Multiple Regression mit voneinander abhängigen Beobachtungen

Random-Effects und Fixed-Effects

Conrad Ziller

Inhalt

Zusammenfassung

Ein Großteil der in der empirisch-vergleichenden Politikwissenschaft verwendeten Datensätze ist räumlich und/oder zeitlich strukturiert. Räumliche und zeitliche Strukturen gehen in der Regel mit statistischen Abhängigkeiten einher, die bei der Datenanalyse mitberücksichtigt werden müssen. Dieser Beitrag stellt *Random-Effects-Modelle* (RE) und *Fixed-Effects-Modelle* (FE) als Analysemethoden für voneinander abhängige Beobachtungen vor. Dabei wird auf den Problemgegenstand eingegangen und die Anwendung von RE- und FE-Modellen erklärt. Darüber hinaus werden Entscheidungsheuristiken und Hinweise für die praktische Anwendung gegeben.

Schlüsselwörter

Random Effects · Fixed Effects · Panelanalyse · Hybrid-Modelle · Mehrebenenanalyse

C. Ziller (✉)
Universität zu Köln, Köln, Deutschland
E-Mail: ziller@wiso.uni-koeln.de

C. Wagemann et al. (Hrsg.), *Handbuch Methoden der Politikwissenschaft*, Springer Reference Sozialwissenschaften, https://doi.org/10.1007/978-3-658-16937-4_28-2

1 Einleitung[1]

Eine zunehmende Anzahl politikwissenschaftlicher Forschungsarbeiten macht Gebrauch von international-vergleichenden Datensätzen. Dabei handelt es sich beispielsweise um Umfragedaten aus dem European Social Survey (ESS), dem International Social Survey Programme (ISSP) oder der Comparative Study of Electoral Systems (CSES). Auch sind Paneldaten und aggregierte Zeitreihendaten zu Ländermerkmalen (*Time-Series-Cross-Sectional-Data*, TSCS[2]) eine beliebte Datengrundlage für empirisch-vergleichende Analysen. Sind Daten räumlich und/oder zeitlich strukturiert bzw. gruppiert, zeigen sich auch oftmals in der Datenanalyse statistische Abhängigkeiten zwischen Beobachtungen.

Werden diese Daten naiv zusammengefasst bzw. *gepoolt* und mithilfe des regressionsanalytischen Verfahrens der kleinsten Quadrate (*Ordinary-Least-Squares*, OLS) analysiert, kann dies zu verzerrten Standardfehlern und/oder Regressionskoeffizienten führen. Der Grund ist, dass die Schätzung der Regressionsparameter im OLS-Verfahren auf spezifischen Annahmen hinsichtlich statistischer Unabhängigkeit von Beobachtungen beruht. Sind diese Bedingungen erfüllt, können OLS-Regressionsmodelle geschätzt werden (siehe hierzu den Beitrag von Seng in diesem Sammelband). Bestehen statistische Abhängigkeiten, können alternativ sogenannte *Random-Effects*- (RE) oder *Fixed-Effects*-Verfahren (FE) genutzt werden. Die Bezeichnungen „fixed“ und „random“ haben in der Literatur diverse Bedeutungen (siehe Gelman 2005, S. 20–21). Sie werden hier zur Kategorisierung von bestimmten Modellarten genutzt. Während Fixed-Effects sich auf Modelle bezieht, die mithilfe von Dummy-Variablen gruppenbezogene Unterschiede in der abhängigen Variable kontrollieren, bezieht sich Random-Effects auf Modelle, die gruppenbezogene Unterschiede über eine Zufallsvariable modellieren.[3]

Ziel dieses Beitrags ist es, eine praxisorientierte Einführung in RE- und FE-Modelle zu leisten, Vor- und Nachteile der einzelnen Verfahren herauszustellen und neuere Entwicklungen und Debatten aufzuzeigen. Für weiterführende technische Details zu RE- und FE-Modellen siehe z. B. Andreß et al. (2013); Rabe-Hesketh und Skrondal (2012); Wooldridge (2009, Kap. 14) und zu FE-Modellen Brüderl und Ludwig (2015). Der Einfachheit halber bezieht sich der Beitrag auf

[1]Ich danke Carl Berning, Hans-Jürgen Andreß, Merlin Schaeffer und den Herausgebern für hilfreiche Kommentare sowie Manuel Diaz Garcia und Erik Wenker fürs Korrekturlesen.

[2]Im Vergleich zu Paneldaten handelt es sich bei TSCS-Daten um Aggregatdaten, bei denen die Anzahl der Zeitpunkte T in der Regel größer ist als die Anzahl der Untersuchungseinheiten N (T > N). Sowohl bei Panel- als auch TSCS-Daten ergeben sich ähnliche Problematiken hinsichtlich statistischer Abhängigkeit (zu Besonderheiten von TSCS-Daten siehe Beck und Katz 1995).

[3]In Bezug auf die Terminologie ist zu beachten, dass auch in RE-Modellen die geschätzten Koeffizienten mitunter als Fixed-Effects bezeichnet werden (z. B. im Rahmen der Mehrebenenanalyse, siehe auch den Beitrag von Pötschke in diesem Band). Dies soll verdeutlichen, dass die Effekte für alle Gruppen in der untersuchten Population als identisch bzw. fix angenommen werden. Hingegen beziehen sich hier die Bezeichnungen RE und FE auf Modellklassen, mit jeweils distinkten Eigenschaften.

kontinuierliche abhängige Variablen und lineare Zusammenhänge zwischen Variablen. RE- und FE-Modelle können darüber hinaus auch andere abhängige Variablen (z. B. binäre Variablen oder Zählvariablen) bei entsprechender Anpassung der Modellspezifikation (z. B. logistisches, ordinal-logistisches oder Poisson-Modell) berücksichtigen (siehe Snijders und Bosker 2012, Kap. 17; Andreß et al. 2013, Kap. 5).

Im Folgenden wird zunächst auf das Problem voneinander abhängiger Beobachtungen eingegangen. Anschließend erläutert der Beitrag die Grundlagen von RE- und FE-Modellen und zeigt mögliche Heuristiken und Entscheidungshilfen zur Modellauswahl. Den Abschluss bilden praktische Hinweise und Einblicke in aktuelle methodische Diskussionen zum Themenbereich.

2 Das Problem voneinander abhängiger Beobachtungen

2.1 Statistische Abhängigkeit

Ein weitverbreitetes Phänomen in der empirischen Politikwissenschaft ist der Umgang mit Daten, die entlang bestimmter übergeordneter Einheiten (auch als Kontexte oder Makroeinheiten bezeichnet) gruppiert – man sagt auch „geclustert“ oder „genestet“ – sind. Gehen wir beispielsweise von im Querschnitt vorliegenden Umfragedaten als Beobachtungsgrundlage aus, können räumliche Strukturen wie Nachbarschaften, Gemeinden, Regionen und Länder relevante Gruppen darstellen.[4] Ebenso stellen über die Zeit hinweg gemessene Daten eine gruppierte Datenstruktur dar. Bei Paneldaten (z. B. wiederholte Messungen derselben Individuen) sind die zeitpunktspezifischen Beobachtungen in den jeweiligen Individuen gruppiert (d. h. Person ist hier das Gruppenmerkmal). Werden nun Standardregressionsverfahren wie OLS-Regression zur Modellierung solcher Datenstrukturen verwendet, kann dies zu verzerrten Standardfehlern und/oder Regressionskoeffizienten führen. Der Grund ist, dass diese Verfahren statistische Unabhängigkeit von Beobachtungen annehmen.

Aber was bedeutet statistische Unabhängigkeit/Abhängigkeit? Allgemein formuliert bedeutet statistische Unabhängigkeit, dass Beobachtungen unabhängig voneinander auftreten und somit jede Beobachtung eine unabhängige Information beinhaltet. Statistische Abhängigkeit hingegen heißt, dass Beobachtungen (innerhalb bestimmter Gruppen) in systematischer Weise zusammenhängen und dadurch Annahmen statistischer Modelle verletzt sind.

Dabei begünstigen bestimmte Datentypen das Auftreten statistischer Abhängigkeiten. So sind Paneldaten zeitlich strukturiert und es ist wahrscheinlich, dass sich

[4]Gruppen haben in der Regel keinen intrinsischen Informationsgehalt, sind austauschbar und beinhalten eine große Anzahl an Kategorien (analog zur Definition höherer Ebenen/Level in der Mehrebenenanalyseliteratur). Merkmale, die diese Kriterien nicht erfüllen (z. B. Geschlecht), gelten als Variable und nicht als Gruppe.

aktuelle und zukünftige Beobachtungen von Merkmalen einer Person ähneln bzw. miteinander korrelieren. Ein Beispiel ist die Mitgliedschaft in einer politischen Partei zum Zeitpunkt t_0, die wahrscheinlich auch noch in t_1 besteht. Ein Grund dafür ist möglicherweise, dass Menschen mit hohem politischem Interesse (durchgängig) Parteimitglieder sind, während solche mit geringem Interesse gar nicht erst eintreten. Bei Querschnittsdaten aus mehrstufigen Zufallsstichroben (z. B. zufällig ausgewählte Personen aus zufällig ausgewählten Ländern) können historische oder institutionelle Länderunterschiede dazu führen, dass sich bestimmte Befragtenmerkmale innerhalb von Ländern ähneln. Bei Querschnittsdaten, die aus einer einfachen Zufallsstichprobe stammen, bestehen – auf die gesamte Stichprobe bezogen – in der Regel keine oder nur geringe statistische Abhängigkeiten. Sollen allerdings zum Beispiel regionale Unterschiede bestimmter Befragtenmerkmale erklärt werden, können (nach entsprechender Einteilung der Daten) auch hier Abhängigkeiten aufgrund der Zugehörigkeit von Befragten zu bestimmten räumlichen Kontexten (z. B. Bundesländer) bestehen.[5] Die Gruppierung der Daten folgt hierbei aus dem theoriegeleiteten Erkenntnisinteresse heraus, bestimmte gruppenbezogene Unterschiede zu erklären – ein Vorgehen, das bei der Analyse von Kontexteffekten in Mehrebenenanalysen üblich ist (siehe u. a. den Beitrag von Pötschke in diesem-Band).

