DISKUSSIONSBEITRAG

aus der Fakultät für
WIRTSCHAFTSWISSENSCHAFTEN

der

UNIVERSITÄT DUISBURG - ESSEN
Campus Essen

Nr. 186
Januar 2011

Verlaufsanalysen (Panelerhebungen) in der Statistik: Warum und wie?

Peter von der Lippe

Universitätsstraße 12
45117 Essen

Peter von der Lippe

Verlaufsanalysen (Panelerhebungen) in der Statistik: Warum und wie?

Zusammenfassung[1]

Man findet in der Literatur mehr oder weniger ausführliche Darstellungen von Problemen der Erhebung und Auswertung von Wiederholungsbefragungen in zwei sehr verschiedenen Bereichen, nämlich der Bevölkerungsstatistik, empirischen Sozialforschung und Meinungsforschung usw. einerseits und in der Ökonometrie andererseits. Dabei entsteht der Eindruck als haben diese beiden Arten der Darstellung überhaupt nichts miteinander zu tun. Im Folgenden wird versucht, was über Panel zu sagen ist in ein beides umfassendes System zu bringen und in einem elementaren und leicht lesbaren Text Zusammenhänge aufzuzeigen. Dabei stellte sich heraus, dass in der Tat nicht sehr viele Berührpunkte zwischen den verschiedenen Arten, sich mit Paneldaten zu beschäftigen zu bestehen scheinen.

1. Was sind Panelerhebungen?

Werden im Zeitablauf (z.B. über einige aufeinander folgende Jahre) die gleichen Einheiten (z.B. Haushalte, Unternehmen, einzelne Personen, z.B. bei Meinungsbefragungen) wiederholt befragt spricht man von einem Panel oder einer Verlaufsanalyse. Die Besonderheit dieser Wiederholungsbefragung ist, dass versucht wird im Zeitablauf (über T Perioden) die gleichen N Einheiten (also den gleichen "Querschnitt" zu erheben. Die Terminologie hat sich in diesem Bereich etwas geändert. Früher unterschied man drei Arten von Wiederholungsbefragungen (vgl. Übersicht 1a)

1. **Zeitreihenanalysen** mit wiederholten Querschnitten (Erhebungen der jeweils zu einer Beobachtungsgesamtheit gehörenden Einheiten ohne besondere Berücksichtigung von Strukturveränderungen durch Ein- und Austritte von Einheiten);
2. wiederholte Befragung der Mitglieder eines strukturell gleich zusammengesetzten sog. "**Panels**" (das im Zeitablauf allerdings aus unterschiedlichen Einheiten bestehen kann; es wird nur darauf geachtet, dass sich die *Struktur* der Gesamtheit in Bezug auf bestimmte Merkmale, z.B. Alter, Geschlecht usw. nicht verändert);

[1] Anlass dieser Arbeit waren Diskussionen im Rahmen meiner Tätigkeit im wiss. Beirat für das (Arzt-) Praxispanel des Zentralinstituts für die kassenärztliche Versorgung (ZI) in Berlin. Ein Teil des folgenden einführenden Textes (Abschn. 2.3 und 4 betreffend) wurde zusammen mit einem Anwendungsbeispiel (chines. Unternehmen) bereits im Febr. 2009 Studenten zum Download auf meiner Homepage zur Verfügung gestellt. Der Anlass hierfür war eine Zusammenarbeit mit Herrn Kollegen Markus Taube bei der Auswertung von Panel-Daten über fast 3000 chinesische Unternehmen. Bei diesen Arbeiten wurde ich von Herrn *Jens Mehrhoff* (Deutsche Bundesbank) sehr unterstützt. Von ihm habe ich viel über Panelökonometrie gelernt.

2. wiederholte Befragung immer der gleichen Einheiten (individualisiertes Erhebungsverfahren, auch [echte] **Längsschnittsanalyse** oder **Verlaufsanalyse** [longitudinal survey] genannt)[2] im Zeitablauf.

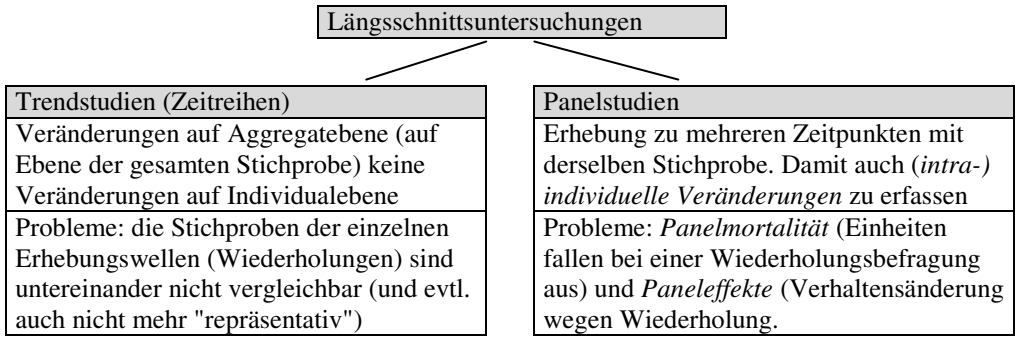
Übersicht 1: Varianten der Erhebung und Darstellung von Abläufen

a) Varianten von Wiederholungserhebung *(früher übliche Differenzierungen)*

	Zeitreihen mit wiederholen Querschnittserhebung)	Panelbefragung	echte Längsschnittserhebung
Untersuchungsgesamtheit	zum Stichtag im Bestand anwesende Einheiten	*strukturell* gleich zusammengesetzte Masse [a)]	*personell* (individuell) identische Einheiten
beobachtet wird	Bestand u. evtl. summarisch die Zu- und Abgänge (netto oder [selten] brutto)	auf Beobachtungszeitpunkt bezogene Merkmale (z.B. Meinungen	individuelle Verweilzeiten [b)] zwischen Veränderungen (Ereignissen)
Daten gleicher Individuen zu versch. Zeiten	werden nicht zusammengeführt; nur Fortschreibung des Bestands	nicht *individuell* zusammengeführt (nur Vergleich der Verteilungen)	werden individuell (für jede Person) zusammengeführt (echte Kohorte)
was ist berechenbar und interpretierbar?	Zeitreihe des Bestands, Durchschnittsbestand und *durchschnittliche* Verweildauer	Erklärung von Veränderungen mit gruppenspezifischen Einflüssen der Panelgesamtheit	auch *Verteilung* der Verweildauer, Bestand kontinuierlich [c)], Schwundquote (Austritte [d)])

a) Ausscheidende Individuen werden durch solche mit gleicher Merkmalsausprägung (hinsichtlich der strukturbestimmenden Merkmale) ersetzt, so dass die Struktur der Gesamtheit in Bezug auf bestimmte Merkmale gleich bleibt. Bei Querschnittsanalysen wird die Gesamtheit der in der Periode t "Anwesenden" betrachtet.
b) Beobachtungen von Ereignissen sind unabhängig von vorgegebenen Beobachtungsintervallen, also kontinuierlich möglich (alle Merkmalsveränderungen - oder allgemein "Ereignisse" - werden mit dem Zeitpunkt zu dem sie eintreffen und der Einheit, die davon betroffenen ist notiert).
c) Der Bestand ist kontinuierlich festzustellen (bei Querschnitten nur diskontinuierlich), wenn die Kohortenanalyse vollständig ist (alle jeweils gegenwärtig existierenden Kohorten umfasst).
d) auch zwischen Zählungsterminen.

b) Varianten von Wiederholungserhebung *(eine inzwischen wohl etwas üblichere Differenzierungen)*



Es ist heutzutage üblich (vgl. Übers. 1b), auch im Falle von Zeitreihen von "Längsschnitt" zu sprechen und keine Unterschiede zwischen den Fällen 2 und 3 zu machen[3] (und in beiden Fällen von Panel zu sprechen). [4]

[2] Das Ideal einer solchen Erhebung ist eine Kohortenanalyse, wie sie z.B. im Nationalen Bildungspanel in Deutschland vorgesehen. In den Jahren 2009 bis 2012 wurden sechs Startkohorten (Neugeborene, 4-jährige Kindergartenkinder, Fünftklässler, Neuntklässler, Studienanfänger und 23- bis 65-Jährige) mit insgesamt mehr als 60.000 Personen gezogen. Die Stichprobenziehungen orientierten sich sowohl an den Übergängen im Bildungssystem als auch an den Übergängen zwischen Bildungssystem und Arbeitsmarkt. Entscheidend ist, dass *dieselben* Panelteilnehmer (panelists) über einen längeren Zeitraum regelmäßig befragt werden.

[3] In der Ökonometrie spricht man im Falle Nr. 3 - also bei auch "personell" gleichbleibenden Gesamtheiten von einem "balanced panel"

[4] Im heute üblichen etwas anderen Sprachgebrauch (Übers. 1b) wird (wie in der Ökonometrie) oft von "Längsschnitt" im Sinne einer Zeitreihe gesprochen, also auch dann wenn sich die Struktur der Anwesenden permanent

Im Unterschied zur Zeitreihenbetrachtung (repeated observations), bei der sich die Struktur der erfassten (Teil-)Gesamtheit zwischen den Erhebungszeitpunkten ändern kann (so dass in den Ergebnissen "echte" und strukturell bedingte Veränderungen schwer oder gar nicht zu unterscheiden sind) gilt es bei einem Panel den Einfluss von Strukturveränderungen auszuschalten indem man versucht die Struktur des Panels konstant zu halten. In der Literatur werden für ein Panel meist folgende "Vorteile aufgeführt, die

1. Betrachtung *individueller* ("intra-individueller") Verläufe und Verweildauern, nicht nur von Veränderungen auf Aggregatebene
2. Befragungen nach Meinungen der gleichen Person zum gleichen Thema im Zeitablauf (Analyse von Meinungsänderungen) oder zur Veränderung des Sozialstatus (Analysen des sozialen Wandels), oder z.B. des Gesundheitszustands (individueller Verlauf einer Krankheit und Therapie) usw. um eine "echte" (vom Einfluss von Strukturveränderungen bereinigte) Veränderung herauszuarbeiten,
3. Analysen, in denen die Zeit selber das interessierende Merkmal ist, z.B. die Zeitpunkte von Übergängen zwischen Zuständen (allgemein: von "Ereignissen") und damit der Verweildauer in best. Zuständen (Dauer von "Episoden" [spells], wie z.B. Arbeitslosigkeit, Krankheit) und bei möglichst vollständiger Kenntnis des jeweiligen Verbleibs einer einmal befragten Einheit, und schließlich die
4. Erfassung von drop-outs (z.B. Studienabbrecher, Studienfachwechsler).

Die Gegenstände hängen untereinander zusammen[5] und sind mit wiederholten Querschnitten nicht sachgerecht zu analysieren.

Man kann diese vier oft genannten Anwendungen unter dem Stichwort "Feststellung individueller Verläufe" zusammenfassen. Nicht erwähnt sind in dieser Aufzählung jedoch weitere Fragestellungen, wie z.B. die Differenzierung zwischen Alters-, Perioden- und Kohorteneffekten oder die in der Ökonometrie mit Panels verbundenen Auswertungsinteressen, wie eine Variable von objektspezifischen (spezifisch für die Einheiten) und periodenspezifischen Einflüssen bestimmt wird und die i.d.R. gar nichts mit der zeitlichen Charakteristik individueller Verläufe zu tun haben.

Bevor wir eine vollständigere Darstellung von Themen versuchen ist kurz auf einige Nachteile einer Panelerhebung hinzuweisen. Die üblicherweise aufgeführten Nachteile sind

- der nicht unerhebliche *Erhebungsaufwand* und die meist lange Dauer von Längsschnittsanalysen[6] sowie die Notwendigkeit wegen des individualisierten Erhebungsverfahrens und der Wiederholung der Befragung *gleicher* Untersuchungseinheiten die Einheiten (Panelteilnehmer) jederzeit "identifizieren" zu können,[7]
- *Panelmortalität*, d.h. Ausfälle von Einheiten aus dem Panel durch Sterblichkeit, Umzüge, Verweigerungen usw.[8] Häufig sind die Ausfälle nicht zufällig, sondern systematisch, das heißt die Panelmortalität bestimmter Bevölkerungs- oder Risikogruppen ist

---

ändert und nicht nur dann, wenn - wie bei der Kohortenanalyse - jeweils eine Kohorte im Zeitablauf betrachtet wird. Früher wurde von Längsschnittsanalysen nur im Sinne einer echten Verlaufsanalyse gesprochen

[5] So ist z.B. die zuverlässige Feststellung von Verweildauern nicht möglich, wenn es nicht gelingt "drop outs" zu erkennen.

[6] Es mag auch sachlich gesehen nicht sehr sinnvoll sein, über sehr lange Zeit immer wieder ein gleichbleibendes Messinstrument (Fragebogen) zu benutzen.

