36 Kausalanalyse mit Paneldaten

Josef Brüderl*

Universität Mannheim

Zusammenfassung. Der Beitrag gibt eine anwendungsorientierte Einführung in die Kausalanalyse mit Paneldaten. Es wird versucht, dem Anwender die Grundlogik der Paneldatenanalyse nahe zu bringen. Im ersten Abschnitt wird eine intuitive Einführung in die Paneldatenanalyse gegeben. Zuerst werden die Vorzüge von Paneldaten für die Kausalanalyse herausgearbeitet, dann werden die grundlegenden Panelregressionsmodelle vorgestellt. Deren „Mechanik" wird schließlich anhand eines fiktiven Datensatzes demonstriert. Im zweiten Abschnitt erfolgt eine präzisere Vorstellung verschiedener (linearer) Panelregressionsmodelle. Diese Modelle werden im dritten Abschnitt eingesetzt, um mit Daten des SOEP 1984–2007 den Effekt einer Heirat auf die Lebenszufriedenheit zu untersuchen. Im vierten Abschnitt werden einige nicht-lineare Panelregressionsmodelle vorgestellt. Der Beitrag endet mit der Diskussion irreführender Argumentationen und suboptimaler Modellklassen.

1 Einführung in die Paneldatenanalyse

Paneldaten erhält man durch die wiederholte Messung derselben Variablen an denselben Untersuchungseinheiten. In der Sozialforschung werden Paneldaten meist mittels eines Panelsurvey erhoben: Hierbei sind die Untersuchungseinheiten Personen und die Erhebungsmethode ist die Befragung. Personen werden (meist) Jahr für Jahr kontaktiert und mit demselben Fragebogen befragt, wobei jeweils die aktuellen Werte der Variablen erhoben werden (prospektiver Panelsurvey). Paneldaten könnten auch retrospektiv erhoben werden, was aber auf Grund von Erinnerungsproblemen bei den Befragten eher problematisch ist. Im Folgenden gehe ich meist implizit davon aus, dass man Panelsurveydaten vorliegen hat.

Paneldaten bieten gegenüber Querschnittsdaten einige gewichtige Vorteile (s. u.), weshalb sie über kurz oder lang die Sozialforschung dominieren werden. Der in Deutschland bekannteste Panelsurvey ist das Sozio-ökonomische Panel (SOEP), welches seit 1984 läuft. Im Jahre 2008 wurden im Rahmen des SOEP etwa 20.000 Personen befragt, davon etwa 2.500 zum 25. male. Aufgrund seiner langen Laufzeit kann das SOEP inzwischen für eine Vielzahl von Fragestellungen verwendet werden. Das SOEP ist ein „Allzweck"-Panel. Inzwischen wurden auch vermehrt spezialisierte, groß angelegte Panelsurveys gestartet. Die erste Welle des „Survey on Health, Ageing and Retirement

* Für hilfreiche Anmerkungen danke ich Vera Troeger und Johannes Huinink. Die Kollegin bzw. der Kollege teilen allerdings nicht jeden meiner Standpunkte.

S. 963–994 in: Christof Wolf & Henning Best, Hg. (2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften

C. Wolf , H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse,*
DOI 10.1007/978-3-531-92038-2_36,

in Europe“ (SHARE) wurde 2004 erhoben, im Jahr 2008 ging die erste Welle des deutschen Beziehungs- und Familienpanels (pairfam) ins Feld und 2009 startete das nationale Bildungspanel (NEPS).

Bei Panelsurveys ist üblicherweise die Zahl der Untersuchungseinheiten (N) groß und die Zahl der Wiederholungsmessungen (T) klein. Mit dem umgekehrten Fall – kleines N und relativ großes T – hat man es häufig bei Analysen über Länder/Regionen zu tun. Hier hat man typischerweise eine Reihe von Makroindikatoren für mehrere Länder über viele Jahre vorliegen („Pooled Cross-Section Time-Series“). Im Fall von $N = 1$ spricht man von Zeitreihendaten. Da es sich insbesondere bei gepoolten Zeitreihendaten auch um Paneldaten handelt, kann man sie auch mit den im Folgenden vorgestellten Verfahren analysieren. Bei solchen „Makro-Paneldaten“ stellen sich jedoch besondere Probleme (z. B. zeitliche und räumliche Autokorrelation), weshalb für deren Analyse spezielle Verfahren entwickelt wurden (siehe Kapitel 40 in diesem Handbuch), die im Folgenden jedoch aus Platzgründen nicht behandelt werden können.

Paneldaten besitzen gegenüber Querschnittsdaten mindestens drei gewichtige Vorteile:

(a) Sie erlauben die Verfolgung der individuellen Dynamik sozialer und psychischer Prozesse (intraindividuelle Veränderung).
(b) Sie erlauben die zeitliche Abfolge von Veränderungen festzustellen (kausale Reihenfolge).
(c) Sie ermöglichen es, das Problem unbeobachteter Heterogenität abzumildern.

Während die Vorteile (a) und (b) offensichtlich sind und auch in jedem Lehrbuch gewürdigt werden, ist Vorteil (c) noch viel zu wenig bekannt. Unbeobachtete Heterogenität ist ein fundamentales Problem nicht-experimenteller Sozialforschung und deshalb ist die Möglichkeit, es mittels Paneldaten abzumildern, nicht hoch genug einzuschätzen. Infolgedessen wird im Folgenden die Betonung auf Verfahren liegen, die das Potential von Paneldaten zur Reduzierung des Problems unbeobachteter Heterogenität ausschöpfen.

Zur programmtechnischen Umsetzung einer Panelanalyse kann man Stata empfehlen. Stata verfügt über ein mächtiges Panelanalyse-Paket (die `xt`-Kommandos). Inzwischen existieren auch einige Einführungen in die Paneldatenanalyse mit Stata (Brüderl 2005; Allison 2009; Cameron & Trivedi 2009).

1.1 Kontrafaktische Kausalanalyse und Paneldaten

Bevor die Verfahren dargestellt werden, ist es hilfreich, sich die Grundidee des kontrafaktischen Kausalitätsbegriffs zu vergegenwärtigen (hier aus didaktischen Gründen stark vereinfachend, präziser in den Kapiteln 2 und 35 in diesem Handbuch). Wir wollen der Einfachheit halber den kausalen Effekt einer dichotomen Variable X auf eine metrische Variable Y betrachten. Der üblichen Terminologie folgend ist X die unabhängige und Y die abhängige Variable (Outcome). X hat die Ausprägungen 0 und 1. In der Literatur bezeichnet man $X = 1$ in Anlehnung an die in Experimenten übliche Terminologie als „Treatment“ und $X = 0$ als „Control“. Der kausale Effekt von X auf der Ebene eines Individuums (Einheitseffekt) ist dann definiert als

$$\Delta_i = Y_{i,t_0}^{X=1} - Y_{i,t_0}^{X=0}, \tag{1}$$

wobei i der Personenindex ist und t_0 einen Zeitpunkt indiziert.[1] Der kausale Effekt ist also die Veränderung der abhängigen Variable, wenn die Person i von „Control“ ($X = 0$) zum „Treatment“ ($X = 1$) wechselt – *und zwar gemessen zu ein und demselben Zeitpunkt* t_0. Das macht konzeptionell zwar sehr viel Sinn, ist aber in der realen Welt so nicht umsetzbar. In der realen Welt kann eine Person faktisch zu einem Zeitpunkt nur einen Zustand einnehmen. Der andere Zustand ist hypothetisch, eben kontrafaktisch. Kausalität ist im kontrafaktischen Modell definiert durch den Vergleich eines faktischen und eines kontrafaktischen Outcomes. Ein Kausaleffekt kann deshalb nie direkt gemessen werden. Er kann in praktisch durchführbaren Untersuchungsdesigns bei Gültigkeit von bestimmten Annahmen nur indirekt erschlossen werden.

In einem (nicht-experimentellen) Querschnittsdesign erschließt man nun den Einheitseffekt aus dem Vergleich verschiedener Personen i und j, die sich nur in X unterscheiden:

$$\Delta_i^B = Y_{i,t_0}^{X=1} - Y_{j,t_0}^{X=0}. \tag{2}$$

Dies ist der so genannte *Between-Schätzer*, wobei die zentrale Annahme ist, dass sich i und j nur in X unterscheiden (keine unbeobachtete Heterogenität). Dies ist in der Realität normalerweise nicht der Fall, weshalb man etwa mittels Regressionen oder Matching-Verfahren für beobachtete Unterschiede zu kontrollieren versucht. In vielen Fällen bleiben allerdings relevante Unterschiede unbeobachtet, weshalb auf Querschnittsdaten beruhende Kausalschlüsse höchst unsicher sind. Da die Sozialforschung bisher überwiegend auf Querschnittsdaten aufbaut, kann man festhalten: *Unbeobachtete Heterogenität ist ein zentrales Problem der nicht-experimentellen Sozialforschung.* Der tiefere Grund hierfür ist, dass der Einheitseffekt aus dem Vergleich verschiedener Personen erschlossen werden muss.

Anders ist die Situation bei einem Paneldesign. Hier kann man den Einheitseffekt aus dem Vergleich derselben Person i zu verschiedenen Zeitpunkten erschließen:

$$\Delta_i^W = Y_{i,t_1}^{X=1} - Y_{i,t_0}^{X=0}. \tag{3}$$

Dies ist der so genannte *Within-Schätzer*. Die zentrale Annahme ist hier, dass ein und dieselbe Person sich über die Zeit nur in X unterscheidet (keine zeitveränderliche unbeobachtete Heterogenität). In den meisten Situationen dürfte die Annahme, dass keine zeitveränderliche intraindividuelle unbeobachtete Heterogenität vorliegt, eher erfüllt sein, als die Annahme, dass keine interindividuelle unbeobachtete Heterogenität vorliegt. Deshalb sind Kausalschlüsse auf der Basis von Paneldaten im Allgemeinen sicherer als Kausalschlüsse anhand von Querschnittsdaten. *Der zentrale „Trick“ ist, dass man bei Vorliegen von Paneldaten Kausalanalyse anhand eines intertemporalen Vergleichs ein und derselben Person betreiben kann (Within-Schätzer) und nicht auf den Vergleich zwischen Personen angewiesen ist (Between-Schätzer).* Das Ganze beruht natürlich auf der Voraussetzung, dass man in den Paneldaten genügend Personen hat,

[1] Der durchschnittliche Kausaleffekt auf Populationsebene kann dann durch Durchschnittsbildung über alle Einheitseffekte berechnet werden.

bei denen sich X über die Zeit ändert. Gibt es keine intraindividuelle Varianz auf X, so kann man den Within-Schätzer nicht einsetzen.

Die Gültigkeit des Within-Schätzers ist allerdings durch Alterseffekte (Y ändert sich aufgrund des Alterungsprozesses) und Periodeneffekte bedroht (Y ändert sich aufgrund exogener Einflüsse). Dieses Problem kann man jedoch durch einen Vergleich mit einer Person j, bei der sich X nicht über die Zeit ändert, in den Griff bekommen:

$$\Delta_i^{\mathrm{DID}} = (Y_{i,t_1}^{X=1} - Y_{i,t_0}^{X=0}) - (Y_{j,t_1}^{X=0} - Y_{j,t_0}^{X=0})\,. \tag{4}$$

Dies ist der so genannte *Difference-in-Differences-Schätzer* (DID). Hier wird der Alters-/ Periodeneffekt anhand der Veränderung der Person j geschätzt und aus dem Within-Schätzer „herausgerechnet". Der DID-Schätzer wird auch zur Berechnung des Kausaleffektes bei Vorliegen von Experimentaldaten (mit Vorher-Nachher-Messung) verwendet. Insofern kann man mit Paneldaten den Kausaleffekt analog zum Experiment berechnen. Natürlich fehlen bei Paneldaten die beiden wesentlichen Merkmale eines Experiments: Randomisierung und kontrollierte Stimulussetzung. Die fehlende Randomisierung wird allerdings durch die Vorher-Messung kompensiert, wodurch eventuelle Unterschiede in der Versuchs- und Kontrollgruppe kontrolliert werden. Dies ist eine deutliche Verbesserung im Vergleich zur Situation beim Vorliegen von Querschnittsdaten. Allerdings ist selbst mit Paneldaten das Fehlen der kontrollierten Stimulussetzung nicht kompensierbar. Gibt es etwa eine Selbstselektion in das Treatment ($P(X = 1)$ steigt, wenn sich Y über die Zeit verändert), so wird auch der anhand von Paneldaten geschätzte Kausaleffekt verzerrt sein. Die Sicherheit von auf Paneldaten basierenden Kausalschlüssen wird also irgendwo zwischen der mit Querschnitts- und Experimentaldaten erzielbaren Sicherheit liegen. Wo immer möglich, sollte man deshalb in der nicht-experimentellen Sozialforschung Paneldaten verwenden.

1.2 Eine intuitive Einführung in Panelregressionsmodelle

Die erhöhte Sicherheit der Kausalschlüsse mit Paneldaten ist allerdings nur dann realisierbar, wenn man statistische Verfahren verwendet, die auf dem Within-Prinzip beruhen. Nicht bei allen Panelregressionsverfahren ist dies der Fall. Im Folgenden sollen einige dieser Verfahren vorgestellt werden. Wir betrachten zuerst den Fall einer linearen Regression mit metrischer abhängiger Variable Y. Um eine Panelregression durchführen zu können, muss man die Daten „poolen", d. h. man legt die Beobachtungen einer Welle in jeweils einer eigenen Datenzeile ab. Der resultierende Datensatz hat mithin $N \cdot T$ Fälle (man sagt auch oft: $N \cdot T$ Personenjahre).[2]

Das einfachste Panelregressionsmodell erhält man, wenn man auf die gepoolten Daten ein lineares Regressionsmodell anwendet (zur linearen Regression siehe Kapitel 24 in diesem Handbuch):

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it},\; i = 1, \ldots, N,\; t = 1, \ldots, T\,, \tag{5}$$

[2] Im Folgenden wird unterstellt, dass die Zahl der Wellen für alle Personen gleich ist (balanciertes Panel). Die Schätzer sind aber leicht für den Fall eines nicht-balancierten Panels verallgemeinerbar (oft muss man T nur durch T_i ersetzen).

wobei i der Personenindex ist und t der Zeitindex. $\mathbf{x}_{it}$ ist ein Vektor, der die gemessenen Werte der K unabhängigen Variablen einer Person i zum Zeitpunkt (in der Welle) t enthält. Die unabhängigen Variablen können zeitkonstant (z. B. Geschlecht, Geburtsjahr) oder zeitveränderlich (z. B. Einkommen, Familienstand) sein. $\boldsymbol{\beta}$ ist ein Vektor mit K Regressionskoeffizienten. u ist ein Fehlerterm, der die üblichen Regressionsannahmen erfüllt. Dann ist der OLS-Schätzer von $\boldsymbol{\beta}$ konsistent (*Pooled-OLS oder POLS*).

