

Klassifizierung von Präferenzen

Eine Anwendung mit dem R Paket `RprobitB`

Lennart Oelschläger

18.11.2021

Angewandte Klassifikationsanalyse
26. Workshop im Kloster Irsee
17.-19.11.2021

Warum?

- Gesellschaftliche Relevanz:
 - Verkehrsmittelwahl (Städteplanung, Umweltschutz)
 - Kaufentscheidungen (Marketing)
 - Lebensplanung (Energieanbieter, Wohnungsmarkt)
 - Produktionsplanung, politische Wahl, u.v.m.

- Unser Interesse:
 - Einflussfaktoren auf Wahlentscheidungen
 - Vorhersage von Wahlentscheidungen
 - heterogene Präferenzen
 - Klassifizierung von Entscheidern

Was?




- 1 Präferenzen
- 2 + Heterogenität
- 3 + Klassifizierung
- 4 = RprobitB

Präferenzen

Wie können wir Wahlentscheidungen und Präferenzen modellieren?

Wahlentscheidung: Entscheidersicht

„Berufstätige wählen das Transportmittel zur Arbeit, das ihren Nutzen maximiert. Der Nutzen sei $-2 \times \text{Fahrzeit} - \text{Fahrtkosten} + \text{Fahrspa\ss}$.“

Alternativen			
Fahrzeit	1	2	4
Fahrtkosten	3	1	0
Fahrspaß	0	-1	4
Nutzen ¹	-5	-6	-4
Wahlentscheidung			✓

¹Beachte: Das Level und die Skala der Nutzenwerte sind irrelevant.

Wahlentscheidung: Modelliersicht

$$\text{Nutzen} = \underbrace{V(\text{Beobachtete Einflüsse})}_{\text{Fahrzeit, Fahrtkosten}} + \underbrace{\text{Unbeobachtete Einflüsse}}_{\text{Fahrspaß, etc.}}$$

$$U = V(X) + \epsilon$$

In unserem Beispiel und u.d.A. $V(X) = X\beta$:

$$\begin{bmatrix} U_{\text{Auto}} \\ U_{\text{Bus}} \\ U_{\text{Roller}} \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 4 & 0 \end{bmatrix} \begin{bmatrix} \beta_{\text{Fahrzeit}} \\ \beta_{\text{Fahrtkosten}} \end{bmatrix} + \begin{bmatrix} \epsilon_{\text{Auto}} \\ \epsilon_{\text{Bus}} \\ \epsilon_{\text{Roller}} \end{bmatrix}$$

Die Verbindung zu der getroffenen Wahl:

$$\text{Wahlentscheidung } (y) = \arg \max \{ U_{\text{Auto}}, U_{\text{Bus}}, U_{\text{Roller}} \}$$

Wahlentscheidung: Modelliersicht

$$\begin{bmatrix} U_{\text{Auto}} \\ U_{\text{Bus}} \\ U_{\text{Roller}} \end{bmatrix} = X \begin{bmatrix} \beta_{\text{Fahrzeit}} \\ \beta_{\text{Fahrtkosten}} \end{bmatrix} + \begin{bmatrix} \epsilon_{\text{Auto}} \\ \epsilon_{\text{Bus}} \\ \epsilon_{\text{Roller}} \end{bmatrix}$$
$$y = \arg \max U$$

Annahmen:

- Der Entscheider wählt nutzenmaximierend ($\arg \max$)
- Linearität ($V(X) = X\beta$)
- $\begin{bmatrix} \epsilon_{\text{Auto}} \\ \epsilon_{\text{Bus}} \\ \epsilon_{\text{Roller}} \end{bmatrix} \sim N(\mu, \Sigma)$ (Probit)

Das Ziel: β , μ und Σ schätzen²

²Nicht alle Elemente von Σ sind identifiziert aufgrund der Invarianz bzgl. Nutzenlevel und -skala.