[7] Dies geschieht meist über ein Nummerungssystem (z.B. Arzt- und Betriebsstättennummer BSNR) oder das Vorhandensein von im Zeitablauf konstanten Identifikationsmerkmalen (z.B. Geburtsjahr und -ort) für die Erhebungseinheiten

[8] Man geht allgemein davon aus, dass das Ausfallrisiko bei Mehrfachbefragungen (die sich zudem oft über viele Jahre erstrecken) erheblich größer ist als bei einmaligen Befragungen.

gegenüber anderen erhöht. Es ist nicht immer einfach, für die Ausfälle ganzer Erhebungseinheiten (unit non-response) Werte für Angaben die sonst erhoben worden wären zu "imputieren" oder Ausfälle durch passende Ersatzeinheiten zu kompensieren. Weiter werden gerne genannt

- *Paneleffekte (panel conditioning)*, d.h. allein auf den Umstand der wiederholten Teilnahme am Panel zurückzuführende Veränderungen wie z.B. Meinungsänderungen aufgrund eines im Verlauf der vielen Befragungen gestiegenen Problembewusstsein, Gewöhnungseffekte, Vertrautheit mit dem jeweils gleichen Interviewer etc. (was alles Erscheinungen sind, die außer im Fall wiederholter Meinungsbefragungen von relativ geringer Bedeutung sein dürften), und - was seltener genannt wird
- die Gefahr eines *Zahlenfriedhofs* weil derartige Befragungen eine sehr große Fülle von Auswertungen ermöglichen, die i.d.R. gar nicht alle genutzt werden.

Echte Längsschnittsanalysen sind aus solchen Gründen auch eher selten. Abgesehen von der Studentenstatistik, bei der es um die Verteilung der Verweildauer, nicht nur die durchschnittliche Verweildauer geht, kennt man diese Erhebungsform vor allem bei demographischen aber auch z.B. medizinischen (Krankheitsverläufe, Therapieerfolg) Betrachtungen, die auf längere Zeiträume angelegt sind. Es gibt solche, den Alterungsprozess einiger Kohorten begleitende Befragungen nach der subjektiv wahrgenommenen Gesundheit und der Inanspruchnahme von Gesundheitsleistungen in England, den USA[9] und neuerdings auch auf EU-Ebene.[10] Meist wird jedoch der bei derartigen Erhebungen erforderliche immense Aufwand gescheut.[11]

Echte Längsschnittsbetrachtungen sind - wie gesagt - sehr aufwändig und deshalb i. d. R. nur geboten, wenn es gilt,

- *individuelle Verläufe*, d.h. Feststellungen wann bei einzelnen Einheiten bestimmte "Ereignisse" (Zustandsänderungen) eintreten und damit wie lange (Dauer) ein Zustand anhält; eine interessierende Verweildauern ist auch die Dauer von "Episoden" , d.h. vorübergehenden Zustände, wie z.B. Armuts- oder Arbeitslosigkeitsepisoden
- Entwicklungen im Zeitablauf *frei von Strukturveränderungen* (durch eine sich ändernde Zusammensetzung der beobachteten Gesamtheit, also durch das Ein- und Austreten von Einheiten) darzustellen oder (anspruchvoller) *Kohorteneffekte* herauszuarbeiten.[12]
- im Rahmen ökonometrischer Modelle nach *Einflussfaktoren* zu differenzieren, die *spezifisch* sind *für* die betrachteten *Objekte* (Einheiten) *und Zeiten* (Zeitpunkte bzw. -räume), wobei die Betrachtung objektspezifischer Einflüsse (fixed oder random effects) üblicher ist als die periodenspezifischer Einflüsse (also der Zeitreihencharakteristika).

Zur Betrachtung individueller Verläufe (erster Punkt) gehört auch die Feststellung von Zeitpunkten des vorübergehenden oder endgültigen Ausscheidens einer Einheit. In vielen Fällen besteht Unsicherheit über einen evtl. nicht erkannten "Schwund" (unbemerktes und unaufgeklärtes Ausscheiden einer Einheit) so dass man die (mit periodischen Querschnitten allein nicht mögliche) Feststellung von "*Schwundquoten*" ebenfalls als ein typisches Anwendungsgebiet von (echten) Längsschnittserhebungen ansehen kann.

---

[9] Die gemeinten Erhebungen heißen English Longitudinal Study of Ageing (ELSA) und Health and Retirement Study (HRS).

[10] Die entsprechende Erhebung heißt SHARE (= Survey of Health, Ageing and Retirement in Europe).

[11] Eine andere Anwendung sind Haushaltsbefragungen wie das SOEP oder SILC der EU. Ferner gibt es Betriebspanel (der Bundesagentur für Arbeit oder auch im Rahmen privater Statistiken).

[12] Ein Kohorten- oder Generationeneffekt liegt vor, wenn sich heute Zwanzigjährige signifikant unterscheiden von Zwanzigjährigen des Jahres 1970 (also der Kohorte g = 1950).

Die Frage nach "Schwund" im Sinne von Studienabbruch kann von erheblicher Bedeutung sein. Unsicherheiten in dieser Hinsicht haben vor allem im Falle des Medizinstudiums Wellen geschlagen. Es scheint erhebliche Uneinigkeit darüber zu bestehen, ob, wann und in welchem Ausmaß es in Zukunft einen Ärztemangel geben könnte, weil offenbar Unklarheit darüber besteht, wie viele Nachwuchsmediziner in welcher Phase ihrer Ausbildung "aussteigen".[13]

Solche hier erforderliche Feststellungen über Studien- und Weiterbildungsabläufe verlangen i.d.R. *Kohortenanalysen* als wichtigste Form einer (echten) Längsschnittserhebung. Eine *Kohorte* ist eine Gruppe von Einheiten (Personen) die hinsichtlich eines oder mehrerer unveränderlicher Merkmale (z.B. Geburtsjahrgang, Beginn des Studiums etc.) gleich zusammengesetzt ist und fortlaufend statistisch beobachtet wird.

Als ein wohl nicht generell einsichtiger Vorteil von Panelerhebungen werden auch oft der größere Stichprobenumfang (TN statt nur N bei einem Querschnitt) und damit die (zumindest bei pool-regression; vgl. Abschn. 2.3) zu erwartende größere Effizienz (kleinere Konfidenzintervalle) der Schätzung genannt. Als weitere eher (ökonometrisch) "technische" Vorteile von Paneldaten werden oft auch die vermutlich geringere Wahrscheinlichkeit einer "omitted variables bias" und der Multikollinearität genannt.

2. Fragestellungen, für die Panelerhebungen erforderlich sind

Panelbefragungen (oder speziell: echte Längsschnitte im Sinne von Kohortenbetrachtungen) werden für sehr verschiedene Fragestellungen durchgeführt. Um besser zu verstehen ob und warum sie bei diesen Fragestellungen vorteilhaft und Querschnitten überlegen sein können ist es nützlich diese Themen und einige damit zusammenhängende Begriffe zu systematisieren und genauer darzustellen.

2.1. Ein und Austrittszeiten, Feststellung individueller Verläufe

In diesem Anwendungen geht es um die Feststellung von Zeiten des Eintritts (Zugangszeit t_{Zi}) in und Austritts (Abgangszeit (t_{Ai}) aus bestimmten Zuständen (z.B. Geburt und Tod, Beginn und Ende des Studiums), allgemein von Zeitpunkten von Zustandsänderungen.[14] Diese Zeiten beziehen sich - und das ist entscheidend - auf Individuen[15] i = 1, 2,..., n. Dies sind Fragestellungen, die dafür sprechen, den Aufwand einer von Längsschnittserhebung in Kauf zu nehmen, die aber z.B. bei Panelbetrachtungen ("Längsschnittsdaten") im Sinne der Ökonometrie nicht von Interesse sind.

Die inflow-outflow Matrix ist eine vollständige Darstellung von Strömen,[16] so dass mit ihr auch eine lückenlose Fortschreibung (perpetual inventory) des Bestands

$$B_t = B_{t-1} + Z_{t-1,t} + A_{t-1,t}$$

möglich ist und nicht irgendwelche Einheiten mit unbekanntem Verbleib verschwinden (aus dem System herausfallen) oder plötzlich Einheiten von unbekannter Herkunft auftauchen.

---

[13] So wird z.B. argumentiert, dass "viel zu hohe 'Verluste' für die Nachwuchsmediziner in den Publikationen der berufsständischen Ärzteorganisationen" angegeben werden, oder es heißt, dass Überlegungen, einen Bachelor in der Medizin einzuführen "um den mutmaßlichen Studienabbrechern einen früheren Abschluss anzubieten" nicht zielführend seien, weil es nicht so viele Abbrecher gibt und die meisten "Ausstiege" erst später stattfinden (nach der Approbation). Außerdem ist auch angesichts der Art der Abbruchgründe nicht mit einem erfolgreichen Bachelorstudium bei diesen betreffenden Personen zu rechnen sei. Vgl. D. Bitter-Suermann, Ärzteschwund/ Ärztemangel, Wo liegen die Probleme? in: Forschung und Lehre 1/2011, S. 42ff.

[14] Damit sind auch die individuellen Verweildauern $d_i = t_{Ai} - t_{Zi}$ gegeben.

[15] Es geht um Einzelpersonen, Haushalte, Unternehmen, aber auch Objekte wie z.B. bei Inbetriebnahme und Verschrottung eines PKWs oder bei der Lebensdauer eines Wohnhauses.

[16] Es ist mit der Übersicht auch offensichtlich, dass die Matrixdarstellung informativer als die Bilanzdarstellung ist.

Solche Probleme (z.B. das nicht aufgeklärte Verschwinden) spielen eine nicht unerhebliche Rolle z.B. bei Studienverläufen.[17]

Die Analyse von Verweildauern und der Ursachen von Bestandsveränderungen sowie die Berechnung von Kennzahlen wie Durchschnittsbeständen, durchschnittliche Verweildauer und Umschlagshäufigkeit sind Gegenstand der *Bestandsanalyse*, einem Teilgebiet der (deskriptiven) Statistik.[18]

Für derartige Betrachtungen sind Längsschnittserhebungen deutlich informativer als periodische Querschnitte (evtl. mit summarischer Feststellung der Gesamtzahl der Zugänge $Z_{t-1,t}$ und Abgänge $A_{t-1,t}$ in einem bestimmten Zeitintervall von t-1 bis t). Periodische Erhebungen des Bestands B_t (t = 0, 1, ..., m) erlauben nur die Berechnung von Nettoströmen $\Delta B_t = B_t - B_{t-1}$, nicht aber die der Bruttoströme (Zu- und Abgänge $Z_{t-1,t}$ und $A_{t-1,t}$[19] wobei $\Delta B_t = Z_{t-1,t} - A_{t-1,t}$), oder anders gesagt (vgl. Übers. 2) man erhält nur die "Randverteilungen" der inflow-outflow Matrix, in der Zu- und Abgänge zu bzw. von "Zuständen" differenziert werden.[20]

Übersicht 2: Matrix- und Bilanzdarstellungen in der Bevölkerungsstatistik

a) Inflow-outflow-Matrix und Bilanz (mit Bestands- und Stromgrößen)

Matrixdarstellung

	1	2	Σ
1	Feld uninteressant*	Geburten, Immigration $Z_{t-1,t}$	
2	Todesfälle, Emigration $A_{t-1,t}$	G	Anfangsbestand B_{t-1}
Σ		Endbestand B_t	

Bilanzdarstellung**

Aktiva	Passiva
B_{t-1}	$A_{t-1,t}$
$Z_{t-1,t}$	B_t

* Totgeburten, Wanderungen zwischen und innerhalb der Länder der übrigen Welt

** aus der Bilanzdarstellung folgt unmittelbar die Fortschreibungsgleichung: $B_m = B_0 + Z_{0m} - A_{0m}$.
Die Bilanzsumme N_{0m} umfasst alle Personen, die im Intervall $[t_0,t_m]$ jemals dem Bestand angehört haben.

b) Inflow-outflow-Matrix auf dem Arbeitsmarkt

	A	N	B	Σ
A	AA	AN	AB	A_0
N	NA	NN	NB	N_0
B	BA	BN	BB	B_0
Σ	A_m	N_m	B_m	

Veränderung der Arbeitslosigkeit A (Bestandsveränderung):
$\Delta A = A_m - A_0 = (NA - AN) + (BA - AB)$
linke Seite: Nettostrom (= ΔA)
rechte Seite: Bruttoströme (NA, AN, BA, AB)

Als Zeilen- und Spaltensummen enthält diese Tabelle den Spalten- bzw. Zeilenvektor der Anfangs- (A_0 N_0 B_0) bzw. der Endbestände (A_m N_m B_m) der Arbeitslosen (A), Beschäftigten (B) und Nichterwerbspersonen (N).

