Ein SAS®-Makro zur Schätzung des Stereotype Regressionsmodells

Oliver Kuß

Abt. Klinische Sozialmedizin, Universitätsklinikum Heidelberg
Bergheimer Str. 58, 69115 Heidelberg
eMail: Oliver_Kuss@med.uni-heidelberg.de

Abstract

Im Fahrwasser der logistischen Regression, einem inzwischen akzeptierten und häufig angewandten Verfahren in der medizinischen Statistik, entwickeln sich auch Regressionsmodelle für andere kategorielle, aber nicht-binäre Zielgrößen wie das Proportional Odds Modell oder das multinomiale logistische Regressionsmodell prächtig. Ein kümmerliches Dasein fristet dagegen immer noch das Stereotype Modell, eine Sonderform des multinomialen logistischen Modells, die mit weniger Parametern auskommt und die Ordinalität der Zielgröße mit Hilfe der Kovariablen modelliert. Hauptgrund dafür dürfte bisher die Nichtverfügbarkeit von adäquater Software gewesen sein. Deshalb bleibt es in der Literatur auch oft bei der bloßen Erwähnung des Modells oder bei Vorschlägen zur informellen Beurteilung von herkömmlichen multinomialen Modellen.
In diesem Beitrag soll das Stereotype Modell kurz vorgestellt, zu den anderen logistischen Modellen abgegrenzt und seine Berechnung mit Hilfe eines SAS®-Makros beschrieben werden. In diesem Makro wird das dem Stereotype Modell zugrundeliegende multinomiale logistische Modell als bedingtes logistisches Modell aufgefaßt und so einer Schätzung mit PROC PHREG zugänglich. Die dazu notwendige alternierende Parameterschätzung wird mit PROC IML realisiert. Zur Illustration dient ein Datensatz zu berufsbedingten Handekzemen bei Auszubildenden in der Automobilindustrie.

Ausgangspunkt

Ausgangspunkt zur Beschäftigung mit dem Stereotype Modell war eine Studie zur Beurteilung von exogenen (v.a. Arbeitsbelastungen) und endogenen (genetische Disposition) Risikofaktoren bei der Entstehung von berufsbedingten Handekzemen in der Automobilindustrie. Dazu wurden zwischen 1990 und 1998 alle 2078 Auszubildenden der AUDI AG, die ihre Ausbildung in Ingolstadt 1990-1994 und Neckarsulm 1991-1994 begonnen hatten, im Rahmen einer prospektiven Kohortenstudie standardisiert untersucht. Untersuchungen erfolgten zu Beginn, nach dem ersten Ausbildungsjahr und am Ende der dreijährigen Ausbildung. Durchgeführt wurde die Studie mit Unterstützung des Bundesministeriums für Bildung und Forschung in Zusammenarbeit von Dr. Ulrich Funke vom Bereich Gesundsheitswesen der AUDI AG und Prof. Dr. Thomas Diepgen und Prof. Dr. Manigé Fartasch von der Dermatologischen Universitätsklinik Erlangen.
Die Zielgröße, das Auftreten eines berufsbedingten Handekzems im Verlauf der Ausbildung, war auf einer dreistufigen ordinalen Skala (kein, leicht, schwer) erhoben worden. Zum Abschluß der Studie lagen die Daten von 1910 Auszubildenden vor. In 167 Fällen war dabei ein schweres und in 103 Fällen ein leichtes Handekzem beobachtet worden, 1640 Auszubildende waren beschwerdefrei geblieben.

Statistische Modelle

Die Relevanz der verschiedenen Risikofaktoren sollte mit Hilfe eines Regressionsmodells beurteilt werden. Dies stellte kein größeres Problem dar, da Regressionsmodelle für kategorielle Zielgrößen in den letzten 20 Jahren sowohl methodisch als auch bezüglich der Umsetzung mit Standardsoftware einem stürmischen Entwicklungsprozess unterworfen waren und inzwischen zur allgemeinen Verfügung stehen.

