Universität Zürich[UZH]

Schätzgenauigkeit von Standardfehlern und deren Einfluss auf die Zuverlässigkeit von Tests 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

5. Mai 2020

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kurzfassung

Diese Arbeit kombiniert eine theoretische Einführung in die hierarchischen linearen Modelle (HLM) und eine Simulationsstudie über deren Schätzgenauigkeit. In der theoretischen Einführung wird erklärt, was hierarchische Daten sind und was für Konsequenzen eine ungenügende Berücksichtigung dieser Datenstrukturen mit sich bringen. Dabei wird konkret auf den problematischen Einsatz von normalen linearen Modellen (LM) zur Analyse von solchen Daten eingegangen und wie ein HLM diese Probleme umgehen kann. In einer Simulationsstudie wurde untersucht, wie sich die Ausprägung der Intraklassen Korrelation (IKK) und das Level der Intervention auf die Schätzgenauigkeit von LMs und HLMs auswirkt. Dabei zeigte sich, dass beide Modelle in allen Bedingungen die Regressionskoeffizienten genau schätzten. Der Standardfehler wurde allerdings von LMs mit zunehmender IKK ungenauer geschätzt, wobei die Schätzung von HLMs genau blieb. Diese ungenaue Schätzung des Standardfehlers durch ein LM führte in Abhängigkeit des Studiendesign zu einer Reduktion der Power oder einer Erhöhung der Fehler Typ 1 Rate. So wurde bei einer Intervention auf Level-1 der Standardfehler durch ein LM bei steigender IKK zunehmend überschätzt. Dies resultierte in einer abnehmenden Power des LMs. Der Standardfehler einer Level-2 Variable wurden von LMs hingegen zunehmend unterschätzt, was schlussendlich in einer erhöhten Fehler Typ 1 Rate resultierte. Dabei litt ein HLM unter keiner dieser Limitationen und wies eine adäqute Fehler Typ 1 Rate sowie eine zuverlässige Power auf. Diese Studie konnte zum einen Ergebnisse aus früheren Studien replizieren und zum anderen aufzeigen, dass ein HLM bei der Analyse von hierarchischen Daten zu genaueren Ergebnissen führt.

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 betrachtet. 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 Schulkindern aus jeweils 100 Schulklassen durchzuführen, als mit einem Schulkind 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 Schulklasse 10 Schulkinder für eine Studie ausgewählt, ist es sehr wahrscheinlich, dass Schulkinder 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 Schulkindes sich auf seine schulische Leistung auswirkt. Aber auch Level-2 Einheiten können sich gegenseitig beeinflussen. Das Klima der Schulklasse 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 Schulklasse) 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 Verfahren 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 der Multilevel Analyse ist das hierarchische lineare Modell, das 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 hierarchische lineare Modell 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:

$$Y_{ij} = \mu + U_j + R_{ij} \tag{1}$$

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:

$$\rho_I = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \tag{2}$$

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) eingesetzt[1]:

$$\rho_I = \frac{12.33}{12.33 + 44} = 0.22 \tag{3}$$

Die daraus resultierende IKK von $\rho_I = 0.22$ weist darauf hin, dass 22% 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.

[1]Die 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:

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

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 Achsenabschnitt (*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.

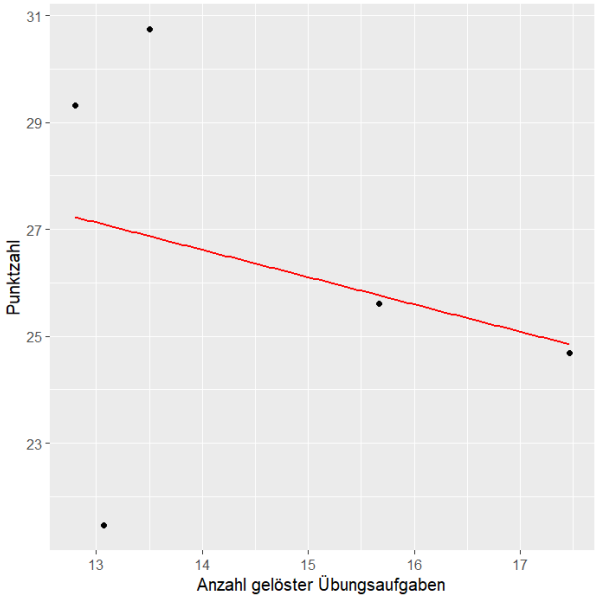


Abbildung 1: Zusammenhang zwischen der durchschnittlich gelösten Anzahl an Übungsaufgaben und der durschnittlich erreichten Punktzahl pro Klasse

2.3.2 Disaggregation

Die zweite Möglichkeit um hierarchsiche 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 voneinander 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

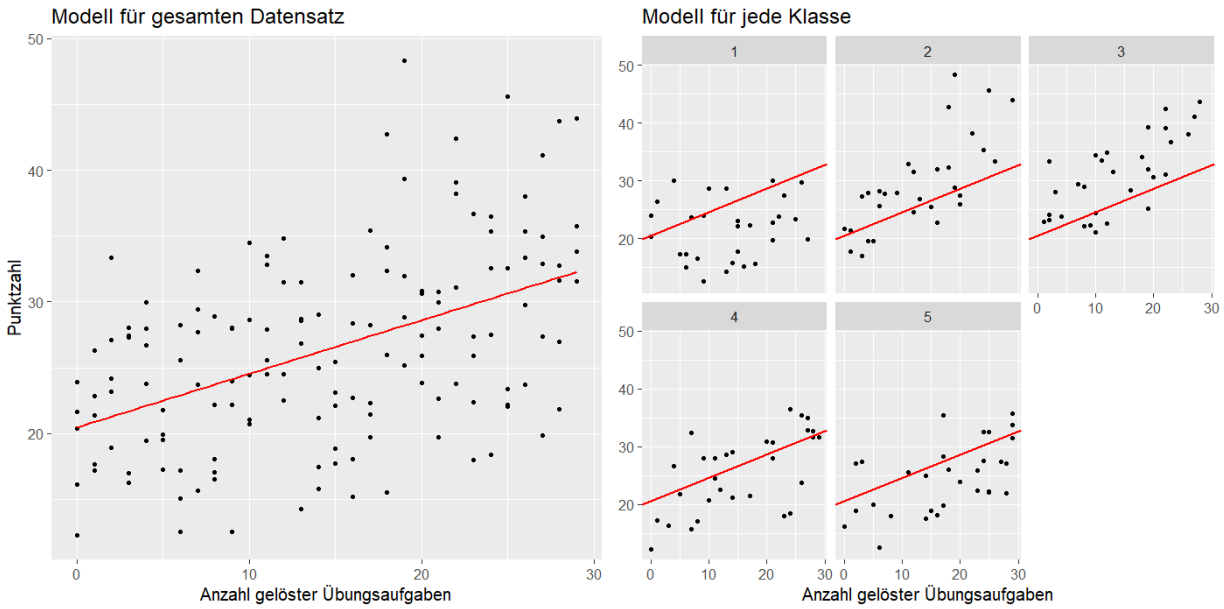


Abbildung 2: Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und erreichte Punktzahl mittels Disaggregation und Anwendung dieses Zusammenhangs auf jede der fünf Klassen

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 besprochenen 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:

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

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:

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

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:

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

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:

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

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}:

$$\beta_{01} = 19.9 - 4.01 = 15.89 \tag{9}$$

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 sowohl Unterschiede in der abhängige Variablen zwischen den Klassen als auch Unterschiede im Einfluss der unabhängigen Variablen auf die abhängige Variable zwischen den Klassen berücksichtigt. Der Regressionskoeffizient β_1 aus dem *Random Intercept* Modell (7) ist nun spezifisch für jede Klasse j und wird durch die zufällige Abweichung U_{1j} erweitert. Dies führt zum folgenden Modell in der hierarchischen Notation:

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

Durch Einsetzen der Gleichungen aus Level-2 in die Gleichung aus Level-1, erhaltet man die folgende Gleichung des *Random Intercept and Slope* Modells:

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

Dabei wurde die Gleichung so umgeformt, dass der erste Teil $\gamma_{00} + \gamma_{10} x_{ij}$ die jeweiligen geschätzten Parametern enthält, die sich nicht zwischen den Klassen verändern und wird 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 zufällige Abweichungen und das Residuum enthält. Der Term $U_{1j} x_{ij}$ beschreibt den gruppenspezifischen Effekt, der die unabhängige Variable x_{ij} auf die abhängige Variable hat. Bezogen auf unser Bei-

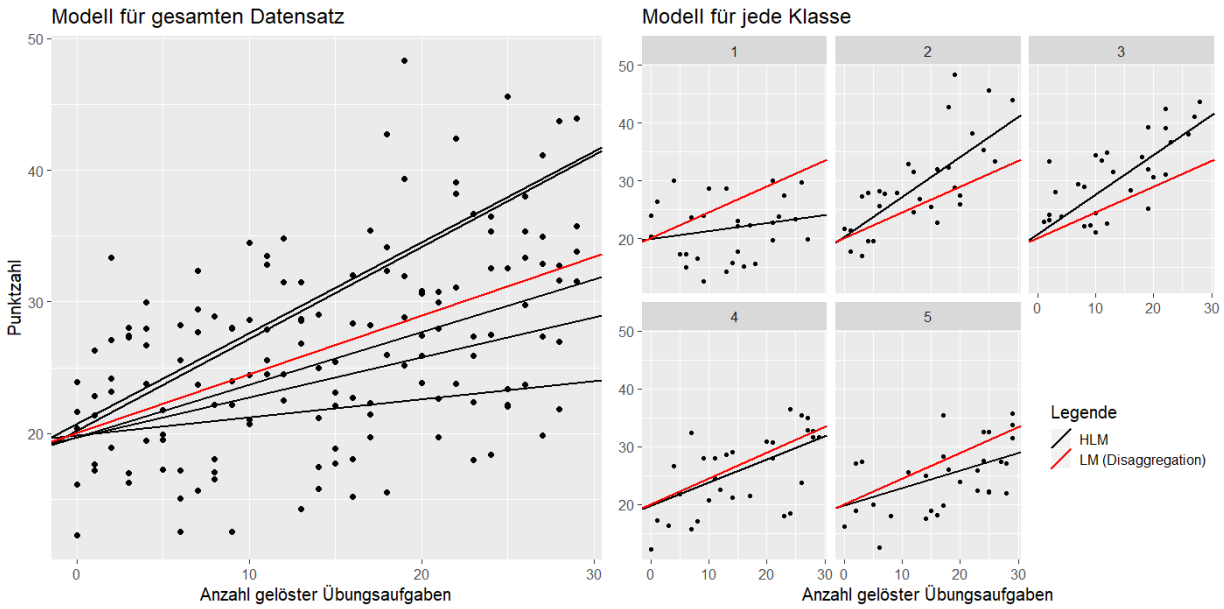


Abbildung 4: Zusammenhang zwischen der Anzahl gelöster Übungsaufgaben und der erreichten Punktzahl unter Berücksichtigung der Klassenzugehörigkeit und deren Effekt auf den Einfluss der Anzahl gelösten Übungsaufgaben. Links: Geraden von LM und HLM im gesamten Datensatz. Rechts: Geraden von LM und HLM auf jede Klasse aufgeteilt.

spiel 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 der Effekt der Klassenzugehörigkeit auf die Einfluss der Anzahl gelösten Übungsaufgaben einfach zu erkennen. Für gewisse Klassen nimmt die Zufallsvariable U_{1j} eine positiven und für andere einen negativen Wert ein. Dies spiegelt sich wiederum in klassenspezifischen Steigungen β_{1j} die höher oder tiefer als die Gesamtsteigung γ_{10} ist. Wie bereits im *Random Intercept* Modell werden auch im *Random Intercept and Slope* Modell nur die Varianzen der zufälligen Abweichungen geschätzt und die Effektive zufällige Abweichung U_{1j} muss ebenfalls berechnet werden. Diese Berechnung verläuft ebenfalls nach dem Prinzip von *shrinkage to the mean*, wobei die Gruppensteigung leicht in Richtung der Gesamtsteigung verschoben wird.

Betrachten wir die Klasse 1 unseres Beispiels. Mit der Schätzung der Varianz der zufälligen Abweichung durch das *Random Intercept and Slope* Modell lässt sich für diese Klasse die zufällige Abweichung von der Gesamtsteigung von $U_{11} = -0.24$ berechnen.

Setzt man nun diesen Wert zusammen mit der Gesamtsteigung $\gamma_{10} = 0.45$ in die Gleichung für β_{1j} aus (10) ein, erhält man die klassenspezifische Steigung β_{11}:

$$\beta_{11} = 0.45 - 0.24 = 0.2 \tag{12}$$

Für jede weitere gelöste Übungsaufgabe eines Schulkindes i aus der Klasse 1 steigt also die erwartete Punktzahl im Mittel um 0.2 Punkte an. Betrachten wir nun die Klasse 2 aus unserem Beispiel, schätzt unser Modell eine positive zufällige Abweichung $U_{12} = 0.2$ von der Gesamtsteigung. Wird dieser Wert wieder in die Gleichung für β_{1j} aus (10) eingesetzt, erhalten wir einen klassenspezifische Steigung von $\beta_{12} = 0.65$. Diese Steigung ist nun grösser als die Gesamtsteigung $\gamma_{10} = 0.45$. Folglich nimmt die erreichte Punktzahl bei einer weiteren gelösten Übungsaufgabe eines Schulkindes i aus der Klasse 1 im Mittel um 0.65 zu. Schulkinder aus Klasse 2 verzeichnen also einen höheren Anstieg in der erreichten Punktzahl bei jeder weiteren gelösten Übungsaufgabe als Schulkinder aus Klasse 1.

