Jahresber Dtsch Math-Ver (2013) 115:51–55
DOI 10.1365/s13291-013-0058-2

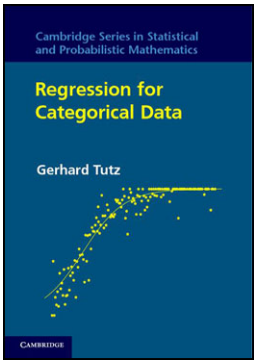
BOOK REVIEW

Gerhard Tutz: „Regression for Categorical Data“

Cambridge University Press, 2012, 572 pp.

Thomas Kneib

Online publiziert: 6. Februar 2013



Lineare Regressionsmodelle stellen eines der wesentlichen Werkzeuge der Angewandten Statistik dar, wenn eine Zielgröße in Abhängigkeit von einer Reihe von erklärenden Variablen dargestellt werden soll. Während im klassischen linearen Modell die Zielgröße typischerweise als normalverteilt oder zumindest als stetig vorausgesetzt wird, ist diese Annahme in vielen praktischen Beispielen nicht mehr haltbar. Insbesondere im Falle kategorialer, also diskreter Zielgrößen müssen geeignete Erweiterungen betrachtet werden. In seinem Buch „Regression for Categorical Data“ behandelt Gerhard Tutz solche Erweiterungen, wobei kategoriale Daten in einem relativ weiten Sinn verstanden werden, so dass neben klassischen Ansätzen der kategorialen Regression für binäre und multinomiale Zielgrößen oder Zähldaten auch beispielsweise Modelle mit Zero-Inflation behandelt werden.

Ein typisches Beispiel eines kategorialen Regressionsmodells lässt sich für die Wahl eines Transportmittels für Fernreisen aus einer vorgegebenen Liste von Alternativen (Flugzeug, Zug, Bus, Auto) formulieren. Ziel der Analyse ist es hier, die Wahrscheinlichkeit für die Wahl eines bestimmten Transportmittels in Abhängigkeit beispielsweise von Preis, Reisezeit und anderen Einflussgrößen zu bestimmen. Da das lineare Modell eine stetige Zielgröße unterstellt, ist es zur Beschreibung der

T. Kneib (✉)
Lehrstuhl für Statistik, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5,
37073 Göttingen, Deutschland
e-mail: tkneib@uni-goettingen.de

kategorialen Zielgröße „gewähltes Transportmittel“ offenbar ungeeignet. Weitere im vorliegenden Buch besprochene Beispiele, die exemplarisch für andere Typen kategorialer Zielgrößen stehen können, sind etwa

- die Schätzung von Kreditausfallwahrscheinlichkeiten mit binärer Zielgröße (Kredit wurde zurückgezahlt oder nicht zurückgezahlt) in Abhängigkeit von Kredithöhe, Laufzeit des Kredits und früherem Zahlungsverhalten des Kunden (binäre Regressionsmodelle),
- die Modellierung der Anzahl insolventer Firmen in einem Monat in Abhängigkeit von der Kalenderzeit, um Konjunkturentwicklungen zu analysieren (Zähldaten-Regression), oder
- die Analyse des Behandlungserfolgs in einer Schmerztherapie nach Knie-Operationen, bei denen die Schmerzintensität auf einer geordneten Skala mit fünf Punkten (von „keine Schmerzen“ bis „starke Schmerzen“) beurteilt wird (Regression für ordinale Zielgrößen).

Neben der Beschreibung geeigneter Klassen von Regressionsmodellen für diese verschiedenen Typen von Zielgrößen werden die Prädiktorstruktur und die zum Schätzen verwendete Methodik im vorliegenden Buch dahingehend erweitert, dass moderne Ansätze der statistischen Regularisierung mit einem besonderen Fokus auf die Prädiktorselektion ebenfalls einbezogen werden können. Schließlich beinhaltet das Buch auch eigene Kapitel zu semiparametrischen Regressionsansätzen für kategoriale Daten, zu baum-basierten Verfahren und zur Prädiktion in kategorialen Regressionsmodellen.

