Universität Innsbruck

Schönheit liegt im Auge des Betrachters:
Ein statistischer Blick auf Topmodels und Professoren

Achim Zeileis

http://eeecon.uibk.ac.at/~zeileis/

Überblick

- Fragestellungen
  - Attraktivität von Germany's Next Topmodel Kandidatinnen
  - Schönheit und Lehrevaluierungen von Professoren
- Motivation
  - Datenmodelle und algorithmische Modelle
  - Bäume und Blätter
- Modellbasiertes rekursives Partitionieren
  - Modellschätzung
  - Tests auf Parameterinstabilität
  - Segmentatierung
  - Stutzung
- Software
- Zusammenfassung

Fragestellungen: Attraktivität von Topmodels






Fragen: Welche dieser Frauen ist attraktiver? *Und:* Wie hängt die Antwort auf diese Frage von Alter, Geschlecht und Kenntnis der TV-Show Germany's Next Topmodel ab?

Fragestellungen: Attraktivität von Topmodels

Quelle: Strobl, Wickelmaier, Zeileis (2011, *Journal of Educational and Behavioral Statistics*). “Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning.”

Germany’s Next Topmodel:

- TV-Casting-Show, moderiert von Heidi Klum.
- Zweite Staffel von März bis Mai 2007 auf ProSieben.
- Finalistinnen: Barbara Meier, Anni Wendler, Hana Nitsche, Fiona Erdmann, Mandy Graff, Anja Platzer (in absteigender Platzierung).

Frage: Wie soll Attraktivität der sechs Topmodels erfasst werden?

- *Rating.* Bewertung jeder Kandidatin auf einer numerischen Skala.
- *Ranking.* Rangfolge der Kandidatinnen.
- *Paarvergleiche.* Für viele/alle möglichen Paare von zwei Kandidatinnen: *Welche dieser zwei Frauen ist attraktiver?*

Fragestellungen: Attraktivität von Topmodels

Hier: Paarvergleiche, da für Menschen oft leichter, schneller und verläßlicher zu beantworten.

Daten:

- Umfrage bei 192 Personen im Sommer 2007 durch Psychologisches Institut der Universität Tübingen.
- Stichprobe war stratifiziert nach Geschlecht und Alter ($\leq$ 30 und $>$ 30 Jahre) mit jeweils 48 Befragten.
- Paarvergleiche für alle 15 mögliche Paare auf Basis von Fotos.
- Zusätzliche Information über jede befragte Person: Alter, Geschlecht, Kenntnis der TV-Show.

Fragestellungen: Schönheit und Lehrevaluierungen

Qualität der Lehre

unbefriedigend … sehr gut

Fragen: Hängt die Lehrevaluierung von Professoren von ihrer Schönheit ab? *Und:* Ist dieser Zusammenhang anders für Professoren unterschiedlichen Alters, Geschlechts, . . . ?

Quelle: Hamermesh & Parker (2005, *Economics of Education Review*). “Beauty in the Classroom: Instructors’ Pulchritude and Putative Pedagogical Productivity.”

Fragestellungen: Schönheit und Lehrevaluierungen

Daten:

- 463 Kurse an der University of Texas at Austin von 2000–2002.
- *Lehrproduktivität:* Durchschnittliche Evaluierung pro Kurs bzgl. der Gesamtbewertung der Qualität der Lehre (auf Skala 1–5). Insgesamt ursprünglich 16.957 Bewertungen.
- *Schönheit:* Rating unabhängiger Jury für alle Kursleiter/innen. Aggregierte und standardisierte Ratings (auf Skala 1–10) von sechs unabhängigen Studierenden (stratifiziert nach Geschlecht und Studienabschnitt) auf Basis von Personal-Fotos.
- Zusätzliche Information über jeden Kurs bzw. Kursleiter/innen: Studienabschnitt, Geschlecht, Minderheit, Tenure, u.ä.

