Marco Giesselmann
Michael Windzio

Regressionsmodelle zur Analyse von Paneldaten

LEHRBUCH

STUDIENSKRIPTEN ZUR SOZIOLOGIE

Springer VS

Studienskripten zur Soziologie

Die Bände „Studienskripten zur Soziologie“ sind als in sich abgeschlossene Bausteine für das Bachelor- und Masterstudium konzipiert. Sie umfassen sowohl Bände zu den Methoden der empirischen Sozialforschung, Darstellung der Grundlagen der Soziologie als auch Arbeiten zu so genannten Bindestrich-Soziologien, in denen verschiedene theoretische Ansätze, die Entwicklung eines Themas und wichtige empirische Studien und Ergebnisse dargestellt und diskutiert werden. Diese Studienskripten sind in erster Linie für Anfangssemester gedacht, sollen aber auch dem Examenskandidaten und dem Praktiker eine rasch zugängliche Informationsquelle sein.

Marco Giesselmann • Michael Windzio

Regressionsmodelle zur Analyse von Paneldaten

Springer VS

Marco Giesselmann
DIW Berlin, Deutschland

Michael Windzio
EMPAS Bremen, Deutschland

ISBN 978-3-531-18694-8
ISBN 978-3-531-18695-5 (eBook)
DOI 10.1007/978-3-531-18695-5

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer VS

Einbandentwurf: KünkelLopka GmbH, Heidelberg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer VS ist eine Marke von Springer DE. Springer DE ist Teil der Fachverlagsgruppe Springer Science+Business Media
www.springer-vs.de

Inhaltsverzeichnis

Vorwort

Dieses Buch beschäftigt sich mit sozialwissenschaftlichen Paneldaten und hat das Ziel, verschiedene Analysemethoden und Techniken unter Berücksichtigung der Anforderungen des empirischen Forschers zu sortieren und aufzuarbeiten. Als *Panel* soll im Folgenden eine Datenstruktur bezeichnet werden, bei der für mehrere Untersuchungseinheiten jeweils mindestens zwei Messungen vorliegen und zusätzlich die Zeitintervalle zwischen den Messpunkten bei allen Versuchseinheiten identisch sind. Wichtige formale Abgrenzungen bestehen zu Querschnittsdaten (*Cross Section Data*), für deren Einheiten jeweils nur eine Messung durchgeführt wird, zu einfachen Zeitreihen (*Time Series Data*), bei denen nur eine Einheit (zu mehreren Zeitpunkten) untersucht wird, zu zeitversetzt erhobenen Querschnittsdaten (*Pooled Cross Section Data*) und schließlich zu Ereignisdaten (*Event History Data*), deren Messpunkte ereignisabhängig sind und somit von Einheit zu Einheit variieren. Da sich Paneldaten und Ereignisdaten darin gleichen, dass Informationen über Einheiten zu mehreren Zeitpunkten vorliegen, werden diese Typen häufig auch unter dem Begriff *Längsschnittdaten* zusammengefasst. Gelegentlich werden zudem gepoolte Querschnittsdaten, aufgrund der unterschiedlichen Erhebungszeitpunkte, zu dieser Familie von Datentypen gezählt[1].

Paneldaten sind in der modernen Sozialwissenschaft eine häufig genutzte empirische Basis zur Überprüfung von Hypothesen. Allerdings ist dieser Trend relativ neu – zumindest außerhalb der Ökonomie. Der wichtigste Grund hierfür ist, dass sich die elementaren sozialwissenschaftlichen Fragestellungen in den letzten Jahrzehnten verändert haben.

So hat sich in der modernen, auf Makrodaten basierenden Politikwissenschaft die ländervergleichende Analyse als empirische Königsdisziplin herausgebildet. Um Effekte institutioneller Veränderung zu messen, werden nicht nur Veränderungen in den Rahmenbedingungen eines Landes im Zeitverlauf, sondern gleichzeitig Unterschiede zwischen Ländern untersucht. Das zu diesem Ansatz korrespondierende Datenformat ist das Panel.

