Teil II

Grundlegende Verfahren der multivariaten Analyse

2 Daten und Skalen

Das „Rohmaterial" für multivariate Analysen sind die (vorhandenen oder noch zu erhebenden) *Daten*. Die Qualität von Daten wird u.a. bestimmt durch die Art und Weise der *Messung*. Daten sind nämlich das Ergebnis von Messvorgängen. Messen bedeutet, dass Eigenschaften von Objekten nach bestimmten Regeln in Zahlen ausgedrückt werden.

Messung

Im wesentlichen bestimmt die jeweils betrachtete Art einer Eigenschaft, wie gut man ihre Ausprägung messen, d.h. wie gut man sie in Zahlen ausdrücken kann. So wird z.B. die Körpergröße eines Menschen sehr leicht in Zahlen auszudrücken sein, seine Intelligenz, seine Motivation oder sein Gesundheitszustand dagegen sehr schwierig.

Die „Messlatte", auf der die Ausprägungen einer Eigenschaft abgetragen werden, heisst *Skala*. Je nachdem, in welcher Art und Weise eine Eigenschaft eines Objektes in Zahlen ausgedrückt (gemessen) werden kann, unterscheidet man Skalen mit unterschiedlichem *Skalenniveau*:

Skalen(niveau)

1. Nominalskala
2. Ordinalskala
3. Intervallskala
4. Ratioskala.

Das Skalenniveau bedingt sowohl den *Informationsgehalt der Daten* wie auch die *Anwendbarkeit von Rechenoperationen*. Nachfolgend sollen die Skalentypen und ihre Eigenschaften kurz umrissen werden.

Nominalskala

Die *Nominalskala* stellt die primitivste Grundlage des Messens dar. Beispiele für Nominalskalen sind

- Geschlecht (männlich – weiblich)
- Religion (katholisch – evangelisch – andere)
- Farbe (rot – gelb – grün – blau ...)
- Werbemedium (Fernsehen – Zeitungen – Plakattafeln).

Nominalskalen stellen also Klassifizierungen qualitativer Eigenschaftsausprägungen dar. Zwecks leichterer Verarbeitung mit Computern werden die Ausprägungen von Eigenschaften häufig durch Zahlen ausgedrückt. So lassen sich z.B. die Farben einer Verpackung wie folgt kodieren:

rot = 1

gelb = 2

grün = 3

Die Zahlen hätten auch in anderer Weise zugeordnet werden können, solange diese Zuordnung eineindeutig ist, d.h. solange durch eine Zahl genau eine Farbe definiert ist. Mit derartigen Zahlen sind daher keine arithmetischen Operationen (wie Addition, Subtraktion, Multiplikation oder Division) erlaubt. Vielmehr lassen sich lediglich durch Zählen der Merkmalsausprägungen (bzw. der sie repräsentierenden Zahlen) Häufigkeiten ermitteln.

Rechenoperationen

Eine *Ordinalskala* stellt das nächsthöhere Messniveau dar. Die Ordinalskala erlaubt die Aufstellung einer Rangordnung mit Hilfe von Rangwerten (d.h. ordinalen Zahlen). Beispiele: Produkt A wird Produkt B vorgezogen, Herr M. ist tüchtiger als Herr N. Die Untersuchungsobjekte können immer nur in eine Rangordnung gebracht werden. Die Rangwerte 1., 2., 3. etc. sagen nichts über die Abstände zwischen den Objekten aus. Aus der Ordinalskala kann also nicht abgelesen werden, um wieviel das Produkt A besser eingeschätzt wird als das Produkt B. Daher dürfen auch ordinale Daten, ebenso wie nominale Daten, nicht arithmetischen Operationen unterzogen werden. Zulässige statistische Maße sind neben Häufigkeiten z.B. der Median oder Quantile.

Ordinalskala

Das wiederum nächsthöhere Messniveau stellt die *Intervallskala* dar. Diese weist gleichgroße Skalenabschnitte aus. Ein typisches Beispiel ist die Celsius-Skala zur Temperaturmessung, bei der der Abstand zwischen Gefrierpunkt und Siedepunkt des Wassers in hundert gleichgroße Abschnitte eingeteilt wird. Bei intervallskalierten Daten besitzen auch die Differenzen zwischen den Daten Informationsgehalt (z.B. großer oder kleiner Temperaturunterschied), was bei nominalen oder ordinalen Daten nicht der Fall ist.

Intervallskala

Oftmals werden – auch in dem vorliegenden Buch – Skalen benutzt, von denen man lediglich annimmt, sie seien intervallskaliert. Dies ist z.B. der Fall bei Ratingskalen: Eine Auskunftsperson ordnet einer Eigenschaft eines Objektes einen Zahlenwert auf einer Skala von 1 bis 7 (oder einer kürzeren oder längeren Skala) zu. Solange die Annahme gleicher Skalenabstände unbestätigt ist, handelt es sich allerdings strenggenommen um eine Ordinalskala.

Intervallskalierte Daten erlauben die arithmetischen Operationen der Addition und Subtraktion. Zulässige statistische Maße sind, zusätzlich zu den oben genannten, z.B. der Mittelwert (arithmetisches Mittel) und die Standardabweichung, nicht aber die Summe.

