehler zu gross geschätzt, verkleinert sich die Prüfgrösse und die Rate in der die Alternativhypothese abgelehnt wird, nimmt zu. Folglich führt eine Unterschätzung des Standardfehlers zu einer erhöhten Fehler Typ 1 Rate und eine Überschätzung zu einer erhöhten Fehler Typ 2 Rate und somit zu einer geringeren Power (Snijders & Bosker, 2012). Unter Power wird die Wahrscheinlichkeit verstanden, einen Effekt zu finden, wenn dieser auch effektiv in der Population vorhanden ist (Scherbaum & Ferreter, 2009). Daher ist eine genau Schätzung des Standardfehlers umso wichtiger, da dieser massgeblich zur Testung des Effekts beiträgt. Da der Standardfehler in einem direkten Zusammenhang mit der Stichprobengrösse steht, ist die Wahl der Stichprobengrösse ein entscheidender Faktor (McNeish, 2014; Snijders & Bosker, 2012). Da bei hierarchischen Daten Beobachtungen aus der selben Gruppe ähnlicher zueinander sind als zu anderen Beobachtungen, verkleinert sich die effektive Stichprobengrösse (Raudenbush & Bryk, 2002). Werden beispielsweise aus 100 Schulklassen 10 Schulkinder ausgewählt, würde das zu einer effektiven Stichprobengrösse von 100 führen. Ein normales lineares Modell würde in diesem Fall aber mit einer Stichprobengrösse von 1000 arbeiten, wohingegen ein hierarchisches lineares Modell mit der effektiven Stichprobengrösse von 100 arbeitet. Folglich können diese beiden Methoden zu unterschiedlichen Standardfehlern und dementsprechend auch zu unterschiedlichen Prüfgrössen für den t Test gelangen.

Neben der Prüfgrösse ist auch die Anzahl an Freiheitsgrade relevant, um die Signifikanz eines Effekts mittels t Test zu überprüfen. Während bei normalen linearen Modellen die Anzahl Freiheitsgrade durch $N - p - 1$ bestimmt wird, wobei N die Stichprobengrösse und p die Anzahl Parameter im Modell sind, ist die Berechnung der Freiheitsgrade bei hierarchischen linearen Modellen nicht eindeutig geklärt und ein aktueller Forschungsgegenstand (McNeish, 2014; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Aufgrund dieser Gegebenheiten ist es also wichtig, dass die verwendete Analyseverfahren der Struktur des Datensatzes gerecht wird, um ungenaue Schätzungen des Standardfehlers

zu vermeiden. McNeish (2014) konnte in seiner Studie zeigen, dass mit zunehmender Intraklassen Korrelation die Schätzung des Standardfehlers bei normalen linearen Modellen ungenauer wird. Diese Erkenntnis entspricht den Berechnungen, die aus dem Artikel von Moerbeek et al. (2003) hervorgehen. In diesem Artikel differenzieren die Autoren zwischen zwei Studiendesigns, die in der Praxis häufig eingesetzt werden. Zum einen kann eine Intervention auf Level-1 durchgeführt werden und die zufällige Zuteilung zur Interventions- und Kontrollgruppe erfolgt auf Stufe der Schulkinder. Zum anderen kann man die Intervention auf Level-2 durchführen, wobei die Zuteilung auf Klassenstufe stattfindet. Moerbeek et al. (2003) konnten in ihrem Artikel zeigen, dass der geschätzte Standardfehler einer Intervention mittels normaler linearer Regression je nach Studiendesign unter- oder überschätzt wird.

(Hier: bei welchem design wird was erwartet)

Da Moerbeek et al. (2003) nur einen einzigen Datensatz simuliert haben, sollten ihre Ergebnisse noch in einer Simulationsstudie überprüft werden.

3.2 Studiendesign

In einer ersten Simulationsstudie wird folglich versucht, die Ergebnisse bezüglich der Effizienz der Schätzung der Regressionskoeffizienten von Mundform und Schults (2002) zu replizieren. Ebenfalls wird die Aussagen von Moerbeek et al. (2003) überprüft, dass die Verwendung von normalen linearen Modellen zur Analyse von hierarchischen Daten je nach Studiendesign zu Über- oder Unterschätzung der Standardfehler führt. Zusätzlich wird die Intraklassen Korrelation variiert, um zu überprüfen, dass diese Über- oder Unterschätzung der Standardfehler bei Zunahme der Intraklassen Korrelation steigt (McNeish, 2014). Gemäss den Ergebnissen von McNeish (2014) und Moerbeek et al. (2003) wird erwartet, dass hierarchische lineare Modelle auch bei Zunahme der Intraklassen Korrelation zu keiner Über- oder Unterschätzung des Standardfehlers führen.

In einer zweiten Simulationsstudie wird anhand realitätsnäheren simulierten Datensätzen versucht aufzuzeigen, wie sich die Power bezüglich einer Intervention zwischen den beiden

Analysemethode bei variierender Interklassen Korrelation und gleichbleibender Stichprobengrösse unterscheidet. Dabei wird erwartet, dass durch die Anwendung von Multilevel Analyse bei einer Randomisierung der Intervention auf Level-1 eine grössere Power erreicht wird, als durch die Analyse mittels normalen linearen Modellen. Diese Erwartung begründet sich dadurch, da gemäss Moerbeek et al. (2003) in dieser Situation die Analyse mit normalen linearen Modellen zu einer Überschätzung des Standardfehlers führt, dies führt wiederum zu einer erhöhten Fehler Typ 2 Rate und folglich zu einer tieferen Power. Bei der Randomisierung der Intervention auf Level-2 wird grundsätzlich eine tiefere Power erwartet, da in diesem Design der Effekt der Intervention nicht vom Effekt der Klassenzugehörigkeit getrennt werden kann (Donner et al., 1981; Hsieh, 1988). Da nach Moerbeek et al. (2003) eine Randomisierung auf Level-2 zu einer Unterschätzung des Standardfehlers führt, wird in diesem Fall erwartet, dass ein normales lineares Modell zu einer höheren aber immernoch tiefen Power führt.

3.3 Simulationsdesign

Um diese Annahmen zu überprüfen werden in beiden Simulationsstudien Daten basierend auf den selben zwei Designs von Moerbeek et al. (2003) generiert. Beim ersten Design handelt es sich um eine Intervention auf Stufe des Schulkindes und die Daten werden anhand der folgenden Gleichung generiert:

Design 1:

$$\begin{aligned}
\text{Level 1:} \quad y_{ji} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\
\text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + U_{0j} \\
\beta_1 &= \gamma_{10}
\end{aligned} \tag{24}$$

Dabei ist ϵ_{ij} das Residuum des i -ten Schulkindes aus der j -ten Klasse. Die Variable x_{ij} gibt an, ob sich das Schulkind i aus der Klasse j in der Interventions- oder Kontrollgruppe befindet. Der Koeffizient β_{0j} beschreibt den Achsenabschnitt, der wiederum durch den Gesamtmittelwert γ_{00} und der zufälligen Abweichung U_{0j} der Klasse j beschrieben wird.

