Zusammenfassung

Multivariate Verfahren I

in der Psychologie

Carlo Michaelis

Wintersemester 2015/2016

Basierend auf dem ersten Teil einer Vorlesung

von Prof. Dr. Manuel Völkle

Dieses Dokument ist unter folgender Lizenz veröffentlicht:
Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

Inhaltsverzeichnis

1 Einführung

1.1 Psychologische Methodenlehre

Eine **Theorie** ist ein System von Aussagen zur (1) Beschreibung, (2) Erklärung, (3) Vorhersage und der (4) Modifikation von (primär menschlichem) Erleben und Verhalten.

In der **psychologischen Methodenlehre** werden psychologische Theorien überprüft.

1.1.1 Kriterien einer Theorie

Eine Theorie muss folgende Kriterien erfüllen:

1. widerspruchsfrei
2. überprüfbar
3. explizit

Dies führt zu einer **empirischen Verankerung**.

1.2 Struktur eines Forschungsvorhabens & Kriterien guter Forschung

In jedem Schritt des Forschungsprozesses können Kriterien formuliert werden:

- Problem
- Sampling: Externe Validität
- Design: Interne Validität
- Messung: Konstruktvalidität
- Analyse und Schlussfolgerung: Schlussfolgerungsvalidität

Die Kriterien werden im Folgenden genauer betrachtet.

1.2.1 Sampling

In einer Messung werden folgende Größen erfasst:

- Personen (Untersuchungseinheit)
- Untersuchungsgegenstände (Variablen), z.B. Kognitive Fähigkeit
- Situationen (Zeitpunkte)

Diese werden gewöhnlich nicht in einer **Vollerhebung** bestimmt. Es wird nur eine **Stichprobe** betrachtet. D.h. nicht alle Personen, nicht alle Aspekte des Untersuchungsgegenstandes und nicht alle Situationen bzw. Zeitpunkte werden erfasst.

Ein Sampling kann in zwei Extremfällen erfolgen bzw. aus einer Mischform dieser Pole gewählt werden:

- Zufalls-Sample (Random Sample)
- Ausgewähltes Sample (Convenience Sample)

Die **externe Validität** beschreibt den *Grad der Generalisierbarkeit* über (1) Personen, (2) Situationen und (3) Untersuchungsgegenstände.

Der Sachverhalt kann in einem **Cattel'schen Datenquader** dargestellt werden.

In der *multivariaten Forschung* werden Zusammenhänge zwischen mehreren abhängigen und unabhängigen Variablen betrachtet. Es werden mehrere Datenquader berücksichtigt.

1.2.2 Untersuchungsdesign

Wird nur eine Stichprobe betrachtet, können mögliche Drittvariablen einen Einfluss haben, die bei einem Sampling u.U. nicht erhoben wurden. Aussagen über Kausalität sind daher oft schwierige. Die **interne Validität** beschreibt den Grad in dem Aussagen über Ursache-Wirkungszusammenhänge möglich sind.

Externe vs. interne Validität

Um eine ideale interne Validität zu erhalten ist es nötig experimentelle Untersuchungen zu machen, bei denen möglichst viele Faktoren konstant gehalten werden. Dies führt meist jedoch zu einer geringeren externen Validität, da die Untersuchung damit konstruierter und weniger auf die Realität anwendbar ist, der Grad der Generalisierbarkeit ist geringer. Umgekehrt weisen Feld-Untersuchungen zwar einen hohe externe Validität auf, ihre Aussagekraft über Ursache-Wirkungszusammenhänge, und damit ihre interne Validität, ist jedoch eher gering.

Kausalität

Bedingungen für Kausalität

Der Philosoph John Stuart Mill (1806-1873) formulierte drei Bedingungen für kausale Vorgänge:

1. Die Ursache muss dem Effekt zeitlich vorausgehen
2. Ursache und Effekt müssen zusammenhängen (kovariieren)
3. Alternative Erklärungsmöglichkeiten für den Ursache-Effekt-Zusammenhang müssen ausgeschlossen sein

Problematisch ist aus methodischer Sicht vor allem die dritte Bedingung. Es ist sehr schwierig *alle* alternativen Erklärungsmöglichkeiten auszuschließen.

Definition der Kausalität

Eine formelle Definition der Kausalität lieferte Paul Holland (1986). Dabei ergibt sich der **Kausaleffekt** durch:

$$Y_{i|d=1} - Y_{i|d=0} \tag{1.1}$$

Dabei bezeichnet Y den Effekt $d = 0$ bzw. $d = 1$ jeweils an der Stelle i (z.B. Personen, Zeitpunkt, etc.). Der Kausaleffekt könnte jedoch nur eindeutig erfasst werden, wenn die beiden subtrahierten Effekte an der gleichen Stelle i erfasst werden. Werden sie nicht an der gleichen Stelle (z.B. gleiche Person, gleicher Zeitpunkt, etc.) erfasst, sollte davon ausgegangen werden, dass sich Y bereits geändert hat. Holland formulierte „**The Fundamental Problem of Causal Inference**".

„It is impossible to observe the value of $Y_{i|d=1}$ and $Y_{i|d=0}$ on the same unit i and, therefore, it is impossible to observe [but not infer] the effect on i."(Holland, 1986, p. 947).

Zur Lösung des Problems können zwei Ansätze gewählt werden:

1. Ansatz, „wissenschaftliche“Lösung

Bezüglich (1) Homogenität und (2) Invarianz werden Annahmen getroffen, die gemeinsam oder einzeln gültig sein müssen:

- *temporal stability* (Zeitliche Stabilität) und *causal transience*
- *unit exchangeability*

In Gleichung 1.1 bezeichne i die Personen und die zusätzliche Variable t die Zeit, dann kann z.B. davon ausgegangen werden, dass die Werte bei einer Messung t genauso auch bei einer früheren Messung $t-1$ auftraten, sofern die gleichen Personen gemessen wurden. $Y_{i,t|d=0}$ ist somit äquivalent zu $Y_{i,t-1|d=0}$. Für de Kausaleffekt ergibt sich:

$$\gamma_i = Y_{i,t|d=1} - Y_{i,t-1|d=0} \tag{1.2}$$

2. Ansatz, Statistische Lösung

Mittels des *Erwartungswertes* werden durchschnittliche kausale Effekte betrachtet.

$$\gamma = \mathbb{E}\left[Y_{i|d=1} - Y_{i|d=0}\right] \tag{1.3}$$

Dabei können folgende Fälle verwendet werden:

1. *Unterschiedliche Personen* i, j zum *gleichen Zeitpunkt* t: $E\left[Y_{i,t=1|d=1} - Y_{j,t=1|d=0}\right]$
2. *Gleiche Personen* i zu *unterschiedlichen Zeitpunkten* t: $E\left[Y_{i,t=1|d=1} - Y_{i,t=0|d=0}\right]$
3. Kombinationen aus den ersten beiden Varianten.

1.2.3 Messung

Bei der Messung werden zwei Kriterien unterschieden:

- **Reliabilität**: Güte der Messung
- **Konstruktvalidität**: Güte des Konstrukts

1. Reliabilität

Die Reliabilität ergibt sich über:

$$r_{tt} = \frac{\sigma^2_{wahr}}{\sigma^2_{gesamt}} \tag{1.4}$$

Während die Gesamtvarianz σ^2_{gesamt} bekannt ist, muss die wahre Varianz σ^2_{wahr} geschätzt werden. Um einen Schätzer zu bestimmen werden folgende Axiome verwendet:

1. $X = T + \epsilon$ (Der gemessene Wert entspricht dem wahren Wert und dem Fehler)
2. $E[\epsilon] = 0$ (Im Mittel verschwindet der Fehler)
3. $Cov(T, \epsilon) = Cov(\epsilon_i, \epsilon_j) = 0$ (Der wahre Wert und der Fehler, sowie die Fehler untereinander sind unabhängig voneinander)

Mit Hilfe der Axiome kann der wahre Wert geschätzt werden:

$$\sigma^2_{wahr} = Cov(X_A, X_B) \overset{(1)}{=} Cov(T + \epsilon_A, T + \epsilon_B) \overset{(3)}{=} Cov(T, T) + 0 + 0 + 0 \tag{1.5}$$

Die **Reliabilität** ist der Anteil der wahren Varianz an der Gesamtvarianz. Er bildet eine notwendige (nicht hinreichende) Voraussetzung für Korrelationen.

Die maximal mögliche Korrelation zwischen zwei Variablen x und y wird durch die Reliabilität begrenzt. Dies wird als **Attenuationskorrektur** bezeichnet:

$$r_{wahr}(x, y) = \frac{r_{beob}(x, y)}{\sqrt{r_{tt}(x) \cdot r_{tt}(y)}} \tag{1.6}$$

2. Konstruktvalidität

Selbst wenn die Messung „technisch" korrekt abläuft, kann es sein, dass das gewählte Konstrukt mit den gewählten Mitteln nicht vollständig erfasst wird. Vor allem wenn — wie in der Psychologie häufig — latente Konstrukte gemessen werden, ist es wichtig abschätzen zu können wie gut das jeweilige Konstrukt erfasst wird und nicht etwa ganz andere Konstrukte gemessen werden.

Egon Brunswik (1903-1955) beschäftigte sich mit der Frage, wie eine Messung konstruiert werden sollte, um sie valide zu gestalten. Ein zentraler Begriff ist dabei die **Symmetrie**. Wird bspw. der Einfluss des latenten Konstruktes „Kognitive Fähigkeit" auf den „Durchschnitt der Schulnote" betrachtet, müssen die gemessenen Aspekte der Konstrukte zueinander passen. Wird beispielsweise zur Erfassung der kognitiven Fähigkeit ein Test zur numerischen Verarbeitung durchgeführt und für die durchschnittliche Schulnote nur die Note im Fach Deutsch verwendet, so werden die Korrelationen gering ausfallen. Vielmehr sollte die verbale Fähigkeit erfasst werden oder alternativ nur die Note im Fach Mathe verwendet werden. Werden alle Fächer für die Durchschnittsnote berücksichtigt, so sollte auch die kognitive Fähigkeit möglichst umfassende Tests enthalten, um vernünftige Zusammenhänge zu erhalten. Da das Prinzip von Brunswik in Form einer Linse dargestellt werden kann, wird das Modell auch als **Linsenmodell** bezeichnet.

Die **Konstruktvalidität** beschreibt den Grad der Übereinstimmung von theoretischem und empirischen Konstrukt.

Auch die Konstruktvalidität begrenzt die maximale Korrelation zweiter Variablen. Gleichung 1.6 kann noch um den Einfluss der Konstruktvalidität $R(x)$, $R(y)$ und unbekannte Effekte e (z.B. nonlineare Effekte, Selektionseffekte, etc.) ergänzt werden:

$$r_{wahr}(x, y) = \frac{r_{beob}(x, y)}{\sqrt{r_{tt}(x) \cdot r_{tt}(y)} \cdot R(x) \cdot R(y)} + e \tag{1.7}$$

Effekte werden wegen mangelnder Validität häufig überschätzt.

1.2.4 Analyse und Schlussfolgerung

Im letzten Schritt, der Analyse und der Schlussfolgerung, welche im Folgenden genauer betrachtet werden, ist die **Schlussfolgerungsvalidität** ein letztes Kriterium. Sie beschreibt die Gültigkeit mit der inhaltliche Schlussfolgerungen aus empirischen Daten gezogen werden.

Begriffe

Theorie, Psychologische Methodenlehre, Kriterien einer Theorie, Kriterien guter Forschung, Sampling, Vollerhebung, Stichprobe, Random Sample, Convenience Sample, externe Validität, Cattel'sches Datenquader, multivariate Forschung, Untersuchungsdesign, interne Validität, Zusammenhang interne/externe Validität, Bedingungen für Kausalität, Definition der Kausalität, Fundamental Problem of Causal Inference, Wissenschaftliche Lösung, Statistische Lösung, Reliabilität, Attenuationskorrektur, Konstruktvalidität, Linsenmodell, Schlussfolgerungsvalidität

2 Multiple lineare Regression

2.1 Grundlagen

2.1.1 Komponenten der Regressionsgleichung

Es werden ausschließlich lineare Regressionsleichungen betrachtet, die Regressionskoeffizienten liegen in erster Potenz vor.

Popluationsregressionsfunktion mit der abhängien Variable y, den unabhängigen Variablen x_k und den **Regressionskoeffizienten** bzw. **Regressionsgewichten** b_k insbesondere der **Regressionskonstante** b_0 (y-Achsenabschnitt der Regressionsgeraden). e ist der Fehler.

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \cdots + b_n x_n + e \tag{2.1}$$

Bei der Stichprobenregressionsfunktion werden die Populationsparameter geschätzte:

$$y = \hat{y} + \hat{e} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_k x_k + \cdots + \hat{b}_n x_n + \hat{e} \tag{2.2}$$

Die Regresionskoeffizienten b_k werden so geschätzt, dass die Summe der quadrierten Differenzen $\hat{e}$ zwischen vorhergesagten und beobachteten Werten $\hat{y}_i$ und y_i über alle Personen N minimal wird. Dies wird als **Methode der kleinsten Quadrate** bezeichnet.

$$\underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left[\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right] = \underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left[\sum_{i=1}^{N}(\hat{e}_i)^2\right] = \underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left[\sum_{i=1}^{N}(y_i-(\hat{b}_0+\hat{b}_1 x_{1,i}+\cdots+\hat{b}_k x_{k,i}))^2\right] \tag{2.3}$$

Für den Spezialfall eines Prädiktors (einer unabhängigen Variable) vereinfacht sich die Gleichung zu:

$$\underset{\hat{b}_0,\hat{b}_1}{arg\,min}\left[\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right] = \underset{\hat{b}_0,\hat{b}_1}{arg\,min}\left[\sum_{i=1}^{N}(y_i-(\hat{b}_0+\hat{b}_1 x_{1,i}))^2\right] \tag{2.4}$$

Ein anschauliches Beispiel ist in Abbildung 1 dargestellt. Für 2 Prädiktoren ergibt sich eine Ebene, für 3 oder mehr Prädiktoren sind grafische Darstellungen nicht mehr anschaulich.

Anmerkung: Für die x_k werden die Begriffe unabhängige Variable & Prädiktor, für y die Begriffe abhängige

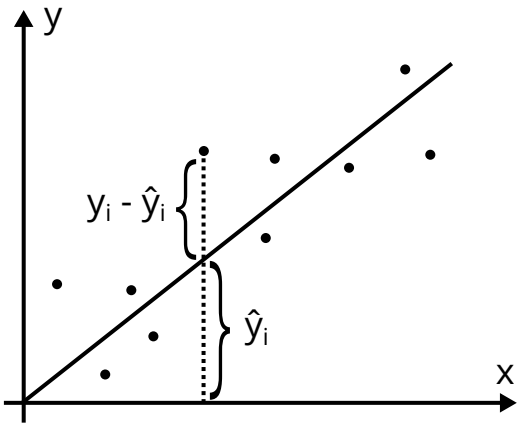


Abbildung 1: Lineare Regression. Mit den echten Werten y_i und den geschätzten Werten $\hat{y}_i$.

Variable & Kriterium im Folgenden synonym verwendet.

2.1.2 Partialregressionsgewichte

Bei einer multiple Regression werden **partielle Regressionsgewichte** verwendet. Die Regressionsgewichte werden im Kontext der anderen Regressionsgewichte berechnet. Im Index wird dies deutlich gemacht, indem der erste Teil des Index den Index des betrachteten Koeffizienten ist, während der zweite Teil des Index den Indizes entsprechen, welche gerade nicht betrachtet werden. Während der Koeffizient eines Prädiktors variiert wird, werden die anderen Prädiktoren konstant gehalten.

$$\hat{y}_i = \hat{b}_{y0.1\ldots k\ldots K} + \hat{b}_{y1.2\ldots k\ldots K} x_{1,i} + \cdots + \hat{b}_{yK.1\ldots k\ldots K-1} x_{K,i} \tag{2.5}$$

Beispieldaten

Die Fragestellung der Beispieldaten soll lauten:
Wie viel Dollar Gehalt verdient ein Professor abhängig von der Anzahl der Jahre im Beruf, der Anzahl der Publikationen, dem Geschlecht und der Anzahl der Zitationen?

Die Regressionskonstante (Intercept) gibt die Höhe des Gehaltes an (unstandardisiert, also in Dollar), ohne Beachtung der anderen unabhängigen Variablen (unter der Annahme alle anderen Koeffizienten wären Null). Der Intercept bildet den Ausgangspunkt. Das Regressionsgewicht einer Variable erhöht den Wert des Gehaltes multipliziert mit der Höhe des jeweiligen Wertes. Z.B. wird vorhergesagt, dass ein Professor pro Jahr $857,01\,\$$ mehr oder z.B. pro Zitation $201,93\,\$$ mehr verdient.

2.1.3 Methode der kleinsten Quadrate (OLS): Bivariate Regression

Bei bivariater Regression wird zunächst der Fall betrachtet, bei dem ein Prädiktor verwendet wird. Dazu wird von der Minimierungsfunktion aus Gleichung 2.4 ausgegangen. Zur Bestimmung des Minimums werden die partiellen Ableitungen nach den Regressionsgewichten gebildet und Null gesetzt.

$$\frac{\partial}{\partial \hat{b}_0}\left[\sum_{i=1}^{N}(y_i - (\hat{b}_0 + \hat{b}_1 x_{1,i}))^2\right] = 2 \cdot \left[\sum_{i=1}^{N} y_i - (\hat{b}_0 + \hat{b}_1 x_{1,i})\right] \cdot (-1) \stackrel{!}{=} 0 \tag{2.6}$$

$$\frac{\partial}{\partial \hat{b}_1}\left[\sum_{i=1}^{N}(y_i-(\hat{b}_0+\hat{b}_1 x_{1,i}))^2\right]=2\cdot\left[\sum_{i=1}^{N}y_i-(\hat{b}_0+\hat{b}_1 x_{1,i})\right]\cdot(-x_{1,i})\stackrel{!}{=}0 \tag{2.7}$$

Anschließend werden aus den zwei erhaltenen Gleichungen die Regressionsgewichte bestimmt.

$$\hat{b}_1=\frac{N\sum_{i=1}^{N}(x_i\cdot y_i)-\sum_{i=1}^{N}x_i\cdot\sum_{i=1}^{N}y_i}{N\sum_{i=1}^{N}x_i^2-(\sum_{i=1}^{N}x_i)^2} \tag{2.8}$$

$$\hat{b}_0=\frac{1}{N}\sum_{i=1}^{N}y_i-\hat{b}_1\cdot\frac{1}{N}\sum i=1^N x_i \tag{2.9}$$

2.1.4 Methode der kleinsten Quadrate (OLS): Multiple Regression

Im allgemeineren Fall werden N Personen betrachtet. Das **Populationsregressionsmodell** kann in Matrixschreibweise wie folgt formuliert werden:

$$\mathbf{y}=\mathbf{X}\cdot\mathbf{b}+\mathbf{e} \tag{2.10}$$

Ausgeschrieben entspricht dies für K Prädiktoren und N Personen:

$$\begin{pmatrix}y_1\\y_2\\\vdots\\y_N\end{pmatrix}=\begin{pmatrix}1&x_{1,1}&\cdots&x_{1,K}\\1&x_{2,1}&\cdots&x_{2,K}\\\vdots&\vdots&\ddots&\vdots\\1&x_{N,1}&\cdots&x_{N,K}\end{pmatrix}\begin{pmatrix}b_0\\b_1\\\vdots\\b_K\end{pmatrix}+\begin{pmatrix}e_1\\e_2\\\vdots\\e_N\end{pmatrix} \tag{2.11}$$

Für das **Stichprobenregressionsmodell** gilt:

$$\mathbf{y}=\mathbf{X}\cdot\hat{\mathbf{b}}+\hat{\mathbf{e}} \tag{2.12}$$

Auch hier wird die Minimierungsfunktion so aufgestellt, dass der quadrierte Fehler $\hat{e}$ minimal wird:

$$\underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left[\sum_{i=1}^{N}(\hat{e}^i)^2\right]=\underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left(\hat{\mathbf{e}}^{\mathbf{T}}\hat{\mathbf{e}}\right)\stackrel{(2.12)}{=}\underset{\hat{b}_0,\ldots,\hat{b}_n}{arg\,min}\left[\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^T\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)\right] \tag{2.13}$$

Der innere Term wird als SSE (*Sum of Squared Errors*) bezeichnet. Nach Ausmultiplizieren kann dieser Term differenziert und Null gesetzt werden, um das Minimum zu bestimmen.

$$\begin{aligned}\frac{\partial SSE}{\partial\hat{\mathbf{b}}}&=\frac{\partial}{\partial\hat{\mathbf{b}}}\left[\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^T\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)\right]\\&=\frac{\partial}{\partial\hat{\mathbf{b}}}\left[\mathbf{y}^{\mathbf{T}}\mathbf{y}-2(\mathbf{y}^{\mathbf{T}}\mathbf{X}\hat{\mathbf{b}})+\hat{\mathbf{b}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\mathbf{b}}\right]\\&=-2\mathbf{X}^{\mathbf{T}}\mathbf{y}+2\mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\mathbf{b}}\stackrel{!}{=}0\end{aligned} \tag{2.14}$$

Der Schätzer $\hat{\mathbf{b}}$ ergibt sich durch Umstellen der Gleichung:

$$\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} \tag{2.15}$$

Durch einsetzen in das Stichprobenregressionsmodell in Gleichung 2.12 können Vorhersagen getroffen werden. Dabei entspricht $\hat{\mathbf{y}}$ der geschätzten Vorhersage. D.h. während $\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \hat{\mathbf{e}}$ die gemessenen Daten beschreibt, beschreibt $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ die vorhergesagten Daten. Für das Modell der geschätzten Werte gilt:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} \overset{(2.15)}{=} \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T y} = \mathbf{H y} \tag{2.16}$$

Dabei beschreibt $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$ den Zusammenhang zwischen den wahren Werten $\mathbf{y}$ und den geschätzten Werten $\hat{\mathbf{y}}$.

