Einführung in die Analyse mit Kovarianzstrukturmodellen

Prof. Dr. Petra Stein
Marc-André Nehrkorn-Ludwig

13. Dezember 2007

Inhaltsverzeichnis

1 Einleitung

Ziel dieser Ausarbeitung ist eine Einführung in die multivariate Datenanalyse durch Kovarianzstrukturmodelle. Hierzu wird in diesem ersten Kapitel zunächst der Begriff erklärt und ein kurzer historischer Überblick über die bisherige Entwicklung dieser Analysetechniken gegeben. Abgerundet wird dieser Abschnitt durch Vorüberlegungen zum weiteren Vorgehen.

1.1 Was sind Kovarianzstrukturmodelle?

In den Sozialwissenschaften werden häufig Hypothesen über nicht direkt beobachtbare Sachverhalte formuliert. Und ebenso häufig vermutet man Zusammenhänge zwischen diesen latenten Konstrukten sowie anderen manifesten Variablen. Ein Soziologe vermutet beispielsweise einen Zusammenhang zwischen der Ausländerfeindlichkeit und dem Erwerbsstatus einer Person. Während der Erwerbsstatus einfach erfragt werden kann, also direkt beobachtbar ist, stellt Ausländerfeindlichkeit ein latentes Konstrukt dar. Dieses muss über eine Reihe von Indikatoren empirisch messbar gemacht werden, etwa über Fragen zur Einstellung der Person gegenüber verschiedenen ethnischen Gruppen.
Kovarianzstrukturmodelle erlauben nun die statistische Überprüfung derartiger Beziehungen. Ausgehend von der inhaltlichen Problemstellung und Hypothesenformulierung kann ein entsprechendes Modell in linearer Gleichungsform aufgestellt werden, das die vermuteten Beziehungen wiedergibt. Diese können dann anhand empirischer Daten überprüft und getestet werden. Damit grenzt sich dieses Vorgehen von explorativen Verfahren ab.[1] Da metrisches oder mindestens ordinales Skalenniveau Voraussetzung für den Gebrauch derartiger Modelle ist, können gerichtete Beziehungen zwischen den Variablen formuliert werden. Ausgangspunkt für eine Analyse derartiger Beziehungen sind Varianzen, Kovarianzen oder auch Korrelationen.[2] Diese werden in entsprechenden Matrizen erfasst. Damit basiert die Analyse auf den Zusammenhängen zwischen den einzelnen manifesten Indikatoren. Überprüft wird nun, inwieweit sich die empirische Kovarianzmatrix mit der theoretischen, also der durch das Modell geschätzten Matrix, reproduzieren lässt: „The goal of SEM analysis is to determine the extent to which the theoretical model is supported by sample data.“[3] Mit der Differenzierung von manifesten und latenten Variablen lassen sich im Modell eine Mess- und eine Strukturebene unterscheiden. Erstere dient der Modellierung der theoretischen Konstrukte (zum Beispiel Ausländerfeindlichkeit) und zweitere der Abbildung der Beziehungen zwischen den latenten Variablen.
Neben dem Begriff Kovarianzstrukturmodelle bezeichnen Strukturgleichungsmodelle oder auch Kovarianzstrukturanalyse den selben Sachverhalt. In der englischsprachigen Literatur findet man den Begriff structural equation modelling, der häufig mit SEM abgekürzt wird.

[1] Vgl. Reinecke (2005): 3-5
[2] siehe hierzu Abschnitt 3
[3] Schumacker und Lomax (2004): 2

1.2 Geschichtliche Entwicklung

Wichtige Voraussetzung der Kovarianzstrukturanalyse war die, auf der von Pearson entwickelten Produktmomentkorrelation (1896) beruhende, Regressionsanalyse. Mit dieser konnten lineare Beziehungen zwischen einer abhängigen manifesten Variablen und mehreren unabhängigen manifesten Variablen untersucht und statistisch getestet werden.[4]
Eine Weiterentwicklung von Regressionsmodellen stellt die Pfadanalyse dar. Diese wurde von dem Genetiker Wright in den 20er und 30er Jahren des 20. Jahrhunderts eingeführt. Auf diese Weise wurden durch die Verbindung mehrerer Regressionsgleichungen Modellierungen von komplexeren Beziehungen zwischen manifesten Variablen möglich. Insbesondere die Berechnung direkter und indirekter Effekte sowie die Möglichkeit, diese Gleichungen simultan zu lösen, sind wichtige Aspekte. Jedoch wurde diese Technik erst in den 50er und 60er Jahren für die Ökonometrie und die empirische Sozialforschung wiederentdeckt.[5, 6]
Die konfirmatorische Faktorenanalyse[7] hat ihre Ursprünge zu Beginn des 20. Jahrhunderts. Aufbauend auf Pearson's Korrelationskoeffizent entwickelte Spearman (1904, 1927) Methoden zur Berechnung von Faktorenmodellen. Dabei wurde unterstellt, dass eine Menge von Items, die miteinander stark korrelieren, einen gemeinsamen unterliegenden Faktor haben. Durch Arbeiten von Howe, Anderson und Rubin sowie Lawley in den 50er Jahren wurde die konfirmatorische Faktorenanalyse entwickelt. In den folgenden Jahren leistete insbesondere Jöreskog einen wesentlichen Beitrag zu deren Weiterentwicklung. Ein wichtiges Ereignis, insbesondere für die weitere Verbreitung dieser Analysetechnik, dürfte die Entwicklung der ersten Software zur Berechnung derartiger Modelle in der Folgezeit gewesen sein.[8]
Kovarianzstrukturmodelle verbinden die Pfad- und Faktorenanalyse miteinander und ermöglichen damit die Analyse der Beziehungen zwischen latenten Variablen. Jöreskog (1973), Keesling (1972) und Wiley (1973) leisteten hierzu wesentliche Beiträge. Mit der durch Jöreskog und van Thillo entwickelten LISREL-Software – *L*inear *S*tructural *REL*ations model – und weiteren inhaltlichen Modifikationen, etwa der Einführung der Simplis-Notation, mit der die Programmnutzung vereinfacht werden konnte, wurde dieser Ansatz einer breiteren wissenschaftlichen Öffentlichkeit zugänglich gemacht. Heute sind neben Lisrel auch andere Programme wie zum Beispiel Amos, EQS und Mplus verbreitet.[9, 10]
Seit dem Aufkommen dieser Methoden wurden sie stetig verfeinert und verbessert. Hierzu zählen unter anderem alternative Schätzverfahren, Techniken zum Gruppenvergleich, Längsschnittanwendungen sowie Verfahren zur Berücksichtigung fehlender Werte.[11]

[4] Vgl. Schumacker und Lomax (2004): 5
[5] Vgl. Schumacker und Lomax (2004): 5-6
[6] Vgl. Reinecke (2005): 7-8
[7] im Englischen *confirmatory factor analysis* (CFA)
[8] Vgl. Schumacker und Lomax (2004): 5
[9] Vgl. Schumacker und Lomax (2004): 6
[10] in den Beispielen dieser Ausarbeitung wird Lisrel verwendet, eine eingeschränkte Studentenversion kann über die Internetseite des Herstellers kostenlos bezogen werden: `http://www.ssicentral.com/lisrel/student.html`
[11] Vgl. Reinecke (2005): 14-18

1.3 Einige Vorüberlegungen

Bevor die intensiver auf die statistische Materie eingegangen werden soll, wird an dieser Stelle auf einige wichtige Aspekte in der Arbeit mit Kovarianzstrukturmodellen hingewiesen.

1. *Kausalität*: Oft wird einer Beziehung zwischen zwei oder mehreren Variablen eine kausale Interpretation gegeben, ohne das die statistischen Modelle diese nahe legen. Daher sollte mit dem Begriff der Kausalität entsprechend vorsichtig umgegangen werden, zumal er an verschiedene Voraussetzungen gebunden ist: (a) den theoretischen Begründungszusammenhang, (b) den empirischen Zusammenhang, (c) die zeitliche Asymmetrie der Variablen und (d) den Ausschluss von Einflüssen durch Drittvariablen.[12]

2. *Modellspezifikation*: Vor dem Einsatz der entsprechenden mathematischen bzw. statistischen Verfahren muss das Modell theoretisch hergeleitet werden. Diese Überlegungen umfassen sowohl die Menge an Variablen, die in das Modell aufgenommen werden soll, als auch die Art und Richtung der Beziehungen der einzelnen Variablen untereinander. Mit diesem theoriegeleiteten Gerüst versucht der Forscher den datengenerierenden Prozess möglichst gut zu erklären. Durch verschiedene Verfahren kann dann die Abweichung der theoretischen Kovarianzmatrix von der empirischen getestet werden. Ein Modell ist zum Beispiel fehlspezifiziert, wenn wesentliche Variablen ausgelassen wurden. Hierdurch kann das Modell systematisch verzerrt werden.[13]

3. *Modellidentifikation*: Das Identifikationsproblem bezieht sich auf die Lösbarkeit der Modellgleichungen. Ein einfaches Beispiel erläutert die Problematik: Man betrachte die Gleichung $X+Y=Z$. Nun wird angenommen, dass $Z=10$ sei. Eine eindeutige Lösung der Gleichung kann nun aber nicht mehr gegeben werden, da $X=3$ und $Y=7$, aber auch $X=6$ und $Y=4$ sein könnte. Übertragen auf Kovarianzstrukturmodelle bedeutet dies, dass bspw. die Parameter nicht mehr zu ermitteln sind, die die Beziehungen zwischen den latenten Variablen angeben. Es können dann Beschränkungen eingeführt werden, so zum Beispiel, dass eine Voruntersuchung ergeben hat, das $X=2$ ist. Nun ist das Modell mit $Y=8$ eindeutig lösbar. „If a structural equation model is not identified, an infinte number of sets of parameters could generate the observed variables.“[14] Demnach kann ein Parameter entweder zur Schätzung freigegeben, auf einen bestimmten Wert fixiert oder auch einem anderen Parameter werden. Man unterscheidet (a) *unteridentifizierte* bzw. *nicht identifizierte*, (b) *gerade noch identifizierte* und (c) *überidentifizierte* Modelle. Unter (a) fallen solche, in denen mindestens ein Parameter und dementsprechend das gesamte Modell

[12] Vgl. Reinecke (2005): 12
[13] Vgl. Schumacker und Lomax (2004): 62-63
[14] Long (1983a): 36

nicht identifiziert sind. (b) und (c) beschreiben identifizierte Fälle, wobei versucht wird (c) zu erreichen, da hier mehr Parameter unfrei sind, als frei zu schätzen und daher statistische Tests möglich werden. Es stehen verschiedene Verfahren zur Verfügung die Identifikation eines Modells zu überprüfen. Eine notwendige aber nicht hinreichende Bedingung ist die „order condition"und eine hinreichende, die „rank condition".[15] Reinecke nennt zur Lösung dieser Problematik die t-Regel, nach der folgende Bedingung erfüllt sein muss:

$$t \leq \frac{1}{2}(p+q)(p+q+1) \tag{1}$$

Mit p als Anzahl der unabhängigen Variablen X und q als Anzahl der unabhängigen Variablen Y sowie t als die Anzahl zu schätzender Parameter. Die Differenz zwischen diesen beiden Ausdrücken entspricht der Anzahl der Freiheitsgrade (degress of freedom – *df*) eines Modells.[16] Zudem sollten die Varianzen der latenten Variablen in den Messmodellen entweder durch die Fixierung einer Faktorladung (etwa auf den Wert 1) oder die Fixierung der Varianz selbst auf eine Konstante (in der Regel ebenfalls 1) skaliert werden.[17] Satow nennt darüber hinaus Anzeichen, die auf ein Identifikationsproblem hindeuten können.[18] nicht positivdefinite Matrizen, abnormale Schätzungen und mathematisch nicht zulässige Werte, wie negative Varianzen[19]. Infolgedessen wird dann eine weitere Schätzung empfohlen, deren iterativer Prozess auf anderen Startwerten beruht.

4. *Modellschätzung*: Ziel der Parameterschätzung ist es, dass man Parameter erhält, die die zu Grunde liegende empirische Stichprobenmatrix S mit der modellimplizierten Kovarianzmatrix Σ möglichst gut reproduzieren. Wie bereits erwähnt ist die Lösung des Gleichungssystems nicht mehr analytisch, sondern nur noch iterativ möglich. Das heißt, dass versucht wird, die Differenz zwischen den beiden Matrizen zu minimieren, indem verschiedene Parameterwerte „ausprobiert"werden. Hierzu stehen verschiedene Diskrepanzfunktionen (fitting functions) zur Verfügung, die sich nach ihren Anwendungsvoraussetzungen und statistischen Eigenschaften unterscheiden: unweigthed least squares (ULS), ordinary least squares (OLS), generalised least squares (GLS) sowie maximum likelihood (ML) um nur einige zu nennen. Nähere Informationen zu den Schätzverfahren werden in Kapitel . . . . . gegeben.[20]

5. *Modellüberprüfung*: Bei diesem Aspekt geht es darum, zu fragen, inwieweit die Daten das theoretisch abgeleitete Modell unterstützen. Globale Fitindizes zeigen die Anpassungsgüte des Gesamtmodells. Um diese zu beurteilen stehen zahlreiche Goodness of Fit Indizes zur Verfügung, die in einem späteren Kapitel detailliert dargestellt werden. Zumeist basieren sie auf dem Vergleich der

[15] Vgl. Schumacker und Lomax (2004): 63-66
[16] Vgl. Reinecke (2005): 52-53
[17] Vgl. Reinecke (2005): 102
[18] Vgl. Satow (1999): 8
[19] Vgl. hierzu auch Byrne (1998): 175
[20] Vgl. Schumacker und Lomax (2004): 66

Matrizen S und Σ. Daneben können die geschätzten Parameter direkt untersucht werden. Die Betrachtung des Vorzeichens gibt Aufschluss darüber, ob das Vorzeichen den Erwartungen entspricht. Zudem sollte der Wert in einem bestimmten theoretisch hergeleiteten Intervall liegen. Darüber hinaus bietet sich die Möglichkeit durch Standardfehler und kritische Werte statistische Tests einzelner Parameter durchzuführen und zu prüfen, ob diese signifikant von Null verschieden sind.[21]

6. *Modellmodifikation*: Zeigt das Modell keinen akzeptablen Fit, so muss modifiziert werden. Eine Verbesserung der Modellanpassung ist allerdings nur ein Aspekt, neben diesem können hieraus signifikante und theoretisch gehaltvolle Parameterschätzungen hervorgehen. Über den Ausschluss von nicht signifikanten Parametern sollte mit Vorsicht entschieden werden. Zum Einen ist zu beachten, dass die statistische Signifikanz von der Stichprobengröße abhängig ist und zum Anderen sollten darüber die theoretischen Überlegungen eine bedeutsame Rolle spielen. Eine andere Möglichkeit bietet sich durch die Analyse der (standardisierten) Residualmatrix, die sich aus der Differenz zwischen der empirischen und der modellimplizierten Kovarianzmatrix ergibt. Große Residuen vieler Variablen deuten auf eine Fehlspezifikation des gesamten Modells hin, während einzelne Residuen sich auf mögliche Fehlerquellen in einzelnen Variablen beziehen. Lisrel stellt so genannte Modifikationsindizes bereit, die die Verbesserung der χ^2-Statistik anzeigen, wenn der betreffende Parameter zur Schätzung freigegeben wird. Die hieraus resultierende Veränderung der Parameter wird ebenfalls ausgegeben. Weiterhin geben verschiedene Determinationskoeffizienten R^2 Informationen darüber, inwieweit die Indikatoren zur Bestimmung einer spezifischen latenten Variablen sinnvoll sind. Diese Maße stehen auch für die Strukturgleichungen zur Verfügung. Neben diesen Optionen bietet EQS den Lagrange Multiplier-Test und die Wald-Statistik an. Ersterer lässt sich als multivariater Modifikationsindex auffassen, während zweitere die statistische Sinnhaftigkeit einer Parameterfreisetzung angibt. In zahlreichen Studien konnte keine optimale Strategie zur Modellmodifikation gefunden werden. Es empfiehlt sich daher ein Modell zunächst theoretisch herzuleiten, es zu testen und zuerst das Mess- und dann das Strukturmodell zu analysieren. Hierbei sollten sämtliche Beurteilungskriterien herangezogen werden. Das Programm TRETAD II von Spirtes, Scheines, Meek und Glymour (1994) sowie Amos in der neueren Version bieten Möglichkeiten zur automatisierten Spezifikationssuche.[22]

1.4 Weiteres Vorgehen - Struktur der Arbeit

Wie ein Blick auf die Historie der Kovarianzstrukturmodelle gezeigt hat, liegen die Ursprünge in der Regressionsanalyse, aus der sich dann die Pfadanalyse entwickelte. Durch die Einbeziehung der konfirmatorischen Faktorenanalyse konnten dann latente Variablen, modelliert durch eine Menge von Indikatoren, in die Analyse auf-

[21] Vgl. Schumacker und Lomax (2004): 69-70
[22] Vgl. Schumacker und Lomax (2004): 70-75

genommen werden. Da diese Konzepte in einem inhaltlichen und formalen Zusammenhang stehen, bildet diese Struktur auch das Vorgehen in dieser Arbeit ab. Bevor jedoch auf die Herleitung und Erklärung der einzelnen Methoden eingegangen werden soll, werden mathematische und statistische Grundlagen wiederholt. Im Bereich der Mathematik ist insbesondere auf die lineare Matrixalgebra einzugehen, da diese das grundlegende analytische Instrumentarium bereitstellt. Unter die statistischen Grundlagen fallen einige Bemerkungen zum Skalenniveau, statistischen Verteilungen und Zusammenhangsmaßen.
Ergänzend werden zu den methodischen Kapiteln Beispiel präsentiert, um einerseits die Anwendung und den Nutzen der präsentierten Methoden vorzuführen und andererseits auf den Aufbau zunächst einfacher Lisrel-Analysen einzugehen. Sofern nicht anders vermerkt, basieren die Beispiele auf den Daten, die mit der Lisrel-Studenten und -Vollversion installiert werden.

2 Mathematische Grundlagen

Um die folgende Darstellung der statistischen Verfahren besser verstehen zu können bietet es sich an, zunächst mit einer kurzen Einführung in die lineare Algebra zu beginnen, da die präsentierten Methoden auf der Matrizenrechnung basieren. Lineare Algebra ist die Lehre von Gleichungsystemen mit Variablen, die in der ersten Potenz stehen (bspw. x^1.

2.1 Vektoren und Matrizen

Eine Matrix ist ein rechteckiges Zahlenschema mit spezifischen Rechenregeln. In gewisser Weise stellen Matrizen eine Verallgemeinerung von Vektoren da, was im Folgenden ersichtlich wird. Eine Matrix besteht aus m Zeilen und n Spalten, man spricht von einer (m,n)-Matrix.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

Durch a wird das jeweilige Element, durch den ersten Subindex m die jeweilige Zeile und durch den zweiten Subindex n die jeweilige Spalte angezeigt. Auf diese Weise ist jede Position innerhalb der Matrix eindeutig zu bestimmen. Ein Spaltenvektor ist demnach eine Matrix mit mehreren Zeilen, aber nur einer Spalte, also eine (m,1)-Matrix. Umgekehrt ist ein Zeilenvektor eine Matrix mit nur einer Zeile, aber mehreren Spalten, man könnte auch sagen (1,n)-Matrix.[23]

[23] Dörsam (2003): 26-27

2.2 Rechenregeln

In diesem Abschnitt werden einige wichtige Rechenregeln beschrieben und mit Beispielen verdeutlicht.

2.2.1 Addition

Matrizen werden addiert, indem man die jeweiligen Elemente addiert. Das bedeutet auch, dass die zusammenzurechnenden Matrizen die gleiche Anzahl an Zeilen und Spalten aufweisen müssen.

$$\begin{pmatrix} 4 & 3 & -1 \\ 8 & 1 & 5 \\ -5 & 2 & 11 \end{pmatrix} + \begin{pmatrix} 2 & 1 & 4 \\ 7 & -3 & 9 \\ 3 & 8 & 5 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 3 \\ 15 & -2 & 14 \\ -2 & 10 & 16 \end{pmatrix} \tag{2}$$

2.2.2 Skalarmultiplikation

Eine Matrix wird mit einem Skalar, also einer reellen Zahl, multipliziert, in dem jedes einzelne Element der Matrix mit der Zahl multipliziert wird. Beispielsweise:

$$a * \begin{pmatrix} 4 & 3 & -1 \\ 8 & 1 & 5 \\ -5 & 2 & 11 \end{pmatrix} = \begin{pmatrix} 4*a & 3*a & -1*a \\ 8*a & 1*a & 5*a \\ -5*a & 2*a & 11*a \end{pmatrix} \tag{3}$$

a kann hierbei zum Beispiel den Wert 1 oder -422 annehmen. In jedem Fall aber ist $a \in \Re$.

2.2.3 Transposition

Eine weitere wichtige Operation im Zusammenhang mit Matrizen und Vektoren ist die Transposition. Man erhält eine transponierte Matrix, in dem man die Zeilen mit Spalten vertauscht. Das Element a_{ij} wird dann zum Element $a_(ji)$, wie am folgenden Beispiel zu sehen ist.

$$\mathrm{A} = \begin{pmatrix} 1 & -2 & 7 \\ 4 & 9 & -1 \\ -3 & 8 & 2 \\ 6 & 3 & 5 \end{pmatrix} \qquad A^T = \begin{pmatrix} 1 & 4 & -3 & 6 \\ -2 & 9 & 8 & 3 \\ 7 & -1 & 2 & 5 \end{pmatrix} \tag{4}$$

Ebenso ist zu erkennen, dass es sich bei A um eine (4,3)-Matrix handelt und bei A^T um eine (3,4)-Matrix. Man kennzeichnet transponierte Matrizen durch ein hochgestelltes T A^T oder auch durch ein Apostroph A'.

