Cross-Level-Inference

Kai Arzheimer | Vorlesung Forschungsmethoden

Übersicht

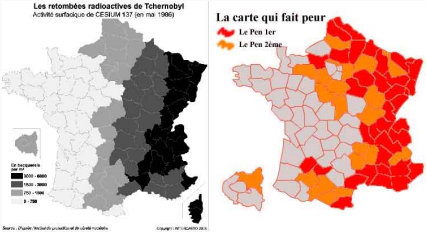
Einführung

- Grundbegriffe und -probleme
- Warum trotzdem Cross-Level Inference?

Moderne Cross-Level Inference

- $R \times C$-Tabellen
- Method of Bounds (Duncan and Davis)
- Goodman Regression (nach King et al.)
- King's Method und neue Entwicklungen
- Software

Zusammenfassung



Einführung

Analyseebenen

Ebene	Mikro		Meso		Makro		...
Typ	global	→	analytisch				
	relational	→	strukturell				
	kontextuell	←	global	→	analytisch		
			relational	→	strukturell		
			kontextuell	←	global	→	
					relational	→	
					kontextuell	←	

→: Aggregation
←: Disaggregation
Quelle: Darstellung nach Hox 2002 und Lazarsfeld/Menzel 1961

Was ist Cross-Level Inference?

- Theoretischer Zusammenhang auf einer Ebene spezifiziert (z.B Individuen)
- Daten auf anderer Ebene vorhanden (z.B. Landkreise)
- Schätzung von Parametern/Test von Hyothesen möglich?
- Cross-Level Inference

Individualistischer Fehlschluß

- Spezifisch
  - Individualkorrelationen nicht auf Makro-Ebene übertragbar
  - Individualebene: Einkommen $\leftrightarrow$ demokratische Werte
  - Makroebene: GDP $\leftrightarrow$ Demokratiequalität?
- Allgemeiner
  - Objekte auf höheren Ebenen (Systeme) können Eigenschaften haben …
  - … die mehr als eine Aggregation individueller Merkmale sind

Ökologischer Fehlschluß

- Zusammenhänge auf Individualebene *müssen* nicht Zusammenhängen auf Makro-Ebene entsprechen
- Robinson 1950:
  - Individuelle Korrelation zwischen foreign born / illiterate + 0.12
  - Auf der Ebene von 9 US-Staaten: -0.62
- Warum?

Robinson 1950 (p.340-341)

Das Ende der Aggregatanalyse?

„The relation between ecological and individual correlations which is discussed in this paper provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot.
While it is theoretically possible for the two to be equal, the conditions under which this can happen are far removed from those ordinarily encountered in data. From a practical standpoint, therefore, the only reasonable assumption is that an ecological correlation is almost certainly not equal to its corresponding individual correlation.“

Aggregatdaten in der Sozialforschung

- Entwicklung der Sozialforschung aus Polizei-, Bevölkerungs- und Wirtschaftsstatistik
- Deutschland, Frankreich (Durkheim, Siegfried)
- (Fast) kein Problembewußtsein (Key)
- Durchsetzung der surveybasierten Individualdatenanalyse, Robinson als Schlußpunkt
- Warnung vor *individualistischem* Fehlschluß (Scheuch)

Historische Anwendungen

- **Keine Individualdaten vorhanden**
- Historisch-quantitative Sozialforschung (fast) vollständig auf Aggregatdaten angewiesen
- Erkenntnisinteresse (meist) auf Mikroebene
- Beispiele
  - Hitlers Wähler
  - Wähler der SPD im Kaiserreich
  - Entwicklung von Parteibindungen im Süden der USA vor WK II
  - Historische Demographie (Wanderungsbewegungen, Heiratsquoten etc.)

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**
- Delinquenz von ethnischen Gruppen

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**
- Delinquenz von ethnischen Gruppen
- Wählerwanderungen RLP 2006 $\rightarrow$2011$\rightarrow$2016$\rightarrow$2021

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**
- Delinquenz von ethnischen Gruppen
- Wählerwanderungen RLP 2006 $\rightarrow$2011$\rightarrow$2016$\rightarrow$2021
- 2nd order election/recall über 5 Jahre?

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**
- Delinquenz von ethnischen Gruppen
- Wählerwanderungen RLP 2006 $\rightarrow$2011$\rightarrow$2016$\rightarrow$2021
- 2nd order election/recall über 5 Jahre?
- Veränderungen in Stimmbezirken, aber ...