Der Koeffizient β_{j1} wird nur durch die Gesamtsteigung γ_{1j} beschrieben. Folglich wird keine klassenspezifische Abweichung der Steigung in der Studie simuliert.

Das zweite Design berücksichtigt Interventionen auf Stufe der Klassen. Dabei werden die Daten nach der folgenden Gleichung generiert:

Design 2:

$$\begin{aligned} \text{Level 1:} \quad y_{ji} &= \beta_{0j} + \epsilon_{ij} \\ \text{Level 2:} \quad \beta_{0j} &= \gamma_{00} + \gamma_{01}z_j + U_{0j} \end{aligned} \tag{25}$$

Wieder beschreibt ϵ_{ij} das Residuum des i -ten Schulkindes aus der j -ten Klasse. Der Achsenabschnitt β_{0j} wird durch den Gesamtmittelwert γ_{00} , der Gesamtsteigung γ_{01} , der Variable z_j und der klassenspezifischen zufälligen Abweichung U_{0j} beschrieben. Die Variable z_j gibt an, zu welcher Interventionsgruppe die j -te Klasse gehört. Da in diesem Design die Intervention auf Stufe der Klasse durchgeführt wird, handelt es sich bei der Intervention um eine Level-2 Variable und wird typischerweise mit z_j und nicht mit x_j bezeichnet. Es wird in beiden Designs angenommen, dass die zufälligen Effekte ϵ_{ij} und U_{0j} voneinander unabhängig sind und einer Normalverteilung folgen. Ebenfalls wird angenommen, dass diese zufällige Effekte einen Mittelwert von Null und eine Varianz von σ_e^2 , resp. τ_0^2 aufweisen.

Gewisse Parameter werden in beiden Simulationsstudien nicht manipuliert. Für diese Parameter wurden die Werte von Moerbeek et al. (2003) und McNeish (2014) verwendet. Die Zuweisung zur Interventionsgruppe wurde für die Variablen x_{ij} und z_j durch die Werte -1 und 1 festgelegt. Dabei steht -1 für die Kontrollgruppe und 1 für die Interventionsgruppe. Der Gesamtmittelwert der Population wurde in beiden Designs auf $\gamma_{00} = 2.34$ festgelegt. Die Gesamtsteigung der Population wurde ebenfalls in beiden Designs beider Simulationsstudien mit $\gamma_{10} = 0.12$ resp. $\gamma_{01} = 0.12$ festgehalten. Für die Varianz des Residuums wurde ein Wert von $\sigma_e^2 = 1.72$ verwendet.

Neben den nicht manipulierten Parametern wurden die Intraklassen Korrelation und die Analysemethode variiert. Bei der IKK gab es insgesamt neun Bedingungen, die in drei Gruppen eingeteilt werden können. Die erste Gruppe entspricht einer IKK von 0.00

und beschreibt einen Datensatz bei der die Klassenzugehörigkeit keinen Einfluss hat. Die zweite Gruppe beinhaltet IKKs die typischerweise in der psychologischen Forschung angetroffen werden und reichen von 0.05 bis 0.25 mit einem Abstand von 0.05 zwischen den jeweiligen IKK Bedingungen (Hedges & Hedberg, 2007; Snijders & Bosker, 2012). Die dritte Gruppe beinhaltet Extremwerte der IKK von 0.30, 0.40 und 0.50. Wie aus der Formel (2) zu entnehmen ist, wird die IKK alleine durch die Varianz des Residuums und durch die Varianz der zufälligen Abweichung bestimmt. Da durch das Studiendesign die Varianz des Residuums und die theoretische IKK vorgegeben ist, lässt sich durch Umformen der Formel (2) die Varianz der zufälligen Abweichung τ_0^2 bestimmen. Anhand dieser Varianz wurden dann Datensätze generiert, die den theoretischen IKKs entsprechen. Für jede dieser Bedingung wurden in jedem Design jeweils 1000 Replikationen simuliert, das zu einer Gesamtanzahl von 18000 Replikationen führte. Dabei wurde jeder einzelne Replikation zum einen mit einer normalen linearen Regression und zum anderen mit einer hierarchischen linearen Regression analysiert.

3.4 Studie 1: Genauigkeit von Schätzparametern

Die Stichprobengröße wurde in der ersten Simulationsstudie über alle Bedingungen konstant gehalten. Dabei wurden wie bei McNeish 300 Klassen mit jeweils 50 Schulkindern mit den oben besprochenen Parametern simuliert (2014). In der Multilevel Literatur wird eine Mindestanzahl von 50 Gruppen empfohlen, damit die Schätzungen der Koeffizienten mittels hierarchischen linearen Modellen genau sind (Maas & Hox, 2005). Mit dieser grossen Stichprobengröße wird sichergestellt, dass Ergebnisse auf die Manipulation der Parameter und nicht auf eine ungenügende Stichprobengröße zurückzuführen sind.

Um die oben genannten Aussagen zu überprüfen, wurden zwei Kennwerte berechnet. Der erste Kennwert ist die relative Abweichung der geschätzten Regressionskoeffizienten $\hat{\gamma}$ von den Populationsmittelwerten γ . Die Stärke dieser Abweichung wird in Prozent ange-

geben (Hoogland & Boomsma, 1998) und nach folgender Formel berechnet:

$$\Delta\hat{\gamma} = \frac{\bar{\hat{\gamma}} - \gamma}{\gamma} \quad (26)$$

Dabei ist $\bar{\hat{\gamma}}$ der Mittelwert aller Regressionskoeffizienten aus einer Bedingung. Diese relative Abweichung wurde in beiden Designs für jede Analysemethode und in jeder IKK Bedingung für den Gesamtmittelwert γ_{00} als auch für die Gesamtsteigung γ_{10} resp. γ_{01} berechnet. Gemäss Hoogland und Boomsma (1998) gelten relative Abweichungen von kleiner als 5% als akzeptabel.

Der zweite Kennwert beschreibt die Genauigkeit der Schätzung des Standardfehlers (engl. *Standard Error*) und berechnet sich aus dem Verhältnis der Abweichung des mittleren Standardfehlers aus einer Bedingung von der Standardabweichung der Regressionskoeffizienten über alle 1000 Replikationen dieser Bedingung geteilt durch dieselbe Standardabweichung (Hoogland & Boomsma, 1998; McNeish, 2014). Die Formel zur Berechnung sieht wie folgt aus:

$$\Delta\widehat{SE}_{\hat{\gamma}} = \frac{\widehat{SE}_{\hat{\gamma}} - \widehat{SD}_{\hat{\gamma}}}{\widehat{SD}_{\hat{\gamma}}} \quad (27)$$

Der berechnete Wert beschreibt also wie bei der relativen Abweichung, um wie viel Prozent der geschätzte Standardfehler vom wahren Populationswert abweicht. Liegen Genauigkeitswerte über 0 gelten die Standardfehler als Überschätzt und liegen die Werte unter 0 werden Standardfehler unterschätzt. Hoogland und Boomsma (1998) bezeichnen jegliche Genauigkeitswerte, die um mehr als 0.10 von 0 abweichen als unakzeptabel. Die Genauigkeit des Standardfehlers wurde wieder in beiden Designs für jede Analysemethode und in jeder IKK Bedingung berechnet.