Zur weiteren Darstellung werden prototypische Daten aus einer fiktiven einmaligen Umfrage unter Bürgerinnen und Bürgern zum Thema Zufriedenheit mit der aktuellen Regierung herangezogen. Diese Umfrage wurde in 20 Ländern durchgeführt und hat einen Stichprobenumfang von 1000 Befragten pro Land (insgesamt 20.000 Befragte). Die abhängige Variable ist Regierungszufriedenheit (y), welche mit einer 11-stufigen Antwortskala (von 0 = „extrem unzufrieden“ bis 10 = „extrem zufrieden“) erfasst wurde. Weiter wird davon ausgegangen, dass sich Befragte innerhalb der Länder systematisch ähneln und deshalb mittlere Unterschiede bezüglich Regierungszufriedenheit im Ländervergleich zu beobachten sind. Formal besteht die Variable Regierungszufriedenheit aus den folgenden Komponenten:

$$y_{ij} = \alpha_0 + u_j + e_{ij}. \tag{1}$$

Jeder Beobachtungswert y_{ij} (einer befragten Person i innerhalb eines Länderkontextes j) ist bestimmt durch einen Gesamtmittelwert α_0, einen länderspezifischen Effekt u_j und einen Residualeffekt e_{ij}. Der länderspezifische Effekt u_j repräsentiert dabei die Abweichung des Ländermittelwerts (von Land j) vom Gesamtmittelwert. Der Residualeffekt e_{ij} repräsentiert wiederum die Abweichung eines gemessenen Wertes (von Befragtenperson i) vom Ländermittelwert (von Land j). Die Gesamtvarianz der Variable Regierungszufriedenheit Var(y_{ij}) unterteilt sich entsprechend in

[5]Neben einer solchen *räumlich-kategorialen* Abhängigkeit, können auch Formen *räumlich-kontinuierlicher* Abhängigkeit existieren, bei der nahe beieinander liegende Beobachtungen sich ähnlich sind und diese Ähnlichkeit mit zunehmender räumlicher Distanz abnimmt. In solchen Fällen kommen spezielle räumliche Regressionsverfahren zum Einsatz, bei denen Merkmale der umliegenden Beobachtungen gewichtet nach ihrer Distanz zueinander einbezogen werden.

die Varianz des Gruppeneffekts σ_u^2 (sogenannte *Between-Group-Variance*) und die Varianz des intragruppalen Fehlerterms bzw. Residuums σ_e^2 (sogenannte *Within-Group-Variance*). Der Anteil der gruppenbezogenen Varianz an der Gesamtvarianz ist dabei die zentrale Maßzahl für statistische Abhängigkeit von Beobachtungen, welche als *serielle Korrelation* (bei Paneldaten auch als Autokorrelation) bezeichnet wird. Sie entspricht dabei dem sogenannten Intraklassen-Korrelationskoeffizienten (ICC)[6] (Andreß et al. 2013, S. 77 f.; Snijders und Bosker 2012, S. 17 ff.). Je größer der ICC, desto größer sind die Ähnlichkeiten innerhalb von Gruppen und desto größer sind folglich die Abweichungen zwischen Gruppen.

Die weiteren Ausführungen beziehen sich darauf, wie verschiedene statistische Methoden mit statistischer Abhängigkeit umgehen. Zentrale Unterschiede bestehen hier vor allem in der Modellierung des Gruppeneffektes u_j, der auch als *unbeobachtete Heterogenität der Gruppenebene* bezeichnet wird. Zunächst wird auf konkrete Auswirkungen statistischer Abhängigkeit im Rahmen des OLS-Regressionsverfahrens eingegangen.

2.2 Verletzung der Annahme unkorrelierter Fehler

Verfahren wie die OLS-Regression bauen auf der Annahme auf, dass jede Beobachtung einen unabhängigen Informationsbeitrag leistet und somit *keine* serielle Korrelation vorliegt. Die Annahme bezieht sich dabei auf serielle Korrelation der Fehler, also die Varianz der abhängigen Variable, die nicht durch die im Modell befindlichen Prädiktoren erklärt wird.[7] Liegt serielle Korrelation vor, können zwar die Koeffizienten einer OLS-Regression unverzerrt[8] sein, die Standardfehler sind hingegen in jedem Fall nicht mehr korrekt geschätzt bzw. verzerrt. Das wirkt sich wiederum auf die Gültigkeit statistischer Testverfahren aus. Da hierbei die Fehler in der Regel positiv korreliert sind, kommt es zu einer Unterschätzung der Standardfehler und entsprechend überschätzten *F*- und *t*-Statistiken (Moulton 1986). Mit anderen Worten: Es werden statistisch signifikante Zusammenhänge gefunden und interpretiert, die gar keine sind. Ursächlich kann serielle Korrelation auf (unbeobachtete) Merkmale der Gruppenebene zurückzuführen sein. Bei Panel- und TSCS-Daten kommen auch intragruppale Faktoren (d. h. zeitveränderliche Prozesse) in Frage.

[6] Formal zeigt diese Maßzahl die Korrelation von Werten zweier zufällig ausgewählter Beobachtungseinheiten derselben (zufällig ausgewählten) Gruppe j an: $Corr\left(y_{ij}, y_{kj}\right) = \frac{Cov\left(y_{ij}, y_{kj}\right)}{\sqrt{Var\left(y_{ij}\right)} * \sqrt{Var\left(y_{kj}\right)}} \triangleq \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$, $i \neq k$.

[7] Das Ausmaß des Fehlers der Beobachtung i darf also keinen Effekt auf das Ausmaß (oder die Richtung) des Fehlers der Beobachtung k haben. Formell ausgedrückt: $Corr(e_i, e_k) = 0$, für alle $i \neq k$. Der Erwartungswert aller paarweisen Produkte von Fehlern ist folglich null.

[8] Unverzerrtheit (auch als Erwartungstreue bezeichnet) bedeutet, dass der Durchschnitt hypothetisch vieler unabhängiger Replikationen einer Schätzung (mit wiederholt gezogenen Stichproben) dem tatsächlichen Populationswert entspricht oder ihm zumindest sehr nahe kommt. Konsistenz bedeutet zudem, dass ein Schätzer mit immer größerem Stichprobenumfang immer genauer wird.

Zur Bestimmung von serieller Korrelation kann ein ICC berechnet werden. Entsprechende Routinen sind in den gängigen Datenanalyseprogrammen implementiert[9] und bei Vorliegen der Varianzparameter kann dieser einfach per Hand berechnet werden. Zudem können modellvergleichende Teststatistiken wie der Breusch-Pagan-LM-Test Aufschluss über das Vorliegen systematischer gruppenbezogener Unterschiede geben. Ob bei Panel- oder TSCS-Daten darüber hinaus serielle Korrelation im intragruppalen Fehlerterm e_{ij} (auch idiosynkratischer Fehlerterm genannt) besteht, kann mit dem Wooldridge-Test für serielle Korrelation bei Paneldaten geprüft werden (Drukker 2003).

Um korrekte Standardfehler trotz serieller Korrelation zu erhalten, kann zunächst einmal für Faktoren, die ursächlich mit der seriellen Korrelation in Zusammenhang stehen, kontrolliert werden. Bei dem oben genannten Beispiel zu individueller Parteienmitgliedschaft könnte ein relevanter Faktor das allgemeine politische Interesse von Personen sein. In Bezug auf Regierungszufriedenheit könnten institutionelle und wirtschaftliche Ländermerkmale relevant sein. Da mitunter nicht für alle relevanten Faktoren geeignete Indikatoren zur Verfügung stehen, korrigieren RE-Modelle für serielle Korrelation aufgrund gruppenbezogener Unterschiede durch die explizite Modellierung des Gruppeneffektes u_j. Gleichermaßen beheben FE-Modelle serielle Korrelation, indem sie gruppenbezogene Abweichungen per Design kontrollieren. Eine weitere Möglichkeit ist die Schätzung von cluster- bzw. panel-robusten Standardfehlern, welche die empirischen Varianzen der Fehler für jede Gruppe berücksichtigen und so allgemein für Formen serieller Korrelation sowie Heteroskedastizität[10] korrigieren. Robuste Standardfehler fallen dadurch in der Regel höher als konventionelle Standardfehler aus (siehe hierzu auch Abschn. 5.2). Speziell für Paneldaten und serielle Korrelation aufgrund zeitveränderlicher Prozesse kommen unter anderem generalisierte Schätzgleichungen in Frage, die für verschiedene Formen korrelierter Fehler korrigieren können.[11]

Zusammengefasst gibt es eine Reihe an wirksamen Gegenmaßnahmen, die serieller Korrelation und entsprechenden Auswirkungen auf Standardfehler und Teststatistiken begegnen. Statistische Abhängigkeit kann sich darüber hinaus auf die Annahme der Exogenität von Prädiktoren auswirken, was einen weiteren Anlass zur Verwendung von insbesondere FE-Modellen darstellt.

[9]Zum Beispiel zeigt in Stata „rho“ im Output des „xtreg var1 var2 etc., re“-Befehls den Anteil unerklärter Varianz der Gruppenebene gegeben das jeweilige Modell an.

[10]Ist die Homoskedastizitätsannahme (einheitliche Verteilung der Fehler) verletzt, beeinträchtigt dies ebenfalls die Unverzerrtheit von Standardfehlern und Gültigkeit von Teststatistiken. Verletzungen dieser Annahme können zum Beispiel mithilfe des White-Tests geprüft werden.

[11]Häufig wird angenommen, dass die serielle Korrelation zwischen zwei aufeinander folgenden Beobachtungszeitpunkten am stärksten ist und mit wachsendem zeitlichem Abstand rasch abnimmt. Diese auch als Autokorrelation erster Ordnung (AR(1)) bezeichnete Fehlerstruktur kann beispielsweise in Stata mit den Befehlen „xtgee var1 var2 etc., corr(ar1)“ oder „xtregar var1 var2 etc.“ modelliert werden.

2.3 Verletzung der Exogenitätsannahme

Eine weitere Annahme des OLS-Regressionsverfahrens, die aufgrund statistischer Abhängigkeit verletzt sein kann, ist die der Exogenität. Sie besagt, dass Fehler und Prädiktorvariablen unkorreliert sein müssen. Nicht-gemessene Merkmale (wie z. B. Persönlichkeitsmerkmale oder genetische Einflüsse), die gutmöglich die zu erklärende Variable beeinflussen, dürfen also keinesfalls mit den gemessenen Merkmalen in der Regressionsgleichung korrelieren (Vaisey und Miles 2017, S. 47). Im Anwendungsfall und unter Bedingungen imperfekter Messung und Verfügbarkeit von Indikatoren, ist diese Annahme mehr als heroisch. Die Auswirkungen einer Verletzung der Exogenitätsannahme können gravierend sein. Im Gegensatz zur Verletzung der Annahme unkorrelierter Fehler ist hier nicht nur die Varianz der OLS-Schätzer (d. h. Standardfehler und damit Teststatistiken) betroffen, sondern die Unverzerrtheit des Schätzers an sich. Ein Koeffizient, bei dem wir uns nicht über die Höhe und Richtung des Effektes sicher sein können, ist ohne substanziellen Informationsgehalt.[12] Die Hauptursache für die Verletzung der Exogenitätsannahme ist ein Auslassen relevanter Variablen (*Omitted-Variables*).[13] Denn wären alle relevanten Variablen ins Modell aufgenommen, wäre die abhängige Variable vollumfänglich erklärt und es verblieben keine Fehler, die mit den Prädiktoren korrelieren könnten. Hingegen können nicht-gemessene bzw. unbeobachtete Eigenschaften von Untersuchungseinheiten als Normalfall in der angewandten empirischen Sozialforschung gelten.

Zur weiteren Illustration wird auf den eingangs beschriebenen ländervergleichenden Umfragedatensatz zur Regierungszufriedenheit zurückgegriffen. Analog zu Theorien, welche die Bewertung von Regierungsleistung mit ökonomischen Motive erklären (Lewis-Beck und Stegmaier 2000; Tilley et al. 2018), wird hier angenommen, dass Personen mit einer positiven Bewertung ihrer persönlichen Einkommenssituation zufriedener mit der momentanen Regierungsleistung sind als Personen mit finanziellen Problemen. Die Bewertung der persönlichen Einkommenssituation wird mit einer ordinal-skalierten Variable x (0 = „komme gar nicht/kaum mit Einkommen aus“, 1 = „komme mit Einkommen aus“, 2 = „lebe komfortabel mit Einkommen“) erhoben. Zur Prüfung des Zusammenhangs werden alle 20.000 Befragten zusammengefasst bzw. gepoolt und mit der Methode der kleinsten Quadrate (OLS-Regression) analysiert. Eine entsprechende linear-additive Regressionsgleichung hat die folgende funktionale Form:

$$y_i = \beta_0 + \beta_1 x_i + e_i. \tag{2}$$

[12]Dabei muss betont werden, dass das Ausmaß an Verzerrung von der Höhe der Korrelation zwischen Prädiktor und Fehlerterm abhängt (siehe dazu die Simulationsstudien in Clark und Linzer 2015 sowie Vaisey und Miles 2017).

