

Analyse von Paneldaten - Ein Abriss ausgewählter Verfahren

Petra Stein, Dawid Bekalarczyk

22. Juli 2010

Inhaltsverzeichnis

1	Einleitung	3
2	Ausgewählte Analyseverfahren für Paneldaten	9
2.1	Elementare Veränderung der Werte einer Variablen im Zeitverlauf	9
2.1.1	Der Korrelationskoeffizient als Indikator für Stabilität und Veränderung?	10
2.1.2	Die einfache Regression einer zeitlich vorangestellten inhaltlich gleichen Variablen auf die Variable zum späteren Zeitpunkt	15
2.2	Eine kurze Einführung in die Pfadanalyse	18
2.3	Eine einfache Pfadanalyse mit Paneldaten	27
2.4	Das Ein-Indikatoren-Modell als Ansatz zur Unterscheidung zwischen Veränderung und mangelnder Reliabilität	31
2.5	Regressionsmodelle für Paneldaten	37
2.5.1	Modelle mit variablen Regressionskonstanten: Fixed- und Random-Effects-Modelle	44
2.5.2	Regressionsmodelle mit Differenzvariablen	83
2.5.3	Modelle mit endogener Dynamik	91
2.6	Lineare Panelmodelle mit latenten Variablen	98
2.7	Anwendung von LISREL auf Paneldaten	111
3	Fazit	118
4	Appendix	119
4.1	Appendix A - Fiktiver Datensatz mit variierender x_2 -Variable	119
4.2	Appendix B - Fiktiver Datensatz zur Pfadanalyse mit Paneldaten	120

4.3	Appendix C - Fiktive Datensätze für eine Regression mit Differenzvariablen	121
4.3.1	Datensatz mit einer eher realistischen Struktur	121
4.3.2	Datensatz mit einer eher unrealistischen Struktur	122
4.4	Appendix D - Datensatz zur Berechnung eines FEM	123
4.4.1	Ursprungsdatensatz	123
4.4.2	Datensatz mit Dummy-Variablen	124
4.5	Appendix E - Korrelations- und Kovarianzmatrix der Indikatoren für ein Strukturgleichungsmodell	125
4.6	Appendix F - Modellzeilen in LISREL für verschiedene Varianten von Strukturgleichungsmodellen	126
4.7	Kurzer Verweis auf Grundlagen der linearen (Regressions-)Analyse	128

Literaturverzeichnis

134

Kapitel 1

Einleitung

Dieses Skript umreißt einige Möglichkeiten, sog. Paneldaten zu analysieren. Diese Möglichkeiten werden oft unter dem Begriff der **Panelanalyse** subsumiert. Ehe auf diesen Begriff eingegangen wird, sollten aber zuerst elementarere Begriffe geklärt werden:

Man redet von **Paneldaten**, als Spezialfall von Längsschnittdaten, wenn sie über sog. **Panelstudien** erhoben werden, denen das Paneldesign unterliegt.¹ Solche Studien haben gemeinsam, dass sie zu mehreren Zeitpunkten dieselben Messungen an den gleichen Objekten vornehmen. Diese drei Aspekte, welche sich auf die Zeit-Dimension, die erhobenen Merkmale und auf die Untersuchungseinheiten beziehen, charakterisieren Panelstudien. Die Messzeitpunkte werden üblicherweise als sog. *Wellen* bezeichnet. Eine Panelstudie mit z.B. 4 Wellen würde bedeuten, dass zu vier Zeitpunkten Messungen an denselben Objekten vorgenommen worden sind. Hat eine Panelstudie m Wellen (W) und k Variablen (V), lässt sie sich bezeichnen als $mWkV$ -Panel.

Wird z.B. Soziologie-Studenten der Universität Duisburg über mehrere Semester immer zu Semesterbeginn derselbe Fragebogen zum Thema „Zufriedenheit mit der Universität Duisburg“ vorgelegt und handelt es sich dabei immer um dieselben Personen, dann spricht man von einer Panelstudie.

Streng genommen sind die oben eingeführten Kriterien zu hart formuliert, denn kleine Abweichungen z.B. der Items im Fragebogen von Zeitpunkt zu Zeitpunkt und eine geringe Fluktuation der Teilnehmer lässt es immer noch zu, den Begriff Panelstudie zu verwenden. In der Umfragepraxis hat es sich

¹Weitere Arten, Längsschnittdaten zu erheben (und die daraus resultierenden Konsequenzen) werden weiter unten anhand der Gegenüberstellung mit der Trendstudie kurz diskutiert, sonst werden sie in diesem Skript nicht thematisiert.

ferner eingebürgert, vom Panel schon dann zu sprechen, wenn dieselben Personen mehrmals befragt werden, auch wenn das Messinstrument von Welle zu Welle völlig unterschiedlich ist. Dies ist nicht zuletzt auf die immer stärkere Verbreitung von sog. „(Online-)Access-Panels“, meist in der Marktforschung, zurückzuführen. Der Betreiber eines Access-Panels rekrutiert einen möglichst großen und heterogenen Pool von Teilnehmern, welche über Bonus- und Incentive-Systeme zu zukünftigen Teilnahmen an diversen Befragungen motiviert werden. Verschiedene Auftraggeber können nun an den Betreiber herantreten, um Befragungen mit einer (meist nach Merkmalen gezielt ausgewählten) Teilmenge der Probanden des Access-Panels durchzuführen. Diese Praxis muss allerdings von der klassischen Vorgehensweise bei der Realisierung eines Paneldesign abgegrenzt werden. Sie wird im Skript nicht weiter berücksichtigt.

An dieser Stelle seien auszugsweise ein paar Beispiele für deutsche Panelstudien, welche nicht gerade zahlreich in der Forschungslandschaft vertreten sind, aufgeführt:

- Sozio-Ökonomisches Panel (DIW Berlin)
- Migrationsentscheidungen im Lebensverlauf und die Relevanz lokaler Bedingungen (Uni Bremen)
- Kriminalität in der modernen Stadt – Jugenddelinquenz und -devianz im Wandel von urbanen Sozialmilieus, Lebens-, Freizeit- und Konsumstilen, ethnisch-kulturellen Orientierungen und sozialer Kontrolle (Uni Münster/Bielefeld)
- Panelstudien innerhalb des DFG-Projektes „Survey Methodology“ – Kooperation mehrerer Universitäten, die sich methodischen Fragen in Bezug auf die Qualität von Umfragedaten widmet (u.a. Uni Duisburg / Bremen)

Das „Sozio-Ökonomisches Panel“ (kurz auch SOEP) ist der mit Abstand größte Paneldatensatz, welcher in Deutschland erhoben wurde und wird. Gestartet 1984 mit einer Stichprobe von knapp 6000 Haushalten (ca. 12.000 Personen) und bis heute fortgeführt, bietet er eine große Bandbreite an Variablen zu sozialen, ökonomischen und demographischen Themengebieten.

Die anderen aufgeführten Beispiele für Panelerhebungen (ausgenommen „Survey Methodology“) stellen eher verhältnismäßig kleine Erhebungen dar, die auf eine bestimmte spezielle Fragestellung ausgerichtet sind (Jugenddelinquenz bzw. Städte-Migration).

Nun soll einführend auf das Hauptthema dieser Arbeit, die *Panelanalyse*, eingegangen werden.

Der Begriff **Panelanalyse** impliziert, dass es eine *einheitliche* Form der Analyse von Paneldaten gibt, die exklusiv nur für solche Daten existiert. Dem ist nicht so. Deshalb ist dieser Begriff etwas irreführend. Zunächst einmal sei erwähnt, dass der Begriff der Panelanalyse als ein weit reichender Sammelbegriff zu verstehen ist. Darüber hinaus muss die eben angesprochene Exklusivität in Frage gestellt werden, denn oft bedient sich die Panelanalyse gängiger statistischer Techniken – z.B. aus dem Bereich der multivariaten Analyse –, welche auf die Besonderheiten von Paneldaten hin modifiziert werden. So finden im Zuge der Analyse bekannte Analyseformen und -modelle wie „Regressions- und Pfadanalyse“, „Strukturgleichungsmodelle“, „Analyse latenter Klassen“ etc. ihren Platz.

Es müssen im Vorfeld also die gängigen Vorüberlegungen stattfinden, welche denen im Falle einmalig erhobener Daten gleichen:

- Welches Skalenniveau weisen meine Variablen auf und welche Verfahren sind auf die vorhandenen Skalenniveaus abgestimmt?
- Bleibe ich auf der Ebene der empirischen Variablen oder vermute ich, dass einige meiner Variablen als Indikatoren zu sehen sind, hinter denen sich latente Größen verbergen?
- Habe ich fundierte Annahmen darüber, ob es gerichtete Zusammenhänge zwischen Variablen gibt, welche es mir erlauben, asymmetrische Analyseverfahren zu verwenden?²

Zu diesen gängigen Vorüberlegungen gesellen sich weitere Aspekte hinzu, die sich speziell auf Paneldaten beziehen. Dazu gehören:

- Die Möglichkeit, Entwicklungen zu analysieren, da Daten zu mehreren Zeitpunkten vorliegen; die Unterscheidbarkeit von Variablen wird somit

²Hier bezieht sich der Autor auf verschiedene Variablen, die zu einem Zeitpunkt gemessen werden - diese Problematik wird natürlich bei der Panelanalyse um weitere Aspekte beträchtlich erweitert (mehr dazu im Verlauf des Texts).

komplexer, da nun Unterscheidungen in Hinblick auf den *Inhalt* der Variablen und in Hinblick auf den *Messzeitpunkt* der Variablen möglich sind³.

- Die Möglichkeit, intraindividuelle Prozesse zu analysieren, da Daten zu mehreren Zeitpunkten an *ein und denselben* Personen gemessen worden sind

Der erste Punkt macht zugleich den Reiz der Panelanalyse als auch ihre Komplexität aus. Denn Veränderung ist aus statischer Perspektive eine emergente Erscheinung, welche neuer Herangehensweisen bei der Analyse bedarf, die sich aus dieser Perspektive (und so arbeitet der Statistiker, wenn er mit den gängigen Methoden Querschnittsdaten analysiert) selbst nicht ergeben. Deswegen, auch wenn die Panelanalyse z.T. auf gewohnte Analyseverfahren zurückgreift, können diese Verfahren durch die Implementierung der Analyse von Prozessen beträchtlich verkompliziert werden.

Der zweite Punkt stellt die Vorteile des Paneldesigns gegenüber anderen Längsschnitterhebungen dar. Bei einer Trendstudie z.B. werden ebenfalls zu mehreren Zeitpunkten die gleichen Merkmale aus der gleichen Grundgesamtheit erhoben, allerdings mit von Zeitpunkt zu Zeitpunkt variierenden Stichproben.

So können Aussagen über Wandel und Stabilität nur auf der Aggregatebene getätigt werden. Datensätze aus Trendstudien können nur getrennt analysiert werden, da eine Verknüpfung der Elemente zwischen den Datensätzen aufgrund der unterschiedlichen Stichprobenszusammensetzung nicht möglich ist. Intraindividuelle Veränderungen sind somit nicht identifizierbar. Dies kann zu einem beträchtlichen Informationsverlust führen.

Man denke an folgendes Beispiel: Es gibt in der Grundgesamtheit zwei politische Parteien, Partei A und Partei B. Alle Individuen aus der Grundgesamtheit gehören einer Partei an und können problemlos jederzeit die Zugehörigkeit wechseln.

Zum ersten Messzeitpunkt gehören Partei A 70% und Partei B 30% der Grundgesamtheit an. Zum zweiten Zeitpunkt findet man dasselbe Verhältnis wieder. Es wird daraus geschlossen, dass sich nichts verändert hat, dass also

³Die bekannte Datenmatrix mit m Objekten und n inhaltlich unterschiedlichen Variablen lässt sich so um eine dritte Dimension t erweitern, welche für die Zeitpunkte steht; es resultieren insgesamt $n \cdot t$ Variablen und - bei einem vollständigen Datensatz $m \cdot n \cdot t$ Daten; betrachtet man nur die $n \cdot t$ -Matrix (t spaltenweise), dann erklärt sich auch die Herkunft des Begriffs, da diese (bei wenigen Zeitpunkten und vielen Variablen) wie ein „Panel“ aussieht.

„alles beim Alten“ ist. Dies mag zwar für die simple Betrachtung von Anteilswerten stimmen. Wenn kein tieferes Verständnis für die Strukturen und Dynamiken innerhalb des Aggregats beansprucht wird, mag dieser Befund ausreichen.

Wird hingegen ein tieferes Verständnis angestrebt, dann müssen Abläufe auf „intra-individueller“ Ebene betrachtet werden. Denn es ist im obigen Beispiel vorstellbar, dass in der Zwischenzeit innerhalb der Parteien eine hohe Mitglieder-Fluktuation vorherrschte. Im Extremfall könnten alle Mitglieder der Partei B zum ersten Zeitpunkt zwischenzeitlich zur Partei A und genug Partei-A-Mitglieder zu Partei B gewechselt sein, so dass dieses 70/30-Verhältnis aufrechterhalten wurde. Die wechselseitigen Wanderungen hätten sich in diesem Fall somit gegenseitig kompensiert – ein Prozess, der anhand des Vergleichs der Anteilswerte nicht sichtbar wird.

An diesem Beispiel wird auch deutlich, dass wenn von „intra-individuellen“ Veränderungen die Rede ist, nicht etwa die Veränderung der Merkmale konkreter Einzelfälle (z.B. die Gehaltsänderung von Herrn XY) im Fokus steht. Vielmehr sind auch hier statistische Kenngrößen auf Aggregatebene relevant - z.B. der Anteil der Wechselwähler im Kontext zweier Bundestagswahlen. Solche Erkenntnisse über Bewegungen innerhalb eines Aggregat können allerdings, wie oben erwähnt, aufgrund der fehlenden Verknüpfbarkeit der Datensätze einer Trendstudie, nur anhand von Paneldaten ermittelt werden.⁴

Neben diesen erweiternden Möglichkeiten der Panelanalyse, welche von großem theoretischen Reiz sein können, gibt es Besonderheiten, welche eher vom Nachteil sind. Diese beziehen sich vor allem auf die Datenerhebung. Es ist einleuchtend, dass eine Erhebung, welche zu mehreren Zeitpunkten durchgeführt wird, aufwändiger, kostenintensiver und fehleranfälliger ist. Vor allem macht sich diese Problematik durch die Eigenheit bemerkbar, an denselben Personen (Objekten) zu mehreren Zeitpunkten Messungen durchzuführen. Personen können nach einem oder mehreren Zeitpunkten ihre Teilnahme verweigern oder sie fallen aus anderen Gründen weg (Panelmortalität). Das Dramatische daran ist, dass solche Ausfälle in Panelstudien eine größere Tendenz als bei Querschnittserhebungen dazu haben, systematisch zu sein. Andererseits, im Falle einer Befragung (welche immer die Gefahr der Reaktivität in sich birgt), könnten Befragte auf die Befragungssituation späterer Zeit-

⁴Natürlich lassen sich in Querschnitts- und Trendstudien Variablenwerte retrospektiv erfragen und mit aktuellen Variablenwerten vergleichen (z.B. Parteiwahl bei der aktuellen und bei der vorherigen Landtagswahl). Allerdings sind solche retrospektiven Messungen hinsichtlich ihrer Reliabilität und Validität ziemlich problematisch.

punkte anders reagieren, als dies im Falle einmaliger Befragung geschehen wäre (Paneleffekte). Daraus könnten verzerrte, invalide Messungen resultieren. Diese Besonderheiten führen zu weiteren Maßnahmen, die Phase der Datenerhebung zu optimieren und an Datensätze mit fehlenden Werten bei der Analyse adäquat heranzugehen.

In diesem Skript wird diese Problematik nicht weiter vertieft, es kann aber auf weiterführende Literatur verwiesen werden: Schnell (2005: Kap. 5.3.2.2.), Arminger (1990: Kap. 4), Hsiao (2005: Kap. 9), Engel (2004: Kap. 5).

Auch treten Probleme in der Datenanalyse auf. Dies betrifft vor allem den Sachverhalt, dass die in zwei verschiedenen Wellen erhobenen Variablenwerte ein und derselben Person zum inhaltlich gleichen Merkmal nicht mehr als zwei *unabhängige* Realisierungen eines Zufallsexperimentes verstanden werden können. Die Annahme der stochastischen Unabhängigkeit ist aber wichtig im Kontext verschiedener statistischer Verfahren. Diese Problematik wird im Zuge einiger Anwendungen im Skript vertieft.

Zum weiteren Verlauf des Skripts:

Wie bereits erwähnt, stellt den Kern dieser Arbeit ein Auszug von Möglichkeiten dar, Paneldaten unter der Berücksichtigung ihrer Besonderheiten zu analysieren. Hierbei wird hauptsächlich auf *lineare Modelle* eingegangen.

Einige Grundlagen, der Regressionsanalyse z.B., müssen vorausgesetzt werden. Hinweise zu diesem Thema für Einsteiger sind im Anhang 4.7 zu finden. Grundlagen der Pfadanalyse, welche an die Regression anknüpft, werden kurz im Kapitel 2.2 abgehandelt.

Es wird ferner gezeigt, dass Paneldaten einerseits dazu genutzt werden, *Entwicklungen zu analysieren*, und andererseits neue Möglichkeiten für *statistisch-theoretische Überprüfungen* von Modellen, z.B. auf Fehlspezifikationen hin, bieten.

Abgerundet wird diese Arbeit mit einem kurzen zusammenfassenden Überblick (Kapitel 3).

Kapitel 2

Ausgewählte Analyseverfahren für Paneldaten

2.1 Elementare Veränderung der Werte einer Variablen im Zeitverlauf

Veränderung, Wandel, Entwicklung, Prozess – alles Begriffe, die sich auf die zeitliche Abfolge von Zuständen, Beschaffenheiten, Strukturen und Konstellationen beziehen. Gerade diese zeitliche Abfolge übt in vielen Wissenschaften (und so ist es auch in den Sozialwissenschaften) einen gewissen Reiz aus, denn viele Phänomene sind von ihrer Geschichte entkoppelt nicht zu verstehen.

Ein trauriges Beispiel stellt hierbei die grausame Vernichtung und Zerstörung von Menschenleben und Materiellem im 2. Weltkrieg dar. Ohne die vorangehende geschichtliche Entwicklung zu studieren, ist das Ausmaß dieser Katastrophe nicht zu begreifen (danach sicherlich auch nicht im Sinne von *Legitimierung*, aber man versteht einige Zusammenhänge besser).

Des Weiteren sind Wissenschaftler stets bemüht, festgestellte Zusammenhänge als Kausalitäten zu interpretieren – und eines der wichtigsten Kausalkriterien ist bekanntermaßen das *zeitliche Vorgehen* einer Ursache, so dass auch hier die Zeitdimension Eingang findet. Der Sozialwissenschaftler steht aber vor dem Dilemma, dass die zeitliche Abfolge von Effekten oft nicht kontrollierbar ist. Denn er arbeitet häufig mit Merkmalen, die sich nicht manipulieren lassen, wie z.B. Geschlecht, Intelligenz, Schichtzugehörigkeit etc., so dass bei der Erforschung von vielen sozialwissenschaftlichen Fragestellungen nicht auf experimentelle Designs zurückgegriffen werden kann.

Auch wenn in Panelstudien im Vergleich zu echten Experimenten keine Möglichkeit besteht, den Stimulus (also die unabhängige Variable) zu manipulieren und die Folgen zu studieren, so führt das Paneldesign durch wiederholte Messungen den zeitlichen Aspekt ein, welcher die Aussagekraft in Hinblick auf Veränderungen deutlich gegenüber der in Querschnittsanalysen verbessert.

Zuerst soll nun gefragt werden, wie eine Veränderung statistisch zu begreifen bzw. zu analysieren ist. Ohne diesen Begriff explizit zu definieren sei außerdem erwähnt, dass dieser Begriff von dem der *Stabilität* zu einem Gegensatzpaar ergänzt wird. Je mehr sich also ein Merkmal über die Zeit verändert, umso weniger Stabilität weist es auf – und umgekehrt.

2.1.1 Der Korrelationskoeffizient als Indikator für Stabilität und Veränderung?

Wenn sich die Werte einer Variablen x über die Zeit verändern, dann kann das verschiedene Ursachen haben. Regressionsanalytisch ausgedrückt können hier verschiedene unabhängige Variablen signifikanten Einfluss auf x haben.

Aber zum Einstieg soll die einfachste Form des Einflusses betrachtet werden: Eine Variable beeinflusst ihre Werte über die Zeit durch sich selbst. Wenn z.B. das Einkommen von Personen zum Zeitpunkt t abgefragt wird, dann kann man erwarten, dass dieses Einkommen zum Zeitpunkt $t + 1$ konstant bleibt. Wird diese Erwartungshaltung erfüllt, so liegt bzgl. des Einkommens Stabilität vor, ist das Gegenteil der Fall, so hat sich etwas geändert.¹

Doch wie macht sich denn Veränderung statistisch bemerkbar? Ein intuitiver Gedanke wäre, den Korrelationskoeffizienten zwischen einer inhaltlich gleichen Variablen zum Zeitpunkt t und dem Zeitpunkt $t + 1$ zu bilden. Je höher er ausfallen würde, umso eher hingen die zwei Variablen zusammen und durch umso mehr Stabilität wären sie folglich gekennzeichnet. Doch dieser Gedanke ist nur auf den ersten Blick plausibel. Denn eine hohe Korrelation *kann* zeitliche Stabilität zum Ausdruck bringen, aber eben nicht nur, sondern ebenso eine *gleichmäßige, proportionale Veränderung*, egal wie stark sie ist.

¹Wobei dann noch lange nicht davon auszugehen ist, dass das Einkommen zum früheren Zeitpunkt selber *Ursache* für den Wandel ist – es ist eher anzuzweifeln. Ein plausibleres Beispiel wäre das „rich-get-richer“-Phänomen, welches impliziert, dass mit zunehmendem materiellen Reichtum Personen immer stärker dazu tendieren, über die Zeit diesen Reichtum weiter auszubauen.

Dies sei am folgenden Beispiel illustriert: Es wird als Variable das Einkommen der Mitarbeiter verschiedener Abteilungen eines Großunternehmens erhoben. Der Datensatz enthält Daten von dem Geringverdiener bis zum Manager mit einem Spitzengehalt. Wiederholt man diese Erhebung im gleichen Unternehmen ein Jahr später und es hat sich kaum etwas bzgl. der Gehälter geändert, so würde der Korrelationskoeffizient nahe bei 1 liegen. Hätte allerdings in der Zwischenzeit das Unternehmen beschlossen, alle Löhne um 20% zu erhöhen, dann würde durch diesen proportionalen Anstieg die Korrelation ebenfalls nahe bei 1 liegen, obwohl sich sehr wohl Einiges geändert hat.

Um diese beiden Sachverhalte voneinander zu trennen, soll nun die Zerlegung der Korrelation einer Variablen x zum Zeitpunkt 1 (x_1) mit derselben Variablen x zum Zeitpunkt 2 (x_2) betrachtet werden.

Da sich die bivariate Korrelation zweier Merkmale aus der Kovarianz dieser Merkmale (geteilt durch das Produkt der Standardabweichung der einzelnen Variablen) errechnen lässt, wird hier zuerst die Zerlegung der Kovarianz der Merkmale betrachtet. Dabei wird eine neue Variable Δx eingeführt, welche die Differenz zwischen x_1 und x_2 und somit die Veränderung der x -Werte eines Individuums zwischen den beiden Zeitpunkten zum Ausdruck bringt ($\Delta \mathbf{x} = \mathbf{x}_2 - \mathbf{x}_1$). Die Zerlegung gestaltet sich wie folgt (Vgl. Kessler 1981: 9):

$$Cov_{x_1 x_2} = Var_{x_1} + Cov_{x_1 \Delta x} \quad (2.1)$$

„Cov“ steht für Kovarianz und „Var“ für Varianz.

Nachweis:

Da $\Delta x_i = x_{2i} - x_{1i}$ und die Kovarianz von x_1 und $x_2 = \frac{\sum_{i=1}^n [(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)]}{n}$ (wobei i = Laufindex der einzelnen Werte von x_1 und x_2 , n = der höchste Wert von i . \bar{x}_1 bzw. \bar{x}_2 sind die arithmetischen Mittel von x_1 bzw. x_2):

$$\begin{aligned}
Cov_{x_1x_2} &= \sum_{i=1}^n [(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)] \\
&= \sum_{i=1}^n [(x_{1i} - \bar{x}_1)((x_{1i} + \Delta x_i) - \overline{(x_1 + \Delta x)})] \\
&= \sum_{i=1}^n [(x_{1i} - \bar{x}_1)((x_{1i} + \Delta x_i) - (\bar{x}_1 + \overline{\Delta x}))] \\
&= \sum_{i=1}^n [(x_{1i} - \bar{x}_1)(x_{1i} + \Delta x_i - \bar{x}_1 - \overline{\Delta x})] \\
&= \sum_{i=1}^n [x_{1i}^2 + x_{1i}\Delta x_i - x_{1i}\bar{x}_1 + x_{1i}\overline{\Delta x} - x_{1i}\bar{x}_1 - \bar{x}_1\Delta x_i + \bar{x}_1^2 + \bar{x}_1\overline{\Delta x}] \\
&= \sum_{i=1}^n [x_{1i}^2 - 2(x_{1i}\bar{x}_1) + \bar{x}_1^2 + x_{1i}\Delta x_i + x_{1i}\overline{\Delta x} - \bar{x}_1\Delta x_i + \bar{x}_1\overline{\Delta x}] \\
&= \sum_{i=1}^n [(x_{1i} - \bar{x}_1)^2 + (x_{1i} - \bar{x}_1)(\Delta x_i - \overline{\Delta x})] \\
&= \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^n [(x_{1i} - \bar{x}_1)(\Delta x_i - \overline{\Delta x})] \\
&= Var_{x_1} + Cov_{x_1\Delta x}
\end{aligned}
\tag{2.2}$$

Da an jeder Stelle der Gleichung durch ein konstantes n dividiert wird, wurde aus Gründen der Übersichtlichkeit die Division durch n ausgelassen.

Hierbei ist der erste Summand Var_{x_1} als der Teil der Kovarianz zu verstehen, welcher die zeitliche Stabilität in den Werten der Variablen X zum Ausdruck bringt. Das ist logisch, wenn man bedenkt, was passieren würde, wenn der zweite Summand auf Null gesetzt wird. Dann ist nämlich die gesamte Kovarianz von x_1x_2 auf die Varianz von x_1 zurückzuführen. Es hätte sich demnach „nichts geändert“.

Der zweite Summand $Cov_{x_1\Delta x}$ schließlich drückt das gemeinsame Variieren der Variablen X zum ersten Zeitpunkt mit der Differenzvariablen aus. Er kann somit als der Teil angesehen werden, welcher die Veränderung zum Ausdruck bringt. So kann die Veränderung einer Variablen über zwei Zeitpunkte

in Hinblick darauf beurteilt werden, welchen Anteil die beiden Summanden an der Kovarianz $Cov_{x_1x_2}$ haben.

Die Korrelation lässt sich nun auf der Basis der eingeführten Kovarianzzerlegung wie folgt darstellen („ r “ steht für den Korrelationskoeffizienten und „ s “ für die Standardabweichung einer Variablen):

$$r_{x_1x_2} = \frac{Var_{x_1} + Cov_{x_1\Delta x}}{s_{x_1} \cdot s_{x_2}} \quad (2.3)$$

Anhand des folgenden Beispiels mit fiktiven Daten soll gezeigt werden, wie drei Variablenpaare jedes Mal einen hohen bivariaten Korrelationskoeffizienten aufweisen, sich aber in der Zusammensetzung ihrer Kovarianz (also in der Stabilität ihrer Werte) stark unterscheiden (Der Datensatz ist im Appendix 4.1 abgelegt).

Korrelationen

		X1	X2WENIG	X2MITTEL	X2VIEL
X1	Korrelation nach Pearson	1	,987**	,999**	,999**
	Signifikanz (2-seitig)	.	,000	,000	,000
	N	20	20	20	20
X2WENIG	Korrelation nach Pearson	,987**	1	,986**	,986**
	Signifikanz (2-seitig)	,000	.	,000	,000
	N	20	20	20	20
X2MITTEL	Korrelation nach Pearson	,999**	,986**	1	,998**
	Signifikanz (2-seitig)	,000	,000	.	,000
	N	20	20	20	20
X2VIEL	Korrelation nach Pearson	,999**	,986**	,998**	1
	Signifikanz (2-seitig)	,000	,000	,000	.
	N	20	20	20	20

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Abbildung 2.1: Korrelationsmatrix

Korrelationen

		X1	X1X2WENI	X1X2MITT	X1X2VIEL
X1	Korrelation nach Pearson	1	,287	,998**	,999**
	Signifikanz (2-seitig)	.	,220	,000	,000
	Quadratsummen und Kreuzprodukte	665,000	34,500	2632,000	16022,000
	Kovarianz	35,000	1,816	138,526	843,263
	N	20	20	20	20
X1X2WENI	Korrelation nach Pearson	,287	1	,291	,286
	Signifikanz (2-seitig)	,220	.	,213	,222
	Quadratsummen und Kreuzprodukte	34,500	21,750	139,000	828,500
	Kovarianz	1,816	1,145	7,316	43,605
	N	20	20	20	20
X1X2MITT	Korrelation nach Pearson	,998**	,291	1	,997**
	Signifikanz (2-seitig)	,000	,213	.	,000
	Quadratsummen und Kreuzprodukte	2632,000	139,000	10461,200	63402,400
	Kovarianz	138,526	7,316	550,589	3336,968
	N	20	20	20	20
X1X2VIEL	Korrelation nach Pearson	,999**	,286	,997**	1
	Signifikanz (2-seitig)	,000	,222	,000	.
	Quadratsummen und Kreuzprodukte	16022,000	828,500	63402,400	386445,8
	Kovarianz	843,263	43,605	3336,968	20339,253
	N	20	20	20	20

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Abbildung 2.2: Kovarianzmatrix

Es wurden hierzu eine $X1$ -Variable und drei „mögliche Partner“, also drei potentielle $X2$ -Variablen gebildet. Wie im Datensatz (s. 4.1) zu sehen ist, weicht die Variable „ $X2$ -WENIG“ kaum von den Werten von $X1$ ab, während die Variable „ $X2$ -MITTEL“ stärkere und „ $X2$ -VIEL“ sehr starke Diskrepanzen aufweist.²

Allerdings handelt es sich hierbei um annähernd proportionale Abweichungen. Dies manifestiert sich in den jeweiligen sehr hohen bivariaten Korrelationen, wie in Abb. 2.1 zu sehen ist.

Nun können die Kovarianzen in Abb. 2.2 (in der Zeile „Kovarianz“) zwischen der $X1$ -Variablen und der jeweiligen Differenzvariablen Δ betrachtet werden. Diese sind schließlich nach Gl. 2.3 Bestandteil der Korrelations- bzw. Kovarianzzerlegung zwischen $X1$ und der jeweiligen $X2$ -Variablen.

Der erste Summand des Zählers in der Gl. 2.3, nämlich die Varianz von $X1$ ($Var_{X1} = 35$), ist für alle drei Fälle konstant. Was variiert, ist hingegen der zweite Summand, nämlich $Cov_{x_1 \Delta x}$, je nachdem, welche $X2$ -Variable mit

²Die drei weiteren Variablen sind als die Differenzvariablen (Δ) von $X1$ und der jeweiligen $X2$ -Variablen definiert.

X_1 gepaart wird. Es ist zu sehen, dass mit immer größerer Diskrepanz diese Kovarianz drastisch zunimmt:

- **Für „ X_2 -WENIG“: 1,8**
- **Für „ X_2 -MITTEL“: 138,5**
- **Für „ X_2 -VIEL“: 843,3**

Mit der Zunahme der Kovarianz zwischen X_1 und Δ nimmt auch die Dominanz dieses Summanden an der Summe im Zähler von Gl. 2.3 zu. Var_{X_1} bleibt schließlich, wie bereits erwähnt, immer gleich.

Wenn also die Kovarianz zwischen zwei Variablen *von der Kovarianz einer der beiden Variablen mit der Differenzvariablen dominiert wird*, dann ist dieser Befund ein Indikator dafür, dass trotz eines möglicherweise hohen Korrelationskoeffizientes *Veränderung* (auf proportionaler Ebene) stattgefunden hat.

Abschließend sei nochmals darauf hingewiesen, dass die Ausführungen hier sich nur auf *relativ hohe Korrelationen* beziehen und somit lediglich unterschieden wird zwischen *hoher Stabilität* und *proportionaler (!!!) Veränderung*. Die Frage nach „Veränderung an sich“ wurde hier nicht behandelt, es ist aber logisch, dass ein niedriger Korrelationskoeffizient auf mangelnde Stabilität und somit auf nicht-lineare Veränderung hindeutet.

2.1.2 Die einfache Regression einer zeitlich vorangestellten inhaltlich gleichen Variablen auf die Variable zum späteren Zeitpunkt

Die vorhin diskutierte Frage der Beziehung zwischen zwei inhaltlich gleichen Variablen gemessen zu zwei verschiedenen Zeitpunkten lässt sich auch in Form einer einfachen bivariaten Regression spezifizieren. In diesem Falle ist eine gerichtete Beziehung zu benennen und zwar mit x_1 als der unabhängigen und x_2 als der abhängigen Variablen. Die Regressionsgleichung lässt sich wie folgt formalisieren (auf den Index für einzelne Werte wurde aus Veranschaulichungsgründen verzichtet):

$$x_2 = a + b_1x_1 + e$$

mit:

a =Regressionskonstante

b_1 =unstandardisierter Regressionskoeffizient der unabhängigen Variablen

e =Fehlerterm

Der Regressionskoeffizient ist im Vergleich zum Korrelationskoeffizienten ein unstandardisiertes und asymmetrisches Maß (letzteres weil die Kovarianz nur durch die Varianz der *unabhängigen* Variablen geteilt wird). Im Falle bivariater Regression lässt er sich einfach berechnen durch:

$$b_{x_1} = \frac{Cov_{x_1x_2}}{Var_{x_1}}$$

Nach der in Kapitel 2.1.1 eingeführten Aufteilung lässt sich die Kovarianz zerlegen. Dies eingesetzt ergibt:

$$b_{x_1} = \frac{Var_{x_1} + Cov_{x_1\Delta x}}{Var_{x_1}} = \frac{Var_{x_1}}{Var_{x_1}} + \frac{Cov_{x_1\Delta x}}{Var_{x_1}} = 1 + \frac{Cov_{x_1\Delta x}}{Var_{x_1}}$$

Der zweite Summand $\frac{Cov_{x_1\Delta x}}{Var_{x_1}}$ ist hierbei nichts anderes als der Regressionskoeffizient der unabhängigen Variablen auf die Differenzvariable Δx , im Folgenden bezeichnet mit \hat{b}_{x_1} . So kann man für die ursprüngliche Gleichung schreiben:

$$x_2 = a + (1 + \hat{b}_{x_1})x_1 + e$$

$$x_2 = a + x_1 + \hat{b}_{x_1}x_1 + e$$

$$x_2 - x_1 = a + \hat{b}_{x_1}x_1 + e$$

$$\boxed{\Delta x = a + \hat{b}_{x_1}x_1 + e} \quad (2.4)$$

Die letzte Gleichung stellt die Regressionsgleichung mit x_1 als der unabhängigen und Δx als der abhängigen Variablen dar. In dieser Gleichung wird thematisiert, inwieweit die Veränderung in der X -Variable (präziser: die Streuung der Differenzvariablen), durch die Werte der X -Variable, gemessen zum ersten Zeitpunkt, erklärt werden kann. Da Folgendes gilt: $\mathbf{b}_{x_1} = \mathbf{1} + \hat{\mathbf{b}}_{x_1}$, kann der Regressionskoeffizient der Δx -Regression aus dem ursprünglichen Koeffizienten b_{x_1} einfach berechnet werden.

Die gerade besprochenen Regressionskoeffizienten sind unstandardisiert, deswegen ist prinzipiell kein Wertebereich anzugeben, der starke Veränderung oder hohe Stabilität ausdrückt. Sie zu standardisieren würde aber im bivariaten Fall wiederum die Werte des Korrelationskoeffizienten liefern, so dass man vor dem ursprünglichen Problem zu Beginn des Kapitels 2.1.1 stünde.

Allerdings kann man etwas über perfekte Stabilität sagen: Ist $b_{x_1} = 1$, so ergibt sich für $\hat{b}_{x_1} = 0$. Geht man von einer fehlerfreien Regression aus, bei welcher der Fehlerterm vernachlässigt werden kann, so wären in dem Fall die Differenzwerte (Werte von Δx) allein durch die Regressionskonstante bestimmt. Gilt weiter für diese $a = 0$, so hätte die Differenzvariable durchgehend die Ausprägung 0 und es hätte sich demzufolge zwischen den Werten von x_1 und x_2 nichts geändert.