3.4.1 Ergebnisse Studie 1

Der erste Kennwert der untersucht wurde, war die relative Abweichung der Regressionskoeffizienten. Sowohl bei einer Intervention auf Level-1 als auch bei einer Intervention auf Level-2 schätzten LM als auch HLM die Regressionskoeffizienten des Achsenabschnittes

und der Steigung in allen IKK Bedingungen mit einer relativen Abweichung von kleiner als $|\Delta\hat{\gamma}| < .05$. Auch die Varianzen dieser relativen Abweichungen war in allen Bedingungen kleiner als $\sigma^2 < .01$. Diese Werte entsprechen in diesem Fall den Ergebnissen von Mundfrom und Schults (2002) und McNeish (2014).

Die Genauigkeit der Schätzung des Standardfehlers der beiden Regressionskoeffizienten $\hat{\gamma}_{00}$ und $\hat{\gamma}_{10}$ bzw. $\hat{\gamma}_{01}$ kann für jede der beiden Methoden aus der Tabelle 5 entnommen werden. Im ersten Simulationsdesign wurden die Standardfehler bei der Verwendung von LM mit einer Genauigkeit von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0511$ bis $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.8028$ geschätzt. Bei der Verwendung von HLM reichten die SE Genauigkeitswerte von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0366$ bis $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = .0201$. Bei einer IKK von .00 zeigte sich bei beiden Methoden eine akzeptable Schätzgenauigkeit des Standardfehlers mit einer Genauigkeit von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0511$ bei LM und $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0366$ bei HLM. Sobald aber die IKK anstieg, wurde die Schätzung bei LM zunehmend ungenauer. Aus Tabelle 5 ist zu entnehmen, dass bereits ab einer IKK von .10 die Anforderungen von Hoogland und Boomsma (1998) nicht mehr erfüllt sind bei Verwendung von LM. Mit Werten von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.4599$ bis $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.8028$ weist LM eine klare Unterschätzung des Standardfehlers auf, die zu einer erhöhten Fehler Typ 1 Rate führt. Vergleicht man die SE Genauigkeit von HLM aus dem ersten Simulationsdesign, erkennt man, dass in keiner IKK Bedingung den Grenzwert von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .10$ überschritten wird. Es wurde sogar ein noch strengeres Kriterium von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .05$ in allen IKK Bedingungen erfüllt. Bei Verwendung von HLM lag im ersten Simulationsdesign also weder eine Unter- noch eine Überschätzung vor. Dementsprechend entstehen bei der Verwendung von HLM keine erhöhten Fehler Typ 1 und Fehler Typ 2 Raten.

Beim zweiten Simulationsdesign weist das LM eine Schätzgenauigkeit von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0181$ bis $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.8014$ auf. Das HLM Schätzte den Standardfehler mit einer SE Genauigkeit von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0241$ bis $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = .0396$. Es zeigte sich bezüglich der Genauigkeit der Schätzung des Standardfehlers des Gesamtmittelwertes $\hat{\gamma}_{00}$ ein ähnliches Bild wie beim ersten Simulationsdesign. Wieder wiesen beide Methoden bei einer IKK von .00 eine genau Schätzung des Standardfehlers von $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0181$ bei LM und $\Delta\widehat{SE}_{\hat{\gamma}_{00}} = -.0027$ bei HLM auf. Diese Genauigkeit nahm bei erhöhter IKK und Verwendung von LM wieder

Tabelle 5: SE Genauigkeit beider Regressionskoeffizienten in beiden Simulationsdesigns und für jede Analyse­methode in allen IKK Bedingungen.

IKK	$\Delta\widehat{SE}_{\widehat{\gamma}_{00}}$				$\Delta\widehat{SE}_{\widehat{\gamma}_{10}}$ bzw. $\Delta\widehat{SE}_{\widehat{\gamma}_{10}}$			
	Design 1		Design 2		Design 1		Design 2	
	LM	HLM	LM	HLM	LM	HLM	LM	HLM
.00	-.0511 ^a	-.0366 ^b	-.0181 ^b	-.0027 ^b	-.0095 ^b	-.0098 ^b	.0056 ^b	.0213 ^b
.05	-.4599	.0009 ^b	-.4494	.0244 ^b	.0312 ^b	.0053 ^b	-.4729	-.0193 ^b
.10	-.5963	-.0191 ^b	-.5900	-.0061 ^b	.0542 ^a	.0001 ^b	-.5790	.0207 ^b
.15	-.6510	.0084 ^b	-.6620	-.0241 ^b	.0599 ^a	-.0226 ^b	-.6534	.0010 ^b
.20	-.6894	.0201 ^b	-.6832	.0396 ^b	.1327	.0135 ^b	-.6896	.0187 ^b
.25	-.7257	-.0029 ^b	-.7180	.0266 ^b	.1720	.0159 ^b	-.7324	-.0258 ^b
.30	-.7506	-.0120 ^b	-.7481	-.0016 ^b	.1565	-.0318 ^b	-.7472	.0016 ^b
.40	-.7783	.0057 ^b	-.7787	.0059 ^b	.3443	.0424 ^b	-.7758	.0191 ^b
.50	-.8028	-.0045 ^b	-.8014	.0041 ^b	.4186	.0044 ^b	-.7943	.0401 ^b

^a $|\Delta\widehat{SE}_{\widehat{\gamma}}| < .10$

^b $|\Delta\widehat{SE}_{\widehat{\gamma}}| < .05$

stark ab. In der vierten Spalte der Tabelle 5 kann man erkennen, dass bei einer IKK von .05 nur noch eine SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.4494$ erreicht wird und schlussendlich bei einer IKK von .50 bis zu einer SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.8014$ abnimmt. Auch diese Werte weisen bei der Verwendung eines LM auf eine Unterschätzung des Standardfehlers hin, die wiederum zu einer erhöhten Fehler Typ 1 Rate führt. Betrachtet man die fünfte Spalte aus Tabelle 5, erkennt man dass auch im zweiten Simulationsdesign die Standardfehler des Gesamtmittelwerts durch ein HLM genau geschätzt werden. In allen Bedingungen wurde der Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ nicht überschritten. Folglich kommt es auch im zweiten Simulationsdesign zu keiner Über- oder Unterschätzung des Standardfehlers des Gesamtmittelwertes $\widehat{\gamma}_{00}$ bei Verwendung von HLM.