[13]Weitere Ursachen für eine Verletzung der Exogenitätsannahme, auch als Endogenenitätsverzerrung (*Endogeneity-Bias*) bezeichnet, sind Messfehler sowie wechselseitige Kausalität.

β_0 ist hier der Intercept (auch als Achsenabschnitt oder Konstante bezeichnet), der den mittleren Wert an Regierungszufriedenheit angibt, wenn x (Einkommenszufriedenheit) null ist. β_1 ist der Koeffizient[14] für den Einfluss der Variable Einkommenszufriedenheit. e repräsentiert die Abweichungen der einzelnen Beobachtungen von der Regressionsgeraden bzw. den Fehlerterm. Im hier illustrierten Fall bezieht sich die Regressionsgleichung auf Individuum i ($i = 1, \ldots, n$). Damit der Koeffizient β_1 aus dem OLS-Verfahren unverzerrt ist, müssen eine Reihe von Annahmen erfüllt sein (Verbeek 2004, S. 16; Wooldridge 2009, S. 47–52). Die wichtigste Annahme ist die der Exogenität, also dass Fehlerterm e einen Erwartungswert von null unabhängig vom Wert der Prädiktorvariable x hat: $E(e|\ x) = 0$ bzw. Fehlerterm und Prädiktorvariable unkorreliert sind.[15]

Auch im vorliegenden Beispiel ist es durchaus möglich, dass aufgrund ausgelassener Variablen die Exogenitätsannahme nicht erfüllt ist. So können nicht-gemessene persönlichkeitsbezogene Unterschiede (z. B. individuelle Verträglichkeit) zwischen Befragten existieren, die sowohl mit persönlicher Einkommenszufriedenheit als auch Regierungszufriedenheit zusammenhängen. Die Folge ist, dass die Variable persönliche Einkommensbewertung nun auch anteilsweise den Effekt des Persönlichkeitsmerkmals Verträglichkeit transportiert. Der Koeffizient der Variable Einkommenszufriedenheit ist folglich verzerrt.[16]

Darüber hinaus können systematische Unterschiede in der abhängigen Variable zwischen Gruppen (unbeobachtete Heterogenität der Gruppenebene), sofern diese nicht durch x erklärt werden, eine Verletzung der Exogenitätsannahme verursachen. Abb. 1 illustriert die Folge unberücksichtigter gruppenbezogener Unterschiede in der abhängigen Variable grafisch. Wie im oben angeführten Beispiel soll die Regierungszufriedenheit mit der persönlichen Einkommensbewertung der Befragten erklärt werden. Betrachtet man vier zufällig ausgewählte Beobachtungseinheiten in Land 1, zeigt sich ein moderater positiver Zusammenhang zwischen persönlicher Einkommenssituation und Regierungszufriedenheit. Ebendieser ist auch in den beiden weiteren Länderkontexten gegeben. Zwischen den Ländern bestehen jedoch erhebliche Unterschiede im Niveau der abhängigen Variable. Befragte in Land 3 haben eine wesentlich höhere durchschnittliche Regierungszufriedenheit als Befragte in Land 1. Werden diese Niveauunterschiede vernachlässigt und eine

[14]Auch als Schätzer, Effekt, Steigungsparameter oder Regressionsgewicht bezeichnet.

[15]Neben der Exogenitätsannahme wird die Linearität der Parameter vorausgesetzt, es darf keine perfekte Multikollinearität zwischen den Variablen herrschen und x muss variable Werte aufweisen (bzw. darf nicht konstant sein). Damit auch die Standardfehler (und damit Teststatistiken) unverzerrt sind, müssen die Fehler einen Erwartungswert von null aufweisen, normalverteilt sein und die Varianz der Fehler muss konstant für alle Ausprägungen von x sein (Homoskedastizität). Zudem darf keine serielle Korrelation vorherrschen. Eine weitere Annahme ist, dass die Beobachtungen eine zufällige Auswahl aus einer bestimmten Grundgesamtheit bzw. Population widerspiegeln, was für die Gültigkeit statistischer Inferenz notwendig ist.

[16]Wäre Verträglichkeit die einzige ausgelassene Variable und gäbe es einen passenden Indikator, den wir in die Gleichung einbeziehen könnten, wäre der Effekt von Einkommenszufriedenheit für Verträglichkeit kontrolliert. Die Exogenitätsannahme wäre somit erfüllt und der Koeffizient für Einkommenszufriedenheit unverzerrt.

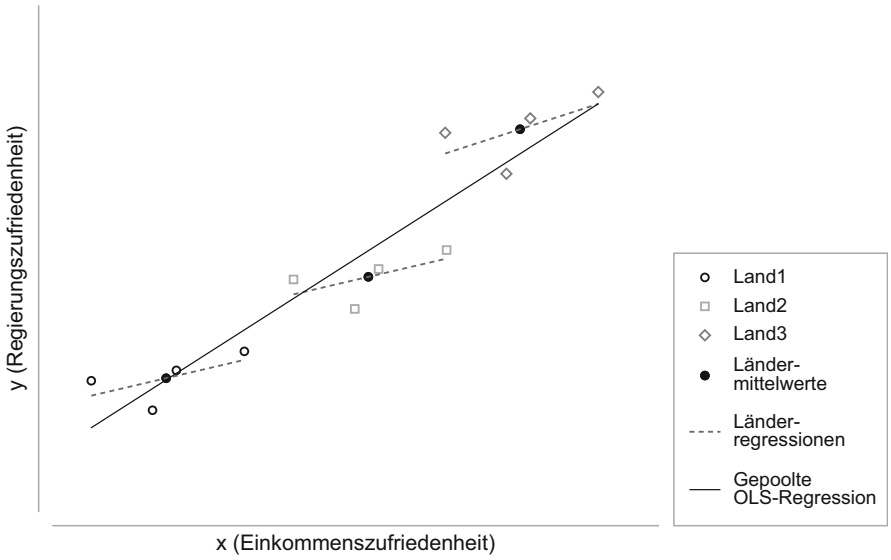

Abb. 1 Gepoolte OLS-Regression mit gruppierten Beobachtungen

OLS-Regression mit den gepoolten Daten geschätzt, ist der Effekt von Einkommenszufriedenheit auf Regierungszufriedenheit β_1 weitaus stärker (steilerer Anstieg der Regressionsgeraden) als dies die Zusammenhänge innerhalb der einzelnen Länder vermuten lassen. Der Grund ist, dass die Mittelwertunterschiede unbeobachtete Ländermerkmale repräsentieren, die nun durch die Individualvariable Einkommenszufriedenheit transportiert werden.

Beispielsweise können wir davon ausgehen, dass die beobachteten Länderunterschiede bezüglich Regierungszufriedenheit auf Unterschiede im länderspezifischen Wirtschaftswachstum zurückzuführen sind. Ohne den Einbezug dieser Variable transportiert die Individualvariable den Effekt des länderspezifischen Wirtschaftswachstums, da in diesem Beispiel auch die persönliche Einkommensbewertung vom Wirtschaftswachstum (und den dadurch steigenden Einkommen) beeinflusst wird. Die Folge ist eine verzerrte Schätzung des Koeffizienten der Variable Einkommenszufriedenheit, in diesem Fall eine Überschätzung.

Es existieren keine direkten Tests, um die Verletzung der Exogenitätsannahme aufgrund ausgelassener Variablen im Modell zu prüfen.[17] Allerdings zeigen gruppenbezogene Mittelwertunterschiede in der abhängigen Variablen unbeobachtete Heterogenität an, die entweder modelliert oder wegkontrolliert werden sollte. Hierfür kommen RE- sowie FE- Modelle in Frage.

[17]Tests wie der Ramsey-RESET-Test zeigen zwar Missspezifikationen im Hinblick auf bereits im Modell vorhandene Variablen an, jedoch nicht, ob weitere ausgelassene Variablen relevant sind. Welche ausgelassenen Variablen relevant sind, sollte sich daher vorrangig nach theoretischen Überlegungen richten.

3 Random-Effects- und Fixed-Effects-Modelle

Bestehen statistische Abhängigkeiten aufgrund gruppenbezogener unbeobachteter Heterogenität, ist in jedem Fall die Annahme unkorrelierter Fehler verletzt und – ohne entsprechende Gegenmaßnahmen – sind Standardfehler und Teststatistiken aus der OLS-Regression verzerrt. RE- und FE-Modelle beheben solche Formen der seriellen Korrelation. Analog zum OLS-Verfahren produzieren bei einer Verletzung der Exogenitätsannahme aufgrund (mit Prädiktorvariablen) korrelierter gruppenbezogener Heterogenität auch RE-Modelle verzerrte Koeffizientenschätzer. FE-Modelle sind hingegen nicht anfällig für korrelierte gruppenbezogene Heterogenität, weisen jedoch andere Einschränkungen auf.

3.1 Random-Effects-Modelle

Die allgemeine funktionale Form des RE-Modell ist gegeben als:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij}. \tag{3}$$

Der Unterschied zum OLS-Modell besteht darin, dass die Gruppenebene über die Notation *j* im Modell berücksichtigt ist. Die wichtigste Erweiterung in der Modellierung ist, dass der Intercept β_0 nun über Gruppen (z. B. Länder) *j* hinweg variieren kann. Einen variablen Intercept zu modellieren, trägt dem Phänomen Rechnung, dass – mit Blick auf das oben genannte Beispiel – in einigen Ländern die Regierungszufriedenheit (y) im Durchschnitt höher ist als in anderen Ländern. Zur weiteren Spezifizierung kann β_{0j} in einen mittleren Intercept über alle Gruppen β_0 und länderspezifische Abweichungen u_j unterteilt werden:

$$\beta_{0j} = \beta_0 + u_j. \tag{4}$$

Beide Teilgleichungen sind in eine integrierbar:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}. \tag{5}$$

Dieses Modell ist äquivalent zu einem *Random-Intercept*-Mehrebenenmodell (auch als hierarchische Modelle oder *Mixed-Models* bezeichnet, siehe auch den Beitrag von Pötschke in diesem Band). u_j ist der Fehlerterm der Gruppenebene (unbeobachtete Heterogenität der Gruppenebene) und als normalverteilte Zufallsvariable konzipiert, bei der die Gruppeneffekte als zufällige Abweichungen vom Gesamtmittelwert interpretiert werden können. *e* repräsentiert wiederum die Abweichungen der in den jeweiligen Gruppen befindlichen Einzelbeobachtungen (vom Gruppendurchschnitt $\beta_0 + u_j$). Die Schätzung der Parameterwerte basiert beim RE-Modell auf der Generalisierten-Kleinste-Quadrate-Methode (*Generalized-Least-Squares*, GLS), bei der von den Beobachtungswerten ein bestimmter Anteil eines jeden Variablenmittelwertes abgezogen wird (auch als *Quasi-Demeaning*

bezeichnet). Die Höhe des Anteils wird von einem Transformationsparameter (*Theta*) bestimmt, der wiederum sowohl vom Ausmaß gruppenbezogener Heterogenität sowie der Anzahl an Perioden (bei Paneldaten) bzw. der durchschnittlichen Anzahl an Fällen pro Gruppe (gruppierte Querschnittsdaten) abhängt (siehe Andreß et al. 2013, S. 154). Alternativ zu GLS findet auch das *Maximum-Likelihood*-Verfahren (ML) Anwendung. Ausgehend von bestimmten Startwerten wird hier diejenige Kombination von Parametern gesucht, die den Wert der Likelihood-Funktion iterativ erhöht, bis dieser maximal ist. Dieser Maximalwert ist gleichzeitig die Kombination an Parametern, für welche die Realisierung der beobachteten Daten am wahrscheinlichsten ist.

