

Modellbildung und Berechnung von Partial Proportional Odds Modellen mit PROC CATMOD

Martina Kron¹, Andrea Ochsmann², Michael Friedrichs²

¹ Abt. Biometrie und Medizinische Dokumentation, Universität Ulm
89075 Ulm
Tel. 0731/50-26905
E-Mail: Martina.kron@medizin.uni-ulm.de

² Integrationsprogramm Arbeit und Gesundheit (IPAG), Essen

1. Einleitung

Soll der Einfluss von Kovariablen auf eine polytome ordinale Zielgröße modelliert werden, so wird versucht, logistische Regressionsmodelle anzupassen. Dabei können mit PROC LOGISTIC leicht Proportional Odds Modelle (POM) [1] und mit PROC CATMOD generalisierte polytome logistische Modelle berechnet werden. Häufig ist jedoch die Proportional Odds Annahme nur für einen Teil der Einflussgrößen erfüllt. Wird dann ein generalisiertes Modell angepasst, so sind in diesem Modell mehr Parameter als nötig zu schätzen. Darüber hinaus sind die geschätzten Parameter häufig instabil, nicht alle Parameter schätzbar und die Ergebnisse schlecht interpretierbar. Daraus ergibt sich der Wunsch, Partial Proportional Odds Modelle (PPOM) [2] zu berechnen. Dafür muss jedoch vorab geklärt werden, welche Odds proportional sind.

2. Modellbildung

Bei einer ordinalen Zielgröße werden in einem POM kumulative Logit-Funktionen verwendet, die auf den kumulierten Wahrscheinlichkeiten basieren. Für eine 4-stufige Zielgröße werden damit 3 kumulative Logit-Funktionen benutzt:

$$\text{logit}(\theta_1) = \log \frac{\pi_2 + \pi_3 + \pi_4}{\pi_1}, \quad \text{logit}(\theta_2) = \log \frac{\pi_3 + \pi_4}{\pi_1 + \pi_2}, \quad \text{logit}(\theta_3) = \log \frac{\pi_4}{\pi_1 + \pi_2 + \pi_3}$$

Für diese 3 Logit-Funktionen werden im POM verschiedene Regressionskonstanten β_{01} , β_{02} und β_{03} , aber die gleichen Regressionsparameter $(\beta_1, \dots, \beta_p)$ angepasst, da davon ausgegangen wird, dass die Logit-Funktionen parallel verlaufen, d.h. die Odds für alle p Einflussgrößen proportional sind. Somit werden bei einer k -stufigen Zielgröße $(k-1)+p$ Modellparameter geschätzt. Sind die Odds von einer Einflussgröße für mindestens zwei Logit-Funktionen nicht proportional, so werden weitere $(k-2)$ Regressionsparameter modelliert und geschätzt. Vereinfachend soll im Macro nicht weiter geprüft werden, ob für einzelne Logit-Funktionen die Proportional Odds Annahme doch erfüllt ist. Die Annahme von Proportional Odds über alle Logit-Funktionen kann für jede einzelne Einflussgröße j ($j=1, \dots, p$) überprüft werden, indem

$$H_0 : (\beta_{j,1}, \dots, \beta_{j,k-2})' = (\beta_{j,2}, \dots, \beta_{j,k-1})'$$

im allgemeinen logistischen Modell mit kumulativen Logit-Funktionen (ohne Proportional Odds Annahme) getestet wird. Bei diesem Test werden simultan alle Parameter, die zur Einflussgröße j gehören, zwischen den $(k-1)$ kumulativen Logit-Funktionen verglichen.

Bei einer k -stufigen Zielgröße werden damit in einem Modell ohne Proportional Odds Annahme $(k-1)$ kumulative Logit-Funktionen und die zugehörige Designmatrix mit $(k-1)(p+1)$ Modellparametern erstellt. Nach Schätzen der Modellparameter kann getestet werden, für welche Einflussgrößen die Proportional Odds Annahme erfüllt ist. Über die Anzahl von Einflussgrößen mit und ohne Proportional Odds Annahme bestimmt sich die Anzahl der Modellparameter und damit die Gestalt der zugehörigen Designmatrix für das PPOM.

3. SAS-Macro

Das oben beschriebene Vorgehen bei der Modellbildung wird in einem SAS-Macro umgesetzt, welches dazu die SAS-Prozedur PROC CATMOD benutzt. Folgende Macro-Variablen müssen dazu an das Macro übergeben werden:

```
%macro ppom(level,alpha1,alpha2,data,y,yformat,beta1,beta2,beta3,...);
```

level	= Anzahl der Stufen von Ziel- und Einflussgrößen
alpha1	= Signifikanzniveau für den Test auf Proportional Odds
alpha2	= Signifikanzniveau für die Tests auf Partial Proportional Odds
data	= SAS-Dataset
y	= Zielgröße
yformat	= Formatzuweisung für die Zielgröße, um gewünschte Richtung des Effekts zu erzielen
beta1,beta2,beta3,...	= Einflussgrößen (für kategorielle Variablen werden automatisch Dummy-Variablen kodiert, wobei die kleinste Ausprägung als Referenz gewählt wird und für jede Dummy-Variable die Proportional Odds Annahme überprüft wird)

