Veränderungsmessung mit dem Linear Logistic Test Model

Stefan Klein, Andreas Frey, Ulrich Gauger
Humboldt-Universität zu Berlin
Institut für Psychologie
10178 Berlin
stefanklein@snafu.de
Andreas.Frey@rz.hu-berlin.de
Gauger@rz.hu-berlin.de

Zusammenfassung

Ein bekanntes Problem der Psychometrie liegt in der Messung von Veränderungen auf der Ebene einzelner Personen. Die Probleme der personenbezogenen Veränderungsmessung können umgangen werden, wenn man sich auf die Messung gruppenspezifischer Veränderungen beschränkt. Ein bekanntes Modell, das für diese Zwecke verwendet wird, ist das sog. Linear Logistic Test Model. In diesem Beitrag wird ein neuentwickeltes SAS-Programm zur Schätzung der Parameter dieses Modells vorgestellt. Zudem kann mit diesem SAS-Programm für einzelne Personen untersucht werden, ob vorgegebene Parameterwerte für diese Personen zutreffen können. Die Möglichkeiten des Programms werden anhand eines Beispiels illustriert.

Keywords: Linear Logistic Test Model, Latent Trait Modelle, Personenfittests, SAS IML, PROC CATMOD, Veränderungsmessung.

1 Einführung: Veränderungsmessung in der klassischen Psychometrie

Dieser Beitrag behandelt Verfahren der Latent-Trait-Analyse, die bei der Messung von Veränderungswerten verwendet werden können.
Ausgangspunkt sind dabei aus mehreren Teilgrößen (z.B. Fragebogenitems)

zusammengesetzte Messwerte. Ziel ist es, aus solchen Teilgrößen ein möglichst reliables und valides Messinstrument zu konstruieren. Die Reliabilität eines Messinstruments wird dabei i.d.R. durch das Verhältnis von interessierender Varianz zu Fehlervarianz gemessen (z.B. durch das bekannte Reliabilitätsmaß Cronbachs α). Validität wird meist über die Korrelation mit einem Außenkriterium sichergestellt.
Bei der Messung von Veränderungswerten benutzt die klassische Psychometrie Differenzwerte

$$X_v^{Diff} = X_{v2} - X_{v1} = T_2 - T_1 + E_2 - E_1$$

Hierbei ist X_v^{Diff} der Differenzenmesswert für die Veränderung, X_{vt} der Messwert zum Zeitpunkt t, T_t der wahre (fehlerfreie) Messwert und E_t der Fehler im Zeitpunkt t. Es wird angenommen , dass die Fehler unkorreliert von den wahren Messwerten sind, sowie dass die Fehler untereinander unkorreliert sind. Dieser Ansatz ist jedoch problematisch:

- Er führt zu einer erhöhten Fehlervarianz:

  $$\begin{aligned} &Var(X_v^{Diff}) = \\ &= Var(E_1) + Var(E_2) + Var(T_1) + Var(T_2) - 2Cov(T_1, T_2) - 2Cov(E_1, E_2) \\ &\approx Var(E_1) + VarE_2 \end{aligned} \quad (1)$$

  Dies gilt immer dann, wenn die wahren Messwerte T_1 und T_2 hoch miteinander korrelieren und somit ein reliables Messinstrument vorliegt (vgl. z.B. „Some Neglected Problems in IRT“ [3]).

- Bei diesem Ansatz wird nicht die Ausgangswertabhängigkeit berücksichtigt: So ist die Zahl richtiger Antworten bei einem Fragebogen nach oben begrenzt. Eine Versuchsperson mit hohem Ausgangsmesswert X_{v1} kann sich nicht so stark verbessern, wie eine Person mit einem niedrigen Ausgangsmesswert. (vgl. z.B. „Probleme der Veränderungsmessung im Rahmen der Evaluationsmethodik“ [7]).

Um diese Probleme zu vermeiden, werden für die Veränderungsmessung zumeist andere Modellklassen vorgeschlagen: Lineare Strukturgleichungsmodelle („Probleme der Veränderungsmessung im Rahmen der Evaluationsmethodik“ [7]), Markov-Modelle oder Latent-Trait-Modelle.

