Ludwig-Maximilians-Universität München
Fakultät für Mathematik, Informatik und Statistik
Institut für Statistik
Sommersemester 2013

Bachelor-Seminar
Nationale und internationale Sozialberichterstattung:
Wichtigste Datenquellen, zentrale Ergebnisse und method(olog)ische Probleme

Der ökologische Fehlschluss

(Vorbereitungmaterial)

Leitung: Prof. Dr. Thomas Augustin
Betreuung: Dr. Marco Cattaneo

Nina Scharrer
Matrikelnr.: 10362621
Bachelor Statistik (FS 6)
NinaScharrer@gmx.de
19.06.2013

Inhaltsverzeichnis

1 Abstract

Eine Hauptaufgabe der nationalen und internationalen Sozialberichterstattung besteht darin, ein Bild über die Lebenssituation der Menschen zu liefern und Problemlagen zu erklären. Von Interesse sind somit meist Aussagen über Personen und Zusammenhänge auf der Individualebene. Aufgrund der Vertraulichkeit und Anonymität liegen die Daten jedoch meist in aggregierter Form vor und es benötigt Methoden der ökologischen Inferenz, um Rückschlüsse auf eine niedrigere Untersuchungsebene ziehen zu können. Hinter dieser Strategie steht das Problem, dass viele unterschiedliche Datenkonstellationen auf der Individualebene das auf der Aggregatebene beobachtete Muster erzeugen können. Ein empirisch bestätigter Zusammenhang zweier Merkmale auf der Aggregatebene kann daher auch auf der Individualebene bestehen, häufig ist dies aber nicht der Fall und man würde bei der Übertragung auf die Individualebene einen ökologischen Fehlschluss begehen. Für die Problematik des ökologischen Fehlschlusses gibt es keine Lösung, aber es wurden statistische Verfahren entwickelt, um die Zusammenhänge auf der individuellen Ebene zu schätzen. Besonders etabliert haben sich dabei die Methode der Ränder von Duncan und Davis (1950), Goodmans ökologische Regression (1950) und das Ecological Inerence Model von King (1997).

2 Einleitung- Untersuchungsebenen und ökologische Inferenz

In der nationalen wie der internationalen Sozialberichterstattung treten häufig methodologische Probleme auf, welche auf die unterschiedlichen Untersuchungsebenen zurückzuführen sind, auf denen Daten erhoben werden können. Man unterscheidet hierbei zwischen Untersuchungseinheiten auf Individual- und Aggregatebene. Informationen über Personen beziehungsweise individuelle Merkmale wie Alter, Einkommen, Wahlabsicht oder Geschlecht (Gehring and Weins, 2010, S.18) sind Individualdaten, da sie einem einzelnen Element der Stichprobe zugeordnet werden können. Werden alle Daten einer Befragung durch Summierung, Bildung von Durchschnitts- oder Anteilswerten oder mit Hilfe anderer Rechenoperationen zu einer größeren Einheit zusammengefasst, so erhält man Aggregatdaten. Diese beinhalten folglich Informationen über eine Gruppe, oder ein Kollektiv. Aussagen über die einzelnen Mitglieder lassen sich nicht mehr problemlos treffen. Damit birgt die Aggregation einen deutlichen Informationsverlust, besonders, da eine Disaggregation der Daten nicht möglich ist. In vielen Fällen ist die Aggregation daher eine bewusst eingesetzte Methode zur Anonymisierung von Individualdaten.
Gerade in der amtlichen Statistik, also auch bei der Sozialberichterstattung, spielt das Verfahren eine wichtige Rolle und es stehen folglich häufig nur Aggregatdaten zur Verfügung. Im Mittelpunkt des Analyseinteresses stehen hingegen meist Zusammenhänge und Prozesse auf der Individualebene. Aussageeinheit und Untersuchungseinheit sind also auf unterschiedlichen Ebenen angesiedelt. Eine Beziehung zwischen Merkmalen auf der Aggregatebene kann durchaus auf die Individualebene übertragbar sein, der Zusammenhang kann sich aber genauso gut umkehren oder gar nicht existieren. (Gschwend, 2006, S.227) Grund dafür ist, dass viele unterschiedliche Datenkonstellationen auf der Individualebene das auf der Aggregatebene beobachtete Muster erzeugen können. (Gschwend, 2006, S.227)
Diese Problematik war schon zu Beginn des 20. Jahrhunderts bekannt und es wurde vereinzelt ökologischen Fragestellungen nachgegangen und Analysen zu dieser Thematik durchgeführt. Doch erst als William S. Robinson im Jahr 1950 seinen Artikel „Ecological Correlations and the Behavior of Individuals“ veröffentlichte, rückte das Problem der ökologischen Inferenz in den Mittelpunkt des wissenschaftlichen Interesses. Robinson bewies mit Hilfe der Gleichung der Kovarianzanalyse, dass man von ökologischer Korrelation, also einer Korrelation bei der die Merkmalsträger Gruppen von Personen sind, wobei als Gruppierungskriterium regionale Einheiten verwendet werden, nicht direkt auf individuelle Korrelation, also eine Korrelation zwischen Variablen, deren Merkmalsträger Individuen sind, schließen kann. Wenn also Zusammenhänge auf der Individualebene mit aggregierten Daten überprüft werden sollen, da Individualdaten nicht zugänglich, nicht vorhanden, oder unzuverlässig sind, ist die direkte Übertragung der Ergebnisse auf die andere Analyseebene meist falsch und man benötigt stattdessen Methoden der ökologischen Inferenz. (Gschwend, 2006, S.227) Robinsons Schlussfolgerung besagt jedoch, dass mit den vorhandenen Methoden der damaligen Zeit

keine ökologische Inferenz möglich sei, um einen sogenannten „ökologischen Fehlschluss“ zu vermeiden, welcher immer dann entsteht, wenn ein auf der Aggregatebene festgestellter Zusammenhang zwischen zwei Merkmalen auf der Individualebene nicht besteht. (Klima, 2011, S.52)