In unserem Beispiel ist es so, dass je höher der klassenspezifische Achsenabschnitt ist, desto höher ist auch die klassenspezifische Steigung. Man spricht hier auch von ei-

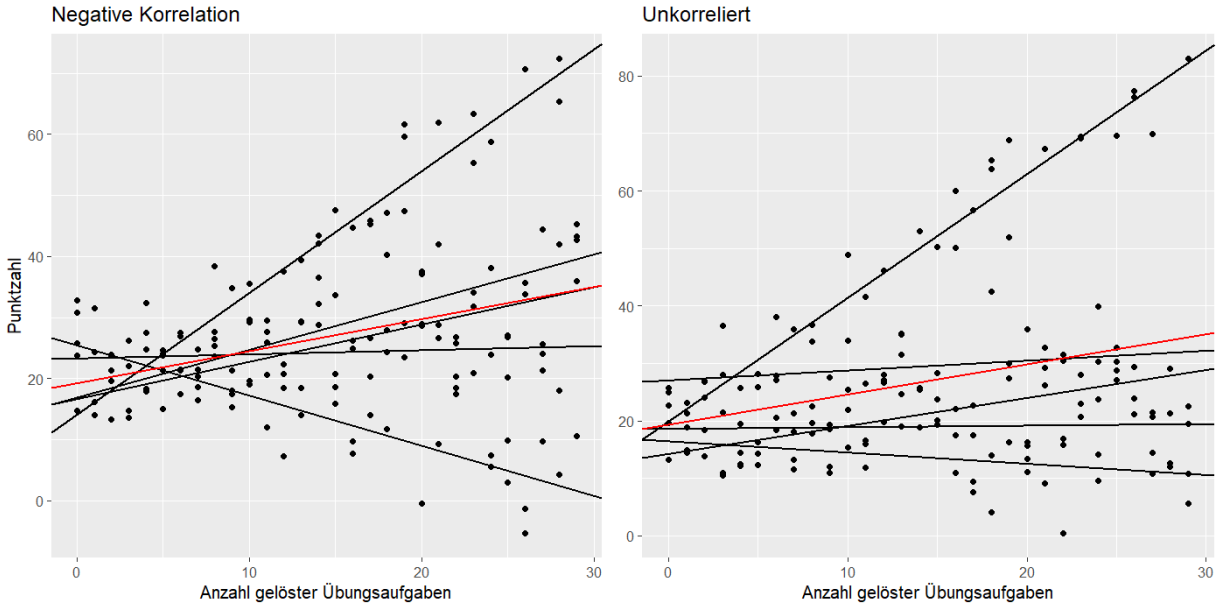


Abbildung 5: Darstellung einer negativen und nicht vorhandenen Korrelationen zwischen Achsenabschnitt und Steigung

ner 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. Auf der linken Seite der Abbildung ist eine negative Korrelation abgebildet. 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. Dieser Zusammenhang ist auf der rechten Seite der Abbildung zu sehen.

In diesem Abschnitt wurde das *Random Intercept and Slope* Modell besprochen und es wurde gezeigt, dass ein Modell, bei dem sowohl der Achsenabschnitt als auch die Steigung zwischen den Klassen variiert, am besten zu unserem Datensatz passt. Im nächsten Abschnitt wird nun besprochen, wie diese Modelle durch die Hinzunahme von weiteren Prädiktoren in ihrer Fähigkeit, Variabilität in der abhängigen Variablen zu erklären, verbessert werden können.

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 HLMs 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 HLMs zu beschreiben, können die Varianzen all dieser Komponenten verringert werden. Um die Varianz des Residuums zu verringern können, wie bei der linearen Regression, weitere Level-1 Variablen in das Modell aufgenommen werden. Da die Varianzen der zufälligen Abweichungen nicht durch Unterschiede innerhalb der Gruppen sondern zwischen den Gruppen entstehen, können diese nicht durch das Hinzufügen 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}:

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01} z_j + U_{0j} \\ & \beta_{1j} = \gamma_{10} + \gamma_{11} z_j + U_{1j} \end{aligned} \tag{13}$$

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. Substituiert man die beiden Koeffizienten in die Level-1 Gleichung erhält man folgende Darstellung dieses HLMs:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + U_{0j} + (\gamma_{10} + \gamma_{11} z_j + U_{1j}) x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + U_{0j} + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + U_{1j} x_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + U_{0j} + U_{1j} x_{ij} + \epsilon_{ij} \end{aligned} \tag{14}$$

Auch wenn es in der hierarchischen Notation einfacher zu erkennen ist, welche Varianz genau durch die Hinzunahme dieser Level-2 Variable verringert wird, erkennt man in dieser Notation einen weiteren wichtigen Zusammenhang. Der Term $\gamma_{11} z_j x_{ij}$ 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 beschrieben werden. Die Anzahl der Fenster erklären also einen Anteil der Variabilität in der Steigung zwischen den Klassen bezüglich des Effekts der Anzahl gelösten Übungsaufgaben auf die erreichte Punktzahl.

In den letzten Abschnitten wurden zwei verschiedene HLMs besprochen und das Prinzip ihrer Anwendung diskutiert. Es ist dabei zu berücksichtigen, dass hier nur die Grundlagen zu den HLMs behandelt wurden. Für eine weitere Vertiefung dieses Themas wird auf die gängige Literatur zur Multilevel Analyse mittels HLMs verwiesen (Gelman & Hill, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Twisk, 2006). Im anschliessenden Abschnitt wird die praktische Anwendung von HLMs in R mit dem Paket `lme4` (Bates et al., 2015) besprochen.

2.5 Anwendung von HLMs 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 Schritt für Schritt ein HLM erstellt und der `summary()`-Output wird interpretiert. Hierbei wird aufgezeigt, wie HLMs aufgebaut und miteinander verglichen werden, um das geeignetste Modell zu identifizieren. Am Ende dieses Abschnittes werden Effektstärkemasse 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. Beide dieser Methoden erzeugen eine *Log-Likelihood* aus der die *Deviance* berechnet werden kann. Die *Deviance* gibt die relative Passung eines Modells zu den Daten an und kann folglich genutzt werden, um zwei Modelle miteinander zu vergleichen (Snijders & Bosker, 2012). Allerdings unterscheidet sich die Bedeutung der *Deviance* zwischen ML und REML. Unter ML gibt die *Deviance* an, wie gut die geschätzten Regressionskoeffizienten und die geschätzte Varianz zu den Daten passt. Wohingegen die *Deviance* unter REML nur angibt, wie gut die geschätzten Varianzen zu den Daten passen. Dieses Argument ist also vor allem dann wichtig, wenn man Modelle vergleichen möchte. Dabei sollte ML für den Vergleich von festen Effekten und REML für den Vergleich von zufälligen Effekten verwendet werden (Peugh, 2010). Die drei Punkte in der Funktion stehen für weitere Argumente, die in `lmer()` eingegeben werden können. Diese Argumente werden aber hier nicht weiter behandelt, da sie für eine erweiterte Anwendung verwendet werden und den Rahmen dieser Arbeit sprengen würden[2]

Der folgende Code zeigt exemplarische Darstellung, wie eine Formel in `lmer()` eingegeben wird:

```
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 gibt es viele Möglichkeiten, wie zufällige Effekte und Gruppen in das Modell

[2]Mit dem Befehl `?lmer` kann die gesamte Dokumentation von `lmer()` angezeigt werden.

aufgenommen werden können. Beispielsweise können mehrere Gruppen definiert werden für die der Achsenabschnitt variiert oder dass sich nur die Steigung zwischen den Gruppen unterscheidet. Da es in diesem Abschnitt nur um die grundlegende Anwendung von `lmer()` geht, werden diese Möglichkeiten hier nicht behandelt. Eine Ausführliche Beschreibung der Anwendung von `lmer()` und dem gesamten Paket von `lme4` kann in Bates et al. (2015) nachgeschlagen werden. In Tabelle 4 ist nun die Syntax für die drei Möglichkeiten aufgeführt, die auch in diesem Abschnitt besprochen sind.

Tabelle 4: Auswahl der möglichen Syntax für `lmer()` nach Bates et al. (2015). Der Buchstabe g beschreibt hier den Gruppenfaktor und x die weiteren zufälligen Effekte.

Formel	Alternative	Bedeutung
(1 | g)	-	Zufälliger Achsenabschnitt
(x | g)	(1 + x | g)	Korrelierter Achsenabschnitt und Steigung
(x || g)	(1 | g) + (0 + x | g)	Unkorrelierter Achsenabschnitt und Steigung

2.5.2 Aufbau und Vergleich von HLMs

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.

Als Erstes berechnen wir die IKK, um den Einfluss der Klassenzugehörigkeit zu quantifizieren. 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:

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

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

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

print(VarCorr(m_leer), comp = "Variance")

##  Groups   Name        Variance
##  klasse   (Intercept) 12.3
##  Residual             44.0
```

Aus diesem Modell können nun die Varianzen für den Achsenabschnitt $\tau^2 = 12.33$ und das Residuum $\sigma^2 = 44$ ausgelesen und in die Formel (2) eingesetzt werden. Daraus ergibt sich die bereits in Abschnitt 2.2 berechnete IKK von $\rho_I = 0.22$. Folglich werden 22% der Variabilität in der erreichten Punktzahl durch die Klassenzugehörigkeit erklärt.

Wir können nun damit anfangen unser leeres Modell mit weitere Prädiktoren aufzubauen, um mehr Variabilität in der abhängigen Variablen zu erklären. In unserem Beispiel fügen wir die Anzahl gelöster 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:

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

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 = TRUE)
```

Dabei ist zu beachten, dass der Term innerhalb der Klammern der Vorgabe aus Tabelle 4 entspricht, um ein *Random Intercept* Modell zu schätzen. Wir haben nun also ein leeres Modell und ein *Random Intercept* Modell mit einem Prädiktor. Möchten wir jetzt überprüfen, ob unser grösseres Modell besser zu den Daten passt als das leere Modell und mehr Varianz aufklärt, kann man wie beim Vergleich von normalen LMs mit dem Befehl `anova()` die Modelle miteinander Vergleichen.

```
anova(m_leer, m_intercept, method = "LRT")

## refitting model(s) with ML (instead of REML)

## Data: beispiel_data
## Models:
## m_leer: punktzahl ~ (1 | klasse)
## m_intercept: punktzahl ~ uebung + (1 | klasse)
##             Df  AIC  BIC logLik deviance Chisq Chi Df
## m_leer       3 1009 1018   -502     1003
## m_intercept  4  952  964   -472      944  59.1      1
##             Pr(>Chisq)
## m_leer
## m_intercept    1.5e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Da sich zwischen diesen beiden Modellen nur feste Effekte unterscheiden, sie aber mittels REML geschätzt wurden, schätzt `anova()` die beiden Modelle nochmals neu mit der ML Methode. Ausserdem ist zu Beachten, dass beim Vergleich von zwei HLMs nicht wie bei normalen LMs ein F Test sondern ein *Likelihood Ratio Test* verwendet wird (Peugh, 2010; Snijders & Bosker, 2012). Dieser Test verwendet die Differenz der *Deviance* als Prüfgrösse

einer χ^2-Verteilung. Bei einem signifikanten Ergebnis passt das komplexere Modell mit mehr Parametern besser zu den Daten und bei einem nicht-signifikaten Test wird das einfachere Modell beibehalten, da es bereits genügend gut zu den Daten passt.

