Schätzung partieller attributabler Risiken in der epidemiologischen Praxis

Rabe C, Ecke J, Lehnert-Batar A, Pfahlberg A, Gefeller O

FAU Erlangen-Nürnberg
Institut für Medizininformatik, Biometrie und Epidemiologie

September 2006, Jahrestagung der GMDS, Leipzig

Besonderheiten des PARs

- PAR zerlegt das gemeinsame attributable Risiko in Komponenten
- Interaktionen und Assoziationen zwischen Expositionsvariablen werden berücksichtigt
- Fälle, die durch das gemeinsame Agieren zweier Faktoren entstanden sind, werden gleichmäßig auf die Faktoren aufgeteilt
- Es entsteht ein Ranking von Faktoren

Beispiel

E_1	E_2	E_3	S_{ijk}	$P(S_{ijk}\|D)$	$P(S_{ijk})$	RR_{ijk}	AR_{ijk}
0	0	0	S_{000}	0.228	0.586	1.00	0.000
0	0	1	S_{001}	0.053	0.034	4.00	0.039
0	1	0	S_{010}	0.074	0.064	3.00	0.050
0	1	1	S_{011}	0.068	0.116	1.50	0.023
1	0	0	S_{100}	0.011	0.014	2.00	0.005
1	0	1	S_{101}	0.218	0.066	8.45	0.192
1	1	0	S_{110}	0.218	0.086	6.50	0.185
1	1	1	S_{111}	0.131	0.034	9.95	0.118

Schichtspezifische AR:

$$AR_{ijk} = P(S_{ijk}|D)\frac{RR_{ijk} - 1}{RR_{ijk}}$$

Beispiel

E_1	E_2	E_3	S_{ijk}	$P(S_{ijk}\|D)$	$P(S_{ijk})$	RR_{ijk}	AR_{ijk}
0	0	0	S_{000}	0.228	0.586	1.00	0.000
0	0	1	S_{001}	0.053	0.034	4.00	0.039
0	1	0	S_{010}	0.074	0.064	3.00	0.050
0	1	1	S_{011}	0.068	0.116	1.50	0.023
1	0	0	S_{100}	0.011	0.014	2.00	0.005
1	0	1	S_{101}	0.218	0.066	8.45	0.192
1	1	0	S_{110}	0.218	0.086	6.50	0.185
1	1	1	S_{111}	0.131	0.034	9.95	0.118

Schichtspezifische AR:

$$AR_{ijk} = P(S_{ijk}|D)\frac{RR_{ijk} - 1}{RR_{ijk}}$$

- Die Summe der schichtspezifischen ARs ergibt das gemeinsame AR
- Das PAR eines Faktors E_k wird aus den Schichten, in denen E_k vorliegt, berechnet
- Schichten, in denen nur E_k vorliegt, werden ganz E_k zugeschrieben
- In den Schichten, in denen mehrere Faktoren vorliegen, muss das ER aufgeteilt werden (nach der Shapley-Regel)

Aufteilung des Exzess-Risikos

In einer Schicht mit nur zwei Faktoren gilt z.B.:

$$RR_{110} - 1 = \underbrace{RR_{100} - 1 + \frac{1}{2}(RR_{110} - RR_{100} - RR_{010} + 1)}_{E_1} +$$
$$\underbrace{RR_{010} - 1 + \frac{1}{2}(RR_{110} - RR_{010} - RR_{010} + 1)}_{E_2}$$

In der Schicht mit drei Faktoren hingegen gilt: E_1 bekommt

$$\frac{1}{3}(RR_{100} - 1) + \frac{1}{6}(RR_{110} - RR_{010}) + \frac{1}{6}(RR_{101} - RR_{001}) +$$
$$\frac{1}{3}(RR_{111} - RR_{011})$$

