Modelle mit variablen Regressionskonstanten – Random- und Fixed-Effects-Modelle

Petra Stein, Dawid Bekalarczyk

8. September 2017

Inhaltsverzeichnis

Kapitel 1

Grundlegendes zu Modellen mit variablen Regressionskonstanten

Dieses Kapitel stellt eine Klasse häufig genutzter Regressionsmodelle vor, welche speziell auf Paneldaten zugeschnitten sind. Das sind sog. Modelle mit variablen Regressionskonstanten. Die Datenstruktur von Paneldaten ist aufgrund der Differenzierung zwischen Personen und Zeitpunkten im Vergleich zu Querschnittsdaten komplexer. Modelle mit variablen Regressionskonstanten gehen in vielen Fällen angemessener mit dieser Komplexität um, als Verfahren, welche diese Differenzierung ignorieren. Außerdem nutzen sie die differenzierte Datenstruktur, um die Qualität von Schätzungen in Regressionsmodellen zu verbessern.

1.1 Varianzzerlegung und Notation bei Paneldaten

Statistisch lässt sich die Unterscheidung zwischen der Personen- und der Zeitebene für *eine* Variable mithilfe der Zerlegung ihrer Varianz in zwei Bestandteile aufgreifen: Der „between variation“ (Varianz zwischen Personen) und der „within variation“ (Varianz zwischen Zeitpunkten, innerhalb von Personen). Bei Querschnittsdaten existiert hingegen nur die between variation. Zu einem Objekt liegt nur eine Angabe vor, so dass keine within variation

O↓ Z →	t_1	t_2	t_3	t_4
1	1,3	1,4	1,4	1,2
2	3	3,1	3,3	3,4
3	8,8	8,4	8,5	8,6
4	5,5	9	3	2,4
5	5,4	8,9	2,8	2,1
6	5,7	9,1	2,6	2,1
7	3	9,9	8	5
8	8	4	2,6	11
9	6,5	2,1	0	8,6

Tabelle 1.1: Werte einer Variablen von Objekten zu verschiedenen Zeitpunkten

bestimmt werden kann.

Tabelle 1.1 zeigt ein Beispiel für 4 Zeitpunkte und 7 Objekte (O=Objekte, Z=Zeitpunkte).

Betrachtet man in der Tabelle 1.1 die Daten für die Fälle 1 bis 3 (zeilenweise), so kann man eine Dominanz der *between variation* gegenüber der *within variation* feststellen. Die Werte der Objekte bleiben (in Relation zu dem objekteigenen Mittelwert, den man sich bei den Daten „dazudenken“ könnte) über die Zeitpunkte relativ konstant, während sich die Werte zwischen den Objekten relativ stark unterscheiden. Dies könnte z.B. eine Gruppe unterschiedlicher Individuen sein, welche bzgl. einer Einstellung jeweils eine eher gefestigte Meinung haben.

Die Fälle 4 bis 6 weisen genau das Gegenteil auf: Die Werte der Objekte variieren von Zeitpunkt zu Zeitpunkt sehr stark. Zwischen den Personen aber sind sie relativ ähnlich. *Within variation* dominiert hier. Dies könnte eine Gruppe von Personen sein, welche sich z.B. bzgl. eines Verhaltensmusters relativ ähnlich sind. Da sich aber zwischen den Messungen starker Wandel vollzieht, äußert sich dieses Verhalten von Zeitpunkt zu Zeitpunkt unterschiedlich (z.B. vor und nach einem Krieg).

Schließlich weisen die Fälle 7 bis 9 beide Arten der Streuung im ähnlichen Ausmaß auf.

Modelle mit variablen Regressionskonstanten greifen diese Varianzzerlegung

auf, welche sich – wie später gezeigt wird – auch auf die Kovarianz mehrerer Variablen erweitern lässt. Um diese Modelle verstehen zu können, muss zunächst einmal eine einheitliche und präzise Darstellung der Individual- und der Zeitebene eingeführt werden. Nur auf diesem Wege können panelanalytische Regressionsmodelle, welche eine Varianzzerlegung vornehmen, formal korrekt dargestellt werden. Dies geschieht, entsprechend der gängigen Vorgehensweise, mithilfe einer Index-Notation, die neben einem aus Querschnittsregressionen bekannten Individuen-Index auch einen Index für Zeitpunkte eingeführt.

Die Ergänzung einer Variablen x mit einem Personenindex „i“ (mit $i = 1, 2, ..., N$) um den Wellenindex „t“ (mit $t = 1, 2, ..., T$), zu x_{it} drückt aus, dass x Werte verschiedener Personen zu verschiedenen Zeiten annehmen kann. $x_{2,4}$ wäre folglich der x-Wert der 2. Person zum 4. Zeitpunkt einer Untersuchung.

Fundamental für das Verständnis des Prinzips von Modellen mit variablen Regressionskonstanten ist die Erkenntnis, dass auch die *Koeffizienten* eines Regressionsmodells mit diesen Indizes versehen werden können. Somit wäre im Falle der einfachen Regression[1] theoretisch das folgende „Maximal-Modell“ denkbar:

$$y_{it} = \alpha_{it} + \beta_{it} x_{it} + \epsilon_{it} \tag{1.1}$$

mit

y_{it} = Wert der abhängigen Variablen für die i-te Person zum Zeitpunkt t

α_{it} = Regressionskonstante für die i-te Person zum Zeitpunkt t

β_{it} = Regressionskoeffizient für die i-te Person zum Zeitpunkt t

x_{it} = Wert der unabhängigen Variablen für die i-te Person zum Zeitpunkt t

ϵ_{it} = Residuum für die i-te Person zum Zeitpunkt t

Der Rückgriff auf die Indizes verkürzt die Darstellung, da die Modellgleichung 1.1 im Grunde für jedes i und t als separate Gleichung ausgeschrieben werden

[1] „einfach“ bezieht sich hier auf den Sachverhalt, dass das Modell inhaltlich gesehen eine einzige unabhängige Variable berücksichtigt – dies lässt sich jedoch bedenkenlos auf eine multiple Regressionskonstruktion übertragen; dies gilt auch für weitere Formulierungen von Regressionsgleichungen in diesem Abschnitt.

könnte:

$$
\begin{aligned}
y_{1,1} &= \alpha_{1,1} + \beta_{1,1}x_{1,1} + \epsilon_{1,1} \\
y_{2,1} &= \alpha_{2,1} + \beta_{2,1}x_{2,1} + \epsilon_{2,1} \\
&\ldots \\
y_{1,2} &= \alpha_{1,2} + \beta_{1,2}x_{1,2} + \epsilon_{1,2} \\
y_{2,2} &= \alpha_{2,2} + \beta_{2,2}x_{2,2} + \epsilon_{2,2} \\
&\ldots \\
y_{N,T} &= \alpha_{N,T} + \beta_{N,T}x_{N,T} + \epsilon_{N,T}
\end{aligned}
\tag{1.2}
$$

Eine solche Modellformulierung ist jedoch weder inhaltlich sinnvoll, noch sind die Koeffizienten mathematisch bestimmbar. Denn es müsste pro Person und Zeitpunkt jeweils ein Regressionskoeffizient und eine Regressionskonstante geschätzt werden. In der Regel dienen Regressionsmodelle aber dazu, Zusammenhangsstrukturen auf Aggregatebene zu untersuchen. „Individuelle“ Koeffizienten, die auch noch von Welle zu Welle variieren, würden diese Funktion nicht erfüllen. So ist es selten für sozialwissenschaftliche Hypothesenprüfungen von Interesse, z.B. den Regressionskoeffizienten von „Herrn Müller“ und seinen persönlichen y-Achsenabschnitt aus dem Jahr 2004 zu kennen.

Mathematisch gesehen lassen sich die Koeffizienten darüber hinaus nicht schätzen, da die Anzahl der Freiheitsgrade negativ ist. Denn es stehen viel zu wenige Informationen zur Schätzung der großen Anzahl von Koeffizienten zur Verfügung.

Das Pendant zu diesem „Maximal-Modell“ auf Querschnittsebene wäre übrigens Gl. 1.3, deren Koeffizienten ebenfalls nicht schätzbar sind:

$$y_i = \alpha_i + \beta_i x_i + \epsilon_i \tag{1.3}$$

Doch auch wenn das in Gl. 1.1 dargestellte „Maximal-Modell“ für Paneldaten lediglich ein Gedankenexperiment darstellt, so sind Modelle mit variablen Regressionsparametern konstruierbar, die inhaltlich plausibel und mathema-

tisch berechenbar sind – im Gegensatz zum Querschnittsmodell in 1.3.[2]

Entscheidend ist hierbei, ob sich die Zusammenhangsstruktur auf einer der beiden Ebenen (Zeiten- oder Personenebene) von der Zusammenhangsstruktur auf der indifferenten Gesamtebene (= Zeiten- + Personenebene) signifikant unterscheidet.

In der Regel nimmt man bei Paneldaten an, dass die Daten einer hierarchischen Struktur folgen, wobei die Personenebene die oberste Ebene darstellt. Einzelbeobachtungen (z.B. die Messung des Einkommens einer bestimmten Person i zum Zeitpunkt t) werden damit als „eingebettet (nested) in Personen“[3] begriffen.[4] Eine solche Sichtweise legt nahe, dass sich die Werte einer Variablen als auch ggf. die Zusammenhänge zwischen mehreren Variablen *innerhalb* einer Person ähnlicher sind, als *zwischen* den Personen. Entsprechend der oben eingeführten Varianz-Terminologie würde die between-variation der within-variation überlegen sein. So kann sich z.B. das Einkommen eines freiberuflichen IT-Spezialisten oder eines Aushilfsarbeiters im Logistik-Bereich natürlich von Messzeitpunkt zu Messzeitpunkt unterscheiden. Aber diese Unterschiede mögen geringer sein, als der Unterschied zwischen dem über die Zeit feststellbaren Durchschnittseinkommen des IT-Spezialisten gegenüber dem korrespondierenden Durchschnittseinkommen des Aushilfsarbeiters. Beide Personen bewegen sich nämlich, trotz aller Schwankungen im Zeitverlauf, auf einem unterschiedlichen Einkommensniveau.

Wird dieser Gedanke auf eine größere Anzahl von Personen übertragen, dann kann von einer Einbettung von Einzelwerten in Personen dann gesprochen werden, wenn die Streuung eines Merkmals (oder Kovariation von zwei Merkmalen) zwischen Personen relativ zur Streuung innerhalb von Personen „bedeutsam“ ist. An späterer Stelle wird gezeigt, welche konzeptionellen Vorteile

[2] Das liegt daran, dass nur eine Ebene i der Personen vorhanden ist. Wird folglich der i-Index aus den beiden Parametern entfernt, dann bleibt keine Variationsquelle für die Parameter über. Sie werden automatisch zu Konstanten. Nur dann sind sie schätzbar. Diese Einschränkungen von Paneldaten gilt allerdings speziell für die Unterscheidung zwischen Individuen und Zeitpunkten. Dass Querschnittsdaten auf andere Weise hierarchisch beschaffen sein können, zeigt die unten gezogene Parallele zur Mehrebenenanalyse – s. Abs. 1.3

[3] Aber auch die umgekehrte Sichtweise ist möglich. Demnach wären Personen eingebettet in Zeitpunkte. Das würde z.B. dann Sinn machen, wenn die Zeitpunkte globale Ereignisse zeitlich markieren, welche auf nahezu alle Personen einen ähnlichen Effekt ausüben - z.B. die Konsequenzen der Eröffnung einer neuen Autobahn auf das Verkehrsverhalten einzelner Auto-Vielfahrer in der Umgebung. Allerdings ist diese Richtung der Einbettung in den Sozialwissenschaften eher selten, da oftmals die Frage nach Unterschieden zwischen Personen entscheidender ist.

[4] Stichprobentheoretisch bedeutet diese Einbettung, dass zunächst aus einer Grundgesamtheit von Personen einzelne Personen mit bestimmten Niveaus auf relevanten Merkmalen zufällig gezogen werden. Dann werden innerhalb von Personen Einzelbeobachtungen zufällig ermittelt. Konzeptionell wird der in einer Einzelbeobachtung festgestellte Messwert (z.B. Einkommen) einer Person i zum Zeitpunkt t dann als „Abweichung“ von dem „Gesamtniveau“ der Person begriffen.

in einem solchen Falle eine Berücksichtigung dieser Einbettung der Datenstruktur eröffnet.

Zur Klärung, ob eine Differenzierung der Datenstruktur nach Personen sinnvoll ist, sollten zuerst theoretische Annahmen herangezogen werden. Im nächsten Schritt sollte eine empirische Überprüfung erfolgen (vgl. z.B. Hsiao 2014: 14ff.). Dabei stellt sich die Frage nach dem Sinn dieser Differenzierung für jede abhängige Variable, derer Streuung mit Regressionsmodellen zu erklären angestrebt wird, neu.

In einem Regressionsmodell lässt sich diese Differenzierung formal berücksichtigen, indem eine Variation der Parameter zwischen Personen zugelassen wird. Im Vergleich zur nicht schätzbaren Maximal-Gleichung 1.1 wird damit ein Modell formuliert, welches mathematisch lösbar ist (wie, wird weiter unten deutlich):

$$y_{it} = \alpha_i + \beta_i x_{it} + \epsilon_{it} \tag{1.4}$$

2) In Gleichung 1.4 werden sowohl individuelle Regressionskonstanten a_i als auch individuelle Regressionskoeffizienten β_i der Variablen x zugelassen. Dieses recht komplexe Modell lässt sich weiter vereinfachen, indem nur eine individuelle Variation der Regressionskonstanten a_i berücksichtigt wird:[5]

$$y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it} \tag{1.5}$$

Dieses Formulierung entspricht exakt dem Modell mit variablen (bzw. genauer: mit zwischen *Individuen* variierenden)[6] Regressionskonstanten. Nur dieser Fall wird hier behandelt. Der Fall in Gl. 1.4 (in der Literatur auch als „Random Coefficient Model“ bekannt) ist als *Erweiterung* des Modells mit variablen Regressionskonstanten zu verstehen. Der interessierte Leser sei hierfür auf weiterführende Literatur verwiesen (z.B. Rabe-Hesketh et al. 2008: Kap. 4-5). Somit wird deutlich, dass Modelle mit variablen Regressionskonstanten eine Subkategorie von Modellen mit variablen Koeffizienten darstellen. Im Folgenden soll gezeigt werden, warum aus-

[5] Der umgekehrte Weg, nur einen individuelle Regressionskoeffizienten β_i zuzulassen, macht hingegen logisch gesehen in den meisten Fällen keinen Sinn. Das würde nämlich bedeuten, dass alle Personen im Falle von $x = 0$ *denselben* geschätzten $y-$Wert inne haben, der lineare Einfluss von x außerhalb von $x = 0$ aber individuell variieren darf.

[6] Im Folgenden sind zur sprachlichen Vereinfachung mit „variablen“ Regressionskonstanten immer „mit zwischen *Individuen* variierenden Regressionskonstanten“ gemeint.

gerechnet dieser Variante eine so hohe Bedeutung zukommt.

Schließlich sei noch auf den Fall verweisen, wenn beide Parameter für alle i konstant sind. Das stellt das sog. „pooled model“ dar – also ein einfaches Regressionsmodell, in dem die hierarchische Datenstruktur ignoriert wird:

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \tag{1.6}$$

Leider finden sich in der Forschungspraxis genügend Beispiele (wenn auch im Zeitverlauf immer seltener), in denen ein solches vereinfachendes Modell unreflektiert angewendet wird. Dabei ist das „pooled model“ in der Paneldatenstruktur-Logik ausschließlich dann sinnvoll, wenn sich vorher in einer statistischen Prüfung herausgestellt hat, dass die Annahme einer hierarchischen Datenstruktur, immer jeweils auf eine bestimmte abhängige Variable y bezogen, nicht haltbar ist. Formal würde das bedeuten, dass für alle i gilt $a_i = 0$. Personen würden sich damit in dem betrachteten Merkmal y nicht mehr sonderlich voneinander unterscheiden. Ein Beispiel hierfür wäre das Einkommen von Mitarbeitern in *einer bestimmten* beruflichen Position, das sehr strikt durch Tarifverträge reguliert ist. Dann würde sich das Einkommen einzelner Personen zwar von Messzeitpunkt zu Messzeitpunkt unterscheiden können (Sonderzahlungen, tariflich vereinbarte Lohnerhöhungen), *zwischen* den einzelnen Personen gäbe es aber kaum Unterschiede. Bei vielen sozialwissenschaftlichen Fragestellungen sind aber signifikante Unterschiede in der abhängigen Variablen zwischen Personen vorhanden. Deshalb ist das „pooled model“ nur selten die korrekte Wahl. Wie später gezeigt wird, führt die Ignoranz der hierarchischen Datenstruktur, wenn sie denn vorhanden ist, zur massiven Verzerrung der Koeffizienten.

1.2 Das „Intercept-Only-Modell“ (Regression ohne unabhängige Variablen)

Um sich dem Verständnis von Regressionsmodellen mit variablen Regressionskonstanten zu nähern, wird zunächst ein Modell betrachtet, in dem keine unabhängigen Variablen vorkommen. Dieses wird als „Intercept-Only-

Modell“ (IO) bezeichnet, wobei mit Intercept hier die *variable* Regressionskonstante α_i gemeint ist. Gleichung 1.5 vereinfacht sich dann zu:

$$y_{it} = \alpha_i + \epsilon_{it} \tag{1.7}$$

Dabei ist es gängiger, die individuelle Regressionskonstante α_i aufzusplitten in eine allgemeine Regressionskonstanten α und die individuelle mittlere Abweichung a_i von ihr. Die individuelle Regressionskonstante ergibt sich dann aus $\alpha_i = \alpha + a_i$. Beide Formulierungen sind identisch, letzteres ermöglicht aber ein besseres Verständnis des Konzepts. Gl. 1.7 wird daher umformuliert zu:

$$y_{it} = \alpha + a_i + \epsilon_{it} \tag{1.8}$$

Was bedeutet ein Regressionsmodell ohne unabhängige Variablen? Das lässt sich am besten an einer einfachen Querschnittsregression verdeutlichen. Sind alle Personen nur einmalig befragt worden, dann entfällt der Zeitindex t. Eine individuelle Regressionskonstante α_i wäre auch statistisch nicht bestimmbar, da pro Person nur ein Wert vorliegt. Im Querschnitt vereinfacht sich daher Gl. 1.8 zu:

$$y_i = \alpha + \epsilon_i \tag{1.9}$$

Da es in diesem Modell keine unabhängigen Variablen gibt, die eine Schätzung von y bedingen könnten, ist die unbedingte Schätzung für alle Personen konstant und lautet α. Und welcher Wert ist der beste Schätzwert, wenn außer y sonst keine Informationen (unabhängige Variablen) vorhanden sind? Es ist das arithmetische Mittel von y – formal dargestellt als $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Das bedeutet für die Querschnittsregression ohne unabhängige Variablen, dass die Regressionskonstante dem arithmetischen Mittel entspricht: $\alpha = \overline{y}$. Das

Residuum ϵ_i ist somit die Abweichung einer Person i vom arithmetischen Mittel. Ein Wert von y lässt sich also zerlegen in den Mittelwert und die Abweichung vom Mittelwert:

$$y_i = \overline{y} + \epsilon_i \tag{1.10}$$

Wird Gl. 1.10 als Schätzung für y, also $\widehat{y}$, dargestellt, dann lässt sich schreiben:[7]

$$\widehat{y}_i = \alpha = \overline{y} \tag{1.11}$$

Übertragen auf das panelanalytische Intercept-only-Modell (Gl. 1.8) lässt sich ein y-Wert in *drei* Komponenten zerlegen: Dem allgemeinen Mittelwert $\overline{y} = \alpha$, der mittleren individuellen Abweichung von diesem allgemeinen Mittelwert a_i und der Abweichung der Einzelmessung von $\alpha + a_i$. Was aber bedeutet eine „mittlere individuelle Abweichung“? Wenn a_i die mittlere individuelle Abweichung von dem allgemeinen Mittelwert ist, dann stellt $\alpha + a_i$ bzw. $\overline{y} + a_i$ umgekehrt den *individuellen Mittelwert* $\overline{y}_i$ dar. Dieser ist für eine Person i definiert als: $\overline{y}_i = \overline{y} + a_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$. Gl. 1.8 lässt sich somit auch schreiben als:

$$\begin{aligned} y_{it} &= \overline{y} + a_i + \epsilon_{it} \\ &= \overline{y}_i + \epsilon_{it} \end{aligned} \tag{1.12}$$

Die mittlere individuelle Abweichung a_i ist folglich die Differenz zwischen allgemeinem und individuellem Mittelwert. Wenn der individuelle Mittelwert als das persönliche „absolute Niveau“ einer Person hinsichtlich y begriffen wird, dann drückt a_i dieses Niveau in Relation zum allgemeinen Durchschnitt über alle Personen dar. Ist a_i positiv, dann befindet sich die Person hinsichtlich y auf einem überdurchschnittlichem Niveau – und umgekehrt.

[7] Eine Erweiterung dieser Gleichung 1.11 um eine unabhängige Variable x ist folgerichtig der Mittelwert von y gegeben x. Dieser durch x bedingte geschätzte Mittelwert von y lautet dann: $\alpha + bx_i$.