Wählerwanderungen und mehr

- **Individualdaten nicht nutzbar/zuverlässig**
- Delinquenz von ethnischen Gruppen
- Wählerwanderungen RLP 2006 $\rightarrow$2011$\rightarrow$2016$\rightarrow$2021
- 2nd order election/recall über 5 Jahre?
- Veränderungen in Stimmbezirken, aber ...
  - Allgemeine Fehlschlußproblematik
  - Zuzug von Wählern
  - Jungwähler
  - Wähler versterben oder ziehen über Grenzen von Stimmbezirken

Moderne Cross-Level Inference

Variablen

- Grundsätzlich kontinuierliche Variablen möglich
- Einfacher für kategoriale Variablen
  - Allgemein $R \times C$ Tabellen
  - Im einfachsten Fall 2×2 Tabellen

Das Standardbeispiel (King 1997)

	T+	T-	Σ
X+	?	?	100
X-	?	?	900
Σ	500	500	1000

- Census × electoral data
- Secret ballot
- Das ganze für mehrere Stimmbezirke in einem Wahlkreis, mehrere Wahlkreise in einem Staat …
- (Möglichst kleine Einheiten, d.h. möglichst wenig Aggregation)

Terminologie

- $1, 2, \cdots, p$ Stimmbezirke (precincts)
- In jedem Stimmbezirke zwei bekannte Größen
  - Zahl bzw. Anteil der wahlberechtigten Schwarzen: X_i
  - Zahl bzw. Anteil der Wahlberechtigten, die wählen: T_i
- Und zwei unbekannte, aber interessante Größen:
  - Beteiligungsquote bei den Schwarzen β_i^b
  - Beteiligungsquote bei den Weißen β_i^w
- Ebenfalls interessant: mittlere Beteiligungsquoten: B^b und B^w
- Seit 1953 *zwei* Ansätze zur Bestimmung dieser Parameter
  - Goodman Regression
  - Method of Bounds (Duncan and Davis)

Duncan & Davis (nach King et al.)

- Ohne Kenntnis von Daten: β_i^b und β_i^w jeweils im Intervall [0;1]
- In manchen Fällen ergeben sich durch die Daten *deterministische* Schranken; entspricht 100%-Konfidenzintervall
- Bsp.:
  - 150 Schwarze in Bezirk (X_i), 87 Wähler (T_i) $\rightarrow$
  - $\beta_i^b \in [0; \frac{87}{150}]$
- Allgemein (da $T_i = X_i\beta_i^b + (1 - X_i)\beta_i^w$) (Achtung, Anteilswerte):

$$\beta_i^b \in \left[\max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right); \min\left(\frac{T_i}{X_i}, 1\right)\right] \tag{1}$$

$$\beta_i^w \in \left[\max\left(0, \frac{T_i - X_i}{1 - X_i}\right); \min\left(\frac{T_i}{1 - X_i}, 1\right)\right] \tag{2}$$

In unserem Beispiel ...

- $T_i = 0.5, X_i = 0.1$
- β_i^b:
  - Untere Schranke: 0 (da Turnout alleine von weißer Bevölkerung produziert sein könnte)
  - Obere Schranke: 1 (da beobachteter Turnout nicht überschritten, wenn alle Schwarzen wählen)
- β_i^w:
  - Untere Schranke: $\frac{0.5-0.1}{1-0.1} = \frac{4}{9}$, da sonst selbst bei hundertprozentiger Wahlbeteiligung der Schwarzen Gesamtwahlbeteiligung nicht zustandekommt
  - Obere Schranke: $\frac{0.5}{1-0.1} = \frac{5}{9}$, da sonst beobachtete Gesamtwahlbeteiligung überschritten, selbst wenn Schwarze gar nicht wählen
- Stimmbezirk informativ *bezüglich Wahlbeteiligung der Weißen*

Wann funktioniert die Method of Bounds?