2.2.4 Matrizenmultiplikation

Zunächst wird das Produkt zweier Vektoren betrachtet. Hier spricht man von einem Skalarprodukt, was nicht mit der Skalarmultiplikation zu verwechseln ist. Skalarprodukt deshalb, weil das Produkt zweier Vektoren eine Zahl ist. Multipliziert man

zwei Vektoren miteinander so bildet man das Produkt der jeweiligen Elemente und addiert diese anschließend auf.

$$\begin{pmatrix} 1 \\ 4 \\ -3 \\ 6 \end{pmatrix} \quad * \quad \begin{pmatrix} 1 \\ -2 \\ 7 \\ 3 \end{pmatrix} \quad = \quad 1 * 1 + 4 * (-2) + (-3) * 7 + 6 * 3 = -22 \tag{5}$$

Zu bachten ist, dass die Anzahl der Zeilen in beiden Vektoren die gleiche ist.

Die Multiplikation zweier Matrizen lässt sich anhand des *Falkschen Schemas* verdeutlichen. Hierzu werden zwei zu mulitplizierende Matrizen in einem Schema diagonal angeordnet, dann entsteht das Produkt unterhalb der Diagonalen. Im folgenden Beispiel werden zwei Matrizen miteinander multipliziert.

$$A = \begin{pmatrix} 4 & 3 & -1 \\ 8 & 1 & 5 \\ -5 & 2 & 11 \end{pmatrix} \qquad B = \begin{pmatrix} 2 & 1 \\ 7 & -3 \\ 3 & 8 \end{pmatrix} \tag{6}$$

$A * B =$

			2	1
			7	-3
			3	8
4	3	-1		
8	1	5		
-5	2	11		

(7)

			2	1
			7	-3
			3	8
4	3	-1	$4 * 2 + 3 * 7 + (-1) * 3 = 26$	$4 * 1 + 3 * (-3) + (-1) * 8 = -13$
8	1	5	$8 * 2 + 1 * 7 + 5 * 3 = 38$	$8 * 1 + 1 * (-3) + 5 * 8 = 45$
-5	2	11	$(-5) * 2 + 2 * 7 + 11 * 3 = 37$	$(-5) * 1 + 2 * (-3) + 11 * 8 = 77$

Die Multiplikation erfolgt also dadurch, dass man die einzelnen Zeilen und Spalten wie Vektoren miteinander multipliziert. Dies ist jedoch nur dann möglich, wenn die zweite Matrix, in diesem Fall B, genauso viele Zeilen wie die erste Matrix (A) Spalten hat. Es ergibt sich also:

$A * B =$

			2	1
			7	-3
			3	8
4	3	-1	26	-13
8	1	5	38	45
-5	2	11	37	77

$$= \begin{pmatrix} 26 & -13 \\ 38 & 45 \\ 37 & 77 \end{pmatrix} \tag{8}$$

Multipliziert man eine Matrix A mit einer *Einheitsmatrix* (I) so ist das Produkt wiederum die Ausgangsmatrix A. In einer Einheitsmatrix stehen entlang der Hauptdiagonalen nur Einsen und alle anderen Element sind Null. Hier ist beispielhaft eine

(3,3)-Einheitsmatrix dargestellt.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{9}$$

Sie ist folglich das neutrale Element der Matrizenmultiplikation. Im Gegensatz zur „normalen "Multiplikation gilt bei den Matrizen das Kommutativgesetz nicht. Demnach ist $A * B \neq B * A$.

2.3 Determinanten

Determinanten werden in der Statistik insbesondere für die Berechnung der Inversen von Matrizen und bei der Überprüfung auf lineare Abhängigkeit[24] von Vektoren genutzt. Determinanten lassen sich nur für quadratische Matrizen berechnen. Es ist dabei relativ einfach die Determinante für eine (2,2)- oder eine (3,3)-Matrix zu berechnen.
Für eine (2,2)-Matrix werden hierzu nur die Elemente entlang der Hauptdiagonalen multipliziert und von diesem Produkt das Produkt der Elemente der Nebendiagonalen abgezogen.

$$A = \begin{pmatrix} 4 & 2 \\ 5 & 9 \end{pmatrix} \qquad detA = det \begin{pmatrix} 4 & 2 \\ 5 & 9 \end{pmatrix} = \begin{vmatrix} 4 & 2 \\ 5 & 9 \end{vmatrix} \tag{10}$$

$$detA = 4 * 9 - 5 * 2 = 26 \tag{11}$$

Zur Berechnung der Determinanten von (3,3)-Matrizen kann die Sarrus'sche Regel angewendet werden, nach der die ersten beiden Spalten einer Matrix einfach hinten an die Determinante „angehangen "werden. Dann verfährt man wie im obigen Fall, nur das es jetzt die Produkte der Elemente der Hauptdiagonalen addiert werden und man von diesen die Produkte der Elemente der Nebendiagonalen abzieht.

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 5 & 9 & 4 \\ 3 & -1 & 7 \end{pmatrix}$$

$$detA = det \begin{pmatrix} 4 & 2 & 1 & |4 & 2 \\ 5 & 9 & 4 & |5 & 9 \\ 3 & -1 & 7 & |3 & -1 \end{pmatrix} = \begin{vmatrix} 4 & 2 & 1 & |4 & 2 \\ 5 & 9 & 4 & |5 & 9 \\ 3 & -1 & 7 & |3 & -1 \end{vmatrix} \tag{12}$$

$$detA = 4*9*7+2*4*3+1*5*(-1)-3*9*1-(-1)*4*4-7*5*2 = 190 \tag{13}$$

Für größere Matrizen dient der *Laplace Entwicklungssatz* der hier jedoch nicht nachvollzogen werden soll.
Mit Bezug auf das Problem linearer Abhängigkeit lässt sich festhalten, dass „eine Determinante [...] gerade dann Null [wird, d. V.], wenn ihre Spaltenvektoren

[24] dieser Begriff wird in Kapitel **??** näher behandelt

(und damit ihre Zeilenvektoren) linear abhängig sind.“[25] Eng hiermit verbunden ist der *Rang* einer Matrix, der die Anzahl der linear unabhängigen Zeilen- und damit auch Spaltenvektoren angibt. Eine Matrix besitzt demnach vollen Rang, wenn ihre Determinante ungleich Null ist.

2.4 Inverse Matrizen

Ein letzter wichtiger Punkt sind inverse Matrizen. Ähnlich wie bei der „normalen “Multiplikation gibt es auch in der linearen Algebra Inversionen. Hat man eine Zahl gegeben, so muss aus der Multiplikation dieser mit ihrer Inversen gerade das neutrale Element resultieren. Gegeben ist beispielsweise die Zahl 3, so muss diese mit 3^{-1} bzw. $\frac{1}{3}$ multipliziert werden, um als Ergebnis Eins zu erhalten. Das neutrale Element der Matrixmultiplikation ist die Einheitsmatrix I. Demnach ist die Inverse der Matrix A definiert als:

$$A * A^{-1} = I \tag{14}$$

Man kann jedoch nicht zu jeder Matrix eine Inverse berechnen. Ist nämlich die Determinante einer Matrix gleich Null und sind damit die Zeilen- und auch Spaltenvektoren linear abhängig, dann ist eine Matrix nicht invertierbar. Eine solche Matrix bezeichnet man als *singulär*. Im anderen Fall, linear unabhängiger Vektoren, bezeichnet man die Matrix als *regulär*.
Es existieren verschiedene Verfahren zur Berechnung von Inversen. Hier wird jedoch nur die Bestimmung durch *adjungierte Matrizen* betrachtet und der *Gauß-Algorithmus* außen vor gelassen.
Geht man von der Matrix A aus dem vorherigen Abschnitt aus, von der man weiß, dass sie invertierbar ist (det A = 190), kann man die Inverse durch folgende Formel berechnen:

$$A^{-1} = \frac{1}{detA} * adj(A) \tag{15}$$

Hier wird auch sofort ersichtlich warum die Determinante nicht Null sein darf, da eine Division durch Null nicht definiert ist. Der Ausdruck *adj(A)* steht für die adjungierte Matrix, die im Folgenden berechnet wird. Zunächst muss die Matrix A transponiert werden.

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 5 & 9 & 4 \\ 3 & -1 & 7 \end{pmatrix} \qquad A^T = \begin{pmatrix} 4 & 5 & 3 \\ 2 & 9 & -1 \\ 1 & 4 & 7 \end{pmatrix} \tag{16}$$

Nun wird die entsprechende Vorzeichenmatrix aufgestellt. Sie enthält ebenso viele Zeilen und Spalten, wie die Matrix A^T. Das erste Element (a_{11}) erhält ein positives Vorzeichen, das zweite Element (a_{12}) ein negatives, das nächste wiederum ein positives Vorzeichen, ...

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix} \tag{17}$$

In diese werden nun die Unterdeterminanten eingetragen. Diese werden aus der transponierten Matrix A^T berechnet, indem man die Zeile und Spalte entsprechend dem Subskript der Unterdeterminante streicht.

[25] Dörsam (2003): 66

$$adj(A) = \begin{pmatrix} +det_{11} & -det_{12} & +det_{13} \\ -det_{21} & +det_{22} & -det_{23} \\ +det_{31} & -det_{32} & +det_{33} \end{pmatrix} \tag{18}$$

Es ergibt sich als Unterdeterminante det_{11}, durch das Streichen der ersten Zeile und ersten Spalte von A^T beispielsweise $9 * 7 - 4 * (-1) = 67$. Die Vorzeichen in der Matrix 18 werden so behandelt, als ob multipliziert würde. Plus mal einer positiven Determinante bleibt also positiv, während plus mal einer negativen Determinante negativ wird. Man erhält also als adjungierte Matrix (in den Klammern stehen die Werte der Unterdeterminanten):

$$adj(A) = \begin{pmatrix} +(67) & -(15) & +(-1) \\ -(23) & +(25) & -(11) \\ +(-32) & -(-10) & +(26) \end{pmatrix} = \begin{pmatrix} 67 & -15 & -1 \\ -23 & 25 & -11 \\ -32 & 10 & 26 \end{pmatrix} \tag{19}$$

Für die Inverse ergibt sich nun:

$$A^{-1} = \tfrac{1}{detA} * adj(A) = \tfrac{1}{190} * \begin{pmatrix} 67 & -15 & -1 \\ -23 & 25 & -11 \\ -32 & 10 & 26 \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} 0,35 & -0,08 & -0,01 \\ -0,12 & 0,13 & -0,06 \\ -0,17 & 0,05 & 0,14 \end{pmatrix}$$

Es wurde auf zwei Nachkommastellen gerundet. Multipliziert man diese Inverse mit der Ausgangsmatrix A, erhält man die Einheitsmatrix I.

3 Statistische Grundlagen

Da diese Übersicht Teil einer Hauptstudiumsveranstaltung ist, wird an dieser Stelle nur in aller Kürze auf die notwendigen statistischen Grundlagen eingegangen, ohne dabei einen Anspruch auf Vollständigkeit zu erheben.

3.1 Variablen

„Variablen können als zusammenfassender Begriff für verschiedene Ausprägungen einer Eigenschaft (den „Variablenwert“angesehen werden [...].“[26] Man unterscheidet zwischen *manifesten*, also direkt beobachtbaren und *latenten*, nicht direkt beobachtbaren, Variablen. Letztere müssen durch geeignete Indikatoren operationalisiert werden.

[26] Schnell, Hill, Esser (1999): 124

3.2 Skalenniveau von Variablen

Das Messniveau einer Variablen ist entscheidend für die Anwendbarkeit statistischer Methoden. Im Regelfall setzen die hier vorgestellten Verfahren metrisches Skalenniveau voraus. Durch neuere Entwicklungen ist es jedoch auch möglich besondere Zusammenhangsmaße für nominale, ordinale und metrische Variablen zu berechnen, auf deren Basis dann Modelle geschätzt werden können. Liegt etwa ordinales Messniveau vor, können Kovarianzen und Korrelationen nicht korrekt berechnet werden. Wie im nächsten Abschnitt beschrieben wird, ist es in einem solchen Fall aber möglich, polychorische oder polyserielle Korrelationen zu berechnen.[27]

3.3 Zusammenhang zwischen Variablen

Da Kovarianzstrukturmodelle auf der Grundlage von Kovarianzen und auch Korrelationen berechnet werden, wird an dieser Stelle nochmals auf diese beiden wichtigen statistischen Konzepte eingegangen. „Der Kovarianz zweier Merkmale entnehmen wir, in welchem Ausmaß die Unterschiedlichkeit der Untersuchungsobjekte, bezogen auf das Merkmal x, der Unterschiedlichkeit der Untersuchungsobjekte im Merkmal y entspricht.“[28]

$$cov(x, y) = \frac{\sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \tag{21}$$

Eine positive Kovarianz bedeutet dann einen positiven Zusammenhang zwischen den beiden Merkmalen. Sind sie voneinander unabhängig resultiert eine Kovarianz von Null. Problematisch erscheint jedoch, dass die Kovarianz nicht normiert ist, also keinen einheitlichen Wertebereiche aufweist. Teilt man die Kovarianz durch die Standardabweichungen der beiden Variablen findet eine Normierung auf einen Wertebereich zwischen $[-1, +1]$ statt. Der *Produkt-Moment-Korrelationskoeffizient* berechnet sich demnach wie folgt:[29]

$$r = \frac{cov(x, y)}{s_x \cdot s_y} \tag{22}$$

Der lineare Zusammenhang zwischen zwei Variablen ist dann perfekt positiv, wenn $r = 1$ ist. Entsprechend gilt für einen perfekten negativen linearen Zusammenhang $r = -1$ und bei Unabhängigkeit $r = 0$.
Liegen Variablen vor, die nicht dem metrischen Skalenniveau entsprechen, müssen polychorische, tetrachorische und polyserielle Korrelationen berechnet werden. Sind beide Merkmale ordinalskaliert berechnet man mit PRELIS die polychorische Korrelation. Weist dahingegen eine Variable metrisches und die andere ordinales Skalenniveau auf, wird die polyserielle Korrelation verwendet. Des Weiteren findet die tetrachorische Korrelation Anwendung, wenn beide Variablen dichotom sind. Ohne auf die technische Details einzugehen, wird bei der polychorischen Korrelation davon ausgegangen, dass den beiden ordinalen Variablen metrische Indikatorvariablen zu Grunde liegen. Jede Ausprägung einer Variablen fällt dann in ein bestimmtes

[27] In Lisrel geschieht dies mit dem Programm PRELIS
[28] Bortz (2005): 203
[29] Vgl. Bortz (2005): 205

Intervall dieser Indikatorvariablen, wobei die Intervalle durch zu schätzende Schwellenwerte getrennt werden. Um dies zu ermöglichen, wird angenommen, dass die Indikatorvariablen normalverteilt sind.[30] [31]

3.4 Statistische Verteilungen

Eine der wichtigsten Verteilungsannahmen mit Hinblick auf Kovarianzstrukturmodelle ist die der *Normalverteilung*. Derartige Verteilungen sind u.a. dadurch gekennzeichnet, dass sie glockenförmig und symmetrisch sind und die Lageparameter auf einen Punkt zusammenfallen. Die Verteilung ist dabei abhängig von den Parametern μ, dem Erwartungswert und σ, der Standardabweichung. Ist der Erwartungswert gleich Null und die Streuung gleich Eins, spricht man von einer *Standardnormalverteilung*. Durch eine *z-Transformation* können sämtliche Normalverteilungen in diese überführt werden.[32] Da Kovarianzstrukturmodelle jedoch multivariate Modelle sind müssen zumeist multivariate Normalverteilungen vorliegen, deren Voraussetzung aber eine univariate Normalverteilung der Variablen ist. Bei einer Verletzung dieser Annahme können Modelle nicht mehr zuverlässig geschätzt und interpretiert werden. Daher muss in einem solchen Fall mit alternativen Schätzmethoden oder Normalisierungsverfahren reagiert werden.[33]

4 Regressionsmodelle

„Die Regressionsanalyse bildet eines der flexibelsten und am häufigsten eingesetzten statistischen Analyseverfahren.“[34] Sie ist ein Instrument, mit dem sich die Beziehungen zwischen einer abhängigen und einer bzw. mehreren unabhängigen Variablen analysieren lassen. Dabei unterscheidet man drei generelle Anwendungsbereiche: (1) Ursachenanalyse, (2) Wirkungsprognosen und (3) Zeitreihenanalyse.[35] In der Regel geht man davon aus, das die Variablen metrisches Messniveau aufweisen. Diese Annahme wird jedoch gerade in den Sozialwissenschaften selten erfüllt, so dass man auch binäre und quasi-metrische Variablen in das Regressionsmodell aufnimmt.
Da diese Analysetechnik von besonderer Bedeutung ist, wird im Folgenden zunächst auf den mathematischen Modellrahmen eingegangen und wesentliche Annahmen und Voraussetzungen besprochen, ohne dabei den Anspruch auf Vollständigkeit zu erheben. Ziel ist weniger eine umfassende Darstellung der Regressionsanalyse als vielmehr die Vermittlung des grundlegenden Prinzips dieser Methode.

---

[30] Vgl. Reinecke (2005): 31-35

[31] Vgl. hierzu auch Jöreskog (2002): Structural Equation Modeling with Ordinal Variables using LISREL

[32] Vgl. Bortz (2005): 73-75

[33] Vgl. Reinecke (2005): 111

[34] Backhaus et al. (2006): 46

[35] Vgl. Backhaus et al. (2006): 48

4.1 Das einfache Regressionsmodell

In diesem Kapitel wird das Regressionsmodell anhand des einfacheren Falles der einfachen linearen Regression dargestellt. Dieses lässt sich dann auf den Fall mehrerer unabhängiger Variablen, die multiple Regression, erweitern.
Das Streudiagramm 2 zeigt einen fingierten Datensatz, der zwei Variablen Alter und Einkommen besitzt. Die 15 Beobachtungen legen die Vermutung nahe, dass es sich um eine positive lineare Beziehung handelt. Mit steigendem Alter geht zumindest tendenziell ein höheres Einkommen einher. Wie lässt sich eine solche Beziehung nun

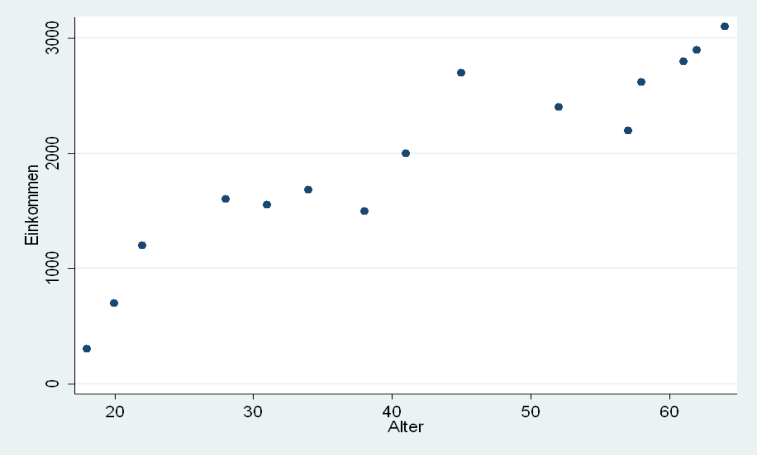

Abbildung 1: Streudiagramm: Alter und Einkommen

statistisch modellieren? Eine Option ist die, eine Gerade in die Punktwolke möglichst gut einzupassen. Dies wird in der folgenden Abbildung dargestellt. Zwei Punkte sind dabei zu erkennen: (1) die Gerade ist stetig und zeigt damit auch Punkte, zu denen keine Beobachtungen vorliegen sowie (2) zwischen der Geraden und den tatsächlichen Beobachtungen bestehen Differenzen. Diese Differenzen sind von entscheiden-

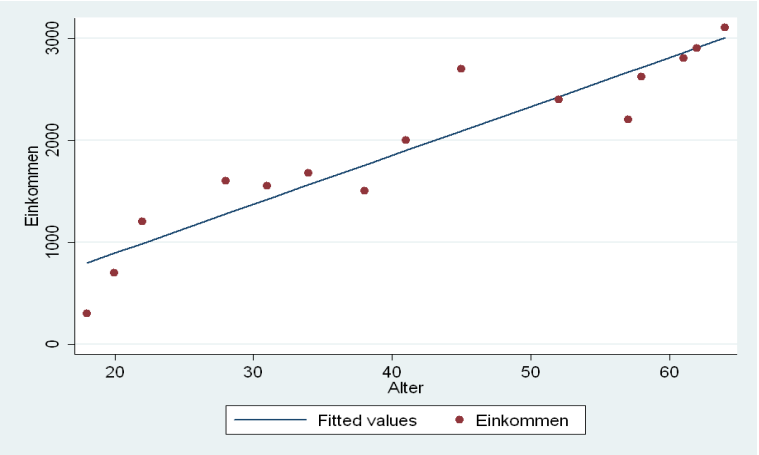

Abbildung 2: Streudiagramm: Alter und Einkommen

der Bedeutung für die Regressionsanalyse und das verwendete Schätzverfahren. Eine

Gerade lässt durch die Schätzgleichung

$$y_t = \pi_1 + \pi_2 x_{2t} + v_t \tag{23}$$

angeben.[36] y_t steht für den empirischen Wert der abhängigen Variablen (Einkommen). Die Variable x_{2t} ist unabhängig (Alter). Der Subindex t steht für die Anzahl der Beobachtungen und ist läuft in diesem Beispiel von t = 1,. . . , 15. Die Parameter π_1 und π_2 sind zuschätzen. Erster gibt im geometrischen Sinne den Schnittpunkt der Geraden mit der y-Achse an. Zweiter stellt die Steigung der Geraden dar. Inhaltlich gibt er an, um wie viel sich die y-Variable ändert, wenn sich die x-Variable um eine Einheit erhöht. Da sich vermuten lässt, dass noch andere Einflüsse außer dem Alter auf das Einkommen wirken, ist die Analyse mit einer Unsicherheit behaftet, die durch den Störterm v_t repräsentiert wird. Für die 15 Beobachtungen ergibt sich damit ein 15-dimensionales Gleichungssystem:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{15} \end{pmatrix} = \begin{pmatrix} 1 & x_{2\,1} \\ \vdots & \vdots \\ 1 & x_{2\,15} \end{pmatrix} * \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ \vdots \\ v_1 5 \end{pmatrix} \tag{24}$$