Betrachtet man dann wieder die ursprüngliche Regression mit x_2 als abhängige Variable, so erscheint dies logisch, denn bei einer Regressionskonstanten $a = 0$, dem Koeffizienten $b_{x_1} = 1$ und einem zu vernachlässigenden Fehlerterm würden sich die x_2 -Werte direkt aus den x_1 -Werten ergeben.

Trotz fehlender Standardisierung lassen sich die Regressionskoeffizienten dann vergleichen, wenn den Variablen immer gleiche Messeinheiten zu Grunde liegen. Bezogen auf das obige Beispiel in Kap.2.1.1 ergeben sich folgende Regressionsgleichungen:

$$X2WENIG = -0,295 + 1,052 \cdot X1 + e$$

$$X2MITTEL = 0,6421 + 4,958 \cdot X1 + e$$

$$X2VIEL = -1,079 + 25,093 \cdot X1 + e$$

Man sieht, wie mit zunehmender Diskrepanz zwischen den x_1 - und x_2 -Werten, also mit zunehmender Kovarianz von x_1 und x_2 der Regressionskoeffizient stetig ansteigt.

Allerdings sei nochmals darauf hingewiesen: Dieser Zusammenhang gilt lediglich dann, wenn die Abweichung der x_2 -Werte von den x_1 -Werten *gleichmäßig, also proportional zunimmt*. Ist dies nicht der Fall, kovariieren die Werte nicht gleichmäßig, so kann die Kovarianz – und in der Folge der Regressionskoeffizient – sehr niedrig ausfallen.

Spätestens dann muss die univariate Ebene verlassen werden, denn es muss

weitere Variablen geben, welche die Veränderung in X beeinflussen.³ Solchen Sachverhalten widmen sich die nächsten Abschnitte.

2.2 Eine kurze Einführung in die Pfadanalyse

Einige der im Verlauf des Skripts vorgestellten Verfahren enthalten Elemente der Pfadanalyse, so dass es zweckmäßig ist, an dieser Stelle ihre elementare Funktionsweise nahe zu bringen.

Die Pfadanalyse ermöglicht es, komplexe Zusammenhänge zwischen mehreren Variablen zu modellieren. In Regressionsmodellen z.B. formuliert man Zusammenhänge zwischen *einer* abhängigen und mehreren unabhängigen Variablen. Je nachdem, wie viele unabhängige Variablen einbezogen werden, nimmt das Modell an Komplexität zu. Nun lassen sich aber darüber hinaus noch komplexere Modelle denken. Solche Modelle können von der Pfadanalyse aufgegriffen werden, sie erweitern den Regressionsansatz um folgende Punkte:

- Es kann mehr als eine abhängige Variable formuliert werden
- Es können Variablen einbezogen werden, welche in Bezug auf einen Teil der anderen Variablen als abhängige, und in Bezug auf einen zweiten Teil der Variablen als unabhängige Variablen fungieren (somit können indirekte Einflüsse zwischen Variablen, die über mehrere zwischenliegende Variablen vermittelt werden, mitmodelliert werden)
- Es werden sowohl gerichtete als auch ungerichtete Beziehungen zugelassen

Da solche Zusammenhänge schnell kompliziert werden können, lassen sie sich über Pfaddiagramme visualisieren. Folgende Spielregeln gelten:

- Zwei zusammenhängende Variablen werden durch Pfeile (auch: Pfade) verbunden

³Und wie schon an anderer Stelle angemerkt: Selbst bei hoher Stabilität kann die Variable X zum Zeitpunkt t_1 nicht unreflektiert als eine kausale Wirkung auf sich selbst zum Zeitpunkt t_2 angenommen werden.

- Bei einer gerichteten Beziehung zeigt der Pfeil auf die abhängige Variable, im Falle der Ungerichtetheit zeigt der Pfeil auf beide Variablen
- Variablen, auf die kein einseitig gerichteter Pfeil zeigt, nennt man **exogene Variablen**, alle anderen bezeichnet man als **endogene Variablen**
- Über den Pfeilen stehende Werte werden **Pfadkoeffizienten** genannt (Zur Bedeutung s.u.)
- Die zwei Subskripte eines Pfadkoeffizienten symbolisieren die zwei Variablen, wobei die abhängige Variable als erstes gelistet wird (p_{yx} würde also bedeuten, dass y die abhängige und x die unabhängige Variable ist); bei ungerichteten Beziehungen ist die Reihenfolge bedeutungslos

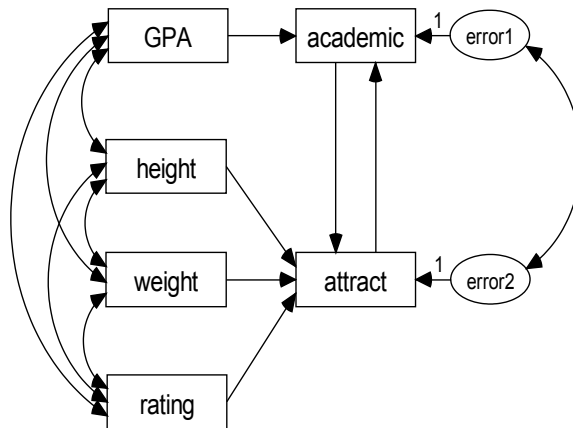
Wie bei Regressionsmodellen auch, werden Annahmen über die Verbindungen der Variablen (durch Pfeile) a priori gemacht, also bevor konkrete Berechnungen beginnen. Auch der Wert mancher Pfadkoeffizienten wird a priori festgelegt (das kann z.B. die Annahme der Unkorreliertheit zwischen Residuen und endogenen Variablen betreffen).

Ein Beispiel für ein Pfaddiagramm mit acht Variablen (sechs gemessene Variablen + zwei Residualvariablen) sei in Abbildung 2.3 vorgestellt. Aus Übersichtlichkeitsgründen wurden die Pfadkoeffizienten nicht abgetragen. Das Diagramm stellt ein Modell von Felson and Bohrnstedt (1979) dar, in dem die von Schulkindern *wahrgenommene physische Attraktivität* und *wahrgenommene Schulleistung* der Mitschüler thematisiert worden ist (Auf die Inhalte dieser Studie wird im Weiteren nicht eingegangen).

Nun ist zu fragen, wie die einzelnen Pfadkoeffizienten berechnet werden und welche Bedeutung sie haben. Zuerst muss man wissen, dass Pfadmodelle als mehrere „miteinander verbundene Regressionsmodelle“ (Engel 1994: 22) zu denken sind.

Die Pfadkoeffizienten drücken die Effekte einzelner Variablen aufeinander, unter Berücksichtigung der Einflüsse weiterer einbezogener Merkmale, aus – es sind sozusagen bereinigte Effekte.

Denn das Problem bivariater Korrelation (nullter Ordnung) zwischen zwei Variablen ist, dass sie einen Zusammenhang ausdrückt, in welchem die Einflüsse weiterer außenstehender Variablen beinhaltet sein können. Diese sollen



GPA = Grade Point Average
height = Deviation of height from mean by grade and sex
weight = Weight adjusted for height
rating = Physical attractiveness rated by children outside class
academic = Perceived academic ability, based on class-mates' ratings
attract = Perceived attractiveness, classmates' ratings

beispiel.pdf

Abbildung 2.3: Beispiel für ein Pfaddiagramm

aus den Pfadkoeffizienten herausgehalten werden; wobei das nur für Variablen möglich ist, welche im Modell aufgenommen werden.

Deswegen ist die schwierigste Aufgabe, überhaupt erst ein geeignetes Modell aufzustellen. Dazu gehört Fingerspitzengefühl. Die darauf folgenden Berechnungen sind dann wiederum standardisierte Prozesse.⁴ Diesen wollen wir uns nun an einem einfachen Beispiel widmen. Es wird folgendes Modell betrachtet (vgl. Opp 1976: 134 ff):

⁴Fragen nach der Modellanpassung werden an dieser Stelle erst mal ausgeklammert.

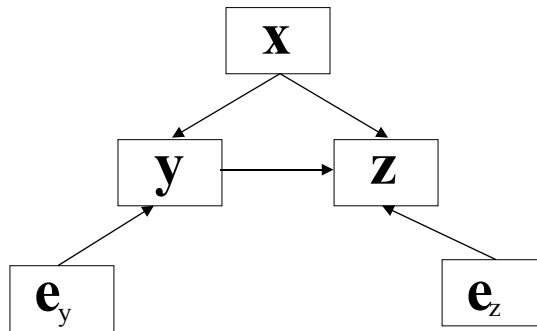


Abbildung 2.4: Pfaddiagramm des Beispiels

x, y und z sind beliebige gemessene Variablen
 e_y ist das zur Variablen y zugehörige Residuum
 e_z ist das zur Variablen z zugehörige Residuum

Zuerst soll davon ausgegangen werden, dass die Variablen z -transformiert sind, um bestimmte Berechnungen zu erleichtern. Die Regressionskonstante z.B. entfällt. Weiterhin wird angenommen, dass die Zusammenhänge zwischen den Variablen linear sind.

Nun lassen sich zwei Regressionsgleichungen aufstellen, die daraufhin ineinander verschachtelt werden. Die Residuen e_y und e_z werden als eigenständige Variablen behandelt, so dass auch ihnen Pfadkoeffizienten zugewiesen werden. Diese Pfadkoeffizienten werden gemäß oben eingeführter Konvention bezeichnet. Die Gleichungen lauten:

$$y = p_{yx}x + p_{ye_y}e_y \quad (2.5)$$

$$z = p_{zx}x + p_{zy}y + p_{ze_z}e_z \quad (2.6)$$

An dieser Stelle schätzt man nicht die einzelnen Parameter wie in der Regressionsanalyse, sondern man versucht, die einzelnen Pfadkoeffizienten durch Umformung und Verschmelzung der Gl. 2.5 und 2.6 zu ermitteln. Hierzu wird zunächst jede Gleichung mit jeder unabhängigen Variablen (außer mit den Residuen) gesondert multipliziert:

$$yx = p_{yx}x^2 + p_{ye_y}e_yx \quad (2.7)$$

$$zx = p_{zx}x^2 + p_{zy}yx + p_{ze_z}e_zx \quad (2.8)$$

$$zy = p_{zx}xy + p_{zy}y^2 + p_{ze_z}e_zy \quad (2.9)$$

Es entstehen drei Gleichungen, da Gl. 2.5 eine und Gl. 2.6 zwei unabhängige Variablen enthält.

Nun wurde bislang zur Vereinfachung auf den Fälle-Index i verzichtet. Trotzdem gilt für jede Variable, dass sie Platzhalter für die Werte einzelner Fälle ist. Im Gegensatz dazu sind die Pfadkoeffizienten für alle Fälle N konstant. Die einzelnen Variablen und Variablenprodukte lassen sich demnach zur Mittelwertsformel erweitern. Daraufhin werden die Gleichungen 2.7, 2.8 und 2.9 modifiziert:

$$\frac{\Sigma(yx)}{N} = p_{yx} \frac{\Sigma(x^2)}{N} + p_{ye_y} \frac{\Sigma(e_yx)}{N} \quad (2.10)$$

$$\frac{\Sigma(zx)}{N} = p_{zx} \frac{\Sigma(x^2)}{N} + p_{zy} \frac{\Sigma(yx)}{N} + p_{ze_z} \frac{\Sigma(e_zx)}{N} \quad (2.11)$$

$$\frac{\Sigma(zy)}{N} = p_{zx} \frac{\Sigma(xy)}{N} + p_{zy} \frac{\Sigma(y^2)}{N} + p_{ze_z} \frac{\Sigma(e_zy)}{N} \quad (2.12)$$

Dem Umstand, dass es sich um standardisierte Variablen handelt, welche also alle einen Mittelwert von Null und eine Standardabweichung von Eins haben, ist zu verdanken, dass die Mittelwerte der Variablenprodukte den bivariaten Korrelationen der jeweiligen zwei Variablen entsprechen.

Bezeichnet man die Korrelation der Variablen i und j als r_{ij} und bedenkt man, dass eine Korrelation einer Variablen mit sich selbst Eins ergibt, dann vereinfachen sich Gleichungen 2.10 - 2.12) wie folgt:

$$r_{yx} = p_{yx} + p_{ye_y}r_{e_yx} \quad (2.13)$$

$$r_{zx} = p_{zx} + p_{zy}r_{yx} + p_{ze_z}r_{e_zx} \quad (2.14)$$

$$r_{zy} = p_{zx}r_{xy} + p_{zy} + p_{ze_z}r_{e_zy} \quad (2.15)$$

Setzt man weiterhin die Annahme, dass die Residuen nicht mit den unabhängigen Variablen korrelieren, dann erfolgt eine weitere Vereinfachung:

$$r_{yx} = p_{yx} \quad (2.16)$$

$$r_{zx} = p_{zx} + p_{zy}r_{yx} \quad (2.17)$$

$$r_{zy} = p_{zx}r_{xy} + p_{zy} \quad (2.18)$$

Der Pfadkoeffizient p_{yx} entspricht der Korrelation r_{yx} . Dies sollte nicht weiter verwundern, denn in diesem hier betrachteten einfachen Modell wird die Variable y nur von x beeinflusst. Die zwei weiteren Korrelationen setzen sich dagegen aus mehreren Summanden zusammen. Um die Pfadkoeffizienten zu bestimmen, muss substituiert werden. Dies ist die Stelle, an der die Verschachtelung der zwei multiplen Regressionsmodelle beginnt.

Man löst Gl. 2.17 nach p_{zx} auf und ersetzt p_{zx} in Gl. 2.18 mit dem Ausdruck $r_{zx} - p_{zy}r_{yx}$. Nun lässt sich die neue Gleichung nach p_{zy} auflösen. Es ergibt sich:

$$p_{zy} = \frac{r_{zy} - r_{zx}r_{yx}}{1 - r_{yx}^2}$$

Analog dazu verfährt man mit p_{zx} , so dass sich im Endeffekt für die drei Pfadkoeffizienten drei Ausdrücke ergeben, welche sich aus den bivariaten Korrelationskoeffizienten zusammensetzen. Sie lauten:

$$p_{yx} = r_{yx} \quad (2.19)$$

$$p_{zy} = \frac{r_{zy} - r_{zx}r_{yx}}{1 - r_{yx}^2} \quad (2.20)$$

$$p_{zx} = \frac{r_{zx} - r_{zy}r_{yx}}{1 - r_{yx}^2} \quad (2.21)$$

Generell lässt sich sagen, dass sich die Pfadkoeffizienten, wie auch Korrelationskoeffizienten, in einem standardisierten, interpretierbaren Intervall

$-1 \leq p \leq +1$ bewegen.

Gleichung 2.19 wurde bereits oben diskutiert. An den anderen Gleichungen 2.20 und 2.21 sieht man, dass die jeweiligen Pfadkoeffizienten die entsprechenden bivariaten Korrelationen bereinigen.

Betrachtet man z.B. für Gl. 2.20 nur den Zähler, dann sieht man, dass von der Korrelation r_{zy} das Produkt der anderen Korrelationen, nämlich $r_{zx}r_{xy}$ subtrahiert wird. Somit sieht man schon hier, dass je dominanter die eigentliche Korrelation der im Pfadkoeffizienten betrachteten Variablen gegenüber den weiteren Korrelationen ist, umso größer (bei Konstanzhaltung des Nenners) der Wert des Pfadkoeffizienten wird.⁵

Geteilt wird in Gl. 2.20 durch den Anteil der **nicht** erklärten Varianz, wenn man eine einfache Regression mit den Variablen xy (also die, welche hier auf z wirken) durchführen würde.⁶ Daraus lässt sich schließen: Je höher die quadrierte Korrelation zwischen den auf z wirkenden Variablen ist (bei Konstanzhaltung des Zählers), umso größer ist p_{zy} .⁷

Somit bleibt soz. als „Bösewicht“ die Korrelation von r_{zx} übrig. Wenn sie hoch ist, ist sie in erster Linie dafür verantwortlich, wenn der Pfadkoeffizient p_{zy} auch bei hohen Werten von r_{zy} eher niedrig ausfällt. Dies ist logisch, da in unserem Beispiel die Variablen x und y um den Einfluss auf z konkurrieren. Betrachten wir den bereinigten Einfluss von y auf z (was ja bei p_{zy} der Fall ist), so wird er höher ausfallen, wenn der Einfluss von x auf z möglichst gering ist – und letzterer manifestiert sich schließlich in r_{zx} .

Nun sei noch auf zwei Extremfälle hingewiesen:

$p_{zy} = r_{zy}$, wenn $r_{xy} = 0$.

Auch dies ist logisch erschießbar, denn in diesem Fall wirken y und x völlig unabhängig voneinander auf z .

Außerdem gilt: $p_{zy} = 0$, wenn $r_{zx}r_{yx} = r_{zy}$.

Die Korrelation von z und y setzt sich in diesem Fall gewissermaßen aus den jeweiligen Korrelationen der beiden Variablen mit x zusammen. Hiernach „hat y keine Wirkung auf z , die unabhängig von dem Einfluss von x ist“ (Opp 1976: 142).

Um diese vorhin diskutierten Sachverhalte zu demonstrieren, sind im Fol-

⁵ Wobei dieses Verhältnis nicht linear ist

⁶ Die Frage nach der Richtung dieser Regression ist irrelevant, denn auch wenn die Regressionsanalyse ein asymmetrisches Verfahren ist, so ist der Determinationskoeffizient wiederum symmetrisch.

⁷ s. Fußnote 4

genden fiktive Beispiele aufgeführt:

r_{zy}	r_{xy}	r_{xz}	Zähler	Nenner	p_{zy}
0,8	0,7	0,7	0,31	0,51	0,61
0,8	0,6	0,6	0,44	0,64	0,68
0,8	0,55	0,65	0,44	0,697	0,63
0,8	0,65	0,55	0,44	0,578	0,76

Tabelle 2.1: Verschiedene Korrelationsstrukturen

Die Tabelle 2.1 zeigt vier Beispiele für unterschiedliche Werte der drei Korrelationskoeffizienten. Die Ausgangskorrelation r_{zy} wurde in allen Beispielen mit 0,8 konstant gehalten, um zu zeigen, wie sich trotz der Invarianz dieser Korrelation der Pfadkoeffizient verändert. Die Beispiele in Zeile 1 und 2 zeigen erst mal die allgemeine Tendenz, dass p_{zy} steigt, wenn die anderen zwei Korrelationen fallen.

Dann wird die Betrachtung weiter differenziert. Die Werte von r_{xy} und r_{xz} werden in Zeile 3 und 4 variiert, allerdings so, dass der Zähler immer gleich bleibt und dem in Zeile 2 entspricht. So hängt p_{zy} nur noch vom Nenner, bzw. der Höhe von r_{yx}^2 ab. r_{yx}^2 hängt wiederum im Falle eines konstanten Zählers von r_{xz} ab, so dass sichtbar wird: **Bei steigenden Werten von r_{xz} , unter Kontrolle weiterer Einflüsse, wird der reine Einfluss von y auf z (p_{zy}) geringer.**

An dieser Stelle bleibt noch zu sagen, dass die eben festgestellten Zusammenhänge sich noch weiter verkomplizieren, wenn das Modell komplexer wird. Deswegen wurde hier ein recht einfaches Modell gewählt, in dem die Beziehungen überschaubar sind, um auf einfachem Wege in das Funktionsprinzip der Pfadanalyse einzuführen.

Nun wurden die Pfadkoeffizienten im obigen Beispiel auf umständlichem Wege ermittelt. Die Berechnung wird im Falle von mehr als drei Variablen und einer steigenden Anzahl an postulierten Beziehungen noch umständlicher, da die Anzahl der Gleichungen, welche denen im Beispiel 2.7 - 2.9 entsprechen, schnell ansteigt.

Es gibt eine Formel, das sog. *Grundtheorem der Pfadanalyse*, welche allgemein gehalten ist und einiges an Rechenarbeit erspart (vgl. Opp 1976: 166 f). Sie lautet:

$$\boxed{r_{ij} = \sum_q p_{iq} r_{qj}} \quad (2.22)$$

Mit dieser Formel lässt sich jede Korrelation zweier beliebiger Variablen x_i und x_j eines Pfadmodells ausdrücken. Der Index q symbolisiert hingegen **jede** Variable, welche auf x_i einen Einfluss ausübt. Es gibt also so viele q -Variablen, wie Pfeile auf x_i zeigen.

Übertragen auf das obige Beispiel lässt sich z.B. r_{zx} wie folgt ausdrücken (wobei es drei q -Variablen gibt – x , y und e_z):

$$r_{zx} = p_{zx} r_{xx} + p_{zy} r_{yx} + p_{ze_z} r_{e_z x} \quad (2.23)$$

Gemäß der Tatsache $r_{xx} = 1$ und der Annahme $r_{e_z x} = 0$ verkürzt sich der rechte Ausdruck der Gleichung 2.23 zu $p_{zx} + p_{zy} r_{yx}$ und entspricht dem obigen Ausdruck in Gleichung 2.17. Analog dazu verfährt man mit den anderen Korrelationen. Danach bleibt es, wieder umzuformen und zu substituieren, um die Pfadkoeffizienten zu bestimmen. Man hat sich aber die vorherigen Rechenschritte ersparen können.

Zum Schluss sei noch erwähnt, dass die Pfadkoeffizienten immer Ausdruck eines sog. **direkten Effekts** sind. Denn auch wenn sie in die Berechnung die Korrelationen benachbarter Variablen einbeziehen, so stehen sie dennoch unmittelbar auf dem kürzesten Wege, der zwei Variablen verbindet, nämlich auf dem direkten Pfad zwischen ihnen.

Es gibt darüber hinaus noch einen sog. **indirekten Effekt**. Das ist ein Effekt, welcher über mehrere dazwischenliegende Variablen verläuft.

Im obigen Beispiel (s. Abb. 2.4) hat die Variable x einen direkten Effekt auf z , da sie durch einen Pfeil unmittelbar verbunden sind. Die Variable x hat aber auch einen indirekten Effekt, welcher über y , also über zwei Pfade verläuft. Dieser indirekte Effekt berechnet sich, indem die Pfadkoeffizienten, welche auf dem „Weg liegen“ multipliziert werden. Da es sich bei diesen Koeffizienten um Werte $< |1|$ handelt, wird bei zunehmender Anzahl an Schritten,

welche zwischen zwei Variablen liegen, das Produkt immer kleiner (da immer mehr Faktoren, welche $< |1|$ sind, miteinander multipliziert werden). Dies ist logisch, denn je „weiter weg“ zwei Variablen voneinander sind, umso weniger können sie indirekt Einfluss aufeinander ausstrahlen. Der Effekt verpufft sozusagen auf dem langen Weg zwischen ihnen.

Es sei dann noch der Begriff des **totalen Effekts** eingeführt. Der totale Effekt erfasst sowohl den direkten als auch den indirekten Effekt. Er lässt sich intuitiv einfach berechnen (E steht für „Effekt“):

$$E_{total} = E_{direkt} + E_{indirekt} \quad (2.24)$$

2.3 Eine einfache Pfadanalyse mit Paneldaten

Diese sehr theoretischen Ausführungen sollen durch ein praxisnahes Beispiel aufgelockert werden. Die Basis stellen manifeste Variablen aus einem Paneldatensatz dar. Der Datensatz ist fiktiv⁸.

Es geht darum, die Hypothese zu testen, wie Lebenszufriedenheit und Gesundheitszustand zusammenhängen. Im ersten Augenblick denkt man intuitiv, dass der Gesundheitszustand eine kausale Wirkung auf die Lebenszufriedenheit haben kann. Aber die umgekehrte Richtung ist ebenfalls denkbar. Schließlich hört man immer wieder von Theorien, in denen behauptet wird, dass eine gute psychische Verfassung (und dazu gehört schließlich eine gewisse Lebenszufriedenheit als Indikator) gesund hält. Die Richtung gilt es anhand von Paneldaten zu testen. Es werden folgende Variablen in die Analyse einbezogen:

- subjektive Lebenszufriedenheit zum Zeitpunkt t : x_1
- subjektive Lebenszufriedenheit zum Zeitpunkt $t + 5$ Jahre: x_2
- subjektiv eingeschätzter Gesundheitszustand zum Zeitpunkt t : y_1
- subjektiv eingeschätzter Gesundheitszustand zum Zeitpunkt $t + 5$ Jahre: y_2
- Residuum der Variablen x_2 : e_{x_2}

⁸s. Appendix 4.2

- Residuum der Variablen y_2 : e_{y_2}

Weiterhin soll angenommen werden, dass die Variablen auf einer 7-stufigen Rang-Skala gemessen worden sind (von 1=“sehr unzufrieden“ bis 7=“sehr zufrieden“) und als quasi-metrische Variablen behandelt werden. Es werden folgende Zusammenhänge postuliert (für ein Beispiel mit gleicher Modellstruktur s. Engel 1994: 25 ff):

- Der Querschnittseffekt $p_{x_1y_1}=r_{x_1y_1}$
- Die sog. „*kreuzverzögerten Effekte*“ $p_{x_2y_1}$ und $p_{y_2x_1}$
- Die sog. „*Stabilitätskoeffizienten*“ $p_{x_2x_1}$ und $p_{y_2y_1}$
- Die Unkorreliertheit der Residuen mit anderen Variablen: $p_{x_2e_{x_2}} = 0$ und $p_{y_2e_{y_2}} = 0$

Eine Visualisierung dieses Modells taucht in einem anderen Zusammenhang in Abb. 2.13 auf.

Im Folgenden wird die Korrelationsmatrix dargestellt. Es sei noch mal darauf hingewiesen, dass der Datensatz vom Autor frei erfunden ist und die Korrelationen sicherlich stark überschätzt sind. Aber gönnen wir uns mal den Luxus, auch mal mit hohen Korrelationen zu rechnen:

Korr.	x_1	y_1	x_2	y_2
x_1		0,74	0,87	0,75
y_1			0,635	0,816
x_2				0,63
y_2				

Tabelle 2.2: Korrelationsmatrix für eine Pfadanalyse

Hier lässt sich die oben eingeführte allgemeine Formel 2.22 anwenden, so dass die Korrelationen der Modell-Variablen wie folgt ausgedrückt werden (wobei Selbstkorrelationen und Korrelationen = 0 bereits rausgerechnet sind):

$$r_{y_1x_1} = p_{y_1x_1} \tag{2.25}$$

$$r_{x_2x_1} = p_{x_2x_1} + p_{x_2y_1}r_{x_1y_1} \quad (2.26)$$

$$r_{x_2y_1} = p_{x_2y_1} + p_{x_2x_1}r_{x_1y_1} \quad (2.27)$$

$$r_{y_2x_1} = p_{y_2x_1} + p_{y_2y_1}r_{x_1y_1} \quad (2.28)$$

$$r_{y_2y_1} = p_{y_2y_1} + p_{y_2x_1}r_{x_1y_1} \quad (2.29)$$

Um die einzelnen Pfadkoeffizienten zu bestimmen, müssen nun die Gleichungen umgeformt und substituiert werden. Es ergibt sich zuerst für die *Stabilitätskoeffizienten*:

$$p_{x_2x_1} = \frac{r_{x_2x_1} - r_{x_2y_1}r_{x_1y_1}}{1 - r_{x_1y_1}^2} \quad (2.30)$$

$$p_{y_2y_1} = \frac{r_{y_2y_1} - r_{y_2x_1}r_{x_1y_1}}{1 - r_{x_1y_1}^2} \quad (2.31)$$

Des Weiteren lassen sich die *kreuzverzögerten Effekte* berechnen:

$$p_{x_2y_1} = \frac{r_{x_2y_1} - r_{x_2x_1}r_{x_1y_1}}{1 - r_{x_1y_1}^2} \quad (2.32)$$

$$p_{y_2x_1} = \frac{r_{y_2x_1} - r_{y_2y_1}r_{x_1y_1}}{1 - r_{x_1y_1}^2} \quad (2.33)$$

Die Ergebnisse für das Beispiel sind in der folgenden Tabelle 2.3 aufgelistet:

Querschnittseffekt	$p_{x_1y_1} = 0,742$
Stabilitätskoeffizienten	$p_{x_2x_1} = 0,89$ $p_{y_2y_1} = 0,586$
Kreuzverzögerte Effekte	$p_{x_2y_1} = -0,03$ $p_{y_2x_1} = 0,31$

Tabelle 2.3: Ergebnisse Pfadanalyse

Ohne sich zu weit aus dem Fenster zu lehnen, möchte der Autor eine kleine

Deutung der Befunde vornehmen:

Zuallererst sieht man an $p_{x_1y_1}$, dass die beiden Variablen x und y im Querschnitt recht stark miteinander zusammenhängen.⁹

Dann ist festzustellen, dass beide Variablen über die verstrichene Zeit relativ stabil in ihren Werten sind, wobei dies in einem stärkeren Maße auf $p_{x_2x_1}$, also auf die Variable „Lebenszufriedenheit“ zutrifft. Während die Werte der Korrelationen $r_{x_2x_1}$ und $r_{y_2y_1}$ recht nah beieinander liegen, macht sich die stärkere „reine“ Stabilität der Variablen x im Vergleich zu y bei den Pfadkoeffizienten deutlicher bemerkbar.

Die Stabilitätskoeffizienten fallen deutlich höher aus als die kreuzverzögerten Effekte. Das liegt in der Natur der Sache, dass gewisse Variablen in einem bestimmten Zeitraum nicht so stark variieren. Dennoch lässt sich ein nach Augenmaß signifikanter Effekt von x_1 nach y_2 ausmachen. Der Wert ist höher als der praktisch nicht vorhandene Effekt $p_{x_2y_1}$. Auch hier ist die Diskrepanz zwischen den beiden Pfadkoeffizienten höher als zwischen den Korrelationskoeffizienten.

Das würde die These stützen, dass die Lebenszufriedenheit durchaus einen Effekt auf den Gesundheitszustand haben kann. Aber auch hier sei nochmals darauf hingewiesen, dass solche Ergebnisse mit Vorsicht zu interpretieren sind. Eine echte Kausalität ist damit noch längst nicht nachgewiesen. Erstens ist das zeitliche Vorgehen der wirkenden Variablen durch ein Paneldesign nicht sicher bestätigt (da keine experimentelle Manipulation der wirkenden Variablen vorliegt), zweitens ist nicht geprüft worden, ob das Modell korrekt spezifiziert ist und somit keine wirkenden Drittvariablen ausgeschlossen worden sind.

Trotz aller Bescheidenheit wären solche Ergebnisse, wenn sie denn einem echten Datensatz zugrunde lägen, für den Statistiker ein kleines Erfolgserlebnis.

Außerdem ließ sich zeigen, dass bereinigte Koeffizienten in der Lage sind, gewisse Relationen zwischen Zusammenhängen von Variablen deutlicher hervorzuheben, als das bei z.B. bivariaten Korrelationskoeffizienten der Fall sein kann.

⁹Das Problem der Multikollinearität soll hier nicht diskutiert werden.

2.4 Das Ein-Indikatoren-Modell als Ansatz zur Unterscheidung zwischen Veränderung und mangelnder Reliabilität

In diesem Abschnitt betrachten wir wieder nur eine inhaltliche Variable – zumindest auf der empirisch-deskriptiven Ebene. Allerdings wird diese Betrachtung um eine sog. *latente Variable* erweitert.

Das Konzept latenter Variablen postuliert, dass sich hinter gemessenen Merkmalen u.U. latente Größen verbergen, welche auf die Messung einen kausalen Einfluss haben. So kann man hinter der Zustimmung eines Befragten zu einem Item wie „Ausländer haben in Deutschland nichts zu suchen“ (empirische Variable) eine starke Ausprägung der latenten Variablen „Ausländerfeindlichkeit“ vermuten.

Stellt man solch eine Verbindung zu latenten Größen her, so werden in diesem Zusammenhang empirische Variablen als *Indikatoren* der latenten Variablen bezeichnet. Die so entstehende Verbindung heißt *Messmodell*.

I.d.R. fungieren ganze Itembatterien als Indikatoren einer einzigen latenten Größe. So wird in einer Befragung, welche u.a. Tendenzen zur Ausländerfeindlichkeit messen soll, normalerweise eine Reihe ähnlicher Items (sog. „multiple Indikatoren“) vorgelegt, und nicht nur, wie im obigen Beispiel, lediglich eins.

Dies hat zum Vorteil, dass einerseits durch Techniken wie die der Itemanalyse unbrauchbare Indikatoren identifiziert und anschließend entfernt werden können und dass andererseits durch mehrere „gute“ Indikatoren, welche als wiederholte Messungen einer latenten Variablen verstanden werden können, die Messung an sich insgesamt weniger fehlerbehaftet wird.

Denn genau das ist ein großes Problem: Jede Messung ist mit einem Fehler behaftet. Dies lässt sich nicht verhindern. Es muss allerdings angestrebt werden, den Fehler möglichst gering zu halten und dafür zu sorgen, dass er unsystematisch ist, er sich also bei wiederholten Messungen soweit wie möglich aufhebt. Das Konzept der latenten Variablen versucht, gerade solche Variablen zu konstruieren, welche messfehlerfrei sind.

Die gerade erwähnten unsystematischen oder anders gesagt: zufälligen Messfehler sind Ausdruck fehlender Präzision bei der Messung. Je präziser ein Messinstrument misst, umso verlässlicher ist es, deswegen bezeichnet man den hier angesprochenen Sachverhalt als die *Frage nach der Reliabilität einer*

Messung / eines Messinstruments.

Wenn nun eine gemessene Variable als einziger Indikator für eine latente Variable betrachtet wird, dann wird hier der einfachste Fall modelliert: Das *Ein-Indikatoren-Modell*. Die latente Variable ist hierbei inhaltlich mit dem Indikator identisch, sie enthält allerdings keine Messfehler. Das ist eine sog. *true score variable*. Geht man von einer Messung im Querschnittsdesign aus, so hat dieses Konzept keine Bedeutung, da sich aus einer einmalig gemessenen Variablen nichts anderes als sie selbst konstruieren lässt. Aussagen über Messfehler sind nicht möglich. Latente Variable und Indikator wären somit redundant.

Dieser Zustand ändert sich, wenn Paneldaten zur Analyse hinzugezogen werden. Denn nun stehen, bei z.B. zwei Zeitpunkten, zwei gemessene Variablen (und zwei latente) der Analyse zur Verfügung. Die zweite Messung kann hierbei als eine Wiederholung der ersten Messung angesehen werden und wiederholte Messungen sind geeignet, um die Reliabilität einer gemessenen Variablen zu schätzen.

Ohne sehr weit abzuschweifen, soll nun kurz auf die Grundlagen der klassischen Testtheorie eingegangen werden, um die Berechnung der Reliabilität plausibel darzustellen.

Jeder gemessene Wert x wird als Summe eines wahren Wertes und eines Messfehlers begriffen:

$$x = \tau + \epsilon \quad (2.34)$$

mit τ =wahrer Wert und ϵ =Messfehler.

Unter der Annahme, dass wahrer Wert und Messfehler in einer Messreihe unkorreliert sind (vgl. Engel 1994: 32), lässt sich die Varianz von x [Var_x] folgendermaßen darstellen:

$$Var_x = Var_\tau + Var_\epsilon \quad (2.35)$$

Je größer der Anteil der Varianz von τ (wahre Varianz) an der Gesamtvarianz ist, umso geringer ist die Varianz von ϵ (Fehlervarianz). Des Weiteren: Je präziser (also: reliabler) eine Messung über mehrere Objekte ist, umso kleiner ist die Fehlervarianz. Somit lässt sich Reliabilität (p) formal definieren

als Anteil der wahren Varianz an der Gesamtvarianz:

$$p_x = \frac{Var_\tau}{Var_x}. \quad (2.36)$$

Das Problem hierbei ist: Die wahren Werte sind meist unbekannt. Ständen sie zur Verfügung, dann wäre das Problem gelöst und man könnte mit ihnen statt mit den gemessenen Werten weiterrechnen. Ebenso ist die wahre Varianz unbekannt.

So bleibt es dem Statistiker, die Reliabilität auf anderem Wege zu schätzen. Zu diesem Zwecke wurden verschiedene Verfahren entwickelt, welche größtenteils darauf basieren, dass mehrere Messungen als *Wiederholungen ein und derselben Messung* zu verstehen sind. Denn wenn ein Messinstrument verlässlich ist, so muss es bei wiederholten Messungen sehr ähnliche Werte liefern.