POLS nützt die Paneldaten nur insofern aus, als durch die erhöhte Fallzahl die Schätzer präziser werden. Die POLS-Schätzer beruhen jedoch auch auf der Between-Variation. Insofern werden die POLS-Schätzer bei Vorliegen von unbeobachteter Heterogenität verzerrt sein. Dann ist nämlich eine wichtige Regressionsannahme verletzt: Die X-Variablen sind mit dem Fehlerterm korreliert – $\mathrm{Cov}(x_{it}, u_{it}) \neq 0$ – und die POLS-Schätzer sind verzerrt. Im Vergleich zur Querschnittsanalyse hat sich kaum etwas verbessert.

Auf dem Weg zu einem hilfreicheren Panelregressionsmodell ist es nützlich, den Fehlerterm in zwei Komponenten zu zerlegen (Fehlerkomponenten Zerlegung): $u_{it} = \alpha_i + \varepsilon_{it}$. Hierbei ist α_i ein personenspezifischer, zeitkonstanter Fehlerterm, ε_{it} erfasst die restlichen unbeobachteten Größen, die über Personen und die Zeit variieren (idiosynkratischer Fehler). α_i könnte (je nach Anwendung) unbeobachtete Merkmale der Personen enthalten, die sich über die Zeit nicht verändern, wie etwa Attraktivität und Intelligenz. Diese Zerlegung ist rein formal immer möglich, allerdings ist sie im Falle von Querschnittsdaten nutzlos, denn man kann die beiden Fehlerkomponenten dann nicht identifizieren. *Mit Paneldaten aber – und das ist ein entscheidender Vorteil von Paneldaten – kann man die personenspezifischen Fehler schätzen bzw. „herausrechnen".* Damit kann dieser Teil der unbeobachteten Heterogenität die Schätzer nicht mehr verzerren. Um das zu verstehen, betrachten wir das Fehlerkomponenten-Modell:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \,. \tag{6}$$

Der Fehler ist in die beiden Komponenten zerlegt und die Konstante ist weggelassen, weil sie mit den personenspezifischen Fehlern kollinear wäre (man kann die α_i auch als personenspezifische Konstanten auffassen). Schätzt man dieses Modell mit POLS, so dürfen die X-Variablen mit *beiden* Fehlertermen nicht korreliert sein. Mit Paneldaten kann man die Situation nun aber verbessern, indem man die α_i herausrechnet. Dazu führt man zuerst die so genannte Between-Transformation aus:

$$\bar{y}_i = \bar{\mathbf{x}}'_i\boldsymbol{\beta} + \alpha_i + \bar{\varepsilon}_i \,. \tag{7}$$

Man bildet für jede Person die Mittelwerte der Daten über die Zeit. Damit beseitigt man die Within-Variation (man hat sie herausgemittelt), übrig bleibt die Between-Variation (die Variation der Mittelwerte zwischen den Personen). Zieht man nun (7) von (6) ab (Within-Transformation), so erhält man:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) \,. \tag{8}$$

In diesem Modell ist von allen Variablen die Between-Variation abgezogen, es bleibt nur die Within-Variation übrig. Ein auf Modell (8) aufsetzendes Regressionsmodell benutzt mithin nur die Within-Variation. In der Terminologie des letzten Unterabschnitts handelt es sich um einen Within-Schätzer. Technisch gesehen sind die personenspezifischen Fehler in Modell (8) heraus gefallen, übrig sind nur die idiosynkratischen Fehler. Wendet man deshalb POLS auf das Modell (8) an, so muss man für eine konsistente Schätzung von $\boldsymbol{\beta}$ nur annehmen, dass die X-Variablen und die idiosynkratischen Fehler unkorreliert sind. Personenspezifische unbeobachtete Heterogenität verzerrt den POLS-Schätzer nach der Within-Transformation nicht mehr.

Genau dies ist der große Vorteil von Paneldaten gegenüber Querschnittsdaten: Während eine Querschnittsregression einen Between-Schätzer liefert, der von personenspezifischer unbeobachteter Heterogenität verzerrt wird, liefert eine Panelregression (nach der Within-Transformation) einen Within-Schätzer, der von personenspezifischer unbeobachteter Heterogenität nicht verzerrt wird.

Die Within-Transformation ist an die Voraussetzung gebunden, dass die α_i zeitkonstante Größen sind, so genannte fixe Effekte. Unter dieser Annahme heißt (6) das Fixed-Effects-Modell (FE-Modell). Durch die Within-Transformation werden die fixen Effekte herausgerechnet und der POLS-Schätzer von (8) ist konsistent auch wenn personenspezifische unbeobachtete Heterogenität in den Ursprungsdaten steckt. Dies nennt man den *Fixed-Effects-Schätzer* (FE-Schätzer). Der FE-Schätzer ist ein Within-Schätzer.

1.3 Ein didaktisches Beispiel

Anhand eines Beispiels mit fiktiven Daten soll im Folgenden gezeigt werden, wie ein FE-Schätzer trotz des Vorhandenseins von personenspezifischer unbeobachteter Heterogenität eine unverzerrte Schätzung liefert (ausführlicher bei Brüderl 2005). Aus didaktischen Gründen belegen wir die fiktiven Daten mit einer inhaltlichen Bedeutung. Wir wollen untersuchen, ob eine Heirat dazu führt, dass Männer anschließend mehr verdienen. Viele Querschnittsstudien haben gezeigt, dass dies der Fall ist. Es wurden verschiedene Kausalmechanismen für diese Beobachtung vorgeschlagen, z. B. könnte durch die Heirat ein Entlastungseffekt eintreten (die Ehefrauen nehmen „lästige“ Hausarbeit ab), weshalb Männer nach der Heirat produktiver arbeiten können. Allerdings liegt der Verdacht nahe, dass es sich hier nicht um einen Kausaleffekt, sondern um ein durch unbeobachtete Heterogenität verzerrtes Ergebnis handelt. Gut verdienende Männer könnten attraktivere Heiratspartner sein, weshalb sie eher Heiraten. Nicht die Heirat erhöht den Verdienst, sondern Männer mit mehr Humankapital selektieren sich in die Ehe (bzw. werden von den Frauen selektiert). Mit Paneldaten und FE-Schätzern konnte die neuere Literatur zeigen, dass tatsächlich kein Kausaleffekt vorliegt, dass das Querschnittsergebnis allein auf unbeobachteter Heterogenität beruht (Ludwig & Brüderl 2009).

Aus didaktischen Gründen konstruieren wir hier dennoch einen Datensatz, in dem ein Kausaleffekt der Heirat vorliegt. Abbildung 1 zeigt die Daten in graphischer Form. [3] Es liegt ein Panel mit sechs Wellen ($T = 6$) für vier Männer ($N = 4$) vor. Die abhängige

[3] Die Daten („Panelanalyse fiktiv.dta“) und ein Auswertungsfile („Panelanalyse fiktiv.do“) können von der Webseite des Handbuchs heruntergeladen werden.

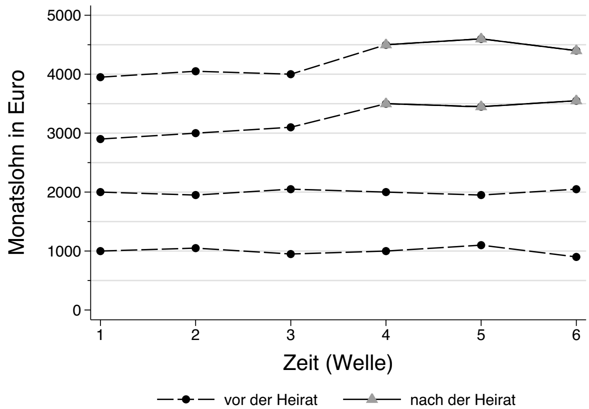

Abb. 1: Die fiktiven Lohnkarrieren von vier Männern

Variable ist der Monatslohn in Euro, die unabhängige Variable ist eine Heirats-Dummy (0 steht für unverheiratet, 1 für verheiratet). Wir haben zwei „Geringverdiener", die während der Beobachtungsdauer nicht heiraten. Ihr Lohn beträgt – bis auf zufällige Fluktuationen – 1000 bzw. 2000 Euro. Andererseits haben wir zwei „Besserverdiener", die zwischen den Wellen drei und vier heiraten. Die Heirat bewirkt, dass der Lohn jeweils um 500 Euro ansteigt. Der in diese Daten eingebaute Kausaleffekt ist also 500 Euro. Zusätzlich ist aber auch noch unbeobachtete Heterogenität eingebaut: Die Heiratenden sind keine Zufallsauswahl, sondern die Heirat hängt systematisch an unbeobachteten Faktoren, die einen hohen Lohn verursachen. Es sind die „Besserverdienenden", die heiraten.

Im Folgenden wollen wir schauen, welche Ergebnisse wir mit diesen Daten anhand einer Querschnittsregression, einer POLS-Regression und einer FE-Regression erhalten (Tabelle 1). Wir schätzen jeweils eine Regression mit dem Lohn als abhängiger Variable und der Heiratsdummy als unabhängiger Variable. Der Regressionskoeffizient der Heiratsdummy gibt uns dann jeweils Auskunft über den „Kausaleffekt" einer Heirat (diese Spezifikation impliziert einen dauerhaften Effekt der Heirat).[4]

Schätzen wir zuerst eine Querschnittsregression nur mit den Daten der vierten Welle ($T = 4$). Das Ergebnis ist ein Regressionskoeffizient von 2500 Euro (das Lohnmittel der nicht Verheirateten ist 1500 Euro, das der Verheirateten 4000 Euro) (Tabelle 1, Spalte 1). Offensichtlich wäre es grob irreführend, dieses Ergebnis als Kausaleffekt einer Heirat zu interpretieren. Unbeobachtete Heterogenität führt zu einer deutlichen Verzerrung des Querschnitts-Schätzers.

Im zweiten Modell verwenden wir die Daten aller Wellen und berechnen den POLS-Schätzer. Das Ergebnis ist 1833 Euro (Tabelle 1, Spalte 2). Es ergibt sich aus dem mittleren Lohn in den verheirateten Personen-Jahren minus dem mittleren Lohn in

[4] Für das Folgende ist es wichtig zu wissen, wie sich der Regressionskoeffizient einer Dummy in einer bivariaten Regression ergibt: Der Koeffizient ist einfach der Mittelwertsunterschied der beiden Dummy-Gruppen auf der abhängigen Variable.

Tab. 1: Vergleich dreier Regressionen mit fiktiven Daten

	(1) OLS Querschnitt	(2) POLS	(3) FE
Konstante	1500	2167	—
Heirat	2500	1833	500
	(707)	(656)	(39)
Zahl Personen	4	4	2
Zahl Personenjahre	—	24	12

Standardfehler in Klammern. In (2) ist der Standardfehler panel-robust geschätzt. In (3) ist das aufgrund von Eigenheiten der Daten nicht möglich.

den nicht verheirateten Personen-Jahren. Die Verzerrung ist zwar geringer, aber immer noch deutlich. Der Grund ist, dass POLS zwar die Within-Variation (die Löhne der „Besserverdiener“ vor und nach der Heirat) nutzt, aber dennoch die Between-Variation dominiert. Die Lehre daraus ist, dass Paneldaten in Verbindung mit Between-Schätzern *nicht* das Problem personenspezifischer unbeobachteter Heterogenität beheben.

Dies gelingt erst mit einem Within-Schätzer: Der FE-Schätzer beträgt 500 Euro (Tabelle 1, Spalte 3), kann also den in die Daten eingebauten Kausaleffekt reproduzieren. Der FE-Schätzer – indem er rein die Within-Variation nutzt – ist nicht durch unbeobachtete Heterogenität verzerrt. Das heißt, *Paneldaten in Verbindung mit dem FE-Schätzer haben das Potential, das Problem personenspezifischer unbeobachteter Heterogenität zu beheben.*

Um die „Mechanik“ des FE-Schätzers zu verstehen, sind in Abbildung 2 die Daten nach der Within-Transformation in ein Streudiagramm abgetragen (mit Zufallsstreuung, um Überdeckungen zu vermeiden). Wie oben erläutert, werden bei der Within-Transformation bei jeder Variable jeweils die personenspezifischen Mittelwerte abgezogen (Mittelwertbereinigung). Die mittelwertsbereinigte Heiratsdummy ist auf der X-Achse abgetragen. Die immer Ledigen haben hier in jeder Welle ein Null stehen. Die Heiratenden haben vor der Heirat eine $-0{,}5$, danach eine $+0{,}5$. Auf der Y-Achse ist der mittelwertsbereinigte Lohn abgetragen. Die Datenpunkte der immer Ledigen streuen leicht um den Punkt (0,0). Sie tragen nichts zur Steigung der Regression bei, denn eine Regression geht immer durch den Punkt $(\bar{x},\bar{y})$ was hier eben (0,0) ist. Die Steigung der Regressionsgerade wird allein vom mittleren Lohn vor und nach der Heirat bestimmt. Der FE-Schätzer ist somit ganz einfach die Differenz des mittleren Lohnes nach der Heirat und vor der Heirat. Er nutzt nur die Within-Variation, indem er nur die Daten der Heiratenden berücksichtigt und einen vorher-nachher Vergleich anstellt.