2 Latent-Trait-Modelle für die Veränderungsmessung

Latent-Trait-Modelle beschreiben die Wahrscheinlichkeit für die Beantwortung eines Items durch eine Versuchsperson mit Hilfe einer latenten Fähigkeitsvariable, dem sog. Trait.

Die hier beschriebenen Modelle sind Abarten des Rasch-Modells, bei dem sich die Lösungswahrscheinlichkeit eines Items alleine durch den latenten Fähigkeitsparameter der Versuchsperson und den Schwierigkeitsparameter des Items berechnen lässt. Fähigkeits- und Schwierigkeitsparameter sind dabei multiplikativ verknüpft, bzw. additiv, wenn man den Logarithmus der Lösungswahrscheinlichkeit betrachtet.
Als Erweiterungen des Raschmodells für die Veränderungsmessung wird v.a. das Linear Logistic Test Model (=LLTM, „A measurement model for the effect of mass media“ [2]) verwendet.
Das LLTM (=Linear Logistic Test Model) ist eine multivariate Erweiterung des gewöhnlichen Raschmodells für dichotome Items, mit dem Ziel, eine von einem Zeitpunkt, bzw. einer Subpopulation abhängende Fähigkeitsstärke messen zu können. Jeder Subpopulaton / jedem Zeitpunkt wird ein eigener Schwierigkeitsparameter zugeordnet. Das LLTM modelliert die Wahrscheinlichkeit einer richtigen Antwort auf folgende Weise:

$$P(X_{vi} = 1) = \frac{exp\,[\theta_v - \beta_i + \delta_{gt}]}{1 + exp\,[\theta_v - \beta_i + \delta_{gt}]} \tag{2}$$

Dabei ist θ_v die Fähigkeit einer Person v, β_i die Schwierigkeit eines Items i, sowie δ_{gt} die Veränderung der Itemschwierigkeit für Subpopulation g und Zeitpunkt t.
Eine gruppenspezifische Veränderung kann durch den Parameter δ_{gt} geschätzt werden, nicht jedoch eine personenspezifische Veränderung. Der Index g bezeichnet dabei eine beliebige Subpopulation.
Für die Item- und Veränderungsparameter dieses Modells ist Conditional Maximum Likelihood- (=CML-) Schätzung möglich, da der Summenscore r_v einer Person v suffizient für den Fähigkeitsparameter ist, und somit aus der Likelihood herausgerechnet werden kann (vgl. auch „The Linear Logistic Test Model“ [4]).
Weniger empfehlenswert ist die Joined Maximal Likelihood-Schätzung, bei der die gesamte Likelihood einer Stichprobe maximiert wird. Diese führt zu nichtkonsistenten Schätzungen (vgl. z.B.„The Linear Logistic Test Model“ [4]).
Nachteilig bei der CML-Schätzung ist, dass diese nicht in Standardpaketen für kategoriale Datenanalyse wie PROC CATMOD oder PROC GENMOD enthalten ist.
Möglich wird eine Schätzung z.B. mit PROC CATMOD dadurch, dass man das LLTM in ein loglineares Modell für eine 2^m-Kontingenztafel umformuliert:

$$\begin{aligned} &log(P(X_v = x_v)) = \\ &= -\sum_t \left(\beta_i \sum_i x_{vit} \right) + \sum_t \sum_g \delta_{gt} + log \left(\frac{P\left(\sum_{i,t} X_{vit} = r_v\right)}{\sum_{(\sum_{v,i} X_{vit} = r_v)} \prod_{v,i} exp\,[\theta_v - \beta_i + \delta_{gt}]} \right) \\ &= -\sum_t \left(\beta_i \sum_i x_{vit} \right) + \sum_t \sum_g \delta_{gt} + u_r \end{aligned} \tag{3}$$

(vgl. hierzu auch „Loglinear Rasch Models for the Analysis of Stability and Change“[8]). Hierbei ist $r_v = \sum_v x_{vit}$ der Summenscore der Person v.
Eine konsistente Schätzung der Modellparameter ist dann mittels Maximum-Likelihood-Schätzung möglich. Nachteilig bei dieser Modellierung ist die Größe des dafür benötigten Datensatzes. Schon bei wenigen Items wächst die Zahl der Zellen in der Kontingenztafel stark an. Bei nur 7 Items pro Zeitunkt und 2 Messzeitpunkten erhält man eine Kontingenztafel mit 2^{14} Zellen.