Ein einfaches und intuitives Verfahren ist die *Test-Retest-Methode*: Ein und dieselbe Messung an derselben Stichprobe wird zu zwei Zeitpunkten durchgeführt. Die Korrelation dieser zwei Messungen gilt als Schätzung für Reliabilität. Eine solche Konzeption ist durch das Paneldesign zu verwirklichen.

Das zentrale Problem dieser Methode, welches oben als Vorteil des Paneldesigns diskutiert wurde, ist die Zeit, welche zwischen zwei Messungen verstreicht. Nur wenn man annimmt, dass sich die wahren Werte einer Messung zwischen zwei Zeitpunkten nicht verändert haben, gilt die Korrelation zwischen den zwei Messwertreihen als unverzerrter Schätzer der Reliabilität.

Aber auch wenn der Statistiker notwendigerweise öfters zur Berechnung gewisser Koeffizienten notwendige Annahmen setzt, so ist die Annahme in diesem Fall sehr fraglich. Es gilt zu fragen:

*Sind Schwankungen zwischen zwei zeitlich versetzten Messungen Ausdruck von **Veränderung** oder von **mangelnder Reliabilität**?*

Diese Unterscheidung wird aufgegriffen, indem im Folgenden als latente Variablen die wahren Werte der gemessenen Variablen x modelliert werden (vgl. Engel 1994: 32 ff). Liegt eine Messung zu zwei Zeitpunkten vor, so lassen sich die Beziehungen im folgenden Pfaddiagramm veranschaulichen¹⁰:

¹⁰Die Bezeichnung der Pfadkoeffizienten in der Graphik weicht etwas ab von dem bisher gewählten Standard, weil dieser Standard mit dem Graphikprogramm nicht zu verwirklichen war; Unterstrich steht für *Index des Pfadkoeffizienten*, τ wurde ausgeschrieben und ϵ mit „error“ bezeichnet.

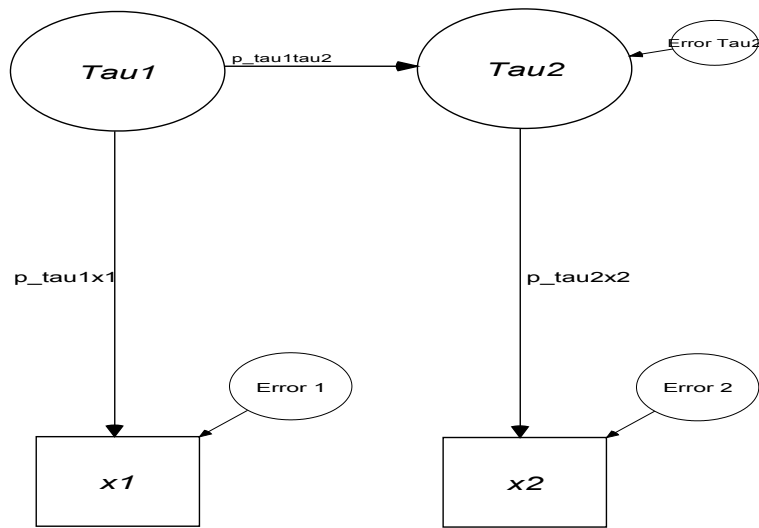


Abbildung 2.5: Pfaddiagramm eines Ein-Indikatoren-Modells

Die Variablen lauten:

x_1 = Die Variable x zum ersten Zeitpunkt

x_2 = Die Variable x zum zweiten Zeitpunkt

τ_1 = Die True-Score-Variable von x zum ersten Zeitpunkt

τ_2 = Die True-Score-Variable von x zum zweiten Zeitpunkt

ϵ_1, ϵ_2 = Die Fehlerterme der Variablen x_1 und x_2

ξ_2 = Die Fehlervariable von $\tau_2 = \sqrt{1 - p_{\tau_2\tau_1}^2}$

Die Fundamental-Gleichung 2.22 lässt sich auch hier anwenden. Es ergeben sich folgende Gleichungen¹¹:

$$r_{x_1x_2} = p_{x_1\tau_1}r_{x_2\tau_1} \quad (2.37)$$

$$r_{\tau_2\tau_1} = p_{\tau_2\tau_1} \quad (2.38)$$

$$r_{x_2\tau_1} = p_{x_2\tau_2}p_{\tau_2\tau_1} \quad (2.39)$$

Durch Umformung und Substitution lässt sich die Test-Retest-Korrelation $r_{x_1x_2}$ nun als Produkt von Pfadkoeffizienten darstellen:

¹¹Die Zerlegung von $r_{x_1\tau_2}$ ist für diese Zwecke unbrauchbar.

$$r_{x_1x_2} = p_{x_1\tau_1}p_{x_2\tau_2}p_{\tau_2\tau_1} \quad (2.40)$$

In diesem Kontext lassen sich $p_{x_1\tau_1}$ und $p_{x_2\tau_2}$ als *Reliabilitätskoeffizienten* und $p_{\tau_2\tau_1}$ als *Stabilitätskoeffizient* begreifen. Setzt man die Annahme, dass die Reliabilität eines Messinstruments über die Zeit stabil bleibt, dann lassen sich $p_{x_1\tau_1}$ und $p_{x_2\tau_2}$ vereinheitlichen zu $p_{x\tau}$.

Gleichung 2.40 lässt sich somit schreiben als:

$$r_{x_1x_2} = p_{x\tau}^2 p_{\tau_2\tau_1} \quad (2.41)$$

Aber auch in diesem Falle hat die Gleichung zu viele, nämlich zwei Unbekannte, so dass eine saubere arithmetische Lösung nicht möglich ist.

Bezieht man allerdings einen dritten Messzeitpunkt ein, dann entstehen neue Möglichkeiten. Unter der Annahme, dass τ_1 keinen direkten Einfluss auf τ_3 besitzt (vgl. Engel 1994: 36), gesellen sich zu der Test-Retest-Korrelation aus Gl. 2.41 zwei weitere in Form von:

$$r_{x_2x_3} = p_{x\tau}^2 p_{\tau_3\tau_2} \quad (2.42)$$

$$r_{x_1x_3} = p_{x\tau}^2 p_{\tau_2\tau_1} p_{\tau_3\tau_2} \quad (2.43)$$

Auch an dieser Stelle muss wieder Umformungsarbeit geleistet werden, die aber schlussendlich in der eindeutigen Bestimmung der Koeffizienten mündet.

$$\begin{array}{|l} p_{x\tau}^2 = \frac{r_{x_1x_2}r_{x_2x_3}}{r_{x_1x_3}} \\ p_{\tau_2\tau_1} = \frac{r_{x_1x_3}}{r_{x_2x_3}} \\ p_{\tau_3\tau_2} = \frac{r_{x_1x_3}}{r_{x_1x_2}} \end{array} \quad (2.44)$$

An dem Reliabilitätskoeffizienten $p_{x\tau}^2$ erkennt man, dass je kleiner die Korrelation von x_1 und x_3 im Verhältnis zu den Korrelationen zwischen benachbarten Zeitpunkten ist, umso größer wird die geschätzte Reliabilität.

Dies ist einleuchtend, denn wenn die Werte des ersten und des zweiten Zeitpunktes schwach miteinander korrelieren, die Werte des dritten Zeitpunktes aber plötzlich wieder viel stärker mit den ersten Werten zusammenhängen, dann liegt dies die Vermutung nahe, dass etwas mit dem Messinstrument nicht stimmt.

Hier ist also stillschweigend die Annahme implementiert, dass sich Veränderung kontinuierlich vollzieht – als Folge der Annahme von Linearität.

Ein Prozess, bei dem sich vom ersten zum zweiten Zeitpunkt vieles wandelt (z.B. Veränderung der Institution Familie), dieser Wandel aber Orientierungslosigkeit nach sich zieht und sich in Folge dessen zum dritten Zeitpunkt eine „Rückentwicklung“ zur Struktur des ersten Zeitpunkts vollzieht (z.B. die „Rückbesinnung auf alte Werte“), würde von diesem Modell nicht identifiziert, sondern als mangelnde Reliabilität gedeutet.

Dieser Sachverhalt wird auch an den Stabilitätskoeffizienten deutlich, da vor allem die Höhe der Korrelation zwischen den entfernten Zeitpunkten t_1 und t_3 (im Zähler stehend) ausschlaggebend für die Beurteilung der Stabilität / Veränderung zwischen benachbarten Zeitpunkten ist.

Auf ein grundlegendes Problem sei an dieser Stelle noch hingewiesen: Man darf nicht vergessen, dass eine solch saubere arithmetische Lösung wie die obige nur unter den getroffenen Annahmen möglich ist. Würde man nicht unterstellen, dass

- die Residuen unkorreliert sind (gerade bei Paneldaten ist dies fragwürdig),
- die Reliabilität über die Messzeitpunkte konstant ist
- und τ_1 nicht direkt auf τ_3 wirkt,

dann müsste man diese Sachverhalte ebenfalls modellieren, so dass sich mehr Unbekannte als Gleichungen ergäben und in Folge eine rein rechnerische Lösung nicht möglich wäre.

Nichtsdestotrotz lässt sich hier eine enorme Bereicherung festhalten, die durch Paneldaten ermöglicht wird:

Ab drei Messzeitpunkten ist es möglich, mit Paneldaten rechnerisch zwischen wahrer Veränderung und mangelnder Reliabilität zu unterscheiden!

Es gibt weitere Ansätze, im Ein-Indikatoren-Modell die Reliabilität von der Stabilität zu trennen (z.B. das Zurückgreifen auf Kovarianzen anstatt auf Korrelationen), außerdem lässt sich das Modell zu einem multiplen Indikatorenmodell erweitern. Für Interessierte sei hier auf die Ausführungen von Engel/Reinecke (1994: 38 ff) verwiesen.

2.5 Regressionsmodelle für Paneldaten

Dieses Kapitel stellt Regressionsmodelle vor, welche speziell auf Paneldaten zugeschnitten sind.

Gemeint sind hier vor allem lineare Modelle für manifeste Variablen. Im Gegensatz zum vorhergehenden Kapitel wird die Ebene latenter Variablen also wieder verlassen. Stattdessen liegt der Fokus darauf, sich mit der Einbeziehung von sich *inhaltlich* unterscheidenden Variablen zu befassen. Dies war im Ein-Indikatoren-Modell nicht der Fall. Es ging letztendlich um *eine* Variable x , gemessen zu mehreren Zeitpunkten und erweitert um ihre eigene True-Score-Variable.

Nachdem also in diesem Kapitel auf „manifeste Ebene“ mehrere in Beziehung zueinander stehende Variablen betrachtet werden¹², kann im nächsten Kapitel dazu übergegangen werden, diese Betrachtungsweise um latente Größen zu erweitern. Tabelle 2.4 gibt hierzu eine Übersicht:

	Nur manifeste Variablen	Manifeste+Latente Variablen
Eine inhaltliche Variable	Kap. 2.1 (elementare Veränderungen)	Kap. 2.4 (Ein-Indikatoren-Modelle)
Mehrere inhaltliche Variablen	Kap. 2.3/2.5 (Pfadmodelle/Regressionsmodelle)	Kap. 2.6 (Strukturgleichungsmodelle)

Tabelle 2.4: Übersicht über die Eigenschaften vorgestellter Modelle

¹²dies geschah zwar bereits in dem Beispiel zur Pfadanalyse, allerdings wurde das Verfahren konventionell angewendet, ohne speziell auf das Paneldesign hin modifiziert zu werden.

Wenn auch dieses Kapitel speziell Regressionsmodellen gewidmet ist, so schärfen die hier vorgestellten Kriterien, nach denen Modelle für Paneldaten unterschieden werden, den Blick für die sich eröffnenden Möglichkeiten, aber auch statistisch-mathematischen Fallen, die *generell* mit Paneldaten einhergehen.

Zunächst einmal muss, wie auch schon an früheren Stellen des Skripts deutlich wurde, eine einheitliche und präzise Darstellung der Individual- und der Zeitebene gesichert sein. Nur auf diesem Wege können die unterschiedlichen Regressionsmodelle formal korrekt dargestellt werden. Dies geschieht in der Regel mithilfe einer Index-Notation, wie sie auch schon aus der Querschnittsregression bekannt ist. Hier wird allerdings neben der Unterscheidung nach Individuen auch ein Index für die Unterscheidung nach Zeitpunkten eingeführt.

Die Ergänzung einer Variablen x um den Personenindex „ i “ (mit $i = 1, 2, \dots, N$) und den Wellenindex „ t “ (mit $t = 1, 2, \dots, T$), zu x_{it} drückt also aus, dass x Werte verschiedener Personen zu verschiedenen Zeiten annehmen kann. $x_{2,4}$ wäre folglich der x -Wert der 2. Person zum 4. Zeitpunkt einer Untersuchung.

Nun lassen sich neben den Variablen auch die *Koeffizienten* eines Regressionsmodells mit diesen Indizes versehen. Somit wäre im Falle der einfachen Regression¹³ theoretisch das folgende „Maximal-Modell“ denkbar:

$$y_{it} = a_{it} + b_{it}x_{it} + \epsilon_{it} \quad (2.45)$$

mit

y_{it} = Wert der abhängigen Variablen für die i -te Person zum Zeitpunkt t

a_{it} = Regressionskonstante für die i -te Person zum Zeitpunkt t

b_{it} = Regressionskoeffizient für die i -te Person zum Zeitpunkt t

x_{it} = Wert der unabhängigen Variablen für die i -te Person zum Zeitpunkt t

ϵ_{it} = Residuum für die i -te Person zum Zeitpunkt t

Der Rückgriff auf die Indizes verkürzt die Darstellung, da die Modellgleichung 2.45 im Grunde für jedes i und t als separate Gleichung ausgeschrieben

¹³ „einfach“ bezieht sich hier auf den Sachverhalt, dass das Modell inhaltlich gesehen eine einzige unabhängige Variable berücksichtigt – dies lässt sich jedoch bedenkenlos auf eine multiple Regressionskonstruktion übertragen; dies gilt auch für weitere Formulierungen von Regressionsgleichungen in diesem Abschnitt.

werden könnte:

$$\begin{aligned}
 y_{1,1} &= a_{1,1} + b_{1,1}x_{1,1} + \epsilon_{1,1} \\
 y_{2,1} &= a_{2,1} + b_{2,1}x_{2,1} + \epsilon_{2,1} \\
 &\dots \\
 y_{1,2} &= a_{1,2} + b_{1,2}x_{1,2} + \epsilon_{1,2} \\
 y_{2,2} &= a_{2,2} + b_{2,2}x_{2,2} + \epsilon_{2,2} \\
 &\dots \\
 y_{N,T} &= a_{N,T} + b_{N,T}x_{N,T} + \epsilon_{N,T}
 \end{aligned} \tag{2.46}$$

Eine solche Modellformulierung ist jedoch weder inhaltlich sinnvoll, noch sind die Koeffizienten mathematisch bestimmbar. Denn es müsste pro Person und Zeitpunkt jeweils ein Regressionskoeffizient und eine Regressionskonstante geschätzt werden. In der Regel dienen Regressionsmodelle aber dazu, Zusammenhangsstrukturen auf Aggregatebene zu untersuchen. „Individuelle“ Koeffizienten, die auch noch von Welle zu Welle variieren, würden diese Funktion nicht erfüllen. So ist es selten für sozialwissenschaftliche Hypothesenprüfungen von Interesse, z.B. den Regressionskoeffizienten von „Herrn Müller“ und seinen persönlichen y-Achsenabschnitt aus dem Jahr 2004 zu kennen.

Mathematisch gesehen lassen sich die Koeffizienten darüber hinaus nicht schätzen, da die Anzahl der Freiheitsgrade negativ ist. Denn es stehen viel zu wenige Informationen zur Schätzung der großen Anzahl von Koeffizienten zur Verfügung.

Doch auch wenn das „Maximal-Modell“ lediglich ein Gedankenexperiment darstellt, so sind doch einige Modelle **mit variablen Regressionsparametern** konstruierbar, die inhaltlich plausibel und mathematisch berechenbar sind.

Entscheidend ist hierbei, ob sich die Zusammenhangsstruktur auf einer der beiden Ebenen (Zeiten- oder Personenebene) von der Zusammenhangsstruktur auf der indifferenten Gesamtebene (= Zeiten- + Personenebene) signifikant unterscheidet.

Zur Veranschaulichung dieses Gedankens sei ein Beispiel aufgeführt: Mal angenommen, es liegen Paneldaten zur „ x = Dauer der Betriebszugehörigkeit (unabhängige Variable)“ und „ y = Produktivität (abhängige Variable)“ von Mitarbeitern einer bestimmten Firma vor. Es wird ein einfaches Regressionsmodell $y_{it} = a + bx_{it} + \epsilon_{it}$ formuliert, welches die Unterscheidung zwischen der Zeiten- und der Personenebene ignoriert. Diese einfache Regressionsgleichung wird im folgenden als „globales Modell“ bezeichnet. Nun sei angenommen, dass die Begutachtung der Koeffizientenschätzung zum Schluss führt, dass kein signifikanter Einfluss von x auf y existiert.

Würden allerdings die Koeffizienten a und b nicht mehr als konstant für einzelne Individuen betrachtet, so müsste das globale Modell reformuliert werden müssen: $y_{it} = a_i + b_i x_{it} + \epsilon_{it}$ (im folgenden: „reformuliertes Modell“). Nun könnte sich zeigen, dass die individuellen b_i -Koeffizienten größtenteils signifikant und positiv sind. Dies würde bedeuten, dass für die meisten Mitarbeiter in der Längsschnittbetrachtung gilt: Je länger ein Mitarbeiter i in der Firma tätig ist, umso produktiver ist derselbe Mitarbeiter i . Dass dieser Zusammenhang zwischen x und y erst im reformulierten Modell zum Vorschein kommt, liegt daran, dass er im globalen Modell durch inter-individuelle Unterschiede verwischt wurde. Denn in dem globalen Modell wurden sowohl Unterschiede zwischen Individuen als auch Unterschiede innerhalb von Individuen im Zeitverlauf undifferenziert berücksichtigt. Besteht zwar im letzteren Falle ein Zusammenhang zwischen x und y , im ersteren Falle aber nicht, so ist es möglich, dass das globale Ergebnis ebenfalls keinen Zusammenhang attestiert (vor allem dann, wenn die Anzahl N der Individuen der Anzahl T der Zeitpunkte deutlich überlegen ist).

Somit ist in diesem Beispiel zu vermuten, dass der auf globaler Ebene nicht anerkannte Einfluss von x auf y auf einen fehlenden Zusammenhang im Vergleich *zwischen* den Personen zurückgeht. Dies könnte bspw. damit erklärt werden, dass in der Firma Arbeitsverhältnisse nicht konsequent in Abhängigkeit der Produktivität aufrechterhalten bzw. beendet werden. Viell. herrscht generell eine hohe Fluktuation in dem Betrieb. Oder es nutzen einige produktive Mitarbeiter ihre Produktivität, um weiter aufzusteigen, während andere produktive Mitarbeiter an einem Aufstieg kein Interesse haben und auf derselben Position verbleiben.

All diese Erklärungsskizzen implizieren *inter*-individuelle Unterschiede mit einer *fehlenden* Systematik zwischen x und y . Diese sind aber derart „mächtig“, dass das *bestehende* Zusammenhangsmuster auf *intra*-individueller Ebene

ne im globalen Modell nicht zum Vorschein kommt. Das intra-individuelle Zusammenhangsmuster ist allerdings in unserem Beispiel über das reformulierte Modell nachgewiesen. Es besteht also ein, wenn auch spezifischer Zusammenhang zwischen x und y . Durchaus lässt sich schlussfolgern, dass mit zunehmender Betriebszugehörigkeitsdauer die Produktivität steigt – nur eben immer im Verhältnis zum Produktivitätsniveau **desselben Individuums** zu einem **früheren Zeitpunkt**. Da das globale Modell dies nicht erkennt, muss es als fehlspezifiziert deklariert werden.

Zur Klärung, ob eine Differenzierung der Regressionsparameter nach i und t sinnvoll ist, sollten in erster Linie theoretische Annahmen herangezogen werden. Es ist aber auch eine rein statistische Überprüfung mithilfe von „F-Tests“ möglich (vgl. Hsiao 2005: 14ff. – wird hier nicht weiter behandelt).

Die möglichen Differenzierungen lässt sich formal veranschaulichen, indem die Maximal-Gleichung 2.45 auf unterschiedlichen Kombinationswegen restringiert wird. Allerdings sind nicht alle Kombinationsmöglichkeiten inhaltlich plausibel. An dieser Stelle seien daher nur die gängigsten Modellformulierungen genannt:

1) Es wird angenommen, dass die Regressionsparameter über den Individuen aber nicht über die Zeit variieren können. Gleichung 2.45 vereinfacht sich zu:

$$y_{it} = a_i + b_i x_{it} + \epsilon_{it} \quad (2.47)$$

2) Ausgehend von Gleichung 2.47 könnte weiter differenziert werden, ob nur die Regressionskonstante oder nur der Regressionsparameter über i variieren kann:

$$y_{it} = a_i + b x_{it} + \epsilon_{it} \quad (2.48)$$

$$y_{it} = a + b_i x_{it} + \epsilon_{it} \quad (2.49)$$

Es sei angemerkt, dass es gängiger ist, 2.48 anzunehmen. Die Begründung hierfür wird in Kap. 2.5.1 geliefert.

3) Es kann letztendlich angenommen werden, dass beide Parameter für alle i und t konstant sind. Dies würde zu dem sog. „pooled model“, also einem

einfachen Regressionsmodell führen, in dem die Panelstruktur ignoriert wird (wie im obigen Beispiel im Falle des „globalen Modells“):

$$y_{it} = a + bx_{it} + \epsilon_{it} \quad (2.50)$$

Die Koeffizienten des in 2.50 formulierten Modells lassen sich mit der „Methode der kleinsten Quadrate“ (KQ) auf gewöhnlichem Wege schätzen. Die Parameter der Modelle 2.47 und 2.48 bedürfen hingegen eines verfeinerten Schätzansatzes. In diesem Kontext wird der variierenden Regressionskonstante a_i ein zentraler Stellenwert beigemessen. Die Unterscheidung, ob a_i als ein zu schätzender Parameter oder als eine Zufallsvariable aufgefasst wird, führt zu unterschiedlichen Vorgehensweisen bei der Schätzung und gipfelt in unterschiedlichen Interpretationen der Parameter. Eng damit verbunden sind zwei Sachverhalte, und zwar 1.) die Frage, welche Quelle der Varianz (Variation der Werte über Individuen vs. über Zeitpunkte) der im Regressionsmodell involvierten Variablen berücksichtigt wird und 2.) die Frage, ob die Annahme fehlender Autokorrelation der Residuen weiterhin plausibel ist, wenn Panel-daten vorliegen.

Die nach diesen Gesichtspunkten zu unterscheidenden sog. „fixed-“ und „random-effects“-Modelle werden im Kapitel 2.5.1 vorgestellt.

Die Frage, ob Regressionsparameter über i und t variieren können, lässt sich ebenso für die *Variablen* eines Regressionsmodells stellen. So gibt es Merkmale, deren Ausprägungen über die Zeit relativ konstant bleiben (z.B. das Geschlecht) und welche, die sich relativ häufig ändern können (z.B. die Einschätzung über das aktuelle politische Geschehen). Auch kann es Merkmale geben, die für Individuen relativ konstant sind, welche aber über die Zeit variieren (z.B. die Inflationsrate eines Landes). I.d.R. ist die Deklaration einer Variablen „ x “ als x , x_i , x_t oder x_{it} im Vergleich zur obigen „Deklarationspflicht“ bei *Regressionsparametern* nicht notwendig. Denn die Werte einer Variablen liegen nun mal vor und ändern sich nicht, je nachdem, ob sie mit einem i - bzw. t -Index versehen werden. Parameter müssen hingegen geschätzt werden. In diesem Falle macht es einen substantiellen Unterschied, ob ein Regressionsparameter über i bzw. t gleichgesetzt wird. Somit können Variablen eines Regressionsmodells bedenkenlos in der „Maximalversion“, also als x_{it} bzw. y_{it} notiert werden.

Der Sachverhalt ändert sich, wenn Annahmen über im Regressionsmodell

nicht-berücksichtigte unabhängige Variablen getroffen werden. In diesem Falle ist eine Unterscheidung sinnvoll, da sich mithilfe von sog. Differenzgleichungen die Einflüsse einer bestimmten Klasse nicht berücksichtigter unabhängiger Variablen eliminieren lassen – nämlich derer, die über die Zeit konstant sind, für die also gilt: $x_{it} = x_i$.

Diese Art von Elimination wird in Kap. 2.5.2 diskutiert. Es sei nur kurz angemerkt, dass in den oben erwähnten „fixed-“ und „random-effects“-Modellen der Beseitigung der Effekte nicht-berücksichtigter Variablen ebenfalls ein zentraler Stellenwert zukommt. Zum Schluss des Kapitels 2.5.2 werden die Vor- und Nachteile dieser Modelle denen des Differenzenmodells gegenübergestellt.

Abschließend werden noch Modelle mit endogener Dynamik angesprochen. Diese Modelle berücksichtigen den Einfluss, den die abhängige Variable zeitversetzt auf sich selbst ausübt. Somit fungiert die inhaltlich gleiche Variable als unabhängige Variable, gemessen vor der Messung der eigentlichen abhängigen Variablen – z.B. in der Form: $y_{it} = a + bx_{it} + cy_{i,t-1} + \epsilon_{it}$, wobei c den zur zeitverzögerten Variablen $y_{i,t-1}$ zugehörigen Regressionskoeffizienten darstellt.

Die mit einer solchen Konstruktion verbundenen Besonderheiten sind Bestandteil des Kapitels 2.5.3.

Insgesamt wird deutlich, dass lineare Regressionsmodelle für Paneldaten nach verschiedenen Kriterien kategorisiert werden können¹⁴ – ein Auszug:

- ***Modelle mit unterschiedlichen Annahmen hinsichtlich der Variation von Regressionsparametern über i und t*** (s. Kap. 2.5.1), mit den Unterthemen:
 - Zerlegung der Varianz der abhängigen Variablen in eine objekt- und eine zeitbezogene Komponente
 - Annahme korrelierender Residuen
- ***Differenzenmodelle, in denen zwischen verschiedenen Typen unabhängiger Variablen unterschieden wird*** (vor allem im Kontext nicht-berücksichtigter Variablen – s. Kap. 2.5.2)

¹⁴hier werden nur Kriterien aufgeführt, welche erst durch die Möglichkeiten, die Paneldaten eröffnen, bestehen. Natürlich lassen sich Regressionsmodelle für Paneldaten ferner auch nach gängigen Unterscheidungskriterien differenzieren, wie z.B. nach Skalenniveau der involvierten Variablen.

- **Modelle mit endogener Dynamik** (also die Frage, ob unter den unabhängigen Variablen eines Modells die abhängige Variable eines früheren Zeitpunkts auftaucht – s. Kap. 2.5.3)

Es sei noch gesagt, dass sich diese Kriterien bei der Formulierung eines Regressionsmodells auch kombiniert berücksichtigen lassen. So stellt sich für Differenzenmodelle und Modelle mit endogener Dynamik ebenfalls die Frage, welche Annahmen hinsichtlich der Variation von Regressionsparametern über i und t getroffen werden.

2.5.1 Modelle mit variablen Regressionskonstanten: Fixed- und Random-Effects-Modelle

In der Einleitung dieses Kapitels 2.5 wurden einige gedankliche Ansätze zu Regressionsmodellen mit Paneldaten angeschnitten:

1. die Frage, inwieweit Regressionsparameter über i und t variieren können,
2. die Frage, inwieweit es notwendig ist, die Varianz einer (abhängigen) Variablen in eine Personen- und eine Zeitkomponente zu zerlegen, und
3. die Frage, ob die Annahme fehlender Autokorrelation der Residuen weiterhin plausibel ist, wenn Paneldaten vorliegen.

Diese Fragestellungen hängen stark miteinander zusammen. Deren gemeinsame analytische Umsetzung führt zur Modellierung der im späteren Verlauf vorgestellten Fixed- und Random-Effects-Modelle. Zunächst werden allerdings diese Gedankengänge vertieft, um den Zusammenhang zwischen ihnen zu verdeutlichen. Dabei wird bewusst auf eine starke mathematische Formalisierung verzichtet (was aber in der Spezifikation der Modelle nachgeholt wird). Denn ehe die konkrete mathematische Umsetzung der Modelle besprochen wird, sollen die sich dahinter verbergenden grundlegenden Ideen nachvollzogen sein. Allerdings müssen an manchen Stellen einige regressions- und varianzanalytischen Grundlagen und Begrifflichkeiten als bekannt vorausgesetzt werden. Im Abschnitt 4.7 des Anhangs werden einige Anregungen und Literaturhinweise angeboten, falls diese Grundlagen nicht vorhanden sind.

2.5.1.1 Die Dekomposition der Varianz einer Variablen bei vorliegenden Paneldaten

Stehen Daten im Panelformat zur Verfügung, so ergeben sich drei Dimensionen,

- Die Dimension der Variablen (definiert über Inhalte),
- Die Dimension der Objekte und
- Die Dimension der Zeitpunkte

Optisch lässt sich dieser Sachverhalt vorstellen, indem z.B. die übliche 2-dimensionale Excel-Tabelle oder SPSS-Datenansicht um eine in den Raum hineinragende Dimension erweitert wird (also die Form eines Quaders annimmt).

Wird *nur eine Variable* betrachtet, so ergeben die Daten wiederum eine 2-dimensionale Tabelle, welche durch Objekte und Zeitpunkte aufgespannt wird. Diese könnte beispielsweise für 4 Zeitpunkte und 7 Objekte so aussehen (O=Objekte, Z=Zeitpunkte):

O↓ Z →	t ₁	t ₂	t ₃	t ₄
1	1,3	1,4	1,4	1,2
2	3	3,1	3,3	3,4
3	8,8	8,4	8,5	8,6
4	5,5	9	3	2,4
5	5,4	8,9	2,8	2,1
6	5,7	9,1	2,6	2,1
7	3	9,9	8	5
8	8	4	2,6	11
9	6,5	2,1	0	8,6

Tabelle 2.5: Werte einer Variablen von Objekten zu verschiedenen Zeitpunkten

Wird beim Vorliegen von Querschnittsdaten eine univariate Verteilung analysiert, so lässt sich nur eine Quelle der Variation feststellen: Werte einer Variablen variieren mit den Beobachtungen, sprich mit den Objekten.

Diese Variations- oder Streuungsquelle erweitert sich bei Paneldaten durch eine zweite Dimension. Nun kann differenziert werden zwischen:

- *Between Variation* – Variation zwischen den Objekten
- *Within Variation* – Variation zwischen den Zeitpunkten

Betrachtet man in der Tabelle 2.5 die Daten für die Fälle 1 bis 3 (zeilenweise), so kann man eine Dominanz der *between variation* gegenüber der *within variation* feststellen. Die Werte der Objekte bleiben über die Zeitpunkte relativ konstant, während sich die Werte zwischen den Objekten relativ stark unterscheiden. Dies könnte z.B. eine Gruppe unterschiedlicher Individuen sein, welche bzgl. einer Einstellung jeweils eine eher gefestigte Meinung haben.

Die Fälle 4 bis 6 weisen genau das Gegenteil auf: Die Werte der Objekte variieren von Zeitpunkt zu Zeitpunkt sehr stark. Zwischen den Personen aber sind sie relativ ähnlich. *Within variation* dominiert hier. Dies könnte eine Gruppe von Personen sein, welche sich z.B. bzgl. eines Verhaltensmusters relativ ähnlich sind. Da sich aber zwischen den Messungen starker Wandel vollzieht, äußert sich dieses Verhalten von Zeitpunkt zu Zeitpunkt unterschiedlich (z.B. vor und nach einem Krieg).

Schließlich weisen die Fälle 7 bis 9 beide Arten der Streuung im ähnlichen Ausmaß auf.

Nun stellen sich die Fragen, wie diese Varianz-Dekomposition mit der Idee einer variablen Regressionskonstanten zusammenhängt und was in diesem Kontext mit fixen bzw. zufälligen Regressionskonstanten gemeint ist. Um diese Fragen zu beantworten, muss der Umweg über die Idee der Varianz- bzw. Kovarianzanalyse gegangen werden.

2.5.1.2 (Ko-)Varianzanalytische Überlegungen auf dem Weg zu variablen Regressionskonstanten

Die Idee, in einer Regressionsgleichung eine variable Regressionskonstante einzubauen, ist nicht speziell im Kontext der Panelanalyse entstanden. Vielmehr ist diese Möglichkeit *immer* gegeben, wenn sich die Fälle eines Datensatzes in überschneidungsfreie Gruppen aufteilen lassen. Dann lassen sich nämlich die Werte einer beliebigen Variablen x in der in Tab. 2.5 vorgeführten zweidimensionalen Form aufführen. In der formalen Darstellung der Werte von x ist die Aufspaltung in zwei Dimensionen durch die Differenzierung der Laufindizes umsetzbar. Der erste Index gibt die Gruppe an, derer ein

Objekt zugehört, der zweite Index erlaubt die Identifizierung des Objektes *innerhalb* der Gruppe. Ein Beispiel für diese Formalisierung könnte die Ergänzung einer Variablen x um die Indizes j mit $j = 1, 2, \dots, n_k$ und k mit $k = 1, 2, \dots, K$ zu x_{jk} sein. Hierbei bezeichnet k den Laufindex der jeweiligen Gruppe mit $K = \text{Anzahl der Gruppen}$. j steht für den Laufindex der Objekte innerhalb einer Gruppe k , n_k kennzeichnet entsprechend die Anzahl der zur Gruppe k gehörenden Objekte. Wird bspw. als eine Variable x die Körpergröße von Basketballspielern 18 verschiedener Vereine deklariert und hat der fünfte Spieler ($j = 5$) der vierten betrachteten Basketballmannschaft ($k = 4$) eine Körpergröße von 212 cm inne, dann ließe sich dieser Wert formal darstellen als: $x_{5,4} = 212$.¹⁵

Nun stellt die in Paneldaten mögliche Unterscheidung zwischen einer Person und einem Zeitpunkt nichts anderes als einen Spezialfall der Einteilung von Objekten in Gruppen dar: Die „Gruppen“ sind die einzelnen Personen ($k = i$)¹⁶ und die „Mitglieder einer Gruppe k “ sind die Werte einer Person k im Zeitverlauf ($j = t$). Auch die umgekehrte Gruppenzuweisung mit $k = t$ und $j = i$ ist möglich. Es wird in den folgenden Ausführungen geklärt, warum es sinnvoll ist, eine Person als eine „Gruppe“ aufzufassen.

Zurück zu der allgemeinen Idee der Gruppierung von Objekten: Kausalanalytisch gesehen ist eine solche Gruppierung dann sinnvoll, wenn vermutet wird, dass die Gruppenzugehörigkeit einen Einfluss auf eine Variable y hat. Genau dies entspricht der Grundidee einer (im einfachsten Falle einfaktoriellen univariaten) Varianzanalyse.¹⁷ Über die Zerlegung der Varianz von y in eine Innergruppen- und eine Zwischengruppen-Komponente wird geprüft, inwieweit die Gruppierung einen Einfluss auf die gesamte Streuung der Variablen y hat. Dieser Einfluss ist umso stärker, je homogener die Gruppenmitglieder einer Gruppe (geringe Innergruppenvarianz) sind und je mehr sich die Mitglieder verschiedener Gruppen im Vergleich unterscheiden (hohe Zwischengruppenvarianz).

Ein Beispiel für einen *mittleren* Einfluss der Gruppierung zwischen bspw. der allgemeinen Lebenszufriedenheit y und dem Familienstand (mit insgesamt K Ausprägungen bzw. Gruppen $k = 1, 2, \dots, K$) könnte auf die folgen-

¹⁵Die Setzung des Kommas zwischen den beiden Indexwerten beugt lediglich der potentiellen Gefahr vor, die beiden Werte fälschlich als einen Zahlenwert „54“ zu lesen.

¹⁶entsprechend der bisherigen Notationen: einerseits s. Beginn von 2.5 (Differenzierung der zwei Panelebenen), und andererseits s. vorheriges Beispiel zur allgemeinen Notation im Falle gruppierter Objekte.