Im oberen Output ist also zu entnehmen, dass unser *Random Intercept* Modell auf einem Signifikanzniveau von 5% besser zu den Daten passt als das leere Modell. Wir möchten nun zusätzlich ein *Random Intercept and Slope* Modell schätzen. Entsprechend des Abschnittes 2.4.2 wird der Gleichung von β_{1j} eine zufällige Abweichung von der mittleren Steigung hinzugefügt:

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_1 \cdot \text{uebungen}_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + U_{0j} \\ & \beta_{j1} = \gamma_{10} + U_{1j} \end{aligned} \tag{17}$$

Diese zufällige Abweichung der Steigung wird nun auch in der Formel in R hinzugefügt:

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

Dabei ist zu beachten, dass hier keine Korrelation zwischen Achsenabschnitt und Steigung angenommen wird, da man erst einmal einfach überprüfen möchte, ob überhaupt ein Modell mit einer variierenden Steigung besser zu den Daten passt als eines ohne. Werden nun zwei Modelle miteinander Verglichen, die sich nur in ihren zufälligen Effekten unterscheiden, können beide Modelle mittels REML geschätzt werden (Peugh, 2010). Um sicherzustellen, dass `anova()` die Modelle nicht nochmals neu mit ML schätzt, kann ihr das Argument `refit = FALSE` hinzugefügt werden. Im folgenden Output ist das Ergebnis des *Likelihood Ratio Tests* aufgeführt:

```
anova(m_intercept, m_uncorr, method = "LRT", refit = FALSE)

## Data: beispiel_data
## Models:
```

```
## m_intercept: punktzahl ~ uebung + (1 | klasse)
## m_uncorr: punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
##             Df AIC BIC logLik deviance Chisq Chi Df
## m_intercept  4 953 965   -473      945
## m_uncorr     5 945 960   -467      935  10.7      1
##             Pr(>Chisq)
## m_intercept
## m_uncorr        0.0011 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der *Likelihood Ratio Test* kommt also zum Ergebnis, dass ein Modell, das eine zufällige Abweichung der Steigung zulässt, bei einem Signifikanzniveau von 5% besser zu den Daten passt. Vergleicht man die beiden Abbildungen 3 und 4 scheint dieses Ergebnis des *Likelihood Ratio Test* glaubhaft zu sein. Wir möchten nun auch noch herausfinden, ob ein Modell, das eine Korrelation zwischen Achsenabschnitt und Steigung annimmt, besser ist als unser unkorreliertes Modell. Dazu wird gemäss der Tabelle 4 ein weiteres Modell geschätzt und mit unserem unkorrelierten Modell verglichen.

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

anova(m_uncorr, m_corr, method = "LRT", refit = FALSE)

## Data: beispiel_data
## Models:
## m_uncorr: punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
## m_corr: punktzahl ~ uebung + (uebung | klasse)
##          Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m_uncorr  5 945 960   -467      935
## m_corr    6 946 964   -467      934  1.19      1       0.27
```

Das Ergebnis des *Likelihood Ratio Tests* zeigt, dass bei einem Signifikanzniveau von 5% das Modell mit einer Korrelation nicht besser zu den Daten passt. Folglich wird in diesem Fall ein *Random Intercept and Slope* Modell beibehalten, das von keiner Korrelation zwischen Achsenabschnitt und Steigung ausgeht.

Bis jetzt haben wir uns nur mit Level-1 Variablen oder der zufälligen Abweichung beschäftigt. Möchten wir dem Modell die Anzahl Fenster im Klassenzimmer als Level-2 Variable hinzufügen, kann man diese Variable mit einer *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, in dem eine Level-2 Variable noch zwischen Level-3 Einheiten variieren könnte. In der Formel würde das Hinzufügen einer Level-2 Variable ohne *Cross-Level* Interaktion wie folgt aussehen:

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

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. Möchte man die Level-2 Variable als *Cross-Level* Interaktion hinzufügen, kann für das + einfach ein $*$ eingesetzt werden. In diesem Modell wird aber die Level-2 Variable einfach als fester Effekt hinzugefügt.

```
m_fix_lvl2 <- lmer(punktzahl ~ uebung + fenster + (uebung || klasse),
        data = beispiel_data, REML = TRUE)
```

Nachdem das neue Modell mit der Level-2 Variable geschätzt wurde, kann es wieder mit dem sparsameren Modell `m_uncorr` verglichen werden. Da sich zwischen den beiden Modellen wieder nur die festen Effekte unterscheiden, werden von `anova()` die Modelle wieder neu mit ML geschätzt. Im folgenden Output ist zu erkennen, dass das komplexere Modell mit einer Level-2 Variable auf einem Signifikanzniveau von 5% nicht besser zu den Daten passt als unser vorheriges Modell. Die Level-2 Variable Anzahl Fenster kann also aus dem Modell ausgeschlossen werden.

```r
anova(m_uncorr, m_fix_lvl2, method = "LRT")

## refitting model(s) with ML (instead of REML)

## Data: beispiel_data
## Models:
## m_uncorr: punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
## m_fix_lvl2: punktzahl ~ uebung + fenster + ((1 | klasse) + (0 + uebung |
## m_fix_lvl2:     klasse))
##            Df AIC BIC logLik deviance Chisq Chi Df
## m_uncorr    5 944 959   -467      934
## m_fix_lvl2  6 942 960   -465      930  3.37      1
##            Pr(>Chisq)
## m_uncorr
## m_fix_lvl2      0.066 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.5.3 Interpretation eines HLMs

Im letzten Abschnitt wurde ein HLM von Grund auf aufgebaut, mit dem Versucht wurde den Zusammenhang zwischen der Anzahl gelösten Übungsaufgaben und der erreichten Punktzahl in einer Mathematikprüfung unter Berücksichtigung der Klassenzugehörigkeit zu erklären. Dabei ergab sich nach mehreren Modellvergleichen, dass ein *Random Intercept and Slope* Modell mit unkorrelierten Achsenabschnitten und Steigungen unter Berücksichtigung der Modellkomplexität am besten zu unseren Daten passt.

Nun geht es darum dieses Modell zu interpretieren. Als erstes kann man mit dem `summary()` Befehl eine Zusammenfassung ausgeben, die man für eine erste Interpretation unseres Modells verwenden kann. Der `summary()`-Output des HLMs ähnelt dem eines LMs. Dabei werden aber die zufälligen und die festen Effekte getrennt aufgeführt.

```r
summary(m_uncorr)

## Linear mixed model fit by REML ['lmerMod']
## Formula:
```

```
## punktzahl ~ uebung + ((1 | klasse) + (0 + uebung | klasse))
##    Data: beispiel_data
##
## REML criterion at convergence: 935
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1103 -0.8016  0.0542  0.7212  2.8707
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  klasse   (Intercept)  0.8560  0.925
##  klasse.1 uebung       0.0646  0.254
##  Residual             26.8972  5.186
## Number of obs: 150, groups:  klasse, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   20.036      0.933   21.47
## uebung         0.446      0.124    3.59
##
## Correlation of Fixed Effects:
##        (Intr)
## uebung -0.312
```

Bei den zufälligen Effekten werden für den Achsenabschnitt als auch für die Steigung die geschätzte Varianzen und die Standardabweichungen. Da in unserem Modell keine Korrelationen geschätzt wurden, wird diese hier nicht angezeigt. 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 eines LMs und lässt sich auch dementsprechend interpretieren. Es werden hier aber keine p-Werte angezeigt, da man sich in der Forschung noch nicht einig darüber ist, wie genau die Anzahl der Freiheitsgrade geschätzt werden soll (Peugh, 2010; Snijders & Bosker, 2012)[3].

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

Bei der Interpretation ist es von Vorteil die Standardabweichung zu verwenden, da diese nicht wie die Varianz quadriert ist und somit der Masseinheit, in unserem Beispiel der Punktzahl, entspricht. Eine Interpretation des Achsenabschnittes unter Berücksichtigung der zufälligen Effekte würde dann wie folgt lauten:

> Die erreichte Punktzahl in der Mathematikprüfung von 95% der Schulkinder liegt zwischen $20.04 \pm 1.96 \cdot 0.93$ Punkten, wenn keine einzige Übungsaufgabe gelöst wurde.

Und für die Steigung:

> Die Zunahme der Punktzahl liegt bei 95% der Schulkinder zwischen $0.45 \pm 1.96 \cdot 0.25$ Punkten pro gelöste Übungsaufgabe.

Dabei ist 0.93 die Standardabweichung des Achsenabschnittes und 0.25 die Standardabweichung der Steigung. Die Zahl 1.96 ist eine Approximation für das 97.5te Perzentil einer Standardnormalverteilung und wird hier verwendet, um den Bereich zu berechnen, der 95% aller Beobachtungen enthält.

2.5.4 Effektstärkemasse von HLMs

Effektstärkemasse gibt es für normale lineare Regressionsmodelle einige, wie zum Beispiel das Bestimmtheitsmass R^2 oder das Cohen's d. Bei HLMs werden Effektstärkemasse in globale und lokale Effektstärkemasse getrennt (Peugh, 2010). Dabei gibt es bei den globalen Effektstärkemasse mehrere Formen zur Berechnung eines Bestimmtheitsmasses R^2. Allerdings herrscht hier keinen Konsens über die bestmögliche Berechnungsform (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). In diesem Abschnitt wird die Berechnung des Bestimmtheitsmasses R^2 von Snijders und Bosker (2012) behandelt. Bei dieser Berechnungsform wird davon ausgegangen, dass bei HLMs auf mehreren Levels die proportional erklärte Varianz berechnet werden kann. Ausgehend von unserem Beispiel mit zwei Levels 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 mit folgender Formel:

$$\begin{aligned} R_1^2 &= 1 - \frac{\text{Gesamtvarianz des Modells mit Prädiktoren}}{\text{Gesamtvarianz des leeren Modells}} \\ &= 1 - \frac{\sigma^2 + \tau_0^2}{\sigma_{leer}^2 + \tau_{0leer}^2} \end{aligned} \tag{19}$$

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 Effektstärkemass nur für *Random Intercept* Modelle berechnet werden. Die Berechnung von R_1^2 für *Random Intercept and Slope* Modellen ist viel aufwändiger, da in diesem Fall auch noch die Varianz der Steigung miteinbezogen werden muss. Gemäss Snijders und Bosker liegen aber die R_1^2 Werte von einem *Random Intercept* Modell normalerweise sehr nahe an den R_1^2 Werten eines *Random Intercept and Slope* Modells, wenn in beiden Modellen die selben festen Effekten vorhanden sind (Snijders & Bosker, 2012). Daher reicht es in dem meisten Fällen völlig aus, die einfache Berechnung von R_1^2 für *Random Intercept* Modelle zu verwenden.

Für unser Beispiel würde das nun bedeuten, dass die Varianzen aus dem *Random Intercept* Modell verwendet werden können, um ein R_1^2 zu berechnen, dass nahe am R_1^2 des *Random Intercept and Slope* Modells `m_uncorr` liegt. Das *Random Intercept* Modell wurde bereits in Abschnitt 2.5.2 geschätzt und ist hier zur Übersicht noch einmal aufgeführt:

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

print(VarCorr(m_intercept), comp = "Variance")

##  Groups   Name        Variance
##  klasse   (Intercept) 15.5
##  Residual             29.3
```

Die Varianzen des leeren Modells können im Abschnitt 2.5.2 entnommen und mit den Varianzen für das *Random Intercept* Modell aus dem oberen Output in die Formel eingesetzt werden:

$$R_1^2 = 1 - \frac{29.29 + 15.46}{44 + 12.33} = 0.21 \tag{20}$$

Folglich kann man nun die Aussage treffen, dass das *Random Intercept* Modell `m_intercept` 21% der Varianz in der erreichten Punktzahl von Schulkindern erklärt. Gemäss Snijders und Bosker (2012) entspricht dieser Wert also ungefähr der Varianzaufklärung unseres *Random Intercept and Slope* Modells `m_uncorr`.

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). Wie bereits in früheren Abschnitten erläutert, wird bei HLMs die Varianz des Residuums sowohl auch die Varianzen des Achsenabschnittes und der Steigung geschätzt. In Abschnitt 2.4.3 wurde angesprochen, dass durch die Hinzunahme von weitere Prädiktoren diese Varianzen gezielt verringert werden können. Durch die Berechnung der PRV, kann nun die Reduktion in der Varianz bei allen drei geschätzten Varianzen des HLMs quantifiziert werden. Die Berechnung erfolgt für alle Varianzen mit der selben Formel:

$$PRV = \frac{Var_0 - Var_1}{Var_0} \tag{21}$$

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 des Residuums 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 = TRUE)

print(VarCorr(m_leer), comp = "Variance")

##  Groups   Name        Variance
##  klasse   (Intercept) 12.3
##  Residual             44.0

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

print(VarCorr(m_intercept), comp = "Variance")