Dies macht deutlich, dass der (einfache) FE-Schätzer die Daten der immer Ledigen nicht nutzt (er benutzt die Kontrollgruppe nicht). Dies ist so in Ordnung, weil die Kontrollgruppe keine Within-Information liefert. Gibt es allerdings Alters- oder Periodeneffekte, dann wird die Nicht-Berücksichtigung der Kontrollgruppe zum Problem. Man kann dieses Problem aber leicht beheben, indem man das Modell um Alters- und

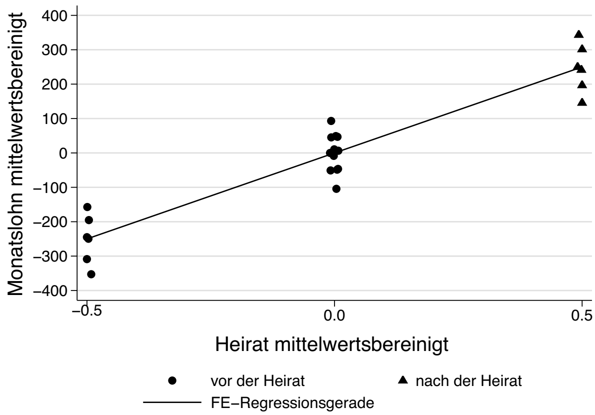

Abb. 2: Die 'Mechanik' einer FE–Regression

Periodenvariablen erweitert (ein Beispiel hierzu findet sich unten). Die Alters- und Periodeneffekte werden unter Einbezug der Daten der Kontrollgruppe geschätzt. *Eine FE-Regression mit Perioden- und Alters-Variablen implementiert im Prinzip einen vorher-nachher Vergleich mit Kontrollgruppe.* Sie ist daher die beste Methode zur Analyse von Paneldaten.

An dieser Stelle wird ein weiterer Punkt deutlich: nicht alle in der Stichprobe enthaltenen Personen tragen zur FE-Schätzung des Effektes des interessierenden Ereignisses (hier der Heirat) bei. Nur die Personen, die während der Bebachtungszeit das Ereignis erlebt haben, tragen zur Schätzung bei. Damit ist die Generalisierbarkeit des FE-Schätzers eingeschränkt. Wenn die Personen mit Ereignis (Treatment) keine Zufallsauswahl aus allen Personen sind, so repräsentiert der FE-Schätzer nur den Effekt in der Gruppe mit Treatment. In der Literatur zur kontrafaktischen Kausalität bezeichnet man deshalb einen solchen Schätzer als ATET (Average Treatment Effect on the Treated, siehe Kapitel 35 in diesem Handbuch). Wie wir unten an einem Beispiel sehen werden, trägt oft nur ein kleiner Teil der Panel-Stichprobe zum FE-Schätzer bei. Das ist in Ordnung so, weil nur der Teil der Daten verwendet wird, der tatsächlich Within-Information liefert. Aber man muss sich bewusst sein, dass dies ein nur mit Vorsicht generalisierbarer ATET-Schätzer ist.

2 Eine präzisere Einführung in die Panelregression

Nach der eher intuitiven Einführung in die Panelregression im letzten Kapitel, sollen im Folgenden die Modelle präziser formuliert werden. Dabei werden auch einige weitere Panelregressionsmodelle eingeführt. Die Darstellung orientiert sich stark an Cameron & Trivedi (2005), wo man auch die hier nicht angeführten Herleitungen nachschlagen kann.

2.1 Das Fehlerkomponenten-Modell

Ausgangspunkt ist das Fehlerkomponenten-Modell (auch Varianzkomponenten-Modell):

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \,. \tag{9}$$

Der idiosynkratische Fehler ε_{it} ist unabhängig, identisch verteilt über i und t. Alle im Folgenden vorgestellten Panelregressionsmodelle machen die *Annahme der strikten Exogenität:* $E[\varepsilon_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}] = 0$ Der idiosynkratische Fehler darf nicht mit den X-Variablen korrelieren (jeweils innerhalb einer Person, über alle Zeitpunkte). D. h. die X-Werte dürfen nicht eine Funktion vergangener, gegenwärtiger oder zukünftiger unbeobachteter Faktoren sein. Trifft diese Annahme nicht zu, so sind die im Folgenden vorgestellten Panelschätzer verzerrt.

α_i ist ein personenspezifischer, zeitkonstanter Fehlerterm (also eine Zufallsvariable). Abhängig von den weiteren Annahmen über den personenspezifischen Fehler kann man zwei grundlegende Modelle unterscheiden. Im *Fixed-Effects-Modell* (FE-Modell) nimmt man an, dass α_i potentiell mit den beobachteten X-Variablen korreliert ist.[5] Unter dieser Annahme ist es nötig, die α_i zu eliminieren. Ansonsten sind die Schätzer von $\boldsymbol{\beta}$ verzerrt.

Das *Random-Effects-Modell* (RE-Modell) hingegen nimmt an, dass die personenspezifischen Fehler α_i unabhängig von den Regressoren sind. Weiterhin wird üblicherweise angenommen, dass sowohl die personenspezifischen, wie auch die idiosynkratischen Fehler unabhängig, identisch verteilt sind und konstante Varianz haben:

$$\begin{aligned} \alpha_i &\sim (\alpha, \sigma^2_\alpha) \\ \varepsilon_{it} &\sim (0, \sigma^2_\varepsilon) \,. \end{aligned} \tag{10}$$

Eine präzisere Bezeichnung wäre allerdings „Random-Intercept"-Modell, weil hier – im Unterschied zum klassischen Regressionsmodell – der Achsenabschnitt eine Zufallsvariable ist.[6] Diese Zufallsvariable muss allerdings unabhängig von den Regressoren sein. Ist dies nicht der Fall, so sind RE-Schätzer verzerrt.

2.2 Die Schätzverfahren

Zur Schätzung von $\boldsymbol{\beta}$ mit Paneldaten gibt es verschiedene Möglichkeiten. Das einfachste Verfahren ist die Anwendung von OLS auf die „gepoolten" Daten (*POLS*). POLS ist konsistent unter der RE-Annahme, aber nicht effizient, denn es liegt Autokorrelation vor (die Fehlerterme sind über die Zeit korreliert, wegen dem personenspezifischen Fehlerterm). Unter der realistischeren FE-Annahme liefert POLS inkonsistente Schätzer.

[5] Die Bezeichnung „Fixed-Effects" geht auf eine Tradition zurück, in der man die α_i als fixe Parameter betrachtete, die geschätzt werden sollten.

[6] Unter diesem Namen ist das RE-Modell in der Mehrebenenanalyse-Literatur bekannt (siehe Kapitel 28 in diesem Handbuch). Paneldaten können als Mehrebenendaten konzeptionalisiert werden: die obere Ebene sind Personen, darunter liegt die Ebene „Zeit". Deshalb gibt es Parallelitäten zwischen Panelmodellen und Mehrebenenmodellen (siehe Rabe-Hesketh & Skrondal 2008).

POLS verwendet sowohl die Variation über die Zeit (Within), als auch die Variation zwischen den Personen (Between). Der *Between-Schätzer* benutzt nur die Variation zwischen den Personen. Man erhält ihn, indem man OLS auf die Daten nach der Between-Transformation anwendet:

$$\bar{y}_i = \bar{\mathbf{x}}_i' \boldsymbol{\beta} + \alpha_i + \bar{\varepsilon}_i \,. \tag{11}$$

Der Between-Schätzer ist unter der RE-Annahme konsistent, nicht aber unter der FE-Annahme. Der Between-Schätzer ist in der Forschungspraxis nicht relevant, aber konzeptionell ist er interessant, da er nur auf der Between-Variation beruht.

Das andere Extrem – ein Within-Schätzer – ist der *FE-Schätzer*. Er nutzt nur die Variation der Daten über die Zeit. Man erhält den FE-Schätzer indem man OLS auf die Daten nach der Within-Transformation anwendet. Dazu zieht man (11) von (9) ab und erhält das Within-Modell:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) \,. \tag{12}$$

Entscheidend ist, dass die unbeobachteten, personenspezifischen Fehler α_i nun eliminiert sind. Deshalb ist der FE-Schätzer sowohl unter der RE-Annahme, wie auch unter der FE-Annahme konsistent. Er ist durch personenspezifische unbeobachtete Heterogenität nicht verzerrt. Damit nützt der FE-Schätzer eine spezielle Eigenschaft von Paneldaten aus, nämlich Kausaleffekte durch den intraindividuellen Vergleich zu schätzen.

Ein „Nachteil" des FE-Schätzers ist allerdings, dass die Koeffizienten zeitkonstanter Variablen nicht geschätzt werden können. Durch die Within-Transformation fallen nicht nur die personenspezifischen unbeobachteten Größen aus dem Modell, sondern auch die personenspezifischen beobachteten Variablen. Dass dem so sein muss, ist leicht verständlich, denn mit konstanten Größen ist kein Within-Vergleich möglich. Wenn sich nichts ändert, kann kein vorher-nachher Vergleich angestellt werden.

Zwei weitere – zum FE-Verfahren äquivalente – Schätzverfahren, verdeutlichen die „Logik" der Within-Schätzung aus anderen Perspektiven. Beim ersten Verfahren fügt man in das Modell für jede Person eine Dummy ein und schätzt die Parameter mit OLS (*LSDV, Least Squares Dummy Variable*). Man schätzt also die α_i in Modell (9) mit, und kontrolliert somit für alle personenspezifischen Merkmale (beobachtete und unbeobachtete). Unbeobachtete Personen-Heterogenität ist deshalb kein Problem mehr. Mit dem LSDV-Schätzer ist es möglich, den FE-Schätzer auch ohne ein spezielles Panel-Programm zu schätzen (allerdings nur bei kleinem N, denn bei großem N überfordert man die meisten Programme durch die vielen Dummies). Weiterhin erhält man Schätzer für die personenspezifischen Effekte, was bei Länder- bzw. Firmenpaneldaten manchmal von inhaltlichem Interesse sein kann.

Ein weiteres äquivalentes Verfahren ist eine Regression mit personenspezifischen Koeffizienten. Dabei berechnet man im Prinzip mit den Beobachtungen einer jeden Person eine eigene Regression. Das gewichtete Mittel der Steigungskoeffizienten dieser Regressionen ist dann der FE-Schätzer. Für Personen, die während ihrer Beobachtungsdauer keine Varianz auf einer unabhängigen Variablen aufweisen, kann keine Regressionsgerade geschätzt werden, weshalb sie nicht in die Berechnung eingehen.

Dies zeigt noch einmal, dass nur die „Treated" zum FE-Schätzer beitragen. Weiterhin macht dieser Schätzer noch einmal klar, dass es sich um eine reine Within-Betrachtung handelt: Es wird gefragt, wie sich Y ändert, wenn sich X ändert und zwar bei einer Person über die Zeit. Das Mittel der Antworten auf diese Frage ist dann der FE-Schätzer.

Ein weiterer Within-Schätzer ist der *FD (First Differences) Schätzer*. Wenn man (9) um eine Periode zeitverzögert, erhält man $y_{i,t-1} = \mathbf{x}'_{i,t-1}\boldsymbol{\beta} + \alpha_i + \varepsilon_{i,t-1}$. Zieht man dies von (9) ab, so erhält man das FD-Modell:

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1})\,. \tag{13}$$

Auch hier sind die α_i eliminiert und damit ist der OLS Schätzer von (13) sowohl unter der RE-, wie auch unter der FE-Annahme konsistent. Wie beim FE-Schätzer ist hier die Frage, wie sich Y ändert, wenn sich X ändert. Anders als beim FE-Schätzer, werden jedoch nur die beiden Y-Werte unmittelbar vor und nach der X-Veränderung betrachtet (der FE-Schätzer bezieht alle verfügbaren Y-Werte in den Davor-Danach-Vergleich ein). Hat man mehr als zwei Wellen zur Verfügung, so ist dies ineffizient. Bei zwei Wellen sind FD- und FE-Schätzer äquivalent. Bei mehr als zwei Wellen unterscheiden sich FE- und FD-Schätzer und man sollte den effizienteren FE-Schätzer verwenden.

Schließlich kann man das RE-Modell (9) und (10) mit GLS (Generalized Least Squares) schätzen. Im RE-Modell sind die Fehler autokorreliert, weshalb OLS ineffizient ist. GLS behebt dieses Problem, indem die Daten entsprechend transformiert werden (Feasible GLS). Der GLS-Schätzer des RE-Modells wird *RE-Schätzer* genannt. Der RE-Schätzer ist konsistent unter der RE-Annahme. Unter der FE-Annahme ist er jedoch verzerrt. Die RE-Transformation ist folgende:

$$\begin{aligned} y_{it} - \lambda\bar{y}_i = (\mathbf{x}_{it} - \lambda\bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (1-\lambda)\alpha_i + (\varepsilon_{it} - \lambda\bar{\varepsilon}_i)\,,& \\ \text{wobei } \lambda = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}\,.& \end{aligned} \tag{14}$$

OLS angewandt auf die so transformierten Daten, liefert den RE-Schätzer. Man beachte, dass für $\lambda = 0$ der RE-Schätzer äquivalent zu POLS ist und für $\lambda = 1$ ist er äquivalent zu FE. Normalerweise wird λ Werte zwischen 0 und 1 annehmen.[7] In diesem Fall ist der RE-Schätzer eine Mischung aus dem Between- und dem FE-Schätzer. Sind die personenspezifischen Fehler und X korreliert, so ist der Between-Schätzer verzerrt. Damit wird auch der RE-Schätzer verzerrt sein. Das Ausmaß der Verzerrung hängt vom Wert des Transformationsgewichtes λ ab. Ist λ in der Nähe von 1, so wird die Verzerrung gering sein. Für lange Panels (T groß) wird λ gegen 1 gehen. Dasselbe gilt, falls die personenspezifische Fehlervarianz deutlich größer ist, als die idiosynkratische Fehlervarianz.

[7] Man beachte, dass das Transformationsgewicht λ von T abhängt. In nicht-balancierten Panels gibt es deshalb verschiedene Transformationsgewichte.