Man kann mit periodischen Bestandszählungen B_0, B_1, B_2, ..., B_m zwar den Durchschnittsbestand $\overline{B}$ und auch die durchschnittliche Verweildauer $\overline{d}$ (zumindest bei einer *geschlossenen*

[17] Ein Student erscheint nicht mehr in der Querschnittserhebung weil er das Studium abgebrochen hat, exmatrikuliert ist (mit oder ohne Examen), das Studienfach oder den Studienort gewechselt hat usw. Vor Einführung von Längsschnitterhebungen in der Studentenstatistik war es schwer bis unmöglich; "Abbrecherquoten" oder "Sickerquoten" (Nicht-Wiederauffinden, obgleich kein Abbruch vorgelegen hat) festzustellen.

[18] Zu diesem Bereich der Statistik vgl. Kap. 12 meiner auf dieser Homepage zum Download bereitstehenden Bücher der Deskriptiven Statistik.

[19] Stromgrößen (flows) haben zwei Subskripte und beziehen sich auf einen Zeitraum, Bestandsgrößen (stocks) haben nur ein Subskript und beziehen sich auf Zeitpunkte (Stichtage).

[20] Im Teil a der Übersicht sind die Zustände 1 = übrige Welt [einschl. dem "Jenseits"] und 2 = Deutschland, und im Teil b werden drei Zustände unterschieden A = Arbeitslosigkeit, N = eine Nichterwerbsperson sein und B = Beschäftigung, dann ist AN der Strom von Personen, die aus der Arbeitslosigkeit in die Nichterwerbstätigkeit [z. B. vom Arbeitslosen zum Rentner] gehen und NA der entgegen gerichtete Strom.

Masse, bei der gilt $B_0 = B_m = 0$) bestimmen, nicht aber die Verteilung der Verweildauer angeben. Um $\overline{B}$ aus den Beständen zu berechnen ist die Zeitmengenfläche F_{0m} zu bestimmen. Es gilt $F_{0m} = \sum_j B_{j-1}\left(t_j - t_{j-1}\right)$, bzw. bei äquidistanten Beobachtungszeitpunkten t_j (mit j = 1, 2, ..., m) $F_{0m} = \frac{1}{2}B_o + B_1 + K + B_{m-1} + \frac{1}{2}B_m$. Dann erhält man $\overline{B}$ indem man F_{0m} durch die Länge m des Beobachtungsintervalls dividiert, also $\overline{B} = F_{0m}/m$.

Deutlich schwieriger ist es, die durchschnittliche Verweildauer $\overline{d}$ (und damit zusammenhängend die [durchschnittliche] Umschlagshäufigkeit U, insbesondere bei einer *offenen* Masse zu schätzen.[21] Eine grobe Schätzung von $\overline{d}$, wenn nur Bestände und Gesamtzugänge und Gesamtabgänge in bestimmten Intervallen bekannt sind ist

(1) $\overline{d} = \frac{2m\overline{B}}{Z_{0m} + A_{0m}}$ (sofern m hinreichend groß ist im Verhältnis zu $\overline{d}$)[22]

während die korrekte Bestimmung per Längsschnittanalyse[23] mit $\overline{d} = \frac{1}{n}\sum_i d_i$ zu bestimmen ist. Für die Umschlagshäufigkeit gilt

(2) $U = \frac{m}{\overline{d}}$ oder $U = \frac{N}{\overline{B}}$.

Das bedeutet: Bei gegebenem m ist U groß (klein) wenn $\overline{d}$ klein (groß) ist und $\overline{d}$ ist klein (groß) wenn Bewegungen N (also Zu- und Abgänge) groß (klein) sind im Verhältnis zum Bestand.[24]

Mit *Längsschnittsdaten* (Verlaufsanalysen) - und nur mit ihnen - sind die Zeiten t_{Zi} und t_{Ai} bekannt und damit sind auch die individuellen Verweildauern $d_i = t_{Ai} - t_{Zi}$ und die (Häufigkeits-) Verteilung von d, nicht nur deren arithmetisches Mittel $\overline{d}$ gegeben und zwar mit

$\overline{d} = \frac{\sum d_i}{N}$ wobei N die Anzahl der Eintritts*fälle* - und damit [bei einer geschlossenen Masse] auch die Anzahl der Austrittsfälle - ist.

Wenn alle Kohorten eines Bestands und alle Zeiten t_{Zi} (Zugänge) und t_{Ai} (Abgänge) erfasst werden sind damit außerdem auch die Bestände B_t *zu allen Zeiten* t (t: stetig) bekannt sowie für beliebige Intervalle (0, m) die kumulierten Zu- und Abgänge Z_{0m} und A_{0m}

Umgekehrt gilt: sind aus *Querschnittsdaten* die Zeitreihen der Bestände B_0, B_1, ..., B_m oder der Zugänge $Z_{0,1}$, $Z_{1,2}$, ..., $Z_{m-1,m}$ (und entsprechend der Abgänge $A_{0,1}$, $A_{1,2}$, ..., $A_{m-1,m}$) kann man dagegen nicht auf Verteilung der Verweildauer schließen sondern nur $\overline{d}$ schätzen.

Anders gesagt: Längsschnittsdaten bieten die volle Information über *individuelle Verläufe* (und daraus abgeleite *kollektive Maßzahlen*); Querschnittsdaten enthalten dagegen weniger

---

[21] Bei einer offenen Masse sind Verweilsummen vor t = 0 und nach t = m hinzuzuschätzen.

[22] Bei einer (für das Intervall [0, m]) geschlossenen Masse ist definitionsgemäß $Z_{0m} = A_{0m}$ weshalb man bei einem einigermaßen konstanten Bestand von z.B. 3000 Studenten an einem Fachbereich und einer Erstsemesterzahl A_{0m}/m von z.B. 300 Studenten (in jedem der m Semester) in grober Näherung auf eine durchschnittliche Dauer des Studiums von 10 Semestern schließen kann.

[23] wenn nur ein Ein- und ein Austritt anzunehmen ist. Die Zahl der Bewegungen N ist größer als die Zahl der Personen n ($N \geq n$) wenn eine Einheit mehrmals ein- und austreten kann.

[24] Beispiel Bevölkerung: Bewegungen (jährlicher Anzahl der Geburten bzw., Sterbefälle) gering relativ zum Bestand. Damit ist die Verweildauer groß. Umgekehrt ist die Situation z.B. der Belegung Krankenhausbetten, viele Aufnahmen und Entlassungen bei relativ geringer Anzahl von Krankenhausbetten.

Information. Man kann von den individuellen Zeitangaben t_{Zi} und t_{Ai} (und die Verteilung von "Dauern" wie $d_i = t_{Ai} - t_{Zi}$) auf Kennzahlen des Kollektivs aus dem die Individuen stammen, also Zeitreihen wie B_t, $Z_{t-1,t}$, $A_{t-1,t}$ sowie $\bar{d}$ und U schließen, aber nicht umgekehrt:

mit Längsschnitten feststellbar	→	mit Querschnitten* feststellbar
individuelle Verläufe: Zu- und Abgangszeiten (t_{Zi} , t_{Ai}), und Dauern d_i		*Zeitreihen* B_t, $Z_{t-1,t}$, $A_{t-1,t}$ und *Maßzahlen* eines Kollektivs $\bar{d}$ und U

* Bestandszählungen in Verbindung mit regelmäßigen Feststellungen über die gesamten Zu- und Abgänge in einem Intervall.

Wir haben es hier also mit einer Einbahnstrasse zu tun (aus links folgt rechts, aber nicht umgekehrt).

Neben der Bestandsanalyse gibt es ein weiteres Gebiet der Statistik, das sich mit der Auswertung entsprechender Zeitangaben und der Konstruktion von Verläufen (z.B. Modelle zur Analyse von Verweildauern) beschäftigt, die *Ereignisanalyse*. Solche Betrachtungen sind auch unter dem Stichwort Ereignisdatenanalyse, Duration- oder Survivalanalyse bekann. Sie haben jedoch wenig gemeinsam mit den anderen hier dargestellten Themen zur Verlausbetrachtung, insbesondere mit dem was üblicherweise in der Ökonometrie unter "Paneldatenanalyse" betrieben wird.

Darüber hinaus gibt in Form der "*Tafelrechnung*" (vor allem in der Bevölkerungsstatistik" seit langem Methoden, aus Querschnittsdaten "unechte" Längsschnitte zu konstruieren.[25] Man spricht von Längsschnitten weil sie Verläufe nach Art eines Längsschnitts darstellen. Andererseits sind sie "unecht" weil die auf aus Querschnitten ermittelten Übergangswahrscheinlichkeiten (transition probabilities; Übergang zwischen Zuständen) beruhen.

Ein kurzer Exkurs zu einer vermuteten Paradoxie bei der durchschnittlichen Verweildauer

Angaben zur Verweildauer können bei einer Längsschnittsanalyse bei Kenntnis der Zu- und Abgangszeit bestimmt werden. Im Falle von Querschnittsdaten (über Bestände) und Abgangsstatistiken können sie erfragt werden. Dabei tritt die in der (politischen) Praxis oft als paradox empfundene Situation auf, dass die durchschnittliche Verweildauer nach der Bestandsstatistik größer ist als nach der Abgangsstatistik und das, obgleich mit der Bestandsstatistik nur die bisherige Verweildauer v_i einer Einheit i festgestellt wird, mit der Abgangsstatistik dagegen die abgeschlossene Verweildauer d_i und für jede Einheit $v_i \leq d_i$ sein muss. Dass gleichwohl für die Mittelwerte $\bar{v} > \bar{d}$ gilt ist damit zu erklären, dass hier die Mittelwerte über verschiedene Gesamtheiten gebildet werden, einmal über den Bestand B_t und zum anderen über die gesamten Abgänge $A_{t-1,t}$ in einem Zeitraum und weil im Bestand aus verständlichen Gründen Einheiten mit besonders langer Verweildauer überrepräsentiert sind.

2.2. Alters-, Perioden- und Kohorteneffekt

Neben der Betrachtung individueller Verläufe ist eine Anwendung der Statistik, die eine doch recht aufwändige Längsschnittsbetrachtung rechtfertigen kann der Versuch, Entwicklungen einer Gesamtheit (also nicht individuelle Verläufe) im Zeitablauf frei von *Strukturveränderungen* darzustellen. Das wird besonders dann relevant wenn bei der betrachteten Gesamtheit (z.B. Unternehmen einer bestimmten Branche) mit raschen Strukturveränderungen (die für die Untersuchungsmerkmale bedeutsam sind) durch Ein- und Austritte(aber auch Fusionen, Aufspaltungen und Schwerpunktverlagerungen) von Unternehmen zu rechnen ist.

Eine weitere, etwas speziellere (hinsichtlich der Daten anspruchsvollere) Fragestellung ist das Herauszuarbeiten sog. *Kohorteneffekte* (Generationeneffekte; Einfluss der Zugehörigkeit zur

[25] Man kann oft aus Querschnittsdaten Übergangswahrscheinlichkeiten (transition probabilities) schätzen und damit "unechte Längsschnittsanalysen herleiten. Ein Beispiel hierfür sind Sterbetafeln (Berechnung von Lebenserwartungen) mit aus Querschnitten ermittelten (einjährigen) Sterbewahrscheinlichkeiten. Hinsichtlich dieser Methoden möchte ich wieder auf Kap. 12 in meinen beiden Büchern zur "Deskriptiven Statistik" sowie auf Kap. 2 meines Buches "Wirtschaftsstatistik" verweisen.

gleichen Kohorte). Man unterscheidet die Generation (Kohorte) g, das Alter x und die Periode t, wobei die drei Größen untereinander zusammenhängen g + x = t. Wie man sieht ist mit jeweils zwei Größen (etwa g = 1985 Geburtsjahr und t = 2010, also die Periode) die dritte Größe (hier x, das Alter x = 25) eindeutig bestimmt.

In Übers. 3 erkennt man auch, dass man nur eine Größe konstant halten kann, z.B. die Kohorte g bei einer Kohortenanalyse oder t bei einer typischen Querschnittsanalyse und damit die beiden anderen Größen variabel sind.

