Programme zur Logitanalyse von kategorialen abhängigen Variablen auf Individualdatenebene

Kühnel, Steffen M.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Historical Social Research, Vol. 20 — 1995 — No. 3, 63–87

Programme zur Logitanalyse von kategorialen abhängigen Variablen auf Individualdatenebene

Steffen Kühnel*

Abstract: **Relations between an dependent categorical variable and independent variables can be analyzed with logit models. The first part of the paper gives an short overview on different logit models including models for binary panel data, ordinal variables and decision trees. The availability of these models in BMDP, LIMDEP, SAS, SPSS, SYSTAT and the free ware statistical system TDA is discussed in the second part Though only few procedures are designed especially to estimate the parameters of logistic models other procedures can be used as well. For example, the conditional logit model or logistic discrete choice model may be estimated by procedures for event history analysis. Exemplaric program setups are given for the procedures 2L in BMD and PHREG in SAS.**

In der klassischen linearen Regression werden die Werte einer abhängigen Variablen Y als lineare Funktionen der Werte von unabhängigen Regressorvariablen X_1, X_2, ... X_K und einer Residualgröße E aufgefaßt[1]

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots b_KX_K + E \qquad (1)$$

Üblicherweise wird angenommen, daß der Mittelwert der Residualgröße Null ist und daß die Residualgröße nicht mit den unabhängigen Variablen korreliert ist. Dann läßt sich Gleichung (1) auch so umschreiben, daß die *bedingten Mittelwerte* $\hat{y}$ der abhängigen Variablen nur durch die Regressionskoeffizienten b_0, b_1, ..., b_K und die unabhängigen Variablen bestimmt werden:

* Address all Communications to Steffen Kühnel, IFAS, Universität zu Köln, Greinstr. 2, D-50939 Köln, email: Kuehnel@WiSo.Uni-Koeln.DE.

[1] Um die Darstellung übersichtlich zu halten, verachte ich auf zusätzliche Indizes, die für die Untersuchungseinheiten in einer Stichprobe stehen. Ich unterscheide in diesem Beitrag auch nicht explizit zwischen Modellparametem und deren Schätzungen. Aus dem jeweiligen Kontext ist ohne Schwierigkeiten ersichtlich, ob es sich bei Regressionskoeffizienten um Parameter (Grundgesamtheitswerte) oder um deren Schätzungen (Stichprobenstatistiken) handelt.

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_K X_K = b_0 + \sum_{k=1}^{K} b_k X_k \qquad (2)$$

Ist eine abhängige Variable nominalskaliert, dann ist die Zuordnung von Zahlen zu den Ausprägungen der Variablen willkürlich. In dieser Situation macht die Berechnung von Mittelwerten offensichtlich wenig Sinn. Bei nominalskalierten abhängigen Variablen sind lineare Regressionsmodelle daher nicht angebracht. Als Ausweg bietet sich an, anstelle der Mittelwerte die Realisierungswahrscheinlichkeiten der Kategorien der abhängigen Variablen als Funktion der Regressorvariablen zu spezifizieren. Dabei müssen zwei Eigenschaften von Wahrscheinlichkeiten berücksichtigt werden. Zum einen gibt es weder negative Wahrscheinlichkeiten noch Wahrscheinlichkeiten größer Eins. Zum anderen muß die Summe der Wahrscheinlichkeiten über alle Kategorien der abhängigen Variablen stets Eins ergeben. Diese beiden Eigenschaften sind nicht von vornherein garantiert, wenn man bei einer abhängigen Variable mit i=l,2,...,I Kategorien die Wahrscheinlichkeit p_i der Kategorie i ähnlich wie im linearen Regressionsmodell als (z.B. lineare) Funktionen von Regressorvariablen X_{ik} auffaßt:[2]

$$p_i = b_{i0} + \sum_{k_i=1}^{K_i} b_{ik} X_{ik} \qquad i = 1,2,\ldots,I \qquad (3)$$

Dieses Problem läßt sich aber durch zwei einfache »Tricks« lösen. Der erste Trick besteht darin, anstelle der linearen Funktionen die Exponentialfunktion zu verwenden:

$$p_i = \exp\left(b_{i0} + \sum_{k_i=1}^{K_i} b_{ik} X_{ik} \right) \qquad i = 1,2,\ldots,I \qquad (4)$$

Durch die Anwendung der Exponentialfunktion exp(...) werden negative Werte in Zahlen zwischen Null und Eins transformiert. Es ist jedoch noch nicht ausgeschlossen, daß Werte größer Eins auftreten. Hier hilft der zweite Trick weiter,

---

[2] **Da im Prinzip für jede Kategorie der abhängigen Variablen eine eigene Wahrscheinlichkeitsgleichung aufgestellt werden kann, werden in Gleichung (3) sowohl die Regressionskoeffizienten b_{ik} als auch die unabhängigen Regressorvariablen x_{ik} durch einen weiteren Index i gekennzeichnet, der angibt, daß sich die Gleichung auf die Kategorie i der abhängigen Variablen bezieht.**

nämlich die Regressionsgleichung für die Kategorie i durch die Summe der Regressionsgleichungen für alle i=1,2,...,I Kategorien zu dividieren:

$$p_i = \frac{\exp\left(b_{i0} + \sum_{k_i=1}^{K_i} b_{ik} X_{ik}\right)}{\sum_{i'=1}^{I} \exp\left(b_{i'0} + \sum_{k_{i'}}^{K_{i'}} b_{i'k} X_{i'k}\right)} \qquad i = 1,2,...,I \qquad (5)$$

Beide Tricks zusammen stellen sicher, daß die Werte tatsächlich als Wahrscheinlichkeiten interpretiert werden können. Gleichung (5) ist die allgemeine Ausgangsgleichung für logistische Regressionsmodelle auf Individualdatenebene. Die Gleichung gibt an, wie sich die Wahrscheinlichkeit p_i der Kategorie i der abhängigen Variablen aus den Werten der unabhängigen Regressorvariablen berechnet.

Die Bezeichnung Logitmodell ergibt sich durch eine Umformung der Gleichung (5), so daß auf der linken Seite die logarithmierten Wahrscheinlichkeitsverhältnisse zweier Kategorien der abhängigen Variablen stehen:

$$\ln\left(\frac{p_i}{p_j}\right) = \left(b_{i0} + \sum_{k_i=1}^{K_i} b_{ik} X_{ik}\right) - \left(b_{j0} + \sum_{k_j=1}^{K_j} b_{jk} X_{jk}\right) \qquad i \neq j \qquad (6)$$

Solche logarithmierten Wahrscheinlichkeitsverhältnisse werden dem englischen Sprachgebrauch folgend als »Logits« bezeichnet Die Anwendung solcher Modelle in der Datenanalyse heißt entsprechend Logitanalyse.