2.3 RE- oder FE-Modell?

Damit stellt sich die Frage, ob man das RE- oder das FE-Modell verwenden soll. Angesichts der Tatsache, dass bei den meisten sozialwissenschaftlichen Fragestellungen personenspezifische unbeobachtete Heterogenität vorhanden sein dürfte, ist die Antwort eigentlich klar: Man sollte das FE-Modell verwenden, weil es den besonderen Vorzug von Paneldaten – die Möglichkeit des Within-Vergleichs – voll umsetzt, und deshalb bei Vorliegen von personenspezifischer unbeobachteter Heterogenität nicht verzerrt ist. Dagegen wird in dieser Situation der RE-Schätzer verzerrt sein, weil er auch den Between-Vergleich mit einbezieht (zumindest solange λ deutlich kleiner 1 ist).

In der Sozialforschung sind diese Vorzüge des FE-Schätzers noch viel zu wenig bekannt, was dazu führt, dass viele Autoren POLS oder RE verwenden. Halaby (2004) wertete Panelstudien aus, die in ASR und AJS von 1990 bis 2003 erschienen sind. Sein Ergebnis ist, dass die Hälfte dieser Artikel keine FE-Schätzer verwendet. Den Autoren und Gutachtern der besten Soziologie-Journale scheint der große Vorzug von Paneldaten – nämlich die Möglichkeit der Within-Schätzung – unbekannt zu sein.

Zwei „Probleme“ des FE-Modells sind es, die manchen Forscher zum RE-Modell greifen lassen. Erstens kann man mit dem FE-Modell die Koeffizienten zeitkonstanter Variablen nicht mitschätzen. Da man es von Querschnittsregressionen gewohnt ist, dass Regressionen möglichst viele Kontrollvariablen enthalten, wollen Forscher auch bei Panelregressionen viele Variablen ins Modell nehmen. Da Kontrollvariablen oft zeitkonstant sind, macht dies das RE-Modell attraktiv. Dies ist aber ein trügerischer Schluss: Bei Querschnittsregressionen ist es sinnvoll, für viele Variablen zu kontrollieren, weil dies das Potential für verzerrte Schätzer reduziert. Bei Panelregressionen ist das Gegenteil der Fall, denn durch die Verwendung des RE-Modells erhöht man das Potential für verzerrte Schätzer. Ein FE-Modell enthält zwar weniger Variablen, es kontrolliert aber implizit für *alle* zeitkonstanten Variablen. Um den Informationsgehalt von Paneldaten voll auszunutzen, ist ein Umdenken erforderlich: Nicht große Regressionsmodelle sind „beautiful“, sondern FE-Modelle mit evtl. wenigen zeitveränderlichen Variablen.

Man muss es eigentlich umgekehrt formulieren: Nicht das FE-Modell hat hier ein Problem, sondern das Beharren auf den Effekten zeitkonstanter Variablen ist das Problem. Auch mit Paneldaten können die Effekte zeitkonstanter Variablen nur verzerrt geschätzt werden, falls unbeobachtete Heterogenität vorliegt. Man kann sogar in Frage stellen, dass dies Kausalanalyse ist, denn wo keine Veränderung, da keine Ursache und Wirkung. Effekte zeitkonstanter Variablen sind keine Kausaleffekte, sondern Korrelate. Mit Querschnittsdaten muss man sich mit der Frage nach den Korrelaten begnügen. Aber mit Paneldaten kann man neue Fragen stellen: Man kann versuchen die hinter den Korrelationen stehenden kausalen Mechanismen aufzuklären. Dabei können Paneldaten sehr hilfreich sein, weil sie die Dynamik der Prozesse in verschiedenen Gruppen beleuchten helfen. Hierzu kann man z. B. Wachstumskurven-Modelle einsetzen (s. u.). Mit der bloßen Frage nach den Effekten zeitkonstanter Variablen schöpft man das Potential von Paneldaten bei weitem nicht aus.

Ein zweiter angeblicher „Vorteil“ wird oft zugunsten des RE-Modells angeführt. Im Vergleich zu einem RE-Schätzer ist der FE-Schätzer weniger effizient (die Standardfehler der FE-Schätzer sind größer). Dies ist leicht zu verstehen: Während der RE-Schätzer die gesamte Variation in den Daten nutzt, stützt sich der FE-Schätzer nur auf die Within-Variation. In Situationen, in denen die Within-Variation im Vergleich zur Between-Variation klein ist, benutzt der FE-Schätzer nur einen kleinen Teil der Variation in den Daten. Da sei der RE-Schätzer doch vorzuziehen, weil er die gesamte Variation nutze. Dieses Argument führt aber in die Irre, denn nicht Effizienz (kleine Standardfehler) sondern Vermeidung von Bias ist das wichtigere Ziel in der Sozialforschung. Wem helfen präzise Schätzer, die aber massiv verzerrt sind? Indem der FE-Schätzer die „kontaminierte" Between-Variation ignoriert, opfert er Effizienz dem Oberziel „Vermeidung von Bias" (Allison 2009, S. 3).

In manchen Situationen mag personenspezifische Heterogenität kein Problem sein, oder das Transformationsgewicht ist nahe 1. Dann – und nur dann – ist die Verwendung des RE-Modells gerechtfertigt und eben auch effizienter. Mit einem so genannten *Hausman-Test* kann man feststellen, ob eine solche Situation vorliegt. Ein Hausman-Test testet ein „immer" konsistentes Modell gegen ein „manchmal" konsistentes aber dann effizientes Modell. Im Panel-Fall ist das immer konsistente Modell das FE-Modell und das manchmal konsistente Modell ist das RE-Modell. Die Grundidee für die Hausman-Teststatistik ist einfach: Man berechnet die standardisierte Differenz der Parameterschätzer. Ist die groß, so weicht das RE-Modell stark vom FE-Modell ab und man muss das FE-Modell verwenden. Seien $\hat{\boldsymbol{\beta}}_{FE}$ die FE-Schätzer und $\hat{\boldsymbol{\beta}}_{RE}$ die RE-Schätzer. Sei $\hat{V}(.)$ die geschätzte Varianz-Kovarianzmatrix der Schätzer. So ergibt sich die Hausman-Teststatistik H als:

$$H = (\hat{\boldsymbol{\beta}}_{RE} - \hat{\boldsymbol{\beta}}_{FE})'[\hat{V}(\hat{\boldsymbol{\beta}}_{RE}) - \hat{V}(\hat{\boldsymbol{\beta}}_{FE})]^{-1}(\hat{\boldsymbol{\beta}}_{RE} - \hat{\boldsymbol{\beta}}_{FE})\,. \qquad (15)$$

Die Nullhypothese ist, dass die Schätzer beider Modelle identisch sind ($H = 0$). H ist unter der Nullhypothese asymptotisch χ^2-verteilt mit K (Zahl der Variablen) Freiheitsgraden. Kann die Nullhypothese nicht abgelehnt werden, so sind die RE-Schätzer nicht durch unbeobachtete Heterogenität verzerrt und den FE-Schätzern wegen ihrer größeren Effizienz sogar vorzuziehen. Muss die Nullhypothese abgelehnt werden, so sind die RE-Schätzer verzerrt und man sollte das FE-Modell verwenden.

2.4 Ein Hybrid-Modell

Wenn aber der Effekt einer zeitkonstanten Variable von zentralem inhaltlichem Interesse ist, dann nützen all die Vorteile des FE-Modells nichts: es liefert diesbezüglich keinen Schätzer. Anstatt nun zum RE-Modell zu greifen, gibt es als Alternative das so genannte Hybrid-Modell. Solche Modelle kombinieren RE- und FE-Modelle und erlauben so, die Effekte zeitkonstanter Variablen mitzuschätzen. Allison (2009) schlägt ein solches Hybrid-Modell vor.[8] Man schätzt ein RE-Modell, wobei allerdings die zeitveränderlichen Variablen in zweifacher Form aufgenommen werden: mittelwertsbereinigt (Within-Transformation) und als Personenmittel (Between-Transformation). Das Hybrid-Modell lautet:

[8] Zwei weitere Hybrid-Modelle (auch „Mixed Models" genannt) erwähnt Halaby (2004, S. 530 ff.).

$$y_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + \mathbf{z}_i'\boldsymbol{\delta} + \alpha_i + \varepsilon_{it}\,, \tag{16}$$

wobei $\mathbf{z}_i$ der Vektor der zeitkonstanten Variablen ist. Die Parameter werden mit GLS geschätzt (RE-Modell). Die Schätzer der Within-Komponente ($\boldsymbol{\beta}$) sind identisch mit den FE-Schätzern. Die Schätzer der Between-Komponente ($\boldsymbol{\gamma}$) sind Between-Schätzer und bei Vorliegen unbeobachteter Heterogenität verzerrt. Deshalb sind sie normalerweise nicht von inhaltlichem Interesse. Liegt keine personenspezifische unbeobachtete Heterogenität vor, so sollte $\boldsymbol{\beta} = \boldsymbol{\gamma}$ sein. Ein Test diesbezüglich kann somit als Alternative zum Hausman-Test eingesetzt werden.

Das attraktive am Hybrid-Modell ist nun, dass es neben den Within-Schätzern auch Schätzer der Effekte der zeitkonstanten Variablen ($\boldsymbol{\delta}$) liefert. Im Gegensatz zum RE-Modell sind diese Schätzer nicht durch die ungenügende Kontrolle der zeitveränderlichen Variablen verzerrt. Das RE-Modell unterstellt ja implizit die – meist falsche – Restriktion $\boldsymbol{\beta} = \boldsymbol{\gamma}$. Insofern sollte man die Hybrid-Schätzer den RE-Schätzern immer vorziehen. Dennoch sind die Effektschätzer der zeitkonstanten Variablen auch im Hybrid-Modell bei Vorliegen von personenspezifischer unbeobachteter Heterogenität verzerrt. Man muss sich – wie schon mehrfach erwähnt – bewusst sein, dass Paneldaten bezüglich der Identifikation der Effekte zeitkonstanter Variablen kaum Vorteile bieten.

2.5 Panel-robuste Standardfehler

Schließlich sind noch ein paar Bemerkungen zu den Standardfehlern notwendig. Die konventionellen Formeln zur Berechnung von Standardfehlern in Regressionsmodellen setzen die Abwesenheit von Heteroskedastizität und Autokorrelation voraus. Da Paneldaten eine Cluster-Struktur aufweisen, ist jedoch ziemlich sicher mit Autokorrelation zu rechnen. Die Beobachtungen einer Person sind nicht unabhängig, weshalb die Fehlerterme innerhalb einer Person korreliert sein werden (Autokorrelation). Die Berücksichtigung personenspezifischer Terme bei RE- und FE-Modellen reduziert zwar das Problem, beseitigt es aber normalerweise nicht ganz. Weiterhin muss man davon ausgehen, dass in den meisten Daten Heteroskedastizität vorliegt. Aus dem letzteren Grund hat es sich inzwischen bei Querschnittsregressionen eingebürgert, die Huber/White-Sandwich-Schätzer der Standardfehler zu verwenden. Die sind robust gegen die Verletzung der Heteroskedastizitätsannahme (robuste Standardfehler). Das Sandwich-Prinzip lässt sich auf den Panelkontext verallgemeinern und man erhält damit panel-robuste Standardfehler, die auch gegen das Vorliegen von Autokorrelation unempfindlich sind. In Stata erhält man panel-robuste Standardfehler mit der Option `vce (cluster id)`, wobei `id` der Personen-Identifier ist.

Man hüte sich aber davor, das Problem verzerrter Standardfehler überzustilisieren.[9] Viele Forschungsartikel beinhalten eine längere methodologische Diskussion über verzerrte Standardfehler bei Paneldaten. Das mündet dann in die Verwendung von RE-Modellen mit panel-robusten Standardfehlern. Verzerrte Standardfehler sind aber nicht

[9] Im Übrigen weisen Angrist & Pischke (2009, Kap. 8) darauf hin, dass auch die robusten Standardfehler in finiten Stichproben verzerrt sein werden. Die geschätzten Standardfehler (egal ob konventionell oder robust) werden nie ganz richtig sein. Ihr Ratschlag: „Don't panic".

das Hauptproblem der Sozialforschung, sondern durch unbeobachtete Heterogenität verzerrte Effektschätzer. Man arbeitet sich an einem zweitrangigen Problem ab und vergisst dabei das eigentliche Problem, welches man mit der Verwendung von FE-Modellen in den Griff bekommen könnte.

3 Ein Beispiel: Der Effekt der Heirat auf die Zufriedenheit

Anhand eines Beispiels mit SOEP-Daten soll in diesem Kapitel der Umgang mit den verschiedenen Panelregressionsmodellen illustriert werden. Ich wähle ein Beispiel aus dem in den letzten Jahren boomenden Bereich der „Happiness-Forschung". Es soll untersucht werden, ob eine (Erst-) Heirat die Lebenszufriedenheit erhöht (der den Boom auslösende Aufsatz ist Lucas et al. 2003). Die meisten Menschen würden wohl antworten „natürlich ja", und einschlägige Auswertungen mit Querschnittsdaten belegen dies auch. Aber der methodisch geschulte Sozialforscher wird bemerken, dass dies auch die Folge von Selbstselektion seine könnte: Vermutlich heiraten glückliche Menschen eher. Da das SOEP jedes Jahr die Zufriedenheitsfrage gestellt hat, ist es für diese Fragestellung ideal geeignet. Für die folgenden Analysen verwende ich die Daten des SOEP 1984–2007 (100 % Version). Damit hat man teilweise sehr lange Zufriedenheitspanels vorliegen und hat genügend Varianz auf der unabhängigen Variable, so dass in der typischen Within-Herangehensweise die Zufriedenheit vor und nach der Heirat innerhalb einer Person verglichen werden kann.