Die *Ratio- (oder Verhältnis)skala* stellt das höchste Messniveau dar. Sie unterscheidet sich von der Invervallskala dadurch, dass zusätzlich ein natürlicher Nullpunkt existiert, der sich für das betreffende Merkmal im Sinne von „nicht vorhanden" interpretieren lässt. Das ist z.B. bei der Celsius-Skala oder der Kalenderzeit nicht der Fall, dagegen aber bei den meisten physikalischen Merkmalen (z.B. Länge, Gewicht, Geschwindigkeit) wie auch bei den meisten ökonomischen Merkmalen (z.B. Einkommen, Kosten, Preis). Bei verhältnisskalierten Daten besitzen nicht nur die Differenz, sondern, infolge der Fixierung des Nullpunktes, auch der Quotient bzw. das Verhältnis (Ratio) der Daten Informationsgehalt (daher der Name).

Ratio(Verhältnis)Skala

Ratioskalierte Daten erlauben die Anwendung aller arithmetischen Operationen wie auch die Anwendung aller obigen statistischen Maße. Zusätzlich sind z.B. die Anwendung des geometrischen Mittels oder des Variationskoeffizienten erlaubt.

Nominalskala und Ordinalskala bezeichnet man als nichtmetrische oder auch kategoriale Skalen, Intervallskala und Ratioskala als metrische Skalen.

In Abbildung 1 sind noch einmal die vier Skalenniveaus mit ihren Merkmalen zusammengestellt.

Zusammenfassend lässt sich sagen: Je höher das Skalenniveau ist, desto größer ist auch der Informationsgehalt der betreffenden Daten und desto mehr Rechenoperationen und statistische Maße lassen sich auf die Daten anwenden.

Informationsgehalt

Es ist generell möglich, Daten von einem höheren Skalenniveau auf ein niedrigeres Skalenniveau zu transformieren, nicht aber umgekehrt. Dies kann sinnvoll sein, um die Übersichtlichkeit der Daten zu erhöhen oder um ihre Analyse zu vereinfachen. So werden z.B. häufig Einkommensklassen oder Preisklassen gebildet. Dabei kann es sich

Skala		Merkmale	Mögliche rechnerische Handhabung
nicht-metrische Skalen	NOMINALSKALA	Klassifizierung qualitativer Eigenschaftsausprägungen	Bildung von Häufigkeiten
	ORDINALSKALA	Rangwert mit Ordinalzahlen	Median, Quantile
metrische Skalen	INTERVALLSKALA	Skala mit gleichgroßen Abschnitten ohne natürlichen Nullpunkt	Subtraktion, Mittelwert
	RATIOSKALA	Skala mit gleichgroßen Abschnitten und natürlichem Nullpunkt	Summe, Division, Multiplikation

Abbildung 1: Skalenniveau

um eine Transformation der ursprünglich ratio-skalierten Daten auf eine Intervall-, Ordinal- oder Nominal-Skala handeln. Mit der Transformation auf ein niedrigeres Skalenniveau ist natürlich immer auch ein Informationsverlust verbunden.

3 Einteilung multivariater Analysemethoden

In diesem Buch werden die nachfolgenden grundlegenden Verfahren behandelt:[2]

Grundlegende Verfahren

Teil II: Grundlegende Verfahren der multivariaten Analyse

1. Regressionsanalyse
2. Zeitreihenanalyse
3. Varianzanalyse
4. Diskriminanzanalyse
5. Logistische Regression
6. Kreuztabellierung und Kontingenzanalyse
7. Faktorenanalyse
8. Clusteranalyse
9. Conjoint-Analyse

Komplexe Verfahren

Teil III: Komplexe Verfahren der multivariaten Analyse

10. Nichtlineare Regression
11. Strukturgleichungsmodelle
12. Konfirmatorische Faktorenanalyse

[2]Einen Überblick über multivariate Analysemethoden geben auch die folgenden Bücher: Hair et al. (2006), Herrmann, A./Homburg, C./Klarmann, M. (Hrsg.) (2008), Norusis SPSS Inc. (2008).

13. Neuronale Netze
14. Multidimensionale Skalierung
15. Korrespondenzanalyse
16. Auswahlbasierte Conjoint-Analyse

Die Entwicklung auf dem Gebiet der multivariaten Analyse hat dazu geführt, dass im Laufe der Zeit die Anzahl der in diesem Buch behandelten Methoden angewachsen ist und wohl auch in der Zukunft weiter zunehmen wird. Um dennoch die Verwendung des Buches zu erleichtern und seinen Umfang nicht ausufern zu lassen, haben wir unsere Internetplattform www.multivariate.de mit der 12. Auflage um einen Downloadbereich erweitert. Hier kann der Verwender dieses Buches die Kapitel 10 bis 16 („Komplexe Verfahren der multivariaten Analyse"), die in dem vorliegenden Buch nur in ihren Grundzügen behandelt werden, in der bewährten Form im Detail nachlesen. Der Abruf dieser Kapitel kann mit Hilfe der E-Mailadresse erfolgen und setzt lediglich die Beantwortung einiger kurzer Fragen voraus.