Varianten von Logitmodellen

Die Ausgangsgleichungen (5) bzw. (6) der Logitmodelle sind sehr allgemein. Spezielle Logitmodelle unterscheiden sich nun dadurch, daß zusätzliche Annahmen über die unabhängigen Variablen oder die Regressionskoeffizienten gemacht werden. Tabelle 1 gibt eine Übersicht über die im folgenden behandelten Modelle. Da die verschiedenen Varianten der Logitmodelle in unterschiedlichen Kontexten entwickelt wurden, gibt es nicht immer eindeutige Bezeichnungen. Ich möchte daher vor der Diskussion von Software zur Logitanalyse zunächst die einzelnen Logitmodelle zumindest kurz skizzieren. Bei den jeweiligen Modellgleichungen kann es Abweichungen zu den in Statistikprogrammen implementierten Formeln geben. So werden in Programmen für

Tabelle 1: Varianten von Logitmodellen

Modellvariante		Kennzeichen	Gleichung
(a)	Binäres Logitmodell	abhängige Variable ist dichotom; daher nur eine Modellgleichung	(7)
(b)	Multinomiales Logitmodell	gleiche Regressorvariablen (X-Variablen) für alle Kategorien der abhängigen Variablen; dabei mehrere Regressionskoeffizienten für jede unabhängige Variable	(9)
(c)	Konditionales Logitmodell	unterschiedliche Regressorvariablen (X-Variablen) für jede Kategorie, aber gemeinsame Regressionskoeffizienten	(10)
(d)	Mischmodell	Kombination aus (b) und (c)	(11)
(e)	Binäres Panelmodell	binäres Logitmodell, bei dem für jeden Fall mehrere Beobachtungen über die Zeit vorliegen	(13)
(f)	Mehrstufiges Logitmodell (Präferenzbaum)	Berücksichtigung von Ähnlichkeiten zwischen Ausprägungen	(14)
(g)	Logitmodell der benachbarten Kategorien	Logitmodell für ordinalskalierte abhängige Variablen; Spezialfall von (c)	(15)
(h)	Logitmodell der sequentiellen Anordnung	Logitmodell für ordinalskalierte abhängige Variablen; Spezialfall von (a)	(16)
(i)	Logitmodell der kumulativen Logits	Logitmodell für ordinalskalierte abhängige Variablen; kann als Schwellenwertmodell aufgefaßt werden	(17)

multinomiale Logitmodelle etwa unterschiedliche Referenzkategorien verwendet. Bei ordinalen Logitmodellen werden gelegentlich Zähler und Nenner der Logitgleichungen ausgetauscht oder die Vorzeichen der Regressionskoeffizienten umgedreht. Solche programmspezifischen Umsetzungen lassen es als nicht sehr sinnvoll erscheinen, ein Programm anzuwenden, ohne sich zuvor ausführlich mit der Logik und genauen Spezifikation der jeweiligen Modelle beschäftigt zu haben.[3]

Ich habe bereits erwähnt, daß die allgemeine Ausgangsgleichung (5) sehr allgemein ist. Da sich die Ausprägungswahrscheinlichkeiten zu Eins summieren, ist es nicht notwendig, für alle Ausprägungen jeweils eine eigene Regressionsgleichungen zu formulieren. Stattdessen kann eine beliebige Kategorie der abhängigen Variablen als *Referenzkategorie* aufgefaßt werden. Deren Wahrscheinlichkeit ergibt sich dann als 1.0 minus der Summe der Wahrscheinlichkeiten der anderen Ausprägungen. Es wird somit eine Regressionsgleichung eingespart. Hat die abhängige Variable nur zwei Ausprägungen, benötigt man nur eine einzige Gleichung. Das allgemeine Logitmodell (5) reduziert sich dann zum *binären Logitmodell:*

$$p_1 = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_KX_K)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_KX_K)} = \frac{1}{1 + \exp(-b_0 - b_1X_1 - b_2X_2 - \ldots - b_KX_K)} \quad (7)$$

Die zweite Kategorie ist Referenzkategorie. Ihre Wahrscheinlichkeit berechnet sich nach:

$$p_2 = 1 - p_1 = \frac{1}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_KX_K)} \quad (8)$$

---

[3] **Leider gibt es noch nicht sehr viele deutschsprachige Monographien zur Logitanalyse. Die empfehlenswerte Arbeit von Maier und Weiss (1990) stellt aufgrund ihrer Orientierung an ökonomische Anwendungen relativ hohe Voraussetzungen hinsichtlich statistischer Vorkenntnisse. Die Arbeit von Urban (1993) ist zwar einfacher, in ihren Darstellung aber nicht immer unproblematisch. Eine Einführung von mir in einem Lehrbuch zur kategorialen Datenanalyse (Andress, u.a., in Vorbereitung) wird vermutlich erst im Laufe des Jahres erscheinen. Es gibt allerdings eine Reihe englischsprachiger Einführungen, die auch ohne große mathematische Vorkenntnisse lesbar sind. Zu nennen ist hier beispielweise Cramer (1991), Demaris (1992) oder Kleinbaum (1994).**

Bei einer abhängigen Variablen mit mehr als zwei Ausprägungen müssen Regressionsgleichungen für mehrere Kategorien formuliert werden. Werden für alle Modellgleichungen dieselben unabhängigen Regressorvariablen X_1, X_2, ..., X_K verwendet, erhält man die Modellgleichungen für das *multinomiale Logitmodell:*

$$p_i = \frac{\exp\left(b_{i0} + b_{i1}X_1 + b_{i2}X_2 + \ldots + b_{iK}X_K\right)}{1 + \sum_{i'=1}^{I-1} \exp\left(b_{i'0} + b_{i'1}X_1 + b_{i'2}X_2 + \ldots + b_{i'K}X_K\right)} \quad ; \quad i = 1, 2, \ldots, I-1$$

$$p_I = \frac{1}{1 + \sum_{i'=1}^{I-1} \exp\left(b_{i'0} + b_{i'1}X_1 + b_{i'2}X_2 + \ldots + b_{i'K}X_K\right)} \qquad (9)$$

In Gleichung (9) ist die letzte Kategorie (I) der abhängigen Variablen die Referenzkategorie. Der Gleichung kann entnommen werden, daß für jede unabhängige Variable jeder Ausprägung der abhängigen Variablen ein eigener Regressionskoeffizienten zugeordnet ist. Eine Ausnahme bildet die Referenzkategorie, für die ja keine Gleichung formuliert wurde und die daher auch keine Regressionskoeffizienten haben kann. Formal ist dies gleichbedeutend mit der willlkürlichen Festsetzung der Regressionskoeffizienten der Referenzkategorie auf die Werte Null (exp(0)=l). Statistisch gesehen ist es beliebig, welche Ausprägung der abhängigen Variablen Referenzkategorie ist. Für die Interpretation der Regressionskoeffizienten eines multinomialen Logitmodells muß man die Referenzkategorie allerdings kennen."