		IO							OIO				
i	t	y_{it}	=	α	+	a_i	+	ϵ_{it}	y_{it}	=	α	+	ϵ_{it}
1	1	58,8	=	47,5	+	15,15	+	-3,85	58,8	=	47,5	+	11,3
1	2	59,6	=	47,5	+	15,15	+	-3,05	59,6	=	47,5	+	12,1
1	3	65,2	=	47,5	+	15,15	+	2,55	65,2	=	47,5	+	17,7
1	4	67	=	47,5	+	15,15	+	4,35	67	=	47,5	+	19,5
2	1	31,7	=	47,5	+	-15,15	+	-0,65	31,7	=	47,5	+	-15,8
2	2	32,1	=	47,5	+	-15,15	+	-0,25	32,1	=	47,5	+	-15,4
2	3	32	=	47,5	+	-15,15	+	-0,35	32	=	47,5	+	-15,5
2	4	33,6	=	47,5	+	-15,15	+	1,25	33,6	=	47,5	+	-13,9

Tabelle 1.2: Zerlegung abhängige Variable IO vs. OIO

Die Zerlegung von y soll nun verdeutlicht werden, indem dem Intercept-only-Modell (IO; Gl. 1.8) die Variante ohne individuelle Regressionskonstante (OIO: $y_{it} = \alpha + \epsilon_{it}$) gegenüber gestellt wird. Dies geschieht am Beispiel eines fiktiven Produktivitätsscores (in %) zweier Mitarbeiter einer Firma zu je vier Zeitpunkten. Person $i = 1$ hat einen individuellen Mittelwert von $\overline{y}_1 = 62,65$, Person $i = 2$ hat einen individuellen Mittelwert von $\overline{y}_2 = 32,35$. Der globale Mittelwert über beide Personen liegt bei $\overline{y} = \alpha = 47,5$. Entsprechend lautet die individuelle Abweichung für die erste Person: $a_1 = 62,65 - 47,5 = 15,15$; bei der zweiten Person: $a_2 = 32,35 - 47,5 = -15,15$. Die Zerlegung der einzelnen $y-$Werte gestaltet sich nun in beiden Modellvarianten entsprechend Abb. 1.2.

Abb. 1.2 verdeutlicht, dass eine Zerlegung von y_{it}, welche die hierarchische Datenstruktur berücksichtigt (IO), deutlich kleinere Restfehler ϵ_{it} im Vergleich zu OIO produziert. Während ϵ_{it} in OIO die Abweichung eines Produktivitätswertes einer Person zu einem bestimmten Zeitpunkt vom *allgemeinen* Mittelwert ausdrückt, steht ϵ_{it} in IO lediglich für die Abweichung eines Produktivitätswertes einer Person zu einem bestimmten Zeitpunkt von *ihrem persönlichen* Mittelwert. Ein Teil der Fehlervarianz aus OIO lässt sich somit

durch unterschiedliche Niveaus der Personen „erklären“.[8] Und es ist schließlich das Ziel von Regressionsmodellen, Restfehler zu minimieren. In dieser Hinsicht ist im oberen Beispiel das die hierarchische Datenstruktur adäquat berücksichtigende Modell IO dem diese Struktur ignorierenden Variante OIO vorzuziehen.

Allgemein ist IO dem Modell OIO dann überlegen, wenn die mittleren Abweichungen a_i vom allgemeinen Mittelwert a statistisch bedeutsam (im Schnitt signifikant von Null verschieden) sind. Diese Frage (formal dargestellt als $a_i = 0$ für alle i) ist identisch mit der Frage, ob sich die individuellen Mittelwerte $\overline{y}_i = \alpha + a_i$ signifikant voneinander unterscheiden. Nur wenn das der Fall ist, dann liegt eine hierarchische Datenstruktur hinsichtlich y vor. Wenn alle Personen dieselbe durchschnittliche Produktivität inne hätten, dann wäre jeder individuelle Mittelwert gleich dem allgemeinen Mittelwert. Alle mittleren Abweichungen a_i würden dann Null betragen. Eine Aufspaltung, wie sie in IO erfolgt, würde gegenüber OIO keine zusätzliche Minimierung der Fehlervarianz herbeiführen.

Um zu prüfen, ob $a_i = 0$ für alle i ist, wird die Relation zwischen den zwei oben eingeführten Varianzen betrachtet: Die Varianz zwischen den Personen ($V(a_i)$) (between variation) in Relation zur Varianz innerhalb von Personen ($V(\epsilon_{it})$) (within variation) setzt. Wäre im einen Extremfall $V(a_i) = 0$, dann gäbe es keine Streuung zwischen den Personen. Das würde zur oben beschriebenen Situation führen: Alle Personen-Mittelwerte $\overline{y}_i$ wären gleich, würden damit dem allgemeinen Mittelwert $\overline{y}$ entsprechen. Alle a_i würden folglich Null betragen. Im anderen Extremfall $V(\epsilon_{it}) = 0$ gäbe es nur Unterschiede zwischen Personen. Der Wert einer einzelnen Person wäre zeitkonstant (wie z.B. bei 99,9% aller Personen das Geschlecht).[9]

Das IO-Modell erfüllt damit hier nicht nur die rein didaktische Funktion zur Verbesserung des Verständnisses der hier vorgestellten Modelle. Es sollte grundsätzlich den Beginn jeder Regressionsanalyse mit Paneldaten markieren, um die Entscheidung zu erleichtern, ob die darauf aufbauenden Modelle *mit echten unabhängigen Variablen* eine Regressionstechnik mit variablen Regressionskonstanten benötigen. Würde sich herausstellen, dass das OIO-Modell völlig ausreicht, dann könnten weiterführende Modelle mit derselben

[8] Diese Aussage muss eingeklammert werden, da sie streng genommen nur für sog. Fixed-Effects-Modelle gilt, welche nur eine Untergruppe der hier betrachteten Modelle darstellen. Die weiteren Ausführungen bringen allerdings Klarheit in dieser Hinsicht.

[9] Die Grundlagen zum genaueren Verständnis dieser Relation werden in Abschnitt 2 behandelt.

abhängigen Variablen auf die Berücksichtigung der hierarchischen Datenstruktur entsprechend Gl. 1.6 verzichten.

Es sollte verdeutlicht werden, dass die Einführung von a_i in die Regressionsgleichung ohne Kovariate (im Gegensatz zu der Form: $y_{it} = \alpha + \epsilon_{it}$) die Berücksichtigung der hierarchischen Datenstruktur von Paneldaten (Einzelbeobachtungen genestet in Personen) formal realisiert. Wenn dieses Prinzip einmal begriffen wurde, dann ist der Weg zu Modellen *mit* unabhängigen Variablen nicht mehr weit. Die Zerlegung eines y-Wertes funktioniert analog, nur dann unter der Bedingung der x-Variablen. Wenn nur eine unabhängige Variable x vorliegt, dann wird Gl. 1.8 erweitert zu:

$$y_{it} = \alpha + \beta x_{it} + a_i + \epsilon_{it} \tag{1.13}$$

In diesem Falle stellt a_i nicht mehr die für Personen gemittelte Abweichung vom *unbedingten* Mittelwert von y dar, sondern von dem geschätzten Mittelwert *gegeben* x – und das ist: $\alpha + \beta x_{it}$. a_i ist also die gemittelte Abweichung von y, wenn x aus y herausgerechnet wird. Was genau damit gemeint ist und welche Vorteile und Probleme eine solche Betrachtung mit sich bringt, wird in Abschnitt 2 deutlich.

1.3 Parallelen zur Mehrebenenanalyse

Zum besseren Verständnis der Grundidee von Modellen mit variablen Regressionskostanten ist es von Vorteil, eine Parallele zur Mehrebenenanalyse zu ziehen. Denn die Idee, eine variable Regressionskonstante in eine Regressionsgleichung einzubauen, ist nicht speziell im Kontext der Panelanalyse entstanden. Vielmehr ist diese Möglichkeit formal immer dann gegeben, wenn sich die Fälle eines Datensatzes in überschneidungsfreie Gruppen aufteilen lassen – und das gilt auch für Querschnittsdaten. *Inhaltlich sinnvoll* ist die Einführung einer variablen Regressionskostanten, wenn erstens vermutet wird, dass die Gruppenzugehörigkeit einen Einfluss auf eine Variable y hat.

Zweitens sollte die Gruppenzugehörigkeit eine eigenständige Ebene in der Datenstruktur darstellen, so dass eine hierarchische Datenstruktur vorliegt.

Ob eine Gruppenzugehörigkeit tatsächlich eine eigenständige Datenebene aufspannt, ist eine theoretische Frage. Hat die Gruppenzugehörigkeit in irgendeiner Form eine integrative Funktion für die Individuen (vgl. Langer 2010: 771)? Ein gutes Anzeichen für eine hierarchische Datenstruktur ist, dass neben Merkmalen der Gruppenmitglieder auch eigenständige Merkmale auf Gruppenebene vorliegen. Ein zutreffendes Beispiel für solche Gruppierungen sind einzelne Schulen als Gruppen, deren Schüler die Gruppenmitglieder darstellen. Die Schulen stellen für Schüler einen integrativen Kontext dar. Außerdem lassen sich leicht sowohl Merkmale auf der Schülerebene (z.B. Schulnoten) als auch auf der Schulebene (z.B. der Führungsstil des Direktors) identifizieren. Antworten hingegen Befragte auf eine Frage mit einer Ganzzahl zwischen 1 und 5, dann macht es weniger Sinn, von einer oberen „Antworten"-Ebene mit fünf „Antwortgruppen" zu sprechen. Die Beantwortung der Frage mag zwar mit einer betrachteten abhängigen Variablen zusammenhängen. Die integrative Funktion der Zugehörigkeit zu einer „Antwortgruppe" darf allerdings angezweifelt werden – so lassen sich auch schwer sinnvolle Merkmale denken, welche nur auf dieser Gruppenebene variieren.

Liegt nun eine hierarchische Datenstruktur mit einer relevanten Gruppierungsvariablen vor, dann müssen statistische Verfahren genauso diese Differenzierung berücksichtigen, wie bei vorliegenden Paneldaten. Außerhalb der Panelanalyse spricht man dann von *mehrebenenanalytischen* Verfahren. Modelle mit variablen Regressionskonstanten sind dabei ein wichtiger Bestandteil der modernen Mehrebenenanalyse.

Somit ist die bei Paneldaten mögliche konzeptionelle Einbettung einer Messung zu einem bestimmten Zeitpunkt in eine Person ein Spezialfall der generellen hierarchischen Einbettung von Objekten in Gruppen: Die „Gruppen" sind die einzelnen Personen i und die „Mitglieder einer Gruppe i" sind die Werte der Person i im Zeitverlauf (it).

Ein Beispiel für eine hierarchische Datenstruktur außerhalb von Paneldaten wäre die Einbettung von Basketballspielern in Basketballmannschaften. Ein relevantes Merkmal y_{it} könnte die Körpergröße von Basketballspielern sein. Hätte z.B. in der vierten betrachteten Basketballmannschaft ($i = 4$) der

fünfte Spieler ($t = 5$) eine Körpergröße von 212 cm inne, dann ließe sich dieser Wert formal darstellen als: $y_{4,5} = 212$.[10] Analog zur Darstellung in Tab. 1.2 könnte die Körpergröße y_{it} eines einzelnen Spielers zerlegt werden in die Durchschnittsgröße aller Spieler über alle Mannschaften hinweg α, der Abweichung a_i des Mannschaftsmittelwertes α_i der Mannschaft, in der der betrachtete Spieler spielt, vom allgemeinen Mittelwert α und schließlich der Abweichung des Spielers ϵ_{it} vom Mannschaftsmittelwert α_i. Entsprechend würde auch in diesem Falle ϵ_{it} im IO-Modell im Vergleich zum OIO verringert werden, wenn sich die Mannschaftsmittelwerte signifikant voneinander unterscheiden ($V(a_i) \neq 0$). Das würde z.B. dann zutreffen, wenn die Mannschaften um die Spieler mit der höchsten Körpergröße konkurrieren und sich einige Mannschaften in der Hinsicht besser durchsetzen, als andere. Die auf dem Transfermarkt erfolgreicheren Mannschaften hätten dann tendenziell eine höhere Durchschnittsgröße.

Hierarchische Datenstrukturen sind also auch jenseits des Panelkontextes vorhanden. Diese allgemeinere Sicht hilft, das Prinzip der Einbettung von Einzelmessungen in Personen besser nachzuvollziehen. Aus dieser Sicht wird auch verständlich, warum z.B. in Stata gar nicht zwischen Verfahren der Mehrebenen- und der Panelanalyse unterschieden wird. Auch die nicht nur in Stata gängige und u.U. anfangs befremdliche Bezeichnung einer Person als „Gruppe“ oder „Cluster“ wird dadurch nachvollziehbar.

Die Parallele zur Mehrebenenanalyse hilft aber auch, zu verstehen, warum individuelle Mittelwerte $\overline{y}_i$ (bzw. individuelle Regressionskonstanten α_i; s. Abs. 1.2) als „Stellvertreter“ für alle zeitkonstanten Merkmale einer Person fungieren: Mal angenommen, es nehmen Schüler von drei Schulen an einem Leistungstest (Leistungswert in %) teil. Die Leistung der Schüler der ersten Schule bilden einen Schul-Mittelwert von $\overline{y}_1 = \alpha_1 = 78\%$. Analog gilt für Schule zwei und drei: $\overline{y}_2 = \alpha_2 = 37\%$; $\overline{y}_3 = \alpha_3 = 55\%$. Schon nach Augenmaß existieren deutliche Unterschiede im „Leistungsniveau“ zwischen den Schulen. Gehen wir weiter davon aus, dass sich die Leistungen zweier Schüler ein und derselben Schule in den meisten Fällen nicht stark voneinander unterscheiden. Die „within variation“ fällt also relativ niedrig aus. Im

[10] Die Setzung des Kommas zwischen den beiden Indexwerten beugt lediglich der potentiellen Gefahr vor, die beiden Werte fälschlich als einen Zahlenwert „54“ zu lesen.

varianzanalytischen Sinne würde die Schule, in der der Schüler eingebettet ist, dann einen relativen hohen Anteil der Varianz von y erklären. Ist also für einen Schüler bekannt, zur welcher der drei Schulen er geht, dann kann man eine deutlich bessere Schätzung seiner Leistung abgeben, als wenn nur der allgemeine Mittelwert $\overline{y}$ bekannt wäre. Dies liegt an der Dominanz der „between variation“ (große Mittelwertsunterschiede) in Relation zur „within variation“.

Wie aber kommen diese Mittelwertsunterschiede selbst zustande? Anscheinend existieren Merkmale auf Schulebene (!), welche diese Unterschiede ausmachen. Möglicherweise ist das Lehrpersonal an Schule 1 besser ausgebildet als an Schule 2 und 3. Vielleicht hat der Direktor an Schule 2 einen zu liberalen Führungsstil, so dass Lerninhalte seitens der Lehrer nicht konsequent im Unterricht umgesetzt werden. Es kann auch sein, dass Schule 1 bei der Aufnahme der Schüler stärker als die beiden anderen Schulen nach Merkmalen selektiert, die mit der zukünftigen Leistung korrelieren. Es ließen sich sicherlich ad hoc weitere potentielle Erklärungsfaktoren benennen. Entscheidend ist aber, dass die Mittelwertsunterschiede all diese relevanten Merkmale auf Schulebene „einfangen“. Da Merkmale auf Schulebene naturgemäß nicht zwischen Schülern ein und derselben Schule variieren können, schlagen sie sich in einer Maßzahl nieder, welche ebenfalls nicht zwischen Schülern ein und derselben Schule sondern nur **zwischen Schulen** variieren kann. Das genau ist der Schulmittelwert.

Bezogen auf die Panelsituation absorbiert ein individueller Mittelwert $\overline{y}_i$ (bzw. die individuelle Regressionskonstante α_i) alle Merkmale, welche für die Person zeitkonstant sind und konstant auf y wirken. Ist y das Einkommen, dann könnten dies Merkmale wie das Geschlecht, die Bildung, die Intelligenz etc. sein. Während das Geschlecht und die Bildung häufig als Variablen vorhanden sind, gilt dies z.B. für die Intelligenz oftmals nicht. Über individuelle Mittelwertsunterschiede könnten aber solche nicht vorhandenen Merkmale implizit eben doch berücksichtigt werden, was das Erklärungspotential eines solchen Modells deutlich vergrößert (und daher ϵ_{it} verringert).

1.4 Zwischenfazit

Es sollte gezeigt werden, dass die Struktur von Paneldaten einer im Vergleich zur Querschnittsregression differenzierteren statistischen Modellierung bedarf. Hierzu wurde zunächst die Zerlegbarkeit der Varianz einer Variablen im Falle von Paneldaten vorgestellt. Es wurde gezeigt, wie diese Zerlegung in der Notation von Regressionsmodellen berücksichtigt werden kann. Es wurde ferner argumentiert, dass im Vergleich zur Querschnittsregression auch die Regressionsparameter über die Personen und Zeitpunkte variieren können. Allerdings ist nicht jede Konstellation mathematisch identifizierbar und logisch sinnvoll. Modelle mit über Individuen variierenden Regressionskonstanten gehören zu den Konstellationen, die für die Analyse von Paneldaten von herausragender Bedeutung sind. Um sich dem Grundgedanken dieser Modelle zu nähern, wurde zunächst eine Regression ohne unabhängige Variablen betrachtet. Daraufhin wurde eine Parallele zur Mehrebenenanalyse gezogen, um zu zeigen, dass die Anwendbarkeit variabler Regressionskonstanten nicht nur auf die durch Personen und Zeitpunkte aufgespannte hierarchische Datenstruktur beschränkt ist. Sie eignet sich immer dann, wenn Einzelmessungen in einen diese Messungen beeinflussenden Kontext eingebettet sind.

Bislang waren die Ausführungen an mehreren Stellen bewusst noch etwas unkonkret, um auf möglichst einfachem Wege das Grundprinzip von Modellen mit variablen Regressionskonstanten zu vermitteln. Die fehlende Konkretheit bezieht sich vor allem auf die schwammige Aussage, dass die über Individuen variierende Regressionskonstante die hierarchische Datenstruktur (Einbettung der Einzelmesswerte in Personen) „berücksichtigt“. Was genau unter „Berücksichtigung“ zu verstehen ist, gilt es im nächsten Kapitel 2 zu präzisieren. Es wird gezeigt, dass es in der Panelanalyse zwei etablierte Arten gibt, die eine unterschiedliche „Berücksichtigung“ vornehmen. Sie hängt davon ab, ob a_i als ein zu schätzender Parameter (fix) oder als eine Zufallsvariable (random) aufgefasst wird. Es wird deutlich werden, dass Vorteile und Probleme von Modellen mit variablen Regressionskonstanten, je nach Modellvariante (fix oder random), unterschiedlich ausgeprägt sind. Schon an dieser Stelle sollen dabei zwei modellvariante-abhängige große Vorteile genannt werden: Erstens lassen sich alle zeitkonstante Merkmale von Personen aus den Regressionsgleichungen herausrechnen, auch wenn sie nicht als

unabhängige Variablen im Modell auftreten. Zweitens wird eine Eigenheit hierarchisch beschaffener Daten statistisch korrekter erfasst. Diese besteht darin, dass zwei (oder mehrere) Messwerte ein und derselben Person einander ähnlicher sein können, als zwei Messwerte von zwei verschiedenen Personen. Dieses Phänomen ist bekannt als sog. „personeninterne Autokorrelation von Residuen“. Die hinsichtlich dieser Vorteile unterscheidbaren „fixed-“ und „random-effects“-Modelle werden im Kapitel 2 vorgestellt.

Kapitel 2

Zwei Varianten der Modelle mit variablen Regressionskostanten: Fixed- (FEM) und Random-Effects-Modelle (REM)

Entscheidend für die konkrete Realisierung von Modellen mit variablen Regressionskostanten ist der Umgang mit den individuellen Abweichung von der Regressionskonstanten, a_i – je nach dem, ob diese als fixe oder zufällige Variablen gesehen werden. Doch was bedeutet diese Unterscheidung?

2.1 Der Unterschied zwischen fixen und zufälligen Variablen

Die Frage nach dem Unterschied zwischen fixen und zufälligen variablen Regressionskonstanten verweist auf eine generellen Sachverhalt in Regressionsmodellen: Einige Variablentypen werden als zufällig[1], andere als fix betrachtet. Um diese gedanklich-konzeptionelle Unterscheidung zu verstehen, muss die inferenzstatistische Grundidee deutlich werden, dass Eigenschaften von Merkmalsträgern und Zusammenhänge zwischen diesen Eigenschaften bereits *vor* einer Datenerhebung bzw. einer Datenanalyse existieren.

[1] Der Begriff „Zufallsvariable“ ist hier synonym mit dem Begriff „zufällige Variable“.

Bevor also überhaupt eine Stichprobe gezogen wird, Daten erhoben werden und eine Analyse durchgeführt wird, existiert eine Grundgesamtheit G. Ferner wird angenommen, in dieser Grundgesamtheit G gilt eine wahre (aber uns unbekannte) und unter allen gängigen Annahmen zu Regressionsmodellen korrekt spezifizierte Regressionsgleichung, welche den Einfluss zwischen einer abhängigen Variablen x und einer abhängigen Variablen y quantifiziert (zur Vereinfachung werden hier die Indizes weggelassen):

$$y = a + bx + \epsilon \tag{2.1}$$

mit

$a =$ Regressionskonstante

$b =$ Regressionsparameter zur Variablen x

$\epsilon =$ Residuum.

Nun wird aus G eine Zufallsstichprobe mit n Elementen gezogen und die Datenerhebung durchgeführt. Es wird (vorübergehend!) angenommen, dass die Werte von y mithilfe eines experimentellen Designs erhoben werden. Im experimentellen Design kann nämlich der Reiz, also die Ausprägung der x-Variablen kontrolliert gesetzt werden. Auch wenn unabhängige Variablen einer linearen Regression als metrisch angenommen werden, so wird hier (auch wieder vorübergehend!) vereinfachend x als eine binäre Variable deklariert, so dass nur zu unterscheiden ist, ob ein Reiz gesetzt wurde ($x = 1$) oder nicht ($x = 0$). In einem Experiment wird ja sozusagen die „Realität in der Grundgesamtheit“ simuliert. Somit sind bereits vor dem Experiment bzw. unabhängig von dessen Durchführung einige Elemente der Grundgesamtheit G mit einem Reiz versehen ($x = 1$, z.B. die Einnahme eines Medikamentes) und andere nicht ($x = 0$, das Medikament wird nicht eingenommen). Anhand dieser Unterscheidung lässt sich G als eine in Subpopulationen zerteile bzw. geschichtete Gesamtheit verstehen. In dem einfachen Fall hier teilt sich G folglich in zwei Gruppen G_1 und G_2 auf, entsprechend der Unterscheidung zwischen $x = 1$ (G_1) und $x = 0$ (G_2). Diese Situation lässt sich problemlos auf eine multiple Regression mit mehreren (metrischen) unabhängigen Variablen erweitern: Demnach wird G gedanklich in so viele Schichten geteilt,

wie Merkmalskombinationen der x-Variablen existieren.