Statistische Modellierung

Breiman (2001, *Statistical Science*) unterscheidet zwei Kulturen statistischer Modellierung.

Datenmodelle: Stochastische Modelle, häufig parametrisch.

- Beispiele: (Verallgemeinerte) lineare Modelle, ...
- Illustration: Bradley-Terry-Modell für Schätzung von Attraktivität in Topmodels-Paarvergleich.

Algorithmische Modelle: Flexible Modelle, datengenerierender Prozess unbekannt.

- Beispiele: Rekursives Partitionieren/Entscheidungsbäume, neuronale Netze, ...
- Illustration: Entscheidungsbaum für Attraktivitätspräferenz von Barbara vs. Anja anhand von Geschlecht und Alter der beurteilenden Person.

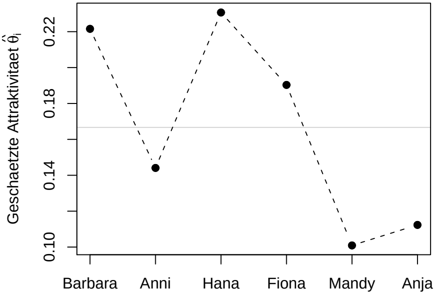
Statistische Modellierung: Datenmodelle

Bradley-Terry-Modell: Wahrscheinlichkeit p_{ij}, dass i gegenüber j vorgezogen wird, parametrisiert durch “Wert” (hier: Attraktivität) θ_i.

$$\begin{aligned} p_{ij} &= \frac{\theta_i}{\theta_i + \theta_j} \\ \text{logit}(p_{ij}) &= \log(\theta_i) - \log(\theta_j) \end{aligned}$$

Vorteil: Paarvergleichsstruktur wird genau parametrisiert.

Nachteil: Abhängigkeit der θ_i von Kovariablen (Geschlecht, Alter, ...) müßte auch genau spezifiziert werden.

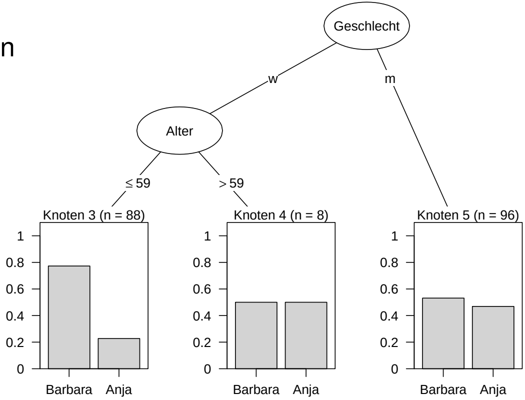


Statistische Modellierung: Algorithmische Modelle

Entscheidungsbaum: Rekursives Aufteilen der Entscheidungen (hier: für attraktivere von zwei Kandidatinnen) bzgl. Kovariablen (hier: Geschlecht und Alter) mit "stärkstem Kontrast" der Entscheidungen.

Vorteil: Abhängigkeit der Attraktivität von Kovariablen wird flexibel "erlernt".

Nachteil: Paarvergleichs-struktur kann nicht leicht ausgenutzt werden.



Motivation: Bäume und Blätter

Ziel: Synthese von parametrischen Datenmodellen und algorithmischen Baummodellen.

- *Baum:* Erlernen des Zusammenhangs von Y mit Kovariablen X_j durch rekursive Partitionierung.
- *Blätter:* Spezialisierte parametrische Modelle für Y, lokal für jeden Teildatensatz.

Modellbasiertes rekursives Partitionieren

Basisalgorithmus:

1. Passe ein Modell für Y an.
2. Erfasse den Zusammenhang zwischen Y und jedem der X_j.
3. Unterteile die Stichprobe entlang dem X_{j^*} mit der höchsten Assoziation: wähle den Bruchpunkt mit der höchsten Verbesserung der Anpassungsgüte.
4. Wiederhole die Schritte 1–3 rekursiv in den Teilstichproben so lange, bis ein Abbruchkriterium erfüllt ist.

