Modellierung ordinaler Präferenzen

mit {RprobitB} am Beispiel des pairfam Panels

Lennart Oelschläger Dietmar Bauer

Universität Bielefeld, Fakultät für Wirtschaftswissenschaften, Lehrstuhl für Ökonometrie

10. November 2022

Worum geht's?

1. Ordered Probit Modell

2. Anwendung am pairfam Panel

3. Computerwerkstatt: {RprobitB}

Worum geht's?

1. Ordered Probit Modell
2. Anwendung am pairfam Panel
3. Computerwerkstatt: {RprobitB}

Motivation

Antwortskalen bei Umfragen

- Kardinalskala: Klassische lineare Regression
  - z.B. Modellierung des Einkommens
- Nominalskala: Wahlmodelle
  - z.B. Modellierung der Wahl eines Verhütungsmittels (Vortrag letztes Jahr)
- Ordinalskala: **Ordered Probit Modell**
  - z.B. Modellierung von Antworten auf einer Likert-Skala

Definition

Gegeben eine Frage mit J geordneten Antwortmöglichkeiten.

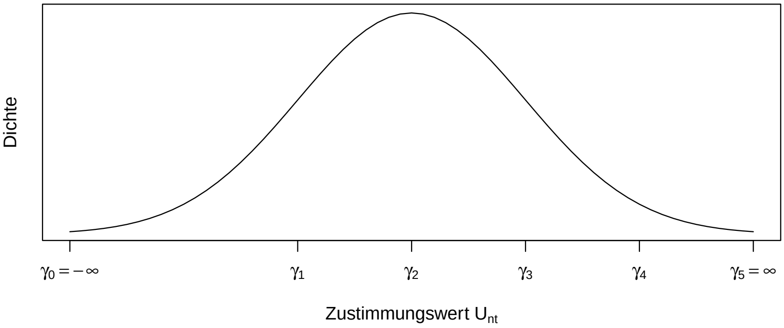
Modelliere für Person $n = 1, \dots, N$ bei der Wahlsituation $t = 1, \dots, T(n)$:

$$U_{nt} = X_{nt}' \beta + \epsilon_{nt}$$

- $U_{nt} \in \mathbb{R}$ ist der stetige Zustimmungswert von n bei t
- $X_{nt} \in \mathbb{R}^P$ ist der Vektor mit P Charakteristiken von n bei t
- $\beta \in \mathbb{R}^P$ ist der zu schätzende Koeffizientenvektor
- $\epsilon_{nt} \in \mathbb{R}$ ist der Fehlerterm für n bei t mit $\epsilon_{nt} \sim N(\mu, \sigma)$ (Probit Modellklasse)

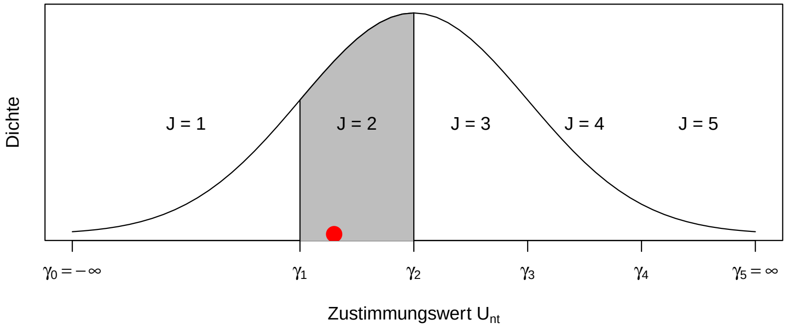
Kategorisierung der Zustimmung

In diesem Beispiel: $J = 5$ Antwortmöglichkeiten



Kategorisierung der Zustimmung

In diesem Beispiel: $J = 5$ Antwortmöglichkeiten



Zustimmungswert und Antwort

Es bezeichne $y_{nt} \in \{1, \dots, J\}$ die Antwort von n bei t. Dann:

$$y_{nt} = \sum_j j \cdot 1\{\gamma_{j-1} < U_{nt} \leq \gamma_j\}$$

mit $\gamma_0 = -\infty$ und $\gamma_J = +\infty$.

Das heißt, Antwortalternative j wird genau dann gewählt, wenn $U_{nt} \in \ (\gamma_{j-1}, \gamma_j]$.

