Dennis Klinkhammer
Alexander Spermann

Einführung in die empirische Kausalanalyse und Machine Learning mit R

utb 5110

Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag · Wien · Köln · Weimar
Verlag Barbara Budrich · Opladen · Toronto
facultas · Wien
Wilhelm Fink · Paderborn
Narr Francke Attempto Verlag / expert verlag · Tübingen
Haupt Verlag · Bern
Verlag Julius Klinkhardt · Bad Heilbrunn
Mohr Siebeck · Tübingen
Ernst Reinhardt Verlag · München
Ferdinand Schöningh · Paderborn
transcript Verlag · Bielefeld
Eugen Ulmer Verlag · Stuttgart
UVK Verlag · München
Vandenhoeck & Ruprecht · Göttingen
Waxmann · Münster · New York
wbv Publikation · Bielefeld

Dennis Klinkhammer, Alexander Spermann

Einführung in die empirische Kausal-analyse und Machine Learning mit R

wbv Media GmbH & Co. KG · Bielefeld

Gesamtherstellung:
wbv Media, Bielefeld
wbv.de

Einbandgestaltung:
Atelier Reichert, Stuttgart

Bestellnummer: utb 5510

ISBN: 978-3-8252-5510-7
e-ISBN: 978-3-8385-5510-2

Online-Angebote oder elektronische Ausgaben sind erhältlich unter www.utb-shop.de

Printed in Germany

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie;
detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Inhalt

"While the individual man is an insoluble puzzle,
in the aggregate he becomes a mathematical certainty.
You can, for example, never foretell what any one man will do,
but you can say with precision what an average number will be up to.
Individuals vary, but the percentages remain constant.
So says the statistician"

Sherlock Holmes

Schnelleinstieg in R

(1) Installation der Basisversion
Die kostenfreie Basisinstallation von R kann für verschiedene Betriebssysteme unter folgenden Links bezogen werden:

Mac: https://cran.r-project.org/bin/macosx/

Linux: https://cran.r-project.org/bin/linux/

Windows: https://cran.r-project.org/bin/windows/base/

(2) Installation einer grafischen Benutzeroberfläche
Da es sich bei R um eine Programmiersprache handelt, wird mit der Basisinstallation primär eine Konsole zur Eingabe von Befehlen installiert. Diese ist für die ausgewiesenen Programmierbeispiele ausreichend. Bei Bedarf kann aber zusätzlich eine grafische Benutzeroberfläche installiert werden:

RStudio: https://rstudio.com/products/rstudio/

(3) R Troubleshooting
Alle Befehle der ausgewiesenen Programmierbeispiele sind entsprechend ausgewiesen und stehen als Programmierbeispiele online zur Verfügung (Teil 5: Weitere Materialien). Oftmals sind Fehlermeldungen auf „Vertipper“ durch die Nutzerinnen und Nutzer zurückzuführen. In seltenen Fällen sind Fehlermeldungen auf zugrunde liegende Inkompatibilitäten zurückzuführen. Sollte weitere Hilfe erforderlich sein, so bietet R folgende Supportseite an:

Weitere Hilfestellungen: https://www.r-project.org/help.html/

(4) Programmierbeispiele auf GitHub
Ausgewiesene Befehle sind als Programmierbeispiele entsprechenden Lernzielen zugeordnet und können kostenfrei auf einem GitHub Repository eingesehen, kopiert und in die R-Konsole übertragen werden:

GitHub Repository: https://github.com/statistical-thinking/econometrics/

(5) Lernvideos zum Schnelleinstieg in R
Für den schnellen Einstieg in R stehen zwei Lernvideos zum allgemeinen Prinzip der Programmiersprache R, aber auch zur Erweiterung des Funktionsumfanges über sogenannte Packages zur Verfügung:

Programmiersprache R: https://youtu.be/JuDKMGj8zR4/

Packages in R: https://youtu.be/aaq6GPANcdY/

Teil 1: Grundlagen der Forschungsmethoden

1 Einführung in die Forschungsmethoden

In diesem Kapitel lernen Sie...

