Methoden der Politikwissenschaft

LEHRBUCH

Joachim Behnke

Logistische Regressionsanalyse

Eine Einführung

PEER REVIEWED

Springer VS

Methoden der Politikwissenschaft

LEHRBUCH

Joachim Behnke

Logistische Regressionsanalyse

Eine Einführung

PEER REVIEWED

Springer VS

Methoden der Politikwissenschaft

Weitere Bände in dieser Reihe
http://www.springer.com/series/12223

In der Schriftenreihe werden kompakte Einführungstexte in grundlegende und avancierte Verfahren der Datenerhebung und Datenauswertung veröffentlicht. Der Schwerpunkt liegt dabei auf standardisierten Techniken, die für politikwissenschaftliche Fragestellungen relevant sind. Die Reihe wendet sich in erster Linie an Studierende und ist auf den Einsatz in der universitären Methodenlehre zugeschnitten. Sie wendet sich darüber hinaus aber auch an Forscherinnen und Forscher, die sich schnell über bestimmte Verfahren informieren wollen, um deren möglichen Nutzen für die eigene Forschung abzuschätzen, oder um die Arbeiten anderer Autorinnen und Autoren besser verstehen und beurteilen zu können. Dem Adressatenkreis entsprechend vermitteln die einzelnen Bände der Reihe ein grundlegendes Verständnis des jeweils dargestellten Verfahrens. Kennzeichnend für die Reihe ist das Prinzip größtmöglicher Anschaulichkeit: Die Verfahren werden jeweils unter Bezugnahme auf ein konkretes Anwendungsbeispiel aus der politikwissenschaftlichen Forschung eingeführt und dargestellt. Besonderes Gewicht wird dabei den Anwendungsvoraussetzungen sowie den in der Praxis auftretenden Schwierigkeiten gewidmet. In den Bänden werden keine Detailprobleme des jeweiligen Verfahrens diskutiert, sondern dafür auf weiterführende Spezialliteratur verwiesen. Die Bände beinhalten ein kommentiertes Literaturverzeichnis, in dem die wichtigsten Lehrbücher und Einführungstexte zum jeweiligen Verfahren kurz vorgestellt werden. Setzt die Anwendung eines Verfahrens die Verwendung von spezieller Erhebungs- bzw. Analysesoftware voraus, wird kurz in diese eingeführt. Ist ein Analyseverfahren im Rahmen der gängigen Statistikpakete verfügbar, so werden die notwendigen Befehle erläutert. Um die Bände möglichst kompakt zu halten, wird die Beschreibung der Software auf einer speziellen Homepage zur Schriftenreihe veröffentlicht.

Joachim Behnke

Logistische Regressionsanalyse

Eine Einführung

Springer VS

Joachim Behnke
Friedrichshafen
Deutschland

ISBN 978-3-658-05081-8 ISBN 978-3-658-05082-5 (eBook)
DOI 10.1007/978-3-658-05082-5

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

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

Vorwort

Das vorliegende Buch wendet sich an all diejenigen, die entweder in der eigenen Forschung oder in der Literatur auf Auswertungsprobleme gestoßen sind, für die die logistische Regression einen geeigneten Ansatz darstellt. Es handelt sich dabei um die Struktur eines statistischen Modells mit einer abhängigen dichotomen Variablen und unabhängigen Variablen, die intervallskaliert sind. Die Anwendungsbereiche der logistischen Regression sind mannigfaltig und finden sich praktisch in jedem Bereich, der empirisch erforscht werden kann, von den Politik- und anderen Sozialwissenschaften bis zur Biologie und Medizin.

Das Buch wendet sich an Leser mit statistischen Grundkenntnissen auf dem Level der multivariaten linearen Regression, sollte aber ansonsten für jeden Leser mit einem rudimentären Grundverständnis für Mathematik verständlich sein. Lesern, die lediglich ein grundlegendes Verständnis des Verfahrens gewinnen wollen, werden vor allem das zweite und das dritte Kapitel ans Herz gelegt. Das zweite Kapitel versucht, die Logik der statistischen Regressionsanalyse im Vergleich zum Modell der linearen Wahrscheinlichkeit zu erläutern. Für Leser, die einen schnelleren Einstieg suchen, kann dieses Kapitel auch übergangen werden. Im dritten Kapitel wird die Form der logistischen Funktion dargestellt und durch eine schrittweise Entwicklung dieser Form versucht, ein intuitives Verständnis für die Funktion selbst zu vermitteln. Das fünfte Kapitel ist insbesondere für Forscher, die selbst eine logistische Regression durchführen wollen, von zentraler Bedeutung, da hier die verschiedenen Möglichkeiten aufgezeigt werden, wie man die Ergebnisse einer logistischen Regression sinnvoll interpretieren und darstellen kann. Im sechsten Kapitel werden verschiedene Goodness-of-Fit-Maße erläutert. Die Kenntnis von dem, was in Kap. 3 vermittelt wird, ist als Grundlage notwendig für das Verständnis des Stoffs in Kap. 5 und 6, ansonsten können alle Kapitel auch für sich gelesen werden.