---

1 Eine ausführliche Dar- und Gegenüberstellung der verschiedenen Längsschnittformate sowie deren Ausformungen in der Praxis bietet Ruspini (2002, Kapitel 2 und 3).

Ähnliches gilt für die empirische Soziologie. Hier rückte mit den Thesen zur Pluralisierung der Gesellschaft in den 1980er Jahren der Lebenslauf in den Fokus der Disziplin: Während sich Probleme der klassischen Soziologie zuvorderst auf Unterschiede *zwischen* Individuen beziehen, versucht die moderne Soziologie zusätzlich, Auswirkungen von intraindividuellen Differenzen bzw. von Ereignissen (Heirat, Geburt eines Kindes, Scheidung, Arbeitslosigkeit etc.) *innerhalb* individueller Lebensläufe zu bestimmen.

Bei solchen Fragestellungen entfalten Paneldaten ihr Potenzial: Sie gestatten die Betrachtung von Auswirkungen der Ereignisse auf der individuellen Ebene (z. B. im Rahmen eines Vorher/Nachher-Vergleichs) und damit die Annäherung an eine experimentelle Analyselogik. Dabei ermöglichen sie die Kontrolle von unbeobachteten, personen- bzw. einheitenspezifischen Merkmalen[2]. Folglich beinhalten sie ein Potenzial zur Absicherung der in Querschnittsanalysen grundsätzlich kritischen Exogenitätsannahme[3].

Gelegentlich ist die Wahl eines Paneldatensatzes als Basis der Datenanalyse aber nicht in der Forschungsfrage (bzw. dem Motiv, unbeobachtete Heterogenität zu kontrollieren) angelegt, sondern einer wissenschaftlichen Infrastruktur geschuldet, die mittlerweile viele zentrale Datensätze im Panelformat bereitstellt. So kann beobachtet werden, dass mehr und mehr sozialwissenschaftliche Studien auf Basis des Sozio-oekonomischen Panels (SOEP) durchgeführt werden, obgleich zur Bearbeitung vieler der untersuchten Fragestellungen aus analytischer Perspektive ein Querschnitts- bzw. gepoolter Querschnittsdatensatz ausreichen würde. Ein Forscher, der beispielsweise den Effekt der sozialen Herkunft auf das Einkommen untersucht, ist auf der Grundlage seiner Fragestellung nicht auf Paneldaten angewiesen, denn schließlich variiert die soziale Herkunft einer Person nicht im Lebenslauf. Da aber die Möglichkeiten zur Messung sozialer Kontexte im SOEP im Vergleich zu verfügbaren Querschnittserhebungen überlegen sind, wird der Forscher nun trotzdem mit Paneldaten arbeiten – und muss sich entsprechend auch mit Methoden zu ihrer Analyse auseinandersetzen. Natürlich

2 Mittlerweile wird in den meisten neueren ökonometrischen Lehrbüchern und methodischen Abhandlungen dieses Potenzial als zentrales Motiv zur Verwendung von Paneldaten anerkannt und Methoden vor dem Hintergrund dieses Motivs aufbereitet und diskutiert (z. B. Allison 2009, Brüderl 2010). Das hier von uns vorgelegte Lehrbuch bewegt sich in der Tradition dieser Ansätze.

3 Korreliert in einem statistischen Modell die erklärende Variable x mit einem unbeobachteten Merkmal, welches einen Einfluss auf y hat, bezeichnet man x als *endogene erklärende Variable*. Wird der Zusammenhang zwischen x und y als Effekt von x auf y interpretiert, so impliziert dieses (u.a.) die Annahme, dass x eine exogene erklärende Variable ist, also *nicht* mit unbeobachteten Merkmalen korreliert. Die praktisch nicht auszuschließende Möglichkeit der Verletzung dieser Annahme ist das grundlegende Problem von Querschnittsstudien und führt immer wieder zu Kritik an dieser Art des empirischen Zugangs und den Wissenschaften, die ihn verwenden.

könnte in solchen Fällen auch einfach eine Welle des Panels als Querschnittsdatensatz verwendet werden. Allerdings ist ein solches Vorgehen häufig ineffektiv, da sich bei Verwendung mehrerer Wellen (und einer entsprechend vergrößerten Stichprobe) Zusammenhänge klarer zeigen und statistisch besser absichern lassen[4].