Monotonie der Grenzen: Anstelle von γ_j, $j = 1, \dots, J-1$, schätze d_j mit

$$\gamma_j = \sum_{i \leq j} \exp(d_i).$$

Normalisierung

Modellgleichung für n bei t:

$$U_{nt} = X'_{nt}\beta + \epsilon_{nt}$$

Die Zustimmungswerte U_{nt} sind invariant bezüglich

- Skala: $U_{nt} \in \ (\gamma_{j-1},\ \gamma_j] \Leftrightarrow c \cdot U_{nt} \in \ (c \cdot \gamma_{j-1},\ c \cdot \gamma_j]$ für alle $c \in \mathbb{R}_+$
  - Lösung: Fixiere zum Beispiel $\sigma = 1$ oder $\mu = 0$
- Level: $U_{nt} \in \ (\gamma_{j-1},\ \gamma_j] \Leftrightarrow c + U_{nt} \in \ (c + \gamma_{j-1},\ c + \gamma_j]$ für alle $c \in \mathbb{R}$
  - Lösung: Fixiere zum Beispiel $\gamma_1 = 0$

Heterogenität und latente Klassen

Anstelle eines konstanten β für jeden Befragten $n = 1, \dots, N$:

$$\beta_n \sim f$$

Die Verteilung f ist eine Modellspezifikation, z.B.

- ein Produkt unabhängiger Marginalverteilungen (z.B. Normalverteilung, Log-Normal Verteilung)
- eine multivariate Verteilung (erlaubt Korrelation zwischen Effekten)
- eine Latente Mischung:

$$\beta_n \sim \sum_{c=1,\dots,C} s_c \cdot \text{MVN}(b_c, \Omega_c)$$

Bayesianische Schätzung

Grundlegende Formel:

$$\underbrace{\textsf{Pr}(\text{Parameter} \mid \text{Daten})}_{\text{A-posteriori-Verteilung}} \propto \underbrace{\textsf{Pr}(\text{Parameter})}_{\text{A-priori-Verteilung}} \cdot \underbrace{\textsf{Pr}(\text{Daten} \mid \text{Parameter})}_{\text{Likelihood}}$$

Für das **Ordered Probit Modell**:

$$\textsf{Pr}(\beta, \sigma, d \mid y) \propto \textsf{Pr}(\beta, \sigma, d) \cdot \underbrace{\textsf{Pr}(y \mid \beta, \sigma, d)}_{\text{Likelihood}}$$

Müssen Likelihood nicht berechnen (zeitintensiv), wenn (U_{nt}) “augmentiert” wird:

$$\textsf{Pr}(\beta, \sigma, d, U \mid y) \propto \textsf{Pr}(\beta, \sigma, d) \cdot \textsf{Pr}(U \mid \beta, \sigma, d) \cdot 1\{y_{nt} = j \Leftrightarrow U_{nt} \in (\gamma_{j-1}, \gamma_j]\}$$

Posterior bestimmen

Methode: Gibbs Sampling

- Gesucht: Zufallszahlen von $f(X,Y)$
- Bekannt: $f(X|Y)$ und $f(Y|X)$
- Algorithmus:
  - Initialisiere X^0 und Y^0
  - Für $r = 1, \ldots, R$: $X^j \sim f(X|Y^{j-1})$ und $Y^j \sim f(Y|X^j)$
- Dann: $(X^r, Y^r) \sim f(X,Y)$ für $r = Q+1, \ldots, R$ ("Burn-in" Phase)

A-posteriori-Verteilung der Modellparameter durch Gibbs Sampling approximieren:

- $U \sim$ Trunkierte Normalverteilung
- $\sigma \sim$ Inverse Wishart-Verteilung (bei konjugierter Prior)
- $\beta \sim$ Multivariate Normalverteilung (bei konjugierter Prior)
- d mittels Metropolis-Hastings (keine direkte Ziehung aus der Marginalverteilung möglich)

Worum geht's?

Das pairfam Panel

Deutsches Beziehungs- und Familienpanel

- Circa 12.000 “Anchors” (Hauptbefragte)
- Auch Befragung von Eltern, Kindern, Partnern
- Durchschnittlich circa 150 Variablen pro Befragung
- In diesem Vortrag:
  - Release 11.0 (11 Erhebungswellen von 2008/09 bis 2018/19)
  - “NA Fälle” ignoriert (MCAR Annahme)

pairfam
Panel Analysis of Intimate Relationships and Family Dynamics

Fragen mit diskreter, ordinaler Antwortskala

Frage 3 Variable val1_	**Zunächst geht es um allgemeine Ansichten zu Familie und Familienleben.** **Bitte sagen Sie mir, wie sehr Sie persönlich diesen Ansichten zustimmen.** *Int.: Liste 1 vorlegen!*