##  Groups   Name        Variance
##  klasse   (Intercept) 15.5
##  Residual             29.3
```

Die Varianz der Residuen des leeren Modells beträgt $\sigma_0^2 = 44$ und die Varianz der Residuen des *Random Intercept* Modells $\sigma_1^2 = 29.29$. Fügen wir diese beiden Werte in die Formel (21) ein erhalten wir die proportionale Reduktion der Varianz des Residuums:

$$PRV_R = \frac{44 - 29.29}{44} = 0.33 \tag{22}$$

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 33% 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.

Wir haben nun zwei Effektstärkemasse von HLMs besprochen. Zum einen das globale Effektstärkemass R_1^2, das dazu genutzt werden kann, um die gesamte erklärte Varianz eines Modells zu beschreiben. Dieses Effektstärkemass ist vor allem dann nützlich, wenn man

aussagen darüber treffen möchte, wie gut ein Modell die vorhandenen Daten beschreibt. Möchte man aber wissen, wie viel Varianz durch einzelne bestimmte Prädiktoren erklärt werden, kann man mit dem lokalen Effektstärkemass der PRV für jeden Prädiktor die aufgeklärte Varianz berechnen. Zusätzlich ist die PRV auch nützlich, um die Varianzen zu identifizieren, die am meisten durch den gewählten Prädiktor reduziert werden.

Im letzten Abschnitt wurde anhand eines einfachen Beispiels aufgezeigt, wie man eine ein HLM in R aufbaut und analysiert. 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 Simulationsstudie 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 normalen linearen Modellen (LM) und hierarchischen linearen Modellen (HLM) gelegt. 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 ist, zum einen das Konzept der HLM zu verstehen und zum anderen werden die Ergebnisse dieser Simulationsstudie interaktiv für die Nutzer dargestellt.

3.1 Herleitung der Forschungsfrage

In der Einleitung wurde bereits erwähnt, dass in der psychologischen Forschung hierarchische Datenstrukturen keine Seltenheit sind und in Tabelle 1 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 Daten-

struktur oder den Möglichkeiten von HLMs nicht bewusst sind (McNeish, 2014). Dies kann dazu führen, dass diese hierarchischen Daten mittels normalen LMs anstelle von HLMs analysiert werden. Simulationsstudien haben aber gezeigt, dass dies nicht zwingend ein Problem darstellen muss (McNeish, 2014; Mundfrom & Schults, 2002). In diesen Simulationsstudien wurden systematisch Datensätze mit unterschiedlichen IKKs generiert, um den Einfluss der Gruppenzugehörigkeit zu variieren. Basierend auf diesen Datensätzen wurden von LMs und HLMs die Regressionskoeffizienten geschätzt und miteinander verglichen. Diese Vergleiche ergaben, dass sowohl HLMs als auch normale LMs die Grösse des Regressionskoeffizienten relativ genau schätzen konnten, unabhängig davon wie gross der Einfluss der Gruppenzugehörigkeit auf die abhängige Variable war (McNeish, 2014; Mundfrom & Schults, 2002). Diese Ergebnisse deuten also darauf hin, dass die Schätzung der Grösse des Effekts einer Intervention oder des Achsenabschnitts bei beiden Methoden relativ nahe am Populationsmittelwert liegt.

Allerdings interessiert man sich in der Forschung oft nicht nur für die Grösse des Effekts, sondern auch ob dieser einen signifikanten Einfluss hat. Um zu überprüfen, ob eine Intervention auch effektiv einen Einfluss auf die Ausprägung der abhängige Variable hat, werden üblicherweise t Tests durchgeführt (Snijders & Bosker, 2012). Die Prüfgrösse des t Tests wird über das Verhältnis zwischen dem geschätzten Regressionskoeffizienten und dessen Standardfehlers bestimmt. Wird beispielsweise ein Standardfehler zu klein geschätzt, steigt die Prüfgrösse an und die Rate mit der die Nullhypothese abgelehnt wird, nimmt zu. Wird der Standardfehler zu gross geschätzt, verkleinert sich die Prüfgrösse und die Rate mit 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 genaue Schätzung des Standardfehlers umso wichtiger, da dieser massgeblich zur Power des Tests beiträgt. Da der Standardfehler in einem direkten Zusammenhang mit der Stichprobengrösse steht und grössere Stichproben zu kleineren Standardfehlern führen, ist die

Wahl der Stichprobengrösse ein entscheidender Faktor (James et al., 2013; Snijders & Bosker, 2012). Bei hierarchischen Daten sind Beobachtungen aus derselben Gruppe ähnlicher zueinander als zu anderen Beobachtungen aus anderen Gruppen. Folglich verkleinert sich die effektive Stichprobengrösse (Raudenbush & Bryk, 2002). Wie stark sich die effektive Stichprobengrösse in einem hierarchischen Datensatz mit zwei Levels verkleinert, hängt zum einen von der Gruppengrösse und der Grösse des Einflusses der Gruppenzugehörigkeit ab. Dies kann mit dem sogenannten *Design Effect* quantifiziert werden und wird mit der folgenden Formel berechnet (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012):

$$Design\ Effect = 1 + (n - 1)\rho_I \tag{23}$$

Dabei wird angenommen, dass die Gruppengrösse n bei allen Gruppen gleich gross und ρ_I die IKK ist. Mit zunehmender IKK und Gruppengrösse erhöt sich folglich der *Design Effect*. Zum Beispiel würde das bei einer IKK von $\rho_I = 0.5$ und einer Stichprobe von $N = 100$ Schulklassen mit jeweils $n = 10$ Schulkindern pro Schulklasse zu einem *Design Effect* von 5.5 führen. Obwohl hier Beobachtungen von 1000 Schulkindern erfasst wurden, erhaltet man aufgrund der hierarchsichen Datenstruktur nur eine effektive Stichprobe von $M_{\text{effektiv}} = 182$:

$$M_{\text{effektiv}} = \frac{Nn}{Design\ Effect} = \frac{1000}{5.5} = 182 \tag{24}$$

Ein normales LM würde in diesem Fall mit einer Stichprobengrösse von 1000 arbeiten, um den Standardfehler zu berechnen und nicht mit der effektiven Stichprobengrösse. Bei HLMs wird diese hierarchische Struktur bei der Berechnung des Standardfehlers berücksichtigt (Snijders & Bosker, 1993). Da ein HLM die hierarchische Struktur berücksichtigt und ein LM nicht, müssten diese beide Modelle zu unterschiedlichen Standardfehlern und dementsprechend auch zu verschiedenen Prüfgrössen für den t Test gelangen.

Diese Erkenntnis bestätigte sich schon in mehreren Simulationsstudien und theoretischen Artikeln (z.B. Guo, 2005; Krull & MacKinnon, 2001; McNeish, 2014; Moerbeek et al., 2003). Dabei ergab sich in diesen Artikeln, dass der Standardfehler von LMs in

Abhängigkeit des Studiendesigns und der Analyseart unter- oder überschätzt wird. Beispielsweise fanden Krull und MacKinnon (2001), dass die Standardfehler eines Mediationseffekts konstant von LMs unterschätzt wurden und folglich zu einer erhöhten Fehler Typ 1 Rate führten. Diese Unterschätzung stieg mit zunehmender IKK sogar noch weiter an (Krull & MacKinnon, 2001). McNeish (2014) fand ebenfalls, dass der Standardfehler des Achsenabschnittes von LMs konstant unterschätzt wurde und bei steigender IKK die Unterschätzung zunahm. Allerdings wurde der Standardfehler nicht immer unterschätzt. Moerbeek et al. (2003) konnten in ihrem Artikel an einem simulierten und an einem realen Datensatz zeigen, dass der Standardfehler eines Interventionseffekts je nach Studiendesign von LMs unter- oder überschätzt wurde. Wird eine Intervention auf Level-1 durchgeführt, d.h. die zufällige Zuweisung zu einer Interventions- oder Kontrollgruppe geschieht auf der Individualebene, können Standardfehler durch LM überschätzt werden und folglich zu einer niedrigeren Power führen (Moerbeek et al., 2003). Wird eine Intervention auf Level-2 durchgeführt und die zufällige Zuweisung zu einer Interventions- oder Kontrollgruppe findet auf der Gruppenebene statt, führte dies wieder zu einer Unterschätzung des Standardfehlers (Moerbeek et al., 2003). In all diesen Studien konnte allerdings gezeigt werden, dass HLMs nicht diesen Limitationen unterlagen und den Standardfehler durchgehend genau geschätzt haben.

Neben der Prüfgrösse ist auch die Anzahl an Freiheitsgrade relevant, um die Signifikanz eines Effekts mittels t Test zu überprüfen. Während bei normalen LMs 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 HLMs nicht eindeutig geklärt und ein aktueller Forschungsgegenstand (McNeish, 2014; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Wie man in den vorherigen Studien erkennen kann, gibt es viele Faktoren, die einen Einfluss auf die Grösse des Standardfehlers haben. Um diese Einflüsse zu untersuchen werden zwei Simulationsstudien durchgeführt. Das Ziel der ersten Simulationsstudie ist es Teile dieser vorherigen Studien zu replizieren. Dabei wird der Einfluss von der IKK und des Studiendesigns auf die Schätzgenauigkeit der Regressionskoeffizienten und des Stan-

dardfehlers von LMs und HLMs untersucht. Gemäss der vorherigen Studien von Krull und MacKinnon (2001) und McNeish (2014) wird erwartet, dass bei zunehmender IKK die Schätzgenauigkeit von LMs abnimmt, bei HLMs aber konstant bleibt. Ausserdem wird davon ausgegangen, dass der Standardfehler des Interventionseffekts bei einer Intervention auf Level-1 gemäss Moerbeek et al. (2003) überschätzt wird, sofern keine Interaktion zwischen Intervention und Gruppenzugehörigkeit angenommen wird. Ebenfalls wird angenommen, dass bei einer Intervention auf Level-2 der Standardfehler des Interventionseffekts durch LMs wieder unterschätzt wird. In beiden Simulationsdesigns wird gemäss Moerbeek et al. (2003) erwartet, dass HLMs eine konstant genaue Schätzung der Standardfehler vorweisen.