Mit der zunehmenden Verbreitung von Paneldaten verschiebt sich auch das benötigte methodische Grundlagenwissen des praktisch arbeitenden Sozialwissenschaftlers. Die dafür erforderlichen Kompetenzen lassen jedoch anhand des vorhandenen, vorwiegend in englischer Sprache vorliegenden Lehrbuchmaterials nur schwer aneignen: Ökonometrische Lehrbuchtexte zur Panelanalyse verlangen ein über die multiple Regressionsanalyse weit hinausweisendes Vorwissen. Dass beim Erlernen von Techniken der sozialwissenschaftlichen Panelanalyse häufig komplizierte Umwege in Kauf genommen werden müssen, hängt auch damit zusammen, dass die Techniken zur Analyse von Längsschnittdaten in der Ökonomie entwickelt wurden. Dabei herrscht zumeist eine *large t, small n*-Situation vor, also eine Datensatzstruktur, die aus einer überschaubaren Anzahl von Einheiten mit vielen Messpunkten besteht[5].

Die Daten sozialwissenschaftlicher Panel weisen allerdings häufig eine *small t, large n*-Struktur auf und erfordern daher alternative Analysemethoden. Zudem variieren viele in den Humanwissenschaften verwendete Merkmale im Zeitverlauf nur geringfügig: Im Gegensatz zu betriebs- oder volkswirtschaftlichen Kennzahlen verändern sich biographische Variablen in der Regel nur an wenigen Punkten im Lebensverlauf eines Individuums und weisen über Jahrzehnte hinweg dieselbe Ausprägung auf. Ein hinreichendes Maß an Variation, welches viele typische wirtschaftswissenschaftliche Methoden voraussetzen, ist daher außerhalb der Ökonomie häufig nicht gegeben.

---

4 Neben den hier genannten Vorteilen (*Kontrolle von Heterogenität*, *Vergrößerung der Stichprobe*) sind weitere Motive zur Verwendung von Paneldaten einschlägig (z. B. Baltagi 2005, Brüderl 2010): Einerseits können mit Paneldaten Ereignisse und Zustände, welche die Verknüpfung von Informationen verschiedener Zeitpunkte erzwingen (z. B. *Langzeitarbeitslosigkeit*, *Scheidung*), zuverlässiger identifiziert werden. Auch können Merkmale, die jeweils nur in bestimmen Lebensphasen valide gemessen werden können, gemeinsam modelliert werden (z. B. *frühkindliche Betreuungssituation* und *Intelligenz im Erwachsenenalter*). Zudem lassen sich mit Paneldaten komplexe dynamische Prozesse (*wie wirkt sich die Einkommenshöhe zu einem bestimmten Zeitpunkt auf das Einkommen im nächsten Jahr aus?*) modellieren. Als weiteres Motiv zur Verwendung von Paneldaten kann die Identifikation von Dynamiken hinter Trends – also beispielsweise Einstiegs- und Ausstiegsraten aus der Armut – betrachtet werden.

5 Dieses offenbart sich bereits in der unter Ökonometrikern prominenten Bezeichnung *Time-Series Cross-Section Data* für Paneldaten.