3.1 Datenaufbereitung

Die Frage nach der Lebenszufriedenheit lautet: „Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?". Die Antworten erfolgen auf einer 11-stufigen Skala von 0 („ganz und gar unzufrieden") bis 10 („ganz und gar zufrieden"). Dies ist eine ordinale Variable, aber die meisten Autoren behandeln sie wie eine intervallskalierte Variable. Deshalb verwende auch ich im Folgenden metrische Regressionsverfahren. Eine weitere wichtige Entscheidung ist die Modellierung des Zeitpfades des Heiratseffektes. Nimmt man an, dass eine Heirat einen sofortigen und dauerhaft konstanten Effekt auf die Zufriedenheit hat, dann genügt es, in das Modell eine Heiratsdummy aufzunehmen. Die Heiratsdummy ist null vor der Erstheirat, in Wellen nach dem Datum der Erstheirat ist sie eins. In unserem Fall liegt aber die Vermutung nahe, dass der Heiratseffekt mit der Zeit verschwinden könnte. Deshalb nehme ich zwei Zeitvariablen zusätzlich in das Modell auf: Jahre seit der Heirat und Jahre seit der Heirat quadriert. Die Heiratsdummy und die beiden Zeitvariablen zusammen können den Zeitpfad des Heiratseffektes sehr flexibel modellieren.

Weiterhin berücksichtige ich drei Kontrollvariablen. Um eventuelle Alterseffekte zu erfassen, wird das Alter in Jahren in die Modelle aufgenommen (aus didaktischen Gründen vorerst nur linear). Da eine Abhängigkeit der Zufriedenheit vom Einkommen zu vermuten ist, ist auch das Haushaltseinkommen in den Modellen (natürlicher Logarithmus). Diese beiden Kontrollvariablen sind zeitveränderlich. Die dritte Kontrollvariable ist zeitkonstant: das Geschlecht.

Die hierzu erforderlichen Datenaufbereitungsschritte sind nicht trivial. Paneldaten sind komplexe Daten, weshalb die Datenaufbereitung den Hauptteil der Datenanalyse in Anspruch nimmt. Aus Platzgründen kann ich hier nicht weiter auf die Datenaufbereitung eingehen, aber der kommentierte Datenaufbereitungsfile (und auch der Datenanalysefile) kann auf der Internetseite des Handbuchs eingesehen werden.

Einige Besonderheiten sind bei der Eingrenzung des Schätz-Samples zu beachten, weshalb ich darauf hier eingehen will (insbesondere weil man diese wichtigen Schritte in keinem Lehrbuch findet). Die Personendatensätze des SOEP beinhalten 376.581 Personenjahre von 47.466 Personen. Es ist nicht sinnvoll, mit all diesen Fällen die Analyse durchzuführen. Die erste Eingrenzung kennt man so auch aus der Querschnittsdatenanalyse: Personenjahre mit fehlenden Werten auf einer der Analysevariablen werden ausgeschlossen (Listwise Deletion). Damit verliert man 5.526 Personenjahre und 379 Personen. Die weiteren Eingrenzungen sind jedoch panelspezifisch: Die Idee ist, das Sample auf die Personenjahre einzugrenzen, die für die Within-Analyse relevant sind. Das sind die Personen (bzw. Personenjahre), die das Ereignis (in unserem Falle eine Erstheirat) potentiell während der Beobachtungsdauer des Panels erleben könnten. Das sind nicht Personenjahre nach der Auflösung einer Ehe (Verwitwung, Trennung, Scheidung). Damit verlieren wir 64.301 Personenjahre und 5.568 Personen. Ebenso tragen nur Personen, die bei der ersten Beobachtung im Panel ledig waren, zum Within-Schätzer bei. Deshalb schließen wir 25.375 am Anfang bereits verheiratete Personen (das sind 61 % der Personen) mit 195.478 Personenjahren aus. Schließlich schließe ich alle Personen aus, die nur mit einem Personenjahr im Datensatz vertreten sind (N=2.605). Damit hat der Analysedatensatz zum Schluss 108.369 Personenjahre von 13.539 Personen.

Diese massiven Dateneingrenzungsschritte sind sicher gewöhnungsbedürftig. Aber wenn man in der Within-Logik denkt, machen sie Sinn. Wie oben erläutert, ist ein FE-Schätzer immer ein ATET-Schätzer, weshalb es keinen Unterschied macht, ob man die „Non-Treated" im Analysedatensatz lässt oder nicht. Allerdings schließe ich nicht alle „Non-Treated" aus: die immer Ledigen bleiben im Analysedatensatz, um zuverlässigere Schätzer für die Alters- und Periodeneffekte (und die weiteren Kontrollvariablen) zu erhalten. Man könnte nun argumentieren, dass die Schätzer der Kontrollvariablen noch präziser würden, wenn man auch z. B. die am Anfang des Panels bereits verheirateten Personen im Analysedatensatz belassen würde. Allerdings bringt man dadurch eher unnötige Heterogenität in den Analysedatensatz, welche die Schätzer der Kontrollvariablen „verzerren" könnte. Angenommen der Alterseffekt ist bei Ledigen ganz anders, als bei Verheirateten. Die Verheirateten werden aufgrund ihrer größeren Zahl die Schätzung des Alterseffektes dominieren, was dazu führen kann, dass der Heiratseffekt verzerrt geschätzt wird. Deshalb macht es Sinn, den Analysedatensatz auf die Personen zu beschränken, die in einem wichtigen Punkt vergleichbar sind: sie können das Ereignis (Treatment) potentiell erleben. In unserem Beispiel sind dies eben die am Anfang des Panels noch Ledigen.

Tab. 2: Regressionsmodelle auf die Lebenszufriedenheit (SOEP 1984–2007)

	(1) POLS	(2) RE	(3) FE	(4) FE Perioden
Heirat	0,403***	0,243***	0,249***	0,249***
Jahre seit Heirat	−0,055**	−0,058***	−0,043***	−0,047***
(Jahre seit Heirat)^2	0,002**	0,002***	0,002***	0,002***
Alter in Jahren	−0,013***	−0,022***	−0,037***	−0,022**
Haushaltseinkommen (ln)	0,283***	0,138***	0,091***	0,091***
Frau	0,066**	0,057**	—	—
(Within-)R^2	0,026	0,014	0,015	0,019
Zahl Personen	13539	13539	13539	13539
Zahl Personenjahre	108369	108369	108369	108369
Zahl Heiraten	3491	3491	3491	3491

Standardfehler panel-robust geschätzt. Modell (4) enthält zusätzlich Periodeneffekte.
*: $p \leq 0{,}05$; **: $p \leq 0{,}01$; ***: $p \leq 0{,}001$

3.2 Schätzergebnisse

Mit unserem Analysedatensatz vergleichen wir nun POLS, RE- und FE-Modell. Die Schätzergebnisse finden sich in Tabelle 2, Modelle 1–3. Von den 13.539 Personen haben 3.491 Personen während der Beobachtungsdauer eine Heirat. Dies sind genügend Ereignisse, um eine verlässliche Within-Schätzung durchführen zu können.

Zuerst gehe ich auf die Interpretation der Kontrollvariablen ein. Der Alterseffekt ist negativ, d. h. mit zunehmendem Alter werden die Deutschen unzufriedener. Der Effekt wird im RE- und im FE-Modell stärker. Allerdings sind hier keine Periodeneffekte kontrolliert, so dass der Alterseffekt verzerrt ist (s. Modell 4). Der positive Einkommenseffekt zeigt, dass wohlhabende Menschen zufriedener sind. Der Effekt ist im FE-Modell allerdings deutlich kleiner, was darauf hindeutet, dass hier zum Teil unbeobachtete Heterogenität vorliegt: Es liegen unbeobachteten Eigenschaften vor, die sowohl die Zufriedenheit als auch das Einkommen erhöhen. Schließlich zeigt sich, dass Frauen etwas zufriedener sind. Da es sich hierbei um eine zeitkonstante Variable handelt, ist der Effekt im FE-Modell nicht schätzbar.

Der eigentlich interessierende Effekt der Heirat ist aufgrund der komplexen Modellierung nur schwer zu interpretieren. Ein Conditional-Effect Plot hilft hier weiter (siehe Abbildung 3). Alle drei Modelle kommen zu dem Ergebnis, dass eine Heirat die Zufriedenheit erhöht, allerdings nur in den ersten Jahren der Ehe. POLS überschätzt den Heiratseffekt deutlich. Der Grund ist vermutlich Selbstselektion. Der FE-Schätzer ist kleiner, aber immer noch signifikant. Es gibt also neben der Selbstselektion auch einen kausalen Effekt der Heirat. Im Honeymoon-Jahr ist die Zufriedenheit um 0,25 Skalenpunkte höher. Dieser Honeymoon-Effekt baut sich über die Jahre rasch ab und ist nach sechs Jahren nicht mehr signifikant von null verschieden.[10] Der RE-Schätzer

[10] Dies zeigt ein Test für Linearkombinationen (s. Analysefile).

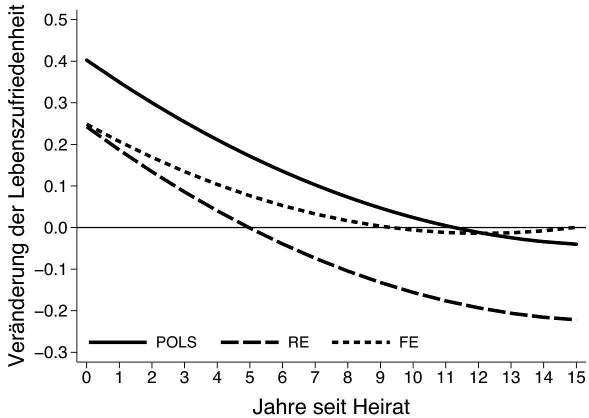

Abb. 3: Effekt der Heirat auf die Lebenszufriedenheit

liegt hier erstaunlicherweise nicht zwischen POLS und FE, sondern ist noch kleiner. Es ist allerdings ziemlich wahrscheinlich, dass er aufgrund von Selbstselektion verzerrt ist. Die inhaltliche Schlussfolgerung ist somit, dass eine (Erst-) Heirat die Lebenszufriedenheit erhöht. Allerdings ist dies ein vorübergehender Effekt: nach fünf Jahren kehrt die Zufriedenheit wieder auf ihren Ausgangswert zurück.[11]

Die Schätzergebnisse haben sich bei Verwendung des FE-Modells deutlich verändert. Insofern scheint Selbstselektion in unserem Anwendungsfall ein Problem darzustellen und man sollte das FE-Modell vorziehen. Eine formale Entscheidungshilfe liefert ein Hausman-Test. Ein Hausman-Test testet die Nullhypothese, dass sich die RE-Schätzer nicht von den FE-Schätzern unterscheiden. In unserem Fall hat die χ^2-Statistik den Wert 645 bei 5 Freiheitsgraden und ist damit hochsignifikant. Die RE-Schätzer weichen damit signifikant von den FE-Schätzern ab und man muss das FE-Modell verwenden.

Die oben angeführten Signifikanztests basieren auf panel-robusten Standardfehlern. Damit ist Heteroskedastizität und Autokorrelation unbestimmter Art innerhalb der Panels zugelassen. Das führt normalerweise zu deutlich größeren Standardfehlern. Schätzt man beispielsweise mit unseren Daten ein vereinfachtes FE-Modell (ohne die „Jahre seit der Heirat“), so beträgt der konventionelle Standardfehler der Heiratsdummy 0,018. Der daraus resultierende t-Wert ist 9,74. Der panel-robuste Standardfehler ist 0,024 und der t-Wert ist nur mehr 7,29. Konventionelle Standardfehler führen zu einer Überschätzung der Signifikanz der Effekte.

Schließlich sei auf den Modellfit eingegangen. In Panelregressionsmodellen kann man verschiedene R^2-Werte berechnen: Within, Between, Overall. Da POLS die Varianzkomponenten nicht trennt, ist hier das Overall-R^2 ausgewiesen: Heirat, Alter und Einkommen zusammen erklären 2,6 % der gesamten Zufriedenheitsvarianz zwischen und innerhalb der Personen. Bei RE- und FE-Modell hat man dagegen die Wahl. Der Within-Logik folgend ist es hier sinnvoll, das Within-R^2 auszuweisen. Im FE-Modell

[11] Dies ist analog zu dem Ergebnis, das Lucas et al. (2003) mit dem SOEP 1984–1998 berichten (wobei sie RE-Modelle verwenden).

erklären die Variablen 1,5 % der Zufriedenheitsvariation innerhalb der Personen. Man beachte hier einen (feinen) Unterschied zur üblichen R^2-Interpretation: Das Within-R^2 gibt den Anteil der erklärten Varianz innerhalb der Personen, die Varianz zwischen den Personen ist völlig ausgeblendet. Die 98,5 % nicht erklärte Varianz sind also nicht durch unbeobachtete Personen-Heterogenität verursacht, sondern durch unterschiedliche Lebenssituationen und Stimmungslagen, in denen sich die Respondenten zum Zeitpunkt der Interviews befinden. Das Overall-R^2 in einem analogen LSDV-Modell fällt viel höher aus und beträgt 44,7 %.

3.3 Ein Modell mit Periodeneffekten

Die bisherigen Modelle kontrollieren nicht für Periodeneffekte. Standardmäßig sollte man dies aber – wie oben bereits ausgeführt – tun. Das Modell ist nun:

$$y_{it} = \mu_t + \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \,. \tag{17}$$

μ_t sind die Periodeneffekte, die ein periodenspezifisches Zufriedenheitsniveau modellieren. Dieses Modell heißt auch „two-way FE-Modell“, denn es enthält nicht nur fixe Effekte für jede Person sondern auch für jede Periode. Die fixen Periodeneffekte werden über Perioden-Dummies mitgeschätzt. Modell 4 in Tabelle 2 enthält die Schätzergebnisse (die Perioden-Effektschätzer sind aus Platzgründen nicht aufgeführt). Wie man sieht, ändern sich in unserem Fall die Schätzergebnisse kaum. Nur der Alterseffekt geht dem Betrag nach deutlich zurück.

Mit der Berücksichtigung der Periodeneffekte handelt man sich allerdings ein methodisches Problem ein: das APC-Problem (Age-Period-Cohort). Es ist nicht möglich, Perioden-, Alters- und Kohorteneffekte in ein lineares Regressionsmodell gleichzeitig aufzunehmen, weil perfekte Kollinearität vorliegt: Das Alter ergibt sich z. B. aus Periode minus Kohorte (siehe Rabe-Hesketh & Skrondal 2008, S. 182 ff.). Mit Querschnittsdaten kann man nur einen der drei Zeiteffekte schätzen, weil die Periode konstant ist (Alter und Kohorte sind kollinear). Mit Kohorten-Paneldaten kann man auch nur einen Zeiteffekt schätzen, weil die Kohorte konstant ist (Alter und Periode sind kollinear). Mit Paneldaten mit mehreren Kohorten kann man zwei der Zeiteffekte schätzen. Meist wird man den Alters- und den Periodeneffekt in das Modell aufnehmen.