3 Der ökologische Fehlschluss

Der Begriff „ökologischer Fehlschluss“ leitet sich von den ökologischen Daten ab, worunter man Daten versteht, die über geographische Gebiete aggregiert wurden. „Ökologisch“ steht in diesem Fall also für „kollektiv“. Die von Robinson eingeführte Bezeichnung hat sich etabliert und beschreibt den falschen Schluss, von einer auf der Aggragatebene empirisch belegten Beziehung auf die entsprechende Beziehung zwischen Individuen. Es wird also zu Unrecht ein Zusammenhang auf der Individualebene angenommen, der auf der Aggregatebene errechnet wurde.
Zur Veranschaulichung der Problematik dient an dieser Stelle das häufig zitierte Beispiel über den Zusammenhang des Ausländeranteils und des Abschneidens rechtsextremer Parteien in einem Wahlbezirk. Häufig findet man eine positive Korrelation zwischen den beiden Merkmalen auf Aggregatebene. Das bedeutet, dass rechtsextreme Parteien in Bezirken mit hohem Ausländeranteil bessere Wahlergebnisse erzielen. Es ist jedoch ein offensichtlicher ökologischer Fehlschluss daraus zu folgern, dass Ausländer Wähler solcher Parteien sind. Die positive Korrelation auf Aggregatebene kommt dadurch zustande, dass wahlberechtigte Nicht-Ausländer in Wahlkreisen mit hohem Ausländeranteil besonders häufig rechtsextreme Parteien wählen.

4 Formalisierung und Notation

Wie beschrieben stellt die ökologische Inferenz den Versuch dar, von Daten einer höheren Aggregatebene auf das Verhalten in einer niedrigeren Aggregatebene zu schließen. Um dieses Problem zu formalisieren, fasst man die aggregierten Daten in einer Kreuztabelle zusammen, wobei die bekannten Informationen den Randwerten entsprechen. Ziel ist es, die unbekannten Werte der inneren Zellen zu bestimmen.
Als Beispiel soll hier das unterschiedliche Wahlverhalten ethnischer Gruppen im US- Bundesstaat Ohio untersucht werden. (vgl. King, 1997) Die betrachteten dichotomen Variablen auf der Individualebene sind also Ethnizität mit den Ausprägungen „Afroamerikaner“, „Weißer“ und Wahlteilnahme mit den Ausprägungen „ja“, „nein“. Jedes Individuum könnte man so genau einer der inneren Zellen zuordnen, doch liegen die Daten hier nur in aggregierter Form vor. Für jeden Wahlbezirk i ist also nur bekannt, wie viele Afroamerikaner beziehungsweise Weiße dort leben und wie viele Personen zur Wahl gegangen sind und wie viele nicht. Meist wird aber nicht mit diesen absoluten Zahlen gearbeitet, sondern mit den jeweiligen

Anteilen an der Gesamtbevölkerung. In diesem Fall stehen also in den Rändern der bekannte Anteil Afroamerikaner X_i und der bekannte Anteil der Wähler Y_i im Wahlkreis i. Da die beiden Variablen die Bevölkerung innerhalb eines Wahlkreises partitionieren, also jedes Individuum jeweils entweder der einen oder der anderen Ausprägung eines Merkmals entspricht und somit die Addition der einzelnen Zellen die Grundgesamtheit ergibt, sind folglich auch die Gegenwahrscheinlichkeiten $1 - X_i$ und $1 - Y_i$ bekannt, welche für den Anteil der weißen Bevölkerung beziehungsweise den Anteil der Nichtwähler in einem Wahlkreis i stehen. Damit ist die Randverteilung der Vierfeldertafel bekannt:

	Wahl	Nicht-Wahl	
Afroamerikaner	β_i^a	$1 - \beta_i^a$	X_i
Weiße	β_i^w	$1 - \beta_i^w$	$1 - X_i$
	Y_i	$1 - Y_i$	

Tabelle 1: Inferenzproblem als Vierfeldertafel

Über den Anteil der Wähler in den Gruppen der Afroamerikaner β_i^a beziehungsweise der Weißen β_i^w liegen hingegen keine Informationen vor. Wir wissen nur, dass die beiden gesuchten Parameter β_i^a und β_i^w im Intervall [0, 1] liegen müssen, da es sich um Anteilswerte handelt. Wichtig an dieser Stelle ist festzuhalten, dass sich in diesem Beispiel die Ränder, nicht wie üblich, aus den Zeilen- und Spaltensummen ergeben. Das liegt daran, dass es sich bei β_i^a und β_i^w um Anteilswerte aus unterschiedlichen Grundgesamtheiten handelt. β_i^a repräsentiert den Anteil der Wähler in der Gruppe der Afroamerikaner, die Grundgesamtheit sind damit alle wahlberechtigten Afroamerikaner in diesem Wahlbezirk. β_i^w hingegen steht für den Anteil der Wahlbeteiligung unter den Weißen. Die Grundgesamtheit sind damit alle weißen Wahlberechtigten in dem Bezirk. Damit gilt im Allgemeinen $\beta_i^a + \beta_i^w = 1$ nicht. Stattdessen gilt aufgrund der Partitionierung, also der Tatsache, dass sich die Wählerschaft in jedem Wahlkreis i anteilig aus Afroamerikanern und Weißen zusammensetzt (Gschwend, 2006, S.228), der Zusammenhang