Wie bereits erläutert führen zu klein geschätzte Standardfehler zu einer erhöhten Fehler Typ 1 Rate. Diese erhöhte Fehler Typ 1 Rate resultiert folglich in auffällig kleinen p-Werten (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). In einer Studie von Guo (2005) wurde gezeigt, dass ein p-Wert eines Effekts, der basierend auf auf einem LM berechnet wurde, halb so gross war, wie der p-Wert eines HLMs. Dies führte dazu, dass ein LM diesen Effekt als signifikant und ein HLM als nicht signifikant identifizierten (Guo, 2005). Aber auch das Gegenteil ist ein Problem. Werden Standardfehler Überschätzt, erhöht sich die Fehler Typ 2 Rate und die Power verkleinert sich. In seinem Artikel erwähnte Guo ebenfalls, dass vor allem bei der Suche nach kleinen Effektstärken oder bei kleineren Stichproben die Gefahr besteht, dass ein LM zu fehlerhaften Ergebnissen gelangt (Guo, 2005). Diese Schätzungenauigkeiten können also zu fehlerhaften Schlussfolgerungen führen und es sollte folglich im Interesse jedes Forschenden sein, diese zu vermeiden.

Folglich wird in einer zweiten kleineren Simulationsstudie untersucht, wie sich die Fehlerraten und die Power dieser beiden Analysemethoden bei einer kleineren Stichprobe und in Abhängigkeit der IKK verhält. Wie in der ersten Simulationsstudie, werden auch in der zweiten Simulationsstudie die beiden Interventionsdesigns von Moerbeek et al. (2003) untersucht. Dabei werden bei einem Interventionsdesign auf Level-1 grundsätzlich eine höhere Power erwartet, da in diesem Design der Effekt der Gruppenzugehörigkeit vom Effekt der Intervention getrennt werden kann (Moerbeek et al., 2000). Ebenfalls wird erwartet, dass die Power von LM bei einer Intervention auf Level-1 mit zunehmender IKK abnimmt, da

angenommen wird, dass es zunehmend zu einer Überschätzung des Standardfehlers kommt und dieser zu einer erhöhten Fehler Typ 2 Rate führt. Die Power eines HLMs sollte aber über alle IKK Bedingungen konstant bleiben, da gemäss den besprochenen Studien ein HLM auch bei zunehmender IKK den Standardfehler genau schätzt (McNeish, 2014). Bei einer Intervention auf Level-2 werden aufgrund des Interventionsdesigns grundsätzlich bei beiden Analysemethoden keine hohe Power erwartet (Moerbeek et al., 2000). Allerdings wird erwartet, dass die Power von LMs höher als die Power von HLMs ist, da gemäss Moerbeek et al. (2003) in dieser Situation die Standardfehler von LMs zunehmend unterschätzt werden und eine erhöhte Fehler Typ 1 Rate diese vermeintlich höhere Power verursacht.

3.2 Simulationsdesign

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

$$\begin{aligned} &\text{Design 1:} \\ \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + U_{0j} \\ & \beta_1 = \gamma_{10} \end{aligned} \tag{25}$$

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:

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \epsilon_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01} z_j + U_{0j} \end{aligned} \tag{26}$$

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 dieselben 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 der gesamten Simulationsstudie mit $\gamma_{10} = 0.12$ für die Gesamtsteigung der Level-1 Variable und $\gamma_{01} = 0.12$ für die Gesamtsteigung der Level-2 Variable festgehalten. Für das Residuum wurde eine Varianz von $\sigma_e^2 = 1.72$ und ein Erwartungswert von 0 verwendet. Diese Werte stammen alle aus der Simulation von Moerbeek et al. (2003), die ihre Werte auf einem Teildatensatz des *Television School and Family Smoking Prevention and Cessation Projekt (TVSFP)* basierten (Flay et al., 1995).

Neben den nicht manipulierten Parametern wurden die Intraklassen Korrelation und die Analysemethode variiert. Bei der IKK gibt 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 zwischen den Gruppen 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 zwischen den Gruppen 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.

3.3 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 viel grösseren Stichprobengrösse wird sichergestellt, dass Ergebnisse auf die Manipulation der Parameter und nicht auf eine ungenügende Stichprobengrösse zurückzuführen sind.

Um die oben getroffenen Annahme zu überprüfen, dass ein HLM als auch ein LM die Regressionskoeffizienten bei sich verändernder IKK genau schätzen, wird die relative Abweichung der geschätzten Regressionskoeffizienten $\widehat{\gamma}$ von den Populationsmittelwerten γ berechnet. Die Stärke dieser Abweichung wird in Prozent angegeben (Hoogland & Booms-

ma, 1998) und nach folgender Formel berechnet:

$$\Delta\widehat{\gamma} = \frac{\overline{\widehat{\gamma}} - \gamma}{\gamma} \tag{27}$$

Dabei ist $\overline{\widehat{\gamma}}$ 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. Alle weiteren Werte die eine Abweichung von mehr als 5% aufweisen gelten folglich als ungenau und sollten nicht verwendet werden.

Um nun auch noch die Annahme zu überprüfen, dass der Standardfehler von HLM auch bei zunehmender IKK genau geschätzt wird und die Schätzung des Standardfehlers eines LMs immer ungenauer wird, muss ein weiterer Kennwert berechnet werden. Dieser Kennwert beschreibt die Genauigkeit der Schätzung des Standardfehlers 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[4]:

$$\Delta\widehat{SE}_{\widehat{\gamma}} = \frac{\overline{\widehat{SE}_{\widehat{\gamma}}} - \widehat{SD}_{\widehat{\gamma}}}{\widehat{SD}_{\widehat{\gamma}}} \tag{28}$$

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 inakzeptabel. Die Genauigkeit des Standardfehlers wurde wieder in beiden Designs für jede Analysemethode und in jeder IKK Bedingung berechnet.

[4]SE ist die Abkürzung von *Standard Error*, der englischen Bezeichnung für Standardfehler

3.3.1 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 $|\Delta\widehat{\gamma}| < .05$. Auch die Varianzen dieser relativen Abweichungen ist 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 des Gesamtmittelwertes $\widehat{\gamma}_{00}$ und der Gesamtsteigung $\widehat{\gamma}_{10}$ für einen Level-1 Prädiktor bzw. $\widehat{\gamma}_{01}$ für einen Level-2 Prädiktor kann für jede der beiden Methoden aus Tabelle 5 entnommen werden. Im ersten Simulationsdesign wurden die Standardfehler des Gesamtmittelwertes bei der Verwendung von LM mit einer Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.05$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.80$ geschätzt. Bei der Verwendung von HLM reichten die SE Genauigkeitswerte von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.04$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = .02$. Bei einer IKK von .00 zeigte sich bei beiden Methoden eine akzeptable Schätzgenauigkeit des Standardfehlers mit einer Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.05$ bei LM und $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.04$ 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 $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.46$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.80$ 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 der Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ überschritten wird. Es wurde sogar ein noch strengeres Kriterium von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .05$ 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 $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} =$

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

	Design 1 (Level-1 Prädiktor)				Design 2 (Level-2 Prädiktor)			
	$\Delta\widehat{SE}_{\widehat{\gamma}_{00}}$		$\Delta\widehat{SE}_{\widehat{\gamma}_{10}}$		$\Delta\widehat{SE}_{\widehat{\gamma}_{00}}$		$\Delta\widehat{SE}_{\widehat{\gamma}_{01}}$	
IKK	LM	HLM	LM	HLM	LM	HLM	LM	HLM
.00	-.05[a]	-.04[b]	-.01[b]	-.01[b]	-.02[b]	-.00[b]	.01[b]	.02[b]
.05	-.46	.00[b]	.03[b]	.01[b]	-.45	.02[b]	-.47	-.02[b]
.10	-.60	-.02[b]	.05[a]	.00[b]	-.59	-.01[b]	-.58	.02[b]
.15	-.65	.01[b]	.06[a]	-.02[b]	-.66	-.02[b]	-.65	.00[b]
.20	-.69	.02[b]	.13	.01[b]	-.68	.04[b]	-.69	.02[b]
.25	-.73	-.00[b]	.17	.02[b]	-.72	.03[b]	-.73	-.03[b]
.30	-.75	-.01[b]	.16	-.03[b]	-.75	-.00[b]	-.75	.00[b]
.40	-.78	.01[b]	.34	.04[b]	-.79	.01[b]	-.78	.02[b]
.50	-.80	-.00[b]	.42	.00[b]	-.80	.00[b]	-.79	.04[b]

Hinweis: $\Delta\widehat{SE}_{\widehat{\gamma}_{00}}$ ist die Schätzgenauigkeit des Standardfehlers des Gesamtmittelwerts. $\Delta\widehat{SE}_{\widehat{\gamma}_{10}}$ ist die Schätzgenauigkeit des Standardfehlers der Gesamtsteigung eines Level-1 Prädiktors und $\Delta\widehat{SE}_{\widehat{\gamma}_{01}}$ ist die Schätzgenauigkeit des Standardfehlers der Gesamtsteigung eines Level-2 Prädiktors. [a] $|\Delta\widehat{SE}_{\widehat{\gamma}}| < .10$, [b] $|\Delta\widehat{SE}_{\widehat{\gamma}}| < .05$