Das FE-Modell (3, in Tabelle 2) enthält einen linearen Altersterm, aber keinen Kohortenterm. Dennoch bricht das Modell zusammen, wenn man Perioden-Dummies (oder einen linearen Periodenterm) einführt. Wieso das, wo doch kein Kohortenterm im Modell ist? POLS und RE haben kein Problem mit einem Modell, das nur Alters- und Periodenterme enthält. Das FE-Modell allerdings schon, weil es ja implizit auch die zeitkonstante Variable „Kohorte“ mitkontrolliert. Innerhalb einer Person ist die Kohorte konstant und Alter und Periode sind perfekt kollinear. Man kann das APC-Problem umgehen, indem man Restriktionen einführt. In unserem Falle liegt es nahe, Perioden-Dummies für Jahre, die ähnliche mittlere Zufriedenheitswerte aufweisen (geschätzt mit POLS), zusammenzufassen. Die Jahre 1984, 1986, 1990 und 1991 weisen die höchsten Zufriedenheitswerte auf, weshalb ich für diese vier Jahre nur eine Dummy berücksichtige. Mit diesen so restringierten Perioden-Dummies gelingt es, das Modell

Tab. 3: Regressionsmodelle auf die Lebenszufriedenheit (SOEP 1984–2007)

	(1) RE	(2) FE	(3) HYBRID
Heirat	0,103***	0,175***	
Heirat (within)			0,175***
Heirat (between)			0,358***
Alter	−0,026***	−0,042***	
Alter (within)			−0,042***
Alter (between)			−0,010***
HHeink	0,123***	0,086***	
HHeink (within)			0,086***
HHeink (between)			0,435***
Frau	0,056**	—	0,074***
Within-R^2	0,013	0,014	0,014
Zahl Personenjahre	108369	108369	108369

*: p $\leq$ 0,05; **: p $\leq$ 0,01; ***: p $\leq$ 0,001; Standardfehler panel-robust geschätzt.

4 zu schätzen. Gegenüber Modell 3 geht der Alterseffekt deutlich zurück, was darauf hindeutet, dass der Alterseffekt ohne Berücksichtigung des Periodeneffektes (dem Betrag nach) überschätzt wird (der Tendenz nach zeigt sich, dass über die Jahre die Zufriedenheit eher zurückgeht). Allerdings muss man darauf hinweisen, dass die Alters- und Periodenschätzer nur dann unverzerrt sind, wenn die Restriktion zutrifft. Ist die Einschränkung (die vier Jahre gleichzusetzen) so nicht zutreffend, so sind die Schätzer verzerrt.

3.4 Ein Hybrid-Modell

Um das oben eingeführte Hybrid-Modell zu demonstrieren, vereinfachen wir die Modellierung. Wir wollen Modelle mit der Heirat, dem Alter, dem Haushaltseinkommen und dem Geschlecht schätzen. Tabelle 3 enthält in den ersten beiden Spalten die Ergebnisse für ein RE- und ein FE-Modell. Die Ergebnisse fallen ähnlich zu den Bisherigen aus. Wie üblich ist der Effekt des Geschlechts im FE-Modell nicht schätzbar. Ein Hybrid-Modell erhält man nun, wenn man jede zeitveränderliche unabhängige Variable in eine Between- und eine Within-Komponente zerlegt. Die Between-Komponente ist das Mittel über alle Beobachtungen einer Variable bei einer Person (Between-Transformation). Die Within-Komponente ist die Variation um das Personenmittel (Within-Transformation). Mit diesen so transformierten zeitveränderlichen Variablen schätzt man dann ein RE-Modell, in das man auch (nicht-transformierte) zeitkonstante Variablen aufnehmen kann.

Tabelle 3 enthält in Spalte 3 die Ergebnisse für unser Beispiel. Man sieht, dass die Schätzer der Within-Komponente exakt die FE-Schätzer reproduzieren. Das Hybrid-Modell erlaubt somit eine FE-Schätzung, wobei aber noch weitere Koeffizienten geliefert werden. Die Between-Komponenten liefern die Between-Schätzer. Die sind allerdings – wie wir wissen – bei Vorliegen von unbeobachteter Heterogenität verzerrt. Eine starke Abweichung von den Within-Schätzern kann deshalb einen Hinweis auf Probleme mit unbeobachteter Heterogenität liefern. In unserem Fall gibt es bei allen drei Variablen deutliche Abweichungen, was noch einmal bestätigt, dass es mit den Happiness-Daten massive Selbstselektionsprobleme gibt. Schließlich liefert das Hybrid-Modell einen Schätzer für den Geschlechtseffekt. Der fällt etwas stärker aus als der RE-Schätzer. Er kann interpretiert werden als der Effekt, unter Kontrolle sowohl der Within- wie auch der Beween-Variation der zeitveränderlichen Variablen. Es muss allerdings betont werden, dass es sich hierbei nicht um einen Within-Schätzer handelt. Bei Vorliegen von unbeobachteter Heterogenität wird der Geschlechtseffekt verzerrt sein.

3.5 Ein Wachstumskurven-Modell

Das Hybrid-Modell erlaubt die Berücksichtigung der Geschlechts-Dummy. Mit Paneldaten kann man aber nicht nur Gruppenunterschiede im Niveau schätzen, sondern auch Gruppenunterschiede in der Dynamik. Die Frage lautet dann: Wie unterscheidet sich die Entwicklung der Zufriedenheit im Lebensverlauf von Männern und Frauen? In der Literatur nennt man das Wachstumskurven-Analyse (Growth Curve Analysis, siehe Kapitel 38 in diesem Handbuch). Wachstumskurven werden meist mit RE-Modellen geschätzt (Random Coefficient Modelle, siehe Rabe-Hesketh & Skrondal 2008, S. 210 ff.). Aber auch in einem FE-Modell kann man Wachstumskurven schätzen und hat damit den Vorteil, dass die Wachstumskurven nicht durch personenspezifische Heterogenität verzerrt geschätzt werden. Im Folgenden soll deshalb kurz erläutert werden, wie man FE-Wachstumskurven schätzen kann.

Wachstumskurven sind über eine Zeitvariable definiert. In unserem Kontext könnten wir Wachstumskurven über die Jahre seit Heirat schätzen und könnten damit etwa untersuchen, wie stark sich der Heiratseffekt zwischen Männern und Frauen unterscheidet (diese Analysen zeigen, dass der Unterschied gering ist). Am häufigsten wird aber die Zeitvariable „Lebensalter" betrachtet. Oben haben wir „Alter" linear im Modell berücksichtigt und damit eine lineare Wachstumskurve mitgeschätzt. Genauere Analysen zeigen, dass eine kubische Modellierung angemessener ist. Die Wachstumskurven modellieren wir deshalb im Folgenden mit „Alter", „Alter^2" und „Alter^3". Die Wachstumskurven-Analyse macht sich nun die Tatsache zu nutze, dass im FE-Modell Interaktionen von zeitkonstanten und zeitveränderlichen Variablen schätzbar sind. Wir können also „Geschlecht" mit den drei Altersvariablen interagieren und in einem FE-Modell die Effekte schätzen. Die numerischen Ergebnisse der Modellschätzung führen wir hier nicht an, aber die geschätzten Wachstumskurven sind in Abbildung 4 eingetragen.

Die durchgezogene Linie ist die Entwicklung der Zufriedenheit über den Lebensverlauf bei Frauen, die gestrichelte bei Männern. Beide Wachstumskurven sind jeweils von einem 95 %-Konfidenzintervall (panel-robust) umgeben. Betrachtet man zuerst den Verlauf der Wachstumskurven, so erkennt man, dass der Rückgang der Zufriedenheit bis etwa 60 annähernd linear verläuft, danach beschleunigt er sich. Das erscheint plausibel, denn aus der Literatur ist bekannt, dass insbesondere kurz vor dem Tod die

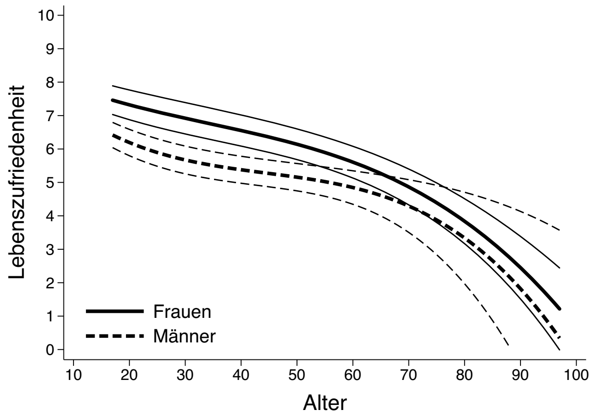

Abb. 4: Zufriedenheit im Lebensverlauf

Lebenszufriedenheit stark abfällt. Nun aber zu den Geschlechtsunterschieden. Aus den obigen Analysen (siehe Tabelle 3) wissen wir, dass Frauen zufriedener sind als Männer. Hier sehen wir noch viel mehr: Die Zufriedenheit ist bei Frauen in jungen Jahren etwa einen Skalenpunkt höher. Da sich die Konfidenzintervalle hier nicht überlappen, ist dieser Unterschied signifikant. In höherem Alter verringert sich der Unterschied und ist ab etwa 50 nicht mehr signifikant. Ein F-Test (nicht angeführt) zeigt, dass sich die Wachstumskurven von Männern und Frauen insgesamt signifikant unterscheiden.[12]

4 Nicht-lineare FE-Modelle

Bisher haben wir nur lineare Panelregressionsmodelle betrachtet. Doch auch für den nicht-linearen Fall sind Panelregressionsmodelle verfügbar. Wie im linearen Fall gibt es Pooled-, RE- und FE-Modelle. Die Vor- und Nachteile gelten analog. Die gepoolten und die RE-Modelle verwenden den Between-Vergleich und sind deshalb bei Vorliegen personenspezifischer unbeobachteter Heterogenität verzerrt. FE-Modelle dagegen liefern in dieser Situation konsistente Schätzer, die auch im nicht-linearen Fall auf dem Within-Prinzip basieren. Aus Platzgründen können im Folgenden nur die wichtigsten Modelle kurz vorgestellt werden (ausführlicher Cameron & Trivedi 2005, Kap. 23; Allison 2009).

[12] Man kann noch einen Schritt weitergehen und personenspezifische Wachstumskurven erlauben. Das ist eine Erweiterung des FE-Modells, bei der man nicht nur personenspezifische Achsenabschnitte zulässt, sondern personenspezifische Wachstumskurven (Fixed Growth Modell, oder FE-IS). Damit kontrolliert man nicht nur für zeitkonstante Heterogenität, sondern (eingeschränkt) auch für zeitveränderliche Heterogenität. Eine Anwendung des FE-IS Modells findet man bei Ludwig & Brüderl (2009).

4.1 Nicht-lineare Panelregressionsmodelle

Bevor die konkreten Modelle vorgestellt werden, beginnen wir mit einigen Bemerkungen zum allgemeinen Modellierungsansatz. In linearen Panelregressionsmodellen – wie wir sie bisher kennen gelernt haben – modelliert man den bedingten Mittelwert als $E[y_{it}|\alpha_i,\mathbf{x}_{it}] = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i$. Der personenspezifische Fehler geht additiv in das Modell ein. Der große Vorteil dieser Modellierung ist, dass man die personenspezifischen Fehler herausdifferenzieren kann (z. B. durch die Within-Transformation).

In nicht-linearen Modellen macht es dagegen meist keinen Sinn, bedingte Mittelwerte zu modellieren. Häufig modelliert man bedingte Dichten: $f(y_{it}|\alpha_i,\mathbf{x}_{it}) = f(y_{it},\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i,\gamma)$. γ ist ein Streuungsparameter. Durch diese Modellierung ist es aber nicht mehr möglich, die personenspezifischen Fehler herauszudifferenzieren. Man muss sie mit den interessierenden Parametern $(\boldsymbol{\beta},\gamma)$ als „Störgrößen" (Nuisance Parameters) mitschätzen. Dies erzeugt das so genannte „Incidental Parameter" (IP) Problem: Mit $N \to \infty$ geht auch die Zahl der zu schätzenden Parameter gegen Unendlich. Dies verletzt eine zentrale Konsistenzbedingung der ML-Schätzung, weshalb die Parameterschätzer inkonsistent sind.

Deshalb ist es im nicht-linearen Falle nicht einfach, FE-Modelle konsistent zu schätzen. Dies ist nur in den Fällen möglich, in denen es gelingt, die Störgrößen durch suffiziente Statistiken aus der Likelihood herauszukonditionieren (Conditional Likelihood). Die Schätzung von FE-Modellen mittels bedingter Likelihood ist in folgenden Fällen möglich:

1. Bei Zähldatenmodellen (Poisson, Negbin),
2. binären Modellen (Logit), und
3. bei Ereignisdatenmodellen (Cox).

Auf die Zähldatenmodelle gehe ich im Folgenden nicht ein, aber das FE-Logit und das FE-Cox Modell will ich vorstellen.[13] Für diese und noch viele weitere nicht-lineare Modelle existieren auch RE-Modelle, worauf ich im Folgenden aber nicht eingehe.