In der Ökonometrie werden Logitmodelle im Kontext von diskreten Entscheidungsmodellen diskutiert. Da sich Bezüge zur Theorie rationaler Entscheidungen herstellen lassen, werden die Modelle dort auch als Zufallsnutzenmodelle bezeichnet. Wir haben gesehen, daß im multinomialen Logitmodell (9) die unabhängigen Variablen jeweils über mehrere Regressionskoeffizienten die Ausprägungswahrscheinlichkeiten der abhängigen Variablen beeinflussen. Im *logistischen Zufallsnutzenmodell* oder *konditionalen Logitmodell* ist es gerade andersherum. Über ein gemeinsames Regressionsgewicht beeinflussen hier die (»alternativenspezifischen«) unabhängigen Variablen X_{ik} nur jeweils eine Kategorie i direkt

[4] Aus Platzgründen kann hier nicht näher auf die Interpretation eingegangen werden (vgl. dazu Kühnel, 1990 u. 1993). Wenn man im Nenner der Logits in Gleichung (6) für die Kategorie j jeweils die Referenzkategorie einsetzt und berücksichtigt, daß alle Regressionskoeffizienten der Referenzkategorie per definitionem Null sind, wird deutlich, daß im multinomialen Logitmodell (9) die Regressionskoeffizienten für die i-te Ausprägung der abhängigen Variablen den Einfluß der unabhängigen Variablen auf das Verhältnis der i-ten Kategorie zur Referenzkategorie mißt.

$$p_i = \frac{\exp\left(X_{i1}b_1 + X_{i2}b_2 + \ldots + X_{iK}b_K\right)}{\sum_{i'=1}^{I} \exp\left(X_{i'1}b_1 + X_{i'2}b_2 + \ldots + X_{i'K}b_K\right)} \quad ; \; i = 1,2,\ldots,I \qquad (10)$$

Im Unterschied zum multinomialen Logitmodell ist es im Zufallsnutzenmodell nicht notwendig (aber auch nicht ausgeschlossen), eine Referenzkategorie zu bilden. Dies liegt daran, daß im konditionalen Logitmodell (10) die in der Regel unbekannten und daher zu schätzenden Regressionskoeffizienten bei gegebenen Werten der unabhängigen Variablen eindeutig festgelegt sind. Auffallend ist weiter, daß keine Regressionskonstante definiert ist. Es ist allerdings möglich, bei insgesamt I Kategorien der abhängigen Variablen bis maximal I-1 Regressionskonstanten in das Modell aufzunehmen. Die Berücksichtigung von Regressionskonstanten ist die einfachste Form der Kombination des multinomialen Logitmodells (9) und des logistischen Zufallsnutzenmodells (10). Die allgemeine Gleichung für das kombinierte Mischmodell lautet

$$p_i = \frac{\exp\left(b_{i0} + \sum_{j=1}^{J} X_j b_{ij} + \sum_{k=1}^{K} X_{ik} b_k\right)}{\sum_{i'=1}^{I-1} \exp\left(b_{i'0} + \sum_{j=1}^{J} X_j b_{i'j} + \sum_{k=1}^{K} X_{i'k} b_k\right) + \exp\left(\sum_{k=1}^{K} X_{Ik} b_k\right)} \quad ; \; i = 1,2,\ldots,I-1$$

$$p_I = \frac{\exp\left(\sum_{k=1}^{K} X_{Ik} b_k\right)}{\sum_{i'=1}^{I-1} \exp\left(b_{i'0} + \sum_{j=1}^{J} X_j b_{i'j} + \sum_{k=1}^{K} X_{i'k} b_k\right) + \exp\left(\sum_{k=1}^{K} X_{Ik} b_k\right)} \qquad (11)$$

Bisweilen wird das Mischmodell (11) als eine eigene Modellvariante aufgefaßt. Da aber grundsätzlich jedes multinomiale Logitmodell auch als ein spezielles logistisches Zufallsnutzenmodell dargestellt werden kann (vgl. Kühnel, 1992), ist es nicht unbedingt notwendig, zusätzlich zur Implementation des logistischen Zufallsnutzenmodells (10) eine eigene Prozedur für die Schätzung der Modelle (9) und (11) bereitzustellen. Umgekehrt kann das logistische Zufallsnutzenmodell (10) aber nur dann als ein spezielles multinomiales Logitmodell geschätzt werden, wenn es der Schätzalgorithmus erlaubt, in Gleichung (9) beliebige Regressionskoeffizienten auf den Wert Null zu fixieren und andere Regressionskoeffizienten gleichzusetzen (Kühnel, 1993). Dies ist aber nur in wenigen Statistikprogrammen (z.B. in TDA) möglich.

Die Bezeichnung »konditionales Logitmodell« findet sich auch im biomedizinischen Kontext. Sie bezieht sich dort allerdings auf eine spezielle Anwendung eines zunächst binären Logitmodells in sogenannten »matched case-control«-Studien. Um nicht beobachtete Faktoren möglichst zu kontrollieren, werden in solchen Studien Probanden aufgrund gemeinsamer Eigenschaften wie Alter und Geschlecht in möglichst homogene Subpopulationen zusammengefaßt (engl: matched). In jeder Subpopulation wird ein Krankheitsfall (case) einem oder mehreren Kontrollfällen (control) gegenübergestellt. Die Unterschiede zwischen den Krankheits- und Kontrollfällen in einer Gruppe werden dann auf die unterschiedlichen Ausprägungen bei den Risikofaktoren zurückgeführt. Für die Analyse wird das Logitmodell aus Gleichung (10) verwendet.

Eine ähnliche Logik wird auch bei der Analyse von Paneldaten mit einer binären abhängigen Variablen angewendet. Die Subpopulationen der »matched case-control«-Studien entsprechen hier einem einzigen Fall, der im Zeitverlauf mehrfach beobachtet wurde. Die Wahrscheinlichkeit der Kategorie Eins zum Meßzeitpunkt t wird in dem Modell für die Untersuchungseinheit i durch ein binäres Logitmodell spezifiziert

$$p_{1ti} = \frac{\exp\left(b_i + \sum_{k=1}^{K} X_{tik} b_k\right)}{1 + \exp\left(b_i + \sum_{k=1}^{K} X_{tik} b_k\right)} \qquad (12)$$