In diesem Verständnis entspricht die zufällige Auswahl einer Person für ein Experiment und ihre Zuordnung in die Experimentalgruppe ($x = 1$) dem Prozess der Ziehung einer Person aus der Subpopulation der Personen, die dem Reiz ($x = 1$) ausgesetzt sind – analog dazu ist die Stichprobenziehung im Falle $x = 0$ zu verstehen. Nun wird eine aus der Population $x = 1$ gezogene Person einem Reiz ausgesetzt und reagiert auf diesen Reiz, produziert also *scheinbar* einen y-Wert. Doch diese Auffassung muss korrigiert werden, wenn angenommen wird, dass in G bereits vor dieser Untersuchung ein fester, wahrer Einfluss von x auf y besteht. Die Gleichung $y = a + bx + \epsilon$ quantifiziert nämlich bereits ***vor*** der Durchführung des Experiments den bestehenden linearen Zusammenhang. b ist also bereits vorhanden (auch wenn uns unbekannt).

Ferner ist es aufgrund der Komplexität der meisten Zusammenhangsstrukturen selten realistisch anzunehmen, dass y immer perfekt durch den Term $a+bx$ erzeugt wird. Daher ist in der Gleichung ein Störterm ϵ enthalten, welcher *zufällige* Abweichungen von der perfekten, aber unrealistischen linearen Zusammenhangsstruktur $y = a+bx$ „einfängt“. Da x im experimentellen Design für eine Messung gesetzt wird und mit $a+bx$ der feste Einfluss von x auf y charakterisiert wird, fängt die Messung von y letztlich die Abweichung von diesem idealen linearen Zusammenhang ein. Es wird demnach, gegeben dem x-Wert bzw. der Subpopulationszugehörigkeit, **das Residuum „gemessen“ bzw. „erfasst“!** Das so erfasste Residuum erzeugt (unter der Bedingung von x) durch die Addition mit $a + bx$ den y-Wert. y ist somit gedanklich als eine lineare Transformation von dem Residuum zu sehen. Aus diesem erweiterten Blickwinkel sollte nun die vertraute Gleichung $y = a+bx+\epsilon$ gelesen werden.

Das Residuum ϵ umfasst die Summe von (z.T. unkalkulierbaren) Einflüssen, welche neben der festen Wirkung von x einen Einfluss auf die Messung von y haben. Da oben angenommen wird, dass der Einfluss von x auf y durch die lineare Gleichung $y = a + bx + \epsilon$ nicht von Annahmeverletzung betroffen ist und folglich korrekt spezifiziert ist, weist das Residuum *keine Systematik* auf. Da zusätzlich die Probanden *zufällig* aus den Subpopulationen G_1 und G_2 gezogen wurden, ist das Zustandekommen der Residualwerte innerhalb der Subpopulationen als ausschließlich zufallsbedingt zu sehen. Daher wird das Residuum als eine Zufallsvariable, gegeben x, verstanden. Da es sich bei den y-Werten lediglich um lineare Transformationen der Residualwerte handelt,

ist folglich auch y ***als eine Zufallsvariable aufzufassen*** (von Auer 2007: 68f.). Damit ist eine klare analytische Unterscheidung zwischen der unabhängigen Variablen x und der abhängigen Variablen y zu treffen: Erstere ist eine nicht-stochastisch fixe Variable, letztere ist eine Zufallsvariable.

Warum ist diese Unterscheidung wichtig? Ausgehend von der Deklaration der Residuen als Zufallsvariablen lassen sich einige Annahmen über Regressionsmodelle formulieren; die Erfüllung bzw. Verletzung dieser Annahmen ist bedeutend für die Einschätzung, ob ein Modell oder Teile des Modells korrekt spezifiziert sind. Die Annahmen über das Residuum als Zufallsvariable, zusammen mit der Annahme des nicht-stochastischen Charakters von x, erlaubt den Nachweis, dass es sich bei den Schätzern nach dem „Kleinste-Quadrate-Prinzip" (KQ-Prinzip) um BLUE-Schätzer[2] handelt (vgl. von Auer 2007: 83, 430). Ohne ins Detail zu gehen sei kurz erwähnt, dass dieser Nachweis deshalb gelingt, weil die Eigenschaft der Nicht-Zufälligkeit von x u.a. an einer bestimmten Stelle eine entscheidende mathematische Umformung erlaubt (vgl. von Auer 2007: 83, 430).[3]

Nun basieren aber sozialwissenschaftliche Studien oftmals nicht auf dem experimentellen Design, sondern entstammen einem Ex-Post-Facto-Design, wie z.B. einer Befragung. Da in einer Befragung die x-Werte nicht als Reize manipuliert werden können, müssen sie streng genommen ebenfalls als stochastische Zufallsvariablen angesehen werden – unter der Annahme, dass der Pool der Befragten durch die Realisation einer Zufallsstichprobe zustande kam. Es lässt sich aber mathematisch nachweisen, dass mit zunehmendem Stichprobenumfang n ($n \to \infty$) die Schätzer eines Regressionsmodells nach der KQ-Methode dennoch die BLUE-Eigenschaft, zumindest asymptotisch besitzen. Da dieses Grundkonzept also im Falle von stochastischen unabhängigen Variablen nicht in sich zusammenbricht, gleichzeitig aber gerade auf der Prämisse von fixen x-Variablen aufbaut, kann weiter konzeptionell zwischen festen unabhängigen Variablen x und der Zufallsvariablen y unterschieden werden – auch wenn Befragungsdaten vorliegen.

Eine Konsequenz aus der Unterscheidung zwischen fixen und zufälligen Va-

[2] BLUE steht für den besten (=effizientesten) Schätzer aus der Gruppe der unverzerrten linearen Schätzer (vgl. allgemein zu den Voraussetzungen der BLUE-Eigenschaft von Auer 2007: 74ff).

[3] Denn es gilt für eine nicht zufällige, fixe Größe x, dass ihr Erwartungswert $E(x) = x$ ist. Diese Vereinfachung gegenüber den Erwartungswerten von Zufallsvariablen ist für die angesprochene mathematische Beweisführung entscheidend.

riablen im Kontext von Regressionsmodellen ist die unterschiedliche Art ihrer Identifizierbarkeit. Fixe Variablen sind gegeben (exogen). Bei fixen Variablen werden *Parameter geschätzt*, mit denen sie sozusagen „von Außen" *systematisch* die abhängige Variable beeinflussen. Zufallsvariablen sind hingegen nicht direkt bestimmbar, da sie eine Vielzahl von idiosynkratischen Einflüssen in sich vereinen und *keinerlei Systematik* unterliegen. Es lassen sich einzig Eingenschaften *ihrer Wahrscheinlichkeitsverteilung* schätzen. Dazu gehört in der Regressionsanalyse insbesondere die Schätzung *der Varianz* des Residuums (welcher zu den Zufallsvariablen gehört). Diese Schätzung ist ein entscheidender Bestandteil bei der Beurteilung, wie gut die unabhängigen Variablen die Gesamtvarianz der abhängigen Variablen aufklären.

Wichtig ist also für die folgenden Ausführungen festzuhalten: Während bei fixen Variablen ihr systematischer Einfluss als Parameter geschätzt wird, lassen sich bei Zufallsvariablen aufgrund der fehlenden Systematik nur Eigenschaften der dahinterliegenden Wahrscheinlichkeitsverteilung schätzen.

2.2 Spezifikation und Koeffizientenschätzung im FEM und REM

Das grundlegende Verständnis des Unterschieds zwischen fixen und zufälligen Variablen soll helfen, die Unterscheidung zwischen random- und fixed-effects-Modellen zu verstehen. Diese bezieht sich auf die Behandlung von a_i.

a_i stellt die individuelle Abweichung von der allgemeinen Regressionskostanten dar. Da dieser Wert für eine einzelne Person konstant ist, nicht über die Zeit variieren kann, umfasst er stellvertretend alle Merkmale, welche für diese Person konstant sind. Etwas vereinfachend gesagt stellt a_i den Einfluss „der Person" selbst auf ihre Einzelwerte dar. Nun kann man a_i, also den „Stellvertreter einer Person", als Koeffizient einer fixen Variablen (der Person) oder als eine Zufallsvariable deklarieren. Im ersten Fall werden die a_i-Werte zu den Koeffizienten unabhängiger Variablen (Personen) gezählt und explizit in die Parameterschätzung involviert. Im letzteren Fall werden sie als eine Komponente des Residuums aufgefasst. Da das Residuum eine Zufallsvariable darstellt, zählt a_i als eine seiner Komponenten auch zu dem Lager der Zufallsvariablen.

Für die folgenden Ausführungen werden vier Vereinfachungen / Einschränkungen vorgenommen, welche aber die Grundidee nicht verändern. Sie vereinfachen aber die formale Darstellung an einigen Stellen:

Erstens beschränken wir uns auf den einfachen Fall, wenn nur eine einzige unabhängige Variable spezifiziert wird. Für eine Erweiterung auf ein multiples Regressionsmodell gilt eine analoge Umsetzung. Allerdings müssen die mathematischen Transformationen mithilfe von Matrizen dargestellt werden. Ausführliche mathematische Darstellungen hierzu finden sich bei Hsiao (2014).

Zweitens arbeiten wir fortan mit zentrierten Variablen. Eine Variable y aus einem Paneldatensatz wird zentriert, indem jeder Wert y_{it} durch seine Differenz vom arithmetischen Mittel $\overline{y} = \frac{1}{n \cdot T} \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it}$ ersetzt wird. Das arithmetische Mittel der zentrierten Variablen liegt dann bei Null. Die absoluten Differenzen zwischen einzelnen y-Werten und folglich die Varianz der zentrierten Variablen bleibt aber (im Gegensatz zur z-Transformation) unberührt. Deshalb sind in Regressionsmodellen mit zentrierten un- und abhängigen Variablen Schätzungen für Regressionskoeffizienten identisch mit denen, wenn keine Zentrierung vorgenommen wurde. Allerdings nimmt die allgemeine Regressionskonstante α den Wert Null an (da nun die Mittelwerte aller Variablen Null sind und daher der geschätzte y-Wert im Falle, wenn alle x-Werte Null sind, ebenfalls Null beträgt). Sie braucht somit nicht mehr in der Darstellung von Regressionsmodellen aufgeführt werden. So vereinfacht sich an manchen Stellen die formale Darstellung, ohne dass die Regressionskoeffizienten dadurch manipuliert werden.

Drittens gehen wir von einem balancierten Paneldatensatz aus. Das bedeutet, dass für jede Person i dieselbe Anzahl an Beobachtungen vorliegt. Statt einer individuell variierenden Anzahl an Beobachtungen T_i können wir daher einfach T schreiben. Die Anzahl aller Messwerte einer Variablen ergibt sich daher simpel aus $n \cdot T$ statt aus $\sum_{i=1}^{n} T_i$. Auch diese Einschränkung vereinfacht stellenweise die Darstellung.

Viertens beschränken wir uns auf metrisch skalierte abhängige Variablen und unterstellen einen linearen Einfluss der unabhängigen Variablen x auf die abhängige Variable. Daher nehmen die vorgestellten Modelle die Form einer linearen Regression an. Für analoge Umsetzungen bei nicht-metrischen abhängigen Variablen (z.B. logistische Regression mit Paneldaten) s. Giesselmann & Windzio (2012: Kap. 7) oder ausführlicher bei Rabe-Hesketh, S. & Skrondal, A. (2012b).

Aus den Einschränkungen resultiert die folgende Ausgangsgleichung (entsprechend der Symbol-Definitionen zur Gl. 1.1):

$$y_{it} = \beta \cdot x_{it} + a_i + \epsilon_{it} \tag{2.2}$$

Die Einführung von a_i in 2.2 ist in erster Linie zur Verringerung und Bereinigung von ϵ_{it} gedacht. Denn a_i fängt sozusagen alle (im Regressionsmodell nicht spezifizierten) zeitkonstanten Einflüsse ab, die eine Person auf y ausübt – ehe sie in ϵ_{it} „landen“ würden. In ϵ_{it} verbleiben dann nur noch die restlichen zeitvarianten Einflüsse einer Person auf y. Schon im Modell ohne unabhängige Variablen (Abs. 1.2) konnte gezeigt werden, dass die Berücksichtigung der individuellen Mittelwerte α_i ϵ_{it} reduzieren konnte (im Falle zentrierter Daten gilt $\alpha_i = a_i$). Da die meisten Regressionsmodelle einen beachtlichen Teil der Varianz von y nicht erklären können, ist zu vermuten, dass im vielen Modellen zeitkonstante Merkmale am Werk sind, die nicht als unabhängige Variablen expliziert wurden (weiter unten auch als „omitted variables“-Problem bezeichnet). Nun wird in Abhängigkeit der Deklaration der a_i-Werte als fix oder zufällig erklärt, wie sie diese zeitkonstante Merkmale „repräsentieren“ und wie sie von ϵ_{it} separiert werden können.

2.2.1 Die Spezifikation des FEM

Werden die a_i-Werte als fix begriffen, dann gehören sie zu den zu schätzenden Koeffizienten einer Regressionsgleichung. Um das nachvollziehen zu können, wird nochmals auf das mehrebenenanalytisches Alternativbeispiel in Abs. 1.3 zurückgegriffen. Dort wurde die Bedeutung von Mittelwertsunterschieden zwischen drei Schulen für die Erklärung der individuellen Leistung der Schüler y_{it} diskutiert. Es wurde gezeigt, dass diese schulspezifischen Mittelwerte alle für die Leistung der Schüler relevanten Merkmale auf Schulebene „repräsentieren“. Wollte man den Einfluss der Schulen in Form von Mittelwertsunterschieden als Koeffizienten in einer Regression modellieren, dann ließe sich das über Dummy-Variablen realisieren. Wenn y_{it} die individuelle Leistung des Schülers t der Schule i, d_1 die Dummy-Variable der ersten, d_2 die Dummy-Variable der zweiten und d_3 die Dummy-Variable der dritten Schule darstellen, dann lautet die Regressionsgleichung:

$$y_{it} = a_1 \cdot d_1 + a_2 \cdot d_2 + a_3 \cdot d_3 + \epsilon_{it} \tag{2.3}$$

Die individuellen Regressionskonstanten a_i sind nun Regressionskoeffizienten der i-ten Schule. a_1 entspricht hierbei dem Mittelwert der ersten Schule, a_2 der zweiten etc. Zu beachten ist, dass hier die a_i-Werte, entgegen der Darstellung in 1.8, nicht die mittlere Abweichung der Schule i vom allgemeinen Mittelwert sondern direkt den Schulmittelwert darstellen. Das ist der in 2.2 eingeführten Zentrierung (welche hier allerdings nur auf y angewendet wird, da sie für Dummy-Variablen nicht zulässig ist) geschuldet. Dadurch entfällt die allgemeine Regressionskonstante, weswegen wiederum keine der drei Schulen als Referenzkategorie deklarieren werden muss (wie dies sonst der Fall ist, wenn die Ausprägungen kategorialer Variablen im Rahmen des allgemeinen linearen Modells als Dummys in die Regressionsgleichung eingehen).

Werden die in Abs. 1.3 aufgeführten Mittelwerte als Schätzungen der Koeffizienten aus Gl. 2.3 eingesetzt, dann ergibt sich:

$$y_{it} = 78 \cdot d_1 + 37 \cdot d_2 + 55 \cdot d_3 + \epsilon_{it} \tag{2.4}$$

ϵ_{it} enthält jetzt für jeden Schüler die Abweichung, welche nicht durch die Schulzugehörigkeit erklärt werden kann. Das ist rechnerisch die Abweichung eines Schülers vom Schulmittelwert seiner Schule. Entsprechend der Differenzierung zwischen den Modellen IO und OIO im Abs. 1.2 ist Gl. 2.4 als eine Realisierung des IO-Modells zu verstehen. Sie ist gegenüber OIO überlegen, wenn ϵ_{it} signifikant verringert wurde. Und da weiter oben festgelegt wurde, dass die Mittelwertsunterschiede zwischen den Schulen signifikant sind, ist dies hier auch der Fall.

Würden nun in Gl. 2.3 zusätzlich „echte“ unabhängige Variablen auf Schülerebene eingeführt werden, z.B. die Durchschnittsnote auf dem letzten Zeugnis, dann würde der Effekt dieser Durchschnittsnote unter Konstanthaltung aller auf Schulebene variierender Merkmale geschätzt. Denn Letztere werden über die Dummy-Variablen im Modell explizit berücksichtigt. Die Effekte a_i der Schuldummys, welche stellvertretend die Effekte dieser Schulmerkmale ausdrücken, sind Koeffizienten des Modells, gehören somit zum *fixen* Teil der Regressionsgleichung. Dieser Umgang mit a_i ist das Grundprinzip des FEM und verleiht ihm seinen Namen.

Übertragen auf Paneldaten lässt sich für jede Person i eine Dummy-Variable bilden. a_i ist entsprechend der Einfluss-Koeffizient der Person i, welche mithilfe ihres „persönlichen“ Dummys d_i als unabhängige Variable in die Regression eingeht. Die Ausgangsgleichung 2.2 mit einer „echten“ unabhängigen Variablen wird wie folgt reformuliert:

$$y_{it} = \beta \cdot x_{it} + \sum_{i=1}^{n} a_i \cdot d_i + \epsilon_{it} \tag{2.5}$$

Das Summenzeichen komprimiert die Darstellung. Für eine ausführliche Formulierung müsste das Produkt $a_i \cdot d_i$ für jede Person i separat aufgeführt werden. Das heißt, dass Gl. 2.5 n Produkte $a_i \cdot d_i$ bzw. $n + 1$ unabhängige Variablen (inkl. x) enthält.[4]

Zur Umsetzung des FEM entsprechend Gl. 2.5 mithilfe von Dummy-Variablen bedarf es dann keines speziellen Schätzverfahrens. Die Koeffizienten der Dummys und „echter“ unabhängiger Variablen werden einfach nach der Kleinste-Quadrate-Methode (KQM) berechnet, die aus der Querschnittsregression bekannt ist. Entscheidend ist, dass neben „echten“ unabhängigen Variablen eben pro Person eine Dummy-Variable eingefügt wird. Die dann mit der KQM geschätzten Koeffizienten der Dummy-Variablen entsprechen der Realisierung von a_i als fixen Bestandteil von Modellen mit variablen Regressionskonstanten. a_i fängt alle zeitkonstanten Einflussfaktoren der Person i auf die Variable y ein. Der Koeffizient einer „echten“ unabhängigen Variablen x stellt dann automatisch eine Schätzung der Einflussstärke von x unter Konstanthaltung aller (!) relevanter zeitinvarianter Merkmale dar. Letztere sind über die Koeffizienten der Dummy-Variablen aus dem Modell herauspartialisiert worden. ϵ_{it} besteht jetzt folglich aus den idiosynkratischen Rest von y, welcher weder durch x noch durch alle relevanten zeitinvarianten Merkmale, repräsentiert durch a_i, erklärt werden kann.

Anhand eines Beispiels soll nun verdeutlicht werden, welche Vorteile das Herauspartialisieren zeitkonstanter Variablen mithilfe fixer a_i bietet: Mal an-

[4] Bei einem Panel mit bspw. 540 Personen würde Gl. 2.5 somit 541 unabhängige Variablen (540 Personendummys und x) enthalten.

genommen, y_{it} ist die Produktivität eines Mitarbeiters i einer bestimmten Firma, gemessen zum Zeitpunkt t (fiktiver Score, gemessen in %). x_{it} sei die Dauer der Betriebszugehörigkeit des Mitarbeiters i zum Zeitpunkt t in Jahren. Wird nun neben x für jede Person i eine eigene Dummy-Variable d_i eingeführt und mit diesen Variablen eine gewöhnliche Regressionsschätzung der Parameter nach KQM berechnet, dann würde der Einfluss der Betriebszugehörigkeit auf die Produktivität *unter Konstanthaltung aller zeitkonstanter relevanter Einflussfaktoren* berechnet. Dazu könnte z.B. die Bildung, das Geburtsjahr, das Geschlecht und die Intelligenz gehören. Nun kann es sein, dass einige dieser für die Produktivität relevanten zeitkonstanten Merkmale auch mit der Betriebszugehörigkeit korrelieren. Das könnte z.B. für das Geburtsjahr der Fall sein: Je höher das Geburtsjahr (in Form des Kalenderjahrs; also je später eine Personen geboren ist), umso niedriger die Produktivität und (!) umso niedriger die Dauer der Betriebszugehörigkeit (kurz: Dauer). Ohne Konstanthaltung des Geburtsjahres würde dann der Effekt der Dauer überschätzt werden. Denn der Erklärungsbeitrag, den sich das Geburtsjahr und die Dauer aufgrund ihrer Korrelation teilen, würde nur der Dauer zugeschrieben werden. Wird aber das Geburtsjahr ebenfalls als Variable berücksichtigt, dann lässt sich der Effekt der Dauer um den Effekt des Geburtsjahres bereinigen. Der „reine Effekt“ des Geburtsjahres, der „reine Effekt“ der Dauer und deren gemeinsamer Effekt werden separierbar. Eine Einbeziehung des Geburtsjahres als unabhängige Variable hilft also, die Qualität der Schätzung des Einflusses der Dauer entscheidend zu verbessern. Nun kann es aber sein, dass das Geburtsjahr als Variable gar nicht vorhanden ist (z.B. aus Datenschutzgründen). Oder der Forscher kommt gar nicht darauf, dass diese Variable einen Effekt auf die Produktivität haben könnte. Dann kann mithilfe fixer a_i der Effekt der Dauer dennoch um den Effekt des Geburtsjahres bereinigt werden – ohne dass das Geburtsjahr explizit bekannt sein muss (!). Denn der Effekt des Geburtsjahres ist in a_i enthalten und wird somit bei expliziter Modellierung von a_i als fixen Teil des Regressionsmodells herausgerechnet.