3 Personenfittests in Latent-Trait-Modellen

Modellgeltungsannahmen werden in einem Latent-Trait-Modell gewöhnlicherweise mit Likelihood-Quotienten-Tests überprüft.
Einen weiteren, interessanten Ansatz zur Überprüfung des Modells bieten die sog. Personenfittests (vgl. z.B. „The Assessment of Person Fit“[6], „Exact person fit Indexes for the Rasch model for arbitrary alternatives“[9]). Diese untersuchen für einzelne Personen, ob das Antwortmuster einer Person kompatibel zu den geschätzten Modellparametern ist.
Wichtig hierbei ist die Kontrolle des Gesamtsignifikanzniveaus. Falls dieses nicht beachtet wird, können zufallsinduzierte Ablehnungen der Nullhypothese auftreten. Ein entsprechendes Verfahren wird z.B. von „An exact and optimal standardized person fit test for assessing consistency with the Rasch model“[5] dargestellt.

Im Rahmen der Veränderungsmessung kann mit Hilfe der Personenfittests untersucht werden, ob die Antwortmuster einzelner Personen bestimmten Veränderungstypen entsprechen. Im hier vorliegenden Beispiel wird die Gültigkeit der geschätzten Veränderungsparameter untersucht. Somit ist ein Rückschluss von den für eine ganze Personengruppe geschätzten Parametern auf einzelne Versuchspersonen möglich.

3.1 Ein SAS-SCL-Programm für die Veränderungsmessung mit Latent-Trait-Modellen

Wie weiter oben erwähnt, existiert für die Modellierung von Veränderungen mit dem LLTM in SAS keine vollkommen zufriedenstellende Lösung. Daher wurde ein SCL-Programm für diesen Zweck entwickelt.
Dem Programm zugrunde liegt eine Datenbankanwendung, in der die wichtigsten Informationen über die zu analysierenden Datensätze gespeichert sind.

Dazu zählen z.B.:

- Informationen über Versuchspersonen,
- Katalogisierung der verwendeten Items,
- Informationen über Datentabellen etc.

Die Schätzung des LLTMs erfolgt durch Maximierung der Conditional Likelihood in einem SAS-IML-Modul. Die Maximierung erfolgt dabei durch eine IML-Optimierungsroutine.
Die Item- und Veränderungsparameter werden in eine SAS-Tabelle der Datenbank ausgegeben.
Die Bestimmung der Modellgüte erfolgt automatisch bei der Schätzung der Modellparameter mittels Likelihood- Quotiententests. Implementiert sind Tests zum Vergleich mit einem Modell ohne Veränderung und zum Vergleich mit einem konstanten Modell (d.h.: alle Parameter nehmen den Wert 0 an).
Bei Personenfittests wird zwischen intervallförmigen und punktförmigen Nullhypothesen unterschieden. Bei ersteren müssen Ober- und Untergrenze des Nullhypothesenintervalls angegeben werden, bei letzteren die Nullhypothese für einen (gewöhnlichen) zweiseitigen Signifikanztest.
Die Personenfittests werden für alle Personen einer Stichprobe durchgeführt. Die Ergebnisse dieser Tests werden in eine temporäre SAS-Tabelle geschrieben. Da randomisierte gleichmäßig beste (statistische) Tests durchgeführt werden, enthält die Ausgabetabelle neben den Grenzen des Ablehnbereichs auch die Wahrscheinlichkeit, mit der die Nullhypothese an den Grenzen des Ablehnbereichs abgelehnt wird.

4 Datenbeispiel: Veränderungsmessung bei einer Subskala des EORTC QLQ-C30

Im folgenden wird anhand eines Fragebogens zur Lebensqualität dargestellt, wie das oben erwähnte Programm eingesetzt werden kann.