4.2 Das FE-Logit Modell

Bei binärer abhängiger Variable ist ein sinnvoller Modellierungsansatz ein Logit-Modell mit personenspezifischen Fehlern:

$$P(y_{it} = 1) = \frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i)}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i)} \,. \tag{18}$$

Dieses Modell ist analog zum für Querschnittsdaten gern verwendeten Logit-Modell (siehe Kapitel 31 in diesem Handbuch). Allerdings ist es erweitert um einen personenspezifischen Fehlerterm. Für diesen kann man wie im linearen Modell eine RE- oder

[13] Auf den ersten Blick kann man natürlich immer ein FE-Modell über die Dummy-Methode schätzen: Man erweitert das nicht-lineare Regressionsmodell um Personen-Dummies. Abgesehen von den praktischen Problemen bei großem N führt dies bei nicht-linearen Modellen aufgrund des IP-Problems leider nicht zu konsistenten Schätzungen. Allerdings haben Simulationsstudien gezeigt, dass die Verzerrung bei langen Panels ($T > 10$) gering ist.

eine FE-Annahme treffen. Das *FE-Logit* ist mittels bedingter Likelihood konsistent schätzbar, denn es existiert eine suffiziente Statistik nämlich $\sum_t y_{it}$ (die Zahl der Einsen im Panel einer Person). Deshalb nennt man dieses Model oft auch „Conditional Logit“. Intuitiv kann man sich das so vorstellen, dass durch die Zahl der Einsen α_i „kontrolliert“ ist, und damit die Schätzer nicht mehr von personenspezifischer Heterogenität verzerrt werden. Dadurch hat man den Vorteil der FE-Methodologie auch im Logit-Modell zur Verfügung: Der Schätzer für $\boldsymbol{\beta}$ ist konsistent auch wenn es zeitkonstante, personenspezifische unbeobachtete Heterogenität gibt (für das Pooled- bzw. RE-Logit gilt dies nicht).[14]

Zwei Besonderheiten des FE-Logit Modells führen gerne zu Verwirrung. Erstens ist es nicht möglich auf $\sum_t y_{it} = 0$ bzw. $\sum_t y_{it} = T$ zu konditionieren. D. h. Personenjahre von Personen, die auf allen Beobachtungen der abhängigen Variable konstant 0 oder 1 aufweisen, fliegen aus der Schätzung. Das kann den Analysedatensatz drastisch reduzieren, wenn auf der abhängigen Variable nicht genügend Varianz ist. Bei vielen Nutzern (und leider auch Gutachtern) entsteht dadurch die irrige Meinung, dass dies ein Stichprobenauswahl-Problem induziert. Dem ist aber nicht so, denn diese Beobachtungen tragen keine Information zur Schätzung der Regressionseffekte bei. Die Effekte sind – analog zum linearen Fall – als ATET-Effekte zu interpretieren.

Zweitens ist die Interpretation der FE-Schätzer ungewohnt (siehe Allison 2009, S. 36 f.). Die Koeffizientenschätzer jeder Regression können in zweifacher Art und Weise interpretiert werden: als subjektspezifischer Effekt (der Effekt bei einer bestimmten Person) und als Populationsdurchschnitts-Effekt (der durchschnittliche Effekt in der Population). Der Populationsdurchschnitts-Effekt ergibt sich durch Aggregation der subjektspezifischen (Marginal-) Effekte. Da in Fehlerkomponenten-Modellen ein personenspezifischer Fehlerterm enthalten ist, unterscheiden sich die beiden Interpretationen potentiell. Bei der Berechnung eines subjektspezifischen Marginaleffektes geht α_i als eine festzulegende Konstante ein. Um den Populationsdurchschnitts-Effekt zu erhalten, muss man aber die Verteilung von α_i in der Population berücksichtigen und darüber integrieren. Bei linearen Modellen unterscheiden sich subjektspezifische und Populationsdurchschnitts-Koeffizienten nicht, weil der subjektspezifische Marginaleffekt unabhängig von α_i ist. Bei nicht-linearen Modellen gilt dies nicht mehr, da im Allgemeinen der subjektspezifische Marginaleffekt abhängig von α_i ist. Welche Koeffizienten ein Modell schätzt, hängt ab von der Modellspezifikation: Fehlerkomponenten-Modelle (FE und RE) weisen subjektspezifische Koeffizienten aus, Pooled-Modelle Populationsdurchschnitts-Koeffizienten.[15] Nun gilt allgemein, dass Populationsdurchschnitts-Koeffizienten (betragsmäßig) kleiner ausfallen, als die sub-

[14] Die Anwendbarkeit des FE-Logit ist auf Situationen beschränkt, in denen jederzeit ein Wechsel von 0 zu 1 und umgekehrt erfolgen kann. Denn gegeben eine bestimmte Zahl von Einsen werden alle möglichen Sequenzmuster bestehend aus 0 und 1 in Abhängigkeit von X verglichen. Die Zahl der möglichen Sequenzmuster ist aber drastisch eingeschränkt, wenn etwa nur ein Wechsel von 0 nach 1 möglich ist, aber nicht zurück. Das ultimative Beispiel hierfür wäre eine Mortalitätsanalyse mit Tod= 1. Allerdings präsentiert Allison (2009, S. 79 ff.) auch für diesen Fall ein FE-Verfahren, dessen Intuition mir jedoch schleierhaft ist.

[15] In der epidemiologischen Literatur wird oft prototypisch für ein „Subject-Specific“ Modell das RE-Logit und für ein „Population-Averaged“ Modell das GEE-Logit (Generalized

jektspezifischen Koeffizienten (Rabe-Hesketh & Skrondal 2008, S. 254 ff.). Dies wird durch die Streuung der α_i in der Population verursacht (Heterogenitätsschrumpfung).[16] Allison (2009, S. 37) führt für den Logit-Fall folgende näherungsweise Beziehung an (falls α_i normalverteilt ist):

$$\beta^* \approx \frac{\beta}{\sqrt{0{,}346V(\alpha_i)+1}}\,, \tag{19}$$

wobei β^* der Populationsdurchschnitts- und β der subjektspezifische Koeffizient ist. Gibt es keine Heterogenität in der Population ($V(\alpha_i) = 0$), so sind beide Koeffizienten gleich. Mit zunehmender Heterogenität „schrumpft“ β^* gegen Null. Ist z. B. $V(\alpha_i) = 9$, so sind die Populationsdurchschnitts-Koeffizienten nur mehr halb so groß.

Für die Kausalanalyse sind die subjektspezifischen Koeffizienten bedeutsamer, da sie den Effekt von X auf die abhängige Variable schätzen. Populationsdurchschnitts-Koeffizienten dagegen hängen von der Verteilung von α_i in der gegebenen Stichprobe ab. Insofern ist es durchaus erwünscht, wenn das FE-Logit die subjektspezifischen Koeffizienten ausweist. Allerdings muss man aufpassen, wenn man die Pooled- und FE-Schätzer vergleicht. Die Pooled-Schätzer können sich nun aus zwei Gründen von den FE-Schätzern unterscheiden: Sie könnten durch unbeobachtete Heterogenität verzerrt sein und sie sind von Heterogenitätsschrumpfung betroffen.[17]

Schließlich muss man bei der Interpretation der (subjektspezifischen) FE-Logit Koeffizienten aufpassen. Die Interpretation der Vorzeichen der Logit-Koeffizienten ist unproblematisch, auch die übliche Odds-Interpretation ist möglich. Aber – wie oben schon angesprochen – die Interpretation als Wahrscheinlichkeitseffekte ist problematisch, da sie von α_i abhängt. Die üblichen Algorithmen setzen $\alpha_i = 0$. Das kann aber eventuell ein sehr spezieller Fall sein und die Effektinterpretation dadurch in die Irre führen.

4.3 FE-Modelle für Ereignisdaten

Schließlich sei noch erläutert, wie man Ereignisdaten in einem Panelkontext analysieren kann. Die üblichen Modelle der Ereignisdatenanalyse (siehe Kapitel 37 in diesem Handbuch) beruhen – wie jede andere Querschnittsregression auch – auf dem Between-Vergleich. Damit sind ihre Schätzer potentiell von unbeobachteter Personenheterogenität bedroht. In der Ereignisdatenanalyse sehr populär sind Modelle für „unbeobachtete Heterogenität“. Viele Anwender glauben, dadurch die durch unbeobachtete Heterogenität verursachten Probleme im Griff zu haben. Das ist ein Irrglaube. Im Prinzip sind dies RE-Modelle, die mit nur einer Beobachtung pro Person geschätzt

Estimating Equations) angeführt. Das GEE-Logit ist ein Pooled-Logit, wobei die Kovarianzmatrix der Fehlerterme unrestringiert ist (eine andere Art robuster Standardfehler).

[16] Die schöne deutsche Terminologie habe ich der Dissertation von Jette Schröder entnommen.

[17] Ein Hybrid-Modell (analog zu dem Vorgehen in Abschnitt 2.4), welches man mit Pooled-Logit statt RE schätzt, liefert FE-Schätzer, die auch von Heterogenitätsschrumpfung betroffen sind. Ein Vergleich mit den normalen FE-Schätzern zeigt dann das Ausmaß der Heterogenitätsschrumpfung.

werden. Die Zufallseffekte sind mit nur einer Beobachtung natürlich nicht identifiziert und insofern verwundert es nicht, dass die Schätzergebnisse dieser Modelle massiv von den getroffenen Annahmen beeinflusst sind. Die in der Ereignisdatenanalyse so genannten Modelle mit unbeobachteter Heterogenität sind somit unbrauchbar.

Ereignisdatenanalyse wird erst wirklich längsschnittlich, wenn man pro Person mehrere Episoden beobachtet hat (Multiple Episodes). Dies ist dann möglich, wenn das betrachtete Ereignis wiederholt auftreten kann (Repeated Events). Z. B. kann man in Gesellschaften mit hoher Scheidungsrate für viele Personen mehrere Ehe-Episoden beobachten. Liegen multiple Episoden vor, so hat man im Prinzip eine Paneldatenstruktur (mehrere Beobachtungen pro Person). Damit kann man auch in der Ereignisdatenanalyse Within-Schätzer verwenden und von deren Eigenschaft – bei Vorliegen personenspezifischer unbeobachteter Heterogenität konsistente Schätzer zu liefern – profitieren.

In neueren Lehrbüchern der Ereignisdatenanalyse werden inzwischen auch Modelle für wiederholte Ereignisse vorgestellt. Dabei werden aber die Prioritäten falsch gesetzt. Erste Priorität gilt den panel-verzerrten Standardfehlern. Deshalb gibt man sich mit RE-Modellen (Frailty-Modelle genannt) und panel-robusten Standardfehlern zufrieden. Das Potential der Within-Schätzung lässt sich mit RE-Modellen aber nicht umsetzen. Bisher noch weitgehend unbekannt ist, dass es auch FE-Modell für Ereignisdaten gibt (siehe Allison 1996, 2009).

Im Prinzip stehen zwei Analyseoptionen zur Verfügung (siehe genauer Brüderl 2008): Modelle für diskrete oder stetige Zeit. Bei jährlich erhobenen Paneldaten liegt es oft nahe, Modelle für diskrete Zeit zu verwenden. Die logistische Regression ist ein in der diskreten Ereignisdatenanalyse sehr häufig verwendetes Modell. Im Falle von wiederholten Ereignissen in diskreter Zeit liegt es deshalb nahe, das oben vorgestellte FE-Logit zu verwenden.

Wurde die Prozesszeit stetig gemessen, so ist ein FE-Cox-Modell verfügbar. Das aus der Ereignisdatenanalyse bekannte Cox-Modell (siehe Kapitel 37 in diesem Handbuch) wird um einen personenspezifischen Fehlerterm erweitert. Das FE-Cox-Modell lautet:

$$r_{ij}(t) = r_0(t)\exp(\mathbf{x}'_{ij}(t)\boldsymbol{\beta} + \alpha_i)\,, \tag{20}$$

wobei $r_{ij}(t)$ die Ereignisrate der Person i in der Episode j ist, die von der Verweildauer t abhängt. $r_0(t)$ ist die Basisrate, die für alle Personen als identisch angenommen wird. Die Effekte der Kovariaten X werden exponentiell parametrisiert, was proportionale Effekte impliziert. Die Kovariaten können zeitveränderlich sein. Die Verweildauer wird in den meisten Anwendungen so definiert sein, dass sie bei jedem Ereignis wieder bei null beginnt (Gap Time Approach). α_i ist ein personenspezifischer, zeitkonstanter Fehlerterm, der potentiell mit den X-Variablen korreliert ist. Ist dies der Fall, so sind z. B. die Pooled-Cox-Schätzer verzerrt. Die FE-Schätzer dagegen wären nicht verzerrt. Aber das FE-Cox-Modell ist aufgrund des IP-Problems nicht so ohne weiteres schätzbar. Mittels eines „Tricks“ gelingt es allerdings, die α_i aus dem Modell zu entfernen. Man kann die α_i in die Basisrate integrieren:

$$r_{ij}(t) = r_{0i}(t)\exp(\mathbf{x}'_{ij}(t)\boldsymbol{\beta})\,. \tag{21}$$

$r_{0i}(t)$ ist nun eine personenspezifische Basisrate. Damit modelliert man nicht nur das personenspezifische Niveau – wie in Standard-FE-Modellen –, sondern sogar personenspezifische Ratenverläufe – ähnlich dem FE-IS Modell. Der Punkt ist nun, dass das Schätzverfahren des Cox-Modells – Partial Likelihood – die Basisrate nicht mitschätzt, wenn man jede Person als eine eigene Schicht definiert (Stata Option `strata(id)`). Eine nach Personen geschichtete Cox-Regression ermöglicht somit eine einfache Within-Schätzung bei wiederholten Ereignissen.

Das oben angeführte Scheidungsbeispiel soll den entscheidenden Punkt verdeutlichen. Man könnte die Vermutung habe, dass Ehen, bei denen sich die Partner in einer Disko kennen lernten, weniger stabil sind. Eine Cox-Regression, bei der man nur eine Ehe pro Person zur Verfügung hat, ist ein Between-Schätzer. Der geschätzte Disko-Effekt beruht ganz und gar auf dem Vergleich verschiedener Personen. Der könnte aber durch unbeobachtete Merkmale, auf denen sich Diskobesucher von anderen Personen unterscheiden, verzerrt sein. Hat man multiple Ehe-Episoden zur Verfügung, so kann man FE-Cox anwenden. Hier beruht der geschätzte Disko-Effekt auf einem Within-Vergleich: Bei ein und derselben Person wird die Dauer einer Disko-Ehe mit der Dauer einer anderen Ehe verglichen, die nicht in der Disko begann. Es sollte einsichtig sein, dass der Within-Vergleich weniger von unbeobachteter Heterogenität bedroht ist. Eine Beispiel-Anwendung des FE-Cox-Modells mit SOEP-Daten findet man in Brüderl (2008).