Dieses Beispiel verweist auf ein allgemeines Problem in Regressionsmodellen, welches mithilfe des FEM-Ansatzes abgeschwächt werden kann: Sobald in einem Regressionsmodell die für die abhängige Variable relevanten Variablen unberücksichtigt bleiben, welche mit den berücksichtigten unabhängigen Variablen korrelieren, dann sind die Koeffizienten der berücksichtigten

unabhängigen Variablen nach oben oder unten verzerrt (vgl. Arminger 1990: 2ff). Dieses auch als „omitted variables“[5]-Phänomen bekannte Problem kann eine gravierende Missinterpretation von Parameterschätzern in Regressionsmodellen herbeiführen. Schließlich ist bei nicht-experimentellen Daten eine solche Korreliertheit zwischen relevanten Einflussfaktoren in vielen Situationen sehr plausibel (vgl. Brüderl 2010). Außerdem bleibt meist ein großer Teil der Varianz der abhängigen Variablen unaufgeklärt, so dass die Existenz von *unberücksichtigten* relevanten Variablen (also „omitted variables“) in solchen Fällen sehr wahrscheinlich ist.

Die explizite Modellierung von a_i als Koeffizienten (fixed effects) hilft nun, das Problem der „omitted variables“ abzuschwächen. Über die Einbeziehung der Personen als „unabhängige Variablen“ werden zumindest alle „omitted variables“ herausgerechnet, welche zeitkonstant sind. Damit sind zwar weiterhin Verzerrungen durch zeit*variable* „omitted variables“ denkbar. Aber immerhin kann man durch die Herauspartialisierung der großen Klasse zeitkonstanter Merkmale die Verzerrung der x-Koeffizienten in vielen Fällen bedeutsam verringern.

Es wurde bislang erläutert, welche Bedeutung der Koeffizient von x unter Kontrolle der a_i inne hat. Nun ist umgekehrt zu klären, welche Bedeutung die einzelnen a_i-Schätzungen haben und wie sie im Falle eingeführter unabhängiger Variablen zu deuten sind. Denn im Gegensatz zum IO-Modell im Abs. 1.2 stehen bei vorhandenen unabhängigen Variablen (wie in Gl. 2.2) die einzelnen a_i-Werte nicht mehr für die individuellen Abweichungen vom allgemeinen Mittelwert sondern für die individuellen Abweichungen vom Mittelwert von y, ***gegeben*** $\mathbf{x}$. Sie stehen also für den Rest der Varianz von y, welcher nicht durch x erklärt werden konnte und auf zeitkonstante Merkmale der Personen zurückgeht. Graphisch ist a_i als der Betrag zu sehen, um den die allgemeine Regressionsgerade $y_{it} = \beta \cdot x_{it}$, je nach Vorzeichen von a_i, parallel nach oben oder nach unten verschoben wird. Sie drückt die Niveauverschiebung aus, die durch zeitkonstante „omitted variables“ (wie z.B. der Bildung) verursacht wird.

Wurde eine Regression mit einer x-Variablen und Personendummys nach

[5] = unberücksichtigte Variablen; in diesem Zusammenhang spricht man auch von unbeobachteter Heterogenität. Es gibt Systematiken in der Streuung des Restfehlers einer Regression, der nach der Schätzung als unerklärter Part über bleibt. Dadurch sind im Restfehler Systematiken enthalten, welche die Annahmen der Zufälligkeit des Residuums verletzen.

der KQM bereits berechnet, kann zur Bestimmung von a_i anstatt auf die geschätzten Regressionskoeffizienten von Personendummys auch auf folgende Formel zurückgegriffen werden (vgl. Hsiao 2014: 33):

$$\widehat{a}_i = \overline{y}_i - \beta \overline{x}_i \tag{2.6}$$

mit

$\widehat{a}_i$ = Geschätzter a_i-Wert

$\overline{x}_i$ = Mittelwert der i-ten Person in Bezug auf die Variable x

Es wird von dem individuellen Mittelwert $\overline{y}_i$ der Term $\beta \overline{x}_i$ abgezogen. Diese Darstellung macht ebenfalls deutlich, dass die individuelle Konstante a_i den individuellen y-Mittelwert darstellt, welcher um den Einfluss von x bereinigt wurde.

Allerdings sind die konkreten a_i-Werte i.d.R. nicht von Interesse. Welche parallele Abweichung ein Herr XY von der allgemeinen Regressionsgeraden $y_{it} = \beta \cdot x_{it}$ hat, spielt meist keine Rolle. Statt der a_i-Werte unter Kontrolle von x ist hingegen umgekehrt nur der Einfluss von x unter Kontrolle von a_i relevant. a_i wird also instrumentalisiert, um die Schätzung von β dahingehend zu verbessern, dass β nicht mehr durch „omitted variables“ verzerrt ist, die zeitkonstant sind und mit x zusammenhängen.

Wenn ein Statistik-Programm ein FEM für uns rechnen soll, dann benötigen wir also keine Auflistung der einzelnen a_i-Werte. Außerdem ist zu bedenken, dass bei einer großen Personenzahl (z.B. $n = 1.000$) insgesamt 1.000 Dummy-Variablen gebildet und 1.001 Koeffizienten (inkl. β) geschätzt werden müssten. Nicht nur, dass eine Auflistung der einzelnen a_i-Werte dann extrem lang ausfallen würde – auch stoßen hier die meisten Statistik-Software-Programme an ihre Grenzen.

Da wir aber die a_i-Werte nicht wirklich brauchen (bzw. auch nachträglich mithilfe der Formel 2.6 rekonstruieren können), können wir einen rechnerischen Trick anwenden, um die gleichen Ergebnisse zu erzielen, ohne 1.001 Koeffizienten schätzen zu müssen (vgl. Hsiao 2014: 32; Brüderl 2010: 967; Giesselmann & Windzio 2012: 43). Dieser Trick basiert darauf, die Effekte von a_i durch eine Verbindung zweier Gleichungen herauszurechnen. Ausgangspunkt ist wieder die Basis-Gleichung entsprechend Gl. 2.2:

$$y_{it} = \beta \cdot x_{it} + a_i + \epsilon_{it} \tag{2.7}$$

Anstatt direkt eine FEM-Schätzung von β mithilfe von Personendummys zu realisieren, lassen sich sich mathematische Umformungen durchführen. Gl. 2.7 wird nun so über Individuen aggregiert, dass pro Individuum eine „individuelle Mittelwertsgleichung“ entsteht:

$$\overline{y}_i = \beta \cdot \overline{x}_i + a_i + \overline{\epsilon}_i \tag{2.8}$$

Wichtig ist an dieser Darstellung, dass der individuelle Mittelwert von a_i eben a_i ist, da a_i nicht innerhalb von Individuen variieren kann (daher braucht man nicht zu schreiben: $\overline{a}_i$).

Nun kann man Gl. 2.8 von Gl. 2.7 subtrahieren. Es ergibt sich:

$$y_{it} - \overline{y}_i = \beta \cdot (x_{it} - \overline{x}_i) + (a_i - a_i) + (\epsilon_{it} - \overline{\epsilon}_i) \tag{2.9}$$

Da $a_i - a_i = 0$ rechnet sich an der Stelle der zeitkonstante Effekt der Individuen heraus. Dies ist hinsichtlich der Schätzung von β völlig äquivalent mit der Einbeziehung der a_i-Werte als Koeffizienten von Personendummys. Da außerdem angenommen wird, dass $\overline{\epsilon}_i = 0$, vereinfacht sich Gl. 2.9 zu:

$$y_{it} - \overline{y}_i = \beta \cdot (x_{it} - \overline{x}_i) + \epsilon_{it} \tag{2.10}$$

Dies ist die sog. „Within-Transformation“ (vgl. Brüderl 2010: 967). Statt mit den ursprünglichen Werten der Variablen y und x wird nun mit den korrespondierenden Differenzen des Werte vom jeweiligen individuellen Mittelwert gearbeitet. Nun kann man diese Differenzen als neue Variablen darstellen: $y^w_{it} = y_{it} - \overline{y}_i$ und $x^w_{it} = x_{it} - \overline{x}_i$ (das w wurde wg. „**W**ithin-Transfomation“ gewählt. Gleichung 2.10 wird reformuliert zu:

$$y^w_{it} = \beta \cdot x^w_{it} + \epsilon_{it} \tag{2.11}$$

Nun kann β mithilfe der KQM geschätzt werden. Aufgrund der Within-Transformation ist die gewöhnliche Schätzung von β nach der KQM eine Realisierung des hier betrachteten FEM. Diese Vorgehensweise ist äquivalent

zur Realisierung des FEM mithilfe von Personen-Dummys, liefert daher ein identisches Schätzergebnis für β. Denn erstens wird mit der KQM in beiden Fällen dieselbe Schätzmethode angewendet. Zweitens entspricht der Fehlerterm ϵ_{it} in Gl. 2.11 dem der Basis-Gleichung 2.7 (bzw. der Gleichung mit Dummy-Variablen 2.5). Die Within-Transformation hat das Residuum nicht verändert. Somit stellt eine Schätzung von β in Gl. 2.11 nach der KQM ebenso den Einfluss von x (nicht: x^w!) unter Kontrolle aller zeitkonstanter Merkmale dar, wie eine Schätzung von β in der Basisgleichung 2.7 nach der KQM und unter expliziter Modellierung von a_i als Koeffizienten von Personendummys.

Da Gl. 2.11 eine einfache Regression mit nur einer unabhängigen Variablen x^w darstellt, ergibt sich folgende Schätzgleichung nach der KQM (Das f symbolisiert die β-Schätzung im ***f***ixed-effects-Modell):

$$\widehat{\beta}_f = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}^w - \overline{x}^w)(y_{it}^w - \overline{y}^w)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}^w - \overline{x}^w)^2} \tag{2.12}$$

Die Formel ähnelt der Schätzformel einer „pooled Regression“, bei der die Variablen x und y nicht within-transformiert wurden. Eine pooled Regression ignoriert die Panelstruktur, da sie auf die Modellierung individueller Regressionskonstanten verzichtet (technisch würde das bedeuten, dass implizit alle a_i in Gl. 2.2 auf Null gesetzt werden, so dass alle zeitkonstanten Einflüsse dem Residuum zugerechnet werden):

$$\widehat{\beta}_p = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x})(y_{it} - \overline{y})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x})^2} \tag{2.13}$$

Der fundamentale Unterschied besteht darin, dass die Schätzung in Gl. 2.12 auf „within-transformierten“ Variablen beruht, die explizit dazu führen, dass a_i nicht nur berücksichtigt, sondern auch aus der Regression herausgerechnet wird. In Gl. 2.13 wird hingegen mit den untransformierten Variablen gearbeitet, so dass die Panelstruktur der Daten nicht berücksichtigt wird. $\widehat{\beta}_p$ ist

somit verzerrt durch alle zeitkonstanten Merkmale, welche auf y einen Einfluss ausüben und mit x korreliert sind.

Um nun die FEM-Schätzung von $\widehat{\beta}_f$ besser nachzuvollziehen zu können, werden y_{it}^w und x_{it}^w in Gl. 2.12 wieder durch ihre ursprüngliche Bedeutung ersetzt. Gl. 2.12 wird dann zu:

$$\widehat{\beta}_f = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}[(x_{it} - \overline{x}_i) - \overline{x}^w][(y_{it} - \overline{y}_i) - \overline{y}^w]}{\sum_{i=1}^{n}\sum_{t=1}^{T}[(x_{it} - \overline{x}_i) - \overline{x}^w]^2} \tag{2.14}$$

Die beiden Ausdrücke $\overline{x}^w$ und $\overline{y}^w$ stellen, jeweils für x und y, den Mittelwert der individuellen Abweichungen vom individuellen Mittelwert dar. Beide nehmen den Wert Null an, da die individuellen Abweichungen schon für eine einzelne Person Null betragen. Der Mittelwert über „Nullen“ beträgt ebenfalls Null:

$$\begin{aligned} \overline{x}^w &= \frac{1}{n \cdot T}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i) = 0 \\ \overline{y}^w &= \frac{1}{n \cdot T}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \overline{y}_i) = 0 \end{aligned} \tag{2.15}$$

Gl. 2.14 vereinfacht sich daher zu:

$$\widehat{\beta}_f = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)^2} \tag{2.16}$$

Gl. 2.16 zeigt, wie in Folge der Umsetzung des FEM-Prinzips die Schätzgleichung von β beschaffen ist, wenn sie mit den ursprünglichen, nicht-transformierten Variablen x und y ausgedrückt wird. Der wesentliche Unterschied

zu Gl. 2.13 besteht darin, dass die Variablenwerte von x und y nun von den ***individuellen*** arithmetischen Mitteln abgezogen werden. Dies ergibt sich rechnerisch eben aus der Tatsache, dass mit a_i *inter*-individuelle Unterschiede herausgerechnet werden. Denn die a_i-Werte stellen das Grundniveau der Person dar. Sind Unterschiede in diesen Grundniveaus (between variation) zwischen Personen eliminiert, dann bleiben zur Schätzung von β als Informationsquelle ***nur*** *intra*-individuelle Unterschiede im Zeitverlauf über. Und das sind eben die Abweichungen der Werte einer Person von *ihrem eigenen* arithmetischen Mittel – also die within variation. $\widehat{b_f}$ ist folglich zu interpretieren als der Betrag, um den der geschätzte y-Wert steigt, wenn x um eine Einheit steigt und (!) alle Unterschiede zwischen den Individuen vorher herausgerechnet wurden. Er entspricht der geschätzten durchschnittlichen Veränderung von y in einer Person, wenn x *bei dieser Person* im Zeitverlauf um eine Einheit steigt. Geometrisch gesehen ist β näherungsweise die durchschnittliche individuelle Regressionsgerade, die sich ergeben würde, wenn man für jede Person ihren eigenen Regressionskoeffizienten β_i berechnen und dann den Mittelwert daraus bilden würde.[6]

Es wird deutlich, dass im FEM ausschließlich die Variation innerhalb von Personen (within variation) zur Schätzung von β genutzt wird. Diese Einschränkung ist zwingend notwendig, um die a_i vollständig herauszurechnen. Die Nicht-Berücksichtigung der Variation *zwischen* Personen (between variation) hat zur Folge, dass keine Koeffizientenschätzungen für perfekt zeitkonstante unabhängige Variablen möglich sind, welche ggf. für den Analysten von Interesse sein könnten. Das lässt sich wieder an der „Within-Transformation“ zeigen (s.o.). Dazu wird in die Ausgangsgleichung 2.7 eine zeitkonstante Variable z_i eingeführt. Der dazugehörige Regressionskoeffizient lautet δ.

$$y_{it} = \beta \cdot x_{it} + \delta \cdot z_i + a_i + \epsilon_{it} \tag{2.17}$$

Nun lässt sich Gl. 2.17 wieder so über Individuen aggregieren, dass pro Individuum eine „individuelle Mittelwertsgleichung“ entsteht:

$$\overline{y}_i = \beta \cdot \overline{x}_i + \delta \cdot z_i + a_i + \overline{\epsilon}_i \tag{2.18}$$

[6] Warum β nicht exakt dem Mittelwert entspricht, wird im alten Skript auf S. 72 erläutert.

Genauso wie der individuelle Mittelwert von a_i a_i entspricht (s.o.), lautet der individuelle Mittelwert von z_i z_i. Im nächsten Schritt wird Gl. 2.18 von Gl. 2.17 subtrahiert. Es ergibt sich:

$$y_{it} - \overline{y}_i = \beta \cdot (x_{it} - \overline{x}_i) + \delta \cdot (z_i - z_i) + (a_i - a_i) + (\epsilon_{it} - \overline{\epsilon}_i) \tag{2.19}$$

Da $\delta \cdot (z_i - z_i) = 0$, $a_i - a_i = 0$ und $\overline{\epsilon}_i = 0$ vereinfacht sich Gl. 2.19 letztlich zu einer Gleichung, die mit Gl. 2.10 exakt identisch ist:

$$y_{it} - \overline{y}_i = \beta \cdot (x_{it} - \overline{x}_i) + \epsilon_{it} \tag{2.20}$$

Es wird deutlich, dass δ nicht schätzbar ist, da es sich aufgrund $\delta \cdot (z_i - z_i) = 0$ ebenso wie a_i herausrechnet. Dass sich der Koeffizient eines zeitkonstanten Merkmals z nicht schätzen lässt, liegt daran, dass der Effekt von z schon in a_i enthalten ist (perfekte Multikollinearität). Da sich a_i vollständig herausrechnet, rechnet sich damit auch der Effekt von z heraus. Das gilt für alle zeitkonstanten Variablen.

Ein weiteres Problem im FEM, das mit den meisten sozialwissenschaftlichen Panelstudien einhergeht, besteht in der relativ kleinen Anzahl T der Zeitpunkte. Die FEM-Schätzung (s. Gl. 2.16) beruht auf der within variation, also auf Differenzen der Einzelwerte vom *individuellen* Mittelwert $\overline{x}_i$ respektive $\overline{y}_i$. Diese individuellen Mittelwerte werden bei kleinem T auf Basis von nur wenigen Messwerten ermittelt. Weist das zugrundeliegende Merkmal auch noch eine hohe within variation auf, dann kann der aus den Daten errechnete Mittelwert einer Person i als Schätzung für ihren „tatsächlichen Mittelwert“ mit einer hohen Unsicherheit behaftet sein. Liegen z.B. pro Person drei Einkommensmessungen vor, so können die individuellen Mittelwerte massiv durch einzelne Ausreißer beeinflusst sein – z.B. wenn eines der gemessenen Einkommen einer Person durch ungewöhnlich viele Überstunden und einmalige Prämien sehr hoch ausfällt. Diese Problematik, die daraus resultiert, dass die gesamte between variation in der Schätzung unberücksichtigt bleibt, lässt sich unter dem Stichwort „mangelnde Effizienz von Schätzern“ subsumieren. Für einen kritischen Kommentar zur Frage, ob dieses Problem ausschlaggebend sein sollte, um auf alternative Modellierungen zurückzugreifen, siehe Abs. 2.3.

Eine solche alternative Modellierung, welche diese potentiellen Nachteile des FEM aufgreift, ist das Random-Effects-Modell (REM). Dort wird a_i auf anderem Wege in die Regressionsgleichung integriert, als im FEM. Dies führt dazu, dass die in dem a_i (so wie es im FEM modelliert wurde) enthaltenen Effekte zeitkonstanter Variablen nicht vollständig herausgerechnet werden. Damit geht einher, dass auch die between variation in die Schätzung einbezogen wird, was wiederum die Schätzung von Effekten zeitkonstanter Variablen erlaubt. Dieser Ansatz wird im folgenden Abschnitt vorgestellt.

2.2.2 Die Spezifikation des REM

Im REM werden a_i nicht mehr, wie im FEM, als Koeffizienten aufgefasst, die es zu spezifizieren gilt. Vielmehr werden die individuellen Regressionskonstanten als zufällige Abweichungen von dem systematischen Teil der Regressionsgleichung begriffen, die somit als Element des Fehlerterms zu modellieren sind. In diesem Sinne ist a_i ein Bestandteil innerhalb des Gesamtfehlers, welcher für eine einzelne Person i konstant ist. Im REM wird versucht, diesen Bestandteil zu identifizieren, indem er vom Restfehler separiert wird. Entsprechend der Ausführungen in Abs. 2.1 erfolgt die Identifikation nicht mehr, wie im FEM, durch Koeffizientenschätzung. Da a_i nun als Zufallsvariablen aufgefasst werden, sind Parameter ihrer Wahrscheinlichkeitsverteilungen zu identifizieren. Insbesondere wird die Varianz von a_i von Bedeutung sein.

Es ergeben sich aus der Behandlung von a_i als Zufallsvariable gegenüber dem FEM zwei praktische Konsequenzen: Dadurch, dass a_i nicht herausgerechnet wird, können erstens auch Koeffizienten zeitinvarianter Merkmale geschätzt werden. Zweitens wird neben der within variation auch die between variation als Variationsquelle zur Schätzung genutzt. Das kann die Effizienz der Schätzer gegenüber dem FEM vergrößern (für einen kritischen Kommentar s. 2.3). In den folgenden Ausführungen soll verdeutlicht werden, *wieso* diese Konsequenzen mit der Konzeption von a_i als Element des Fehlerterms und somit als Zufallsvariable einhergehen und auf welche Weise diese Konzeption die bei Paneldaten vorliegende hierarchische Datenstruktur berücksichtigt (Stichwort: Autokorrelation – s.u.).

Zunächst soll gezeigt werden, wie a_i als Zufallsvariable formal dargestellt und

welche Konsequenz diese Darstellung für das Gesamtresiduum hat. Aufbauend auf diesen Erkenntnissen wird die Schätzung von β im REM vorgestellt.

Entsprechend der Ausführungen im FEM soll auch im REM Gl. 2.2 die Ausgangsgleichung darstellen – zur Erinnerung:

$$y_{it} = \beta \cdot x_{it} + a_i + \epsilon_{it} \tag{2.21}$$

Im FEM wurde a_i als fixer Bestandteil deklariert. Das Residuum stellte folglich ϵ_{it} dar. Dies ist der idiosynkratische Rest, wenn neben der echten unabhängigen Variablen x mit a_i auch alle unbeobachteten zeitkonstanten Merkmale der Personen aus y herausgerechnet werden. Im REM gehört hingegen a_i zum Residuum. Das Residuum besteht folglich aus den zwei Bestandteilen $a_i + \epsilon_{it}$. Diese zwei Teile lassen sich zu einem „Gesamt-Residuum“ u_{it} zusammenfassen:

$$u_{it} = a_i + \epsilon_{it} \tag{2.22}$$

Somit lässt sich Gl. 2.21 reformulieren:

$$y_{it} = \beta \cdot x_{it} + u_{it} \tag{2.23}$$

Oben wurde bereits erläutert, dass aufgrund der hierarchischen Struktur von Paneldaten angenommen werden kann, dass sich die Werte einer Variablen und auch ggf. die Zusammenhänge zwischen mehreren Variablen *innerhalb* einer Person ähnlicher sind, als *zwischen* den Personen. In der Regressionsgleichung 2.23 resultiert aus einer solchen Annahme, dass die Residuen u_{it} einer Person i korreliert sind. Würde in Gl. 2.23 die Variable x wegfallen, dann entsprächen die Residuen u_{it} den Abweichungen der Werte einer Person i zum Zeitpunkt t vom *allgemeinen* Mittelwert (ähnlich dem OIO-Modell – s.o.), welcher hier aufgrund der Zentrierung Null beträgt. Wenn y bspw. das zentrierte Einkommen darstellt, dann ist u_{it} der Wert, mit dem eine Person i zum Zeitpunkt t vom Einkommensdurchschnitt nach oben oder nach unten abweicht. Liegen zwei Einkommensabweichungen u_{i1} und u_{i2} ein und derselben Person i zu zwei Zeitpunkten $t = 1$ und $t = 2$ vor, dann können diese sich zwar voneinander unterscheiden. Aber es kann nicht mehr davon ausgegangen werden, dass diese zwei Werte völlig unabhängig von-

einander sind. Schließlich wird das Einkommen ein und derselben Person nicht von Zeitpunkt zu Zeitpunkt völlig unsystematisch variieren, sondern von dem „Gesamt-Einkommensniveau“ dieser Person abhängen. Stichprobentheoretisch ausgedrückt handelt es sich bei der Teilnahme ein und derselben Person an einer Panelstudie zu zwei verschiedenen Zeitpunkten und somit bei der „Ziehung“ ihrer zwei Residuen nicht um zwei unabhängige Zufallsexperimente. Diese Werte sind stochastisch nicht unabhängig und somit korreliert. Dieselbe Logik gilt bei einer vorhandenen x-Variablen, nur ist dann u_{it} die Abweichung einer Person i zum Zeitpunkt t *von der Regressionsgeraden*.