Logistische Regression

Der bekannteste Vertreter der Familie der kategoriellen Regressionsmodelle ist das logistische Regressionsmodell. Dieses schätzt Modelle mit binären Zielgrößen durch die Modellgleichung

$$\log\left(p \big/ 1-p\right) = \alpha + \beta' x ,$$

wobei $p = P(Y = 1 \mid x)$ die bedingte Wahrscheinlichkeit für das Eintreten des Zielereignisses Y, gegeben die Kovariablen x, ist. Der Parameter α bildet einen konstanten Term ab, der unabhängig von den Kovariablen ist, die Parameter β messen den Einfluß der Kovariablen x. Das logistische Regressionsmodell ist zu einem Standardwerkzeug in der medizinischen Statistik geworden, wenn es um die Auswertung von binären Zielgrößen geht. Dies liegt vor allem an der leichten Interpretierbarkeit der geschätzten Parameter und an der Möglichkeit, sowohl prospektive als auch retrospektive Beobachtungsstudien damit auszuwerten.
Die Schätzung von logistischen Regressionsmodellen mit SAS® ist unproblematisch, dafür steht eine Vielzahl von Prozeduren zur Verfügung (vgl. Kuß, 1999).
Im vorliegenden Fall ist das logistische Regressionsmodell jedoch weniger geeignet, da damit eine Dichotomisierung der dreistufigen Zielgröße verbunden wäre und dies einen Verlust an Information über den Schweregrad mit sich bringen würde.

Proportional Odds Modell

Eine Möglichkeit zur Analyse des vorliegenden Datensatzes unter Beibehaltung des ordinalen Skalenniveaus der Zielgröße bietet das Proportional Odds Modell. Es kann als natürliche Verallgemeinerung des logistischen Modells gesehen werden. Der Unterschied liegt darin, dass jetzt eine kumulierte Wahrscheinlichkeit $p_j = P(Y \le j \mid x),\ j = 1,\ldots, J\text{-}1$ in Abhängigkeit von den Kovariablen dargestellt wird, wobei J die Anzahl der Ausprägungen der Zielgröße ist:

$$\log\left(p_j \big/ 1-p_j\right) = \alpha_j + \beta' x$$

Im Prinzip liegen also J-1 Modellgleichungen vor, wobei mittels der Definition der p_j in jeder Gleichung eine dichotomisierte Zielgröße erzeugt wird, die wie in einem herkömmlichen logistischen Modell modelliert wird. Dadurch ergeben sich in der Modellgleichung des Proportional Odds Modells J-1 konstante Terme α_j. Die Parameter β messen auch hier den Einfluß der Kovariablen, wobei dieser jedoch als unabhängig von j angenommen wird, d.h. als konstant über alle J-1 möglichen Dichotomisierungen der Zielgröße.
Zur Schätzung der Parameter des Proportional Odds Modells stellt SAS® die Prozeduren PROC LOGISTIC und PROC PROBIT zur Verfügung, auch eine Schätzung mit PROC CATMOD ist möglich.

Das Proportional Odds Modell ist unter der Annahme einer zugrundeliegenden stetigen Zielgröße hergeleitet, die lediglich klassifiziert beobachtet werden kann. Die Parameterschätzer aus dem Proportional Odds Modell sind dabei asymptotisch äquivalent zu denen, die man aus einer linearen Regression erhalten hätte, wenn die zugrundeliegende stetige Zielgröße direkt zu beobachten gewesen wäre. Dies macht das Modell für unsere Zwecke aber weniger geeignet, da ein kontinuierlicher Schweregrad für Handekzeme klinisch keinen Sinn macht. Ein Handekzem muss vielmehr als ein multidimensionales Phänomen angesehen werden, wobei mögliche Dimensionen die Ausdehnung, die Intensität (Rötung, Nässen, Austrocknung), die Häufigkeit des Auftretens, aber auch subjektive Symptome wie Juckreiz oder Schlaflosigkeit sind.