$$Y_i = \beta_i^a \cdot X_i + \beta_i^w \cdot (1 - X_i). \tag{1}$$

Für jeden Wahlkreis gibt es also genau eine Gleichung mit zwei unbekannten Parametern. Das sind zu wenige Informationen, um die inneren Zellwerte der Kreuztabelle eindeutig zu bestimmen. Würden wir beispielhaft von einem Bezirk i ausgehen, bei dem der Anteil der Afroamerikaner an der Wählerschaft 20 Prozent beträgt ($X_i = 0,2$) und die Gesamtwahlbeteiligung 60 Prozent beträgt ($Y_i = 0,6$), dann wäre es möglich, dass die Hälfte aller Afroamerikaner dieses Bezirks ($\beta_i^a = 0,5$) und 62,5 Prozent aller Weißen ($\beta_i^w = 0,625$) zur Wahl gegangen sind. (vgl. Tabelle 2) Es wäre aber genauso gut möglich, dass bei unverändertem X_i und Y_i, nur jeder vierte Afroamerikaner ($\beta_i^a = 0,25$), dafür aber 11 von 16 Weißen ($\beta_i^w = 0.6875$) an der Wahl Teil genommen haben. (vgl. Tabelle 3)
Da die eindeutige Bestimmung der gesuchten Parameter nicht möglich ist, bleibt nur, durch statistische Verfahren die Zusammenhänge der individuellen Ebene zu schätzen. Dabei müs-

	Wahl	Nicht-Wahl	
Afroamerikaner	$0,5$	$0,5$	**0,2**
Weiße	$0,625$	$0,375$	**0,8**
	0,6	**0,4**	**1**

Tabelle 2: Fundamentale Unbestimmtheit (Möglichkeit 1)

	Wahl	Nicht-Wahl	
Afroamerikaner	$0,25$	$0,75$	**0,2**
Weiße	$\frac{11}{16}$	$\frac{5}{16}$	**0,8**
	0,6	**0,4**	**1**

Tabelle 3: Fundamentale Unbestimmtheit (Möglichkeit 2)

sen häufig zusätzliche Annahmen getroffen werden oder es können nur bestimmte Bereiche ermittelt werden, in denen der wahre Wert liegen muss. Ohne Hinzuziehung solcher Verfahren sollten sich Aussagen immer nur auf die Ebene der Untersuchungseinheit beziehen.

5 Methoden der ökologischen Inferenz

Auch wenn die Verwendung von ökologischen Daten in den Sozialwissenschaften nach Robinsons Artikel 1950 zunächst zurückging, so bewirkte die Publikation nach einiger Zeit, dass mehr denn je versucht wurde, neue Durchbrüche in der ökologischen Inferenz zu erzielen, um doch noch Schlüsse von ökologischen Daten auf individuelles Verhalten ziehen zu können. So wurden viele neue Ansätze für Schätzverfahren für die interessierenden Parameter β_i^a und β_i^w entwickelt, doch nur wenige konnten sich auch unter Verwendung wahrer Daten behaupten. (King, 1997, S.8) 1953 entstanden schließlich zwei konkurrierende Ansätze, welche über einen langen Zeitraum das Standardinstrument der ökologischen Inferenz wurden: Duncan und Davis stellten ihre Methode der Ränder vor, Goodman seine ökologische Regression. (Klima, 2011, S.54) 1997 veröffentlichte schließlich Gary King sein Buch „A solution to the Ecological Inference Problem“, in dem er seine Methode der ökologischen Inferenz vorstellte. Dabei verband er die Methode der Ränder von Duncan und Davis mit einem Regressionsansatz und regte die methodische Diskussion damit neu an. (Klima, 2011, S.55)

5.1 Methode der Ränder von Duncan und Davis

Ziel dieser Methode ist es, die internen Zellenbesetzung anhand der minimal und maximal möglichen Ausprägungen der Anteilswerte einzuschränken. Um diese Randwerte der Bandbreite zu bestimmen, wird (1) nach β_i^w aufgelöst und wir erhalten die Gleichung

$$\beta_i^w = \frac{Y_i}{1 - X_i} - \frac{X_i}{1 - X_i} \cdot \beta_i^a \tag{2}$$

Für jeden Wahlkreis ergibt sich also eine Gerade, bei der Achsenabschnitt ($\frac{Y_i}{1-X_i}$) und Steigung ($-\frac{X_i}{1-X_i}$) bekannt sind. Da $X_i \in$ [0,1] ist, bleibt das negative Vorzeichen der Steigung stets erhalten. Folglich ist β_i^a maximal, wenn $\beta_i^w = 0$ und minimal, wenn $\beta_i^w = 1$. β_i^w nimmt seinen Maximalwert an, wenn $\beta_i^a = 0$ und wird minimal, wenn $\beta_i^a = 1$. (vgl. Rechnung (1), (2) im Anhang) Auf das Beispiel der Wählerschaft in Ohio übertragen bedeutet dies, dass der Anteil der Wähler in der Gruppe der Afroamerikaner maximal wird, wenn unter den Weißen keiner zur Wahl geht. Minimal hingegen wird dieser Anteil, wenn jeder der wahlberechtigten weißen Bevölkerung seinen Stimmzettel abgibt. Dieser Zusammenhang lässt sich auch auf den Anteil der Wähler in der weißen Bevölkerung übertragen.
Schließlich ergibt sich allgemein für die Schranken nach der Methode der Ränder folgende Gleichungen:

$$\beta_i^a \in [max[(0, \frac{Y_i - (1 - X_i)}{X_i}), min(\frac{Y_i}{X_i}, 1)] \tag{3}$$

$$\beta_i^w \in [max[(0, \frac{Y_i - X_i}{1 - X_i}), min(\frac{Y_i}{1 - X_i}, 1)] \tag{4}$$