Um die intra-personale Korreliertheit der Residuen (im Folgenden kurz: personeninterne Autokorrelation) u_{it} formal darstellen zu können, ist eben eine Aufsplittung von u_{it} entsprechend Gl. 2.22 notwendig – und zwar in einen Bestandteil a_i, der für eine Person konstant ist und eine Bestandteil ϵ_{it}, welcher den Rest „auffängt“. Gelingt eine Separation des personenkonstanten Fehlers a_i (Gesamt-Niveau der Person), dann hätte man mit ϵ_{it} einen Bestandteil isoliert, welcher selbst nicht personenintern autokorreliert ist.

Wie bereits erläutert, lassen sich Zufallsvariablen über Parameter ihrer Wahrscheinlichkeitsverteilung identifizieren. Entsprechend wäre es von Vorteil, wenn man schon u_{it} und ihre Bestandteile nicht auf direktem Wege identifizieren kann, zumindest die personeninterne Autokorrelation[7] zu bestimmen. Das lässt sich mithilfe die Aufsplittung in Gl. 2.22 realisieren. Als Zwischenschritt auf dem Weg zur Korrelation wird die *Kovarianz* von zwei Residualvariablen u_{i1} und u_{i2} einer Person i (mit $t = 1$ und $t = 2$) berechnet:

$$\begin{aligned} Cov(u_{i1}u_{i2}) &= \sum(u_{i1}u_{i2}) \\ &= \sum[(a_i + \epsilon_{i1})(a_i + \epsilon_{i2})] \\ &= \sum[a_i^2 + a_i\epsilon_{i2} + \epsilon_{i1}a_i + \epsilon_{i1}\epsilon_{i2}] \\ &= \sum a_i^2 + \sum a_i\epsilon_{i2} + \sum \epsilon_{i1}a_i + \sum \epsilon_{i1}\epsilon_{i2} \\ &= \sum a_i^2 \end{aligned} \tag{2.24}$$

Folgende Annahmen, abgeleitet aus gewöhnlichen linearen Regressionsmo-

[7] Eine Korrelation zweier Zufallsvariablen ist ein zentraler Parameter ihrer gemeinsamen Wahrscheinlichkeitsverteilung.

dellen nach der KQM, liegen den Umformungen in 2.24 zugrunde:[8]

- u_it, a_i und ϵ_{it} besitzen einen Erwartungswert von 0, daher vereinfacht sich die Gleichung der Kovarianz zur Summe der Residuenprodukte $\sum(u_{i1}u_{i2})$ (es muss somit keine Differenz vom Erwartungswert[9] dargestellt werden)
- a_i und ϵ_{it} sind miteinander unkorreliert, deshalb gilt für $\sum a_i\epsilon_{i2} = 0$ und für $\Sigma\epsilon_{i1}a_i = 0$
- Die Fehler ϵ_{it} sind untereinander unkorreliert, daraus resultiert $\sum \epsilon_{i1}\epsilon_{i2} = 0$

Die letzte Zeile von 2.24 liefert einen eindeutigen Wert für die personeninterne Autokorrelation: $\sum a_i^2$. $\sum a_i^2$ ist nichts anderes, als die Varianz von a_i, im Folgenden bezeichnet mit $V(a_i)$.[10] Folglich lautet der Erwartungswert der Kovarianz zweier u_{it}-Variablen (wenn nur ihre t-Werte verschieden sind): $V(a_i)$. Entsprechend der Homoskedastizitäts-Annahme ist $V(a_i)$ für alle i konstant, so dass diese Varianz auch vereinfacht als $V(a)$ dargestellt werden kann.

Um im nächsten Schritt auch die korrespondierende *Korrelation* bestimmen zu können, wird die Benennung der Varianz von u_{it} sowie seiner Bestandteile benötigt. Aufgrund der getroffenen Annahme, dass a_i und ϵ_{it} unkorreliert sind, folgt aus der Aufteilung des Fehlertermes $u_{it} = a_i + \epsilon_{it}$ folgende Aufteilung der Fehler*varianz*:

$$V(u_{it}) = V(a_i) + V(\epsilon_{it}) \tag{2.25}$$

Entsprechend der Homoskedastizitäts-Annahme sind die einzelnen Varianzen für i und t konstant. Folglich vereinfacht sich Gl. 2.25 zu:

$$V(u) = V(a) + V(\epsilon) \tag{2.26}$$

Nun kann $Cor(u_{i1}u_{i2})$, also die Korrelation zwischen u_{i1} und u_{i2} berechnet werden (vgl. Rabe-Hesketh & Skrondal 2008: 59):

[8] Eine ausführliche Einführung in die Annahmen einer linearen Regression findet sich bei Auer (2007).

[9] Diese würde formal wie folgt aussehen: $\sum[(u_{i1} - E(u_{i1}))(u_{i2} - E(u_{i2}))]$.

[10] Im Folgenden steht V(.) immer für die Varianz von „“.

$$
\begin{aligned}
Cor(u_{i1}u_{i2}) &= \frac{Cov(u_{i1}u_{i2})}{\sqrt{V(u_{i1})} \cdot \sqrt{V(u_{i2})}} \\
&= \frac{V(a_i)}{\sqrt{V(a_i) + V(\epsilon_{i1})} \cdot \sqrt{V(a_i) + V(\epsilon_{i2})}} \\
&= \frac{V(a)}{\sqrt{V(a) + V(\epsilon)} \cdot \sqrt{V(a) + V(\epsilon)}} \\
&= \frac{V(a)}{V(a) + V(\epsilon)}
\end{aligned}
\tag{2.27}
$$

Dieser Ausdruck für die personeninterne Autokorrelation kann aufgrund der Homoskedastizitäts-Annahme (Schritt 3 in 2.27) für alle t verallgemeinert werden. Er wird in der Literatur als ρ (Aussprache: „Rho“) bezeichnet. Sind p und q zwei beliebige Werte, die t annehmen kann, und gilt $p \neq q$, dann ist ρ:

$$
\rho = Cor(u_{ip}u_{iq}) = \frac{V(a)}{V(a) + V(\epsilon)} \tag{2.28}
$$

Die Autokorrelation zweier Residuen ein und derselben Person entspricht also dem Anteil der Varianz von a_i an der Varianz des Gesamtfehlers $V(u) = V(a)+V(\epsilon)$. Das ist plausibel, denn a_i fängt alle zeitkonstanten, nicht als unabhängige Variable modellierten Einflüsse einer Person auf. a_i kann folglich nur zwischen Personen streuen. Je größer der Anteil dieser between variation des Residuums an der Gesamt-Streuung des Residuums, umso ähnlicher müssen sich im Umkehrschluss die Residuen ein und derselben Person in Relation zu den Residuen verschiedener Personen sein – umso stärker ist somit die personeninterne Autokorrelation. Schließlich dominieren dann die Unterschiede zwischen den Personen gegenüber den Unterschieden innerhalb von Personen. Ein hoher ρ-Wert signalisiert folglich, dass unter den relevanten unberücksichtigten Einflussfaktoren („omitted variables“ / unbeobachtete Heterogenität; s.o.) insbesondere zeitkonstante Faktoren einen bedeutsamen Einfluss auf y ausüben.

Durch eine Aufspaltung des Gesamtfehlers (2.22) in eine Komponente, welche

nur zwischen Personen variiert und einem Rest, der auch zwischen Zeitpunkten variieren kann, wurde es möglich, die Korrelation der Residuen ein und derselben Person zu spezifizieren. Das ist nötig, da bei Paneldaten die Annahme plausibel ist, dass diese Korrelation $\neq 0$ ist (s.o.). Damit ist aber auch die Annahme der KQM von unkorrelierten Residuen verletzt. Bei der Schätzung von β in Gl. 2.23 wird daher eine alternative Methode benötigt, welche Autokorrelationen zulässt. Diese Anforderung erfüllt die sog. „generalized-least-squares"-Methode (GLS-Methode). An dieser Stelle wird auf die Darstellung mathematischer Details verzichtet.[11] Wichtig ist, dass der GLS-Schätzer von β unter der Annahme berechnet wird, dass die Korrelation zwischen zwei verschiedenen Residuen derselben Person ρ beträgt.[12] Das führt zu einer Differenzierung der Informationen, welche aus der within und welche aus der between variation von x und y (bzw. aus der within und between *co*variation zwischen x und y) stammen. Somit wird im REM die hierarchische Datenstruktur berücksichtigt, nach der Einzelmessungen in Personen eingebettet werden. Denn durch die Annahme personeninterner Autokorrelationen wird ein „Bezug zwischen Messungen jeweils einer Person" (vgl. Giesselmann & Windzio 2012: 28) hergestellt. Formal wird diese Differenzierung durch ein Gewicht umgesetzt, mit denen between- und within-Variationsinformationen in die Schätzung eingehen.

Ehe die GLS-Schätzformel inkl. dieses Gewichts vorgestellt wird, wird als Zwischenschritt die Schätzung des sog. „Between-Effects-Modells" (BEM) eingeführt. Das BEM schätzt β auf aggregierter Individualebene, ähnlich der Darstellung in Gl. 2.8:

$$\overline{y}_i = \beta \overline{x}_i + \epsilon_i \tag{2.29}$$

Indem pro Person nur ihre individuellen Mittelwerte in die Schätzung eingehen, wird bei BEM ausschließlich die between variation betrachtet. Eine Schätzung von β nach der KQM ($\widehat{\beta}_b$; das kleine b steht für „between") ergibt:

[11] Die Herleitung der GLS-Schätzung zeigt Hsiao (2014: 35ff.)

[12] Außerdem wird angenommen, dass die Korrelation zwischen zwei verschiedenen Residuen zwei verschiedener Personen Null beträgt.

$$\widehat{\beta}_b = \frac{\sum_{i=1}^{n}(\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y})}{\sum_{i=1}^{n}(\overline{x}_i - \overline{x})^2} \tag{2.30}$$

Diese Darstellung ist für die Erläuterung des REM wichtig, da sich Elemente der BEM-Formel 2.30 in der Schätzung des REM wiederfinden. Denn wie bereits gesagt, enthält die REM-Schätzung sowohl Elemente der within variation (aus der FEM-Schätzung) als auch der between variation (aus der BEM-Schätzung). Diese gehen mithilfe eines Gewichts G unterschiedlich gewichtet in die Schätzung ein. Mithilfe des GLS-Verfahrens, welches die personeninterne Autokorrelation der Residuen berücksichtigt, ergibt sich folgende Schätzgleichung.

$$\widehat{\beta}_r = \frac{\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i) + G \cdot \sum_{i=1}^{n}(\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y})}{\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)^2 + G \cdot \sum_{i=1}^{n}(\overline{x}_i - \overline{x})^2} \tag{2.31}$$

mit:

$G = \frac{V(\epsilon)}{V(\epsilon)+T \cdot V(a)}$

Die Größe G stellt in Gl. 2.31 das Gewicht dar, mit dem die Komponenten der BEM-Schätzung (rote Farbe) gegenüber den Komponenten der FEM-Schätzung (blaue Farbe) gewichtet werden. Somit kann der Schätzer des REM $\widehat{\beta}_r$ als ein gewichteter Durchschnitt aus dem Schätzer des FEM- und dem des BEM-Modells gesehen werden. Da die FEM-Komponenten mit dem impliziten Gewicht 1 eingehen, während $G \leq 1$ ist, bestimmt das Gewicht G, wie stark *zusätzlich* zu den within Informationen, welche im FEM *ausschließlich* verarbeitet werden, auch die between Informationen hinzugezogen werden.

Denn ein möglicher Nachteil der FEM-Schätzung basiert darauf, dass die individuellen Mittelwerte auf Basis von nur wenigen Zeitpunkten und ggf. zusätzlich einer hohen within variation berechnet wurden. Diese errechneten

individuellen Mittelwerte als Schätzungen der „wahren“ individuellen Mittelwerte sind dann mit einer gewissen Unsicherheit verbunden (vgl. Giesselmann & Windzio 2012: 83).

G schätzt das Ausmaß dieser Unsicherheit, indem es erstens (wenn man sich das T zunächst wegdenkt) den Anteil der Varianz des Rest-Residuums $V(\epsilon)$ an der Gesamtvarianz $V(u)$ bestimmt (eine alternative Darstellung dieser Gewichtung auf Basis der Standardfehler von $\widehat{\beta}_f$ und $\widehat{\beta}_b$ findet sich bei Rabe-Hesketh et al. 2012a: 148).[13] Zweitens wird im Nenner $V(a)$ mit der Anzahl der Zeitpunkte multipliziert, so dass auch die Anzahl der Messwerte, welche pro Individuum zur Berechnung eines individuellen Mittelwertes zur Verfügung stehen, berücksichtigt wird.

Bei konstantem T steigt folglich mit steigendem Anteil der within Fehlervarianz $V(\epsilon)$ die Unsicherheit der Schätzung im FEM. Denn bei einer hohen Unsicherheit muss es verhältnismäßig viele unberücksichtigte Einflüsse geben, die innerhalb von Personen relativ starke Schwankungen der Messwerte hervorrufen. Umgekehrt: Je geringer die between Fehlervarianz (kleines $V(a)$ im Verhältnis zu $V(\epsilon)$) ausfällt, umso größer ist die aus der between covariation zwischen x und y zusätzlich gewonnene Varianzaufklärung in Bezug auf y – umso größer fällt dann das Gewicht G aus. Bei konkret realisierten FEM-/REM-Berechnungen auf Basis von Survey-Daten seien aber Zweifel angebracht, ob ein relativ großes $V(\epsilon)$ tatsächlich nur für die Unsicherheit bzw. fehlende Effizienz eines auf Basis von within (co-)variation (FEM) errechneten Schätzers steht, und ob die zusätzlich gewonnene Varianzaufklärung durch die Hinzunahme von between Informationen auf unverzerrte kausale Effekte der berücksichtigten unabhängigen Variablen zurückgeht. Für einen kritischen Kommentar zu diesen Fragen siehe Abs. 2.3.

Bleibt man zunächst bei der Logik, dass ein relativ großer Anteil der within-

[13] Zur Herleitung der Schätzung dieser Fehlervarianzen vgl. Hsiao 2014: 38; die Formeln sind in Abs. 2.4 nachzulesen. Dem aufmerksamen Leser mag aufgefallen sein, dass die Schätzung zirkulär ist. Schließlich wird in der Schätzgleichung die Kenntnis des Gewichts G vorausgesetzt. G setzt wiederum die Kenntnis der Fehlervarianzen voraus, die man aber erst dann schätzen kann, wenn die Parameter des REM (hier: β) bekannt sind. Diese Zirkularität lässt sich dadurch auflösen, dass $V(\epsilon)$ im FEM aus mathematischen Gründen dieselbe Größe annimmt, wie im REM. Man kann also auf die Schätzung im FEM zurückgreifen. Die Schätzung von $V(a)$ setzt hingegen, bei bekanntem $V(\epsilon)$, keine Kenntnis von Parameterschätzern mehr voraus – s. 2.4. Damit lassen sich alle relevanten Größen eindeutig identifizieren. Wie in der Formelsammlung am Ende von Abs. 2.4 nachzuvollziehen ist, besteht die Schätzung von $V(a)$ im Wesentlichen aus der Fehlervarianz des BEM – sie lässt sich als between Fehlervarianz begreifen. Dass die Formel nicht *ausschließlich* aus der Fehlervarianz des BEM besteht, liegt daran, dass das BEM within Unterschiede nicht, wie man annehmen könnte, (spiegelverkehrt zum FEM) vollständig herausrechnet. Sie werden lediglich „geglättet“ (vgl. Giesselmann & Windzio 2012: 94). Dadurch ist in der BEM-Fehlervarianz auch within-Fehlervarianz enthalten. Um die reine between Fehlervarianz zu erhalten, muss diese daher von der within-Fehlervarianz separiert werden. Dies geschieht, indem in der Formel $\frac{\hat{V}(\epsilon)}{T}$ abgezogen wird (vgl. StataCorp 2011: 472). $V(a)$ ist also nicht, wie im FEM, die Varianz der *totalen* individuellen Abweichungen von der FEM-Regressionsgeraden (die ja im FEM auch total herausgerechnet werden). Es ist die Varianz der individuellen Unterschiede, welche durch die gemeinsame Kovariation von x und y auf between-Ebene nicht erklärt werden kann. Die individuelle Fehlervarianz im REM ($V(a)$) umfasst somit nur einen Teil der totalen individuellen Varianz im FEM ($V(a)_f$): $V(a) \leq V(a)_f$.

Fehlervarianz $V(\epsilon)$ am Gesamtresiduum für die Unsicherheit von within-Schätzern steht, dann ergibt sich die folgende Relation:

Tendiert G gegen 0, dann liegt dies an der Dominanz von $V(a)$ und folglich an der Dominanz *zeitinvarianter* individueller Abweichungen (die im BEM nicht erklärt werden können). Der REM-Schätzer $\widehat{\beta}_r$ konvergiert dann gegen den FEM-Schätzer $\widehat{\beta}_f$. Denn die verhältnismäßig geringe within variation des Restfehlers lässt den Schluss zu, dass die individuellen Mittelwerte in der FEM-Schätzung mit nur geringer Unsicherheit behaftet sind. Umgekehrt kann man sagen: Die between Informationen sind nicht allzu gut geeignet, um diese Unsicherheit zu reduzieren. Im Extremfall $G = 0$ reduziert sich Gl. 2.31 sogar ganz auf den „blauen Teil“. Dieser sehr unwahrscheinliche Fall wäre gegeben, wenn die within variation von x die within variation von y vollständig aufklären würde.

Tendiert G gegen 1, dann dominieren mit $V(\epsilon)$ die Abweichungen *innerhalb* von Personen, welche über die Zeit variieren. Die Anteile des FEM-Schätzers und des BEM-Schätzers am REM-Schätzer sind dann nahezu „gleichgewichtig“. Im Extremfall $G = 1$ brauchte man nicht weiter zwischen i und t zu differenzieren. Es würde die einfache KQ-Methode des pooled-Modells (s.o., Gl. 2.13) ausreichen. Somit konvergiert mit steigendem G der Schätzer $\widehat{\beta}_r$ gegen den Schätzer des pooled-Modells, $\widehat{\beta}_p$ (Gl. 2.13). Die Konstellation $G = 1$ ist als die maximale Erweiterung der in der FEM-Schätzung verarbeiteten within Informationen um between-Informationen zu verstehen. Die within-Fehlerstreuung ist dann in Relation zur between-Fehlerstreuung so stark, dass die für die FEM-Schätzung berechneten individuellen Mittelwerte mit einem sehr hohen Ausmaß an Unsicherheit behaftet sind. Umgekehrt kann man sagen: Die between Informationen sind dann sehr gut geeignet, um diese Unsicherheit zu reduzieren. Im Falle von $G = 1$ würden sie maximal genutzt werden.

Eine alternative Darstellung der REM-Schätzung ist die Gewichtung der im FEM ungewichtet verwendeten individuellen Mittelwerte durch ein Gewicht λ. Dieses Gewicht, welches zwischen 0 und 1 liegt, lässt sich unmittelbar aus dem Gewicht G errechnen (vgl. Giesselmann & Windzio 2012: 84):

$$\lambda = 1 - \sqrt{G} = 1 - \sqrt{\frac{V(\epsilon)}{V(\epsilon) + T \cdot V(a)}} \tag{2.32}$$

Je kleiner λ wird, umso ein geringeres Gewicht wird den individuellen Mittelwerten zugesprochen, was wiederum zu einer größeren Berücksichtigung der between-Variation in der Schätzung führt. Es findet eine Art Anpassung der individuellen Mittelwerte, die bekanntlich von Person zu Person streuen, am Gesamtmittelwert statt. λ wird dabei besonders klein, wenn der Anteil der within-Fehlervarianz an der Gesamtvarianz steigt bzw. der Anteil der between-Fehlervarianz an der Gesamtvarianz fällt – was wiederum mit einer Unsicherheit der individuellen Mittelwerte und einer guten Adjustierbarkeit dieser durch between-Informationen (bzw. durch den Gesamtmittelwert) assoziiert wird.

Zur Verdeutlichung wird die im FEM vorgestellte Within-Transformation mithilfe von λ modifiziert. Ausgangspunkt ist wieder die Basisgleichung, wobei hier im REM-Kontext a_i zum Fehlerterm gehört:

$$y_{it} = \beta \cdot x_{it} + a_i + \epsilon_{it} \tag{2.33}$$

Nun lässt sich wieder eine Mittelwerts-Aggregation über Individuen durchführen. Die Mittelwerte werden allerdings mit λ gewichtet:

$$\lambda \cdot \overline{y}_i = \beta \cdot \lambda \cdot \overline{x}_i + \lambda \cdot a_i + \lambda \cdot \overline{\epsilon}_i \tag{2.34}$$

Dann lässt sich analog zur Within-Transformation eine Differenzgleichung aus 2.33 und 2.34 bilden:

$$y_{it} - \lambda \cdot \overline{y}_i = \beta \cdot (x_{it} - \lambda \cdot \overline{x}_i) + (a_i - \lambda \cdot a_i) + (\epsilon_{it} - \lambda \cdot \overline{\epsilon}_i) \tag{2.35}$$

Das ist die sog. GLS-Transformation. Es wird deutlich, dass die Within-Transformation einen Spezialfall der GLS-Transformation darstellt. Dieser tritt ein, wenn $\lambda = 1$. Das wiederum kann nur dann der Fall sein, wenn $V(\epsilon) = 0$, wenn also die Fehlervarianz einzig auf Unterschiede zwischen den Personen zurückgeht. Dann wären die im FEM verwendeten individuellen Mittelwerte nicht mehr mit Unsicherheit behaftet. Schließlich wäre dann der

within Zusammenhang zwischen x und y perfekt – ein Ergebnis, welches kaum auf Basis unsicherer Schätzungen individueller Mittelwerte herauskommen kann.

