Schätzung kumulativer Wahrscheinlichkeiten geordneter kategorialer Daten bei festgelegter Kategorienanzahl und unterschiedlich leeren Klassen

Karen Höltl
Martin-Luther-Universität Halle-Wittenberg,
Institut für Agrar- und Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
Halle (Saale)
karen.hoeltl@landw.uni-halle.de

Katrin Thamm
Martin-Luther-Universität Halle-Wittenberg,
Institut für Agrar- und Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
Halle (Saale)
katrin.thamm@landw.uni-halle.de

Norbert Mielenz
Martin-Luther-Universität Halle-Wittenberg,
Institut für Agrar- und Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
Halle (Saale)
norbert.mielenz@landw.uni-halle.de

Joachim Spilke
Martin-Luther-Universität Halle-Wittenberg,
Institut für Agrar- und Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
Halle (Saale)
joachim.spilke@landw.uni-halle.de

Zusammenfassung

Die statistische Analyse geordneter kategorialer Daten nach dem Schwellenwertmodell kann in SAS mit der Prozedur GLIMMIX durchgeführt werden. Interessieren im Zuge der Auswertung auch die Wahrscheinlichkeiten innerhalb der Kategorien, so muss ihre Berechnung innerhalb der Prozedur GLIMMIX über ESTIMATE-Statements erfolgen. Dies führt jedoch zur falschen Zuordnung der Ergebnisse, wenn die Kategorienanzahl im Vorfeld festgelegt ist und die auszuwertenden Datensätze nicht für alle Kategorien Beobachtungen aufweisen. Im Beitrag wird anhand von zwei Datensätzen gezeigt, wie bei Nutzung der Prozeduren GLIMMIX und IML die Berechnung kumulativer Wahrscheinlichkeiten bei Beachtung der tatsächlich beobachteten Kategorien flexibel erfolgen kann. Die Berechnung erfolgt auf Basis der von der Prozedur GLIMMIX bereitgestellten Schätzwerte für die Stutzungspunkte, Behandlungseffekte, deren Kovarianzmatrix sowie tatsächlich besetzten Kategorien.

Schlüsselwörter: kategoriale Daten, Schwellenwertmodell, kumulative Wahrscheinlichkeiten, GLIMMIX, IML

1 Einleitung und Problemstellung

Die Analyse geordneter kategorialer Merkmale wie Boniturdaten führt zur Anwendung der Multinomialverteilung. Ihre Auswertung kann durch die Nutzung des Schwellenwertmodells [1] erfolgen, das den generalisierten linearen Modellen zuzuordnen ist. Dieses Modell bietet die Möglichkeit, die Wahrscheinlichkeiten für einzelne Kategorien sowie die kumulativen Wahrscheinlichkeiten zu schätzen. In SAS ist die Umsetzung u. a. in der Prozedur GLIMMIX realisiert. Das Schwellenwertmodell basiert auf der Schätzung von Stutzungspunkten und Behandlungseffekten. Für m Kategorien und a Behandlungen führt dies zu m-1 Stutzungspunkten und a Behandlungseffekten.
Beispiele für Boniturdaten im landwirtschaftlichen Versuchswesen sind die Bewertung von Pflanzenbeständen bezüglich ihres Pilzbefalls oder Lagerverhaltens. Sollen Versuchsergebnisse mehrerer Jahre oder verschiedener Pflanzenbestände miteinander verglichen werden oder bezieht sich die Auswertung auf die Analyse von Simulationsstudien, so ist die Kategorienanzahl im Allgemeinen über einen vorgegebenen Boniturschlüssel festgelegt. Das bedeutet jedoch nicht zwangsläufig, dass jede Kategorie besetzt sein muss.
Liegt das Interesse bei der Auswertung auf der Schätzung der kumulativen Wahrscheinlichkeiten für die einzelnen Kategorien, so kann ihre Berechnung innerhalb der Prozedur GLIMMIX durch ESTIMATE-Statements angewiesen werden. Das ESTIMATE-Statement erfordert die Vorgabe eines Koeffizientenvektors, dessen Struktur von der Anzahl beobachteter Kategorien und damit Stutzungspunkten sowie der Behandlungseffekte abhängt. Das führt zu Problemen, falls im Rahmen von Routineauswertungen (bspw. PIAF = Planungs-, Informations- und Auswertungssystem für Feldversuche) oder Simulationen von einer vorgegebenen Kategorienanzahl ausgegangen und entsprechend die Koeffizientenvektoren in den ESTIMATE-Statements aufgebaut sind, in den Beobachtungen aber Kategorien fehlen. Leere Kategorien führen aber zu einer Reduktion der Stutzungspunkte und in den ESTIMATE-Statements sind dann teilweise zu viele Koeffizienten vorgegeben. Das führt zur Reduzierung gültiger Ergebnisse aus den ESTIMATE-Statements und dazu, dass die Ergebnisse nicht den richtigen Kategorien zugeordnet werden. Fehlt beispielsweise Note 7, so bezieht sich das Ergebnis des zugehörigen ESTIMATE-Statements auf Note 8, ohne dass das aus den Ergebnissen abzulesen ist.
Für die Auswertung von Datensätzen mit leeren Kategorien bei festgelegter Kategorienanzahl muss also ein Weg gefunden werden, die Schätzung der kumulativen Wahrscheinlichkeiten auf die tatsächlich vorhandene Kategorienanzahl anzupassen. Dies kann mittels Programmierung innerhalb der Prozedur IML erfolgen und wird im Folgenden anhand von zwei Beispieldatensätzen demonstriert.