Die Diagonalelemente von $\mathbf{H}$ werde als **Leverage** bezeichnet. Der Leverage gibt an wie weit der Wert vom Mittel aller Prädiktoren $\mathbf{X}$ entfernt ist. Er bildet das „Potential“ eines Wertes auf die Regressionsgerade Einfluss zu nehmen. Der Wert hat daher eine Art „Hebelwirkung“ auf die Regression. Für einen einzelnen Prädiktor kann der Leverage wie folgt berechnet werden:

$$h_i = \frac{1}{N} + \frac{1}{N-1}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 \tag{2.17}$$

2.1.5 Standardisierte Regressionsgewichte

Die **standardisierten Regressionsgewichte** werden durch **z-Standardisierung** aller Variablen vor der Regressionsanalyse erreicht. Für einen Prädiktor $\mathbf{X}$ gilt:

$$\mathbf{Z} = \frac{\mathbf{X} - \mu}{\sigma} \tag{2.18}$$

Ist die Steigung eines Regressionsgewichtes beispielsweise 0.5, so wird pro Erhöhung um ein entsprechendes z, die Ausgangsvariable um eine halbe Standardabweichung erhöht sein.

2.1.6 Multiples R und Determinationskoeffizient

Der Zusammenhang bzw. die Korrelation zwischen den beobachteten Werten $\mathbf{y}$ und den vorhergesagten Werten $\hat{\mathbf{y}}$ kann wie folgt berechnet werden.

$$R = Cor(y, \hat{y}) = \frac{Cov(y, \hat{y})}{\sqrt{Var(y)\,Var(\hat{y})}} \tag{2.19}$$

Für zwei Prädiktoren entspricht die Korrelation der *Pearson Produkt Moment Korrelation*. Im Fall mehrerer Prädiktoren wird die Korrelation als **multiples R** bezeichnet. Wird R quadriert wird dies als **Determinationskoeffizient** (auch **Bestimmtheitsmaß**) R^2 bezeichnet. Er ist definiert mit der Variation von $\mathbf{y}$, die mit SS_{total} (sum of squares) bezeichnet wird und der Variation der Residuen $SS_{residual}$:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2.20}$$

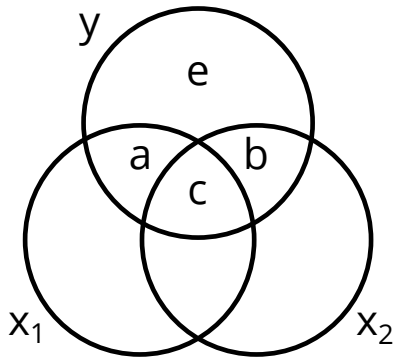


Abbildung 2: Venn-Diagramm zur Veranschaulichung der Varianzanteile. a entspricht der Varianz, die isoliert nur die Variable x_1 beiträgt. b bezeichnet diejenige Varianz, die von x_2 aufgeklärt wird. c wird von beiden Variablen x_1 und x_2 gemeinsam beigetragen. Der Anteil e entspricht der Residualvarianz, die von keiner Variablen erklärt werden kann.

Der Determinationskoeffizient ist somit ein Schätzer für die *aufgeklärte Varianz*. Der Schätzer ist jedoch nicht *erwartungstreu*, d.h. im Mittel entspricht der Koeffizient nicht der aufgeklärten Varianz in der Population ρ^2. Insbesondere bei kleinen Stichproben überschätzt R^2 auf Grund der Quadrierung die aufgeklärte wahre Varianz ρ^2. Zur Korrektur kann der **Adjusted $\mathbf{R^2}$** (Korrigiertes R^2) verwendet werden, welcher den Anteil der aufgeklärten Varianz erwartungstreu angibt. Die Anzahl der Prädiktoren K und der Stichprobenumfang (Anzahl der Personen) N werden zur Korrektur verwendet.

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K-1} \tag{2.21}$$

2.1.7 Partial- und Semipartialkorrelation

Die **quadrierte Semipartialkorrelation** gibt an wie stark der Prädiktor k isoliert zur Gesamtvarianz beiträgt. Es gilt $k = 1, \ldots, K$, wobei K der Anzahl der Prädiktoren entspricht.

$$sr_k^2 = R^2_{1\ldots k\ldots K} - R^2_{1\ldots k-1,k+1\ldots K} \tag{2.22}$$

Die **quadrierte Partialkorrelation** gibt an wie stark der Prädiktor k isoliert zur Gesamtvarianz beiträgt, wobei der Wert ins Verhältnis zur unaufgeklärten Varianz (**Residualvarianz**) gesetzt wird. D.h. die Semipartialkorrelation wird durch die Varianz geteilt, die durch *keine* Variable erklärt werden kann.

$$pr_k^2 = \frac{sr_k^2}{1 - R^2_{1\ldots k-1,k+1\ldots K}} \tag{2.23}$$

Die Verhältnisse können in einem **Venn-Diagramm** (*Ballantines*) veranschaulicht werden. Ein Beispiel ist in Abbildung 2 dargestellt. Die quadrierten Korrelationen können aus dem Venn-Diagramm anschaulich abgeleitet werden.

- Korrelation $\quad r^2_{y,x_1} = a + c \quad r^2_{y,x_2} = b + c$
- Semipartielle Korrelation $\quad sr^2_{y,x_1} = a \quad sr^2_{y,x_2} = b$
- Partielle Korrelation $\quad pr^2_{y,x_1} = \frac{a}{a+e} \quad pr^2_{y,x_2} = \frac{b}{b+e}$
- Multiple Korrelation $\quad R^2_{y,x_1x_2} = a + b + c$

2.1.8 Statistische Inferenz

Für die Parameterschätzer können Standardfehler berechnet werden. Sie geben die Güte des Schätzers an.

Standardfehler der Regressionsgewichte

$$SE_{b_k} = \frac{sd_y}{sd_k} \cdot \sqrt{\frac{1}{1-R_k^2}} \cdot \sqrt{\frac{1-R_y^2}{N-K-1}} \tag{2.24}$$

Dabei wird $1 - R_k^2$ als **Toleranz** bezeichnet. Sie beschreibt die unaufgeklärte Varianz der jeweiligen Variablen. $1-R_y^2$ wird als **Indeterminationskoeffizient** bezeichnet. Er bildet das Inverse des Determinationskoeffizienten R_y^2.

Der Standardfehler kann für verschiedene Verfahren verwendet werden:

1.) Das Regressionsgewicht $\hat{b}_k$ kann unter Annahme eines bestimmten Signifikanzniveaus α auf Signifikanz geprüft werden ($H_0 : b_k = 0$). Der t-Wert für den **Hypothesentest** wird mit dem Standardfehler SE_{b_k} bestimmt.

$$t_{1-\alpha,df} = \frac{\hat{b}_k}{SE_{b_k}} \quad \text{mit den Freiheitsgraden } df = N - K - 1 \tag{2.25}$$

2.) Auf Basis des Standardfehlers kann das **Konfidenzintervall** bestimmt werden:

$$C = \left[\hat{b}_k - t_{1-\alpha,df} \cdot SE_{b_k};\, \hat{b}_k + t_{1-\alpha,df} \cdot SE_{b_k}\right] \tag{2.26}$$

Wird das Niveau z.B. $\alpha = 0,05$ gewählt, dann wird mit der Gleichung 2.26 das 90%-Konfidenzintervall angegeben.

Die *Interpretation des Konfidenzintervalls* ist streng genommen wie folgt: Werden mehrere Messungen durchgeführt und deren Konfidenzintervalle berechnet, so liegt der wahre Wert bei 90% der bestimmten Konfidenzintervalle innerhalb des Intervalls. Die Annahme, der wahre Wert würde mit 90%-er Wahrscheinlichkeit innerhalb des Intervalls einer Messung liegen, ist nicht korrekt.

Gesamt aufgeklärte Varianz und F-Test

Die gesamt aufgeklärte Varianz wird mittels des **F-Tests** auf Signifikanz geprüft. Die Nullhypothese lautet:

$$H_0 : \rho^2 = 0 \quad \text{äquivalent zu} \quad H_0 : b_1 = \cdots = b_K = 0 \tag{2.27}$$

Die Prüfgröße (F-Wert) berechnet sich wie folgt:

$$F_{df_1,df_2} = \frac{R^2}{1-R^2}\frac{N-K-1}{K} \quad \text{mit } df_1 = K \text{ (Zählerfr.)},\ df_2 = N - K - 1 \text{ (Nennerfr.)} \tag{2.28}$$

Dabei ist N die Größe der Stichprobe und K die Anzahl der Prädiktoren.

Die Alternativhypothese H_1 lautet, dass mindestens ein $b_k \neq 0$ ist.

2.1.9 Annahmen und Prüfung der multiplen linearen Regression (MLR)

Der multiplen linearen Regression (MLR) liegen *5 Annahmen* zu Grunde.

1. Linearität

Der Zusammenhang zwischen unabhängigen und abhängigen Variablen muss linear sein.

Dies bedeutet, dass die Regressionsterme linear additiv verknüpft sein müssen. Der Zusammenhang der Variablen muss jedoch *nicht* linear sein. Erfüllt ist die Bedingung bspw. für:

- $y = b_0 + b_1 x_1 + e$
- $y = b_0 + b_1 \ln(x_1) + e$
- $y = b_0 + b_1 x_1^2 + e$

Zur anschaulichen Überprüfung eigenen sich **Residualplots** (x-Achse: Vorhergesagte Werte, y-Achse: Residuen). Der Verlauf muss hier ungefähr konstant bleiben.

2. Exogene Prädiktoren

Die Prädiktoren müssen perfekt reliabel sein, d.h. die Prädiktoren $\mathbf{X}$ müssen unabhängig vom Fehler $\mathbf{e}$ sein. Die Messung muss perfekt durchgeführt werden, damit die Messgrößen nicht von der Messung beeinflusst werden.

$$\mathbb{E}(e_i|\mathbf{X}) = 0 \tag{2.29}$$

Zur Überprüfung gibt es mehrere Möglichkeiten, z.B.:

- Aufnahme weiterer Variablen ($\rightarrow$ Pfadanalyse, siehe Abschnitt 2.2)
- Exogenität des Prädiktors durch Wahl des Versuchsdesigns
- Reliabilität berechnen/kontrollieren

Anmerkung: Die Berechnung der Korrelation zwischen $\mathbf{X}$ und $\mathbf{e}$ zur Überprüfung ergibt keinen Sinn. Per Annahme ist diese Korrelation Null.

3. Homoskedastizität der Residuen

Die Varianz der Residuen muss gleich sein, über den gesamten Bereich der vorhergesagten Werte. Mit anderen Worten: Die Varianz des y-Wertes an einer bestimmten Stelle muss mit den y-Werten an allen anderen Stellen übereinstimmen. Die Varianz entspricht dem *quadrierten Standardschätzfehler*:

$$Var(\hat{\mathbf{b}}) = \mathbb{E}\left[(\mathbf{X^T X})^{-1}\mathbf{X^T e e^T X}(\mathbf{X^T X})^{-}1\right] = (\mathbf{X^T X})^{-1}\mathbf{X^T}\mathbb{E}[\mathbf{e e^T}]\mathbf{X}(\mathbf{X^T X})^{-}1 \tag{2.30}$$

Dabei ist $\mathbb{E}[\mathbf{e e^T}] = \sigma^2\mathbf{1}$ die Homoskedastizitätsannahme.

Wie bei der 1. Annahme lässt sich auch diese gut mit Hilfe von **Residualplots** beurteilen. Die Varianz der Residuen sollte sich im Verlauf nicht ändern.

4. Unabhängigkeit der Residuen

Der Wert des Residuums eines Wertes, darf nicht von dem Wert des Residuums eines anderen Wertes abhängen. Es gilt, dass die einzelnen Werte **i.i.d.**, d.h. „independent and identically distributed“ Verteilt sind.

Verletzt ist diese Annahme häufig bei (1) *Clustering* oder (2) *serieller Abhängigkeit.*

Ob eine Abhängigkeit vorliegt kann u.a. durch folgende Verfahren geprüft werden:

- Theoretische Gründe (z.B. Schüler in einer Klasse, die Residuen werden über die Schüler mit der Klasse zusammenhängen)
- Ausmaß des Clusterings (mittels Intraklassenkorrelation, hier nicht behandelt)
- Ausmaß serieller Abhängigkeit (Prüfung z.B. durch Autokorrelation, → longitudinal data analysis)

5. Normalverteilung der Residuen

Die Residuen sollten normalverteilt sein. Für große Stichproben ist die Annahme in der Praxis häufig nicht wichtig auf Grund asymptotischer Normalverteilung (*Zentraler Grenzwertsatz*). Die Parameterschätzung bleibt unverzerrt. Problematisch ist es aber in (1) kleinen Stichproben. Außerdem indiziert eine fehlende Normalverteilung häufig (2) andere Probleme der Modellspezifikation.

Zur Überprüfung können z.B. (1) Histogramme oder (2) QQ-Plots eingesetzt werden.

Im Zweifelsfall, d.h. wenn es Probleme mit der Normalverteilungsannahme gibt (z.B. bei kleinen Stichproben), kann auf verteilungsfreie, d.h. nicht-parametrische Verfahren zurück gegriffen werden.

Begriffe

Komponenten der Regressionsgleichung, Methode der kleinsten Quadrate, Partielle Regressionsgewichte, Populationsregressionsmodell, Stichprobenregressionsmodell, SSE, Bivariate Regression, Multiple Regression, Leverage, z-Standardisierung, Standardisierte Regressionsgewichte, Determinationskoeffizient, Person-Produkt-Moment Korrelation, Multiples R, Adjusted R^2, Semipartialkorrelation, Bestimmtheitsmaß, Partialkorrelation, Residualvarianz, Venn-Diagramm, Standardfehler, Standardfehler Regressionsgewicht, Toleranz, Indeterminationskoeffizient, Hypothesentest, Konfidenzintervall, F-Test, Gesamt aufgeklärte Varianz, 5 Annahmen MLR, Linearität, Residualplots, Exogene Prädiktoren, Homoskedastizität, Quadrierter Standardschätzfehler, Unabhängigkeit Residuen, iid, Clustering, Serielle Abhängigkeit, Normalverteilung Residuen, QQ-Plots

2.2 Pfadanalyse

Mit der multiplen Regression können nicht nur Variablen vorhergesagt, sondern auch Hypothesen zum Zusammenhang von Variablen geprüft werden. Dies wird als **Pfadanalyse** bezeichnet. Auf Basis theoretischer Überlegungen werden **Kausalmodelle** (Causal Models) erstellt, welche die Zusammenhänge zwischen den Variablen angeben.

Der Kausalbegriff ist nur bedingt korrekt. Ein Kausalzusammenhang wird im Kontext einer Pfadanalyse so verstanden, dass ein kausaler Zusammenhang angenommen und entsprechend berechnet wird. Ob ein wirklicher kausaler Zusammenhang vorliegt, ist damit nicht gewährleistet.

Die Auswertung der Pfadanalyse erfolgt in zwei Schritten:

1. Zur Auswertung werden häufig **Pfaddiagramme** verwendet. Hiervon werden die *Modellgleichungen* abgeleitet und die *Pfadkoeffizienten* bzw. Regressionskoeffizienten bestimmt.
2. Das Modell wird an die Daten angepasst (*Modellfit*). Je nach Ergebnis muss das Modell u.U. angepasst werden.

Die Pfadanalyse kann als Vorläufer von Strukturgleichungsmodellen mit latenten Variablen verstanden werden, siehe Abschnitt 4.

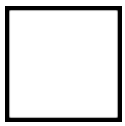
2.2.1 Hierarchischer F-Test

Der **hierarchische F-Test** bildet die Grundlage der Pfadanalyse. Der F-Test wurde (siehe Abschnitt 2.1.8) eingesetzt um Varianzunterschiede zu testen. Der F-Test kann im Kontext der Pfadanalyse auch zur Testung von Unterschieden in **Sets** (Variablengruppen) verwendet werden. Die Prüfgröße (F-Wert) berechnet sich wie folgt:

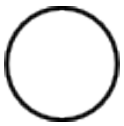
$$F = \frac{R^2_{Y,AB} - R^2_{Y,A}}{1 - R^2_{Y,AB}} \cdot \frac{N - k_A - k_b - 1}{k_B} \tag{2.31}$$

Dabei ist k_A die Anzahl der Variablen im Prädiktorset A, k_B die Anzahl der Variablen im Prädiktorset B und N die Stichprobengröße. $R^2_{Y,AB}$ ist die multiple Korrelation mit den Sets A und B, sie beschreibt wie gut A und B die Varianz von Y erklären können. $R^2_{Y,A}$ ist die multiple Korrelation mit Set A, sie beschreibt wie gut A die Varianz von Y erklären kann. Der Zählerterm $R^2_{Y,AB} - R^2_{Y,A}$ beschreibt wie gut B die Varianz von Y erklären kann (aber auch unter Berücksichtigung von A), der Nennerterm $1 - R^2_{Y,AB}$ beschreibt den Anteil der Varianz, der durch A und B *nicht* erklärt werden kann.

2.2.2 Pfaddiagramme und Notation



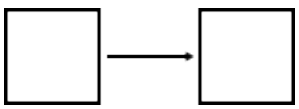
(a) Manifeste Variable (beobachtet)



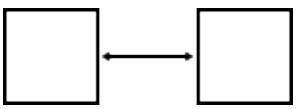
(b) Latente Variable (nicht beobachtet)

(c) Konstante (Mittelwerte/Regress.)

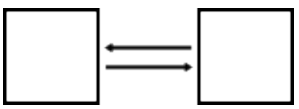
Abbildung 3: Symbole für Variablen bei der Pfadanalyse. Latente Variablen werden auch als „Faktoren“ oder „Fehlerterme“ bezeichnet. Konstanten dienen zur Darstellung von Mittelwerten und Regressionskoeffizienten.