Hier: Segmentierung (3) von parametrischen Modellen (1) mit additiver Zielfunktion auf Basis von Parameterinstabilitätstests (2) und zugehöriger statistischer Signifikanz (4).

1. Modellschätzung

Modelle: $\mathcal{M}(Y, \theta)$ mit (potentiell) multivariaten Beobachtungen Y und k-dimensionalem Parametervektor θ.

Parameterschätzung: $\widehat{\theta}$ durch Optimierung einer additiven Zielfunktion $\Psi(Y, \theta)$ auf n Beobachtungen Y_i ($i = 1, \dots, n$):

$$\widehat{\theta} \quad = \quad \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta).$$

Spezialfälle: Maximum Likelihood (ML), (gewichtete) kleinste Quadrate (OLS und WLS), Quasi-ML, und weitere M-Schätzer.

Zentraler Grenzwertsatz: Wenn es einen wahren Parameter θ_0 gibt und gewisse Regularitätsannahmen gelten, dann ist $\widehat{\theta}$ asymptotisch normalverteilt mit Mittel θ_0.

1. Modellschätzung

Idee: Oft ist es nicht sinnvoll ein einziges globales Modell $\mathcal{M}(Y, \theta)$ an *alle n* Beobachtungen anzupassen. Möglicherweise kann man aber eine Partition bzgl. Kovariablen $X = (X_1, \dots, X_l)$ finden, so dass ein passendes Modell für jede Zelle der Partition existiert.

Werkzeug: Erfassung von Parameterinstabilitäten bzgl. der Partitionierungsvariablen $X_1, \dots X_l)$.

Schätzfunktion: Modellabweichungen können erfasst werden durch Model deviations can be captured by

$$\psi(Y_i, \widehat{\theta}) = \frac{\partial \Psi(Y, \theta)}{\partial \theta} \Big|_{Y_i, \widehat{\theta}}$$

d.h. die Beiträge zum *Gradienten*.

2. Tests auf Parameterinstabilität

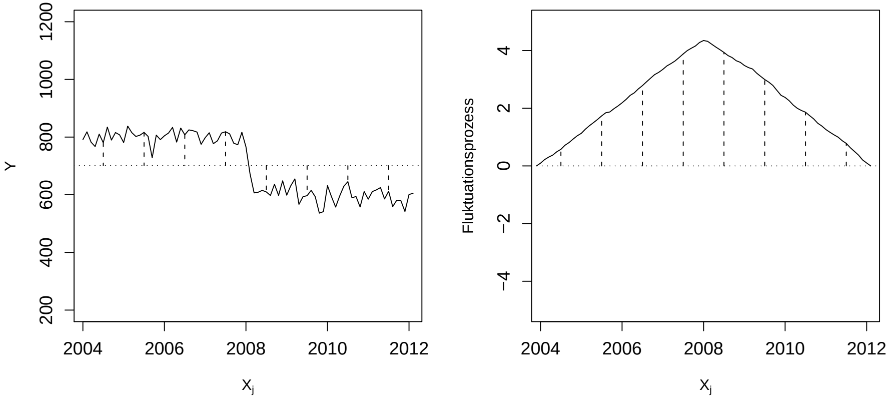
Verallgemeinerte M-Fluktuationstests können Instabilitäten in $\widehat{\theta}$ über eine Anordnung bzgl. X_j erfassen.

Grundlage: Empirischer Fluktuationsprozess von kumulativen Abweichungen über eine Anordnung $\sigma(X_{ij})$.

$$W_j(t, \widehat{\theta}) \quad = \quad \widehat{V}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi(Y_{\sigma(X_{ij})}, \widehat{\theta}) \qquad (0 \le t \le 1)$$

Funktionaler zentraler Grenzwertsatz: Unter Parameterstabilität gilt $W_j(\cdot) \xrightarrow{\mathrm{d}} W^0(\cdot)$, wobei W^0 eine k-dimensionale Brownsche Brücke ist.