...die Gemeinsamkeiten zwischen der **Denkweise von Sherlock Holmes** und der formalen Herangehensweise in der Statistik kennen

...einfache **formale Beziehungen statistischer Variablen** zu formulieren

Der Beginn eines Hochschulstudiums markiert für viele Studierende die erste bewusste Auseinandersetzung mit *Forschung* und den dazugehörigen Forschungsmethoden, obschon die zugrunde liegenden Prinzipien des forschenden Blickes durchaus auch im Alltag und somit außerhalb des Hochschulstudiums zur Anwendung kommen können. Einer der wesentlichen Unterschiede zwischen Forschung im Hochschulstudium und wissenschaftlichen Kontext sowie „Forschung" im Alltag besteht in der Systematisierung des Vorgehens hinsichtlich der Suche nach neuen Erkenntnissen, deren Dokumentation und Veröffentlichung. Insbesondere die Dokumentation und Veröffentlichung ermöglichen im Rahmen des Hochschulstudiums und des wissenschaftlichen Kontextes einen Diskurs, wie er im Alltag in der Regel nicht geführt werden kann, da sowohl die neuen Erkenntnisse als auch die forschungsmethodische Herangehensweise zu diesen Erkenntnissen über die Dokumentation und Veröffentlichung diskutiert werden können.

Dass die systematische Suche nach neuen Erkenntnissen im Zusammenspiel mit einer geeigneten Dokumentation und Veröffentlichung auch im Alltag von großer Bedeutung sein kann, verdeutlicht ein Blick auf die Werke von Arthur Conan Doyle und dessen Romanfigur *Sherlock Holmes*. In Chapter Two des Romans „A Study in Scarlet" werden unter anderem die forschungsmethodische Herangehensweise des berühmten Detektivs und die dabei oftmals zugrunde liegenden und zu fokussierenden *Zusammenhänge* des Alltags anschaulich beschrieben: *From a drop of water [...] a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other*. Mit diesen Zeilen beschreibt Arthur Conan Doyle zunächst das sogenannte induktiv schließende Vorgehen, bei dem ausgehend von etwas Besonderem auf das Allgemeine geschlossen werden kann. Mittels induktiv schließenden Vorgehens können Gesetzmäßigkeiten aufgeschlüsselt, dokumentiert und schließlich als Theorien veröffentlicht werden. Die dokumentierten und veröffentlichten Theorien ermöglichen darauf aufbauend ein deduktiv schließendes Vorgehen, bei dem entsprechend vom Allgemeinen auf das Besondere geschlossen werden kann.

Kleiner Exkurs: Im Kontext der Forschungsmethoden spricht man bei einem induktiv schließenden Vorgehen in der Regel von qualitativen Erhebungsverfahren

und bei einem deduktiv schließenden Vorgehen von quantitativen Erhebungsverfahren. Die wesentlichen Unterscheidungskriterien und Anwendungskontexte qualitativer und quantitativer Erhebungsverfahren werden im nachfolgenden Abschnitt zu den grundlegenden Begriffen und Definitionen noch im Detail dargestellt.

Betrachtet man mit diesem Vorwissen die forschungsmethodische Herangehensweise und die daraus resultierenden Fähigkeiten des Sherlock Holmes, dann wird ersichtlich, mit welcher Raffinesse er die ihm zugetragenen Kriminalfälle löst. So begegnet er in den Romanen nicht einfach nur anderen Personen, sondern analysiert und systematisiert die vielen zu diesen Personen gehörenden Details: *By a man's finger-nails, by his coat-sleeve, by his boots, by his trouser-knees, by the callosities of his forefinger and thumb, by his expression, by his shirtcuffs – by each of these things a man's calling is plainly revealed.* Beispielsweise können die Beschaffenheit der Fingernägel sowie die Hornhaut an den Händen einer Person einen ersten Eindruck von der gesellschaftlichen Stellung dieser Person vermitteln – ein ganz ähnliches Prinzip verwenden beispielsweise Wahrsagerinnen und Wahrsager auf Jahrmärkten, die auf anscheinend wundersame Weise richtig zu erkennen scheinen, ob eine Person einen körperlich anstrengenden Beruf ausübt oder doch eher einer Schreibtischtätigkeit nachgeht.