Um die Abhängigkeit der Messungen im Zeitverlauf zu berücksichtigen und unbeobachtete zeitkonstante Störgrößen auszuschalten, ist im Modell anstelle einer gemeinsamen Regressionskonstanten für jeden Fall eine eigene Konstante b_i vorgesehen. Dabei entsteht jedoch das Problem, daß zu viele Regressionskonstanten vorkommen und diese daher nicht geschätzt werden können. Man kann nun jedoch anstelle einer Realisierung der abhängigen Variablen zum Zeitpunkt t das sich bei insgesamt T Meßzeitpunkten einstellende Antwortmuster der abhängigen Variablen betrachten. Die Wahrscheinlichkeit eines spezifischen Antwortmusters i sei p_i. Wenn man nun die bedingte Wahrscheinlichkeit $p_{i|c}$ betrachtet, daß sich das Antwortmuster i einstellt, wenn insgesamt bei den T Meßzeitpunkten c mal die Ausprägung Eins vorkommt, kürzen sich die individuenspezifischen Regressionskonstanten heraus:

$$p_{i|c} = \frac{\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} Y_{ti} X_{tik} b_k\right)}{\sum_{i' \in c} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} Y_{ti'} X_{ti'k} b_k\right)} \qquad (13)$$

Die in der Gleichung auftretenden Ausdrücke Y_{it} weisen den Wert Eins auf, wenn im Antwortmuster i zum Zeitpunkt T die Kategorie 1 bei der abhängigen Variablen realisiert wird. Wird im Antwortmuster i zum Zeitpunkt t dagegen die Kategorie Null realisiert, ist Y_{it} ebenfalls Null.

In allen bisher betrachteten Logitmodellen ist die Anordnung oder Reihenfolge der Kategorien der abhängigen Variablen beliebig. In *mehrstufigen (geschachtelten) Logitmodellen* oder *Präferenzbäumen* wird dagegen die Gesamtmenge der Kategorien der abhängigen Variablen in Gruppen zusammengefaßt. Eine Begründung hierfür kann beispielweise darin bestehen, daß die in einer Gruppe zusammmengefaßten Kategorien nicht gemessene Gemeinsamkeiten aufweisen, die sie von den Kategorien einer anderen Gruppe unterscheiden. In dem Modell werden die bedingten Wahrscheinlichkeiten p_{ig} einer Ausprägung i gegeben die Gruppe g als ein Logitmodell mit erklärenden Variablen X_{igk} modelliert. Auf der Ebene der Gruppen werden die Gruppenwahrscheinlichkeiten p_g ebenfalls als ein Logitmodell aufgefaßt. Die unabhängigen Variablen Z_{gt} variieren nur zwischen den Gruppen, weisen also für alle Ausprägungen in einer Gruppe den gleichen Wert auf. Von besonderer Bedeutung ist dabei eine spezielle Regressorvariable S_g, die die Ergebnisse der Modellschätzung auf der unteren Ebene auf die höhere Ebene transformiert. Die Werte dieser Variablen werden auch als Inklusivwerte, der zugehörige Regressionskoeffizient als Inklusivwertparameter bezeichnet. Wenn p_{ig} die (unbedingte) Wahrscheinlichkeit der Realisierung der Kategorie i in der Gruppe g bezeichnet, ergibt sich hierfür

$$p_{ig} = p_{i|g} \; p_g$$

$$= \frac{\exp\left(\sum_{k=1}^{K} X_{igk} b_k\right)}{\sum_{i'=1}^{I_g} \exp\left(\sum_{k=1}^{K} X_{i'gk} b_k\right)} \cdot \frac{\exp\left(S_g b_g + \sum_{t=1}^{T} Z_{gt} b_t\right)}{\sum_{g'=1}^{G} \exp\left(S_{g'} b_{g'} + \sum_{t=1}^{T} Z_{g't} b_t\right)} \qquad (14)$$

$$\text{mit: } S_g = \ln\left(\sum_{i=1}^{I_g} \exp\left(\sum_{k=1}^{K} X_{igk} b_k\right)\right)$$

Die Modellogik kann ohne Schwierigkeiten auf mehr als zwei Ebenen verallgemeinert werden. Die Regressionskoeffizienten lassen sich in jeder Ebene als einfache Logitmodelle schätzen. Da aber die Inklusivwerte S_g berechnete Regressorvariablen sind, in die die Schätzergebnisse der jeweils tieferen Ebene

einfließen, sind die Formeln für die Standardschätzfehler der Regressionskoeffizienten und damit alle inferenzstatistischen Aussagen auf den höheren Ebenen ungültig. Spezielle Prozeduren für geschachtelte Logitmodelle sind in der Lage, entweder korrigierte Standardfehler zu berechnen oder das gesamte Modell konsistent und effizient zu schätzen.

Eine andere Form der Anordnung von Kategorien ergibt sich bei ordinalskalierten abhängigen Variablen. Hier lassen sich die Kategorien entlang einer Dimension in eine Rangordnung bringen. In *ordinalen Logitmodellen* wird versucht, diese zusätzliche Rangordnungsinformation zu berücksichtigen. Ausgangspunkt der Modellformulierung sind Logits. Bei dem *Logitmodell der benachbarten Kategorien* werden die Logitgleichungen für jeweils zwei aufeinanderfolgende Kategorien i und i+1 formuliert:

$$\ln\left(\frac{p_{i+1}}{p_i}\right) = b_{0i} + \sum_{k=1}^{K} X_k b_k \qquad i=1,2,\ldots,I-1 \qquad (15)$$

In dem *Logitmodell der sequentiellen Anordnung* wird dagegen eine Kategorie der abhängigen jeweils mit allen größeren kontrastiert

$$\ln\left(\frac{p_{Y>i}}{p_i}\right) = b_{0i} + \sum_{k=1}^{K} X_k b_k \qquad i=1,2,\ldots,I-1 \qquad (16)$$

Die Summe der Kategorien kleiner oder gleich i wird schließlich bei dem Modell der *kumulativen Logits* der Summe der Kategorien größer i gegenübergestellt

$$\ln\left(\frac{p_{Y>i}}{p_{Y\leq i}}\right) = b_{0i} + \sum_{k=1}^{K} X_k b_k \qquad i=1,2,\ldots,I-1 \qquad (17)$$

Ein Unterschied zu den Logitmodellen für nominalskalierte abhängige Variablen besteht darin, daß zwar bei insgesamt I Kategorien I-1 Regressionskonstanten geschätzt werden. Die Regressionsgewichte der unabhängigen Variablen weisen aber für sämtliche modellierten Logits gleiche Werte auf. Dies erhöht zum einen die Effizienz der Parameterschätzung. Zusammen mit Restriktionen für die Regressionskonstanten wird gleichzeitig sichergestellt, daß die Beziehung zwischen den unabhängigen Variablen und der erklärenden Variablen monoton steigend oder fallend ist. Es muß allerdings erwähnt werden, daß diese Restriktion nicht immer eingehalten wird. Insbesondere das Modell der sequentiellen Anordnung wird oft so dargestellt daß auch die Regressionsgewichte bei jeder der I-1 Logitgleichungen variieren können (vgl. Ludwig-Mayerhofer, 1992).