Multinomiale logistische Regression

Eine letzte Möglichkeit ist schließlich, die Ordinalität der Zielgröße zu vernachlässigen und diese als nominal skaliert anzunehmen. Für diese Zwecke steht eine weitere Verallgemeinerung des logistischen Regressionsmodells zur Verfügung, das multinomiale logistische Regressionsmodell. Die Modellgleichung für dieses Modell lautet

$$\log\left(p_j \Big/ p_1 \right) = \alpha_j + \beta_j' x \ ,$$

mit $p_j = P(Y = j \mid x)$, $j = 2,..., J$ und $\alpha_1 = 0$, $\beta_1 = 0$.
Auch hier liegt wieder ein System von J-1 Modellgleichungen vor, wobei in jeder dieser Gleichungen die Wahrscheinlichkeit p_j (in diesem Falle, die Wahrscheinlichkeit, dass die Zielgröße die Ausprägung j hat), in Beziehung zu einer Referenzwahrscheinlichkeit p_1 gesetzt wird. Im multinomialen Modell wird für jede dieser Modellgleichungen ein eigener konstanter Term α_j und eine eigene Menge von Parametern β_j geschätzt, die Parameter für die Referenzkategorie werden aus Identifizierbarkeitsgründen auf Null gesetzt. Die Interpretation der Parameter muß nicht notwendigerweise im Vergleich zur Referenzkategorie durchgeführt werden, es können auch Odds für den Vergleich von zwei beliebigen Katgorien j und j' berechnet werden.
Die Schätzung von multinomialen logistischen Modellen erlaubt SAS® mit Hilfe von PROC CATMOD (generalisiertes Modell) und PROC PHREG (Discrete Choice Modell).
Auch das multinomiale Modell erscheint in unserem Fall als wenig geeignete Alternative. Zum einen wird man nur ungern die Ordinalität der Zielgröße unberücksichtigt lassen, zum anderen wird eine große Anzahl von Parametern geschätzt, was in unserem Fall angesichts der großen Fallzahl weniger relevant ist, aber bei kleinen Datensätzen, bei großen J oder bei einer großen Anzahl von Kovariablen durchaus zum Problem werden kann. Desweiteren ist zu erwarten, dass in den Parametern ein großer Anteil an redundanter Information enthalten ist.

Stereotype Regression

Das Stereotype Modell wurde von Anderson, 1984 vorgeschlagen. Bei der Namensgebung bezog er sich auf einen Arzt, der die Schwere einer Krankheit beurteilen muss und dabei nie nur einen isolierten Aspekt der Krankheit berücksichtigt, in unserem Beispiel etwa die Ausdehnung des Handekzems, sondern von jedem Schweregrad der Krankheit einen *Stereotyp* hat, in dem die verschiedenen Aspekte der Krankheit zusammengefasst werden.

Anderson geht sogar noch einen Schritt weiter und schlägt ein neues Skalenniveau vor, um multidimensionale Zielgrößen zu beschreiben. Er spricht dabei von "assessed responses", also beurteilten Zielgrößen, die von exakt gemessenen zu unterscheiden sind.
Das Stereotype Modell liefert letztendlich die Lösung für die Probleme mit den oben beschriebenen kategoriellen Regressionsmodellen in unserem Anwendungsbeispiel. Es baut auf dem multinomialen Modell auf, berücksichtigt aber trotzdem noch die Ordinalität der Zielgröße. Zusätzlich wird auch noch die Anzahl der geschätzten Parameter im Vergleich zum multinomialen Modell reduziert.
Diese Reduktion der Parameterzahl erreicht man durch Einführung von neuen Parametern θ_j mittels

$$\beta_j = \theta_j \beta .$$

Durch Einsetzen in die Modellgleichung des multinomialen logistischen Modells erhält man die Modellgleichung für das Stereotype Modell zu:

$$\log\left(p_j \Big/ p_1 \right) = \alpha_j + \theta_j \beta' x$$

wobei $p_j = P(Y = j \mid x)$, $j = 2,..., J$, $\alpha_1 = 0$, $\theta_1 = 0$ und $\theta_J = 1$.
Die Wahrscheinlichkeit p_j für das Eintreten des Zielereignisses wird hier wieder im Vergleich zu einer Referenzkategorie modelliert, auch die konstanten Terme α_j und die Parameter β kennen wir aus den oben beschriebenen Regressionsmodellen. Die β werden hier wie im Proportional Odds Modell als konstant über alle Kategorien der Zielgröße angenommen.
Den entscheidenden Unterschied machen jedoch die Parameter θ_j aus. Betrachten wir den Vergleich zweier Erfolgswahrscheinlichkeiten p_j und $p_{j'}$ durch

$$\log\left(p_j \Big/ p_{j'} \right) = \alpha_j - \alpha_{j'} + (\theta_j - \theta_{j'}) \beta' x ,$$

so sehen wir, dass das Odds für eine Beobachtung, in Zielkategorie j anstatt in Zielkategorie j' zu fallen, nur dann stark von den Kovariablen beeinflußt wird, wenn θ_j und $\theta_{j'}$ verschieden sind. Je größer dabei die Differenz der θ_j, desto größer auch der Beitrag der Kovariablen zum Unterschied zwischen p_j und $p_{j'}$.
Liegt nun eine Ordinalität der Zielgröße vor, wie in unserem Beispiel, so wird man erwarten, dass bei einem Durchlauf von 2 bis J über die Kategorien der Zielgröße der Einfluss der Kovariablen auf die Zielgröße immer größer wird. Das heißt aber bei (über die j) konstantem β, dass auch die Differenzen θ_j-θ_1 immer größer werden und sich dadurch eine Anordnung der θ_j ($\theta_1 \leq ... \leq \theta_j$) ergibt. Diese Anordnung bildet dabei gerade die Ordinalität der Zielgröße ab.
Das Reizvolle am Stereotype Modell ist, dass diese Ordinalität der Zielgröße nicht vorausgesetzt wird wie im Proportional Odds Modell, sondern mittels der θ_j geschätzt wird und deshalb auch getestet werden kann. Desweiteren sind auch Tests auf Ununterscheidbarkeit der θ_j möglich, wodurch entschieden werden kann, ob eventuell Kategorien der Zielgröße zusammengefasst werden sollten.
Zu beachten ist ferner, dass die Ordinalität der Zielgröße im Stereotype Modell bezüglich der Kovariablen abgebildet wird und nicht bezüglich einer zugrundeliegenden stetigen Zielgröße wie im Proportional Odds Modell.