Betrachtet man nun die SE Genauigkeiten der Gesamtsteigung $\widehat{\gamma}_{10}$ bzw. $\widehat{\gamma}_{10}$ aus Tabelle 5, reicht im ersten Simulationsdesign die Schätzgenauigkeit von LM von $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = -.0095$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = .4186$. Die Schätzgenauigkeit des HLM reichte von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.0318$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = .0424$. Beide Methoden weisen wieder bei einer IKK von .00 eine genaue Schätzung des Standardfehlers auf. Das LM verzeichnete eine abnehmende Genauigkeit

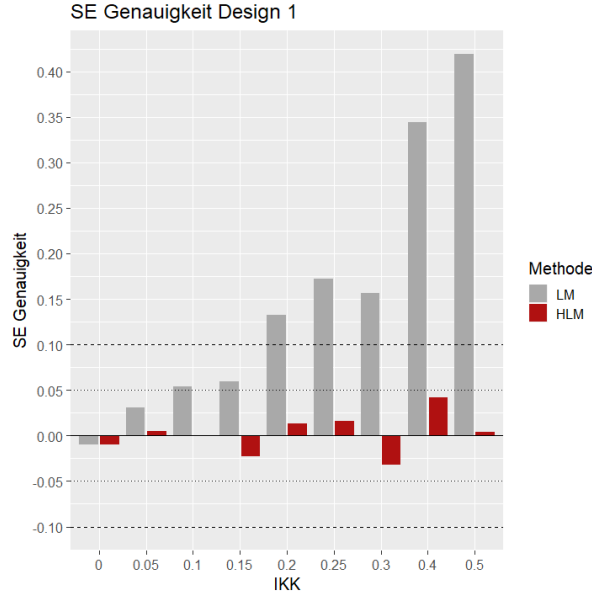


Abbildung 7: Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im ersten Simulationsdesign.

bei der Schätzung des Standardfehlers sobald die IKK zunahm, so dass der Standardfehler zunehmend überschätzt wurde. Allerdings überschritten die Schätzungen von LM bis zu einer IKK von .15 den Grenzwert von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .10$ nicht. Mit einer SE Genauigkeit von $\Delta\widehat{SE}_{\hat{\gamma}_{10}} = .0599$ lag die SE Genauigkeit von LM bei einer IKK von .15 immer noch im von Hoogland und Boosmam (1998) als akzeptabel definierten Bereich. Ab einer IKK von .20 wurde von LM im ersten Simulationsdesign der Standardfehler so stark überschätzt, dass er nicht mehr im akzeptablen Bereich lag. Diese Überschätzung des Standardfehlers mittels LM erreichte dann bei einer IKK von .50 mit $\Delta\widehat{SE}_{\hat{\gamma}_{10}} = .4186$ den höchsten Wert. Grundsätzlich wies das LM also eine zunehmende Überschätzung des Standardfehlers auf, das zu einer erhöhten Fehler Typ 2 Rate führt, die schlussendlich in einer niedrigeren Power resultiert. In Abbildung 7 wird dieser Zusammenhang noch einmal graphisch dargestellt. Wie in Abbildung 7 zu erkennen und aus Tabelle 5 zu entnehmen ist, bleiben die geschätzten Standardfehler des HLM immer hinnerhalb des akzeptablen Bereiches und erfüllen auch in allen Bedingungen das noch strengere Kriterium von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .05$. Die Schätzungen des Standardfehlers sind daher auch noch bei einer hohen IKK von .50 mit $\Delta\widehat{SE}_{\hat{\gamma}_{10}} = .0044$ im

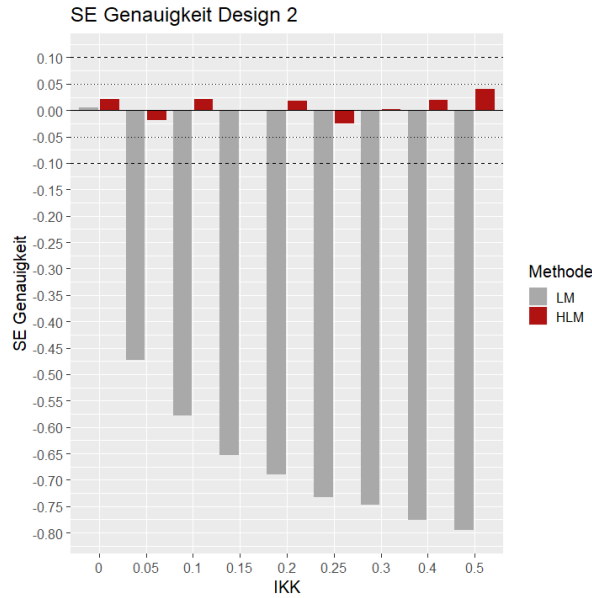


Abbildung 8: Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im zweiten Simulationsdesign.

akzeptablen Bereich. Folglich kommt es bei der Analyse mit einem HLM zu keiner Unter- oder Überschätzung des Standardfehlers und führt dementsprechend zu keiner erhöhten Fehler Typ 1 oder Fehler Typ 2 Rate, die in einer tieferen Power resultiert.

Beim zweiten Simulationsdesign wurden die Standardfehler der Gesamtsteigung durch das LM mit einer Genauigkeit von $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = .0056$ bis $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = -.7943$ geschätzt. Das HLM schätzte die Standardfehler mit einer Genauigkeit von $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = -.0258$ bis $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = .0401$. Beide Analysemethoden weisen wieder eine genaue Schätzung des Standardfehlers bei einer IKK von .00 auf. Betrachtet man die Entwicklung der Schätzung des Standardfehlers durch das LM, erkennt man, dass die Standardfehler wieder mit zunehmender IKK unterschätzt werden. Bei einer IKK von .05 wird mit $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = -.4729$ der Grenzwert für eine akzeptable Abweichung bereits überschritten. Diese Unterschätzung steigt mit zunehmender IKK weiter an, bis hin zu einer SE Genauigkeit von $\Delta \widehat{SE}_{\hat{\gamma}_{01}} = -.7943$ bei einer IKK von 0.50. In diesem Fall resultiert diese Unterschätzung des Standardfehlers wieder in einer erhöhten Fehler Typ 1 Rate. In Abbildung 8 sind wieder beide Verläufe

über die verschiedenen IKK Bedingungen abgebildet. Dabei lässt sich zum einen direkt die Unterschätzung des Standardfehlers durch das LM erkennen, aber auch dass das HLM wieder eine sehr genaue Geschätzung des Standardfehlers aufweist. Diese Werte sind in der letzten Spalte der Tabelle 5 abgebildet und zeigen, dass bei Verwendung von einem HLM die Standardfehler wieder in allen Bedingungen den Grenzwert von $|\Delta \widehat{SE}_{\hat{\gamma}}| > .10$ nicht überschreiten und sogar kleiner als das noch strengere Kriterium $|\Delta \widehat{SE}_{\hat{\gamma}}| > .05$ sind. Es entsteht folglich wieder weder eine Unter- noch Überschätzung des Standardfehlers bei der Verwendung eines HLM in diesem Simulationsdesign und dies resultiert demnach in keiner erhöhten Fehler Typ 1 oder Fehler Typ 2 Rate.