Interpretation von β

Die geschätzten Werte für β bilden Präferenzen ab und stellen Substitutionsverhalten dar. Zum Beispiel:

$$\begin{bmatrix} \beta_{\text{Fahrzeit}} \\ \beta_{\text{Fahrtkosten}} \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

- Größere Sensitivität bezüglich Fahrzeit als Fahrtkosten
- "*Value of Time*": Die Extrakosten, die ein Entscheider bereit ist zu zahlen, um Zeit zu sparen.

$$\left| \frac{\beta_{\text{Fahrzeit}}}{\beta_{\text{Fahrtkosten}}} \right| = 2$$

Wahlwahrscheinlichkeiten

Alternative i wird gewählt, wenn

$$U_i > U_j \quad \forall j \neq i.$$

Die Wahrscheinlichkeit, dass Alternative i gewählt wird beträgt somit:

$$\begin{aligned} P_i &= \text{Prob}(U_i > U_j \quad \forall j \neq i) \\ &= \int 1(U_i > U_j \quad \forall j \neq i) \phi(\epsilon) d\epsilon \end{aligned}$$

Modellschätzung

1. Frequentistisch: Maximierung der Likelihood Funktion

$$\arg \max_{\{\beta, \mu, \Sigma\}} \prod_{n, t, j} 1(y_{nt} = j) P_{ntj}$$

🗨️ Approximation notwendig, numerisch aufwendig

2. Bayesianisch: Bestimmung der A-posteriori-Verteilung

$$\text{Prob}(\beta, \mu, \Sigma \mid y, X) \propto \text{Prob}(y, X \mid \beta, \mu, \Sigma) \times \text{Prob}(\beta, \mu, \Sigma)$$

👍 Berechnung der Likelihood Funktion nicht notwendig
(Konzept: *data augmentation*)

+ Heterogenität

Wie können wir das Modell erweitern, um heterogene Präferenzen zu erfassen?

Heterogene Präferenzen

„Berufstätige wählen das Transportmittel zur Arbeit, das ihren Nutzen maximiert. Der Nutzen sei $-2 \times \text{Fahrzeit} - \text{Fahrtkosten} + \text{Fahrspa\ss}$.“

Das impliziert:

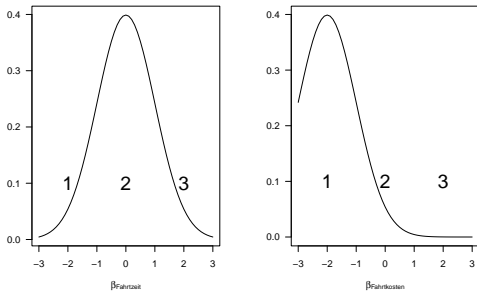
$$\beta = \begin{bmatrix} \beta_{\text{Fahrzeit}} \\ \beta_{\text{Fahrtkosten}} \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

Warum sollte jeder Entscheider das gleiche β haben?

Idee:

- β ist entscheiderspezifisch, also ein β_n pro Entscheider n
- Ausprägungen sind durch eine Verteilung bestimmt, also $\beta_n \sim f$
- Die Parameter einer parametrischen Verteilung f können ebenfalls geschätzt werden

Mischverteilung

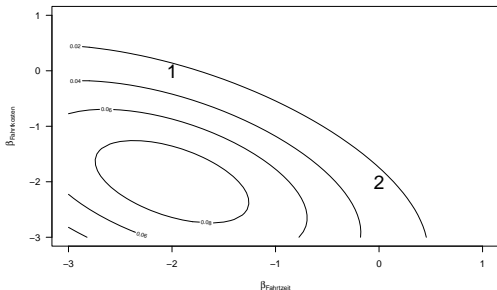


Typ 1: Je weniger Fahrtzeit / Fahrtkosten, desto besser.

Typ 2: Fahrtzeit / Fahrtkosten haben keinen Einfluss auf meine Entscheidung.

Typ 3: Je mehr Fahrtzeit / Fahrtkosten, desto besser.

Mischverteilung mit Korrelation



Typ 1: Für eine kürzere Fahrtzeit bezahle ich gerne mehr Geld.

Typ 2: Die Fahrt kann ruhig länger dauern, dafür soll sie aber nicht teuer sein.

+ Klassifizierung

Wie können wir das Modell erweitern, um Entscheider zu klassifizieren?

Latente Klassen

- Die Entscheider seien bzgl. ihrer Präferenzen in C Klassen einzuteilen.
- Sei $z_n \in \{1, \dots, C\}$ die Klasse von Entscheider n mit $\text{Prob}(z_n = c) = s_c$.
- Dann sei $\beta_n \sim N(b_{z_n}, \Omega_{z_n})$.