Die Koeffizienten der logistischen Regression werden nicht wie die der linearen Regression mit der Methode der kleinsten Quadrate geschätzt, sondern mit der sogenannten Maximum-Likelihood-Methode. Dieses Schätzverfahren ist sehr grund-

legend und wird inzwischen bei vielen Standardverfahren angewandt. Während die Methode der kleinsten Quadrate analytischer Natur ist, d. h. die Parameterwerte werden als Lösungen bestimmter Gleichungen gefunden, sind Maximum-Likelihood-Verfahren computativ, d. h. sie werden auf sehr aufwändige Weise errechnet. Die immer größere Verbreitung von Maximum-Likelihood-Verfahren in der Statistik ist daher vor allem auf die immer größere Rechenkapazität auch einfacher PCs zurückzuführen. Für diejenigen, die die Logik eines Maximum-Likelihood-Schätzverfahrens genauer kennenlernen wollen, ist das vierte Kapitel gedacht. In ihm werden auch bestimmte Algorithmen erläutert, die bei der logistischen Regression häufig angewandt werden. Für Leser, die lediglich an der Anwendung einer logistischen Regression interessiert sind, oder an den Grundkenntnissen, die notwendig sind, um die Ergebnisse einer logistischen Regression richtig „lesen" zu können, ist das vierte Kapitel entbehrlich und kann übersprungen werden. Es kann aber auch gelesen werden, nachdem man sich mit den Kap. 2, 3 und 5 ein grundlegendes Verständnis der logistischen Regression verschafft hat.

Florian Bader, Martin Valdés-Stauber und der externe Gutachter Henning Best haben das Skript mit großer Sorgfalt gelesen und mir mit vielen Anmerkungen und Verbesserungsvorschlägen geholfen, den Text zu verbessern. Ihnen gilt hierfür mein herzlicher Dank. Verbliebene Fehler und stilistische Ungeschicklichkeiten gehen auf mein persönliches Konto.

Auf der Webseite www.springer.com/springer+vs/politik/book/978-3-658-05081-8 findet sich die Syntax für alle Beispielrechnungen im Text in den Statistikpaketen R, SPSS und Stata.

Friedrichshafen, im April 2014 Joachim Behnke

Inhaltsverzeichnis

Einführung

1

Das wohl wichtigste und am meisten verwendete statistische Analysemodell in den Sozialwissenschaften ist das lineare Regressionsmodell. Seine Vorteile liegen klar auf der Hand: Es ist einfach zu berechnen und – womöglich noch wichtiger – die Ergebnisse sind einfach zu interpretieren. Es scheint daher naheliegend, dieses Modell überall dort anzuwenden, wo einem sinnvollen Einsatz nichts entgegensteht. Im linearen Regressionsmodell wird die Variation einer sogenannten abhängigen oder bewirkten Variablen durch die Variation sogenannter unabhängiger oder bewirkenden Variablen erklärt. Veränderungen der Ausprägung der abhängigen Variablen können also durch Veränderungen der Ausprägungen der unabhängigen Variablen erklärt werden. Im einfachsten Fall ist dieser funktionale Zusammenhang linearer Art, d. h. dasselbe Ausmaß einer Veränderung bei der unabhängigen Variablen ruft immer dasselbe Ausmaß der Veränderung bei der abhängigen Variablen hervor, eine doppelt so große Veränderung der unabhängigen Variablen eine doppelt so große Veränderung der abhängigen usw. Variablen, bei denen man die Differenzen von gemessenen Werten zueinander ins Verhältnis setzen kann, werden als intervallskaliert bezeichnet. Wenn die unabhängige Variable kontinuierlich über eine große Bandbreite von Werten verteilt ist, d. h. viele verschiedene Ausprägungen annehmen kann, muss dementsprechend auch die abhängige Variable über eine große Anzahl von Werten streuen, wenn sie durch die unabhängige Variable erklärt werden soll. Die typische lineare Regressionsanalyse besteht also aus kontinuierlich verteilten, intervallskalierten unabhängigen und abhängigen Variablen, wie wenn man z. B. das Gewicht eines Menschen durch seine Körpergröße erklären bzw. schätzen möchte.