Da die Parameter π_1 und π_2 unbekannt sind, müssen sie geschätzt werden. Ihre Schätzungen werden mit $\hat{p_1}$ und $\hat{p_2}$ bezeichnet. Damit ergibt sich die Regressionsgerade als der Erwartungswert von y

$$\hat{y_t} = \hat{p_1} + \hat{p_2} x_{2t}. \tag{25}$$

Das Streudiagramm hat gezeigt, dass eine geschätzte Gerade von den tatsächlichen Beobachtungen abweicht. Diese Abweichungen, als vertikale Differenz zwischen tatsächlicher Beobachtung y_t und vorhergesagtem Wert $\hat{y_t}$, wird als Residuum bezeichnet und dient als Schätzwert für den Störterm v_t. Demnach gilt:

$$\hat{v_t} = y_t - \hat{y_t} \tag{26}$$

Ein intuitiver Weg, die Gerade möglichst gut anzupassen, ist es, diese Residuen zu minimieren, also möglichst kleine Differenzen zu erhalten. Dies ermöglicht die OLS-Schätzmethode. Formal ausgedrückt, bedeutet dieses Minimierungsproblem:

$$\sum \hat{v_t}^2 \rightarrow^{Min!} p_1, p_2 \tag{27}$$

Damit wird die Summe der quadrierten Residuen minimiert. Führt man eine konkrete Schätzung durch, so schreibt man für $\hat{p_1}$ p_1 und für $\hat{p_2}$ p_2. Wodurch das Minimierungsproblem unter Beachtung der vorangegangenen Gleichungen in folgende Formulierung übergeht:

$$S = \sum \hat{v_t}^2 = \sum (y_t - \hat{y_t})^2 = \sum (y_t - p_1 - p_2 x_{2t})^2 \rightarrow^{Min!} p_1, p_2 \tag{28}$$

[36] Die folgenden Ausführungen orientieren sich an Assenmacher (2002): 81-88

In den folgenden Schritten werden die entsprechenden mathematischen Operationen und Umformungen vorgenommen, um die Funktion nach p_1 und p_2 aufzulösen. Zunächst wird hierzu die obige Funktion nach dem dritten Summenzeichen nach beiden Parametern partiell abgeleitet.

$$\frac{\partial S}{\partial p_1} = \sum -2(y_t - p_1 - p_2 x_{2t}) = 0 \tag{29}$$

$$\frac{\partial S}{\partial p_2} = \sum -2x_{2t}(y_t - p_1 - p_2 x_{2t}) = 0 \tag{30}$$

Durch die Division beider Gleichungen durch -2 und das Auflösen der beiden Klammern erhält man die *Normalgleichungen*:

$$\sum y_t = Tp_1 + p_2 \sum x_{2t} \tag{31}$$

$$\sum y_t x_{2t} = p_1 \sum x_{2t} + p_2 \sum x_{2t}^2 \tag{32}$$

Dividiert man Normalgleichung I durch T erhält man:

$$\bar{y} = p_1 + p_2 \bar{x}_{2t} \tag{33}$$

$$\text{bzw.} \qquad p_1 = \bar{y} - p_2 \bar{x}_{2t} \tag{34}$$

Damit ist die Schätzgleichung für π_1 bestimmt. Setzt man nun Gleichung 34 in die Normalgleichung II ein folgt:

$$\sum y_t x_{2t} = (\bar{y} - p_2 \bar{x}_2) \sum x_{2t} + p_2 \sum x_{2t}^2 \tag{35}$$

$$\sum y_t x_{2t} = \bar{y} \sum x_{2t} - p_2 \bar{x}_2 \sum x_{2t} + p_2 \sum x_{2t}^2 \tag{36}$$

$$\sum y_t x_{2t} - \bar{y} \sum x_{2t} = -p_2 \bar{x}_2 \sum x_{2t} + p_2 \sum x_{2t}^2 \tag{37}$$

$$\sum y_t x_{2t} - \bar{y} \sum x_{2t} = p_2 (\sum x_{2t}^2 - \bar{x}_2 \sum x_{2t}) \tag{38}$$

$$\sum (y_t - \bar{y}) x_{2t} = p_2 \sum (x_{2t} - \bar{x}_2) x_{2t} \tag{39}$$

Durch einige letzte Umformungen ist dann die Schätzgleichung für π_2 gefunden:

$$p_2 = \frac{\sum (y_t - \bar{y}) x_{2t}}{\sum (x_{2t} - \bar{x}_2) x_{2t}} = \frac{\frac{1}{T} \sum (y_t - \bar{y})(x_{2t} - \bar{x}_2)}{\frac{1}{T} \sum (x_{2t} - \bar{x}_2)^2} \tag{40}$$

$$p_2 = \frac{cov(y_t, x_{2t})}{var(x_{2t})} \tag{41}$$

Diese *unstandardisierten* Koeffizienten sind jedoch abhängig von der jeweiligen Maßeinheit. In einem multiplen Regressionsmodell können die Parameter ihrer Größe nach nicht verglichen werden. Daher werden im nächsten Abschnitt die *standardisierten* β-Koeffizienten vorgestellt.

4.2 Das multiple Regressionsmodell

In der praktischen Arbeit werden häufiger komplexere Modelle bestimmt, so dass mehrere unabhängige Variablen (*Regressoren*) aufgenommen werden. Die zu schätzende Gleichung lautet dann:

$$y_t = \pi_1 + \pi_2 x_{2t} + \ldots + \pi_K x_{Kt} + v_t \tag{42}$$

Der Subindex zeigt die jeweilige Beobachtung an und läuft von t = 1, ..., T. T entspricht dem Stichprobenumfang. Durch den Laufindex k werden die verschiedenen Regressoren angegeben. K entspricht dann der Anzahl der unabhängigen Variablen des Modells. Dieser Index läuft von k = 2, ..., K. Die Variable x_{1t} ist auf den Wert 1 normiert, um das Absolutglied der Regressionsgleichung zu erhalten.
Sind für die unbekannten Parameter entsprechende Werte geschätzt, kann die Gleichung überführt werden in:

$$\hat{y}_t = p_1 + p_2 x_{2t} + \ldots + p_K x_K t \tag{43}$$

Für T Beobachtungen und K Regressoren resultiert daraus folgendes Gleichungssystem

$$\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} 1 & x_{2\,1} & \ldots & x_{K\,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2\,T} & \ldots & x_{K\,T} \end{pmatrix} * \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ \vdots \\ v_T \end{pmatrix} \tag{44}$$

Einfacher lässt sich dieses Gleichungssystem in der Matrixschreibweise formulieren:

$$y = X\pi + v \tag{45}$$

y und v sind (T,1)-Vektoren, π ist ein (K,1)-Vektor und die Matrix X ist eine (T,K)-Matrix. Wendet man wiederum das Prinzip der Minimierung der Summe der quadrierten Residuen an, ergibt sich als OLS-Schätzer für die gesuchten Parameter:

$$p = (X'X)^{-1}X'y \tag{46}$$

Da auch hier eine Matrizeninversion notwendig ist, darf die Determinante der Matrix X nicht gleich Null und die Spaltenvektoren der Matrix nicht linear abhängig sein. Sind die Spaltenvektoren voneinander linear abhängig bedeutet dies, dass die Daten redundant sind und Informationen nicht mehr eindeutig einer bestimmten Variablen zugeordnet werden können. Dieses Problem bezeichnet man auch als Multikollinearität. Während perfekte lineare Abhängigkeit dazu führt, dass das Modell nicht mehr geschätzt werden kann, kommt es bei einer hohen Multikollinearität u.a. zu verzerrten Werten des Determinationskoeffizienten.
Bei mehreren Regressionskoeffizienten liegt die vergleichende Betrachtung der absoluten Parameter nahe. Wie bereits angedeutet ist dies mittels der unstandardisierten Koeffzienten nicht möglich, da diese von ihrer jeweiligen Maßeinheit abhängen. Um dimensionslose Koeffizienten zu erhalten kann er standardisiert werden. Diese β-Werte ergeben sich nach folgender Formel:

$$\beta_k = p_k * \frac{s_{x_k}}{s_y} \tag{47}$$

Der β-Wert ergibt sich also indem man den Regressionskoeffizienten p_k mit der Standardabweichung des k-ten Regressors s_{x_k} mulitpliziert und dieses Produkt durch die Standardabweichung der unabhängigen Variable y s_y teilt. Wird die Analyse mit bereits standardisierten Variablen durchgeführt sind die Schätzungen der Parameter gleich den entsprechenden β-Werten.[37]

[37] Vgl. Backhaus et al. (2006): 62

4.3 Annahmen der linearen Regression

Dem linearen Regressionsmodell liegen verschiedene Annahmen zu Grunde, die bei einer Anwendung geprüft und beachtet werden sollten.[38]

1. Richtige Spezifikation: Linearität in den Parametern, Auswahl und Anzahl der Variablen
2. Der Störterm v_t hat einen Erwartungswert von Null
3. Der Störterm und die Regressoren sind unkorreliert
4. Die Varianz der Störterme ist konstant, Freiheit von Heteroskedastizität
5. Die Störterme sind unkorreliert, Freiheit von Autokorrelation
6. Die Regressoren sind linear unabhängig, keine perfekte Multikollinearität
7. Die Störterme sind normalverteilt

4.4 Modellprüfung

Abschließend werden einige statistische Methoden vorgestellt mit denen das Modell geprüft werden kann. Man unterscheidet hier insbesondere globale und koeffizientenbezogene Prüfmaße. Ein wichtiges Kriterium zur Beurteilung der Anpassung der Regressionsgeraden an die empirischen Daten stellt das Bestimmtheitsmaß oder auch Determinantionskoeffizient genannt dar. Dieses setzt bei den Residuen, also den Abweichungen der vorhergesagten Werte von den empirischen Werten, an. Die Gesamtstreuung der tatsächlichen Beobachtungen lässt sich zerlegen in einen erklärten und einen nicht erklärten Teil der Streuung. Die Gesamtstreuung wird als Quadratsumme der Differenz von tatsächlichen Beobachtungen der abhängigen Variable und ihrem arithmetischen Mittel. Die Zerlegung erfolgt so:[39]

$$\sum (y_t - \bar{y})^2 = \sum (\hat{y}_t - \bar{y})^2 + \sum (y_t - \hat{y}_t)^2 \tag{48}$$

Der zweite der Gleichung stellt den erklärten Varianzanteil dar, während der letzte Summand die unerklärte Varianz abbildet. Hieraus lässt sich ein Maß ableiten, das die erklärte Streuung in Relation zur gesamten Streuung setzt:

$$R^2 = \frac{\sum (\hat{y}_t - \bar{y})^2}{\sum (y_t - \bar{y})^2} = \frac{\text{erklärte Streuung}}{Gesamtsstreuung} = 1 - \frac{\sum (y_t - \hat{y}_t)^2}{\sum (y_t - \bar{y})^2} \tag{49}$$

R^2 ist normiert auf einen Wertebereich zwischen 0 und 1. 0 bedeutet, das keine Varianz erklärt wird. Im Gegensatz dazu bedeutet ein Wert von 1, dass 100% der Varianz durch das Modell erklärt werden.

[38] Vgl. Backhaus et al. (2006): 79ff
[39] Vgl. Backhaus et al.(2006): 66f

Da der Determinationskoeffizient von der Anzahl der Regressoren beeinflusst wird, sollte im multiplen Regressionsmodell das korrigierte R^2_{korr} betrachtet werden:

$$R^2_{korr} = R^2 - \frac{K(1-R^2)}{T-K-1} \tag{50}$$

Mittels der F-Statistik lässt sich klären, ob das für die Stichprobe geschätzte Modell auch für die Grundgesamtheit Gültigkeit besitzt. Die Nullhypothese dieses Tests ist, dass alle Regressionskoeffizienten gleich Null sind: $H_0 : \pi_1 = \pi_2 = \ldots = \pi_K = 0$. Damit besteht in der Grundgesamtheit kein kausaler Zusammenhang zwischen der abhängigen und den unabhängigen Variablen des Modells. Aus der Stichprobe klässt sich dann ein empirischer F-Wert berechnen:[40]

$$F_{emp} = \frac{\frac{\sum(\hat{y}_t - \bar{y})^2}{K}}{\frac{\sum(y_k - \hat{y}_k)^2}{T-K-1}} \tag{51}$$

Somit werden die erklärte und die nicht erklärte Streuung jeweils dividiert durch ihre Freiheitsgrade ins Verhältnis gesetzt. Dieser empirische Wert wird dann mit dem entsprechenden Wert der theoretischen F-Statistik verglichen (dabei sind die Anzahl der Freiheitsgrade und die Vertrauenswahrscheinlichkeit zu berücksichtigen). Ist der empirische Wert größer als der theoretische spricht man von einer signifikanten Abweichung von Null und hat dementsprechend einen signifikanten Zusammenhang vorliegen. Die Nullhypothese würde dann zugunsten der Alternativhypothese verworfen.
Eine weitere Möglichkeit zur Beurteilung der Modellgüte ist der Standardfehler, der sich aus der Summe der quadrierten Residuen, die durch die Anzahl der Freiheitsgrade (T - K - 1) dividiert wird ergibt. Er gibt den mittleren Fehler der Regression an.
Mit Hilfe des t-Test können einzelne Regressionskoeffizienten auf ihre Signifikanz getestet werden. Die Vorgehensweise verläuft analog zum F-Test, wobei sich der empirische t-Wert wie folgt berechnen lässt:[41]

$$t_{emp} = \frac{p_k - \pi_k}{s_{p_k}} \tag{52}$$

Da auch hier von einer Nullhypothese H_0: $\pi_k = 0$ ausgegangen wird, lässt sich die obige Gleichung vereinfachen.

$$t_{emp} = \frac{p_k}{s_{p_k}} \tag{53}$$

Zuletzt lässt sich das Konfidenzintervall für einen bestimmten Parameter bilden. Dieses gibt an, in welchem Bereich der Wert des unbekannten Populationsparameters liegt. Je größer dieses Intervall bei einer gewählten Vertrauenswahrscheinlichkeit ist, desto unbestimmter ist die Schätzung.

$$p_k - t \cdot s_{p_k} \leq \pi_k \leq p_k + t \cdot s_{p_k} \tag{54}$$

[40] Vgl. Backhaus et al. (2006): 69f
[41] Vgl. Backhaus et al. (2006): 74f

5 Pfadanalytische Modelle

Wie bereits beschrieben geht die Pfadanalyse auf Arbeiten des Genetikers Wright zurück. Dabei versuchte er gerichtete Beziehungen zwischen Variablen zu analysieren. Der Regressionsanalyse ähnlich geht es darum, den Einfluss der unabhängigen auf die abhängigen Variablen zu quantifizieren und Hypothesen empirisch zu testen. Jedoch können die Beziehungen zwischen den Variablen wesentlich vielfältiger sein.[42] In einem Regressionsmodell wird darüber hinaus lediglich eine abhängige Variable betrachtet, in der Pfadanalyse hingegen können verschiedene Unabhängige formuliert werden. Dabei ist es möglich das eine Variable x_1 in Bezug auf die Variable x_2 als Regressor wirkt, aber wiederum durch eine Variable x_3 erklärt wird. Ein wesentliches Mittel, um sich derartige Beziehungen zu veranschaulichen sind Pfaddiagramme.
Um derartige Diagramme besser lesen zu können seien hierzu vorab einige Konventionen getroffen:

1. Exogene Variablen werden mit dem Buchstaben X bezeichnet. Eine Variable ist dann exogen, wenn keine andere auf sie einwirkt.

2. Endogene Variablen werden mit dem Buchstaben Y bezeichnet. Eine Variable ist dann endogen, wenn mindestens eine andere Variable auf sie einwirkt.

3. Ein gerichteter Pfeil $\rightarrow$ zeigt eine asymmetrische Kausalbeziehung an. Demzufolge wirkt X auf Y ($X \rightarrow Y$).

4. Eine gebogene Linie oder ein gerade Linie mit Pfeilen an beiden Enden steht für eine ungerichtete Beziehung und symbolisiert eine Korrelation zwischen den Variablen ($X \leftrightarrow Y$).

5. Zwei entgegengesetzte Pfeile stehen für eine Rückkoppelung zwischen zwei Variablen ($X \overrightarrow{\leftarrow} Y$).

6. Die Subindizes der Pfadkoeffzienten sind derart gestaltet, dass zuerst die abhängige und dann die unabhängige Variable angegeben wird (p_{yx}).

7. Variablen in Ovalen sind latente Variablen.

8. Variablen in Rechtecken sind manifeste Variablen.

Die Unterscheidung zwischen latenten und manifesten Variablen ist hier nur der Vollständigkeit halber aufgeführt, spielt aber im Moment keine Rolle, da in diesem Kapitel lediglich manifeste Variablen erfasst werden. Pfadmodelle ohne Wechelwirkungen oder Rückkoppelungen (a) bezeichnet man als *rekursive* Modelle. Treten hingegen solche Wechselwirkungen oder Rückkoppelungen auf so spricht man von *nicht-rekursiven* Pfadmodellen. Weiterhin unterscheidet man *saturierte* (vollständige) und *nicht-saturierte* (unvollständige) Modelle. Ist zum Beispiel jede Variable mit jeder anderen über einen Pfad verbunden bezeichnet man dieses Modell als saturiert. Die Pfadanalyse stellt eine Verschachtelung von verschiedenen Regressionsgleichungen dar. Aus diesem Grund müssen zum Einen die Annahmen des Regressionsmodells erfüllt sein. Zum Anderen sind zusätzlich folgende Annahmen zu treffen:

[42] Vgl. Reinecke (2005): 45

1. Die Indikatoren wurden messfehlerfrei erhoben
2. Die Residuen der einzelnen Regressionsgleichungen sind unkorreliert
3. Die Residuen und die exogenen Variablen sind unkorreliert

5.1 Rekursive Pfadmodelle

Je nach Modellspezifikation können einfache Pfadmodelle (mit drei Variablen) durch die bereits besprochene bivariate oder multiple Regression analysiert werden. „Basieren die empirischen Informationen auf Korrelationskoeffizienten, dann sind die standardisierten Regressionskoeffizienten die Pfadkoeffizienten im Modell.“[43]
Komplexere Pfadmodelle hingegen verlangen den Einsatz mehrere Regressionsgleichungen. Beispielhaft sei hier ein Fall mit vier Variablen betrachtet. Die Beziehungen zwischen diesen Variablen werden durch das folgende Pfaddiagramm dargestellt.
Demnach wirken X_1, X_2 und Y_1 auf die Variable Y_2, wobei Y_1 von den Variablen

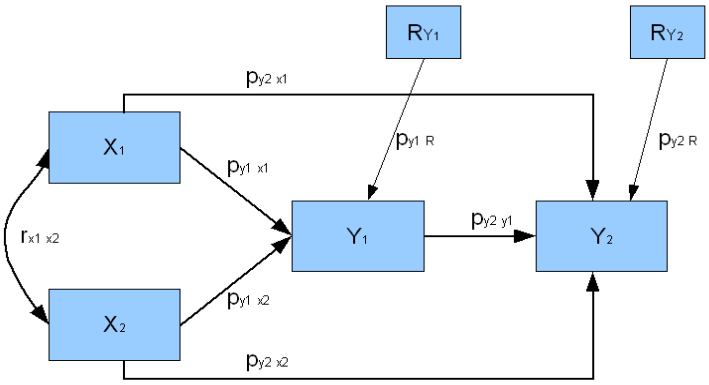

Abbildung 3: Pfaddiagramm: Pfadanalyse mit vier Variablen

X_1 und X_2 beeinflusst wird, die wiederum korreliert sind. Um diesen Sachverhalt zu formalisieren werden zwei Regressionsgleichungen benötigt:[44]

$$Y_1 = p_{Y_1 X_1} X_1 + p_{Y_1 X_2} X_2 + p_{Y_1 R_{Y_1}} R_{Y_1} \tag{55}$$

$$Y_2 = p_{Y_2 X_1} X_1 + p_{Y_2 X_2} X_2 + p_{Y_2 Y_1} Y_1 + p_{Y_2 R_{Y_2}} R_{Y_2} \tag{56}$$

Die Terme R_{Y_1} und R_{Y_2} symbolisieren die Residuen der jeweiligen Strukturgleichungen. Im Folgenden wird angenommen, dass alle Variablen standardisiert sind. Nun werden die Pfadkoeffizienten der Gleichung 56 ermittelt, um die Berechnung zu verdeutlichen.
In einem ersten Schritt multipliziert man die Gleichung 56 jeweils separat mit ihren unabhängigen Variablen. Für die Multiplikation mit X_1 ergibt sich Gleichung 57 und

[43] Reinecke (2005): 46
[44] die folgenden Ausführungen orientieren sich an Reinecke (2005): 47ff

für X_2 Gleichung 58:

$$Y_1X_1 = p_{Y_1X_1}X_1^2 + p_{Y_1X_2}X_2X_1 + p_{Y_1R_{Y_1}}R_{Y_1}X_1 \tag{57}$$

$$Y_1X_2 = p_{Y_1X_1}X_1X_2 + p_{Y_1X_2}X_2^2 + p_{Y_1R_{Y_1}}R_{Y_1}X_2 \tag{58}$$

Aus Gründen der Vereinfachung wurden bislang die Laufindizes der Variablen nicht angegeben. Man sollte sich jedoch vergegenwärtigen, dass jede Variable eine Reihe von Beobachtungen repräsentiert. Daher können für die beiden Gleichungen Mittelwerte gebildet werden:

$$\frac{\sum(Y_1X_1)}{N} = p_{Y_1X_1}\frac{\sum(X_1^2)}{N} + p_{Y_1X_2}\frac{\sum(X_2X_1)}{N} + p_{Y_1R_{Y_1}}\frac{\sum(R_{Y_1}X_1)}{N} \tag{59}$$

$$\frac{\sum(Y_1X_2)}{N} = p_{Y_1X_1}\frac{\sum(X_1X_2)}{N} + p_{Y_1X_2}\frac{\sum(X_2^2)}{N} + p_{Y_1R_{Y_1}}\frac{\sum(R_{Y_1}X_2)}{N} \tag{60}$$

