SSOAR

Open Access Repository
www.ssoar.info

Über einige Anwendungs- und Interpretationsprobleme "anspruchsvoller" Schätzverfahren: Entgegnung auf den Beitrag von Langeheine

Jagodzinski, Wolfgang

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Jagodzinski, W. (1987). Über einige Anwendungs- und Interpretationsprobleme "anspruchsvoller" Schätzverfahren: Entgegnung auf den Beitrag von Langeheine. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 20, 56-63. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-205411

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Über einige Anwendungs- und Interpretationsprobleme „anspruchsvoller" Schätzverfahren

(Entgegnung auf den Beitrag von Langeheine)

von Wolfgang Jagodzinski

Es freut mich sehr, daß die am Ende meines Aufsatzes geäußerte Anregung, sich mit den statistischen Modellen für qualitative Daten intensiver zu beschäftigen, so prompt aufgegriffen worden ist. LANGEHEINE präsentiert nicht nur einen interessanten Ansatz, der bei großem N dem von mir angewandten OLS-Schätzverfahren unbedingt vorzuziehen ist, er kommt auch zu anderen Schlußfolgerungen als ich. Zwar weichen unsere Ergebnisse nicht allzu stark voneinander ab, wenn man das Panel insgesamt betrachtet, denn nach LANGEHEINE wählen über 40% in mindestens einer Umfrage eine unzutreffende - d.h. ihrer Einstellung nicht entsprechende - Antwortsequenz, während es nach meinen eigenen Schätzungen über 50% der Respondenten sind. Unsere Ergebnisse differieren jedoch dramatisch, wenn man auf die einzelnen Panelwellen blickt: Nach LANGEHEINE schwankt der Anteil der unzuverlässigen Antworten zwischen 14.6% und 21%, nach meinen eigenen Schätzungen liegt er in jeder Befragung in der Nähe von 50%. LANGEHEINE behauptet zudem, daß harte statistische Kriterien zwischen seinem Modell M4 und meinem Modell M3 diskriminieren: Während M4 gut mit den Daten verträglich ist, sei M3 unhaltbar. Demgegenüber vertrete ich zwei diametral entgegengesetzte Thesen, die nachfolgend kurz begründet werden sollen:

> These 1: In Anbetracht der geringen Zellenhäufigkeiten ist das von LANGEHEINE benutzte L^2 weder als deskriptives noch als inferenzstatistisches Anpassungsmaß geeignet und kann folglich nicht zwischen den Modellen M2-M4 diskriminieren.
>
> These 2: LANGEHEINE überschätzt in seinem Modell M4 den Anteil der zuverlässigen Mischtypantworten ganz erheblich und gelangt vor allem deshalb zu völlig anderen Ergebnissen als ich.

1.) LANGEHEINE zieht als Anpassungsfunktion die Likelihood-Ratio-Chi-Quadrat-Statistik L^2 heran, wobei

$$(1) \qquad L^2 = 2 * \sum_i f_{bi} * \ln(f_{bi}/f_{ei}).$$

Dabei sind f_{bi} und f_{ei} die beobachteten und die unter dem Modell erwarteten Häufigkeiten in der i-ten Zeile von Tabelle 1 (nachfolgend beziehe

ich mich der Einfachheit halber ausschließlich auf die Tabellen in LANGEHEINEs Beitrag). Solange die Häufigkeiten nicht aggregiert werden, läuft der Index i von 1 bis 27. Im allgemeinen nimmt L^2 umso höhere Werte an, je stärker beobachtete und erwartete Häufigkeiten auseinanderklaffen. Dies gilt jedoch nicht, wenn die beobachtete Häufigkeit in der i-ten Zeile Null beträgt; weil der Logarithmus von Null nicht definiert ist, wird der i-te Summand in (1) auf Null gesetzt, es wird also so getan, als stimmten beobachtete und erwartete Häufigkeiten exakt überein. Dies geschieht, wenn die Daten in Tabelle 1 nicht aggregiert werden, [1] in 9 von 27 Fällen. Ein Anpassungsmaß, das ein Drittel aller Abweichungen ignoriert, ist m.E. selbst für eine rein deskriptive Analyse unbrauchbar.