(a) Gerichtete Beziehung („Kausal“-Pfad)



(b) Ungerichtete Beziehung (Kovarianz/Korrelation)



(c) Reziproke/Feedback Beziehung (Nicht-Rekursive Modelle)

Abbildung 4: Relationen zwischen Variablen bei der Pfadanalyse.

Zur Erstellung von Pfaddiagrammen gibt es bestimmte Konventionen für die verwendeten Symbole. Die üblichen Symbole sind in Abbildung 3, die Relationen in Abbildung 4 dargestellt. Die grafische und mathematische Repräsentation ist häufig äquivalent. Oft dient sie jedoch auch der einfachen Illustration.

Sprechweise: „Regression von Y auf X“ entspricht der „Vorhersage von Y durch X“

Variablenarten

- **Manifeste/Latente Variablen**: Direkt beobachtete / nicht direkt beobachtete Variablen
- **Endogene Variablen**: Variablen welche vorhergesagt werden (und nicht andere Variablen erklären)
- **Exogene Variablen**: Variablen welche vorhersagen (und nicht durch andere erklärt werden)
- **Mediatorvariablen**: Variablen welche Effekte vermitteln (von einer/mehrere Variablen auf eine/mehrere Variablen)
- **Konfundierende Variablen**: Variablen welche in Wirklichkeit (zumindest teilweise) einen Effekt bewirken. Werden diese nicht berücksichtigt, kommt es zu Scheineffekten (siehe unten).
- **Moderierende Variablen**: Variablen welche in Abhängigkeit zu einem Effekt stehen. Ändert sich die Variable, so ändert sich der Zusammenhang zwischen den ursprünglich betrachteten Variablen.
- **Residualvariablen**: Latente Variablen welche die nicht-erklärte Varianz beschreiben

Die Beziehungen einiger Variablenarten sind in Abbildung 5 dargestellt.

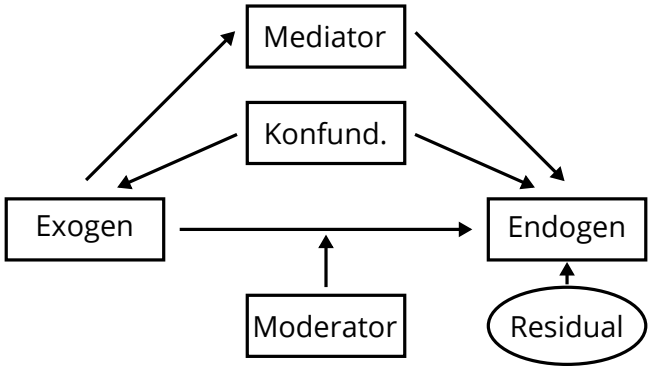


Abbildung 5: Variablenarten: Die Beziehung zwischen endogener und exogener Variabler mit Mediatorvariable, Moderatorvariable und konfundierender Variable, sowie einer latenten Residualvariable.

Effektarten

- **Direkter Effekt**: Effekt von einer Variablen auf eine andere Variable (unter Kontrolle aller anderen)
- **Indirekter Effekt**: Effekt von einer Variablen auf eine andere Variable, über eine/mehrere andere Variablen (Mediatoreffekt)
- **Totaler Effekt**: Summe aller indirekter Effekte addiert mit dem direkte Effekt einer Variablen
- **Zero-Order-Effekt**: Effekt von einer Variablen auf eine andere Variable (unter Ignorierung aller anderen)
- **Scheineffekt**: Differenz zwischen Zero-Order-Effekt und totalem Effekt

Die Effekte sind in Abbildung 6 an Beispielen dargestellt.

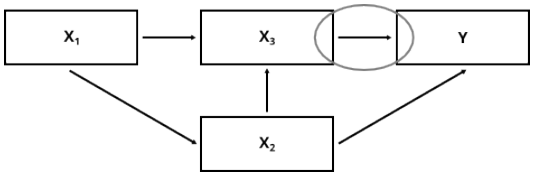
Rekursive/Nicht-Rekursive Pfadmodelle

Rekursive Pfadmodelle erfüllen zwei Bedingungen:

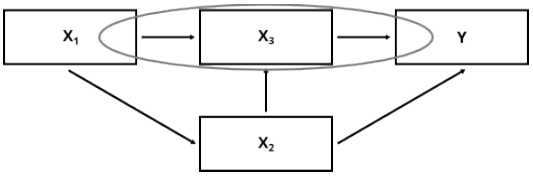
1. Hierarchisch: Kausaler Fluss geht nur in eine Richtung
2. Wechselseitig unkorrelierte Residualterme (Residualterme hängen nicht miteinander zusammen)

In der folgenden Betrachtung wird sich auf rekursive Pfadmodelle beschränkt. Zur Betrachtung nicht-rekursiver Modelle (z.B. mit Feedback-Schleifen) ist unter anderen Voraussetzung möglich.

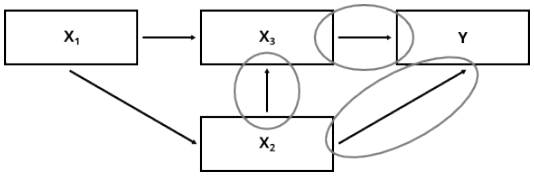
Die Effekte in Pfadmodellen können im (1) *hierarchischen* und im (2) *simultanen* Ansatz bestimmt werden. Die Ansätze sollen anhand eines Beispiels betrachtet werden. Wieder wird das Beispiel von oben betrachtet. Es soll



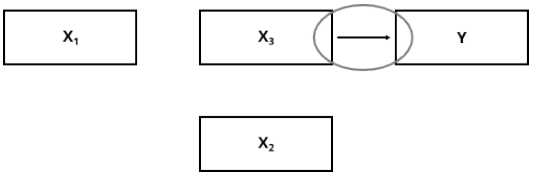
(a) Direkter Effekt von X_3 auf Y



(b) Indirekter Effekt von X_1 via X_3 auf Y



(c) Totaler Effekt von X_2 auf Y



(d) Zero-Order-Effekt von X_3 auf Y

Abbildung 6: Effekte zwischen Variablen bei der Pfadanalyse.

untersucht werden welche Faktoren das Gehalt eines Professors bestimmen. Das zu untersuchende Modell ist in Abbildung 7 dargestellt.

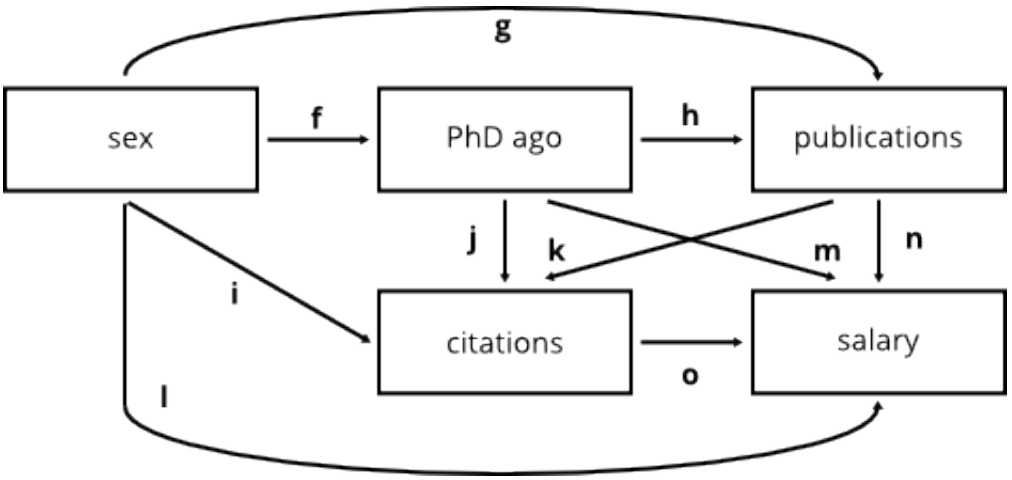


Abbildung 7: Beispiel-Modell der Einflüsse auf das Gehaltes eines Professors in der Pfadanalyse.

2.2.3 Hierarchischer Ansatz

Es sind viele verschiedene Einflüsse von Drittvariablen möglich. Um herauszufinden welche Drittvariablen eine Rolle spielen, wird im **hierarchischen Ansatz** ein Prädiktor nach dem anderen hinzugenommen (hierarchisch). Im Beispiel in Abbildung 7 wäre das Vorgehen wie folgt:

1. sex → salary
2. sex & PHD ago → salary
3. sex, PHD ago & publications → salary
4. sex, PHD ago, publications & citations → salary

In dem Modell ergeben sich folgende Effekte:

- Zero-Order-Effekt: $r_z = r_{sex,salary}$
- Direkter Effekt: $r_d = r_l$
- Indirekte Effekte: r_c mit $c \in \{f, g, h, i, j, k, m, n, o\}$
  - via citations: $i \cdot o$

  - via publications: $g \cdot n$ und $g \cdot k \cdot o$
  - via PhD ago: $f \cdot m$, $f \cdot h \cdot n$, $f \cdot j \cdot o$ und $f \cdot h \cdot k \cdot o$
- Totaler Effekt: $R = \sum r_{\text{indirekt}} + r_l = r_{io} + r_{gn} + r_{gko} + r_{fm} + r_{fhn} + r_{fjo} + r_{fhko} + r_l$

Da überwiegend totale Effekte erfasst werden, erfolgt die Berechnung der indirekten Effekte durch Subtraktion.

2.2.4 Simultaner Ansatz

Im **simultanen Ansatz** wird die abhängige Variable variiert. In jedem Schritt wird eine unabhängige Variable mehr hinzugenommen, während gleichzeitig eine andere abhängige Variable gewählt wird. Es wird in der Reihenfolge des Modells vorgegangen. Im Beispiel in Abbildung 7 wäre das Vorgehen wie folgt:

1. sex → PHD ago
2. sex & PHD ago → publications
3. sex, PHD ago & publications → citations
4. sex, PHD ago, publications & citations → salary

Die Effekte entsprechen denen des hierarchischen Ansatzes. Da die Effekte jedoch immer direkt erfasst werden, erfolgt die Berechnung der indirekten Effekte durch Multiplikation direkter Effekte.

Begriffe

Pfadanalyse, Kausalmodell, Pfaddiagramme, Notation, Modellgleichungen, Pfadkoeffizienten, Auswertung der Pfadanalyse, Hierarchischer F-Test, Sets, Sprechweise Regression, Variablenarten, Effektarten, Rekursive Pfadmodelle, Hierarchischer Ansatz, Simultaner Ansatz

2.3 Kategoriale Prädiktoren

Bei der multiplen linearen Regression wurden bisher nur *quantitativ-kontinuierliche* Prädiktoren betrachtet. Es können jedoch auch *qualitativ-kategoriale* Prädiktoren betrachtet werden. Die Begriffe der „Kategorie“ der „Ausprägung“ und der „Stufe“ werden synonym verwendet.

Untersucht werden dafür Unterschiede in Erwartungswerten. Es wird also ein ähnlicher Ansatz verfolgt wie bei einer Varianzanalyse. Hier wird jedoch ein regressionsanalytischer Ansatz verwendet. (1) Beobachtungseinheiten (z.B. Personen) unterscheiden sich (2) kategorial (z.B. eine/keine Teilnahme an einem Kurs) und werden bzgl. eines (3) Kriteriums (z.B. Leistung) untersucht.

2.3.1 Kodierung kategorialer Variablen

Definition

Für eine kategoriale Variable mit k Ausprägung (wobei $k \in 1, \ldots, K$) gilt:

- $K \geq 2$
- Erschöpfend: Jeder Beobachtungseinheit (z.B. Person) wird eine Ausprägung zugewiesen
- Ausschließlich: Es wird nur genau eine Ausprägung zugewiesen

Die kategorialen Variablen werden mit **Indikatorvariablen** bzw. **Dummyvariablen** beschrieben. Diese können nur zwei Werte annehmen (z.B. 0 für nicht ausgeprägt und 1 für ausgeprägt). Ausreichend sind $1 - K$ Indika-

torvariablen. Da die kategorialen Variablen *ausschließlich* sind, ergibt sich die K-te Indikatorvariable aus den jeweils anderen.

Beispiel

Die Hypothese lautet, dass ein kognitives Training die kognitive Leistungsfähigkeit verbessert. Dafür wird eine dreistufige unabhängige Variable verwendet. Es erfolgt ein Prätest, die Intervention und ein Posttest. Gemessen wird die kognitive Leistungsfähigkeit (abhängige Variable). Die unabhängige Variable hat folgende drei Stufen:

- Kontrollgruppe (KG): Passiv
- Experimentalgruppe 1 (EG_1): Fitnesstraining
- Experimentalgruppe 2 (EG_2): Gedächtnistraining

Da kategoriale Variablen per Definition *ausschließlich* sind, kann eine Person nur in einer der drei Gruppen teilnehmen, nicht in mehreren. Daher können folgende Indikatorvariablen verwendet werden:

- x_1 bildet EG_1 ab
- x_2 bildet EG_2 ab

Die Indikatorvariable x_3 für die Kontrollgruppe KG ergibt sich aus den beiden anderen und ist nicht notwendig. Wenn z.B. $x_1 = 0$ und $x_2 = 0$ (wenn die Person nicht in EG_1 und nicht in EG_2 ist), dann muss $x_3 = 1$ sein (dann muss die Person in KG sein).

Kodierung

Zur Kodierung können folgende gängige Verfahren angewendet werden:

1. Dummy Kodierung
2. Effekt Kodierung
3. Kontrast Kodierung

Nach der Kodierung können die Variablen in einer multiplen Regression verwendet werden. Dabei kann die Art der Kodierung einen Einfluss auf die Resultate haben. Für die Einflüsse gilt:

- Keinen Einfluss auf statistische Prüfung des Gesamteffekts (R, R^2 und F-Test)
- Einfluss auf Interpretation der Regressionskoeffizienten
- Einfluss welche Vergleiche inferenzstatistisch abgesichert sind (??)

2.3.1.1 Dummy Kodierung

Bei der **Dummy Kodierung** wird eine der k Ausprägungen (von Insgesamt K Ausprägungen) als **Referenzkategorie** festgelegt. Diese bekommt auf allen Indikatorvariablen (Merke: Anzahl ist $K - 1$) den Wert 0. Die anderen Kategorien werden mit den Indikatorvariablen erschöpfend und ausschließlich belegt. D.h. genau eine Ausprägung erhält 1, alle anderen 0. Wird KG als Referenzkategorie gewählt, so gilt für das Beispiel:

Training	x_1	x_2
KG	0	0
EG_1	1	0
EG_2	0	1

Die Interpretation der unstandardisierten Regressionskoeffizienten erfolgt wie folgt:

- $\hat{b}_0$: Wert für $\hat{y}$ unter der Voraussetzung, dass $x_1, x_2 = 0$
- $\hat{b}_1$: Änderung für $\hat{y}$ unter der Voraussetzung, dass x_1 verändert und x_2 konstant gehalten wird
- $\hat{b}_2$: Änderung für $\hat{y}$ unter der Voraussetzung, dass x_2 verändert und x_1 konstant gehalten wird

Damit können die Regressionskoeffizienten wie folgt bestimmt werden:

$$\hat{y}_{KG} = \hat{b}_0 + \hat{b}_1(x_1 = 0) + \hat{b}_2(x_2 = 0) = \hat{b}_0 \tag{2.32a}$$

$$\hat{y}_{EG_1} = \hat{b}_0 + \hat{b}_1(x_1 = 1) + \hat{b}_2(x_2 = 0) = \hat{b}_0 + \hat{b}_1 \tag{2.32b}$$

$$\hat{y}_{EG_2} = \hat{b}_0 + \hat{b}_1(x_1 = 0) + \hat{b}_2(x_2 = 1) = \hat{b}_0 + \hat{b}_2 \tag{2.32c}$$

Durch Umformen der Gleichungen 2.32 ergeben sich die Regressionskoeffizienten:

$$\hat{b}_0 = \hat{y}_{KG} \tag{2.33a}$$

$$\hat{b}_1 = \hat{y}_{EG_1} - \hat{b}_0 = \hat{y}_{EG_1} - \hat{y}_{KG} \tag{2.33b}$$

$$\hat{b}_2 = \hat{y}_{EG_2} - \hat{b}_0 = \hat{y}_{EG_2} - \hat{y}_{KG} \tag{2.33c}$$

Anwendung

Die Dummy Kodierung kann verwendet werden, wenn der Effekt von EG_1 und EG_2 in Referenz zur Kategorie KG interessiert (Referenzkategorie).

2.3.1.2 Effekt Kodierung

Die **Effekt Kodierung** kann (1) *ungewichtet* und (2) *gewichtet* erfolgen.

Ungewichtete Effekt Kodierung

Bei der **ungewichteten Effekt Kodierung** wird eine der k Ausprägungen als **Basiskategorie** definiert. Die Auswahl der Basiskategorie wird durch *inhaltliche Kriterien* bestimmt, die Wahl hat keine Auswirkung auf statistische Effekte. Die Basiskategorie (im Beispiel KG) erhält den Wert -1. Die anderen Kategorien (im Beispiel EG_1 und EG_2) bekommen jeweils 0 oder 1 zugewiesen. Dabei darf eine Kategorie nur auf einer Indikatorvariable eine 1 besitzen. Die Indikatorvariable darf ebenso nur für eine kategoriale Variable eine 1 besitzen.

Training	x_1	x_2
KG	-1	-1
EG_1	1	0
EG_2	0	1

Die Interpretation der unstandardisierten Regressionskoeffizienten erfolgt analog zur Interpretation in Abschnitt 2.3.1.1, sie werden auf Grund der unterschiedlichen Indikatorvariablen anders gebildet:

$$\hat{y}_{KG} = \hat{b}_0 + \hat{b}_1(x_1 = -1) + \hat{b}_2(x_2 = -1) = \hat{b}_0 - \hat{b}_1 - \hat{b}_2 \tag{2.34a}$$

$$\hat{y}_{EG_1} = \hat{b}_0 + \hat{b}_1(x_1 = 1) + \hat{b}_2(x_2 = 0) = \hat{b}_0 + \hat{b}_1 \tag{2.34b}$$

$$\hat{y}_{EG_2} = \hat{b}_0 + \hat{b}_1(x_1 = 0) + \hat{b}_2(x_2 = 1) = \hat{b}_0 + \hat{b}_2 \tag{2.34c}$$

Durch Umformen der Gleichungen 2.34 ergeben sich die Regressionskoeffizienten:

$$\hat{b}_0 = \hat{y}_{KG} + \hat{b}_1 + \hat{b}_2 \overset{(2.35b,2.35c)}{=} \frac{1}{3}\left(\hat{y}_{KG} + \hat{y}_{EG_1} + \hat{y}_{EG_2}\right) \tag{2.35a}$$

$$\hat{b}_1 = \hat{y}_{EG_1} - \hat{b}_0 \overset{(2.35a)}{=} \hat{y}_{EG_1} - \frac{1}{3}\left(\hat{y}_{KG} + \hat{y}_{EG_1} + \hat{y}_{EG_2}\right) \tag{2.35b}$$

$$\hat{b}_2 = \hat{y}_{EG_2} - \hat{b}_0 \overset{(2.35a)}{=} \hat{y}_{EG_2} - \frac{1}{3}\left(\hat{y}_{KG} + \hat{y}_{EG_1} + \hat{y}_{EG_2}\right) \tag{2.35c}$$

Gewichtete Effekt Kodierung

Die Ausprägungen kategorialer Variablen kommen meist nicht gleich oft vor. Um diese Schwankungen in der Auswertungen berücksichtigen zu können, kann eine **gewichtete Effekt Kodierung** verwendet werden.