¹⁷An dieser Stelle kann keine erschöpfende Abhandlung über die Varianzanalyse erfolgen; es sei auf Backhaus (2006) und auf das uni-interne Skript zu multivariaten Analyseverfahren (Stein, Pavetic, Noack) verwiesen.

de fiktive Konstellation zurückgehen: Unter den ledigen Personen ($k = 1$) weisen die meisten einen ähnlich niedrigen Zufriedenheitswert auf, sind also eher unzufrieden. Dies äußert sich in einer geringen Innergruppen-Varianz der Gruppe $k = 1$ bei einem verhältnismäßig niedrigen Gruppenmittelwert \bar{y}_1 ¹⁸. Hingegen sind unter den verheirateten Personen ($k = 2$) die meisten auf einem ähnlich hohen Niveau zufrieden¹⁹ – somit auch hier eine geringe Innergruppen-Varianz bei einem allerdings verhältnismäßig hohen Gruppenmittelwert \bar{y}_2 . Nur die Gruppe der Geschiedenen ($k = 3$) folgt nicht dem gruppenhomogenen Trend, was sich statistisch an einer größeren Innergruppen-Varianz²⁰ erkennen lässt – bei einem im Gruppenvergleich „moderaten“ Gruppenmittelwert \bar{y}_3 nahe dem Gesamtmittelwert \bar{y} ($\bar{y}_1 < \bar{y}_3 \approx \bar{y} < \bar{y}_2$).

Die relative Heterogenität der Zufriedenheit in der Geschiedenen-Gruppe ist dafür verantwortlich, dass der Zusammenhang zwischen Familienstand und Lebenszufriedenheit nur moderat ausfällt. Im Extremfall wäre dieser Zusammenhang dann perfekt, wenn innerhalb jeder Familienstand-Gruppe absolute Homogenität vorherrschen würde, also die Zufriedenheit jedes Gruppenmitglieds dem Gruppenmittelwert entspräche ($y_{1k} = y_{2k} = \dots = y_{n_k k} = \bar{y}_k$), und sich die Gruppenmittelwerte aber voneinander unterscheiden würden, im Extremfall: $\bar{y}_1 \neq \bar{y}_2 \neq \dots \neq \bar{y}_K$.²¹ In diesem Falle wäre die vollständige Varianz von y auf die Gruppenunterschiede zurückzuführen.

Übertragen auf Paneldaten würde dieser perfekte Zusammenhang bedeuten, dass die Werte einer abhängigen Variablen y im Falle $k = i$ für jedes Individuum über die Zeit konstant bleiben, sich aber zwischen den Individuen unterscheiden. Bevor diese Feststellung weiter vertieft wird, soll erst kurz auf eine Erweiterung der Varianzanalyse eingegangen werden:

Die unabhängige Variable, welche die Gruppierung definiert, kann ein beliebiges Skalenniveau annehmen. Am sinnvollsten ist die Anwendung auf nominalskalierte Variablen (wie eben Familienstand) und ordinale Variablen, welche nicht als quasi-metrisch behandelt werden können. Für (quasi-)metrische Variablen eignet sich hingegen die Regressionsanalyse eher, da in

¹⁸ Allgemein ist das arithmetische Mittel einer Gruppe \bar{y}_k definiert als $\frac{1}{n_k} \sum_{j=1}^{n_k} y_{jk}$. Da hier $k = 1$, wird der Mittelwert bezeichnet als \bar{y}_1 . Der allgemeine Mittelwert von y lautet entsprechend: $\frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} y_{jk}$ ($n = n_1 + n_2 + \dots + n_K =$ Anzahl aller Objekte)

¹⁹ Wahrscheinlich ist dieses Beispiel sehr unrealistisch.

²⁰ die vereinfachend darauf zurückgeführt werden könnte, dass einige Personen eine Scheidung als Befreiungsakt und andere als Verlust empfinden.

²¹ Der Extremfall ist unter Vorliegen der ersten Bedingung nicht zwingend für einen perfekten Zusammenhang notwendig. Es sollten aber zumindest zwischen *einigen* Gruppen Mittelwertsunterschiede bestehen, sonst würden *alle* Werte von y demselben Wert entsprechen.

der Schätzung von Regressionskoeffizienten der volle mathematische Gehalt dieser Variablen ausgenutzt wird. So werden z.B. bei der metrischen Variablen „ $x = \text{Körpergröße in cm}$ “ zwei Ausprägungen $x = 157$ und $x = 163$ aus varianzanalytischer Sicht lediglich als zwei verschiedene Gruppen betrachtet. Es wird aber nicht, wie in der Regressionsanalyse, die Information der mathematischen Wertigkeit dieser Ausprägungen und des mathematischen Abstandes zwischen 157 und 163 verarbeitet. Ferner führt bei metrischen Variablen mit vielen Ausprägungen eine zu feine Gruppierung dazu, dass die meisten Gruppen nur sehr schwach besetzt sind. Die aus einer Stichprobe errechnete Innergruppen-Varianz für eine solchen schwach besetzte Gruppe kann dann keine zuverlässige Schätzung der entsprechenden Subpopulations-Varianz darstellen.

Aber im Prinzip folgen sowohl die Regressions- als auch die univariate Varianzanalyse demselben Grundgedanken: der Aufklärung der Varianz einer abhängigen Variablen. Im Endeffekt unterscheiden sie sich, vereinfachend formuliert, lediglich durch skalenniveau-abhängige Zulassungsbeschränkungen in Hinblick auf unabhängige Variablen. Es liegt daher nahe, beide Ansätze zu einem allgemeinen linearen Modell zu verknüpfen. Ausgehend von der linearen Regression würde dies über die Aufnahme nominalskalierter Variablen in Form von Dummy-Variablen (pro Gruppe ein Dummy)²² funktionieren. Wird wiederum die Varianzanalyse als Ausgangspunkt gewählt, so lassen sich metrische unabhängige Variablen in Form von sog. Kovariaten berücksichtigen – man spricht dann von der *Kovarianzanalyse*. Beide Ansätze führen zu identischen Ergebnissen. Wird z.B. nur eine Gruppierungsvariable und eine abhängige Variable betrachtet, dann beziehen sich die über die OLS-Schätzung berechneten Regressionskoeffizienten der Dummy-Variablen auf die korrespondierenden Gruppenmittelwerte der abhängigen Variablen²³. Anhand dieser Verbindung wird deutlich, dass im Endeffekt ein und dieselbe Sachlage aus zwei verwandten mathematischen Perspektiven untersucht wird.²⁴

Angelehnt an die Terminologie der Varianzanalyse lässt sich im einfachen

²²Technisches Detail: Um die OLS-Schätzung mathematisch zu ermöglichen, muss ein Dummy in der Formulierung der Regressionsgleichung weggelassen werden; die entsprechende Gruppe wird dann durch die Regressionskonstante repräsentiert.

²³Auch diese Aussage muss technisch präzisiert werden. Denn der Gruppenmittelwert der in der Gleichung „weggelassenen Gruppe“ stellt die Referenz dar und bestimmt den Wert der Regressionskonstanten. Die anderen Regressionskoeffizienten entsprechen den *Abweichungen* des Gruppenmittelwertes von dieser Referenz.

²⁴Diese Verknüpfung wird bspw. in STATA mit der Unteroption „regress“ innerhalb des Befehls „anova“ für eine Varianzanalyse unterstrichen; diese Unteroption erzeugt *innerhalb einer Varianzanalyse* einen Output mit Regressionskoeffizienten der Dummy-Variablen zu der in der Varianzanalyse spezifizierten Gruppierungsvariablen.

Fall einer Gruppierungsvariablen und einer abhängigen Variablen y , entsprechend obiger Notation folgende Gleichung aufstellen:

$$y_{jk} = a_k + \epsilon_{jk} \quad (2.51)$$

mit $a_k =$ Mittelwert von y in der Gruppe k , formal aufschlüsselbar:

$$a_k = \bar{y}_k = \bar{y} - \Delta_k = \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} y_{jk} - \frac{1}{n_k} \sum_{j=1}^{n_k} (\bar{y} - y_{jk}) \quad (2.52)$$

Der Gruppenmittelwert lässt sich somit trivial als die Differenz des Gesamtmittelwertes von Δ_k schreiben, wobei formal gilt $\Delta_k = \frac{1}{n_k} \sum_{j=1}^{n_k} (\bar{y} - y_{jk})$. Inhaltlich steht Δ_k eben für den Effekt der Gruppe k , entspricht also dem Wert, um den der Gruppenmittelwert von dem Gesamtmittelwert abweicht.

Die varianzanalytische Gleichung 2.51 kann nun um unabhängige metrische Variablen erweitert werden, was letztlich zu der oben diskutierten Verbindung beider Ansätze führt. Im einfachen Fall der Ergänzung um *eine* unabhängige Variable x gilt:

$$y_{jk} = a_k + bx_{jk} + \epsilon_{jk} \quad (2.53)$$

mit $b =$ der zu schätzende Einflusskoeffizient der Variablen x .

Diese Gleichung entspricht exakt einer in der Einleitung dieses Kapitels vorgestellten Variante von panelanalytischen Regressionsmodellen mit unterschiedlich variierenden Regressionsparametern, nämlich Gl. 2.49, in der die Regressionskonstante über die Individuen i variiert und der Regressionskoeffizient über i und t konstant ist.

Unabhängig von Detailfragen in Bezug auf die Schätzung der Parameter von 2.53 können die Vorteile einer solchen kovarianzanalytischen Modellierung benannt und auf die Paneldaten-Situation übertragen werden: Aus varianzanalytischer Perspektive kann durch die Hinzunahme von Kovariaten der Einfluss der Gruppierung auf y unter Kontrolle eben dieser Kovariate abgeschätzt werden. Wird zu dem obigen Beispiel als Kovariate x das persönliche Nettoeinkommen der Probanden hinzugenommen, so ließe sich der Einfluss des Familienstandes auf die *um den Einkommenseinfluss bereinigte* Lebenszu-

friedenheit quantifizieren. Beispielsweise würde der Effekt eliminiert, der entsteht, wenn Personen durch eine Heirat automatisch aufgrund des Wechsels der Steuerklasse einen höheren Nettoverdienst haben.²⁵ Der vom Familienstand ausgehende Einfluss bezieht sich also soz. auf die Lebenszufriedenheit, aus der der Einkommenseffekt herausgerechnet ist.

Aus einer regressionsanalytischen Perspektive kann umgekehrt genauso die Wirkung des Einkommens unter Kontrolle des Familienstandes abgeschätzt werden. Und genau diese Perspektive ist für die Betrachtung von Paneldaten von großem Reiz. Die Möglichkeit, Gruppenzugehörigkeits-Effekte herauszurechnen, erfordert nämlich lediglich die Kenntnis der Gruppenzugehörigkeit eines Objektes. Das Wissen um die konkreten Mechanismen und Motivationen, welche zu dieser Mitgliedschaft führen, ist nicht notwendig. Wenn nun bei vorliegenden Paneldaten ein Individuum eine Gruppe darstellt, dann lassen sich mit Gl. 2.53 entsprechend die *individuellen zeitkonstanten* Effekte kontrollieren. So ist es nicht mehr notwendig, zeitkonstante Variablen, die einen Einfluss auf y haben könnten, zu entdecken, Daten zu ihnen zu erheben und sie explizit mitzumodellieren. Sie rechnen sich automatisch (!) heraus.

Um zu verstehen, warum diese Effekte als zeitkonstant deklariert werden, muss man sich die Grundidee der Varianzanalyse nochmals verdeutlichen: Es werden die Werte der abhängigen Variablen y der Mitglieder einer Gruppe k nämlich immer in Bezug zum Gruppenmittelwert \bar{y}_k gesetzt. Der Gruppenmittelwert ist somit die Quantität, welche, als konstanter Bezugspunkt für alle Gruppenmitglieder einer Gruppe, diese analytisch miteinander verbindet. Die „Mitglieder“ im Paneldatenkontext sind einzelne Zeitpunkte. Ihre einzelnen Werte sind also als Abweichungen vom Gruppenmittelwert \bar{y}_k zu sehen. Der Gruppenmittelwert wiederum ist in diesem Kontext ein für ein Individuum (=Gruppe) k *konstanter* Wert im Zeitverlauf. Diese Abweichungen sind zwar als zeitliche Schwankungen aufzufassen, deren Basis stellt aber mit \bar{y}_k immer der individuelle, zeitunveränderliche Mittelwert dar. Dieser lässt sich begreifen als eine Art zeitkonstante Ausgangslage des Individuums, von derer aus sich zeitpunkt- bzw. ereignisbedingte Schwankungen ergeben.

Im oberen Beispiel zu Beginn des Kapitels 2.5 wurde der Zusammenhang zwischen der Länge der Betriebszugehörigkeit und Produktivität im Zeitver-

²⁵ wobei sich hier natürlich die kritische Frage stellt, ob nicht dieser Effekt in Hinblick auf die Lebenszufriedenheit durch ein oft damit einhergehendes gleichzeitiges Abfallen des Nettoäquivalenzeinkommens sogar überkompensiert wird.

lauf diskutiert. Angewendet auf Gl. 2.53 und den damit verbundenen rechnerischen Bereinigsmöglichkeiten könnten zeitlich relativ unveränderliche (und oftmals nicht erhobene) und die Produktivität beeinflussende Merkmale wie Intelligenz, Geschlecht, körperliche Eigenschaften und sonstige (auch genetische) Dispositionen kontrolliert werden. Diese Merkmale markieren unterschiedliche Ausgangslagen der Individuen, die ihre Arbeitsleistung mitbeeinflussen. Der in Gl. 2.53 geschätzte Koeffizient b würde dann den Effekt der Länge der Betriebszugehörigkeit im Zeitverlauf unter Konstanzhaltung dieser individuell unterschiedlichen Ausgangslagen angeben. Auf diese Art und Weise lassen sich also relativ zeitkonstante Effekte, die eigentlich unbekannt sind, explizit mitmodellieren und führen unter der Annahme, dass diese zeitkonstanten Variablen in ihrer Summe²⁶ einen signifikanten Effekt auf y haben, zu einer im Vergleich zur Querschnittsregression besseren Modellspezifikation.

Es müssen abschließend noch wichtige Punkte noch angemerkt werden: Die Konstanzhaltung inter-individueller Unterschiede führt dazu, dass letztlich **nur** die „within-variation“ der abhängigen Variablen berücksichtigt wird. Denn die „between-variation“ wird über die individuellen Regressionskonstanten „herausgerechnet“. Damit gehören die individuellen zeitkonstanten Ausgangslagen zu den unabhängigen Variablen. Wie weiter unten erläutert wird, gehören konzeptionell unabhängige Variablen in einem Regressions- bzw. Kovarianzanalyse-Modell zu den fixen, bzw. gesetzten und nicht zu den Zufallsvariablen. Daher charakterisiert die hier erläuterte Anwendung der Kovarianzanalyse auf Paneldaten die Idee der *fixed-effects*-Modelle. Folgerichtig ist nun klären, was den Unterschied zwischen fixed- und random-effects-Modellen ausmacht.

2.5.1.3 Der Unterschied zwischen fixen und zufälligen variablen Regressionskonstanten

Die Frage nach dem Unterschied zwischen fixen und zufälligen variablen Regressionskonstanten verweist auf einen generellen Sachverhalt in Regressionsmodellen: Einige Variablentypen werden als zufällig, andere als fix betrachtet. Um diese gedanklich-konzeptionelle Unterscheidung zu verstehen, muss

²⁶ Somit muss das Kriterium nicht erfüllt sein, dass alle zeitkonstanten Variablen einzeln einen signifikanten Effekt haben. Es muss bei einzelner Betrachtung dieser Variablen noch nichtmals *eine* Variable einen signifikanten Effekt haben.

die inferenzstatistische Grundidee deutlich werden, dass Eigenschaften von Merkmalsträgern und Zusammenhänge zwischen diesen Eigenschaften bereits *vor* einer Datenerhebung bzw. einer Datenanalyse existieren.

Bevor also überhaupt eine Stichprobe gezogen wird, Daten erhoben werden und eine Analyse durchgeführt wird, existiert eine Grundgesamtheit G . Ferner wird angenommen, in dieser Grundgesamtheit G gilt eine wahre (aber uns unbekannt) und unter allen gängigen Annahmen zu Regressionsmodellen korrekt spezifizierte Regressionsgleichung, welche den Einfluss zwischen einer abhängigen Variablen x und einer abhängigen Variablen y quantifiziert:

$$y = a + bx + \epsilon \quad (2.54)$$

mit

a = Regressionskonstante

b = Regressionsparameter zur Variablen x

ϵ = Residuum.

Nun wird aus G eine Zufallsstichprobe mit n Elementen gezogen und die Datenerhebung durchgeführt. Es wird (vorübergehend!) angenommen, dass die Werte von y mithilfe eines experimentellen Designs erhoben werden. Im experimentellen Design kann nämlich der Reiz, also die Ausprägung der x -Variablen kontrolliert gesetzt werden. Auch wenn unabhängige Variablen einer linearen Regression als metrisch angenommen werden, so wird hier (auch wieder vorübergehend!) vereinfachend x als eine binäre Variable deklariert, so dass nur zu unterscheiden ist, ob ein Reiz gesetzt wurde ($x = 1$) oder nicht ($x = 0$). In einem Experiment wird ja sozusagen die „Realität in der Grundgesamtheit“ simuliert. Somit sind bereits vor dem Experiment bzw. unabhängig von dessen Durchführung einige Elemente der Grundgesamtheit G mit einem Reiz versehen ($x = 1$, z.B. die Einnahme eines Medikamentes) und andere nicht ($x = 0$, das Medikament wird nicht eingenommen). Anhand dieser Unterscheidung lässt sich G als eine in Subpopulationen zerteile bzw. geschichtete Gesamtheit verstehen. In dem einfachen Fall hier teilt sich G folglich in zwei Gruppen G_1 und G_2 auf, entsprechend der Unterscheidung zwischen $x = 1$ (G_1) und $x = 0$ (G_2). Diese Situation lässt sich problemlos auf eine multiple Regression mit mehreren (metrischen) unabhängigen Variablen erweitern: Demnach wird G gedanklich in so viele Schichten geteilt,

wie Merkmalskombinationen der x -Variablen existieren.

In diesem Verständnis entspricht die zufällige Auswahl einer Person für ein Experiment und ihre Zuordnung in die Experimentalgruppe ($x = 1$) dem Prozess der Ziehung einer Person aus der Subpopulation der Personen, die dem Reiz ($x = 1$) ausgesetzt sind – analog dazu ist die Stichprobenziehung im Falle $x = 0$ zu verstehen. Nun wird eine aus der Population $x = 1$ gezogene Person einem Reiz ausgesetzt und reagiert auf diesen Reiz, produziert also *scheinbar* einen y -Wert. Doch diese Auffassung muss korrigiert werden, wenn angenommen wird, dass in G bereits vor dieser Untersuchung ein fester, wahrer Einfluss von x auf y besteht. Die Gleichung $y = a + bx + \epsilon$ quantifiziert nämlich bereits *vor* der Durchführung des Experiments den bestehenden linearen Zusammenhang. b ist also bereits vorhanden (auch wenn uns bekannt).

Ferner ist es aufgrund der Komplexität der meisten Zusammenhangsstrukturen selten realistisch anzunehmen, dass y immer perfekt durch den Term $a + bx$ erzeugt wird. Daher ist in der Gleichung ein Störterm ϵ enthalten, welcher *zufällige* Abweichungen von der perfekten, aber unrealistischen linearen Zusammenhangsstruktur $y = a + bx$ „einfängt“. Da x im experimentellen Design für eine Messung gesetzt wird und mit $a + bx$ der feste Einfluss von x auf y charakterisiert wird, fängt die Messung von y letztlich die Abweichung von diesem idealen linearen Zusammenhang ein. Es wird demnach, gegeben dem x -Wert bzw. der Subpopulationszugehörigkeit, **das Residuum „gemessen“ bzw. „erfasst“!** Das so erfasste Residuum erzeugt (unter der Bedingung von x) durch die Addition mit $a + bx$ den y -Wert. y ist somit gedanklich als eine lineare Transformation von dem Residuum zu sehen. Aus diesem erweiterten Blickwinkel sollte nun die vertraute Gleichung $y = a + bx + \epsilon$ gelesen werden.

Das Residuum ϵ umfasst die Summe von (z.T. unkalkulierbaren) Einflüssen, welche neben der festen Wirkung von x einen Einfluss auf die Messung von y haben. Da oben angenommen wird, dass der Einfluss von x auf y durch die lineare Gleichung $y = a + bx + \epsilon$ nicht von Annahmeverletzung betroffen ist und folglich korrekt spezifiziert ist, weist das Residuum *keine Systematik* auf. Da zusätzlich die Probanden *zufällig* aus den Subpopulationen G_1 und G_2 gezogen wurden, ist das Zustandekommen der Residualwerte innerhalb der Subpopulationen als ausschließlich zufallsbedingt zu sehen. Daher wird das Residuum als eine Zufallsvariable, gegeben x , verstanden. Da es sich bei den y -Werten lediglich um lineare Transformationen der Residualwerte handelt, ist folglich auch y *als eine Zufallsvariable aufzufassen* (von Auer

2007: 68f.). Damit ist eine klare analytische Unterscheidung zwischen der unabhängigen Variablen x und der abhängigen Variablen y zu treffen: Erstere ist eine nicht-stochastisch fixe Variable, letztere ist eine Zufallsvariable.

Dieser Gedanke lässt sich sofort auf das multiple Regressionsmodell erweitern. Es muss in Bezug auf unabhängige Variablen lediglich im Plural gesprochen werden. Demnach *sind* die unabhängigen Variablen als nicht-stochastisch zu verstehen. Zur Vereinfachung wird ab sofort die Konvention eingeführt, mehrere unabhängige Variablen mit einem transponierten Spaltenvektor zu bezeichnen. Für K unabhängige Variablen gilt entsprechend: $\mathbf{x}' = (x_1, \dots, x_K)$. Auch die Größe b muss damit einhergehend zu einem transponierten Spaltenvektor mit zu den Elementen von \mathbf{x}' korrespondierenden Regressionsparametern $\mathbf{b}' = (b_1, \dots, b_K)$ erweitert werden. Folglich lässt sich die einfache Regressionsgleichung $y = a + bx + \epsilon$ im multiplen Falle schreiben als $y = a + \mathbf{b}'\mathbf{x} + \epsilon$

Warum ist diese Unterscheidung wichtig? Ausgehend von der Deklaration der Residuen als Zufallsvariablen lassen sich einige Annahmen über Regressionsmodelle formulieren; die Erfüllung bzw. Verletzung dieser Annahmen ist bedeutend für die Einschätzung, ob ein Modell oder Teile des Modells korrekt spezifiziert sind. Die Annahmen über das Residuum als Zufallsvariable, zusammen mit der Annahme des nicht-stochastischen Charakters von \mathbf{x}' , erlaubt den Nachweis, dass es sich bei den Schätzern nach dem „Kleinste-Quadrat-Prinzip“ (KQ-Prinzip) um BLUE-Schätzer²⁷ handelt (vgl. von Auer 2007: 83, 430). Ohne ins Detail zu gehen sei kurz erwähnt, dass dieser Nachweis deshalb gelingt, weil die Eigenschaft der Nicht-Zufälligkeit von \mathbf{x}' u.a. an einer bestimmten Stelle eine entscheidende mathematische Umformung erlaubt (vgl. von Auer 2007: 83, 430).²⁸

Nun basieren aber sozialwissenschaftliche Studien oftmals nicht auf dem experimentellen Design, sondern entstammen einem Ex-Post-Facto-Design, wie z.B. einer Befragung. Da in einer Befragung die x -Werte nicht als Reize manipuliert werden können, müssen sie streng genommen ebenfalls als stochastische Zufallsvariablen angesehen werden – unter der Annahme, dass der Pool der Befragten durch die Realisation einer Zufallsstichprobe zustande

²⁷ BLUE steht für den besten (=effizientesten) Schätzer aus der Gruppe der unverzerrten linearen Schätzer (vgl. allgemein zu den Voraussetzungen der BLUE-Eigenschaft von Auer 2007: 74ff).

²⁸ Denn es gilt für eine nicht zufällige, fixe Größe x , dass ihr Erwartungswert $E(x) = x$. Diese Vereinfachung gegenüber den Erwartungswerten von Zufallsvariablen ist für die angesprochene mathematische Beweisführung entscheidend.

kam. Es lässt sich aber mathematisch nachweisen, dass mit zunehmendem Stichprobenumfang n ($n \rightarrow \infty$) die Schätzer eines Regressionsmodells nach der KQ-Methode dennoch die BLUE-Eigenschaft, zumindest asymptotisch besitzen. Da dieses Grundkonzept also im Falle von stochastischen unabhängigen Variablen nicht in sich zusammenbricht, gleichzeitig aber gerade auf der Prämisse von fixen x -Variablen aufbaut, kann weiter konzeptionell zwischen festen unabhängigen Variablen x und der Zufallsvariablen y unterschieden werden – auch wenn Befragungsdaten vorliegen.

2.5.1.4 Die Bedeutung der Unterscheidung zwischen fixen und zufälligen variablen Regressionskonstanten für Regressionsmodelle mit Paneldaten (fixed- und random-effects-Modelle)

Warum war dieser Gedankenexkurs derart wichtig? Weil er dazu verhelfen soll, die Unterscheidung zwischen random- und fixed-effects-Modellen zu verstehen. Die mit dieser Unterscheidung einhergehende Frage stellt sich in Bezug auf die Behandlung der variablen Konstante a_i :²⁹ Soll a_i als eine fixe (F-Fall) oder als eine Zufallsvariable (Z-Fall) aufgefasst werden? Auf Basis obiger Ausführungen lässt sich diese Frage konkretisieren:

- F-Fall – a_i wird zu den unabhängigen Variablen gezählt und explizit in die Parameterschätzung involviert
- Z-Fall – a_i wird als eine Komponente des Residuums gesehen; da das Residuum eine Zufallsvariable darstellt, zählt a_i als eine seiner Komponenten auch zu dem Lager der Zufallsvariablen

Aufbauend auf dieser Konkretisierung ist nun zu fragen, welche Konsequenzen die Unterscheidung zwischen dem F- und dem Z-Fall für die Interpretation der Regressionsparameter dieser Modelle hat.

Für die folgenden Ausführungen soll, entsprechend der Gleichungen 2.53 und 2.48 und der obigen Erweiterung auf den Fall *multipler* Regression, folgende Ausgangsgleichung gelten:

$$y_{it} = \mathbf{b}'\mathbf{x}_{it} + a_i + \epsilon_{it} \quad (2.55)$$

²⁹ i stellt entsprechend obiger Notationen, z.B. Gl. 2.48, den Laufindex für die einzelnen Objekte in der Querschnittsbeobachtung dar.

mit

i = Laufindex für die einzelnen Objekte in der Querschnittsbetrachtung

t = Laufindex für die einzelnen Zeitpunkte (Wellen)

y = Abhängige Variable

$\mathbf{b}' = (b_1, \dots, b_K)$ = Transponierter Spaltenvektor mit K Regressionsparametern

$\mathbf{x}' = (x_1, \dots, x_K)$ = Transponierter Spaltenvektor mit K unabhängigen Variablen

a = Regressionskonstante

ϵ = Residuum, allerdings nur dann, wenn ein fixed-effects-Modell vorliegt.

Mit u_{it} wird ferner eine weitere Größe eingeführt:

$$u_{it} = a_i + \epsilon_{it} \quad (2.56)$$

u_{it} stellt das Residuum dar, wenn ein random-effects-Modell vorliegt. Somit lässt sich Gl. 2.55 reformulieren:

$$y_{it} = \mathbf{b}'\mathbf{x}_{it} + u_{it} \quad (2.57)$$

Worin die Unterscheidung zwischen u_{it} im random- und ϵ_{it} im fixed-effects-Modell begründet ist, wird weiter unten deutlich.

IM F-FALL (fixed-effects-Modell) gehen, wie oben erläutert, die Regressionskonstanten der einzelnen Personen explizit in die Parameterschätzung ein. Damit werden die Regressionsparameter von \mathbf{x}'_{it} unter der Kontrolle der individuellen Ausgangslagen geschätzt und entsprechend interpretiert. M.a.W. rechnen sich inter-individuellen Unterschiede völlig heraus. Es wird nur die within-variation als Variations- und somit Informationsquelle für die Berechnung des Einflusses von \mathbf{x}'_{it} auf y_{it} genutzt. Ein Regressionskoeffizient in einem fixed-effects-Modell sagt demnach aus, wie sich y_{it} gegeben \mathbf{x}'_{it} im Zeitverlauf verändert, *wenn inter-individuelle Ausgangslagen konstant*

gehalten werden. Dies stellt eine einseitige Fokussierung der zeitlichen Entwicklung zwischen \mathbf{x}'_{it} und y_{it} in den Vordergrund. Die explizite Modellierung von unabhängigen Variablen, die zeitkonstant sind und die a priori einen Effekt auf y_{it} haben, ist nicht möglich – allein schon mathematisch nicht (s.u.). Es können also keine Effekte zeitunveränderlicher Merkmale, die vielleicht für die Analyse von Bedeutung sind, berechnet werden, da sie sich vorher schon durch die explizite Modellierung von a_i als fixe Variable implizit herausgerechnet haben.

IM Z-FALL (random-effects-Modell) werden die Regressionsparameter von \mathbf{x}'_{it} nicht unter der expliziten Konstanthaltung der individuellen Unterschiede berechnet. Die a_i -Werte stellen in diesem Sinne keine unabhängigen Variablen dar. a_i wird im random-effects-Modell vielmehr als Komponente des Residuums aufgefasst und gehört somit zu den Zufallsvariablen. Sie stellt, gegeben \mathbf{x}'_{it} , die zufällige Abweichung von der Regressionsgleichung $y_{it} = \mathbf{b}'\mathbf{x}_{it}$ dar, welche für ein konkretes Individuum i über die Zeit konstant ist. In diesem Sinne enthält a_i , analog zu den fixed-effects-Modellen, ebenfalls zeitunveränderliche Merkmale eines Individuums, welche eine Art Ausgangslage bzw. Ausgangsniveau konstituieren. Nur wird dieser zeitkonstante Individualeffekt nicht explizit als zu schätzender Parameter in die Schätzung involviert, sondern als zu minimierender „Rest“, welcher einen perfekten Zusammenhang zwischen \mathbf{x}'_{it} und y_{it} „stört“.

Die Extraktion einer zeitkonstanten Komponente aus dem Residuum u_{it} , nämlich eben a_i , hat zur Konsequenz, dass eine zentrale Annahme der KQ-Methode verletzt wird: Die Unkorreliertheit der Residuen untereinander. Folglich gilt nicht mehr die Annahme, dass die Kovarianz von Residuen zweier verschiedener Zeitpunkte $t = r$ und $t = s$ (mit $r \neq s$) ϵ_{ir} und ϵ_{is} Null beträgt, wenn i in beiden Fällen gleich ist.

Diese Erkenntnis ist nicht nur mathematisch, sondern auch intuitiv nachvollziehbar. Schließlich können z.B. im Falle einer Befragung zwei Teilnahmen ein und derselben Person zu zwei verschiedenen Zeitpunkten nicht als stochastisch unabhängig betrachtet werden. Folglich kann auch in einem Regressionsmodell das Residuum einer konkreten Person i zum Zeitpunkt $t = r$ nicht als unkorreliert mit dem Residuum derselben Person zum Zeitpunkt $t = s$ (mit $r \neq s$) aufgefasst werden. Wird diese Erkenntnis im Zuge der Schätzung der Regressionsparameter ignoriert, dann sind die resultierenden KQ-Schätzer nicht mehr effizient und verlieren somit die BLUE-Eigenschaft

(vgl. Hsiao 2005: 35; von Auer 2007: 74ff). Folglich muss das Vorgehen bei der KQ-Schätzung modifiziert werden, indem die Korreliertheit der Residuen von gleichen Personen zu verschiedenen Zeitpunkten mithilfe von a_i explizit in der Varianz-Kovarianzmatrix der Residuen berücksichtigt wird (s.u.). Dies führt zu der sog. „generalized-least-squares“-Schätzung (GLS-Schätzung). Die so geschätzten Regressionsparameter machen das random-effects-Modell aus. Ohne dieses Prinzip mathematisch zu erläutern sei gesagt, dass auf diesem Wege die Schätzer die BLUE-Eigenschaft erreichen. Der Zusammenhang zwischen x_{it} und y_{it} wird auf diesem Wege „korrekter“ berechnet, da im Gegensatz zum z.B. pooled-Modell (s. Gl. 2.50) der panelspezifischen Struktur der Daten über die Einbeziehung einer zeitinvarianten Fehlerkomponente Rechnung getragen wird.

Nun stellt sich die Frage nach der Interpretation der GLS-Schätzer im random-effects Modell: Es lässt sich mathematisch zeigen, dass auf dem Weg zur Schätzung sowohl die within-variation als auch die between-variation der abhängigen Variablen verarbeitet wird (vgl. Hsiao 2005: 37f.): Der GLS-Schätzer stellt einen gewichteten Durchschnitt aus dem „within-Schätzer“ (dies ist der Schätzer des fixed-effects-Modells) und dem „between-Schätzer“ (dies ist der Schätzer des sog. between-effects-Modells)³⁰ dar. Das Gewicht ist davon abhängig, wie groß der Anteil der Varianz der Fehlerkomponente a_i an der Gesamtvarianz der Residuen ist. Das random-effects-Modell ist somit als eine flexible (da sich automatisch gewichtende) Kompromisslösung zwischen dem fixed-effects-Modell (es wird nur die within-Variation der y -Werte im Zeitverlauf unter der Kontrolle inter-individueller Unterschiede berücksichtigt) und dem pooled-Modell (s. Gl. 2.50; die Unterscheidung zwischen within- und between-Variation wird völlig ignoriert; die einzelnen Varianzen werden einfach aufsummiert) gesehen.

Zusammenfassend lässt sich kontrastieren:

Das fixed-effects-Modell behandelt die Regressionskonstante a_i als eine fixe unabhängige Variable. Dies hat zur Konsequenz, dass sich inter-individuelle Unterschiede bei der Schätzung der Parameter von \mathbf{x}'_{it} vollständig herausrechnen und daher nur die within-variation der abhängigen Variablen y_{it}

³⁰Das between-effects-Modell ignoriert die within-variation völlig: Zunächst wird pro involvierte Variable \mathbf{x}' und y für jede Person i ihr eigenes arithmetisches Mittel entlang der Messzeitpunkte errechnet; z.B. für y : $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ (mit T = letzter Zeitpunkt). Schließlich wird mit diesen neuen „Mittelwertvariablen“ eine „normale“ Regression nach der KQ-Methode durchgeführt.

berücksichtigt wird.

Das random-effects-Modell konzeptualisiert a_i als Zufallsvariable und als Komponente des Residuums. a_i gehört somit *nicht* zu den unabhängigen Variablen. Bei der Schätzung der Parameter von \mathbf{x}'_{it} wird folglich, wie auch auch bei der Regression des pooled-Modells, sowohl die within- als auch die between-variation der abhängigen Variablen y_{it} verwertet. Im Gegensatz zum pooled-Modell werden aber beide Variationsquellen analytisch differenziert. Diese Differenzierung wird erreicht, indem mit der Extraktion von a_i als eine zeitinvariante Residualkomponente des „Gesamtresiduums“ einer der für Paneldaten-Regressionen bedeutendsten Eigenschaft Rechnung getragen wird: Der Korreliertheit von zeitversetzten Residuen ein und derselben Person.

Nachdem nun die Unterschiede zwischen dem fixed- und dem random-effects-Modell geklärt sind, stellt sich die Frage, nach welchen Kriterien in einer konkreten Analysesituation eines der beiden Ansätze gewählt werden soll. Um diese Frage zu beantworten wird zunächst pro Verfahren die konkrete Spezifikation und Parameterschätzung anhand eines Beispiels vorgestellt. Schließlich werden Vor- und Nachteile diskutiert.

2.5.1.5 Spezifikation der einzelnen Modelle und Schätzung der Koeffizienten

In diesem Abschnitt soll zunächst die inhaltliche Bedeutung von a_i in fixed-effects-Modellen (ab sofort abgekürzt: FEM) und random-effects-Modellen (ab sofort abgekürzt: REM) konkretisiert werden. Es werden die Ausgangsgleichungen formuliert. Zur Vereinfachung enthalten diese nur jeweils eine x -Variable. Sie lassen sich aber problemlos auf den multiplen Fall ³¹ Dann wird die Schätzung der Koeffizienten vergleichend vorgestellt.

Der nächste Abschnitt diskutiert die Vor- und Nachteile der FEM- und REM-Modelle. Im abschließenden Abschnitt soll dann das Verständnis dieser theoretischen Ausführungen anhand eines Beispiels vertieft werden.