Für das obige Zahlenbeispiel erhält man damit
$\beta_i^a \in [0; 1]$
$\beta_i^w \in [0,5; 0,75]$
(vgl. Rechnung (3) im Anhang)
Der Anteil der Wähler in der Gruppe der Afroamerikaner lässt sich also in diesem Fall durch die Methode der Ränder nicht weiter eingrenzen. Es wäre möglich, dass keiner der Afroamerikaner zur Wahl gegangen ist, dass die gesamte schwarze Bevölkerung an der Wahl teil nahm, aber auch jedes Zwischenergebnis wäre möglich. In diesem Wahlbezirk wäre somit über die Wahlbeteiligung der Afroamerikaner anhand der Methode der Ränder keine genauere Aussage möglich. Bei der weißen Bevölkerung hingegen legt die Methode der Ränder die Untergrenze auf $0,5$ fest. Das bedeutet, dass mindestens 50 Prozent der Weißen zur Wahl gegangen sein müssen, da sonst die Gesamtwahlbeteiligung nicht zustande käme, selbst wenn alle Afroamerikaner zur Urne gegangen wären. Die Obergrenze liegt bei $0,75$, das heißt, dass nicht mehr als 75 Prozent der Weißen gewählt haben können, da sonst die beobachtete Gesamtwahlbeteiligung überschritten würde, selbst wenn keiner der Afroamerikaner gewählt hätte. Somit schränkt die Methode der Ränder in diesem beispielhaften Wahlbezirk den in-

teressierenden Anteil der weißen Wähler auf das Intervall $[0,5; 0,75]$ ein, indem der wahre Wert liegen muss. Obwohl also in diesem fiktivem Wahlkreis i nur Aggredatdaten in Form der anteiligen Wahlbeteiligung und Aufsplittung in Afroamerikaner und Weiße gegeben sind, lässt sich mit Hilfe der Methode der Ränder angeben, dass zwischen 50 und 75 Prozent der weißen Bevölkerung in diesem Wahlkreis ihre Stimme abgegeben haben.

5.1.1 Tomographien

Die Methode der Ränder lässt sich grafisch durch sogenannte Tomographien darstellen. Dabei wird die bereits zuvor hergeleitete Geradengleichung (2) in ein Einheitsquadrat $[0,1] \times [0,1]$ eingezeichnet. Es ergibt sich also für jeden Wahlkreis eine Gerade mit spezifischen Achsenabschnitt und stets negativer Steigung, auf der jede gültige Parameterkombination liegt. Je länger also die Gerade im Einheitsquadrat, desto mehrere Möglichkeiten gibt es für β_i^a und β_i^w und desto weniger Informationen kann somit mit Hilfe der Methode der Ränder gewonnen werden. Eine Diagonale (E) bedeutet also minimalen Informationsgehalt, da sowohl β_i^a, als auch β_i^w das gesamte Intervall von $[0,1]$ einnehmen können. Verlaufen die Linien hingegen über Ecken (A und B), bedeutet dies, dass beide Parameter auf ein kleinere Intervall eingeschränkt werden können. In Abbildung (C) liegt ein hoher Informationsgehalt für β_i^a vor, wohingegen β_i^w alle Werte zwischen 0 und 1 annehmen kann, in Abbildung (D) ist der umgekehrt Fall gegeben. (Gschwend, 2006, S.230)

plots.pdf

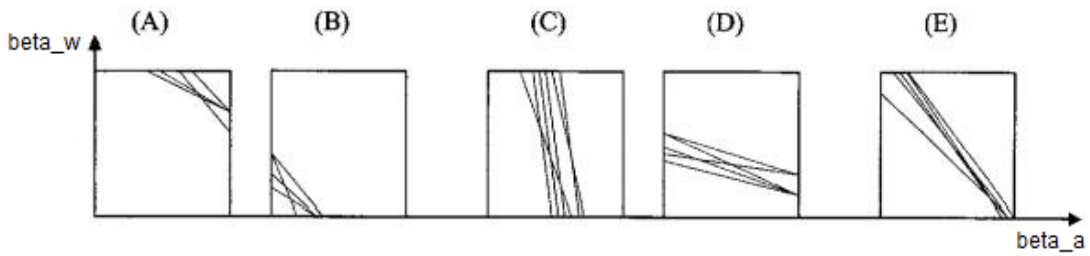
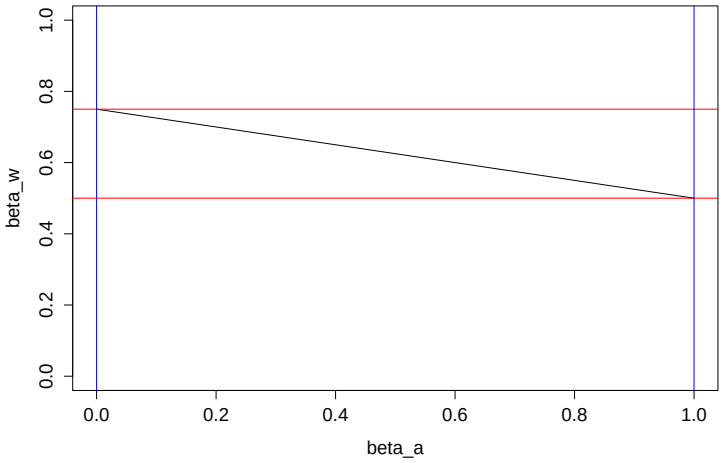