Übersicht 3: Alters-, Personen- und Generationeneffekte[26]

a) Alter, Periode, Generation (Kohorte)

Geburtsjahr (Kohorte)	Beobachtungszeit (Periode)				
	t-2	t-1	t	t+1	t+2
g-1	A1		B1		
g	A2	C1	C2	B2	C3
g+1	A3				B3

Betrachtet man einen Zeit*punk*t, etwa den 1.7.2010, so besteht *ein Altersjahrgang* (etwa die 20 - Jährigen) genau genommen aus *zwei Kohorten* (bzw. Teilen von Kohorten): Personen, die zwischen dem 1.7.89 und dem 31.12.89 geboren wurden (89-er Kohorte) und Personen, die zwischen dem 1.1.90 und dem 30.6.90 geboren wurden (Teil der 90-er Kohorte).

b) Vergleiche von Personengruppen

verglichene Gruppen	die Personen sind	Erhebungstyp, bzw. Erhebungen	Unterschied zwischen den Personen ist zu deuten als
A1, A2, A3	gleichzeitig im Bestand	*eine* Querschnittserhebung	Alters- *und* Generationen-(oder Kohorten-) effekt (A, K)
A2, C1, C2, B2, C3	aus der gleichen Generation	Kohorten- oder Verlaufsanalyse	Perioden- *und* Alterseffekt (A, P)
B1, B2, B3	gleich alt	Vergleich *verschiedener* Querschnitte*	Perioden und Generationeneffekt (P, K)

* kein übliches Verfahren in der Bevölkerungsstatistik, engl. auch time lag analysis genannt im Unterschied zu transversal - (Querschnitt) und longitudinal approach (Längsschnitt).

c) Interpretation der Effekte

Effekt	Einflussfaktoren	Beispiel
A: Alterseffekt	biologische Prozesse (Wirkung des Älterwerdens)	Sterbewahrscheinlichkeit der 60-jährigen ist "naturgemäß" größer als die der 30-jährigen
K: Kohorten-(Generationen)-effekt	Ereignisse, die von den Personen jeweils gleichaltrig erlebt (erfahren) wurden	Die Nachkriegsgeneration hat andere Einstellungen als die im oder vor dem Zweiten Weltkrieg geborenen Menschen, weil sie nicht geprägt ist durch die Kriegserfahrung
P: Periodeneffekt	Zeitgeist, aktuelle Lebensbedingungen; Wirkung aktueller Ereignisse	zur Zeit der "Wende" in der DDR haben die Menschen anders gedacht als es gleichaltrige Personen früher oder später taten

Der Name (Alters-, Perioden- oder Generationeneffekt) betrifft die jeweils konstant gehaltene Variable. In dem Maße, in dem sich Statistiken (Kennzahlen) der Generation g und g* zum Zeitpunkt t, bzw. im Zeitablauf t = 0, 1, ... unterscheiden liegt demnach ein Generationeneffekt vor. Unterscheiden sich die relevanten Merkmale im Alter x vom Alter x* trotz gleicher Generation g, dann liegt ein Alterseffekt vor.

[26] Identisch mit Übers.2.4 in meinem Buch "Wirtschaftsstatistik".

Generationeneffekte treten i.d.R. nur zwischen zeitlich etwas weiter auseinander liegenden Kohorten auf. Der Geburtsjahrgang (oder z.B.: Eheschließungsjahrgang) g = 1985 wird sich kaum von g* = 1986 unterscheiden. Da mithin hinreichend verschiedene Kohorten jeweils über relativ lange Zeiträume zu betrachten sind, sind entsprechende Anwendungen der Längsschnittsanalyse zeitaufwändig und daher auch eher selten. Ein Beispiel für solche Anwendungen von Längsschnitten ist die Betrachtung der "Fruchtbarkeit"[27] (Häufigkeit und zeitliche Verteilung von Geburten) von Frauenkohorten. Kennzeichnend für eine Kohorte ist, dass ihre Einheiten zur gleichen Zeit in den Bestand eintreten (z.B. gleiches Geburtsjahr; eine ex-ante oder prospektive Kohorte) oder austreten (ex-post oder retrospektive Kohorte).

Bei den in der Ökonometrie üblichen Panels wird dagegen eine Gesamtheit von z.B. Unternehmen im Zeitablauf betrachtet, die idealerweise - hinsichtlich der Struktur - gleich bleibt, aber nicht notwendig eine Kohorte von im Jahre t = 0 gegründeten Unternehmen darstellen muss, sondern sehr wohl auch Unternehmen verschiedenen Alters umfassen kann. Hier geht es also nicht um die Feststellung von Kohorteneffekten wohl aber um die allgemeinere und weniger anspruchvolle Aufgabe, Struktureffekte auszuschalten.

2.3. Erklärung einer Variable y mit objekt- und perioden-spezifischen Einflüssen

Bei dieser vor allem in der Ökonometrie üblichen Betrachtung von Paneldaten (Panel wird hierbei eher in einem weiteren Sinne verstanden) wird eine Variable y durch eine Regressionsfunktion mit K Regressoren x_1, ..., x_K und meist (über den Beobachtungszeitraum) konstanten Regressionskoeffizienten "erklärt".[28] Da die beobachteten Variablen x_1, ..., x_K (es können auch 0-1 Variablen sein) sowohl eine Querschnitts- (Objekt-) als auch eine Längsschnitt- (Perioden-) Dimension haben ist es sinnvoll, hier mit mehreren Subskripten zu arbeiten. In der (üblicherweise linearen) Regressionsfunktion

$$(3) \qquad y_{jt} = \alpha_{j1} + \beta_{1jt} x_{1jt} + ... + \beta_{Kjt} x_{Kjt} + \varepsilon_{jt}$$

bezeichnet j = 1,...,N die Objektdimension, t = 1,..., T die Periode und k = 1, ..., K den Regressor. Die Unterschiedlichkeit von x_{11t} und x_{12t} , x_{13t} usw. (entsprechend von x_{21t} und x_{22t}, x_{23t} ... x_{2Nt} beim Regressor x_2) ist Ausdruck der beobachteten (durch Regressoren explizit berücksichtigten) Heterogenität (der N Objekte). Der Hinweis dürfte nützlich sein, weil es üblich ist, die in bestimmten Modellen angenommene Unterschiedlichkeit (weil objektspezifisch differenziert) der Koeffizienten α_1, α_2, ..., α_N als *nichtbeobachtete Heterogenität* zu interpretieren.

Die Besonderheit von Modellen der Panel-Ökonometrie ist der Versuch, bei der "Erklärung" von y sowohl objekt- als auch periodenspezifische Einflüsse explizit (oder implizit über die Störgröße) zu berücksichtigen. In der sehr allgemeinen Form von Gl. 3 ist die Regression von y auf den Vektor [x_1 ... x_K] nicht zu schätzen,[29] weshalb es üblich ist bestimmte Restriktionen einzuführen.

Ein (fast) ***nicht-restringiertes Modell*** läge vor bei einer getrennten Schätzung für jedes Objekt auf Basis der als T (K+1)-Tupel (y_{j1}, x_{1j1}, ...,x_{Kj1}, y_{j2}, x_{1j2}, ...,x_{Kj2}, ... y_{jT}, x_{1jT}, ...,x_{KjT}) für jedes Objekt (jede Einheit) j = 1, …, N vorliegenden Daten mit

---

[27] Der Begriff ist in der deutschen, ganz und gar nicht aber in der angloamerikanischen Bevölkerungsstatistik inzwischen verpönt.

[28] Es gibt auch Modelle in denen die Koeffizienten als Realisationen von Zufallsvariablen aufgefasst werden. Weitere Hinweise auf S. 9 der Habilschrift von Martin Spiess.

[29] Es wären N(K+1)T Koeffizienten (α und K Koeffizienten β für jede der N Einheiten zu jeder der T Perioden) zu schätzen und das bei nur NT Wertetupeln (genauer (K+1)-Tupel y_{jt}, x_{1jt}, …, x_{Kjt}), also genauso vielen Beobachtungen.

(4) $y_{jt} = \alpha_{j1} + \beta_{j1} x_{1jt} + ... + \beta_{jK} x_{Kjt} + \varepsilon_{jt}$ ($\beta_{jkt} = \beta_{jk}$ für alle t = 1,..., T und k = 1, ..., K)[30]
und dem Vektor $\boldsymbol{\varepsilon}'_j = [\varepsilon_{j1} \quad \varepsilon_{j2} \quad \Lambda \quad \varepsilon_{jT}]$ der T Störgrößen.[31]

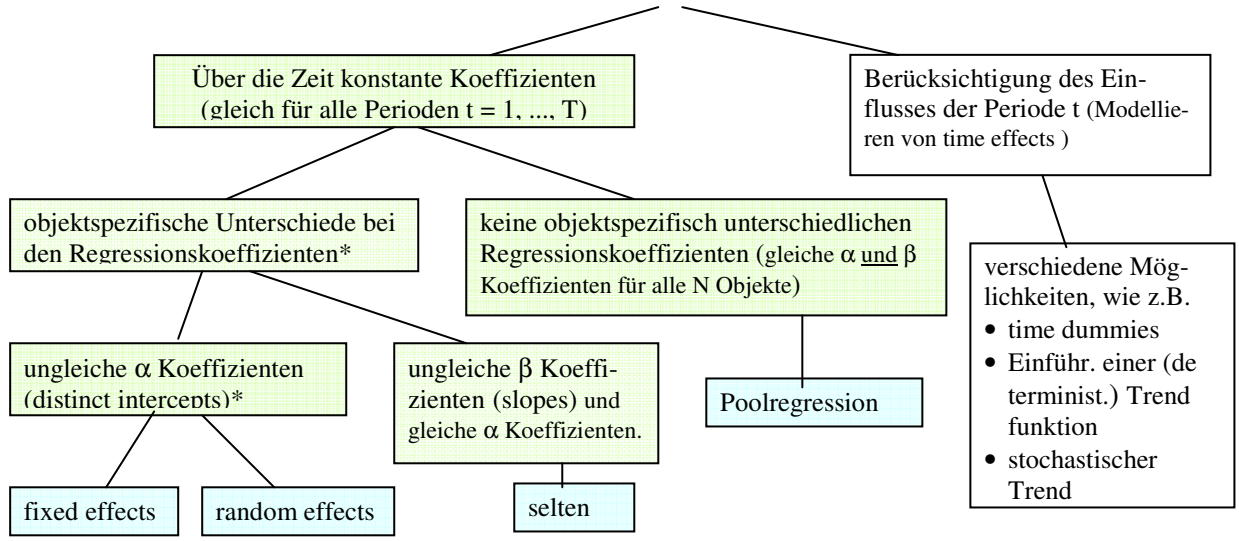
Die nicht-erklärte (residuale) Variation (Summe von Abweichungsquadraten) $S_{\hat{\varepsilon}\hat{\varepsilon}} = \sum_{j=1}^{N} \hat{\boldsymbol{\varepsilon}}_j' \hat{\boldsymbol{\varepsilon}}_j = \sum_{j=1}^{N} \sum_{t=1}^{T} \hat{\varepsilon}_{jt}^2$ ist eine wichtige Größe für einen Vergleich mit stärker restringierten Modellen (bei denen die Nullhypothese H_0 Gleichsetzung [Konstanz] bestimmter Parameter bedeutet). In einem F-Test wird dann jeweils $S_{\hat{\varepsilon}\hat{\varepsilon}}$ mit der notwendig nicht kleineren Summe von Abweichungsquadraten $S^0_{\hat{\varepsilon}\hat{\varepsilon}}$ (das Superskript 0 bedeutet: bei Geltung von H_0) verglichen (Das entspricht der zweiten Stufe in dem in Übers. 5 beschriebenen zweistufigen Verfahren.

H_0 heißt also konstante, nicht perioden- wohl aber noch objektspezifische Koeffizienten, weshalb $S^0_{\hat{\varepsilon}\hat{\varepsilon}} = S_{\hat{\varepsilon}\hat{\varepsilon}}$ statt $S^0_{\hat{\varepsilon}\hat{\varepsilon}} > S_{\hat{\varepsilon}\hat{\varepsilon}}$ zu erwarten wäre.

Wir betrachten nun verschiedene Möglichkeiten wie man bestimmte Restriktionen einführen und entsprechend verschiedene Modelle unterscheiden kann (vgl. Übersicht 4).
Man kann für alle Objekte (und Perioden) gleiche Regressionskoeffizienten $\beta_1, ..., \beta_K$ oder (was seltener geschieht) einen gleichen Regressionskoeffizienten α annehmen. Sind sowohl die α als auch die β Koeffizienten für alle Objekte gleich, so spricht man von ***pool regression***. In diesem Modell

(5) $y_{jt} = \mathbf{x}'_{jt}\boldsymbol{\beta} + \varepsilon_{jt}$ mit den Vektoren $\mathbf{x}'_{jt} = [1 \quad x_{1jt} \quad \Lambda \quad x_{Kjt}]$, $\boldsymbol{\beta}' = [\alpha \quad \beta_1 \quad \Lambda \quad \beta_K]$

Übersicht 4: Einige Modelle der Panel-Ökonometrie



* unterschiedliche α Koeffizienten (α_1, …, α_N) aber gleiche β Koeffizienten

werden praktisch die Daten in einem Topf geworfen und es wird kein Unterschied gemacht ob es N Objekte sind, die T mal beobachtet worden sind oder ob NT Objekte einmal beobachtet

[30] Kürzer geschrieben: ∀ t, k.
[31] Unter diesen Voraussetzungen sind nur noch N(K+1) Koeffizienten zu schätzen, statt bisher die T-fache Anzahl.

wurden.[32] Es ist das restriktivste Modell in Übersicht 4 und ignoriert Heterogenität sowohl in der Zeit als auch in der Querschnitts- (oder Objekts-) Dimension.