2 Schwellenwertmodell

Im Schwellenwertmodell wird der Zusammenhang zwischen der kategorialen Merkmalsausprägung und einer stetigen nichtbeobachtbaren Zufallsgröße, der so genannten "latenten Variablen", über Schwellenwerte hergestellt.

Sei y_{ij} die j-te Beobachtung (j = 1, ... ,r) von Behandlung i (i = 1, ..., a).

Weiter sei Y_{ij} eine Zufallsgröße, deren Realisierungen y_{ij} die Beobachtungen des Versuches darstellen.

Sei $Y_{ij} = k$ falls $Z_{ij} \in [\theta_{k-1}, \theta_k]$; k = 1, ..., m und für die Schwellenwerte θ_k gilt:

$\theta_k > \theta_{k-1}$ für k = 1, ..., m mit $\theta_0 = -\infty$ und $\theta_m = +\infty$.

Die latente Variable Z sei normalverteilt und es gelte $Z_{ij} = \eta_i + \varepsilon_{ij}$ mit $E(\varepsilon_{ij}) = 0$ und $Var(\varepsilon_{ij}) = \sigma^2$.

Für ihren Erwartungswert (linearen Prädiktor) gelte:

$E(Z_{ij}) = \eta_i$ mit $\eta_i = \mu + \alpha_i$, mit $\alpha_i =$ Effekt der i-ten Behandlung.

Die kumulativen Wahrscheinlichkeiten werden dann wie folgt modelliert:

$$q_{ijk} = Pr(Y_{ij} \le k) = Pr(Z_{ij} \le \theta_k) = Pr\left(\frac{Z_{ij} - \eta_i}{\sigma} \le \frac{\theta_k - \eta_i}{\sigma}\right) = \Phi\left(\frac{\theta_k - \eta_i}{\sigma}\right)$$

mit $\Phi(z)$ Verteilungsfunktion der Standardnormalverteilung.

Da der Quotient $\frac{\theta_k - \eta_i}{\sigma}$ nicht eindeutig identifizierbar ist, wird gefordert: $Var(Z_{ij}) = \sigma^2 = 1$.

3 Auswertung mit der SAS Prozedur GLIMMIX

3.1 Beispieldatensätze

Beispiel sind zwei Datensätze mit je 20 Behandlungen und r = 4 Wiederholungen sowie ein auf 9 Kategorien festgelegter Boniturschlüssel. Dabei weist der zweite Datensatz zwei leere Kategorien (Kategorie 1 und 6) auf. Tabelle 1 zeigt die absoluten Häufigkeiten der aufgetretenen Kategorien über alle Behandlungen. Für beide Datensätze sollen mit einem einheitlichen SAS-Programm die kumulativen Wahrscheinlichkeiten für die

Kategorien der Behandlung 1 geschätzt werden. Durch den vorgegebenen Boniturschlüssel mit 9 Kategorien existieren im Modell die 8 Stutzungspunkte θ_1 bis θ_8.

