BestMasters

Michael Seitz

Intervalldaten und generalisierte lineare Modelle

Springer Spektrum

BestMasters

Mit „BestMasters“ zeichnet Springer die besten Masterarbeiten aus, die an renommierten Hochschulen in Deutschland, Österreich und der Schweiz entstanden sind. Die mit Höchstnote ausgezeichneten Arbeiten wurden durch Gutachter zur Veröffentlichung empfohlen und behandeln aktuelle Themen aus unterschiedlichen Fachgebieten der Naturwissenschaften, Psychologie, Technik und Wirtschaftswissenschaften.
Die Reihe wendet sich an Praktiker und Wissenschaftler gleichermaßen und soll insbesondere auch Nachwuchswissenschaftlern Orientierung geben.

Michael Seitz

Intervalldaten und generalisierte lineare Modelle

Mit einem Geleitwort von Prof. Dr. Thomas Augustin

Springer Spektrum

Michael Seitz
München, Deutschland

BestMasters
ISBN 978-3-658-08745-6 ISBN 978-3-658-08746-3 (eBook)
DOI 10.1007/978-3-658-08746-3

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

Springer Spektrum

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Geleitwort

Seit längerem herrscht in der angewandten Statistik ein Unbehagen, dass viele gängige statistische Verfahren eigentlich voraussetzen, dass die Daten in einer Qualität und Präzision vorliegen, die oft nicht gewährleistet werden kann. Gerade in großen Studien zu komplexen Themen treten häufig nicht nur Antwortverweigerungen auf, sondern die erhobenen Daten sind zusätzlich durch Antworteffekte verzerrt und unpräzise. Ein charakteristisches Beispiel sind *Intervalldaten*: In Surveys runden die Befragten oft grob und bevorzugen bestimmte "attraktive Werte". Ferner wird bei der Konzeption von Fragebögen häufig empfohlen, Angaben zu heiklen Fragen (wie etwa dem Einkommen) von vornherein nur in Intervallen zu erheben, um Verweigerungen zu vermeiden, oder es werden von primären Verweigerern wenigstens Intervallangaben erbeten. Bei ereignisanalytischen Modellen kommt zusätzlich oft noch das der durch individuelle Beobachtungszeitpunkte induzierte Problem der Intervallzensierung hinzu.

Sensibilisiert dafür, dass eine solche "Defizitäten" negierende, naive Anwendung von statistischen Verfahren unter Umständen zu gravierenden Verzerrungen der Ergebnisse – und damit zu folgenschweren inhaltlichen Fehlinterpretationen – führen kann, wurden in der Literatur einerseits genauere Kriterien gewonnen, wann mit starken Verzerrungen zu rechnen ist, sowie entsprechende Korrekturverfahren entwickelt, um die Verzerrungen durch Messfehler, Vergröberungen und fehlende Daten zu kompensieren.

Diese theoretisch sehr leistungsfähigen Korrekturmethoden stellen einen wichtigen Fortschritt dar; ihre praktische Relevanz bleibt aber mit dem Makel behaftet, dass die Korrekturen typischerweise auf weitreichenden Zusatzannahmen über den "Defizitäts"- prozess" beruhen, die empirisch nicht überprüfbar und oft auch inhaltlich nicht begründbar sind. Deshalb findet in den letzten Jahren insbesondere im Bereich der systematisch fehlenden Daten eine prinzipiell andere Vorgehensweise immer mehr Anklang, die v.a. unter dem Begriff partielle Identifikation bekannt wurde: Man verzichtet bewusst auf die vermeintliche Präzision von unter Einbezug von artifiziellen Zusatzannahmen erzielten Modellschätzungen und lässt "mengenwertige Ergebnisse" zu, indem man die Menge aller mit den Daten (und inhaltlich unbezweifelter Zusatzannahmen) verträglichen Modelle betrachtet. Diese so erzielten

Ergebnisse sind zwar häufig unpräzise, aber durch die Art ihrer Gewinnung glaubwürdiger und zuverlässiger, und, wie sich mittlerweile in vielfältigen Studien gezeigt hat, oft immer noch präzise genug, um die eigentlichen substanzwissenschaftlichen Fragestellungen beantworten zu können.