Aufteilung des Exzess-Risikos

In einer Schicht mit nur zwei Faktoren gilt z.B.:

$$RR_{110} - 1 = \underbrace{RR_{100} - 1 + \frac{1}{2}(RR_{110} - RR_{100} - RR_{010} + 1)}_{E_1} + \underbrace{RR_{010} - 1 + \frac{1}{2}(RR_{110} - RR_{010} - RR_{010} + 1)}_{E_2}$$

In der Schicht mit drei Faktoren hingegen gilt: E_1 bekommt

$$\frac{1}{3}(RR_{100} - 1) + \frac{1}{6}(RR_{110} - RR_{010}) + \frac{1}{6}(RR_{101} - RR_{001}) + \frac{1}{3}(RR_{111} - RR_{011})$$

- D.h., das *PAR* ist der erwartete *marginale* Anteil, den jeder Faktor zum gemeinsamen *AR* beiträgt
- Das Risiko in jeder einzelnen Schicht wird evaluiert
- Interaktionen (d.h. Abweichungen vom additiven Modell) sind wesentlicher Aspekt bei der Berechnung

Schätzung

Zur Schätzung des PARs (Shapley) in multinomialverteilten Daten betrachte:

$$PAR(E_l) = \frac{1}{L!P(D=1)} \sum_{r=1}^{2^{L-1}} P(Z_r, E_l = 1) \times$$

$$\sum_{q=1}^{2^{L-1}} (L - r^+ - 1)! \left[P(D=1|Z_{r \circ q}, E_l = 1) - P(D=1|Z_{r \circ q}, E_l = 0) \right].$$

D.h. man benötigt

- alle bedingten Wahrscheinlichkeiten $P(D|E_1, ...E_L)$
- alle Schichtwahrscheinlichkeiten $P(E_1, ...E_L)$
- $P(D)$

Im Falle zusätzlicher Confounder muss $P(D|E_1, ...E_L)$ noch adjustiert werden:

$$P(D|E_1, ...E_L) = \sum_{k=1}^{K} P(D|E_1, ...E_L, C_k)P(C_k)$$

D.h. man benötigt

- alle bedingten Wahrscheinlichkeiten $P(D|E_1, ...E_L)$
- alle Schichtwahrscheinlichkeiten $P(E_1, ...E_L)$
- $P(D)$

Im Falle zusätzlicher Confounder muss $P(D|E_1, ...E_L)$ noch adjustiert werden:

$$P(D|E_1, ...E_L) = \sum_{k=1}^{K} P(D|E_1, ...E_L, C_k)P(C_k)$$

Beispiel: NAEVAC-Daten

Querschnittsstudie, bei der die Anzahl der Naevi bei 6-7-jährigen Kindern bestimmt wurde. Risikofaktoren:

- Konstitution (Haarfarbe, Augenfarbe, Sommersprossen, FP)
- UV-Exposition (Südurlaube, Sonnencreme, Freibad, Sonnenbrand)

Dazu kommen Confounder (Alter, Geschlecht, Untersuchungsort, Körperoberfläche)

ORs und Prävalenzen

Faktor	OR	$P(E)$
bl. Augen	1.50	0.46
gr. Augen	1.80	0.21
br. Haar	3.92	0.51
bl. Haar	4.99	0.40
ro. Haar	1.52	0.02
Sommerspr.	1.26	0.17
FP	1.04	0.33
Freibad	1.35	0.76
Urlaub	1.24	0.46
Sonnenbr.	1.16	0.38
Freien	1.09	0.62

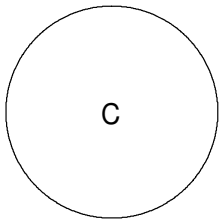
Dimension der Kontingenztafel aus der man schätzt:

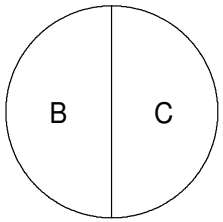
- nur Exposition: 768
- Dis und Exp : 1536
- mit Confoundern: 55296

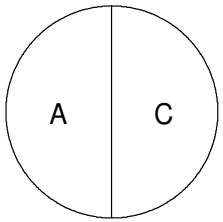
Anzahl der Beobachtungen: 1911!