Es gibt zwei Arten objektspezifische Differenzierungen des Absolutglieds (intercept) einzuführen. Ist in

(6) $y_{jt} = \alpha_j + \beta_1 x_{1jt} + ... + \beta_K x_{Kjt} + \varepsilon_{jt}$ mit $\alpha_j = \alpha + \mu_j$ und

wobei μ_j ein zu schätzender Parameter (feste aber unbekannte Größe der Grundgesamtheit) liegt ein **fixed effects** (FE) Modell vor. Bei **random effects** (RE) sind die $\mu_1, ..., \mu_N$[33] keine konstante Größen sondern N Realisationen einer Zufallsvariable (die Schätzwerte $\hat{\alpha}_1,...,\hat{\alpha}_N$ sind natürlich immer, auch bei fixed effects Zufallsvariabeln), d.h. bei RE gilt

(7) $y_{jt} = \alpha + \beta_1 x_{1jt} + \ldots + \beta_K x_{Kjt} + (\mu_j + u_{jt}) = \alpha + \Sigma\beta_k x_{kt} + v_{jt}$ mit $v_{jt} = \mu_j + u_{jt}$

wobei μ und u und deshalb auch v Zufallsvariablen sind.

Beim random effects Modell hat die Störgröße v_{jt} (total disturbance) zwei Komponenten (sie ist also eine Summe von zwei Zufallsvariablen):[34]

- die (zeitkonstante individuen- oder objektspezifische)[35] "individual error component" μ_j, für die üblicherweise $E(\mu_j) = 0$ angenommen wird,[36] und die verantwortlich für unterschiedliche Absolutglieder (random intercepts) $\alpha_j = \alpha + \mu_j$, ist und
- die zeit- *und* objektabhängigen auch "idiosynkratische" Fehlerkomponente genannte Zufallsvariable u_{jt} (j = 1, …, N und t = 1, …, T).

Die zufällig schwankenden Größen $\mu_1, \ldots, \mu_N$ mit $E(\mu_j) = 0$ als Ausdruck einer "unobserved heterogeneity" (Inbegriff aller über die Zeit konstanter aber für das jeweilige Objekt spezifischer, nicht beobachtbarer Einflüsse).[37]

Das Modell verlangt es also, über die (Wahrscheinlichkeits-) Verteilung von v_{jt} und über die Kovarianzen von v_{jt} mit x_{ijt} Annahmen zu machen. Es gibt beim random effects model erheblich mehr Schätzprobleme als beim fixed effects model, weil die Autokovarianz[38] von v_{jt} und v_{jt*} (wegen der Gemeinsamkeit von μ_j über alle t) auch unter den im Folgenden aufgelisteten Annahmen[39] nicht verschwindet.[40]

	individual error component μ_j	idiosynkratische Zufallsvariable u_{jt}
Standardannahmen	$E(\mu_j) = 0,\ V(\mu_j) = \sigma^2_\mu$	$E(u_{jt}) = 0,\ V(u_{jt}) = \sigma^2_u\ (\forall\ j, t)$
uncorrelated across individuals	$E(\mu_j\mu_m) = 0$	$E(u_{jt}u_{mt*}) = 0$ für $j \neq m$ und $t \neq t^*$
Die μ_j sind nicht mit den u_{jt} korreliert	$E(\mu_j u_{mt}) = 0\ (\forall\ j, m, t)$	

[32] Gleichwohl dürfte dies einen nicht unerheblichen Unterschied ausmachen bezüglich der Eigenschaften der Störgrößen ε_t (t = 1, ..., T) wie z.B. Homoskedastizität oder keine Autokorrelation.

[33] Und damit natürlich auch $\alpha_1 = \alpha + \mu_1, \ldots, \alpha_N = \alpha + \mu_{1N}$.

[34] Man (z. B. Murray) spricht *deshalb* auch vom "*error components*" Modell (synonym mit random effects).

[35] als "unbeobachtete Heterogenität"

[36] Somit gilt auch $E(\alpha_j) = \alpha = \text{const.}$, d. h. die α_j schwanken zufällig um α. Damit ist μ_j eine *zufällige* objektspezifische Abweichung des Objekts j vom allgemeinen Durchschnitt. Für das Objekt m gilt entsprechend $v_{mt} = \mu_m + u_{mt}$.

[37] Wird eine bestehende unobserved heterogeneity vernachlässigt (also fälschlich H_0: $\alpha_1 = \ldots = \alpha_N = \alpha$ angenommen obgleich dies nicht gilt), so entsteht ein omitted variables bias.

[38] Mit t, t* = 1, …, T und $t \neq t^*$.

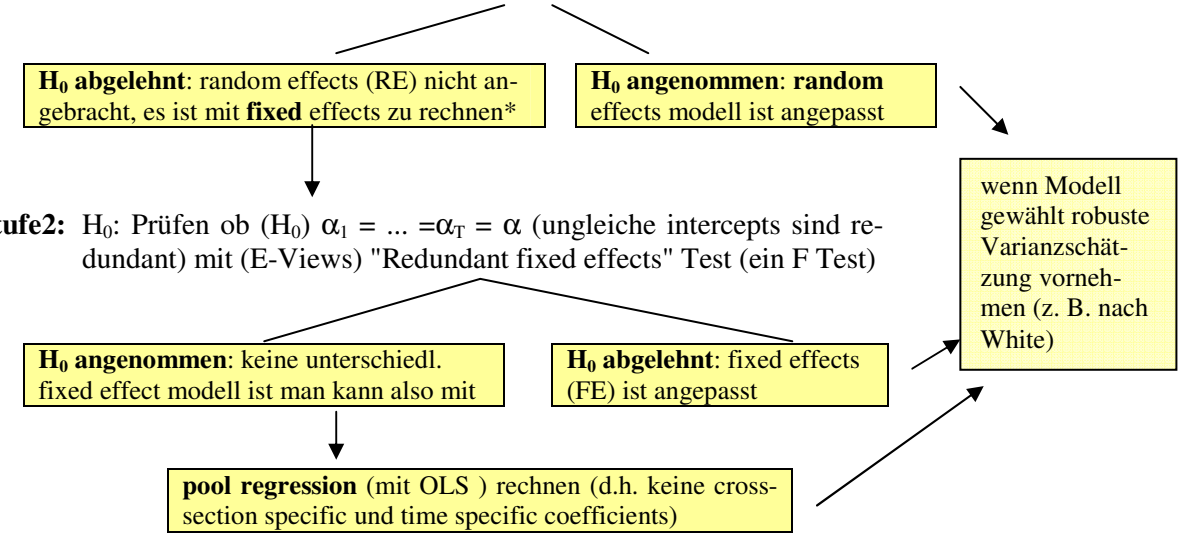
[39] Nach Murray, S. 683.

[40] Vielmehr gilt nach Murray, S. 691: $E(v_{jt}v_{jt*}) = E(\mu_j + u_{jt})(\mu_j + u_{jt*}) = V(\mu_j) = V(\mu)$. Es sind Schätzungen der Varianz von μ nötig, und man spricht hier (wie auch sonst bei Heteroskedastizität und/oder Autokorrelation) von der geschätzten verallgemeinerten Methode der kleinsten Quadrate (= feasible generalized least squares FGLS).

Neben diesen Standardannahmen bezüglich der NT Störgrößen u_{jt} (j = 1, ..., N und t = 1,..., T) ist auch die Annahme der Exogenität aller Regressoren x_{ijt} also $E(x_{ijt}v_{jt}) = 0$ ($\forall$ i, j, t mit i = 1, …, K) wichtig. Dies kann aber nicht ohne weiteres angenommen werden, denn μ_m (bzw. μ_j) ist nicht notwendig für jede Periode t "uncorrelated with explanators" für x_{ijt}.

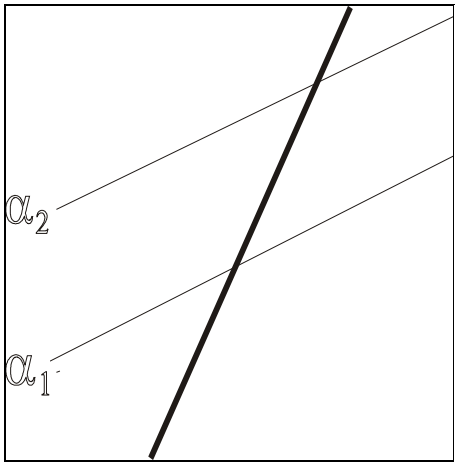
Übersicht 5: Ablaufschema

Stufe1: H_0: Störgrößen nicht mit Regressoren korreliert (Hausman Test: Ausschließen von Endogenität, bei E-Views : "correlated random effects")



* RE wäre verzerrt wenn Störgröße mit Regressoren korreliert.

Zur Modellwahl: bei einem long panel (T groß, N klein) sind die Unterschiede zwischen RE und FE gering, aber bei einem short panel können sie beträchtlich sein. Eine Schätzung von FE ist immer konsistent (sofern die Modellannahmen zutreffen und auf der rechten Seite keine verzögert endogenen Variablen erscheinen), auch wenn das wahre Modell RE oder pooled (dann nicht effizient wegen überflüssiger Schätzung einer Varianz) ist; umgekehrt gilt: die Schätzung von RE ist inkonsistent wenn das wahre Modell pooled oder FE ist.



Die nebenstehende Graphik macht deutlich, warum man in der Tat eine systematische Überschätzung der Steigung β in der Regression $y_t = \alpha + \beta x_t + u_t$ erhält, wenn man tatsächlich ein FE Modell mit N = 2 Objekten hat und damit zwei parallele Regressionsfunktionen mit unterschiedlichen "intercepts" α_1 und α_2 und den entsprechenden Punkten um die beiden Regressionsgeraden im Streuungsdiagramm. Schätzt man fälschlich das Pool-Modell durch den gesamten Punkthaufen, so erhält man eine mit der dickeren Linie angedeutete Regressionsfunktion deren Steigung β offensichtlich größer ist.

Gilt $E(x_{ijt}\mu_m) = 0$ nicht, ist also die objektspezifische Zufallsvariable μ_j (die individuelle Fehlerkomponente) mit den Regressoren (explanators) x_{ijt} korreliert, dann liegt das Modell der fixed effects und nicht das (speziellere) random effects Modell vor.[41]

[41] Zum Testen der Annahme der Unkorreliertheit – und damit der Konsistenz einer OLS Schätzung – ist der *Hausman Spezifikationstest* anzuwenden; vgl. Murray, S.701.

Es ist deshalb sinnvoll nach einem zweistufigen Verfahren gem. Übers. 5 vorzugehen.

Im Abschnitt 4 dieses Papiers wird kurz auf die mit diesen und weiteren Modellen verbundenen Schätzprobleme eingegangen.

Als Vorteil von Paneldaten gilt nicht nur der größere Stichprobenumfang NT statt N bzw. T Beobachtungen (und damit größere Zahl der Freiheitsgrade) sondern auch, dass die Regressoren x_1, x_2, … weniger von Multikollinearität betroffen sein könnten, weil jede Variable nicht nur zwischen den Einheiten, sondern auch im Zeitablauf variieren kann. Das macht es auch möglich eine sich nicht in den konkreten Werten der Regressoren x_{1jt}, x_{1j^*t}, …, x_{2jt}, x_{2j^*t},… *explizit* ausdrückende Ungleichheit ("Heterogenität") von zwei Objekt j bzw. j* zu messen in Gestalt von unterschiedlichen Konstanten (intercepts) α_1, …, α_N. Somit wird eine nicht-beobachtete Heterogenität zwischen Querschnittseinheiten (was auch als Vermeidung einer omitted variables bias aufgefasst werden kann) modelliert. Diese Heterogenität käme sonst (bei reinen Querschnittsdaten mit nur einer statt T Beobachtungen für jedes Objekt oder auch bei der "pool regression") nur in der Störgröße zum Ausdruck.

3. Probleme der Durchführung von Panelerhebungen

Im Folgenden werden kurz einige in der Literatur häufig diskutierte praktische Probleme im Zusammenhang mit Panel behandelt werden. Die in Abschn. 2.3 begonnene Darstellung mehr ökonometrischer Probleme tritt dabei in den Hintergrund und wird erst in Abschn. 4 wieder aufgegriffen. Es ist naheliegend, dass bei der sich oft über einen längeren Zeitraum hinziehenden praktischen Durchführung von Wiederholungs-Befragungen (Erhebungen) Probleme entstehen mit der Aufrechterhaltung der Antwortbereitschaft der Befragten (allgemein die vergleichsweise höhere Belastung [response burden] der Befragten durch die Befragung), der Sicherstellung vergleichbarer Bedingungen ("Kontrolle" der übrigen [nicht expliziten] Einflüsse) und der Zusammenführung von Daten des gleichen Objekts zu verschiedenen Zeitpunkten (von verschiedenen Erhebungswellen).