$-.02$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.80$ auf. Das HLM schätzte den Standardfehler mit einer SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.02$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = .04$. Es zeigte sich bezüglich der Genauigkeit der Schätzung des Standardfehlers des Gesamtmittelwertes $\widehat{\gamma}_{00}$ ein ähnliches Bild wie beim ersten Simulationsdesign. Wieder wiesen beide Methoden bei einer IKK von .00 eine genau Schätzung des Standardfehlers von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.02$ bei LM und $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.00$ bei HLM auf. Diese Genauigkeit nahm bei erhöhter IKK und Verwendung von LM wieder stark ab. In der sechsten Spalte von Tabelle 5 kann man erkennen, dass bei einer IKK von .05 nur noch eine SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.45$ erreicht wird und schlussendlich bei einer IKK von .50 bis zu einer SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.80$ 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 siebte 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 $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = -.01$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = .42$. Die Schätzgenauigkeit des HLM reichte von $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = -.03$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{00}} = .04$. Beide Methoden weisen wieder bei einer IKK von .00 eine genaue Schätzung des Standardfehlers auf. Das LM verzeichnete eine abnehmende Genauigkeit 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 $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ nicht. Mit einer SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = .06$ 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 $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = .42$ 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 6 wird dieser Zusammenhang noch einmal graphisch dargestellt. Wie in Abbildung 6 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 $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .05$. Die Schätzungen des Standardfehlers sind daher auch noch bei einer hohen IKK von .50 mit $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = .00$ im 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 $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = .0056$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = -.79$ geschätzt. Das HLM schätzte die Standardfehler mit einer Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = -.03$ bis $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = .04$. Beide Analysemethoden weisen wieder eine genaue Schätzung des Standardfehlers bei einer

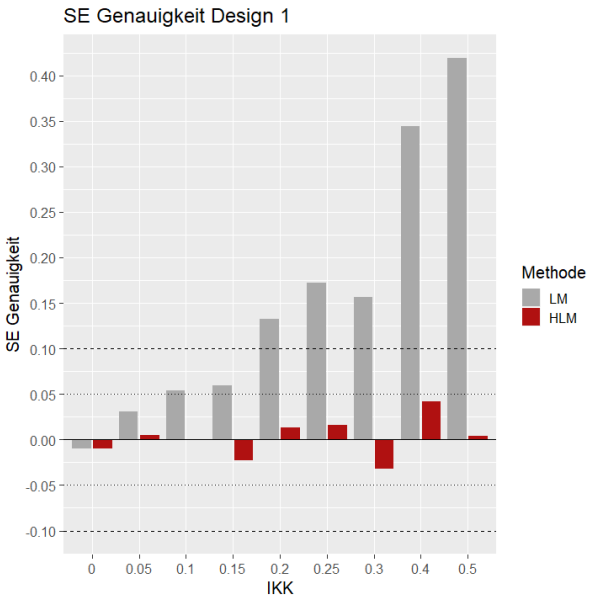


Abbildung 6: Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im ersten Simulationsdesign. SE: *Standard Error*, LM: Lineares Modell, HLM: Hierarchisch lineares Modell, IKK: Intraklassen Korrelation

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 $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = -.47$ der Grenzwert für eine akzeptable Abweichung bereits überschritten. Diese Unterschätzung steigt mit zunehmender IKK weiter an, bis hin zu einer SE Genauigkeit von $\Delta\widehat{SE}_{\widehat{\gamma}_{01}} = -.79$ bei einer IKK von 0.50. In diesem Fall resultiert diese Unterschätzung des Standardfehlers wieder in einer erhöhten Fehler Typ 1 Rate. In Abbildung 7 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 $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ nicht überschreiten und sogar kleiner als das noch strengere Kriterium $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .05$ sind. Es entsteht folglich wie-

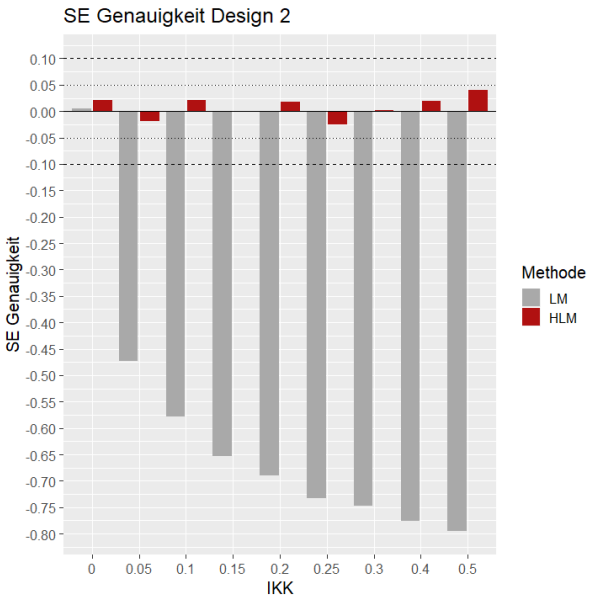


Abbildung 7: Genauigkeit der Schätzung des Standardfehlers der Gesamtsteigung für jede Methode in allen IKK Bedingungen im zweiten Simulationsdesign. SE: *Standard Error*, LM: Lineares Modell, HLM: Hierarchisch lineares Modell, IKK: Intraklassen Korrelation

der 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.

3.3.2 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 Simulati-

onsdesigns bei Verwendung von LM zunehmend unterschätzt. Nur bei einer IKK von .00 schätzte das LM den Standardfehler genau. Das bedeutet, dass ein LM den Standardfehler nur dann genau schätzt, wenn die zwischen den Gruppen keine zufällige Abweichung besteht und die Gruppenzugehörigkeit folglich keinen Einfluss auf die abhängige Variable hat. 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 mit einem Level-1 Prädiktor wurde sobald die Gruppenzugehörigkeit einen Einfluss hatte, der Standardfehler mit steigender IKK zunehmend überschätzt. Beim zweiten Simulationsdesign mit einem Level-2 Prädiktor 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 Analysemethode 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 gestützt.

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 werden 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 Anaylse 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 Fehlerraten oder die effektive Power dieser beiden Analysemethoden. 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 Fehlerraten und die Power dieser beiden Methoden über die verschiedenen IKK Bedingungen verändert.

3.4 Studie 2: Zuverlässigkeit von LM und HLM

In der zweiten Studie wurde untersucht, wie zuverlässig LMs und HLMs einen Effekt einer Intervention finden. Die Fehler Typ 1 Rate und die Power sind zwei mögliche Indikatoren für diese Zuverlässigkeit einer Methode. Wie bereits erwähnt, entsteht ein Fehler Typ 1 dann, wenn ein Test fälschlicherweise zu einem signifikanten Ergebnis gelangt. Das würde beispielsweise bedeuten, dass ein Test eine Intervention als effektiv identifiziert, obwohl sie das in Wirklichkeit gar nicht ist. Die Power bezeichnet die Fähigkeit einen Effekt zu finden, wenn er auch wirklich vorhanden ist. Damit die Fehler Typ 1 Rate als auch die Power untersucht werden kann, wurden in dieser Simulationsstudie die Datensätze einmal mit einer effektiven Intervention und einmal mit einer ineffektiven Intervention simuliert. Unter

einer ineffektiven Intervention wird hier eine Intervention verstanden, die keinen Effekt auf die abhängige Variable hat. Das bedeutet, dass bei der Simulation der ineffektiven Intervention die Gesamtsteigung des Level-1 Prädiktors γ_{10} und die des Level-2 Prädiktors γ_{10} auf 0 gesetzt wurden. Betrachtet man die Gleichungen aus Abschnitt 3.2, kann man erkennen, dass mit diesen gewählten Werten für γ_{10} bzw. γ_{10} die Level-1 Variable x_{ij} bzw. die Level-2 Variable z_j keinen Einfluss mehr auf die Ausprägung der abhängigen Variable y_{ij} hat.

Die Fehlerraten und die Power stehen in einem direkten Zusammenhang mit der Stichprobengrösse, so dass eine steigende Stichprobengrösse zu einer tieferen Fehlerrate und einer höheren Power führt (Snijders, 2005). Folglich würde eine so grosse Stichprobe aus Studie 1 wahrscheinlich zu keinen Unterschieden in den Fehlerraten oder 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 wurde 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 Werte entsprechen nun den festgelegten Werte aus der Simulation von Moerbeek et al. (2003), die aus dem Datensatz des *TVSFP* entnommen wurden (Flay et al., 1995). Diese Reduktion führte zu einer Stichprobengrösse von insgesamt 840 Beobachtungen.

Wie bereits in der Herleitung der Forschungsfrage beschrieben, wurde der Effekt der Intervention mit einem t Test überprüft, dessen Prüfgrösse sich aus dem Verhältnis des Geschätzten Regressionskoeffizienten und dem dazugehörigen Standardfehler berechnet. 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 zur Berechnung der Freiheitsgrade von hierarchischen linearen Modellen verwendet wird (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

Um nun die Fehlerraten und die Power zu berechnen, wurde für alle Datensätze 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. In einem Datensatz mit einer effektiven Intervention entspricht diese prozentuale Häufigkeit der Power. In einem Datensatz mit einer ineffektiven Intervention gibt diese Häufigkeit der Fehler Typ 1 Rate an.

3.4.1 Ergebnisse Studie 2

Die Ergebnisse der zweiten Simulationsstudie werden in Tabelle 6 abgebildet. Es ist zu beachten, dass die Fehler Typ 1 Rate und die Power nur für die Gesamtsteigung des Level-1 Prädiktors $\widehat{\gamma}_{10}$ bzw. die des Level-2 Prädiktors $\widehat{\gamma}_{10}$ berechnet wurde, weil diese Regressionskoeffizienten auch den Effekt der Intervention abbilden. Als erstes werden die Datensätze besprochen, die mit $\widehat{\gamma}_{10} = 0.12$ bzw. $\widehat{\gamma}_{10} = 0.12$ simuliert wurden. Dies entspricht hier einer effektiven Intervention und wird folglich dazu verwendet, um die Power der beiden Methoden zu untersuchen.