Diese Gleichungen lassen sich vereinfachen, wenn man bedenkt, dass sich der Korrelationskoeffizient als das Produkt zweier z-transformierter Variablen auffassen lässt. Um dies zu verdeutlichen wird hier ein kurzer Zwischenschritt[45] unternommen. Die Kovarianz der Variablen X und Y wird wie folgt berechnet:

$$cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N} \tag{61}$$

In dieser Formel bezeichnet N den Stichprobenumfang und i läuft von 1, ..., n. Teilt man diese durch das Produkt der Standardabweichungen der beiden Variablen erhält man den Korrelationskoeffizienten r_{xy}:

$$r_{xy} = \frac{cov(x,y)}{s_x s_y} \tag{62}$$

Setzt man die Gleichung der Kovarianz in Formel 62 ein, dann folgt:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N s_x s_y} \tag{63}$$

$$r_{xy} = \frac{1}{N}\sum\left(\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}\right) \tag{64}$$

Da N eine Konstante ist kann sie in Gleichung 64 vor das Summenzeichen gezogen werden. Außerdem sind auch s_x und s_y konstant, wodurch man sie in die Summe hineinziehen kann, zumal N^{-1} ausgeklammert wurde. Damit wurde der Korrelationskoeffizient als das Produkt zweier standardisierter Variablen gezeigt.
Die Gleichungen 59 und 60 lassen sich dann wie folgt vereinfachen:

$$r_{Y_1X_1} = p_{Y_1X_1}r_{X_1X_1} + p_{Y_1X_2}r_{X_2X_1} + p_{Y_1R_{Y_1}}r_{R_{Y_1}X_1} \tag{65}$$

$$r_{Y_1X_2} = p_{Y_1X_1}r_{X_2X_1} + p_{Y_1X_2}r_{X_2X_2} + p_{Y_1R_{Y_1}}r_{R_{Y_1}X_2} \tag{66}$$

Es wurde angenommen das die Residuen nicht mit den unabhängigen Variablen der Strukturgleichungen korrelieren, wodurch die entsprechenden Produkte in den

[45] Vgl. Reinecke (2005): 30f

beiden Gleichungen wegfallen ($r_{R_{Y_1}X_1} = 0$ und $r_{R_{Y_1}X_2} = 0$). Da die Korrelation einer Variablen mit sich selbst perfekt, also gleich Eins, ist ($r_{X_1X_1} = 1$ und $r_{X_2X_2} = 1$), vereinfachen sich die Gleichungen zu:

$$r_{Y_1X_1} = p_{Y_1X_1} + p_{Y_1X_2}r_{X_2X_1} \tag{67}$$
$$r_{Y_1X_2} = p_{Y_1X_1}r_{X_2X_1} + p_{Y_1X_2} \tag{68}$$

Stellt man nun Gleichung 67 nach $p_{Y_1X_1}$ um und setzt diese dann in Gleichung 68 ein, resultiert:

$$r_{Y_1X_2} = (r_{Y_1X_1} - p_{Y_1X_2}r_{X_2X_1})r_{X_2X_1} + p_{Y_1X_2} \tag{69}$$

Diese Gleichung lässt sich wie folgt umformen, um den Pfadkoeffizienten $p_{Y_1X_2}$ zu erhalten:

$$r_{Y_1X_2} = r_{Y_1X_1}r_{X_2X_1} - p_{Y_1X_2}r^2_{X_2X_1} + p_{Y_1X_2} \tag{70}$$
$$r_{Y_1X_2} = r_{Y_1X_1}r_{X_2X_1} + p_{Y_1X_2}(1 - r^2_{X_2X_1}) \tag{71}$$
$$r_{Y_1X_2} - r_{Y_1X_1}r_{X_2X_1} = p_{Y_1X_2}(1 - r^2_{X_2X_1}) \tag{72}$$
$$p_{Y_1X_2} = \frac{r_{Y_1X_2} - r_{Y_1X_1}r_{X_2X_1}}{(1 - r^2_{X_2X_1})} \tag{73}$$

Um nun auch den Pfadkoeffizienten $p_{Y_1X_1}$ zu erhalten, kann Gleichung 73 in die nach $p_{Y_1X_1}$ aufgelöste Formel 67 eingesetzt werden:

$$p_{Y_1X_1} = r_{Y_1X_1} - \frac{r_{Y_1X_2} - r_{Y_1X_1}r_{X_2X_1}}{(1 - r^2_{X_2X_1})}r_{X_2X_1} \tag{74}$$
$$p_{Y_1X_1} = r_{Y_1X_1} - \frac{r_{Y_1X_2}r_{X_2X_1} - r_{Y_1X_1}r^2_{X_2X_1}}{(1 - r^2_{X_2X_1})} \tag{75}$$

Am Beispiel der letzten Gleichung soll der Pfadkoeffizient näher betrachtet werden. Zunächst ist festzustellen, dass die eigentliche bivariate Korrelation um den Bruch korrigiert wird. Damit trägt man der Tatsache Rechnung, dass durch die Korrelation der beiden X-Variablen und den $\text{Pfad}_{Y_1X_2}$ die bivariate Korrelation der Variablen X_1 und Y_1 beeinflusst wird. Im Nenner des Bruches findet man die nicht erklärte Varianz, als die Differenz aus 1 und der quadrierten Korrelation der die Variable Y_1 erklärenden Variablen X_1 und X_2. Hält man den Nenner konstant, führen stärkere indirekte Einflüsse zu einem absolut betrachtet kleineren Pfadkoeffizienten. Ebenso führt unter Konstanthaltung des Zählers eine höhere quadrierte Korrelation der beiden erklärenden Variablen zu einem absolut betrachtet niedrigeren Pfadkoeffizienten. Nimmt man an, dass X_1 und X_2 unkorreliert sind, so ergibt sich $r_{X_2X_1} = 0$. Infolgedessen würde der Pfadkoeffizient aus Gleichung 75 der bivariaten Korrelation entsprechen.
Die Pfadkoeffizienten der Residualvariablen lassen sich über die Varianzzerlegung ermitteln:

$$R^2_{Y_1} = p_{Y_1X_1}r_{Y_1X_1} + p_{Y_1X_2}r_{Y_1X_2} \tag{76}$$
$$R^2_{Y_2} = p_{Y_2X_1}r_{Y_2X_1} + p_{Y_2X_2}r_{Y_2X_2} + p_{Y_2Y_1}r_{Y_2Y_1} \tag{77}$$

Über diese Gleichungen lässt sich durch das Einsetzen der entsprechenden Werte der erklärte Varianzanteil berechnen. Die Residualpfadkoeffizienten ergeben sich dann aus dem Anteil der nicht erklärten Varianz:

$$p_{Y_1 R_{Y_1}} = \sqrt{1 - R^2_{Y_1}} \tag{78}$$

$$p_{Y_2 R_{Y_2}} = \sqrt{1 - R^2_{Y_2}} \tag{79}$$

Um eine Identifizierbarkeit des Modells zu gewährleisten, müssen in der Praxis zumeist Restriktionen in das Modell aufgenommen werden. Möglich ist etwa, die Wirkung einer abhängigen Variablen auf sich selbst gleich Null zu setzen. Zudem kann es sich anbieten, die inhaltlich nicht weiter spezifizierten Residualvariablen auf den Wert Eins zu skalieren. Weitere Restriktionen ergeben sich aus der theoretischen Vorarbeit. Eine praktische Orientierungshilfe bietet die eingangs erwähnte t-Regel. Die Modellschätzung erfolgt zum Beispiel über die Maximum-Likelihood-Methode, die eine simultane Lösung des Gleichungssystems ermöglicht. Während sie bei gerade identifizierten Modell zu einer eindeutigen Lösung kommt, wird bei überidentifizierten Modellen iterativ verfahren, indem eine Fitfunction minimiert wird.[46]
Neben dem Modell, welches als Beispiel behandelt wurde, lassen sich noch weitere Spezifikationen und Wirkungszusammenhänge zwischen den Variablen unterstellen und es können daher verschiedene saturierte Alternativmodelle formuliert werden. Von Nachteil ist hier jedoch, dass diese gerade identifiziert sind und somit immer eine perfekte Anpassung zeigen. Durch die Einführung von Modellrestriktionen ist es möglich, Freiheitsgrade zu gewinnen und den Informationsgehalt zu steigern. Zum Einen kann man dabei theoretischen Überlegungen folgen und zum Anderen empirische Indikatoren, wie das Signifikanzniveau einzelner Parameter zum Beispiel, heranziehen. Zudem werden χ^2-Differenztests möglich, um hierarchische Modelle zu vergleichen.[47]

Effektzerlegung
Korrelationen können in direkte und indirekte Effekte zerlegt werden. Während die Pfadkoeffizienten die direkten Korrelationen wiedergeben, ergeben sich die indirekten Effekte durch die Multiplikation der einzelnen Pfadkoeffizienten. Bezogen auf das Beispiel ergibt sich demnach folgende Zerlegung für den Effekt der Variablen X_1 auf die Variable Y_1:[48]

$$r_{Y_1 X_1} = \underbrace{p_{Y_1 X_1}}_{\text{direkter Effekt}} + \underbrace{p_{Y_1 X_2} \cdot p_{X_2 X_1}}_{\text{indirekter Effekt}} \tag{80}$$

Auf diese weise lassen sich auch alle anderen Korrelationen zerlegen. Exemplarisch wird dies an der Korrelation $r_{X_1 Y_2}$ dargestellt:

$$r_{X_1 Y_2} = p_{Y_2 X_1} + p_{Y_2 X_2} \cdot p_{X_2 X_1} + p_{Y_2 Y_1} \cdot p_{Y_1 X_1} + p_{Y_2 Y_1} \cdot p_{Y_2 X_1} \cdot p_{X_2 X_1} \tag{81}$$

Durch diese Zerlegung kann die Beziehungstruktur offengelegt werden. Hierdurch ist es möglich Variablen zu erkennen, die einen direkten Beitrag zur Varianzaufklärung leisten und solche, bei denen dies nur indirekt der Fall ist.

[46] Vgl. Reinecke (2005): 52-54
[47] Vgl. Reinecke (2005): 56-58
[48] Vgl. Reinecke (2005): 50ff

5.2 Nicht-rekursive Pfadmodelle

„Ein Pfadmodell wird als nicht-rekursiv bezeichnet, wenn mindestens eine *direkte* oder *indirekte* Rückwirkung zwischen zwei Variablen besteht.“[49] Ein solches Modell wird durch Abbildung **??** dargestellt.[50]
Zu sehen ist ein nicht-rekursives Pfadmodell mit einer direkten Rückwirkung zwischen den Variablen Y_1 und Y_2. Formal lässt sich das Modell durch folgende Gleichungen ausdrücken:

$$Y_1 = \gamma_{11} \cdot X_1 + \beta_{12} \cdot Y_2 + \zeta_1 \tag{82}$$
$$Y_2 = \gamma_{22} \cdot X_2 + \beta_{21} \cdot Y_1 + \zeta_2 \tag{83}$$

γ_{11}, γ_{22}, β_{12} und β_{21} stellen die Pfadkoeffizienten dar. ϕ_{21} ist die Kovarianz zwischen den beiden X-Variablen. ζ_1 und ζ_2 stehen für die Residuen. Insgesamt enthält die empirische Kovarianzmatrix damit 10 Elemente, auf deren Basis 9 Parameter zu schätzen. Damit ist das Modell mit einem Freiheitsgrad überidentifiziert.
Hinsichtlich der kausalen Interpretation des Modells ist es nun problematisch, dass die beiden Y-Variablen jeweils abhängige und unabhängige Variablen sind. Nunmehr kann man nicht-rekursive Pfadmodelle lediglich als eine Annäherung an kausalanalytische Modelle betrachten, da hier keine zeitliche Differenz zwischen Ursache und Wirkung vorliegt.
Die Entscheidung für oder gegen ein rekursives bzw. ein nicht-rekursives Modell sollte wohl überlegt sein. Insbesondere können alternative Modelle nach dem oben beschriebenen Verfahren gegeneinander getestet werden. Durch die Einführung einer intervenierenden Variablen Y_3 kann ein indirekter Effekt von Y_2 auf Y_1 modelliert werden.
pfaddiagramm 60
Daraus folgen die nachstehenden Strukturgleichungen:

$$Y_1 = \gamma_{11} \cdot X_1 + \beta_{13} \cdot Y_3 + \zeta_1 \tag{84}$$
$$Y_2 = \gamma_{22} \cdot X_2 + \beta_{21} \cdot Y_1 + \zeta_2 \tag{85}$$
$$Y_3 = \beta_{32} \cdot Y_2 + \zeta_3 \tag{86}$$

Da das Modell auf Kovarianzen basiert, sind die resultierenden Parameter nicht standardisiert. Dies entspricht der eigentlichen Forderung nach einer Kovarianzmatrix, da auf der Grundlage von Korrelationen falsche Standardfehler sowie Verzerrungen einiger Gütemaße entstehen können. Es ist jedoch möglich die unstandardisierten Koeffizienten zu standardisieren:

$$\beta_{ij}^s = \beta_{ij} \cdot \frac{\sigma_{Y_j}}{\sigma_{Y_i}} \tag{87}$$
$$\gamma_{ij}^s = \gamma_{ij} \cdot \frac{\sigma_{X_j}}{\sigma_{X_i}} \tag{88}$$

Dabei steht der Index i für die jeweilige Zielvariable (Wirkung) und das Subskript j für die ursächliche Variable. σ ist Standardabweichung der Variablen. Die Standardisierung erfolgt also dadurch, dass man den unstandardisierten Pfadkoeffizienten

[49] Reinecke (2005): 58
[50] Vgl. zu den folgenden Ausführungen Reinecke (2005): 58-63

mit dem Verhältnis der Standardabweichungen der jeweiligen unabhängigen und abhängigen Variablen multipliziert.

5.3 Beispiel

Im Folgenden wird beispielhaft ein rekursives Pfadmodell besprochen, das im Wesentlichen der Modellstruktur entspricht, die in Kapitel 5.1 vorgestellt wurde. Das Beispiel wurde aus einem Skript zur Pfadanalyse von Wolfgang Langer entnommen.[51] Es basiert auf dem „General Social Survey 1978“, mit dem u.a. der Alkoholkonsum der amerikanischen Bevölkerung analysiert werden sollte.
In die Analyse gehen vier manifeste Variablen ein. Zwei exogene Variablen *Alter* und *Bildung* sowie zwei endogene Variablen *Häufigkeit des Kneipenbesuchs* und *Alkoholmenge*. Ein Pfaddiagramm dient der Strukturierung und Darstellung der vermuteten Beziehungen und Wirkungszusammenhänge. Es wird folglich unterstellt,

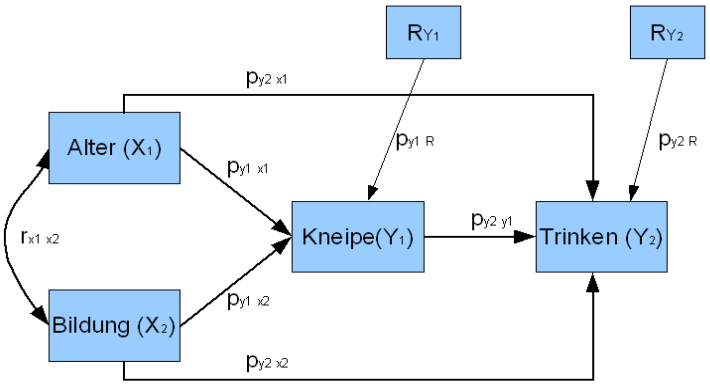

Abbildung 4: Pfaddiagramm: Beispiel zur Pfadanalyse mit vier Variablen

das die Häufigkeit des Kneipengangs einerseits vom Alter und andererseits vom Bildungsniveau abhängig ist. Zugleich korrelieren Alter und Bildung. Des Weiteren wird angenommen, dass die konsumierte Menge an Alkohol durch das Alter, die Bildung und die Anzahl der Kneipenbesuche beeinflusst wird.
Das Modell lässt sich aber auch in Gleichungsform darstellen:

$$Y_1 = p_{Y_1 X_1} X_1 + p_{Y_1 X_2} X_2 + p_{Y_1 R_{Y_1}} R_{Y_1} \tag{89}$$

$$Y_2 = p_{Y_2 X_1} X_1 + p_{Y_2 X_2} X_2 + p_{Y_2 Y_1} Y_1 + p_{Y_2 R_{Y_2}} R_{Y_2} \tag{90}$$

Langer gibt neben dieser Modellstruktur die Korrelationsmatrix der manifesten Variablen an:

1.000			
-0.310	1.000		
0.145	-0.340	1.000	
0.248	-0.283	0.536	1.000

[51] Vgl. Langer (2002a): 5ff

Nun kann das Modell in ein LISREL-File überführt werden, um die Pfadkoeffizienten zu schätzen.[52] Hierzu wird die einfachere SIMPLIS-Notation verwendet, wobei die farbigen Anmerkungen als ergänzende Hinweise zur Erstellung eines SIMPLIS-Projektes zu verstehen sind:

!Pfadanalyse Der Titel eines SIMPLIS-Files wird mit einem ! eingeleitet
Observed Variables: Bildung Alter Kneipe Trinken Beobachtete Variablen in korrekter Reihenfolge
Correlation Matrix:Eine Korrelationsmatrix ist symmetrisch
1.000
-0.310 1.000
0.145 -0.340 1.000
0.248 -0.283 0.536 1.000

Sample Size: 1521 Größe der Stichprobe
Equations:Formulierung der Beziehungen im Modell
Kneipe = Bildung Alter
Trinken = Bildung Alter Kneipe

Path diagram Hierdurch wird ein Pfaddiagramm angefordert

end of problem Beendet die Syntax

Auf einige Grundlagen zur Erstellung einer solchen Syntax wird im Anhang A eingegangen. Nachdem diese Befehlszeihlen in LISREL eingegeben und die Berechnungen durchgeführt wurden, gibt das Programm folgendes Pfaddiagramm aus: Die Pfadko-

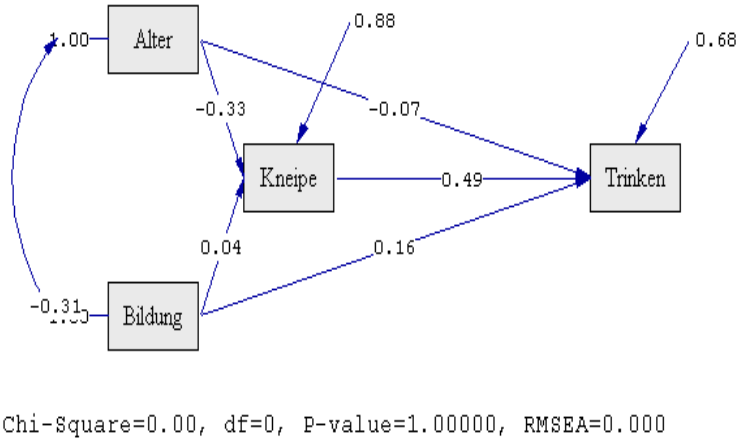

Abbildung 5: Pfaddiagramm: Beispiel zur Pfadanalyse mit vier Variablen mit Pfadkoeffizienten nach ML-Schätzung (nach eigenen Berechnungen)

effizienten geben nun den reinen Einfluss einer Variablen auf eine andere wieder, ohne dabei Effekte über Drittvariablen zu berücksichtigen. Demnach hat zum Beispiel die Häufigkeit des Kneipengangs mit einem standardisierten Pfadkoeffizienten von 0,49

[52] dies könnte bspw. auch mittels SPSS oder Stata durchgeführt werden

den stärksten Effekt auf den Alkoholkonsum. Tendenziell geht zudem mit steigendem Alter ein verminderter Alkoholkonsum einher, wie der direkte Pfadkoeffizient mit einem Wert von -0,07 anzeigt. Der indirekte Effekt des Alters über die Anzahl der Kneipenbesuche auf das Trinkverhalten bestimmt sich dagegen wie folgt:[53]

$$r_{Y_1X_1} = \underbrace{p_{Y_1X_1}}_{direkter\ Kausaleffekt} + \underbrace{p_{Y_1X_1} \cdot p_{Y_2Y_1}}_{indirekter\ Kausaleffekt} + \underbrace{r_{X_1X_2} \cdot p_{Y_2X_2} + r_{X_1X_2} \cdot p_{Y_1X_2} \cdot p_{Y_2Y_1}}_{korrelative\ Effekte} \tag{91}$$

$$r_{Y_1X_1} = \underbrace{-0,07}_{direkter\ Kausaleffekt} + \underbrace{(-0,33) \cdot 0,49}_{indirekter\ Kausaleffekt} + \underbrace{(-0,31) \cdot 0,16 + (-0.31) \cdot 0,04 \cdot 0,49}_{korrelative\ Effekte} = -0,287$$

Bei der ausgegebenen LISREL-Lösung ist zu beachten, dass die Eingabematrix eine Korrelationsmatrix war und daher die Pfadkoeffizienten bereits standardisiert sind. Wäre die Eingabe über eine Kovarianzmatrix erfolgt, müsste in der Zeile *LISREL Output* der Befehl *SC* angegeben werden, um die vollständig standardisierte Lösung zu erhalten.
Auf die weitere Interpretation des Outputs wird an dieser Stelle nicht eingegangen.