3.1. Panelmortalität (panel attrition)

Aufgrund von Panelmortalität (panel attrition)[42] kann bei weiteren "Wellen" (Wiederholungen) die Anzahl der permanent im Panel befindlichen Einheiten stark zusammenschrumpfen, so dass das "balanced panel" im Endeffekt sehr klein werden kann, weil es über einen langen Zeitraum kaum "matched pairs" gibt.[43] Panelmortalität ist meist dann kein Problem (und durch "Hochrechnung" auszugleichen) wenn die damit verbundenen unit nonresponse Fälle zufällig verteilt sind. Häufig sind die Ausfälle aber systematisch (korreliert mit Merkmalen der Einheiten).[44] Die Probleme Panelmortalität und Nichtbeantwortung (die auch schon in der ersten Befragungswelle einer Wiederholungsbefragung auftreten kann und nicht – wie im Fall

[42] Beim sog. SOEP des DIW wurden anfänglich (1984) 9.527 Haushalte befragt, von denen jedoch nur 5921 Haushalte teilnahmen. Hiervon nahmen nach 16 Jahren im Jahr 2000 immerhin noch 4060 Haushalte teil. Der Vergleich der Zahlen (entnommen aus der Habilitations-Schrift von Martin Spiess S.11f) stellt eher eine Untergrenze der panel attrition (Panelabnutzung) dar, weil Haushalte auch aufgespalten und (seltener) zusammengelegt werden können. Deutlich schlechter scheint die Bilanz beim "Familiensurvey" des Deutschen Jugendinstituts (www.dji.de) zu sein, das mit einer Befragung von 10.043 Familien (1988) begann. An der zweiten Welle (1994) nahmen hiervon nur noch 4.997 und bei der dritten Befragungswelle (2000) nur noch 2.002 Haushalte teil. Längsschnittserhebungen

[43] Man kann sich in den Veröffentlichungen natürlich auf die Einheiten beschränken, die in allen Erhebungen oder in jeweils m aufeinanderfolgenden Erhebungen (etwa m =2 oder m > 2) gleichermaßen in der Beobachtungsgesamtheit vertreten waren.

[44] Das Problem ist gut beschrieben mit "disproportionate dropout".

der Panelmortalität – bereits in früheren Wellen befragte Einheiten betrifft[45]) sind miteinander verwandt und entsprechend werden auch ähnliche Methoden zum Umgang hiermit angewendet. Das damit verbundene Problem wird meist als eine Gefährdung oder Verringerung der "Repräsentativität" gesehen.

Das nicht aufgeklärte Verschwinden einer Einheit aus einer Wiederholungsbefragung kann die Darstellung von Verläufen verzerren. So kann z.B. die längsschnittanalytische Betrachtung der Sterblichkeit (oder allgemein eine Ereignisanalyse[46]) die Sterblichkeit unterschätzen wenn Einheiten nach ihrem Ausscheiden sterben und die Zeitpunkte solcher Todesfälle nicht durch "Verbleibstudien" (Nachforschungen was aus ausgeschiedenen Einheiten geworden ist) aufgeklärt werden. Das Problem wird besonders dann virulent wenn anzunehmen ist, dass Kontaktverlust oder Verweigerung (und damit Ausscheiden) mit der Morbidität und damit der Sterbewahrscheinlichkeit korreliert ist. Wenn alte und kranke Personen weniger bereit zur Teilnahme sind[47] und deshalb unterrepräsentiert sind, dann ist auch zu erwarten, dass Statistiken zu Merkmalen die mit dem Alter und Gesundheitszustand korrelieren verzerrt sind.

Für eine mögliche Korrektur (oder Gewichtung) der *beobachteten* Daten und/oder Imputation (Konstruktion von Ersatzwerten) *nicht beobachteter* Daten um Ausfälle auszugleichen sind die Mittelwerte $\bar{y}_r$ der response (r) Gruppe und $\bar{y}_n$ der non-response (n) Gruppe sowie die entsprechenden Stichprobenumfänge $n = n_r + n_n$ entscheidend. Nennt man $\bar{y}_{r+n}$ den um $\bar{y}_n$ korrigierten Mittelwert $\bar{y}_r$, dann ist $\bar{y}_{r+n} - \bar{y}_r$ der prinzipiell unbekannte und nur zu schätzende non-response-bias. Ein naheliegender Gedanke ist, die bekannte Größe $\bar{y}_r$ durch eine Regressionsfunktion mit den Regressoren x_{1ri}, x_{2ri}, ... $i = 1,...,n_r$ zu schätzen und diese Funktion bei der Schätzung von $\bar{y}_n$ mit den Werten x_{1nj}, x_{2nj}, ... $j = 1, ..., n_n$ (oder anderen Variablen x_1, x_2, ...) zu Grunde zu legen.

Man kann dem Problem des Ausfallens von Einheiten (die in früheren "Wellen" teilgenommen haben zu lösen versuchen durch verschiedene Verfahren, die hier nicht vollständig aufgezählt werden und Varianten von missing-data Techniken darstellen:

- Nichtbeachten unvollständig repräsentierter Einheiten (Eliminieren der [früher tatsächlich gegebenen] Antworten der Verweigerer; complete case analysis) was auf der Annahme $\hat{\bar{y}}_n = \bar{y}_r$ hinausläuft;
- Einführung einer geeigneten Gewichtung ("redressment") der gegebenen Antworten um die fehlenden Antworten auszugleichen;[48]
- Ersatz von n_n fehlenden Werten y_{n1}, y_{n2}, y_{n3} .... durch entsprechende Mittelwerte $\bar{y}$ (single imputation, statt $\bar{y}$ kann man auch der Median oder Modus nehmen) für die Gesamtheit der non-response group, oder durch verschiedene Mittel $\bar{y}_1$, $\bar{y}_2$, ... für Teilgesamtheiten dieser Gruppe
- Schätzen von n_n fehlenden Werten y_{n1}, y_{n2}, y_{n3} ... oder Mittelwerten $\bar{y}_1$, $\bar{y}_2$, ... mit Regressionsfunktion aufgrund von Werten x_1, x_2, ... der Stichprobe oder anderer Daten (z.B. früherer Erhebungen).

---

[45] Der Umstand, dass eine Einheit früher einmal *teilgenommen* hat erleichtert natürlich die Imputation. Im Gegensatz dazu sind bei der non-response Problematik i.d.R. überhaupt keine Erkenntnisse über die Antwortverweigerer zu erhalten.

[46] Die Ereignisanalyse fragt – wie oben bereits erwähnt - wann und (im Falle der Sterblichkeit irrelevant) wie oft bestimmte Ereignisse bei Personen (allgemein: bei Einheiten) auftreten.

[47] Das ist das typische Problem der panel attrition "that panel members of certain demographics ... may disproportionately opt out

[48] Die für die Bestimmung wichtige Wahrscheinlichkeit $p(A_i)$ (mit $0 \leq p(A_i) \leq 1$) des Ausfalls A einer Einheit i wird oft mit demographischen, sozioökonomischen und anderen Variablen modelliert (logistische Regression).

Um die Panelmortalität möglichst gering zu halten, ist es nötig, das Panel regelmäßig zu "pflegen" (was sehr aufwändig ist), d. h. es müssen wiederholte Kontaktversuche unternommen werden (callbacks), die Adresskartei muss aktualisiert und Ausfälle müssen durch geeignete Ersatzpersonen kompensiert werden. Es ist möglich, die Wahrscheinlichkeit der panel attrition durch ein entsprechendes Stichprobendesign zu verringern. Eine Möglichkeit ist z.B. ein alternierendes Panel (ein erster Teil wird in der 1., 3., usw. Welle, der andere in der 2., 4. usw. Welle befragt) oder ein Rotationssystem (nach jeder Welle scheidet ein Teil des Panels aus und wird einen hinzukommenden Teil ersetzt) und es gibt auch Varianten, die beide Möglichkeiten miteinander kombinieren, indem z.B. Teile des Panels wiederholt und in jeder Welle befragt werden, andere dagegen nur einmal oder alternierend).

3.2. Paneleffekte (panel conditioning)

Ein zweites Problem, das neben panel attrition, v. a. bei Befragungen nach Meinungen (und nur bei ex-ante Kohorten) auftritt ist der Paneleffekt, wonach die Wiederholung (weitgehend) gleicher Befragungen selbst die Ursache für einer beobachteten und zu analysierenden Veränderung ist (z.B. die Änderung einer Meinung nach einer Neubesinnung *aufgrund* der wiederholten Befragung). Derartige Effekte (von "wave specific responses",[49] was auch "repeat measures effect" genannt wird) sind wohl im Falle von Unternehmenspanel im Unterschied zu Meinungsbefragungen weniger zu erwarten.

Gravierender dürfte dagegen ein ebenfalls mit der *ungewöhnlich langen Dauer des Projekts* einer Längsschnitterhebung zusammenhängendes Problem sein: wenn die Ergebnisse der Untersuchung vorliegen, können sich die Verhältnisse bereits so weit geändert haben, dass die ursprüngliche Fragestellung nicht mehr aktuell (oder nur noch historisch interessant)ist.

Ein ebenfalls ernstzunehmendes Problem dürften die meist *nicht* wirklich *ausgenutzten* vielfältigen *Möglichkeiten der statistischen Auswertung* von Paneldaten (im Sinne aufwändiger *echter* Längsschnitte) sein. Es fragt sich dann ob der deutlich höhere Aufwand gerechtfertigt ist und ob man nicht evtl. nur geringfügig schlechtere Auswertungsmöglichkeiten auf weniger aufwändige und für die Befragten stärker schonende Weise, insbesondere durch Auswertung vorhandener administrativer Register erhalten kann. Auch in der amtlichen Statistik ist für die Zukunft deutlich der Weg zu einer immer stärker "registerbasierten" also sekundärstatistischen Datenbeschaffung vorgezeichnet. Es wird immer schwieriger Akzeptanz für Primärstatistiken zu erreichen, was bei *Wiederholungs*-befragungen umso mehr gilt.

3.3. Identifizierbarkeit der Erhebungseinheiten

Panelerhebungen machen es nötig, die zu erhebenden Einheiten so eindeutig und unveränderlich zu identifizieren, dass sie in neuen Erhebungswellen wiedergefunden werden können, bzw. ihr Antwortverhalten zurückverfolgt werden kann. Ein solches "retrieval system" verlangt die Vergabe von Ordnungsnummern (z.B. Personen oder Betriebskennzeichen) oder die Konstruktion eines Identifikationsmerkmals, was in der Praxis nicht einfach zu realisieren ist.[50] Hinzu kommt, dass das eine Identifikation erfordernde individualisierte Erhebungsver-

[49] Um das zu vermeiden werden z.B. in verschiedenen Paneldesigns nicht in jeder Welle die gleichen Kohorten betrachtet (alternierende und rotierende Systeme). Bei Ausfällen (non-response, drop out, missing data) wird in der Literatur auch gern eine Fallunterscheidung dergestalt gemacht, ob der Ausfall vom Wert der Untersuchungsvariable oder anderer Variablen (covariates) abhängig ist (covariate dependent drop out) oder unsystematisch und daher voll und ganz zufällig ist (missing completely at random).

[50] Bei Einführung der Hochschulverlaufsstatistik (Befragung von Studenten) wurde ein 15 stelliges Merkmal gebildet aufgrund des Namensangfangs, Geschlechts, Geburtsdatums, Geburtslands und Geburtsorts (was natürlich voraussetzt, dass bei wiederholter Befragung jeweils die Frage nach den entsprechenden Hilfsmerkmalen auch [korrekt] beantwortet wird). Gleichwohl gab es bei verschiedenen Auswertungen bis zu 17% unklare Zuordnungen ("unpaarige Fälle"). "Das zeigt deutlich, dass bei der Erhebung des Identifikationsmerkmals besonde-

fahren bei den Befragten Zweifel an der Gewährleistung des Datenschutzes aufkommen lassen,[51] weil evtl. mehr als sonst Merkmale (Adresse, Geburtsdatum etc.) erfasst werden müssen, die nicht Gegenstand der Untersuchung sind, sondern nur dazu dienen, eine Identifikation oder evtl. nötige Rückfragen zu erleichtern. Eine Panelerhebung kann damit deutlich mehr Probleme mit der organisatorischen Vorbereitung und Gewinnung von Akzeptanz als einmalige Erhebungen mit sich bringen.

4. Schätzmethoden bei Modellen für Panelerhebungen in der Ökonometrie

Neben den im Abschnitt 2.3 kurz behandelten statischen Panelmodellen gibt es auch ***dynamische***[52] ***Modelle***, in denen verzögert endogene Variablen auftreten, etwa in der Form

(8) $y_{jt} = \alpha + \gamma_1 y_{j,t-1} + \gamma_2 y_{j,t-2} + \ldots + \gamma_p y_{j,t-p} + \beta_1 x_{1jt} + \ldots + \beta_K x_{Kjt} + \varepsilon_{jt}$ oder

Eine Dynamisierung der Regression ist in dieser expliziten Art möglich, aber auch mit timedummies oder implizit über die Annahme eines stochastischen Prozesses (z.B. autoregressiver Prozess) für die Störgröße ε_{jt}. Mit "dynamischen Modellen" ist meist der erste Fall gemeint. Auf diese Modelle, die in der Regel mit der Generalized Method of Moments (GMM Methode) geschätzt werden, kann hier nicht weiter eingegangen werden.