Wie auch in OLS-Modellen nutzen RE-Modelle zur Schätzung der Modellparameter sowohl Varianz zwischen Gruppen (Between-Group-Variance) als auch Varianz innerhalb von Gruppen (Within-Group-Variance).[18] Dabei legen OLS-Modelle mehr Gewicht auf die Varianz zwischen Gruppen als RE-Modelle. Der Grund ist, dass im OLS-Modell die gesamte Varianz zwischen Gruppen in der Variable x enthalten ist, während im RE-Modell ein Teil dieser Varianz durch das Quasi-Demeaning eliminiert bzw. im Fehlerterm u_j gebunden ist (Greene 2003, S. 295 f.). Folglich sind die Ergebnisse einer RE-Schätzung in der Regel zwischen denen einer OLS-Schätzung und einer auf Within-Variance basierenden FE-Schätzung (siehe Abschn. 3.2) angesiedelt. Das Ausmaß, in dem die RE-Schätzung hin zum FE-Schätzer gewichtet ist, hängt dabei maßgeblich von der durchschnittlichen Gruppengröße ab. Je höher die Anzahl an Perioden (Paneldaten) bzw. die durchschnittliche Fallzahl pro Gruppe (Querschnittsdaten), desto stärker entspricht die RE-Schätzung einer auf Within-Variance basierenden FE-Schätzung (analog zur Berechnung des Transformationsparameters Theta).

In Abb. 2 sind die relevanten Komponenten eines RE-Modells abgetragen. Auch hier ist Regierungszufriedenheit die abhängige Variable und die Bewertung des persönlichen Einkommens der Prädiktor. War der Regressionskoeffizient aus dem OLS-Verfahren (Abb. 1) sehr nahe am Between-Schätzer der Gruppendurchschnitte, ist der RE-Koeffizient β_1 nun stärker in Richtung der Effekte der einzelnen Länder (bzw. des Within-Schätzers) gewichtet. Der Effekt der Einkommenszufriedenheit ist nunmehr teilweise von der Gruppenebene entkoppelt und repräsentiert somit – verglichen mit dem OLS-Schätzer – stärker den unverzerrten, d. h. um unbeobachtete Heterogenität bereinigten, Effekt.

Analog zur Mehrebenenanalyse können in RE-Modellen Variablen der Gruppenebene (auch als Makro- oder Kontextvariablen bezeichnet) aufgenommen werden, um gruppenspezifische Abweichungen vom Gesamtmittelwert u_j zu erklären.[19]

[18] Analog dazu bezieht sich der sogenannte Between-Schätzer auf eine Regression, die nur die Gruppendurchschnitte einbezieht. Der Within-Schätzer bezieht sich auf Regressionen mit Beobachtungen der jeweiligen Gruppen. Gemittelte Within-Schätzer sind äquivalent zu Koeffizienten aus FE-Modellen (siehe Abschn. 3.2).

[19] Werden Variablen der Gruppenebene in ein OLS-Modell aufgenommen, resultieren daraus in der Regel Schätzer mit zu geringen Standardfehlern, was wiederum zu invaliden Teststatistiken führt (Moulton 1986).

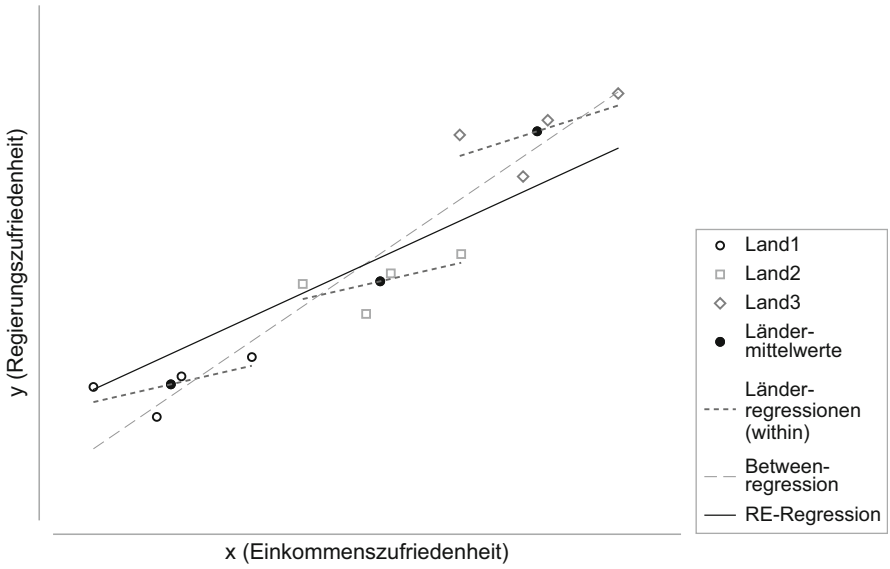

Abb. 2 RE-Modell mit gruppierten Beobachtungen

Nehmen wir das länderspezifische Wirtschaftswachstum als Kontextvariable z_j auf, erwarten wir eine Verringerung der gruppenspezifischen Abweichungen (im Vergleich zu einem Modell ohne diese Variable):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + u_j + e_{ij} \tag{6}$$

Darüber hinaus ist es möglich, die Effekte von Intra-Gruppenvariablen (hier Individualvariablen) über Gruppen bzw. Kontexte hinweg variieren zu lassen (sogenannte *Random-Slopes*). Diese Variationen können in einem weiteren Schritt durch eine Gruppenvariable unter Verwendung sogenannter *Cross-Level-Interaktionen* erklärt werden.

Als politikwissenschaftliches Anwendungsbeispiel untersuchen Hakhverdian und Mayne (2012) den Effekt von Bildung auf das politische Vertrauen von Bürgerinnen und Bürgern (im europäischen Vergleich). Darüber hinaus sind die Autoren daran interessiert, ob der Effekt von Bildung über Länderkontexte variiert (=Random Slope) und ob mögliche Unterschiede durch Grade an politischer Korruption (als Ländermerkmal) erklärt werden können (=Cross-Level-Interaktion). Dafür nutzen Sie Umfragedaten aus dem European Social Survey und finden, dass 27 Prozent der Varianz im politischen Vertrauen auf Länderunterschiede zurückzuführen sind (=serielle Korrelation aufgrund gruppenbezogener Heterogenität).

Dies (sowie die Absicht Ländermerkmale zu testen) rechtfertigt die Verwendung eines RE- bzw. Mehrebenenmodells. Bereits 10 Prozent des gruppenbezogenen Varianzanteils werden von Individualvariablen im Modell erklärt. Ländermerkmale und die Cross-Level-Interaktion erklären weitere gruppenbezogene Varianz, so dass

im finalen Modell noch drei Prozent der verbleibenden Gesamtvarianz auf nicht-modellierte Länderunterschiede zurückzuführen sind. Als Ergebnis zeigen die Autoren, dass der Effekt von Bildung auf politisches Vertrauen über Länder variiert und mit dem Ländermerkmal Korruption interagiert. Bildung ist negativ mit politischem Vertrauen assoziiert in Ländern mit hoher Korruption (je mehr Bildung desto weniger Vertrauen), während ein positiver Zusammenhang in Ländern mit geringer Korruption besteht (je mehr Bildung desto mehr Vertrauen).

Im Hinblick auf Modellannahmen gehen auch RE-Modelle davon aus, dass Fehlerterme und (intragruppale sowie gruppenbezogene) Prädiktorvariablen unkorreliert sind (Exogenitätsannahme).[20] Ist diese Annahme nicht erfüllt, sind auch die Schätzer im RE-Modell verzerrt. Wieder stellt das Auslassen von relevanten Variablen die Hauptursache für eine Verletzung der Exogenitätsannahme dar. Im angeführten Beispiel könnte eine ausgelassene gruppenbezogene Variable ein weiterer, wirtschaftsbezogener Faktor, wie Arbeitslosenquoten, sein. Zusammenfassend kann festgehalten werden, dass RE-Modelle im Vergleich zu OLS-Verfahren gruppenbezogene Varianz über einen zusätzlichen Fehlerterm der Gruppenebene modellieren. RE-Modelle erlauben es zudem, Variablen der Gruppenebene in die Schätzung einzubeziehen, nehmen aber ebenso wie OLS-Verfahren Unkorreliertheit zwischen (gruppenbezogener) unbeobachteter Heterogenität und Prädiktorvariablen an.

3.2 Fixed-Effects-Modelle

Im Gegensatz zu RE-Modellen ist bei FE-Modellen eine Korrelation zwischen Prädiktoren und gruppenbezogenem Fehlerterm irrelevant, da sie die komplette unbeobachtete Heterogenität der Gruppenebene bereits per Design eliminieren. Die einfachste Variante eines FE-Modells ist die Schätzung einer OLS-Regression mit einer Dummy-Variablen D_j für jede im Modell vorhandene Gruppe j (und das darin gruppierte Individuum i), so dass $D_{j[i]} = 1$, wenn Individuum i Gruppe j angehört und $D_{j[i]} = 0$ falls nicht.[21]

[20] Zudem werden auch in RE-Modellen Linearität der Parameter, keine perfekte Multikollinearität, Fehlerterme (auch der Fehlerterm der Gruppenebene) mit einem Erwartungswert von null, normalverteilte Fehlerterme, konstante Varianz der Fehlerterme, keine serielle Korrelation der Fehler sowie Unkorreliertheit der Fehler (u, e) untereinander angenommen. Untersuchungseinheiten sollten eine zufällige Auswahl aus einer Grundgesamtheit sein. Darüber hinaus wird angenommen, dass auch die Untersuchungseinheiten der Gruppenebene eine zufällige Auswahl aus einer Grundgesamtheit (bzw. einem Universum von Gruppen) sind. Wenn Länder die Gruppeneinheit darstellen, ist diese Annahme nicht plausibel. Daher sollten die Ergebnisse in diesem Fall nicht als Vorhersagen über zugrundeliegende Populationen, sondern nur in Bezug auf die Länderauswahl im jeweiligen Sample interpretiert werden. Ob man in diesem Fall für die Parameter der Gruppenebene überhaupt statistische Inferenz durchführen sollte, ist eine offene Frage.