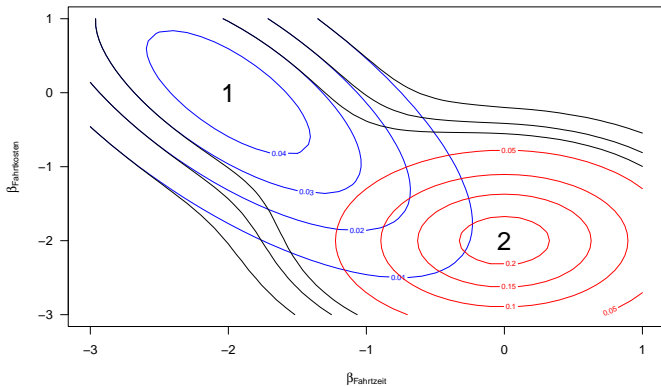
Insgesamt schätzen wir die Mischverteilung

$$\sum_{c=1, \dots, C} s_c N(b_c, \Omega_c).$$

Jede Klasse c ist charakterisiert durch

- s_c , der Klassenanteil,
- b_c , die mittleren Sensitivitäten,
- Ω_c , die Korrelationen der Wahlattribute.

Beispiel

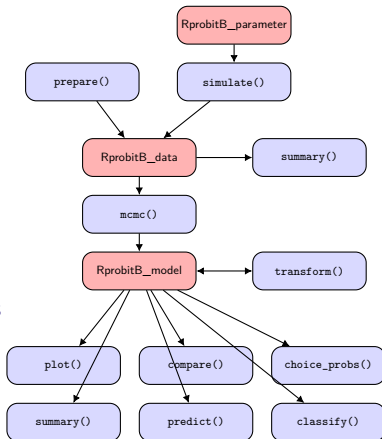


= RprobitB

Unsere Implementierung der Methodik in R.



- CRAN:
CRAN.R-project.org/package=RprobitB
- Webseite:
loeschlaeger.github.io/RprobitB
- Code:
github.com/loeschlaeger/RprobitB



Computerwerkstatt

Paket laden:

```
> install.packages("RprobitB")  
> library("RprobitB")
```

Anwendungen:

1. Zugverbindungswahl ("Train" Datensatz aus dem R Paket mlogit)
2. Simulierte Wahldaten
3. Verhütungsmittelwahl (Beziehungs- und Familienpanel "pairfam")

Datensatz

- $N = 235$ Entscheider mit $T = 5$ bis $T = 19$ Wahlen (*stated preferences*)
- $J = 2$ Zugverbindungsalternativen (A und B)
- Wahlattribute: Fahrkosten in Gulden Cents (`price`), Fahrzeit in Minuten (`time`), Komfort in drei Kategorien (`comfort`), Anzahl Umstiege (`change`)

Download:

```
> data("Train", package = "mlogit")
```

Fahrtkosten von Gulden Cents in Euro umwandeln:

```
> Train$price_A = Train$price_A / 100 * 2.2  
> Train$price_B = Train$price_B / 100 * 2.2
```

Modellformel und Datenvorbereitung

```
> form = choice ~ price | 1 | time + comfort + change
```

- Vorne stehen alternativ-spezifische Variablen mit einem generischen Koeffizienten.
- In der Mitte stehen alternativ-konstante Variablen.
- Hinten stehen alternativ-spezifische Variablen mit alternativ-spezifischen Koeffizienten.

```
> data = prepare(form = form, choice_data = Train)  
> summary(data)
```

Modellschätzung via Markov Chain Monte Carlo Simulation

```
> m1 = mcmc(data)
> summary(m1)
```

Den Preiskoeffizienten auf -1 fixieren:

```
> scale = list(parameter = "a", index = 1, value = -1)
> m2 = transform(m1, scale = scale)
> summary(m2)
> plot(m2)
```

Konvergenz des Gibbs Samplers überprüfen:

```
> plot(m2, "trace")
> plot(m2, "acf")
> m3 = transform(m2, Q = 3)
> plot(m3, "acf")
```

Wahlvorhersage

Datensatz in Trainings- und Testteil aufteilen:

```
> data = prepare(form = form, choice_data = Train,  
                 test_prop = 0.3)
```

Out-of-sample Vorhersage:

```
> m4 = mcmc(data$train)  
> predict(m4, data$test)
```