2. Tests auf Parameterinstabilität



2. Tests auf Parameterinstabilität

Teststatistik: Durch ein skalares Funktional $\lambda(W_j)$ kann erfasst werden, ob die Fluktuation zu groß ist.

Prüfverteilung: Asymptotische Verteilung von $\lambda(W^0)$.

Spezialfälle: Klasse umfasst bekannte Tests für unterschiedliche Modelle. Bestimmte Funktionale λ sind besonders intuitiv für numerische bzw. kategoriale X_j.

Vorteil: Modell $\mathcal{M}(Y, \widehat{\theta})$ muss nur einmal geschätzt werden. Empirische Schätzfunktionen $\psi(Y_i, \widehat{\theta})$ werden dann nur umgeordnet und Partialsummen gebildet.

2. Tests auf Parameterinstabilität

Numerische Partitionierungsvariablen: Erfasse Instabilität durch supLM-Statistik.

$$\lambda_{\sup LM}(W_j) \quad = \quad \max_{i=\underline{i},\dots,\bar{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n} \right)^{-1} \left\| W_j \left(\frac{i}{n} \right) \right\|_2^2 .$$

Interpretation: Maximierung von LM-Statistiken für eine einzelne Parameterveränderung über alle möglichen Bruchpunkte $[\underline{i}, \bar{\imath}]$.

Grenzverteilung: Supremum eines quadrierten, k-dimensionalen bedingten Bessel-Prozesses.

2. Tests auf Parameterinstabilität

Kategoriale Partitionierungsvariablen: Erfasse Instabilität durch χ^2-Statistik.

$$\lambda_{\chi^2}(W_j) \quad = \quad \sum_{c=1}^{C} \frac{n}{|I_c|} \left\| \Delta_{I_c} W_j \left(\frac{i}{n} \right) \right\|_2^2$$

Eigenschaft: Invariant gegen Umordnung der C Kategorien sowie der Beobachtungen innerhalb der Kategorien.

Interpretation: Erfasst Instabilität bei Aufteilung in C Kategorien.

Grenzverteilung: χ^2 mit $k \cdot (C - 1)$ Freiheitsgraden.

3. Segmentierung

Ziel: Vollständige Segmentierung in $b = 1, \dots, B$ Teilmodelle entlang der Variablen X_j mit der höchsten Parameterinstabilität. Lokale Optimierung von

$$\sum_b \sum_{i \in I_b} \Psi(Y_i, \theta_b).$$

$B = 2$: Vollständige Suche der Ordnung $O(n)$.

$B > 2$: Vollständige Suche ist von Ordnung $O(n^{B-1})$, kann aber durch dynamische Programmierung der Ordnung $O(n^2)$ ersetzt werden. Verschiedene Verfahren (bspw. Informationskriterien) können B adaptiv wählen.

Hier: Binäre Partitionierung.

4. Stutzung

Stutzung: Soll Überanpassung verhindern (*Pruning*).

Pre-Pruning: Internes Abbruchkriterium. Beende Partitionierung, wenn keine signifikanten Parameterinstabilitäten mehr vorliegen.

Post-Pruning: Lasse großen Baum wachsen und stutze die Verzweigungen, die keine Verbesserung bringen (bspw. durch Kreuzvalidierung oder Informationskriterien).

Hier: Pre-Pruning auf Basis Bonferroni-korrigierter p-Werte der Fluktuationstests.

Topmodel-Modell

Fragestellung: Skalierung von Attraktivitätseinstufungen.

Modell: Paarvergleich via Bradley-Terry.