Internetplattform

Unabhängig von der obigen Zweiteilung nehmen wir im Folgenden eine Einordnung der multivariaten Analysemethoden nach ihrem Anwendungsbezug vor. Dabei sei jedoch betont, dass eine *überschneidungsfreie Zuordnung* der Verfahren zu praktischen Fragestellungen nicht immer möglich ist, da sich die Zielsetzungen der Verfahren z.T. überlagern. Versucht man jedoch eine Einordnung der Verfahren nach anwendungsbezogenen Fragestellungen, so bietet sich eine Einteilung in primär *strukturen-entdeckende Verfahren* und primär *strukturen-prüfende Verfahren* an. Diese beiden Kriterien werden in diesem Zusammenhang wie folgt verstanden:

Strukturen-prüfende Verfahren

1. *Strukturen-prüfende Verfahren* sind solche multivariaten Verfahren, deren primäres Ziel in der *Überprüfung von Zusammenhängen* zwischen Variablen liegt. Der Anwender besitzt eine auf sachlogischen oder theoretischen Überlegungen basierende Vorstellung über die Zusammenhänge zwischen Variablen und möchte diese mit Hilfe multivariater Verfahren überprüfen.

   Verfahren, die diesem Bereich der multivariaten Analyse zugeordnet werden können, sind die lineare und nichtlineare Regressionsanalyse, die Zeitreihenanalyse, die Varianzanalyse, die Diskriminanzanalyse, die Kontingenzanalyse sowie die Logistische Regression, Strukturgleichungsmodelle und die Conjoint-Analyse zur Analyse von Präferenzstrukturen.

Strukturen-entdeckende Verfahren

2. *Strukturen-entdeckende Verfahren* sind solche multivariaten Verfahren, deren Ziel in der *Entdeckung von Zusammenhängen* zwischen Variablen oder zwischen Objekten liegt. Der Anwender besitzt zu Beginn der Analyse noch keine Vorstellungen darüber, welche Beziehungszusammenhänge in einem Datensatz existieren.

   Verfahren, die primär eingesetzt werden, um mögliche Beziehungszusammenhänge aufzudecken, sind die Faktorenanalyse, die Clusteranalyse, die Multidimensionale Skalierung, die Korrespondenzanalyse und die Neuronalen Netze.

3.1 Strukturen-prüfende Verfahren

Die strukturen-prüfenden Verfahren werden primär zur Durchführung von *Kausalanalysen* eingesetzt, um herauszufinden, ob und wie stark sich z.B. das Wetter, die Bodenbeschaffenheit sowie unterschiedliche Düngemittel und -mengen auf den Ernteertrag auswirken oder wie stark die Nachfrage eines Produktes von dessen Qualität, dem Preis, der Werbung und dem Einkommen der Konsumenten abhängt.

Kausalanalysen

Voraussetzung für die Anwendung der entsprechenden Verfahren ist, dass der Anwender *a priori (vorab)* eine sachlogisch möglichst gut fundierte Vorstellung über den Kausalzusammenhang zwischen den Variablen entwickelt hat, d.h. er weiß bereits oder vermutet, welche der Variablen auf andere Variablen einwirken. Zur Überprüfung seiner (theoretischen) Vorstellungen werden die von ihm betrachteten Variablen i.d.R. in *abhängige* und *unabhängige* Variablen eingeteilt und dann mit Hilfe von multivariaten Analysemethoden an den empirisch erhobenen Daten überprüft. Nach dem Skalenniveau der Variablen lassen sich die grundlegenden strukturen-prüfenden Verfahren gemäß Abbildung 2 charakterisieren.

Hypothesen

		UNABHÄNGIGE VARIABLE	
		metrisches Skalenniveau	nominales Skalenniveau
ABHÄNGIGE VARIABLE	metrisches Skalennivau	Regressionsanalyse, Zeitreihenanalyse	Varianzanalyse, Regression mit Dummies
	nominales Skalenniveau	Diskriminanzanalyse, Logistische Regression	Kontingenzanalyse Auswahlbasierte Conjoint-Analyse

Abbildung 2: Grundlegende strukturen-prüfende Verfahren

Regressionsanalyse

Die Regressionsanalyse ist ein außerordentlich vielseitiges und flexibles Analyseverfahren, das sowohl für die *Beschreibung* und *Erklärung von Zusammenhängen* als auch für die *Durchführung von Prognosen* große Bedeutung besitzt. Sie ist damit sicherlich das wichtigste und am häufigsten angewendete multivariate Analyseverfahren. Insbesondere kommt sie in Fällen zur Anwendung, wenn Wirkungsbeziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen untersucht werden sollen. Mit Hilfe der Regressionsanalyse können derartige Beziehungen quantifiziert und damit weitgehend exakt beschrieben werden. Außerdem lassen sich mit ihrer Hilfe Hypothesen über Wirkungsbeziehungen prüfen und auch Prognosen erstellen.

Erklärung und Prognose

Ein Beispiel bildet die Frage, ob und wie die Absatzmenge eines Produktes vom Preis, den Werbeausgaben, der Zahl der Verkaufsstätten und dem Volkseinkommen abhängt. Sind diese Zusammenhänge mit Hilfe der Regressionsanalyse quantifiziert und empirisch bestätigt worden, so lassen sich Prognosen (What-if-Analysen) erstellen, die beantworten, wie sich die Absatzmenge verändern wird, wenn z.B. der Preis oder die Werbeausgaben oder auch beide Variablen zusammen verändert werden.