[21] Es werden J-1 Dummies geschätzt, da eine Gruppe als Referenzkategorie ausgelassen wird, um das Modell zu identifizieren.

$$y_{ij} = \beta_0 + \sum_{j=1}^{J-1} \beta_j D_{j[i]} + \beta_1 x_{ij} + e_{ij}. \quad (7)$$

Die Dummy-Variablen absorbieren die gesamte Varianz zwischen Gruppen und folglich können die Prädiktorvariablen nicht mehr mit Gruppenunterschieden in der abhängigen Variable korrelieren. Die Kontrolle der gesamten Varianz zwischen den Gruppen führt dazu, dass der Regressionskoeffizient nur aus Varianzen innerhalb der Gruppen (Within-Group-Variance) geschätzt wird. Bei Umfragedaten in mehreren Ländern, die im Querschnitt vorliegen, können Niveauunterschiede in der abhängigen Variable durch den Einbezug von Länder-Dummies absorbiert werden. Eine Schätzung von Koeffizienten basiert folglich auf Varianzen innerhalb der Länder. Unbeobachtete Heterogenität zwischen Ländern, die beispielsweise aus historischen oder institutionellen Unterschieden resultiert, wird somit per Design durch das „Herausrechnen" des Durchschnittsniveaus für jedes Land (mithilfe der Dummy-Variablen) kontrolliert. Alternativ zur Regression mit Dummy-Variablen, können FE-Modelle auch über eine Zentrierung der Variablen am jeweiligen Gruppenmittelwert spezifiziert werden.[22] Diese Prozedur wird auch als *Mean-Differencing* bzw. *Demeaning* bezeichnet und führt zu äquivalenten Ergebnissen wie bei der Verwendung von Dummy-Variablen (Andreß et al. 2013, S. 133 ff.):

$$\left(y_{ij} - \bar{y}_j\right) = \beta_1 \left(x_{ij} - \bar{x}_j\right) + \left(e_{ij} - \bar{e}_j\right). \quad (8)$$

Abb. 3 nimmt Bezug auf das oben entwickelte Beispiel und verdeutlicht die Schätzung eines FE-Modells, mit dem Länderunterschiede in der abhängigen Variable Regierungszufriedenheit absorbiert werden. Der geschätzte Koeffizient der Einkommenszufriedenheit β_1 ist der Durchschnitt der Steigungsparameter aus den drei Länderregressionen. Im Vergleich zu OLS-Regression und RE-Modell ist der Schätzer β_1 kleiner, entspricht jedoch dem unverzerrten Effekt, da Gruppenunterschiede in der abhängigen Variablen, die in unserem Beispiel mit Wirtschaftswachstum zusammenhingen und von der Prädiktorvariablen transportiert wurden, kontrolliert sind. Übrig bleibt ein Variablenzusammenhang auf Individualebene, der frei von unbeobachteter Heterogenität der Gruppenebene ist. Dabei ist zu betonen, dass zwar unbeobachtete Heterogenität der Gruppenebene keine Rolle mehr spielt, unberücksichtigte relevante Individualmerkmale (bzw. unbeobachtete Heterogenität innerhalb von Gruppen) gleichwohl den Zusammenhang konfundieren und dadurch zu möglichen Verzerrungen führen können. Entsprechend sollten relevante Kontrollvariablen innerhalb der Gruppen mitberücksichtigt werden.[23]

[22] Programmierte FE-Routinen in Statistikprogrammen (z. B. „xtreg var1 var2 etc., fe" in Stata) nutzen in der Regel diese Variante.

[23] Neben Exogenität der Prädiktorvariablen werden auch im FE-Modell Linearität, keine perfekte Multikollinearität und bestimmte Eigenschaften der Fehler (Erwartungswert von null, Normalverteilung, konstante Varianz, keine serielle Korrelation) angenommen. Auch müssen die Beobachtungseinheiten eine zufällige Auswahl aus der Grundgesamtheit darstellen (bei Paneldaten zumindest zu Beginn des Beobachtungszeitraums), damit statistische Testverfahren gültig sind.

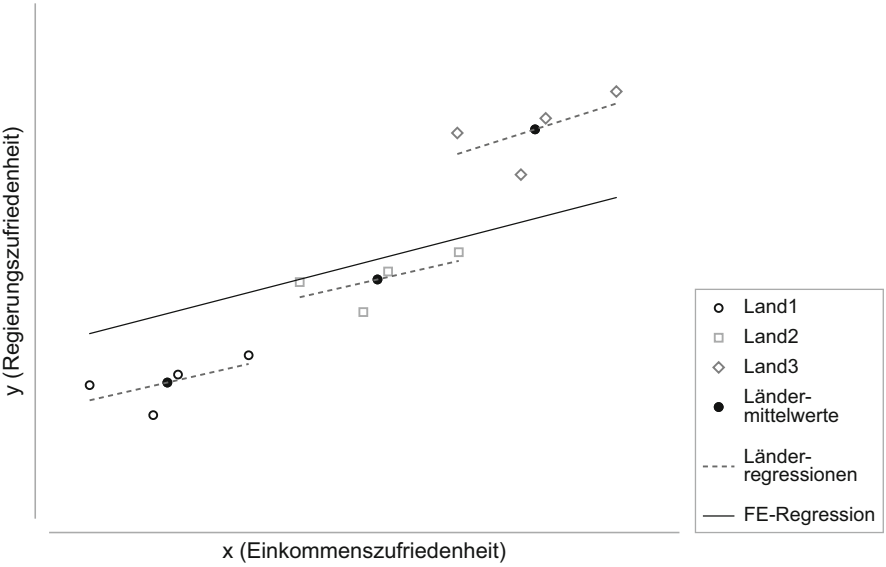

Abb. 3 FE-Modell mit gruppierten Beobachtungen

Ein Beispiel aus der vergleichenden Politikwissenschaft ist die Studie von Ziller und Schübel (2015), die den Zusammenhang zwischen persönlicher Korruptionserfahrung, politischem Vertrauen und der Wahl rechtspopulistischer Parteien untersucht. Für die Untersuchung des Individualzusammenhangs analysieren die Autoren Umfragedaten aus 12 europäischen Ländern (European Social Survey). 21 Prozent der Varianz der abhängigen Variable rechtspopulistisches Wählen sind dabei auf Länderunterschiede zurückzuführen. Um für unbeobachtete Heterogenität auf Länderebene, die möglicherweise die Individualzusammenhänge verzerrt, zu kontrollieren, werden Länder-Dummies in die Analysen einbezogen. Ein Ergebnis der Studie ist, dass individuelle Korruptionserfahrung systematisch mit geringerem politischem Vertrauen zusammenhängt, was wiederum die Wahrscheinlichkeit, eine rechtspopulistische Partei zu wählen, erhöht.

Zusammengefasst bieten FE-Modelle im Vergleich zu RE-Modellen den Vorteil, verzerrte Schätzer aufgrund unbeobachteter Heterogenität der Gruppenebene zu vermeiden. Bei Paneldaten absorbieren die fixen Effekte dabei die Varianz zwischen Personen, bei ländervergleichenden Daten die Varianz zwischen Ländern. Dadurch können Variablen der Gruppenebene nicht einbezogen werden. Diese und weitere Einschränkungen werden daher auch in den folgenden Ausführungen zu Kriterien der Modellauswahl mitberücksichtigt.

4 Kriterien für die Modellauswahl

4.1 Auswahlheuristik und Hausman-Test

Liegt gruppenbezogene unbeobachtete Heterogenität vor, ist möglicherweise die Exogenitätsannahme aufgrund ausgelassener relevanter Variablen verletzt. Ob gruppenbezogene Heterogenität in der abhängigen Variable besteht, kann mithilfe des ICC bestimmt werden. Möglicherweise führen dann die ins Modell einbezogenen Variablen bereits zu einer (fast) vollständigen Kontrolle gruppenbezogener Heterogenität und der ICC ist (nahe) null. In einem solchen Fall könnte ein OLS-Modell mit Anpassung der Standardfehler für serielle Korrelation geschätzt werden (siehe Van der Brug et al. 2007 als Anwendungsbeispiel). Oftmals sind jedoch nicht alle relevanten Variablen verfügbar. RE-Modelle modellieren gruppenbezogene Varianz mit einem zusätzlichen Fehlerterm, nehmen aber dessen Unkorreliertheit mit Variablen im Modell an. Ein Auslassen relevanter Prädiktorvariablen der Gruppenebene führt zu einer Verletzung dieser Annahme und verzerrte Regressionskoeffizienten sind die Folge. FE-Modelle kontrollieren per Design für gruppenbezogene Heterogenität. Finden sich Unterschiede in den Regressionskoeffizienten zwischen RE- und FE-Modell, deutet dies auf korrelierte gruppenbezogene Heterogenität bzw. eine Verletzung der Exogenitätsannahme hin.[24]

Der *Hausman-Test* (Hausman 1978) formalisiert einen Vergleich von RE- und FE-Modell, indem der Standardfehler der Differenz der Parameterschätzungen beider Modelle berechnet wird. Ein signifikantes Ergebnis weist die Nullhypothese über die Unabhängigkeit von gruppenbezogenem Fehlerterm und Prädiktorvariablen (und somit beide Modelle zu gleichen Ergebnissen kommen) zurück und gibt Anlass zur Schätzung eines FE-Modells. Ist der Test hingegen nicht signifikant, ist eine solche Abhängigkeit vernachlässigbar und es kann ein RE-Modell geschätzt werden. Nichtsdestotrotz ist selbst bei nicht-signifikantem Ergebnis nicht auszuschließen, dass auch geringe Korrelationen zwischen gruppenbezogenem Fehlerterm und Prädiktorvariable bestehen.[25]

[24]Wie bei den Ausführungen zu RE-Modellen betont wurde, hängt das Ausmaß, in dem eine RE-Schätzung in Richtung FE-Schätzer gewichtet ist, von der Anzahl an Perioden (Paneldaten) bzw. der durchschnittlichen Fallzahl pro Gruppe ab. Bei Querschnittsdaten mit hoher Fallzahl pro Gruppe (ländervergleichende Umfragedaten) werden daher die Schätzer von RE- und FE-Modellen kaum wesentlich voneinander abweichen. Im Falle von Paneldaten (insbesondere mit geringer Anzahl an Wellen), basieren RE-Modelle stärker auf Between-Group-Variance, was Abweichungen im Modellvergleich wahrscheinlicher macht.

[25]Clark und Linzer (2015, S. 405 f.) finden in einer Simulationsstudie, dass bei Vorliegen selbst kleinster Korrelationen zwischen gruppenbezogenem Fehlerterm und den Prädiktorvariablen FE-Modelle der Verwendung von OLS- oder RE-Modellen vorzuziehen sind. Nur in einem (praktisch eher unwahrscheinlichen) Szenario, in dem die korrelierte unbeobachtete Heterogenität nicht allzu groß, die Fallzahl innerhalb der Gruppen gering (<20) und die Varianz innerhalb der Gruppen sehr gering waren (auch als *Slow-Moving* oder *Sluggish-Data* bezeichnet), wiesen RE-Modelle eine geringere Verzerrung als FE-Schätzer auf.

Es stellt sich also die Frage, unter welchen Umständen es überhaupt sinnvoll ist, RE-Modelle zu schätzen. Zwei Hauptgründe können hier angeführt werden, eine geringere Effizienz von FE-Modellen und eine fehlende Möglichkeit, Gruppenvariablen im FE-Modell zu schätzen.

4.2 Geringe Effizienz von FE-Modellen

Geht man vom eher untypischen Fall aus, dass keine korrelierte unbeobachtete Heterogenität vorliegt und der Hausman-Test keine systematischen Unterschiede zwischen beiden Modellen findet, ist die Verwendung von RE-Modellen empfohlen. Der Grund ist, dass FE-Modelle zur Schätzung von Koeffizienten nur Varianz innerhalb von Gruppen nutzen und sie daher weniger effizient sind, d. h. ihr Standardfehler größer ist. Hinzu kommt, dass mit der Schätzung einer Dummy-Variable pro Gruppe eine nicht unerhebliche Anzahl an Freiheitsgraden verloren geht. Auch beim Mean-Differencing gehen Freiheitsgrade für jeden der anhand der Daten geschätzten Mittelwerte verloren. Die Folge geringer Freiheitsgrade ist geringe statistische Power, um signifikante Effekte zu zeigen. Je nach Datenstruktur, sind geringe Effizienz und Power mehr oder weniger relevant. Bei der Analyse von vergleichenden Umfragedaten mit Ländern als Gruppenebene und umfangreichen Stichproben besteht in der Regel genügend Varianz innerhalb der Länder und der Einbezug von 25 oder 30 Dummy-Variablen führt zu keinem nennenswerten Powerverlust. Bei Paneldaten mit vielen Befragten (d. h. einer hohen Anzahl an fixen Effekten) und geringer Variabilität gemessener Merkmale über die Zeit, stellt sich die Problematik schon eher.