4.1 Der EORTC QLQ-C30: ein Fragebogen zur Lebensqualität von Krebspatienten

Zunächst stellen wir kurz den EORTC QLQ-C30 (=European Organization for Research and Creatment of Cancer Quality of Life Questionaire) vor (vgl. dazu auch „The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology“[1]).

Der EORTC QLQ-C30 besteht aus 30 Items, die 9 verschiedenen Traits zugeordnet werden können. Diese Teilskalen können in 5 funktionale Skalen (physisch, sozial, Rollenverhalten, kognitiv, emotional), 3 Symptomskalen(Schmerzen, Müdigkeit, Erbrechen und Übelkeit) und eine globale Lebensqualitätsskala unterteilt werden.
„The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology“[1] berichtet, dass bei allen Teilskalen außer der Skala Rollenverhalten zufriedenstellende Werte für Cronbachs α vorliegen ($\alpha \geq 0.70$). Bei der Skala Rollenverhalten liegt Cronbachs α lediglich bei 0.5.
Ebenfalls untersucht wurden die Korrelationen zwischen Teilskalen. Zwischen der physischen Skala und der Skala „Rollenverhalten“ wird eine Produkt-Moment-Korrelation von ca. 0.60 berichtet.
Insgesamt wird der EORTC QLQ-C30 als reliabel und valide bezeichnet.
In dem folgenden Beispiel werden die 7 Items der physischen und der Rollenverhaltensskala zusammen verwendet. Dies erscheint v.a. deswegen gerechtfertigt, weil diese beiden Skalen inhaltlich sehr ähnlich sind. Letzteres spiegelt sich auch in der Korrelation zwischen den Teilskalen wieder.
Da mit dem LLTM nur dichotome Items verarbeitet werden können, wurden die verwendeten Items nachträglich dichotomisiert. Die Scores 1 und 2 erhalten den neuen Wert 0, die Scores 3 und 4 den neuen Wert 1.

4.2 Analyse der Stichprobe mit dem Linear Logistic Test Model

Analysiert wird eine Stichprobe von 295 Patienten der Charité Berlin, bzw. angeschlossener Kliniken, die zwischen Dezember 1999 und Juni 2000 erhoben wurde. Bei 221 von diesen 295 Patienten liegen vollständige Datensätze vor, die übrigen Patienten gehen nicht in die Analyse ein.
Die Stichprobe kann in stationäre und ambulante Personen unterteilt werden. Untersucht werden soll, ob bei stationären und ambulanten Patienten Unterschiede in der Veränderung der Lebensqualität auftreten.
Dazu wird ein LLTM mit 7 Items pro Zeitpunkt, 2 Untersuchungsgruppen und 2 Zeitpunkten angepasst.
Folgende Itemparameter wurden geschätzt: Wie man an Tabelle 1 erkennen kann, sind die Itemschwierigkeiten in der Subskala physisch weit gestreut, während in der Subskala Rollenverhalten nur geringe Unterschiede vorliegen.
Für die ambulanten Patienten wurde eine Veränderung von -0.13 geschätzt, für die stationären eine Veränderung von 0.08.
Der Likelihood-Quotiententest für die Nullhypothese „Keine Veränderung in beiden Gruppen“ ergibt die mit 2 Freiheitsgraden χ^2-verteilte Testgröße 0.89. Die Nullhypothese dieses Tests kann daher nicht zum Signifikanzniveau 0.05 abgelehnt werden.