5 Häufige Fehler

In diesem Abschnitt sollen suboptimale Modelle der Paneldatenanalyse, die dennoch häufig verwendet werden, besprochen werden. Weiterhin soll auf die Grenzen des FE-Ansatzes eingegangen werden.

5.1 Suboptimale Panelregressionsmodelle

Es gibt verschiedene statistische Traditionen, in deren Rahmen Panelmodelle vorgeschlagen wurden: LISREL-Modelle (Kapitel 29 in diesem Handbuch), MANOVA-Modelle (Kapitel 19 in diesem Handbuch), Mehrebenen-Modelle (Kapitel 28 in diesem Handbuch), Wachstumskurven-Modelle (Kapitel 38 in diesem Handbuch). Eine große Vielzahl solcher Panelmodelle wird z. B. bei Rabe-Hesketh & Skrondal (2008) beschrieben. Die meisten dieser Modelle sind allerdings RE-Modelle. Man muss sich bei Verwendung dieser Modelle immer fragen, ob dies nun ein Within- oder ein Between-Schätzverfahren ist. Diese fundamentale Unterscheidung geht in mancher Terminologie leicht verloren. Oft werden hoch komplexe, eindrucksvolle Modelle postuliert, die aber im Grunde RE-Modelle sind.

Besonders irreführend sind die so genannten dynamischen Panelmodelle.[18] Als „dynamisch" bezeichnet man Panelmodelle mit verzögert endogener Variable auf der rechten Seite, z. B.:

[18] Der Name ist irreführend, denn auch die bisher vorgestellten, „statischen" Modelle sind dynamisch, weil sie die zeitliche Struktur in den Paneldaten modellieren.

$$y_{it} = \delta y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \,. \tag{22}$$

Dieses Modell hat intuitiven „Appeal“, da es die „Beharrungstendenz“ von Paneldaten explizit modelliert. $\hat{\delta}$ gibt Auskunft über das Ausmaß der Beharrungstendenz.[19] Dynamische Modelle waren historisch die ersten Modelle, die zur Analyse von Paneldaten vorgeschlagen wurden. Eine Verallgemeinerung – das Cross-Lagged-Modell – wurde (und wird) gar als der Königsweg der Panelanalyse gesehen. Aber dynamische Modelle haben ein großes Problem: Durch die Berücksichtigung der verzögert endogenen Variable sind die Schätzer von sowohl δ wie auch $\boldsymbol{\beta}$ verzerrt (sogar unter der RE-Annahme). Deshalb kann man dynamische Modelle nicht mit den „einfachen“, bisher vorgestellten Modellen schätzen. Man benötigt komplexe Ökonometrie (siehe Cameron & Trivedi 2005, Kap. 22.5). Die für diese komplexen Verfahren nötigen Annahmen können aber nicht überprüft werden, weshalb in der Praxis unklar ist, ob die Schätzer durch die komplexen Verfahren tatsächlich besser werden. Der Hauptvorteil von Paneldaten – die Möglichkeit das Problem unbeobachteter Heterogenität mit relativ einfachen Methoden in den Griff zu bekommen – geht durch die Benutzung von dynamischen Modellen verloren. Insbesondere Rogosa (1988) zeigt, dass die weit verbreiteten Argumente für dynamische, Cross-Lagged- und Strukturgleichungs-Modelle Mythen sind. Seiner Meinung nach sind diese Modelle „useless“.[20]

5.2 Grenzen von FE-Modellen

Paneldaten bieten durch die Möglichkeit der Within-Schätzung große Vorzüge im Rahmen der Kausalanalyse. Allerdings sind auch Paneldaten in Verbindung mit FE-Modellen kein Allheilmittel:

1. FE-Modelle sind nicht für jede Fragestellung geeignet und
2. Endogenität und
3. Panel-Attrition können auch FE-Schätzer verzerren.

Eine Within-Schätzung ist im Prinzip ein Davor-Danach-Vergleich und ist deshalb nur dann durchführbar, wenn es eine Veränderung gibt. Eine Within-Schätzung ist nur mit zeitveränderlichen X-Variablen durchführbar. Selbst bei zeitveränderlichen Variablen gibt es manchmal Probleme, wenn im Paneldatensatz nicht genug X-Variation ist (dann sind die FE-Schätzer sehr unpräzise). Paneldaten sind ideal, um die „Effects of Events“ zu untersuchen. Sie helfen allerdings nicht bei der Schätzung des Kausaleffektes zeitkonstanter Variablen. Die Probleme diesbezüglich sind die Gleichen wie bei Querschnittsdaten. Dessen sollte man sich bewusst sein: Paneldaten helfen nichts für die Identifizierung von z. B. Geschlechtseffekten.

[19] Auch die statischen Modelle modellieren eine Beharrungstendenz, nämlich über den personenspezifischen Fehlerterm α_i. Unbeobachtete Heterogenität erzeugt hier eine „Spurious State Dependence“. In dynamischen Modellen wird eine zweite, „kausale“ Komponente der Beharrungstendenz modelliert (True State Dependence).

[20] Allerdings zeigt Allison (2009, Kap. 6), dass man FE-Modelle auch in einem Strukturgleichungsansatz schätzen kann. In diesem Rahmen kann man auch dynamische Modelle schätzen. Die dadurch mögliche Kombination von FE-Methodologie und dynamischen Modellen könnte diese Einschätzung relativieren.

Wie oben erwähnt, beruht die Konsistenz des FE-Schätzers auf der Annahme strikter Exogenität. Der idiosynkratische Fehler darf nicht mit den X-Variablen korreliert sein. Liegt Endogenität vor, so ist auch der FE-Schätzer verzerrt. Verschiedene Ursachen von Endogenität sind denkbar: Messfehler, zeitveränderliche unbeobachtete Heterogenität und umgekehrte Kausalität.

Eine Ursache können Messfehler in X sein. In einer Regression mit nur einem X werden Messfehler den Regressionskoeffizienten nach unten verzerren (Attenuation Bias). Bei mehreren X-Variablen ist die Richtung der Verzerrung unklar. Die durch Messfehler verursachte Verzerrung wird durch die FD- und FE-Schätzung sogar noch verstärkt. Dies könnte man als Argument für POLS und RE werten. Allerdings sind deren Schätzer durch unbeobachtete Heterogenität verzerrt. Ich würde behaupten, dass unbeobachtete Heterogenität in der Sozialforschung das größere Problem ist und man deshalb Modelle verwenden sollte, die unter unbeobachteter Heterogenität konsistente Schätzer liefern.

Eine weitere Ursache für die Verletzung der Annahme strikter Exogenität kann zeitveränderliche Heterogenität sein. Im Beispiel von Abschnitt 3 könnte man etwa argumentieren, dass mit einer Heirat oft ein Lohnanstieg einhergeht (Marital Wage Premium, siehe Ludwig & Brüderl 2009), der den Anstieg der Zufriedenheit erzeugt. Das würde den Schätzer des Heiratseffektes verzerren. Abhilfe schafft man leicht, indem man den Lohn im Modell kontrolliert – wie wir es in Abschnitt 3 getan haben. Hat man die Heterogenität verursachende Variable aber nicht gemessen, so hat man ein Problem. Die nun unbeobachtete zeitveränderliche Heterogenität kann den FE-Schätzer verzerren. Analoges kann im Fall von umgekehrter Kausalität passieren. Wenn eine Veränderung in Y die Veränderung in X verursacht, dann sind auch die FE-Schätzer verzerrt. Angenommen eine Heirat hat keinen Effekt auf die Zufriedenheit. Aber ein zufälliger Anstieg der Zufriedenheit führt dazu, dass man eher heiratet, dann wird der FE-Schätzer einen positiven Heiratseffekt schätzen.

Gegen diese Art von Problemen gibt es kein einfaches Mittel. Theoretisch hilft die Instrumentalvariablen (IV) Methode (siehe Cameron & Trivedi 2005, Kap. 22). Dazu benötigt man (mindestens) eine Instrumentalvariable, welche mit x_{it} korreliert ist, aber nicht mit ε_{it}. Letzteres kann man aber nicht überprüfen. Deshalb ist es Treu und Glauben überlassen, ob der IV-Schätzer konsistent ist. Zusätzlich sind IV-Schätzer nicht robust. Wenn die IV-Annahmen nicht stimmen, dann kann der IV-Schätzer sogar noch schlimmer verzerrt sein, als der konventionelle Schätzer. Besonders problematisch ist, dass man kaum beurteilen kann, welcher der in der Literatur berichteten IV-Schätzer nun „gut“ und welcher „schlecht“ ist. In Forschungsgebieten, in denen IV-Schätzer exzessiv genutzt werden, findet man deshalb oft extrem divergierende Schätzer und es ist unklar, welche Schätzer nun die „richtigen“ sind. Die IV-Methode führt deshalb zu einem furchtbaren Durcheinander und ich rate deshalb von ihr dringend ab. Analoges kann man für LISREL-Modelle konstatieren, die im Zusammenhang mit umgekehrter Kausalität (und Messfehlern) auch gerne verwendet werden.

Schließlich sei noch auf das Problem der Panel-Attrition eingegangen. Wenn Personen nicht-zufällig aus dem Panel ausscheiden, so kann das die Schätzungen verzerren. Attrition ist ein weit verbreitetes Phänomen in Panelstudien, weshalb es gerne als das Hauptproblem der Panelanalyse stilisiert wird. Das muss man allerdings differenzierter sehen, als das die meisten Autoren tun. Attrition ist ein Spezialfall der Stichprobenauswahl. Aus dieser Literatur weiß man, dass eine Auswahl, die von beobachteten Variablen abhängt, leicht korrigiert werden kann, indem man in den Regressionsmodellen diese Variablen kontrolliert. Im Panelkontext hat man es sogar noch besser: Selbst wenn Attrition von unbeobachteten zeitkonstanten Variablen abhängt, kann man die Verzerrung leicht korrigieren. Denn das FE-Modell kontrolliert ja auch für diese Variablen. Dies ist ein weiterer Vorteil des FE-Schätzers: Er ist nicht verzerrt durch Attrition, die von unbeobachteten zeitkonstanten Variablen abhängt (dies gilt leider nicht für die nicht-linearen FE-Modelle). Allerdings wird Attrition, die über zeitveränderliche Variablen läuft, unter Umständen auch den FE-Schätzer verzerren. Dass bestimmte Ereignisse – Umzug und Scheidung etwa – zu erhöhter Attrition führen, ist bekannt. Unter welchen Umständen das zu Verzerrungen in welchem Ausmaß führt ist allerdings bisher kaum erforscht. In der Literatur wurden dennoch – sozusagen auf Verdacht – Modelle zur Korrektur des Attrition-Bias vorgeschlagen. Hier gilt aber wieder analog, was ich oben über die anderen komplexen Modelle gesagt habe: Ob die komplexe Ökonometrie die Sache besser macht, ist unklar. Auf alle Fälle – das zeigt die Erfahrung – produziert komplexe Ökonometrie ein großes Durcheinander. Warum dann also den Aufwand komplexer Modelle betreiben? Mein Rat deshalb – und das sehen inzwischen auch manche Ökonometriker so (z. B. Angrist & Pischke 2009): „Keep it simple".

6 Literaturempfehlungen

Angrist & Pischke (2009) geben eine schöne Einführung in die modernen Methoden der Kausalanalyse (Regression, Matching, Within-Schätzer, Regression Discontinuity). Allison (1994) und Halaby (2004) plädieren vehement für den Einsatz von Fixed-Effects Modellen zur Analyse von Paneldaten. Rogosa (1988) „zerpflückt" die Argumente, die gegen Change-Score Modelle (z. B. alle Within-Modelle) vorgebracht werden. Eine anwendungsorientierte Einführung in die Panelregression geben Brüderl (2005) und Allison (2009). Formaler, aber immer noch leicht verständlich, sind die Kapitel 13 und 14 von Wooldridge (2003). Die Kapitel 21–23 von Cameron & Trivedi (2005) bieten eine fortgeschrittene Einführung in die Panelregression. In Cameron & Trivedi (2009) zeigen die Autoren, wie man Panelregressionen mit Stata umsetzt. Rabe-Hesketh & Skrondal (2008) geben eine Einführung in die Paneldatenanalyse mit Stata aus der Perspektive der Mehrebenenanalyse. FE-Modelle für Ereignisdaten werden in Allison (1996, 2009) und Brüderl (2008) behandelt. Ludwig & Brüderl (2009) und Lucas et al. (2003) sind Anwendungsbeispiele für die Panelregression.

Literaturverzeichnis

Allison, P. D. (1994). Using Panel Data to Estimate the Effects of Events. *Sociological Methods and Research*, 23, 174–199.

Allison, P. D. (1996). Fixed-Effects Partial Likelihood for Repeated Events. *Sociological Methods and Research*, 25, 207–222.

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks: Sage.

Angrist, J. D. & Pischke, J. (2009). *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Brüderl, J. (2005). *Panel Data Analysis*. Letzter Zugriff 28.05.2010: `http://www2.sowi.uni-mannheim.de/lsssm/lehre.html`.

Brüderl, J. (2008). *Event History Analysis*. Letzter Zugriff 28.05.2010: `http://www2.sowi.uni-mannheim.de/lsssm/lehre.html`.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Cameron, A. C. & Trivedi, P. K. (2009). *Microeconometrics Using Stata*. College Station: Stata Press.

Halaby, C. (2004). Panel Models in Sociological Research. *Annual Review of Sociology*, 30, 507–544.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Reexamining Adaption and the Set Point Model of Happiness. *Journal of Personality and Social Psychology*, 84, 527–539.

Ludwig, V. & Brüderl, J. (2009). *The Male Marital Wage Premium: Further Results on an Enduring Puzzle*. Mannheim: unveröffentlichtes Manuskript.

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*. College Station: Stata Press, 2. Auflage.

Rogosa, D. (1988). Myths about Longitudinal Research. In K. W. Schaie (Hg.), *Methodological Issues in Aging Research* (S. 171–209). New York: Springer.

Wooldridge, J. M. (2003). *Introductory Econometrics: A Modern Approach*. Mason: Thomson.