Nur für das Modell (17) der kumulativen Logits muß ein eigener Schätzalgorithmus implementiert werden. Das Logitmodell (16) der sequentiellen Anordnung läßt sich dagegen als Produkt von binären Logitmodellen (7) darstellen und kann daher mit jedem Programm zur binären Logitanalyse geschätzt werden. Das Modell der benachbarten Kategorien (15) kann als ein spezifisches logistisches Zufallsnutzenmodell (10) spezifiziert und geschätzt werden. Auch hierfür ist also kein eigenes Schätzprogramm notwendig.

Trotz der Notwendigkeit eines eigenen Schätzprogramms wird von den drei ordinalen Modellen das der kumulativen Logits am häufigsten eingesetzt. Dieses Modell erlaubt nämlich eine interessante Interpretation, nach der die ordinale abhängige Variable eine ungenaue Messung einer unbeobachteten metrischen Variablen ist. Die Kategorie i der ordinalen beobachteten Variablen wird danach genau dann realisiert, wenn die unbeobachtete metrische Variable zwischen zwei Schwellenwerten (thresholds) hegt. Die eigentlich interessierende Variable ist die unbeobachtete Variable. Bei der folgenden Übersicht über Implementationen von Logitmodellen in Statistiksoftware beziehe ich mich daher bei ordinalen Modellen nur auf das Modell (17) der kumulativen Logits.

Allen Logitmodellen gemeinsam ist, daß sie als Spezialfälle der allgemeinen Gleichungen (5) bzw. (6) aufgefaßt werden können. Eine weitere Gemeinsamkeit ist, daß die Regressionskoeffizienten i.a. nach der *»Maximum Likelihood«* -Methode (ML-Methode) geschätzt werden. Dabei werden die Koeffizienten so bestimmt, daß die beobachteten Stichprobenwerte bei diesen Schätzwerten mit einer maximalen Wahrscheinlichkeit auftreten können. Die ML-Methode hat günstige statistische Eigenschaften. Die Schätzungen sind bei größeren Stichproben in etwa um die tatsächlichen Populationswerte normalverteilt, wobei sich auch die Standardfehler bzw. Varianzen und Kovarianzen schätzen und in sogenannten Wald-Statistiken für inferenzstatistische Tests nutzen lassen. Darüber hinaus lassen sich Likelihood-Ratio-Tests (LR-Tests) und Lagrangian-Multiplier-Tests (LM-Tests) anwenden (vgl. Buse, 1982). Voraussetzung für inferenzstatistische Anwendungen ist allerdings, daß die Modelle korrekt spezifiziert sind, also die Beziehung zwischen der abhängigen Variablen und den Regressorvariablen korrekt wiedergegeben ist. Bei einem fehlspezifizierten Modell sind die geschätzten Standardfehler dagegen verzerrt. Es ist in dieser Situation jedoch möglich, korrigierte Standardfehler zu schätzen (vgl. White, 1982).[5]

---

[5] Bei fehlspezifizierten Modellen sind nach der Korrektur der Standardfehler nur Wald-Tests asymptotisch korrekt. LM-Test und LR-Test sind nicht anwendbar.

Variablendefinition und Aufbau der Datenmatrix

Nach der kurzen Darstellung verschiedener Logitmodelle kann ich nun auf die Umsetzung in Statistiksoftware eingehen. Zunächst ein eher technischer Hinweis. In den meisten Statistikprogrammen werden die zu analysierenden Daten so angeordnet, daß alle Informationen einer Untersuchungseinheit in einer Zeile der Datenmatrix stehen. Jede Zeile der Matrix steht also für einen Fall. Bei einigen Prozeduren zur Logitanalyse wird dagegen vorausgesetzt, daß in jeder Zeile der Datenmatrix die Informationen eines Falles für nur jeweils eine Ausprägung der abhängigen Variablen stehen. Zur Verdeutlichung ein Beispiel. Es sei angenommen, daß ein logistisches Zufallsnutzenmodell (10) geschätzt werden soll. Die abhängige Variable Y habe die drei Ausprägungen 1,2,3. Die Wahrscheinlichkeiten sollen durch die drei (alternativenspezifischen) unabhängigen Variablen X_{11}, X_{21}, X_{31} erklärt werden. Für diese Variablen ist das gemeinsame Regressionsgewicht b_1 zu schätzen. Zusätzlich sollen zwei Regressionskonstanten b_{01} und b_{02} für die beiden ersten Kategorien geschätzt werden. Die Modellgleichungen ergeben sich dann nach:

$$
\begin{aligned}
p_1 &= \frac{\exp(b_{10} + b_1 X_{11})}{\exp(b_{10} + b_1 X_{11}) + \exp(b_{20} + b_1 X_{21}) + \exp(b_1 X_{31})} \\
p_2 &= \frac{\exp(b_{20} + b_1 X_{21})}{\exp(b_{10} + b_1 X_{11}) + \exp(b_{20} + b_1 X_{21}) + \exp(b_1 X_{31})} \qquad (18) \\
p_3 &= \frac{\exp(b_1 X_{31})}{\exp(b_{10} + b_1 X_{11}) + \exp(b_{20} + b_1 X_{21}) + \exp(b_1 X_{31})}
\end{aligned}
$$

Bei der üblichen fallweisen Anordnung der Daten sind alle Informationen für einen Fall in einer Zeile der Datenmatrix gespeichert. Tabelle 2 zeigt hierzu ein Beispiel für drei exemplarische Fälle. Die erste Spalte der Tabelle enthält die Fallnummer (»IDNR«). In der letzten Spalte ist die Konstante »One« zur Schätzung der Regressionskonstanten aufgeführt. In einigen Programmen muß die Konstante explizit als Regressorvariable in die Modellgleichung aufgenommen werden, in anderen werden Konstanten automatisch hinzugefügt.

Wenn ein Programm zur Schätzung der Koeffizienten in Gleichung (18) voraussetzt, daß die Zeilen der Datenmatrix für eine einzelne Ausprägung der abhängigen Variablen stehen, werden aus den drei Zeilen in Tabelle 2 insgesamt neun Zeilen. Das Ergebnis ist in Tabelle 3 festgehalten.

Zusätzlich oder als Alternative zur abhängigen Variablen Y wird in den Prozeduren dabei meist eine 0/1-kodierte Dummyvariable D_j verlangt. Steht die

Tabelle 2: Datenstruktur mit fallweiser Anordnung

IDNR	Y	X_{11}	X_{21}	X_{31}	One
1	2	1.0	3.0	2.0	1
2	3	2.0	1.5	3.2	1
3	2	2.0	1.9	1.5	1
...	...	...	...	...	...

Tabelle 3: Datenstruktur mit alternativenorientierter Anordnung

IDNR	y	d_y	$x_{.1}$	One_1	One_2
1	2	0	1.0	1	0
1	2	1	3.0	0	1
1	2	0	2.0	0	0
2	3	0	2.0	1	0
2	3	0	1.5	0	1
2	3	1	3.2	0	0
3	2	0	2.0	1	0
3	2	1	1.9	0	1
3	2	0	1.5	0	0
...	...	...	...	...	...