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 8 sind auf der linken Seite die einzelnen Power-Werte jeder IKK Bedingung und für jede Analysemethode abgebildet. Dabei lässt sich erkennen, dass die Power von einem LM bis zu einer IKK von .10 noch einigermassen mit der Power eines HLMs mithalten kann. Sobald aber die IKK grösser als .10 ist lässt sich eine kontinuierliche Abnahme der Power beobachten. Betrachtet man die Schätzgenauigkeit des Standardfehlers des LMs an dieser Stelle, erkennt man, dass ab einer IKK von .15 mit $\Delta\widehat{SE}_{\widehat{\gamma}_{10}} = 0.1167$ den Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ (Hoogland & Boomsma, 1998) überschritten wird und folglich die Schätzung des Standardfehlers als nicht mehr akzeptabel gilt. Dies führt zu einer zu hohen Fehler Typ 2 Rate, die schlussendlich in dieser tieferen Power resultiert. Die Power unter Verwendung des HLMs reicht von .75 bis .78 und bleibt gemäss Abbildung 8 über alle Bedingungen

Tabelle 6: Power und Fehler Typ 1 Rate in beiden Simulationsdesigns für jede Analysemethode in allen IKK Bedingungen.

	Power				Fehler Typ 1 Rate			
	Design 1		Design 2		Design 1		Design 2	
IKK	LM	HLM	LM	HLM	LM	HLM	LM	HLM
.00	.76	.76	.74	.72	.05	.05	.05	.04
.05	.72	.75	.71	.53	.05	.06	.12	.06
.10	.73	.76	.64	.38	.04	.05	.17	.05
.15	.72	.78	.64	.34	.04	.05	.23	.05
.20	.67	.75	.56	.24	.02	.04	.27	.04
.25	.69	.78	.56	.21	.02	.05	.29	.04
.30	.62	.78	.57	.19	.02	.05	.36	.06
.40	.55	.77	.56	.13	.01	.05	.43	.06
.50	.44	.76	.55	.10	.01	.06	.46	.04

Hinweis: Bei Power wurden Datensätze mit einer Gesamtsteigung von $\widehat{\gamma}_{10} = 0.12$ bzw. $\widehat{\gamma}_{10} = 0.12$ simuliert. Bei der Fehler Typ 1 Rate mit $\widehat{\gamma}_{10} = 0$ bzw. $\widehat{\gamma}_{10} = 0$. IKK: Intraklassenkorrelation, LM: Lineares Modell, HLM: Hierarchisch lineares Modell

relativ konstant. Die Schätzgenauigkeit der Standardfehler mittels HLM überschreitet in keiner IKK Bedingung den Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$. Bei einer Analyse mit einem 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 LMs von .55 bis .74 und nimmt wieder mit steigender IKK ab. Die Power des HLMs reicht im zweiten Simulationsdesign von .10 bis .72 und verzeichnete ebenfalls eine Abnahme der Power mit zunehmender IKK. Auf der rechten Seite der Abbildung 8 ist dieser Verlauf noch einmal Visualisiert. Dabei kann man erkennen, dass die Abnahme der Power bei einer zunehmenden IKK bei einer Analyse mit einem HLM stärker als bei einer Analyse mit einem LM ist. Bei einer IKK von .50 erreichte ein HLM nur noch eine Power von .10, wobei ein LM eine Power von .55 erreichte. Die Schätzgenauigkeiten des Standardfehlers werden ähnlich wie in Studie 1 von dem LM ab einer IKK von .05 stark unterschätzt und überschreiten in allen weiteren IKK Bedingungen den Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$. Diese Unterschätzung der Standardfehler durch das LM führt zu ei-

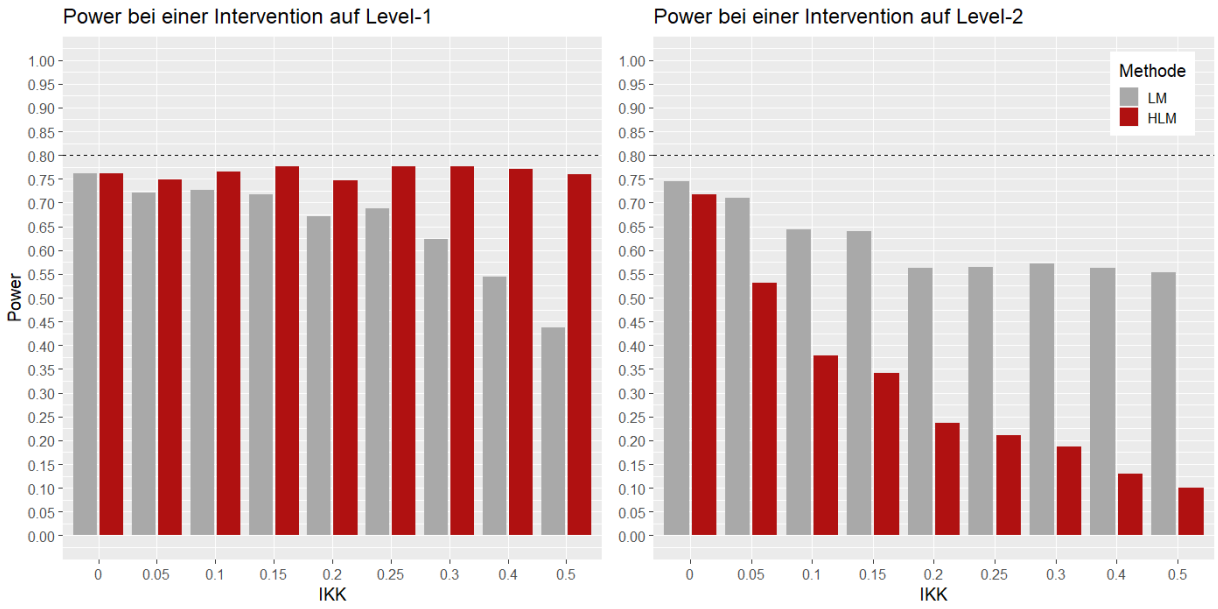


Abbildung 8: Statistische Power von LM und HLM in den verschiedenen IKK Bedingungen in beiden Studiendesigns. LM: Lineares Modell, HLM: Hierarchisches lineares Modell

ner Inflation der Prüfgrösse und schlussendlich zu dieser vermeintlich besseren Power. Die Schätzung der Standardfehler bleibt bei der Verwendung eines HLMs im akzeptablen Bereich und überschreiten den Grenzwert in keiner IKK Bedingung. Folglich bildet ein HLM in dieser Situation die Power adäquater als ein LM ab.

Die im Anschluss besprochenen Ergebnisse, beziehen sich auf die simulierten Datensätze mit einer ineffektiven Intervention, die mit einer Gesamtsteigung von $\widehat{\gamma}_{10} = 0$ bzw. $\widehat{\gamma}_{10} = 0$ simuliert wurden. In den letzten vier Spalten der Tabelle 6 sind die Fehler Typ 1 Raten beider Simulationsdesigns aufgeführt. Im ersten Simulationsdesign mit einer Intervention auf Level-1, zeigt sich, dass die Fehler Typ 1 Rate bei einem LM mit zunehmender IKK fortlaufend abnimmt. Diese Reduktion der Fehler Typ 1 Rate von .05 auf 0.01 ist auf die Überschätzung des Standardfehlers in diesem Studiendesign zurückzuführen. Die Fehler Typ 1 Rate des HLMs im ersten Studiendesign reicht von .04 bis .06. Diese Entspricht ungefähr den Erwartungen, die man bei einem Signifikanzniveau von 5% hat.

Betrachtet man nun die Fehler Typ 1 Raten des zweiten Studiendesigns, kann man erkennen, dass die Fehler Typ 1 Rate des LMs mit zunehmender IKK von .05 bis .46 ansteigt.

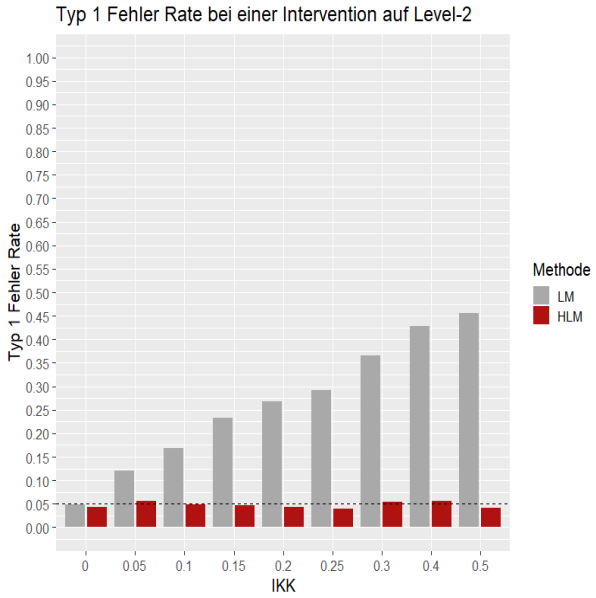


Abbildung 9: Fehler Typ 1 Rate von LMs und HLMs bei einer ineffektiven Intervention auf Level-2. LM: Lineares Modell, HLM: Hierarchisches lineares Modell

Dieser Anstieg der Fehler Typ 1 Rate ist auf die Inflation der Prüfgrösse zurückzuführen, die aufgrund der Unterschätzung des Standardfehlers entsteht. Bereits ab einer IKK von .05 wird mit .12 das Signifikanzniveau von 5% überschritten. Im Vergleich dazu liegen die Fehler Typ 1 Raten des HLMs wieder in einem Bereich von .04 bis .06. In Abbildung 9 ist der Verlauf der Fehler Typ 1 Raten von LMs und HLMs noch einmal visualisiert.

3.4.2 Diskussion Studie 2

Die zweite Simulationsstudie untersuchte, wie genau sich die Fehlerraten und die Power bezüglich eines Effekts einer Intervention zwischen den beiden Analysemethoden unterscheidet und vor allem wie sich diese beiden Faktoren über die verschiedenen IKK Bedingungen verändern. 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 ein LM verwendet wird. Wird an Stelle von einem 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. Vergleicht man die Fehler Typ 1 Raten dieser beiden Modelle, bestätigten sich ebenfalls die Erwartungen. Im ersten Studiendesign nahmen die Fehler Typ 1 Rate des LMs aufgrund der zunehmenden Überschätzung des Standardfehlers bei ansteigender IKK fortlaufend ab. Genau das gegenteilige Bild zeigte sich beim zweiten Studiendesign. Hier stieg die Fehler Typ 1 Rate des LMs bei zunehmender IKK stark an. Die Fehler Typ 1 Rate des HLMs blieb in beiden Studiendesign in einem akzeptablen Bereich.

Interessiert man sich nun 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, sollte man sich nicht von den Ergebnissen auf der rechten Seite der Abbildung 8 täuschen lassen. Auch wenn es so scheint, dass ein LM eine weniger starke Reduktion der Power erfährt als ein HLM, sollte in dieser Situation trotzdem ein HLM zur Analyse verwendet werden. Da in der ersten Studie gezeigt wurde, dass in dieser Situation der Standardfehler von einem LM stark unterschätzt wird, kann diese schwache Abnahme der Power auf die Inflation der Prüfgrösse und die Zunahme der signifikanten

Tests zurückgeführt werden. Diese Unterschätzung des Standardfehlers und die Inflation der Prüfgrösse ist vor allem bei der Fehler Typ 1 Rate in Abbildung 9 gut zu beobachten. Obwohl die Intervention in dieser Studienanordnung keinen Effekt auf die abhängige Variable hatte, gelang es dem LM bei zunehmender IKK schlechter den Standardfehler genau zu schätzen und fand schlussendlich bei einer IKK von .50 in fast der Hälfte aller Fälle einen fälschlicherweise signifikanten Effekt der Intervention.

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).

3.5 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 Fehler Typ 1 Rate und Power von LMs und HLMs zu untersuchen. Auch wenn die Regressionskoeffizienten von beiden Methoden genau geschätzt wurden, konnten die Studien zeigen, dass ein HLM auch bei hoher Abhängigkeit der Gruppenzugehörigkeit genaue Schätzungen der Standardfehler liefert, wohingegen ein LM bei steigender IKK eine immer stärkere Verzerrung der Schätzung aufweist. Dementsprechend führten Analysen mit einem HLM auch zu keiner Verzerrung der Power bei der Testung von Interventionseffekten.

Die aus dieser Simulationsstudie resultierende Erkenntnis, dass ein HLM einen klaren Vorteil gegenüber einem 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 LMs und HLMs zu untersuchen. Damit können klare Empfehlungen für die Forschung mit hierarchi-

schen 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. Ein Beispiel dafür wären Interventionen mit Familien, bei denen es schwierig ist gewisse Interventionen nur bei einzelne Familienmitgliedern durchzuführen, ohne dass andere Familienmitglieder davon beeinflusst werden. Wie man in der zweiten Studie beobachten konnte, verringerte sich die Power des HLMs 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, die Power aber 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. Dabei könnte ebenfalls untersucht werden, ob es ein optimales Verhältnis zwischen Gruppengrösse und Gruppenanzahl gibt, um die Power eines HLMs zu maximiere.

Die aktuelle Studie unterlag mehreren Limitationen. Zum einen wurde nur ein sehr einfaches Modell zur Simulation der Daten verwendet. Dabei gab es keine *Cross-Level* Interaktion zwischen einem Level-1 und einem Level-2 Prädiktor. 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 *Cross-Level* Interaktionen in Bezug auf 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 eine Stichprobe gewählt werden sollte, um eine genaue Schätzung und eine akzeptable Power zu erhalten. Da die Wahl der Stichprobengrösse in der Forschung oft eine Kostenfrage ist, könnte durch eine weitere Simulationsstu-