Tabelle 1: Schätzwerte für die Itemparameter

Itemnr.	Schätzwert	Skala
1	2.73852	Physisch
2	1.36144	Physisch
3	-1.29693	Physisch
4	-1.10375	Physisch
5	-3.30540	Physisch
6	0.91619	Rollenverhalten
7	0.68993	Rollenverhalten

Die mit 8 Freiheitsgraden χ^2-verteilte Testgröße für die Nullhypothese „Kein Einfluss der Itemparameter“ ergibt den Wert 867.0. Die Nullhypothese kann in diesem Fall daher abgelehnt werden.
Schließlich werden für alle beteiligten Probanden Personenfittests zur Nullhypothese $\delta_{gt} = 0$ zum Signifikanzniveau 0.1 durchgeführt. Auch hier zeigt sich nur eine äußerst geringe Veränderung. Bei nur 3 Personen lehnen die randomisierten Personenfittests die Nullhypothese ab, bei weiteren 7 Personen fällt die Testgröße auf die Grenze des Ablehnereichs und kann mit der Wahrscheinlichkeit 0.8 abgelehnt werden. Bei den restlichen Personen sind die Unterschiede zwischen dem Antwortverhalten zu gering, um gesicherte Aussagen treffen zu können.
Angesichts der geringen Häufigkeit von Personen mit schlechtem Fit muss man die erreichten Ablehnungen der Nullhypothese wohl als zufällig bewerten. Zum Vergleich: Aufgrund des Signifikanzniveaus von $\alpha = 0.1$ kann man bei 221 durchgeführten Tests ca. 22 Ablehnungen der Nullhypothese erwarten.

5 Zusammenfassung

Es wurde ein SAS / SCL-Programm vorgestellt, mit dessen Hilfe menügesteuert die Schätzwerte für LLTMs bestimmt, sowie optimale Personenfittests für das LLTM durchgeführt werden können.
Besonders hervorgehoben muss dabei die Möglichkeit der Formulierung intervallförmiger Nullhypothesen des Typs

$$k_u \geq \delta_{gt} \geq k^o$$

wobei k_u und k^o die Grenzen des Annahmebereichs sind. Diese ermöglichen die Angabe eines interessierenden Intervalls, in dem ein bestimmter Parameter liegen darf, und sind somit wesentlich praktischer anzuwenden als die punktförmigen Nullhypothesen.

Im vorhandenen Datensatz kann die Nullhypothese „Keine Veränderung“ zwar bei einigen Probanden abgelehnt werden, insgesamt ist die Zahl der zurückweisungen der Nullhypothese(n) jedoch zu niedrig.
Festzuhalten bleibt somit, dass Personenfittests und Likelihoodquotiententests bei diesem Datensatz zum gleichen Ergebnis kommen.

Literatur

[1] Aaronson, Neil K. , Ahmedzai, Sam , Bergman, Bengt et al. . (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality -of-Life Instrument for Use in International Clinical Trials in Oncology. Journal of the National Cancer Institute , **85**, 365-376 .

[2] Fischer, Gerhard H. (1972). A measurement model for the effect of mass media. Acta Psychologica , **36**, 207-220 .

[3] Fischer, Gerhard H. (1995). Some Neglected Problems in IRT. Psychometrika , **60**, 459-487 .

[4] Fischer, Gerhard H. (1995). The Linear Logistic Test Model. In Rasch Models. Foundations, Recent Developments and Applications, Editors: Gerhard H. Fischer and Ivo W. Molenaar, New York Berlin Heidelberg, Springer, 131-155.

[5] Klauer, Karl Christoph. (1991). An exact and optimal standardized person fit test for assessing consistency with the Rasch model. Psychometrika , **56**, 213-228.

[6] Klauer, Karl Christoph. (1995). The Assessment of Person Fit. In Rasch Models. Foundations, Recent Developments and Applications, Editors: Gerhard H. Fischer, Ivo W. Molenaar, New York Berlin Heidelberg, Springer, 97-110.

[7] Krause, Bodo (1997). Probleme der Veränderungsmessung im Rahmen der Evaluationsmethodik. In Empirische Evaluationsmethoden Band 1: Veränderungsmessung und Interventionsevaluation, Herausgeber: Bodo Krause und Peter Metzler, Berlin, ZeE-Publikationen, 7-25.

[8] Meiser, Thorsten. (1996). Loglinear Rasch Models for the Analysis of Stability and Change. Psychometrika , **61**, 629-645.

[9] Ponocny, Ivo. (2000). Exact person fit Indexes for the Rasch model for arbitrary alternatives. Psychometrika , **65**, 29-42.