Parametertests und Modellvergleiche in Strukturgleichungsmodellen

Um mit Strukturgleichungsmodellen inhaltliche Hypothesen zu prüfen, genügt es meist nicht, nur ein einziges, den Hypothesen entsprechendes Modell aufzustellen, an Daten zu prüfen und anhand des Modellfits über die Hypothesen zu entscheiden. Oft möchte man ausserdem

- wissen, ob vermutete Effekte im Modell tatsächlich signifikant sind, und/oder
- verschiedene Modelle vergleichen, z. B. Modelle, die sich aus alternativen Hypothesen ergeben.

Dafür ist es nützlich, nicht nur deskriptive Gütekriterien (wie CFI, RMSEA, etc.) zu betrachten, sondern exakt zu testen, ob Parameter von Null verschieden sind, bzw. welches von mehreren möglichen Modellen besser zu den Daten passt. Einzelne Parameter auf Signifikanz zu prüfen ist dabei ein Sonderfall des Vergleichs von mehreren Modellen, nämlich der Vergleich eines Modells mit dem fraglichen Parameter gegen ein alternatives Modell ohne diesen Parameter.

1 Test von Parametern auf Signifikanz

Um die Signifikanz von Parametern in Strukturgleichungsmodellen zu testen, gibt es grundsätzlich drei Wege (vgl. Bollen, 1989, S. 295f.; Buse, 1982), je nachdem, ob nur das Modell mit dem fraglichen Parameter analysiert wurde (z-Test), nur das Modell ohne den fraglichen Parameter (Modifikationsindex/Langrange Multiplier-Test), oder ein Modell mit und eines ohne den Parameter (χ^2-Differenztest/Likelihood Ratio-Test).

1.1 z-Test

Zu jedem im Modell geschätzten Parameter wird auch sein Standardfehler geschätzt, der angibt, wie genau die Parameterschätzung ist: Der Standardfehler ist die (geschätzte) Streuung von Schätzungen des gleichen Parameters, wenn man viele Stichproben der gleichen Grösse untersuchen würde. Je grösser die Stichprobe, desto genauer ist die Schätzung, desto kleiner also unter sonst gleichen Umständen der Standardfehler.

Wenn die Daten multivariat normalverteilt sind und die Stichprobe gross, ist der Quotient aus einem geschätztem Parameter $\hat{\pi}$ und seinem Standardfehler $\hat{\sigma}_{\hat{\pi}}$ ein standardnormalverteilter Wert (z-Wert). Dieser Quotient kann daher anhand der z-Verteilung auf

Signifikanz geprüft werden. Bei zweiseitigem Test mit $\alpha = 5\%$ sind z-Werte betraglich grösser als der kritische z-Wert von 1.96 signifikant:

$$\left| \frac{\hat{\pi}}{\hat{\sigma}_{\hat{\pi}}} \right| = |z| \geq z_{\text{crit}} = 1.96 \quad \Rightarrow \quad \text{Parameter } \pi \text{ signifikant}$$

Ist der z-Test signifikant, kann man davon ausgehen, dass der entsprechende Parameter in der Population ungleich null ist.

Ausgegeben werden solche z-Tests von Strukturgleichungsmodell-Programmen standardmässig für alle im Modell geschätzten Pfadkoeffizienten (Faktorladungen, Beziehungen zwischen latenten Variablen), z. T. auch für Varianzen. Ein verwandter, allgemeinerer Test, mit dem auch mehrere Parameter zugleich geprüft werden können, ist der *Wald*-Test (in Strukturgleichungsmodell-Programmen nicht standardmässig verfügbar, bei Test eines einzelnen Parameters äquivalent zum z-Test).

1.2 Modifikationsindex (Lagrange Multiplier-Test)

Zu *nicht* im Modell enthaltenen Parametern kann geschätzt werden, inwieweit sich bei Aufnahme dieser Parameter ins Modell die Anpassung zwischen Modell und Daten voraussichtlich verbessern wird (im Sinne einer Verminderung des χ^2-Wertes). Diese Schätzungen werden standardmässig in Strukturgleichungsmodell-Programmen für einzelne Parameter als Modifikationsindices ausgegeben, da sie Hinweise liefern können, ob ein schlecht zu den Daten passendes Modell bei Aufnahme zusätzlicher Parameter (Modifikation des Modells) besser mit den Daten zu vereinbaren wäre.

Unter der Bedingung, dass die Daten multivariat normalverteilt sind und die Stichprobe gross, sind Modifikationsindices χ^2-verteilt (mit einem Freiheitsgrad) und können so auf Signifikanz geprüft werden (hier wieder $\alpha = 5\%$ vorausgesetzt):

$$\text{Modifikationsindex} \geq \chi^2_{\text{crit, } df=1} = 3.84 \quad \Rightarrow \quad \text{Parameter signifikant}$$

Ist ein Modifikationsindex signifikant, kann man davon ausgehen, dass das Modell bei Aufnahme des entsprechenden Parameters besser mit den Daten zu vereinbaren wäre, so dass man indirekt schliessen kann, dass der fragliche Parameter in der Population ungleich null ist.