Wenn man weiß, auf welche Details es zu achten gilt und welche Bedeutung diesen Details zugeschrieben werden kann, dann lässt sich das Vorgehen des Sherlock Holmes für eine bestimmte Anzahl N an Details als *formale Beziehung statistischer Variablen* wie folgt festhalten:

$$y = x_1 + x_2 + x_3 + \ldots + x_N$$

Dabei steht y stellvertretend für eine sogenannte *abhängige Variable* und x_n für die entsprechend mit n durchnummerierten *unabhängigen Variablen* bis zur letzten Variable x_N. Für das oben genannte Beispiel bedeutet dies, dass es möglich ist auf die gesellschaftliche Stellung y einer Person zu schließen, ohne diese direkt beobachten zu müssen. Dabei ist die abhängige Variable y in der Regel austauschbar, wodurch sich jeweils neue formale Beziehungen mit entsprechend einschlägigen unabhängigen Variablen x_n ergeben. Die Beziehungen zwischen y und x_n müssen dabei nicht zwingend linear verlaufen, sodass eine entsprechende Flexibilität in der Auswahl der abhängigen und unabhängigen Variablen vorliegt. Dabei gilt es jedoch einige Voraussetzungen sowie mögliche Limitationen mitzudenken, auf die an gegebener Stelle im Detail hingewiesen werden soll.

Dennoch funktioniert diese Vorgehensweise bereits im Alltag erstaunlich gut, etwa wenn man das Wartezimmer einer Arztpraxis betritt und nicht nur nach einem freien Stuhl Ausschau hält, sondern auch nach angenehmen Personen, die links und rechts von diesem freien Stuhl bereits Platz genommen haben. Manchmal nimmt man aber auch neben den falschen Personen Platz und setzt sich freiwillig wieder um. Dies kann das Resultat fehlender Achtsamkeit hinsichtlich der Details oder falscher Schlussfolgerungen sein. So bleibt schließlich auch in den Sherlock Holmes-Romanen nicht unerwähnt, dass Holmes' forschungsmethodische Herangehensweise und die daraus resultierenden Fähigkeiten *[...] can only be acquired by long and*

patient study. Insbesondere die Forschung im Hochschulstudium und im wissenschaftlichen Kontext erfordert neben einer intensiven Auseinandersetzung mit den dokumentierten und veröffentlichten Theorien eine umfassende Kenntnis einschlägiger Forschungsmethoden, um relevante Variablen identifizieren, die Beziehungen zwischen abhängigen und unabhängigen Variablen möglichst präzise nachzeichnen und neue Erkenntnisse generieren zu können.

Eine intensive Auseinandersetzung mit den Grundlagen der Forschungsmethoden ist daher – für achtsame Forscherinnen und Forscher – eine gute Voraussetzung, um falschen Schlussfolgerungen, nicht nur im Wartezimmer einer Arztpraxis, wirksam entgegenzuwirken.

Notizen:

2 Ziele der empirischen Forschung

In diesem Kapitel lernen Sie...

...das **Beschreiben, Erklären und Vorhersagen** als Ziele der empirischen Forschung kennen

...zwischen der **univariaten, bivariaten und multivariaten Statistik** zu unterscheiden

In den Romanen über *Sherlock Holmes* geht es oftmals darum, menschliches Erleben, Verhalten und Handeln folgerichtig und zielorientiert aufzuschlüsseln. Die in den Romanen zum Ausdruck kommende Zielorientierung ist mit den Zielen der *empirischen Forschung* insofern deckungsgleich, als dass das menschliche Erleben, Verhalten und Handeln aus forschungsmethodischer Perspektive tatsächlich bis zu einem gewissen Grad *beschrieben, erklärt und manchmal sogar vorhergesagt* und gegebenenfalls auch verändert werden kann. Diese Möglichkeiten erfordern allerdings eine entsprechende Präzision in der Anwendung univariater, bivariater und multivariater Statistik, auf welche die nachfolgenden Ausführungen daher ebenfalls vorbereiten sollen.