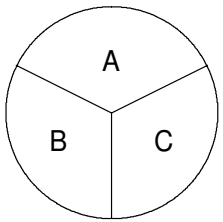
Übliches Vorgehen, z.B. auch bei AR_{adj}

- schätze $P(D|E_1, ...E_L, C_k)$ aus logistischer Regression
- Problem: welches Modell?
- Nur Haupteffekte (d.h. multiplikatives Risikomodell) widerspricht Konzept des PARs! Schichten werden evtl. nicht genau genug geschätzt!





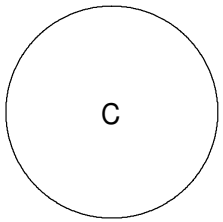


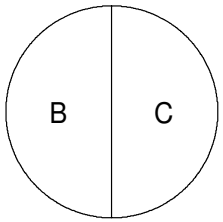


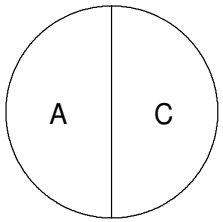
		$P(D\|S)$		
A	B	$A*B$	$A+B$	$P(S)$
0	0	0.25	0.08	0.1
0	1	0.25	0.29	0.4
1	0	0.25	0.33	0.2
1	1	0.75	0.71	0.3

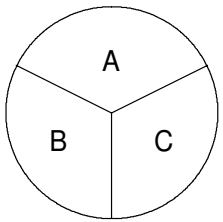
ARs und PARs

		A*B	A+B
PAR	A	0.21	0.39
	B	0.21	0.43
AR	A	0.42	0.46
	B	0.42	0.49
AR	(gem)	0.42	0.82









		$P(D\|S)$		
A	B	$A*B$	$A+B$	$P(S)$
0	0	0.25	0.08	0.1
0	1	0.25	0.29	0.4
1	0	0.25	0.33	0.2
1	1	0.75	0.71	0.3

ARs und PARs

		A*B	A+B
PAR	A	0.21	0.39
	B	0.21	0.43
AR	A	0.42	0.46
	B	0.42	0.49
AR	(gem)	0.42	0.82

- Vorschlag Variablenselektion: Interaktionen, die das PAR eines Faktors um mehr als z.B. 10% ändern, werden in einem stepwise Verfahren in Modell aufgenommen.
- Variablenselektion basierend auf AIC
- Variablenselektion p-Wert basiert
- Alternativ: Boosting

Boosting

- Algorithmus aus der Machine Learning Literatur
- Ursprünglich zur Klassifikation (AdaBoost)
- Vorteil: Variablenselektion während Anpassung möglich
- Gute Eigenschaften in hohen Dimensionen

Boosting

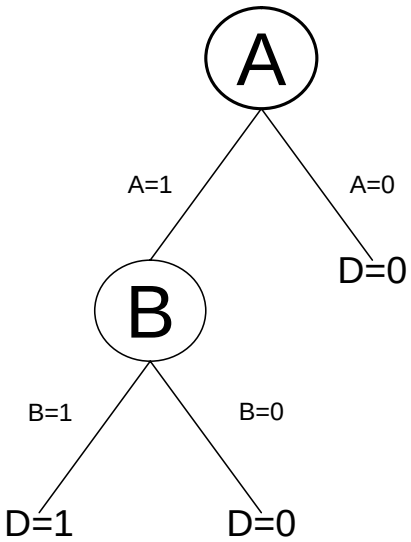
$$F(x) = a_0 + \sum_{m=1}^{M} a_m \cdot f_m(x)$$

- $F(x) = y$ (Zielvariable)
- M Größe des Ensembles bzw. Anzahl der Boosting-Iterationen
- $f(x)$ Baselearner, z.B. Decision trees oder einfache univariate Modelle

In jeder Boosting-Iteration m wird ein Baselearner gesucht, der die Residuen des Ensembles bis zur $(m-1)$ten Iteration am besten erklärt. Das entspricht einer additiven schrittweisen Anpassung. Die Koeffizienten a_m werden entsprechend der Verlustfunktion optimiert.