3.4.2 Diskussion Studie 1

Mit der ersten Studie dieser Arbeit wurde überprüft, wie sich die Schätzgenauigkeit von Regressionskoeffizienten und Standardfehlern zwischen LM und HLM bei der Analyse von hierarchischen Daten unterscheidet, wie sich diese Schätzungen bei variierender IKK verhalten und ob die Ebene der Intervention einen Einfluss auf die Schätzgenauigkeit hat.

Dabei konnte gezeigt werden, dass beide Methoden die Regressionskoeffizienten des Gesamtmittelwertes und der Gesamtsteigung über alle Bedingungen genau schätzten. Unterschiede in der Schätzgenauigkeit gab es erst bei der Schätzung der Standardfehler. So wurde der Standardfehler des Gesamtmittelwertes bei steigender IKK in beiden Simulationsdesigns bei Verwendung von LM zunehmend unterschätzt. Nur in der Bedingung ohne Einfluss der Gruppenzugehörigkeit schätzte das LM den Standardfehler genau. Das HLM schätzte hingegen in allen Bedingungen und in beiden Simulationsdesigns die Standardfehler des Gesamtmittelwertes genau. Bei der Schätzung des Standardfehlers der Gesamtsteigung gab es bei der Verwendung von LM einen Unterschied bezüglich des Simulationsdesigns. Im ersten Simulationsdesign wurde sobald die Gruppenzugehörigkeit einen Einfluss hatte, der Standardfehler mit steigender IKK zunehmend überschätzt. Beim zweiten Simulationsdesign wurde der Standardfehler bei steigender IKK zunehmend unterschätzt. Bei der Analyse mittels HLM wurden wieder in allen Bedingungen und in beiden Simulationsdesigns die

Standardfehler der Gesamtsteigung genau geschätzt.

Diese Ergebnisse bedeuten, dass es je nach Fokus der Forschungsfrage einen entscheidenden Einfluss hat, ob man ein LM oder ein HLM verwendet. Interessiert man sich nur für die Ausprägung der Regressionskoeffizienten spielt die Wahl der Analyseverfahren also keine grosse Rolle. Zu dieser Erkenntnis kamen auch schon Autoren aus früheren Studien (McNeish, 2014; Mundfrom & Schults, 2002; Osborne, 2000) und diese wurde mit der vorliegenden Simulationsstudie noch einmal bestätigt.

Eine weitere Schlussfolgerung, die aus diesen Ergebnissen gezogen werden kann, ist dass es sich empfiehlt hierarchische Daten mittels HLM zu analysieren um Ungenauigkeiten in der Schätzung des Standardfehlers und schlussendlich höhere Fehler Typ 1 oder Fehler Typ 2 Raten zu verhindern. Werden trotz dem Vorhandensein hierarchischer Strukturen mit LM gearbeitet, kann es je nach Studiendesign zu massiven Unter- oder Überschätzung der Standardfehler kommen und folglich zu verzerrten Studienergebnissen führen. Bei einer Unterschätzung des Standardfehlers laufen gemäss dieser Ergebnisse Forschende Gefahr, aufgrund der erhöhten Fehler Typ 1 Rate Effekte zu finden, die in Wahrheit gar nicht vorhanden sind. Diese Unterschätzung tritt vor allem dann auf, wenn Interventionen auf Level-2 durchgeführt wird und wurde ebenfalls bereits von Moerbeek et al. (2003) mathematisch aufgezeigt. Zu einer Überschätzung des Standardfehlers und folglich zu einer erhöhten Fehler Typ 2 Rate kommt es, wenn bei einer Intervention auf Level-1 die hierarchische Struktur missachtet und mit einem LM analysiert wird. Dabei spielt es gemäss Moerbeek et al. (2003) eine Rolle, ob von einer Interaktion zwischen Intervention und Gruppenzugehörigkeit ausgegangen wird oder nicht. Wird von keiner Interaktion ausgegangen, wie in unserer Simulationsstudie, werden Standardfehler von LM zunehmend überschätzt. Wird hingegen bei der Analyse eine Interaktion zwischen Intervention und Gruppenzugehörigkeit angenommen, kann es in Abhängigkeit der Varianzkomponenten und der Gruppengrössen zu einer Unter- oder Überschätzung des Standardfehlers führen (Moerbeek et al., 2003).

Bis jetzt wurden allerdings nur die Schätzgenauigkeit der Regressionskoeffizienten und der Standardfehler besprochen und nicht die Effektive Power dieser Beiden Analyseverfahren. Auch wenn aufgrund der Ergebnisse aus der ersten Simulationsstudie hervorgeht,

dass eine Verwendung von HLM bei der Analyse von hierarchischen Daten zu genaueren Schätzungen und weniger Verzerrung der Prüfgrösse führt als LM, wäre es interessant zu beobachten wie sich die Power dieser beiden Methoden über die verschiedenen IKK Bedingungen verändert.

3.5 Studie 2: Statistische Power von HLM

In der zweiten Studie wurde untersucht, wie zuverlässig LM und HLM einen Effekt einer Intervention finden, wenn dieser Effekt auch effektiv vorhanden ist. Die Fähigkeit einen Effekt zu finden, wenn er auch wirklich vorhanden ist wird Power genannt. Die Power steht in einem direkten Zusammenhang mit der Stichprobengrösse, so dass eine steigende Stichprobengrösse zu einer höheren Power führt (Snijders, 2005). Folglich würde eine so grosse Stichprobe aus Studie 1 wahrscheinlich zu keinen Unterschieden in der Power zwischen den beiden Methoden führen. Allerdings ist es in der Praxis oft nicht möglich, eine solche grosse Stichprobe von insgesamt 15000 Beobachtungen zu erheben. Daher wird in dieser zweiten Studie eine etwas reduzierte und praxisnähere Stichprobengrösse für die beiden Studiendesigns simuliert. Die Anzahl simulierter Klassen wurde folglich auf 70 und die Klassengrösse auf 12 reduziert. Diese Reduktion führte zu einer Stichprobengrösse von insgesamt 840 Beobachtungen.