Bei dem Buch des Kollegen Tutz handelt es sich um eine englische Übersetzung und erhebliche Erweiterung des 2000 im Oldenbourg Verlag erschienen Buchs „Analyse Kategorialer Daten“. Im Vergleich zu der deutschen Version hat sich der Umfang des Buchs deutlich erhöht (um nahezu 150 Seiten), so dass der Leser einen deutlich erweiterten und aktualisierten Inhalt vorfindet. Insbesondere die Methoden der Regularisierung sowie das Kapitel zur Prädiktion sind Ergänzungen, die das englischsprachige Buch auch für Besitzer des deutschen Bandes empfehlenswert machen.

Das Buch ist insgesamt auf einem mittleren mathematischen Niveau angesiedelt und setzt im Wesentlichen Basiswissen in Linearer Algebra und Wahrscheinlichkeitsrechnung voraus. Grundlagen des linearen Modells werden zu Beginn des Buchs kurz wiederholt, Hintergrundwissen zu linearen Modellen ist aber sicherlich dennoch empfehlenswert. Zielgruppe des Buchs sind Statistiker, Forscher unterschiedlicher Anwendungsbereiche mit solider statistischer Ausbildung sowie Studierende der Statistik, Mathematik oder anderer Fachgebiete mit quantitativem Schwerpunkt.

Ergänzend zum Buch werden auf der Homepage http://www.stat.uni-muenchen.de/~tutz/catdata eine Reihe von Datensätzen zur Verfügung gestellt. Darüber hinaus sind die Datensätze sowie R-Code für zahlreiche der durchgeführten Analysen in Form eines R-Pakets erhältlich. SAS-Code steht für einige ausgewählte Anwendungen zur Verfügung, wobei insbesondere die moderneren Erweiterungen und Verfahren nicht für SAS erhältlich sind, da entsprechende Implementationen in SAS bisher nicht vorhanden sind.

Kapitel 1 des Buchs bietet eine kurze Einleitung, in der eine Reihe von Beispielen vorgestellt und einige grundlegende Konzepte besprochen werden. Ergänzend wird

ein kurzer Überblick zu linearen Modellen geboten, um die für das weitere Verständnis notwendigen Grundlagen bereitzustellen bzw. zu wiederholen.

Die anschließenden Kapitel 2 bis 7 bilden einen größeren Block zu parametrischen Regressionsmodellen für verschiedene Typen von univariaten Zielgrößen. Als erstes grundlegendes Beispiel wird in Kapitel 2 („Binary Regression: The Logit Model“) ausführlich das Logit-Modell vorgestellt, mit dem sich binäre Zielgrößen wie die Kreditwürdigkeit eines Bankkunden beschreiben lassen. Insbesondere werden verschiedene Herleitungsmöglichkeiten des logistischen Regressionsmodells (latente Variablen, Bayes-Klassifikation durch quadratische Diskriminanzanalyse) sowie die Kodierung und Interpretation von verschiedenen Kovariablentypen behandelt. Kapitel 3 („Generalized Linear Models“) führt anschließend den allgemeinen Rahmen generalisierter linearer Modelle ein, die das Logit-Modell, das gewöhnliche lineare Modell und eine ganze Reihe weitere im Buch behandelte Modelle als Spezialfälle beinhalten. Damit lassen sich in einem vereinheitlichten Ansatz die Maximum Likelihood-Schätzung und weitere Methoden der statistischen Inferenz entwickeln, die dann unmittelbar für viele Spezialfälle zur Verfügung stehen. Als Erweiterung werden auch Verfahren der Quasi-Likelihood-Schätzung vorgestellt. Kapitel 4 („Modeling of Binary Data“) wendet sich dann erneut der Modellierung binärer Zielgrößen zu und bettet diese in die in Kapitel 3 entwickelte Methodik ein. Speziell wird die Maximum Likelihood-Schätzung ausführlich dargestellt ebenso wie Verfahren zur Beurteilung der Modellanpassung und der Modelldiagnose. Ebenfalls betrachtet werden verschiedene Möglichkeiten, die Kovariablen in einen geeigneten linearen Prädiktor zu übersetzen (insbesondere im Fall kategorialer Kovariablen) sowie den Erklärungsgehalt verschiedener Kovariablen zu beurteilen. Während sich Kapitel 4 weiterhin auf das Logit-Modell konzentriert, werden in Kapitel 5 („Alternative Binary Regression Models“) andere Formen der Regression für binäre Zielgrößen betrachtet. Dies beinhaltet sowohl die Diskussion alternativer Linkfunktionen als auch die simultane Schätzung der Link-Funktion mit den Kovariableneffekten und die Berücksichtigung von Überdispersion.

