planung &analyse

Zeitschrift für Marktforschung und Marketing www.planung-analyse.de 5/2013 D11700F

Forschung

Interne Faktoren nutzen: Innovationsforschung

Report

Unerlässliches Handwerkszeug: Software

Special

Kundenorientierung vor Ort: Handel

Schwerpunkt

Forschung für die Marke

Statistik KOMPAKT

Kategoriale Variablen in Regressionsmodellen

Regressionsmodelle sind nicht beschränkt auf metrische unabhängige Variablen. Kategoriale Variablen wie Geschlecht, Beruf etc. können Berücksichtigung finden, wenn ihre Ausprägungen als Zahlen dargestellt werden. Eine gängige Vorgehensweise ist die Dummy-Codierung.

Dummy-Codierung unabhängiger dichotomer Variablen

Es soll untersucht werden, welchen Einfluss neben dem Preis das Schalten einer Werbung auf den monatlichen Absatz besitzt. Die lineare Regressionsfunktion ist somit

$$Absatzmenge = b_0 + b_1 \cdot Preis + b_2 \cdot Werbung$$

Während der Preis eine metrische Variable ist, weist die Werbung nur zwei Kategorien auf: es wurde Werbung (zu Beginn eines Monats) geschaltet oder nicht. Um diese Einflussgröße im Regressionsmodell zu berücksichtigen, sind beiden Ausprägungen Zahlen zuzuordnen. Folgt man der Dummy-Codierung, ist einer Referenzkategorie der Wert 0 und der anderen Kategorie der Wert 1 zu geben. In diesem Beispiel bietet es sich an, als Referenzkategorie den Verzicht auf Werbung festzulegen. Der Regressionskoeffizient b_2 gibt dann genau die Menge an, um die sich der Absatz durch das Schalten einer Werbung gegenüber der Referenzkategorie „keine Werbung" bei konstantem Preis verändert.

Dummy-Codierung unabhängiger Variablen mit mehr als zwei Kategorien

Es wird zusätzlich differenziert, ob eine TV- oder Print-Werbung geschaltet wurde. Insofern sind drei Kategorien zu unterscheiden. Damit

	W_1	W_2	W_3
TV-Werbung	1	0	0
Print-Werbung	0	1	0
keine Werbung	0	0	1

Abbildung 1: Dummy-Codierung

bedarf es zur Codierung der zwei Variablen W(erbung)$_1$ und W(erbung)$_2$ (siehe Abbildung 1).

Die Kombination der Variablen mit den Ausprägungen $W_1 = 1$, $W_2 = 0$ repräsentiert somit TV-Werbung, $W_1 = 0$, $W_2 = 1$ Print-Werbung und $W_1 = 0$, $W_2 = 0$ keine Werbung. Durch diese Kombinationen sind alle drei Kategorien eindeutig definiert. Eine dritte Variable W_3 wäre nicht nur redundant, sondern würde zu exakter Multikollinearität führen, so dass das Regressionsmodell nicht schätzbar wäre. „Keine Werbung" ist auch hier die Referenzkategorie, da für diese beide Codiervariablen gleich 0 sind. In der entsprechenden Regressionsfunktion

$$Absatzmenge = b_0 + b_1 \cdot Preis + b_2 \cdot W_1 + b_3 \cdot W_2$$

quantifizieren b_2 und b_3 die Wirkungen der TV- bzw. Print-Werbung auf die Absatzmenge im Vergleich zur Referenzkategorie. Die Differenz zwischen b_2 und b_3 gibt an, um wie viel sich die Wirkung einer Werbung zwischen den beiden Medien unterscheidet.

Interaktionseffekte mit kategorialen Variablen

Die Interpretation der Regressionskoeffizienten geht davon aus, dass keine Mehrfachnennungen für die kategoriale Variable vorliegen. Das heißt, es darf im selben Monat nicht in TV und Print geworben worden sein. Um auch den Effekt gemeinsamer Werbung in beiden Medien zu bestimmen, ist eine eigene zusätzliche Kategorie „TV & Print" zu berücksichtigen (siehe Abbildung 2).

	TV & Print als eigene Kategorie			TV & Print als Interaktionseffekt		
	W_1	W_2	W_3	TV	Print	TV&Print
TV-Werbung	1	0	0	1	0	0
Print-Werbung	0	1	0	0	1	0
TV & Print-Werbung	0	0	1	1	1	1
keine Werbung	0	0	0	0	0	0

Abbildung 2: Dummy-Codierung bei Mehrfachnennungen

Alternativ lassen sich TV-Werbung (ja/nein) und Print-Werbung (ja/nein) als zwei eigenständige dichotome Variablen auffassen. Geht man davon aus, dass gleichzeitige Werbung in beiden Medien nicht additiv wirkt, besteht zwischen diesen ein Interaktionseffekt. Dieser kann durch Aufnahme des Produkts der beiden Variablen im Modell abgebildet werden:

$$Absatzmenge = b_0 + b_1 \cdot Preis + b_2 \cdot TV + b_3 \cdot Print + b_4 \cdot TV \cdot Print$$

Die Wirkung gemeinsamer Werbung ist dann gleich der Summe der Einzeleffekte und des Interaktionseffekts ($b_2 + b_3 + b_4$). Diese entspricht dem Regressionskoeffizienten der Codiervariable W_3 aus Abbildung 2, falls TV & Print-Werbung als eigene Kategorie dargestellt wird.

Ebenso kann ein Interaktionseffekt zwischen Werbung und Preis, das heißt zwischen einer kategorialen und einer metrischen Variable, im Modell berücksichtigt werden. Im einführenden Beispiel der Dummy-Codierung der dichotomen Variable Werbung wird ihr Produkt mit dem Preis in die Regressionsfunktion aufgenommen:

$$Absatzmenge = b_0 + b_1 \cdot Preis + b_2 \cdot Werbung + b_3 \cdot Preis \cdot Werbung$$

Angenommen, der Zusammenhang zwischen Preis und Absatzmenge ist negativ ($b_1 < 0$), so bedeutet ein negativer Koeffizient b_3, dass die Wirkung des Preises auf den Absatz bei Schalten einer Werbung stärker ist als ohne Werbung. Ein positiver Koeffizient deutet dagegen auf geringere Preissensibilität hin.

Neben der Dummy-Codierung sind die Effekt- und die Kontrast-Codierung übliche Vorgehensweisen. Die Art der Codierung beeinflusst zwar die Regressionskoeffizienten und deren Interpretation. Das Bestimmtheitsmaß und damit Ergebnisse der Prüfungen der Signifikanz von Verbesserungen des Bestimmtheitsmaßes infolge der Berücksichtigung weiterer Variablen oder von Interaktionseffekten sind davon jedoch unabhängig. ◄

Johannes Lüken und **Dr. Heiko Schimmelpfennig**, Experten für Mulitvariate Analyse bei IfaD, Institut für angewandte Datenanalyse GmbH.

In Ausgabe 6/2013: Discrete Choice Modelle

► Literatur

Cohen, J.; Cohen, P.; West, S. G.; Aiken, L. S.: Interactions With Categorical Variables, In: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3. Auflage, Mahwah, New Jersey, 2003, S. 354-389.

Eid, M.; Gollwitzer, M.; Schmitt, M.: Multiple Regressionsanalyse, In: Statistik und Forschungsmethoden, 2. Auflage, Weinheim, Basel, 2011, S. 648-677.