Abbildung 1: Idealtypische Muster von Tomographien, Quelle: (Gschwend, 2006, S.229)

Für das Zahlenbeispiel mit $X_i = 0,2$ und $Y_i = 0,6$ ergibt sich die Gleichung
$\beta_i^w = 0,75 - 0,25 \cdot \beta_i^a$ und somit die folgende grafische Darstellung:



Tomo R.pdf

Abbildung 2: Zahlenbeispiel als Tomographie

Die schwarze Gerade mit der negativen Steigung repräsentiert alle möglichen Kombinationen für die Parameter β_i^a und β_i^w. Die vertikalen, blauen Linien begrenzen den Wertebereich für β_i^a, die horizontalen, roten Linien schränken die Möglichkeiten für β_i^w auf das Intervall $[0,5; 0,75]$ ein. Die Schnittpunkte der Gerade mit dem Einheitsquadrat entsprechen also den ermittelten Duncan-Davis Rändern.

5.1.2 Streu-Kreuz-Diagramm

Neben Tomographien stellt auch die Abwandlung eines einfachen Streudiagramms von Y_i auf X_i eine weitere Möglichkeit dar, den Informationsgehalt der geschätzten Ränder grafisch abzubilden. Dabei wird ein Einheitsquadrat durch die beiden Diagonalen in informative und weniger informative Bereiche eingeteilt. Für Wahlkreise, die im westlichen Sektor des sogenannten Streu-Kreuz-Diagramm liegen, liefert die Methode der Ränder viel Informationen über β_i^w. β_i^a wird durch die Methode hingegen nicht eingeschränkt. Im östlichen Sektor gilt genau der gegenteilige Informationsgehalt. Wahlkreise im südlichen Sektor haben von oben eingeschränkte Ränder für β_i^w und β_i^a, wohingegen im nördlichen Sektor beide Parameter von unten eingeschränkt werden. (Gschwend, 2006, S.231)

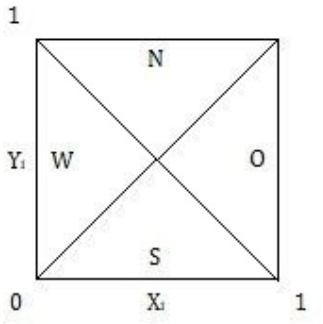


Abbildung 3: Streukreuzdiagramm, Quelle: (Gschwend, 2006, S.231)

Auf das Beispiel der Ethnizität und der Wählerschaft übertragen, wäre somit auf der x-Achse des Streudiagramms der Anteil der Afroamerikaner abgetragen und auf der y-Achse die anteiligen Wahlbeteiligung. Jeder Punkt im Streudiagramm würde dabei einen bestimmten Wahlbezirk in Ohio repräsentieren und aufgrund der jeweiligen Lage im Einheitsquadrat ließe sich sofort eine Aussage darüber treffen, ob die Methode der Ränder die beiden Parameter β_i^a (Anteil der Wähler in der Gruppe der Afroamerikaner) und β_i^w (Anteil der Wähler in der Gruppe der Weißen) einschränkt.
Der beispielhaften Wahlbezirk i, in dem 20 Prozent der Wähler Afroamerikaner sind und und insgesamt eine Wahlbeteiligung von 60 Prozent gegeben ist, würde einem Punkt im westlichen Sektor des Streudiagramms entsprechen, da durch die Methode der Ränder viel Information über den Parameter β_i^w gegeben ist, während β_i^a nicht eingeschränkt wird.

5.2 Goodmans ökologische Regression

Im Gegensatz zu Duncan und Davis Methode der Ränder war Goodmans Lösungsvorschlag zum Inferenzproblem ein Regressionsansatz. Ausgangspunkt bildet auch hier der bereits zuvor erläuterte Zusammenhang $Y_i = \beta_i^a \cdot X_i + \beta_i^w \cdot (1 - X_i)$ (1). Während die Methode der Ränder keine zusätzlichen Informationen oder Annahmen voraussetzt, muss bei Goodmans Regression die sogenannte Konstanz-Annahme getroffen werden. Diese besagt, dass Schwankungen zwischen den einzelnen Bezirken unsystematisch sind und daher von einem spezifischen Fehlerterm ϵ_i absorbiert werden. Da die Region also laut Annahme keinen Einfluss auf den Zusammenhang der beiden Merkmale haben darf, sind die Parameter β_i^a und β_i^w für jeden Bezirk im Durchschnitt gleich und es gilt somit:
$\mathbb{E}(\beta_i^a) = \beta^a$ und $\mathbb{E}(\beta_i^w) = \beta^w$. (Gschwend, 2006, S.229)
Damit ergibt sich die neue Gleichung

$$Y_i = \beta^a \cdot X_i + \beta^w \cdot (1 - X_i) + \epsilon_i, \tag{5}$$

bei der nun ein Fehlerterm ϵ_i aufgenommen wird und der Index i bei den Betas entfällt, da die beiden Parameter β^a und β^w unabhängig vom Kreis i sind. Sie werden für jeden Wahlkreis, sogar für jedes Individuum, als gleich angenommen. Gleichung (5) lässt sich schließlich einfach in das gewohnte Regressionsformat mit Intercept umformen (vgl. Rechnung (4) im Anhang) und man erhält:

$$Y_i = \beta^w + (\beta^a - \beta^w) \cdot X_i + \epsilon_i \tag{6}$$

Diese Modellformel des einfachen Goodman-Modells kann schließlich mit Hilfe einer linearen Regression unverzerrt geschätzt werden. Aus der Formel ist ersichtlich, dass die Parameter direkt interpretierbar sind. Sie stellen einerseits die individuellen Wahrscheinlichkeiten und gleichzeitig die erwartete relative Häufigkeit in der Population eines Keises dar. (Klima, 2011, S.73)