Datenzeile für die Ausprägung, die tatsächlich auftritt, hat die Dummyvariable den Wert Eins, ansonsten den Wert Null. Die drei Regressorvariablen X_{11}, X_{21} und X_{31}, die sich auf verschiedene Ausprägungen der abhängigen Variable beziehen, im logistischen Zufallsnutzenmodell aber einen gemeinsamen Regressionskoeffizienten (b_1) aufweisen, werden bei dieser alternativenorientierten Anordnung der Daten untereinander in eine Spalte geschrieben. Umgekehrt wird die Regressionkonstante »One« durch zwei 0/1-kodierte Variablen ersetzt, die den Wert Eins bei allen Zeilen aufweisen, die erste bzw. zweite Ausprägung der abhängigen Variablen repräsentieren.

Statistikprogramme für Logitanalysen

Nach diesen allgemeinen Hinweisen will ich nun auf die Realisierung der Modelle in verschiedenen Softwareprodukten eingehen. Es sei vorweg erwähnt, daß nicht der Anspruch erhoben werden kann, alle auf dem Markt verfügbaren Produkte einzubeziehen. Ich beschränke mich hier auf sechs Programmpakete: BMDP, LIMDEP, SAS, SPSS, SYSTAT und TDA. Mit BMDP, SAS und SPSS sind die drei »großen« Programmsysteme aufgenommen, von denen zumindest eines an jedem Universitätsrechenzentrum verfügbar sein sollte. SYSTAT war eines der ersten Systeme für PC-Anwender unter MS-DOS. Die Herstellerfirma wurde kürzlich von SPSS aufgekauft, was möglicherweise Auswirkungen auf zukünftige SPSS-Versionen haben wird. LIMDEP ist ein Programm, das sehr stark in der Umsetzung ökonometrischer Modelle ist TDA war ursprünglich auf Ereignisanalysen beschränkt, umfaßt inzwischen aber eine Vielzahl weiterer Modelle. Im Unterschied zu allen anderen Programmen wird es kostenlos zur Verfügung gestellt.

Von allen Systemen gibt es in relativ kurzer Zeit neue Versionen. Mit Einschränkungen des jeweiligen Angebots an Logitmodellen ist nicht zu rechnen. Ältere Versionen haben aber möglicherweise einen eingeschränkten Leistungsumfang. Ich beziehe mich im folgenden bei BMDP auf die Version BMDP/DYNAMIC 7.0 für MS-DOS mit integrierten DOS-Extender, bei LIMDEP auf die 386-Version 6.0 mit MS-DOS-Extender, bei SAS auf die Version 6.08 für Windows, bei SPSS auf die Version 6.0 für Windows, bei SYSTAT auf das Zusatzmodul LOGIT in der Version 2.01 für MS-DOS und bei TDA auf die Version 5.7 für MS-DOS mit DOS-Extender.

Die einzelnen Logitmodelle sind auf unterschiedliche Weise in den Programmsystemen realisiert. Zu einigen Modellen gibt es spezielle Prozeduren. Alternativ können einige Logitmodelle auch mit Prozeduren geschätzt werden, die in erster Linie für andere Analysemodelle gedacht sind. So sind Prozeduren zur Schätzung von binären und ordinalen Probitmodellen in der Regel auch in der Lage, die korrespondierenden Logitmodelle (7) und (17) zu schätzen. Wenn in einem Programmsystem sowohl eine eigene Prozedur für Logitanalysen zur Verfügung steht als auch die Schätzung über ein andere Prozedur, erwähne ich nur die Logitprozedur. Falls ein Logitmodell nur über eine Prozedur verfügbar ist, die nicht speziell für Logitanalysen implementiert wurde, wird diese Prozedur genannt. Da in solchen Fällen oft unklar ist wie ein Logitmodell angefordert wird, gebe ich zusätzlich jeweils ein Beispiel mit den entsprechenden Kommandos.

Mit Ausnahme von TDA stehen in allen Systemen auch Prozeduren zur Minimierung bzw. Maximierung einer benutzerdefinierten Funktion zur Verfügung.[‘] Theoretisch steht damit erfahrenen Anwendern mit guten Statistik-

---

[‘] **In TDA wird dies erst in der nächsten Version der Fall sein.**

kenntnissen der Weg offen, beliebige Logitmodelle mit Hilfe dieser allgemeinen Prozedur zu schätzen. Da hierzu jedoch die Umsetzung der jeweiligen Likelihoodfunktionen und bisweilen auch der ersten Ableitungen nach den Modellparametern als benutzerdefinierte Funktionen vorausgesetzt wird, verzichte ich auf eine nähere Beschreibung dieser Vorgehensweise. Die im folgenden genannten Prozeduren benötigen für ihre Anwendung allein die Kenntnis der Syntaxregeln des jeweiligen Programmsystems.

a) BMDP

BMDP stellt für Logitanalysen zwei eigene Prozeduren zu Verfügung. Das binäre Logitmodell (7) kann über die Prozedur *LR* geschätzt werden. Die abhängige Variable sollte 0/1-kodiert sein. Referenzkategorie ist der Wert 0. Für das multinomiale Logitmodell (9) steht die Prozedur *PR* zur Verfügung. Die Ausprägungen der abhängigen Variablen werden über den Befehl »CODES« in der Sektion »/GROUP« spezifiziert Die hierbei zuletzt spezifizierte Kategorie ist Referenzkategorie. Mit der Prozedur PR können über den Befehl »TYPE=ORDINAL« zudem die Regressionskoeffizienten des Modells (17) für kumulativen Logits geschätzt werden. Mit »TYPE=EQU« wird das Logitmodell (15) der benachbarten Kategorien geschätzt

Das logistische Zufallsnutzenmodell (10) ist nicht als eigene Prozedur implementiert. Die Modellparameter können jedoch mit dem Modul *2L* für Modelle zur Ereignisanalyse geschätzt werden, wenn die Daten alternativenorientiert angeordnet sind. Angenommen, die Daten seien wie in Tabelle 3 angeordnet, und es lägen Informationen für 100 Fälle in freiem Format in der Datei »tabelle3.dat« vor. Das Modell aus Gleichung (18) kann dann mit folgenden Anweisungen geschätzt werden:

```
/INPUT      VARIABLES - 6.
            FORMAT = FREE.
            FILE= 'tabelle3.dat'.
/VARIABLE   NAMES = idnr, y, dy, xl, onel, one2.
/GROUP      CODES(idnr) = 1 TO 100.
/FORM       TIME=y.
            STATUS=dy.
            RESPONSE=l.
/REGRESSION COVARIATES=xl,onel,one2.
            STRATA=idnr.
/END
```

b) LIMDEP

LIMDEP stellt für alle angesprochenen Logitmodelle eigene Prozeduren zur Verfügung. Für das binäre und multinomiale Logitmodell (7) bzw. (9) ist die

f) TDA

TDA stellt für alle Logitmodelle das Kommando *LOGIT* zur Verfügung. Das binäre Logitmodell (7) wird durch die Angabe »LOGIT=l« aktiviert. Zur Schätzung des Modells (17) der kumulativen Logits wird das Kommando »LOGIT=2« verwendet. Mit »LOGIT=3« wird das Mischmodell (11) angefordert, das sowohl das multinomiale Logitmodell (9) wie das logistische Zufallsnutzenmodell (10) als Speziallfall enthält. Referenzkategorie ist die erste Kategorie. Schließlich können auch binäre logistische Panelmodelle nach Gleichung (13) über die Anweisung LOGIT=4 geschätzt werden. Die Zahl der Panelwellen ist unbeschränkt. Bei den Daten wird stets eine fallweise Anordnung angenommen.