Wie bereits in der Herleitung der Forschungsfrage beschrieben, wurde der Effekt des Treatments mit einem t Test überprüft. Die Prüfgrösse berechnet sich aus folgender Formel:

$$t = \frac{\hat{\gamma}}{\widehat{SE}_{\hat{\gamma}}} \quad (28)$$

Dabei ist $\hat{\gamma}$ der jeweilige geschätzte Regressionskoeffizienten und $\widehat{SE}_{\hat{\gamma}}$ der dazugehörige Standardfehler. Die Anzahl Freiheitsgrade wurde bei normalen linearen Modellen mittels der bekannten Formel $N + p - 1$ berechnet. Bei den hierarchischen linearen Modellen wurde die Satterthwaite Methode verwendet, um die Anzahl Freiheitsgrade zu bestimmen (1941). Die Satterthwaite Methode ist eine der in der Forschung diskutierten Methoden, die häufig

Tabelle 6: Statistische Power in beiden Simulationsdesigns für jede Analysemethode in allen IKK Bedingungen.

IKK	Statistische Power			
	Design 1		Design 2	
	LM	HLM	LM	HLM
.00	.76	.76	.74	.72
.05	.72	.75	.71	.53
.10	.73	.76	.64	.38
.15	.72	.78	.64	.34
.20	.67	.75	.56	.24
.25	.69	.78	.56	.21
.30	.62	.78	.57	.19
.40	.54	.77	.56	.13
.50	.44	.76	.55	.10

zur Berechnung der Freiheitsgrade von hierarchischen linearen Modellen verwendet wird (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Um nun die Power zu berechnen, wurde in jeder Bedingung und für jede Analysemethode die Anzahl an Tests, die auf einem Signifikanzniveau von 5% signifikant wurden, durch die Anzahl Replikationen pro Bedingung geteilt. Dies ergibt die prozentuale Häufigkeit, bei der die Analysemethode in der gegebenen IKK Bedingung einen signifikanten Effekt gefunden hat. Da im Simulationsdesign die Intervention in der Tat einen Effekt hat, entspricht diese prozentuale Häufigkeit der Power dieser Analysemethode.

3.5.1 Ergebnisse Studie 2

Die Ergebnisse der zweiten Simulationsstudie werden in Tabelle 6 abgebildet. Es ist zu beachten, dass nur die Power für die Gesamtsteigung $\hat{\gamma}_{10}$ bzw. $\hat{\gamma}_{10}$ berechnet wurde, weil diese Regressionskoeffizienten auch den Effekt der Intervention abbilden. Im ersten Simulationsdesign in dem die Intervention auf Level-1 durchgeführt wurde, reicht die Power des LM von .44 bis .76. Dabei ist zu beachten, dass bei einer IKK von .00 die höchste Power von .76 erreicht wird und mit zunehmender IKK abnimmt. In Abbildung 9 sind die einzelnen

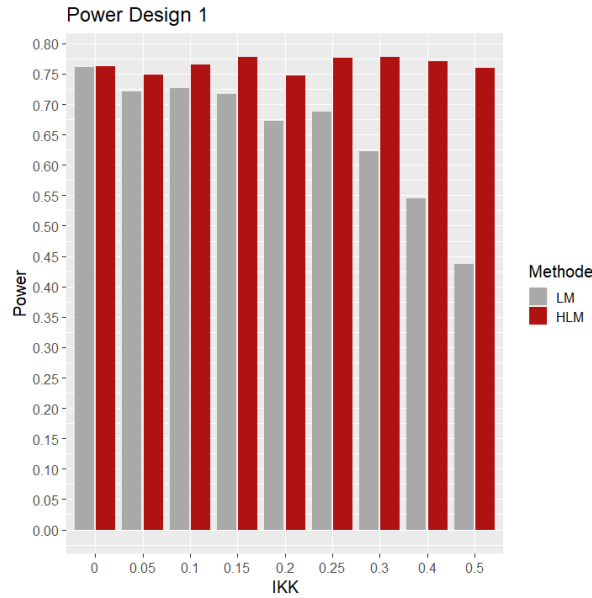


Abbildung 9: Statistische Power von ML und HLM in den verschiedenen IKK Bedingungen bei einer Intervention auf Level-1.

Power-Werte jeder IKK Bedingung und für jede Analysemethode abgebildet. Dabei lässt sich erkennen, dass die Power von LM bis zu einer IKK von .10 noch einigermaßen mit der Power von HLM mithalten kann. Sobald aber die IKK grösser als .10 ist lässt sich eine kontinuierliche Abnahme der Power beobachten. Betrachtet man die Schätzgenauigkeit des Standardfehlers von LM an dieser Stelle, erkennt man, dass ab einer IKK von .15 mit $\Delta\widehat{SE}_{\hat{\gamma}_{10}} = 0.1167$ den Grenzwert von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .10$ (Hoogland & Boomsma, 1998) überschritten wird und folglich die Schätzung des Standardfehlers als nicht mehr akzeptabel gilt. Dies führt zu einer zu hohen Fehler Typ 1 Rate, die schlussendlich in dieser tieferen Power resultiert. Die Power unter Verwendung des HLM reicht von .75 bis .78 und bleibt gemäss Abbildung 9 über alle Bedingungen relativ konstant. Die Schätzgenauigkeit der Standardfehler mittels HLM überschreitet in keiner IKK Bedingung den Grenzwert von $|\Delta\widehat{SE}_{\hat{\gamma}}| > .10$. Bei einer Analyse mit HLM kommt es in diesem Fall also weder zu einer Über- noch einer Unterschätzung des Standardfehlers.

Im zweiten Simulationsdesign bei dem die Intervention auf Level-2 durchgeführt wurde, reicht die Power des LM von .55 bis .74 und nimmt wieder mit steigender IKK ab. Die Power

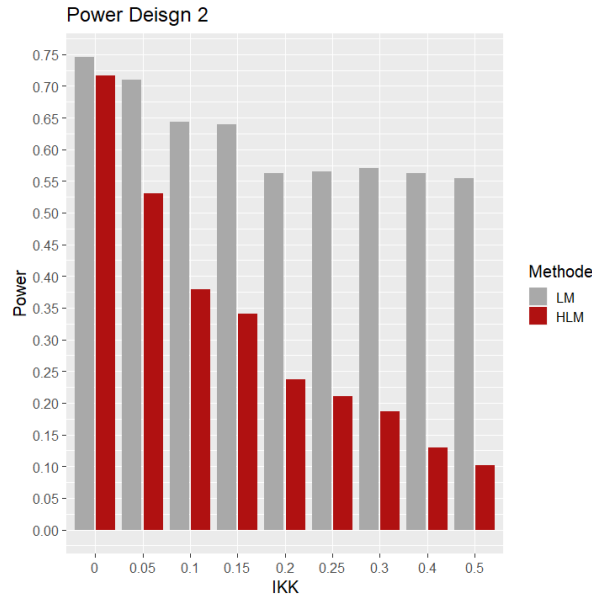


Abbildung 10: Statistische Power von ML und HLM in den verschiedenen IKK Bedingungen bei einer Intervention auf Level-2.

des HLM reicht im zweiten Simulationsdesign von .10 bis .72 und verzeichnete ebenfalls eine Abnahme der Power mit zunehmender IKK. In Abbildung 10 ist dieser Verlauf noch einmal Visualisiert. Dabei kann man erkennen, dass die Abnahme der Power bei einer zunehmenden IKK bei einer Analyse mittels HLM stärker als bei einer Analyse mittels LM ist. Bei einer IKK von .50 erreichten HLM nur noch eine Power von .10 wobei LM eine Power von .55 erreichten. Die Schätzgenauigkeiten des Standardfehlers werden ähnlich wie in Studie 1 von LM ab einer IKK von .005 stark unterschätzt und überschreiten in allen weiteren IKK Bedingungen den Grenzwert von $|\Delta \widehat{SE}_{\hat{\gamma}}| > .10$. Diese Unterschätzung der Standardfehler durch LM führt zu dieser vermeintlich besseren Power, da die Fehler Typ 1 Rate durch die Unterschätzung erhöht ist. Die Schätzung der Standardfehler bleibt bei der Verwendung von HLM im akzeptablen Bereich und überschreiten den Grenzwert in keiner IKK Bedingung.