In Relation zur Gefahr verzerrter Koeffizienten, ist verminderte Effizienz von FE-Modellen ein geringer Preis. Gleichzeitig sollte bei der Verwendung von FE-Modellen ein Bewusstsein über mögliche Auswirkungen der Datenstruktur, wie geringe durchschnittliche Fallzahl oder geringe Variabilität in den Gruppen, herrschen. Impliziert ein Hausman-Test, dass FE- und RE-Modelle zu äquivalenten Ergebnissen führen, sollte daher zwar ein RE-Modell geschätzt werden. Im Anwendungsfall wird dies jedoch (insbesondere bei Paneldaten) selten der Fall sein und die Verwendung von FE-Modellen bleibt der Maßstab für die Analyse gruppierter Daten.

4.3 Einbeziehung von Gruppenmerkmalen und Hybrid-Modelle

Da in FE-Modellen jegliche Gruppenunterschiede absorbiert sind, können keine Einflüsse von Gruppeneigenschaften modelliert und geschätzt werden, obwohl hier möglicherweise ein theoriebezogenes Erkenntnisinteresse besteht. Bei ländervergleichenden Umfragedaten könnte dies der Einfluss des Ländermerkmals Demokratisierungsgrad sein. Bei Paneldaten kann der Einfluss von Personenmerkmalen wie Geschlecht oder ethnischer Hintergrund (zeitkonstante Gruppeneigenschaften)

nicht im Rahmen von FE-Modellen geschätzt werden.[26] Je nach Erkenntnisinteresse kann dies eine enorme Einschränkung bedeuten. Eine Umgehung dieser Hürde ist das sogenannte Hybrid-Modell, das im Folgenden einführend vorgestellt werden (siehe Allison 2009, S. 23 ff. und Andreß et al. 2013, S. 157 ff.).

Hybrid-Modelle sind im Grunde RE- bzw. Mehrebenenmodelle, die über eine Zentrierung bzw. Desaggregation von Prädiktorvariablen eine gleichzeigte Schätzung von Between-Bestandteilen, was den Ergebnissen einer Regression durch Gruppendurchschnitte („Between-Regression") entspricht, und Within-Bestandteilen, was den Ergebnissen einer FE-Regression entspricht, erlauben. So können Variablen der Gruppenebene mit ins Modell aufgenommen werden und gleichzeitig bleiben die FE-Vorteile für Variablen mit Varianz innerhalb von Gruppen gewahrt. Bezogen auf ländervergleichende Umfragedaten im Querschnitt, bei denen Befragte in Ländern gruppiert sind, können so einerseits Individualzusammenhänge frei von unbeobachteter Heterogenität der Gruppenebene (entspricht FE-Modell) und gleichzeitig Gruppenmerkmale (z. B. Demokratisierungsgrad) geschätzt werden. Bei Paneldaten, bei denen Messzeitpunkte in Personen gruppiert sind, werden Zusammenhänge über die Zeit (z. B. ob sich eine Veränderung in der Nutzung sozialer Medien auf eine Veränderung im politischen Engagement auswirkt) als FE-Schätzer modelliert und gleichzeitig können zeitkonstante Personenmerkmale, wie das Geschlecht einer Person, mit einbezogen werden.

Die technische Illustration erfolgt anhand von Paneldaten, bei denen sich Between-Varianz auf zeitkonstante Unterschiede zwischen Befragten j bezieht und Within-Varianz auf zeitveränderliche Merkmale, die zum Zeitpunkt i (d. h. „innerhalb" von Befragten über die Zeit) gemessen wurden. Ein Hybridmodell ist gegeben als:

$$y_{ij} = \beta_0 + \beta_1\left(x_{ij} - \bar{x}_j\right) + \beta_2\bar{x}_j + \beta_3 z_j + u_j + e_{ij}. \tag{9}$$

Wobei x_{ij} eine zeitveränderliche Variable ist und $\bar{x}_j$ der Durchschnittswert dieser Variable für Person j. Dazu wird im Vorfeld eine Verrechnung des Durchschnittes mit den jeweiligen Originalwerten der Variable durchgeführt $\left(x_{ij} - \bar{x}_j\right)$.[27] Folglich repräsentiert β_1 den Within-Schätzer, der dieselben Eigenschaften wie ein Schätzer eines FE-Modells aufweist (d. h. keine unbeobachtete Heterogenität der Gruppenebene). β_2 repräsentiert den Between-Schätzer, wobei dieser nicht identisch mit Schätzern eines RE-Modells ist, sondern einer Regression mit Gruppendurchschnitten entspricht (siehe Abb. 2). Für die abhängige Variable ist keine Transformation

[26]Zwar können Gruppeneigenschaften mit Variablen innerhalb von Gruppen interagiert werden. Allerdings lässt sich damit nicht der allgemeine Einfluss einer Gruppeneigenschaft abbilden, sondern ob und wie stark der Effekt einer Intra-Gruppenvariable entlang eines bestimmten Gruppenmerkmales variiert.

[27]Dies entspricht einer Zentrierung am Gruppenmittelwert. Darüber hinaus wurden weitere Varianten der Desaggregation von Prädiktorvariablen vorgeschlagen, insbesondere wenn Prädiktorvariablen einen zeitlichen Trend aufweisen (siehe Curran und Bauer 2011 und Wang und Maxwell 2015).

notwendig, da die transformierten Variablen die entsprechende Varianzstruktur in der abhängigen Variablen adressieren (Wang und Maxwell 2015). Genuin zeitkonstante Einflüsse wie die Variable z_j können mit ins Modell einbezogen werden. β_3 repräsentiert den Between-Schätzer eines solchen zeitkonstanten Merkmals wie z. B. dem Geschlecht. Hybrid-Modelle können zur Modellierung von Paneldaten mit 3 oder mehr Beobachtungszeitpunkten herangezogen werden. Für Datenstrukturen mit lediglich 2 Beobachtungswellen können RE- oder FE-Modell geschätzt werden.

Weichen Within- und Between-Schätzer einer bestimmten Variablen im Modell voneinander ab, stellt sich die Frage, wie diese Abweichung interpretiert werden kann. Unter statistischen Gesichtspunkten liegt die Vermutung nahe, dass korrelierte unbeobachtete Heterogenität die Ursache für die Abweichung des Between-Schätzers vom unverzerrten Within-Schätzer ist. Andererseits werden Between-Effekte mitunter als langfristige, zeitlich konstante Zusammenhänge und Within-Effekte als kurzfristige, über die Zeit variierende Zusammenhänge interpretiert.[28] Diese Argumentation wäre streng genommen jedoch nur korrekt, wenn wir sicher sein könnten, dass der Between-Effekt nicht durch korrelierte unbeobachtete Heterogenität verzerrt ist. Mit Sicherheit kann dies allerdings nur angenommen werden, wenn die Between-Group-Variance absorbiert würde, womit wir wieder beim Within- bzw. FE-Schätzer sind. Es scheint daher sinnvoll, Between- und Within-Effekte einer bestimmten zeitveränderlichen Variablen nicht als inhaltlich unterschiedlichen sondern als unkontrollierten versus kontrollierten Effekt aufzufassen und vor allem den Within-Effekt zu interpretieren.

Insgesamt bleibt festzuhalten, dass Hybrid-Modelle eine sehr flexible Modellklasse darstellen, die es erlauben den Einfluss von Gruppenmerkmalen zu schätzen. Auch können Random-Slopes und Cross-Level-Interaktionen mit ins Modell aufgenommen werden. Je nach Fragestellung sollte allerdings beachtet werden, dass Between-Effekte keine sogenannten Kontexteffekte darstellen (Snijders und Bosker 2012, S. 56 ff.). Kontexteffekte sind Effekte von Gruppenmerkmalen unter Kontrolle relevanter Individualzusammenhänge und kompositioneller Unterschiede. So ist der Kontexteffekt einer aggregierten Variable (d. h. Mittelwerte pro Gruppe) die Differenz zwischen dem Between- und Within-Effekt, also den Gruppenunterschieden die über den Individualzusammenhang hinaus bestehen. Werden im Hybridmodell durch die Variablenzentrierung Between- und Within-Komponenten getrennt geschätzt, müsste bei Interesse am Kontexteffekt dieser im Nachhinein berechnet werden. Eine Alternative ist der Einbezug von unzentrierten Intragruppenvariablen (siehe auch Andreß et al. 2013, S. 157 ff.; Enders und Tofighi 2007).

[28]Zum Beispiel untersuchen Schmidt-Catran und Spies (2016), ob sich Zuwanderung in Deutschland auf die Unterstützung für wohlfahrtsstaatliche Maßnahmen auswirkt. Dabei desaggregieren die Autoren die Variable Zuwanderung in eine Within- und Between-Komponente. Die Ergebnisse des Hybrid-Modells zeigen, dass nur der Within-Schätzer einen signifikanten (negativen) Einfluss auf wohlfahrtsstaatliche Unterstützung hat, was als Beleg für die Relevanz zeitlicher Veränderungen (im Vergleich zum langfristigen Gesamtniveau) gewertet wird.

5 Hinweise für die praktische Anwendung

5.1 Wechselseitige Kausalität und „parallele Trends“

Alle in diesem Beitrag vorgestellten Modellspezifikationen nehmen an, dass die Prädiktorvariable der abhängigen Variable kausal vorgelagert ist, was über die Exogenitätsannahme ableitbar ist. Für das Vorliegen von Kausalität muss neben der Korrelation, auch ein Ausschluss von Alternativerklärungen (über Drittvariablenkontrolle) bestehen und in der kausalen Anordnung muss die Prädiktorvariable der abhängigen Variable (zeitlich) vorgelagert sein (Bollen 1989). Das Vorgelagertsein ist für manche Variablenzusammenhänge recht eindeutig (z. B. beeinflusst politisches Vertrauen nicht das Alter einer Person), während es für andere wiederum weniger eindeutig ist (z. B. die Frage, ob politisches Vertrauen soziales Vertrauen beeinflusst). Da FE-Modelle mit Paneldaten zeitliche Abweichungen vom Gruppenmittelwert modellieren, wurde hier mitunter eine Modellierung kausaler Ordnung angenommen. Jedoch wird auch im FE-Modell mit Paneldaten Kausalität nicht modelliert, sondern ist a priori vorausgesetzt (Morgan und Winship 2007; Vaisey und Miles 2017). Bestehen wechselseitige kausale Einflüsse zwischen Prädiktor und abhängiger Variable, ist die Exogenitätsannahme verletzt und verzerrte Koeffizienten sind die Folge. Hier kommt der Einsatz von verzögerten abhängigen (*lagged dependent*) Variablen in Betracht, die effektiv für den Anteil an Endogenität kontrollieren sollen. Allerdings sind diese Modelle, genauso wie OLS- und RE-Modelle, anfällig für eine Verletzung der Exogenitätsannahme aufgrund korrelierter unbeobachteter Heterogenität (Vaisey und Miles 2017). Eine Lösung ist die Kombination von verzögerten abhängigen Variablen und Fixed-Effects im Rahmen von Strukturgleichungsmodellen (siehe auch den Beitrag von Berning in diesem Band), insbesondere sogenannte *Autoregressive-Cross-Lagged-Panelmodelle* mit fixen Effekten (Allison et al. 2017; Hamaker et al. 2015). Inwieweit und unter welchen Umständen sich die dabei verwendeten verzögerten Prädiktorvariablen zur Identifikation kausaler Effekte eignen, ist wiederum eine aktuelle Diskussion in der Methodenforschung (Bellemare et al. 2017).