Schätzung des Stereotype Modells

Ein großes Problem im Zusammenhang mit dem Stereotype Modell war in der Vergangenheit, dass kein routinemäßig anwendbares Makro oder Beispielprogramm in einem Standard-Software-Paket vorlag, um die notwendigen Schätzungen durchzuführen.
In der Literatur findet man eine Fülle von Einzellösungen (Anderson, 1984, DiPrete, 1990) oder aber Vorschläge, herkömmliche multinomiale Modelle informell auf die Eigenschaften des Stereotype Modells zu prüfen (Greenwood/Farewell, 1988). Im Übersichtsartikel über ordinale Regressionsmodelle von Ananth/Kleinbaum, 1997 wird das Stereotyp Modell zwar erwähnt, aber nicht geschätzt, weil keine adäquate Software vorlag.
Abhilfe schaffen die SAS®-Makros von John Hendrickx (vgl. auch Hendrickx/Ganzeboom, 1998), die auf dessen Homepage (http://baserv.uci.kun.nl/~johnh/mcl/) mitsamt ausführlichen Beschreibungen erhältlich sind. Die Makros leisten im Prinzip, ausgehend vom Arbeitsgebiet von J. Hendrickx, der Mobilitätsanalyse, sogar einiges mehr und sind in der Lage, noch komplexere Modelle als das Stereotype Modell zu schätzen.
Das Makro beruht statistisch auf der Tatsache, dass das multinomiale logistische Modell als konditionales logistisches Modell aufgefasst werden kann und zwar deshalb, weil die Likelihoodfunktion im multinomialen logistischen Modell äquivalent zur partiellen Likelihoodfunktion von Breslow im Proportional-Hazard-Modell ist. Dadurch wird das multinomiale Modell einer Schätzung mit PROC PHREG zugänglich. Die für die Darstellung als konditionales Modell notwendige Umordnung des Datensatzes wird mit Hilfe des Makros %MCLGEN vorgenommen.
Da im Stereotype Modell nicht-lineare Einschränkungen bezüglich der Parameter gemacht werden und diese multiplikativ ins Modell eingehen, müssen θ und β iterativ alternierend geschätzt werden. Das heißt konkret, man startet mit einem sinnvollen Startwert für die β, nimmt diese dann als fest und nicht-zufällig an und schätzt mit Hilfe von PROC PHREG die θ. Im nächsten Schritt nimmt man diese geschätzen θ als gegeben an und berechnet in einem neuen Durchlauf von PROC PHREG neue verbesserte Schätzer für die β. Mit diesen findet man dann wieder neue Schätzer für die θ usw. Dieser Iterationsprozess wird bis zur Konvergenz fortgeführt. Die Übergabe der Parameter und die wiederholten Aufrufe von PROC PHREG werden im Makro %MCLEST mit PROC IML realisiert, ein Startwert wird ebenfalls automatisch ermittelt.
Ein Problem dieser Methode der Schätzung ist, dass die geschätzten Standardfehler der Parameter und damit Konfidenzintervalle und Parametertests nicht gültig sind. Man muss sich in diesem Fall mit einer α-Adjustierung oder mit bootstrap-korrigierten Konfidenzintervallen behelfen.

Ergebnisse zur Beispielstudie

Zur Illustration wird ein Modell mit drei binären Kovariablen für den Beispieldatensatz berechnet. Bei den Kovariablen handelt es sich um zwei endogene Risikofaktoren, das Vorliegen eines anamnestischen Handekzems bzw. eines Beugenekzems, das heißt also von Ekzemen, die bereits vor der Ausbildung beobachtet worden waren und einen exogenen Risikofaktor, die Dauer der täglich am Arbeitsplatz verrichteten Feuchtarbeit (</>= 3h).
Es ergeben sich folgende Ergebnisse für das multinomiale Modell und das Sterotype Modell:

	Multinomiales Modell		Stereotype Modell
	OR (95%-CI) kein/leicht	OR (95%-CI) kein/schwer	OR (95%*-CI)
Feuchtarbeit	2.42 (1.47, 3.98)	2.31 (1.54, 3.47)	2.33 (1.67, 3.24)
Anamn. HE	6.06 (2.78, 13.23)	5.41(2.75, 10.65)	5.55 (3.11, 9.90)
Anamn. BE	3.25 (1.48, 7.11)	3.42 (1.79, 6.54)	3.33 (1.90, 5.74)
θ_2	.	.	1.03 (0.75,1.24)