1.3 χ^2-Differenztest (Likelihood Ratio-Test)

Am allgemeinsten und oft auch am aussagekräftigsten ist es, zwei Modelle – sowohl ein Modell mit dem fraglichen Parameter oder mehreren fraglichen Parametern, als auch eines ohne – zu analysieren und die Modellanpassung mittels χ^2-Differenztest zu vergleichen. Allgemein wird dies auch als Likelihood Ratio-Test bezeichnet, da die Differenz von χ^2-Werten dem (logarithmierten) Quotienten der Likelihoods („Wahrscheinlichkeit“ der Daten bei Gültigkeit des Modells) entspricht. Dies wird im folgenden Abschnitt vorgestellt.

2 Vergleich geschachtelter Modelle: χ^2-Differenztest

In vielen Situationen kann man mittels χ^2-Differenztest direkt testen, welches von zwei alternativen Modelle besser mit den Daten übereinstimmt. Dies funktioniert beim Vergleich von sog. hierarchisch geschachtelten Modellen (nested models), bei denen ein Modell aus dem anderen durch zusätzlich geschätzte Parameter hervorgeht. Beispiele:

- Vergleich eines Modells mit zusätzlichen Pfaden gegen ein ansonsten identisches Modell ohne diese Pfade: Besteht zwischen zwei latenten Variablen ein Effekt oder nicht? Besteht neben einem indirekten (über eine dritte Variable vermittelten) Effekt auch ein direkter Effekt zwischen zwei Variablen?
- Vergleich eines Modell, welches einen Zusammenhang zwischen zwei latenten Variablen enthält, gegen ein Modell, welches keinen Zusammenhang zwischen diesen Variablen zulässt: Sind die Faktoren einer konfirmatorischen Faktorenanalyse korreliert oder nicht?
- Vergleich eines Modells mit einer zusätzlichen Ladung einer manifesten Variable auf einer latenten Variable gegen ein Modell ohne diese Ladung: Misst eine manifeste Variable ausschliesslich eine einzige latente Variable, oder misst sie auch noch Aspekte einer weiteren latenten Variable?

Durchführung des χ^2-Differenztests

Zur Durchführung des Modelldifferenztests wird die Differenz der χ^2-Werte der beiden fraglichen Modelle gebildet, sowie die Differenz der Freiheitsgrade der beiden Modelle. Testet man nur einen einzelnen Parameter, unterscheiden sich die beiden Modelle also nur um einen Freiheitsgrad.

$$\begin{aligned} \chi^2_{\text{diff}} &= \chi^2_{\text{klein}} - \chi^2_{\text{gross}} \\ df_{\text{diff}} &= df_{\text{klein}} - df_{\text{gross}} \end{aligned}$$

Dabei bezeichnet *klein* das kleinere Modell mit weniger Parametern und daher mehr Freiheitsgraden, *gross* das grössere Modell mit mehr Parametern und daher weniger Freiheitsgraden (andere häufig zu findende Bezeichungen sind *restricted*, d. h. weniger freie Parameter, und *unrestricted*, d. h. mehr freie Parameter, oder 0 für das kleinere Nullhypothesen-Modell mit weniger Parametern und 1 für das grössere Alternativhypothesen-Modell mit mehr Parametern).

Für normalverteilte Daten und grosse Stichproben ist die Differenz χ^2_{diff} selbst χ^2-verteilt mit Freiheitsgraden df_{diff}. Von Hand kann man die Differenz χ^2_{diff} mit einer χ^2-Tabelle auf Signifikanz prüfen, oder mit der entsprechenden R-Funktion `pchisq()`, die zu einem χ^2-Wert und Freiheitsgrad den zugehörigen Wert der kumulativen χ^2-Verteilungsfunktion ausgibt. Standardmässig wird dabei vom unteren Ende der χ^2-Verteilung (lower tail, bei 0) ausgegangen. Für einen p-Wert muss daher in R eine der folgenden beiden Alternativen verwendet werden:

```
1 - pchisq(chisq.klein - chisq.gross, df.klein - df.gross)
pchisq(chisq.klein - chisq.gross, df.klein - df.gross, lower.tail=FALSE)
```

Hat man zwei geschachtelte Modelle mit lavaan analysiert, ist es am einfachsten, die `anova()`-Funktion zum Modellvergleich zu verwenden, die beim Vergleich zweier analysierter Strukturgleichungsmodelle automatisch einen χ^2-Differenztest durchführt (hier z. B. für zwei als R-Objekte abgespeicherte, analysierte Modelle mit Namen `fitted.model.klein` und `fitted.model.gross`) :

```
anova(fitted.model.klein, fitted.model.gross)
```

Die Reihenfolge der Modelle ist hierbei egal. Man muss allerdings selbst darauf achten, dass es sich tatsächlich um hierarchisch geschachtelte Modelle handelt – dies kann die `anova()`-Funktion nicht in jedem Fall erkennen.