Spezielle Programrneigenschaften

Die Tabelle listet zunächst einige Eigenschaften zur Modellspezifikation auf. In der Regel ist die Zahl der Ausprägungen der abhängigen Variablen für alle Untersuchungseinheiten gleich. Bei einigen Anwendungen des logistischen Zufallsnutzenmodells kann die Anzahl der Ausprägungen aber auch variieren. Entsprechende Möglichkeiten sind in der LIMDEP-Prozedur DISCRETE und im SYSTAT-Modul LOGIT bei alternativenorientierter Anordnung vorgesehen. Wenn eine unabhängige Variable nominalskaliert ist, kann ihr Einfluß auf die Ausprägungen der abhängigen Variablen analog zur Vorgehensweise bei der Varianzanalyse über Dummy-Variablen untersucht werden. Einige Prozeduren generieren automatisch entsprechende Dummy-Variablen, wenn eine Regressorvariable als kategorial deklariert wird. Eine Vereinfachung der Modellspezifikation ergibt sich auch, wenn ein Programm die Spezifikation von Interaktionseffekten erlaubt. Ist dies nicht der Fall, müssen stattdessen für jeden Interaktionseffekt eigene Variablen erzeugt und als zusätzliche Regressorvariablen in das Modell aufgenommen werdea

Eine weitere Option bezieht sich auf die Auswahl der Regressorvariablen. Die Suche nach einem Logitmodell kann unter Umständen erleichtert werden, wenn eine Prozedur aus einer Menge möglicher Regressorvariablen die Variablen herausfindet, die die Ausprägungen der abhängigen Variablen am besten prognostizieren. Eine solche automatische Modellselektion folgt der gleichen Logik wie bei der schrittweisen Regression.Für binäre Logitmodelle gibt es eine umfangreiche Literatur zur Residuenanalyse (vgl. Hosmer/Lemeshow, 1989). Die dort genannten Statistiken können in BMDP, SAS, SPSS und SYSTAT angefordert werden. Verallgemeinerungen für abhängigen Variablen mit mehr als zwei Ausprägungen stehen allerdings noch aus.

Neben den Regressionskoeffizienten ist ein Anwender oft auch an den durch das Modell geschätzten Wahrscheinlichkeiten der Ausprägungen der abhängigen Variablen interessiert Einige Prozeduren erlauben das Abspeichern dieser

Tabelle 4: Eigenschaften der Logitprozeduren

Eigenschaft	BMDP LR	BMDP PR	BMDP 2L	LIMDEP Logit	LIMDEP Discrete	LIMDEP Ordered
Variable Kategorienzahl					+	
Kategoriale Regressorvariablen	+	+				
Interaktionseffekte	+	+				
Automatische Selektion	+	+	+			
Residuenanalyse	+					
Ausgabe von Wahrscheinlichkeiten	+	+		(+)	+	+
Schätzerkovarianzen	+	+	+	+	+	+
Korrektur bei Fehlspezifikation				(+)	(+)	(+)
Korrektur bei mehrstufigen Modellen					+	
Wald-Test				+	+	+
LM-Test				(+)	(+)	(+)
LR-Test				(+)	(+)	(+)
Fitfunktion	ln L	ln L	ln L	ln L	ln L	ln L

(+) bedeutet, daß die entsprechende Option über zusätzliche Anweisungen realisierbar ist.

Fortsetzung Tabelle 4

Eigenschaft	SAS			SPSS Windows		
	Logistic	Catmod	Phreg	Logistic Regression	Makro Clogit	Makro Ologit
Variable Kategorienzahl						
Kategoriale Regressorvariablen		+		+		
Interaktionseffekte	+	+		+		
Automatische Selektion	+			+		
Residuenanalyse	+			+		
Ausgabe von Wahrscheinlichkeiten	+	+		+	+	
Schätzerkovarianzen	+	+	+	+	+	+
Korrektur bei Fehlspezifikation						
Korrektur bei mehrstufigen Modellen						
Wald-Test		+	+		+	
LM-Test					+	
LR-Test						
Fitfunktion	ln L	ln L	ln L	-2 ln L	-2 ln L	-2 ln L

(+) bedeutet, daß die entsprechende Option über zusätzliche Anweisungen realisierbar ist.

Fortsetzung Tabelle 4

Eigenschaft	SYSTAT Logit	TDA Logit
Variable Kategorienzahl	+	
Kategoriale Regressorvariablen	+	
Interaktionseffekte	+	
Automatische Selektion	+	
Residuenanalyse	+	
Ausgabe von Wahrscheinlichkeiten	+	
Schätzerkovarianzen	+	+
Korrektur bei Fehlspezifikation	+	+
Korrektur bei mehrstufigen Modellen		
Wald-Test	+	+
LM-Test	+	+
LR-Test		
Fitfunktion	ln L	ln L

(+) bedeutet, daß die entsprechende Option über zusätzliche Anweisungen realisierbar ist.

Wahrscheinlichkeiten. Anderenfalls ist es notwendig, die Wahrscheinlichkeiten mit Hilfe der Modellgleichungen zu berechnen.