Betrachten wir zunächst das multinomiale Modell. Als Referenzkategorie wurde die Kategorie „kein HE" gewählt. Für alle drei Kovariablen gilt analog: Bei Vorliegen des Risikofaktors ist das Odds Ratio, ein berufsbedingtes Handekzem zu erleiden, signifikant erhöht. Allerdings ist kein Unterschied in den Odds Ratios bezüglich des Schweregrad dieses Handekzems zu sehen, die Odds Ratios sind für die Kategorien "leichtes HE" und "schweres HE" nahezu identisch. Die Kovariablen haben also keinerlei prognostischen Wert, was den zu erwartenden Schweregrad des berufsbedingten Handekzems angeht.
Das Stereotyp Modell fasst diesen Sachverhalt in eleganter Art und Weise zusammen. Die geschätzten Odds Ratios vermitteln zwischen denen aus dem multinomialen Modell. Die Konfidenzintervalle für die Odds Ratios sind enger, aber aus oben genannten Gründen wohl zu optimistisch und deshalb mit einem *) gekennzeichnet. Die Tatsache, dass sich keine echte Ordinalität in der Zielgröße bezüglich der Kovariablen zeigt und die Odds Ratios für ein leichtes HE sich kaum von denen für ein schweres HE unterscheiden, wird vom Parameter θ_2 abgebildet. Dieser ist mit 1.03 praktisch gleich θ_3, das aus Identifizierbarkeitsgründen im Modell auf 1 gesetzt wird.

Fazit/Ausblick

Das Stereotype Modell stellt eine nützliche Alternative zu den bekannten Regressionsmodellen mit mehrstufigen Zielgrößen dar. Besonders attraktiv im medizinischen Kontext erscheint die dem Modell zugrundeliegende Annahme, dass das beobachtete Zielereignis keine exakt gemessene singuläre Größe ist, sondern auf einer Bewertung und Gewichtung mehrerer Aspekte beruht. Dies bildet den Prozeß der Beurteilung des Schweregrads einer Krankheit ungleich realistischer ab. Greenland, 1994 liefert ein weiteres inhaltliches Argument für das Stereotype Modell: Er zeigt, dass der Prozess des Durchlaufens von aufeinanderfolgenden Stadien einer Krankheit (auch dies kann in unserem Anwendungsbeispiel angenommen werden) in ganz natürlicher Weise mit einem Stereotyp Modell modelliert werden kann.
Die dargestellten statistischen Probleme (vgl. auch Holtbrügge/Schumacher, 1988) müssen ernst genommen werden. Sie sollten jedoch nicht den Anlass liefern, die methodische Auseinandersetzung mit dem Modell zu beenden, sondern angesichts der Eigenschaften des Modells eher ein Anstoß für zukünftige Arbeit sein.
Am Schluß soll ein Zitat von Greenland, 1994 stehen:

"Nonetheless, the stereotype model provides a useful alternative to other models [...], and thus deserves equal coverage in general discussions of ordinal regression....."

und ich möchte ergänzen:

".... vor allem jetzt, wo auch allgemein zugängliche Software verfügbar ist."

Literatur

Ananth CV, Kleinbaum DG. Regression Models for Ordinal Responses: A Review of Methods and Applications. *Int J Epidemiol,* 26, 1323-1333, 1997.

Anderson, JA. Regression and Ordered Categorical Variables. *J R Statist Soc B*, 46, 1-30, 1984.

DiPrete T. Adding Covariates to Loglinear Models for the Study of Social Mobility. *Am Sociol Rev*, 55, 757-773, 1990.

Greenland S. Alternative Models for Ordinal Logistic Regression. *Stat Med*, 13, 1665-1677, 1994.

Greenwood C, Farewell V. A Comparison of Regression Models for Ordinal Data in an Analysis of Transplanted-Kidney Function. *Can J Stat*, 16, 325-335, 1988.

Hendrickx J, Ganzeboom HBG. Occupational Status Attainment in the Netherlands, 1920-1990. A Multinomial Logistic Analysis. *Europ Sociol Rev*, 14, 387-403, 1998.

Holtbrügge W, Schumacher M. A Comparison of Regression Models for the Analysis of Ordered Categorical Data. *Appl Statist*, 40, 249-259, 1991.

Kuß O. Logistische Regression in SAS®. In: Ortseifen C (Ed.). Proceedings der 3. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), 25./26. Februar 1999, Ruprecht-Karls-Universität Heidelberg, 147-154, 1999.