6 Konfirmatorische Faktorenanalyse

Die bisherige Darstellung beruhte auf der Annahme, dass nur manifeste Variablen für die Analyse von Interesse sind. Diese Ebene wird nun und im folgenden Kapitel verlassen. In der Regel sind gerade in den Sozialwissenschaften latente, also nicht direkt beobachtbare, Sachverhalte von Bedeutung. Um diese zu erfassen, nutzt man bestimmte Indikatoren, denen eine gemeinsame latente Variable – ein *Faktor* – unterstellt wird. Die explorative und konfirmatorische Faktorenanalyse sind mögliche Verfahren, um solche nicht direkt beobachtbaren Strukturen zu entdecken. Während man aber bei der explorativen Analyse keine Vorstellungen über die Beziehungen zwischen den Indikatoren und dem latenten Konstrukt haben muss, verlangt die konfirmatorische Faktorenanalyse konkrete theoretische Überlegungen. Mit ihr kann eine vermutete Struktur bestätigt oder falsifiziert werden. Im Rahmen der explorativen Faktorenanalyse wird demnach *nur* eine Zahl von Faktoren und manifesten Variablen festgelegt. Dabei werden insbesondere alle direkt beobachteten Variablen von allen gemeinsamen Faktoren beeinflusst. Durch die konfirmatorische Faktorenanalyse ist es dem Forscher möglich, Restriktionen in das Modell einzuführen, um somit die Beziehungsstruktur zu modellieren. Anhand der Stichprobendaten kann dann der vermutete datengenerierende Prozess bestätigt oder abgelehnt werden – daher auch konfirmatorisch.[54]
Long nennt u.a. drei Hauptanwendungsbereiche der konfirmatorischen Faktorenanalyse:[55]

1. Messmodelle für latente Variablen unter expliziter Berücksichtigung von Messfehlern der Indikatoren

[53] Vgl. Langer (2002a): 14
[54] Vgl. Long (1983a): 11-15
[55] Vgl. Long (1983a): 17

2. Multiple Indikatorenmodelle mit denen die Korrelationen zwischen den gemeinsamen Faktoren bestimmt werden können

3. Mulitmethod-Multitraid-Modelle mit denen der Faktor auf verschiedene Weise gemessen wurde, um den Einfluss von Methodeneffekten auszuschalten

Wie auch für die späteren Strukturgleichungsmodelle lassen sich Faktorwerte für die Beobachtungen aus den Indikatoren berechnen. Hierzu werden die z-standardisierten Indikatoren mit den Faktorbetagewichten multipliziert. In LISREL werden die nach der Anderson-Rubin-Methode berechneten Faktorbetagewichte durch den Befehl *FS* in der Outputzeile angefordert.[56]

6.1 Modellspezifikation

Nachdem die Methode grob skizziert wurde, dient dieses Kapitel der formalen Präzisierung. Um ein konfirmatorisches Faktorenmodell zu spezifizieren sind eine Reihe von Annahmen zu treffen:[57]

1. Anzahl der Faktoren

2. Anzahl der Indikatoren

3. Varianzen und Kovarianzen unter den Faktoren

4. Beziehungen zwischen Indikatoren und Faktoren

5. Beziehungen zwischen uniquen Faktoren (Messfehlern) und Indikatoren

6. Varianzen und Kovarianzen zwischen den uniquen Faktoren

Das grundlegende mathematische Vorgehen lässt sich am besten wie folgt beschreiben: „Each observed variable is conceptualized as a linear function of one or more factors.“[58] Dieser Sachverhalt lässt sich formal derart ausdrücken:[59]

$$x = \Lambda\xi + \delta \tag{92}$$

Wobei x ein (q,1)-Vektor der manifesten Variablen, Λ eine (q,s)-Faktorladungsmatrix, ξ ein (s,1)-Vektor der latenten Variablen und δ ein (q,1)-Vektor der Residuen oder auch uniquen Faktoren ist. q steht für die Anzahl der beobachteten Variablen und s für die Anzahl der vermuteten latenten Konstrukte, wobei angenommen wird, dass $q > s$. Betrachtet man diese Gleichung vor dem Hintergrund der vorangegangenen Kapitel, so fällt die Ähnlichkeit zur Regressionsanalyse auf. Auffällig ist aber, das kein Absolutglied vorhanden ist. Dieser ist hier gleich Null, da die Variablen als Abweichungen von ihrem Mittelwert in die Analyse eingehen, wodurch die Berechnungen im Folgenden stark vereinfacht werden. Die Matrix der *Faktorladungen* ist von

[56] Vgl. Langer (2002b): 64
[57] Vgl. Long (1983a): 18
[58] Long (1983a): 22
[59] Vgl. Long (1983a): 22ff

besonderer Bedeutung und kann analog zu dem Steigungskoeffizienten der Regressionsgleichung interpretiert werden. Eine Faktorladung gibt demnach die erwartete Änderung der manifesten Variablen infolge des Anstiegs des latenten Konstrukts um eine Einheit an. Ebenso wie in der Regressionsanalyse, kann auch hier keine perfekte Vorhersage angenommen werden, was durch den Fehlerterm δ zum Ausdruck kommt.
Dadurch, dass die Variablen in Abweichung von ihrem Mittelwert gemessen werden, beträgt der Erwartungswert von x und ξ, sowie δ Null: $E(x) = 0$, $E(\xi) = 0$ und $E(\delta) = 0$. Die Erwartungswerte werden als (q,1)- bzw. (s,1)-Vektoren angegeben, deren Element Null sind. Der daraus resultierende Vorteil für die Berechnungen liegt darin, dass die Kovarianzmatrix, auf der das Modell basiert, als Erwartungswert der Produkte der Vektoren ausgedrückt werden kann:

$$COV(U,V) = E[(U-\mu)(V-\nu)] = E[(u+\mu-\mu)(v+\nu-\nu)] = E(uv) = COV(u,v) \quad (93)$$
$$\text{mit} \qquad u = U - \mu \qquad \text{und} \qquad v = V - \nu$$

Dies zeigt, dass zwei Zufallsvariablen U und V die gleiche Kovarianz aufweisen, wie die Variablen u und v, die als U und V in Abweichung von ihren Mittelwerten μ und ν gemessen werden.
Bezogen auf die Kovarianzmatrix ermöglicht dieser Zusammenhang folgende Umformungen. Dabei geht man von einem Vektor q aus, der n Zeilen aufweist und nur aus Zufallsvariablen besteht. Für den Erwartungswert des Vektors q gilt: $E(q) = 0$. Die Matrix Q sei definiert als $Q = E(qq^T)$. Ein Element dieser Matrix wird mit q_{ij} bezeichnet, wobei i die Zeile und j die Spalte des Elements angibt. Geht man im folgenden Beispiel von einem dreizeiligen Vektor q aus, dann resultiert:

$$qq^T = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \begin{pmatrix} q_1 & q_2 & q_3 \end{pmatrix} = \begin{pmatrix} q_1q_1 & q_1q_2 & q_1q_3 \\ q_2q_1 & q_2q_2 & q_2q_3 \\ q_3q_1 & q_3q_2 & q_3q_3 \end{pmatrix} \quad (94)$$

Für Q gilt dann:

$$Q = E(qq^T) = \begin{pmatrix} E(q_1q_1) & E(q_1q_2) & E(q_1q_3) \\ E(q_2q_1) & E(q_2q_2) & E(q_2q_3) \\ E(q_3q_1) & E(q_3q_2) & E(q_3q_3) \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix} \quad (95)$$

Nun wurde angenommen, das die Elemente q_i als Abweichung von ihrem Mittelwert gemessen wurden. Daher gilt $q_{ij} = COV(q_i, q_j)$ und $q_{ii} = COV(q_i, q_i) = VAR(q_i)$. Dadurch kann Q wiederum anders geschrieben werden:

$$Q = \begin{pmatrix} VAR(q_1) & COV(q_1,q_2) & COV(q_1,q_3) \\ COV(q_2,q_1) & VAR(q_2) & COV(q_2,q_3) \\ COV(q_3,q_1) & COV(q_3,q_2) & VAR(q_3) \end{pmatrix} \quad (96)$$

Diese Matrix ist eine *Varianz-Kovarianzmartrix*, die als Grundlage der Berechnungen genutzt wird. Es handelt sich dabei um eine symmetrische Matrix, so dass sie zumeist

als *untere Dreiecks-Matrix* angegeben wird:

$$Q = \begin{pmatrix} VAR(q_1) & 0 & 0 \\ COV(q_2, q_1) & VAR(q_2) & 0 \\ COV(q_3, q_1) & COV(q_3, q_2) & VAR(q_3) \end{pmatrix} \quad (97)$$

$$\text{oder auch nur } Q = \begin{pmatrix} VAR(q_1) & & \\ COV(q_2, q_1) & VAR(q_2) & \\ COV(q_3, q_1) & COV(q_3, q_2) & VAR(q_3) \end{pmatrix} \quad (98)$$

Man unterscheidet weiterhin zwischen einer solchen *Stichproben-* und *Populationsmatrix* sowie im Laufe der Schätzungen einer *theoretischen* oder auch *geschätzten Varianz-Kovarianzmatrix*. Die Populationsmatrix wird mit dem griechischen Buchstaben Σ bezeichnet und ist zumeist unbekannt. Stattdessen ist in der praktischen Forschung nur eine Stichprobe verfügbar. Die hieraus resultierende Varianz-Kovarianzmatrix[60] wird mit S bezeichnet. Σ ergibt sich theoretisch wie folgt:

$$\Sigma = E(xx^T) \quad (99)$$

Und ist damit eine (q,q)-Kovarianzmatrix, die demnach ebenfalls symmetrisch ist. Jedes Element abseits der Hauptdiagonalen dieser Matrix ist dann die Kovarianz zwischen den Indikatoren x_i und x_j. Diese wird mit

$$\sigma_{ij} = E(x_i x_j) \quad (100)$$

Es wurde gezeigt, wie die Kovarianzmatrix aus den empirischen Daten zu gewinnen und formal zu notieren ist. Problematisch bleibt dabei jedoch, wie aus der Gleichung 92 eine Schätzung der unbekannten Koeffizienten möglich ist. Obwohl diese Gleichung an eine multiple Regression erinnert, kann sie nicht auf gleiche Weise gelöst werden, da eine latente Variable auf der Seite der Unabhängigen steht. Aus diesem Grund muss die Populationskovarianzmatrix näher untersucht werden:[61]

$$\Sigma = E(xx^T) = E[(\Lambda\xi + \delta)(\Lambda\xi + \delta)^T] \quad (101)$$

Die Gleichung 101 ergibt sich, wenn man berücksichtigt, dass sich der Vektor der Indikatoren (x) durch Gleichung 92 bestimmen lässt. Zudem wurde Σ als der Erwartungswert des Produktes aus dem Vektor x und seiner Transponierten formuliert. Durch das Einsetzen von 92 in Gleichung 99. Durch elementare Umformungen kann dann folgende Form gewonnen werden:[62]

$$\Sigma = E[(\Lambda\xi + \delta)(\Lambda^T\xi^T + \delta^T)] \quad (102)$$

$$\Sigma = E[\Lambda\xi\xi^T\Lambda^T + \Lambda\xi\delta^T + \delta\xi^T\Lambda^T + \delta\delta^T] \quad (103)$$

$$\Sigma = E[\Lambda\xi\xi^T\Lambda^T] + E[\Lambda\xi\delta^T] + E[\delta\xi^T\Lambda^T] + E[\delta\delta^T] \quad (104)$$

[60] im Folgenden wird nur der Ausdruck Kovarianzmatrix als hierzu synonym gebraucht

[61] Vgl. Long (1983a): 33f

[62] zu beachten sind die Rechenregeln für das Rechnen mit transponierten Matrizen für die erste Gleichung, wobei für die zweite Gleichung Ausmultipliziert wurde und für die dritte Gleichung ist darauf hinzuweisen, dass der Erwartungswertoperator ein linearer Operator ist

Da Λ geschätzt werden muss, enthält sie keine Zufallsvariablen und stellt somit eine Konstante dar und kann Ausgeklammert werden:

$$\Sigma = \Lambda E[\xi\xi^T]\Lambda^T + \Lambda E[\xi\delta^T] + E[\delta\xi^T]\Lambda^T + E[\delta\delta^T] \tag{105}$$

Definiert man nun die $E[\xi\xi^T]$, die Kovarianzmatrix der unabhängigen latenten Variablen, als Φ und $E[\delta\delta^T]$, die Kovarianzmatrix der uniquen Faktoren, als Θ und nimmt zudem an, dass δ und xi unkorreliert sind, vereinfacht sich der Sachverhalt:

$$\Sigma = \Lambda\Phi\Lambda^T + \Theta \tag{106}$$

Diese Gleichung wird als *Kovarianzgleichung* bezeichnet. Mit dieser Formulierung ist es möglich Schätzungen durchzuführen, da die Populationskovarianzmatrix und damit die Stichprobenmatrix[63] in die unbekannten Parameter zerlegt wurde.

6.2 Identifikation

Bevor ein Modell geschätzt werden kann, muss sichergestellt werden, dass es *eine eindeutige* Lösung gibt, da die geschätzten Werte ansonsten beliebig sind und keine Aussagekraft besitzen. Während es in der Literatur einige Verfahren gibt, die eine Identifikation sichern, soll an dieser Stelle auf eine eher praktische Regel verwiesen werden, die bereits im Kapitel 1.3 besprochen wurden. Es erscheint jedoch ratsam, insbesondere auf den Aspekt der Skalierung der latenten Variablen einzugehen, da dieser neben seiner Bedeutung für die Identifizierbarkeit des Modells auch eine wesentliche inhaltliche Bedeutung für die Modellformulierung hat.
Ist die Skalierung der latenten Variablen nicht vorgenommen worden, kann nicht zwischen der Varianz des latenten Konstruktes und den Ladungen der Indikatoren auf diesen differenziert werden. Die Elemente der Faktorladungsmatrix Λ und die Diagonalelemente von Φ können nicht gleichzeitig geschätzt werden.
Im folgenden Beispiel bezeichnet x eine Indikatorvariable[64] und λ die Faktorladung auf die latente Variable ξ. Damit ergibt sich folgende Faktorgleichung:[65]

$$x = \lambda\xi + \delta \tag{107}$$

Multipliziert man diese Gleichung mit sich selbst[66] und wendet auf dieses Produkt den Erwartungswertoperator an, dann folgt:

$$xx = \lambda^2\xi\xi + 2\lambda\xi + \delta\delta \tag{108}$$

$$E(xx) = \lambda^2 E[\xi\xi] + 2E[\lambda\xi] + E[\delta\delta] \tag{109}$$

$$VAR(x) = \lambda^2 VAR(\xi) + VAR(\delta) \tag{110}$$

Dabei wurde angenommen, dass die latente Variable und der unique Faktor nicht korrelieren, wodurch der entsprechende Erwartungswert gleich Null wird. Um zu zeigen, warum die Faktorladung und die Varianz nicht identifiziert sind, nimmt man

[63] die in der Praxis als Schätzung für Σ dient
[64] und damit keinen Vektor
[65] Vgl. Long (1983a): 50ff
[66] mit der ersten binomischen Formel: $(a+b)^2 = a^2 + 2ab + b^2$

nun eine zweite latente Variable an, die mit ξ^* bezeichnet wird. Diese unterscheidet sich von ξ durch ihre Skalierung von ξ: $\xi^* = \alpha\xi$ mit $\alpha \neq 0 \wedge 1$. Long nennt hier beispielsweise Dollar und Cent, die sich durch $\alpha = 100$ unterscheiden, da ein Dollar 100 Cent entspricht. Diesem Skalierungsunterschied wird durch die entsprechende Faktorladung Rechnung getragen: $\lambda^* = \frac{\lambda}{\alpha}$. Es resultiert also folgende Faktorgleichung, die aber mit der ursprünglichen (107)identisch ist:

$$x = \lambda^*\xi^* + \delta \tag{111}$$

$$x = \frac{\lambda}{\alpha}(\alpha\xi) + \delta \tag{112}$$

$$x = \lambda\xi + \delta \tag{113}$$

Die Zerlegung der Gleichung 111 in die einzelnen Varianzbestandteile erfolgt analog zu dem obigen Vorgehen. Dadurch lässt sich zeigen, dass auch die Varianzen identisch sind:

$$VAR(x) = \lambda^{*2}VAR(\xi^*) + VAR(\delta) \tag{114}$$

$$VAR(x) = (\frac{\lambda}{\alpha})^2VAR(\alpha\xi) + VAR(\delta) \tag{115}$$

$$VAR(x) = \frac{\lambda^2}{\alpha^2}\alpha^2VAR(\xi) + VAR(\delta) \tag{116}$$

$$VAR(x) = \lambda^2VAR(\xi) + VAR(\delta) \tag{117}$$

Damit kann unterschieden werden, ob x durch ξ und δ mit λ oder mit ξ^* und δ sowie der Faktorladung λ^* erzeugt wurde. Dieses Problem könnte durch entsprechende Restriktionen vermieden werden.

6.3 Schätzung

An dieser Stelle wird die Schätzung der unbekannten Parameter vorgestellt. Im Rahmen des nächsten Kapitels wird auf eine derartige Darstellung verzichtet, da das Vorgehen zu dem hier präsentierten analog verläuft.
Bevor auf die technischen Details eingegangen wird, soll zunächst die intuitive Idee des Schätzverfahrens benannt werden: Ziel ist es, die unbekannten Parameter so zu schätzen, dass sie die Stichprobenkovarianzmatrix *bestmöglich* reproduzieren. Den Grad, in dem dieses Ziel erreicht ist, also die Differenz minimal wird, geben die *Fitting Functions* an. Es wurde bereits gezeigt, das die Populationskovarianzmatrix mit den gesuchten Populationsparametern durch die Kovarianzgleichung 106 verbunden ist. Die Elemente dieser Gleichung werden durch die folgenden Verfahren geschätzt. Für diese Schätzungen schreibt man:[67]

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Phi}\hat{\Lambda}^T + \hat{\Theta} \tag{118}$$

Ziel ist nun, die Werte der unbekannten Parameter so zu bestimmen, das die Differenz zwischen S und $\hat{\Sigma}$ minimal wird. Die Menge der möglichen Werte wird durch die

[67] Vgl. Long (1983a): 56ff

Restriktionen eingeschränkt. Die verbleibenden zulässigen Werte werden durch einen Asterisk gekennzeichnet:

$$\Sigma^* = \Lambda^* \Phi^* \Lambda^{T*} + \Theta^* \tag{119}$$

Für eine Fitting Function die den möglichst minimalen Abstand gewissermaßen operationalisiert schreibt man $F(S; \Sigma^*)$ oder auch $F(S; \Lambda^*, \Phi^*, \Theta^*)$. Der Wert dieser Funktion soll minimal werden.
Die folgenden Kapitel beschreiben einige der bedeutsamsten Fitting Functions oder Diskrepanzfunktionen.

6.3.1 Maximum-Likelihood(ML)-Diskrepanzfunktion

Die ML-Diskrepanzfunktion wird durch folgende Formel angegeben:[68]

$$F_{ML} = log||\Sigma^*|| + tr(S\Sigma^{-1}) - log||S|| - (p+q) \tag{120}$$

$||\Sigma^*||$ bzw. $||S||$ bezeichnen die Determinanten der modellimplzierten (geschätzten) Kovarianzmatrix bzw. der Stichprobenkovarianzmatrix. Damit ist bereits jetzt klar, dass keine dieser Determinanten kleiner oder gleich Null sein darf. Für Werte kleiner als Null ist ein Logarithmus nicht definiert und aus einer Determinanten von Null können wir auf lineare Abhängigkeit der Spaltenvektoren einer Matrix schließen. Die Funktion $tr(\ldots)$ heißt Spur einer Matrix und bezeichnet die Summe der Diagonalelemente. p und q geben die Anzahl der manifesten Variablen x und y an. Im Rahmen der konfirmatorischen Faktorenanalyse ist p jedoch gleich Null, da keine strukturellen Beziehungen zwischen den latenten Variablen modelliert werden. Die Funktion ist dann gleich Null wenn die modellimplizierte Matrix Σ^* genau der empirischen Matrix S entspricht. Ist das Modell überidentifiziert, so muss es iterativ gelöst werden. Ein Minimum ist dann gefunden, wenn die ersten Ableitungen der Elemente der zu schätzenden Parametervektoren und -matrizen gleich Null sind und es möglich ist, die zweiten Ableitungen zu berechnen.
Long verdeutlicht das Prinzip dieser Diskrepanzfunktion.[69] Betrachtet man zunächst die Spur des Matrizenproduktes so stellt man fest, je ähnlicher sich die Matrizen werden, dass sich das Produkt einer Einheitsmatrix nähert. Die Spur einer Einheitsmatrix entspricht aber gerade der Anzahl q der manifesten Variablen. Ebenso nähern sich mit schwindender Divergenz die beiden Logarithmen an, wobei deren Differenz Null wird, wenn Σ^* gleich S ist. Daher resultiert bei identischen Matrizen eine Funktionswert von Null.
Der ML-Schätzer zeichnet sich durch seine asymptotische Konsistenz und Effizienz aus, wenn angenommen werden kann, dass die Indikatorvariablen multivariat normalverteilt sind. Das bedeutet, dass je größer die Stichprobe ist, der *wahre* Wert immer besser geschätzt und die Varianz des Schätzer kleiner wird sowie die Verteilung des Schätzers zunehmend einer Normalverteilung folgt.[70] Zudem ist der ML-Schätzer skaleninvariant, wodurch die Berechnung auf Grundlage von Korrelationen zu denselben Ergebnissen führt, wie auf der Basis von Kovarianzen. Eine Verletzung der Verteilungsannahmen verlangt die Berechnung *robuster* Teststatistiken.[71]

[68] Vgl. Reinecke (2005): 109-110, wobei die Notation angepasst wurde
[69] Vgl. Long (1983a): 59
[70] Vgl. Long (1983a): 59
[71] Vgl. Reinecke (2005): 109-110

6.3.2 Unweighted-Least-Squares(ULS)-Diskrepanzfunktion

Die ULS-Diskrepanzfunktion minimiert die Quadratsummen der Elemente in der Residualmatrix. Damit wird intuitiv die Nähe des Verfahrens zur OLS-Schätzmethode klar, die im Rahmen der linearen Regression vorgestellt wurde. Formal drückt sich dies so aus:[72]

$$F_{ULS} = \frac{1}{2} tr(S - \Sigma^*)^2 \tag{121}$$

Die Differenz zwischen S und Σ^* ergibt die Residualmatrix der Varianzen und Kovarianzen. Sind beide Matrizen identisch, so nimmt die Diskrepanzfunktion den Wert Null an, wie unmittelbar ersichtlich wird, wenn man sich vergegenwärtigt, das die Spur einer Matrix, die lediglich Nullen enthält, auch gleich Null ist.
Die ULS-Diskrepanzfunktion setzt keine Verteilungsannahmen an die manifesten Variablen voraus und dabei zu konsistenten, aber nicht effizienten Schätzern. Ein weiterer wesentlicher Nachteil ist die Abhängigkeit der Schätzungen von der Skalierung der Variablen. Daher bietet es sich an derartige Schätzungen nur auf Basis von Korrelationen durchzuführen, da diese dimensionslos sind. Zudem sind die Teststatistiken mit Vorsicht zu betrachten, da diese zumeist von einer Normalverteilung ausgehen.[73]

6.3.3 Generalized-Least-Squares(GLS)-Diskrepanzfunktion

Mit der GLS-Diskrepanzfunktion wird die Funktion 122 verallgemeinert. In dieser wird unterstellt, dass alle Elemente der Residualmatrix die gleiche Streuung besitzen. Diese Annahme wird nun aufgegeben, indem die Residualmatrix mit einer Gewichtungsmatrix (W) multipliziert wird:[74]

$$F_{WLS} = \frac{1}{2} tr[(S - \Sigma^*)W^{-1}]^2 \tag{122}$$

Zumeist ist $W^{-1} = S^{-1}$. Es wird also mit der inversen der Stichprobenmatrix gewichtet. Das Prinzip der Schätzung ist analog zur ULS-Diskrepanzfunktion zu verstehen. GLS-Schätzer haben die gleichen asymptotischen Eigenschaften wie ML-Schätzer. Außerdem sind sie skaleninvariant.