val1i1	Eltern und Kinder sollten sich ein Leben lang gegenseitig unterstützen
val1i2	Man sollte heiraten, wenn man mit einem Partner auf Dauer zusammenlebt.
val1i3	Frauen sollten sich stärker um die Familie kümmern als um ihre Karriere.
val1i4	Männer sollten sich genau so an der Hausarbeit beteiligen wie Frauen.
val1i5	Ein Kind unter 6 Jahren wird darunter leiden, wenn seine Mutter arbeitet.
val1i6	Kinder leiden oft darunter, dass sich ihre Väter zu sehr auf die Arbeit konzentrieren.
val1i7	Die Ehe ist eine lebenslange Verbindung und sollte nicht beendet werden.
val1i8	Man sollte spätestens dann heiraten, wenn ein Kind da ist.

	1	2	3	4	5	
Stimme überhaupt nicht zu	❒	❒	❒	❒	❒	Stimme voll zu

Weiß nicht .... ❒-1

Keine Angabe .... ❒-2

Fragen mit diskreter, ordinaler Antwortskala

Frage 3	
Variable val1_	**Zunächst geht es um allgemeine Ansichten zu Familie und Familienleben. Bitte sagen Sie mir, wie sehr Sie persönlich diesen Ansichten zustimmen.**
	Int.: Liste 1 vorlegen!

val1i1	Eltern und Kinder sollten sich ein Leben lang gegenseitig unterstützen
val1i2	Man sollte heiraten, wenn man mit einem Partner auf Dauer zusammenlebt.
val1i3	Frauen sollten sich stärker um die Familie kümmern als um ihre Karriere.
val1i4	Männer sollten sich genau so an der Hausarbeit beteiligen wie Frauen.
val1i5	Ein Kind unter 6 Jahren wird darunter leiden, wenn seine Mutter arbeitet.
val1i6	Kinder leiden oft darunter, dass sich ihre Väter zu sehr auf die Arbeit konzentrieren.
val1i7	Die Ehe ist eine lebenslange Verbindung und sollte nicht beendet werden.
val1i8	Man sollte spätestens dann heiraten, wenn ein Kind da ist.

	1	2	3	4	5	
Stimme überhaupt nicht zu	❐	❐	❐	❐	❐	Stimme voll zu

Weiß nicht .... ❐-1

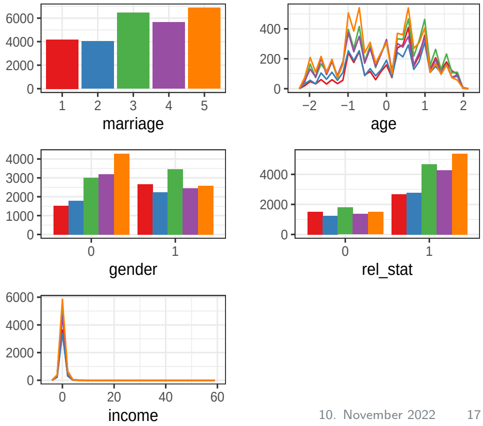
Keine Angabe .... ❐-2

Modelldaten (standardisiert)

```
summary(data)
```

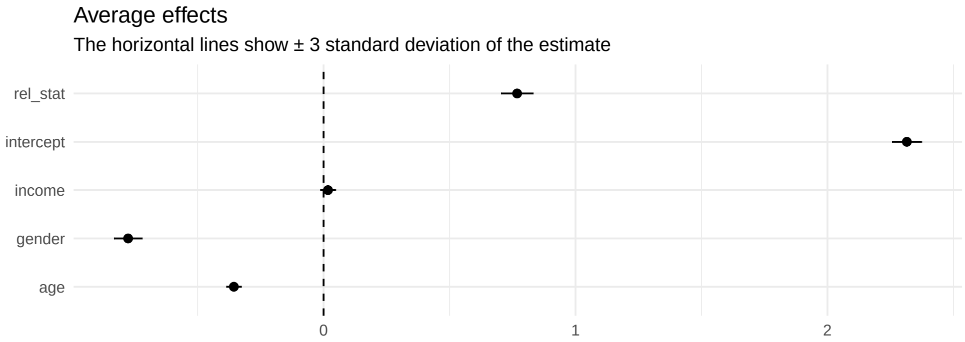
```
##                       count
## deciders              11135
## choice occasions        1-6
## total choices         27207
## alternatives              5
## - '1'                  4186
## - '2'                  4025
## - '3'                  6466
## - '4'                  5649
## - '5'                  6881
```

```
plot(data, by_choice = TRUE)
```