Um die Interpretation der Parameter zu verdeutlichen, soll an dieser Stelle das fiktive Zahlenbeispiel aus Ohio weiter vertieft werden. Gegeben sei noch immer $X_i = 0,2$ und $Y_i = 0,6$. Ergäbe sich beispielsweise bei einer linearen Regression für einen Wahlkreis $\beta^w = 0,65$, so ließe sich daraus zum einen schließen, dass die Wahlbeteiligung in der weißen Bevölkerung in diesem Wahlkreis 65 Prozent beträgt, zum anderen bedeutet dieser Wert, dass die individuelle Wahrscheinlichkeit eines Weißen zur Wahl zu gehen, in diesem Bezirk 65 Prozent beträgt. Es ließe sich hier also von aggregieren Daten auf individuelles Verhalten schließen, aber nur, wenn die Annahmen des linearen Modells erfüllt sind. Um berechenbare Schätzer zu erhalten, dürfen die Regressoren nicht zu stark miteinander linear abhängen. Diese sogenannte Multikollinearitätsannahme spielt bei der Methode der Ränder aber keine Rolle, da wir nur einfache Modelle mit einer Einflussgröße X_i betrachten. Eine weitere Annahme des linearen Modells ist die Homoskedastizität, welche besagt, dass die Varianz der Störterme ϵ_i nicht von der erklärenden Variable X_i abhängen dürfen. Zudem muss noch von einer Normalverteilung der Residuen und einem annähernd linearen Zusammenhang zwischen X_i und Y_i ausgegangen werden können. Erst bei Gültigkeit dieser Annahmen ist die Anwendung des linearen Modells, welches ja die Grundlage der ökonomischen Regression bildet, sinnvoll. Das größte Problem der ökologischen Regression ist jedoch die Gültigkeit der Konstanz-Annahme. Nur wenn diese erfüllt ist, erhält man interpretierbare Parameter und der Schluss von der aggregierten auf die Individualebene ist möglich. Da keine individuellen Daten gegeben sind, muss die Annahme theoretisch gerechtfertigt werden. Um also Goodmans Modell auf das Beispiel der ethischen Unterschiede beim Wahlverhalten anwenden zu dürfen, muss man davon ausgehen können, dass in jedem Wahlbezirk von Ohio der Anteil der Wähler in der afroamerikanischen und der weißen Bevölkerung ungefähr gleich groß ist. In Wahrheit wird das Wahlverhalten jedoch mit Sicherheit von der Umgebung geprägt. Es liegt die Vermutung nahe, dass Schwarze in nahezu weißer Nachbarschaft stärker integriert sind und daher eher an der Wahl teilnehmen, als Afroamerikaner in homogener Nachbarschaft. Damit wäre die Annahme nicht erfüllt und das Verfahren eigentlich nicht anwendbar. Ignoriert man an dieser Stelle die Ungültigkeit der Annahme, so können die mit dem Modell berechneten

Parameterschätzer unrealistische Werte, nämlich kleiner Null oder größer Eins, annehmen. Grund dafür ist der vorliegende Aggregationsbias, welcher auftritt, wenn die Parameter β^a und β^w von einer der Kovariablen X_i oder Y_i abhängen. (Klima, 2011, S.82) Das wäre hier zum Beispiel der Fall, wenn es eine hohe Wahlbeteiligung afroamerikanischer Wähler nur in Wahlkreisen gäbe, in denen es viele afroamerikanische Wahlberechtigte gibt. (Gschwend, 2006, S.231) Auf diesen Kritikpunkt ging insbesondere Freedman in seinem Nachbarschaftsmodell ein, welches auf der Annahme basiert, dass innerhalb der Wahlkreise die Ethnizität keinen Einfluss auf das Wahlverhalten hat und einzig der Absicht dient, das Vertrauen in Goodmans Regressionsmodell zu schwächen. (King, 1997, S.43) Doch Goodman war sich von vornherein über die Bedeutung und der damit verbundenen Probleme der Annahmen in seinem Modell bewusst und warnte davor, sein Modell nicht anzuwenden, solange die Annahmen nicht gegeben sind. Trotz dieser Empfehlung wurde Goodmans ökologische Regression häufig genutzt und wurde damit für eine lange Zeit zum Standardmodell für ökologische Inferenz-Probleme. (Klima, 2011, S.87)