Boosting und Interaktionen

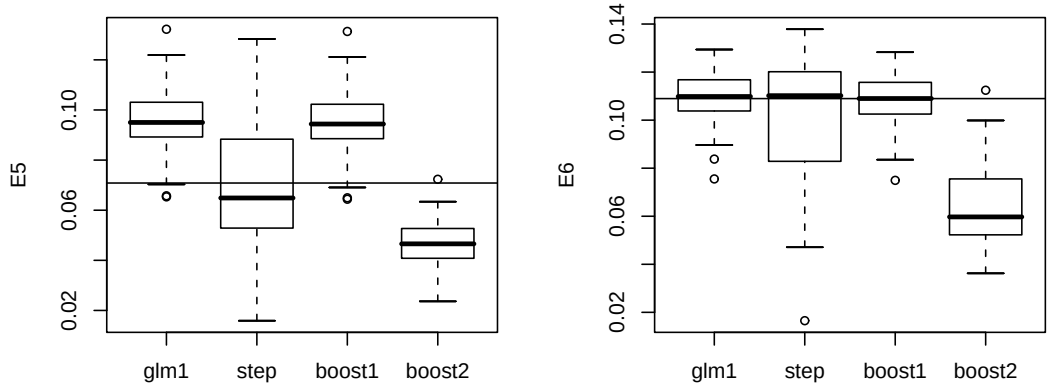
- Interaktionen können durch die Baselearner aufgenommen werden
- Bei binären Variablen sind Interaktionen auf natürlicher Art durch Bäume bzw. durch Boolsche Kombinationen darstellbar:



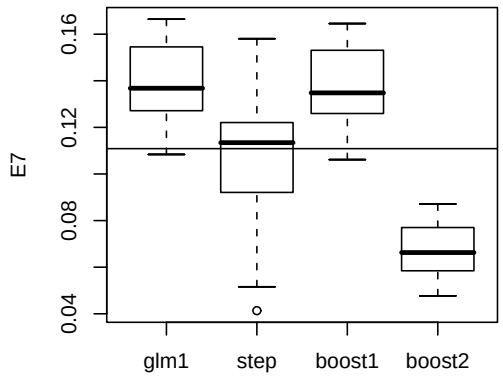
Ergebnisse von verschiedenen Modellen

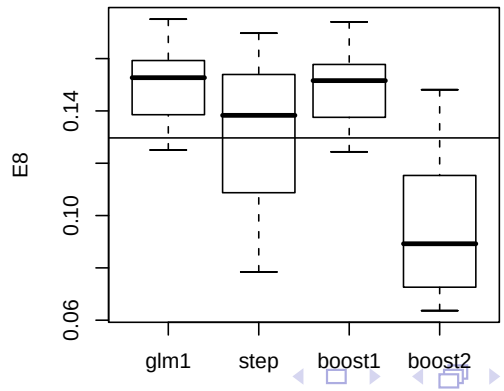
		Stepwise				Boosting	
Faktor	Logistic	Step1	Step2	p	AIC	HE	IA
bl. Aug	0.07	0.10	0.07	0.07	0.07	0.08	0.05
gr. Aug	0.05	0.05	0.01	0.05	0.05	0.06	0.03
br. Haar	0.26	0.23	0.21	0.26	0.26	0.20	0.06
bl. Haar	0.26	0.19	0.18	0.28	0.24	0.21	0.14
r. Haar	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Sommerspr.	0.02	0.02	0.02	0.00	0.01	0.02	0.01
Fitzp.	0.01	0.01	0.04	0.01	0.03	0.00	0.00
Freibad	0.09	0.09	0.13	0.10	0.11	0.09	0.05
Urlaub	0.04	−0.03	−0.06	0.02	−0.01	0.03	0.01
Sonnenbrand	0.02	0.06	0.02	0.04	0.04	0.02	0.02
Draussen	0.02	0.08	0.10	0.02	0.02	0.01	0.01