Beispiel: Ermittlung arbeitsbedingter Gesundheitsgefahren

Aus den Testdatensätzen einer Betriebskrankenkasse soll der Einfluss der kognitiven Arbeitsbelastung (1=gering, 2=mittel, 3=hoch) und des Alters (1=<30, 2=30-50, 3=>50 Jahre) auf die Einnahmedauer eines Medikaments in Tagen bestimmt werden. Es handelt sich um eine 4-stufige ordinale Zielgröße (0, 1-42, 43-120, >120 Tage) und zwei 3-stufige Einflussgrößen. Zur Anpassung eines POM oder eines PPOM wird zunächst ein generalisiertes logistisches Regressionsmodell mit kumulativen Logit-Funktionen angepasst. Dazu wird das Macro wie folgt aufgerufen.

```
***Macro-Aufruf***
```

```
%ppom(4 3 3, 0.1, 0.1, data=dat.test5_6M, y=ddd_kl, yformat=ddd_kl., beta1=indkog, beta2=altgrup);
```

```
***Auszug aus dem Output***
```

```
***Überprüfen der Proportional Odds Annahme***
```

Analysis of Contrasts	Contrast	DF	Chi-Square	Pr > ChiSq
	Test auf PO	8	14.05	0.0805
	Test: beta1_1:	2	0.89	0.6407
	Test: beta1_2:	2	0.19	0.9085
	Test: beta2_1:	2	5.33	0.0697
	Test: beta2_2:	2	10.20	0.0061

Test auf PO gibt den p-Wert zum Test über die Proportional Odds Annahme für alle Einflussgrößen ($p=0.0805$), die offensichtlich nicht erfüllt ist. Bei Betrachtung der Proportional Odds Annahme für jeweils 2 Stufen einer jeden Einflussgröße zeigt sich, dass diese für beide Dummy-Variablen, die zur Einflussgröße „kognitive Arbeitsbelastung“ gehören, erfüllt und für beide Dummy-Variablen, die zur Einflussgröße „Alter“ gehören, nicht erfüllt sind. Beispielsweise beträgt der p-Wert zur Proportional Odds Annahme 0.0697 für den Vergleich von Altersgruppe 2 mit 1 und 0.0061 für den Vergleich von Altersgruppe 3 mit 1.

Resultierendes PPOM

Response Profiles		Response DDD_KL		

		1	keine	
		2	1-42	
		3	43-120	
		4	über 120	
Source	ProbChiSq	OR	KI95_U	KI95_O

alpha1	<.0001	0.11704	0.06488	0.2111
alpha2	0.0352	0.67454	0.46760	0.9731
alpha3	<.0001	2.78022	1.83314	4.2166
beta1_1	0.0800	1.32887	0.96656	1.8270
beta1_2	0.4198	1.17350	0.79559	1.7309
beta2_1_1	0.1038	1.46573	0.92468	2.3234
beta2_1_2	0.1505	1.32441	0.90295	1.9426
beta2_1_3	0.0011	2.73875	1.49764	5.0084
beta2_2_1	0.4823	1.28112	0.64185	2.5571
beta2_2_2	0.0205	1.95031	1.10841	3.4317
beta2_2_3	<.0001	5.16506	2.51684	10.5997

Die geschätzten Odds Ratios mit zugehörigen 95% Konfidenzintervallen in der obigen Tabelle sind wie folgt zu interpretieren, z.B. beschreibt

- alpha1 das Grundrisiko für die 1. Logit-Funktion $\text{logit}(\theta_1) = \ln \frac{\pi_2 + \pi_3 + \pi_4}{\pi_1}$, d.h. für die Personen mit mindestens 1 Tag Einnahme des Medikaments (Response 2, 3, oder 4) gegenüber keiner Einnahme (Response 1),
- beta1_1 den Vergleich von mittlerer zu geringer kognitiver Arbeitsbelastung in allen 3 Logit-Funktionen,
- beta1_2 den Vergleich von hoher zu geringer kognitiver Arbeitsbelastung in allen 3 Logit-Funktionen,
- beta2_1_2 den 1. Vergleich zur Altersklassenreferenz „<30 Jahre“ (Effekt von Altersklasse 2 zu 1) in der 2. Logit-Funktion und
- beta2_2_3 den 2. Vergleich zur Altersklassenreferenz (Effekt von Altersklasse 3 zu 1) in der 3. Logit-Funktion.

4. Bemerkungen

- Falls einzelne Kombinationen von Ausprägungen der Einflussgrößen zu gering besetzt sind, kann kein Modell angepasst werden.
- Alle Modellparameter werden durch PROC CATMOD im Macro mit Weighted Least Squares und nicht wie in PROC LOGISTIC mit Maximum Likelihood geschätzt.
- Falls die Proportional Odds Annahme erfüllt ist, wird im Macro automatisch ein POM gerechnet. Falls die Proportional Odds Annahme auch partiell für keine Einflussgröße (bzw. keine Dummy-Variable) erfüllt ist, wird im Macro ein generalisiertes ordinales logistisches Modell gerechnet.
- Das Macro läuft erst unter SAS-Version 8.1 und kann abgerufen werden unter www.uni-ulm.de/uni/fak/medizin/biodok/abteilung/ppom_macro.htm

Literatur

- [1] McCullagh P.: Regression models for ordinal data (with discussion). J R Stat Soc B 42:109-142 (1984)
- [2] Peterson B., Harrell F.E.: Partial proportional odds models for ordinal response variables. Appl Statist 39:205-217 (1990)