Im Unterschied zur ungewichteten Effekt Kodierung wird bei der gewichteten Effekt Kodierung ein anderer Wert für die Basiskategorie gewählt. Während bei der ungewichteten Methode eine -1 gesetzt wird, wird bei der gewichteten Methode ein Verhältnis gesetzt, welches sich wie folgt berechnet:

$$-\frac{N_{\text{Kategorie}}}{N_{\text{Basis}}} \tag{2.36}$$

Dabei ist $N_{\text{Kategorie}}$ die Anzahl der Ausprägungen der betrachteten Kategorie. N_{Basis} entspricht der Anzahl der Ausprägungen der Basiskategorie.

Training	x_1	x_2
KG	$-\frac{N_{EG_1}}{N_{KG}}$	$-\frac{N_{EG_2}}{N_{KG}}$
EG_1	1	0
EG_2	0	1

Wären bspw. folgende Teilstichprobengrößen gegeben

$N_{EG_1} = 20$, $N_{EG_2} = 17$ und $N_{KG} = 25$

dann wären die Quotienten in der Basiskategorie mit $-\frac{N_{EG_1}}{N_{KG}} = -\frac{4}{5}$ und $-\frac{N_{EG_2}}{N_{KG}} = -\frac{17}{25}$ gegeben.

Die unstandardisierten Regressionskoeffizienten werden nun mit den gewichteten Indikatorvariablen der Basiskategorie gebildet. Die Regressionskonstante $\hat{b}_0$ entspricht dem gewichteten Gesamtmittelwert der Ausprägungen (der kategorialen Variable). Die weiteren Regressionskonstanten $\hat{b}_1$, $\hat{b}_2$, usw. entsprechen der Differenz zwischen dem Mittelwert der Ausprägung (der betrachteten kategorialen Variable) und dem gewichteten Gesamtmittelwert aller Ausprägungen.

Anwendung

Die Effekt-Kodierung kann verwendet werden, wenn der Effekt der Einzelwerte im Vergleich zum Gesamtschätzer (z.B. Gesamtmittelwert) über die Einzelkategorien hinweg interessiert. Der Gesamtmittelwert ist der Mittelwert über alle k Ausprägungen der kategorialen Variable.

2.3.1.3 Kontrast Kodierung

Bei der **Kontrast Kodierung** werden die Kategorien in *drei* **Sets** eingeteilt. Die Sets werden mit u, v und w bezeichnet. Die Kategorien in Set u und v werden verglichen, während Set w vom Vergleich ausgeschlossen wird. Die Kategorien in den jeweiligen Sets erhalten folgende Werte:

- Kategorien in Set u: $-\frac{N_v}{N_u+N_v}$
- Kategorien in Set v: $+\frac{N_u}{N_u+N_v}$
- Kategorien in Set w: 0

Dabei entspricht N_u bzw. N_v der Anzahl der Kategorien innerhalb des Sets.

Bei der Konstruktion müssen zwei Regeln beachtet werden:

- $\sum_{k=1}^{k=K} c_{k,j} = 0$, d.h. die Summe der Werte aller Kategorien muss Null sein
- $\sum_{k=1}^{k=K} c_{k,j} c_{k,j'} = 0$, d.h. die Summe der Produkt-Werte-Paare aller Kategorien muss Null sein

Beispiel

Training	x_1	x_2
KG	$-\frac{2}{3}$	0
EG_1	$\frac{1}{3}$	$\frac{1}{2}$
EG_2	$\frac{1}{3}$	$-\frac{1}{2}$

In diesem Fall besteht für die Indikatorvariable x_1 das Set u aus KG (damit ist $N_u = 1$) und das Set v aus EG_1 und EG_2 (damit ist $N_v = 2$). Für x_2 ist das Set u einfach EG_1 ($N_u = 1$), v ist EG_2 (damit ist $N_v = 1$) und w ist KG. Unter Verwendung der Formeln ergeben sich die Werte in der Tabelle.

Die Regressionskonstante $\hat{b}_0$ entspricht dem ungewichteten Gesamtmittelwert der kategorialen Variable für alle Kategorien des Prädiktors. Dies entspricht der ungewichteter Effektkodierung. Die Regressionsgewichte $\hat{b}_1$, $\hat{b}_2$, etc. entsprichen der Differenz zwischen dem ungewichteten Mittelwert in Set u und dem ungewichteten Mittelwert in Set v.

Anwendung

Die Kontrast-Kodierung kann verwendet werden, wenn bestimmte Kategorien direkt miteinander verglichen werden sollen.

2.3.1.4 Überblick Kodierungsarten

Die Kodierarten lassen sich vereinfacht wie folgt beschreiben:

- Dummy-Kodierung: Unterschiede zwischen Gruppen
- Effekt-Kodierung: Gruppen werden zusammengenommen und mit anderen zusammengenommenen Gruppen verglichen
- Kontrast-Kodierung: Mehrere Gruppen werden zusammen genommen (gemittelt) und mit einer Basisgruppe verglichen

Eine Gewichtung kommt in Frage, wenn die Gruppen unterschiedlich groß sind bzw. genauer, wenn die Größenunterschiede berücksichtigt werden sollen.

Wichtig ist, dass sich das Kodierschema an der *inhaltlichen Fragestellung* und der Interpretation orientiert. Kodierschemata ohne sinnvolle Interpretation der Koeffizienten sollten nicht angewendet werden.

2.3.2 Allgemeines Lineares Modell (ALM)

Mit Hilfe des **Allgemeinen Linearen Modells** (ALM) können viele inferenzstatistische Verfahren verallgemeinert werden. Dafür müssen zwei Voraussetzungen erfüllt sein:

- Für die Variablensätze müssen multiple Korrelationen und Regressionen berechenbar sein
- Nominale Merkmale müssen durch Indikatorvariablen abgebildet werden können

2.3.2.1 t-Test

Werden die Gruppenzugehörigkeiten effektkodiert abgebildet, kann ein t-Test für unabhängige Stichproben damit durchgeführt werden. Hierfür wird eine einfache Regression durchgeführt.

$$H_0 : \mu_1 = \mu_2 \; ; \quad H_1 : \mu_1 \neq \mu_2 \tag{2.37}$$

2.3.2.2 ANOVA

In einer einfaktoriellen ANOVA (analysis of variance, Varianzanalyse) wird eine unabhängige Variable mit p Stufen in Beziehung zu einer abhängigen Variable gesetzt. Werden die p Stufen mit $p-1$ Indikatorvariablen kodiert, kann die ANOVA mit einer Regression berechnet werden (hier nicht weiter betrachtet, siehe https://de.wikipedia.org/wiki/Varianzanalyse). Der F-Test der Varianzanalyse und der F-Test für multiple Korrelationen (zwischen Kriterium und Indikatorvariable) stimmt überein.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p \; ; \quad H_1 : \mu_i \neq \mu_j \tag{2.38}$$

2.3.2.3 ANCOVA

In einer ANCOVA (analysis of covariance, Kovarianzanalyse) werden Fehlervarianzen reduziert (und damit die Teststärke bzw. Power). Nicht relevante unabhängige Variablen sollen ausgeblendet werden. Bei der ANCOVA wird eine lineare Regression verwendet. Daher muss zusätzlich zu (1) normalverteilten Residuen und der (2) Homoskedastizität auch eine (3) lineare Abhängigkeit der abhängigen von den unabhängigen Variablen erfüllt sein.

Begriffe

Kategoriale Prädiktoren, Eigenschaften, Indikatorvariable, Dummyvariable, Einflüsse, Dummykodierung, Referenzkategorie, Effekt-Kodierung (Gewichtet/Ungewichtet), Basiskategorie, Kontrastkodierung, Sets, Anwendungen der Kodierungsarten, Allgemeines lineares Modell, t-Test, ANOVA, ANCOVA

3 Logistische Regression

3.1 Allgemeines und verallgemeinertes lineares Modell

Das **allgemeine lineare Modell** wurde bereits in Gleichungen 2.1 und 2.10 beschrieben. Im allgemeinen linearen Modell wird von einem *normalverteilten* Fehlerterm ausgegangen, der zu einer ebenfalls normalverteilten abhängigen Variablen führt.

Im **verallgemeinerten linearen Modell** (generalized linear model, GLM) wird die Voraussetzung an die Verteilung des Fehlerterms gelockert. Der Fehlerterm kann im verallgemeinerten linearen Modell aus der Exponentialfamilie kommen (z.B. Normalverteilung, Binomialverteilung, Poisson-Verteilung, Gamma-Verteilung, etc.). Da der Parameter abhängig von der zu Grunde liegenden Verteilung ist (z.B. Normalverteilung: $N(\mu, \sigma)$ hat Parameter μ und σ, Binomialverteilug: $Bin(n, p)$ hat Parameter n und p, etc.), ist der Parameter nun kein fester Wert mehr, da die Verteilung beliebig sein kann. Der Parameter μ wird zu einer Funktion, welche als **Link Funktion** g bezeichnet wird. Unter *iid*-Annahme (unabhängig und gleichverteilt, wobei mit gleichverteilt gemeint ist, dass die einzelnen Variablen der gleichen Verteilung folgen) gilt:

$$\mathbb{E}[\mathbf{y}] = \mu = g^{-1}(\mathbf{Xb}) \tag{3.1}$$

Auch die Varianz wird zu einer Funktion, dabei wird V als **Varianz Funktion** bezeichnet. Diese ist vom Mittelwert μ abhängig.

$$Var[\mathbf{y}] = \phi V(\mu) = V(g^{-1}(\mathbf{Xb})) \tag{3.2}$$

Dabei wird der *lineare Prädiktor* häufig mit $\eta = \mathbf{Xb}$ bezeichnet.

Das *allgemeine lineare Modell* kann als Spezialfall des *verallgemeinerten linearen Modells* betrachtet werden. Die Link Funktion wird dann als *Identity Link* bezeichnet. Es gilt:

$$g(\mu) = \mathbf{Xb} = \mu \tag{3.3a}$$

$$V(\mu) = 1 \tag{3.3b}$$

Auch ab Abschnitt 3.3 betrachtete logistische Regression ist eine Variante des GLM. Hier wird die *Binomiale Link Funktion* verwendet:

$$g(\mu) = \mathbf{Xb} = \ln\left(\frac{\mu}{\mathbf{1} - \mu}\right) \tag{3.4a}$$

$$V(\mu) = \mu(\mathbf{1} - \mu) \tag{3.4b}$$

3.2 Diskriminanzanalyse

Wird allgemein eine *dichotome abhängige Variable* betrachtet, kann auch hier das lineare Modell angenommen werden. Es gilt dabei jedoch zu überprüfen, ob die entsprechenden Annahmen (siehe Abschnitt 2.1.9) noch gelten. Beispielsweise ist die *Homoskedastizitätsannahme* verletzt. Auch *Linearität* und *Normalverteilung* sind nicht unbedingt erfüllt. Zusätzlich sind abhängige Variablen, welche nicht eindeutig 0 oder 1 sind schwierig zu interpretieren, v.a. dann wenn die Werte z.B. größer als 1 werden.

Zur Klassifikation kann die **Diskriminanzanalyse** verwendet werden, die versucht zwischen zwei Gruppen möglichst gut zu unterscheiden bzw. zu diskriminieren. Mathematisch ist dieses Modell äquivalent zum *linear probability model*, auch hier werden normalverteilte Fehler angenommen. Entsprechend liegen die bereits genannten Nachteile vor. Es muss gelten:

- Normalverteilte Prädiktoren
- Homogene Kovarianzmatrix

Da die Voraussetzungen meist nicht erfüllt sind, wird meist eine *logistische Regression* durchgeführt.

3.3 Drei Formen der logistischen Regressionsgleichung

In der **logistischen Regression** wird die *Wahrscheinlichkeit* $\hat{p}_i$ angegeben, dass ein dichotomer Fall eintritt bzw. dass eine der beiden Gruppen zutrifft. Dazu wird eine nicht-lineare S-förmige Funktion verwendet. Es handelt sich meist um eine *probit* oder *logit* Funktion. Die Modelle dazu werden als *Probitmodell* bzw. *Logitmodell* bezeichnet (hiervon ist der Begriff der logistischen Regression abgeleitet). Im Folgenden wird nur das Logitmodell betrachtet. Die Regression dann folgt prinzipiell folgender Funktion (wobei $e \approx 2,71828$ der eulerschen Zahl entspricht):

$$f(x) = \frac{e^x}{1 - e^x} \tag{3.5}$$

Die **Logistische Regressionsgleichung** lautet dann konkret wie folgt:

$$\hat{p} = P(y = 1|x_1 \dots x_K) = \frac{1}{1 + \exp\left[-(b_0 + b_1 x_1 + \cdots + b_k x_k + \cdots + b_K x_K)\right]} \tag{3.6}$$

Umgeschrieben gilt auch:

$$\hat{p} = P(y = 1|x_1 \dots x_K) = \frac{\exp\left(b_0 + b_1 x_1 + \cdots + b_k x_k + \cdots + b_K x_K\right)}{1 + \exp\left(b_0 + b_1 x_1 + \cdots + b_k x_k + \cdots + b_K x_K\right)} \tag{3.7}$$

Die logistische Regressionsgleichung kann in drei Formen beschrieben werden:

$$\hat{p} = \frac{1}{1 + e^{-z}} \qquad \text{(Wahrscheinlichkeit)} \tag{3.8a}$$

$$\frac{\hat{p}}{1 - \hat{p}} = e^z \qquad \text{(Odds)} \tag{3.8b}$$

$$\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = logit = z \qquad \text{(Logit)} \tag{3.8c}$$

Wobei $z = (b_0 + b_1x_1 + \cdots + b_kx_k + \cdots + b_Kx_K)$, $\hat{p}$ entspricht der geschätzten Wahrscheinlichkeit und $1 - \hat{p}$ der Gegenwahrscheinlichkeit.

3.3.1 Beispiel: Beförderung

Ob eine Person befördert wird oder nicht entspricht einer dichotomen abhängigen Variable. Eine Vorhersage soll auf Basis der Publikationen getroffen werden.

Logit

Für die Logit ergibt sich dann beispielsweise:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = logit = (b_0 + b_1x_1) = (-4,82 + 0,299x_1) \tag{3.9}$$

Für keine Publikationen beträgt die logit $-4,82$, mit jeder weiteren Publikation erhöht sich die logit um $0,299$. Diese Daten sind an sich nur wenig anschaulich. Anschaulicher ist es mit den *Odds* zu arbeiten.

Odds

Die *Odds* (die als **Chance** interpretiert werden kann) erhalten wir über:

$$e^{\ln(logit)} = \exp\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \frac{\hat{p}}{1-\hat{p}} = e^z = e^{(b_0+b_1x_1)} = e^{b_0}e^{b_1x_1} \tag{3.10}$$

In diesem Fall geben die Odds ein Verhältnis an, zwischen der geschätzten Wahrscheinlichkeit *für eine* Beförderung zu der geschätzten Wahrscheinlichkeit *für keine* Beförderung. Die Werte liegen im Intervall $[0, \infty)$. Mit keinen Publikationen ($x_1 = 0$) ergibt sich im konkreten Beispiel:

$$ODD_0 = e^{(b_0+b_1x_1)} = e^{(-4,82+0,299\cdot 0)} \approx 0,008 \tag{3.11}$$

Gut interpretieren lässt sich der so genannte **Odds-Ratio**. Dieser gibt die Veränderung der Odds bzw. das Verhältnis von einem Odd zum darunter liegenden Odd an. Unter der Annahme von Linearität reicht es den Odd einmal für zwei Werte zu bestimmen, um den gesamten Verlauf vorhersagen zu können.

Der erste Odd wurde bereits in Gleichung 3.11 berechnet. Der zweite Odd ergibt sich mit einer Publikation ($x_1 = 1$):

$$ODD_1 = e^{(b_0+b_1x_1)} = e^{(-4,85+0,299\cdot 1)} = 0,011 \tag{3.12}$$

Der Odd-Ratio ergibt sich aus dem Quotien:

$$OR = \frac{ODD_1}{ODD_0} = \frac{0,011}{0,008} \approx 1,35 \tag{3.13}$$

Der Odd-Ratio ergibt sich in diesem Fall bereits aus dem ersten Regressionsgewicht:

$$OR = \frac{ODD_1}{ODD_0} = \frac{e^{(b_0+b_1\cdot 1)}}{e^{(b_0+b_1\cdot 0)}} = \frac{e^{b_0}e^{b_1}}{e^{b_0}} = e^{b_1} = e^{0,299} \approx 1,35 \tag{3.14}$$

Inhaltlich bedeutet das: Mit jeder Publikation mehr steigt die Wahrscheinlichkeit befördert zu werden um $1,35$. Wie bereits angesprochen bleibt der Odds-Ratio im linearen Fall *konstant*, während die Odds selbst ein multiplikatives Verhalten aufweisen. Prinzipiell gilt: Je kleiner b_1, desto geringer ist der Effekt.

Wahrscheinlichkeit

Mit Hilfe der *Wahrscheinlichkeit* kann angegeben werden, wie wahrscheinlich es ist, dass ein dichotomes Kriterium mit einer bestimmten Ausprägung (0 oder 1) eintritt. Die Wahrscheinlichkeit berechnet sich nach Gleichung 3.7. Beispielhaft soll berechnet werden wie hoch die Wahrscheinlichkeit für eine Beförderung ($y = 1$) ist, wenn die Person 10 Publikationen hat ($x_1 = 10$).

$$\hat{p} = P(y = 1|x_1 = 10) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(-4,82+0,299\cdot 10)}} \approx 0,1387 \tag{3.15}$$

Mit Hilfe des Quotienten aus b_0 und b_1 kann bestimmt werden, welchen Wert der Prädiktor annehmen muss, damit eine Wahrscheinlichkeit von 50% erreicht wird. Im Beispiel wäre dies:

$$\frac{-b_0}{b_1} = \frac{4,82}{0,299} \approx 16 \tag{3.16}$$

Diese Berechnung lässt sich nur durchführen, wenn es nur einen Prädiktor gibt.

Bei der Interpretation der Wahrscheinlichkeiten sollte sorgfältig vorgegangen werden. Die *Erhöhung der Wahrscheinlichkeit* um mehrere Prozent, muss nicht unbedingt ein großer Effekt sein, wenn die zu Grunde liegende Wahrscheinlichkeit sehr gering ist. Beispielsweise bedeutet eine Erhöhung der Wahrscheinlichkeit von $0,7\%$ auf 1% bereits eine Erhöhung um 34%.

3.4 Parameterschätzung mit Maximum-Likelihood

Die Koeffizienten der logistischen Regression kann nicht mit Hilfe der *Methode der kleinsten Quadrate* bestimmt werden, da die Voraussetzungen (wie oben angesprochen) nicht erfüllt sind. Zur Schätzung kann das **Maximum-Likelihood**-Verfahren verwendet werden. Maximum Likelihood beschreibt die gegebenen Daten unter der Voraussetzung eines bestimmten Modells. Die Parameter werden so bestimmt, dass die Wahrscheinlichkeit maximal wird (daher maximum likelihood), dass die Daten mit Hilfe des Modells beschrieben werden können. Genauer wird nach einem Parameter gesucht, der die Daten möglichst plausibel beschreibt.

Beispielsweise ist die Wahrscheinlichkeit für eine Person $i = 1$ mit dem Wert $x_1 = 1$ den Wert $y = 1$ zu erzielen gegeben durch (wobei $y \in \{0, 1\}$):

$$P(y = 1|x_{i=1} = 1) = \frac{e^{(b_0+b_1)}}{1 + e^{(b_0+b_1)}} \tag{3.17}$$

Die Wahrscheinlichkeit für Person $i = 2$ den Wert $y = 0$ zu erzielt, wäre:

$$P(y = 1|x_{i=2} = 1) = 1 - \frac{e^{(b_0+b_1)}}{1 + e^{(b_0+b_1)}} \tag{3.18}$$

Unter der Voraussetzung *iid* verteilter Variablen (unabhängig und gleichverteilt) können die Wahrscheinlichkeiten einfach multipliziert werden. Im allgemeinen Fall kann die Likelihood-Funktion L (für die logistische Regression) wie folgt formuliert werden:

$$L = \prod_{i=1}^{N} P(y = 1|x_i)^{y_i} \cdot [1 - P(y = 1|x_i)]^{(1-y_i)} \tag{3.19}$$

Dabei ist y_i der realisierte Wert, der jeweiligen Person (also 0 oder 1), entsprechend ist nur der vordere oder der hintere Term relevant je Person. Die Wahrscheinlichkeiten aller Personen werden multipliziert, sie hängen von b_0 und b_1 ab. Nun werden b_0 und b_1 so gewählt, dass L maximal wird.