Mischverteilung

```
> data = simulate(form = choice ~ price | 0 | time,  
                 N = 100, T = 10, J = 3,  
                 alternatives = c("car", "bus", "scooter"),  
                 re = c("price", "time"), C = 3,  
                 seed = 1)  
  
> m5 = mcmc(data, latent_classes = list(C = 3))  
> summary(m5)  
> plot(m5, "mixture")  
> predict(m5)  
> classify(m5)
```

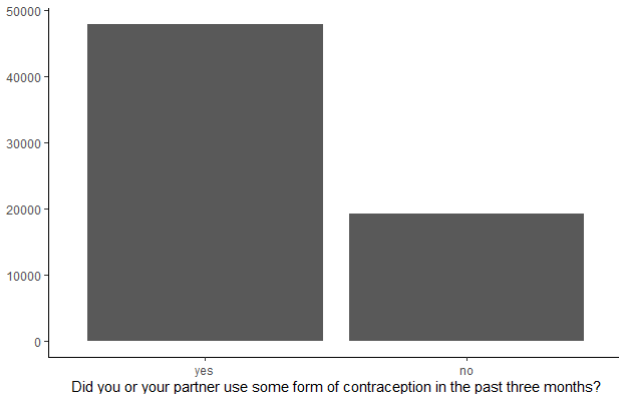
Update der Klassenanzahl:

```
> m6 = mcmc(data,  
            latent_classes = list(update = TRUE,  
                                   C = 5,           # initial no  
                                   epsmin = 0.05,   # remove  
                                   epsmax = 0.8,    # split  
                                   distmin = 0.1)) # join
```

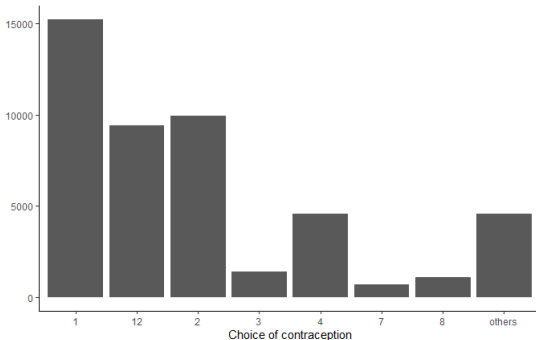
Datensatz

- "Panel Analysis of Intimate Relationships and Family Dynamics", kurz "pairfam"
- Längsschnittstudie in Deutschland seit 2008
- Jährlich wiederholte Befragungen von über 12.000 bundesweit zufällig ausgewählten Personen
- www.pairfam.de

Verhütungsmittleinsatz



Alternativen



1 = Pille, 2 = Kondom, 3 = Hormonpräparate, 4 = Spirale, 7 = Sterilisation der Frau, 8 = Sterilisation des Mannes

Einfluss von Alter und Partnerschaft

- Entscheider: 7534 mit je 1 - 11 Wahlen
- Alternativen: 1. Pille (11515), 2. Kondom (8899)
- Alter: 15 - 48
- Partnerschaft: Single (5468), Beziehung (14946)

```
> data = prepare(form = choice ~ 0 | age + relstat + 1,  
                 choice_data = data, standardize = "age_1")  
> fit = mcmc(data)
```

Einfluss von Alter und Partnerschaft

```
> summary(fit)
Legend of linear coefficients:
      name      re
1   age_1 FALSE
2 relstat_1 FALSE
3   ASC_1 FALSE
```

```
Parameter statistics:
      mean      sd      R^
alpha
1  -0.02    0.00    1.00
2   0.85    0.02    1.00
3   0.08    0.04    1.00

Sigma
1,1    1.00    0.00    1.00
```

```
> predict(fit)
      predicted
true      1      2
1  9548 1967
2  5398 3501
```

Weitere mögliche Wahlattribute

Alternativen-spezifisch

Eingriff in den Sexualakt
Biologischer / permanenter Eingriff
Regelmäßige Anwendung
Preis
...

Entscheider-spezifisch

Einkommen
Ost- / Westdeutschland
Anzahl Kinder
Anzahl Sexualpartner
Kinderwunsch
...

Danke für die Aufmerksamkeit!

Ich freue mich sehr über:

- Fragen und Anregungen zur Methodik und zum R Paket
- Ideen zur Modellierung der Verhütungsmittelwahl
- Datensätze über Wahlen

 lennart.oelschlaeger@uni-bielefeld.de

 [l_oelschlaeger](https://twitter.com/l_oelschlaeger)

 oilbat.de