Die *Mehrebenenanalyse* wird ebenfalls als Methode zur Analyse sozialwissenschaftlicher Paneldaten angeboten. Dieser Begriff ist dabei bis heute nicht eindeutig definiert – in der Regel ist aber die Analyse geschachtelter Datenstrukturen mit der sog. *Random Effects*-Technik gemeint. Hierbei handelt es sich um ein Verfahren, dessen Ursprung in der Biometrie liegt. Es wird in den Sozialwissenschaften immer dann verwendet, wenn die Einbettung von Individuen in einen (zumeist sozialen) Kontext von Bedeutung für die Ausprägung eines abhängigen Merkmals ist[6]. Obgleich Paneldaten ebenfalls als geschachtelte Daten aufgefasst und dementsprechend auch Techniken der Mehrebenenanalyse angewendet werden können, greifen diese die Bedürfnisse des mit Paneldaten operierenden Wissenschaftlers nur bedingt auf. Schließlich stehen beim mathematisch anspruchsvollen *Random Effects*-Verfahren die Modellierung der Variationseigenschaften des abhängigen Merkmals im Vordergrund, nicht die in der Panelanalyse zumeist benötigte Modellierung des Längsschnittes oder die panelspezifische Korrektur der Teststatistik.

Techniken, die den Längsschnitt als Prozess intraindividueller Veränderungen modellieren und korrekte Teststatistiken liefern, sind zumeist einfacher zu erlernen. So ist aus unserer Sicht die Wahl und Ausführung der richtigen Techniken zur Analyse sozialwissenschaftlicher Paneldaten schon in der sorgfältig formulierten Fragestellung angelegt. Diese Techniken können als einfache Erweiterungen der multiplen Regressionsmethode aufgefasst und dargestellt werden, was wir mit dem vorliegenden Lehrbuch demonstrieren wollen. Wir verzichten daher bei der didaktischen Aufarbeitung des Stoffes zumeist auch auf Darstellungen in der Matrixform. Eine Ausnahme stellen die in Kapitel 8 erörterten Strukturgleichungsmodelle dar.

Schwerpunktmäßig behandelt dieses Buch die Analyse von Paneldatensätzen, bei denen es viele Untersuchungsobjekte ($n>100$) und wenige Beobachtungspunkte ($t < 20$) gibt. Diese so genannte *large n, small t*-Situation kennzeichnet das Format der meisten Personen- oder Haushaltsdatensätze und wird in diesem Buch als Normalfall behandelt. Die umgekehrte, *small n, large t*-Situation liegt häufig bei der Analyse von Aggregatdaten vor, beispielsweise wenn Länder oder Firmen die Einheiten der Untersuchung bilden. Spezifische methodische Herausforderungen der Analyse solcher Daten werden zwar skizziert, spielen bei der Strukturierung des Buches und der Diskussion der Methoden aber eine untergeordnete Rolle.

6 Zum Beispiel: Die Leistung eines Schülers hängt nicht nur vom eigenen sozialen Status ab, sondern auch vom sozialen Status der Schülerinnen und Schüler, die ihn umgeben. Oder: Die Wahlwahrscheinlichkeit des Individuums für eine extreme Partei hängt nicht nur von dem eigenen Beschäftigungsstatus ab, sondern auch von der Arbeitslosenquote im Wohngebiet des Individuums.

Somit wäre der Inhalt und die Grenzen des in diesem Buch behandelten Stoffes definiert: Es geht um die verschiedenen Möglichkeiten, regressionsbasierte Verfahren auf sozialwissenschaftliche Paneldaten anzuwenden, sowie um die Diskussion der Eignung dieser Verfahren in verschiedenen empirischen Situationen. Die Ausarbeitung dieser Themen adressiert insbesondere Anfänger der Panelanalyse. Gleichzeitig wollen wir mit dem Buch den wissenschaftlichen Diskurs um die Eigenschaften und die Angemessenheit verschiedener Methoden befruchten.

Die einzelnen Abschnitte sind dabei wie folgt aufgebaut: In der Einführung stellen wir zunächst unsere Notation vor. Anschließend wiederholen wir das einfache Regressionsverfahren mit Querschnittsdaten und erklären den grundlegenden Unterschied zu Regressionsverfahren mit Paneldaten. Schließlich wird die Regressionsgleichung, als Ausgangspunkt der Datenanalyse, so erweitert, dass sie die Eigenschaften von Paneldaten aufgreift.