Zunächst zum FEM:

Als erstes soll die Ausgangsgleichung präzisiert werden, indem Gl. 2.55 um

³¹Zur Übertragung auf den multiplen Fall vgl. Hsiao 2005.

die für Individuen i konstante Regressionskonstante a erweitert wird. a ist aus der Querschnittsregressionen bekannt. Sie wurde bislang aus Gründen der Übersichtlichkeit weggelassen. Die Gleichung mit nur einer x -Variablen lautet dann (entsprechend der Notation in Gl. 2.55):

$$y_{it} = a + a_i + bx_{it} + \epsilon_{it} \quad (2.58)$$

Eigentlich stellt auch schon Gl. 2.58 eine multiple Regression dar, da neben x die einzelnen a_i -Größen unabhängige Variablen darstellen (s.o.). Wird aber Gl. 2.58 für eine konkrete Person i betrachtet, dann lässt sich entsprechend dem Fall einer einfachen Regression eine Regressionsgerade zwischen x und y vorstellen.

Die Größe a_i ergibt dann in Addition mit der Regressionskonstante a die Stelle, an der die Regressionsgerade der i -ten Person die y -Achse schneidet (an der also $x = 0$ gilt).

Ferner ist zu bedenken, dass b für alle Personen i konstant ist. Daher stellen alle individuellen Regressionsgeraden der n Personen *parallel* zueinander liegende Geraden dar. Der a_i -Wert gibt an, um wieviel y -Einheiten die Regressionsgerade im Vergleich zur „Basis“ a verschoben wird. Jedes Individuum besitzt somit soz. seine *eigene Regressionskonstante*. Der *individuelle Schnittpunkt* mit der y -Achse ergibt sich aus $a + a_i$.³² Die relative Höhe der Regressionsgeraden markiert die „Ausgangslage“ eines Individuums (vgl. das graphische Beispiel in Kohler 2008: 252).

Nun wurde bereits oben im Kontext der kovarianzanalytischen Überlegungen erwähnt, dass eine Möglichkeit zur Bestimmung von a_i darauf basiert, für jedes Individuum eine Dummy-Variable zu bilden und sie zu den unabhängigen Variablen des Modells zu zählen. Umfasst die Stichprobe n Individuen, so werden $n - 1$ Dummy-Variablen benötigt.³³ Die Koeffizienten der Dummy- und der x -Variablen können dann mit der gewöhnlichen KQ-Methode geschätzt werden.

Weist ein FEM-Modell k x -Variablen auf, so umfasst es insgesamt

³² Wird in vielen Regressionsmodellen a durch Zentrierung der Variablen weggelassen, so bedeutet dies nichts anderes, als $a = 0$. Die eben erwähnten Zusammenhänge gelten allerdings genauso; in diesem Modell ist es sinnlos, mit zentrierten Variablen zu arbeiten, da durch die Einführung von Dummy-Variablen, welche per Definition *nicht zentriert werden können*, im endgültigen Modell sowieso eine Regressionskonstante $a \neq 0$ auftaucht.

³³ Von n wird 1 abgezogen, da der letzte bzw. irgendein Fall (in anderen Kontexten wird dieser Fall als *Referenzkategorie* bezeichnet) sich automatisch dadurch ergibt, dass alle Dummy-Variablen die Ausprägung Null aufweisen; der n -te Dummy wäre somit redundant.

$k + n - 1$ unabhängige Variablen. Die Anzahl der zu schätzenden Parameter übersteigt im Falle einer Querschnittsregression die überhaupt vorhandene Anzahl n der Fälle. Daher wären, wenn Daten nur zu einem Zeitpunkt vorliegen würden ($T = 1$), die Parameter von 2.58 nicht schätzbar, da nicht genügend Informationen vorhanden wären.

Somit kann ein individueller Effekt nur dann errechnet werden, wenn *mehrere Werte* eines Individuum i zur Verfügung stehen. Mehrere Werte können wiederum nur dann gegeben sein, wenn Messungen *zu mehreren Zeitpunkten* vorliegen. In einem Querschnittsdatensatz würde aber der einzige y -Messwert der Person i automatisch den individuellen Effekt (bzw. dem Regressionskoeffizienten der Dummy-Variable) ausmachen und das System wäre somit redundant.³⁴ Erst mit der Einführung der durch wiederholte Messungen zustande kommenden *within variation* erhalten die Individualeffekte a_i eine inhaltliche Daseinsberechtigung (s.o.).

Es ist offensichtlich, dass mit einem hohen n auch die Anzahl der zu schätzenden a_i -Koeffizienten steigt. So kann die Berechnung eines FEM-Modells sehr aufwendig werden, da es, wie oben erwähnt, $k + n - 1$ Regressionskoeffizienten zu schätzen gilt. Wenn z.B. $n = 1.000$, dann müssten entsprechend Gl. 2.58 1.001 Koeffizienten ermittelt werden. Wie allerdings weiter unten ausgeführt wird, ist eine Berechnung der Dummy-Koeffizienten nicht zwingend notwendig.

Die a_i -Werte lassen sich zudem relativ leicht ermitteln. Wie oben im Kontext kovarianzanalytischer Überlegungen erwähnt wurde, beziehen sich die Dummy-Koeffizienten eines Individuums i auf seinen individuellen Mittelwert \bar{y}_i . Konkret ergibt die KQ-Schätzung für Gl. 2.58 (vgl. Hsiao 2005: 33):

$$\hat{a}_i = \bar{y}_i - a - b\bar{x}_i \quad (2.59)$$

mit

\hat{a}_i = Geschätzter a_i -Wert (hier nach der KQ-Methode)

\bar{x}_i = Mittelwert der i -ten Person in Bezug auf die Variable x

³⁴Dies würde sich rechnerisch darin bemerkbar machen, dass im Falle von Querschnittsdaten alleine die lineare Verknüpfung von Individual-Dummy-Variablen die abhängige Variable vollständig erklären würde – unabhängige Variablen wären überflüssig (und deren Koeffizienten könnten wg. $k + n - 1 > n$ gar nicht erst berechnet werden), das Modell würde allerdings auch nichts aussagen.

Es wird also von dem individuellen Mittelwert \bar{y}_i der Term $(a + b\bar{x}_i)$ abgezogen. Die individuelle Konstante a_i ist somit der individuelle y -Mittelwert, welcher um den Einfluss von x bereinigt wurde.

Durch einen rechnerischen Trick (vgl. Hsiao 2005: 32) lässt sich b schätzen, ohne dass die einzelnen a_i -Werte vorher bestimmt sein müssen. Somit ist man eben *nicht* darauf angewiesen, $k + n - 1$ Koeffizienten zu bestimmen. I.d.R. werden die a_i -Werte auch nicht zur Interpretation gebraucht. Man interessiert sich hauptsächlich für den Einfluss von x auf y , wenn die individuellen zeitinvarianten Merkmale herausgerechnet sind. Eine aufwendige Bestimmung aller Dummy-Koeffizienten macht nur in seltenen Fällen Sinn – z.B. wenn Daten aus einer sehr kleinen Grundgesamtheit vorliegen und ihre Elemente so spezifisch sind, dass ihre a_i -Werte einzeln analysiert werden sollen.³⁵ Bei wenigen Fällen ist aber wiederum die Anzahl $k + n - 1$ relativ klein.

Bevor die Schätzungen des b -Koeffizienten vorgestellt werden, werden die wichtigsten Größen in Bezug auf die Variablen x und y nochmals tabellarisch übersichtlich dargestellt (die übrigen Größen werden entsprechend obiger Ausführungen, z.B.: 2.55, 2.56 und 2.57 definiert. Ferner stehen n und T respektive für den letzten Wert von i bzw. t).

	Kontext: Variable x	Kontext: Variable y
Variable variiert über i (Objekte) und t (Zeitpunkte):	x_{it}	y_{it}
Arithmetisches Mittel des Objektes i	\bar{x}_i	\bar{y}_i
Globales arithmetisches Mittel (über alle i und t)	\bar{x}	\bar{y}

Unter Verwendung von 2.59 bzw. des oben erwähnten rechnerischen Tricks kann nun nach der KQ-Methode b aus Gl. 2.58 geschätzt werden (warum die Formel in blauer Farbe erscheint, wird an späterer Stelle geklärt):

³⁵ z.B. wenn die Elemente der Grundgesamtheit zusammengenommen die Menge aller Bundesländer Deutschlands bilden und die Ausgangsniveaus der einzelnen Bundesländer miteinander verglichen werden sollen.

$$\widehat{b}_f = \frac{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2} \quad (2.60)$$

mit

\widehat{b}_f = Schätzer von b aus Gl. 2.58 nach der KQ-Methode. Das f symbolisiert das **f**ixed-effects-Modell.

Die Formel erinnert an die Schätzung des Regressionskoeffizienten einer einfachen Regression von y auf x im Falle von Querschnittsdaten: $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. Der wesentliche Unterschied ist, dass im Falle des fixed-effect-Schätzers aus Gl. 2.60 die Variablenwerte von den **individuellen** arithmetischen Mitteln abgezogen werden. Dies ergibt sich rechnerisch eben aus der Tatsache, dass mit a_i *inter*-individuelle Unterschiede herausgerechnet wurden. Somit bleibt zur Schätzung von b **nur** die Informationsquelle über, die sich aus *intra*-individuellen Unterschieden im Zeitverlauf ergibt. Dies sind die Abweichungen der Werte einer Person von ihrem eigenen arithmetischen Mittel, welche die „within variation“ charakterisieren. Die „between variation“ wird entsprechend nicht berücksichtigt.

Die Schätzung \widehat{b}_f ist also zu interpretieren als der Betrag, um den der geschätzte y -Wert steigt, wenn x um eine Einheit steigt und (!) alle Unterschiede zwischen den Individuen vorher herausgerechnet wurden. Er gibt soz. den durchschnittlichen intra-individuellen Veränderungswert von y unter der Bedingung von x im Zeitverlauf wieder.

Es sein nochmals darauf hingewiesen, dass es sich bei \widehat{b}_f um einen gewöhnlichen Schätzer nach der KQ-Methode handelt. Würden, statt des Rückgriffs auf Gl. 2.59, die individuellen Regressionskonstanten als Regressionskoeffizienten von Dummy Variablen berechnet werden, so ergäbe sich die übliche KQ-Gleichung des multivariaten Falls zur Schätzung der Koeffizienten – in Matrizenform: $\mathbf{b} = (X'X)^{-1}X'y$. \mathbf{b} = Spaltenvektor mit den Regressionskoeffizienten; \mathbf{y} = Spaltenvektor mit den Werten der abhängigen Variablen y (die Werte differenziert nach i und t werden einfach untereinander geschrieben); X = Matrix mit den Werten der unabhängigen Variablen. Jede unabhängige Variable

stellt innerhalb der Matrix einen Spaltenvektor dar. Wichtig ist, dass die Matrix aus $k + n - 1$ Spalten besteht: $n - 1$ Spaltenvektoren stehen für die Dummy-Variablen und k Spalten für die „echten“ unabhängigen Variablen. Im hier diskutierten Falle $k = 1$ entspricht dann der letzte Wert des Vektors \mathbf{b} dem Schätzer \hat{b}_f aus Gl. 2.60.

Schätzung im REM:

Wie oben gezeigt, berücksichtigt das **FEM** nur die in Paneldaten enthaltene „within-variation“. Das andere Extrem, in dem nur die Informationen der „between-variation“ genutzt werden, existiert ebenfalls: Das between-effects-Modell (BEM). Warum dieses Modell im Kontext des **REM** vorgestellt wird, wird im weiteren Verlaufe deutlich.

Wie oben in einer Fußnote erwähnt, ignoriert das **BEM** die within-variation völlig: Zunächst wird pro involvierte Variable x und y für jede Person i ihr eigenes arithmetisches Mittel entlang der Messzeitpunkte errechnet, also \bar{x}_i und \bar{y}_i . Schließlich wird mit diesen neuen „Mittelwertsvariablen“ eine „normale“ Regression nach der KQ-Methode durchgeführt.

Sinnvoll ist diese Anwendung, wenn das Ausmaß der „within-variation“ inhaltlich uninteressant ist und statistisch gesehen gering ausfällt. Die Zusammenfassung mehrerer Werte einer Person i zu ihrem individuellen Mittelwert verbessert dann die Reliabilität der Daten im Vergleich zur Querschnittsregression. Denn es kann angenommen werden, dass sich kleine Zufallsschwankungen in den Werten einer Person i im Zeitverlauf gegenseitig herausrechnen. Ist hingegen eine klare Systematik in der zeitlichen Entwicklung der y -Werte anzunehmen, dann ist das BEM nicht geeignet.

Die Regressionsgleichung des BEM im einfachen Falle einer Variablen x lässt sich ausdrücken als:

$$\bar{y}_i = a + b\bar{x}_i + \epsilon_i \quad (2.61)$$

Die Schätzung \hat{b}_b (das kleine b steht für „between“) des Koeffizienten im BEM lautet dann (die Wahl der Farbe rot wird weiter unten erläutert):

$$\hat{b}_b = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2} \quad (2.62)$$

Nun zurück zum **REM**: Wie oben erläutert wurde, wird im REM a_i nicht

mehr als fixe Größe sondern als Zufallsvariable aufgefasst, die eine Komponente des Residuums darstellt. Dies schlägt sich auch in der Ausgangsgleichung des REM nieder. Angelehnt an die Gleichungen 2.55, 2.56 und 2.57 lautet sie:

$$y_{it} = bx_{it} + a_i + \epsilon_{it} \quad (2.63)$$

Wenn also a_i zum Residuum gehört, dann lassen sich die zwei Komponenten $a_i + \epsilon_{it}$ zusammenfassen zu u_{it} . Eingesetzt in 2.63 ergibt sich:

$$y_{it} = bx_{it} + u_{it} \quad (2.64)$$

Oben wurde bereits erläutert, dass die Residuen u_{it} im REM nicht als unkorreliert angenommen werden können. Damit wird eine zentrale Annahme verletzt, auf der die KQ-Schätzung basiert. Die Korreliertheit von u_{it} lässt sich exemplarisch für die Kovarianz von zwei Residualvariablen u_{i1} und u_{i2} (mit $t = 1$ und $t = 2$) zeigen:

$$\begin{aligned} Cov(u_{i1}u_{i2}) &= \sum (u_{i1}u_{i2}) \\ &= \sum [(a_i + \epsilon_{i1})(a_i + \epsilon_{i2})] \\ &= \sum [a_i^2 + a_i\epsilon_{i2} + \epsilon_{i1}a_i + \epsilon_{i1}\epsilon_{i2}] \\ &= \sum a_i^2 + \sum a_i\epsilon_{i2} + \sum \epsilon_{i1}a_i + \sum \epsilon_{i1}\epsilon_{i2} \\ &= \sum a_i^2 \end{aligned} \quad (2.65)$$

Folgende Annahmen, abgeleitet aus gewöhnlichen linearen Regressionsmodellen nach der KQ-Methode, liegen den Umformungen in 2.65 zugrunde:

- u_{it} , a_i und ϵ_{it} besitzen einen Erwartungswert von 0, daher vereinfacht sich die Gleichung der Kovarianz zur Summe der Residuenprodukte $\sum (u_{i1}u_{i2})$
- a_i und ϵ_{it} sind miteinander unkorreliert, deshalb gilt für $\sum a_i\epsilon_{i2} = 0$ und für $\sum \epsilon_{i1}a_i = 0$
- Die Fehler ϵ_{it} sind untereinander unkorreliert, daraus resultiert $\sum \epsilon_{i1}\epsilon_{i2} = 0$

Mit der letzten Zeile von 2.65 wird also die oben getroffene Annahme, dass die Residuen im REM nicht unkorreliert sind, mathematisch nachgewiesen. $\sum a_i^2$ ist nichts anderes, als die Varianz von a_i , bezeichnet mit $V(a_i)$. Folglich lautet der Erwartungswert der Kovarianz zweier u_{it} -Variablen (wenn ihre t -Werte verschieden sind): $V(a_i)$.

Weiterhin kann, unter der getroffenen Annahme, dass a_i und ϵ_{it} unkorreliert sind, aus der Aufteilung des Fehlerterms $u_{it} = a_i + \epsilon_{it}$ die Aufteilung der Fehlervarianzen abgeleitet werden:

$$V(u_{it}) = V(a_i) + V(\epsilon_{it}) \quad (2.66)$$

Entsprechend der Homoskedastizitäts-Annahme sind die einzelnen Varianzen für i und t konstant. Folglich vereinfacht sich Gl. 2.66 zu:

$$V(u) = V(a) + V(\epsilon) \quad (2.67)$$

Gleichungen 2.65 und 2.67 konstituieren die Ausgangslage des **REM**, dessen Koeffizienten nun mit einem alternativen Schätzverfahren bestimmt werden müssen. Dann die KQ-Methode setzt die Unkorreliertheit von u_{it} voraus. Stattdessen muss auf die sog. „generalized-least-squares“-Methode (GLS) zurückgegriffen werden, mithilfe derer verschiedene Annahmen hinsichtlich der Korrelationsstruktur von Residuen berücksichtigt werden können. Die Herleitung der GLS-Schätzung zeigt Hsiao (2005: 35ff.). Hier liegt der Fokus auf dem Verständnis der resultierenden Formel des GLS-Schätzers: Bei der Betrachtung dieses Ergebnisses ist es nämlich wichtig, die Komponenten zu erkennen, welche einerseits in dem **FEM**-Schätzer \hat{b}_f (blaue Farbe) und andererseits in dem **BEM**-Schätzer \hat{b}_b (rote Farbe) vorkommen. Der Koeffizient aus Gl. 2.64 kann mit der GLS-Methode durch \hat{b}_r wie folgt geschätzt werden:

$$\hat{b}_r = \frac{\frac{1}{T} \sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + G \cdot \sum_{i=1}^n (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\frac{1}{T} \sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2 + G \cdot \sum_{i=1}^n (\bar{x}_i - \bar{x})^2} \quad (2.68)$$

mit:

$$G = \frac{V(\epsilon)}{V(\epsilon) + T \cdot V(a)}$$

Die Größe G stellt in Gl. 2.68 den Faktor dar, mit dem die Komponenten

der **BEM**-Schätzung (rote Farbe) gegenüber den Komponenten der **FEM**-Schätzung (blaue Farbe) gewichtet werden. Somit kann der Schätzer des **REM**-Modells \hat{b}_r als ein gewichteter Durchschnitt aus dem Schätzer des **FEM**- und dem des **BEM**-Modells gesehen werden. Die Gewichtung hängt von G und somit von dem Anteil der Varianz des nicht-zeitkonstanten Residuums $V(\epsilon)$ an der Gesamtvarianz $V(u)$ ab.³⁶

Tendiert G gegen 0, dann liegt dies an der Dominanz von $V(a)$ und folglich an der Dominanz *zeitinvarianter* individueller Abweichungen. Der **REM**-Schätzer \hat{b}_r konvergiert dann gegen den **FEM**-Schätzer \hat{b}_f . Gl. 2.68 reduziert sich dann im Extremfall $G = 0$ auf den „blauen Teil“.

Tendiert G gegen 1, dann dominieren mit $V(\epsilon)$ die Abweichungen, welche über die Zeit *und* die Individuen variieren. Die Anteile des **FEM**-Schätzers und des **BEM**-Schätzers am **REM**-Schätzer sind dann „gleichgewichtig“, so dass man eigentlich nicht weiter zwischen i und t zu differenzieren braucht. Es würde die einfache KQ-Methode des pooled-Modells (s.o.) ausreichen. Somit würde der Schätzer \hat{b}_r gegen den Schätzer des pooled-Modells, \hat{b}_p konvergieren, mit:

$$\hat{b}_p = \frac{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x})(y_{it} - \bar{y})}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x})^2} \quad (2.69)$$

2.5.1.6 Vor- und Nachteile und sinnvolle Anwendungsgebiete der **REM** und **FEM**-Modelle

Vorteile des **REM** sind u.a. darin zu sehen, dass beide Quellen der Variation von y (between- und within-Variation) genutzt werden. Damit verbunden bietet das **REM** eine „mittlere“ Lösung zwischen **FEM** und **BEM** einerseits und **FEM** und dem pooled-Modell andererseits. Wenn $V(a)$ signifikant von Null verschieden ist, dann ist das **REM** auch korrekter spezifiziert als das pooled-Modell nach der KQ-Methode.

In welchem Falle eine Entscheidung zugunsten des **REM** oder **FEM** fallen sollte, ist auch an den Entstehungsmechanismus der Daten geknüpft. Stellen die im Datensatz enthaltenen Objekte die Realisierung einer *Zu-*

³⁶Zur Schätzung der Fehlervarianzen vgl. Hsiao 2005: 38

fallsstichprobe dar, dann unterliegen die durch a_i charakterisierten Ausgangslagen der Objekte selbst einer zufälligen Auswahl. In diesem Falle wäre eigentlich das **REM** vorzuziehen (warum „eigentlich“, wird gleich erläutert). Denn es ist wenig einleuchtend, mithilfe von a_i die *zufällig* aus einer größeren Grundgesamtheit gezogenen Objekte als *fixe* unabhängige Größen zu modellieren.

Entlang dieser Logik eignet sich **FEM** mehr für spezifische Objekte einer Grundgesamtheit, deren Auswahl keinem Zufallsprozess unterliegt. Dies ist vor allem dann gegeben, wenn Vollerhebungen zu kleinen Grundgesamtheiten vorliegen – z.B. wenn über alle Mitglieder des Bundestages Daten zu ihren politischen Aktivitäten vorliegen würden. Da die Struktur einer kleinen, speziellen Grundgesamtheit stärker von den Spezifika ihrer einzelnen Mitglieder abhängen kann, ist es sinnvoller, die Ausgangslagen dieser Mitglieder als fixe unabhängige Variablen zu konzeptualisieren.

Allerdings wird dieser Gedankenstrang verkompliziert durch einen Nachteil des **REM**: Aus der Unterscheidung zwischen fixen und zufälligen Größen ergibt sich nämlich generell eine für Regressionsmodelle logische Annahme: Die Unkorreliertheit zwischen Residuen (Zufallsgrößen) und unabhängigen Variablen (fixe Größen). Wenn nun unter den in a_i zusammengefassten zeitinvarianten Einflussgrößen eine oder mehrere Variablen (z.B. das Geschlecht oder die Intelligenz) besonders in ihrem Einfluss auf y dominieren, dann ist es nicht unplausibel anzunehmen, dass diese Einzelgrößen auch mit den *unabhängigen* Variablen korreliert sind. Die oben erwähnte Annahme wäre dann verletzt.

In diesem Falle kann doch das **FEM** die bessere Alternative sein – auch wenn die Objekte einer Zufallsstichprobe entstammen. Dann steht nämlich a_i nicht mehr für eine zufällig gezogene Person, deren Ausgangslage sich aus vielen kleinen, einzeln genommen unbedeutenden Zufallseinflüssen zusammensetzt. Vielmehr repräsentiert der a_i -Wert dann eine oder wenige bedeutsame Einzelvariablen, die auf diesem Wege zurecht als fixe Größen berücksichtigt werden können.

Wie so oft in der Statistik verbleibt die Entscheidung **FEM** vs. **REM** in der Angemessenheit des Anwenders. Dabei sollten vor allem inhaltlich plausible Kriterien zur Rate gezogen werden.

2.5.1.7 FEM vs. REM – ein Beispiel

Die konkrete Anwendung von FEM und REM soll nun anhand eines einfachen Beispiels vollzogen werden: Es wurde ein sehr kleiner fiktiver Datensatz vom Autor konstruiert. Dieser Datensatz³⁷ weist folgende Eckdaten auf:

- eine abhängige Variable y_{it} : *Körpergewicht einer Person i zum Zeitpunkt t*
- eine unabhängige Variable x_{it} : *Fettgehalt des Essens, welche eine Person i zum Zeitpunkt t täglich durchschnittlich zu sich nimmt*
- Messungen an vier Personen ($n = 4$; alle Personen weisen dasselbe Alter und dieselbe Körpergröße auf, so dass der „*body-mass-index*“ **nur** von dem Körpergewicht abhängt)
- Messungen zu drei Zeitpunkten ($T = 3$)

Es soll hier der allseits bekannten Hypothese nachgegangen werden, dass das Ausmaß des Essens von fetthaltigen Gerichten das Gewicht positiv beeinflusst.

Würde hier die Besonderheit ignoriert, dass ein Paneldatensatz vorliegt, dann ließe sich mit den Daten eine einfache „pooled“-Regression, entsprechend der Schätzung in Gl. 2.69 berechnen.

Schon nach Augenmaß lässt sich ein leichter positiver Zusammenhang zwischen beiden Variablen feststellen. Die geschätzte Gleichung lautet:

$$y_{it} = -19 + 1,635x_{it} + e_{it} \quad (2.70)$$

mit $r^2 = 0,66$

Der Regressionskoeffizient ist positiv und der Determinationskoeffizient ist mit 66% der erklärten Varianz von y relativ hoch. Dies stützt die Hypothese eines positiven Zusammenhangs.

Im nächsten Schritt soll nun das FEM nach Gl. 2.58 berechnet werden. Aus Veranschaulichungsgründen wird mit der Dummy-Variante gearbeitet.

³⁷ abgelegt unter Appendix 4.4, in Tab. 4.1

In diesem Fall werden drei Dummy-Variablen ($n - 1 = 4 - 1 = 3$) eingeführt, so dass Gl. 2.58 ausgedrückt wird als:

$$y_{it} = a + a_1 + a_2 + a_3 + bx_{it} + \epsilon_{it} \quad (2.71)$$

Um die Dummy-Variablen kenntlich zu machen, wird nun statt a_i $a_i d_i$ geschrieben, wobei a_i für den individuellen Regressionskoeffizienten i -ter Einheit und d_i für die entsprechende Dummy-Variable steht:

$$y_{it} = a + a_1 d_1 + a_2 d_2 + a_3 d_3 + bx_{it} + \epsilon_{it} \quad (2.72)$$

Die Datenstruktur inkl. Dummy-Variablen ist in Appendix 4.4 unter Tab. 4.2 abgelegt. Es ist zu sehen, dass auch die letzte Person des Datensatzes, Person D, im Modell aufgenommen wird – und zwar durch die Konstellation „alle Dummy-Variablen nehmen den Wert Null an“. Ihr a_n ergibt sich somit rechnerisch aus der Konstanten a .

Diese Datenstruktur, so wie sie in Tab. 4.2 zu sehen ist, wird nun einer einfachen linearen Regression nach der KQ-Methode unterzogen (z.B. in SPSS). Es ergibt sich folgende Gleichung:

$$y_{it} = 142 - 45,4d_1 - 32,31d_2 - 10,67d_3 - 0,702x_{it} + \epsilon_{it} \quad (2.73)$$

mit $r^2 = 0,9969$

I.d.R. sind die Koeffizienten einzelner Individuen nicht von Interesse. Hingegen sehr von Belang sind der Koeffizient b und r^2 . Es stellt sich die Frage, wie sich diese Koeffizienten unter der Konstanthaltung individueller zeitinvarianter Effekte verändern.

Man sieht, dass das Bestimmtheitsmaß drastisch gestiegen ist (um mehr als 20%). Die Berücksichtigung inter-individueller Ausgangslagen in Form unabhängiger Variablen steigert also die erklärte Varianz der abhängigen Variablen auf rund 99%.

Nun ist zu fragen, wie b im Kontext des Beispiels zu interpretieren ist und warum der Koeffizient im Vergleich zum einfachen Regressionsmodell hier sein Vorzeichen geändert hat.³⁸

³⁸Fragen nach der Signifikanz von Koeffizienten sollen hier ausgeklammert werden, da schließlich ein erfundener und kleiner Datensatz vorliegt

Wie oben mehrfach erwähnt, ist ein Regressionskoeffizient einer unabhängigen Variablen x im FEM zu verstehen als Einfluss von x auf y bereinigt um intra-individuellen Ausgangslagen, die in den Term a_i absorbiert werden. Wie an Gl. 2.60 deutlich wird, werden sowohl im Zähler als auch im Nenner des Schätzers nur Abweichungen der x - und y -Werte von den *individuellen* arithmetischen Mitteln berechnet und aufsummiert. Wäre die Varianz der x -Variablen für alle Personen i konstant, dann würde der FEM-Koeffizient sogar exakt dem Durchschnittswert aller n *individuellen* Regressionskoeffizienten der Regressionen von y auf x entsprechen. Denn man könnte bei konstanter x -Varianz statt $\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2$ schreiben: $n \cdot C$. C ist hierbei der für alle i konstante Zähler der individuellen x -Varianz. Formel 2.60 würde sich vereinfachen zu:

$$\hat{b}_f = \frac{1}{n} \cdot \frac{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{C} \quad (2.74)$$

Die Übereinstimmung in der x -Varianz ist im Beispiel für die ersten beiden Personen A und B gegeben:

Zuerst wird für diese Personen eine Regression mit jeweils drei Wertepaaren $(x_t|y_t)$ errechnet – sozusagen als Analyse des Zusammenhangs der Werte *eines* Individuums, welche *über mehrere* Zeitpunkte streuen. Es ergibt sich für den Regressionskoeffizienten der Variablen x mit y als abhängige Variable

- für Person A: $b = -0,643$,
- für Person B: $b = -0,929$.

Es ist hier also in beiden Fällen ein negativer Zusammenhang zwischen x und y zu sehen (in beiden Fällen ist r^2 relativ hoch).

Wird nun ein FEM *nur für diese beiden Personen* gerechnet, dann ergibt sich in diesem Modell für den Regressionskoeffizienten der Variablen x : $b = -0,786$. Dieser Wert ist eben nichts anderes, als das ***arithmetische Mittel der Regressionskoeffizienten für Person A und B***. An dieser Stelle wird nochmals deutlich, dass sich das FEM **nur** der *within variation* bedient.

Nun soll der Vorzeichenwechsel im FEM gegenüber dem pooled-Modell angesprochen werden. Denn die individuellen Zusammenhänge weisen eine

negative Richtung auf,³⁹ während das Gesamtsystem einen positiven Trend verzeichnet.

Es gilt zuerst tendenziell: „Je fettiger das Essen, umso mehr Gewicht“. Für eine einzelne Person gilt aber entlang der Zeitachse: „Je fettiger das Essen, umso **weniger** (!) Gewicht“. Dies klingt widersprüchlich. Plausibel erklärbar (reine statistische Mathematik wird hier zunächst ausgeklammert) wird dieses Zusammenhangssystem allerdings, wenn man die kausale Richtung auf der Individualebene umkehrt. Gehen wir nun davon aus, dass x die abhängige Variable ist. So lautete die Hypothese auf der Individualebene: „Je mehr Gewicht, desto weniger fettig das Essen“.

Es könnte hier also ein rekursiver Prozess derart stattfinden, dass wenn Personen relativ zu ihrem persönlichen Ausgangsgewicht zunehmen, sie auf die Zunahme mit der Reduktion von fettigem Essen reagieren. Umgekehrt, wenn Personen es schaffen, relativ zu ihrem Ausgangsgewicht ein paar Kilo abzunehmen, dann werden sie nachlässig bei der bewussten Ernährung, und essen wieder fettiger. Beide Fälle sind Formulierungen eines negativen Zusammenhangs zwischen x und y .

Auch wenn diese Erklärung sicherlich nicht ganz der Realität entspricht, sehr vereinfachend ist und, wenn überhaupt, dann nur auf bestimmte Personengruppen zutrifft, so soll sie doch ein allgemeines Phänomen veranschaulichen: Die Zusammenhangsstruktur von x und y kann sich völlig unterscheiden zwischen der Betrachtung entlang der Personen i einerseits und entlang der zeitlichen Entwicklung t andererseits. Es lässt sich festhalten:

Das Gesamtsystem weist einen bestimmten Zusammenhang zwischen zwei Variablen auf. Dieser Zusammenhang besitzt aber auf der Individualebene entlang der Zeitachse, im Schnitt über alle Individuen, eine ganz andere Struktur. Solch eine Unterscheidung von „Zusammenhangsebenen“ ist, wie gezeigt, durch das Vorhandensein von Paneldaten analysierbar.

Im Kontext dieses Beispiels wird nochmals die Reduziertheit des FEM deutlich: Wird die Analyse auf die Variation innerhalb einzelner Individuen reduziert, dann sagt das Modell prinzipiell nichts darüber aus, wodurch die Unterschiede zwischen den Individuen zustande kommen. Warum wiegt z.B. eine Person mehr als eine andere (bei kontrollierter Körpergröße)? Das „fetti-

³⁹ dies lässt sich am Datensatz bereits mit dem Auge erkennen, wenn nur die Wertepaare eines Individuums betrachtet werden

ge Essen“ als erklärende Variable reicht bei Leibe nicht aus, denn ausgehend von dem einfachen Modell in Gl. 2.70 bleibt noch ein großer Prozentsatz unerklärter Varianz über.

Ein sehr hoher Wert von r^2 im FEM darf nicht darüber hinweg täuschen, dass durchaus wichtige Einflussgrößen nicht im Modell enthalten sein könnten. Denn ein solch hoher r^2 -Wert ist schließlich durch die Einführung künstlicher Dummy-Variablen zustande gekommen und die individuellen Effekte sind zwar in die Rechnung zugunsten einer Erhöhung von r^2 eingeflossen, müssen aber noch selbst erklärt werden. Es wäre also zu überlegen, ob die unterschiedlichen individuellen Ausgangslagen a_i von wenigen einflussreichen zeitinvarianten Variablen dominiert werden, oder es sich dabei um eine Gemengelage aus vielen, im einzelnen betrachtet unbedeutenden Einflüssen handelt.

Zur technischen Durchführung ist anschließend noch zu sagen, dass die Berechnung eines FEM mit z.B. STATA einfach realisierbar ist. Dort ist dieses Modell explizit als Analyseoption implementiert, so dass die Bildung von Dummy-Variablen nicht manuell zu tätigen ist (im Gegensatz zu SPSS).

Der zu analysierende Datensatz muss in STATA lediglich in das sog. „lange Format“ gebracht werden und es muss definiert werden, welche Variable die einzelnen Personen und welche die Zeitpunkte definiert (zur Durchführung dieser Schritte vgl. Kohler 2008: 245f).

Ist dies getätigt, dann reicht ein Befehl aus, um das FEM zu berechnen:

xtreg y x, fe

Die Bezeichnungen y und x sind hierbei Platzhalter für beliebige Variablen. Wichtig ist, dass die abhängige Variable in der Reihenfolge zuerst notiert wird. Daraufhin können beliebig viele unabhängige Variablen (aber nicht (!!!) die Dummy-Variablen, nur die „inhaltlichen“ unabhängigen Variablen) folgen.

Ähnlich einfach lässt sich die GLS-Schätzung des REM-Modells in STATA umsetzen. Die etwas abschreckend wirkende Formel 2.65 muss also nicht eigenhändig programmiert werden. Stattdessen wird einfach nur in der oben dargestellten Befehlszeile der Ausdruck „fe“ durch „re“ ersetzt.

Entsprechend der Gl.2.64 lautet auch für das hier behandelte Beispiel die REM-Ausgangsgleichung (mit $u_{it} = a_i + \epsilon_{it}$):

$$y_{it} = bx_{it} + u_{it} \quad (2.75)$$

Die GLS-schätzung nach Gl. 2.65 ergibt einen Koeffizient von $b = -0,578$. Es zeigt sich, dass auch in einem Verfahren, in dem *sowohl* die *within*- als auch die *between-variation* in die Analyse einfließen, der Koeffizient auf einen negativen Zusammenhang zwischen „Gewicht“ und „Fetthaltigem Essen“ hindeutet. Dies ist eine Tendenz, welche erst unter Beachtung der Panelstruktur zum Vorschein gekommen ist, da, wie bereits in Gl. 2.70 gezeigt wurde, eine einfache Regression eher einen positiven Zusammenhang vermuten lässt.

Rechnerisch ergibt sich der negative Koeffizient aus der Tatsache, dass das Gewicht G aus der Schätzformel 2.68 mit 0,00349 sehr klein ausfällt. Folglich wird die Between-Variation von y bzw. die Between-Kovariation von x und y zugunsten der Within-(Ko-)Variation deutlich runtergewichtet. Da der Zusammenhang zwischen x und y aus der „Within-Perspektive“ negativ ist, setzt sich letztlich aufgrund der starken Abwertung der „Between-Perspektive“ das negative Vorzeichen bei der REM-Koeffizientenschätzung durch.

Der kleine Gewichtungswert von G ist wiederum, wie unterhalb der Formel 2.68 zu sehen ist, auf einen in Relation zu $V(\epsilon)$ hohen Schätzwert der „Between-Fehlervarianz“ $V(a)$ zurückzuführen (mehr zu den Schätzformeln s.u.).⁴⁰ In den Daten manifestiert sich dieser Unterschied darin, dass die absoluten Unterschiede *zwischen* den Personen (relativ hohe Between-Varianz) deutlich größer sind, als die Unterschiede innerhalb der Personen (relativ niedrige Within-Varianz). Dies gilt sowohl in Hinblick auf die Varianzen von x und y als auch für die Kovarianz zwischen den beiden Variablen. Da mit einer relativ hohen Between-(Ko-)Varianz auch hohe absolute Residualwerte „drohen“, wird insgesamt die Between-Varianz in der Schätzung mithilfe von G heruntergewichtet.