Beispiel

Die Regressionsanalyse ist prinzipiell anwendbar, wenn sowohl die abhängige als auch die unabhängigen Variablen metrisches Skalenniveau besitzen. Dies ist der klassische Fall, wobei die Beziehungen zwischen unabhängigen und abhängigen Variablen

Dummy Variablen

auch nicht linear sein können. Durch Anwendung der sog. *Dummy-Variablen-Technik* lassen sich aber auch qualitative (nominal skalierte) Variable in die Regressionsanalyse einbeziehen und deren Anwendungsbereich somit ausweiten. Dummy-Variable sind binäre Variable, die nur die Werte 0 oder 1 annehmen. Stellen wir uns vor, es sollen die Einflüsse verschiedener Produkteigenschaften auf das Kaufverhalten von Konsumenten untersucht werden. Die Dummy-Variable q_1 würde dann in allen Fällen, bei denen das Produkt eine rote Verpackung hat, den Wert 1 annehmen, und wenn dies nicht der Fall ist, den Wert 0.

$$q_1 = \begin{cases} 1 & \textit{falls} \quad \textit{Farbe} = \textit{rot} \\ 0 & \textit{sonst} \end{cases}$$

In analoger Weise lassen sich auch eine Dummy-Variable q_2 für die Farbe Gelb und eine Dummy-Variable q_3 für die Farbe Grün definieren. Wenn allerdings nur Verpackungen in den drei Farben Rot, Gelb und Grün vorkommen, so wäre eine der drei Dummies überflüssig. Denn wenn $q_1 = 0$ und $q_2 = 0$ gilt, so muss zwangsläufig $q_3 = 1$ gelten. Die drei Farben lassen sich also eindeutig mittels der zwei Dummies (q_1, q_2) beschreiben: rot = (1, 0), gelb = (0, 1), grün = (0, 0). Generell gilt, dass sich eine nominale Variable mit n Ausprägungen durch $n - 1$ Dummy-Variablen ersetzen lässt.

Die Bedeutung von Dummy-Variablen liegt darin, dass sie sich wie metrische Variable behandeln lassen. Somit lassen sich mit ihrer Hilfe auch nominal skalierte Variable in eine Regressionsanalyse einbeziehen. Dies gilt aber generell nur für die unabhängigen Variablen und nicht für die abhängige Variable. Nachteilig ist, dass sich dadurch u.U. die Zahl der Variablen und der damit verbundene Kodierungs- und Rechenaufwand stark erhöht. Deshalb kann in solchen Fällen die Anwendung einer Varianzanalyse einfacher und übersichtlicher sein.

Zeitreihenanalyse

Prognose

Die Zeitreihenanalyse dient neben der Beschreibung und Erklärung der zeitlichen Entwicklung einer Variablen insbesondere auch deren Prognose, d.h. der Schätzung von Werten dieser Variablen für zukünftige Zeitpunkte oder Perioden. Jede weitreichende Entscheidung basiert auf Prognosen. Die Zeitreihenanalyse ist daher für die Stützung von Entscheidungsproblemen jeglicher Art von großer Wichtigkeit. Für die Produktions- und Absatzplanung eines Herstellers ist z.B. von Wichtigkeit, wie sich seine Absatzmenge oder das Volumen seines Marktes langfristig entwickeln werden oder welchen periodischen Schwankungen diese Größen unterworfen sind. Neben anderen Verfahren bildet insbesondere die zuvor behandelte Regressionsanalyse ein wichtiges Instrument zur Durchführung von Zeitreihenanalysen. Sie ermöglicht die Erstellung von Punktprognosen sowie auch die Berechnung von Prognosefehlern und von Prognoseintervallen, innerhalb derer das vorhergesagte Ereignis mit einer festgelegten Wahrscheinlichkeit liegen wird.

Varianzanalyse

Experimente

Werden die unabhängigen Variablen auf nominalem Skalenniveau gemessen und die abhängigen Variablen auf metrischem Skalenniveau, so findet die Varianzanalyse Anwendung. Dieses Verfahren besitzt besondere Bedeutung für die *Analyse von Experimenten*, wobei die nominalen unabhängigen Variablen die experimentellen Einwirkungen repräsentieren. So kann z.B. in einem Experiment untersucht werden, welche

Wirkung alternative Verpackungen eines Produktes oder dessen Plazierung im Geschäft auf die Absatzmenge haben.

Diskriminanzanalyse

Gruppenunterschiede

Klassifizierung

Ist die abhängige Variable nominal skaliert, und besitzen die unabhängigen Variablen metrisches Skalenniveau, so findet die Diskriminanzanalyse Anwendung. Die Diskriminanzanalyse ist ein Verfahren zur *Analyse von Gruppenunterschieden.* Ein Beispiel bildet die Frage, ob und wie sich die Wähler der verschiedenen Parteien hinsichtlich soziodemografischer und psychografischer Merkmale unterscheiden. Die abhängige nominale Variable identifiziert die Gruppenzugehörigkeit, hier die gewählte Partei, und die unabhängigen Variablen beschreiben die Gruppenelemente, hier die Wähler.