Wenn aber $V(\epsilon) \neq 0$, dann rechnet sich a_i aufgrund von $a_i - \lambda \cdot a_i \neq 0$ im Gegensatz zum FEM nicht mehr vollständig heraus. Das liegt daran, dass nun nicht mehr die Differenz eines Messwerts vom *ungewichteten* individuellen Mittelwert gebildet wird. Nach Umformung ergibt sich:

$$y_{it} - \lambda \cdot \overline{y}_i = \beta \cdot (x_{it} - \lambda \cdot \overline{x}_i) + \epsilon_{it} + (1 - \lambda) \cdot a_i \tag{2.36}$$

Das Residuum besteht nun aus dem Ausdruck $\epsilon_{it} + (1 - \lambda) \cdot a_i$. Es wird deutlich, dass dieses Residuum der GLS-Transformation komplexer ist, als das Residuum der Within-Transformation. Es entspricht nicht mehr ϵ_{it}, welches das Residuum im FEM darstellt, wenn a_i (durch Differenzbildung oder mithilfe von Personendummys) vollständig herausgerechnet wird. Das Residuum besteht hingegen aus den beiden Bestandteilen ϵ_{it} und a_i, reflektiert also die Idee des REM.

Die geschätzte individuelle Fehlervarianz $V(a)$ determiniert, als Bestandteil von λ, mit welchem Gewicht die individuelle Fehlerkomponente a_i in das Residuum eingeht. Das Gewicht beträgt $1 - \lambda$, es gilt folglich: Je größer $V(a)$ in Relation zu $V(\epsilon)$, mit einem umso kleineren Gewicht geht a_i in die REM-Gleichung ein (und umgekehrt). Denn λ wird bei steigendem $V(a)$ größer, das Gewicht von a_i, $1 - \lambda$, daher *kleiner*. Das liegt daran, dass bei einen verhältnismäßig hohem $V(a)$ die between-Informationen relativ ungeeignet sind, um die „Unsicherheit“ des FEM-Schätzers zu reduzieren. Je größer also $V(a)$, umso weniger werden in der REM-Schätzung die within-Komponenten aus der FEM-Schätzung um between Informationen erweitert (umso kleiner G). Bei einem entsprechend hohen λ dominiert in der REM-Schätzung die FEM-Komponente, was wiederum dazu führt, dass die between-Unterschiede (und damit auch die between-Fehlervarianz) zum Großteil herausgerechnet werden. Es verbleibt ein kleiner Teil der between-Fehlervarianz in der Gleichung, welcher über die nur geringe Anreicherung der REM-Schätzung mit Informationen aus der between covariation zwischen x und y auch nur geringfügig in das Residuum „hinein transportiert“ wird. Ein Großteil der between-Fehlervarianz wird hingegen durch die Dominanz der FEM-Komponente an der REM-Schätzung absorbiert.

Fällt $V(a)$ hingegen relativ klein aus, dann fällt das Gewicht $1 - \lambda$ relativ groß aus. Die between covariation zwischen x und y kann individuelle Niveauunterschiede relativ gut erklären (was aber in Bezug auf Kausalität skeptisch betrachtet werden sollte – s. 2.3), wird daher mit einem recht hohen Gewicht G in der REM-Schätzung berücksichtigt. Dadurch aber, dass die Informationen aus der between covariation zwischen x und y so stark in die REM-Schätzung eingebunden werden, wird ein sehr hoher Anteil der between-Fehlervarianz „mitgeschleppt“ bzw. in das Residuum „hinein transportiert“. λ fällt wiederum klein aus, das führt zu einer „Entwertung“ der individuellen Mittelwerte in der Schätzung, somit wird nur ein geringer Teil der between-Unterschiede herausgerechnet. Das gilt dann auch für between-Unterschiede, welche nicht durch die between covariation zwischen x und y erklärt werden. Das wäre im FEM anders gewesen. Schließlich werden dort alle between-Unterschiede von y herausgerechnet – unabhängig davon, ob sie durch die between covariation von x und y erklärt werden können, oder nicht.

Schließlich kann man die GLS-transformierten Variablen aus Gl. 2.36 als neue Variablen darstellen: $y_{it}^g = y_{it} - \lambda \cdot \overline{y}_i$ und $x_{it}^g = x_{it} - \lambda \cdot \overline{x}_i$ (das g wurde wg. „**G**LS-Transfomation“ gewählt). Auch der Fehlerterm lässt sich zusammenfassen zu: $u_{it} = \epsilon_{it} + (1 - \lambda) \cdot a_i$. Die GLS-transformierte Gl. 2.36 lässt sich dann komprimierter darstellen als:

$$y_{it}^g = \beta \cdot x_{it}^g + u_{it} \tag{2.37}$$

Nun kann β mithilfe der KQM geschätzt werden. Aufgrund der GLS-Transformation ist die gewöhnliche Schätzung von β anhand der GLS-transformierten Variablen nach der KQM identisch mit der in Gl. 2.31 dargestellten REM-Schätzung von β nach der GLS-Methode:

$$\widehat{\beta}_r = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}^g - \overline{x}^g)(y_{it}^g - \overline{y}^g)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}^g - \overline{x}^g)^2} = \frac{\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i) + G \cdot \sum_{i=1}^{n}(\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y})}{\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)^2 + G \cdot \sum_{i=1}^{n}(\overline{x}_i - \overline{x})^2} \tag{2.38}$$

Es wurde deutlich, dass im REM im Vergleich zum FEM a_i als Bestandteil des Re-

siduums modelliert wird. Indem die personeninterne Autokorrelation in der Schätzung berücksichtigt wird, erhält man eine Schätzung, welche je nach Stärke dieser Autokorrelation zusätzlich zu den in der FEM-Schätzung verarbeiteten within Informationen auch auf between Informationen zurückgreift. Denn die Stärke der Autokorrelation und die Gewichtung hängen unmittelbar zusammen: Bei konstantem und kleinem T ist das Gewicht G näherungsweise umgekehrt proportional zu ρ. Steigt also die personeninterne Autokorrelation ρ, dann fällt das Gewicht G und umgekehrt. Ein aus dieser Vorgehensweise resultierender praktischer Vorteil gegenüber dem FEM liegt darin, dass nun auch Koeffizienten von zeitinvarianten unabhängigen Variablen schätzbar werden. Schließlich werden die a_i nicht aus dem Modell völlig herausgerechnet, sondern als ein Bestandteil des Residuums (zufälliger Part) in der Schätzung von dem „systematischen“ Rest separiert. Genau darin liegt aber auch die Gefahr des REM: Wenn dieser systematische Part z.B. aus einer unabhängigen Variablen x besteht, dann wird fälschlicherweise der Einfluss von x unter- bzw. überschätzt, sobald es zeitkonstante nicht berücksichtigte Einflüsse gibt, die mit x korrelieren. Solche Verzerrungen kann man in den meisten auf Survey-Daten basierten Modellen annehmen.

Während also durch das Herauspartialisieren von a_i im FEM das „omitted variables“-Problem zumindest in Bezug auf zeitkonstante Merkmale eliminiert wurde, ist es im REM noch vorhanden. Das REM kann zwar mithilfe der vorgenommenen Modellierung der personeninternen Autokorrelation, im Gegensatz zur „pooled Regression“, die hierarchische Struktur von Paneldaten korrekter berücksichtigt werden.[14] Das „omitted variables“-Problem wird dadurch aber höchstens abgemildert.

2.3 REM und FEM im Vergleich

Die beiden Abschnitte 2.2.1 und 2.2.2 beschränkten sich nicht nur die mathematische Darstellung der Schätzgleichungen von REM und FEM. Es wurden auch die mit den jeweiligen Verfahren verbundenen Konsequenzen angeschnitten, die an dieser Stelle gezielt diskutiert werden:

Es wurde deutlich, dass durch das Herauspartialisieren von a_i die Schätzungen im FEM unter Kontrolle aller zeitkonstanter „omitted variables“ erfolgen. Koeffizienten der im Modell aufgenommenen unabhängigen Variablen sind somit nicht mehr durch zeitkonstante „omitted variables“ verzerrt, welche mit den aufgenommenen unabhängigen Variablen korreliert sind.

[14] Die Effizienz der Schätzung wird verbessert, da im REM im Falle eines hohen ρ berücksichtigt, dass wiederholte Informationen von ein und derselben Person einen niedrigeren Informationsgehalt aufweisen, als Informationen verschiedener Personen.

Verzerrungen aufgrund des „omitted-variables"-Problem in Regressionsmodellen sind in der Sozialforschung sehr ernst zu nehmen, da dort meist mit Survey-Daten gearbeitet wird, wo im Vergleich zu echten Experimenten Korrelationen unter den unabhängigen Variablen in der Regel zu erwarten sind. Diese Korrelationen lassen stark vermuten, dass es auch Korrelationen dieser unabhängigen Variablen mit weiteren für die abhängige Variable relevanten Eigenschaften gibt, die nicht erhoben bzw. im Modell nicht berücksichtigt wurden („omitted variables"). Die Verbesserung der Qualität von Schätzern, welche über das FEM zumindest im Hinblick auf *zeitkonstante* „omitted variables" erreicht wird, nimmt somit im Kontext *allgemeiner* Bemühungen zur Verbesserung der Qualität von Schätzern in Regressionsmodellen einen zentralen Stellenwert ein.

Diese Verbesserung wird im REM nicht bzw. nur zum Teil erreicht. Um Effekte zeitkonstanter und quasi-zeitkonstanter Merkmale schätzbar zu machen (für eine Diskussion der Notwendigkeit, Effekte zeitkonstanter Merkmale schätzen zu können, s.u.) bzw. um die Schätzung auf eine breitere Datenbasis zu stellen, werden, zusätzlich zu den vollwertig (implizites Gewicht $= 1$) im Modell inkludierten within-Informationen, die aus der between (co)variation stammenden Informationen mit einem Gewicht $G \leq 1$ (s.o.) in die Schätzung aufgenommen. Das Gewicht G hängt von dem Verhältnis zwischen der within- und der between-Fehlervarianz ab, welches sich durch die Modellierung personeninterner Autokorrelation ergibt. Dadurch werden aber die a_i im REM nicht vollständig aus dem Modell herauspartialisiert. So ist weiterhin eine Verzerrung der Koeffizienten durch zeitkonstante „omitted variables" vorhanden, welche mit den unabhängigen Variablen korreliert sind. Der entscheidende Vorteil des FEM, ein so großes Problem von Regressionsmodellen entscheidend reduzieren zu können, geht im REM wieder verloren.

Zudem basiert die Logik des Gewichts G, mit dem im REM between-Informationen zusätzlich zu within-Informationen aufgenommen werden, nicht darauf, die Verzerrung durch zeitkonstante „omitted variables" gering zu halten. Hingegen wird, wie in der Regressionstheorie im Allgemeinen, vom „Idealzustand" ausgegangen. In diesem Zustand sind die Residuen tatsächlich zufällig, weisen also keinerlei Systematik auf. In einem solchen Falle ist es nachvollziehbar, dass das Gewicht davon abhängt, wie viel zusätzliche Erklärkraft between-Informationen mit sich bringen bzw. wie groß die Fehlervarianz auf within-Ebene in Relation zur Fehlervarianz auf between-Ebene ausfällt. Da aber in der Praxis des Arbeitens mit Survey-Daten diese Erklärkraft auf between-Ebene eben häufig überschätzt ist, ist man entsprechend häufig weit von diesem Idealzustand entfernt. So ist es z.B. möglich, dass das Gewicht G relativ hoch ausfällt, gleichzeitig das Ausmaß der Verzerrung durch between-Informationen ebenfalls hoch ist. In einem solchen Fall würden gerade die

Informationen mit einem hohen Gewicht Eingang in die Schätzung eines REM-Koeffizienten finden, die besonders „großen Schaden“ anrichten.

Immerhin fällt das Ausmaß der Verzerrung durch unberücksichtigte zeitkonstante Drittvariablen im REM geringer aus, als in der pooled Regression. Denn das Gewicht G, mit dem between-Informationen in der Schätzung berücksichtigt werden, fällt im REM i.d.R. deutlich kleiner als 1 aus, während es in der pooled Regression immer implizit bei 1 liegt. Hinsichtlich der Intensität der aus dieser Quelle stammenden Verzerrung liegt das REM also zwischen FEM und der pooled Regression.

Häufig aufgeführtes Argument gegen das FEM ist der Verweis darauf, dass das FEM weniger effizient ist, als das REM, da es nur einen Teil der Informationen in den Daten nutzt. Gerade bei einer kleinen Anzahl an Zeitpunkten pro Person werden zudem die individuellen Mittelwerte, die den Bezugspunkt zur Differenzbildung in der Berechnung der within-Kovarianz (s.o.) darstellen, aus einer dünnen Datenlage abgeleitet. Im REM wird dieser Nachteil dadurch berücksichtigt, dass die between variation desto stärker in der Schätzung verarbeitet wird (Gewicht G steigt), je weniger Messzeitpunkte es pro Person (im unbalancierten Fall: im Schnitt) gibt. Brüderl (2010: 975 ff., Hervorhebungen im Original) führt hierzu aus, dass „nicht Effizienz (kleine Standardfehler) sondern Vermeidung von Bias [gleichzusetzen mit dem „omitted-variables“-Problem; Stein, P. & Bekalarczyk, D.] ... das wichtigere Ziel in der Sozialforschung [ist]. 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'“. Dieser Argumentation möchten wir uns anschließen. Wie oben ausgeführt, ist schließlich gerade bei Survey-Daten davon auszugehen, dass between-Informationen „kontaminiert“ sind.

Wie schon oben dargelegt, wird im FEM für diesen großen Vorteil der „Bias-Reduktion“ der Preis gezahlt, dass die Schätzung von Effekten zeitinvarianter Merkmale verunmöglicht wird. Schließlich wird nur die „within (co-)variation“ zur Schätzung herangezogen, wodurch das Unterschiede *zwischen* Personen vollständig herausgerechnet werden. Ein damit verwandtes Problem tritt bei *quasi*-zeitinvarianten Merkmalen (Merkmale mit geringer within variation) auf. Die Schätzung im FEM wird dann nur auf diejenigen wenigen Fälle der Stichprobe beschränkt, welche echte within variation aufweisen. Eine solche Schätzung wird daher aufgrund der geringen Fallzahl im Vergleich zum REM ineffizient. Außerdem ist der Schluss von der Gesamt-Stichprobe auf die angestrebte Grundgesamtheit beeinträchtigt, wenn diese Untergruppe nicht repräsentativ für die Gesamtstichprobe ist. Ein Beispiel wäre der Einfluss des höchsten Schulabschlusses in einem Erwachsenensample, dessen Schätzung nur auf denjenigen beruht, die auf dem zweiten Bildungsweg einen

(höheren) Schulabschluss nachgeholt haben.

Ob es im Einzelfall wirklich ein Problem darstellt, dass Effekte zeitkonstanter Merkmale gar nicht und Effekte quasi-zeitkonstanter Merkmale nur unter großer Unsicherheit geschätzt werden können, sollte aber gut überlegt werden. Will man z.B. in einem Regressionsmodell das Geschlecht oder, im Falle von Migranten, das Geburtsland als Kontrollvariable einführen, weil aus vorangegangener empirischer Forschung bedeutende Effekte bekannt sind, stehen diese Variablen aber nicht im Zentrum der eigenen Fragestellung, dann kann weiterhin auf das FEM zurückgegriffen werden. Schließlich werden dort durch das Herausrechnen der Effekte aller zeitkonstanten Merkmale auch die Effekte von Geschlecht und Geburtsland herausgerechnet. Eine Kontrolle mit diesen beiden Merkmalen findet also ohnehin statt. Umgekehrt betrachtet werden eben nicht nur die Effekte von einzelnen ausgewählten zeitkonstanten Merkmalen wie Geschlecht und Geburtsland herauspartialisiert, sondern *zusätzlich* die Effekte aller anderen zeitkonstanten Merkmale, deren Wirkung dem Forscher u.U. gänzlich unbekannt sein kann.[15]

Manchmal steht nicht der Effekt einer zeitkonstanten Variablen wie dem Geschlecht an sich im Vordergrund, die zeitkonstante Variable fungiert hingegen als *Moderator* für den Effekt einer zeitveränderlichen Variablen. Beispielsweise könnten geschlechterspezifische Unterschiede in Bezug auf den Effekt der Anzahl der Kinder auf den Stundenlohn interessieren. Eine solche Fragestellung wird in Regressionsmodellen i.d.R. mithilfe eines Interaktionsterms umgesetzt. Auch dies lässt sich im FEM realisieren. Zwar kann nicht der konditionale Haupteffekt der zeitkonstanten Variablen (hier: Effekt des Geschlechts für die Kinderzahl 0) geschätzt werden, da die zeitkonstante Variable natürlich auch dann zeitkonstant ist, wenn die Interaktionspartnervariable Null beträgt. Aber der Effekt des Interaktionsterms selbst kann geschätzt werden, da die Multiplikation einer zeitkonstanten Variablen mit einer zeitvarianten wiederum eine zeitvariante Variable ergibt.

Interaktionseffekte bringen zwar im FEM generell das Problem mit sich, dass sie keine reinen within-Schätzer darstellen, wenn einfach nur die aus Multiplikation beider Variablen errechnete Interaktionsvariable der within-Transformation unterzogen wird. Durch zusätzliche Transformationen lassen sich aber within-Schätzer erreichen (Giesselmann & Schmidt-Catran 2016).

Steht hingegen ein zeitkonstantes Merkmal selbst in der eigenen Fragestellung im Vordergrund, so dass die Werte des Regressionskoeffizienten für dieses Merkmal explizit benötigt werden, dann muss auf das REM (oder das Hybrid-Modell, s. 2.5)

[15] Diese Argumentation kann auch auf quasi-zeitkonstante Merkmale angewendet werden. Hier stellt sich ebenfalls die Frage, ob der entsprechende Regressionskoeffizient explizit benötigt wird. Ist dies nicht der Fall, dann ist es zulässig, beim FEM zu bleiben, dann sollte aber auf die Interpretation des betroffenen Regressionskoeffizienten weitestgehend verzichtet werden.

ausgewichen werden. Beispiele wären die Prüfung von Unterschieden zwischen Migrantengruppen (operationalisiert über das eigene Geburtsland oder das der Eltern) unter Kontrolle von aufnahmelandspezifischen Ressourcen in einer migrationstheoretischen Fragestellung (z.B: Kalter 2006) oder die Frage, ob das Geschlecht auch unter Kontrolle von Bildung/beruflicher Qualifikation einen signifikanten Effekt auf die Einkommenshöhe ausübt (z.B.Pollmann-Schult 2009).

In der Logik von Giesselmann & Windzio (2012) würde es sich bei diesen Beispielen um Querschnittsfragestellungen handeln, die dadurch gekennzeichnet sind, dass Effekte zeitkonstanter Merkmale im Zentrum stehen. Die Bezeichnung zielt darauf ab, dass solche Merkmale wie Geschlecht und Geburtsland keine kausalen Prozesse dadurch ausüben können, dass sie sich innerhalb einer Person ändern. Der „Längsschnittcharakter“ fehlt somit. Bei Längsschnittfragestellungen stehen hingegen Einflussfaktoren im Vordergrund, die über eine intraindividuelle Änderung ihrer Werte Veränderungen auf der abhängigen Variablen auslösen können – also zeitveränderliche Variablen.

Giesselmann & Windzio (2012: 108) bieten u.a. entlang dieser Unterscheidung zwischen Querschnitts- und Längsschnittfragestellungen einen vereinfachten Entscheidungsbaum an, um zur Wahl der panelanalytischen Modellvariante zu kommen. Die Angemessenheit des gewählten Verfahrens, insbesondere die Frage, ob bei den zentralen unabhängigen Variablen „genug“ within-variation vorliegt oder ob der Einflusskoeffizient einer zeitkonstanten Variablen benötigt wird, ist letztlich eine theoretische Frage. Auch technische Hilfsmittel, wie statistische Testverfahren, dienen lediglich als Entscheidungsstütze und sollten nicht als Entscheidungs-Automatismus eine theoriegeleitete Auseinandersetzung des Forschers mit der Methodenwahl ersetzen (vgl. hierzu die Anwendung und die Grenzen des populären Hausman-Tests bei Giesselmann & Windzio 2012: 109-113).

Zusammenfassend ist aus der Perspektive des Sozialforschers, der mit Survey-Daten arbeitet, zu empfehlen, das FEM dem REM aufgrund dieser wertvollen Möglichkeit zur Reduktion der Verzerrung von Regressionskoeffizienten wenn möglich vorzuziehen. Diese Empfehlung gilt, wenn eine echte Längsschnittfragestellung vorliegt – wenn also die zentralen unabhängigen Variablen ausreichend within variation aufweisen und das Wissen um die Werte von Koeffizienten zeitkonstanter unabhängiger Variablen irrelevant für die Fragestellung ist. Dies ist in der Sozialforschung häufig der Fall, da häufig Kausalhypothesen auf Individualebene geprüft werden. Schließlich impliziert Kausalität die Wirkung zeit*veränderlicher* Merkmale.

Nur aus Mangel an den dazu passenden within Informationen wird im Falle von Querschnittdaten auf between Informationen ausgewichen. Auf Basis der Unterschiede zwischen Personen wird auf kausale Prozesse innerhalb von Personen ge-

schlossen. Da aber im Falle von Paneldaten within Informationen vorliegen, können diese statt der between Informationen genutzt werden, womit man näher an der Modellierung kausaler Prozesse ist.