3.5.2 Diskussion Studie 2

Die zweite Simulationsstudie untersuchte, wie genau sich die Power bezüglich eines Effekts einer Intervention zwischen den beiden Analysemethoden unterscheidet und vor allem wie sich die Power über die verschiedenen IKK Bedingungen verändert. Wie in der ersten Studie wurde auch in der zweiten Studie zwischen zwei Simulationsdesigns unterschieden bei denen die Intervention zum einen auf Level-1 und zum anderen auf Level-2 durchgeführt wurde.

Im ersten Studiendesign bestätigte sich die Vermutung aus der vorherigen Studie, dass bei einer Intervention auf Level-1 ohne Interaktion zwischen Gruppenzugehörigkeit und Intervention die Power mit zunehmender IKK abnimmt, wenn bei der Analyse LM verwendet wird. Wird an Stelle von LM ein HLM verwendet bleibt die Power über alle IKK Bedingungen Konstant. Im zweiten Simulationsdesign zeigte sich allerdings eine Abnahme der Power bei beiden Analysemethoden bei zunehmender IKK. Diese Abnahme war bei der Verwendung von LM geringer als bei der Verwendung von HLM.

Interessiert sich nun eine Forschungsfrage für den Effekt einer Intervention, lässt sich aus diesen Ergebnissen zwei Schlussfolgerungen ziehen. Zum einen ist es Ratsam, wenn immer möglich die Intervention auf Level-1 durchzuführen, da die Power im zweiten Simulationsdesign bei beiden Analysemethoden mit steigender IKK stark abnahm. Diese Erkenntnis stimmt mit den Aussagen aus der Literatur überein, dass eine Intervention auf Level-2 es erschwert, den Effekt der Intervention von dem Effekt der Gruppenzugehörigkeit zu trennen, da in diesem Studiendesign keine Beobachtungen aus Kontroll- und Interventionsgruppe in der selben Level-2 Einheit vorhanden sind (Cleary et al., 2012; Moerbeek et al., 2000).

Zum anderen sollte grundsätzlich ein HLM zur Analyse von hierarchischen Daten verwendet werden, da sich die Power bei einer Intervention auf Level-1 auch bei hoher IKK nicht verschlechtert. Ist es den Forschenden jedoch nicht möglich eine Intervention auf Level-1 durchzuführen, könnte man aus den Ergebnissen dieser zweiten Studie schliessen, dass ein LM im zweiten Studiendesign eine bessere Power als das HLM vorzuweisen. Da in

der ersten Studie gezeigt wurde, dass in dieser Situation der Standardfehler vom LM stark unterschätzt wird, kann diese schwache Abnahme der Power auf die erhöhte Fehler Typ 1 Rate und die Inflation der signifikanten Tests zurückgeführt werden. Da es bei der Analyse mittels HLM zu keiner Verzerrung der Schätzung des Standardfehlers kommt, kann der Power von HLM mehr Vertrauen geschenkt werden. Mit dieser Information könnten nun Forschende die Analysemethode als Ursache für diese tiefe Power ausschliessen und nach anderen Gründe suchen, die zu dieser tiefen Power führen (z.B. Studiendesign oder Stichprobengrösse).

In der folgenden abschliessenden Diskussion dieser Arbeit werden die bereits besprochenen Ergebnisse noch einmal kurz aufgegriffen und es werden mögliche weitere Forschungsfragen, sowie Limitationen dieser Simulationsstudie besprochen.

3.6 Abschliessende Diskussion und Shiny App

Mit diesen beiden Studien wurde das Ziel verfolgt, die Unterschiede zwischen der Schätzgenauigkeit von Regressionskoeffizienten und deren Standardfehlern sowie die daraus resultierende Power von LM und HLM zu untersuchen. Auch wenn die Regressionskoeffizienten von beiden Methoden genau geschätzt wurden, konnten die Studien zeigen, dass HLM auch bei hoher Abhängigkeit der Gruppenzugehörigkeit genaue Schätzungen der Standardfehler liefern, wohingegen ein LM bei steigender IKK eine immer stärkere Verzerrung der Schätzung aufweist. Dementsprechend führten Analysen mittels HLM auch zu keiner Verzerrung der Power bei der Testung von Interventionseffekten.

Die aus dieser Simulationsstudie resultierende Erkenntnis, dass HLM einen klaren Vorteil gegenüber LM bei der Analyse von hierarchischen Daten vorweisen, ist keine Neuheit und wurde in einigen Studien bereits gefunden (McNeish, 2014; Moerbeek et al., 2003; Mundfrom & Schults, 2002; Osborne, 2000). Dennoch konnte diese Studie, die in Moerbeek et al. (2003) verwendeten Studiendesigns, in eine Simulationsstudie integrieren, um den Einfluss dieser beiden Interventionsformen auf die Schätzgenauigkeit von LM und HLM zu untersuchen. Damit können klare Empfehlungen für die Forschung mit hierarchischen

Daten abgeleitet werden. Beispielsweise sollte wenn immer möglich eine Intervention auf Level-1 durchgeführt werden, um Effekte der Gruppenzugehörigkeit von Effekten der Intervention trennen zu können.

In manchen Fällen ist es allerdings nicht möglich die Intervention auf Level-1 durchzuführen und benötigen daher eine Intervention auf Level-2, um eine Kontamination der Interventionsgruppe zu vermeiden. Wie man in der zweiten Studie beobachten konnte, verringerte sich die Power des HLM bei einer Intervention auf Level-2 bei einer Zunahme der IKK enorm. Weiterführende Simulationsstudien könnten nun untersuchen, welche Möglichkeiten es gibt diese Power zu verbessern. Da in der aktuellen Simulationsstudie nur die IKK variiert wurde und bekanntlicher Weise die Power mit steigender Stichprobengrösse zunimmt, könnten weitere Studien untersuchen, wie gross eine Stichprobe sein müsste, damit ein HLM auch bei einer Intervention auf Level-2 eine akzeptable Power liefert.