Tabelle 1: Absolute Häufigkeiten der aufgetretenen Kategorien der Beispieldatensätze über alle Behandlungen

Kategorie	Datensatz 1	Datensatz 2
1	2	-
2	24	24
3	18	18
4	8	14
5	10	17
6	10	-
7	3	3
8	3	1
9	2	3
Summe:	80	80

3.2 Berechnung von Wahrscheinlichkeiten mit dem ESTIMATE-Statement

Die Schätzung der Stutzungspunkte und der Behandlungseffekte wird mit der Prozedur GLIMMIX ausgeführt. Sollen die kumulativen Wahrscheinlichkeiten der einzelnen Kategorien geschätzt werden, so muss dies über das ESTIMATE-Statement unter Vorgabe eines Koeffizientenvektors erfolgen. Im Falle der hier vorliegenden Kategorienanzahl 9 sind die Koeffizientenvektoren der ESTIMATE-Statements für die Behandlung 1 entsprechend den nachfolgenden Anweisungen in PROC GLIMMIX zu definieren. Für die Kategorie 9 muss kein ESTIMATE-Statement definiert werden, da die kumulative Wahrscheinlichkeit für die letzte Kategorie gemäß Definition 1 entspricht.

```
proc glimmix data=              ;
  CLASS Behandlung;
  MODEL boni= Behandlung
  DIST=multinomial
  LINK=cumprobit;

  ESTIMATE
  "1_Behandlung1" INTERCEPT 1                 Behandlung 1 /ILINK;
  ESTIMATE
  "2_Behandlung1" INTERCEPT 0 1               Behandlung 1 /ILINK;
          ⋮           ⋮    ⋮⋮                    ⋮     ⋮    ⋮
  ESTIMATE
  "7_Behandlung1" INTERCEPT 0 0 0 0 0 0 1     Behandlung 1 /ILINK;
  ESTIMATE
  "8_Behandlung1" INTERCEPT 0 0 0 0 0 0 0 1   Behandlung 1 /ILINK;
RUN;
```

Die Tabelle 2 zeigt die so erhaltenen Ergebnisse für die Schätzungen der kumulativen Wahrscheinlichkeiten innerhalb der Kategorien für die Beispieldatensätze. Dabei erfolgte eine Transformation der Schätzwerte im linearen Prädiktor in Wahrscheinlichkeiten nach

$$q_{ijk} = \Pr(Y_{ij} \le k) = \Pr(Z_{ij} \le \theta_k) = \Pr\left(\frac{Z_{ij} - \eta_i}{\sigma} \le \frac{\theta_k - \eta_i}{\sigma}\right) = \Phi\left(\frac{\theta_k - \eta_i}{\sigma}\right)$$

mit $\Phi(z)$ Verteilungsfunktion der Standardnormalverteilung mit der Anweisung der Option ILINK (vergl. Abschnitt 2).
Für den ersten Datensatz stimmen die im Ergebnis des ESTIMATE-Statements erhaltenen Kategorien mit den tatsächlichen Kategorien überein. Die Ergebnisse für Datensatz 2 zeigen dagegen eine falsche Zuordnung der Kategorien sowie fehlende kumulative Wahrscheinlichkeiten für die Kategorien 1 und 6. Aufgrund der Nichtbesetzung der ersten und sechsten Kategorie (Tabelle 1) sind tatsächlich nur 7 Kategorien vorhanden und mit der Prozedur GLIMMIX werden infolge nur m-1=6 Stutzungspunkte geschätzt. Der erste geschätzte Stutzungspunkt trennt eigentlich die zweite und dritte Kategorie voneinander. Bei Nutzung des Koeffizientenvektors im ESTIMATE-Statement wird die kumulative Wahrscheinlichkeit der zweiten Kategorie jedoch der ersten Kategorie zugeordnet. Die folgenden 3 Wahrscheinlichkeiten werden falsch der nächst niedrigeren Klasse zugewiesen. Dieser Fehler in der Zuordnung findet sich in analoger Weise auch für die kumulative Wahrscheinlichkeit der siebten Kategorie.