Gelegentlich wird als Entscheidungshilfe REM vs. FEM auch auf den Entstehungsmechanismus der Daten verwiesen. Stellen die im Datensatz enthaltenen Personen die Realisierung einer *Zufalls*stichprobe dar, dann unterliegen die durch a_i charakterisierten Ausgangslagen der Objekte selbst einer zufälligen Auswahl. Es wurde mit a_i sozusagen zufällig eine persönliche Abweichung von der Regressionsgeraden aus der Grundgesamtheit aller möglichen persönlichen Abweichungen gezogen. In diesem Falle, der auf die meisten Panelstudien zutrifft (!), wäre eigentlich das REM vorzuziehen.

Denn nur die im REM vorgenommene Modellierung der a_i und somit „der Personen" als zufällig erlaubt Schlussfolgerungen von einer Zufallsstichprobe von Personen auf die dazugehörige Grundgesamtheit von Personen. Entlang dieser Logik eignet sich FEM eher für spezifische Objekte einer Grundgesamtheit, deren Auswahl keinem Zufallsprozess unterliegt. Dies ist vor allem dann gegeben, wenn Vollerhebungen zu kleinen Grundgesamtheiten vorliegen – z.B., wenn zu allen Mitgliedern des Bundestages Daten zu ihren politischen Aktivitäten vorliegen würden.

Allerdings wird die Relevanz dieser Unterscheidung entschärft durch ein abstrakteres Verständnis der Inferenzpopulation: Wird der Einfluss einer zeitveränderlichen Variablen x auf eine ebenfalls zeitveränderliche abhängige Variable y untersucht, dann interessiert in der Regel die Frage nach *intra*-individuellen Veränderungen: Wie wird eine Veränderung von x bei Personen auf *ihren* y-Wert wirken? Dass Antworten auf solche Fragen in einer Querschnittsregression notgedrungen nur aus between Informationen abgeleitet werden können, ändert nichts daran, dass eigentlich intra-individuelle Änderungen interessieren.

Wie oben bereits dargestellt, nennen Giesselmann & Windzio (2012) solche Forschungsfragen „Längsschnitt-Fragestellungen". In dieser Logik ist auch die Erweiterung des REM gegenüber dem FEM zu sehen: Kern der REM-Schätzung sind im Falle einer zeitveränderlichen unabhängigen Variablen x ebenfalls die within-Informationen. Die between-Informationen werden lediglich *zusätzlich* herangezogen, um die Effizienz der Schätzung zu verbessern.

In diesem Verständnis haben inferenzstatistische Tests in panelanalytischen Modellen für Längsschnittfragestellungen nicht die Aufgabe, auf eine große Grundgesamtheit von Personen, sondern *von Prozessen* zu schließen. Die entscheidende Frage ist: Kann ich die mithilfe von Paneldaten ermittelte intra-individuelle Auswirkung von

x auf y auf nicht in den Daten enthaltenen *Auswirkungen* verallgemeinern? Diese Auswirkungen können zwar auf andere Personen, aber genauso auch auf dieselben Personen bezogen werden. So gesehen ist die Anwendung von FEM mit Paneldaten vereinbar, bei denen die Teilnahme von Personen über eine Zufallsstichprobenziehung zustande kommt. Denn die Personen werden lediglich instrumentalisiert, um *Prozesse* zu untersuchen.

Zusätzlich ist, wie oben intensiv diskutiert wurde, im Falle von Survey-Daten stark anzuzweifeln, dass die mit den in die Stichprobe gezogenen Personen einhergehenden a_i-Werte unsystematisch (Erwartungswert Null) sind – auch wenn die Personen selbst nach einem Zufallsmechanismus in die Stichprobe gelangt sind.

Bei Querschnittsfragestellungen entfällt hingegen eine Entscheidung REM vs. FEM, da diese aufgrund der Kollinearität zeitinvarianter Merkmale mit den herausgerechneten a_i sowieso nicht mit dem FEM untersucht werden können.

Als Beispiel zur Verdeutlichung der Vor- und Nachteile von FEM und REM fungiert ein kleiner Datensatz in Tab. 2.1. Es ist eine Erweiterung des in 1.2 aufgeführten Beispiels auf fünf Personen und um potentielle unabhängige Variablen. Die abhängige Variable ist weiterhin der fiktive Produktivitätsscore in % („Produktivität“) dieser fünf Mitarbeiter einer Firma, gemessen zu je vier Zeitpunkten. Außerdem sind in Tab. 2.1 Werte von drei weiteren Variablen aufgeführt, wobei insbesondere die Dauer der Betriebszugehörigkeit („Dauer“) im Mittelpunkt der folgenden Ausführungen sein wird.[16]

Zuerst interessiert den Forscher also der Einfluss der Dauer auf die Produktivität. Entsprechend der Terminologie von Giesselmann & Windzio (2012) handelt es sich dabei um eine „Längsschnittfragestellung“, da ja beide Variablen zeitveränderlich sind. Schon nach Augenmaß lässt sich eine gegenläufige Zusammenhangsstruktur auf beiden Ebenen feststellen: Während auf der between-Ebene ein negativer Zusammenhang zu verzeichnen ist, scheint auf der within-Ebene eine zunehmende Dauer die Produktivität zu verbessern. Wird eine Regression der Produktivität auf die Dauer berechnet, ergeben sich die in Tab. 2.2 aufgeführten, nach Modellvariante differenzierten Schätzergebnisse.

Im pooled Modell wird die hierarchische Datenstruktur ignoriert. Da die between (co-)variation im Vergleich zur within (co-)variation deutlich stärker ausgeprägt ist, setzt sich der Zusammenhang durch, welcher auf between

[16] Die jeweils erste Spalte von Produktivität und Dauer gibt die Einzelmessungen an. Die jeweils zweite den individuellen Mittelwert.

i	t	**Produktivität**		**Dauer**		**Einarb**	**Abi**
1	1	58,8	62,65	0	1,5	1	1
	2	59,6		1		1	1
	3	65,2		2		1	1
	4	67		3		1	1
2	1	44	47,45	2	3,5	1	1
	2	45,6		3		1	1
	3	49		4		1	1
	4	51,2		5		1	1
3	1	42,1	42,33	7	8,5	1	0
	2	42,3		8		1	0
	3	42,3		9		1	0
	4	42,6		10		1	1
4	1	31,7	32,65	10	11,5	0	0
	2	32,1		11		0	0
	3	32		12		0	0
	4	33,6		13		0	0
5	1	24	24,53	18	19,5	0	0
	2	24,5		19		0	0
	3	24,7		20		0	0
	4	24,9		21		0	0

Tabelle 2.1: Beispiels-Datensatz

Unabhängige Variable	**FEM**	**pooled**	**REM1**	**REM2**
Dauer	1,304 ***	-1,836 ***	-0,0899	0,909 **
Einarb	-	-	-	32,379 ***

* p<.1; ** p<.05; *** p<.01

Tabelle 2.2: Regression der Produktivität in panelanalytischen Varianten

Ebene dominiert – und dieser ist negativ. Man könnte also (voreilig) daraus schließen, dass die Dauer einen negativen Einfluss auf die Produktivität ausübt. Heißt das, dass Mitarbeiter mit zunehmender Dauer der Betriebszugehörigkeit eine immer schlechtere Produktivität entwickeln? Ist das ein Zeichen von abnehmender Motivation? Ruhen sich Mitarbeiter auf ihrem Status aus, je länger sie „dabei“ sind? Mit solchen Schlussfolgerungen sollte man vorsichtig sein. Schaut man sich die FEM-Schätzung an, dann kehrt sich nämlich das Vorzeichen um. Es wird positiv. Bei alleiniger Betrachtung der within (co-)variation herrscht also ein positiver Einfluss der Dauer auf die Produktivität. Und entsprechend der Logik von Längsschnittfragestellungen ist genau das entscheidend: Wie wird sich die Produktivität eines Mitarbeiters im Verlaufe *seiner* Betriebszugehörigkeit entwickeln? Die deutlichen between-Unterschiede zwischen den Mitarbeitern sprechen hingegen weniger für einen kausalen negativen Einfluss der Dauer als mehr dafür, dass hier noch andere unberücksichtigte zeitkonstante Größen am Werk sind, welche diesen negativen Einfluss auf between-Ebene erzeugen. Diese sind im FEM herausgerechnet, weswegen der mit positivem Vorzeichen versehene Schätzer als weniger verzerrt gesehen werden kann, als der negativ gepolte pooled Schätzer. Und die REM-Lösung (REM1) ist in seiner Verzerrtheit zwischen der pooled- und dem FEM-Schätzung zu sehen: Diese hat, wie der pooled-Schätzer, ebenfalls ein negatives Vorzeichen. Dessen Betrag fällt aber deutlich niedriger aus, so dass der Koeffizient nicht mehr signifikant von Null verschieden ist. Das Gewicht G im REM-Schätzer fällt mit 0,0232 zwar recht niedrig aus. Das heißt, dass between-Informationen als Ergänzung zu den within Informationen nur einen relativ kleinen Beitrag zur REM-Schätzung leisten. Die within Informationen (individuellen Mittelwerte) sind als relativ zuverlässig zu deuten. Der REM-Schätzer ist somit näher am FEM- als am pooled-Schätzer. Dass er nicht positiv und signifikant ausfällt, liegt daran, dass die between-Unterschiede zwischen den Personen so hoch ausfallen und einen gegenläufigen Zusammenhang zwischen x und y aufweisen. Auch wenn diese between-Unterschiede mit nur einem geringen Gewicht in die Schätzung eingehen, erzeugen sie daher eine massive Diskrepanz zwischen dem REM- und dem FEM-Schätzer.

In solchen Fällen ist das FEM vorzuziehen. Denn wie bereits ausgeführt, signalisiert das deutlich abweichende Schätzergebnis im FEM, dass pooled- und REM-Schätzer beide massiv durch zeitkonstante „omitted variables“ verzerrt sind – wenn auch der REM-Schätzer im Vergleich zum pooled-Schätzer

schwächer verzerrt zu sein scheint.

Ein Vorzeichenwechsel unter „Kontrolle“ zeitkonstanter „omitted variables“ im FEM deutet in diesem Beispiel darauf hin, dass einige der relevantesten zeitkonstanten „omitted variables“ mit der Dauer *negativ* und mit der Produktivität *positiv* korreliert sind. Eines solcher Merkmale könnte ein erst seit ein paar Jahren eingeführtes intensives Einarbeitungsprogramm („Einarb“) bei der Einstellung von Mitarbeitern sein: Von diesem Programm könnten Mitarbeiter profitieren, welche erst seit wenigen Jahren im Betrieb sind (negativer Zusammenhang zwischen Einarb und Dauer). Die „betriebsjungen“ Mitarbeiter, welche eine solche intensive Einarbeitung genossen haben (im Datensatz die Personen 1-3), weisen *eben aufgrund ihrer Teilnahme an diesem Einarbeitungsprogramm* eine deutlich höhere individuelle Durchschnittsproduktivität auf (positiver Zusammenhang zwischen Einarb und Produkt). Einarb ist somit für die negative Scheinkorrelation zwischen Dauer und Produktivität auf between Ebene mit verantwortlich. Im FEM wurde der Einfluss von Einarb herausgerechnet, der Koeffizient ist also nicht mehr durch die negative Korrelation zwischen Einarb und Dauer verzerrt – wie das im REM- und pooled-Modell der Fall ist. Deutlich wird die Verzerrung an dem RE-Modell, in welchem die Variable Einarb explizit als zusätzliche unabhängige Variable eingeführt wird (REM2): Der Koeffizient wird (im Vergleich zu REM1) positiv und signifikant. Die Qualität der Schätzung verbessert sich deutlich, der REM-Koeffizient nähert sich dem FEM-Koeffizienten. Dass es immer noch eine Diskrepanz zwischen dem FEM- und dem REM2-Schätzer gibt, liegt vermutlich daran, dass es neben Einarb weitere relevante zeitkonstante Variablen gibt, deren Effekte im FEM automatisch herausgerechnet wurden, im REM hingegen nicht.

In dieser Längsschnittfragestellung (Einfluss der Dauer) ist das FEM die bessere Wahl. Würde hingegen die Frage nach dem Einfluss von Einarb im Zentrum stehen, dann läge eine Querschnittsfragestellung vor. Diese ließe sich mit dem FEM nicht beantworten, da Einarb ein zeitkonstantes Merkmal ist, welches innerhalb von Personen nicht streut, sich daher als Bestandteil von a_i vollständig aus der Regressionsgleichung herausrechnen würde. Hier kommt also nur das REM in Frage, welches aufgrund der korrekten Modellierung personeninterner Autokorrelationen gegenüber dem pooled Modell die bessere Wahl ist.

Problematisch ist das FEM, wenn der Effekt einer *quasi*-zeitkonstanten Grö-

ße im Vordergrund steht. Hier wird dies exemplarisch an der *quasi*-zeitkonstanten Variablen „Abi“ (Abi = 0: kein Abitur; Abi = 1: Abitur) verdeutlicht. Die Frage, ob das Abitur die Produktivität verbessert, wird vermutlich in den meisten Fällen eher als Querschnittsfrage verstanden: Sind Personen mit Abitur produktiver als andere Personen (und nicht: Führt das Erlangen des Abiturs bei einer Person dazu, dass diese Person produktiver wird?)? Würde man eine FEM-Regression der Produktivität auf Abi rechnen, dann würde lediglich die dritte Person berücksichtigt, da sie die einzige Person ist, bei der sich der Abi-Status im Zeitverlauf verändert. Sie hat vermutlich das Abitur auf einer Abendschule neben ihrem Job nachgeholt. Der FEM-Schätzer basiert auf den Informationen nur dieser einen Person und ist mit 0,3666662 identisch mit dem Ergebnis einer simplen Regression der Produktivität auf Abi, wenn alle Personen aus dem Datensatz entfernt werden, außer die dritte. Das kann aber nicht im Sinne des Forschers sein, wenn er den Einfluss des Abiturs auf die Produktivität als Querschnittsfrage formuliert hat. Denn dann ist ja gerade die between (co-)variation von Interesse, welche im FEM vollkommen ausgeklammert wird. Bei Querschnittsfragestellungen eignet sich also das REM eher. Allerdings stellt sich das „omitted-variables“-Problem im REM in gleicher Schärfe, wie bei Längsschnittfragestellungen.

2.4 Umsetzung in Stata und zusätzliche Erläuterungen zum Output

Die Berechnung eines FEM oder REM ist mit Stata einfach realisierbar. Dort sind diese Modelle explizit als Prozeduren implementiert und können mit einzeiligen Analysebefehlen angefordert werden. Die Bildung von Dummy-Variablen oder die within-Transformation im FEM bzw. die GLS-Transformation im REM müssen somit nicht manuell getätigt werden.

Allerdings müssen einige Schritte vorgenommen werden, ehe ein Paneldatensatz von Stata auch als panel- bzw. mehrebenenanalytischer Datensatz verstanden wird. Die meisten Roh-Datensätze aus Panelstudien sind nicht automatisch „Stata-ready“ beschaffen. So muss der Datensatz insb. mit dem Befehl „*reshape*“ in das sog. „lange Format“ gebracht werden. Es muss ferner mit „*xtset*“ definiert werden, welche Variable die einzelnen Personen und welche die Zeitpunkte identifiziert (zur Durchführung dieser Schritte vgl. Kohler

2008: 245f).

Ist dies getätigt, dann reicht ein Befehl aus, um ein FEM oder REM zu berechnen:

xtreg y x, fe

xtreg y x[, re]

Die erste Zeile steht für das FEM, die zweite für das REM. Die eckige Klammer signalisiert, dass ihr Inhalt auch weggelassen werden kann. Die Bezeichnungen y und x sind hierbei Platzhalter für beliebige Variablen. Wichtig ist, dass die abhängige Variable in der Reihenfolge zuerst notiert wird. Daraufhin können beliebig viele unabhängige Variablen (aber im Falle des FEM nicht (!!!) die Dummy-Variablen, nur die „inhaltlichen" unabhängigen Variablen) folgen.

An dieser Stelle sollen noch einige interessante Größen im Stata-Output der RE- und FE-Modelle besprochen werden. Es folgt eine Auflistung der Berechnungsformeln der wichtigsten Maßzahlen / Schätzungen.

Zunächst sind immer drei Arten von Determinationskoeffizienten angegeben. **„R-sq: within"** bezieht sich auf den Anteil der within-Varianz von y, welche durch die within-Varianz von x erklärt wird. Das Pendant dazu ist **„R-sq: between"**, bei der nur die between(Ko-)Variation von x und y berücksichtigt wird. **„R-sq: overall"** ist der Anteil der gesamten Varianz von y, welche durch x erklärt wird. Sie entspricht dem Determinationskoeffizienten in der pooled Regression, in der zwischen der Zeit- und der Personenebene nicht weiter unterschieden wird.

Die dazugehörigen Formeln in der unteren Auflistung zeigen, dass die verschiedenen Determinationskoeffizienten alle demselben Prinzip unterliegen: Es wird die jeweilige quadrierte Kovarianz von x und y in Relation gesetzt zu dem Produkt der korrespondierenden Varianzen beider Variablen. Dies entspricht der Quadrierung des Zählers und des Nenners des Korrelationskoeffizienten, aus dem sich ja im einfachen Falle einer unabhängigen Variablen direkt der Determinationskoeffizient errechnen lässt (eben durch die Quadrierung).

Einschränkend ist zu sagen, dass diese intuitiv zugänglichen Formeln nur für den Fall gelten, dass eine einzige unabhängige Variable vorliegt (und dass ferner die Analysestichprobe die Form eines balancierten Panels annimmt). In einem solchen Fall ist die Berechnung der drei Determinationskoeffizienten unabhängig davon möglich, ob mit den Daten ein RE- oder ein FE-Modell gerechnet wurde. Dies wird an der generellen Logik dieser drei Determinationskoeffizienten deutlich, die für Modelle mit beliebig vielen unabhängigen Variablen gilt:

Mit der Schätzung eines RE- oder eines FE-Modells erhält man, wie in der gewöhnlichen OLS-Regression, eine Schätzgleichung, mit der mithilfe der Werte der unabhängigen Variablen Vorhersagen über y-Werte getroffen werden können – und zwar für einzelne Personen i zu einzelnen Zeitpunkten t.[17] Solche Vorhersagewerte lassen sich in einer Variablen abspeichern, so dass für jede Person zu jedem Zeitpunkt der tatsächliche y_{it}-Wert dem geschätzten y_{it}-Wert ($\widehat{y}_{it}$) zugeordnet werden kann.

Wird die Korrelation zwischen y_{it} und $\widehat{y}_{it}$ berechnet und quadriert, erhält man den aus der OLS-Regression bekannten quadrierten multiplen Korrelationskoeffizienten, der gleichzeitig dem Determinationskoeffizienten entspricht, würde y_{it} auf $\widehat{y}_{it}$ regressiert werden. Mit dieser Größe wird ausdrückt, wie „nahe“ die Vorhersagewerte $\widehat{y}_{it}$ an den tatsächlichen Werten von y_{it} sind. Diese Größe entspricht dem „R-sq: overall“ im REM- oder FEM-Output von Stata.

Bei dieser Vorgehensweise wird die Paneldatenstruktur, also die Einbettung der Werte einer Person im Zeitverlauf in die Person selbst, ignoriert. Eine entsprechende Berücksichtigung findet dagegen in der Berechnung von „R-sq: within“ und „R-sq: between“ statt. Die Prozedur ähnelt der oben dargestellten für „R-sq: overall“, allerdings werden im Vorfeld sowohl die tatsächlichen y_{it}-Werte, als auch die geschätzten $\widehat{y}_{it}$-Werte between- bzw. within-transformiert.

Konkret werden für die Berechnung von „R-sq: between“ zunächst y_{it} und $\widehat{y}_{it}$ zu individuellen Mittelwerten verdichtet. „R-sq: between“ entspricht dann der quadrierten Korrelation zwischen den auf Basis des geschätzten Modells (REM oder FEM) vorhergesagten und den tatsächlichen individuellen Mittelwerten. Vereinfachend gesagt kann „R-sq: between“ eine Antwort auf die Frage geben, wie gut das geschätzte Modell individuelle Mittelwerte vorhersa-

[17] Dabei werden zur Vorhersage nur die explizit spezifizierten unabhängigen Variablen hinzugezogen, nicht zusätzlich die individuellen Regressionskonstanten (was in Stata der Option „xb“ und nicht „xbu“ des Befehls „predict“ nach „xtreg“ entspricht – s. „h xtreg postestimation“)

gen kann (eine Frage, die auch dann gestellt werden kann, wenn es, wie beim FEM, gar nicht beabsichtigt ist, individuelle Mittelwerte vorherzusagen).

Für „R-sq: within“ werden y_{it} und $\widehat{y}_{it}$ within-transformiert, indem jeweils die Differenzen der Einzelwerte von den individuellen Mittelwerten gebildet werden. Die quadrierte Korrelation zwischen den tatsächlichen und den vorhergesagten within-transformierten Werten ist „R-sq: within“. Der Wert steht für das Ausmaß, mit dem sich die Schwankungen von y innerhalb von Personen durch das Regressionsmodell vorhersagen lassen.

Zu erwähnen ist, dass sich „R-sq: overall“ nicht additiv aus den beiden anderen Größen zusammensetzt. In dem oberen Beispiel (Abschnitt 2.3) ist er sogar niedriger, als das between- bzw. within-Bestimmtheitsmaß. Dies liegt daran, dass auf der between-Ebene ein positiver und auf der within-Ebene ein negativer Zusammenhang zwischen x und y besteht. Auf der undifferenzierten Gesamtebene heben sich diese gegenläufigen Zusammenhänge teilweise auf, so dass „R-sq: overall“ niedriger ausfällt.

Fällt ferner „R-sq: between“ deutlich höher aus, als „R-sq: within“, kann dies als weiterer Hinweis dafür gedeutet werden, dass unberücksichtigte zeitkonstante Drittvariablen die Koeffizienten bestehender Variablen verzerren, sobald between-Informationen in der Schätzung berücksichtigt werden (wie dies im Gegensatz zum FEM beim REM der Fall ist). Keinesfalls sollte in einem solchen Fall voreilig der Schluss gezogen werden, dass ein Modell, das auf within-Informationen beruht, wie das FEM, ungünstig ist, da ja schließlich das Erklärpotential auf between-Ebene höher ist. Im Falle unberücksichtigter zeitkonstanter Drittvariablen wäre dies ein Trugschluss.