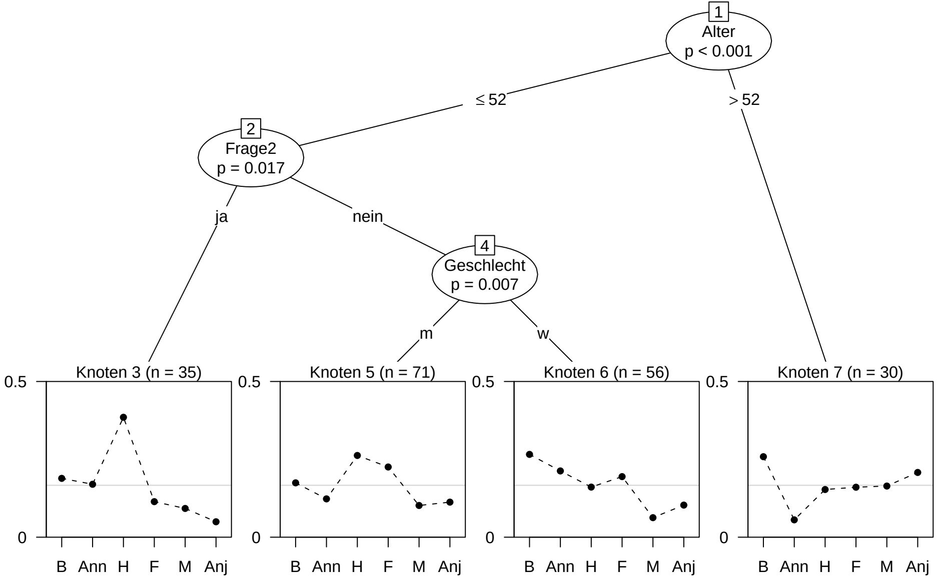
- Paarvergleiche für Attraktivität von Germany's Next Topmodel Finalistinnen: Barbara, Anni, Hana, Fiona, Mandy, Anja.
- Kovariablen: Geschlecht, Alter, Kenntnis der TV-Show.
- Kenntnis erfasst durch Ja/Nein-Fragen: (1) Erkennen Sie die Frauen? Kennen Sie die Sendung? (2) Haben Sie sie regelmäßig angeschaut? (3) Haben Sie das Finale gesehen? Wissen Sie wer gewonnen hat?

Topmodel-Modell






Topmodel-Modell



Schöne Professoren

Fragestellung: Zusammenhang von Schönheit und Lehrevaluierung von Professoren

Modell: Lineare Regression per WLS

- für durchschnittliche Lehrevaluierung pro Kurs (auf Skala 1–5),
- erklärt durch standardisierte Schönheitsbewertung sowie Faktoren Geschlecht, Minderheit, Tenure, u.ä.,
- gewichtet mit Anzahl Studenten pro Kurs.

Schöne Professoren

	Alle	Männer	Frauen
(Konstante)	4.216	4.101	4.027
Schönheit	0.283	0.383	0.133
Geschlecht (= w)	−0.213		
Minderheit	−0.327	−0.014	−0.279
Muttersprache $\neq$ Englisch	−0.217	−0.388	−0.288
Tenure-Track	−0.132	−0.053	−0.064
Erster Abschnitt	−0.050	0.004	−0.244
R^2	0.271	0.316	

(Bemerkung: Nur Kurse mit mehr als 1 Leistungspunkt.)