LANGEHEINE mildert das Problem insofern etwas ab, als er die Zellen mit einem Erwartungswert kleiner 1 zusammenfaßt. Beim Black & White-Modell M2 in Tabelle 1 werden also die sechs erwarteten Häufigkeiten von 0.45 zu einem Wert aufaddiert und ebenso die damit korrespondierenden beobachteten Häufigkeiten. Die Zahl der Nullzellen vermindert sich infolgedessen zwar um drei, aber immer noch bleiben bei der Berechnung von L^2 sechs von 22 Zellen unberücksichtigt.

Will man die Modellanpassung statistisch testen, führen nicht nur Nullzellen, sondern auch kleine erwartete Häufigkeiten zu Komplikationen, da die Testgröße L^2 bei kleinen Stichproben höchstens approximativ chiquadratverteilt ist. Dabei ist keineswegs geklärt, welchen Mindestwert die erwarteten Häufigkeiten erreichen bzw. überschreiten müssen, damit eine zufriedenstellende Approximation gewährleistet ist. Ähnlich wie beim Chiquadrat von PEARSON gilt im allgemeinen als ausreichend, daß bei mehreren Freiheitsgraden keine erwartete Häufigkeit die Zahl 5 unterschreitet. Sind zahlreiche Zellen vorhanden, so wird diese Bedingung von einigen Autoren noch etwas abgeschwächt. LANCASTER (1969) etwa erachtet die Approximation in solchen Fällen auch dann noch für akzeptabel, wenn ein Drittel bis ein Viertel der erwarteten Häufigkeiten zwischen 1 und 5 variiert. Nun genügt aber nicht einmal die verkleinerte Tabelle zu Modell M2 dem liberalisierten Kriterium von LANCASTER, denn in mehr als der Hälfte der 22 Zellen beträgt die erwartete Häufigkeit 1.81. Die Anwendung des Chiquadrat-Anpassungstests ist daher nicht gerechtfertigt.

M.E. führen die vorliegenden Simulationsstudien (vgl. etwa LARNTZ 1978; KOEHLER und LARNTZ 1980; KOEHLER 1986; weitere Nachw. bei

LANGEHEINE 1986) zu keiner anderen Beurteilung. Zwar wurden in solchen Studien auch Tabellen mit sehr niedrigen erwarteten Häufigkeiten analysiert, doch ging es meist um einfache Hypothesentests und nicht um komplizierte Modellanpassungstests, bei denen vorab mehrere Parameter geschätzt werden. Entfernte Ähnlichkeit mit der hier behandelten Konstellation hat vielleicht die Studie von KOEHLER (1986), die u.a. den Chiquadrat-Anpassungstest für hierarchische log-lineare Modelle zum Gegenstand hat, wobei sich die Simulationen freilich nur auf Modelle zur Überprüfung vollständiger statistischer Unabhängigkeit erstrecken. Was die Chiquadrat-Approximation von L^2 (bei KOEHLER: Gk^2) anbelangt, so fällt das Resumee sehr eindeutig aus: "The accuracy of the chi-squared approximation for G_{K}^{2} in sparse tables is generally unacceptable for testing the fit of log-linear modeis" (KOEHLER 1986: 490). Nun wäre es gewiß falsch, solche Ergebnisse vorschnell zu generalisieren. Solange jedoch einschlägige Simulationsstudien fehlen, scheint es mir ein Gebot der Vorsicht, auf die Anwendung des Chiquadrat-Anpassungstests zu verzichten, wenn mehr als 50% der erwarteten Häufigkeiten unter 2 liegen und mehr als ein Viertel der beobachteten Häufigkeiten gleich 0 sind, wie dies beim Black & White-Modell M2 der Fall ist.