3.3 Berechnung der Wahrscheinlichkeiten mittels Programmierung in der Prozedur IML

Die Berechnung der kumulativen Wahrscheinlichkeiten bei festem Boniturschlüssel und flexibler Beachtung der tatsächlichen Klassenbesetzung erfolgt bei Nutzung der Prozedur IML in mehreren Schritten. Als Erstes müssen im Vorfeld verschiedene Ergebnisse

Tabelle 2: Schätzungen der kumulativen Wahrscheinlichkeiten (kum. WK) der Kategorien der Beispieldatensätze für die Behandlung 1 mittels ESTIMATE-Statements und mittels Programmierung in IML (Wahrscheinlichkeiten der Kategorien bei Beachtung der tatsächlichen Kategorien) sowie Schätzungen der Wahrscheinlichkeiten (WK) je Kategorie mittels Programmierung in IML

	Datensatz 1			
wahre Kategorie	Kategorie gemäß ESTIMATE	kum. WK gemäß ESTIMATE	WK gemäß IML	kum. WK gemäß IML
1	1	0.0301	0.0301	0.0301
2	2	0.4135	0.3834	0.4135
3	3	0.6763	0.2628	0.6763
4	4	0.7791	0.1028	0.7791
5	5	0.8853	0.1062	0.8853
6	6	0.9644	0.0791	0.9644
7	7	0.9819	0.0175	0.9819
8	8	0.9946	0.0127	0.9946
9	9[#]	1.0000[#]	0.0054	1.0000[#]

	Datensatz 2			
wahre Kategorie	Kategorie gemäß ESTIMATE	kum. WK gemäß ESTIMATE	WK gemäß IML	kum. WK gemäß IML
1			0.0000	0.0000
2	1	0.6137	0.6137	0.6137
3	2	0.8370	0.2233	0.8370
4	3	0.9353	0.0983	0.9353
5	4	0.9947	0.0594	0.9947
6			0.0000	0.9947
7	5	0.9984	0.0037	0.9984
8	6	0.9991	0.0007	0.9991
9	7[#]	1.0000[#]	0.0009	1.0000[#]

[#]: per Definition der letzten Kategorie der Multinomialverteilung

der Prozedur GLIMMIX als Datei ausgelesen und teilweise durch DATA STEPS für die Nutzung innerhalb der Prozedur IML aufbereitet werden. Zu den benötigten Ergebnissen zählen neben der Information über die tatsächlich aufgetretenen Kategorien die Schätzung der Stutzungspunkte und Behandlungseffekte sowie die Kovarianzmatrix der festen Effekte. Als nächster Schritt erfolgt die eigentliche IML Prozedur. Hier werden die kumulativen und Einzelwahrscheinlichkeiten für alle Kategorien sowie deren Standardfehler berechnet. Hierbei besteht in IML die Möglichkeit, die Schätzungen der Stutzungspunkte und somit auch die berechneten Wahrscheinlichkeiten gezielt ihrer tatsächlichen Kategorie zuzuordnen. Für leere Kategorien werden innerhalb der IML Prozedur keine Wahrscheinlichkeiten geschätzt. Um leere Kategorien in der Ergebniszusammenstellung mit einer Einzelwahrscheinlichkeit von 0 bzw. einer unveränderten kumulativen Wahrscheinlichkeit aufzufüllen, werden an die IML-Prozedur weitere DATA STEPS angeschlossen. Für die Erzeugung übersichtlicher Grafiken können mittels zusätzlicher DATA STEPS, ausgehend von den Ergebnissen der IML-Prozedur, die Differenzen der kumulativen Wahrscheinlichkeiten je Kategorie inklusive ihrer Standardfehler für verschiedene Behandlungen berechnet werden.
Die Programmstruktur im SAS Code ist in der folgenden Übersicht zusammengefasst.

Programmstruktur des SAS-Codes zur flexiblen Berechnung von Wahrscheinlichkeiten für die Kategorien einer Multinomialverteilung mit leeren Kategorien bei festgelegter Kategorienanzahl

1. Schritt PROC GLIMMIX

```
proc glimmix DATA=  ;
  CLASS Behandlung;
  MODEL boni = Behandlung / s covb
  DIST=multinomial
  LINK=cumprobit;
  ODS OUTPUT PARAMETERESTIMATES = solution
                        (KEEP=Bsp boni Behandlung estimate);
  ODS OUTPUT COVB = covb(DROP = effect boni Behandlung row );
  ODS OUTPUT responseprofile = profile ;
RUN;
```

2. Schritt PROC IML

Teilschritte:

1. Einlesen der von PROC GLIMMIX erzeugten Dateien.
2. Abfrage der verfügbaren Stutzungspunkte entsprechend der tatsächlichen Kategorien.
3. Erzeugung des Koeffizientenvektors für die Berechnung der kumulativen Wahrscheinlichkeiten.
4. Schätzung der kumulativen Wahrscheinlichkeiten sowie deren Standardfehler und der Wahrscheinlichkeiten je Kategorie.