In diesem Abschnitt werden nur Schätzmethoden für einige relativ einfache und inzwischen sehr allgemein bekannte Modelle für Paneldaten dargestellt und es wird abschließend nur ein kurzer Ausblick auf weitere Möglichkeiten gegeben.

4.1. Schätzung im fixed und random effects Modell

Bei der "***pool regression***" wird nicht davon Gebrauch gemacht, dass vom gleichen Objekt j bzw. j* (z.B. dem gleichen Unternehmen j bzw. j*) im Zeitablauf mehrere Beobachtungen existieren,[53] so dass sich jede nicht mit den konkreten Werten der Regressoren x_{1jt}, x_{1j*t}, …, x_{2jt}, x_{2j*t},… explizit erfasste Ungleichheit (also die nicht beobachtbare "Heterogenität") der Objekte nur in den Störgrößen u_{j1}, u_{j2}, …, u_{jT} bzw. u_{j*1}, u_{j*2}, …, u_{j*T} niederschlagen kann. Es ist deshalb möglich dass u_{jt}, mit x_{1jt}, x_{2jt}, … korreliert (und entsprechend u_{j*t}, mit den Werten x_{1j*t}, x_{2j*t},… der Regressoren x_1, x_2,…). Wenn das der Fall ist, dann ist die Schätzung des Pool-Modells mit OLS verzerrt und inkonsistent.[54] Es ist dann ein weniger restriktives Modell für die Daten zu suchen.

---

re Sorgfalt zu verwenden ist … denn sonst wird der Aufwand für die Zusammenführung unverhältnismäßig hoch." Ferner "muß alles vermieden werden, was die Unpaarigkeitsprozentsätze über die z. Z. noch rein organisatorisch-technisch bedingten Relationen steigen lassen könnte. Technisch bedingte Unpaarigkeit muss vom echten Ausscheiden aus dem Kreis der Studenten, also der Aufgabe des Studiums, unbedingt getrennt werden." Der Grund ist, dass sonst die Bestimmung einer Sicker- oder Schwundquote erschwert ist, die zu ermöglichen ja zunächst einmal ein Hauptgrund für die Wahl des recht anspruchsvollen Erhebungsverfahrens einer Kohortenanalyse war. Vgl. L. Herberger, Praktische Erfahrungen mit Verlaufsstatistiken, Allgemeines Statistisches Archiv, Bd. 57 (1973), S. 54ff (69).

[51] Auch das war natürlich ein Problem bei der Einführung der Hochschulverlaufsstatistik.

[52] Der Ausdruck "dynamisch" ist im Zusammenhang mit Paneldaten erklärungsbedürftig, weil es ja kennzeichnend für Paneldaten ist, dass Daten über die gleichen Einheiten *zu verschiedenen Zeitpunkten* vorliegen (also in diesem Sinne "dynamisch" sind). Im Zusammenhang mit (ökonometrischen) Panel-Modellen ist damit gemeint, dass auf der rechten Seite der Regressionsgleichung "lagged dependent variables" also $y_{j,t-1}$, $y_{j,t-2}$, …. auftreten. Das Besondere dieser Modelle ist, dass die Regressoren nicht mehr als strikt exogen angenommen werden können, was für einige Schätzverfahren aber vorauszusetzen ist.

[53] Es wird, wie gesagt, kein Unterschied gemacht ob es N Objekte sind, die T mal beobachtet worden sind oder ob NT Objekte einmal beobachtet wurden.

[54] OLS setzt nicht-stochastische Regressoren voraus oder wenn sie stochastisch sind, dass sie strikt exogen (X_{kt} ist weder mit vergangenen noch kontemporären oder zukünftigen Störgrößen u korreliert).

Dass sich die Individualität (uniqueness) eines Objekts (und damit die Heterogenität der Objekte) in einer Zufallsvariable widerspiegelt gilt auch für die Störgröße v_{jt} im ***random effects model REM*** (oder auch *error components model ECM)*, die ja aus zwei Komponenten besteht, der zeitkonstanten individuen- oder objektspezifischen Komponente μ_j, für die das oben Gesagte (Auffangen der Individualität, Ausdruck der non-observed heterogeneity und damit evtl. Korreliertheit mit den Regressoren) gilt, und der "idiosynkratischen" Fehlerkomponente u_{jt}, Aus den oben (Abschn 2.3) aufgeführten Annahmen $v_{jt} = \mu_j + u_{jt}$ und $E(\mu_j u_{jt}) = 0$ folgt auch $E(v_{jt}) = 0$ und für die Varianz

(9) $$V(v_{jt}) = E\left(v_{jt}^2\right) = \sigma_\mu^2 + \sigma_u^2 .$$

Wäre $\sigma_\mu^2 = 0$ gäbe es keinen Unterschied zum pool regression model

Da μ (anders als u) nicht abhängig ist von t gilt für die Autokovarianz

(10) $$E(v_{jt}v_{js}) = E(\mu_j + u_{jt})(\mu_j + u_{js}) = \sigma_\mu^2 \quad (t \neq s)$$

egal, wie weit t und s von einander entfernt sind. Wenn REM anzunehmen ist,[55] dann sollte wegen (10) nicht mit OLS geschätzt werden sondern mit der verallgemeinerten Methode also GLS (generalized least squares), bzw. weil die Varianzen $(\sigma_\mu^2 + \sigma_u^2)$ und Kovarianzen σ_μ^2 der $T \times T$ Matrix (Kovarianzmatrix der Störgrößen) $\mathbf{\Omega} = E(\mathbf{vv}')$

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_u^2 + \sigma_\mu^2 & \sigma_\mu^2 & \Lambda & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_u^2 + \sigma_\mu^2 & \Lambda & \sigma_\mu^2 \\ M & M & O & M \\ \sigma_\mu^2 & \sigma_\mu^2 & \Lambda & \sigma_u^2 + \sigma_\mu^2 \end{bmatrix} = \sigma_u^2 \mathbf{I}_T + \sigma_\mu^2 \mathbf{i}_T \mathbf{i}_T' .$$

erst noch mit $(\hat{\sigma}_\mu^2 + \hat{\sigma}_u^2)$ und $\hat{\sigma}_\mu^2$ zu schätzen sind, mit der feasible generalized least squares (FGLS) Methode. OLS wäre zwar konsistent (vgl. Übers. 6) was aber als asymptotische Eigenschaft in der üblichen Situation "large N small T", also einem "short panel" nicht zum Tragen kommt, aber OLS ist – im Unterschied zu FGLS – nicht effizient (vgl. auch Übers. 6).

Ist aufgrund des Hausman Tests das ***fixed effects model (FEM)*** statt ein REM zu schätzen, so sind vor allem zwei Schätzverfahren üblich

- "within groups estimation" (WG-Verfahren) und
- das Least squares dummy variables (LSDV) Verfahren.

WG: "within groups estimation"

Beim WG Verfahren wird von allen Variablen jeweils der (über die Zeit, also alle T Perioden gerechneten) Mittelwert abgezogen und auch diese zentrierten oder "de-meaned" Variablen wird OLS angewendet. Durch die Zentrierung fällt das intercept α_j und auch alle anderen möglichen zeit-konstanten (und objektspezifische) Einflüsse in

$$\mathbf{y}_j = \begin{bmatrix} y_{j1} \\ M \\ y_{jT} \end{bmatrix} = \begin{bmatrix} 1 \\ M \\ 1 \end{bmatrix} \cdot \alpha_j + \begin{bmatrix} x_{1j1} & \Lambda & x_{Kj1} \\ x_{1j2} & \Lambda & x_{Kj2} \\ M & O & M \\ x_{1jT} & \Lambda & x_{KjT} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ M \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_{j1} \\ M \\ u_{jT} \end{bmatrix} \text{ oder } \mathbf{y}_j = \mathbf{i}\alpha_j + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{u}_j$$

[55] Das heißt: wenn die H_0 (Gleichheit des fixed und random effect Modells) beim Hausman Test verworfen wird.

weg. Zu schätzen ist danach $\tilde{\mathbf{y}}_j = \tilde{\mathbf{X}}_j\boldsymbol{\beta} + \tilde{\mathbf{u}}_j$ mit

$$\tilde{\mathbf{y}}_j = \begin{bmatrix} y_{j1} - \bar{y}_j \\ M \\ y_{jT} - \bar{y}_j \end{bmatrix} \text{ und } \tilde{\mathbf{X}}_j = \begin{bmatrix} x_{1j1} - \bar{x}_{1j} & \Lambda & x_{Kj1} - \bar{x}_{Kj} \\ M & O & M \\ x_{1jT} - \bar{x}_{1j} & \Lambda & x_{KjT} - \bar{x}_{Kj} \end{bmatrix}.$$

Für die Schätzung von $\boldsymbol{\beta}$ werden wieder die Vektoren und Matrizen gestapelt zum NT × 1 Vektor $\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \\ M \\ \tilde{\mathbf{y}}_N \end{bmatrix}$ (und entsprechend $\mathbf{u}_w$ [w wegen within]) und der NT × K Matrix $\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{X}}_1 \\ \tilde{\mathbf{X}}_2 \\ M \\ \tilde{\mathbf{X}}_j \end{bmatrix}$,

so dass $\boldsymbol{\beta}$ wie folgt

$$(11) \qquad \hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}.$$

mit OLS zu schätzen ist womit dann die β-Koeffizienten im FEM gegeben ist.[56] Für jedes j muss gelten[57]

$$(11a) \qquad \bar{y}_j = \hat{\alpha}_j + \hat{\beta}_1\bar{x}_{1j} + ... + \hat{\beta}_K\bar{x}_{Kj},$$

so dass sich im zweiten Schritt nach den Koeffizienten $\hat{\beta}_k$ (k = 1, …, K) auch die objektspezifischen α Koeffizienten schätzen lassen.

Eine ähnlich, ebenfalls nicht selten betrachtete Methode ist die first difference method, d.h. die Anwendung von OLS auf $\Delta y_{jt} = y_{jt} - y_{j,t-1}$ und analog $\Delta x_{kjt} = x_{kt} - x_{kj,t-1}$ mit den Regressoren k = 1,…,K. Durch die Differenzenbildung entfallen die zeitinvarianten und objektspezifischen Einflüsse wie die α_j (j = 1, …, N). Wie man leicht sieht ist jedoch

$$E(\Delta u_{jt}\Delta u_{j,t-1}) = E(u_{jt} - u_{j,t-1})(u_{j,t-1} - u_{j,t-2}) = E(u_{j,t-1})^2 \neq 0,$$

so dass der Differenzenoperator eine autokorrelierte Störgröße erzeugt, auch wenn u nicht autokorreliert ist.

LSDV: "Least squares dummy variables"

Die Schätzung des FEM nach dieser Methode läuft darauf hinaus, dass man für die Unterschiedlichkeit der α Koeffizienten der N Objekte N–1 Dummy (0-1) Variablen $D_2, D_3, \ldots D_N$ wie folgt einführt $D_2 = \begin{cases} 1 & \text{wenn } j = 2 \\ 0 & \text{sonst} \end{cases}$, …, D_N mit den Koeffizienten Δ_j für diese N–1 Regressoren. Man erhält die α Koeffizienten mit $\alpha_1 = \alpha$, $\alpha_2 = \alpha + \Delta_2$, …, $\alpha_N = \alpha + \Delta_N$ und die Modellgleichung

$$(12) \qquad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ M \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \Lambda & \mathbf{0} \\ \mathbf{i} & \mathbf{i} & \Lambda & \mathbf{0} \\ M & M & O & M \\ \mathbf{i} & \mathbf{0} & \Lambda & \mathbf{i} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \Delta_2 \\ M \\ \Delta_N \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ M \\ \mathbf{X}_N \end{bmatrix} \cdot \boldsymbol{\beta}^L + \begin{bmatrix} \mathbf{u}_1 \\ M \\ \mathbf{u}_N \end{bmatrix} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}^L + \mathbf{u}.$$

[56] Sie ist konsistent aber evtl. nicht effizient. Die Schätzung des FE Modells ist immer konsistent möglich, auch wenn das "wahre" zugrundeliegende Modell das der pooled regression oder das REM ist.

[57] Die Regressionsfunktion geht durch den Schwerpunkt (das arithmetische Mittel).

Man erhält den Vektor $\hat{\boldsymbol{\beta}}^{\#} = \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}}^{L} \end{bmatrix}$ der Least squares dummy variables (LSDV) Schätzwerte $\hat{\boldsymbol{\alpha}}$ (mit Schätzwerten für die N Größen, α, Δ_2, …, Δ_N) und $\hat{\boldsymbol{\beta}}^{L}$ (K Parameter β) gem. Gl. 14 aufgrund der Inversion einer Blockmatrix $\mathbf{X}^{\#\prime}\mathbf{X}^{\#}$.