Simulation

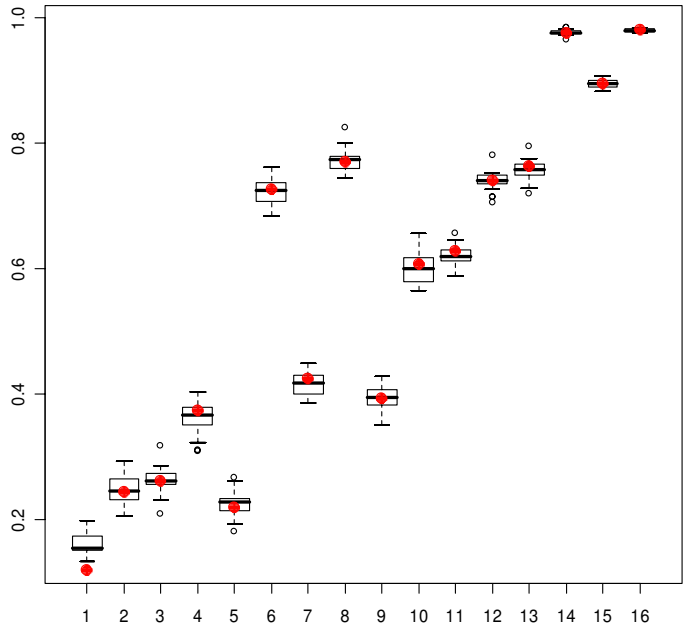

E5
0.02
0.06
0.10
glm1
step
boost1
boost2

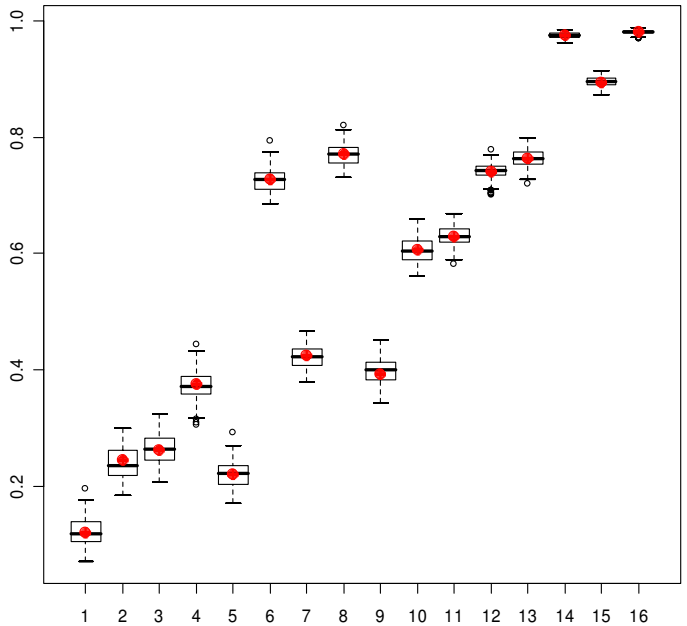
E6
0.02
0.06
0.10
0.14
glm1
step
boost1
boost2


E7
0.04
0.08
0.12
0.16
glm1
step
boost1
boost2


E8
0.06
0.10
0.14
glm1
step
boost1
boost2

Schichtwahrscheinlichkeiten in Simulation





Zusammenfassung

- Schätzen des PARs mit vielen Faktoren problematisch, dafür sind große Datensätze nötig
- Geeignete Modellselektion
- Stepwise Verfahren leiden unter großer Variabilität
- Verschiedene Modelle sollten untersucht werden, der Fit alleine deckt Unterschiede nicht auf

Literatur

- Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. J Clin Epidemiol 1995;48:645-655.
- Land M, Gefeller O. A game-theoretic approach to partitioning attributable risks. Biometrical J 1997;39:777-792.
- Bühlmann P. Boosting for high-dimensional linear models. Ann Statist 2006;34:559-583.