Die Arbeit von Herrn Seitz ist genau diesem neuem Paradigma einer zuverlässigen Inferenz unter unvollständiger Information verpflichtet. In vielen Situationen erscheint es bei Intervalldaten äußerst bedenklich, anzunehmen, der Vergröberungsmechanismus sei nichtinformativ zufällig, und so sind Intervalldaten ein natürlicher Anwendungsbereich für Methoden der partiellen Information. In der Tat gibt es eine Reihe von entsprechenden Ansätzen für lineare Regressionsmodelle unter Intervalldaten, aber überraschenderweise praktisch keine Literatur zu verallgemeinerten linearen Modellen i. e. S., wie sie sich mittlerweile in der statistischen Modellierung als Standardmethode durchgesetzt haben. Herr Seitz präsentiert hier einen sehr allgemeinen Ansatz, der sich im Prinzip auch noch allgemeiner auf beliebige unverzerrte Schätzgleichungen ausdehnen lässt. Nach einer eher als Kontrollsituation dienenden direkten Lösung im Fall eines eindimensionalen Parameters wird die allgemeine Aufgabe konsequent als Optimierungsproblem über die Parameter gesehen, wobei die Intervalldaten in natürlicher Weise lineare Restriktionen formulieren. Um genau die Menge aller Maximum-Likelihood-Schätzer als zulässige Lösungen zu erhalten, wird die Bedingung "jeweiliger Wert der Scorefunktion = 0" zunächst als nichtlineare Nebenbedingung eingebracht. Die Arbeit entwickelt sodann mehrere Methoden, wie dieser Ansatz konkret operational gemacht werden kann. Eine Idee unter anderen besteht darin, diese Nebenbedingungen mit einem Pönalisierungstrick in die Zielfunktion mitaufzunehmen, so dass nun ein nichtlineares Optimierungsproblem über einem konvexen Polyeder entsteht. Die verschiedenen Methoden werden v.a. für die Exponentialregression konkret implementiert, verglichen und sorgfältig evaluiert.

Die Ergebnisse der Arbeit von Herrn Seitz erlauben erstmalig die konkrete Implementation von Methoden der partiellen Identifikation in generalisierten linearen Modellen unter Intervalldaten und erweitern damit den möglichen Anwendungsbereich dieses neuen methodischen Paradigmas substantiell. Sie sind ein wichtiger Meilenstein auf dem Weg zu einem allgemeinen Rahmen für eine zuverlässige statistische Modellierung komplexer Daten.

Prof. Dr. Thomas Augustin

Vorwort

Mit der Theorie der partiellen Identifizierung wird Unsicherheit in den beobachteten Daten auf die Schätzung der Modellparameter übertragen. Im Kontext der generalisierten Regression mit Intervalldaten erhält man für die Parameterschätzer so keine skalaren Werte, sondern ebenfalls Intervalle. Im Allgemeinen kann dies als Optimierungsproblem mit Nebenbedingungen formuliert werden. Die Herausforderung besteht hier insbesondere darin, die globalen Extrema zu bestimmen. Für Spezialfälle sind bereits Lösungen aus der Literatur bekannt: Bei der linearen Regression mit skalarer abhängiger Variable kann das Optimierungsproblem analytisch gelöst werden. Für eindimensionale Parameter in generalisierten linearen Modellen wird in der vorliegenden Arbeit ein neuer Ansatz realisiert und untersucht: Optimiert man die Score-Funktion mit festem Parameter über die beobachteten Intervalldaten, so lassen sich dadurch die Extrema des Parameterschätzers bestimmen. Dieser Ansatz liefert zuverlässig die richtige Lösung. Als allgemeine Herangehensweise kann der Parameterschätzer direkt numerisch optimiert werden. Daneben wird ein alternativer Ansatz verwendet: Um das Intervall der zulässigen Schätzer zu erhalten, kann die Score-Funktion als Strafterm in die Zielfunktion integriert werden. Die Methoden werden für das lineare Modell und das generalisierte lineare Modell mit Exponentialverteilung und log-Link umgesetzt. Bei der Untersuchung von Simulationsbeispielen zeigt sich, dass die allgemeinen Verfahren teilweise nur lokale Extrema finden. Da die Zielfunktionen aber mitunter sehr hochdimensional sind, kann die numerische Optimierung nicht an verschiedenen Stellen systematisch neugestartet werden. Auf Grund dieser Problematik werden heuristische Methoden vorgeschlagen, bei denen iterativ in den Ecken der Datenintervalle neugestartet wird. Diese liefern durchwegs die größten Intervalle der zulässigen Parameter und damit die beste Lösung. Die Methode wird auf ein Anwendungsbeispiel aus der Volkswirtschaft angewandt: Vorerst wird ein Regressionsmodell für den Zusammenhang zwischen dem Bruttoinlandsprodukt und den Ausgaben für Forschung und Entwicklung verschiedener Länder entwickelt. Anschließend werden Intervalle für die Daten konstruiert und die entsprechenden Parameterintervalle berechnet. Mit den vorgestellten und praktisch umgesetzten Verfahren lassen sich oft die exakten Lösungen und in allen untersuchten Fällen sehr gute Approximationen finden. Des Weiteren wird demonstriert, dass

die Ansätze grundsätzlich zur Lösung des Problems führen, wenn die Extrema der Optimierungsprobleme richtig bestimmt werden.

Michael Seitz

Inhaltsverzeichnis

Abbildungsverzeichnis