Im nächsten, zweiten Kapitel des Buches werden Regressionsmethoden für Paneldaten vorgestellt. Zunächst führen wir ein grundlegendes Kriterium zur Auswahl der angemessenen Methode ein, nämlich die Form der Fragestellung: *Liegt eine Querschnitts- oder Längsschnittfragestellung vor?* Im Folgenden gehen wir zunächst auf Längsschnittfragestellungen ein: Was ist der Grund dafür, eine Frage im Längsschnitt zu formulieren? Warum können Längsschnittfragestellungen nur unbefriedigend mit Querschnittsdaten beantwortet werden? Und schließlich: Wie muss das einfache Regressionsverfahren erweitert werden, um das in Paneldaten angelegte Potenzial zur getreuen Beantwortung von Längsschnittfragestellungen zu realisieren? Schließlich folgt die Darstellung dieser erweiterten Regressionsverfahren, nämlich das *First Differences*-Verfahren sowie mehrere Varianten des *Fixed Effects*-Verfahrens.

Sodann gehen wir im dritten Kapitel auf Fälle ein, in denen das Motiv zur Verwendung von Paneldaten nicht in der Längsschnittfragestellung liegt, sondern eine Querschnittsfragestellung bearbeitet werden soll. Zunächst beschreiben wir solche Fälle. Dann stellen wir Verfahren zur Analyse von Querschnittsfragestellungen mit Paneldaten vor, nämlich, als relativ simple Möglichkeit, *Regressionsschätzungen mit korrigiertem Standardfehler* sowie die etwas komplizierteren, aber in einigen Fällen auch fruchtbareren *Random Effects*-Verfahren. Zudem diskutieren wir die Anwendungsmöglichkeiten der in der Praxis selten zu beobachtenden *Between Regression*.

Vor dem Hintergrund des im dritten Kapitel erläuterten *Random Effects*-Verfahrens nehmen wir im vierten Kapitel die Modellierung einer Längsschnittfragestellung wieder auf und diskutieren die Eignung des *Random Effects*-

Verfahrens zu deren Analyse. Dabei widmen wir uns auch dem zunehmend an Popularität gewinnenden *Hybrid Verfahren*, welches die grundlegenden Eigenschaften von *Random Effects* und *Fixed Effects* vereinigt.

Im fünften Kapitel werden die Schritte zur richtigen Auswahl der Technik zusammengefasst und um eine weitere Dimension, nämlich statistische Entscheidungshilfen, erweitert. Im Mittelpunkt steht hier der in der Praxis häufig zur Bestimmung des korrekten Verfahrens angewendete *Hausman-Test*.

Im daran anschließenden Kapitel 6 werden Probleme bei der Analyse von Paneldaten behandelt, die eine über die grundlegenden Verfahren hinausweisende Technik erzwingen. Solche Probleme entstehen dadurch, dass die Abhängigkeitsmuster zwischen den Messungen von Paneldaten in einigen Fällen so komplex sind, dass sie durch die vorgestellten Methoden nicht hinreichend modellierbar sind. Wir beschränken uns hierbei auf die Darstellung der Probleme, nennen mögliche Behandlungsmöglichkeiten und verweisen auf fortgeschrittene Lehrbücher.

Kapitel 7 beschäftigt sich mit Panelregressionen für dichotome abhängige Variablen. Zunächst wird auch hier das einfache Regressionsverfahren zur Berechnung solcher Modelle wiederholt, die *logistische Regression*. Anschließend wird das Konzept der logistischen Regression erweitert und es werden Verfahren vorgestellt, die analoge Eigenschaften zu den linearen Regressionen für Paneldaten besitzen. Dabei wird die besondere Herleitung des *Fixed Effects*-Verfahrens ausführlicher beschrieben und es werden einige weitere Eigenheiten der binären logistischen Regression für Paneldaten erläutert.