An dieser Stelle sollen noch einige interessante Größen im STATA-Output der REM- und FEM-Modelle besprochen werden. Es folgt eine Auflistung der Berechnungsformeln der wichtigsten Maßzahlen / Schätzungen.

Zunächst sind unabhängig von der Modellwahl immer drei Arten von Determinationskoeffizienten angegeben. „**R-sq: within**“ bezieht sich auf den Anteil der Within-Varianz von y , welche durch die Within-Varianz von x erklärt wird. Das Pendant dazu ist „**R-sq: between**“, bei der nur die Between(Ko-

⁴⁰ $\hat{V}(\epsilon) = 0,994260^2 = 0,98855$ und $V(a) = 9,7009^2 = 94,107$

)Variation von x und y berücksichtigt wird. „**R-sq: overall**“ ist der Anteil der gesamten Varianz von y , welche durch x erklärt wird. Sie entspricht dem Determinationskoeffizienten in der pooled Regression, in der zwischen der Zeit- und der Personenebene nicht weiter unterschieden wird.

Die dazugehörigen Formeln in der unteren Auflistung zeigen, dass die verschiedenen Determinationskoeffizienten dem selben Prinzip unterliegen: Es wird die jeweilige quadrierte Kovarianz von x und y in Relation gesetzt zu dem Produkt der korrespondierenden Varianzen beider Variablen. Dies entspricht der Quadrierung des Zählers und des Nenners des Korrelationskoeffizienten, aus dem sich ja im einfachen Falle einer unabhängigen Variablen direkt der Determinationskoeffizient errechnen lässt (eben durch die Quadrierung).

Wichtig ist, dass sich „R-sq: overall“ nicht additiv aus den beiden anderen Größen zusammensetzt. In dem Beispiel ist er sogar niedriger, als das Between- bzw. Within-Bestimmtheitsmaß. Dies liegt daran, dass auf der Between-Ebene ein positiver und auf der Within-Ebene ein negativer Zusammenhang zwischen x und y besteht. Auf der undifferenzierten Gesamtebene heben sich diese gegenläufigen Zusammenhänge teilweise auf, so dass „R-sq: overall“ niedriger ausfällt. Auch daran wird deutlich, dass es sinnvoll sein kann, bei Paneldaten die Zeiten- und die Objektebene gesondert zu betrachten.

Für die Koeffizienten lassen sich Standardfehler berechnen und ein Signifikanz-Test durchführen. Die Vorgehensweise und die Interpretation ist völlig deckungsgleich mit der einer gewöhnlichen Regression. Dies gilt auch für die F-Teststatistik (oberer der beiden F-Werte) im **FEM** und die „Wald chi²“-Statistik im **REM**. Beide testen die Nullhypothese, inwieweit *alle* Koeffizienten des Modells aus einer Population kommen, in der alle korrespondierenden Parameter dem Wert 0 entsprechen.

Im **FEM** ist ferner eine zweite F-Teststatistik angegeben: „**F test that all $u_i=0$** “. Mithilfe dieser wird geprüft, ob die individuellen Regressionskonstanten a_i in ihrer Gesamtheit aus einer Population kommen, in welcher all diese Konstanten Null betragen. Würde diese Nullhypothese beibehalten werden, dann wäre die Einführung von a_i in die Regression unnötig. Die individuellen Regressionskonstanten würden keinen signifikanten Erklärungsbeitrag leisten. In dem Beispiel kann allerdings die Nullhypothese verworfen werden. Schließlich gibt es signifikante Gewichtsunterschiede zwischen den

Personen (Between-Variation).

Die Größe „**corr(u_i, Xb)**“ entspricht der Korrelation zwischen den individuellen Regressionskonstanten einerseits und der mithilfe von x geschätzten y -Werte andererseits. Letztere sind, wie in einer gewöhnlichen Regression, gegeben über die lineare Kombination aus unabhängigen Variablen und Regressionskoeffizienten – hier im einfachen Falle: $\hat{y}_{it} = \hat{b}_f \cdot x_{it}$. Die Korrelation besagt, wie stark die gemeinsame Wirkung der x -Variablen mit den individuellen Ausgangslagen korreliert. Bei Vorliegen nur einer Variablen x vereinfacht sich der Sachverhalt: Der Betrag⁴¹ der Korrelation „**corr(u_i, Xb)**“ entspricht dann der bivariaten Korrelation zwischen x und a_i . Im oberen Beispiel gibt sie also an, wie stark das Ausmaß fetthaltigen Essens mit den individuellen „Gewichtsniveaus“ zusammenhängt. Dieser Zusammenhang ist mit -0,8934 recht stark, was auch durch den hohen „R-sq: between“ unterstrichen wird. Schließlich werden für die Berechnung von „R-sq: between“ die intra-individuellen Streuungen zu den Größen \bar{x} und \bar{y} zusammengefasst, welche konzeptionell diese Ausgangslagen zum Ausdruck bringen. Somit beeinflusst das fetthaltige Essen signifikant das Gewichtsniveau, auf dem sich Personen befinden.

Ein hoher „**corr(u_i, Xb)**“-Wert spricht gegen die Anwendung des **REM**. Denn im **REM** werden die individuellen Ausgangslagen zu der als Zufallsvariable aufgefassten Fehlerkomponente gezählt, welche bei gegebenem x -Wert einen Erwartungswert von 0 hat. Der Fehler wird als zufällig um die perfekte Beziehung $y = b_r \cdot x$ streuend angenommen, ist folglich mit x unkorreliert. Es gilt a priori „**corr(u_i, Xb)**=0“ Diese Annahme ist allerdings anzuzweifeln, wenn sich im **FEM**, also bei der Behandlung von a_i als fixe Größe, doch eine relativ hohe Korrelation „**corr(u_i, Xb)**“ zeigt. Für eine Diskussion dieser Problematik wird auf Hsiao (2005) verwiesen.

Die Angaben zu „**sigma_u**“, „**sigma_e**“ und „**rho**“ ($= \rho$) werden zunächst für den **REM**-Fall erläutert:

Im **REM** stehen sie für die Schätzungen der Standardabweichungen der beiden Fehlerkomponenten. „**sigma_u**“ ist die Schätzung für die Wurzel aus $V(a)$; „**sigma_e**“ entspricht der Schätzung für die Wurzel aus $V(\epsilon)$. Um deren Bedeutung zu verstehen, wird der Sachverhalt zuerst auf eine einfache Regression heruntergebrochen: In der normalen linearen Regression nach der

⁴¹Die Multiplikation mit \hat{b}_f kann lediglich das Vorzeichen der Korrelation von x und a_i ändern, aber nicht die Stärke

KQ-Methode existiert bekanntlich nur ein Fehlerterm. Für diesen Fehlerterm wird eine über alle Beobachtungen – gegeben der x -Werte – konstante Standardabweichung angenommen. Diese Größe wird als σ bezeichnet. Es lässt sich zeigen, dass auch y als abhängige (Zufalls-)Variable⁴² dieselbe, über alle Beobachtungen (wieder gegeben x) konstante Standardabweichung σ inne hat (vgl. Auer 2007: 81).

Um die konkrete Bedeutung von σ nachvollziehbar zu machen, soll bspw. eine einfache Regression des Körpergewichtes y auf die Körpergröße x mit der Gleichung $y_i = a + bx_i + \epsilon$ betrachtet werden. Nun könnte man sich die Verteilung des Gewichtes y gedanklich vorstellen, wenn man *nur aus einer Subpopulation von Personen mit einer festgelegten Körpergröße*, z.B. $x = 165\text{cm}$, unendlich oft Personen per Zufall auswählt und deren Gewicht bestimmt. σ entspricht dann der Streuung des Gewichtes eben in dieser Subpopulation $x = 165$ (Erwartungswert ist $a + b \cdot 165$). Genau das ist gemeint, wenn von „gegeben x “ die Rede ist. Ferner wird σ unter der Annahme von Homoskedastizität als gleich für *alle* Werte deklariert, die x annehmen kann. Bspw. beträgt die Streuung von y gegeben $x = 203,32\text{cm}$ ebenfalls σ .

σ^2 ist entsprechend die Varianz, und zwar sowohl des Residuums als auch der Variablen y . Im REM entspricht σ^2 der Größe $V(u_{it})$ aus Gl. 2.66. Dieselbe Gleichung zeigt auch die Aufteilung von $V(u_{it})$ in die Varianzen der beiden Fehlerkomponenten. Folglich entspricht die auf x konditionierte Varianz $V(y_{it})$ der Zufallsvariablen y :

$$V(y_{it}) = V(u_{it}) = V(a_i) + V(\epsilon_{it}) \quad (2.76)$$

Diese Varianzen sind in der Regel unbekannt, können aber aus den Daten geschätzt werden (s. untere Auflistung). Die von STATA errechneten Schätzungen erhält man über die Quadrierung der Standardabweichungen der Fehler „sigma_u“ bzw. „sigma_e“. Die Größe ρ setzt nun die Varianz der individuellen, über die Zeit konstanten Fehler $V(a_i)$ in Relation zur Varianz des gesamten Fehlers $V(u_{it})$. Da es sich bei den STATA-Angaben um Schätzungen handelt, ist auch die Angabe „rho“ selbst eine Schätzung mit der Formel:

$$\hat{\rho} = \frac{\hat{V}(a_i)}{\hat{V}(a_i) + \hat{V}(\epsilon)} = \frac{\hat{V}(a_i)}{\hat{V}(u_{it})} = \frac{\hat{V}(a_i)}{\hat{V}(y_{it})} \quad (2.77)$$

⁴² y ist konzeptionell ebenfalls eine Zufallsvariable, da sie sich aus der linearen Transformation des zufälligen Residuums ergibt; s. 2.5.1.3

Der letzte Term zeigt, dass es sich bei ρ um den geschätzten Anteil der Varianz von y (!!!) handelt, welcher mit $\widehat{V}(a_i)$, also der Between-Variation von y erklärt werden kann. Auch diese Aussage gilt wieder unter Kontrolle von x .

In dem Beispiel ist $\widehat{\rho}$ mit 0,9896 sehr hoch, was besagt, dass fast die vollständige Streuung der y -Werte auf die Unterschiede *zwischen* den Individuen (Between-Variation) zurückgeht. Auch hier zeigt sich also, wie oben im Zusammenhang mit der Größe G erläutert, dass die absoluten Unterschiede *zwischen* den Personen (relativ hohe Between-Varianz) deutlich größer sind, als die Unterschiede innerhalb der Personen (relativ niedrige Within-Varianz).

Rabe-Heskath et al. (2008: 58) zeigen ferner, dass ρ gleichzeitig auch die Korrelation der y -Werte (gegeben x) einer Person zu zwei verschiedenen Zeitpunkten darstellt. Der hier hoch ausfallende Wert besagt also, dass es sehr wahrscheinlich ist, dass das Gewicht einer Person i bei einer zweiten Messung ähnlich (in Relation zu den Unterschieden *zwischen* Personen) dem Gewicht derselben Person i zum ersten Messzeitpunkt sein wird.

Im **FEM** ist die Interpretation von ρ im Prinzip äquivalent zu der im **REM**. Allerdings muss hier beachtet werden, dass a_i nicht zu den Zufallsvariablen gehört, sondern eine fixe Größe darstellt. Daher wird $V(a_i)$ nicht aus den Daten geschätzt, sondern anhand der Daten eindeutig errechnet.

In den folgenden Tabellen sind die Formeln einiger hier besprochener Größen aufgeführt. Zwei Einschränkungen sind zu machen: Sie gelten 1. nur für den einfachen Fall mit lediglich einer unabhängigen Variablen x . 2. sind sie nur korrekt beim Vorliegen eines balancierten Paneldatensatzes – sprich, alle Personen haben an allen Wellen teilgenommen. Ist mindestens eine dieser zwei Bedingungen nicht gegeben, dann verkomplizieren sich zwar die Formeln, das den Größen unterlegte Grundprinzip bleibt allerdings gleich.

Die im Kap. 2.5 eingeführte Größen und Symbole werden als bekannt vorausgesetzt.

Determinationskoeffizienten und Korrelationen		
STATA-Bez.	Formel	Erläuterung
R-sq: within	$\frac{[\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)]^2}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i)^2 \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y}_i)^2}$	Anteil der Within-Varianz von y , die durch x erklärt wird
R-sq: between	$\frac{[\sum_{i=1}^n (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})]^2}{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2 \sum_{i=1}^n (\bar{y}_i - \bar{y})^2}$	Anteil der Between-Varianz von y , die durch x erklärt wird
R-sq: overall	$\frac{[\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x})(y_{it} - \bar{y})]^2}{\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x})^2 \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y})^2}$	Anteil der gesamten Varianz von y , die durch x erklärt wird
corr(u_i, Xb)	$\frac{\sum_{i=1}^n \sum_{t=1}^T (a_i - \bar{a})(\tilde{y}_{it} - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^n \sum_{t=1}^T (\tilde{y}_{it} - \bar{\tilde{y}})^2}}$	Korrelation zwischen den individuellen Regressionskonstanten und den geschätzten y -Werten, gegeben x . Die Formel bezieht sich nur auf das FEM ! Im REM wird der Korrelationswert 0 per Annahme festgelegt.

Fehlervarianzen im FEM		
STATA-Bez.	Formel	Erläuterung
sigma_e	$\sqrt{\widehat{V}(\epsilon)} = \sqrt{\frac{\sum_{i=1}^n \sum_{t=1}^T \epsilon^2}{n(T-1)-1}}$	Schätzformel für die Standardabweichung des Residuums (analog zum σ in der einfachen Regression)
sigma_u	$\sqrt{V(a_i)} = \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n-1}}$	Standardabweichung der individuellen Regressionskonstanten. Da diese als fixe Werte betrachtet werden, ist dies keine Schätzformel, sondern eine eindeutige Berechnung!
rho	$\widehat{\rho} = \frac{V(a_i)}{V(a_i) + \widehat{V}(\epsilon)}$	Geschätzter Anteil der „Between“-Varianz von y an der gesamten Varianz von y (gegeben x) / Geschätzte Korrelation zwischen zwei y -Werten ein und der selben Person (gegeben x)

Fehlervarianzen im REM		
STATA-Bez.	Formel	Erläuterung
sigma_e	$\sqrt{\widehat{V}(\epsilon)} = \sqrt{\frac{\sum_{i=1}^n \sum_{t=1}^T \epsilon^2}{n(T-1)-1}}$	Schätzformel für die Standardabweichung des Residuums (analog zum σ in der einfachen Regression)
sigma_u	$\sqrt{\widehat{V}(a)} = \sqrt{\frac{1}{n-2} \left[\frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \sum_{i=1}^n (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2} \right] - \frac{\widehat{V}(\epsilon)}{T}}$	Schätzformel für die Standardabweichung der über Individuen konstanten Fehlerkomponente a_i
rho	$\widehat{\rho} = \frac{\widehat{V}(a)}{\widehat{V}(a) + \widehat{V}(\epsilon)}$	Geschätzter Anteil der „Between“-Varianz von y an der gesamten Varianz von y (gegeben x) / Geschätzte Korrelation zwischen zwei y -Werten ein und der selben Person (gegeben x)

mit:

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$$

$$\tilde{y}_{it} = x_{it} \cdot \hat{b}_f$$

$$\bar{\tilde{y}} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \tilde{y}_{it}$$

2.5.2 Regressionsmodelle mit Differenzvariablen

Im Kontext von Modellen mit variablen Regressionskonstanten wurde das Problem von *Fehlspezifikationen in Regressionsmodellen* thematisiert. Dies ist ein sehr bedeutsames Thema. Reelle komplexe Vorgänge zu modellieren, soz. in eine konsistente mathematische Struktur zu übersetzen, ist nämlich ein schwieriges Unterfangen. In vielen Regressionsmodellen kann die Varianz einer abhängigen Variablen (und somit ein komplexer Sachverhalt) nur unzureichend erklärt werden. Es sind also oft zusätzlich andere, unberücksichtigte Größen (sog. *omitted variables*), welche nicht in das Modell aufgenommen wurden, am Werke. Die Nicht-Berücksichtigung dieser Variablen ist eine zentrale Quelle für solche Modell-Fehlspezifikationen.

Dies wäre weniger tragisch, wenn dadurch nicht ein großes Problem aufträte: Für den Fall, dass solche unberücksichtigten Größen mit modellierten unabhängigen Variablen korrelieren, werden die Schätzer der Regressionsparameter inkonsistent (vgl. Arminger 1990: 2ff).

Es wurde in Kap. 2.5.1 gezeigt, dass mithilfe einer variablen Regressionskonstanten a_i eine bestimmte Klasse solcher unberücksichtigter Größen gebündelt, und je nach Modellansatz, herausgerechnet (FEM) oder als spezielle Fehlerkomponente explizit modelliert (REM) werden kann: Es handelt sich um die Klasse *zeitinvarianter* Merkmale. Eine alternative Methode der „unsichtbaren“ Berücksichtigung zeitinvarianter nicht erhobener Variablen ist die Bildung von Differenzenmodellen.

Um diesen Ansatz zu veranschaulichen, soll an dieser Stelle noch mal die bei vorliegenden Paneldaten möglichen Differenzierungen hinsichtlich der Variation von unabhängigen Variablen und ihren Regressionskoeffizienten (bzw. Effekten) wiederholt werden:

- $\mathbf{b}_t \mathbf{x}_{it}$ – Unabhängige Variablen, deren Werte **und** Effekte auf die abhängige Variable *mit den Zeitpunkten variieren* [**Fall A**]
- $\mathbf{b} \mathbf{x}_i$ – Unabhängige Variablen, deren Werte **und** Effekte auf die abhängige Variable *über die Zeit konstant sind* [**Fall B**]
- $\mathbf{b}_t \mathbf{x}_i$ – Unabhängige Variablen, deren Werte über die Zeit konstant sind, aber deren Effekte auf die abhängige Variable mit den Zeitpunkten variieren [**Fall C**]
- $\mathbf{b} \mathbf{x}_{it}$ – Unabhängige Variablen, deren Effekte über die Zeit konstant sind, aber deren Werte mit den Zeitpunkten variieren [**Fall D**]

Die Zuordnung einer Variablen zu einem dieser Fälle geschieht nach theoretischer Erwägung. Zwar kann man aus der Betrachtung von Datensätzen Indizien für die Zuordnung bestimmter Variablen gewinnen, aber alleine aufgrund dessen keine klare Entscheidung treffen. Schließlich haben wir es bei empirischen Datensätzen mit Variablen zu tun, welche behaftet sind mit stichprobentheoretischen, messtheoretischen und sonstigen Fehlern.

Zentral für die folgenden Ausführungen ist die *Variablen-Effekte-Konstellation des Falls B*. Diese Konstellation könnte bspw. für Variablen zutreffen, die bereits im Zuge der FEM-Modelle diskutiert wurden – wie genetische Dispositionen, die Schichtzugehörigkeit, das Geschlecht oder die Intelligenz. Manche dieser Variablen können zwar intraindividuellen Veränderungen unterliegen, solche Veränderungen sind aber für eher kurze Zeiträume sehr unwahrscheinlich (z.B. bei der Schichtzugehörigkeit).

Das Problem ist, dass solche Variablen zwar einen Effekt auf andere Variablen haben können, aber nicht immer erhebbar sind. Während z.B. das Geschlecht bei den meisten Untersuchungen problemlos erhoben wird, gestaltet sich dies bei der Intelligenz schwieriger.

Wirkt eine solche Variable auf eine abhängige Variable in einem Regressionsmodell und wird sie nicht modelliert, so können die Koeffizienten weiterer unabhängiger Variablen u.U. nur inkonsistent geschätzt werden.

Als einfaches Beispiel sei folgende Regressionsgleichung eingeführt (zur Vereinfachung wird der Objekteindex weggelassen und die Regressionskonstante auf Null gesetzt):

$$y_t = b_t x_t + cz + e_t \quad (2.78)$$

mit

y = die abhängige Variable des Modells,

x = eine unabhängige Variable des Falls A,

b = der zu x zugehörige Regressionskoeffizient,

z = eine unabhängige Variable des Falls B (z.B. Intelligenz),

c = der zu z zugehörige Regressionskoeffizient,

e = das Residuum,

t = der Index für Zeitpunkte.

Im Appendix 4.3 unter Abb. 4.2 ist ein fiktiver Datensatz abgelegt mit den Variablen x und y zum Zeitpunkt $t = 1$ und $t = 2$ und der nicht über die Zeit variierenden Variablen z .

Betrachtet man im Folgenden die Variablen eines Querschnitts (hier $t = 1$) und berechnet die lineare Regression der abhängigen Variablen y_1 nach der KQ-Schätzung unter Einbeziehung von x_1 und z (alle Variablen sind z -transformiert), so gelangt man zu der Lösung:

$$y_1 = 0,423x_1 + 0,437z + e \quad (2.79)$$

mit $r^2 = 0,625$.

In dem Falle, dass die Variable z nicht zur Verfügung stünde, würde sich das Modell lediglich auf x_1 als abhängige Variable beschränken. Es ergäbe sich:

$$y_1 = 0,724x_1 + e \quad (2.80)$$

mit $r^2 = 0,525$.

Hier ist zu sehen, dass die Variable z einen signifikanten Erklärungsbeitrag leistet (+10% der erklärten Varianz von y_1).

Vor allem sieht man aber, wie sich b_1 , also der zu x_1 gehörende Regressionskoeffizient, ändert, wenn z einbezogen wird. Die beiden Variablen sind

schließlich nicht unkorreliert ($r_{x_1z} = 0,69$), so dass z auch einen Einfluss auf die Schätzung von b_1 ausübt. Folglich muss b_1 aus Gl. 2.80 als konsistenter Schätzer für den wahren Regressionskoeffizienten β_1 in Frage gestellt werden.
43

Es wurde hier also gezeigt, dass das Modell unter Einbeziehung von z „vollständiger“ ist, als das Modell in Gl. 2.80, somit die Konsistenz der Schätzung von b_1 verbessert wird. Hat man allerdings nur Querschnittsdaten zur Verfügung und die Variable z nicht erhoben, dann besteht keine Möglichkeit, b_1 konsistenter zu schätzen, als in Gl. 2.80.

Dies ändert sich, wenn die Variablen im Paneldesign zu einem zweiten Zeitpunkt erhoben worden sind.

Nun hat man nämlich zwei Querschnittsgleichungen und somit ein Gleichungssystem mit folgenden Gleichungen zur Verfügung (Schreibweise orientiert an Gl. 2.78):

$$y_1 = b_1x_1 + cz + e_1 \quad (2.81)$$

$$y_2 = b_2x_2 + cz + e_2 \quad (2.82)$$

Über Subtraktion der beiden Gleichungen wird der Effekt von z eliminiert:

$$y_2 - y_1 = b_2x_2 - b_1x_1 + e_{2-1} \quad (2.83)$$

bzw.

$$\boxed{\Delta \mathbf{y} = \mathbf{b}_2\mathbf{x}_2 - \mathbf{b}_1\mathbf{x}_1 + \mathbf{e}_{2-1}} \quad (2.84)$$

mit $e_{2-1} = e_2 - e_1$

Auf das Beispiel von oben angewendet, gesellt sich zu der Querschnittsregression aus Gl. 2.79 zu $t = 1$ eine zweite Gleichung zu $t = 2$:

$$y_2 = 0,474x_2 + 0,435z + e \quad (2.85)$$

⁴³ der Schätzer aus Gl. 2.79 allerdings auch, da die Variation von y_1 mit $r^2 = 0,625$ noch lange nicht vollständig erklärt ist - nur an irgendeiner Stelle sind die Grenzen des Machbaren erreicht

mit $r^2 = 0,694$.

Bildet man hier die Differenz, entsprechend Gl. 2.84, erhält man:

$$\Delta y = 0,474x_2 - 0,423x_1 + e_{2-1} \quad (2.86)$$

Es sei also festzuhalten, dass man mithilfe der Verschmelzung von Regressionsgleichungen zu zwei Zeitpunkten, über die Bildung von Δy , die Querschnitts-Parameter b_1 und b_2 besser schätzen kann, da zumindest die Effekte von nicht-modellierten Variablen des Falls B ausgeschaltet werden können – wohlgermerkt **nur** für Variablen des Falls B!

Weiterhin sei gesagt, dass die Bildung von Differenzgleichungen auf mehrere Variablen verschiedener Art (also verschiedener Fälle) anwendbar ist. Wäre die Variable x in Gl. 2.84 eine des Falls D, dann würde sich eine noch einfachere Differenzgleichung ergeben:

$$\boxed{\Delta \mathbf{y} = \mathbf{b}(\mathbf{x}_2 - \mathbf{x}_1) + \mathbf{e}_{2-1}} \quad (2.87)$$

Hätten wir hingegen in einem Regressionsmodell die Variable jedes Falls einmal vertreten, würde also die Gleichung für einen Querschnitt so aussehen:

$$y_t = b_t x_t + cz + d_t v + f w_t + e_t \quad (2.88)$$

mit

v = Variable des Falls C,

d = Regressionskoeffizient der Variablen v ,

w = Variable des Falls D,

f = Regressionskoeffizient der Variablen w ,

und alle anderen Größen entsprechend Gl. 2.78

– dann würde sich die Gleichung, gebildet aus den Differenzen der Querschnitte $t = 1$ und $t = 2$, wie folgt verkomplizieren:

$$\Delta y = \mathbf{b}_2 \mathbf{x}_2 - \mathbf{b}_1 \mathbf{x}_1 + (\mathbf{d}_2 - \mathbf{d}_1) \mathbf{v} + \mathbf{f}(\mathbf{w}_2 - \mathbf{w}_1) + \mathbf{e}_{2-1} \quad (2.89)$$

Auch an dieser Stelle ist zu sehen, dass bei der Bildung von Differenzgleichungen nur die Variable z entfällt, bzw. *entfallen kann*. Somit müssen weitere Variablen, wie x , v und w bekannt sein. Sind sie es nicht, dann kann das Problem von inkonsistenten Schätzungen auch durch Paneldaten nicht behoben werden.

Außerdem sei kurz erwähnt, dass das Arbeiten mit Differenzgleichungen aus statistisch-mathematischer Sicht dann problematisch ist, wenn diese Differenzen sehr klein sind, also nur geringfügige Veränderungen in den Variablenwerten über die Zeit stattgefunden haben.⁴⁴

Die Schätzung der in den hier vorgestellten Modellen enthaltenen Koeffizienten soll nun exemplarisch für das Modell aus Gl. 2.84 angesprochen werden. Möglich sind zwei Varianten: Man berechnet die Differenzvariable Δy und bildet anstelle der Variablen x_1 die Gegenwert-Variable ($-x_1$), indem man alle Werte von x mit (-1) multipliziert. So lässt sich Gl. 2.84 ebenfalls schreiben als:

$$\Delta y = b_2 x_2 + b_1 (-x_1) + e_{2-1} \quad (2.90)$$

Durch die Bildung von ($-x_1$) ist die additive Verknüpfung der Regressionskoeffizienten hergestellt und nun können diese Koeffizienten konventionell, z.B. in STATA nach der KQ-Methode geschätzt werden.

Eine andere Möglichkeit besteht darin, mit Gl. 2.83 zu arbeiten. Nach Addition beider Seiten der Gleichung mit y_1 und unter Gebrauch der oben eingeführten ($-x_1$)-Variablen, lässt sich diese Gleichung wie folgt schreiben:

$$y_2 = y_1 + b_2 x_2 + b_1 (-x_1) + e_{2-1} \quad (2.91)$$

Das Modell in dieser Form lässt sich z.B. im Programm LISREL implementieren. Hierbei werden die Indikatoren x und y als identisch mit den latenten Variablen ξ und η definiert (da in diesen Modell keine latenten Variablen auftauchen, LISREL aber die Spezifikation von ξ und η verlangt). Des Wei-

⁴⁴Zur Kritik und zu Problemen bei der Bildung von Differenzvariablen s. Engel (1994: 19); Arminger (1990: 70f)

teren wird die Variable y_1 zu den unabhängigen Variablen gezählt, weil sie auf der rechten Seite der Gl. 2.91 steht. Da sie aber keine „echte“ unabhängige Variable darstellt, wird der zu ihr zugehörige Regressionskoeffizient nicht zur Schätzung freigesetzt, sondern auf den Wert Eins fixiert. In Matrixschreibweise und entsprechend der LISREL-Symbolik sieht die Gl. 2.91 nun wie folgt aus:

$$(\eta_1) = (1 \quad b_2 \quad b_1) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + (\zeta) \quad (2.92)$$

mit $y_2 = \eta_1$,
 $y_1 = \xi_1$,
 $x_2 = \xi_2$,
 $(-x_1) = \xi_3$ und
 $e_{2-1} = \zeta$

Solch eine Vorgehensweise ist zunächst einmal gewöhnungsbedürftig, lässt aber viel Raum für Variation. So können in LISREL die unterschiedlichsten linearen Modelle implementiert werden: Parameter können z.B. auf gewünschte Werte fixiert, mit anderen Parametern gleichgesetzt oder zur Schätzung völlig freigesetzt werden – und solche „Optionsvielfalt“ erlaubt eben auch die Implementierung der hier besprochenen Modelle.

An dieser Stelle soll der Gebrauch von LISREL nicht weiter vertieft werden. Dies geschieht in Kap. 2.7. Des Weiteren kann hier auf zahlreiche Statistik-Literatur verwiesen werden – so z.B. auf das Buch von Arminger (1990: 72ff), in welchem ein ähnliches Beispiel, wie das obige, und weitere Beispiele zur Panelanalyse ausführlich, inkl. LISREL-Anwendung, erläutert werden.

Es sei noch etwas zu dem Beispiel-Datensatz gesagt: Die Konstruktion einer Variablen z , welche die Eigenschaften vereinigt, sowohl auf y_2 als auch auf y_1 den gleichen Effekt auszuüben, ist nicht einfach – vor allem dann nicht, wenn sich die beiden Variablen x_2 und x_1 in ihren Werten unterscheiden und die Einflüsse imperfekt sind. Dieses Ziel wurde daher nur annähernd erreicht (dies sieht man daran, dass der Koeffizient c in Gl. 2.79 nicht mit dem in Gl. 2.85 identisch ist). Deshalb erbringt eine Schätzung der Koeffizienten allein anhand von Gl. 2.84 nicht exakt dieselben Werte, wie sie unter Gl. 2.86 aus der Differenz der Querschnittsgleichungen errechnet worden sind.

Für den Datensatz sprechen allerdings die „realistischen Bedingungen“. Gemeint ist, dass hier Daten konstruiert worden sind, die ein Regressionsmodell erzeugen, in dem die Varianz der abhängigen Variablen nicht vollständig erklärt wird und in dem die Werte der Regressionskoeffizienten nicht Null oder Eins entsprechen.

Für den Leser, der den Sachverhalt mit Daten nachrechnen will, welche zwar unrealistisch sind, dafür den Vorteil aufweisen, dass die Rechnung „voll aufgeht“, ist ein weiterer Datensatz unter *Appendix 4.3* (in Abb. 4.3) abgelegt. Das Differenzenmodell auf Basis dieses Datensatzes wird in Kap. 2.7 mit LISREL berechnet.

Zum Schluss soll noch kurz erläutert werden, warum das *Differenzenmodell* als weitere Alternative zu den **FEM**- und **REM**-Modellen vorgestellt wurde. Im Abschnitt 2.5.1.6 wurde nämlich deutlich, dass sowohl **FEM**- als auch **REM**-Modelle mit Vor- und Nachteilen behaftet sind. Die Anwendung des **FEM** ist nicht immer angemessen. Das **REM** besitzt hingegen einen Nachteil, welcher in Differenzenmodellen nicht gegeben ist: Es muss in Differenzenmodellen nicht explizit die Annahme der Unkorreliertheit zwischen z und den unabhängigen Variablen des Modells gesetzt werden. z gehört nämlich, obwohl im Modell nicht explizit benannt, wie im **FEM** zu den fixen unabhängigen Variablen.

Allerdings stellt sich die Frage, ob z ein bestimmtes Merkmal zum Ausdruck bringt, oder lediglich der Platzhalter für einen Bündel von einzeln betrachtet insignifikant wirkenden Merkmalen ist, in der gleichen Schärfe wie bei der Frage nach der Angemessenheit der Anwendung eines **FEM**-Modells.

Ein weiterer Nachteil des Differenzenmodells manifestiert sich darin, dass das Rechnen mit Differenzenvariablen, also mit Variablen, welche eine *Änderung zwischen zwei Zeitpunkten* zum Ausdruck bringen, nicht unproblematisch ist. Wenn nur wenig Veränderung zu verzeichnen ist, die Differenzenvariablen also Werte nahe Null annehmen, dann sinkt die Reliabilität der Differenzenvariablen (vgl. Engel 1994: 19). Bezogen auf die Parameterschätzungen „kann der Effizienzverlust erheblich sein“ (Arminger 1990: 75).

Es muss im Einzelfall bei der Entscheidung für oder gegen eines der beiden Verfahren abgewogen werden, welche der Nachteile am ehesten zu akzeptieren sind.

2.5.3 Modelle mit endogener Dynamik

Bislang wurden, bis auf die einführenden Beispiele in Kap. 2.1, Modelle betrachtet, welche keine endogene Dynamik aufweisen. *Endogen* meint, dass eine Variable zu einem Zeitpunkt t_1 einen *Einfluss auf sich selbst* zum Zeitpunkt t_2 ausübt.

Wird dieser Einfluss in einem Regressionsmodell berücksichtigt, so spricht man von einem *autoregressiven Prozess*.

Wird des Weiteren lediglich die endogene Dynamik einer Variablen zwischen *benachbarten Zeitpunkten* vermutet, so liegt ein autoregressiver Prozess **erster Ordnung** vor (bei z.B. drei Wellen und der Variablen y_t mit $t = 1, 2$ und 3 werden also die Einflüsse $y_1 \rightarrow y_2$ und $y_2 \rightarrow y_3$, aber nicht $y_1 \rightarrow y_3$ angenommen). Dieser Fall lässt sich allerdings erweitern zu einem autoregressiven Prozess **n -ter Ordnung**.

Es seien nun y_1 und y_2 betrachtet. y_2 wird hierbei als eine abhängige Variable eines beliebigen Regressionsmodells verstanden:

Im Prinzip kann sich in solch einem Regressionsmodell y_1 ganz normal zu den unabhängigen Variablen gesellen.

Nehmen wir also an, wir hätten eine unabhängige Variable x_2 , dann könnte ein einfaches dynamisches Regressionsmodell wie folgt aussehen (mit c und b als Regressionskoeffizienten der unabhängigen Variablen y_1 respektive x_2 ; $e =$ Fehlerterm; Darstellung wieder unter Verzicht der Konstanten und des Personenindizes):

$$\boxed{y_2 = cy_1 + bx_2 + e} \quad (2.93)$$

Es steht der einfachen Schätzung eines solchen Modells (und einer Erweiterung um weitere Regressoren) nichts im Wege.

Allerdings gibt es *Besonderheiten dynamischer Modelle*, welche an dieser Stelle diskutiert werden sollten:

Zuerst stellt sich die Frage nach *erhöhter Multikollinearität*. Multikollinearität liegt dann vor, wenn unabhängige Variablen eines linearen Regressionsmodells untereinander zusammenhängen – sowohl in Form bivariater als auch

partieller Korrelationen. Perfekte Multikollinearität würde die Berechnung der Regressionskoeffizienten unmöglich machen, sie liegt bei empirischen Daten aber eher selten vor.⁴⁵

Allerdings lässt sich Multikollinearität in abgestufter Form denken, so dass man von einer *relativ hohen Multikollinearität* sprechen kann. Diese verhindert zwar nicht die Berechnung einer Regression, kann aber die Effizienz von Schätzern gefährden und folglich den Standardfehler dieser Schätzer erhöhen.

Weniger mathematisch ausgedrückt lässt sich sagen, dass es bei sehr hoher Multikollinearität schwer ist, die reinen Einflüsse einzelner unabhängiger Variablen auf die abhängige Variable trennscharf zu bestimmen.

Wenn man nun ein Regressionsmodell bestimmt, dann geht man i.d.R. von der Annahme aus, dass die unabhängigen Variablen einen Einfluss auf die zu erklärende Variable ausüben. Werden zwei einfache Querschnittsregressionen zu zwei Zeitpunkten $t = 1$ und $t = 2$ betrachtet, dann ergeben sich folgende Gleichungen:

$$y_1 = a_1 + bx_1 + e_1 \quad (2.94)$$

$$y_2 = a_2 + bx_2 + e_2. \quad (2.95)$$

Wir gehen davon aus, dass in Gl. 2.95 die Variable x_2 mit y_2 korreliert ist. Wird nun ein dynamisches Modell, wie in Gl. 2.93, gebildet, dann kommt die Annahme hinzu, dass y_1 mit y_2 korreliert ist – und zwar auf eine solche Art und Weise, dass damit auch eine signifikante Korrelation von x_2 mit y_1 sehr wahrscheinlich wird.