5.3 Kings Ecological-Inference Modell

Im Jahr 1997 stellte schließlich Gary King seinen Lösungsvorschlag für die ökologische Inferenz vor, welcher die beiden zuvor beschriebenen Verfahren kombiniert. Kings sogenanntes Ecological-Inference Modell (kurz: EI-Modell) schätzt, genauso wie die ökologische Regression von Goodman, die interessierenden Parameter β^a und β^w, jedoch ist für die Schätzung nur ein bestimmter Wertebereich zugelassen, der durch die Methode der Ränder bestimmt wird. Dieser Methode liegt zu Grunde, dass die Parameter β_i, nicht wie bei Goodman gleich sind, sondern in den verschiedenen Wahlkreisen ähnlich und somit innerhalb fester Grenzen zufällig um einen gemeinsamen Wert schwanken. Dabei ist die Abweichung von diesem Wert in einem bestimmten Wahlkreis i bestimmbar und es lassen sich zugehörige Konfidenzintervalle aufstellen. (Gschwend, 2006, S.230) Unlogische Lösungen können sich bei diesem Verfahren nicht ergeben, so lange die folgenden Modellannahmen erfüllt sind:

1. Ähnlichkeitsannahme
   Die Parameter β_i sollen in allen Wahlkreisen i ähnlich sein, also einer gemeinsamen Verteilung entspringen. β_i^a und β_i^w sind innerhalb bestimmbarer Grenzen normalverteilt und (β_i^a, β_i^w) folglich trunkiert bivariat normalverteilt.

2. Es darf kein Aggregationsbias vorliegen.

3. Es darf keine räumliche Abhängigkeit zwischen den Kreisen i bestehen.

(Gschwend, 2006, S.230)

Beim EI-Modell werden in einem ersten Schritt die Parameter der trunkierten bivariaten Normalverteilung, nämlich Mittelwert-Vektor und Varianz-Matrix von (β_i^a, β_i^w) mit Maximum Likelihood geschätzt. Es gilt:

$$\mathbb{E}\begin{pmatrix}\beta_i^a \\ \beta_i^w\end{pmatrix} = \begin{pmatrix}\mathbb{B}^a \\ \mathbb{B}^w\end{pmatrix} = \mathbb{B} \qquad \text{und} \qquad \mathbb{V}\begin{pmatrix}\beta_i^a \\ \beta_i^w\end{pmatrix} = \begin{pmatrix}\sigma_a^2 & \sigma_{aw} \\ \sigma_{aw} & \sigma_w^2\end{pmatrix} = \Sigma \tag{7}$$

Damit ergibt sich für jeden Wahlkreis i:

$$\beta_i^a = \mathbb{B}_a + \epsilon_{i,a} \qquad \text{und} \qquad \beta_i^w = \mathbb{B}_w + \epsilon_{i,w} \tag{8}$$

Diese Formeln drücken formal aus, dass die wahren Parameter einen Kreises um den Erwartungswert über alle Kreise schwanken. Jeder Kreis hat dabei seinen eigenen Fehler um diesen Erwartungswert. Die Erwartungswerte $\mathbb{B}_a$ und $\mathbb{B}_w$ entsprechen dabei aber nicht den gesuchten Parametern β^a und β^w, sondern dem Mittel über die Kreise. (Klima, 2011, S.89) Nachdem nun also die geschätzte trunkierte bivariate Normalverteilung aufgestellt wurde, werden schließlich in einem zweiten Schritt die interessierenden Parameter auf Kreisebene, β_i^a und β_i^w, und ihre Konfidenzintervalle durch (bayesianische) Simulation ermittelt. Die Software, die eigens zur Schätzung von Kings Modell entwickelt wurde, steht kostenlos auf seiner Homepage zur Verfügung.
In seinem Buch „A Solution to the Ecological Inference Problem“ verifizierte King sein Modell mit Hilfe eines Datensatzes, der das Wahlverhalten von Afroamerikanern und Weißen in 275 Wahlbezirken aus vier südlichen U.S. Staaten umfasst, denn es waren hierbei auch die wahren Werte der zu schätzenden Parameter β_i^a und β_i^wbekannt, da in diesen südlichen Staaten die Ethnizität bei Abgabe der Wahlzettel erfasst wurde. Folglich gilt diese Methode im Vergleich zu den anderen Vorgehensweisen als Evolution in der ökologischen Inferenz. Dennoch ist sie nur problemfrei anzuwenden, wenn die drei Bedingungen gegeben sind. Gute Schätzwerte für die Parameter erhält man folglich nur, wenn kein Aggregationsbias, keine räumliche Korrelation, keine schwerwiegende Verletzung der Verteilungsannahme und dazu die grundlegende Vermutung der Ähnlichkeit der Kreise bei den untersuchten Daten vorliegt (Klima, 2011, S.109). Bei reellen Daten treten aber meist die Verletzung einer oder mehrerer dieser Modellannahmen auf und es lässt sich daher keine Aussage darüber treffen, wie gut die Punktschätzer und die ermittelten Konfidenzintervalle tatsächlich sind.

6 Fazit und Ausblick

Während die Methode der Ränder nur ein Intervall liefert, in dem die interessierenden Parameter liegen, ergeben sich bei Goodmans ökologischer Regression und Kings EI Modell für β^a und β^w exakte Werte. Doch die modellbasierten Ansätze hängen stark von ihren Annahmen ab und es ist noch nicht bekannt, inwiefern diese Methoden gegenüber solchen Annahmeverletzungen robust sind. Auch wenn das King'sche Modell einen großen Fortschritt in der Problematik der ökologischen Inferenz darstellt, kann dieses Verfahren nicht als die endgültige Lösung und damit das methodologische Problem als gelöst angesehen werden. Kings Ecological-Inference Modell ist höchstens eine Lösung. Aufgrund des hohen Informationsverlustes, der auf die Aggregation der Individualdaten zurückzuführen ist und nie wieder ausgeglichen werden kann, ist eine eindeutige Lösung des Inferenzproblems nicht möglich. Dennoch versuchen Wissenschaftler ununterbrochen neue Ansätze zu finden und bestehende Methoden weiter zu entwickeln.