Die aktuelle Studie unterlag mehreren Limitationen. Es wurde nur ein sehr einfaches Modell zur Simulation der Daten verwendet. Dabei gab es keinen *Cross-Level* Interaktion zwischen der Intervention und der Gruppenzugehörigkeit. Da in realen Datensätzen solche *Cross-Level* Interaktion wahrscheinlich vorhanden sind, können Ergebnisse aus dieser Simulationsstudie nicht auf solche Situationen angewandt werden. Daher wäre es interessant, wenn weitere Simulationsstudien diese *Cross-Level* Interaktionen und diese Aufteilung in zwei Interventionsdesigns untersuchen könnten. Eine weitere Limitation dieser Simulationsstudie ist, dass nur die IKK variiert wurde. So erhält man zwar einen Einblick, wie sich die Messmethode über verschiedene Einflussstärken der Gruppenzugehörigkeit verhält, es gibt aber keine Einsicht darüber, wie gross Stichproben gewählt werden sollten, um genaue Schätzungen und eine akzeptable Power zu erhalten. Da es in der Forschung oft auch eine Kostenfrage ist, wie gross Stichproben sind, könnte durch eine weitere Simulationsstudie die Effektstärke der Intervention und verschieden grosse Stichproben berücksichtigt klare Empfehlungen erarbeitet werden, mit welcher Mindeststichprobe eine gute Power erreicht wird. Als abschliessende Limitation kann die Wahl des Grenzwertes für die Schätzgenauigkeit des Standardfehlers genannt werden. Dieser Grenzwert von $|\Delta \widehat{SE}_{\hat{\gamma}}| > .10$ wurde von Hoogland

und Boosma (1998) festgelegt. Wie man aber in Abbildung 9 erkennen kann, sinkt die Power von LM bereits bei einer IKK von .05. Bei Dieser IKK wird bereits die Schätzgenauigkeit von $|\widehat{\Delta SE_{\hat{\gamma}}}| > .05$ überschritten. Weitere Simulationsstudien könnten untersuchen, ob dieser Grenzwert von $|\widehat{\Delta SE_{\hat{\gamma}}}| > .10$ gerechtfertigt ist oder ob ein strengerer Grenzwert nötig ist, um genaue von ungenauen Analysemethoden zu trennen.

Die aktuelle Studie konnte also zeigen, dass es sich bei hierarchischen Daten und geeignetem Interventionsdesign lohnt ein HLM anstelle eines LM zu verwenden, da diese Modelle die nötigen Parameter genau schätzen und zu einer konstanten Power führen. Diese Ergebnisse und eine kleine Einführung zu HLM können in einer interaktiven Form in der im Laufe dieser Arbeit programmierten Shiny App abgerufen werden. Dabei wurde die Shiny App ebenfalls in zwei Teile unterteilt, wobei der erste Teil sich der Theorie widmet und Nutzer selber einen hierarchischen Datensatz generieren, den sie dann mit LM oder HLM analysieren können. Für jede Analysemethode werden Regressionsgeraden, Residuenplots, Q-Q Plots und die Zusammenfassung aus R abgebildet, damit die Nutzer einen direkten Vergleich dieser Methoden erhalten. Im zweiten Teil der Shiny App wird kurz die Forschungsfrage und das Studiendesign vorgestellt und anschliessend können die Nutzer für jede Teilstudie die Ergebnisse jeder Bedingung anzeigen lassen und erhalten dazu in einer Infobox die dazugehörige Erklärungen und Implikationen. Die im Anhang aufgeführten Screenshots geben eine weitere Übersicht über den hier kurz vorgestellten Aufbau und Inhalt der Shiny App.

4 Literaturverzeichnis

- Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67 (1), 1–48. doi: 10.18637/jss.v067.i01
- Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2019). shiny: Web Application Framework for R [Software-Handbuch]. Zugriff auf <https://CRAN.R-project.org/package=shiny> (R package version 1.3.2)
- Cleary, P. D., Gross, C. P., Zaslavsky, A. M. & Taplin, S. H. (2012, 05). Multilevel Interventions: Study Design and Analysis Issues. *JNCI Monographs*, 2012 (44), 49-55. Zugriff auf <https://doi.org/10.1093/jncimonographs/lgs010> doi: 10.1093/jncimonographs/lgs010
- Donner, A., Birkett, N. & Buck, C. (1981, 12). Randomization by clusters: Sample size requirements and analysis. *American Journal of Epidemiology*, 114 (6), 906-914. doi: 10.1093/oxfordjournals.aje.a113261
- Dorman, J. P. (2008). The effect of clustering on statistical tests: an illustration using classroom environment data. *Educational Psychology*, 28 (5), 583–595.
- Field, A., Miles, J. & Field, Z. (2013). *Discovering statistics using r* (Reprinted Aufl.). Los Angeles: Sage.
- Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. United Kingdom: Cambridge University Press. (Includes bibliographical references (pages 575-600) and indexes)
- Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29 (1), 60-87. doi: 10.3102/0162373707299706
- Hoogland, J. J. & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26 (3), 329-367. doi: 10.1177/0049124198026003003
- Hsieh, F. Y. (1988). Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine*, 7 (11), 1195-1201. doi: 10.1002/sim.4780071113
- Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82 (13), 1–26. doi: 10.18637/jss.v082.i13
- Maas, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1 (3), 86-92. doi: 10.1027/1614-2241.1.3.86

- 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.
- Moerbeek, M., van Breukelen, G. J. & Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, 56 (4), 341 - 350. doi: [https://doi.org/10.1016/S0895-4356\(03\)00007-6](https://doi.org/10.1016/S0895-4356(03)00007-6)
- Moerbeek, M., van Breukelen, G. J. P. & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, 25 (3), 271-284. doi: 10.3102/10769986025003271
- Mundfrom, D. J. & Schults, M. (2002). A monte carlo simulation comparing parameter estimates from multiple linear regression and hierarchical linear modeling. *Multiple Regression Viewpoints*, 28, 18–21.
- Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation*, 7 (1), 1.
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48 (1), 85-112. doi: <https://doi.org/10.1016/j.jsp.2009.09.002>
- R Core Team. (2019). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf <https://www.R-project.org/>
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Bd. 1). Sage.
- Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38 (2), 337–341.
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6 (5), 309 - 316. doi: <https://doi.org/10.1007/BF02288586>
- Scherbaum, C. A. & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, 12 (2), 347-367. doi: 10.1177/1094428107308906
- Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In *Encyclopedia of statistics in behavioral science*. American Cancer Society. doi: 10.1002/0470013192.bsa492
- Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. Aufl.). Los Angeles: SAGE.
- Twisk, J. W. R. (2006). *Applied multilevel analysis: A practical guide for medical researchers*. Cambridge University Press. doi: 10.1017/CBO9780511610806

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.

5 Anhang

A Shiny App Screenshots

B R Code