- (Fast) homogene Bezirke sind informativ bezüglich der dominanten Gruppe (aber nicht bezüglich der anderen)
- Bezirke mit sehr hohem/sehr niedrigem Turnout sind interessant, weil sie das Intervall effektiv beschränken (möglicherweise für beide Gruppen)
- Gemischte Bezirke mit mittlerer Wahlbeteiligung sind nicht informativ, da β_i^w **und** β_i^b zwischen 0; 1 liegen können

Goodman Regression: Ansatz

- Schätzt durchschnittliche Wahlbeteiligung (B^b, B^w) auf Grundlage der Informationen über die Stimmbezirke
- Basiert auf „accounting identity“: $T_i = X_i \beta_i^b + (1 - X_i)\beta_i^w$
- Regression (ohne Konstante) von T_i auf Anteile von Schwarzen und Weißen in Stimmbezirken um B^b, B^w (durchschnittliche Beteiligungsraten) zu schätzen
- Unverzerrte Schätzung möglich, wenn Wahlbeteiligungsraten **nicht** mit Bevölkerungsanteilen variieren

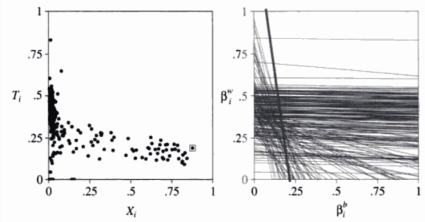
Probleme mit Goodman's Regression

- Umgebungsunabhängige Beteiligungsraten extrem unwahrscheinlich
  - Schwarze in homogen schwarzer Nachbarschaft $\rightarrow$ vermutlich depriviert, niedrige Beteiligung
  - Schwarze in ethnisch gemischter Umgebung $\rightarrow$ Integration oder Polarisation, hohe Beteiligung
  - Schwarze in fast homogen weißer Nachbarschaft $\rightarrow$ integriert oder marginalisiert, hohe oder niedrige Beteiligung
- Goodman's Regression funktioniert nur, wenn es keine kontextuellen Effekte des Merkmals X gibt
- Hinter X stehen individuelle und soziale Variablen und Prozesse $\rightarrow$ extrem unwahrscheinlich

King's (EI) method: Grundgedanke

- „Bifurkation“ der ökologischen Regression in zwei Stränge (Duncan/Davis vs. Goodman) nicht sinnvoll
- Durch Method of Bounds läßt sich ein großer Teil des Regressionsproblems lösen $\rightarrow$ erhöht die Chancen, in zweitem Schritt gute Lösung zu finden
- Kombiniert statistische und deterministische Information über das Problem

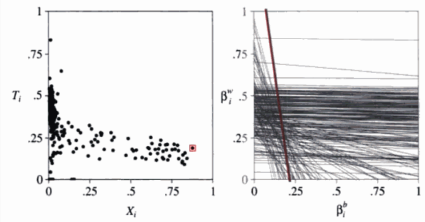
1. Schritt: Bounds



Quelle: King et al. 2004, p.5

- Goodman: T Funktion von X und β_i^w, β_i^b
- β_i^w, β_i^b über lineare Funktion miteinander verbunden

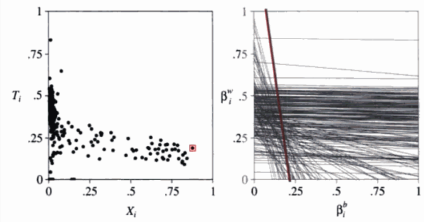
1. Schritt: Bounds



Quelle: King et al. 2004, p.5

- Goodman: T Funktion von X und β_i^w, β_i^b
- β_i^w, β_i^b über lineare Funktion miteinander verbunden
- D. h. aus möglichen Werten für ein β folgt möglicher Wert für anderes
- Linie statt Punkt als Folge der Aggregation

1. Schritt: Bounds



Quelle: King et al. 2004, p.5

- Goodman: T Funktion von X und β_i^w, β_i^b
- β_i^w, β_i^b über lineare Funktion miteinander verbunden
- D. h. aus möglichen Werten für ein β folgt möglicher Wert für anderes
- Linie statt Punkt als Folge der Aggregation

- Informationsgehalt
  - Flache Linien
  - Steile Linien
  - Linien, die Ecke unten links oder oben rechts abschneiden
  - Diagonale

2. Schritt: Statistische Analyse

- Betas für precint i als Ziehung aus einer (bzw. zwei) Verteilungen
- Schätzung präzisieren, indem Information über *alle* Fälle genutzt wird (ähnlich wie Regression)
- Annahmen
  - Kombinationen von Betas bilden *ein* Cluster (und nicht mehrere)
  - Keine räumliche Autokorrelation von T (unter Kontrolle von X)
  - X_i unabhängig von β_i^w und β_i^b

Anwendungen haltbar?