In den Kapiteln 1 bis 7 wird das klassische Regressionsmodell für Paneldaten erweitert und für dichotome abhängige Variablen generalisiert. Das Konzept der Regression basiert in diesen Kapiteln auf der Vorhersage einer abhängigen Variablen durch einen Satz unabhängiger Variablen. In Kapitel 8 wird nun das Regressionsmodell zu einem *System von Gleichungen* verallgemeinert. Es wird in knapper Form das Grundprinzip der Strukturgleichungsmodelle (*Structural Equation Modeling*, SEM) dargestellt und auf die Möglichkeit der Schätzung auch von *Fixed Effects*-Modellen mittels SEM hingewiesen. Dieser Ansatz ermöglicht zudem die Schätzung *latenter Wachstumsmodelle* und eine überaus flexible Modellierung paralleler Prozesse. In diesem Buch wird SEM als alternativer Ansatz zu der zuvor diskutierten Panelregression vorgestellt. In der Regel ist die Panelregression zentraler Bestandteil ökonometrischer Lehrbücher zur Längsschnittanalyse, während insbesondere in der Psychologie Paneldaten häufig mit SEM ausgewertet werden. Leider scheinen beide Perspektiven eher unvermittelt nebeneinander zu stehen bzw. sich distanziert zueinander zu verhalten. Lehrbücher, die beide Perspektiven berücksichtigen, sind daher selten (Allison 2009). Sicher können wir die Kluft zwischen beiden Perspektiven nicht

überwinden. Wenngleich eine fundierte Einführung in SEM den Rahmen dieses Buches gesprengt hätte (vgl. dazu Reinecke 2005; Geiser 2010), möchten wir in der Tradition von Singer und Willet (2003) sowie Allison (2009) aber zumindest ausdrücklich auf die Möglichkeiten der Strukturgleichungsmodelle für die Analyse von Paneldaten hinweisen.

Zur didaktischen Motivation der Methoden sowie zur beispielhaften Illustration ihrer Implikationen arbeiten wir in allen Kapiteln mit verschiedenen (teils echten, teils konstruierten) Beispieldatensätzen. Diese werden jeweils dann, wenn sie erstmals verwendet werden, kurz eingeführt. Außerdem haben wir als Ergänzung zu diesem Lehrbuch eine Homepage eingerichtet, von welcher die Datensätze geladen und so unsere Beispielanalysen reproduziert und erweitert werden können.

```
http://www.barkhof.uni-bremen.de/~mwindzio/lebensz.dta
http://www.barkhof.uni-bremen.de/~mwindzio/CPDS.dta
http://www.barkhof.uni-bremen.de/~mwindzio/growth1.dat
```

Alle im Buch verwendeten Dateien sowie die Analysesyntax und ado-files finden sich komprimiert unter:

```
http://www.barkhof.uni-bremen.de/~mwindzio/panel.zip
```

Die Analysedatei „lebensz.dta“ wird im ersten und dritten Teil des Buches (Kapitel 1-6 und Kapitel 8) verwendet und stellt einen voll anonymisierten und verfremdeten[7], auf wenige Variablen limitierten Ausschnitt des Sozio-oekonomischen Panels (SOEP) dar. Das SOEP ist eine jährliche Befragung von Personen in mehreren Tausend Haushalten in Deutschland, die seit 1984 durchgeführt wird. Information zum Originaldatensatz, seinen Inhalten, der Struktur sowie den Bezugsbedingungen können auf der Homepage der SOEP-Abteilung des DIW Berlin abgerufen werden.

```
www.diw.de/soep
```

„CPDS.dta“ enthält dagegen Informationen über Länder und liegt den Analysen im zweiten Teil des Buches zugrunde (Kapitel 7).

Mit Ausnahme der Strukturgleichungsmodelle wurden sämtliche Analysen mit dem Statistik-Programmpaket STATA durchgeführt; auf den entsprechenden Programmiercode wird jeweils in einer Fußnote verwiesen. Im STATA-Format

7 Die Verfremdungsprozedur beruht im Kern auf einem Algorithmus von Kohler (2003).