Eine weitere Annahme, die für die Identifikation unverzerrter Effekte eine Rolle spielt, ist die der strikten Exogenität (bei FE-Modellen mit Paneldaten auch als Annahme „paralleler Trends“ bezeichnet), welche sich auf die Unabhängigkeit der Prädiktoren vom idiosynkratischen Fehlerterm e_{ij} bezieht.[29] Zur Illustration nehmen wir an, dass der Effekt von individueller Arbeitslosigkeit auf Einstellungen zum Wohlfahrtsstaat untersucht werden soll. Dazu wird ein Panel-Individualdatensatz mit fünf Erhebungszeitpunkten als gegeben angenommen, in dem bei einigen der Befragten das Ereignis Arbeitslosigkeit während des Erhebungszeitraums eintritt (ein Wie-

[29]Bei Paneldaten bezieht sich der idiosynkratischen Fehler auf unbeobachtete zeitveränderliche Prozesse. Bei gruppierten Querschnittsdaten aus einer Umfrage, welche sich in der Regel weitaus weniger für die Identifikation kausaler Effekte eignen, bezieht sich der idiosynkratische Fehler auf unbeobachtete Befragtenmerkmale.

dereintritt ins Berufsleben wird hier einmal vernachlässigt). Wir schätzen ein FE-Modell und entledigen uns möglicher unbeobachteter Heterogenität, die mit zeitkonstanten Personenmerkmalen (wie Geschlecht) in Zusammenhang steht. Die Ergebnisse zeigen einen positiven statistisch signifikanten Zusammenhang: Werden Personen arbeitslos, befürworten sie mehr Sozialleistungen. Damit dieser Effekt unverzerrt ist, dürfen die zeitlichen Trends der Variablen „Befürwortung Sozialleistungen" für diejenigen, die arbeitslos wurden, nicht von denen derer, die nicht arbeitslos wurden, abweichen. Gäbe es unterschiedliche Verlaufskurven im Vorfeld des Ereignisses Arbeitslosigkeit (z. B. aufgrund von Unterschieden im Gesundheitszustand von Personen), könnten diejenigen, die in unserem Beispiel arbeitslos wurden, aufgrund der divergierenden Trends (und damit verbundenen unbeobachteten Prozessen, z. B. Verschlechterung des Gesundheitszustands durch Krankheit) systematisch häufiger arbeitslos geworden sein. Sie hätten also auch im kontrafaktischen Fall der Nicht-Arbeitslosigkeit eine stärkere Befürwortung von Sozialleistungen entwickelt und die beobachteten Präferenzunterschiede sind somit nicht kausal auf das Ereignis Arbeitslosigkeit zurückzuführen. Neben dem Einbezug zeitveränderlicher Kontrollvariablen, können heterogene Trends im FE-Modell mithilfe von individualspezifischen Steigungsgeraden (Brüderl und Ludwig 2015, S. 336 ff.) oder im Hybrid-Modell mithilfe einer Interaktion zwischen gruppenbezogenem Durchschnitt und Zeit-Dummies (Vaisey und Miles 2017, S. 56 f.) modelliert werden.

5.2 Robuste Standardfehler

Wie eingangs im Problemaufriss erläutert, korrigieren cluster-robuste Standardfehler für serielle Korrelation (sowie Heteroskedastizität) und fallen daher oftmals größer als konventionelle Standardfehler aus. Aufgrund der einfachen Anwendung und generellen Wirksamkeit sind cluster-robuste Standardfehler weit verbreitet. Gleichzeitig gibt es eine Reihe möglicher Komplikationen, die bei der Anwendung berücksichtigt werden sollten. Zunächst sind robuste Standardfehler nur asymptotisch korrekt und sollten daher nicht in Samples mit geringer Fallzahl angewendet werden. Darüber hinaus bezieht sich das Verfahren ausschließlich auf die Schätzung von Standardfehlern. Korrelierte unbeobachtete Heterogenität der Gruppenebene aufgrund ausgelassener Variablen und entsprechend verzerrte Regressionskoeffizienten werden hier also nicht korrigiert. Aus diesem Grund argumentieren King und Roberts (2015), dass eine Abweichung von konventionellen und robusten Standardfehlern auf das Vorliegen serieller Korrelation und/oder Heteroskedastizität hinweist, was vorrangig Anlass für eine bessere Modellspezifikation (z. B. Einbezug zusätzlicher relevanter Prädiktoren oder Interaktionen) geben sollte. Eine weitere Problematik ist, dass robuste Standardfehler bei bestimmten Korrelationsstrukturen zwischen Prädiktor und Fehler auch kleiner ausfallen können als konventionelle Standardfehler. Angrist und Pischke (2009) schlagen daher vor, die jeweils höchsten Standardfehler zu nutzen, um konservative Signifikanztests berichten zu können. Zudem wurde darauf hingewiesen, dass die Anwendung cluster-robuster Standardfehler nur optimal funktioniert, wenn die Anzahl der Gruppen hoch (60 und mehr)

und die Anzahl der Beobachtungen innerhalb der Gruppen relativ einheitlich ist (Cameron und Miller 2015).

Zusammenfassend empfiehlt sich die Verwendung cluster-robuster Standardfehler eher als ergänzende Maßnahme zur Verwendung von RE- und FE-Modellen, insbesondere bei Paneldaten. RE- und FE-Modelle korrigieren dabei serielle Korrelation aufgrund von Gruppeneffekten und die robusten Standardfehler zusätzlich Formen serieller Korrelation aufgrund zeitveränderlicher Faktoren und/oder Heteroskedastizität.

5.3 Mehrfache statistische Abhängigkeiten

Ein klassisches Beispiel für mehrfache Abhängigkeiten sind Beobachtungen, die in mehreren räumlichen Kontexten hierarchisch gruppiert sind, wie Schüler in Klassen, die wiederum in Schulen gruppiert sind, oder Befragte in Nachbarschaften, die wiederum in Gemeinden und Regionen bzw. Ländern gruppiert sind.[30] Wieviel Varianz auf die jeweiligen Gruppenstrukturen im Einzelnen zurückzuführen ist, kann über den ICC ermittelt werden. Modelliert werden solche mehrfachen Abhängigkeiten mit RE- bzw. Mehrebenenmodellen, die eine separate Zufallsvariable für jede Gruppe beinhalten.

Darüber hinaus können räumliche und zeitliche Abhängigkeiten gleichzeitig bestehen. Auch hier können RE-Modelle sowie eine Kombination aus RE-Modellen und Dummy-Variablen für einzelne Gruppierungen Verwendung finden (siehe Schmidt-Catran und Fairbrother 2016). Zum Beispiel analysieren Ziller und Helbling (2017), wie sich zeitliche Veränderungen in Antidiskriminierungspolitik und Bürgerwissen über Gleichberechtigung auf politische Unterstützung auswirken. Die dafür verwendeten Daten sind folgendermaßen strukturiert: Befragte aus wiederholten europäisch-vergleichenden Querschnittsbefragungen sind in Länderzeitpunkten gruppiert, die wiederum in Ländern gruppiert sind. Die Autoren schätzten nun ein RE-Modell, in dem Befragte in Länderzeitpunkten gruppiert sind. Um mögliche Verzerrungen der Regressionskoeffizienten durch unbeobachtete Heterogenität zwischen Ländern zu vermeiden, werden Dummy-Variablen für Länder verwendet (analog zum FE-Ansatz).

Bei Paneldaten können zusätzlich zur Kontrolle gruppenbezogener Heterogenität mittels FE-Modellen, weitere Formen zeitbezogener oder räumlicher serieller Korrelation relevant sein, für die mit cluster-robusten oder Driscoll-Kraay-Standardfehlern korrigiert werden kann.

[30] Bei einer Einbettung in mehrere *nicht*-hierarchische Kontexte kommen sogenannte *Cross-Classified*-Mehrebenenmodelle in Betracht (Snijders und Bosker 2012, S. 205 ff.).

5.4 Anzahl der Gruppen in RE-Modellen

Besteht ein inhaltliches Interesse an der Modellierung von Variablen der Gruppenebene, können RE-Modelle geschätzt werden. Dabei stellt sich die Frage, ob die Anzahl der Gruppen für die Verlässlichkeit der Schätzergebnisse eine Rolle spielt. Die Literatur zur Mehrebenenanalyse hat darauf hingewiesen, dass Schätzer für Gruppenvariablen erst in Modellen ab 30 Gruppen und mehr zu verlässlichen Ergebnissen führen (z. B. Bryan und Jenkins 2016; Stegmueller 2013). In Modellen mit geringer Gruppenanzahl (<20) und bei komplexeren Designs (z. B. mehrere Kontexteffekte oder Interaktionen), ist von Verzerrungen der Parameterschätzungen, insbesondere der Standardfehler, auszugehen. Da eine Erhöhung der Anzahl an Gruppen oftmals nicht umsetzbar ist (z. B. Analyse von Ländern ohne machbare Datenerweiterung), wurde von einigen Autoren die Verwendung bayesianischer Schätzmethoden (anstatt Maximum-Likelihood) vorgeschlagen (Stegmueller 2013).

In der Tat nimmt mit zunehmender Modellkomplexität die Anzahl der Freiheitsgrade ab und das Risiko von Ausreißern[31] zu. Nichtsdestotrotz zeigen Elff et al. (2016), dass die Koeffizientenschätzer zum Einfluss von Gruppenvariablen aus dem Maximum-Likelihood-Verfahren (ML) generell unverzerrt sind, auch für eine geringe Anzahl an Gruppen (z. B. 15). Bei einer geringen Anzahl an Gruppen sollte jedoch das Restricted Maximum-Likelihood-Verfahren (REML) angewandt werden, welches Varianzparameter unter Einbezug der durch die Koeffizientenschätzung verbrauchten Freiheitsgrade schätzt. Die Schätzung der Standardfehler und Konfidenzintervalle ist dadurch verlässlicher als im klassischen ML-Verfahren.[32] Zudem sollte die Berechnung von p-Werten auf einer t-Verteilung mit entsprechender Anpassung der im Modell vorhandenen Freiheitsgrade basieren. Das bedeutet in der Praxis, dass bei einer geringen Anzahl an Gruppen (z. B. 15) die Standardfehler der Koeffizienten der Gruppenebene substanziell geringer sein müssen, um auf die gleichen p-Werte (und inferenzstatistischen Folgerungen) wie bei Schätzern aus Modellen mit einer höheren Anzahl an Gruppen (z. B. 40) zu kommen.

Als Empfehlung lässt sich ableiten, dass RE-Modelle mit einer geringen Anzahl an Gruppen mittels REML-Verfahren geschätzt werden. Darüber hinaus sollten p-Werte von Kontexteffekten bei geringer Anzahl an Gruppen (<20) nicht mehr direkt aus dem Datenanalyseprogramm übernommen, sondern manuell mit der für die im Modell vorhandenen Freiheitsgrade entsprechenden t-Verteilung abgeleitet werden. Die Anzahl der Freiheitsgrade bezieht sich hierbei auf die Gruppenebene und kann über die Formel $n - l - 1$ (Anzahl der Gruppen – Anzahl der Koeffizienten der Gruppenebene – 1) angenähert werden. Darüber hinaus sollte bei geringer Anzahl an Gruppen möglichen einflussreichen Fällen größere Beachtung beigemessen werden, indem Ergebnisse grafisch aufbereitet, Residuendiagnostiken durchge-

[31] Als Ausreißer werden einflussreiche Fälle verstanden, die Koeffizientenschätzer maßgeblich beeinflussen und unter deren Exklusion sich die Interpretation der Ergebnisse verändern würde.