Ein weiteres Anwendungsgebiet der Diskriminanzanalyse bildet die *Klassifizierung von Elementen.* Nachdem für eine gegebene Menge von Elementen die Zusammenhänge zwischen der Gruppenzugehörigkeit der Elemente und ihren Merkmalen analysiert wurden, lässt sich darauf aufbauend eine Prognose der Gruppenzugehörigkeit von neuen Elementen vornehmen. Derartige Anwendungen finden sich z.B. bei der Kreditwürdigkeitsprüfung (Einstufung von Kreditkunden einer Bank in Risikoklassen) oder bei der Personalbeurteilung (Einstufung von Außendienstmitarbeitern nach erwartetem Verkaufserfolg).

Logistische Regression

Gruppenzugehörigkeit

Ganz ähnliche Fragestellungen, wie mit der Diskriminanzanalyse können auch mit dem Verfahren der logistischen Regression untersucht werden. Hier wird die *Wahrscheinlichkeit* der Zugehörigkeit zu einer Gruppe (einer Kategorie der abhängigen Variablen) in Abhängigkeit von einer oder mehrerer unabhängiger Variablen bestimmt. Dabei können die unabhängigen Variablen sowohl nominales als auch metrisches Skalenniveau aufweisen. Über die Analyse der Gruppenunterschiede hinaus kann z.B. auch das Herzinfarktrisiko von Patienten in Abhängigkeit von ihrem Alter und ihrem Cholesterin-Spiegel ermittelt werden. Da zur Schätzung der Eintrittswahrscheinlichkeiten der Kategorien der abhängigen Variabeln auf die (s-förmige) logistische Funktion zurückgegriffen wird, gehört dieses Verfahren zu den *nicht-linearen Analyseverfahren.*

Kreuztabellierung und Kontingenzanalyse

Kreuztabelle

Eine weitere Methodengruppe, die der Analyse von Beziehungen zwischen ausschließlich nominalen Variablen dient, wird als Kontingenzanalyse bezeichnet. Hier kann es z.B. darum gehen, die Frage nach dem Zusammenhang zwischen Rauchen (Raucher versus Nichtraucher) und Lungenerkrankung (ja, nein) statistisch zu überprüfen. Die Überprüfung erfolgt dabei auf der Basis von in Form einer Kreuztabelle (Kontingenztabelle) angeordneten Daten. Mit Hilfe weiterführender Verfahren, wie der sog. Logit-Analyse, lässt sich weiterhin auch die Abhängigkeit einer nominalen Variablen von mehreren nominalen Einflussgrößen untersuchen (vgl. hierzu auch das Verfahren der logistischen Regression).

Conjoint-Analyse

Bei den bisher aufgezeigten Verfahren wurde nur zwischen metrischem und nominalem Skalenniveau der Variablen unterschieden. Ein Verfahren, bei dem die abhängige Variable häufig auf ordinalem Skalenniveau gemessen wird, ist die Conjoint-Analyse. Insbesondere lassen sich mit Hilfe der Conjoint-Analyse ordinal gemessene Präferenzen und auch Auswahlentscheidungen (auswahlbasierte CA) analysieren. Ziel ist es dabei, den *Beitrag einzelner Merkmale* von Produkten oder sonstigen Objekten *zum Gesamtnutzen* bzw. zur Kaufentscheidung bzgl. dieser Objekte herauszufinden. Einen Anwendungsbereich bildet die Gestaltung neuer Produkte. Dazu ist es von Wichtigkeit, den Einfluss oder Beitrag alternativer Produktmerkmale, z.B. alternativer Materialien, Formen, Farben oder Preisstufen, auf die Nutzenbeurteilung zu kennen.

Analyse von Präferenzen

Bei der Conjoint-Analyse muss der Forscher vorab festlegen, welche Merkmale in welchen Ausprägungen berücksichtigt werden sollen. Hierauf basierend wird sodann ein Erhebungsdesign gebildet, im Rahmen dessen Präferenzen, z.B. bei potenziellen Käufern eines neuen Produktes, gemessen werden. Auf Basis dieser Daten erfolgt dann die Analyse zur Ermittlung der Nutzenbeiträge der berücksichtigten Merkmale und ihrer Ausprägungen. Die Conjoint-Analyse bildet damit also eine *Kombination aus Erhebungs- und Analyseverfahren.*

Nichtlineare Regression

Durch die Nichtlineare Regression wird das Anwendungsspektrum der Regressionsanalyse erheblich erweitert. Es lassen sich nahezu beliebige Modellstrukturen schätzen. Anwendungen finden sich z.B. im Rahmen der Werbewirkungsforschung (Abhängigkeit der Werbeerinnerung von der Zahl der Werbekontakte, Abhängigkeit der Absatzmenge von der Höhe des Werbebudgets) oder in der Marktforschung bei der Untersuchung des Wachstums von neuen Produkten. Die Nichtlineare Regression ist allerdings mit einer Reihe von Schwierigkeiten verbunden. Der Rechenaufwand ist um ein Vielfaches größer als bei der traditionellen Regressionsanalyse, da iterative Algorithmen für die Berechnung der Schätzwerte verwendet werden müssen. Ob diese Algorithmen konvergieren, hängt u.a. davon ab, welche Startwerte der Untersucher vorgibt. Es werden somit auch erhöhte Anforderungen an den Untersucher gestellt. Ein weiterer Nachteil ist, dass die statistischen Tests, die bei der linearen Regressionsanalyse zur Prüfung der Güte des Modells oder der Signifikanz der Parameter verwendet werden, für die nichtlineare Regression nicht anwendbar sind. Der Untersucher sollte daher, wenn möglich, der linearen Regressionsanalyse den Vorzug geben, Wie gezeigt werden wird, lassen sich auch mit Hilfe der linearen Regressionsanalyse vielfältige nichtlineare Problemstellungen behandeln.