- In Anwendungen unrealistisch, Lockerungen möglich, Modell robust
- Erweiterungen, z. B. Wahlen in Mehrparteiensystemen
- Nützlichkeit des Modells nicht unumstritten
- Aktive Forschung
- Hervorragende Chancen, sich selbst in den Fuß zu schießen

Software

- Goodman's Regression im Prinzip mit jedem Statistikprogramm möglich
- Einzelne spezialisierte Programme oft als DOS (.EXE) Software (zusehends historisch interessant)
- Aktuelle Entwicklung in der Regel in R oder BUGS
- Teil von Kings Modellen via Zelig (in R) verfügbar

Anwendung: Konfession und CDU-Wahl in Rheinland-Pfalz, 1947-96

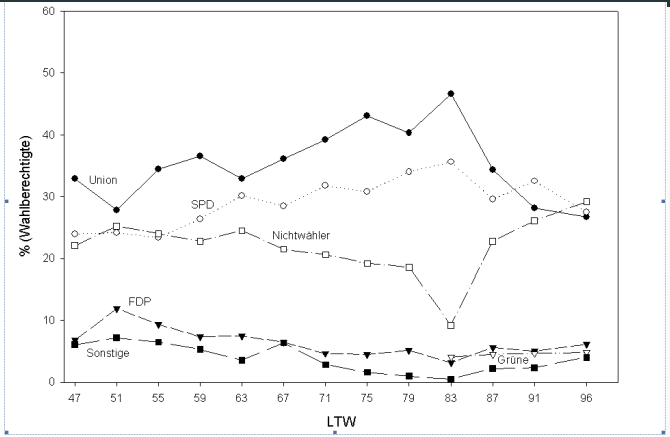


Abbildung 5: Ergebnisse der Landtagswahlen 1947 bis 1996 (Wahlberechtigte)

Anwendung: Konfession und CDU-Wahl in Rheinland-Pfalz, 1947-96

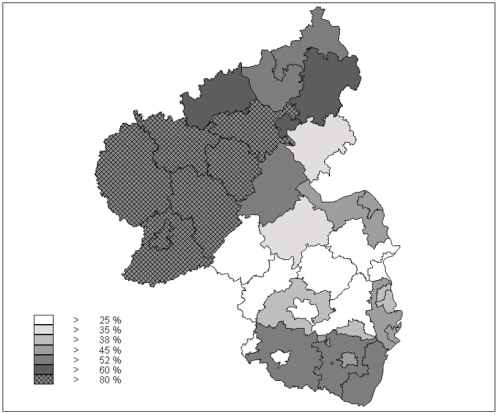


Abbildung 6: Katholikenanteil in den Kreisen und kreisfreien Städten 1987

Anwendung: Konfession und CDU-Wahl in Rheinland-Pfalz, 1947-96

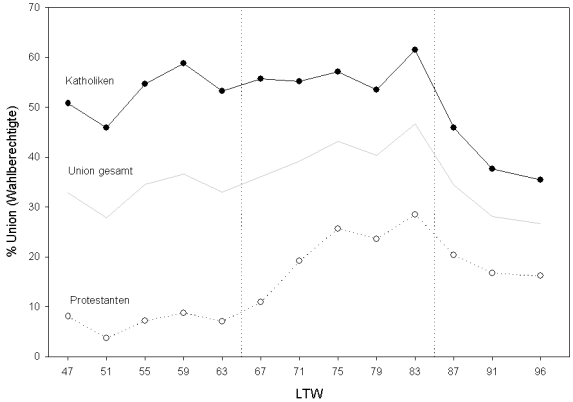


Abbildung 8: Entwicklung des konfessionellen Wahlverhaltens zugunsten der Union 1947-1996 (ökologische Regression nach King)

Zusammenfassung

Zusammenfassung

- Nach Möglichkeit Entsprechung von Niveau der theoretischen Aussagen/Niveau der Analyse
- Cross-Level Inference ist immer problematisch
- Häufigste Form: Rückschluß von Aggregatzusammenhängen auf individuelle Zusammenhänge
- Durch Aggregation geht *immer* Information verloren
- Ansätze
  1. Method of Bounds: Deterministischer Schluß, in welchem Bereich Parameter liegen müssen
  2. Goodman's Regression: Stochastischer Schluß auf mittlere Parameter, *wenn keine Kontexteffekte von X*
  3. King's Method (EI) und Erweiterungen: Nutzen möglichst viel Information, *stochastische* Aussagen über Parameter