Modell (einfach)

```r
formula <- marriage ~ 1 + age + gender + rel_stat + income
```



Grenzwerte für die Zustimmung

Geschätzte Modellparameter γ:

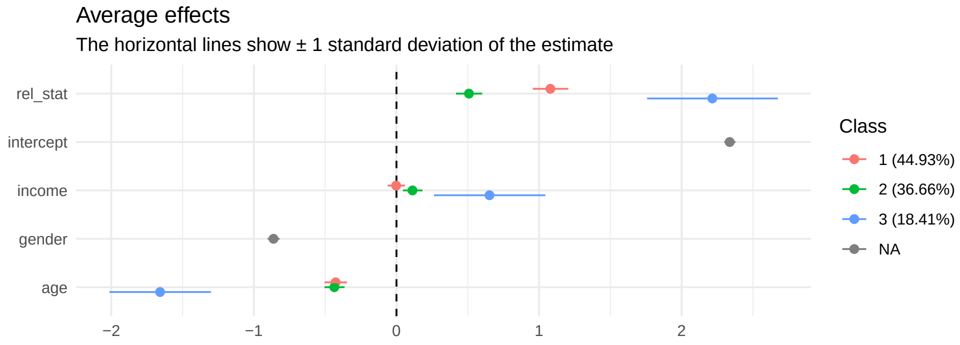
```
## gamma_0 gamma_1 gamma_2 gamma_3 gamma_4 gamma_5
##    -Inf    0.00    0.94    1.71    2.39     Inf
```

Beispiel:

- $\hat{\beta}_{\text{rel\_stat}} = 0.77$
- eine Person würde mit Eintritt in eine Beziehung immer von Antwortkategorie 4 zu 5 wechseln (ceteris paribus)
- aber nicht zwangsläufig von 2 zu 3 (ceteris paribus)

Modell (latente Klassen)

```
formula <- marriage ~ 1 + age + gender + rel_stat + income
re <- c("age", "rel_stat", "income")
C <- 3
```



Analyse der latenten Klassen

```
classes <- classification(mod)
```

```
head(classes, n = 3)
```

```
##                1      2      3 est
## 111000 0.3720 0.3892 0.2388   2
## 174000 0.5504 0.4432 0.0064   1
## 423000 0.4168 0.5328 0.0504   2
```

```
get_cov(mod, id = "111000")
```

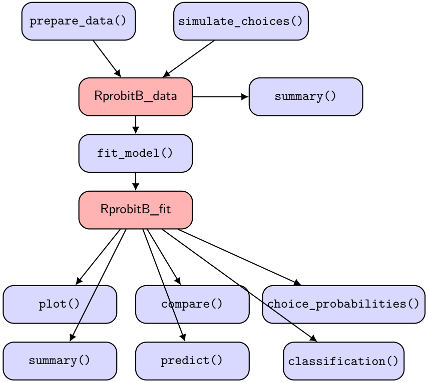
```
##          id    wave      age gender rel_stat     income marriage
## 7 111000 2014/15 1.303499      1        0 -0.7610623        1
```

Worum geht's?

Das {RprobitB} R Paket

RprobitB

modelling heterogeneous choices



Bayesianische Schätzung von Probit Wahlmodellen

- CRAN.R-project.org/package=RprobitB
- loelschlaeger.de/RprobitB
- github.com/loelschlaeger/RprobitB

Computerwerkstatt

Paket installieren und laden:

```
install.packages("RprobitB")
packageVersion("RprobitB") # sollte >= 1.1.2 sein
library("RprobitB")
```

File mit R Code herunterladen und in R öffnen:

loelschlaeger.de/files/code/aka22_rprobitb.R

UNIVERSITÄT BIELEFELD
Faculty of Business Administration and Economics

Danke für die Aufmerksamkeit!

Zusammengefasst:

- Modellierung von ordinalskalierten Antworten
- Klassifizierung von Befragten mittels latenter Klassen
- Modell Selektion wurde nicht behandelt

Ich freue mich über:

- Fragen und Anregungen zum Modell und zum R Paket
- Methodenvergleiche
- Datensätze über ordinale Wahlsituationen

loelschlaeger.de/talks

loelschlaeger@uni-bielefeld.de