Zur Maximierung der Likelihood-Funktion L wird meistens zunächst der Logarithmus gebildet. Diese Funktion wird als **Log-Likelihood** $LL = \ln(L)$ bezeichnet. Dieser Trick vereinfacht die Rechnung meist sehr stark, da die entsprechenden Variablen in den meisten Verteilungen in einer e-Funktion enthalten sind.

Trotz der Vereinfachung durch die Logarithmierung ist das Maximum in vielen Fällen nicht analytisch zu finden, es kommen iterative (numerische) Verfahren zum Einsatz. Diese Verfahren setzen einen Startwert für die Parameter b_0 und b_1, der häufig auch selbst gewählt werden kann. Anschließend wird die Likelihood-Funktion bestimmt. Die Parameter werden systematisch variiert und für jedes Set von Parametern wird die Likelihood-Funktion berechnet (für jedes Set von Parametern wird eine *Iteration* durchgeführt). Anschließend wird aus allen Ergebnissen die Likelihood-Funktion verwendet, für die der Wert der Funktion maximal wurde. Die darin enthaltenen Parameter (welche zuvor eingesetzt wurden), werden anschließend für das Modell verwendet. Konkret können viele unterschiedliche Algorithmen verwendet werden, die sich z.B. darin unterscheiden, nach welcher Systematik Parameter eingesetzt werden.

3.5 Deviance

Die **Deviance** gibt die Abweichung des gewählten Modells (mittel Maximum Likelihood) zu einem theoretisch perfekten Modell an. Mit Hilfe der Deviance kann also die Modellgüte (lack of fit) bestimmt werden. Es können zunächst folgende Likelihoods definiert werden:

$L_{perfekt} = 1$	Maximum Likelihood im perfekten Modell (für gegebene Daten)
L_{null}	Maximum Likelihood in einem Modell, dass nur die Regressionskonstante (Intercept) enthält (*Nullmodell*), entspricht niedrigst möglicher Likelihood
L_K	Maximum Likelihood für ein gewöhnliches Modell mit Intercept und K Prädiktoren

Die Deviance ist wie folgt definiert:

$$Dev = -2 \ln\left(\frac{L_0}{L_1}\right) = -2\left[\ln(L_0) - \ln(L_1)\right] \tag{3.20}$$

Dabei entspricht L_0 einer Likelihood-Funktion zu einem beliebigen Modell. In Bezug dazu bezeichnet L_1 eine Likelihood-Funktion mit einem Modell, welches weniger sparsam (ineffizienter) ist. Es können nun auf Basis der

obigen genannten Likelihood-Funktionen folgende mögliche Deviance angegeben werden:

$$D_{null} = -2\left[\ln(L_{null}) - \ln(L_{perfekt})\right] = -2\,LL_{null} \qquad \text{(Null Deviance)} \tag{3.21a}$$

$$D_K = -2\left[\ln(L_K) - \ln(L_{perfekt})\right] = -2\,LL_K \qquad \text{(Model Deviance)} \tag{3.21b}$$

Dabei ist LL die Log-Likelihood-Funktion. Das Konzept der *Null Deviance* ist vergleichbar mit der Gesamtvarianz der abhängigen Variablen, das beste und schlechteste Modell werden verglichen. Die *Model Deviance* wird auch als *Residual Deviance* bezeichnet. Das Konzept der *Model Deviance* ist vergleichbar mit der Residualvarianz. Durch eine Aufnahme von weiteren Prädiktoren (Erhöhung von K) sollen die Daten besser ins Modell passen bzw. die Deviance sinken.

3.6 Pseudo R

Der Determinationskoeffizient R^2 (Anteil aufgeklärte Varianz am Kriterium) lässt sich nicht auf die logistische Regression übertragen. Es kann jedoch ein **Pseudo $\mathbf{R^2}$** formuliert werden, das Maß gibt jedoch *nicht* die aufgeklärte Varianz an und ist damit nicht direkt mit dem Determinationskoeffizient vergleichbar.

Es können übliche Varianten des Pseudo R^2 formuliert werden:

- Normed Fit Index
- Cox and Snell Index
- Nagelkerke Index

3.6.1 Normed Fit Index

Der **Normed Fit Index $\mathbf{R^2_L}$** ist wie folgt definiert:

$$R^2_L = \frac{D_{null} - D_K}{D_{null}} \tag{3.22}$$

Dabei ist $D_{null} = -2LL_{null}$ und $D_K = -2LL_K$ (siehe Abschnitt 3.5).

Konzeptuell (nicht mathematisch) ist R^2_L mit der Varianzaufklärung zumindest vergleichbar. D_{null} kann in Analogie zur totalen Varianz SS_{total} betrachtet werden. Ebenso kann D_K mit der Residualvarianz SS_{res} verglichen werden. Durch Umformung der Definition von R^2 (vgl. Gleichung 2.20) wird die Analogie zu R^2_L ersichtlich:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = \frac{SS_{total} - SS_{residual}}{SS_{total}} \tag{3.23}$$

Der Normed Fit Index liegt im Bereich $0 \leq R^2_L \leq 1$ Problematisch ist, dass R^2_L nicht monoton ansteigt mit dem Anstieg des Odds-Ratio.

3.6.2 Cox and Snell Index

Der **Cox and Snell Index $\mathbf{R^2_{CS}}$** ist wie folgt definiert:

$$R^2_{CS} = 1 - \exp\left(\frac{D_{null} - D_K}{N}\right) \tag{3.24}$$

Der Cox and Snell Index liegt im Bereich $0 \leq R^2_{CS} < 1$. D.h. der maximale Wert ist nicht 1, er hängt von D_{null} ab und liegt bei:

$$R^2_{CS_{max}} = 1 - \exp\left(\frac{D_{null}}{N}\right) \tag{3.25}$$

3.6.3 Nagelkerke Index

Der **Nagelkerke Index $\mathbf{R^2_N}$** normiert den Cox and Snell Index an seinem Maximum, so dass dieser im Bereich von $0 \leq R^2_N \leq 1$ liegt.

$$R^2_N = \frac{R^2_{CS}}{R^2_{CS_{max}}} \tag{3.26}$$

3.7 Statistische Inferenz

3.7.1 Likelihood Ratio Test

Um zwei Modell vergleichen zu können, kann der **Likelihood Ratio (LR) Test** angewendet werden. Dabei wird auf die Deviance (als Maß für den Modellfit, siehe Abschnitt 3.5) zurückgegriffen. Es wird genutzt, dass die Differenz zweier Deviance mit Modellen mit K und $K + M$ Prädiktoren χ^2 verteilt ist. Dabei ist das zweite Modell um M Prädiktoren erweitert, enthält jedoch immer noch die K Prädiktoren des ersten Modells.

$$\chi^2 = D_K - D_{K+M} \qquad \text{mit } df = M \tag{3.27}$$

3.7.2 z-Test und Wald-Test

Während der Likelihood Ratio Test zwar ein optimale Test ist, gibt es noch alternative Tests die Anwendung finden. Dazu gehören der **z-Test** (in R) und der **Wald-Test** (in SPSS). Die Nullhypothese der Tests geht davon aus, dass ein Regressionskoeffizient Null ist:

$$H_0 : b_k = 0 \tag{3.28}$$

z-Test

Zur Bestimmung des z-Wertes wird der geschätzte Parameter durch seinen geschätzten Standardfehler geteilt. Die Prüfgröße z ist asymptotisch standardnormalverteilt.

$$z = \frac{\hat{b}_k}{SE_{b_k}} \qquad \text{mit } z \sim N(0, 1) \tag{3.29}$$

Wald-Test

Der Wald-Test ergibt sich aus dem Quadrat der Prüfgröße des z-Tests. Die Prüfgröße z^2 ist asymptotisch χ^2-Verteilt.

$$z^2 = \frac{\hat{b}_k^2}{SE_{b_k}^2} \qquad \text{mit } z \sim \chi^2\,,\ df = 1 \tag{3.30}$$

3.8 Standardisierung

In der multiplen Regression wurde eine Standardisierung über zwei alternative Wege erreicht:

- z-Standardisierung aller Variablen vor der Berechnung der Regression
- Standardisierung der Regressionskoeffizienten mittels $\beta_k = b_k \frac{\sigma_k}{\sigma_y}$

Bei der logistischen Regression ist letzterer Ansatz problematisch. Es wäre nötig die Standardabweichung σ des Logits zu berechnen. Die sich ergebenden Parameter könnten inhaltlich nicht interpretiert werden.

Für die logistische Regression kann daher eine **Semistandardisierung** angewendet werden. Es werden nur die unabhängigen Variablen (Prädiktoren) z-standardisiert, nicht die abhängigen Variablen (Kriterium).

Da die Berechnung nur mit Umwegen und die Interpretation nur schwer möglich ist, wird auf standardisierte Regressionskoeffizienten in der logistischen Regression häufig verzichtet. Wenn Programme standardisierte Koeffizienten angeben, sollten diese inhaltlich geprüft werden. Sollen standardisierte Koeffizienten angegeben werden, sollten diese hinreichend eingeführt werden, um die Ergebnisse nachvollziehbar zu halten.

3.9 Bemerkungen

Annamhen der logistischen Regression

Die logistische Regression setzt eine korrekte Spezifikation des Regressions-Modells voraus. Die Spezifikationen sind analog zur multiplen Regression (vgl. Abschnitt 2.1.9). Zusätzlich muss für die Maximum Likelihood Schätzung die Stichprobe hinreichend groß sein. Die Schätzer besitzen ihre Eigenschaften (z.B. Verteilung) nur asymptotisch.

Mehrkategoriale Variablen

Die grundlegenden Konzepte der logistischen Regression lassen sich auch auf mehrkategoriale Variablen erweitern (hier nicht weiter betrachtet). Dazu zählen:

- Normalskalierte abhängige Variable: Multinomiale logistische Regression
- Ordinalskalierte abhängige Variablen: Ordinale logistische Regression

Begriffe

Verallgemeinertes lineares Modell, Link-Funktion, Varianz-Funktion, Linearer Prädiktor, Spezialfall ALM, linear probability model, Diskriminanzproblem, Logistische Regression, logit, probit, Logistische Regressionsgleichung, Drei Formen der logistischen Regression, Odds, Chance, Odds-Ratio, Maximum-Likelihood, Log-Likelihood, Deviance, Null-Deviance, Modell-Deviance, Residual-Deviance, Pseudo R^2, Normed Fit Index, Cox and Snell Index, Nagelkerke Index, Eigenschaften des Pseudo R^2, Modelltests, Likelihood Ratio Test, z-Test, Wald-Test, Standardisierung, Semistandardisierung, Annahmen der log. Regression, Mehrkategoriale Variablen

4 Strukturgleichungsmodelle

4.1 Grundlagen

Bisher wurden *multiple* Modelle betrachtet. Diese zeichnen sich durch mehrere Prädiktoren, aber nur ein Kriterium aus. In *multivariaten* Modellen werden zusätzlich mehrere Kriterien betrachtet. Strukturgleichungsmodelle sind multivariate Modelle.

Die Strukturgleichungsmodell verallgemeinern die bisher betrachteten Verfahren. Dazu werden für ein multivariates Modell mit vielen Variablen die unbekannten Parameter im Vektor $\boldsymbol{\theta}$ betrachtet.

Es wird die **modellimplizierte Kovarianzmatrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$** definiert. Diese hängt von den unbekannten Parametern $\boldsymbol{\theta}$ ab.

Die Abweichung zwischen der genannten *modellimplizierten* Kovarianzmatrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ und der *empirischen* Kovarianzmatrix $\boldsymbol{\Sigma}$ wird minimiert. Durch die Minimierung werden die unbekannten Parameter $\boldsymbol{\theta}$ geschätzt. Im Idealfall gilt:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} \tag{4.1}$$

Das gleiche Prinzip gilt auch für die Mittelwerte:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu} \tag{4.2}$$

Beispiel

Es wird betrachtet wie gut eine Person Flugzeuge in einer Simulation navigieren kann, auf Basis der kognitiven Leistungsfähigkeit (z.B. Intelligenz). Das Intercept beträgt $4,6682$. Der Anstieg (Slope) für die kognitive Leistung g beträgt $1,3871$. D.h. eine Person mit durchschnittlicher kognitiver Fähigkeit manövriert $4,6682$ Flugzeuge korrekt. Eine Person, die eine Standardabweichungen über der durchschnittliche kognitiver Fähigkeit besitzt, navigiert $1,3871$ Flugzeuge mehr korrekt. Das Regressionsmodell lautet wie folgt:

$$y_i = b_0 + b_1 \cdot x_i + e_i \quad \text{mit} \quad \hat{y}_i = \hat{b}_0 + \hat{b}_1 \cdot x_i \tag{4.3}$$

Die Kovarianzmatrix besitzt folgende Form:

$$\mathbf{\Sigma} = \begin{pmatrix} Var(y) & Cov(y,x) \\ Cov(x,y) & Var(x) \end{pmatrix} = \begin{pmatrix} 3,813 & 0,927 \\ 0,927 & 0,668 \end{pmatrix} \tag{4.4}$$

Die Einträge der Kovarianzmatrix lassen sich durch Einsetzen des Modells $y = b_1 x + e$ und mit Hilfe der Rechenregeln für die Kovarianz beschreiben als:

$$\begin{aligned} \mathbf{\Sigma} &= \begin{pmatrix} Var(y) & Cov(y,x) \\ Cov(x,y) & Var(x) \end{pmatrix} = \begin{pmatrix} Var(b_1 x + e) & Cov(b_1 x + e, x) \\ Cov(x, b_1 x + e) & Var(x) \end{pmatrix} \\ &= \begin{pmatrix} Var(b_1 x) + Var(e) & b_1 \cdot Cov(x,x) \\ b_1 \cdot Cov(x,x) & Var(x) \end{pmatrix} \\ &= \begin{pmatrix} b_1^2 \cdot Var(x) + Var(e) & b_1 \cdot Var(x) \\ b_1 \cdot Var(x) & Var(x) \end{pmatrix} \end{aligned} \tag{4.5}$$

Daraus folgt u.a. $Cov(y,x) = b_1 \cdot Var(x)$, damit kann der Koeffizient b_1 bestimmt werden:

$$b_1 = \frac{Cov(y,x)}{Var(x)} = \frac{0,927}{0,668} \approx 1,39 \tag{4.6}$$

Mit Hilfe der Gleichung $Var(y) = b_1^2 \cdot Var(x) + Var(e)$ und dem nun bekannten Koeffizienten b_1 kann die Varianz $Var(e)$ des Fehlers bestimmt werden:

$$\begin{aligned} Var(e) &= Var(y) - b_1^2 \cdot Var(x) \\ &= 3,813 - 1,39^2 \cdot 0,668 \approx 2,52 \end{aligned} \tag{4.7}$$

Mit $Var(y)$ und $Var(e)$ lässt sich nun auch der Determinationskoeffizient bestimmen (vgl. Gleichung 2.20):

$$R^2 = 1 - \frac{Var(e)}{Var(y)} = 1 - \frac{2,52}{3,813} \approx 0,34 \tag{4.8}$$

Durch Betrachtung der Mittelwertstruktur $\boldsymbol{\mu}$ lässt sich b_0 bestimmen. Zum einen gilt:

$$\boldsymbol{\mu} = \begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} 4,649 \\ -0,014 \end{pmatrix} \tag{4.9}$$

Zum anderen gilt hingegen:

$$\boldsymbol{\mu} = \begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b_0 + b_1 \cdot \bar{x} \\ \bar{x} \end{pmatrix} \tag{4.10}$$

Die obere Gleichung $\bar{y} = b_0 + b_1 \cdot \bar{x}$ lässt sich nach b_0 umstellen:

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 4,469 - 1,39 \cdot (-0,014) \approx 4,668 \end{aligned} \tag{4.11}$$

Die Ergebnisse dieses Verfahrens ergibt identische Ergebnisse zum Kleinste-Quadrate-Verfahren (OLS).

4.1.1 LISREL: Messodell und Strukturmodell

LISREL (linear structural relations) ist eine Software zur grafischen Darstellung von Strukturgleichungsmodellen. Gleichzeitig bezeichnet LISREL die entsprechende Notation.

Bei Strukturgleichsmodellen wird unterschieden zwischen (1) **Messmodell** und (2) **Strukturmodell**, welche in Abbildung 8 skizziert sind.

Das *Messmodell* umfasst einen Faktor ξ (für exogene Messmodelle) bzw. η (für endogene Messmodelle) und seine zugehörigen Indikatoren x bzw. y. Den Indikatoren sind entsprechende Fehler δ bzw. ϵ zugeordnet. Das *Strukturmodell* beschreibt die Zusammenhänge der Faktoren ξ bzw. η.

Zusammengefasst werden folgende Symbole eingeführt:

	Exogen	Endogen
Latente Variable / Faktor	ξ	η
Indikatoren	X	Y
Fehler	δ	ϵ

Mit Λ wird die *Faktorladungsmatrix*, dass heißt der Zusammenhang von ξ und X bzw. η und Y, bezeichnet. Γ und B geben die Zusammenhänge zwischen dem exogenen Faktoren ξ und dem endogenen Faktor η, respektive zwischen dem endogenen Faktor η und einem anderen endogenen Faktor η an. Der Fehler der Faktoren wird allgemein mit ζ bezeichnet.

Annahmen

Die Annahmen des *Messmodells* sind:

- $\mathbb{E}[\boldsymbol{\eta}] = 0$, $\mathbb{E}[\boldsymbol{\xi}] = 0$
- $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, $\mathbb{E}[\boldsymbol{\delta}] = 0$
- $\boldsymbol{\epsilon}$ unkorreliert mit $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ und $\boldsymbol{\delta}$
- $\boldsymbol{\delta}$ unkorreliert mit $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ und $\boldsymbol{\epsilon}$

Ebenso lauten die Annahmen des *Strukturmodells*:

- $\mathbb{E}[\boldsymbol{\eta}] = 0$, $\mathbb{E}[\boldsymbol{\xi}] = 0$, $\mathbb{E}[\boldsymbol{\zeta}] = 0$
- $\boldsymbol{\zeta}$ unkorreliert mit $\boldsymbol{\xi}$
- $\mathbf{I} - \mathbf{B}$ nicht singulär

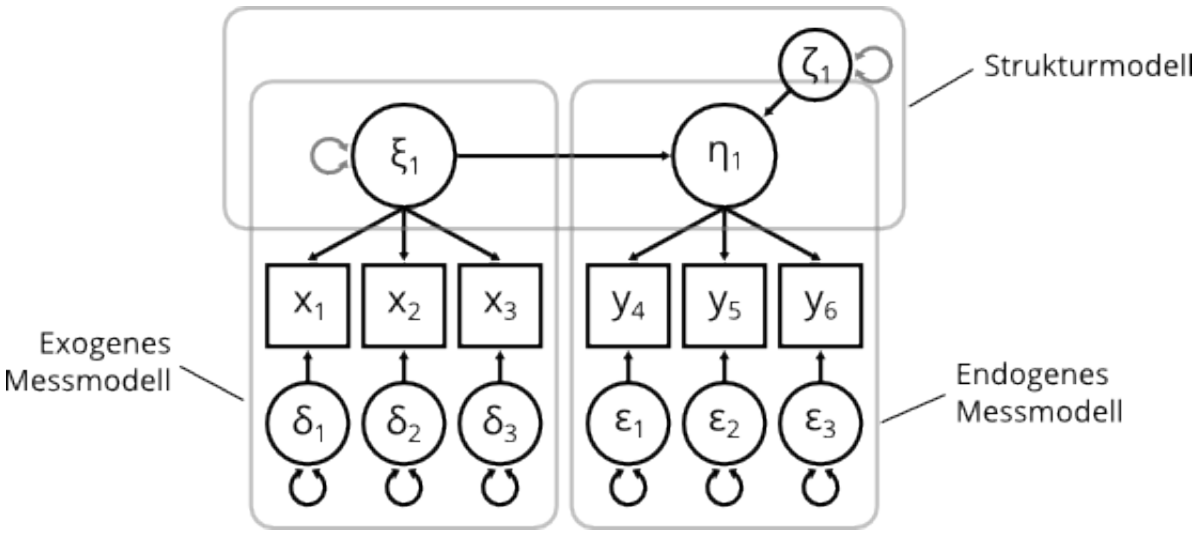


Abbildung 8: Strukturgleichungsmodell mit LISREL-Notation dargestellt. Messmodelle beschreiben einen Faktor und die zugehörigen Indikatoren, Strukturmodelle beschreiben die Zusammenhänge zwischen den Faktoren.