Die Wahrscheinlichkeit für hohe Multikollinearität wird somit in einem dynamischen Modell, wie in Gl. 2.93, gesteigert (weil hier x_2 und y_1 zu den unabhängigen Variablen gehören).

Ein gewisser Grad an Multikollinearität ist allerdings bei empirischen Variablen normal⁴⁶ und hat keine schwerwiegenden Folgen. Es existiert auch kein Grenzwert, ab dem Multikollinearität in ihrer Höhe inakzeptabel wird. Nichtsdestotrotz sollte beim Arbeiten mit dynamischen Modellen stärker auf dieses Phänomen geachtet werden als sonst.

Ein weiteres Problem dynamischer Modelle knüpft an das der erhöhten

⁴⁵Sie würde z.B. auftauchen, wenn in einer Regression mit Dummy-Variablen die Referenzkategorie auch in Form einer Variablen in die Berechnung einfließen würde; denn die Referenzkategorie ist von den restlichen Dummies exakt linear abhängig

⁴⁶Völlige Abwesenheit von Multikollinearität würde man mit Hilfe von orthogonalen Faktoren im Zuge der Berechnung einer explorativen Faktorenanalyse erzeugen

Multikollinearität an. Es betrifft das REM. Wie bereits in Kap. 2.5.1.6 diskutiert wurde, ist ein REM nur unter der Annahme berechenbar, dass die Fehlerkomponente a_i mit unabhängigen Variablen aus dem REM unkorreliert ist.

Dies kann in einem dynamischen Modell nicht mehr behauptet werden. Schließlich verbergen sich hinter a_i unabhängige Variablen, welche eine zeitkonstante Wirkung auf die abhängige Variable haben, aber nicht erhoben wurden.

Wird die Fehlervarianzzerlegung aus in die zwei hier eingeführten Regressionsgleichungen 2.94 und 2.95 implementiert, dann ergibt sich (hier muss zur Unterscheidung zwischen a und a_i wieder mit dem Personenindex i gearbeitet werden):

$$y_{i1} = a_1 + bx_{i1} + a_i + \epsilon_{i1} \quad (2.96)$$

$$y_{i2} = a_2 + bx_{i2} + a_i + \epsilon_{i2} \quad (2.97)$$

mit

y_{i1} und y_{i2} = abhängige Variablen zum Zeitpunkt $t = 1$ bzw. $t = 2$

a_1 und a_2 = über i konstante Regressionskonstanten zum Zeitpunkt $t = 1$ bzw. $t = 2$

b ist der über i und t konstante Regressionskoeffizient der unabhängigen Variablen x_{i1} und x_{i2}

ϵ_{i1} und ϵ_{i2} = die auch über t variierenden Fehlerkomponenten, welche zusammen mit a_i den Gesamtfehler e_{i1} bzw. e_{i2} aus Gl. 2.94 und 2.95 ergeben (nur dort wurde auf die Notation des i -Indizes verzichtet)

Wird ferner in dem dynamischen Modell aus Gl. 2.93 y_1 durch den Ausdruck auf der rechten Seite von Gl. 2.96 ersetzt und die Fehleraufteilung vollzogen, ergibt sich:

$$y_{i2} = c(a_1 + bx_{i1} + a_i + \epsilon_{i1}) + bx_{i2} + a_i + \epsilon_{i2} \quad (2.98)$$

bzw.:

$$y_{i2} = c \cdot a_1 + c \cdot bx_{i1} + c \cdot a_i + c \cdot \epsilon_{i1} + bx_{i2} + a_i + \epsilon_{i2} \quad (2.99)$$

Die Variable a_i taucht hier sowohl als unabhängige Variable in der Form $c \cdot a_i$ als auch als Fehlerkomponente $e_{i2} = a_i + \epsilon_{i2}$ auf. Logischerweise ist a_i aus $c \cdot a_i$ einerseits und aus $e_{i2} = a_i + \epsilon_{i2}$ andererseits mit sich selbst korreliert, so dass die Prämisse verletzt wird, a_i dürfe nicht mit unabhängigen Variablen aus dem Modell korrelieren.

Somit ist die Berechnung eines REM mit endogener Dynamik nicht möglich. In dynamischen Modellen bleibt damit lediglich die Möglichkeit, a_i zu involvieren, indem ein Differenzenmodell, vergleichbar mit dem aus Kap. 2.5.2, berechnet wird. Für eine Anwendung vgl. Arminger (1990: 129 ff).

An dieser Stelle soll die Implementierung eines dynamischen Modells in LISREL vorgestellt werden. Hierbei wird von drei Wellen ausgegangen, da das Vorhandensein dreier Zeitpunkte die Implementierung, im Vergleich zu Daten aus zwei Wellen, etwas verkompliziert.

Betrachtet wird wieder das Beispiel des Abschnitts 2.5.1.7, dessen Datensatz im App. unter Tab. 4.1 zu finden ist. Es werden nun die Variablen einzelner Zeitpunkte als eigenständige Variablen gehandhabt. In das Modell finden 5 Variablen Eingang:

- y_1 : Gewicht zum Zeitpunkt $t = 1$
- y_2 : Gewicht zum Zeitpunkt $t = 2$
- y_3 : Gewicht zum Zeitpunkt $t = 3$
- x_2 : Fetthaltiges Essen zum Zeitpunkt $t = 2$
- x_3 : Fetthaltiges Essen zum Zeitpunkt $t = 3$

Zwei Regressionsmodelle werden aufgestellt, welche anschließend ineinander verschachtelt werden:

$$y_2 = ay_1 + bx_2 + e_2 \quad (2.100)$$

$$y_3 = ay_2 + bx_3 + e_3 \quad (2.101)$$

Die Panelstruktur und eine endogene Dynamik erster Ordnung werden durch folgende Aspekte berücksichtigt:

- Das Gewicht zum zweiten Zeitpunkt gilt als abhängig von dem Gewicht zum ersten Zeitpunkt
- Das Gewicht zum dritten Zeitpunkt gilt als abhängig von dem Gewicht zum zweiten Zeitpunkt
- Es wird angenommen, dass der endogene Effekt (symbolisiert mit a) zwischen dem ersten dem zweiten Zeitpunkt dem zwischen dem zweiten und dem dritten Zeitpunkt gleicht
- Es wird ferner angenommen, dass der Effekt der unabhängigen Variablen x (symbolisiert mit b) über die Zeit konstant ist

Um das Modell nun in LISREL zu implementieren, müssen wieder im Vorfeld Messmodelle gebildet werden, welche eine Gleichsetzung von Indikatoren und latenten Variablen erlauben. Diese sehen wie folgt aus:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} \qquad \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ x_1 \\ x_2 \end{pmatrix}$$

Somit gilt:

$$\eta_1 = y_2$$

$$\eta_2 = y_3$$

$$\xi_1 = y_1$$

$$\xi_2 = x_2$$

$$\xi_3 = x_3$$

Man beachte, dass aufgrund der Tatsache, dass auf y_1 keine weitere Variable im Modell einwirkt, sie zu den exogenen (also unabhängigen) ξ -Variablen hinzugezählt wird.

Die Variable y_2 gesellt sich hingegen zu den endogenen Variablen. Sie wirkt zwar auf y_3 , empfängt auf der anderen Seite aber auch einen Einfluss von y_1 .

Diese unterschiedlichen Modell-Platzierungen von y_1 und y_2 haben zur Folge, dass der endogene Effekt a sowohl in der Parameter-Matrix B als auch in Γ jeweils einmal vertreten ist.

So lässt sich nun das Strukturmodell in Matrizenschreibweise darstellen:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} a & b & 0 \\ 0 & 0 & b \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad (2.102)$$

In der Modell-Spezifikation mit LISREL muss vor allem beachtet werden, dass die Parameter a und b doppelt vertreten sind und somit bei der Schätzung gleichgesetzt werden müssen. An dieser Stelle soll die LISREL-Zeile der Modellspezifikation wiedergegeben werden:

```

MO NY=2 NX=3 BE=FU,FI GA=FU,FI TD=FU,FI TE=FU,FI
FR BE(2,1)
FR GA(1,2)GA(1,1)GA(2,3)
EQ BE(2,1)GA(1,1)
EQ GA(1,2)GA(2,3)

```

Da hier mit fiktiven Daten gearbeitet wird, lässt sich die Stichprobengröße beliebig manipulieren. Dies sollte an dieser Stelle auch getan werden, denn durch die Aufspaltung der Variablen nach Zeitpunkten stehen pro Variable lediglich vier Fälle zur Verfügung – eine zu knappe Zahl, um Regressionskoeffizienten adäquat zu schätzen. Der Autor hat sich entschieden, den Datensatz zu ver Hundertfachen, so dass sich eine Stichprobengröße von $n = 400$ ergibt.

Aus diesem Grunde werden inferenzstatistische Maßzahlen nicht betrachtet, wir verbleiben auf der deskriptiven Ebene.

Wichtig sind die beiden Regressionskoeffizienten:

$$\begin{array}{|c|} \hline \mathbf{a} = \mathbf{1.19} \\ \hline \mathbf{b} = \mathbf{-0.41} \\ \hline \end{array} \quad (2.103)$$

Hier zeigt sich das gleiche Phänomen, wie bei dem FEM und REM. Im Vergleich zu einem einfachen Regressionsmodell, in dem die unterschiedlichen Zeitpunkte ignoriert werden, verändert sich das Vorzeichen des Regressionskoeffizienten b in panelanalytischen Modellen. Der Wert wird negativ, was auf einen negativen Zusammenhang zwischen y und x schließen lässt.

Der Einfluss des Gewichtes aus einem früheren Zeitpunkt hat hingegen, wie intuitiv zu erwarten ist, einen positiven Effekt auf das eigene Gewicht zu einem späteren Zeitpunkt.

Weiterhin ist interessant, dass die Determinationskoeffizienten für beide abhängigen Variablen durch die Einbeziehung der endogenen Dynamik sehr hoch werden – für y_2 mit $r^2 = 1,00$ perfekt und für y_3 mit $r^2 = 0,98$ fast perfekt.

Solch hohe Werte dürfen aber in der Euphorie nicht unkritisch betrachtet werden. Denn es stellt sich nun die Frage, inwieweit ein endogener Einfluss einer Variablen auf sich selbst zum späteren Zeitpunkt kausal interpretiert werden kann.

Vielmehr ist zu vermuten, dass sich hinter dem endogenen Einfluss weitere exogene Einflussgrößen verbergen könnten, welche hier implizit mitmodelliert wurden.

Weist z.B. eine Person in einer langen Zeitperiode ein relativ hohes Gewicht auf und wird diese Zeitperiode in einzelne Zeitabschnitte aufgespaltet, dann korreliert das Gewicht sicherlich stark mit sich selbst zwischen verschiedenen Zeitabschnitten – damit ist aber nicht (ursächlich) erklärt, warum eine Person „von vorneherein“ übergewichtig ist, bzw. welche Faktoren es verhindern, dass sie ihr Gewicht reduziert.

Sicherlich stellt sich das Problem der *kausalen Interpretation* grundsätzlich

immer in statistischen Zusammenhangsanalysen. Allerdings sollte sich der Analyst im Kontext der Analyse endogener Dynamik in Panelmodellen diesem Thema besonders kritisch widmen.

Für weitere Ansätze zur Analyse von dynamischen Panelmodellen sei auf Arminger (1990: Kap. 7, 8) verwiesen.

2.6 Lineare Panelmodelle mit latenten Variablen

In diesem Kapitel sollen kurz mögliche Modelle für Paneldaten vorgestellt werden, welche „echte“ latente Variablen beinhalten.

Es ist durchaus von Vorteil, die Anwendung von LISREL bereits an früheren Stellen des Skripts eingeführt zu haben, da hiermit für den Leser das Denken auf der Ebene von Strukturgleichungen nicht gänzlich neu ist.

Der Unterschied besteht nun darin, dass wir es nicht mehr mit einfachen Gleichsetzungen zwischen Indikator und latenter Variablen zu tun haben, sondern mit Messmodellen, in denen Indikatoren und latente Variablen über eine korrelative Beziehung miteinander verbunden sind.

Diese Korrelationen, auch *Ladungen* (λ : Klein-Lambda) genannt, müssen in den meisten Modellvarianten (von LISREL oder anderen Programmen) geschätzt werden, so dass die Anzahl der zu schätzenden Parameter steigt – und damit auch die Möglichkeit für Fehlspezifikationen im Gesamtmodell, da nun auch die Daseinsberechtigung von Beziehungen zwischen einzelnen Indikatoren und latenten Variablen in Frage gestellt werden kann.

Es müssen folglich in einem typischen Strukturgleichungsmodell drei Submodelle aufgestellt werden:

- Das Messmodell, welches die manifesten y -Variablen mit den latenten endogenen η -Variablen verbindet
- Das Messmodell, welches die manifesten x -Variablen mit den latenten exogenen ξ -Variablen verbindet
- Das Strukturmodell, welches die Beziehung der latenten Variablen η und ξ untereinander spezifiziert

Veranschaulicht könnte ein Strukturgleichungsmodell nun wie folgt aussehen:

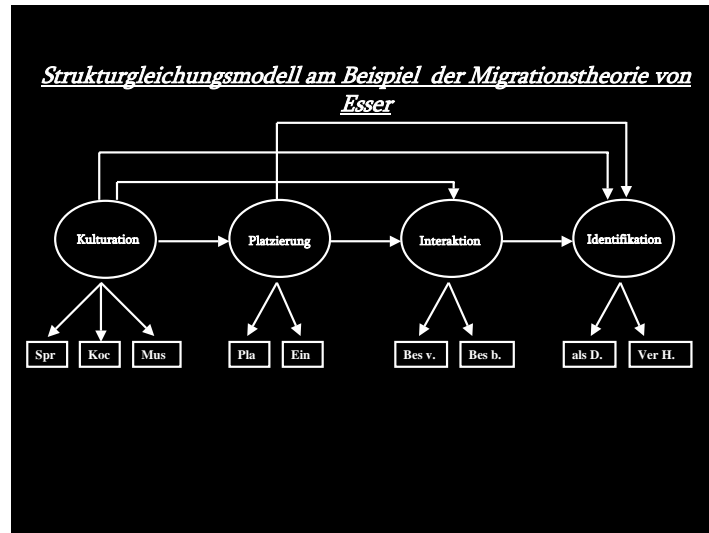


Abbildung 2.6: Veranschaulichung eines Strukturgleichungsmodells

Auf die Inhalte dieses Modells soll an dieser Stelle nicht eingegangen werden, es dient lediglich einer allgemeinen Veranschaulichung.

Festgehalten werden sollte, dass diese Graphik eine Erweiterung der im Kap. 2.2 vorgestellten Methode der Visualisierung von Modellen über *Pathdiagramme* darstellt. Es wird, zusätzlich zu den Punkten aus Kap. 2.2, konventionell unterschieden zwischen:

- **Kreisen** für *latente Variablen* und
- **Rechtecken** für *Indikatoren*.

In diesem Beispiel gibt es auf der Ebene latenter Variablen nur eine ξ -Variable (Kreis ganz links: „Kulturation“), da nur auf diese Variable keine Pfeile zeigen, und drei η -Variablen („Platzierung“, „Interaktion“ und „Identifikation“).

Jede latente Variable wird indirekt gemessen bzw. erzeugt durch zwei bis drei Indikatoren (Rechtecke).

Aufgrund der Komplexität solcher Modelle muss der Analyst oft bei der Anwendung auf eine inhaltliche Fragestellung eine **Vielzahl** ähnlicher Modelle testen, bis er zu einem akzeptablen Modell gelangt. Das Auffinden und

Aufstellen eines tauglichen Modells ist somit meistens ein iterativer Prozess und keine standardisierte Prozedur „nach Vorschrift“.

Ansonsten lassen sich für Paneldaten im Kontext linearer Strukturgleichungsmodelle ähnliche Modellunterscheidungen vornehmen, wie die bereits im Zuge von Regressionsmodellen (vgl. Kap. 2.5) vorgestellten.

Auch auf der Ebene von Strukturgleichungen lässt sich z.B. eine Fehlervarianzzerteilung vornehmen, korrelierende Residuen (sowohl für Messmodelle als auch für Strukturmodelle) spezifizieren und zwischen dynamischen und statischen Modellen differenzieren.

An dieser Stelle wird ein Grundgerüst für ein einfaches Strukturgleichungsmodell vorgestellt und ein paar Modellvarianten besprochen.

Es muss hierbei grundlegendes Wissen in Bezug auf Strukturgleichungsmodelle vorausgesetzt werden. Für eine ausführliche Einführung und Vertiefung sei vor allem auf das Buch von Reinecke (2005), aber auch auf frühere Stellen dieses Skripts (z.B. auf die Ausführungen zur Pfadanalyse, zum Ein-Indikator-Modell und zu LISREL-Modellimplementierungen) verwiesen.

Das nun aufgegriffene Beispiel bezieht sich auf einen kleinen Ausschnitt aus dem Integrationsansatz von Esser (1999). In diesem Ansatz wird versucht, den Prozess der Integration von Migranten zu systematisieren. Zu diesem Prozess gehören u.a. vier Stufen der sozialen Integration, welche aufeinander aufbauen. Diese vier Stufen lassen sich als latente Variablen auffassen, wie in Graphik 2.6 dargestellt.

Eine theoretische Auseinandersetzung mit den Inhalten dieses Ansatzes wäre hier fehl am Platze, es sollen viel mehr statistisch-theoretische Ausführungen auf eine inhaltliche Fragestellung angewendet werden. Das Beispiel dient also lediglich der Illustration.

Eine erste Stufe des Integrationsprozesses ist nach Esser die sog. *Kulturation*. Sie umfasst vor allem die *Sprachkompetenz*. Letztere lässt sich auch als latente Variable verstehen. Diese übt einen Einfluss auf die sog. *Identifikation* aus – ebenfalls eine latente Größe, welche die emotionale Verbundenheit des Migranten mit dem Aufnahmeland angibt.

Es soll hier somit die Hypothese untersucht werden, ***ob Sprachkompetenzen*** (exogene Variable) ***eine positive Wirkung auf die Identifikation***

(endogene Variable) *haben*. Beide Konstrukte sollen nun über Indikatoren aus dem SOEP-Datensatz (zu diesem Datensatz s. einführende Bemerkungen in der Einleitung des Skripts) indirekt messbar gemacht werden. Es ergibt sich folgende einfache Messkonstellation:

LATENTE VARIABLE	INDIKATOR
<i>Sprachkompetenzen</i> (SP)	<i>Deutsch schreiben</i> (Schr)
	<i>Deutsch sprechen</i> (Spre)
<i>Identifikation</i> (IDT)	„ <i>sich als Deutscher fühlen</i> “ (s.a.D.f.)

Tabelle 2.6: Verbindung zwischen Indikatoren und latenten Variablen

Wie zu sehen ist, werden *Sprachkompetenzen* durch zwei Indikatoren greifbar gemacht, während die *Identifikation* aus Gründen der Vereinfachung lediglich mit einem Indikator gleichgesetzt wird.

Alle drei Variablen sind ordinalskaliert mit fünf Ausprägungen und werden hier als quasi-metrisch aufgefasst. Sie reflektieren *Selbsteinschätzungen* ausländischer Personen.

Nun wird noch zwischen zwei Zeitpunkten differenziert: 2001 und 2003. Dadurch ergeben sich vier latente Variablen und sechs Indikatoren, wie folgende Graphik illustriert:

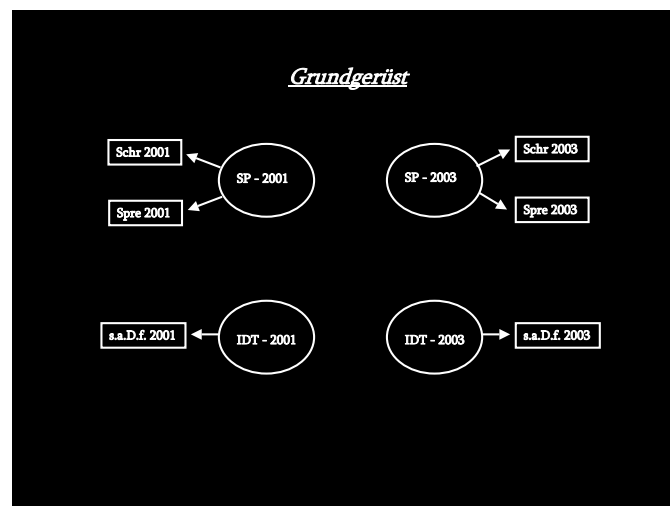


Abbildung 2.7: Grundgerüst

Der Vorteil des SOEP-Paneldatensatzes liegt darin, dass er eine überproportionale Ausländer-Stichprobe beinhaltet. Nach einigen Datenaufbereitungsschritten konnte vom Autor die Korrelations- und die Kovarianzmatrix der sechs Indikatoren, welche als Basis für weitere Berechnungen zur Verfügung steht, ermittelt werden. In ihr sind 391 gültige Fälle erfasst. Alle bivariaten Korrelationen sind auf dem 5%-Niveau signifikant von Null verschieden. Ferner sind alle Zusammenhänge positiv.

Die beiden Matrizen befinden sich im Anhang 4.5, wobei die Korrelationsmatrix in der Graphik 4.4 und die Kovarianzmatrix in der Graphik 4.5 aufgelistet ist.

Nun sollen einige Varianten vorgestellt werden, welche die *Identifikation* als abhängige und die *Sprachkompetenzen* als unabhängige Variablen vorsehen. Das oben vorgestellte Grundgerüst soll somit mit „Leben gefüllt werden“. Es wird in jedem Modell dem Vorteil, dass Paneldaten vorliegen, Rechnung getragen. Die Varianten stellen allerdings nur einen Auszug möglicher Modelle dar.

Bei den ersten Modellen wird in der visuellen Darstellung auf die Fehlerterme aus Veranschaulichungsgründen verzichtet. Es gilt die Annahme, dass sowohl die Residuen aus den Messmodellen als auch aus dem Strukturmodell untereinander unkorreliert sind.

MODELLVARIANTE A (*statisches Modell mit kreuzverzögertem Effekt*)

Hier wird ein sehr einfaches statisches Modell dargestellt. Es wird lediglich der kreuzverzögerte Effekt zwischen den beiden latenten Variablen spezifiziert:

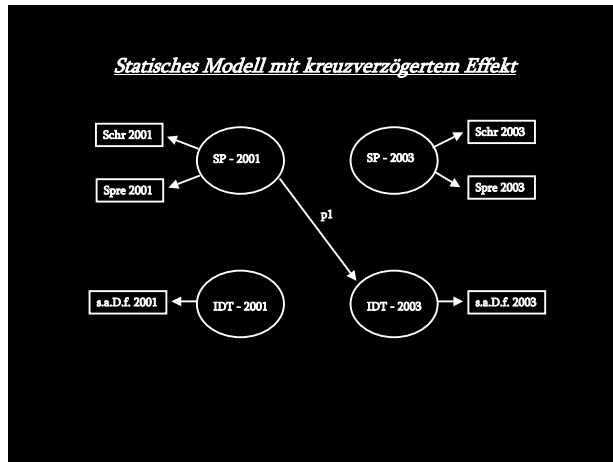


Abbildung 2.8: Modellvariante A

Die Strukturgleichung lautet:

$$(\text{IDT2003}) = (p_1) (\text{SP2001}) + (\zeta_1)$$

mit den Messmodellen:

$$\begin{pmatrix} \text{Schr2001} \\ \text{Spre2001} \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} (\text{SP2001}) + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \quad \text{IDT2003} = \text{s.a.d.f.2003}$$

MODELLVARIANTE B (*Zwei Querschnittsmodelle*)

Als nächstes werden die beiden Querschnitts-Modelle für 2001 und 2003 betrachtet:

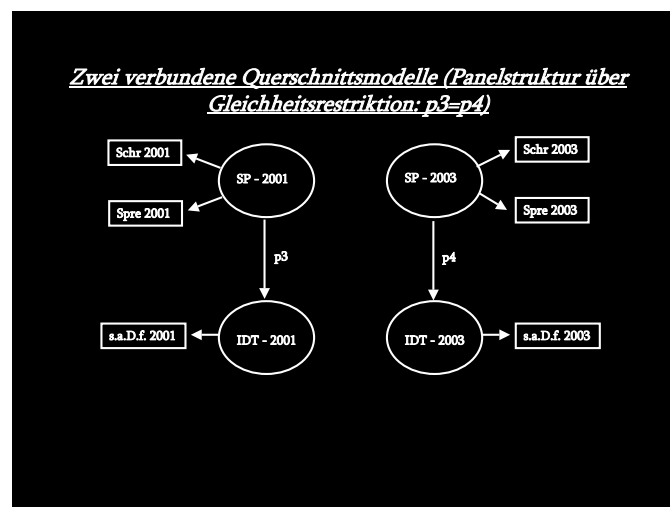


Abbildung 2.9: Modellvariante B

Die Strukturgleichung lautet:

$$\begin{pmatrix} \text{IDT2001} \\ \text{IDT2003} \end{pmatrix} = \begin{pmatrix} p_3 & 0 \\ 0 & p_4 \end{pmatrix} \begin{pmatrix} \text{SP2001} \\ \text{SP2003} \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

mit den Messmodellen:

$$\begin{pmatrix} \text{Schr2001} \\ \text{Spre2001} \\ \text{Schr2003} \\ \text{Spre2003} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & \lambda_3 \\ 0 & \lambda_4 \end{pmatrix} \begin{pmatrix} \text{SP2001} \\ \text{SP2003} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} \quad \begin{pmatrix} \text{s.a.d.f.2001} \\ \text{s.a.d.f.2003} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \text{IDT2001} \\ \text{IDT2003} \end{pmatrix}$$

Hierbei ist zu beachten, dass die Panelstruktur in diesem Modell dann Eingang findet, wenn angenommen wird, dass der Effekt der *Sprachkompetenz* auf die *Identifikation* zu beiden Zeitpunkten gleich ist. Formal geschieht dies über: $\mathbf{p}_3 = \mathbf{p}_4$. An dieser Stelle werden die beiden Querschnittsmodelle miteinander verbunden und somit die Panelstruktur der Daten berücksichtigt.

MODELLVARIANTE C (*Einfaches Modell mit endogener Dynamik*)

In dieser Modellvariante wird eine endogene Dynamik implementiert. Das heißt, dass neben dem Einfluss der *Sprachkompetenz* auch eine Eigenwirkung der latenten Variablen *Identifikation* auf sich selbst (von 2001 auf 2003) zugelassen wird. Dieses Modell lässt sich wie folgt veranschaulichen:

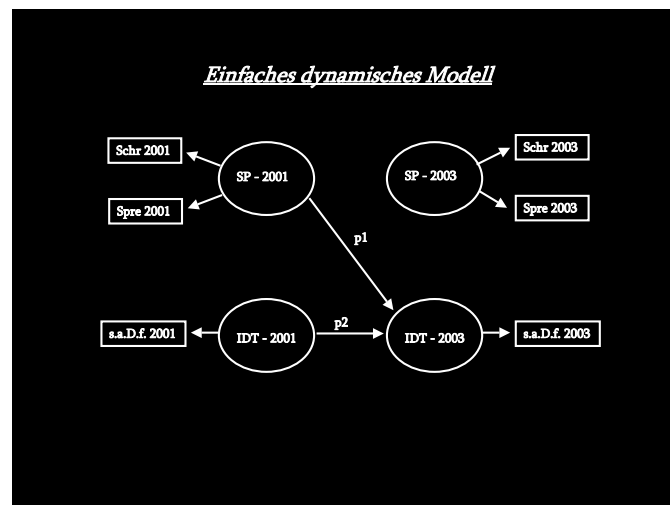


Abbildung 2.10: Modellvariante C

Die Strukturgleichung lautet:

$$(\text{IDT2003}) = (p_1 \quad p_2) \begin{pmatrix} \text{SP2001} \\ \text{IDT2001} \end{pmatrix} + (\zeta_1)$$

mit den Messmodellen:

$$\begin{pmatrix} \text{Schr2001} \\ \text{Spre2001} \\ \text{s.a.D.f.2001} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \text{SP2001} \\ \text{IDT2001} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} \quad \text{IDT2003} = \text{s.a.d.f.2003}$$

Man beachte, dass somit die latente Variable *Identifikation 2001* zu den exogenen ξ -Variablen gezählt wird.

MODELLVARIANTE D („Vollgepacktes Modell“)

Leider geschieht es immer wieder, dass z.B. ein Mangel an fundierten Informationen den Forscher dazu veranlasst, ein Modell unreflektiert „vollzupacken“. Dies ist zwar an sich nicht verwerflich, muss aber keine Verbesserung der Modellanpassung an die Realität bedeuten.

Ein gutes Modell ist u.a. dadurch gekennzeichnet, dass es aus der komplexen Realität Sachverhalte und Zusammenhänge abstrahiert und gleichzeitig vereinfacht darstellt. Ein „vollgepacktes“ Modell kann sich zwar mit einer höheren Wahrscheinlichkeit der Realität anpassen, verliert aber u.U. an Aussagekraft, da sich keine einfachen Zusammenhänge aus ihm ablesen lassen. Ferner kann in solchen Modellen das Phänomen der Multikollinearität steigen, so dass Schätzer ineffizient werden.

Folgendes Beispiel soll nun betrachtet werden (wobei dieses Modell nur „vollgepackt“ in Relation zu den vorherigen Modellen ist – es ist immer noch von einer relativen Einfachheit gegenüber Strukturgleichungsmodellen in der Praxis gekennzeichnet):

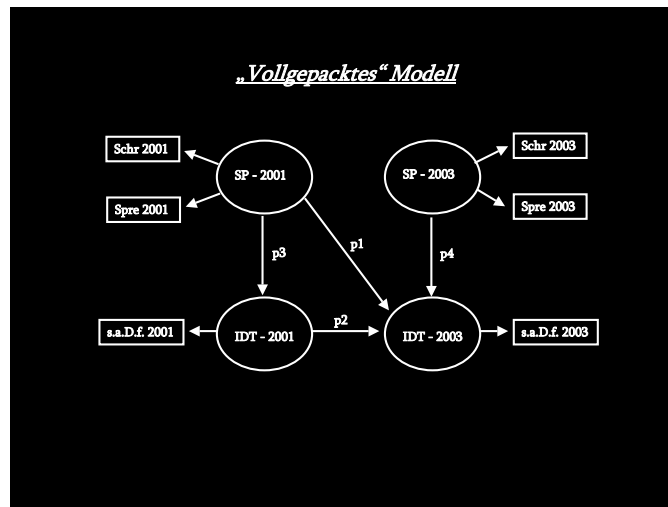


Abbildung 2.11: Modellvariante D

Die Strukturgleichung lautet:

$$\begin{pmatrix} \text{IDT2001} \\ \text{IDT2003} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ p_2 & 0 \end{pmatrix} \begin{pmatrix} \text{IDT2001} \\ \text{IDT2003} \end{pmatrix} + \begin{pmatrix} p_3 & 0 \\ p_1 & p_4 \end{pmatrix} \begin{pmatrix} \text{SP2001} \\ \text{SP2003} \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

mit den Messmodellen:

$$\begin{pmatrix} \text{Schr2001} \\ \text{Spre2001} \\ \text{Schr2003} \\ \text{Spre2003} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ 0 & \lambda_3 \\ 0 & \lambda_4 \end{pmatrix} \begin{pmatrix} \text{SP2001} \\ \text{SP2003} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} \quad \begin{pmatrix} \text{s.a.d.f.2001} \\ \text{s.a.d.f.2003} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \text{IDT2001} \\ \text{IDT2003} \end{pmatrix}$$

Auch hier sollen die beiden Querschnittseffekte gleichgesetzt werden: $p_3 = p_4$.

MODELLVARIANTE E („Vollgepacktes Modell“ mit Autokorrelation der Messfehler)

Bislang wurde eine in diesem Skript oftmals angeschnittene Eigenheit von Paneldaten ignoriert: Autokorrelation der Fehler inhaltlich gleicher Variablen von gleichen linearen Modellen zu verschiedenen Zeitpunkten.

Es ist nämlich wahrscheinlicher, dass die Residuen einer Person in einem linearen Modell sich ähneln, als die Residuen zweier Personen, welche zufällig in einem Datensatz nebeneinander liegen (und als die Residuen einer

Person, welche sich aus verschiedenen linearen Submodellen eines übergeordneten Modells ergeben). Auf Aggregatebene manifestiert sich eine solche tendenzielle Ähnlichkeit in der Autokorrelation von Residuen.

Diese Autokorrelation soll nun im „vollgepackten“ Modell implementiert werden. Es werden dabei nur die Fehler in der Ladungsmatrix Λ_x betrachtet, da die y -Variablen in der Matrix Λ_y durch die Gleichsetzung mit den Indikatoren nicht mit Residualtermen behaftet sind. Folgende Abbildung ist erweitert um die Residuen δ und die korrelativen Beziehungen zwischen ihnen:

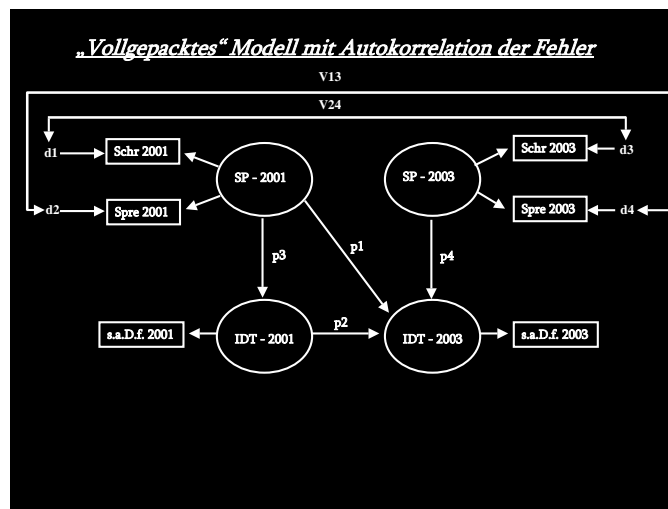


Abbildung 2.12: Modellvariante E

Der Buchstabe „d“ steht hierbei für den Fehler δ , so dass z.B. „d3“ δ_3 entspricht. Der Buchstabe „V“ steht für v , das ein Element der Varianz-Kovarianzmatrix der Messfehler von x , also Θ_δ , repräsentiert. So sei definiert: $V13 = v_{13}$ und $V24 = v_{24}$.

Die Varianz-Kovarianzmatrix Θ_δ ist somit unter Berücksichtigung der Korrelation der Fehler *inhaltlich gleicher Variablen zu verschiedenen Zeitpunkten* gegeben durch:

$$\Theta_\delta = \begin{pmatrix} v_{11} & & & \\ 0 & v_{22} & & \\ v_{13} & 0 & v_{33} & \\ 0 & v_{24} & 0 & v_{44} \end{pmatrix}$$

Alle Modelle wurden in LISREL implementiert und gerechnet. Im Anhang

4.6 befinden sich die Zeilen der jeweiligen Modellspezifikationen.

Hier sollen tabellarisch die wichtigsten LISREL-Schätzer und χ^2 für die jeweiligen Modellvarianten ausgegeben werden:

Modellvarianten	p_1	p_2	p_3	p_4	λ_1	λ_2	λ_3	λ_4	r^2	χ^2
Mod. A	0,44	–	–	–	1,05	0,84	–	–	ID03 : 0,19	0
Mod. B	–	–	0,5	0,5	1,08	0,82	1,09	0,86	ID01 : 0,21 ID03 : 0,22	214,2
Mod. C	0,17	0,52	–	–	1,01	0,88	–	–	ID03 : 0,44	2,65
Mod. D	–0,27	0,56	0,48	0,48	1,09	0,82	1,09	0,86	ID01 : 0,2 ID03 : 0,46	68,58
Mod. E	–0,28	0,54	0,49	0,49	0,99	0,83	1,02	0,91	ID01 : 0,21 ID03 : 0,46	5,03

Insgesamt lässt sich sagen, dass sich im Wesentlichen ein Zusammenhang zwischen *Sprachkompetenzen* und *Identifikation* an den Daten bestätigt.

Alle in den verschiedenen Modellvarianten geschätzten Koeffizienten sind auf dem 5%-Niveau signifikant von Null verschieden. Diese Koeffizienten sind unstandardisiert, aber dadurch, dass die Maßeinheiten für alle Variablen und in allen Modellen gleich sind (5-stufige Rangskalen), sind Vergleiche zwischen den Modellen möglich.