Ein wichtiger Ansatzpunkt ist die Erweiterung der Modelle auf Problemstellungen, die sich anstatt in (2×2) Tabellen nur in $(R \times K)$ Tabellen darstellen lassen. Eine mögliche Fragestellung für ein solches Problem wäre zum Beispiel das unterschiedliche Wahlverhalten von Afroamerikanern und Weißen in Ohio, wobei in diesem Fall nicht die dichotome Variable „Wahlteilnahme“ mit den Ausprägungen „ja“ und „nein“ im Fokus steht, sondern die unterschiedlichen Parteien, also zum Beispiel Partei A, Partei B, Partei C und Partei D. Derzeit lassen sich solche Fragestellungen nur lösen, indem Zeilen und Spalten sukzessive geschickt zu (2×2) Tabellen zusammengefügt werden und damit mehrere Modelle gerechnet werden. In dem Beispiel zum Wahlverhalten würde die nominale Variable „gewählte Partei“ somit zuerst dichotomisiert, sodass sie schließlich nur die Ausprägungen „Partei A gewählt“ und „nicht Partei A gewählt“ hat. Damit wird dann ein Modell gerechnet und der Vorgang schließlich mit der dichotomisierten Variable mit den Ausprägungen „Partei B gewählt“ und „nicht Partei B gewählt“ fortgesetzt.

7 Anhang

1. Intervall von β_i^a:
   Maximum: Sei $\beta_i^w = 0$
   $0 = \frac{Y_i}{1-X_i} - \frac{X_i}{1-X_i} \cdot \beta_i^a \Leftrightarrow -\frac{Y_i}{1-X_i} = -\frac{-X_i}{1-X_i} \cdot \beta_i^a \Leftrightarrow \beta_i^a = \frac{\frac{-Y_i}{1-X_i}}{\frac{-X_i}{1-Xi}} \Leftrightarrow \beta_i^a = \frac{Y_i}{X_i}$
   Minimum: Sei $\beta_i^w = 1$
   $1 = \frac{Y_i}{1-X_i} - \frac{X_i}{1-X_i} \cdot \beta_i^a \Leftrightarrow 1 - \frac{Y_i}{1-X_i} = -\frac{X_i}{1-X_i} \cdot \beta_i^a \Leftrightarrow \beta_i^a = \frac{1 - \frac{Y_i}{1-X_i}}{\frac{-X_i}{1-Xi}} \Leftrightarrow$
   $\beta_i^a = \frac{(1-X_i)-Y_i}{1-X_i} \cdot \frac{1-X_i}{-X_i} \Leftrightarrow \beta_i^a = \frac{(1-X_i)-Y_i}{-X_i} \Leftrightarrow \beta_i^a = \frac{Y_i-(1-X_i)}{X_i}$

2. Intervall von β_i^w:
   Maximum: Sei $\beta_i^a = 0$
   $\beta_i^w = \frac{Y_i}{1-X_i} - \frac{X_i}{1-X_i} \cdot 0 \Leftrightarrow \beta_i^w = \frac{Y_i}{1-X_i}$
   Minimum: Sei $\beta_i^a = 1$
   $\beta_i^w = \frac{Y_i}{1-X_i} - \frac{X_i}{1-X_i} \cdot 1 \Leftrightarrow \beta_i^w = \frac{Y_i}{1-X_i} - \frac{X_i}{1-X_i} \Leftrightarrow \beta_i^w = \frac{Y_i-X_i}{1-X_i}$

3. $\beta_i^a \in [max(0; \frac{0,6-0,8}{0,2}); min(\frac{0,6}{0,2}; 1)] = [max(0; -1); min(3; 1)] = [0; 1]$
   $\beta_i^w \in [max(0; \frac{0,6-0,2}{0,8}); min(\frac{0,6}{0,8}; 1)] = [max(0; 0,5); min(0,75; 1)] = [0,5; 0,75]$

4. $Y_i = \beta^a \cdot X_i + \beta^w \cdot (1 - X_i) + \epsilon_i \Leftrightarrow Y_i = \beta^a \cdot X_i + \beta^w - \beta^w \cdot X_i + \epsilon_i \Leftrightarrow$
   $Y_i = \beta^w + (\beta^a - \beta^w) \cdot X_i + \epsilon_i$

5. R-Code für die graphische Darstellung des Zahlenbeispiels (Tomographie):

```
x <- c(0,1)
y <- c(0.75, 0.5)
plot(x,y, type="l", ylim=c(0,1), xlab="beta_w", ylab="beta_a")
abline(h=0.75, col="red")
abline(h=0.5, col="red")
abline(v=0, col="blue")
abline(v=1, col="blue")
```

Literatur

Gehring, U. W. and C. Weins (2010). *Grundkurs Statistik für Politologen und Soziologen* (fifth ed.). VS Verlag für Sozialwissenschaften.

Gschwend, T. (2006). Ökologische Inferenz. In *Methoden der Politikwissenschaft: Neuere qualitative und quantitative Analyseverfahren*, pp. 227–237. Nomos.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

Klima, A. (2011). Analysen von Wahlergebnissen in Deutschland 1924-1933: Räumlich-zeitliche Analyse und ökologische Inferenz. Master's thesis, Ludwig-Maximilians-Universität München.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review 15*, 351–357.