Kapitel 6 („Regularization and Variable Selection for Parametric Models“) wendet sich dann der Problematik der Variablenselektion und der Regularisierung im Falle hochdimensionaler Vektoren von erklärenden Variablen zu. Beispiele hierzu ergeben sich insbesondere im Genetikbereich, da hier mit modernen Technologien (High-Throughput-Experimente) sehr viele Einflussgrößen wie etwa Genexpressions-Niveaus oder Informationen zur Allelhäufigkeit von Einzelnukleotid-Polymorphismen erhoben werden können. Dagegen ist in diesen Beispielen die Anzahl der Beobachtungen typischerweise weiterhin relativ gering, so dass häufig die Zahl der Einflussgrößen die Zahl der Beobachtungen deutlich übersteigt und klassische Schätzansätze damit nicht mehr verwendet werden können. Entsprechend werden Penalisierungsansätze zur Regularisierung eingeführt, die es erlauben, die Komplexität eines Modells zu kontrollieren und somit auch hochdimensionale Modelle der Schätzung mit relativ geringen Fallzahlen zugänglich zu machen. Dieser Zweig der Statistik hat in den letzten Jahren eine rege Entwicklung erlebt und wird hier sehr schön in Ihrem aktuellen Stand und in breitem Überblick wiedergegeben. Neben Penalisierungsansätzen werden auch indirekte Regularisierungsansätze wie Boosting behandelt. Besonderen Raum nimmt die Regularisierung kategorialer Einflussgrößen

ein, mit deren Hilfe beispielsweise automatisiert Kategorien einer Einflussgröße zusammengefasst werden können. Dabei muss insbesondere die unterschiedliche Skalierung der Einflussgrößen (ordinal versus nominal) berücksichtigt werden. Kapitel 7 („Regression Analysis of Count Data“) beschließt dann den Block parametrischer Modelle für univariate Einflussgrößen mit einer Behandlung von Regressionsmodellen für Zählgrößen. Neben der ausführlichen Betrachtung der klassischen, loglinearen Poisson-Regression werden auch Erweiterungen wie Regressionsmodelle, basierend auf der negativen Binomialverteilung, oder Modelle mit Zero-Inflation behandelt.

Die beiden anschließenden Kapitel 8 („Multinomial Response Models“) und 9 („Ordinal Response Models“) wenden sich dann Modellen für mehrkategoriale Zielgrößen zu und unterscheiden dabei nach der Skalierung der Zielgröße. Kapitel 8 behandelt insbesondere das multinomiale Logit-Modell, aber auch eine Reihe von Erweiterungen beispielsweise für Paarvergleiche sowie die regularisierte Schätzung multinomialer Logit-Modelle. Kapitel 9 dagegen konzentriert sich auf spezielle Modelle für ordinale Zielgrößen wie das kumulative und das sequentielle Modell.