Ist der χ^2-Differenztest signifikant, so weist das grössere Modell (mit mehr frei zu schätzenden Parametern) eine signifikant bessere Anpassung an die Daten auf als das kleinere Modell (in dem die fraglichen Parameter nicht enthalten, bzw. fixiert/gleichgesetzt sind). Es lohnt sich also für bessere Modellanpassung, das grössere Modell anzunehmen und die fraglichen Parameter mit zu schätzen. Ist der χ^2-Differenztest dagegen nicht signifikant, so besteht kein Unterschied in der Anpassung der beiden Modelle an die Daten, so dass man aus Sparsamkeitsgründen das kleinere Modell annehmen und auf die fraglichen Parameter verzichten kann.

Praktische Anwendung

Theoretisch sind z-Tests, Modifikationsindices und χ^2-Differenztests für einzelne Parameter äquivalent, wenn die Daten multivariat normalverteilt sind und die Stichprobe gross ist. In der Praxis können die Tests durch Nichtnormalität der Daten und kleinen Stichprobenumfang aber unterschiedlich ausfallen. Da beim χ^2-Differenztest potentiell verzerrende Einflüsse in *beide* χ^2-Werte eingehen, besteht hier eine grössere Chance, auch bei Verletzung von Voraussetzungen aussagekräftige (Differenz-)Ergebnisse zu erhalten, als beispielsweise bei z-Tests, in die Standardfehlerschätzungen eingehen, die bei Nichtnormalität verzerrt (West, Finch & Curran, 1995) und in verschiedenen Situationen nicht optimal sein können (Gonzalez & Griffin, 2001; Neale & Miller, 1997). Generell sind χ^2-Differenztests daher für die Signifikanzprüfung wichtiger Parameter eine sinnvolle Wahl.

χ^2-Differenztests haben allerdings grundsätzlich die gleichen Stärken und Schwächen wie gewöhnliche χ^2-Tests, die einzelne Modelle auf Anpassung an die Daten prüfen: Die Ergebnisse sind direkt abhängig vom Stichprobenumfang, so dass in hinreichend grossen Stichproben auch minimale Unterschiede zwischen zwei Modellen signifikant werden.

Nicht möglich ist ein χ^2-Differenztest zwischen Modellen, die nicht hierarchisch auseinander hervorgehen und *unterschiedliche* Parameter enthalten, z. B. zwischen einem Modell, in dem x_1 ein Indikator der latenten Variablen ξ_1 ist, verglichen mit einem alternativen Modell, in dem x_1 *statt dessen* Indikator von ξ_2 ist. Möglichkeiten für diesen Fall werden im folgenden Abschnitt besprochen.

3 Vergleich nicht geschachtelter Modelle

Möchte man Modelle vergleichen, die nicht hierarchisch geschachtelt sind, bei denen also keines der Modelle durch zusätzliche Parameter direkt aus dem anderen hervorgeht, so gibt es folgende Möglichkeiten:

- Testen der nicht hierarchisch geschachtelten Modelle jeweils gegen ein gemeinsames Obermodell, in welches alle zu vergleichenden Modelle geschachtelt sind: Beispielsweise würde für die Frage, ob x_1 ein Indikator der latenten Variablen ξ_1 ist oder *statt dessen* ein Indikator von ξ_2, ein gemeinsames Obermodell *beide* Ladungen (von x_1 sowohl auf ξ_1 als auch auf ξ_2) enthalten. Gegen dieses Obermodell könnten dann die beiden alternativen, zu vergleichenden Modelle jeweils mittels χ^2-Differenztest getestet werden. Würde man idealerweise feststellen, dass sich nur für ein Modell ein signifikanter Unterschied gegenüber dem gemeinsamen Obermodell ergibt, kann über diesen Umweg zwischen den nicht geschachtelten Modellen entschieden werden.

- Deskriptiver Vergleich anhand von Modellgütekriterien: Modelle mit gleicher Anzahl Freiheitsgrade können anhand beliebiger deskriptiver Modellgütekriterien wie RMSEA, CFI verglichen werden. Bei Modellen mit ungleicher Anzahl Freiheitsgrade (unterschiedlich vielen Parametern) sollten nur Kriterien verwendet werden, die die Freiheitsgrade bzw. Parameteranzahl direkt berücksichtigen. Dies tut beispielsweise das Akaike Information Criterion AIC:

  $$\text{AIC} = \chi^2 + 2\,t \quad (t = \text{Anzahl der Parameter des Modells})$$

  Für gute Anpassung zwischen Modell und Daten sollte AIC möglichst klein sein, insofern würde man sich zwischen mehreren, nicht geschachtelten Modellen für das Modell mit niedrigstem AIC entscheiden.

Literatur

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley & Sons.

Buse, A. (1982). The likelihood ratio, Wald, and Langrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157. doi:10.2307/2683166

Gonzalez, R. & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, *6*, 258–269. doi:10.1037/1082-989X.6.3.258

Neale, M. C. & Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics*, *27*, 113–120. doi:10.1023/A:1025681223921

West, S. G., Finch, J. F. & Curran, P. J. (1995). Structural equation modeling with nonnormal variables: Problems and remedies. In R. H. Hoyle (Hrsg.), *Structural equation modeling: Concepts, issues and applications* (pp. 37–55). Thousand Oakes, CA: Sage.