Natürlich könnte man noch weitere Zellen zusammenfassen. Beim Black & White-Modell M2 etwa könnte man zusätzlich auch noch die Häufigkeiten jener zwölf Zellen addieren, in denen der Erwartungswert 1.81 beträgt. Damit würden - und so könnte man diesen Schritt zu rechtfertigen versuchen - nicht nur alle Nullzellen beseitigt, es verbliebe auch nur noch eine einzige Zelle mit einer erwarteten Häufigkeit unter 5. Man braucht die statistisch-technische Problematik des Vorgehens gar nicht zu erörtern, denn eines ist auf Anhieb erkennbar: Je mehr Zellen zusammengelegt werden, desto mehr Möglichkeiten hat man, durch geschickte Aggregation den Modellfit zu verbessern. Im konkreten Fall stimmt die Summe der zwölf beobachteten Häufigkeiten (=23) mit der Summe der zwölf erwarteten Häufigkeiten (=21.72) schon recht gut überein. Durch eine einfache und zudem noch formal legitimierte Addition wären mit einem Schlage die meisten Unstimmigkeiten zwischen Modell und Daten beseitigt! Das L^2 würde nicht einmal die Zahl der Freiheitsgrade erreichen und mithin eine exzellente Anpassung an die Daten signalisieren - weit besser als jene, die LANGEHEINE für sein Modell M4 berichtet. Um nicht mehr Konsistenz zwischen Modell und Daten zu suggerieren, als tatsächlich existiert, habe ich in meiner Studie auf solche Aggregationen ganz verzichtet. Wie immer man sich aber zu dieser

Frage stellen mag, eine Konsequenz scheint mir ganz unvermeidlich: Analysiert man die Daten auf dem Aggregationsniveau, für das sich LANGEHEINE entschieden hat, so ist das L^2 sowohl als deskriptives wie auch als inferenzstatistisches Anpassungsmaß ungeeignet; faßt man weitere Zellen zusammen, so paßt bereits das Black & White-Modell ausgezeichnet zu den Daten. In keinem Fall leistet L^2 das, was es nach LANGEHEINE leisten soll, nämlich die Modelle M2 und M3 als empirisch nicht haltbar zurückzuweisen. Die erste These ist damit begründet.

2.) Die eben vorgetragenen Einwände richten sich gegen die Verwendung von L^2 als Anpassungsmaß, sie richten sich noch nicht gegen die von LANGEHEINE präsentierten Modelle. Viele dieser Modelle reproduzieren die beobachteten Häufigkeiten in Tabelle 1 keinesfalls schlechter als mein Modell M3, sie scheinen mir aber aus methodischen wie aus inhaltlichen Gründen inakzeptabel. Die nachfolgende Diskussion konzentriert sich ausschließlich auf das von LANGEHEINE letztlich favorisierte Modell M4 (vgl. LANGEHEINE 1987 S.53). Nach Tabelle 4 beträgt der Anteil der Mischtypen in diesem Modell .578. Von diesen 152*. 578 ~ 88 Personen, deren Einstellungen sich voraussetzungsgemäß[2] während des Untersuchungszeitraums nicht ändern, sollen ca. 45% (= . 722*. 785*. 789*100), also etwas 39 Personen, in allen drei Wellen zuverlässig antworten. Tatsächlich wählen, wie man der ersten Spalte in Tabelle 1 entnehmen kann, nur 38 Befragte in allen drei Wellen eine Mischtyp-Antwort, selbst dieser Anteil wird also leicht überschätzt. Doch ist das nicht der entscheidende Mangel von M4.

Entscheidend ist vielmehr, daß m.E. nicht jeder, der dreimal hintereinander eine Mischtyp-Antwort gibt, als zuverlässig antwortender, stabiler Mischtyp klassifiziert werden kann. Wer etwa in der ersten Welle 'Bekämpfung steigender Preise' und 'Partizipation' an erster und zweiter Stelle nennt, in der zweiten Welle 'Meinungsfreiheit' und 'Ruhe und Ordnung' und in der dritten Welle 'Ruhe und Ordnung` und 'Partizipation', wählt zwar drei "gemischte Antwortkombinationen", aber er antwortet nicht zuverlässig, denn ein Rückschluß auf die (als stabil vorausgesetzten) subjektiven politischen Präferenzen ist gerade nicht möglich. Zu einer abweichenden Beurteilung könnte man nur gelangen, wenn man von den in der Frage angesprochenen Politikinhalten gänzlich abstrahieren und nur noch darauf abstellten wollte, daß der Betreffende stets eine materialistische mit einer postmaterialistischen Antwort kombiniert. Die beiden materialistischen Ziele wären ebenso gegen-

einander austauschbar wie die beiden postmaterialistischen. Vielleicht lassen sich auch für ein solches behavioristisches Verständnis des Befragtenverhaltens in INGLEHARTs Schriften Anhaltspunkte finden, mir scheint es jedoch inadäquat. Daher bin ich davon ausgegangen, daß ein stabiler Mischtyp nur dann zuverlässig antwortet, wenn er in allen drei Wellen die gleiche Antwortkombination wählt. Erste und zweite Priorität mögen wechseln, doch müssen stets dieselben beiden Ziele genannt werden. Das Kriterium noch enger zu fassen, schien mir nicht sinnvoll, da ja auch für die Klassifikation als Materialist und Postmaterialist unerheblich ist, in welcher Reihenfolge die beiden einschlägigen Ziele genannt werden.