In vielen Fällen, die in den Sozialwissenschaften untersucht werden, besteht das zu untersuchende Phänomen jedoch in einer Charakteristik oder einer Eigenschaft, die entweder vorhanden oder nicht vorhanden ist, die abhängige, zu erklärende Variable ist also dichotomer Natur und kann nur zweierlei Ausprägungen annehmen. Typische dichotome Variablen in den Sozialwissenschaften sind z. B. Arbeits-

J. Behnke, *Logistische Regressionsanalyse,* Methoden der Politikwissenschaft,
DOI 10.1007/978-3-658-05082-5_1, © Springer Fachmedien Wiesbaden 2015

losigkeit, die Mitgliedschaft in einer bestimmten Vereinigung, z. B. in der OECD, die Verfolgung einer bestimmten Politik, die Teilnahme an einer Wahl, Ehestatus, die Kinderlosigkeit eines Ehepaars, das Vorliegen einer bestimmten Organisationsstruktur in einem Betrieb, z. B. das Vorhandensein eines Betriebsrats etc. Um das Vorhandensein bzw. Nichtvorhandensein dieser Eigenschaft zu messen, sind im Prinzip alle symbolischen Darstellungen dieser beiden Ausprägungen denkbar, die diese unterscheidbar machen, üblicherweise werden sie aber mit 0 und 1 vercodet. Der Wert 1 bedeutet, dass die betreffende Eigenschaft vorhanden ist, der Wert 0 zeigt hingegen ihr Fehlen an. Der Vorteil der Zuweisung dieser numerischen Werte besteht darin, dass der Mittelwert der Variablen dann auch der relativen Häufigkeit des Auftretens des Werts 1 entspricht. Die relative Häufigkeit wiederum kann auch als Wahrscheinlichkeit interpretiert werden. Nehmen z. B. 80 % aller Wahlberechtigten an einer Wahl teil, dann kann daraus geschlossen werden, dass ein beliebig ausgewählter wahlberechtigter Bürger mit einer Wahrscheinlichkeit von 0,8 an der Wahl teilnimmt, bzw. – genauer – dass es sich bei ihm mit einer Wahrscheinlichkeit von 0,8 um einen Bürger handelt, der zur Wahl geht. Während jeder einzelne Fall immer den Wert 0 oder 1 bezüglich der Ausprägung der abhängigen Variablen besitzt, weist eine Gruppe von Fällen eine Verteilung von 0- und 1-Werten auf, die durch den Mittelwert prägnant beschrieben werden kann. Der Mittelwert einer solchen Gruppe, die hinsichtlich aller relevanten unabhängigen Variablen dieselben Werte aufweisen, kann dann als die Wahrscheinlichkeit, gewissermaßen im Sinne einer Disposition, interpretiert werden, mit dem ein Mitglied dieser Gruppe, das all die entsprechenden Eigenschaften aufweist, den Wert 1 der abhängigen Variablen aufweist. Wenn von allen Männern zwischen 20 und 30 mit einem Hochschulabschluss z. B. 30 % verheiratet wären, dann könnte man aus dem Vorliegen dieser Informationen bei einem konkreten Fall schließen, dass er mit einer Wahrscheinlichkeit von 0,3 verheiratet ist.

Logistische Regressionsmodelle sind statistische Analyseverfahren, die für diese Art von Untersuchungen angewandt werden können bzw. sich für eine Analyse von Daten und Zusammenhängen der beschriebenen Art besonders gut eignen. Neben den klassischen linearen Regressionsmodellen zählen logistische Regressionsanalysen, oft auch als Logit-Modelle bezeichnet, inzwischen zu den Standardverfahren in den Sozialwissenschaften.

Der Aufbau des Buches gliedert sich auf folgende Weise. Im zweiten Kapitel wird die einer logistischen Regressionsanalyse zugrundeliegende Logik unter Rückgriff auf das klassische Modell der linearen Regression und dem Modell der linearen Wahrscheinlichkeit erläutert. Der Vorteil der logistischen Regressionsanalyse wird hier vor allem durch die Betonung der Mängel linearer Modelle zur Analyse von Zusammenhängen mit dichotomen abhängigen Variablen herausgearbeitet. Im

dritten Kapitel gehe ich dann auf die Ableitung bzw. theoretische Begründung der konkreten Logitfunktion ein, die bei logistischen Regressionen angewandt wird. Die Koeffizienten in logistischen Regressionen werden mit Hilfe des Maximum-Likelihoods-Verfahrens gewonnen. Das vierte Kapitel geht auf die spezifische Logik bzw. Vorgehensweise dieser Schätzmethode ein. Dabei geht es hier vor allem darum, ein grundlegendes Verständnis dieses Verfahrens zu wecken. Leser, die in erster Linie an der Anwendung bzw. dem Wissen über die angemessene Interpretation von Ergebnissen einer logistischen Regression interessiert sind, können dieses Kapitel auch auslassen bzw. erst einmal überspringen. Im fünften Kapitel werden die verschiedenen Darstellungsformen der Ergebnisse einer logistischen Regression diskutiert und verschiedene Möglichkeiten der Interpretation aufgezeigt. Das sechste Kapitel schließlich beendet das Buch mit der Thematik von Signifikanztests bzw. der Problematik der Modellauswahl.