Schöne Professoren

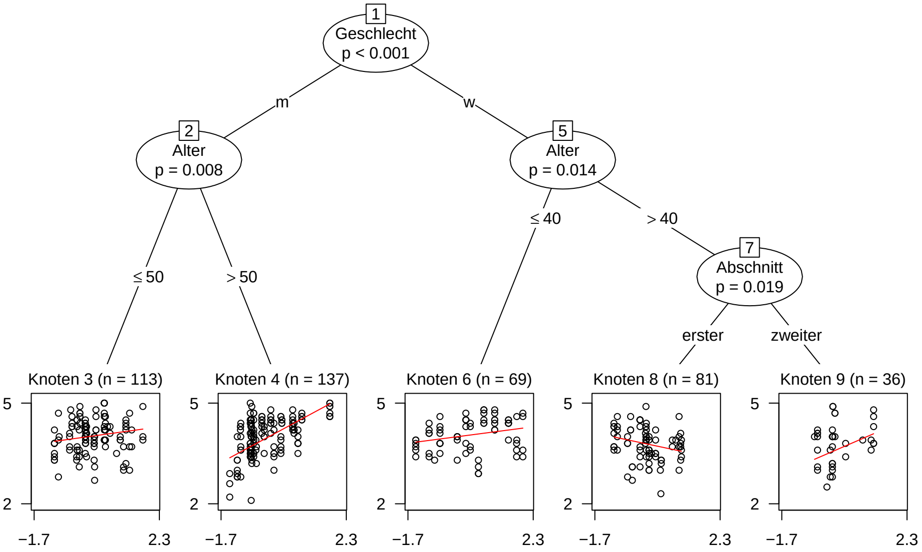
Hamermesh & Parker:

- Modell mit allen Variablen (als Haupteffekte),
- Verbesserung durch Aufteilung nach Geschlecht,
- kein Zusammenhang mit Alter (linear oder quadratisch).

Hier:

- Modell für Lehrevaluierung erklärt durch Schönheit,
- übrige Variablen als Partitionierungsvariablen,
- adaptive Ermittlung von Zusammenhang und Interaktionen.

Schöne Professoren



Schöne Professoren

Rekursive Partitionierung:

	(Konstante)	Schönheit
3	3.997	0.129
4	4.086	0.503
6	4.014	0.122
8	3.775	−0.198
9	3.590	0.403

Modellvergleich:

Modell	R^2	Parameter
gesamt	0.271	7
getrennt nach Geschlecht	0.316	12
rekursiv partitioniert	0.382	10 + 4

Software

Alle Methoden sind in verschiedenen Paketen für das Statistik-System R implementiert.

R selbst und alle Pakete können frei unter der GPL (General Public License) vom R-Archiv CRAN (Comprehensive R Archive Network) bezogen werden:

- Bäume/rekursives Partitionieren: **party**,
- Strukturbruchtests: **strucchange**,
- Bradley-Terry-Regression und Bäume: **psychotree**.

`http://www.R-project.org/`

`http://CRAN.R-project.org/`

Zusammenfassung

Schönheit liegt im Auge des Betrachters:

- Zusammenhänge oft nichtlinear und mit Interaktionen.
- Für statistische Erfassung flexible Modelle erforderlich.

Modellbasiertes rekursives Partitionieren:

- Synthese von klassischen parametrischen Datenmodellen und algorithmischen Baumverfahren.
- Anwendbar auf allgemeine Klasse parametrischer Modelle.
- Alternative zu traditioneller Modellspezifikation, insbesonders bei Variablen deren Einfluss nicht klar ist.

Ausblick auf weitere Anwendungen:

- Unterschiedliche Fragen-Schwierigkeiten in Rasch-Modellen (Beispiel: SPIEGEL Studenten-PISA).
- Unterschiedliche Treatment-Effekte in Teilgruppen.

Referenzen

Zeileis A, Hornik K (2007). “Generalized M-Fluctuation Tests for Parameter Instability.” *Statistica Neerlandica*, **61**(4), 488–508. `doi:10.1111/j.1467-9574.2007.00371.x`

Zeileis A, Hothorn T, Hornik K (2008). “Model-Based Recursive Partitioning.” *Journal of Computational and Graphical Statistics*, **17**(2), 492–514. `doi:10.1198/106186008X319331`

Strobl C, Wickelmaier F, Zeileis A (2011). “Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning.” *Journal of Educational and Behavioral Statistics*, erscheinend. URL `http://eeecon.uibk.ac.at/~zeileis/papers/Strobl+Wickelmaier+Zeileis-2010.pdf`

Strobl C, Kopf J, Zeileis A (2010). “Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben.” In Trepte S, Verbeet M (eds), *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test*, 255–272. VS Verlag, Wiesbaden.