6.3.4 Weighted-Least-Squares(WLS)-Diskrepanzfunktion

In der Praxis ist die Annahme einer Multinormalverteilung häufig nicht haltbar. Dies ist zum Beispiel auf extrem schiefe oder gewölpte Verteilungen zurückzuführen. Ebenso sollte bei manifesten Variablen mit kategorialem Skalenniveau auf die WLS-Diskrepanzfunktion zurückgegriffen werden. Aus einer Verletzung der Verteilungsannahme können verzerrte Parameterschätzungen, Standardfehler, z-Werte und Test resultieren. Die WLS-Methode ist dabei eine mögliche Option im Umgang mit derartigen Daten.[75]

$$F_{WLS} = [s - \sigma^*]^T W^{-1} [s - \sigma^*] \tag{123}$$

[72] Vgl. Reinecke (2005): 110, wobei die Notation angepasst wurde
[73] Vgl. Reinecke (2005): 110
[74] Vgl. Reinecke (2005): 110, wobei die Notation angepasst wurde
[75] Vgl. Reinecke (2005): 111ff, wobei die Notation angepasst wurde

s bezeichnet den Vektor der nicht redundanten Elemente der empirischen Kovarianzmatrix und σ steht für die entsprechenden Elemente der modellimplizierten Kovarianzmatrix. In der praktischen Arbeit dient die *asymptotische Varianz-Kovarianzmatrix* als Gewichtungsmatrix. Dabei berechnet sich die asymptotische Kovarianz zwischen den empirischen Kovarianzen nach folgender Formel:[76]

$$ACOV(s_{ij}, s_{gh}) = N^{-1}(\sigma_{ijgh} - \sigma_{ij}\sigma_{gh}) \tag{124}$$

$$\text{mit} \quad \hat{\sigma}_{ijgh} = \frac{1}{N}\sum(x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)(x_{gt} - \bar{x}_g)(x_{ht} - \bar{x}_h) \tag{125}$$

$$\text{und} \quad \hat{\sigma}_{ij} = \frac{1}{N}\sum(x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j) \tag{126}$$

$$\text{und} \quad \hat{\sigma_{gh}} = \frac{1}{N}\sum(x_{gt} - \bar{x}_g)(x_{ht} - \bar{x}_h) \tag{127}$$

Die asymptotische Kovarianz ergibt sich demnach aus der Differenz zwischen dem vierten Moment σ_{ijgh} und dem Produkt der Populationskovarianzen.
Dadurch nutzt die WLS-Diskrepanzfunktion die Verteilungsinformationen, die durch die asymptotische Kovarianzmatrix bereitgestellt werden und muss nicht auf etwaige Verteilungsannahmen zurückgreifen. Mit ihr ist es möglich, Teststatistiken zu berechnen, die χ^2 zur Grundlage haben. Auch die Standardfehler sind dann problemlos zu berechnen. Nachteilig ist die Erfordernis einer großen Stichprobe und der gesteigerte Rechenaufwand.

6.4 Modellprüfung

Bisher wurde der formale Aufbau eines konfirmatorischen Faktorenmodells beschrieben und grundsätzliche Schätzmethoden besprochen. Wurde das Modell aufgestellt und geschätzt gilt es aber die Anpassung des Modells zu prüfen. Eine Auswahl der entsprechenden Statistiken und Verfahren werden nun präsentiert. Die formale Darstellung ist nicht erschöpfend, vielmehr geht es darum, grundlegende Prinzipien zu verdeutlichen.[77] Primär sind zwei sind zwei Situationen zu unterscheiden: (1) Modellevaluation und (2) Modellvergleich. Punkt (1) wird in den Abschnitten 6.4.1, 6.4.2, 6.4.3, 6.4.4, 6.4.5 sowie 6.4.6 und Punkt (2) in den Abschnitten .... behandelt. Bei der Beurteilung der Modellanpassung anhand der dargestellten Maße ist Vorsicht geboten, da zu beachten ist, dass diese nicht die Möglichkeit äquivalenter Modelle berücksichtigen. Insofern sollten zusätzlich Alternativmodelle miteinander verglichen werden.

6.4.1 χ^2-Statistik

Die χ^2-Statistik prüft, ob die modellimplizierte Kovarianzmatrix mit der Populationskovarianzmatrix übereinstimmt, woraus sich folgende Nullhypothese ergibt: $H_0 : \Sigma = \Sigma^*$. Da Σ unbekannt ist, muss sie durch die Stichprobenkovarianzmatrix

[76] Vgl. Reinecke (2005): 112

[77] dabei wird nur auf die formale Darstellung entsprechender Maße für die ML-Methode eingegangen

geschätzt werden. Das heißt, dass man in der Nullhypothese davon ausgeht, das die geschätzte Kovarianzmatrix nicht (statistisch signifikant) von der empirischen Matrix abweicht. In der Alternativhypothese geht man dagegegen davon aus, dass die geschätzte statistisch signifikant von der empirischen Matrix abweicht. Die χ^2-Statistik kann dabei Werte von Null für saturierte Modelle bis zu einem maximalen Wert bei einem vollkommen unabhängigen Modell. Insbesondere ist er dann gleich Null, wenn also beide Matrizen identisch sind. Ein saturiertes Modell ist damit nicht anzustreben, da dieses immer einer perfekten Fit aufweist.[78] Der χ^2-Wert ergibt sich, indem man den minimalen Wert der Diskrepanzfunktion mit dem um Eins verminderten Stichprobenumfang multipliziert. Damit wird auch schon ein erster Kritikpunkt offensichtlich. χ^2 ist offensichtlich abhängig von der Stichprobe, wodurch zumeist ein signifikanter Wert ausgegeben wird. Zudem ist dieser Test an eine Reihe von sehr restriktiven Voraussetzungen geknüpft:[79]

1. Multinormalverteilung der manifesten Variablen bzw. eine optimale Gewichtungsmatrix: Schiefe oder auch gewölbte Verteilungen können zu verzerrten Standardfehlern, χ^2- und z-Werten führen. Es gibt eine Reihe von korrigierten χ^2-Statistiken, die einen korrigierten Wert ausweisen.

2. Großer Stichprobenumfang: Eine eindeutige Empfehlung kann hierzu nicht ausgesprochen werden, jedoch sollte die Stichprobe mehr als 100 Beobachtungen enthalten um zulässige χ^2-Werte aus einer ML-Schätzung zu erhalten.

3. Die Schätzung erfolgt auf Grundlage einer Kovarianzmatrix

4. Die Nullhypothese ($H_0 : \Sigma = \Sigma^*$) geht von einer exakten Übereinstimmung aus, die aber in der Regel nicht erfüllt ist.

6.4.2 Root Mean Squared Error of Approximation (RMSEA)

Der RMSEA misst die durchschnittliche Abweichung zwischen der Populationskovarianzmatrix und der Matrix, die als an diese bestangepasste Matrix gilt. Die von Browne und Cudecks enwickelte *population discrepancy function* $\hat{F}_0$ bestimmt diese Abweichung:

$$RMSEA = \sqrt{\frac{\hat{F}_0}{df}} \tag{128}$$

$\hat{F}_0$ stellt den minimalen Wert der Diskrepanzfunktion dar, der dementsprechend gleich Null ist, wenn die beiden Matrizen identisch sind. Folglich hat der RMSEA einen unteren Grenzwert von Null. Da es möglich ist ein Vertrauensintervall zu bestimmen, kann folgende Nullhypothese getestet werden: $H_0 : RMSEA \leq 0,05$. Für einen Wert kleiner bzw. gleich 0,05 spricht man von einer guten Anpassung. Ein Wert zwischen 0,05 und 0,08 weist ein akzeptable Anpassung aus.[80]

[78] Vgl. Schumacker und Lomax (2004): 82
[79] Vgl. Reinecke (2005): 116-118
[80] Vgl. Reinecke (2005): 119-120

6.4.3 Goodness of Fit Index (GFI)

Der GFI gibt den Anteil der Varianz von San, der durch das Kausalmodell, also die modellimplizierte Kovarianzmatrix erklärt wird. In der Notation von Reinecke[81] berechnet sich der GFI nach folgender Formel:

$$GFI_{ML} = 1 - \frac{tr[(\Sigma^{*-1}S - I)^2]}{tr[(\Sigma^{*-1}S)^2]} \tag{129}$$

Langer bietet hierzu eine allgemeinere und alternative Notation [82]:

$$GFI = 1 - \frac{F[S, \Sigma^*]}{F[S, \Sigma^0]} \tag{130}$$

Damit wird das geschätzte Modell zum Nullmodell ins Verhältnis gesetzt. Je schlechter demnach die Anpassung des Nullmodells ist, desto kleiner wird der Bruch, wenn man den Zähler konstant hält. Genauso gilt, dass je geringer die Abweichung der Schätzung von der Stichprobe ist, desto kleiner wird der Bruch, wenn der Nenner konstant gehalten wird. Der GFI hat einen Wertebereich von Null bis Eins, wobei Eins bedeutet, dass das Modell 100% der Varianz erklärt. Der GFI sollte für eine gute Anpassung größer oder gleich 0,95 sein.

6.4.4 Adjusted Goodness of Fit Index (AGFI)

Der AGFI korrigiert den GFI um die Freiheitsgrade des Modells. Dabei berechnet er sich wie folgt:[83]

$$AGFI_{ML} = 1 - [\frac{(p+q)(p+q+1)}{2df}](1 - GFI_{ML}) \tag{131}$$

Wie bereits erwähnt, ist p im Falle der konfirmatorischen Faktorenanalye gleich Null. Das Modell weist dann eine gute Anpassung auf, wenn der AGFI größer oder gleich 0,90 ist. Auch er bewegt sich zwischen Null und Eins.

6.4.5 Root Mean Square Residuals (RMR)

Der RMR ist eine Statistik, die die Abweichungen der Elemente der modellimplizierten Kovarianzmatrix von der empirischen Kovarianzmatrix zusammenfasst. Die Elemente der Residualmatrix sind positiv, wenn die entsprechende Kovarianz durch das Modell unterschätzt wurde. Dementsprechend sind sie negativ, bei einer Überschätzung der Kovarianz.[84]

$$RMR = \left[2 \sum_{i=1}^{p+q} \sum_{j=1}^{i} \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{(p+q)(p+q+1)} \right]^{\frac{1}{2}} \tag{132}$$

[81] Vgl. Reinecke (2005): 121, wobei die Notation angepasst wurde
[82] Vgl. Langer (o.A.): 6, wobei die Notation angepasst wurde
[83] Vgl. Reinecke (2005): 121
[84] Vgl. Reinecke (2005): 122-123

Der Zähler des Bruches gibt die quadrierten Abweichungen an und setzte diese ins Verhältnis zu den manifesten Variablen. Die durchschnittlichen Residuen sind umso geringer, je kleiner der RMR ist. Problematisch ist jedoch, dass dieses Maß von der Skalierung der Variablen und dem Stichprobenumfang beeinflusst wird.

6.4.6 Standardized Root Mean Square Residuals (SRMR)

Um die Nachteile des RMR auszugleichen wurde der SRMR entwickelt. Hier werden die Residuen $s_{ij} - \hat{\sigma}_{ij}$ zunächst durch die Standardabweichungen der entsprechenden manifesten Variablen dividiert, woraus standardisierte Residuen resultieren. Der SRMR weist einen Wert von Null aus, sofern das Modell perfekt angepasst ist. Ein Wert von 0,05 oder kleiner gilt als Indikator für einen guten Fit. Akzeptabel sind Werte von 0,10 und kleiner.

6.4.7 Likelihood-Ratio(LR)-Test

Dieser und die folgenden Indizes dienen dem Modellvergleich. Der LR-Test, auch χ^2-Differenztest genannt, bildet die Differenz der χ^2-Werte eines restringierten F_r und eines weniger oder nicht restringierten Modells F_u.[85] Voraussetzung für die Anwendung des Tests ist, das es sich um *hierarchisch geschachtelte (nested)* Modelle handelt, dabei geht ein Modell durch die Festsetzung eines Parameters aus dem anderen hervor.

$$LR = (N-1)(F_r - F_u) \tag{133}$$

$$LR = (N-1)F_r - (N-1)F_u \tag{134}$$

$$LR = \chi_r^2 - \chi_u^2 \tag{135}$$

Diese Differenz ist wiederum χ^2-verteilt. Das unrestringiertere Modell weist dabei weniger Freiheitsgrade auf, da mehr Parameter zur Schätzung freigegeben sind. Ist die Differenz signifikant, dann sollte das weniger restringierte Modell bevorzugt werden, da durch die freie Schätzung des Parameters eine bessere Anpassung erreicht wird.

6.4.8 Lagrange Multiplier (LM)-Test

Der LM-Test zeigt die erwartete Verbesserung der χ^2-Statistik an, die bei der Fixierung eines weiteren Parameters im Modell eintritt. Damit ist er im univariaten Fall identisch mit dem LR-Test bei einem Freiheitsgrad. Der LM-Test wird durch Lisrel als *modification index* bereitgestellt. Es sollte darauf geachtet werden, dass man bei der Spezifikationssuche jedoch nicht ausschließlich diesen Empfehlungen folgt, sondern die Modellmodifikation theoriegeleitet durchführt. Zudem sollte jeweils nur ein Parameter fixiert werden, da Lisrel nur die univariate Version des LM-Tests bietet und der Effekt einer Festlegung mehrerer Parameter vorab nicht zu bestimmmen ist.[86]

[85] Vgl. Reinecke (2005): 123-124

[86] Vgl. Reinecke (2005): 124

6.4.9 Normed Fit Index (NFI) Nonnormed Fit Index (NNFI)

Entscheidend für dieses und die folgenden Maße sind die Begriffe *Nullmodell* oder auch *independence model*. Hierdurch wird ein Modell bezeichnet, in dem die Kovarianzen Null sind. Die Indikatoren also untereinander nicht korrelieren. In einem solchen Modell sind damit nur die Varianzen der Indikatoren zu schätzen, alle anderen Parameter sind auf den Wert Null fixiert.
Der NFI und der NNFI berechnen sich nach folgenden Formeln:[87]

$$NFI = \frac{\chi^2_{null} - \chi^2_{model}}{\chi^2_{null}} \tag{136}$$

$$NFI = \frac{\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{model}}{df_{model}}}{\frac{\chi^2_{null}}{df_{null}} - 1} \tag{137}$$

$$\tag{138}$$

Der NFI nimmt Werte zwischen Null und Eins an. Da er jedoch von der Größe der Stichprobe abhängig ist, kann dies zu Situationen führen, in denen der Wert Eins, selbst bei korrekter Spezifikation nicht erreicht wird. Man spricht von einem guten Fit, wenn der NFI einen Wert von 0,95 oder größer erreicht. Eine akzeptable Modellanpassung wird durch einen Wert ab 0,90 angezeigt. Der NNFI korrigiert den NFI um die Freiheitsgrade des jeweiligen Modells. Er ist auch unter der Bezeichnung TLI, Tucker-Lewis Index, bekannt. Auch dieses Maß nimmt gewöhnlich Werte zwischen Null und Eins an. Die Grenzwerte für eine gute bzw. akzeptable Modellanpassung lauten 0,97 bzw. 0,95. Er wird weniger durch die Stichprobe beeinflusst, bestraft dafür aber komplexe Modelle, mit vielen frei zu schätzenden Parametern.

6.4.10 Comparative Fit Index (CFI)

Der CFI berechnet sich nach folgender Formel:[88]

$$CFI = 1 - \left[\frac{(\chi^2_{model} - df_{modell}}{\chi^2_{null} - df_{null}}\right] \tag{139}$$

Damit basiert er auf dem RNI von McDonald und Marsh, vermeidet dabei aber eine Unterschätzung der Anpassung in kleinen Stichproben. Er kann Werte zwischen Null und Eins annehmen. Liegt ein Wert nahe Eins, bedeutet dies eine bessere Modellanpassung. Als Schwellenwerte sind 0,97 bzw. 0,95 zu nennen.

6.4.11 Parsimony Goodness of Fit Index (PGFI) Parsimony Normed Fit Index (PNFI)

Dieser und die folgenden Indizies dienen der Beurteilung der Modellsparsamkeit. Modellsparsamkeit bezeichnet dabei den Versuch, ein Modell so zu formulieren, dass

[87] Vgl. Schumacker und Lomax (2004): 83
[88] Vgl. Schumacker und Lomax (2004): 84

es mit möglichst wenigen zu schätzenden Koeffizienten eine möglichst gute Modellanpassung erreicht.
Aufbauend auf dem GFI und NFI werden diese beiden Indizes wie folgt definiert:[89]

$$PGFI = \frac{df_{modell}}{df_{null}} GFI \tag{140}$$

$$PNFI = \frac{df_{modell}}{df_{null}} NFI \tag{141}$$

Auch hier variieren die Maße zwischen Null und Eins. Die Werte tendieren umso mehr zur Eins, desto sparsamer das Modell ist. Mit anderen Worten, ein Modell, dass einen guten Modellfit, der durch den GFI oder NFI angezeigt wird, aufweist und dabei mehr Freiheitsgrade besitzt, tendiert gegen einen Wert von Eins.

6.4.12 Akaike Information Criterion (AIC) Consistent Akaike Information Criterion (CAIC)

Mit dem AIC und dem CAIC ist es möglich auch Modelle miteinander zu vergleichen, die in einer nicht hierarchischen Relation stehen. Sie sind als rein deskriptive Maße zu verstehen.[90]

$$AIC = \chi^2_{model} + 2t \tag{142}$$

$$CAIC = \chi^2_{model} + (1 + lnN)t \tag{143}$$

t entspricht der Anzahl der frei zu schätzenden Parameter und N dem Stichprobenumfang. Es ist dann das Modell vorzuziehen, das den niedrigeren AIC- bzw. CAIC-Wert aufweist. Dabei werden zum Einen Anpassungsfehler (χ^2) und zum Anderen Schätzfehler (t) berücksichtigt. Mit der Anzahl von freien Parametern steigt auch der AIC bzw. CAIC, womit Modellsparsamkeit „belohnt "wird. Der CAIC berücksichtigt darüber hinaus die Komplexität des Modells. Consistent bezieht sich auf die damit verbundene Annahme, dass das „wahre "Modell gewählt wird, wenn die Stichprobengröße gegen unendlich strebt. Diese Annahme ist jedoch aus verschiedenen Gründen problematisch. Infolgedessen sollte das AIC bevorzugt werden.

6.4.13 Expected Cross Validation Index (ECVI)

Für kleine Stichproben sollte der ECVI herangezogen werden, um Modell miteinander zu vergleichen. Im Gegensatz zum AIC und CAIC lässt dieser die Berechnung eines Konfidenzintervalls zu.[91]

$$ECVI = \frac{\chi^2_{model}}{N} + \frac{2t}{N} \tag{144}$$

Bei einem Modellvergleich sollte das Modell gewählt werden, welches den niedrigeren ECVI aufweist.