Exogenes Messmodell

In Matrixschreibweise wird das exogene Messmodell wie folgt beschrieben:

$$\mathbf{X} = \mathbf{\Lambda}\,\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.12}$$

Ausgeschrieben ergibt sich:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \dots & \lambda_{nk} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix} \tag{4.13}$$

Dabei ist n die Anzahl der Indikatoren und k die Anzahl der Faktoren. Sollen Einflüsse von bestimmten Indikatoren nur von bestimmten Faktoren berücksichtigt werden, so werden die entsprechenden Einträge Null gesetzt.

Beispiel

In diesem Fall liegen 4 Indikatoren und 2 exogene Faktoren vor. Jeweils nur 2 Indikatoren laden auf einen der exogenen Faktoren. D.h. $x_1 \wedge x_2 \to \xi_1$ mit δ_1 und δ_2, sowie $x_3 \wedge x_4 \to \xi_2$ mit δ_3 und δ_4. Das Modell wird wie folgt formuliert:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ 0 & \lambda_{32} \\ 0 & \lambda_{42} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{pmatrix} \tag{4.14}$$

Endogenes Messmodell

In Matrixschreibweise wird das endogene Messmodell wie folgt beschrieben:

$$\mathbf{Y} = \mathbf{\Lambda}\,\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{4.15}$$

Ausgeschrieben ergibt sich:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \dots & \lambda_{nk} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{4.16}$$

Strukturmodell

Das Strukturmodell beschreibt die Zusammenhänge der Faktoren. Für den Spezialfall, dass ausschließlich der Zusammenhang zwischen exogenen Faktoren und endogene Faktor angegeben werden soll, wobei ein Faktor nicht als exogener und endogener Faktor gleichzeitig auftritt, lautet das Modell in Matrixschreibweise:

$$\boldsymbol{\eta} = \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.17}$$

Ausgeschrieben ergibt sich:

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \dots & \gamma_{1k} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \dots & \gamma_{nk} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{pmatrix} \tag{4.18}$$

Dabei ist n die Anzahl der endogenen Faktoren und k die Anzahl der exogenen Faktoren.

Wenn endogene Variablen η wieder auf endogene Variablen wirken (also eine Variable endogene und exogene Variable gleichzeitig ist), dann wird η wieder auf der rechten Seite der Gleichung eingebracht und wir können das allgemeine Strukturmodell formulieren:

$$\boldsymbol{\eta} = \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\zeta} \tag{4.19}$$

Ausgeschrieben ergibt sich:

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \dots & \gamma_{1k} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \dots & \gamma_{nk} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} + \begin{pmatrix} \beta_{11} & \dots & \beta_{1k} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \dots & \beta_{nk} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_h \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{pmatrix} \tag{4.20}$$

Dabei ist n die Anzahl der endogenen Faktoren, die am Ende betrachtet werden, k die Anzahl der exogenen

Faktoren und h die Anzahl der endogenen Faktoren, die wieder als exogene Faktoren wirken.

Beispiele: Spezialfall

Beispiel mit zwei exogenen Faktoren ξ auf einen endogenen Faktor η (d.h. $\xi_1 \wedge \xi_2 \rightarrow \eta$), wobei die ξ die Indikatoren x besitzen:

$$\eta = \begin{pmatrix} \gamma_1 & \gamma_2 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \zeta \tag{4.21}$$

Beispiel mit zwei exogenen Faktoren ξ und zwei endogenen Faktoren η, wobei $\xi_1 \rightarrow \eta_1$ und $\xi_2 \rightarrow \eta_2$:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{4.22}$$

Beispiel: Allgemeiner Fall

Als Beispiel sollen wieder zwei exogene ξ und zwei endogene η betrachtet werden. Dabei soll zusätzlich zu den Zusammenhängen $\xi_1 \rightarrow \eta_1$ und $\xi_2 \rightarrow \eta_2$ (wie im vorigen Beispiel) auch der Zusammenhang zwischen zwei endogenen Faktoren $\eta_1 \rightarrow \eta_2$ gelten.

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{4.23}$$

4.1.2 Modellimplizierte Kovarianzmatrix

Allgemeine Kovarianzen

Zusammenhänge zwischen den Variablen können in einer Kovarianzmatrix beschrieben werden.

Im *Messmodell* werden zusammenhänge zwischen den Fehlertermen betrachtet. Die Kovarianzmatrixen sind:

- $\boldsymbol{\Theta_\delta} = \mathbb{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^T\right]$
- $\boldsymbol{\Theta_\epsilon} = \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]$

Für unkorrelierte Fehler gilt:

$$\boldsymbol{\Theta_\delta} = \begin{pmatrix} \sigma^2_{\delta_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{\delta_n} \end{pmatrix} \quad , \quad \boldsymbol{\Theta_\epsilon} = \begin{pmatrix} \sigma^2_{\epsilon_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{\epsilon_n} \end{pmatrix} \tag{4.24}$$

Im *Strukturmodell* werden folgende Kovarianzmatrizen betrachtet:

- $\boldsymbol{\Phi} = \mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right]$
- $\boldsymbol{\Psi} = \mathbb{E}\left[\boldsymbol{\zeta}\boldsymbol{\zeta}^T\right]$

Beispiel: Exogene Faktoren

Werden z.B. zwei exogene Faktoren ξ_1 und ξ_2 betrachtet, so ergibt sich die Kovarianzmatrix wie folgt:

$$\boldsymbol{\Phi} = \begin{pmatrix} Var(\xi_1) & Cov(\xi_1, \xi_2) \\ Cov(\xi_2, \xi_1) & Var(\xi_2) \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \tag{4.25}$$

Wobei $\phi_{12} = \phi_{21}$, d.h. eine Kovarianzmatrix ist symmetrisch, da die einzelnen Kovarianzen der Zufallsvariablen symmetrisch sind, d.h. $Cov(X, Y) = Cov(Y, X)$.

Beispiel: Endogene Faktoren

Analog zu den exogenen Faktoren, ergibt sich die Kovarianzmatrix für zwei endogene Faktoren ζ_1 und ζ_2 wie folgt:

$$\boldsymbol{\Psi} = \begin{pmatrix} Var(\zeta_1) & Cov(\zeta_1, \zeta_2) \\ Cov(\zeta_2, \zeta_1) & Var(\zeta_2) \end{pmatrix} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{4.26}$$

Auch hier gilt Symmetrie und somit $\psi_{12} = \psi_{21}$.

Modellimplizierte Kovarianzmatrix

In der Einführung des aktuellen Abschnitts wurde bereits deutlich, dass es das Ziel ist die Abweichung zwischen der **modellimplizierten Kovarianzmatrix $\boldsymbol{\Sigma(\theta)}$** und der *empirischen Kovarianzmatrix* $\boldsymbol{\Sigma}$ zu minimiert. Zur Formulierung der modellimplizierten Kovarianzmatrix beginnen wir mit vier Komponenten der Matrix:

$$\boldsymbol{\Sigma(\theta)} = \boldsymbol{\Sigma} \begin{pmatrix} \boldsymbol{\Sigma_{YY}}(\boldsymbol{\theta}) & \boldsymbol{\Sigma_{YX}}(\boldsymbol{\theta}) \\ \boldsymbol{\Sigma_{XY}}(\boldsymbol{\theta}) & \boldsymbol{\Sigma_{XX}}(\boldsymbol{\theta}) \end{pmatrix} \tag{4.27}$$

1. Exogener Teil

$$\begin{aligned} \boldsymbol{\Sigma_{XX}}(\boldsymbol{\theta}) &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = \mathbb{E}\left[(\boldsymbol{\Lambda_X}\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\Lambda_X}\boldsymbol{\xi} + \boldsymbol{\delta})^T\right] \\ &= \mathbb{E}\left[\boldsymbol{\Lambda_X}\boldsymbol{\xi}(\boldsymbol{\Lambda_X}\boldsymbol{\xi})^T + \boldsymbol{\Lambda_X}\boldsymbol{\xi}\boldsymbol{\delta}^T + \boldsymbol{\delta}(\boldsymbol{\Lambda_X}\boldsymbol{\xi})^T + \boldsymbol{\delta}\boldsymbol{\delta}^T\right] \\ &= \mathbb{E}\left[\boldsymbol{\Lambda_X}\boldsymbol{\xi}\boldsymbol{\xi}^T\boldsymbol{\Lambda_X}^T + \boldsymbol{\Lambda_X}\boldsymbol{\xi}\boldsymbol{\delta}^T + \boldsymbol{\delta}\boldsymbol{\xi}^T\boldsymbol{\Lambda_X}^T + \boldsymbol{\delta}\boldsymbol{\delta}^T\right] \\ &= \boldsymbol{\Lambda_X}\mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right]\boldsymbol{\Lambda_X}^T + 0 + 0 + \mathbb{E}\left[\boldsymbol{\delta}\boldsymbol{\delta}^T\right] \\ &= \boldsymbol{\Lambda_X}\boldsymbol{\Phi}\boldsymbol{\Lambda_X}^T + \boldsymbol{\Theta_\delta} \end{aligned} \tag{4.28}$$

Die letzten Schritte folgen aus der Annahme $\mathbb{E}[\boldsymbol{\xi}] = 0$ und den Kovarianzmatrizen für die Fehler.

2. Endogener Teil

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{YY}}(\boldsymbol{\theta}) &= \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^T\right] = \mathbb{E}\left[(\boldsymbol{\Lambda}_{\mathbf{Y}}\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\Lambda}_{\mathbf{Y}}\boldsymbol{\eta} + \boldsymbol{\epsilon})^T\right] \\ &= \cdots = \boldsymbol{\Lambda}_{\mathbf{Y}}\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^T\right]\boldsymbol{\Lambda}_{\mathbf{Y}}^T + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}}\end{aligned} \tag{4.29}$$

Der Erwartungswert $\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^T\right]$ kann nicht zusammengefasst werden, da $\boldsymbol{\eta}$ auch als Funktion der exogenen Faktoren ausgedrückt werden kann. $\boldsymbol{\eta}$ muss zunächst mit Hilfe der Strukturmodell-Gleichung 4.19 umgeformt werden. Wir beginnen mit:

$$\boldsymbol{\eta} - \boldsymbol{B}\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.30a}$$

$$(\mathbf{I} - \mathbf{B})\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.30b}$$

$$\boldsymbol{\eta} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.30c}$$

Nun kann der Erwartungswert $\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^T\right]$ mit Gleichung 4.30c bestimmt werden mit:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^T\right] &= \mathbb{E}\left[\left((\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right)\left((\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right)^T\right] \\ &= \mathbb{E}\left[\left((\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})\right)\left(((\boldsymbol{\Gamma}\boldsymbol{\xi})^T + \boldsymbol{\zeta}^T)(\mathbf{I} - \mathbf{B})^{-1^T}\right)\right] \\ &= \mathbb{E}\left[(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})(\boldsymbol{\xi}^T\boldsymbol{\Gamma}^T + \boldsymbol{\zeta}^T)(\mathbf{I} - \mathbf{B})^{-1^T}\right] \\ &= (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbb{E}\left[\boldsymbol{\xi}\boldsymbol{\xi}^T\right]\boldsymbol{\Gamma}^T + \mathbb{E}\left[\boldsymbol{\zeta}\boldsymbol{\zeta}^T\right])(\mathbf{I} - \mathbf{B})^{-1^T} \\ &= (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B})^{-1^T}\end{aligned} \tag{4.31}$$

Einsetzen in Gleichung 4.29 liefert:

$$\boldsymbol{\Sigma}_{\mathbf{YY}}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_{\mathbf{Y}}(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B})^{-1^T}\boldsymbol{\Lambda}_{\mathbf{Y}}^T + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \tag{4.32}$$

3. Kovarianzen zwischen endogenem und exogenem Anteil

Die Kovarianzen können mit Hilfe analoger Berechnung erhalten werden. Für $\boldsymbol{\Sigma}_{\mathbf{YX}}(\boldsymbol{\theta})$ gilt:

$$\boldsymbol{\Sigma}_{\mathbf{YX}}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_{\mathbf{Y}}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\mathbf{X}}^T \tag{4.33}$$

Der letzte Anteil ergibt sich durch transponieren $\boldsymbol{\Sigma}_{\mathbf{XY}}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{\mathbf{YX}}(\boldsymbol{\theta})^T$, damit folgt:

$$\boldsymbol{\Sigma}_{\mathbf{XY}}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_{\mathbf{X}}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T(\mathbf{I} - \mathbf{B})^{-1^T}\boldsymbol{\Lambda}_{\mathbf{Y}}^T \tag{4.34}$$

Resultat

Die modellimplizierten Kovarianzmatrix berechnet sich abschließend mit Gleichungen 4.28, 4.32, 4.33 und 4.34 wie folgt:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} \begin{pmatrix} \boldsymbol{\Lambda_Y}(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1^T}\boldsymbol{\Lambda}_\mathbf{Y}^T + \boldsymbol{\Theta_\epsilon} & \boldsymbol{\Lambda_Y}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma\Phi\Lambda}_\mathbf{X}^T \\ \boldsymbol{\Lambda_X\Phi\Gamma}^T(\mathbf{I}-\mathbf{B})^{-1^T}\boldsymbol{\Lambda}_\mathbf{Y}^T & \boldsymbol{\Lambda_X\Phi\Lambda}_\mathbf{X}^T + \boldsymbol{\Theta_\delta} \end{pmatrix} \tag{4.35}$$

4.1.3 Exkurs: RAM-Notation

Neben der *LISREL*-Notation gibt es weitere Methoden zur Beschreibung von Strukturgleichungsmodellen. Ähnlich wie bei LISREL sind diese häufig an Software gebunden. Übliche Notationen sind (in Klammern die zugehörige Software):

- LISREL (LISREL)
- Bentler-Weeks (EQS)
- RAM (RAMONA)

Im Folgenden wird eine Alternative zur LISREL-Notation, die **RAM-Notation** betrachtet. Unterschieden wird dort nur zwischen drei Matrizen **S**, **A** und **F** und einem Vektor **m**. Die Matrizen und der transponierte Vektor enthalten in den ersten Spalten die latenten Variablen, die letzten Spalten enthalten die beobachteten Variablen. Das Beispiel aus Abbildung 8 würde in RAM-Notation wie folgt definiert werden:

$$\mathbf{S} = \begin{bmatrix} \phi & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \psi & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{\delta_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\delta_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\delta_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\epsilon_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{\epsilon_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{\epsilon_3} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_6 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.36}$$

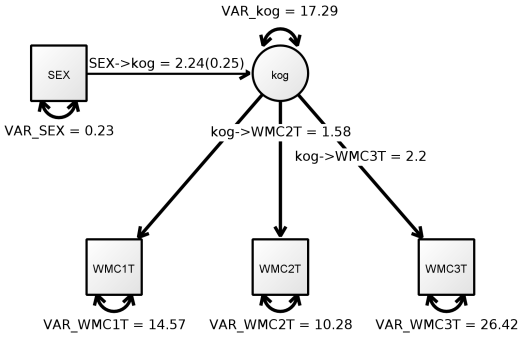
Dabei bezeichnen ϕ und ψ die (Ko-)Varianzen der latenten Variablen und σ^2 die der beobachteten (manifesten) Variablen. Diese Kovarianzmatrix **S** ist *symmetrisch*.

Die λ entstammen der Faktorladungsmatrix $\boldsymbol{\Lambda}$. γ ist der Zusammenhang zwischen den Faktoren ξ und η. Diese Zusammehangsmatrix **A** ist *asymmetrisch*.

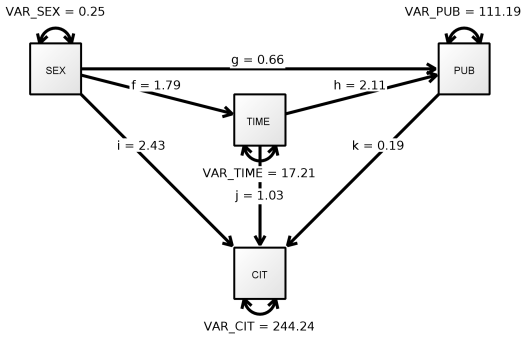
Die Matrizen **F** bezeichnet eine Filter-Matrix, der Vektor **m** gibt die Mittelwerte an. Beide werden hier nicht weiter betrachtet.

4.1.4 Software

- **Ωnyx**: Sehr benutzerfreundlich für einfache Modelle
- **OpenMX**: Sehr komplex, Programmierkenntnisse nötig
- **MPlus**: Wenige Befehle reichen aus um komplexe Modelle zu berechnen, trifft viele Default-Entscheidungen
- **LISREL**: Klassiker, von dem auch die Notation abgelietet ist
- Weitere: **R**, **AMOS**, **SAS**, **Mx**, **sem()**, **EQS**, **lavaan**



(a) Effekt von Geschlecht auf kognitive Leistungsfähigkeit



(b) Effekte auf Anzahl der Zitationen von Professoren

Abbildung 9: Strukturgleichungsmodelle mit Onyx

Beispiele: Onyx

Beispiel 1: Einfluss des Geschlechts auf kognitive Leistungsfähigkeit. Dabei ist die kognitive Leistungsfähigkeit ein Faktor bzw. eine latente Variable, welcher durch drei beobachtete Variablen beschrieben wird. Es wird ein *Messmodell* (Faktor) und ein *Strukturodell* verwendet (Einfluss Geschlecht auf kognitive Leistung). Das Beispiel ist in Abbildung 9a dargestellt. Eine Erhöhung im Geschlecht um 1, führt zu einer erhöhten Leistung um $2,24$ (bzw. standardisiert $0,25$).

Beispiel 2: Auch das *Pfadmodell* mit den Professorengehältern lässt sich mit Onyx modellieren. Das Beispiel ist in Abbildung 9b dargestellt. Eine Erhöhung um 1 des Geschlechts erhöht bspw. die Anzahl der Zitationen um $2,43$ und die Anzahl der Publikationen um $0,66$.

Begriffe

Multiple vs. Multivariate Modelle, Modellimplizierte Kovarianzmatrix, Empirische Kovarianzmatrix, LISREL, Messmodell, Strukturmodell, Notation, Variablen der SEM, Annahmen, Exogenes/Endogenes Messmodell, Alternative Notationen, RAM, Software, Interpretation

4.2 Angewandte Strukturgleichungsmodelle

4.2.1 Modellspezifikation

Wir spezifizieren Strukuturgleichungsmodelle mit der eingeführten LISREL-Notation. Die Notationsregel ist analog zu Pfaddiagrammen.