Werbewirkungsforschung

Wachstumsmodelle

Strukturgleichungsmodelle und Konfirmatorische Faktorenanalyse

Die bisher betrachteten Analysemethoden gehen davon aus, dass alle Variablen in der Realität beobachtbar und gegebenenfalls auch messbar sind. Bei vielen theoriegestützten Fragestellungen hat man es aber auch mit nicht beobachtbaren Variablen zu tun, sog. *hypothetischen Konstrukten* oder *latenten Variablen.* Beispiele hierfür sind psychologische Konstrukte wie Einstellung und Motivation oder soziologische Konstrukte wie Kultur und soziale Schicht. In solchen Fällen kann die Analyse von Strukturgleichungen zur Anwendung kommen.

Latente Variable

Zur Behandlung von Strukturgleichungsmodellen wird in diesem Buch auf das Programmpaket AMOS (Analysis of Moment Structures) zurückgegriffen, das Datenmatrizen aus SPSS analysieren und Ergebnisse mit SPSS austauschen kann.[3] Mit Hilfe von AMOS lassen sich komplexe Kausalstrukturen überprüfen. Insbesondere können Beziehungen mit mehreren abhängigen Variablen, mehrstufigen Kausalbeziehungen und mit nicht beobachtbaren (latenten) Variablen überprüft werden. Der Benutzer muss, wenn er latente Variable in die Betrachtungen einbeziehen will, zwei Modelle spezifizieren:

AMOS

Messmodell

- Das *Messmodell*, das die Beziehungen zwischen den latenten Variablen und geeigneten Indikatoren vorgibt, mittels derer sich die latenten Variablen indirekt messen lassen. Die empirische Überprüfung erfolgt mit Hilfe der Konfirmatorischen Faktorenanalyse (KFA).

KFA

- Das *Strukturmodell*, welches die Kausalbeziehungen zwischen den latenten Variablen vorgibt, die letztlich dann zu überprüfen sind.

Strukturmodell

Die Variablen des Strukturmodells können alle latent sein, müssen es aber nicht. Ein Beispiel, bei dem nur die unabhängigen Variablen latent sind, wäre die Abhängigkeit der Absatzmenge von der subjektiven Produktqualität und Servicequalität eines Anbieters.

Auswahlbasierte Conjoint-Analyse

CBC

Kaufsimulation

Während bei der traditionellen Conjoint-Analyse zwecks Analyse von Nutzenstrukturen die Präferenzen von Probanden bezüglich alternativer Objekte (Stimuli) auf ordinalem Skalenniveau gemessen werden (mittels Ranking- oder Ratingskalen), erfolgt bei der Auswahlbasierten Conjoint-Analyse (Choice-Based Conjoint) nur eine Abfrage von Auswahlentscheidungen. Aus einer Menge von Alternativen (Choice Set) muss der Proband nur jeweils die am meisten präferierte Alternative auswählen, wobei meist auch die Option besteht, keine der Alternativen zu wählen. Dies ist für ihn nicht nur einfacher, sondern kommt auch seinem realen Entscheidungsverhalten (z.B. in Kaufsituationen) sehr viel näher, als das Ranking oder Rating aller Alternativen im Choice Set. Die erhöhte Realitätsnähe wird allerdings mit einem Verlust an Information erkauft, da bei dieser Vorgehensweise die Präferenz nur noch auf nominalem Skalenniveau gemessen wird. Zur Schätzung der Nutzenbeiträge einzelner Merkmale (Teilnutzenwerte) muss daher eine anderes Schätzverfahren verwendet werden. Während bei der traditionellen Conjoint-Analyse die Schätzung meist durch Regression mit Dummy-Variablen erfolgt, kommt bei der Auswahlbasierten Conjoint-Analyse die Maximum-Likelihood-Methode zur Anwendung. Dabei wird dem Verhalten der Probanden ein probabilistisches Entscheidungsmodell zugrundegelegt. Wegen des geringeren Informationsgehalts ist es meist nur möglich, die Teilnutzenwerte aggregiert zu schätzen, während es bei der traditionellen Conjoint-Analyse üblich ist, sie individuell für jeden Probanden zu schätzen.

[3] Bis zur 9. Auflage wurde bei der Behandlung von Strukturgleichungsmodellen auf das Programm LISREL (LInear Structural RELationships) zurückgegriffen.

3.2 Strukturen-entdeckende Verfahren

Die hier den strukturen-entdeckenden Verfahren zugeordneten Analysemethoden werden primär zur *Entdeckung von Zusammenhängen* zwischen Variablen oder zwischen Objekten eingesetzt. Es erfolgt daher vorab durch den Anwender *keine* Zweiteilung der Variablen in abhängige und unabhängige Variablen, wie es bei den strukturen-prüfenden Verfahren der Fall ist.