Dass die Werte der drei Determinationskoeffizienten im Falle einer einzigen unabhängigen Variablen unabhängig davon sind, ob ein REM oder ein FEM vorliegt, ist damit zu erklären, dass sich die Regressionsgleichung zwischen REM und FEM nur durch den Regressionskoeffizienten für x und die allgemeine Regressionskonstante unterscheiden. Somit stehen y-Vorhersagewerte auf Basis des REM im proportionalen Verhältnis zu den y-Vorhersagewerten auf Basis des FEM, was in beiden Fällen zu einer identischen Korrelation mit y (und damit zu einer identischen quadrierten Korrelation) führt. Auch die Korrelation der between- bzw. within-transformierten Versionen von y mit ihren Vorhersagependants fällt in diesem Fall unabhängig davon aus, ob ein REM oder FEM vorliegt.

Ab einer zweiten unabhängigen Variablen ist dies aufgrund der gestiegenen Komplexität der Schätzgleichung für $\widehat{y}_{it}$ nicht mehr der Fall. Ab dann kann sich die Korrelation zwischen y_{it} und $\widehat{y}_{it}$ je nach Modellvariante unterscheiden. Es wird also möglich, festzustellen, wie gut die jeweilige Modellvariante die y-Werte in unterschiedlich transformierten Varianten vorhersagen kann. Auch hier sollte die Frage kritisch gestellt werden, ob die Erklärkraft auf einzelnen Ebenen tatsächlich von den unabhängigen Variablen oder teilweise auch von unberücksichtigten Drittvariablen ausgeht.

Für die Koeffizienten lassen sich Standardfehler berechnen und ein Signifikanz-Test durchführen. Die Vorgehensweise und die Interpretation ist völlig deckungsgleich mit der einer gewöhnlichen Regression. Dies gilt auch für die F-Teststatistik (oberer der beiden F-Werte) im FEM und die „ Wald chi2“-Statistik im REM. Beide testen die Nullhypothese, inwieweit *alle* Koeffizienten des Modells aus einer Population kommen, in der alle korrespondierenden Parameter dem Wert 0 entsprechen.

Im FEM ist ferner eine zweite F-Teststatistik angegeben: **„F test that all u_i=0“**.[18] Mithilfe dieser wird geprüft, ob die individuellen Regressionskonstanten a_i in ihrer Gesamtheit aus einer Population kommen, in welcher all diese Konstanten Null betragen. Das Test-Ergebnis ist im Falle einer Regression ohne unabhängige Variablen identisch mit dem gewöhnlichen F-Test einer Regression, wenn das FEM mithilfe von Dummy-Variablen spezifiziert wird. Kann die Nullhypothese beibehalten werden, dann ist die Einführung von a_i in die Regression unnötig. Die individuellen Regressionskonstanten würden keinen signifikanten Erklärungsbeitrag leisten.

Das Pendant zu diesem F-Test ist im REM (nach der GLS-Methode) der sog. „Lagrange multiplier test for random effects“ von Breusch und Pagan (1980). Er ist in Stata als „Postestimation Command“[19] implementiert und kann mit dem Befehl „xttest0“ angefordert werden.

Die Größe **„corr(u_i, Xb)“** entspricht der Korrelation zwischen den individuellen Regressionskonstanten a_i einerseits und der mithilfe von x geschätz-

[18] In Stata werden die individuellen Regressionskonstanten a_i mit „u_i“ bezeichnet.

[19] Das ist ein Befehl, der nach einer in Stata vorgenommenen Modellrechnung angefordert werden kann. Er bezieht sich immer auf die zuletzt durchgeführte Modellrechnung, derer Schätzergebnisse von Stata immer intern automatisch gespeichert werden.

ten y-Werte andererseits. Letztere sind, wie in einer gewöhnlichen Regression, gegeben über die lineare Kombination aus unabhängigen Variablen und Regressionskoeffizienten – hier im einfachen Falle: $\widehat{y}_{it} = \widehat{b} \cdot x_{it}$. Die Korrelation besagt, wie stark die (gemeinsame Wirkung der) x-Variablen mit den individuellen Ausgangslagen (korreliert) korrelieren. Bei Vorliegen nur einer Variablen x vereinfacht sich der Sachverhalt: Der Betrag[20] der Korrelation „corr(u_i, Xb)“ entspricht dann der bivariaten Korrelation zwischen x und a_i. Diese Korrelation lässt sich nur im FEM berechnen, da dort a_i als fixe Größe behandelt wird, folglich als „fester“ Wert vorliegt. Im REM gilt hingegen a priori „corr(u_i, Xb)=0“. Denn im REM werden die individuellen Ausgangslagen a_i zu der als Zufallsvariable aufgefassten Fehlerkomponente gezählt, welche bei gegebenem x-Wert einen Erwartungswert von 0 hat. Der Fehler wird als zufällig um die perfekte Beziehung $y = b_r \cdot x$ streuend angenommen, ist folglich mit x unkorreliert. Im REM ist somit „corr(u_i, Xb)“ nicht berechenbar, denn man kann über Zufallsvariablen lediglich a priori Annahmen treffen, aber keine deskriptiven Statistiken zu ihnen berechnen.

Ein hoher „corr(u_i, Xb)“-Wert im FEM spricht gegen die Anwendung des REM. Wie schon mehrfach ausgeführt, ist die Annahme „corr(u_i, Xb)=0“ verletzt, wenn relevante zeitkonstante Variablen im Modell fehlen, die mit x korreliert sind. Eine hohe Korrelation „corr(u_i, Xb)“ im FEM ist als Indiz dafür zu sehen, dass eine solche Annahmeverletzung vorliegt.

Die Angaben zu **„sigma_u“** ($= \sqrt{V(a)}$), **„sigma_e“** ($= \sqrt{V(\epsilon)}$) und **„rho“** ($= \rho$) geben die im REM eingeführte Varianz der Fehlerkomponenten und ihr Verhältnis zueinander wieder (s. Gl. 2.26 und 2.27 sowie dazugehörige Ausführungen). ρ sagt aus, wie stark der zeitkonstante Anteil des Residuums am Gesamtresiduum ausfällt, wie ähnlich sich folglich die Residuen ein und derselben Person sind. Ein hoher ρ-Wert deutet also an, dass vor allem zeitkonstante Merkmale im Modell fehlen.

ρ lässt sich ebenfalls im FEM berechnen, macht dort aber interpretativ wenig Sinn, da die Varianz einer festen Größe a_i ($V(a_i)_f$, dessen Quadratwurzel in Stata ebenfalls als **„sigma_u“** bezeichnet wird) in Beziehung zu einer Fehlervarianz, nämlich der von ϵ gesetzt wird (vgl. Giesselmann & Windzio 2012: 47).

[20] Die Multiplikation mit $\widehat{b}_f$ kann lediglich das Vorzeichen der Korrelation von x und a_i ändern, aber nicht die Stärke

Allerdings soll hier kurz auf einen Spezialfall eingegangen werden, nämlich wenn ρ im FEM deutlich höher ausfällt, als im REM:

Ein im REM vorherrschender niedrigerer ρ-Wert im Vergleich zu FEM ist auf ein niedrigeres $V(a)$ in Relation zu $V(a_i)_f$ zurückführen. Denn die within-Reststreuung $V(\epsilon)$ nimmt in beiden Modellen immer die gleiche Größe an. $V(a)_f$ im FEM zeigt an, wie viel Varianz a_i gegeben x aufweist. Das sind also die totalen inter-individuellen Unterschiede in y, welche auch nach Kontrolle des auf Basis der within covariation zwischen x und y errechneten Einflusses von x über bleiben. $V(a)$ im REM steht hingegen für den Rest dieser between Unterschiede von y, welche, neben der within covariation zwischen x und y, auch durch eine zusätzliche Berücksichtigung der *between* covariation zwischen x und y nicht aufgeklärt werden konnten. Je kleiner also $V(a)$ im Vergleich zu $V(a_i)_f$, umso größer der Anteil der im FEM nach Herausrechnen des auf *within* Informationsbasis geschätzten Einflusses von x verbleibenden *between* Unterschiede, welche durch die *between* covariation zwischen x und y erklärt werden können. Eine hohe Diskrepanz zwischen $V(a_i)_f$ und $V(a)$ deutet somit daraufhin, dass die between covariation zwischen x und y anders beschaffen ist, als die korrespondierende within covariation. Wie kann sonst der within Zusammenhang zwischen x und y so viel inter-individuelle Varianz von y unerklärt lassen, welche die *between* covariation zwischen x und y wiederum erklären kann?

Mal angenommen, es gibt einen Fall mit recht starker und einen mit einer eher schwachen Diskrepanz zwischen $V(a_i)_f$ und $V(a)$.[21] Der Fall mit der starken Diskrepanz würde sich dann ergeben, wenn die within- und die between covariation zwischen x und y unterschiedlich beschaffen wären. Das ist im Beispiel in Abs. 2.3 der Fall, bei dem es zwischen x und y auf between Ebene einen negativen und auf within Ebene einen positiven Zusammenhang gibt. Der FEM-Schätzer von x ist folglich positiv, durch den negativen between Zusammenhang weichen aber die individuellen Durchschnittsleistungen gegeben x (a_i) relativ stark von der within Regressionsgeraden ab. Der im FEM ausschließlich betrachtete within Zusammenhang zwischen x und y ist folglich nicht so gut in der Lage, den gegenläufigen between Zusammenhang zwischen x und y „einzufangen“. $V(a_i)_f$ fällt deutlich höher aus, als der Rest dieser individuellen Abweichungen $V(a)$, wenn auch

[21] λ kann übrigens in beiden Fällen gleich sein. Denn λ setzt die beiden Komponenten *des REM*, $V(a)$ und $V(\epsilon)$, ins Verhältnis. So kann es sein, dass die between Beziehung zwischen x und y in beiden Fällen gleich stark ist. Auch kann die within Beziehung zwischen x und y in beiden Fällen gleich stark sein. In beiden Fällen ist λ bzw. G und somit der Gewichtungsfaktor der between-Informationen im REM gleich. Die between Informationen sind somit in beiden Fällen gleich gut geeignet, um die FEM-Schätzkomponenten in der REM-Schätzung zu ergänzen.

die between covariation zwischen x und y betrachtet worden wäre (wie im REM). Wären hingegen die Zusammenhänge zwischen x und y auf beiden Ebenen ähnlich, dann würde der Anteil der between variation von y im FEM niedriger ausfallen, welcher durch den *within* Zusammenhang zwischen x und y *nicht*, dafür aber durch den *between* Zusammenhang zwischen x und y erklärt werden kann. $V(a_i)_f$ wäre somit näher an $V(a)$ dran, da die in der between covariation zwischen x und y enthaltene Information als Ergänzung zu den within Informationen deutlich redundanter ausfallen würde.

Festzuhalten ist, dass eine relativ große Diskrepanz zwischen $V(a_i)_f$ und $V(a)$ darauf hindeutet, dass die Zusammenhangsstruktur zwischen x und y auf beiden Ebenen unterschiedlich ausfällt. Wie bereits erläutert, sind im Sinne einer Längsschnittfragestellung between-Unterschiede, welche durch den within Zusammenhang zwischen x und y nicht eliminiert wurden, erklärungsbedürftig. Eine hohe Diskrepanz zwischen $V(a_i)_f$ und $V(a)$ kann, wie im Beispiel in Abs. 2.3 der Fall, ein Hinweis auf zeitkonstante „omitted variables“ sein, welche positiv (negativ) mit der abhängigen und negativ (positiv) mit den unabhängigen Merkmalen korrelieren. Das REM ist bei einer hohen Diskrepanz zwischen $V(a_i)_f$ und $V(a)$ nicht etwa die bessere Wahl, da es die im Vergleich zur within covariation deutlich anders beschaffene between covariation zwischen x und y mehr berücksichtigt. Im Gegenteil. Mit der Berücksichtigung der between covariation zwischen x und y „schleicht“ sich im REM die korrespondierende inter-individuelle Fehlervarianz $V(a)$ ins Residuum. Und gerade eine solche gegenläufige Zusammenhangsstruktur zwischen x und y auf beiden Ebenen deutet eher daraufhin, dass zeitkonstante relevante Variablen fehlen, welche diese between covariation zwischen x und y erklären (somit auch mit x korreliert sind und daher die Schätzung von β_r verzerren). Das äußert sich auch in den Koeffizienten: Wenn nämlich der Zusammenhang zwischen x und y auf between-Ebene anders beschaffen ist, als auch within-Ebene, gleichzeitig aufgrund eines nur mäßig hohen Gewichts λ between-Informationen relativ stark in die REM-Schätzung eingehen, dann entsteht eine bedeutsame Diskrepanz zwischen dem FEM- und dem REM-Schätzer. Entsprechend der Logik des Hausman-Tests (Giesselmann & Windzio 2012: 109-113, Rabe-Hesketh et al. 2012b: 157f.) spricht eine solche Diskrepanz immer dafür, dass der REM-Schätzer verzerrter ist, als der FEM-Schätzer (und nicht umgekehrt). Im FEM erfährt man zwar auch nicht, *welche* zeitkonstante Variablen

relevant sind. Aber immerhin kann, im Vergleich zum REM, deren gemeinsamer Effekt herausgerechnet werden. Die Schätzung des Effektes von x ist durch solche Variablen nicht mehr verzerrt.

In den folgenden Tabellen sind die Formeln von einigen der hier besprochenen Größen aufgeführt. Sie gelten nur für den eingeschränkten Fall, welcher zu Beginn des Absatzes 2.2 eingeführt wird. Die im Verlaufe dieses Skripts eingeführten Größen und Symbole werden als bekannt vorausgesetzt.

Determinationskoeffizienten und Korrelationen		
Stata-Bez.	**Formel**	**Erläuterung**
R-sq: within	$\frac{[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)(y_{it}-\overline{y}_i)]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\overline{x}_i)^2\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\overline{y}_i)^2}$	Anteil der within-Varianz von y, die durch x erklärt wird
R-sq: between	$\frac{[\sum_{i=1}^{n}(\overline{x}_i-\overline{x})(\overline{y}_i-\overline{y})]^2}{\sum_{i=1}^{n}(\overline{x}_i-\overline{x})^2\sum_{i=1}^{n}(\overline{y}_i-\overline{y})^2}$	Anteil der between-Varianz von y, die durch x erklärt wird
R-sq: overall	$\frac{[\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\overline{x})(y_{it}-\overline{y})]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\overline{x})^2\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\overline{y})^2}$	Anteil der gesamten Varianz von y, die durch x erklärt wird
corr(u_i, Xb)	$\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(a_i-\overline{a})(\widetilde{y}_{it}-\overline{\widetilde{y}})}{\sqrt{\sum_{i=1}^{n}(a_i-\overline{a})^2}\sqrt{\sum_{i=1}^{n}\sum_{t=1}^{T}(\widetilde{y}_{it}-\overline{\widetilde{y}})^2}}$	Korrelation zwischen den individuellen Regressionskonstanten und den geschätzten y-Werten, gegeben x. Die Formel bezieht sich nur auf das FEM! Im REM wird der Korrelationswert 0 per Annahme festgelegt.

Fehlervarianzen im FEM		
Stata-Bez.	**Formel**	**Erläuterung**
sigma_e	$\sqrt{\widehat{V}(\epsilon)} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon^2}{n(T-1)-1}}$	Schätzformel für die Standardabweichung des Residuums (analog zum σ in der einfachen Regression)
sigma_u	$\sqrt{V(a_i)} = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \overline{a})^2}{n-1}}$	Standardabweichung der individuellen Regressionskonstanten. Da diese als fixe Werte betrachtet werden, ist dies keine Schätzformel, sondern eine eindeutige Berechnung!
rho	$\widehat{\rho} = \frac{V(a_i)}{V(a_i) + \widehat{V}(\epsilon)}$	Geschätzter Anteil der Varianz von a_i an der Summe der Varianz von a_i und ϵ

Fehlervarianzen im REM

Stata-Bez.	Formel	Erläuterung
sigma_e	$\sqrt{\widehat{V}(\epsilon)} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\epsilon^2}{n(T-1)-1}}$	Schätzformel für die St.abw. des Rest-Residuums (muss für das Gewicht G bzw. λ schon *vor* der REM-Schätzung bekannt sein; hierzu kann diese Größe aus dem FEM entnommen werden)
sigma_u	$\sqrt{\widehat{V}(a)} = \sqrt{\{\sum_{i=1}^{n}(\overline{y}_i-\overline{y})^2 - \frac{[\sum_{i=1}^{n}(\overline{x}_i-\overline{x})(\overline{y}_i-\overline{y})]^2}{\sum_{i=1}^{n}(\overline{x}_i-\overline{x})^2}\}\frac{1}{n-2} - \frac{\widehat{V}(\epsilon)}{T}}$	Schätzformel für die Standardabweichung der über Individuen konstanten Fehlerkomponente a_i
rho	$\widehat{\rho} = \frac{\widehat{V}(a)}{\widehat{V}(a)+\widehat{V}(\epsilon)}$	Geschätzte Korrelation zwischen zwei Residuen u ein und derselben Person / Geschätzter Anteil der „between“-Varianz des Residuums an der gesamten Varianz des Residuums

mit:

$\overline{a} = \frac{1}{n} \sum_{i=1}^{n} a_i$

$\widetilde{y}_{it} = x_{it} \cdot \widehat{b}_f$

$\overline{\widetilde{y}} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \widetilde{y}_{it}$

2.5 Ausblick auf weiterführende Verfahren

to be continued...

Tabellenverzeichnis

Literaturverzeichnis

[1] Allison, P. D.: *Fixed Effects Regression Models.* Sage, 2009

[2] Andress, H. ; Golsch, K. ; Schmidt, A. W.: *Applied Panel Data. Analysis for Economic and Social Surveys.* Springer, 2014

[3] Arminger, G. ; Müller, F. : *Lineare Modelle zur Analyse von Paneldaten.* Westdeutscher Verlag, 1990

[4] Auer, L. v.: *Ökonometrie. Eine Einführung.* Springer Verlag, 2007

[5] Backhaus, K. ; Erichson, B. ; Plinke, W. ; Weiber, R. : *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 13. Auflage.* Springer Verlag, 2011

[6] Baltagi, B. H.: *Econometric analysis of panel data.* Wiley, 2008

[7] Benninghaus, H. : *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler. 11. Auflage.* VS Verlag, 2007

[8] Bortz, J. ; Schuster, C. : *Statistik für Human- und Sozialwissenschaftler. 7. Auflage.* Springer Verlag, 2010

[9] Brüderl, R. : Kausalanalyse mit Paneldaten. In: Wolf, C. (Hrsg.) ; Best, H. (Hrsg.): *Handbuch der sozialwissenschaftlichen Datenanalyse.* VS Verlag, 2010, S. 963–994

[10] Breusch, T. ; Pagan, A. R.: The Lagrange multiplier test and its applications to model specification in econometrics. In: *Review of Economic Studies* 47 (1980), S. 239–253

[11] Engel, U. ; Reinecke, J. : *Panelanalyse. Grundlagen - Techniken - Beispiele.* Walter de Gruyter Verlag, 1994

[12] Faulbaum, F. : Panelanalyse im Überblick. In: *ZUMA-Nachrichten* (1988), Nr. 23, S. 26–44

[13] Frees, W. : *Longitudinal and Panel Data.* University Press, 2004

[14] Giesselmann, M. ; Schmidt-Catran, A. : *Wechselwirkungen in Fixed Effects Analysen.* Unveröffentlichte Präsentationsfolien, 38. Kongress der Deutschen Gesellschaft für Soziologie in Bamberg, 2016

[15] Giesselmann, M. ; Windzio, M. : *Regressionsmodelle zur Analyse von Paneldaten.* VS Verlag, 2012

[16] Hsiao, C. : *Analysis of Panel Data. Third Edition.* University Press, 2014

[17] Kalter, F. : Auf der Suche nach einer Erklärung für die spezifischen Arbeitsmarktnachteile von Jugendlichen türkischer Herkunft. Zugleich eine Replik auf den Beitrag von Holger Seibert und Heike Solga: "Gleiche Chancen dank einer abgeschlossenen Ausbildung? (ZfS 5/2005)". In: *Zeitschrift für Soziologie* 35 (2006), S. 144–160

[18] Kessler, R. ; Greenberg, D. : *Linear Panel Analysis. Models of Quantitative Change.* Academic Press, 1981

[19] Kohler, U. ; Kreuter, F. : *Datenanalyse mit STATA. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung.* Oldenbourg Verlag, 2008

[20] Langer, W. : *Mehrebenenanalyse. Eine Einführung für Forschung und Praxis. 2. Auflage.* VS Verlag, 2009

[21] Langer, W. : Mehrebenenanalyse mit Querschnittsdaten. In: Wolf, C. (Hrsg.) ; Best, H. (Hrsg.): *Handbuch der sozialwissenschaftlichen Datenanalyse.* VS Verlag, 2010, S. 963–994

[22] Opp, K. ; Schmidt, P. : *Einführung in die Mehrvariablenanalyse. Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen.* Rowohlt, 1976

[23] Pollmann-Schult, M. : Geschlechterunterschiede in den Arbeitswerten. Eine Analyse für die alten Bundesländer 1980-2000. In: *Zeitschrift für ArbeitsmarktForschung* 42 (2009), S. 140–154

[24] Rabe-Hesketh, S. ; Skrondal, A. : *Multilevel and longitudinal modeling using Stata. Second Edition.* StataCorp LP, 2008

[25] Rabe-Hesketh, S. ; Skrondal, A. : *Multilevel and longitudinal modeling using Stata. Third Edition, Volume I: Continuous Responses.* StataCorp LP, 2012a

[26] Rabe-Hesketh, S. ; Skrondal, A. : *Multilevel and longitudinal modeling using Stata. Third Edition, Volume II: Categorical Responses, Counts, and Survival.* StataCorp LP, 2012b

[27] Schnell, R. ; Hill, P. ; Esser, E. : *Methoden der empirischen Sozialforschung. 9. Auflage.* Oldenbourg, 2011

[28] StataCorp (Hrsg.): *Longitudinal-Data/Panel-Data Reference Manual.* Stata Press, 2011

[29] Verbeek, M. : *A Guide To Modern Econometrics. 3. Edition.* Wiley, 2008