4.2.2 Modellidentifikation

Für ein spezifiziertes Modell kann die *empirische Kovarianzmatrix* $\boldsymbol{\Sigma}$ angegeben werden, der Parameter $\boldsymbol{\theta}$ kann anschließend durch Minimierung bzgl. der *modellimplizierte Kovarianzmatrix* $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ gewonnen werden.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{arg\,min}\,[\boldsymbol{\Sigma},\, \boldsymbol{\Sigma}(\boldsymbol{\theta})] \tag{4.37}$$

Die *Anzahl der unbekannten Parameter* $\boldsymbol{\theta}$ in der *modellimplizierten Kovarianzmatrix* $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ und die *Anzahl der bekannten Varianzen und Kovarianzen* in der *empirische Kovarianzmatrix* $\boldsymbol{\Sigma}$ können sich unterscheiden. Dies

führt bei der Minimierung zu Resultaten, die wie folgt unterschieden werden können:

Typ	Freiheitsg.	Beschreibung
just identified	$df = 0$	Es gibt genau eine Möglichkeit alle Parameter $\boldsymbol{\theta}$ als eineindeutige Funktion von $\boldsymbol{\Sigma}$ zu beschreiben.
over identified	$df > 0$	Es gibt mehrere Möglichkeiten alle Parameter $\boldsymbol{\theta}$ als eineindeutige Funktion von $\boldsymbol{\Sigma}$ zu beschreiben.
under identified	$df < 0$	Es gibt keine Möglichkeiten alle Parameter $\boldsymbol{\theta}$ als eineindeutige Funktion von $\boldsymbol{\Sigma}$ zu beschreiben (in diesem Fall existiert keine Lösung).

Bestimmte Annahmen bzw. Beschränkungen (**constraints**) sind nötig, um eine Berechnung der Modelle zu gewährleisten. Zwei wichtige Constraints sind:

1. Die Diagonale von $\mathbf{B}$ muss Null sein. Variablen dürfen keine Effekte auf sich selbst haben.
2. Das Regressionsgewicht der latenten Variablen $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$, $\boldsymbol{\delta}$ und $\boldsymbol{\epsilon}$ wird auf 1 fixiert. Latente Variablen können nicht direkt beobachtet werden, ihnen muss eine Skala zugewiesen werden.

Skalierung

Aus der zweiten Annahme folgt eine notwendige Skalierung. Bisher wurde die Skala einer anderen Variablen übernommen, indem der Effekt auf 1 fixiert wurde (siehe Abbildung 10a). Es gilt:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} \tag{4.38}$$

Dabei wird die mittlere Matrix (Einheitsmatrix) als **Skalierungsmatrix** bezeichnet.

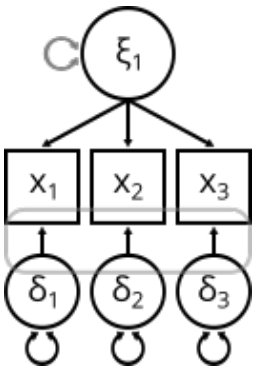
Um auch die *latenten Variablen* zu skalieren, gibt es zwei Möglichkeiten

- Fixierung des Effekts der ersten manifesten Variable auf 1 (prinzipiell kann jede Variable gewählt werden), siehe Abbildung 10b
- Fixierug der Varianz einer latenten Variablen ϕ, siehe Abbildung 10c

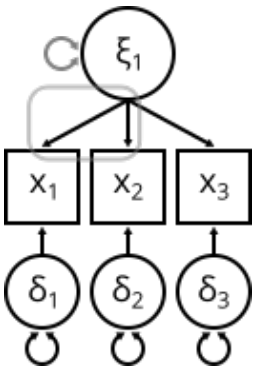
Identifikationsregeln

Das in Abbildung 10 beschriebene Modell ist *just identified*. Die Identifikation eines Modells wird gezeigt, indem *algebraisch* nachgewiesen wird, dass es genau eine Möglichkeit gibt alle Parameter $\boldsymbol{\theta}$ als eineindeutige Funktion von von $\boldsymbol{\Sigma}$ zu beschreiben (vgl. Lösungstyp *just identified*). Dieser Beweis ist bei größeren Modell jedoch sehr aufwändig.

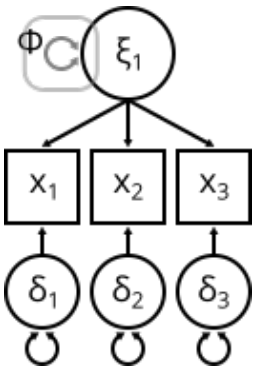
Um einen Beweis umgehen zu können, können Identifikationsregeln formuliert werden.



(a) Skalierung der manifesten Variablen durch Fixierung der Effekt auf x auf 1.



(b) Skalierung der latenten Variablen durch Fixierung des Effektes einer manifesten Variablen auf 1.



(c) Skalierung der latenten Variablen durch Fixierung der Varianz ϕ. Alternative für 10b.

Abbildung 10: Skalierung von Variablen

Regel	**Anwendung**	**Bedingung**	**Notw.**	**Hinr.**
t-Rule	Alle	$t \leq \frac{1}{2}(p+q)(p+q+1)$	ja	nein
Null **B** rule	Beob. Var.	$\mathbf{B} = 0$	nein	ja
Recursive Rule	Beob. Var.	**B** Dreiecksmatrix, $\mathbf{\Psi}$ Diagonalmatrix	nein	ja
Three-Indicator-Rule	Messmodelle	$n \geq 1$, $\mathbf{\Theta}$ Diagonalmatrix Mind. ein Elem. pro Zeile von $\mathbf{\Lambda}$ ist nicht Null Drei oder mehr Indikatoren pro Faktor	nein	ja
Two-Indicator-Rule Regel 1	Messmodelle	$n > 1$, $\mathbf{\Theta}$ Diagonalmatrix $\phi_{ij} \neq 0 \; \forall \; i, j$ Mind. ein Elem. pro Zeile von $\mathbf{\Lambda}$ ist nicht Null Zwei oder mehr Indikatoren pro Faktor	nein	ja
Two-Indicator-Rule Regel 2	Messmodelle	$n > 1$, $\mathbf{\Theta}$ Diagonalmatrix $\phi_{ij} \neq 0$ für mind. ein Paar mit $i \neq j$ (In jeder Zeile von ϕ_{ij} mind. ein Elem. nicht Null) Mind. ein Elem. pro Zeile von $\mathbf{\Lambda}$ ist nicht Null Zwei oder mehr Indikatoren pro Faktor	nein	ja
Two-Step-Rule	Allg. SEM	1. Ursprüngliches Modell in Messmodell umformulieren, eliminiere **B**, $\mathbf{\Lambda}$ und $\mathbf{\Psi}$, Identifiziere Messmodell 2. Identifiziere Strukturmodell (Annahme: Keine Messfehler der zu Grunde liegenden Beobachtungen)	nein	ja

Bei der t-Rule entspricht q der Anzahl der Indikatoren exogener latenter Variablen und p der Anzahl der Indikatoren endogener latenter Variablen. t ist die Anzahl der zu schätzenden Parameter.

Notwendige Bedingung bedeutet dabei: Tritt die Bedingung ein, so ist das Modell auf jeden Fall identifiziert. Aus einem identifizierten Modell folgt aber nicht, dass die Bedingung erfüllt ist.
Eine hinreichende Bedingung bedeutet: Ist das Modell identifiziert, so ist die Bedingung auf jeden Fall erfüllt. Aus einer erfüllten Bedingung folgt aber noch nicht, dass das Modell identifiziert ist.

Da es keine Bedingung gibt, die notwendig und hinreichend gleichzeitig ist, kann eine eindeutige Aussage nicht

immer getroffen werden. Vor allem dann, wenn keine der Bedingungen zutrifft, muss das Modell daher nicht abgelehnt werden, es könnte immer noch lösbar sein. Folgende Regeln können helfen:

- Identifikationsregeln treffen zu: Modell ist identifiziert
- Keine Identifikationsregel trifft zu: Algebraische Lösung versuchen
- Identifikationsregeln gibt Warnung: Weitere Indizien prüfen (z.B. positiv definite Hesse-Matrix)
- Testen unterschiedlicher Startwerte für die numerische Berechnung. Sind die Ergebnisse immer wieder ähnlich, dann ist das Modell wahrscheinlich identifiziert.
- Reproduzierte Kovarianzmatrix speichern und als Input für neue Schätzung verwenden. Bei einem identifizierten Modell sind die Parameter identisch.

4.2.3 Modellschätzung

Nach der Identifikation erfolgt die **Schätzung** des Modells. Dabei wird die Distanz zwischen den Matrizen $\boldsymbol{\Sigma}$ und $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ minimiert. Zunächst muss die **Distanz** dafür definiert werden. Es werden vier Schätzverfahren betrachtet:

1. Unweighted Least Squares (ULS)

Bei der Methode der **Unweighted Least Squares** werden kleinsten Quadrate ungwichtet gebildet, analog zur Methode der kleinsten Quadrate (ordinary least squares) bei der linearen Regression (vgl. Abschnitt 2.1). Er kann für normalverteilte Variablen verwendet werden.

$$F_{ULS} = \frac{1}{2} \operatorname{tr}\left[(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\theta}))^2\right] \tag{4.39}$$

Die Differenz der beiden Matrizen ergibt die Residualmatrix, welche minimiert wird. Wegen der Symmetrie der Matrix, erhalten die Kovarianzen (sie kommen immer doppelt in der Matrix vor) ein größeres Gewicht.

Vorteile

- Einfach zu berechnen
- Algebraische Lösung leichter zu ermitteln
- Konsistenter Schätzer von $\boldsymbol{\theta}$

Nachteile

- Effizientere Schätzer sind möglich
- Nicht skaleninvariant und nicht skalenfrei
- Verteilungsannahmen nötig (für Standardfehler und χ^2-Teststatistiken)

2. Maximum Likelihood (ML)

Der häufigst verwendete Schätzer ist der **Maximum Likelihood Schätzer**. Dabei wird θ so gewählt, dass die Wahrscheinlichkeit für das Auftreten der empirischen Kovarianzmatrix $\boldsymbol{\Sigma}$ maximal wahrscheinlich ist.

$$F_{ML} = \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - \ln|\boldsymbol{\Sigma}| - (p+q) \tag{4.40}$$

Dabei geben p und q die Anzahl der manifesten Variablen an. Es gilt $F_{ML} \to 0$ je plausibler der Parameter die empirische Kovarianzmatrix erklären kann.

Die Maximum Likelihood Methode wird unter der *Annahme* normalverteilter Zufallsvariablen angewendet.

Lösungen

Die Parameter werden über Ableiten und Nullsetzen gewonnen (und durch Prüfung der hinreichenden Bedingung, d.h. zweite Ableitung ist positiv definit). Da nur für einfache Modelle analytischen Lösungen existieren, werden meist numerische Optimierungsverfahren eingesetzt. Dabei gilt:

1. Startwerte: Es müssen günstige Startwerte gewählt werden. Dies kann erfolgen durch:

- A priori Wissen (Theorie)
- A priori nicht-iterative Verfahren (ULS)
- OLS Regression für jede einzelne Gleichung
- Einfaches probieren mehrerer Werte (Versuch und Irrtum)
- Daumenregeln und Empfehlungen

2. Optimierungsverfahren: Es muss ein geeignetes Optimierungsverfahren gefunden werden. In einfachen Fällen liefert das Verfahren eine monotone Funktion (d.h. die Werte sind von Schritt zu Schritt monoton fallend). Dies ist aber häufig nicht der Fall. Es kommen z.B. folgende Verfahren in Frage:

- Newton-Raphson
- Fletcher-Powell
- Gauss-Newton

3. Konvergenzkriterium: Es wird ein Kriterium zum Abbruch des Verfahrens benötigt, da das Verfahren in den meisten Fällen unendlich weiter rechnet und immer genauere Ergebnisse liefert. Ab einem gewissen Punkt ist jedoch kein Mehrwert mehr gegeben, das Ergebnis ist genügend genau. D.h. ist von einem Iterationsschritt zum nächsten der Zugewinn die Anpassung (engl. Fit) der Funktion kleiner als Δ, wird das Verfahren abgebrochen. Verschiedene Programme zur Berechnung von SEM-Modellen verweden unterschiedliche Kriterien. Diese sollten beachtet und zu Vergleichszwecken u.U. selbst gewählt werden.

Prinzipiell wäre es auch möglich das Verfahren in seiner Rechenzeit oder in der Anzahl der Iterationsschritte zu beschränken.

Vorteile

- Erwartungsreuer Schätzer (der Parameter und der Standardfehler)
- Asymptotisch konsistent
- Asymptotisch effizient
- Parameterschätzer folgen für große N einer Normalverteilung
  (Verhältnis Parameter/Standardfehler folgt Standardnormalverteilung)
- Skaleninvariant und skalenfrei

Nachteile

- Bei Abweichungen von Normalverteilung ist Schätzer u.U. nicht effizient
- Bei Abweichungen von Normalverteilung führen χ^2-Werte u.U. zu erhöhtem α-Fehler

(Robuste Verfahren, wie z.B. MLR oder Satorra-Bentler scaled χ^2, sind dann bessere Verfahren)
- Bei Nicht-Normalverteilung sollten andere Schätzer in Betracht gezogen werden (z.B. Weighted Least Squares (WLS))

3. (Generally) Weighted Least Squares (WLS)

Der **Weighted Least Squares (WLS)** Schätzer ist eine Erweiterung des ULS-Schätzers. Er findet Anwendung, wenn (1) keine Normalverteilung vorliegt oder (2) die manifesten Variablen kategoriales Skalenniveau besitzen.

$$F_{WLS} = [\boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta})]^T \ \mathbf{W}^{-1} \ [\boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta})] \tag{4.41}$$

Dabei ist $\mathbf{W}$ die quadratische k-dimensionale **Gewichtsmatrix**, wobei $k = \frac{1}{2}(p+q)(p+q+1)$ (Anzahl der nicht-redundanten Elemente in $\boldsymbol{\Sigma}$). Der Vektor $\boldsymbol{\sigma}$ enthält alle nicht-redundanten Elemente der empirischen Kovarianzmatrix $\boldsymbol{\Sigma}$, ebenso enthält der Vektor $\boldsymbol{\sigma}(\boldsymbol{\theta})$ alle nicht-redundanten Elemente der modellimplizierten Kovarianzmatrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Zusammengefasst wird der Abstand $\boldsymbol{\sigma} - \boldsymbol{\sigma}(\boldsymbol{\theta})$ zwischen den nicht-redundanten Elementen in $\boldsymbol{\Sigma}$ mit $\mathbf{W}$ gewichtet.

Für $\mathbf{W} = \mathbf{I}$ entspricht der WLS-Schätzer dem ULS-Schätzer. Die Elemente von $\mathbf{W}$ (angegeben mit w_{ijgh}) werden mit den empirischen Daten, d.h. mit σ_{ij} und σ_{gh} bestimmt. w_{ijgh} soll dabei proportional zu einem Konsistenzen Schätzer von σ_{ij} und σ_{gh} sein.

$$w_{ijgh} = (N-1)\ Cov(\sigma_{ij}, \sigma_{gh}) \tag{4.42}$$

Vorteile

- Wenige Voraussetzungen
- Auch für kategoriale/ordinale Daten verwendbar
- Für große Stichproben sind χ^2-Werte und Standardfehler korrekt

Nachteile

- Gewichtsmatrix $\mathbf{W}$ wird sehr groß
  (für 10 Variablen wird die Dimensionen $k = 55$ mit 1540 nicht-redundanten Elementen)
- Große Stichproben erforderlich um gute Parameterschätzungen zu erhalten
- Für komplexe Modelle, kleine Stichprobenanzahl und/oder Normalverteilung eher ungeeignet
  (Wahl einfacherer Modelle angebracht)

4. Generalized Least Squares (GLS)

Eine Verallgemeinerung der ULS und WLS-Schätzer bildet der **Generalized Least Squares (GLS)** Schätzer. Hierbei wird die Matrix $\mathbf{V}$ verwendet, um das Modell zu formulieren.

$$F_{GLS} = \frac{1}{2}\text{tr}\left[\left([\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\ \mathbf{V}^{-1}\right)^2\right] \tag{4.43}$$

Für den Spezialfall $\mathbf{V} = \mathbf{W}$ ergibt sich der WLS-Schätzer, für $\mathbf{V} = \mathbf{I}$ ergibt sich der ULS-Schätzer.

4.2.4 Modellevaluation

Nach der Schätzung der Parameter sollte eine Prüfung des Modells erfolgen. Zu Bestimmung der **Güte** des Modells gibt es bei Strukturgleichungsmodellen viele verschiedene **Gütemaße**. Keines der Gütemaße kann in jedem Fall die Güte des Modells gut bestimmen. Je nach Wahl des Modells sind unterschiedliche Kriterien zu wählen. Ein allgemein anerkanntes allumfängliches Gütemaß existiert derzeit nicht. Es können zwei Klassen von Gütemaßen unterschieden werden, das sind zum einen **deskriptive Gütemaße**, zum anderen **inferenzstatistische Gütemaße**.

Inferenzstatistische Gütemaße

Der $\boldsymbol{\chi^2}$**-Test** (auch Likelihood-Ratio-Test) ist der einzige Test, der sich eindeutig in die Klasse der inferenzstatistischen Gütemaße einordnen lässt. Er testet auf eine signifikante Abweichung der empirischen und der modellimplizierten Kovarianzmatrix, $\boldsymbol{\Sigma}$ und $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Das Modell kann akzeptiert werden, wenn die Abweichung *nicht-signifikant* ist.

Der χ^2-Wert leitet sich aus der Anpassungsfunktion des Maximum-Likelihood-Ansatzes her.

$$\chi^2(df) = (N-1)\, F(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})) \tag{4.44}$$

Die Freiheitsgrade werden definiert über:

$$df = \frac{1}{2}(p+q)(p+q+1) - t \tag{4.45}$$

Dabei ist t die Anzahl der zu schätzenden Parameter und p und q die Anzahl der manifesten Variablen.

Mit dem χ^2-Test lassen sich neben der Prüfung einer signifikanten Abweichung der Kovarianzmatrizen auch s.g. **genestete Modelle** evaluieren. Genestete Modelle sind solche, in denen ein Modell in einem anderen enthalten ist.

Nachteile

- Annahme einer Normalverteilung und hinreichend großer Stichprobengröße
- Abhängigkeit von der Anzahl der Parameter: Abnahme des χ^2-Wertes mit steigender Parameterzahl (je komplexer das Modell)
- Abhängigkeit von der Stichprobengröße: Je höher die Stichprobe, desto eher wird der Test signifikant (die Teststärke β nimmt zu), d.h. die Modelle werden eher verworfen

Deskriptive Gütemaße

Deskriptive Gütemaße haben sich maßgeblich wegen der Einschränkungen des χ^2-Tests etabliert. Da die Verteilungseigenschaften der meisten alternativen Gütemaße unbekannt ist, besitzen sie ausschließlich deskriptiven Charakter. Es werden im Folgenden drei unterschiedliche Klassen deskriptiver Gütemaße betrachtet.

1. Measures of overall fit

Measures of overall fit geben an, wie gut das Modell global mit den empirischen Daten zusammenpasst. Hierbei können verschiedene Maße verwendet werden.

(a) Häufig verwendet wird die **Root Mean Square Error of Approximation (RMSEA)**. Zu Grunde liegt die Annahme, dass die Nullhypothese H_0 (Modelle sind gleich) eines exakten Modells mit hinreichend großer Stichprobe immer abgelehnt werden kann. Die RMSEA wählt einen Ansatz bei dem die **Nonzentralität** bestimmt wird. Die Nonzentralität gibt an wie weit der wahre Wert vom geschätzen Wert abweicht (unter Annahme der Nullhypothese). Je größer der Nonenzentralitätsparameter $\lambda = \chi^2 - df$, desto mehr verändert sich die zu Grunde liegende Verteilung. Diese Änderung sollte nicht zu groß ausfallen. Für die RMSEA gilt:

$$\epsilon = \sqrt{\frac{\lambda}{df(N-1)}} \tag{4.46}$$

Für den RMSEA-Wert kann ein Konfidenzintervall bestimmt werden. Für die Interpretation gilt:

- $\epsilon = 0$ Perfektes Modell (exact fit)
- $\epsilon < 0,80$ Gutes Modell
- $0,80 < \epsilon < 1,00$ Mittelmäßiges Modell (mediocre fit)
- $\epsilon > 1,00$ Schlechtes Modell, Ablehnung des Modells

Eigenschaften der RMSEA sind eine (1) Belohnung *sparsamer Modelle* und eine (2) relativ gute *Unabhängigkeit von der Stichprobengröße*.