Faktorenanalyse

Variablenbündelung

Die Faktorenanalyse findet insbesondere dann Anwendung, wenn im Rahmen einer Erhebung eine Vielzahl von Variablen zu einer bestimmten Fragestellung erhoben wurde, und der Anwender nun an einer Reduktion bzw. *Bündelung der Variablen* interessiert ist. Von Bedeutung ist die Frage, ob sich möglicherweise sehr zahlreiche Merkmale, die zu einem bestimmten Sachverhalt erhoben wurden, auf einige wenige „zentrale Faktoren" zurückführen lassen. Ein einfaches Beispiel hierzu bildet die Verdichtung der zahlreichen technischen Eigenschaften von Kraftfahrzeugen auf wenige Dimensionen, wie Größe, Leistung und Sicherheit.

Positionierung

Einen wichtigen Anwendungsbereich der Faktorenanalyse bilden *Positionierungsanalysen.* Dabei werden die subjektiven Eigenschaftsbeurteilungen von Objekten (z.B. Produktmarken, Unternehmen oder Politiker) mit Hilfe der Faktorenanalyse auf zugrundeliegende Beurteilungsdimensionen verdichtet. Ist eine Verdichtung auf zwei oder drei Dimensionen möglich, so lassen sich die Objekte im Raum dieser Dimensionen grafisch darstellen. Im Unterschied zu anderen Formen der Positionierungsanalyse spricht man hier von faktorieller Positionierung.

Clusteranalyse

Objektbündelung

Während die Faktorenanalyse eine Verdichtung oder Bündelung von Variablen vornimmt, wird mit der Clusteranalyse eine *Bündelung von Objekten* angestrebt. Das Ziel ist dabei, die Objekte so zu Gruppen (Clustern) zusammenzufassen, dass die Objekte in einer Gruppe möglichst ähnlich und die Gruppen untereinander möglichst unähnlich sind. Beispiele sind die Bildung von Persönlichkeitstypen auf Basis der psychografischen Merkmale von Personen oder die Bildung von Marktsegmenten auf Basis nachfragerelevanter Merkmale von Käufern.

Überprüfung

Zur Überprüfung der Ergebnisse einer Clusteranalyse kann die Diskriminanzanalyse herangezogen werden. Dabei wird untersucht, inwieweit bestimmte Variablen zur Unterscheidung zwischen den Gruppen, die mittels Clusteranalyse gefunden wurden, beitragen bzw. diese erklären.

Neuronale Netze

Substitut

Neuronale Netze werden heute in der Praxis in zunehmendem Maße sowohl ergänzend zu den klassischen multivariaten Methoden eingesetzt, als auch in den Fällen, in den die klassischen Methoden versagen. Anwendungsgebiete sind Klassifikationen von Objekten, Prognosen von Zuständen oder Probleme der Gruppenbildung. Insofern bestehen hinsichtlich der Aufgabenstellungen Ähnlichkeiten zur Diskriminanzanalyse und zur Clusteranalyse. Die Methodik neuronaler Netze lehnt sich an biologische Informationsverarbeitungsprozesse im Gehirn an (daher der Name). Es werden künstliche neuronale Netze gebildet, die in der Lage sind, selbständig aus Erfahrung

Gehirn

zu lernen. Insbesondere vermögen sie, komplexe Muster in vorhandenen Daten (z.B. Finanzdaten, Verkaufsdaten) zu erkennen und eröffnen so eine sehr einfache Form der Datenanalyse. Besonders vorteilhaft lassen sie sich zur Behandlung von schlecht strukturierten Problemstellungen einsetzen.

Nicht-Linearität

Innerhalb neuronaler Netze werden künstliche Neuronen (Nervenzellen) als Grundelemente der Informationsverarbeitung in Schichten organisiert, wobei jedes Neuron mit denen der nachgelagerten Schicht verbunden ist. Dadurch lassen sich auch hochgradig nicht-lineare und komplexe Zusammenhänge ohne spezifisches Vorwissen über die etwaige Richtung und das Ausmaß der Wirkungsbeziehungen zwischen einer Vielzahl von Variablen modellieren.

Überwachtes vs. unüberwachtes Lernen

Zum Erlernen von Strukturen wird das Netz zunächst in einer sog. *Trainingsphase* mit beobachteten Daten „gefüttert". Dabei wird unterschieden zwischen Lernprozessen, bei denen die richtigen Ergebnisse bekannt sind und diese durch das Netz reproduziert werden sollen (*überwachtes Lernen*), und solchen, bei denen die richtigen Ergebnisse nicht bekannt sind und lediglich ein konsistentes Verarbeitungsmuster erzeugt werden soll (*unüberwachtes Lernen*). Nach der Trainingsphase ist das Netz konfiguriert und kann für die Analyse neuer Daten eingesetzt werden.

Multidimensionale Skalierung

Positionierung

Den Hauptanwendungsbereich der Multidimensionalen Skalierung (MDS) bilden Positionierungsanalysen, d.h. die *Positionierung von Objekten im Wahrnehmungsraum* von Personen. Sie bildet somit eine Alternative zur faktoriellen Positionierung mit Hilfe der Faktorenanalyse.