[89] Vgl. Reinecke (2005): 126, wobei die Notation angepasst wurde
[90] Vgl. Reinecke (2004): 128, wobei die Notation angepasst wurde
[91] Vgl. Reinecke (2004): 128, wobei die Notation angepasst wurde

6.5 Beispiel - Einfaches konfirmatorisches Faktorenmodell

In diesem Beispiel geht es zunächst darum, den praktischen Aufbau eines einfachen Messmodells für zwei latente Variablen kennen zu lernen. Zu diesem Zweck wird sich in diesem Kapitel eines Modells bedient, welches in der LISREL-Hilfe selbst als Beispiel angegeben ist. In den Ausführungen wird jedoch deutlich von der LISREL-Hilfe abgewichen.[92] Das Datenmaterial ist einer Studie von Holzinger und Swineford aus dem Jahr 1939 entnommen, in der Daten über 26 psychologische Tests an 145 Schulkinder der siebten und achten Klasse einer Chicagoer Schule gesammelt wurden. Mit sechs dieser Tests wurde die empirische Grundlage für dieses Beispiel geschaffen. Gemessen werden im Folgenden zwei latente Konstrukte (ξ_1, ξ_2): *visual perception* und *verbal ability*. Hierzu werden jeweils drei Indikatoren verwendet. Während die LISREL-Hilfe auf den Modellaufbau mit Hilfe eines Pfaddiagramms eingeht, wird hier ein eher formalerer Zugang gewählt, um auf die vorherigen Inhalte Bezug nehmen zu können. Im Modell werden die Indikatoren (x) mit *vis perc*, *cubes*, *lozenges*, *par comp*, *sen comp* und *wordmean* bezeichnet. Ohne auf die nähere theoretische Bedeutung einzugehen, wird nun der formale Modellbau entsprechend der vorgestellten Notation präsentiert. Hierzu wird Gleichung 92 angepasst:

$$\underbrace{\begin{pmatrix} vis\ perc \\ cubes \\ lozenges \\ par\ comp \\ sen\ comp \\ wordmean \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{pmatrix}}_{\Lambda} \cdot \underbrace{\begin{pmatrix} visual\ perception \\ verbal\ ability \end{pmatrix}}_{\xi} + \underbrace{\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{pmatrix}}_{\delta} \tag{145}$$

Die Matrix der Kovarianzen und Varianzen der latenten Konstrukte weist die folgende Form auf:

$$\Phi = \begin{pmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{pmatrix} \tag{146}$$

Zudem muss noch die Matrix der Messfehlervarianzen formuliert werden:

$$\Theta_\delta = \begin{pmatrix} \theta_{\delta_1} & & & & & \\ 0 & \theta_{\delta_2} & & & & \\ 0 & 0 & \theta_{\delta_3} & & & \\ 0 & 0 & 0 & \theta_{\delta_4} & & \\ 0 & 0 & 0 & 0 & \theta_{\delta_5} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{\delta_6} \end{pmatrix} \tag{147}$$

Durch das Einsetzen der so definierten Matrizen in die Gleichung 106 kann das theoretische Modell in eine schätzbare Form gebracht werden, indem die Populationskovarianzmatrix so zerlegt wird, dass sie in Abhängigkeit der gesuchten Parameter formuliert werden kann. Die Elemente der Matrix Θsind unterhalb und oberhalb der Hauptdiagonalen Null, da zunächst davon ausgegangen wird, dass die Messfehler

[92] Vgl. LISREL-Hilfe: *Example: confirmatory factor analysis*

untereinander unkorreliert sind.
Nun ist mittels der t-Regel die Identifizierbarkeit des Modells zu überprüfen. Bei sechs manifesten Variablen enthält die empirische Kovarianzmatrix S 21 Elemente, von denen 15 Kovarianzen und 6 Varianzen sind. Da eine zweite notwendige Bedingung zur Modellidentifikation die Skalierung der latenten Variablen ist, werden die Faktorladungen λ_{11} und λ_{41} auf den Wert Eins fixiert. Damit sinkt die Anzahl der unbekannten Modellparameter auf 13: vier Faktorladungen, drei Elemente der Matrix Φ und sechs Elemente der Matrix Θ_δ. Es folgt somit, dass t gleich 13 und damit kleiner als die Anzahl der bekannten Werte ist. Das Modell weist demzufolge acht Freiheitsgrade auf und ist damit überidentifiziert.[93]
Da in diesem Beispiel keinerlei Rohdaten gegeben sind, können die Verteilungseigenschaften der Variablen nicht überprüft werden. Jedoch zeigte ein Vergleich der Schätzergebnisse unter Verwendung einer ML-Schätzung mit denen, die aus einer Korrektur der ML-Schätzung mit der asymptotischen Varianz-Kovarianzmatrix resultierten, keine Abweichungen. Demnach wird in der folgenden Syntax die Schätzmethode nicht weiter spezifiziert, da die ML-Diskrepanzfunktion standardmäßig von LISREL verwendet wird. Die Syntax lässt sich wie folgt formulieren:

!Six Psychological Variables-A Confirmatory Factor Analysis Titel
Observed variables: manifeste Variablen
'VIS PERC' CUBES LOZENGES 'PAR COMP' 'SEN COMP' WORDMEAN
Correlation Matrix: Korrelationsmatrix
1.000
0.318 1.000
0.436 0.419 1.000
0.335 0.234 0.323 1.000
0.304 0.157 0.283 0.722 1.000
0.326 0.195 0.350 0.714 0.685 1.000
Sample Size: 145 Stichprobengröße
Latent Variables: Visual Verbal latente Variablen
Equations: Modellgleichungen
'VIS PERC' - LOZENGES = Visual erste latente Variable
'VIS PERC' = 1*Visual Zur Skalierung auf Eins fixiert
'PAR COMP' - WORDMEAN = Verbal zweite latente Variable
'PAR COMP' = 1*Verbal Zur Skalierung auf Eins fixiert
Number of decimals = 4 Angabe der Dezimalstellen
LISREL Output: MI Zusätzliche Outputoptionen, mit MI werden die Modifikationsindizes angefordert
Path Diagram Anforderung eines Pfaddiagramms
End of Problem Ende der Syntax

Nach Eingabe und Ausführung der Befehlssyntax gibt LISREL folgendes Pfaddiagramm aus, welches die standardisierten Schätzungen der jeweiligen Koeffizienten wiedergibt:
Bereits durch das Pfaddiagramm werden erste Teststatistiken ausgegeben. Neben

[93] $df = 21 - 13 = 8$

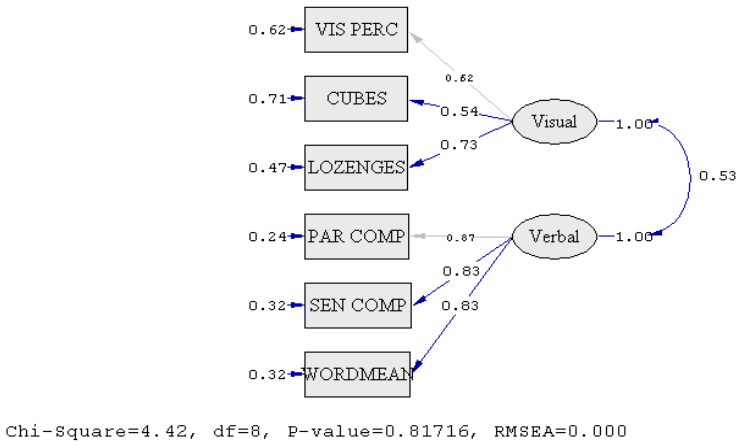

Abbildung 6: Pfaddiagramm: Konfirmatorische Faktorenanalyse mit zwei latenten Variablen

dem χ^2-Wert wird der zugehörige p-Wert und der RMSEA ausgewiesen. χ^2 deutet mit einem Wert von 4,42 auf eine statistisch nicht signifikante Alternativhypothese hin, so dass eine gute Modellanpassung angenommen werden kann. Auch der signifikante[94] RMSEA mit einem Wert von 0,000 weist eine gute Anpassung aus. Auch der GFI (0,9899), der AGFI (0,9734) und der SRMR-Index (0,02931) weisen in die gleiche Richtung.
Im Anhang **??** ist der gesamte Output enthalten. Unter den geschätzten Parametern werden dabei jeweils zwei Werte angezeigt. In Klammern steht der geschätzte Standardfehler, unter dem der t-Wert angegeben wird. Die t-Werte können zur Prüfung der Parameter auf ihre Signifikanz genutzt werden. Es zeigt sich, dass alle geschätzten Parameter signifikant von Null verschieden sind. Die Modifikationsindizes deuten nicht darauf hin, dass das Modell durch die Freisetzung weiterer Parameter weiter verbessert werden könnte.

7 Kovarianzstrukturmodelle

Kovarianzstrukturmodelle verbinden die konfirmatorische Faktorenanalyse mit Pfadmodellen. Sie erlauben damit die Analyse der Beziehungen zwischen latenten Variablen. Um die Darstellung auf eine einheitliche Basis zu stellen, wird im Folgenden nochmals, wenn auch in gebotener Kürze, auf die grundlegende formale Darstellung des Mess- und Strukturmodells eingegangen.

7.1 Messmodell

In einem Kovarianzstrukturmodell unterscheidet man, wie im nächsten Kapitel dargestellt wird, unabhängige (ξ) und abhängige (η) latente Variablen. Demzufolge sind auch die Indikatoren zu unterscheiden. Mit x werden die manifesten Variablen der unabhängigen latenten Konstrukte und mit y die Indikatoren der abhängigen laten-

[94] der p-Wert für den Test der Nullhypothese: RMSEA $< 0{,}05$ beträgt 0,9243

ten Variablen bezeichnet.[95]
Damit weicht das folgende Messmodell, von dem in Kapitel 6 präsentierten, ab. Der Vektor ξ enthält s Faktoren, die durch q manifeste Variablen bestimmt werden. Diese Variablen sind im Vektor x enthalten. Die r abhängigen latenten Konstrukte werden durch den Vektor η abgebildet, wobei die zugehörigen Indikatoren im Vektor y zusammengefasst sind. Dementsprechend lassen sich die folgenden Faktorgleichungen formulieren:

$$x = \Lambda_x \xi + \delta \tag{148}$$

$$y = \Lambda_y \eta + \epsilon \tag{149}$$

Die Matrizen Λ_x und Λ_y beinhalten die jeweiligen Faktorladungen und sind von den Dimensionen (q,s) bzw. (p,r). Die einzelnen Faktorladungen von x_i auf ξ_j zum Beispiel werden durch λ_{ij}^x dargestellt. Messfehler bzw. unique Faktoren sind in den Vektoren δ bzw. ϵ enthalten. Es werden folgende Annahmen getroffen:

1. Es wird angenommen, dass die Variablen als Abweichungen von ihrem arithmetischen Mittel gemessen werden.
2. Die Faktoren und Messfehler sind unkorreliert.
3. Die Messfehler über die Gleichungen korrelieren nicht miteinander.
4. Die Faktoren und Messfehler sind über die Gleichungen unkorreliert.

Es gilt:

$$E(x) = E(\delta) = 0 \quad E(\xi) = 0 \tag{150}$$

$$E(y) = E(\epsilon) = 0 \quad E(\eta) = 0 \tag{151}$$

$$E(\xi\delta^T) = 0 \quad \text{bzw.} \quad E(\delta\xi^T) = 0 \tag{152}$$

$$E(\eta\epsilon^T) = 0 \quad \text{bzw.} \quad E(\epsilon\eta^T) = 0 \tag{153}$$

$$E(\delta\epsilon^T) = 0 \quad \text{bzw.} \quad E(\epsilon\delta^T) = 0 \tag{154}$$

$$E(\xi\epsilon^T) = 0 \quad \text{bzw.} \quad E(\epsilon\xi^T) = 0 \tag{155}$$

$$E(\eta\delta^T) = 0 \quad \text{bzw.} \quad E(\delta\eta^T) = 0 \tag{156}$$

Der Erwartungswert des Produktes des Vektors ξ mit seiner transponierten Form ergibt die (s,s)-Kovarianzmatrix Φ: $\Phi = E(\xi\xi^T)$. Die Kovarianzen der x-Messfehler sind in der Matrix Θ_{delta} enthalten, analog für ϵ Θ_ϵ in einer (q,q)-Matrix. Die Kovarianz der η wird durch die (r,r)-Matrix $COV(\eta)$ abgebildet. Da die manifesten Variablen korrelieren dürfen, muss ihre Kovarianz spezifiziert werden: $COV(x, y) = \Sigma_{xy}$ von der Ordnung (q,p). Die Kovarianz der Faktoren wird in der (s,r)-Matrix $COV(\xi, \eta)$. Im Rahmen der Modellspezifikation sind dann Restriktionen in den Parametermatrizen zu implementieren: $\Lambda_x, \Lambda_y, \Phi, \Theta_\delta$ und Θ_ϵ.
Damit können die Parameter jedoch noch nicht geschätzt werden. Hierzu ist es notwendig die beobachteten Kovarianzen und Varianzen mit den zu schätzenden Werten zu verbinden. Die beobachtete Populationsmatrix Σ ergibt sich wie folgt:

$$\Sigma = E\left[\begin{bmatrix}y\\x\end{bmatrix}\begin{bmatrix}y\\x\end{bmatrix}^T\right] = \left[\begin{array}{c|c} E(yy^T) & E(yx^T) \\ \hline E(xy^T) & E(xx^T)\end{array}\right] \tag{157}$$

[95] Vgl. Long (1983b): 19-25

Durch das Einsetzen von x und y sowie einiger Umformungen und der Umsetzung der Modellannahmen ergibt sich:

$$\Sigma = \left[\frac{E\left[(\Lambda_y\eta + \epsilon)(\Lambda_y\eta + \epsilon)^T\right] | E\left[(\Lambda_y\eta + \epsilon)(\Lambda_x\xi + \delta)^T\right]}{E\left[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \epsilon)^T\right] | E\left[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)^T\right]} \right] \tag{158}$$

$$\Sigma = \left[\frac{\Lambda_y COV(\eta)\Lambda_y^T + \Theta_\epsilon | \Lambda_y COV(\eta, \xi)\Lambda_x^T}{\Lambda_x COV(\xi, \eta)\Lambda_y^T | \Lambda_x \Phi \Lambda_x^T + \Theta_\delta} \right] \tag{159}$$

Damit wurde die Populationskovarianzmatrix derart zerlegt, dass sie eine Funktion der Faktorladungen, der Kovarianzen der latenten Variablen sowie der Messfehler ist.

7.2 Strukturgleichungsmodell

Zunächst wird aus Gründen der Vereinfachung angenommen, dass die latenten Variablen als beobachtet gelten. Diese Annahme wird im nächsten Kapitel wieder aufgegeben.[96]
Ein Strukturgleichungsmodell lässt die Analyse der (kausalen) Beziehungen zwischen Variablen zu. Endogene Variablen werden durch das Modell erklärt, indem sie von anderen endogenen oder exogenen Variablen abhängig sind. Wie auch schon weiter oben beschrieben wurde, ist η ein (r,1)-Vektor abhängiger Variablen und ξ einen (s,1)-Vektor exogener unabhängiger Variablen. Zwischen diesen wird eine lineare Beziehung angenommen:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{160}$$

Die (r,r)-Matrix B bildet die Beziehungen unter den endogenen Variablen ab. Γ ist eine (r,s)-Matrix deren Elemente den Einfluss der jeweiligen exogenen auf die entsprechende endogene Variable angeben. ζ ist der (r,1)-Vektor der Residuen. Die Elemente der Hauptdiagonalen der Matrix B sind gleich Null, da angenommen wird, dass eine abhängige Variable nicht auf sich selbst wirkt. Durch die Fixierung eines Elementes der beiden Matrizen auf Null kann festgelegt werden, das eine Variable auf die andere keinen Einfluss hat. Dabei wird erneut angenommen, da die Variablen in Abweichung zu ihrem arithmetischen Mittel gemessen werden. Die exogenen Variablen und die Messfehler korrelieren nicht miteinander. Es ist sinnvoll Gleichung 160 so umzustellen, das sie nach endogenen und exogenen Variablen sortiert ist:

$$\eta - B\eta = \Gamma\xi + \zeta \tag{161}$$

$$(I - B)\eta = \Gamma\xi + \zeta \tag{162}$$

$$\ddot{B}\eta = \Gamma\xi + \zeta \tag{163}$$

Für die weiteren Umformungen muss eine weitere Annahme getroffen werden, die fordert, dass $\ddot{B}$ invertierbar und somit regulär ist. Dies entspricht der Forderung, dass keine Gleichung im Modell redundant ist. Die Kovarianz der Residuen wird in einer (r,r)-Matrix Ψ abgebildet. Die der exogenen in der (s,s)-Matrix Φ. Um nun wiederum zu einer schätzbaren Form zu gelangen müssen folgende Umformungen

[96] Vgl. Long (1983b): 25-28

herangezogen werden.[97]Zunächst sollten auf der linken Seite der Gleichung 163 nur die endogenen Größen stehen. Hierzu wird die Gleichung mit der Inversen von $\ddot{B}$ multipliziert:

$$\eta = \ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\zeta \tag{164}$$

Diese Form wird als *reduzierte* Form bezeichnet. Aus den Annahmen folgte, dass $E(\eta) = 0$, wodurch folgende Umformungen möglich werden, wenn man Gleichung 164 in die Definition der Kovarianz für η einsetzt:

$$COV(\eta) = E(\eta\eta^T) = E\left[(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\zeta)(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\zeta)^T\right] \tag{165}$$

$$E(\eta\eta^T) = E\left[(\ddot{B}^{-1}\Gamma\xi\xi^T\Gamma^T\ddot{B}^{-1^T}) + (\ddot{B}^{-1}\Gamma\xi\zeta^T\ddot{B}^{-1^T}) + (\ddot{B}^{-1}\zeta\xi^T\Gamma^T\ddot{B}^{-1^T} + (\ddot{B}^{-1}\zeta\zeta^T\ddot{B}^{-1^T})\right] \tag{166}$$

Diese Gleichung lässt sich entsprechend vereinfachen, indem die Annahme ausgenutzt wird, dass die exogenen Variablen und die Residuen unkorreliert sind.

$$COV(\eta) = \ddot{B}^{-1}\Gamma\Phi\Gamma^T\ddot{B}^{-1^T} + \ddot{B}^{-1}\Psi\ddot{B}^{-1^T} \tag{167}$$

$$COV(\eta) = \ddot{B}^{-1}(\Gamma\Phi\Gamma^T + \Psi)\ddot{B}^{-1^T} \tag{168}$$

Damit ist es gelungen die Kovarianzen der endogenen Variablen in Abhängigkeit der Parameter $\ddot{B}^{-1}$, Γ, Φ und Ψ zu formulieren. Nun verbleibt noch eine unbekannte Größe: die Beziehung zwischen den endogenen und exogenen Variablen. Diese wird durch die Kovarianz $COV(\eta, \xi)$ ausgedrückt. Unter Beachtung der Annahmen lässt sich dann folgende Zerlegung mit Hilfe der reduzierten Form, die für η eingesetzt wird, durchführen:

$$COV(\eta, \xi) = E(\eta\xi^T) = E\left[(\ddot{B}^{-1}\Gamma\xi + \ddot{B}^{-1}\zeta)\xi^T\right] \tag{169}$$

$$E(\eta\xi^T) = E\left[\ddot{B}^{-1}\Gamma\xi\xi^T + \ddot{B}^{-1}\zeta\xi^T\right] \tag{170}$$

$$E(\eta\xi^T) = \ddot{B}^{-1}\Gamma\Phi \tag{171}$$

Auch die Kovarianz $COV(\eta, \xi)$ konnte somit in Abhängigkeit der Strukturparameter formuliert werden. Infolgedessen ist es auch möglich die Populationskovarianzmatrix derart zu zerlegen, dass sie von diesen abhängig ist. Diese (r+s, r+s)-Matrix definiert sich wie folgt:

$$\Sigma = \left[\begin{array}{c|c} COV(\eta) & COV(\eta, \xi) \\ \hline COV(\xi, \eta) & COV(\xi) \end{array}\right] = \left[\begin{array}{c|c} \ddot{B}^{-1}(\Gamma\Phi\Gamma^T + \Psi)\ddot{B}^{-1^T} & \ddot{B}^{-1}\Gamma\Phi \\ \hline \Phi\Gamma^T\ddot{B}^{-1^T} & \Phi \end{array}\right] \tag{172}$$

Hinsichtlich der Parametermatrizen lassen sich verschiedene typische Fälle unterscheiden. Für B sind dies insbesondere drei Arten: (1a) B ist eine Diagonalmatrix, so dass die endogenen Variablen nur von exogenen beeinflusst werden; (2a) B ist eine Dreiecksmatrix, so dass sich die endogenen Variablen gegenseitig beeinflussen können (wenn η_i eta_j beeinflusst, kann aber η_j nicht auf η_i wirken); (3a) B ist sowohl über als auch unter der Hauptdiagonalen unbeschränkt, so dass sich die endogenen

[97]Vgl. Long (1983b): 33-35

Variablen simultan beeinflussen können. Des Weiteren gibt es zwei Typen von Fehlermatrizen: (1b) Φ ist eine Diagonalmatrix, so dass alle Fehler unkorreliert sind; (2b) Φ ist eine symmetrische und nichtdiagonale Matrix, so dass die Fehler mindestens zweier Gleichungen korreliert sind.[98]

7.3 Kovarianzstrukturgleichungsmodell

In diesem Abschnitt werden die beiden zuvor diskutierten Modelle zusammengeführt. Damit können die Beziehungen zwischen unbeobachteten Konstrukten untersucht werden ohne davon ausgehen zu müssen, dass diese messfehlerfrei erfasst wurden. Die Modellgleichungen bleiben im Wesentlichen dieselben, die bereits aus den vorangegangenen Kapiteln bekannt sind. Aus diesem Grund werden in der folgenden Übersicht die wichtigsten Notationen und Annahmen wiederholt:[99]

Matrix/ Vektor	Dimension	Mittelwert	Kovarianz	Dimension	Beschreibung
η	(r,1)	0	$COV(\eta) = E(\eta\eta^T)$	(r,r)	latente endogene Variable
ξ	(s,1)	0	$\Phi = E(\xi\xi^T)$	(s,s)	latente exogenen Variable
ζ	(r,1)	0	$\Psi = E(\zeta\zeta^T)$	(r,r)	Residuen
B	(r,r)	–	–	–	direkter Einfluss von η auf η
$\ddot{B}$	(r,r)	–	–	–	$\ddot{B} \equiv (I - B)$
Γ	(r,s)	–	–	–	direkter Einfluss von ξ auf η
x	(q,1)	0	$\Sigma_{xx} = E(xx^T)$	(q,q)	Indikatoren der ξ
Λ_x	(q,s)	–	–	–	Faktorladungen von x auf ξ
δ	(q,1)	0	$\Theta_\delta = E(\delta\delta^T)$	(q,q)	Messfehler von x
y	(p,1)	0	$\Sigma_{yy} = E(yy^T)$	(p,p)	Indikatoren der η
Λ_y	(p,r)	–	–	–	Faktorladungen von y auf η
ϵ	(p,1)	0	$\Theta_\epsilon = E(\epsilon\epsilon^T)$	(p,p)	Messfehler von y

Strukturgleichungen:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{173}$$

$$\ddot{B}\eta = \Gamma\xi + \zeta \tag{174}$$

Faktorgleichungen:

$$x = \Lambda_x\xi + \delta \tag{175}$$

$$y = \Lambda_y\eta + \epsilon \tag{176}$$

Annahmen:

[98] Vgl. Long (1983b): 35 für eine genauere Diskussion
[99] Vgl. Long (1983b): 57

1. Die Variablen werden als Abweichungen von ihrem Mittelwert gemessen: $E(\eta) = E(\zeta) = 0$; $E(\xi) = 0$; $E(x) = E(\delta) = 0$; $E(y) = E(\epsilon) = 0$)
2. Die Faktoren und Messfehler sind unkorreliert: $E(\xi\delta^T) = 0$ bzw. $E(\delta\xi^T) = 0$; $E(\eta\epsilon^T) = 0$ bzw. $E(\epsilon\eta^T) = 0$; $E(\xi\epsilon^T) = 0$ bzw. $E(\epsilon\xi^T) = 0$; $E(\eta\delta^T) = 0$ bzw. $E(\delta\eta^T) = 0$
3. Die Messfehler und Residuen sind unkorreliert über die Gleichungen: $E(\delta\epsilon^T) = 0$ bzw. $E(\epsilon\delta^T) = 0$; $E(\zeta\delta^T) = 0$ bzw. $E(\delta\zeta^T) = 0$; $E(\zeta\epsilon^T) = 0$ bzw. $E(\epsilon\zeta^T) = 0$
4. Die exogenen Variablen und Residuen sind unkorreliert: $E(\xi\zeta^T) = 0$ bzw. $E(\zeta\xi^T) = 0$
5. Keine der Strukturgleichungen ist redundant: $\ddot{B}^{-1} = (I - B)^{-1}$ existiert.