Insgesamt 21 Befragte wählen in der ALLBUS-Retest-Studie in allen drei Wellen die gleiche "Mischkombination", wobei mit Abstand am häufigsten 'Meinungsfreiheit' und 'Ruhe und Ordnung' kombiniert werden. Auf diesen Tatbestand hatte ich in meinem Aufsatz ausdrücklich hingewiesen. Akzeptiert man also das von mir vorgeschlagene Kriterium, so überschätzt LANGEHEINE den Anteil der zuverlässigen Mischtypantworten ganz erheblich. Wenn sich dies nicht in einer schlechten Modellanpassung bemerkbar macht, so nur deshalb, weil die entsprechenden Antwortsequenzen in Tabelle 1 nicht gesondert ausgewiesen sind. Der Tabelle ist nur zu entnehmen, daß 38 Personen dreimal hintereinander eine Mischtyp-Antwort geben, nicht aber, ob sie in allen drei Wellen die gleiche "Mischkombination" gewählt haben oder nicht. LANGEHEINEs Modell paßt also nicht zu den Daten. Diese Fehlspezifikation ist auch der Hauptgrund, weshalb die Schätzungen der Modelle M3 und M4 so weit auseinanderliegen.

3.) Damit erweist sich LANGEHEINEs Kritik in meinen Augen als wenig stichhaltig. Weder eignet sich L^2, zwischen den Modellen M2-M4 zu diskriminieren, noch ist sein Modell M4, wenn man alle in der Stichprobe enthaltenen Informationen auswertet, mit den Beobachtungen verträglich. In gewisser Weise stellen das Black & White-Modell M2 und LANGEHEINEs Modell M4 Extrempositionen dar. Nach M4 wird die beobachtete Fluktuation ausschließlich durch Antwortunsicherheit von Personen mit festen Einstellungen bzw. - im konkreten Fall - mit stabilen politischen Zielvorstellungen hervorgerufen, nach M2 dagegen ausschließlich durch Meinungslosigkeit. M.E. gibt es plausible Gründe, weshalb Befragte bezüglich der Rangordnung der Zielprioritäten keine feste Meinung haben, wenn sich dies in der ALLBUS-Retest-Studie auch nicht empirisch belegen läßt. Meinungslos

könnten zum einen Personen sein, für die politische Themen im allgemeinen oder/und die in der Frage angesprochenen politischen Ziele im besonderen eine sehr geringe Zentralität haben. Meinungslosigkeit könnte aber auch bei Individuen auftreten, die sich für Politik interessieren, die aber einige der in der Frage erwähnten Politikziele gleichermaßen positiv bewerten. Sie werden durch das Item gezwungen (forced choice), als gleichrangig erachtete Ziele in eine Rangordnung zu bringen, wodurch bei ihnen eine Form von Cross-Pressure erzeugt wird.[3] Wenn solche Personen in ihrem Urteil schwankend werden, so entspricht das den Vorhersagen einiger Konsistenztheorien.[4]

Meinungslosigkeit findet also gerade bei Forced-Choice-Items eine einfache und anschauliche Interpretation. Daß daneben bei Befragten mit klaren Zielprioritäten auch Antwortunsicherheit entstehen kann, wird bereits in meinem Modell M3 postuliert. Dieses Modell läßt sich als ein Versuch deuten, zwischen den beiden Extremen M2 und M4 zu vermitteln. Wenn Antwortunsicherheit hier auf die erste Welle beschränkt wurde, so hatte das allein methodisch-technische Gründe. Ein angemessenes Modell müßte also, wie ich bereits in den Schlußbemerkungen angedeutet hatte, neben Meinungslosigkeit auch Antwortunsicherheit in allen drei Wellen zulassen - es wäre, wenn man so will, eine Fortentwicklung der Modelle M2-M4.