Ähnlichkeiten

Im Unterschied zur faktoriellen Positionierung werden bei Anwendung der MDS nicht die subjektiven Beurteilungen von Eigenschaften der untersuchten Objekte erhoben, sondern es werden nur wahrgenommene globale Ähnlichkeiten zwischen den Objekten erfragt. Mittels der MDS werden die diesen Ähnlichkeiten zugrundeliegenden Wahrnehmungsdimensionen abgeleitet. Wie schon bei der faktoriellen Positionierung lassen sich sodann die Objekte im Raum dieser Dimensionen positionieren und grafisch darstellen. Die MDS findet insbesondere dann Anwendung, wenn der Forscher keine oder nur vage Kenntnisse darüber hat, welche Eigenschaften für die subjektive Beurteilung von Objekten (z.B. Produktmarken, Unternehmen oder Politiker) von Relevanz sind.

Beziehung zu anderen Verfahren

Zwischen der Multidimensionalen Skalierung und der Conjoint-Analyse besteht sowohl inhaltlich wie auch methodisch eine enge Beziehung, obgleich wir sie hier unterschiedlich zum einen den strukturen-entdeckenden und zum anderen den strukturenprüfenden Verfahren zugeordnet haben. Beide Verfahren befassen sich mit der Analyse psychischer Sachverhalte und bei beiden Verfahren können auch ordinale Daten analysiert werden, weshalb sie z.T. auch identische Algorithmen verwenden. Ein gewichtiger Unterschied besteht dagegen darin, dass der Forscher bei Anwendung der Conjoint-Analyse bestimmte Merkmale auszuwählen hat.

Korrespondenzanalyse

Visualisierung

Die Korrespondenzanalyse dient, wie auch die Faktorenanalyse und die Multidimensionale Skalierung (MDS), zur Visualisierung komplexer Daten. Sie wird daher in der Marktforschung ebenfalls zur Durchführung von Positionierungsanalysen verwendet. Insbesondere kann sie als ein Verfahren der multidimensionalen Skalierung von nominal skalierten Variablen charakterisiert werden. Sie ermöglicht es, die Zeilen und

Spalten einer zweidimensionalen Kreuztabelle (Kontingenztabelle) grafisch in einem gemeinsamen Raum darzustellen.

Beispiel

Beispiel: Gegeben sei eine Häufigkeitstabelle, deren Zeilen Automarken betreffen und in deren Spalten wünschenswerte Merkmale von Autos (z.B. hohe Sicherheit, schönes Design) stehen. Die Zellen der Matrix sollen beinhalten, mit welcher Häufigkeit ein bestimmtes qualitatives Merkmal den verschiedenen Automarken im Rahmen einer Käuferbefragung zugeordnet wurde. Marken und Merkmale lassen sich sodann mit Hilfe der Korrespondenzanalyse in einem gemeinsamen Raum als Punkte darstellen. Dadurch lässt sich dann erkennen, wie die Automarken relativ zueinander und in Bezug auf die Merkmale von den Käufern beurteilt werden. Für die Korrespondenzanalyse spielt es dabei *keine* Rolle (im Unterschied zur Faktorenanalyse), welche Elemente in den Zeilen und welche in den Spalten angeordnet werden.

Vorteil

Ein besonderer Vorteil der Korrespondenzanalyse liegt darin, dass sie kaum Ansprüche an das Skalenniveau der Daten stellt. Die Daten müssen lediglich nichtnegativ sein. Die Korrespondenzanalyse kann daher auch zur Quantifizierung qualitativer Daten verwendet werden. Da sich qualitative Daten leichter erheben lassen als quantitative Daten, kommt diesem Verfahren eine erhebliche praktische Bedeutung zu.

3.3 Zusammenfassende Betrachtung

Fehlinterpretation

Die vorgenommene Zweiteilung der multivariaten Verfahren in strukturen-prüfende und strukturen-entdeckende Verfahren kann keinen Anspruch auf Allgemeingültigkeit erheben, sondern kennzeichnet nur den vorwiegenden Einsatzbereich der Verfahren. So kann und wird auch die Faktorenanalyse zur Überprüfung von hypothetisch gebildeten Strukturen eingesetzt, und viel zu häufig werden in der empirischen Praxis auch Regressions- und Diskriminanzanalyse im heuristischen Sinne zur Auffindung von Kausalstrukturen verwendet. Diese Vorgehensweise wird nicht zuletzt auch durch die Verfügbarkeit leistungsfähiger Rechner und Programme unterstützt. Der gedankenlose Einsatz von multivariaten Verfahren kann leicht zu einer Quelle von Fehlinterpretationen werden, da ein statistisch signifikanter Zusammenhang keine hinreichende Bedingung für das Vorliegen eines kausal bedingten Zusammenhangs bildet. („Erst denken, dann rechnen!") Es sei daher generell empfohlen, die strukturen-prüfenden Verfahren auch in diesem Sinne, d. h. zur empirischen Überprüfung von theoretisch oder sachlogisch begründeten Hypothesen, einzusetzen. In Abbildung 3 sind die oben skizzierten multivariaten Verfahren noch einmal mit jeweils einem Anwendungsbeispiel zusammengefasst.