Wiederum muss es gelingen die Populationskovarianzmatrix in Abhängigkeit von den Parametern zu formulieren. Auf Grund der Annahmen ist es möglich folgende Gleichung aufzustellen, wobei $\left[\frac{y}{x}\right]$ ein (p+q,1)-Vektor ist. Im zweiten Schritt werden die Faktorgleichungen für x und y eingesetzt.[100]

$$\Sigma = E\left[\left[\frac{y}{x}\right]\left[\frac{y}{x}\right]^T\right] = E\left[\frac{yy^T|yx^T}{xy^T|xx^T}\right] \tag{177}$$

$$\Sigma = E\left[\frac{(\Lambda_y\eta + \epsilon)(\Lambda_y\eta + \epsilon)^T|(\Lambda_y\eta + \epsilon)(\Lambda_x\xi + \delta)^T}{(\Lambda_x\xi + \delta)(\Lambda_y\eta + \epsilon)^T|(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)^T}\right] \tag{178}$$

$$\tag{179}$$

Multipliziert man nun die Klammern aus, wendet den Erwartungswertoperator und die Definitionen und Annahmen an ergibt sich:

$$\Sigma = E\left[\frac{\Lambda_y\ddot{B}^{-1}(\Gamma\Phi\Gamma^T + \Psi)\ddot{B}^{-1^T}\Lambda_y^T + \Theta_\epsilon| \qquad \Lambda_y\ddot{B}^{-1}\Gamma\Phi\Lambda_x^T}{\Lambda_x\Phi\Gamma^T\ddot{B}^{-1^T}\Lambda_y^T \qquad\qquad |\Lambda_x\Phi\Lambda_x^T + \Theta_\delta}\right] \tag{180}$$

Mit diesen Gleichungen ist es möglich, aus den Varianzen und Kovarianzen der manifesten Variablen die gesuchten Parameter zu schätzen, wenn das Modell identifiziert ist. Da das Prinzip der Schätzungen sich nicht ändert, wird auf diese nicht noch einmal eingegangen. In einem iterativen Prozess werden dann die Parameter gesucht, die die Stichprobenkovarianzmatrix am besten reproduzieren. Auch die Gütemaße zur Beurteilung der Modellanpassung und der Modellsparsamkeit sowie zum Modellvergleich sind analog zur reinen konfirmatorischen Faktorenanalyse zu verwenden.

7.4 Beispiel - Ein einfaches Kovarianzstrukturgleichungsmodell

Das Beispiel für ein einfaches Kovarianzstrukturgleichungsmodell mit kontinuierlichen Daten basiert auf dem Datensatz *Students.psf*, der im Unterordner *Tutorial*

[100] Vgl. Long (1983b): 58-59

der LISREL-Installation enthalten ist.[101] Insgesamt stehen damit 194 Beobachtungen mit zehn Variablen zur Verfügung. Die Daten wurden an einer High-School in Bainbridge, Georgia, erhoben und beziehen sich auf folgende Hypothesen: 1. Es wird angenommen, dass der sozioökonomische Status eines Studenten und seine Einstellungen zum familiären und schulischen Bereich auf das Selbstvertrauen wirken; 2. Es wird davon ausgegangen, dass die Einstellung zur Arbeitsmoral durch das Selbstvertrauen, den sozioökonomischen Status sowie den Attitüden zum familiären und schulischen Umfeld beeinflusst wird. In der folgenden Tabelle sind die manifesten Variablen wiedergegeben:
Es gibt nun zwei Möglichkeiten die Daten zur Schätzung eines Modells zu nutzen.

Variable	Erklärung
AVG_SES	Index des sozioökonomischen Statuses
AVGP_AGE	durchschnittliches Alter der Eltern
GPA	Notendurchschnitt
S_ES_S	Einstellung zur Schule
S_SE_H	Einstellung zu häuslichen Umfeld
S_SE_P	Selbstvertrauen
CAF	Einstellung zum Vater
CAM	Einstellung zur Mutter
TOTAL_OW	Index der Arbeitsmoral
AVGP_EDU	durchschnittliches Ausbildungsniveau der Eltern

Tabelle 1: Variablen im einfachen Kovarianzstrukturgleichungsmodell

Zum Einen können sie direkt als Rohdaten in LISREL eingelesen werden. Zum Anderen kann mittels PRELIS eine Kovarianzmatrix generiert werden, was im Folgenden durchgeführt werden soll. Da der Datensatz keine Missings enthält gestaltet sich dieser Schritt relativ einfach. Mit der folgenden Syntax lässt sich die Kovarianzmatrix in die Datei *beispiel.cov* schreiben. Zudem werden Verteilungstests angefordert, die darauf hinweisen, das die Annahme einer multivariaten Normalverteilung erfüllt ist.

!PRELIS-File für eine Kovarianzmatrix Titel
SY='C:\ LISREL 8.8 Student Examples\ TUTORIAL\ STUDENTS.PSF' Angabe des Dateipfades
OU MA=CM SM=beispiel.cov WP XT mit MA=CM wird die Kovarianzmatrix als Ausgabematrix festgelegt; WP sorgt für eine erweiterte Breite der Outputdatei zur besseren Lesbarkeit; XT fordert die Verteilungstests an

[101] Vgl. hierzu auch `http://www.ssicentral.com/lisrel/techdocs/Session3.pdf`, abgerufen am 8. November 2007, 12.58Uhr

Die Datei *beispiel.cov* kann nun in LISREL eingelesen werden. Folgende Syntax ist dazu notwendig:

! Students-Beispiel
Observed variables: AVG_SES AVGP_AGE GPA S_SE_S S_SE_H S_SE_P CAF CAM TOTAL_OW AVGP_EDU
Covariance Matrix from File: beispiel.cov die Kovarianzmatrix wird aus einer Datei geladen
Sample Size: 194
Latent Variables: SES HOME_ATT APPITUDE SCHOOLATT SEPEERS WETHIC
Equations:
AVG_SES AVGP_AGE AVGP_EDU = SES konfirm. Gleichung für die latente Variable ξ_1
AVG_SES = 1*SES
S_SE_H CAF CAM = HOME_ATT konfirm. Gleichung für die latente Variable ξ_2
S_SE_H = 1*HOME_ATT GPA = 1*APPITUDE konfirm. Gleichung für die latente Variable ξ_3
S_SE_S = 1*SCHOOLATT konfirm. Gleichung für die latente Variable ξ_4
S_SE_P = 1*SEPEERS konfirm. Gleichung für die latente Variable η_1
TOTAL_OW = 1*WETHIC konfirm. Gleichung für die latente Variable η_2
SEPEERS = HOME_ATT SCHOOLATT APPITUDE SES Strukturgleichung
WETHIC = SEPEERS HOME_ATT SCHOOLAT APPITUDE SES Strukturgleichung
set the error variance of GPA to 0.0
set the error variance of S_SE_S to 0.0
set the error variance of S_SE_P to 0.0
set the error variance of TOTAL_OW to 0.0
LISREL Output: SS RS MI
Path diagram
End of Problem

Damit wird klar, dass das hier formulierte Modell bereits weitaus komplexer ist, als die bisherigen. Die ersten beiden latenten Variablen *SES* und *HOME_ATT* werden durch eine Reihe von Indikatoren gemessen. Wohingegen die latenten Konstrukte *APPITUDE*, *SCHOOLATT*, *SEPEERS* und *WETHIC* durch jeweils einen Indikator direkt abgebildet werden. Durch die Strukturgleichungen werden die Beziehungen zwischen den latenten Variablen modelliert, wobei die endogenen latenten Konstrukte, wie auch schon bei der Regression mit manifesten Variablen auf der linken Seite der Gleichung stehen. Es ist in diesem Beispiel zu beachten, das die Variable *SPEERS* zum Einen endogen ist, zum Anderen aber auf die Variable *WETHIC* als Regressor wirkt. Da vier Variablen direkt durch jeweils einen Indikator abgebildet werden, müssen die Fehlervarianzen dieser Indikatoren auf den Wert Null gesetzt werden.
Das Modell ist dann mit 24 Freiheitsgraden überidentifiziert. Mit zehn manifesten Variablen enthält die Stichprobenkovarianzmatrix 55 nicht redundante Elemente[102].

[102] dies sind die Elemente auf und unterhalb der Hauptdiagonalen, da diese Matrix symmetrisch ist

Auf dieser Grundlage werden vier Faktorladungen, ein Element der B-Matrix, acht Elemente der Matrix Γ, zehn Parameter der Matrix Φ und zwei Elemente der Matrix Ψ sowie insgesamt sechs Messfehlerkovarianzen in der Matrix Θ_δ. Aus der Differenz der bekannten und unbekannten Werte lässt dann die Anzahl der Freiheitsgrade bestimmen.
Nach dem Durchlauf der Syntax gibt LISREL folgendes Pfaddiagramm mit den standardisierten Werten aus: Wie die Kriterien zur Modellevaluation zeigen, weist

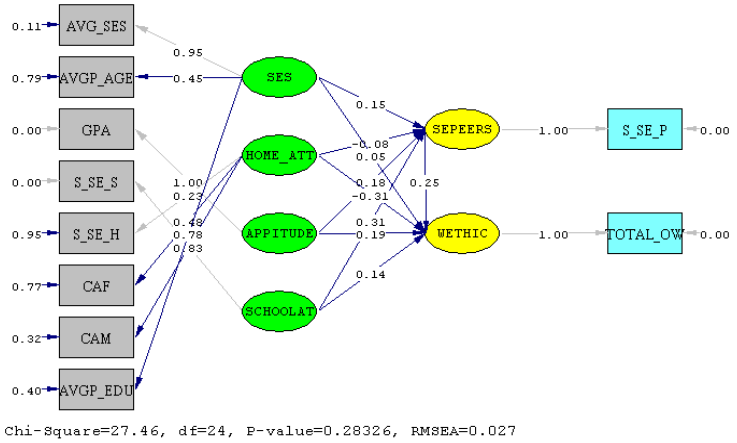

Abbildung 7: Pfaddiagramm: Strukturgleichungsmodell mit kontinuierlichen Variablen

das Modell eine gute Anpassung auf. Auch die Modifikationsindizes zeigen kaum eine signifikante Verbesserungsmöglichkeiten. Zu Übungszwecken soll dennoch eine Modifikation des Modells vorgenommen werden. In der Regel sollte die Modellmodifikation theoretischen Überlegungen folgen, wobei die Modifikationsindizes ein guter Hinweis auf eine Fehlspezifikation des Modells liefern können. Ohne hier jedoch eine theoretische Herleitung der Veränderung des Modells herzuleiten, wird das Element betrachtet, dass den höchsten Modifikationsindex aufweist. Dieses ist das Element (5,4) der Matrix Λ_x mit einem Wert von 6,24. Dieser Empfehlung folgend, wird in der Syntax die Korrelation der Messfehler der Variablen *S_SE_H* und *SCHOOLATT* zur Schätzung freigegeben:

! Students-Beispiel 2
Observed variables: AVG_SES AVGP_AGE GPA S_SE_S S_SE_H S_SE_P CAF CAM TOTAL_OW AVGP_EDU
Covariance Matrix from File: beispiel.cov die Kovarianzmatrix wird aus einer Datei geladen
Sample Size: 194
Latent Variables: SES HOME_ATT APPITUDE SCHOOLATT SEPEERS WETHIC
Equations:
AVG_SES AVGP_AGE AVGP_EDU = SES konfirm. Gleichung für die latente Variable ξ_1
AVG_SES = 1*SES
S_SE_H CAF CAM = HOME_ATT konfirm. Gleichung für die latente Variable ξ_2

S_SE_H = 1*HOME_ATT GPA = 1*APPITUDE konfirm. Gleichung für die latente Variable ξ_3
S_SE_S = 1*SCHOOLATT konfirm. Gleichung für die latente Variable ξ_4
S_SE_P = 1*SEPEERS konfirm. Gleichung für die latente Variable η_1
TOTAL_OW = 1*WETHIC konfirm. Gleichung für die latente Variable η_2
SEPEERS = HOME_ATT SCHOOLATT APPITUDE SES Strukturgleichung
WETHIC = SEPEERS HOME_ATT SCHOOLAT APPITUDE SES Strukturgleichung
set the error variance of GPA to 0.0
set the error variance of S_SE_S to 0.0
set the error variance of S_SE_P to 0.0
set the error variance of TOTAL_OW to 0.0
let the errors between S_SE_H and SCHOOLATT correlate
LISREL Output: SS RS MI
Path diagram
End of Problem

Die Güte des Modells konnte durch die Freisetzung einer Fehlerkorrelation verbessert werden. Sowohl die Maße zur Beurteilung der Modellanpassung, wie etwa der GFI, als auch die Maße zum Vergleich von Modellen und der Modellsparsamkeit belegen eine Verbesserung des Modells.

A LISREL SIMPLIS-Syntax

Die SIMPLIS-Syntax unterscheidet sich in wesentlichen Punkten von der normalen LISREL-Notation, die weitaus formaler orientiert ist. Den groben Aufbau einer solchen Syntax sollten die Beispiele bereits verdeutlicht haben. Im Folgenden wird der Aufbau daher allgemeiner skizziert und entsprechende Literaturhinweise gegeben. Die Syntax lässt sich in mehrere Blöcke unterteilen:[103]

1. *Titel*: Die Titelzeile wird mit einem Ausrufezeichen begonnen, um zu vermeiden, das LISREL diesen als Befehl interpretiert. Bei Gruppenvergleichen ist zuvor ohne das Ausrufezeichen, die Gruppe zu benennen (Group1: ...).

2. *Observed variables:* Hier werden die manifesten Variablen angegeben, wobei die Reihenfolge einzuhalten ist, die der Datensatz vorgibt, da ansonsten Verwechselungen bei der Interpretation der Ergebnisse drohen. Enthält ein Variablenname ein Leerzeichen, so ist dieser in Hochkommata zu setzen. Alternativ kann der Befehl auch lauten: *Observed variables from File ...*

3. *Covariance Matrix*: Es folgt die manuelle Eingabe der Kovarianzen. Da diese Matrix symmetrisch ist, muss lediglich die untere Dreiecksmatrix (inklusive der Hauptdiagonalen) eingegeben werden. Hierbei ist zu beachten, dass LISREL

[103] es kann hier nicht auf alle inhaltlichen und methodischen Problemstellungen eingegangen werden, daher wird zu diesem Zweck auf die entsprechende Literatur verwiesen; diese Aufstellung ist infolgedessen eher als eine erste inhaltliche Einführung zu verstehen

Dezimalstellen anhand eines Punktes erkennt (0.312 und nicht 0,312). Alternativ kann diese Matrix auch extern eingelesen werden. Dies geschieht durch den alternativen Befehl *Covariance Matrix from File ....*

4. *Correlation Matrix*: Basiert die Schätzung auf Korrelationen, so sollte dieser Befehl gelten. In diesem Fall sind aber auch die Mittelwerte und Standardabweichungen anzugeben, wenn Verzerrungen zu erwarten sind. Dies geschieht direkt im Anschluss durch den Befehl *Means from File ...* und *Standard deviations from File ....* Wurden vorab spezielle Zusammenhangsmaße für ordinale Variablen mit PRELIS berechnet, so ist die resultierende Matrix ebenfalls eine Korrelationsmatrix, wobei Mittelwerte und Standardabweichungen nicht angegeben werden müssen.

5. *Asymptotic Covariance Matrix from File ...*: Wird zur Schätzung zum Beispiel die WLS-Diskrepanzfunktion herangezogen, so muss zusätzlich die asymptotische Kovarianzmatrix angegeben werden. Da diese relativ schnell sehr groß wird, wird sie aus einer externen Datei eingelesen.

6. *Sample Size* = Angabe der Stichprobengröße.

7. *Laten Variables:* An dieser Stelle müssen die latenten Konstrukte angegeben werden. Alternativ ist es auch möglich diese, nach dem üblichen Muster, aus einer Datei auszulesen.

8. *Equations:* Auf diesen Befehl folgend die Modellgleichungen. Diese bestimmen zum Einen die Beziehungen der Indikatoren zu den latenten Variablen und zum anderen die der latenten Konstrukte untereinander, wodurch die kausalen Beziehungen festgelegt werden. Bei ersterer Form ist zu beachten, dass das latente Konstrukt auf der rechten Seite des Gleichheitszeichens steht. Werden dagegen kausale Beziehungen angegeben, steht die abhängigen Variable auf der linken und die Regressoren auf der rechten Seite der Gleichung. In diesem Block wird auch die Skalierung der Variablen vorgenommen. Über die Befehle *set* und *let* lassen sich die Modellrestriktionen in die Syntax einbinden. Es können dann beispielsweise Fehler auf einen bestimmten Wert fixiert, mit anderen korreliert oder auch mit anderen gleichgesetzt werden.

9. *Options:* Mit diesem optionalen Befehl lassen sich über eine Reihe von Schlüsselwörtern besondere Aktionen anfordern, wie etwa die Anzahl der Dezimalstellen (ND=4 zum Beipsiel) oder aber auch ME=WLS, wodurch die Schätzmethode auf WLS gesetzt wird.

10. *LISREL Output:* Hier bietet sich die Möglichkeit zusätzliche Outputs anzufordern, wie zum Beispiel die Anforderung einer vollständig standardisierten Lösung über die Angabe des Schlüsselwortes SC.

11. *Path Diagram* Ebenfalls optinal ist der Aufruf eines Pfaddiagramms, das im Folgenden weiter verarbeitet werden kann.

12. *End of Problem* Hiermit wird die Syntax abgeschlossen.

SSICentral bietet auf der Homepage eine Reihe von Publikationen zum Umgang mit der SIMPLIS- aber auch der LISREL-Syntax. Da es eine Vielzahl von Optionen, Schlüsselwörtern und Anwendungen gibt, empfiehlt es sich im konkreten Fall vorab diese Dokumente zu sichten:

1. SIMPLIS: `http://www.ssicentral.com/lisrel/techdocs/SIMPLISSyntax.pdf`, abgerufen 9. November, 15.17 Uhr
2. PRELIS: `http://www.ssicentral.com/lisrel/techdocs/IPUG.pdf`, abgerufen 9. November, 15.18 Uhr
3. Allgemeine LISREL-Einführung: `http://www.ssicentral.com/lisrel/techdocs/GSWLISREL.pdf`, abgerufen 9. November, 15.18 Uhr
4. LISREL-Matrix-Notation: `http://www.ssicentral.com/lisrel/techdocs/LISRELSyntax.pdf`, abgerufen 9. November, 15.20 Uhr

Literatur

W. Assenmacher. *Einführung in die Ökonometrie*. Oldenbourg Wissenschaftsverlag, München, 2002.

J. Bortz. *Statistik*. Springer Verlag, Heidelberg, 2005.

B. M. Byrne. *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

P. Dörsam. *Mathematik anschaulich dargestellt für Studierende der Wirtschaftswissenschaften*. PD-Verlag, Heidenau, 2003.

K. G. Jöreskog. Structural equation modeling with ordinal variables using lisrel, 2002. URL `http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf`. , abgerufen 15.11.2007, 18.35 Uhr.

W. Plinke R. Weiber K. Backhaus, B. Erichson. *Multivariate Analysemethoden - Eine Anwendungsorientierte Einführung*. Springer Verlag, Berlin, Heidelberg, 2006.

W. Langer. Strategien zur beurteilung der modellanpassung von strukturgleichungsmodellen, o.A. URL `http://www.soziologie.uni-halle.de/langer/lisrel/skripten/lisfit2.pdf`. , abgerufen 16.11.2007, 19.35 Uhr.

W. Langer. Einführung in die konfirmatorische faktoren- und pfadanalyse mit lisrel, 2002b. URL `http://www.soziologie.uni-halle.de/langer/lisrel/skripten/lisrelmodelle.pdf`. , abgerufen 16.11.2007 19.25 Uhr.

W. Langer. Grundlagen der explorativen pfadanalyse nach wright (1921, 1934), 2002. URL `http://www.soziologie.uni-halle.de/langer/lisrel/skripten/pfadwright2.pdf`. , abgerufen 16.11.2007 19.13 Uhr.

J. S. Long. *Covariance Structure Models - An Introduction to LISREL*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-034. Sage Publications, Beverly Hills, 1983b.

J. S. Long. *Confirmatory Factor Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-033. Sage Publications, Beverly Hills, 1983a.

R. G. Lomax R. E. Schumacker. *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2004.

E. Esser R. Schnell, P. B. Hill. *Methoden der empirischen Sozialforschung*. Oldenbourg Wissenschaftsverlag, München, 1999.

J. Reinecke. *Strukturgleichungsmodelle in den Sozialwissenschaften*. Oldenbourg Wissenschaftsverlag, München, 2005.

L. Satow. Lisrel einführung, 1999. URL `http://dtserv2.compsy.uni-jena.de/ss2005/metheval_uj/sem/content.nsf/Pages/399EE502E469BAADC1256FD6004C6819/$FILE/satow_1999_lisrel.pdf.` , abgerufen 15.11.2007, 18.24 Uhr.

SSIcentral. Fitting structural equation models to complete continuous data, o.A. URL `http://www.ssicentral.com/lisrel/techdocs/Session3.pdf.` , abgerufen 08.11.2007, 12.58 Uhr.