[32] Ein Grund für die dennoch häufige Verwendung von ML ist, dass Modellvergleiche mit Devianztests wie dem Likelihood-Ratio-Test auf ML-Schätzungen beruhen müssen.

führt und/oder explizite Tests auf Ausreißer berechnet werden (Snijders und Bosker 2012, S. 161 ff.).

6 Fazit

In der empirisch-vergleichenden Politikwissenschaft ist die Analyse von gruppierten bzw. geclusterten Datenstrukturen weit verbreitet. RE-, FE- und Hybrid-Modelle ermöglichen die Modellierung von Datenstrukturen mit statistisch voneinander abhängigen Beobachtungen. Der vorliegende Beitrag nahm zwei Konsequenzen statistischer Abhängigkeit in den Blick. Zum einen kann serielle Korrelation zu verzerrten Standardfehlern und invaliden Teststatistiken führen. Die Kontrolle relevanter Variablen sowie die Verwendung von RE- und FE-Modellen können hier Abhilfe verschaffen. Zum anderen führen statistische Abhängigkeiten zu gruppenbezogenen Unterschieden, die eine Verletzung der Exogenitätsannahme und verzerrte Koeffizienten nach sich ziehen können. RE-Modelle modellieren zwar unbeobachtete Heterogenität der Gruppenebene mit einem zusätzlichen Fehlerterm und können den Einfluss von Gruppenmerkmalen schätzen, nehmen aber weiterhin Unkorreliertheit zwischen gruppenbezogenem Fehlerterm und den im Modell befindlichen Prädiktorvariablen an. FE-Modelle machen diese Annahme nicht, da sie per Design alle gruppenbezogene Heterogenität kontrollieren.

Abschließend sollen die wichtigsten Schritte hin zur geeigneten Modellspezifikation nachvollzogen werden. In der Praxis sollte man sich zunächst einen detaillierten Überblick über die Struktur der vorliegenden Daten machen. Liegt eine räumliche und/oder zeitliche Gruppierung der Daten vor? Was sind relevante Gruppen (z. B. Länder, Regionen, Städte oder Individuen bei Paneldaten) und können die Daten entsprechend eingeteilt werden? Liegt eine mehrfache Gruppierung vor?

Daraufhin sollte ein ICC berechnet werden, um das Ausmaß an statistischer Abhängigkeit bzw. serieller Korrelation in der abhängigen Variable abzuschätzen. Selbst bei kleinen Anteilen gruppenbezogener Unterschiede an der Gesamtvarianz (z. B. 2 Prozent), ist die Anwendung von RE- oder FE-Modellen zu empfehlen. Ein Hausman-Test gibt Aufschluss inwieweit Ergebnisse aus RE- und FE-Modellen voneinander abweichen. Ist dieser nicht signifikant, kann aus Effizienzgründen ein RE-Modell geschätzt werden. Ist er hingegen signifikant, sollte ein FE-Modell geschätzt werden. Bei der Anwendung von FE-Modellen sollte ein Bewusstsein darüber herrschen, ob die relevanten Variablen hinreichend Varianz innerhalb von Gruppen aufweisen und ob die kausale Anordnung der Variablen plausibel ist. Bei Panel- und TSCS-Daten kommt hinzu, dass Tests für serielle Korrelation aufgrund zeitveränderlicher Faktoren durchgeführt werden sollten und gegebenenfalls zusätzlich mit robusten Standardfehlern gearbeitet werden muss. Darüber hinaus sollten in Modellen mit zeitveränderlichen Daten Dummy-Variablen für Zeitpunkte einbezogen werden, um so für nicht-beobachtete zeitbezogene Faktoren, die allen Beobachtungen gemein sind (z. B. exogene Schocks oder Trends), zu kontrollieren.

Besteht das zentrale Forschungsinteresse an dem Einfluss von Gruppenmerkmalen, wie im Rahmen der Mehrebenenanalyse (siehe den Beitrag von Pötschke in

diesem Sammelband), kann entsprechend ein RE-Modell oder Hybrid-Modell geschätzt werden. Um mögliche Verzerrungen aufgrund unbeobachteter Heterogenität zu vermeiden, sollte auf den Einbezug relevanter Kontrollvariablen (insbesondere auch der Gruppenebene) geachtet werden. Ziel ist eine möglichst umfangreiche Reduktion gruppenbezogener Abweichungen, wobei hier ein Bewusstsein über Konsequenzen einer geringen Anzahl an Gruppen, vorhandener Freiheitsgrade, möglicher Kollinearitäten zwischen Prädiktorvariablen und einflussreicher Fälle auf Gruppenebene herrschen sollte. Darüber hinaus können hierbei Random-Slopes und Cross-Level-Interaktionen die Modellanpassung an die Daten verbessern und theoretisch fundierte Hypothesen testen.

7 Kommentierte Literaturhinweise

Eine Einführung in RE- und FE-Modelle mit Blick auf Längsschnittdaten bieten Wooldridge (2009) (Kap. „Advanced Panel Data Methods“) sowie Andreß et al. (2013). Für beide Referenzen existieren Web-Ressourcen, von denen man die Syntaxen der angeführten Beispiele (z. B. in Stata oder R) herunterladen kann. Eine umfangreiche und detaillierte Einführung in RE- und FE-Modelle (sowie die technische Umsetzung in Stata) findet sich bei Rabe-Hesketh und Skrondal (2012). Hier werden zahlreiche Anwendungsbeispiele sowohl für gruppierte Querschnittsdaten (Mehrebenendaten) als auch Längsschnittdaten gezeigt. Die Autoren besprechen neben kontinuierlichen abhängigen Variablen auch kategoriale und Zählvariablen. Sollen gruppierte Querschnittsdaten unter Verwendung von RE-Modellen in Form von Mehrebenenmodellen ausgewertet werden, empfiehlt sich Snijders und Bosker (2012) als praxisorientierte Einführung. Einblicke in Grundlagen kausaler Inferenz bieten Morgan und Winship (2007). Eine entsprechende Einordnung von RE- und FE-Modellen ist bei Vaisey und Miles (2017) zu finden.

Literatur

Allison, Paul D. 2009. *Fixed effects regression models*. Los Angeles: Sage.

Allison, Paul D., Richard Williams, und Enrique Moral-Benito. 2017. Maximum likelihood for cross-lagged panel models with fixed effects. *Socius* 3:1–17.

Andreß, Hans-Jürgen, Katrin Golsch, und Alexander W. Schmidt. 2013. *Applied panel data analysis for economic and social surveys*. Heidelberg: Springer Science & Business Media.

Angrist, Joshua D., und Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton/Oxford: Princeton University Press.

Beck, Nathaniel, und Jonathan N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89(3): 634–647.

Bellemare, Marc F., Takaaki Masaki, und Thomas B. Pepinsky. 2017. Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics* 79(3): 949–963.

Bollen, Kenneth A. 1989. *Structual equations with latent variables*. New York: Wiley.

Brüderl, Joself, und Volker Ludwig. 2015. Fixed-effects panel regression. In *The SAGE handbook of regression analysis and causal inference*, Hrsg. H. Best und C. Wolf, 327–357. Los Angeles: Sage.

Bryan, Mark L., und Stephen P. Jenkins. 2016. Multilevel modelling of country effects: A cautionary tale. *European Sociological Review* 32(1): 3–22.
Cameron, A. Colin, und Douglas L. Miller. 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 50(2): 317–372.
Clark, Tom S., und Drew A. Linzer. 2015. Should I use fixed or random effects? *Political Science Research and Methods* 3(2): 399–408.
Curran, Patrick J., und Daniel J. Bauer. 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology* 62:583–619.
Drukker, David M. 2003. Testing for serial correlation in linear panel-data models. *Stata Journal* 3 (2): 168–177.
Elff, Martin, Jand P. Heisig, Merlin Schaeffer, und Susumu Shikano. 2016. No need to turn Bayesian in multilevel analysis with few clusters: How frequentist methods provide unbiased estimates and accurate inference. *Working Paper.* https://doi.org/10.31235/osf.io/z65s4. Zugegriffen am 08.01.2018.
Enders, Craig K., und Davood Tofighi. 2007. Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods* 12(2): 121.
Gelman, Andrew. 2005. Analysis of variance – Why it is more important than ever. *The Annals of Statistics* 33(1): 1–53.
Greene, William H. 2003. *Econometric analysis*, 5. Aufl. Upper Saddle River: Prentice Hall.
Hakhverdian, Armen, und Quinton Mayne. 2012. Institutional trust, education, and corruption: A micro-macro interactive approach. *The Journal of Politics* 74(3): 739–750.
Hamaker, Ellen L., Rebecca M. Kuiper, und Raoul P. P. P. Grasman. 2015. A critique of the cross-lagged panel model. *Psychological Methods* 20(1): 102–116.
Hausman, Jerry A. 1978. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* 46(6): 1251–1271.
King, Gary, und Margaret E. Roberts. 2015. How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis* 23(2): 159–179.
Lewis-Beck, Michael S., und Mary Stegmaier. 2000. Economic determinants of electoral outcomes. *Annual Review of Political Science* 3(1): 183–219.
Morgan, Stephen L., und Christopher Winship. 2007. *Counterfactuals and causal analysis: Methods and principles for social research*. Cambridge, UK: Harvard University Press.
Moulton, Brent R. 1986. Random group effects and the precision of regression estimates. *Journal of Econometrics* 32(3): 385–397.
Rabe-Hesketh, Sophia, und Anders Skrondal. 2012. *Multilevel and longitudinal modeling using Stata*. College Station: Stata Press Publication.
Schmidt-Catran, Alexander W., und Malcolm Fairbrother. 2016. The random effects in multilevel models: Getting them wrong and getting them right. *European Sociological Review* 32(1): 23–38.
Schmidt-Catran, Alexander W., und Dennis C. Spies. 2016. Immigration and welfare support in Germany. *American Sociological Review* 81(2): 242–261.
Snijders, Tom A. B., und Roel J. Bosker. 2012. *Multilevel analysis. An introduction to basic and advanced multilevel modeling*, 2. Aufl. Los Angeles: Sage.
Stegmueller, Daniel. 2013. How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science* 57(3): 748–761.
Tilley, James, Anja Neundorf, und Sara Hobolt. 2018. When the pound in people's pocket matters: How changes to personal financial circumstances affect party choice. *The Journal of Politics* 80 (2): 555–569.
Vaisey, Stephen, und Andrew Miles. 2017. What you can – and can't – do with three-wave panel data. *Sociological Methods & Research* 46(1): 44–67.
Van der Brug, Wouter, Cees Van der Eijk, und Mark Franklin. 2007. *The economy and the vote: Economic conditions and elections in fifteen countries*. Cambridge, MA: Cambridge University Press.
Verbeek, Marno. 2004. *A guide to modern econometrics*. Chichester: Wiley.

Wang, Lijuan Peggy, und Scott E. Maxwell. 2015. On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods* 20(1): 63–83.

Wooldridge, Jeffrey M. 2009. *Introductory econometrics: A modern approach*. Boston: Cengage Learning.

Ziller, Conrad, und Marc Helbling. 2017. Antidiscrimination laws, policy knowledge, and political support. *British Journal of Political Science*. First View: 1–18. https://doi.org/10.1017/S0007123417000163.

Ziller, Conrad, und Thomas Schübel. 2015. „The pure people“ versus „the corrupt elite“? Political corruption, political trust and the success of radical right parties in Europe. *Journal of Elections, Public Opinion and Parties* 25(3): 368–386.