Während die Kapitel zu parametrischen Regressionsmodellen die Annahme eines linearen Prädiktors aufrecht erhalten, befassen sich die Kapitel 10 („Semi- and Non-Parametric Generalized Regression“) und 11 („Tree-Based Methods“) mit Verfahren, die eine allein datengesteuerte Bestimmung der funktionalen Form von Kovariableneinflüssen zulassen. Kapitel 10 konzentriert sich dabei auf Glättungsverfahren basierend auf Basisfunktionen und entwickelt hierzu eine entsprechende penalisierte Schätzmethodik. Dabei werden auch moderne Verfahren wie Boosting zur automatischen Wahl der Glattheit der zu schätzenden Funktionen herangezogen und Methoden für funktionale Daten behandelt. Kapitel 11 verwendet dagegen Regressions- und Klassifikationsbäume, die insbesondere dann Vorteile aufweisen, wenn komplexe Interaktionsformen in den Daten vorliegen.

Die verbleibenden Kapitel 12 bis 15 behandeln jeweils spezielle Fragestellungen der kategorialen Regression, die sich nicht unmittelbar einem der drei bisher skizzierten Themenblöcken zuordnen lassen. Kapitel 12 („The Analysis of Contingency Tables: Log-linear and Graphical Models“) beschäftigt sich mit der Analyse von Kontingenztabellen insbesondere mit Hilfe log-linearer Modelle. Hierzu werden in aufsteigender Form ausgehend vom einfachsten Fall einer Kontingenztafel für zwei Merkmale komplexere Modelle für mehr Merkmale entwickelt. Kapitel 13 („Multivariate Response Models“) wendet sich dann Modellen mit multivariater Zielgröße zu, wobei insbesondere Markov-Übergangsmodelle und marginale Modelle, basierend auf generalisierten Schätzgleichungen, behandelt werden. Kapitel 14 („Random Effects Models and Finite Mixtures“) entwickelt dann einen alternativen Ansatz für multivariate Zielgrößen und gruppierte Daten, basierend auf gemischten Modellen mit zufälligen Effekten. Neben parametrischen Modellen mit normalverteilten zufälligen Effekten und den entsprechenden Schätzverfahren werden auch semiparametrische Erweiterungen und Modelle mit Mischverteilunsspezifikationen für die zufälligen Effekte betrachtet.

Das finale Kapitel 15 („Prediction and Classification“) behandelt dann ausführlich die Probleme der Vorhersage der Zielgröße für neue Beobachtungen. Neben den grundlegenden Konzepten der Vorhersage und der Optimalität der Bayes-

Klassifikation werden eine ganze Reihe auch moderner Klassifikationsverfahren ausführlich vorgestellt und diskutiert. In einer Reihe von Anhängen werden benötigte Verteilungen, einige grundlegende mathematisch-statistische Konzepte, die Schätzung unter Nebenbedingungen, Informationskriterien, sowie Verfahren der numerischen Integration zusammengestellt. Jedes Kapitel beinhaltet Aufgaben zur vertiefenden Beschäftigung mit dem Inhalt.

Vergleicht man das Buch mit konkurrierenden Publikationen mit ähnlichem Fokus, so fällt insbesondere der Einbezug moderner Erweiterungen aus dem Bereich der Regularisierung auf. Insgesamt stellt das Buch eine Reihe sehr aktueller Entwicklungen vor und macht diese im Kontext der kategorialen Regression sehr gut zugänglich. Darüber hinaus zeichnet sich das Buch durch eine große Breite aus, so dass neben den Kernthemen der kategorialen Regression auch eine ganze Reihe angrenzender Themengebiete behandelt werden und somit auch dem mit den Grundprinzipien der kategorialen Regression vertrauten Leser zahlreiche neue Anregungen zur Verfügung gestellt werden. Das Buch erfüllt damit voll und ganz den vom Autor angestrebten Zweck, Statistikern und angewandten Forschern ebenso wie Studierenden das Gebiet der kategorialen Regression zu erschließen und ansprechend vorzustellen.