Übersicht 6: Modelle und Schätzverfahren

das "wahre" Modell der Grundgesamtheit	Schätzverfahren		
	OLS	LSDV*	FGLS
pool a)	effizient	konsistent b)	konsistent
fixed effects	inkonsistent c)	effizient	inkonsistent d)
random effects	konsistent	konsistent	effizient e)

* Least squares dummy variables (man beachte, dass LSDV nie inkonsistent sein kann)

a) Alle α Koeffizienten in der Grundgesamtheit gleich, daher $V(\mu) = 0$.

b) Weniger effizient als FGLS (es gehen N–1 Freiheitsgrade wegen der N–1 Dummies verloren) während bei FGLS wegen der Schätzung von $V(\mu)$ nur ein Freiheitsgrad verloren geht.

c) Wegen omitted variables bias.

d) Weil die nicht dem wahren Modell entsprechende Störgröße nicht uncorrelated with explanators (x_{ijt}) ist.

e) Gilt für GLS nicht FGLS (nur konsistent).

Auf der Hauptdiagonale (farblich markiert) findet man die Kombinationen, in denen Modell und Schätzverfahren jeweils genau aufeinander abgestimmt sind (d. h. es wird das für den betrachteten Fall entwickelte Schätzverfahren angewendet).

Hierzu definieren wir die Blockmatrix[58] mit $\mathbf{X}^{\#} = \begin{bmatrix} \mathbf{D} & \mathbf{X}^{*} \end{bmatrix}$ der und den $(N+K) \times 1$ Blockvektor $\boldsymbol{\beta}^{\#} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}^{*} \end{bmatrix}$, dann ist Gl. 12 wie folgt kompakt darzustellen

(13) $\mathbf{y} = \mathbf{X}^{\#}\boldsymbol{\beta}^{\#} + \mathbf{u} = \begin{bmatrix} \mathbf{D} & \mathbf{X}^{*} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}^{L} \end{bmatrix} + \mathbf{u}$ so dass mit $\underset{K \times NT}{\mathbf{X}^{\#\prime}} = \begin{bmatrix} \mathbf{D}' \\ \mathbf{X}^{*\prime} \end{bmatrix}$ die OLS Schätzung zu

(14) $\hat{\boldsymbol{\beta}}^{\#} = \left(\mathbf{X}^{\#\prime}\mathbf{X}^{\#}\right)^{-1}\mathbf{X}^{\#\prime}\mathbf{y} = \begin{bmatrix} \mathbf{D}'\mathbf{D} & \mathbf{D}'\mathbf{X}^{*} \\ \mathbf{X}^{*\prime}\mathbf{D} & \mathbf{X}^{*\prime}\mathbf{X}^{*} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{X}^{*\prime} \\ \mathbf{D}' \end{bmatrix} \cdot \mathbf{y}$

führt. Wir verzichten darauf, den recht komplizierten Ausdruck für die inverse Blockmatrix anzuschreiben. Es lohnt sich aber, die Dimensionen der Matrix (Vektor) Produkte zu betrachten: $\mathbf{D}'\mathbf{D}$ ist eine $N \times N$ Diagonalmatrix mit jeweils NT (Skalar) in der Hauptdiagonale, also $\begin{bmatrix} NT & 0 & \Lambda & 0 \\ 0 & NT & \Lambda & 0 \\ M & M & O & M \\ 0 & 0 & \Lambda & NT \end{bmatrix}$, entsprechend gilt $\underset{N \times K}{\mathbf{D}'\mathbf{X}^{*}}$, $\underset{(k-1) \times N}{\mathbf{X}^{*\prime}\mathbf{D}}$ und $\underset{(k-1) \times (k-1)}{\mathbf{X}^{*\prime}\mathbf{X}^{*}}$ und somit $\underset{(N+K) \times (N+K)}{\mathbf{X}^{\#\prime}\mathbf{X}^{\#}}$ für

[58] Die Blockmatrix $\mathbf{X}^{\#}$ ist "partitioned by columns". Die Dimensionen sind bei den Matrizen (Vektoren) wie folgt

D	**D'**	$\mathbf{X}^{*}$	$\boldsymbol{\beta}^{*}$	**y** und **u**	**i** und **0**	$\boldsymbol{\alpha}$
$NT \times N$	$N \times NT$	$NT \times K$	$K \times 1$	$NT \times 1$	$T \times 1$	$N \times 1$

die Momentenmatrix. Ferner ist $\underset{(k-1)\times 1}{\mathbf{X}^{*\prime}\mathbf{y}}$ und $\underset{N\times 1}{\mathbf{D}'\mathbf{y}}$ so dass $\begin{bmatrix}\mathbf{X}^{*\prime}\\ \mathbf{D}'\end{bmatrix}\cdot\mathbf{y}=\begin{bmatrix}\mathbf{X}^{*\prime}\mathbf{y}\\ \mathbf{D}'\mathbf{y}\end{bmatrix}$ ein Spaltenvektor mit N+K Zeilen, so wie auch $\hat{\boldsymbol{\beta}}^{\#}$.

4.2. Ausblick

Im FE und RE Modell sind unterschiedliche (objektspezifische) α Koeffizienten vorgesehen während die β Koeffizienten nicht objekt- und periodenspezifisch sind. Es ist unüblich ein "umgekehrtes" Modell mit unterschiedlichen (objektspezifischen) β Koeffizienten aber (für alle Objekte) gleichen α Koeffizienten zu betrachten. Es dürfte auch nicht einfach sein unterschiedliche Koeffizienten β_{k1}, β_{k2}, …, β_{kN} (bei k = 1, …, K Regressoren) in Analogie zur Messung nicht-beobachteter Heterogenität der N Objekte (in Gestalt unterschiedlicher Koeffizienten α_1, α_2, …, α_N zu interpretieren.

Es gibt auch Modelle, bei denen (sich auf alle Objekte gleichermaßen auswirkende) zeitliche Veränderungen (time fixed effects) modelliert werden. Auch hier kann mit Dummy Variablen ($D_s = 1$ für t = s und 0 für $t \neq s$) gearbeitet werden und es wäre dann $y_{jt} = \alpha_j + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \gamma_t D_t + u_{jt}$ mit $\alpha_j = \alpha + \mu_j$ zu schätzen (Modelle mit time fixed effects "control for variables that are constant across entities but evolve over time" [Stock/Watson, S. 361]).[59] Damit hat man neben den üblichen objektspezifischen Dummies bei der LSDV Schätzung im FEM auch periodenbezogene Dummies (man spricht dann auch von "two-way-fixed effects" oder "entity and time fixed effects"). Zu viele Dummy Variablen sind nicht zu empfehlen weil damit ein Verlust an Freiheitsgraden und evtl. auch Probleme mit Multikollinearität verbunden sind.

Weitere Verallgemeinerungen (und Verkomplizierungen) sind Betrachtungen in denen wesentliche Modellannahmen gelockert werden, wie etwa heteroskedastische und autokorrelierte Störgrößen oder eine Korrelation zwischen den stochastischen objekt- (individuen-) spezifischen Einflüssen, d.h. Lockerung der Annahme $E(\mu_j\mu_m) = 0$ (j m = 1, …, N) auf S. 12 oben.

Bemerkenswert sind auch Versuche für die abhängige Variable (Zielvariable) y andere als metrisch skalierte Variablen zuzulassen, insbesondere eine dichotome Variable (mit oder ohne einer zugrundeliegenden Ordnungsrelation), ferner ordinale, diskrete[60] oder auch gestutzte (truncated) und zensierte (censored) Zielvariablen sowie REM mit weiteren zufälligen Effekten insbesondere auch variablen Koeffizienten.[61] Für Panel mit großem T werden zunehmend Methoden der Zeitreihenanalyse angewendet. Es ist wichtig, unit root und cointegration Hypothesen zu testen. Die entsprechenden Tests sind (wegen unobserved heterogeneity) komplizierter als bei reinen Zeitreihendaten und es gibt inzwischen viel Literatur zu diesen Problemen.

Für die allgemeiner gehaltenen und v.a. auch für dynamische Panelmodelle gewinnt die Schätzung mit der generalized method of moments (GMM) an Bedeutung. GMM ist ein Oberbegriff für nichtlineare und lineare Schätzverfahren, u. a. auch der Methode der Instrumentalvariablen (IV) und der verallgemeinerten (GLS) und gewöhnlichen (ordinary) Methode der kleinsten Quadrate (OLS).

---

[59] In Stock/Watson, S. 354 wird der Modelltyp mit dem einfachen Fall von nur T = 2 eingeführt. Man kann dann auch y und alle x Variablen in Differenzen Δy und Δx_i ausdrücken und regressieren ("before and after" regression). Im Anhang wird gezeigt, dass solche Varianten des Panel Modells in EViews möglich sind. Sie werden hier jedoch nicht weiter behandelt.

[60] darunter insbesondere Zähldaten (count data), die in Form von Häufigkeit (Fallzahlen) bestehen.

[61] Zu einem Überblick vgl. A. Hamerle, Panel Modelle für qualitative Daten, Allgemeines Statistisches Archiv Bd. 78 (1994), S. 1 – 19.

Im bekannten Modell einer Regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ beruht die Schätzung mit OLS von $\boldsymbol{\beta}$ auf der Inversion der Matrix $\mathbf{X}'\mathbf{X}$ mit Momenten $\sum_t x_{kt}^2$ (und Produktmomenten $\sum_t x_{kt} x_{rt}$ zwischen Regressoren x_k und x_r) bei einschränkenden Annahmen über die Matrix $\boldsymbol{\Omega} = E(\mathbf{u}\mathbf{u}')$.[62] Man erhält bekanntlich $\hat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$ durch Minimierung von Σu^2.[63] Hierin enthält $\mathbf{X}'\mathbf{y}$ die Produktmomente $\sum_t y_t x_{kt}$. Es liegt nahe, die Methode in der Weise zu verallgemeinern, dass man weniger Restriktionen für $\boldsymbol{\Omega}$ vorsieht (GLS), sowie andere Variablen (z.B. "Instrumente") $\mathbf{Z}$ und Gewichtungsmatrizen $\mathbf{W}$ einführt.

So ist beispielsweise $\hat{\boldsymbol{\beta}}^{IV} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})$ oder der Schätzer bei der verallgemeinerten (Generalized) IV Methode $\hat{\boldsymbol{\beta}}^{GIV} = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{Z}'\mathbf{y})$. Mit der GMM wird die Verallgemeinerung noch etwas weiter getrieben. Man kann umgekehrt auch zeigen, dass spezielle Fälle der GMM Methode, wie etwa OLS aufgefasst werden können als die Bestimmung eines Vektors $\hat{\boldsymbol{\beta}}^{OLS}$ so dass für Matrizen (wie $\mathbf{X}'\mathbf{X}$)) und Vektoren (wie $\mathbf{X}'\mathbf{y}$) von Momenten gilt $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{X}'\mathbf{y}$. Dieser Zusammenhang zwischen Momenten (bekannt als "Normalgleichungen") zu fordern oder zu sagen, $\hat{\boldsymbol{\beta}}^{OLS}$ wird so bestimmt, dass Σu^2 als Funktion von $\hat{\boldsymbol{\beta}}$ minimiert wird, ist gleichbedeutend.

Abschließend sei bemerkt, dass die eingangs geäußerte Vermutung, dass nämlich die verschiedenen methodischen Entwicklungen, die heutzutage alle unter dem Namen "panel" oder "Längsschnitt" firmieren nicht sehr viele Berührpunkte zu haben scheinen, sich bestätigt haben dürfte. Es fällt schwer, Verbindendes und Gemeinsames in puncto Terminologie und Fragestellungen zu finden bei so unterschiedlichen Gegenständen wie Populationsdynamik[64] oder Verlaufsanalysen (z.B. bei "medical follow up studies", also bei den erwähnten Schätzungen von hazard rates und survival functions) einerseits und ökonometrischen Modellen andererseits.

---

[62] Man beachte, dass der Vektor $\boldsymbol{\beta}$ auch den Koeffizienten α neben $\beta_1, \ldots, \beta_K$ enthält.

[63] GMM beruht anders als OLS nicht auf einer Minimierung von Σu^2 sondern auf einer Wahl von solchen Zahlenwerten für die Koeffizienten mit denen bestimmte Bedingungen für Momente (etwa $\Sigma x_k u = 0$) erfüllt sind.

[64] Hierbei geht es um Modelle, die nicht nur Bestands- und Stromgrößen betrachten, sondern auch Erscheinungsformen der sog. "Eigendynamik" kennen (bekannte Phänomene dieser Art sind z.B. in der Bevölkerungsstatistik der durch Alterung [Älterwerden] entstehende "Pythoneffekt" oder der durch die Fruchtbarkeit entstehende "Echoeffekt").