Selbst wenn die Koeffizienten eines solchen Modells schätz- und interpretierbar wären, so ließe sich doch nicht zeigen, daß es mit den ALLBUS-Retest-Daten wesentlich besser verträglich ist als etwa M3. Noch viel weniger ließe sich ausschließen, daß ganz anders spezifizierte Modelle gleich gut oder besser zu den Daten passen. Auf die vieldeutigen Beziehungen zwischen Modellen und Daten hatte ich ja schon in der Einleitung[5] meines Aufsatzes hingewiesen. Sogar die den Modellen M2 und M4 zugrundeliegenden beiden Extrempositionen lassen sich durch Zusatzannahmen etwa über systematische Antworttendenzen bei (einigen) Meinungslosen oder über kompliziertere Formen des Meßfehlers bei (einigen) Wertträgern weitgehend immunisieren. Sieht man einmal von den oben (unter 2.)) beschriebenen Komplikationen ab, so illustrieren die von LANGEHEINE vorgestellten Modelle das Problem beobachtungsäquivalenter oder fast beobachtungsäquivalenter Lösungen auf eindrucksvolle Weise.

Anmerkungen

1 Bei Berechnung des jeweils ersten L^2, das LANGEHEINE für ein Modell berichtet, werden sämtliche 27 Zellen berücksichtigt.

2 Im Unterschied zu LANGEHEINE scheint mir die Annahme stabiler Orientierungen in der ALLBUS-Retest-Studie vergleichsweise unproblematisch, da die erste und dritte Befragung nur ca. acht Wochen auseinanderliegen.

3 Die Vergabe von Prioritäten mag besonders schwer fallen, wenn Zielkonflikte oder Tradeoffs nicht vorhanden sind bzw. nicht erkannt werden: Warum soll man überhaupt Preisstabilität und Meinungsfreiheit rangordnen, wenn die Verwirklichung des einen Ziels die des anderen nicht beeinträchtigt!

4 Zugunsten von Forced-Choice-Items ist oft vorgebracht worden, sie seien nicht so schief verteilt wie Fragen nach der Wichtigkeit von Valenzissues. Diese Verteilungsprobleme entstünden wahrscheinlich auch bei den im Postmaterialismusitem genannten Politikzielen, denn auch hier würden die meisten Befragten die Verwirklichung eines jeden der vier Ziele für 'sehr wichtig[1] oder 'wichtig' erachten. Andererseits wird gerade an der ALLBUS-Retest-Studie ein Nachteil von Forced Choice deutlich: Weil für Befragte, die die verschiedenen Ziele als gleichrangig erachten, eine Cross-Pressure-Situation entsteht, sind die von ihnen genannten Rangordnungen äußerst instabil. Personen unter Cross-Pressure haben genaugenommen keine Meinung darüber, wie die Politikziele rangzuordnen sind.

5 Dort ging es primär um die Frage, ob sich der Status der latenten Variablen eindeutig bestimmen läßt.

Literatur

JAGODZINSKI, W. (1986): "Black & White statt LISREL? Wie groß ist der Anteil der Zufallsantworten beim Postmaterialismusindex?" ZA-Information 19, 30-51.

KOEHLER, K.J. (1986): "Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables." Journal of the American Statistical Association 81, 483-493.

KOEHLER, K.J. und LARNTZ, K. (1980): "An Assessment of Several Asymptotic Approximations for Sparse Multinomials." Journal of the American Statistical Association 75, 336-344.

LANCASTER, H.O. (1969): "The Chi-Squared Distribution." New York: Wiley.

LANGEHEINE, R. (1986): "Log-lineare Modelle." In J. v. KOOLWIJK und M. WIEKEN-MAYSER (Hrsg.), Techniken der empirischen Sozialforschung. Bd. 8: Kausalanalyse. München: Oldenbourg.

LANGEHEINE, R. (1987): "Black & White, anfängliche Antwortunsicherheit, Mover-Stayer, Third Force oder was? Ein paar weitere Überlegungen zu Jagodzinski's Analyse des Postmaterialismus Panels." ZA-Information 20 (der vorstehende Beitrag).

LARNTZ, K. (1978): "Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics." Journal of the American Statistical Association 73, 253-263.

Prof. Dr. Wolfgang Jagodzinski
Universität Bremen
Fachbereich 8
Studiengang Soziologie
2800 Bremen 33