Nicht weiter ein gehe ich in dieser Einführung auf spezifische Probleme, die man vor allem der Diagnostik zurechnet, also Gesichtspunkten der Analyse wie die Behandlung besonders einflussreicher Fälle bzw. von Ausreißern, die Formulierung des angemessenen funktionalen Zusammenhangs und Kollinearitätsprobleme. Ebenfalls behandle ich Interaktionseffekte nicht vertiefend bzw. als eigenen thematischen Abschnitt. Die Gründe für die „Vernachlässigung" dieser Aspekte sind einerseits, dass eine ausführliche Behandlung dieser Themen den Rahmen einer Einführung sprengen würde, vor allem aber, dass der Umgang mit diesen Problematiken in der Regel keine spezifische Vorgehensweise bei der logistischen Regression nahelegt, sondern man hier im Wesentlichen analog wie bei linearen Modellen vorgehen kann (vgl. Field et al. 2012, S. 338 ff.; Fox 2008, S. 412 ff.; Jaccard 2001)[1]. Die inhaltliche Beschränkung ist Konsequenz der von mir verfolgten Gestaltungsabsicht, sich bei dieser Einführung vor allem auf die Darstellung von Konzepten und Problemen zu konzentrieren, die sich bei der logistischen Regression auf charakteristische Weise von denen einer linearen Regression unterscheiden.

[1] Allerdings erfordert die Interpretation von Interaktionseffekten bei logistischen Regressionen einiges mehr an Fingerspitzengefühl und Hintergrundwissen als bei linearen Regressionen.

Lineare Regression und das Modell der linearen Wahrscheinlichkeit

2

Um die besonderen Eigenschaften einer logistischen Regression zu beschreiben, ist es sinnvoll, diese mit dem Alternativmodell der linearen Regressionsanalyse zu vergleichen, denn gerade im Abgleich mit der linearen Regressionsanalyse können die besonderen Stärken des Logit-Modells prägnant herausgearbeitet werden. Die logistische Regressionsanalyse kann daher am besten als statistische, methodische Antwort auf Schwächen und Probleme der linearen Regressionsanalyse begriffen werden, die sich ergeben, wenn die abhängige Variable dichotomer Natur ist.

Das typische (bivariate) Regressionsmodell besteht aus einer unabhängigen und einer abhängigen Variablen, die beide intervallskaliert sind (und idealerweise kontinuierlich verteilt). Im Regressionsmodell wird unterstellt, dass der (kausale) Zusammenhang zwischen unabhängiger und abhängiger Variablen in Form einer linearen Funktion dargestellt werden kann[1]:

$$Y = \beta_0 + \beta_1 X \qquad \text{GL (2.1)}$$

Diese lineare Funktion entspricht dem kausalen Prozess, in dem eine Veränderung von X eine Veränderung von Y bewirkt. Der konkrete Wert einer einzelnen Ausprägung von Y allerdings wird in der Regel durch weitere, zusätzliche Einflussfaktoren bestimmt. Des Weiteren bleibt noch ein Messfehler zu berücksichtigen,

[1] Ich begnüge mich hier und im Folgenden mit der Darstellung des bivariaten Zusammenhangs zwischen der abhängigen Variablen und einer einzigen unabhängigen Variablen. Grundsätzlich sind alle Erörterungen, die im Folgenden bezüglich des bivariaten Falls gemacht werden, auf den multivariaten Fall verallgemeinerbar. Ich bevorzuge die Darstellung des bivariaten Falls aus Gründen der Einfachheit und der Didaktik, insbesondere, weil so auch graphische Darstellungen möglich sind. Ich verschiebe die explizite Diskussion der multivariaten Analyse auf das 5. Kapitel, da hier die zu erörternden Konzepte nur sinnvoll diskutiert werden können, wenn man den multivariaten Charakter eines Modells explizit berücksichtigt.

J. Behnke, *Logistische Regressionsanalyse,* Methoden der Politikwissenschaft,
DOI 10.1007/978-3-658-05082-5_2, © Springer Fachmedien Wiesbaden 2015