Zunächst einmal lässt sich feststellen, dass die sich im Messmodell der „Sprachvariablen“ befindlichen Ladungen sich als relativ stabil über verschiedene Modellvarianten und über die beiden Zeitpunkte hin erwiesen haben. Auch die hier nicht aufgelisteten Determinationskoeffizienten der Gleichungen in den Messmodellen waren meist sehr hoch angesiedelt ($r^2 > 0,7$). Die größte Veränderung zeigte sich im Modell mit implementierter Autokorrelation der Messfehler – was nicht weiter verwunderlich sein dürfte, da dieser Eingriff speziell die Ladungsmatrix Λ_x betrifft.

An dieser Stelle sei kurz auf die Maßzahl χ^2 eingegangen. In jedem Strukturgleichungsmodell wird durch die Modellvorgaben und die Konstellationen zwischen Variablen eine theoretische Kovarianzmatrix erzeugt. In dieser Matrix werden die Kovarianzen der manifesten Variablen reproduziert.

Schließlich wird diese theoretisch reproduzierte Kovarianzmatrix mit der empirischen Ausgangsmatrix verglichen. Je ähnlicher sich beide sind, umso eher ist das Modell in der Lage, sich der in der Empirie gewonnene Datenstruktur, also pathetisch ausgedrückt: der „Realität“ anpassen.

Die Maßzahl χ^2 basiert auf dem Vergleich zwischen theoretischer und empirischer Kovarianzmatrix. Je niedriger der Wert, umso ähnlicher sind sich beide Matrizen und somit umso besser ist das Modell geeignet, die Realität zu beschreiben.

Dies sollte an dieser Stelle an Ausführungen reichen. Vertiefende Notationen zur Modellbeurteilung sind zu finden in Arminger (2005: 115-129).

Angewendet auf die Statistiken der hier aufgeführten einzelnen Modellvarianten ist zunächst festzustellen, dass im einfachen Modell (Mod. A) mit kreuzverzögertem Effekt die Anpassung perfekt ist mit $\chi^2 = 0$. Das liegt allerdings nicht daran, dass das Modell perfekt im Sinne der Erklärungskraft ist, was an dem Determinationskoeffizienten zu sehen ist. Es ist lediglich durch sein Einfachheit so gestrickt, dass sich rein mathematisch die Kovarianzmatrix perfekt reproduzieren lässt⁴⁷. Der χ^2 -Wert wird also erst für komplexere Modelle relevant.

Interessant ist dieser Wert für das „Zwei-Querschnittsmodell“, der mit $\chi^2 = 214,2$ enorm hoch ist. Anscheinend sind die beiden Querschnitts-Modelle nicht wirklich kompatibel.

Aber auch nachträgliche Berechnungen des Autors, in denen von der Gleichheitsrestriktion $p_3 = p_4$ abgesehen wurde und somit p_4 als eigenständiger Koeffizient zur Schätzung freigegeben wurde, ändern an dem hohen χ^2 -Wert nichts. Generell hat dieses Modell eher schlecht abgeschnitten – auch von der Erklärungskraft her.

Ein weiterer eindrucksvoller Befund manifestiert sich in dem Vergleich zwischen dem χ^2 -Wert des vollgepackten Modells *ohne* Autokorrelation der Fehler (Mod. D) und dem des vollgepackten Modells *mit* Autokorrelation der Fehler (Mod. E): $\chi^2 = 68,58$ vs. $\chi^2 = 5,03$. Auch wenn sonstige Koeffizienten sich nicht wesentlich verändert haben, so ist die Einführung der korrelativen

⁴⁷ Dies trifft auch auf einfache Regressions- und Pfadmodelle zu

Beziehung zwischen Fehlern zu verschiedenen Zeitpunkten in der Lage, das Modell wesentlich besser an die empirischen Daten anzupassen. Es wird somit der „Panelstruktur“ gerechter.

Am besten scheint das einfache dynamische Modell (Mod. C) abzuschneiden: Ein geringer χ^2 -Wert, ein relativ hoher r^2 -Wert und inhaltlich gesehen ein plausibles Modell.

Auffallend ist, dass der kreuzverzögerte Effekt durch die Einführung der endogenen Dynamik deutlich kleiner geworden ist – ein Hinweis für Fehlspezifikation im „kreuzverzögerten Modell“ (Mod. A).

In dem „vollgepackten Modell“ (Mod. D und E) kehrt sich sogar die Richtung des Zusammenhangs des kreuzverzögerten Effekts um – dieser wird negativ. Es wäre zu diskutieren, ob dies eher den kreuzverzögerten Effekt in Frage stellt oder eher auf ein zu „vollgepacktes Modell“ hinweist, indem die einzelnen Koeffizienten nicht mehr effizient, mit minimaler Varianz, geschätzt werden.

Dieses Kapitel sollte einen kleinen Auszug darstellen, wie man sich an die Verknüpfung dieser beiden komplexen Ebenen:

- Ebene der Paneldatenstruktur und
- Ebene von Strukturgleichungsmodellen

herantasten kann.

Natürlich sind verschiedene Erweiterungen und die Modellierung komplexerer Modelle möglich. Es sei z.B. auf die Möglichkeit der Implementierung eines REM (s. Kap. 2.5.1) oder eines Differenzenmodells (s. Kap. 2.5.2) auf der Ebene von Strukturgleichungen verwiesen, nachzulesen bei Arminger (1990).

2.7 Anwendung von LISREL auf Paneldaten

Es sei an dieser Stelle noch etwas zu LISREL, also der Software, in der komplexe lineare Modelle zur Berechnung implementiert werden können, gesagt:

48

⁴⁸Der Autor arbeitet mit LISREL 8.53

Zum ersten Mal in diesem Skript tauchte eine Anwendung in Kap. 2.5.2, also im Zuge der Berechnung eines Differenzenmodells auf. Von LISREL wurde im weiteren Verlauf des Skripts Gebrauch gemacht im Zusammenhang mit immer komplexer werdenden Modellen.

Dies soll nicht den Eindruck erwecken, als sei LISREL **nur** in der Lage, komplexe Strukturgleichungsmodelle zu berechnen. Ebenso lassen sich mit diesem Programm einfache Regressions- und pfadanalytische Gleichungen modellieren, welche keine latenten Variablen beinhalten.

Zwei offensichtliche Gründe gibt es, warum dieses Programm im Kontext verhältnismäßig einfacher Modelle eher selten zur Anwendung kommt:

Erstens liegt dies an der Umständlichkeit der Implementierung solcher Modelle. Eine lineare Regression lässt sich mit STATA oder SPSS ohne viel Aufwand berechnen. In LISREL hingegen muss „getrickst“ werden. Denn es wird immer auf der Ebene von Strukturgleichungen modelliert – und solche Strukturgleichungen sehen eine Spezifikation latenter Variablen vor. Um nun einfache Modelle ohne latente Variablen zu implementieren, müssen die beobachteten Variablen aus einem Regressionsmodell als latente Variablen definiert und mit sich selbst als Indikatoren gleichgesetzt werden. Dies erfordert Kenntnisse in Hinblick auf Strukturgleichungsmodelle und auf die Matrizenrechnung (welche nicht zwingend gegeben sind, wenn man sich mit der Regressionsanalyse auskennt) und bringt einen gewissen Aufwand mit sich.

Zweitens liegt es daran, dass in vielen einfachen Modellen eine eindeutige arithmetische Lösung bei der Schätzung von Koeffizienten errechenbar ist. In solchen Fällen ist es sicherlich nicht begründbar, eine eindeutige Lösung gegen eine einzutauschen, welche über die Annäherung durch iterativ arbeitende Algorithmen, wie im Falle der *Maximum-Likelihood-Schätzung* bei LISREL, bestimmt wird (auch wenn sich hierbei die Schätzungen der beiden Methoden z.T. nur minimal um Nachkomma-Stellen unterscheiden).

Nichtsdestotrotz schadet es nicht, sich bei der Aneignung des Umgangs mit LISREL auch mal an einfachen Modellen zu versuchen. Denn ein in LISREL eingepflegtes bi- oder multivariates Regressionsmodell lässt sich auf seine Richtigkeit hin mit anderen Standard-Statistik-Programmen prüfen. Es kann somit für das Vorankommen beim Erlernen von LISREL förderlich sein, die ersten Schritte auf einer solch einfachen Ebene zu tätigen.

Ein Nachweis für die Anwendbarkeit von LISREL auf einfache Modelle soll an dieser Stelle geliefert werden. Hierzu wird das Pfadmodell aus Kap. 2.3 betrachtet.

Wie bereits mehrmals erwähnt, müssen zuerst die beobachteten Variablen mit Bezeichnungen für latente Variablen versehen werden.

Auf der Seite der endogenen Variablen wird definiert:

$$\begin{aligned}x_2 &= \eta_1 \\y_2 &= \eta_2\end{aligned}$$

Die exogenen Variablen sind gegeben durch:

$$\begin{aligned}x_1 &= \xi_1 \\y_1 &= \xi_2\end{aligned}$$

Die Messmodelle müssen so konstruiert sein, dass die Indikatoren den latenten Variablen entsprechen. Diese Forderung ist erfüllt, wenn die Ladungsmatrizen Einheitsmatrizen entsprechen. So kann für die endogenen Variablen formuliert werden:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \quad (2.104)$$

Das Äquivalent für die exogenen Variablen sieht wie folgt aus:

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \quad (2.105)$$

Nun lässt sich das Strukturgleichungsmodell in Matrizen Schreibweise definieren:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} p_{x_2x_1} & p_{x_2y_1} \\ p_{y_2x_1} & p_{y_2y_1} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad (2.106)$$

Die Schreibweise der zu schätzenden Stabilitätskoeffizienten und kreuzverzögerten Effekte ist hierbei mit der aus Kap. 2.3 identisch. Die Matrix, welche nur Nullen enthält (i.d.R. symbolisiert mit B), hat die Funktion, gerichtete oder ungerichtete Beziehungen zwischen endogenen Variablen zu formulieren.

Da solche Beziehungen hier nicht modelliert werden, entspricht B einer Nullmatrix, so dass dieser Block wegfällt und keine η -Variablen auf der rechten Seite der Strukturgleichung auftauchen. Kompakt geschrieben nimmt deshalb diese Gleichung diese Form an:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} p_{x_2x_1} & p_{x_2y_1} \\ p_{y_2x_1} & p_{y_2y_1} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \quad (2.107)$$

An dieser Stelle muss elementares Wissen bzgl. der LISREL-Syntaxsprache vorausgesetzt werden. Weiter unten befinden sich diesbezüglich Verweise auf nützliche Quellen im Internet. Außerdem sei das an mehreren Stellen bereits zitierte Buch von Arminger (1990) empfohlen, in welchem verschiedene panelanalytische Modelle in LISREL berechnet und ausführlich dargestellt werden. Hier wird nur die Zeile der Modellspezifikation wiedergegeben. Zuerst die ausführliche Variante:

```
MO NY=2 NX=2 NE=2 LY=FU,FI GA=FU,FR PS=DI TE=DI,FI
ST 1 LY(1,1) LY(2,2)
```

In der Modellformulierung wird explizit gesagt, dass es 2 η -Variablen gibt und das Messmodell aus Gl. 2.104 wird spezifiziert. Diese Angaben sind nicht einmal nötig, da beim ihrem Weglassen LISREL diese Zusammenhänge automatisch annimmt.

Solange die Anzahl der endogenen latenten Variablen der Anzahl ihrer Indikatoren entspricht – wie in diesem Falle –, ist der Ausdruck **NE=2** redundant. Solange ferner die Ladungsmatrix aus Gl. 2.104 eine Einheitsmatrix ist, braucht sie nicht in LISREL spezifiziert zu werden (wie hier durch: **LY=FU; ST 1 LY(1,1) LY(2,2)**). Man beachte allerdings: Wenn die Anzahl NE angegeben wird, dann muss auch LY definiert werden.

Nun kann man dieses Modell gekürzt wie folgt darstellen:

```
MO NY=2 NX=2 GA=FU,FR PS=DI TE=DI,FI
```

Die Schätzung der Koeffizienten nach der Maximum-Likelihood-Methode ergibt:

Stabilitätskoeffizienten	$p_{x_2x_1} = 0,88$ $p_{y_2y_1} = 0,68$
Kreuzverzögerte Effekte	$p_{x_2y_1} = -0,03$ $p_{y_2x_1} = 0,28$

Der Querschnittseffekt $p_{x_1y_1} = 0,742$ wurde hier nicht explizit aufgeführt, da dieser der Korrelation $r_{x_1y_1}$ entspricht und nicht mit LISREL geschätzt wurde.

Es ist festzustellen, dass hier die maximum-likelihood-basierten Schätzer der eindeutigen arithmetischen Lösungen aus Kap. 2.3 sehr ähnlich sind. Die größte Diskrepanz taucht bei $p_{y_2y_1}$ auf. Die Differenz beider Koeffizienten liegt bei $|0,09|$. Sonst sind die LISREL-Schätzungen viel näher an der arithmetischen Lösung. Insgesamt lässt sich somit dieses Vorgehen als akzeptabel einstufen.

Ein weiterer Vorteil beim Arbeiten mit LISREL liegt darin, dass der simple Befehl **PD** ein Pfaddiagramm ausgibt, und zwar inkl. der mit Koeffizientenwerten beschrifteten Pfeile. Dieser Befehl sollte einmalig zwischen zwei der vier typischen Komponenten der LISREL-Syntax stehen, also zwischen

1. der Kopfzeile mit Labels der Indikatoren,
2. der Kovarianz- oder Korrelationsmatrix der Indikatoren,
3. der Modellspezifikation,
4. oder der Output-Anforderungen.

In diesem Beispiel sieht das Pfaddiagramm, vom Autor etwas bearbeitet, wie folgt aus:

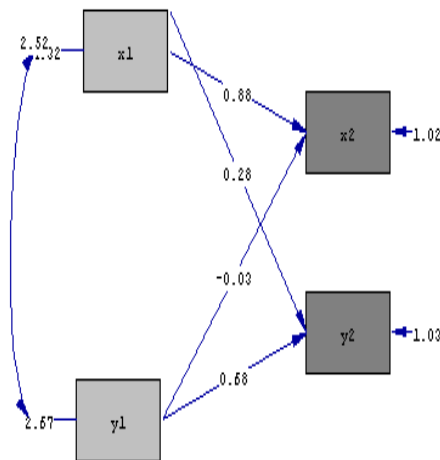


Abbildung 2.13: Pfaddiagramm zum Beispiel

Soviel sei zu dem Pfadmodell gesagt.

Ferner soll kurz gezeigt werden, wie sich das Differenzenmodell aus Kap. 2.5.2, spezifiziert in Gl. 2.92, in LISREL implementieren lässt.

Hierzu soll der Datensatz aus Appendix 4.3 (s. Abb. 4.3.2) der Analyse unterzogen werden. Wie bereits im Kontext der Bildung von Regressionsmodellen mit Differenzenvariablen gesagt worden ist, hat dieser fiktive Datensatz einen ziemlich unrealistischen Charakter. Dies äußert sich darin, dass die Varianz der abhängigen Variablen zu 100% erklärt wird ($r^2 = 1$) und alle Regressionskoeffizienten den Wert 1 betragen. Insofern reduziert sich die Gleichung 2.91 auf:

$$y_2 = y_1 + x_2 - x_1 \tag{2.108}$$

Nichtsdestotrotz ist es interessant, ein solches Modell in LISREL zu implementieren und die aus dem künstlichen Datensatz resultierenden Parameterschätzungen mit den vermuteten Koeffizienten ($b_1, b_2 = 1$) zu vergleichen.

Die Modellspezifikation nimmt die folgende Form an:

```
MO NY=1 NX=3 GA=FU,FI PS=DI TE=DI,FI
ST 1.0 GA(1,1)
FR GA(1,2) GA(1,3)
```

Mit „**ST 1.0 GA(1,1)**“ wird der Schätzer der unechten abhängigen Variablen y_1 auf Eins fixiert (s. Gl. 2.92 und 2.91 und ferner Erläuterungen im

Text an dieser Stelle).

Mit „**FR GA(1,2) GA(1,3)**“ werden b_1 und b_2 zur Schätzung frei gegeben. Es bestätigt sich mit LISREL: $b_1, b_2 = 1$.

Diese Ausführungen sollten einen Einblick in die vielfältigen Möglichkeiten von LISREL verschaffen. Durch die Option, viele Schätzer in komplexen Modellen auf bestimmte Werte zu fixieren, freizugeben oder mit anderen Schätzern gleichzusetzen, hat der Forscher einen weiten Spielraum, seine Modelle zu testen und durch kleine Variationen zu versuchen, ihre Anpassung an reale Daten zu verbessern.

Zum Schluss des Kapitels noch ein paar Verweise auf Quellen im Internet, welche für LISREL-Einsteiger nützlich sind:

Download der LISREL-Studentenversion

<http://www.ssicentral.com/lisrel/student.html>

LISREL-Hilfen für Einsteiger

http://user.uni-frankfurt.de/~cswerner/sem/free_fix.pdf

<http://www.soziologie.uni-halle.de/langer/lisrel/skripten/lisrel83.pdf>

<http://www.ssicentral.com/lisrel/techdocs/SIMPLISSyntax.pdf>

Kapitel 3

Fazit

Dieses Skript zeigte einen kleinen Ausschnitt aus einem riesigen Statistik-Subuniversum: dem der linearen Analyse von Paneldaten. Gewiss ist das Hantieren mit Daten, welche sich zusätzlich zu herkömmlichen Querschnittsdaten noch an der Zeitachse entlang differenzieren, kein einfaches Unterfangen. Vor allem nicht, wenn diese Daten für voraussetzungsvolle statistisch-theoretische Überprüfungen genutzt oder mit der abstrakten Analyse latenter Größen konfrontiert werden.

Aber aus eigener Erfahrung lässt sich sagen, dass man sich Schritt für Schritt in dieses „Statistik-Paradigma“ hineindenken kann und damit merkt, wie wertvoll die Panelanalyse für (sozial-)wissenschaftliche Fragestellungen ist.

Somit sei gehofft, dass dieses Skript einen leichten, aber dennoch fundierten Einstieg in diese Thematik erlaubt und die Idee der Wichtigkeit solcher Analysen ein Stück weit fördert – trotz hohen Aufwands auf der Ebene der Datenerhebung und relativ hoher Komplexität auf der Ebene der Datenanalyse.

Kapitel 4

Appendix

4.1 Appendix A - Fiktiver Datensatz mit variierender x_2 -Variable

X1	X2WENIG	X2MITTEL	X2VIEL	X1X2WENI	X1X2MITT	X1X2VIEL
1	1	5	25	0	4	24
2	2	10	50	0	8	48
3	3	15	75	0	12	72
4	5	22	100	1	18	96
5	5	25	125	0	20	120
6	6	30	150	0	24	144
7	7	35	172	0	28	165
8	8	40	200	0	32	192
9	9	50	225	0	41	216
10	10	50	250	0	40	240
11	9	55	275	-2	44	264
12	12	60	300	0	48	288
13	13	65	325	0	52	312
14	14	66	350	0	52	336
15	15	75	375	0	60	360
16	19	80	400	3	64	384
17	17	86	411	0	69	394
18	18	90	450	0	72	432
19	22	95	475	3	76	456
20	20	100	515	0	80	495

Abbildung 4.1: Fiktiver Datensatz 1

4.2 Appendix B - Fiktiver Datensatz zur Pfadanalyse mit Paneldaten

```
x1y1x2y2
1 4 2 4
2 2 2 2
3 3 3 3
4 2 4 1
5 5 5 5
6 4 6 6
7 7 7 7
1 1 1 2
2 2 2 2
3 3 3 3
4 4 3 4
5 5 5 4
5 6 5 4
7 4 7 7
1 1 1 1
2 4 2 5
3 3 3 3
4 4 4 4
5 5 5 5
6 3 7 2
7 7 7 7
1 1 1 1
2 2 7 2
1 3 3 3
4 4 4 5
5 5 5 5
6 5 6 4
7 4 7 7
```

4.3 Appendix C - Fiktive Datensätze für eine Regression mit Differenzenvariablen

4.3.1 Datensatz mit einer eher realistischen Struktur

<u>y1</u>	<u>y2</u>	<u>x1</u>	<u>x2</u>	<u>z</u>
-1,65190	-1,48008	-,35277	-1,53310	1,30319
-1,51470	-1,34354	-1,65398	-1,39523	-1,58728
-1,37750	-1,07046	-1,50940	-1,25736	-1,44964
-1,24030	-1,07046	1,23759	-1,11949	-1,31200
-1,10310	-1,20700	-1,22024	-,98163	-1,31200
-,96590	-,79739	-1,07567	-,98163	-1,03672
-,82869	-,66085	-,93109	-,70589	-,21087
-,69149	-,52431	-,78651	-,56802	-,76143
-,55429	-,38777	-1,36482	-,43015	-,62379
-,41709	-,25123	-,49735	-,29228	-,48615
-,27989	-,11469	-,35277	-,15441	-,34851
,40612	,02185	-,20819	-,01654	-,21087
,54332	-,38777	-,06361	,12132	-,34851
,13171	-,93392	,08096	,25919	-1,17436
,26891	,43146	,22554	,39706	,20206
,40612	,56800	,37012	,53493	,33970
,54332	,70454	,51470	,67280	,47734
,68052	,84108	,65928	-1,25736	,61498
,81772	,97762	,80386	,94854	,75262
-,00549	1,11416	,94844	1,08641	,89026
1,09212	1,25069	1,09302	1,22427	1,02791
1,22932	1,38723	-,49735	1,36214	1,16555
1,36652	-,52431	1,38217	1,50001	1,06920
1,50372	1,66031	1,52675	1,63788	1,44083
1,64092	1,79685	1,67133	,94854	1,57847

Abbildung 4.2: Fiktiver Datensatz 2

Die Zahlen 1 und 2 hinter x und y stehen für 1. und 2. Zeitpunkt.

4.3.2 Datensatz mit einer eher unrealistischen Struktur

<u>y1</u>	<u>x1</u>	<u>z</u>	<u>y2</u>	<u>x2</u>
1,00	1,00	,00	3,00	3,00
2,00	1,00	1,00	6,00	5,00
3,00	4,00	-1,00	9,00	10,00
4,00	4,00	,00	12,00	12,00
5,00	3,00	2,00	15,00	13,00
6,00	5,00	1,00	18,00	17,00
7,00	10,00	-3,00	21,00	24,00
8,00	11,00	-3,00	24,00	27,00
9,00	8,00	1,00	27,00	26,00
10,00	7,00	3,00	30,00	27,00
11,00	6,00	5,00	33,00	28,00
12,00	7,00	5,00	36,00	31,00
13,00	8,00	5,00	39,00	34,00
14,00	8,00	6,00	42,00	36,00
15,00	9,00	6,00	45,00	39,00
16,00	11,00	5,00	48,00	43,00
17,00	10,00	7,00	51,00	44,00
18,00	10,00	8,00	54,00	46,00
19,00	9,00	10,00	57,00	47,00
20,00	17,00	3,00	60,00	57,00

Auch hier stehen die Zahlen hinter den Variablen für den 1. bzw. 2. Zeitpunkt

Abbildung 4.3: Fiktiver Datensatz zum selber Rechnen

4.4 Appendix D - Datensatz zur Berechnung eines FEM

4.4.1 Ursprungsdatensatz

Person	Zeitpunkt	Variable Y	Variable X
A	1	60	53
A	2	61	51
A	3	62	50
B	1	71	55
B	2	73	54
B	3	70	57
C	1	86	64
C	2	84	69
C	3	83	70
D	1	95	66
D	2	97	64
D	3	100	63

Tabelle 4.1: Ursprungsdatensatz für ein FEM

Variable X ist gemessen in Kilogramm, die Werte von Variable Y basieren auf einer vom Autor erfundenen metrischen Skala von „0= Tagesration des Essens enthielt kein Fett“ bis „100= Tagesration des Essens bestand nur aus Fett“.

4.4.2 Datensatz mit Dummy-Variablen

P	Z	Y	X	D1	D2	D3
A	1	60	53	1	0	0
A	2	61	51	1	0	0
A	3	62	50	1	0	0
B	1	71	55	0	1	0
B	2	73	54	0	1	0
B	3	70	57	0	1	0
C	1	86	64	0	0	1
C	2	84	69	0	0	1
C	3	83	70	0	0	1
D	1	95	66	0	0	0
D	2	97	64	0	0	0
D	3	100	63	0	0	0

Tabelle 4.2: Datensatz für ein FEM inkl. Dummy-Variablen

mit

P=Person

Z=Zeitpunkt

Y, X stehen für die inhaltlichen Variablen

D1-D3 stehen für die drei Dummy-Variablen

4.5 Appendix E - Korrelations- und Kovarianzmatrix der Indikatoren für ein Strukturgleichungsmodell

KORRELATIONSMATRIX

	schr2001	spre2001	schr2003	spre2003	s.a.D.f.2001	s.a.D.f.2003
schr2001	1					
spre2001	0.771	1				
schr2003	0.81	0.67	1			
spre2003	0.734	0.76	0.79	1		
s.a.D.f.2001	0.38	0.43	0.36	0.4	1	
t s.a.D.f.2003	0.38	0.38	0.38	0.4	0.65	1

Abbildung 4.4: Die Korrelationsmatrix der sechs Indikatoren

KOVARIANZMATRIX

	schr2001	spre2001	schr2003	spre2003	s.a.D.f.2001	s.a.D.f.2003
schr2001	1.436					
spre2001	0.888	.922				
schr2003	1.178	.7878	1.47			
spre2003	0.847	.7043	.9365	.9274		
s.a.D.f.2001	0.49	.445	.471	.4193	1.15	
t s.a.D.f.2003	0.49	.390	.494	.4262	.733	1.12

Abbildung 4.5: Die Kovarianzmatrix der sechs Indikatoren

4.6 Appendix F - Modellzeilen in LISREL für verschiedene Varianten von Strukturgleichungsmodellen

Die Reihenfolge der manifesten Variablen in LISREL wurde wie folgt festgelegt:

schr2001 spre2001 schr2003 spre2003 s.a.D.f.01 s.a.D.f.03

MODELLVARIANTE A

SE
6 1 2/
MO NY=1 NX=2 NE=1 NK=1 BE=FU,FI GA=FU,FR LX=FU,FR
LY=FU,FR TE=FU,FI TD=DI
ST 1.0 LY(1,1)

MODELLVARIANTE B

SE
5 6 1 2 3 4/
MO NY=2 NX=4 NE=2 NK=2 BE=FU,FI GA=FU,FI LX=FU,FI
LY=FU,FI TE=FU,FI TD=DI
ST 1.0 LY(1,1) LY(2,2)
FR LX(1,1) LX(2,1) LX(3,2) LX(4,2)
FR GA(1,1)
EQ GA(1,1) GA(2,2)

MODELLVARIANTE C

SE
6 1 2 5/
MO NY=1 NX=3 NE=1 NK=2 BE=FU,FI GA=FU,FR LX=FU,FI
LY=FU,FR TE=FU,FI TD=DI
ST 1.0 LY(1,1)

FR LX(1,1) LX(2,1)
ST 1.0 LX(3,2)

MODELLVARIANTE D

SE
5 6 1 2 3 4/
MO NY=2 NX=4 NE=2 NK=2 BE=FU,FI GA=FU,FI LX=FU,FI
LY=FU,FI TE=FU,FI TD=DI
ST 1.0 LY(1,1) LY(2,2)
FR LX(1,1) LX(2,1) LX(3,2) LX(4,2)
FR GA(1,1) GA(2,1)
EQ GA(1,1) GA(2,2)
FR BE(2,1)

MODELLVARIANTE E

SE
5 6 1 2 3 4/
MO NY=2 NX=4 NE=2 NK=2 BE=FU,FI GA=FU,FI LX=FU,FI
LY=FU,FI TE=FU,FI TD=FU,FI
ST 1.0 LY(1,1) LY(2,2)
FR LX(1,1) LX(2,1) LX(3,2) LX(4,2)
FR GA(1,1) GA(2,1)
EQ GA(1,1) GA(2,2)
FR BE(2,1)
FR TD(1,1) TD(2,2) TD(3,3) TD(4,4) TD(3,1) TD(4,2)
FR TE(1,1) TE(2,2) TD(1,2)

4.7 Kurzer Verweis auf Grundlagen der linearen (Regressions-)Analyse

In einführenden Statistik-Veranstaltungen wird im Kontext der Einführung in die Regressionsanalyse oft der Schwerpunkt auf die **Berechnung von Koeffizienten einer einfachen Regressionsgleichung** gelegt. Dies ist schließlich der erste Schritt zum differenzierten Verständnis von statistischen linearen Modellen. Allerdings sollte man nicht nach diesem Schritt stehen bleiben.

Die Errechnung von Regressionsgleichung erfolgt i.d.R. mit moderner Software augenblicklich. Dies ist komfortabel. Aber die eigentlich wesentliche Arbeit des Forschers **beginnt erst an dieser Stelle**. Es muss nämlich geprüft werden, inwieweit das Modell „gelingen“ ist.

Hierzu wird sowohl das Modell als Ganzes, als auch einzelne Koeffizienten, auf deskriptiver und inferenzstatistischer Ebene, auf ihre Güte (Unverzerrtheit, Effizienz, Erwartungstreue etc.) untersucht.

Ferner sind in linearen Modellen einige Annahmen / Voraussetzungen implizit eingebaut, deren Kenntnis oftmals stillschweigend vorausgesetzt wird. An dieser Stelle sollen einige solcher Voraussetzungen kurz aufgelistet werden – denn nur ihre Kenntnis erlaubt dem Forscher die Überprüfung, ob diese für sein Modell auch erfüllt sind.

Zuerst sei die allgemeine Formel einer multivariaten Regressionsgleichung mit J unabhängigen Variablen gegeben:

$$y_i = a + \sum_{j=1}^J b_j x_{ji} + e_i \quad (4.1)$$

mit

i = Index einzelner Objekte

j = Index einzelner unabhängiger Variablen

y = abhängige Variable

x_j = unabhängige Variable

a = Regressionskonstante

b_j = Regressionskoeffizient der j -ten Variablen

e = Residual- bzw. Fehlervariable

Folgende Annahmen müssen u.a. erfüllt werden:

1. **Linearität in den Parametern:** In linearen Modellen müssen die Parameter rein additiv verknüpft sein; nicht-lineare Zusammenhänge *zwischen den Variablen* hingegen lassen sich oft durch geeignete Transformationen linearisieren
2. **Erwartungswert der Residuen $\mu = 0$:** Sonst können systematische Messfehler vorliegen oder wichtige Einflussgrößen im Modell nicht berücksichtigt worden sein
3. **Normalverteilung der Residuen:** Wird für inferenzstatistische Techniken vorausgesetzt
4. **Homoskedastizität:** Residuen müssen eine konstante Varianz (z.B. in Teilgruppen) aufweisen; diese Annahme kann vor allem bei Längsschnittdaten gefährdet sein, wenn Varianzen der Residuen zu verschiedenen Zeitpunkten verglichen werden
5. **Keine Autokorrelation der Residuen:** Ebenfalls ein großes Problem von Längsschnittdaten, da eine korrelative Beziehung von Residuen gleicher Objekte zu verschiedenen Zeitpunkten wahrscheinlich ist
6. **Keine Korrelation zwischen Residuen und unabhängigen Variablen:** Eine Verletzung dieser Annahme deutet auf das Vorhandensein wichtiger nicht-modellierter Einflussgrößen (also auf eine Modellfehlspezifikation) hin, welche mit in das Modell aufgenommenen Variablen korrelieren
7. **Keine zu hohe Multikollinearität:** Unabhängige Variablen dürfen untereinander nicht zu starke korrelative Beziehungen erster und weiterer Ordnung aufweisen

Diese Auflistung erhebt nicht den Anspruch, erschöpfend zu sein und in die Tiefe zu gehen. Sie dient lediglich einer ersten Orientierung.

Für eine Vertiefung in diese Thematik – sowohl bezogen auf die Theorie als auch auf Techniken zur Aufdeckung und Behebung von Voraussetzungsverletzungen – sei der Leser auf wertvolle Literaturhinweise verwiesen: **von Auer (2007) / Kohler (2008: Kap. 8.3) / Backhaus (2006: 78-94)**.

Einem Leser, welcher noch „vor dem ersten Schritt“ steht, also keine Vorkenntnisse zur Regressionsanalyse mit sich bringt, sei einleitend **Benninghaus (2005: Kap. 7) und Backhaus (2006: 46-78)** empfohlen.

Abbildungsverzeichnis

2.1	Korrelationsmatrix	13
2.2	Kovarianzmatrix	14
2.3	Beispiel für ein Pfaddiagramm	20
2.4	Pfaddiagramm des Beispiels	21
2.5	Pfaddiagramm eines Ein-Indikatoren-Modells	34
2.6	Veranschaulichung eines Strukturgleichungsmodells	99
2.7	Grundgerüst	101
2.8	Modellvariante A	103
2.9	Modellvariante B	104
2.10	Modellvariante C	105
2.11	Modellvariante D	107
2.12	Modellvariante E	108
2.13	Pfaddiagramm zum Beispiel	116
4.1	Fiktiver Datensatz 1	119
4.2	Fiktiver Datensatz 2	121
4.3	Fiktiver Datensatz zum selber Rechnen	122
4.4	Die Korrelationsmatrix der sechs Indikatoren	125
4.5	Die Kovarianzmatrix der sechs Indikatoren	125

Tabellenverzeichnis

2.1	Verschiedene Korrelationsstrukturen	25
2.2	Korrelationsmatrix für eine Pfadanalyse	28
2.3	Ergebnisse Pfadanalyse	29
2.4	Übersicht über die Eigenschaften vorgestellter Modelle	37
2.5	Werte einer Variablen von Objekten zu verschiedenen Zeit- punkten	45
2.6	Verbindung zwischen Indikatoren und latenten Variablen	101
4.1	Ursprungsdatensatz für ein FEM	123
4.2	Datensatz für ein FEM inkl. Dummy-Variablen	124

Literaturverzeichnis

- [1] ARMINGER, G. ; MÜLLER, F. : *Lineare Modelle zur Analyse von Panel-daten*. Westdeutscher Verlag, 1990
- [2] AUER, L. v.: *Ökonometrie. Eine Einführung*. Springer Verlag, 2007
- [3] BACKHAUS, K. ; AL. et: *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 11. Auflage*. Springer Verlag, 2006
- [4] BENNINGHAUS, H. : *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler. 10. Auflage*. VS Verlag, 2005
- [5] ENGEL, U. ; REINECKE, J. : *Panelanalyse. Grundlagen - Techniken - Beispiele*. Walter de Gruyter Verlag, 1994
- [6] ESSER, H. : Inklusion, Integration und ethnische Schichtung. In: *Journal für Konflikt- und Gewaltforschung* (1999), Nr. 1, S. 5–34
- [7] FAULBAUM, F. : Panelanalyse im Überblick. In: *ZUMA-Nachrichten* (1988), Nr. 23, S. 26–44
- [8] FREES, W. : *Longitudinal and Panel Data*. University Press, 2004
- [9] HSIAO, C. : *Analysis of Panel Data. Second Edition*. University Press, 2003
- [10] JÖRESKOG, K. ; SÖRBOM, D. : *Structural Equation Modeling with the SIMPLIS Command Language*. SSI, 1993
- [11] KESSLER, R. ; GREENBERG, D. : *Linear Panel Analysis. Models of Quantitative Change*. Academic Press, 1981

- [12] KOHLER, U. ; KREUTER, F. : *Datenanalyse mit STATA. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung.* Oldenbourg Verlag, 2008
- [13] LONG, J. : *Covariance Structure Models. An Introduction to LISREL.* Sage Publications, 1983
- [14] OPP, K. ; SCHMIDT, P. : *Einführung in die Mehrvariablenanalyse. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen.* Rowohlt, 1976
- [15] PFEIFER, R. ; SCHMIDT, A. : *LISREL. Die Analyse komplexer Strukturgleichungsmodelle.* Gustav Fischer Verlag, 1987
- [16] RABE-HESKETH, S. : *Multilevel and longitudinal modeling using Stata.* StataCorp LP, 2008
- [17] REINECKE, J. : *Strukturgleichungsmodelle in den Sozialwissenschaften.* Oldenbourg Verlag, 2005
- [18] SCHNELL, R. ; HILL, P. ; ESSER, E. : *Methoden der empirischen Sozialforschung. 7. Auflage.* Oldenbourg, 2005
- [19] WEEDE, E. ; JAGODZINSKI, W. : Einführung in die konfirmatorische Faktorenanalyse. In: *Zeitschrift für Soziologie* 6 (1977), Nr. 3, S. 315–333