die, die diese Punkte berücksichtigt, klare Empfehlungen für die Praxis erarbeitet werden. Als abschliessende Limitation kann die Wahl des Grenzwertes für die Schätzgenauigkeit des Standardfehlers genannt werden. Dieser Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ wurde von Hoogland und Boosma (1998) festgelegt. Wie man aber auf der linken Seite der Abbildung 8 erkennen kann, sinkt die Power eines LMs bereits bei einer IKK von .05. Bei dieser IKK wird bereits die Schätzgenauigkeit von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .05$ überschritten. Weitere Simulationsstudien könnten untersuchen, ob dieser Grenzwert von $|\Delta\widehat{SE}_{\widehat{\gamma}}| > .10$ gerechtfertigt ist oder ob ein strengerer Grenzwert benötigt wird, um genaue von ungenauen Analysemethoden zu trennen.

Die Ergebnisse dieser Simulationsstudie 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 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.

Die aktuelle Studie konnte zeigen, dass es sich bei hierarchischen Daten und einem geeigneten Interventionsdesign lohnt ein HLM anstelle eines LMs zu verwenden, da dieses Modell die nötigen Parameter genau schätzen und zu einer konstanten Power führen. Ebenfalls konnte gezeigt werden, dass bei der Verwendung eines HLMs die Fehler Typ 1 Rate konstant gehalten werden kann, auch wenn der Einfluss der Gruppenzugehörigkeit ansteigt. Dies gibt Forschenden mehr Sicherheit bei der Testung von Effekten und führt zu glaubhafteren Ergebnissen. Zukünftige Simulationsstudien könnten sich mit weiteren Aspekten beschäftigen, die einen Einfluss auf die Schätzgenauigkeit haben.

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. doi: 10.1093/jncimonographs/lgs010

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.

Flay, B., Miller, T., Hedeker, D., Siddiqui, O., Britton, C., Brannon, B., ... Dent, C. (1995). The television, school, and family smoking prevention and cessation project: Viii. student outcomes and mediating variables. *Preventive Medicine*, *24* (1), 29 - 40. doi: 10.1006/pmed.1995.1005

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)

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, *27* (6), 637 - 652. doi: 10.1016/j.childyouth.2004.11.017

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

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Bd. 112). Springer.

Krull, J. L. & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36* (2), 249-277. (PMID: 26822111) doi: 10.1207/S15327906MBR3602_06

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: 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: 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: 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. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18* (3), 237-259. doi: 10.3102/10769986018003237

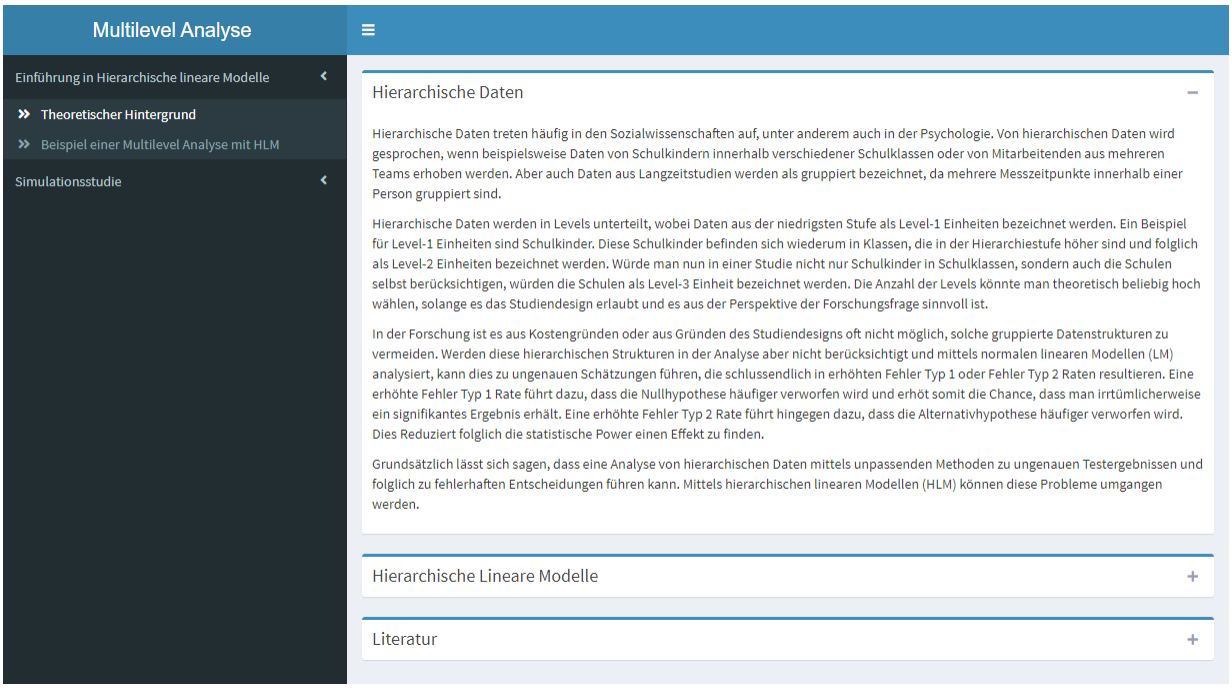
Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. Aufl.). Los Angeles: SAGE.

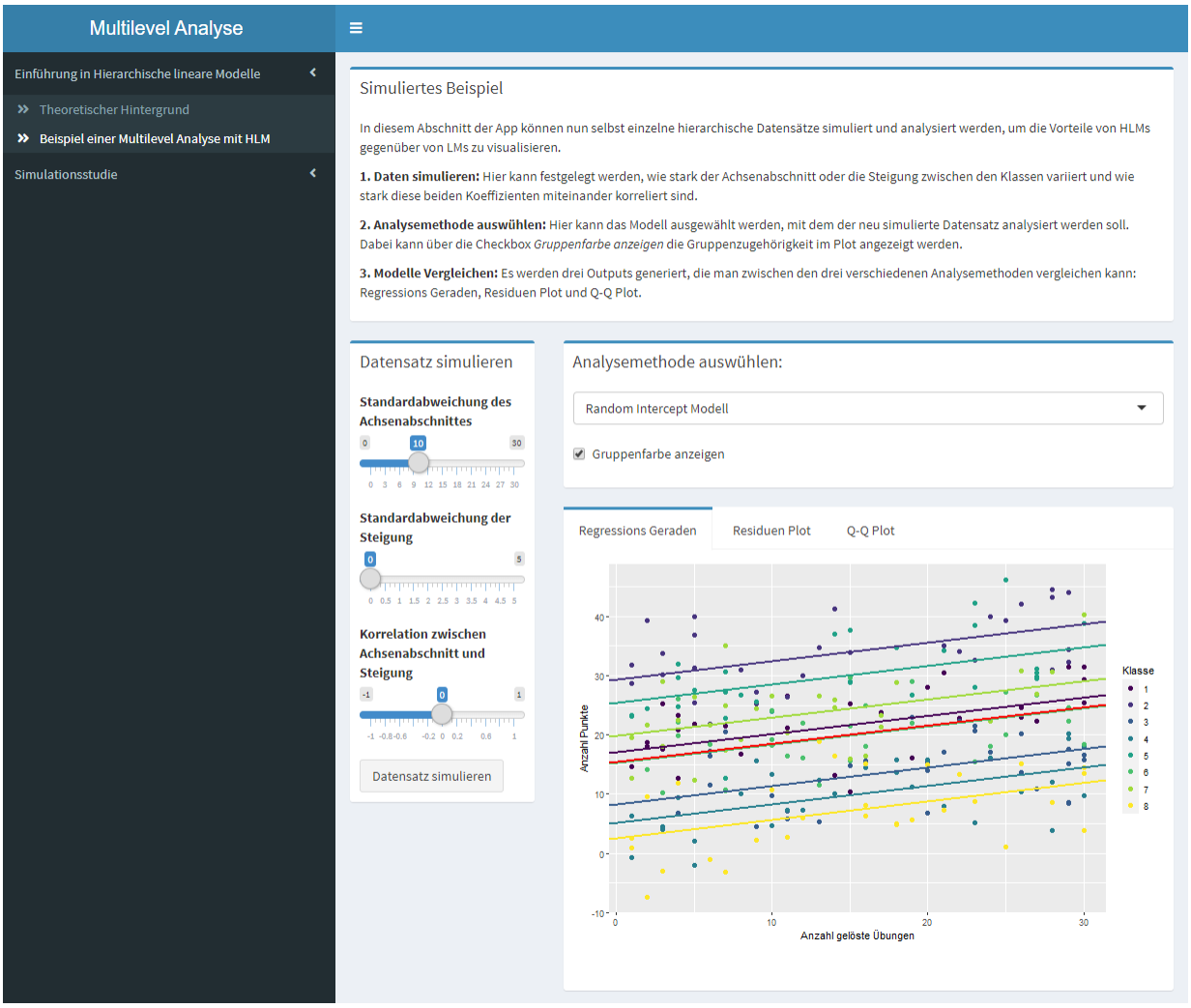
Twisk, J. W. R. (2006). *Applied multilevel analysis: A practical guide for medical researchers*. Cambridge University Press. doi: 10.1017/CBO9780511610806

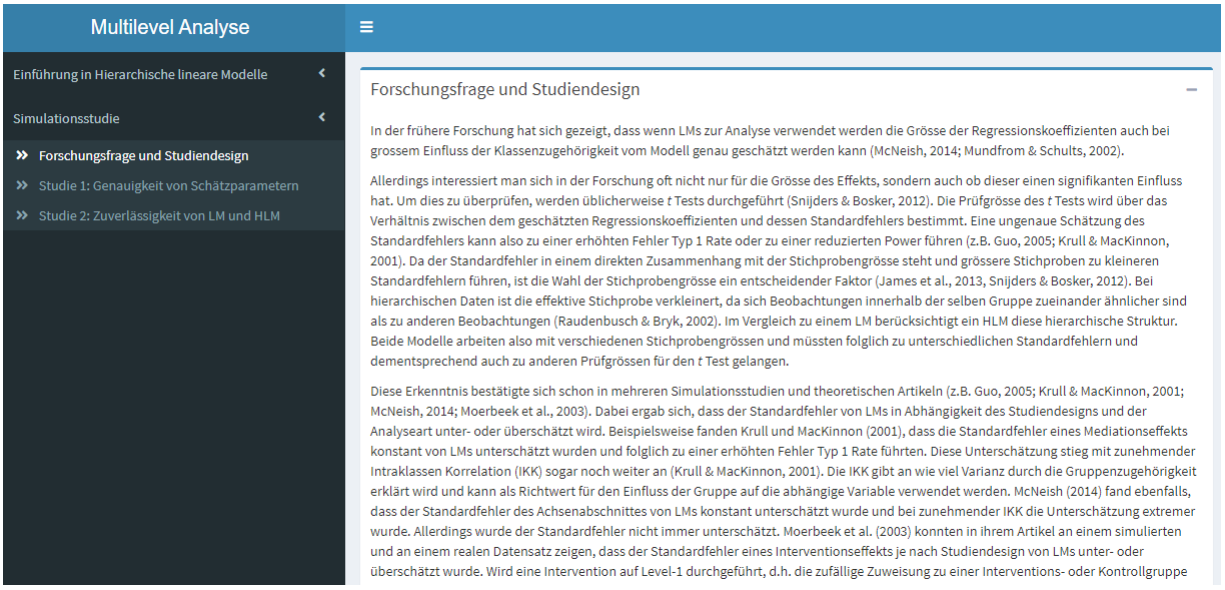
Woltman, H., Feldstain, A., MacKay, J. C. & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, *8* (1), 52–69.

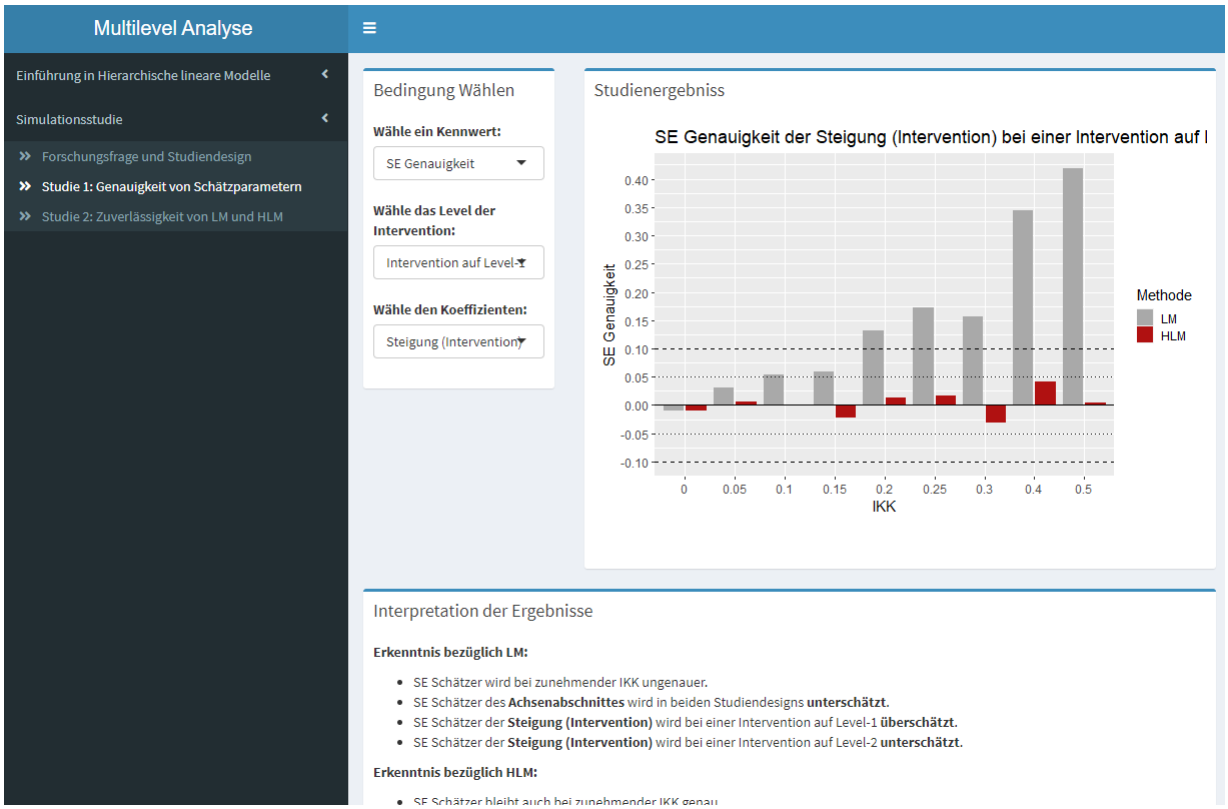
5 Anhang

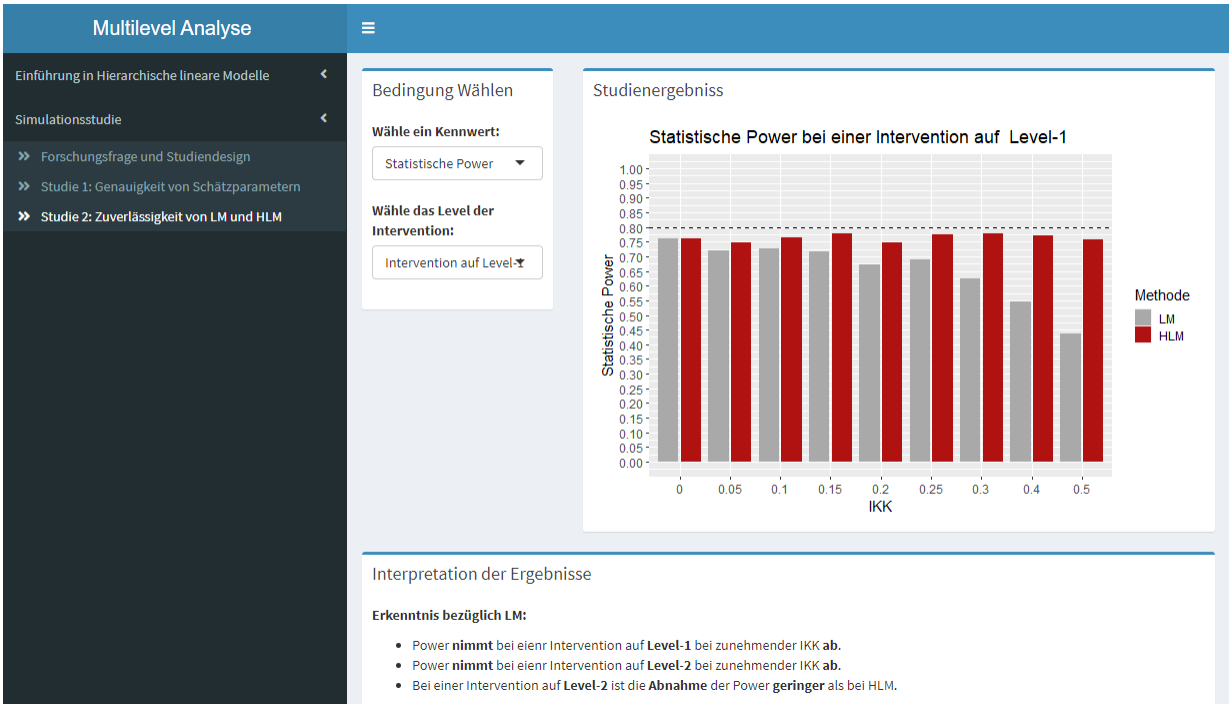
Shiny App Screenshots




Multilevel Analyse
Einführung in Hierarchische lineare Modelle
Theoretischer Hintergrund
Beispiel einer Multilevel Analyse mit HLM
Simulationsstudie
Simuliertes Beispiel
In diesem Abschnitt der App können nun selbst einzelne hierarchische Datensätze simuliert und analysiert werden, um die Vorteile von HLMs gegenüber von LMs zu visualisieren.
1. Daten simulieren: Hier kann festgelegt werden, wie stark der Achsenabschnitt oder die Steigung zwischen den Klassen variiert und wie stark diese beiden Koeffizienten miteinander korreliert sind.
2. Analysemethode auswühlen: Hier kann das Modell ausgewählt werden, mit dem der neu simulierte Datensatz analysiert werden soll. Dabei kann über die Checkbox Gruppenfarbe anzeigen die Gruppenzugehörigkeit im Plot angezeigt werden.
3. Modelle Vergleichen: Es werden drei Outputs generiert, die man zwischen den drei verschiedenen Analysemethoden vergleichen kann: Regressions Geraden, Residuen Plot und Q-Q Plot.
Datensatz simulieren
Standardabweichung des Achsenabschnittes
Standardabweichung der Steigung
Korrelation zwischen Achsenabschnitt und Steigung
Datensatz simulieren
Analysemethode auswühlen:
Random Intercept Modell
Gruppenfarbe anzeigen
Regressions Geraden
Residuen Plot
Q-Q Plot
Anzahl Punkte
Anzahl gelöste Übungen
Klasse


Multilevel Analyse
Einführung in Hierarchische lineare Modelle
Simulationsstudie
Forschungsfrage und Studiendesign
Studie 1: Genauigkeit von Schätzparametern
Studie 2: Zuverlässigkeit von LM und HLM
Forschungsfrage und Studiendesign
In der frühere Forschung hat sich gezeigt, dass wenn LMs zur Analyse verwendet werden die Grösse der Regressionskoeffizienten auch bei grossem Einfluss der Klassenzugehörigkeit vom Modell genau geschätzt werden kann (McNeish, 2014; Mundfrom & Schults, 2002).
Allerdings interessiert man sich in der Forschung oft nicht nur für die Grösse des Effekts, sondern auch ob dieser einen signifikanten Einfluss hat. Um dies zu überprüfen, werden üblicherweise t Tests durchgeführt (Snijders & Bosker, 2012). Die Prüfgrösse des t Tests wird über das Verhältnis zwischen dem geschätzten Regressionskoeffizienten und dessen Standardfehlers bestimmt. Eine ungenaue Schätzung des Standardfehlers kann also zu einer erhöhten Fehler Typ 1 Rate oder zu einer reduzierten Power führen (z.B. Guo, 2005; Krull & MacKinnon, 2001). Da der Standardfehler in einem direkten Zusammenhang mit der Stichprobengrösse steht und grössere Stichproben zu kleineren Standardfehlern führen, ist die Wahl der Stichprobengrösse ein entscheidender Faktor (James et al., 2013, Snijders & Bosker, 2012). Bei hierarchischen Daten ist die effektive Stichprobe verkleinert, da sich Beobachtungen innerhalb der selben Gruppe zueinander ähnlicher sind als zu anderen Beobachtungen (Raudenbusch & Bryk, 2002). Im Vergleich zu einem LM berücksichtigt ein HLM diese hierarchische Struktur. Beide Modelle arbeiten also mit verschiedenen Stichprobengrössen und müssten folglich zu unterschiedlichen Standardfehlern und dementsprechend auch zu anderen Prüfgrössen für den t Test gelangen.
Diese Erkenntnis bestätigte sich schon in mehreren Simulationsstudien und theoretischen Artikeln (z.B. Guo, 2005; Krull & MacKinnon, 2001; McNeish, 2014; Moerbeek et al., 2003). Dabei ergab sich, dass der Standardfehler von LMs in Abhängigkeit des Studiendesigns und der Analyseart unter- oder überschätzt wird. Beispielsweise fanden Krull und MacKinnon (2001), dass die Standardfehler eines Mediationseffekts konstant von LMs unterschätzt wurden und folglich zu einer erhöhten Fehler Typ 1 Rate führten. Diese Unterschätzung stieg mit zunehmender Intraklassen Korrelation (IKK) sogar noch weiter an (Krull & MacKinnon, 2001). Die IKK gibt an wie viel Varianz durch die Gruppenzugehörigkeit erklärt wird und kann als Richtwert für den Einfluss der Gruppe auf die abhängige Variable verwendet werden. McNeish (2014) fand ebenfalls, dass der Standardfehler des Achsenabschnittes von LMs konstant unterschätzt wurde und bei zunehmender IKK die Unterschätzung extremer wurde. Allerdings wurde der Standardfehler nicht immer unterschätzt. Moerbeek et al. (2003) konnten in ihrem Artikel an einem simulierten und an einem realen Datensatz zeigen, dass der Standardfehler eines Interventionseffekts je nach Studiendesign von LMs unter- oder überschätzt wurde. Wird eine Intervention auf Level-1 durchgeführt, d.h. die zufällige Zuweisung zu einer Interventions- oder Kontrollgruppe


Multilevel Analyse
Einführung in Hierarchische lineare Modelle
Simulationsstudie
Forschungsfrage und Studiendesign
Studie 1: Genauigkeit von Schätzparametern
Studie 2: Zuverlässigkeit von LM und HLM
Bedingung Wählen
Wähle ein Kennwert:
SE Genauigkeit
Wähle das Level der Intervention:
Intervention auf Level-1
Wähle den Koeffizienten:
Steigung (Intervention)
Studienergebniss
SE Genauigkeit der Steigung (Intervention) bei einer Intervention auf l
SE Genauigkeit
IKK
Methode
LM
HLM
Interpretation der Ergebnisse
Erkenntnis bezüglich LM:
SE Schätzer wird bei zunehmender IKK ungenauer.
SE Schätzer des Achsenabschnittes wird in beiden Studiendesigns unterschätzt.
SE Schätzer der Steigung (Intervention) wird bei einer Intervention auf Level-1 überschätzt.
SE Schätzer der Steigung (Intervention) wird bei einer Intervention auf Level-2 unterschätzt.
Erkenntnis bezüglich HLM:


Multilevel Analyse
Einführung in Hierarchische lineare Modelle
Simulationsstudie
Forschungsfrage und Studiendesign
Studie 1: Genauigkeit von Schätzparametern
Studie 2: Zuverlässigkeit von LM und HLM
Bedingung Wählen
Wähle ein Kennwert:
Statistische Power
Wähle das Level der Intervention:
Intervention auf Level-1
Studienergebniss
Statistische Power bei einer Intervention auf Level-1
Statistische Power
IKK
Methode
LM
HLM
Interpretation der Ergebnisse
Erkenntnis bezüglich LM:
Power nimmt bei einer Intervention auf Level-1 bei zunehmender IKK ab.
Power nimmt bei einer Intervention auf Level-2 bei zunehmender IKK ab.
Bei einer Intervention auf Level-2 ist die Abnahme der Power geringer als bei HLM.