Die letzten Zeilen von Tabelle 4 beziehen sich auf die Möglichkeiten von statistischen Tests. Angegeben ist, ob eine Teststatistik in einer Prozedur angefordert werden kann. Wald-Tests beziehen sich dabei auf Tests beliebiger linearer Kontraste. Ist die Option nicht verfügbar, können entsprechende Tests auch selber berechnet werden, falls die Varianzen und Kovarianzen der Regressionskoeffizienten (Schätzerkovarianzen) angefordert werden kann. Wenn die White-Korrektur verfügbar ist, sind Wald-Tests auch bei leicht fehlspezifizierten Modellen möglich. Korrigierte Standardfehler sind auch für Tests in mehrstufigen Modellen notwendig. Die LM-Teststatistik ist nur selten implementiert. LR-Tests werden ebenfalls meist nur bei schrittweiser Modellselektion direkt berechnet. Sie sind aber im Prinzip in allen Programmen verfügbar, da für ihre Berechnung nur die Fitfunktionen der Modellschätzungen benötigt werden, die von sämtlichen Programmen ausgegeben werden. Zu beachten ist allerdings, ob die Programme als Fitfunktion die Log-Likelihoodfunktion, die negative Log-Likehoodfunktion oder das Zweifache der negativen Log-Likelihoodfunktion berechnen. Die LR-Teststatistik ist die Differenz der zweifachen negativen Log-Likelihoodfunktion zweier Modelle, die sich nur dadurch unterscheiden, daß in einem Modell zusätzliche Restriktionen spezifiziert sind. Restriktionen müssen meist indirekt spezifiziert werden (vgl. Kühnel, 1992). Eine Ausnahme ist TDA. Lineare Restriktionen werden hier bereits bei der Schätzung berücksichtigt.

Resumé

Tabelle 5 enthält eine Übersicht über die Implementation von Logitmodellen in den betrachteten sechs Programmsystemen. Nicht explizit aufgeführt sind die beiden ordinalen Logitmodelle der benachbarten Kategorien (15) und der sequentiellen Anordnung (16), die über das binäre bzw. konditionale Logitmodell geschätzt werden können. Ebenfalls nicht berücksichtigt sind mehrstufige Logitmodelle (14). Auch sie können im Prinzip stets über mehrfache Anwendung des logistischen Zufallsnutzenmodells geschätzt werden. Korrigierte Standardfehler berechnet aber nur LIMDEP. In allen hier betrachteten Programmsystemen sind die meisten Varianten der Logitmodellen verfügbar. Einschränkungen gibt es beim binären Panelmodell (13). Binäre Panelmodelle mit mehr als zwei Wellen können nur mit LIMDEP und TDA geschätzt werden.[*]

[*] Bei zwei Wellen reduziert sich das Modell (14) zu einem binären Logitmodell nach Gleichung (8). Bei Modellen, die eine variable Anzahl von Ausprägungen der unabhängigen Variablen erlauben, können binäre Panelmodelle auch als Spezialfall kon-

Tabelle 5: Verfügbarkeit von Logitmodellen in Statistikprogrammen

Prozedurname in:	Logitmodell: binär	multinomial	Zufallsnutzen	kumulierte Logits	Panel
BMDP	LR	PR	2L	PR	-
LIMDEP	LOGIT	LOGIT	DISCRETE	ORDERED	LOGIT
SAS	LOGISTIC	CATMOD	PHREG	LOGISTIC	-
SPSS	LOGISTIC REGRESSION	Makro MLOGIT	Makro CLOGIT	Makro OLOGIT	-
SYSTAT	LOGIT	LOGIT	LOGIT	-	-
TDA	logit=1	logit=3	logit=3	logit=2	logit=4

Als Resumé dieser Übersicht bleibt festzuhalten, daß der Anwendung von Logitmodellen auf Individualdatenebene von Seiten der verfügbaren Statistiksoftware wenig entgegensteht. Die wichtigsten Modellvarianten stehen in allen Programmsystemen zur Verfügung.

Vertriebsadressen

Die in diesem Beitrag diskutierten Programme können unter folgenden Adressen angefordert werden:

BMDP: BMDP Statistical Software, Cork Technology Park, Model Farm Road, Cork, Ireland.

LIMDEP: Econometric Software Inc. 43 Maple Avenue, Bellport, NY 11713, USA.

SAS: SAS Institute Inc. SAS Campus Drive, Cary, NC 27513, USA.

SPSS: SPSS Inc. 444 N. Michigan Avenue, Chicago, IL 60611, USA.

---

ditionaler Logitmodelle spezifiziert werden. Dazu müssen die in Gleichung (13) auftretenden Produkte $Z_{in}=Y_{n}*X_{in}$ berechnet werden. Die Variablen sind dann als erklärende Variablen in das Modell aufzunehmen.

SYSTAT: SYSTAT, Inc. 1800 Sherman Avenue, Evanston, IL 60201-3793, USA.

TDA: Dr. Götz Rohwer, Universität Bremen, FB 8, EMPAS, Postfach 330 440, D-28334 Bremen.

TDA ist auch über FTP auf dem FTP-Server der Universität Bremen (Internetadresse: FTP.Uni-Bremen.DE) verfügbar. Das Programm und die Dokumentation in Form von Postscript-Dateien steht unter pup/uni-Bremen/Institutes/ZeS/TDA. Login-name ist »anonymous«, Password ist die e-mail adresse des Benutzers.

Aufgeführt habe ich jeweils die zentrale Adresse des Herstellers des Programmsystems. Die großen Anbieter haben eigene Vertriebe für Deutschland oder Europa. Die Programme werden z.T. auch über Vertreiber von Schul- und Universitätssoftware wie z.B. iec ProGAMMA (P.O. Box 841, 9700 AV Groningen, Niderlande) angeboten.

Literatur

Andreß, H.-J., Hagenaars, J.A. und Kühnel, S.M. (in Vorbereitung), Multivariate Analyse kategorialer Daten. Modelle und Anwendungen. Frankfurt: Campus.

Buse, A. (1982), The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. American Statistician, 36: 153-157.

Cramer, J.S. (1991), The Logit Model: An Introduction for economics. London: E. Arnold.

Demaris, A. (1992), Logit Modeling. Practical Applications. Newbury Park: Sage.

Hosmer, D.H. und Lemeshow, S.L. (1989), Applied Logistic Regression. New York: Wiley.

Kleinbaum, D.G. (1994), Logistic regression. A self-learning text. New York: Springer.

Long, J.S. (1987), A Graphical Method for the Interpretation of Multinomial Logit Analysis. Sociological Methods and Research, 15: 420-446.

Ludwig-Mayerhofer, W. (1992), Statistik-Software zur Schätzung von Regressions-Modellen für ordinale abhängige Variablen. ZA-Information, 31: 93-99.

Kühnel, S. (1990), Lassen sich mit SPSS*-Matrix anwenderspezifische Analyseprobleme lösen? Ein Anwendungstest am Beispiel der multinomialen logistischen Regression. ZA-Information, 27: 89-109.

Kühnel, S. (1992), Sparsame Modellierung mit logistischen Zufallsnutzenmodellen. ZA-Information, 31: 70-92.

Kühnel, S.-M. (1993), Zwischen Boykott und Kooperation. Frankfurt a.M.: P. Lang.

Maier, G. u. Weiss, P. (1990), Modelle diskreter Entscheidungen. Wien u.a.: Springer.

Urban, D. (1993), Logit-Analyse. Stuttgart u.a: G. Fischer.

White, H. (1982), Maximum Likelihood Estimation of Misspecified Models. Econometrica, 50: 1-25.