3. Schritt Ergebnisbearbeitung

Teilschritte:

1. Erzeugung einer vollständigen Kategorienliste.
2. Zuordnung der in Schritt 2 geschätzten Wahrscheinlichkeiten zu den Kategorien entsprechend des vollständigen Boniturschlüssels.
3. Berechnung der Differenzen der kumulativen Wahrscheinlichkeiten je Kategorie inklusive ihrer Standardfehler für verschiedene Behandlungen.

Die Programmierung innerhalb der IML Prozedur liefert die in Tabelle 2 dargestellten Ergebnisse für die Schätzungen der kumulativen Wahrscheinlichkeiten der Kategorien. Die Ergebnisse für den Datensatz 1 stimmen wie erwartet mit denen der Berechnung mittels der ESTIMATE-Statements überein. Für den Datensatz 2 stimmt jetzt die Zuordnung der geschätzten kumulativen Wahrscheinlichkeiten zu ihren Kategorien und die leeren Kategorien tragen mit der Einzelwahrscheinlichkeit von 0 zur kumulativen Wahrscheinlichkeit bei. Die Abbildung 1 zeigt für den Datensatz 1 die Differenzen der kumulativen Wahrscheinlichkeiten je Kategorie für die Behandlung 1 und 2 inklusive ihrer P=0.95 Konfidenzintervalle. Da das Konfindenzintervall der Differenz in jeder Kategorie die Null beinhaltet, ist für den Datensatz 1 kein signifikanter Unterschied zwischen der Behandlung 1 und 2 nachzuweisen. Dieses Ergebnis stimmt mit dem Test der Behandlungsdifferenz mittels ESTIMATE-Statements überein (hier nicht gezeigt).