(b) Alternativ zur RMSEA kann das **Root Mean Residual (RMR)** verwendet werden. Da das RMR jedoch skalenabhängig ist, können keine vergleichenden Aussagen (z.B. zu anderen Modellen oder Erfahrungswerten) abgeleitet werden.

$$RMR = \sqrt{2\,\frac{\sum_{i=1}^{p+q}\sum_{j=1}^{i}(s_{ij} - \hat{\sigma}_{ij})^2}{(p+q)(p+q+1)}} \tag{4.47}$$

(c) Um eine Vergleichbarkeit zu gewährleisten wird beim **Standardized Root Mean Residual (SRMR)** eine Normierung an der Standardabweichung vorgenommen.

$$SRMR = \sqrt{2\,\frac{\sum_{i=1}^{p+q}\sum_{j=1}^{i}(\frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}})^2}{(p+q)(p+q+1)}} \tag{4.48}$$

Für das SRMR gilt $SRMR < 0,05$ als gute Anpassung (*good fit*) und $SRMR < 0,10$ als akzeptable Anpassung (*acceptable fit*).

2. Measures based on model comparisons

Bei **Measures based on model comparisons** wird zwischen dem betrachteten Modell und einem *Referenzmodell* (*Baseline Model*) verglichen (Analoger Ansatz zur logistischen Regression vgl. Abschnitt 3.5). Als Referenzmodell wird üblicherweise das **Independence Model** oder das **Null Model** gewählt. Im Independence Model werden alle Variablen als unkorreliert angenommen. Im Null Model werden ebenfalls alle Variablen

als unkorreliert angenommen, zusätzlich sind alle Parameter null (insbesondere die Varianzen).

(a) Der **Normed Fit Index (NFI)** (vgl. Abschnitt 3.6.1; auch als Bentler Bonett Index bezeichnet) gibt Werte im Bereich (Range) 0 bis 1 an.

$$NFI = \frac{\chi_i^2 - \chi_t^2}{\chi_i^2} = 1 - \frac{\chi_t^2}{\chi_i^2} = 1 - \frac{F_t}{F_i} \tag{4.49}$$

Dabei bezeichnet i das Independence Model und t das betrachtete Modell (*target model*). Ein Problem des NFI ist die Abhängigkeit von der Stichprobengröße N (siehe NNFI/TLI). Für die Güte können folgende Kriterien (**Cut offs**) verwendet werden:

- $> 0,95$ Gute Anpassung (good fit)
- $> 0,9$ Akzeptable Anpassung (acceptable fit)
- $< 0,9$ Schlechte Anpassung (poor fit)

(b) Der **Non-Normed Fit Index (NNFI)**, der auch als TLI bezeichnet wird, wird analog zum NFI gebildet. Das Gütemaß wird jedoch so gewählt, dass es weniger abhängig vom Stichprobenumfang N ist. Der Bereich der Werte (Range) reicht für den NNFI auch über 1 hinaus (keine *natürliche Range*).

$$NNFI = \frac{\frac{\chi_i^2}{df_i} - \frac{\chi_t^2}{df_t}}{\frac{\chi_i^2}{df_i} - 1} = \frac{\frac{F_i^2}{df_i} - \frac{F_t^2}{df_t}}{\frac{F_i^2}{df_i} - \frac{1}{N-1}} \tag{4.50}$$

Die Empfehlungen für *Cut offs* liegen bei:

- $> 0,97$ Gute Anpassung (good fit)
- $> 0,95$ Akzeptable Anpassung (acceptable fit)
- $< 0,95$ Schlechte Anpassung (poor fit)

(c) Der **Comparative Fit Index (CFI)** arbeitet wieder mit dem Nonzentralitätsparameter $\lambda = \chi^2 - df$ (vgl. RMSEA). Der CFI ist (1) weniger abhängig vom Stichprobenumfang N und (2) bevorzugt sparsame Modelle.

$$CFI = \frac{\lambda_i - \lambda_t}{\lambda_i} \tag{4.51}$$

Die Empfehlungen für *Cut offs* liegen bei:

- $> 0,97$ Gute Anpassung (good fit)
- $> 0,95$ Akzeptable Anpassung (acceptable fit)
- $< 0,95$ Schlechte Anpassung (poor fit)

3. Measures of model parsimony

Bei **Measures of model parsimony** wird schwerpunktmäßig das Kriterium der *Sparsamkeit* betrachtet. Einfache Modelle sollen besser bzw. komplexe Modelle schlechter bewertet werden. Je mehr Freiheitsgrade das Modell in Relation zum Independence Model besitzt, desto einfacher ist es. Mit anderen Worten: Je weniger

Parameter im Vergleich zum Independence Model verwendet werden, desto einfacher ist das Modell.

(a) Der **Parsimony Normed Fit Index (PNFI)** ist eine Modifikation des NFI. Es wird ein zusätzlicher Strafterm für komplexe Modelle eingeführt.

$$PNFI = \frac{df_t}{df_i} NFI \tag{4.52}$$

(b) Das **Akaike Information Criterion (AIC)** korrigiert den χ^2-Wert bezüglich der zu schätzenden Parameter. Je mehr Parameter geschätzt werden und desto komplexer das Model damit ist, desto geringer fällt der χ^2-Wert aus. Mit dem AIC können auch nicht-genestete Modelle deskriptiv verglichen werden. Der Wert ist jedoch nicht normiert und kann daher nicht für Vergleiche mit anderen Modellen, die auf anderen Datensätzen basieren, verwendet werden.

$$AIC = \chi^2 - 2\,df \tag{4.53}$$

Anmerkung: Softwarepakete implementieren das Gütemaß teilweise unterschiedlich.

4.2.5 Modellinterpretation

Die Interpretation der gewählten Modelle orientiert sich maßgeblich an der zu Grunde liegenden theoretischen Fragestellung. Mögliche Modelltypen sind:

- Pfadmodelle
- Konfirmatorische Faktorenanalyse
- Strukturmodelle
- Latente Wachstumskurvenmodelle
- Komplexe multilevel SEM
- SEM zur Zeitreihenanalyse
- Zeitstetigen Modellen
- etc.

Begriffe

Modellspezifikation, Modellidentifikation, Typen der Modellidentifikation, Constraints, Skalierung, Skalierungsmatrix, Skalierung latenter Variablen, Identifikationsregeln (Insbesondere: Wenn keine Regel zutrifft?), Modellschätzung, Distanzdefinitionen, Unweighted Least Square (ULS) (Vorteile/Nachteile), Maximum-Likelihood Schätzer (Annahmen, Numerische Optimierung, Startwert, Verfahren, Konvergenzkriterium, Vorteile/Nachteile), Weighted Least Squares (WLS) (Anwendung, Vorteile/Nachteile), Generalized Least Squares (GLS), Modellevaluation, Güte, Klassen von Gütemaßen, Inferenzstatistische Gütemaße, χ^2-Test (Nachteile), Genestete Modelle, Deskriptive Gütemaße, Measures of overall fit, Root mean square of approximation (Nonzentralität, Cut Offs, Eigenschaften), Measures based on model comparison, Referenzmodell, Independence Model, Null Model, Normed Fit Index (NFI) (Range, Cut Offs), Non-Norm Fit Index (NNFI, TLI) (Range, Cut Offs), Comparative Fit Index (CFI) (Eigenschaften, Cut Offs), Measures of model parismony, Parismony Normed Fit Index (PNFI), Akaike Information Citerion (AIC), Modellinterpretation, Modelltypen

5 Vertiefungen

5.1 Kausalität

Ob aus einer Regression kausale Informationen entnommen werden können, hängt von folgenden Kriterien ab:

- Definition der Kausalität
- Modellvorstellung des zu Grund liegenden Prozesses
- Experimentelles Design

5.1.1 Kausalität bei Isolation

Ein kausaler Zusammenhang liegt nur dann sicher vor, wenn unter (experimenteller) Veränderung einer unabhängigen Variablen, eine Veränderung der abhängigen Variablen bewirkt. Dabei muss eine **Isolation** erfolgen. D.h. es muss gewährleistet werden, dass nur die eine unabhängige Variable Einfluss nimmt. Liegt dann ein Zusammenhang vor, muss die Einflussrichtung geprüft werden. Kann eine eindeutige Richtung gefunden werden, kann von einem kausalen Zusammenhang ausgegangen werden. Zusammengefasst müssen folgende Bedingungen erfüllt sein:

- Isolation der abhängigen Variable
- Zusammenhang zwischen den Variablen
- Richtung des Zusammenhangs

Definition: Isolation

Isolation liegt vor, wenn es entweder (1) keine weiteren Einflussgrößen gibt oder wenn (2) die weiteren Einflussgrößen keinen Effekt auf den untersuchten Zusammenhang haben z.B. durch *Konstanthalten des Einflusses*. Bei der Konstanthaltung eines Einflusses wird darauf geachtet, dass ein Einfluss über den zu untersuchenden Zeitraum konstant bleibt.

Definition: Kausalität bei Isolation

Es kann eine kompakte Definition formuliert werden:

Liegt Isolation einer Variablen Y_1 vor und gibt es einen Zusammenhang zwischen Y_1 und einer Variablen X_1 und liegt eine Richtung vor, d.h. geht die Veränderung von X_1 der Variablen Y_1 voraus, dann ist der Zusammenhang **kausal**. Dabei ist X_1 die **Ursache** von Y_1. Y_1 ist die **Wirkung**.

5.1.2 Bedeutung des Fehlerterms

Ein deterministischer linearer Zusammenhang kann nicht exakt berechnet werden, da die Anzahl der unabhängigen Variablen praktisch nie vollständig erfasst werden kann (vgl. Abschnitt 1.2.3 und 2.1.1). Alle nicht erfassten Variablen werden in einem *Fehlerterm* ζ zusammengefasst.

$$Y_i = \alpha_i + \gamma_{i1} X_1 + \gamma_{i2} X_2 + \cdots + \gamma_{iN} X_N \tag{5.1a}$$

$$Y_i = \alpha_i + \gamma_{i1} X_1 + \epsilon_i \tag{5.1b}$$

Dabei ist $\epsilon_i = \gamma_{i2}X_2 + \cdots + \gamma_{iN}X_N$. Unter perfekter Isolation wird für den Fehlerterm kein Einfluss erwartet, da alle weiteren Größen außer der betrachteten Größe X_1 ausgeschlossen wurden. Dann gilt:

$$Y_i = \alpha_i + \gamma_{i1}X_1 \tag{5.2}$$

Unter idealer Isolation und einer idealen Messung ist der Fehlerterm nicht vorhanden. Unter zusätzlicher Voraussetzung einer idealen Modellierung (keine Residuen), geben die gemessenen Werte den wahren Zusammenhang direkt wieder. Kommt es zu einer Abweichung, kann das also im Umkehrschluss vor zwei (drei) Ursachen haben:

- Keine perfekte Isolation
- Keine perfekte Messung
- (Keine perfekte Modellierung)

5.1.3 Kausalität bei Pseudoisolation

Während in klassischen Naturwissenschaften eine annähernd ideale Isolation häufig unter Laborbedingungen möglich ist, ist es in Sozial- und Humanwissenschaften häufig nicht möglich eine Isolation der betrachteten Variablen Y zu erreichen (d.h. Y darf ausschließlich vom betrachteten X abhängen). Es können folgende Fälle denkbar sein:

- Isolation nicht möglich: Nicht alle Einflussgrößen sind bekannt/kontrollierbar
- Isolation nicht umsetzbar: Forschungsressourcen nicht vorhanden
- Isolation nicht wünschenswert: Ethische Gründe verhindern einen solchen Ansatz

In einigen Fällen gibt es begründete Argumente, weshalb von kausalen Zusammenhängen ausgegangen werden kann, auch wenn keine perfekte Isolation vorliegt. V.a. dann, wenn das Experiment (1) gut **statistisch kontrolliert** ist (Kontrollierte Drittvariableneffekte) und eine (2) gut **geplante/umgesesetzte Randomisierung** erfolgt.

Definition: Pseudoisolation

Sind Fehler und Prädiktoren unkorreliert, d.h. es muss gelten:

$$Cov(X_i, \epsilon) = 0 \tag{5.3}$$

In diesem Fall heißen die Prädiktoren **pseudoisoliert**.

Ausführlich

Ein beliebiges lineares Modell kann in zwei Teile geteilt werden (vgl. Abschnitt 5.1.2):

$$Y_i = \alpha_i + \gamma_{i1}X_1 + \cdots + \gamma_{iK}X_K + \gamma_{i(K+1)}X_{(K+1)} + \cdots + \gamma_{iL}X_L \tag{5.4a}$$

$$Y_i = \alpha_i + \gamma_{i1}X_1 + \cdots + \gamma_{iK}X_K + \epsilon_i \tag{5.4b}$$

$X_{(K+1)}, \ldots, X_L$ entsprichen den **omitted variables** (nicht-modellierte Prädiktoren), welche den Fehlerterm ϵ_i bilden. Diese müssen für Pseudoisolation unkorreliert mit den modellierten Prädiktoren $X_1, \ldots, X_K$ sein, d.h.

es muss gelten:

$$Cov(X_k, X_l) = 0 \quad \text{mit } k = 1, \ldots K \text{ und } l = (K+1), \ldots, L \tag{5.5}$$

Daraus folgt unmittelbar Gleichung 5.3.

Ein hinreichendes Kriterium für die Unkorreliertheit von den X_k und den X_l ist die Abwesenheit von (1) **omitted confounder** (Moderator-Effekt von X_l) oder (2) **omitted mediator** (Mediator-Effekt von X_l), d.h. es liegt perfekte statistische Kontrolle vor.

Interpretation bei Pseudoisolation

Bei einer Regression von Y_i auf $X_1, \ldots, X_K$ müssen folgende Bedingungen erfüllt sein:

- Die Prädiktoren $X_1, \ldots, X_K$ sind pseudoisoliert von den omitted variables
- Die Beeinflussungsrichtung verläuft von X_k zu Y_i für alle $k = 1, \ldots, K$
- Der berechnete Steigungskoeffizient $\hat{\gamma}_{ik}$ ist signifikant

Sind die Bedingungen erfüllt ist $\hat{\gamma}_{ik}$ der **direkte kausale Effekte** von X_k auf Y_1, gegeben alle anderen Prädiktoren sind konstant. Verändert sich X_k um einen Wert, dann gibt $\hat{\gamma}_{ik}$ die Veränderung in Y_i an, sofern alle anderen Prädiktoren sich nicht verändern. Die Veränderung wird aber nicht um exakt $\hat{\gamma}_{ik}$ erfolgen, da ϵ_i unbekannt ist und einen Einfluss auf jede Änderung hat, auch wenn der Erwartungswert von ϵ_i Null ist (d.h. $\mathbb{E}[\epsilon] = 0$).

5.1.4 Randomisierung

Die **Randomisierung** ist eine *Zuordnung* bzw. ein *Assignment* bei der Versuchspersonen per Zufall in Gruppen eingeteilt werden. Die Korrelationen zwischen den Störvariablen und den betrachteten Variablen sollte nicht signifikant von Null abweichen (keine Korrelation). Damit wäre die Bedingung für Pseudoisolation erfüllt.

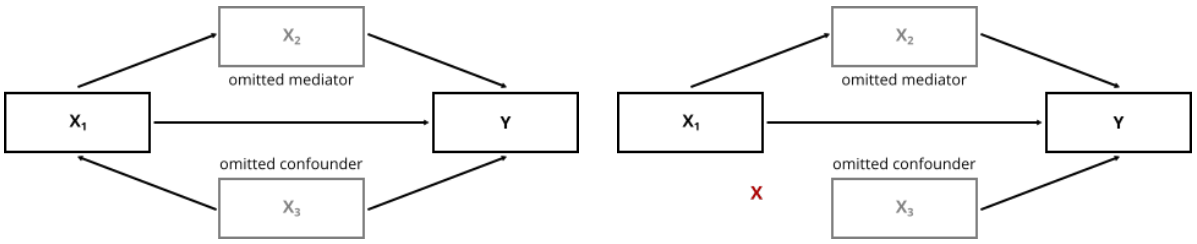
Wird beispielsweise der Effekt einer Intervention (betrachtete Variable X_1) überprüft, so sollte die Intervention erst *nach* einer Randomisierung erfolgen. Die anderen Einflussvariablen X_k (z.B. Konfliktmanagement, Geschlecht, etc.) *vor* der Randomisierung sind dann unabhängig von der Intervention. Vor der Randomisierung sind alle Einflussvariablen auf die unterschiedlichen Gruppen gleichermaßen verteilt.

Die Variablen, die nicht isoliert werden, können wie folgt Einfluss nehmen:

- Kofundierung: Die nicht-betrachteten Variablen X_k sind Ursache für die betrachtete Variable X_1
- Mediation: Die nicht-betrachteten Variablen X_k sind Wirkung der betrachteten Variable X_1
- Unabhängigkeit: Die nicht-betrachteten Variablen X_k sind unabhängig von der betrachteten Variable X_1
- Feedback: Die nicht-betrachteten Variablen X_k und die betrachtete Variable X_1 beeinflussen sich gegenseitig (Ursache und Wirkung)

Auf Grund der Randomisierung haben die nicht-betrachteten Variablen (omitted variables) während der Intervention keinen Einfluss auf die betrachtete Variable. Umgekehrt bleibt ein Einfluss jedoch bestehen. Die betrachtete Variable kann weiterhin die nicht-betrachteten Variablen beeinflussen. Mit der Randomisierung wird also *nur* die *Konfundierung* ausgeschlossen, nicht aber die *Mediation*. Der Sachverhalte ist in Abbildung 11 dargestellt.

Falls die Kovarianz zwischen X_1 und X_k nach dem Test Null ist $Cov(X_1, X_k) = 0$, kann X_k ein (1a) Confounder



(a) Beziehung ohne Randomisierung (b) Beziehung mit Randomisierung

Abbildung 11: Beziehung der Intervention X_1 zum Kriterium Y, mit *omitted mediator* X_2 und *omitted confounder* X_3, während der Durchführung der Intervention.

(der jedoch keinen Einfluss hatte) sein oder X_1 und X_k sind (1b) generell voneinander unabhängig. Falls nach dem Test $Cov(X_1, X_k) \neq 0$, ist X_k ein (2a) Mediator oder es liegt eine (2b) Feedbackbeziehung vor.

Wichtig ist also: Randomisierung stellt Pseudoisolation zwischen betrachteter Variable und omitted confounders her, nicht jedoch zwischen betrachteter Variable und omitted mediators!

Interpretation bei Randomisierung

Der Steigungskoeffizient γ_{11} ist der **erwartete totale kausale Effekte** von X_1 auf Y_1, wenn folgende Bedingungen zutreffen:

- Zufällige Zuordnung der Personen in die Interventionsstufen X_1
- Gerichteter Zusammenhang von X_1 nach Y_1
- Der Steigungskoeffizient γ_{11} ist signifikant

Begriffe

Kriterien für kausale Informationen, Isolation, Kausalität bei Isolation, Ursache, Wirkung, Bedeutung des Fehlerterms, Ursache einer Abweichung, Gründe fehlender Isolation, Argumente für Kausalität bei fehlender Isolation, Pseudoisolation, omitted variables, hinreichendes Kriterium für Unkorreliertheit, omitted confounder, omitted mediator, Kausalität bei Pseudoisolation, Direkter kausaler Effekt, Randomisierung, Bedingung für Pseudoisolation, Einflüsse nicht-isolierter Variablen, Ausschluss von Konfundierung/Mediation?