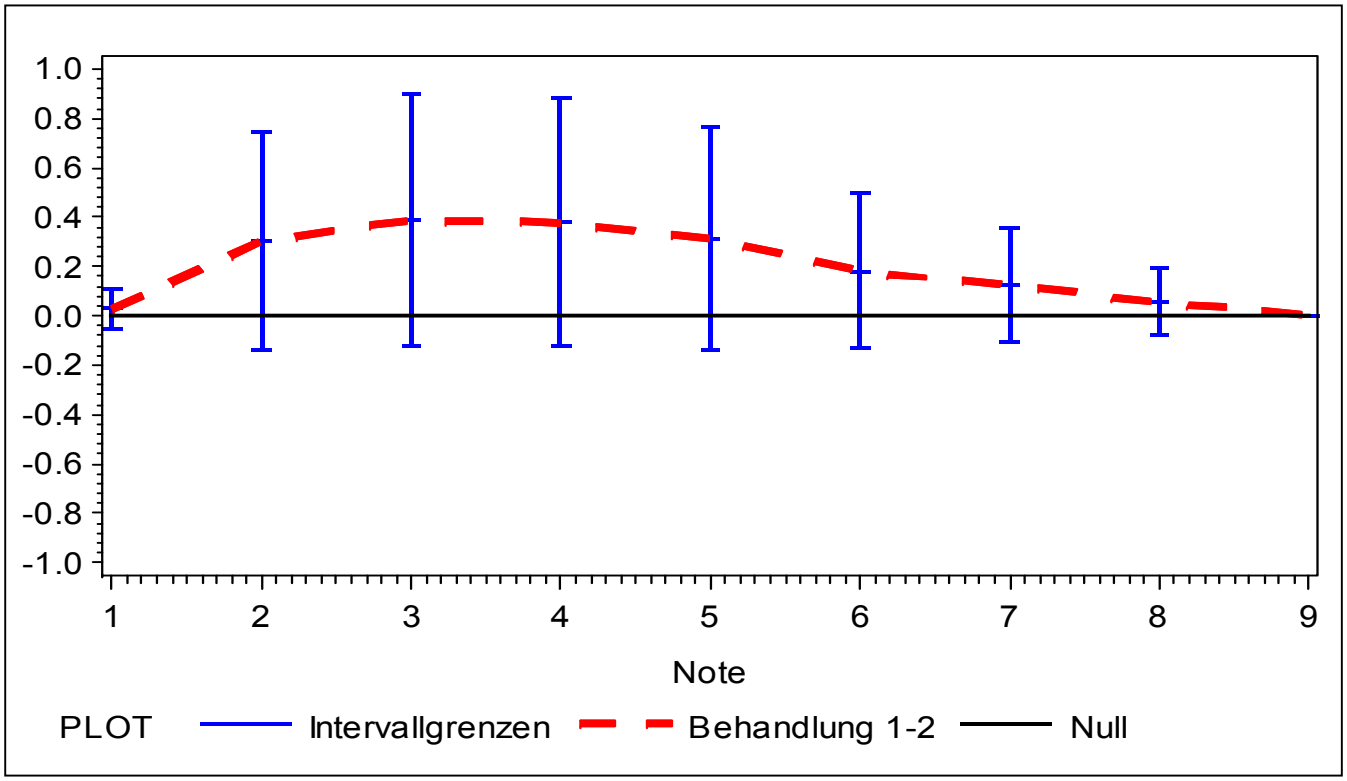

Abbildung 1: Differenzen der kumulativen Wahrscheinlichkeiten der Behandlung 1 und 2 für den Datensatz 1 sowie ihr P=0.95 Konfidenzintervall

4 Schlussfolgerungen

Für multinomialverteilte Daten erfordert die Schätzung von kumulativen Wahrscheinlichkeiten für die Kategorien in der SAS Prozedur GLIMMIX die Berechnung über ESTIMATE-Statements. Diese Statements erfordern die Vorgabe eines Koeffizientenvektors, dessen Struktur von der Anzahl beobachteter Kategorien und damit der Anzahl der Stutzungspunkte sowie der Behandlungseffekte abhängig ist.
Bezieht sich die Auswertung auf verschiedene Versuche oder Simulationsuntersuchungen und damit stets auf mehrere Simulationsläufe mit einem vorgegebenen Boniturschlüssel, so ist damit auch die Struktur des Koeffizientenvektors festgelegt. Die rechentechnische Umsetzung bei Nutzung von ESTIMATE führt jedoch zu ungültigen Ergebnissen, wenn in der Stichprobe leere Kategorien auftreten. Um diese Fehler zu beheben, müssten für diese Fälle die ESTIMATE-Statements manuell an die tatsächlich aufgetretenen Kategorien angepasst werden und in einem nachfolgenden DATA STEP die korrekte Zuordnung der Wahrscheinlichkeiten zu den richtigen Kategorien erfolgen. Im Sinne der oben beschriebenen Anwendungen erfordert dies einen hohen Zeitaufwand und entsprechende Kenntnisse. Die in diesem Beitrag mit Hilfe der IML Prozedur beschriebene Vorgehensweise zeigt eine Möglichkeit, die Berechnung der Wahrscheinlichkeiten innerhalb der Kategorien unabhängig von eventuell auftretenden leeren Kategorien zu schätzen und stellt damit ein wertvolles Hilfsmittel zur automatisierten Auswertung kategorialer Daten dar.
Für die Auswertung von Simulationen weist die beschriebene Methode weitere Vorteile auf. Sie bietet die Möglichkeit, innerhalb der IML Prozedur die Berechnung von Konfidenzintervallen für die kumulativen Wahrscheinlichkeiten sowie die Berechnung des MSE (Mean Squared Error) anzuweisen. Ebenso ist die Berechnung der Wahrscheinlichkeiten je Kategorie auf diese Weise möglich. Natürlich sind zur Lösung der hier beschriebenen Problematik auch alternative Herangehensweisen, wie die Nutzung eines Makros zur Erstellung der ESTIMATE-Statements, denkbar. Die Entscheidung zur Nutzung von IML lag in unserem Fall auch darin begründet, dass weiterführende Informationen, wie beispielsweise Eigenwerte oder Definitheit der Kovarianzmatrix der festen Effekte, berechnet bzw. geprüft werden können. Die Erzeugung von Grafiken zur Veranschaulichung von Behandlungsdifferenzen je Kategorie liefert für signifikante Behandlungsunterschiede die zusätzliche Information darüber, in welcher Kategorie die Unterschiede zu finden sind. Werden Behandlungsdifferenzen mittels ESTIMATE-Statement auf Signifikanz geprüft, kann hierüber keine Aussage getroffen werden, da als Ergebnis nur der Signifikanztest im linearen Prädiktor, also eine Aussage über alle Kategorien hinweg, erhalten wird.

Literatur

[1] McCulloch, C.E.; Searle, S.R.: Generalized, Linear and Mixed Models, New York, NY: John Wiley & Sons, Inc. (2001)