Beim *Beschreiben* des menschlichen Erlebens, Verhaltens und Handelns stehen Angaben zu den Erscheinungsformen und Merkmalen im Vordergrund der Analyse. Dabei können die oftmals unterschiedlichen Erscheinungsformen und Merkmale benannt, geordnet, klassifiziert sowie definiert werden. Ferner können Angaben zu deren Häufigkeiten und den zugrunde liegenden Häufigkeitsverteilungen innerhalb der Erscheinungsformen und Merkmale getroffen werden. Die verschiedenen Möglichkeiten zum Beschreiben einzelner Variablen y sowie x_n werden als *deskriptive Statistik* bezeichnet und präzisieren das Verständnis von der jeweiligen Variable.

Daran anschließend steht das *Erklären* des menschlichen Erlebens, Verhaltens und Handelns im Vordergrund der Analyse. Hierzu erlauben die Verfahren der *bivariaten Statistik* Angaben über die unterschiedlichen Bedingungsverhältnisse von Variablen sowie deren gegenseitige Abhängigkeiten, was formal als $y \sim x_n$ dargestellt werden kann. Anders als beim Beschreiben stehen nicht die einzelnen Variablen y oder x_n im Fokus der Analyse, sondern der kausale Zusammenhang zwischen diesen beiden Variablen. Darüber hinausgehende kausale Zusammenhänge mehrerer erklärender Variablen sind nicht Bestandteil der bivariaten Statistik. Deshalb werden beim Erklären zwar Effektstärken zwischen y und x_n ausgewiesen, diese können aber in einem oftmals realistischeren Modell mit mehreren unabhängigen Variablen durchaus anders ausfallen. Alkoholkonsum (x_n) wirkt beispielsweise auf die Blutalkoholkonzentration (y) ein. Die Effektstärke lässt sich zwar annäherungsweise über die beiden Variablen ermitteln, hängt aber in der Regel ebenfalls vom allgemeinen Gesundheitszustand, vom Alter sowie vom Geschlecht und weiteren unabhängigen Variablen ab, die ebenfalls Einfluss auf die Effektstärke nehmen können.

Unter Kenntnis und Vorlage entsprechender Variablen, also ausgehend von einer theoretisch fundierten und somit angemessenen Anzahl einschlägiger unabhängiger Variablen in Bezug auf eine abhängige Variable, lassen sich ebenfalls *Vorhersagen* treffen. Diese werden auch als Prognosen bezeichnet und können mittels *multivariater Statistik* nachgezeichnet werden. Bezogen auf das zuvor genannte Beispiel zur bivariaten Statistik bedeutet dies, dass der Alkoholkonsum (x_1), der allgemeine Gesundheitszustand (x_2) sowie das Alter (x_3) und Geschlecht (x_4) eine näherungsweise Prognose der Blutalkoholkonzentration (y) ermöglichen, ohne diese direkt messen zu müssen. Folglich stellt bereits auch die forschungsmethodische Herangehensweise von *Sherlock Holmes* aus der *Einführung in die Forschungsmethoden* die multivariate Zusammensetzung mehrerer unabhängiger Variablen und einer abhängigen Variable exemplarisch dar. Die Auswahl ungeeigneter unabhängiger Variablen, also jener Variablen ohne theoretisch fundierten Zusammenhang mit der abhängigen Variable, kann die Prognose allerdings verfälschen, weshalb hier entsprechend achtsam verfahren werden sollte.

Manchmal kann ebenfalls das *Verändern* des menschlichen Erlebens, Verhaltens und Handelns in den Fokus von Forscherinnen und Forschern rücken. Das ist beispielsweise dann der Fall, wenn die Wirkung von Interventionen im Kontext der Evaluationsforschung nachgezeichnet werden soll. Wenn man die kausalen Zusammenhänge zwischen mehreren unabhängigen Variablen und einer abhängigen Variable aufgeschlüsselt hat, kann man Variationen aufseiten der unabhängigen Variablen vornehmen und somit unter Umständen gezielt Einfluss auf die abhängige Variable nehmen.

Notizen:

3 Grundlegende Begriffe und Definitionen

In diesem Kapitel lernen Sie...

...wie über **Forschungsfragen** die Ziele der empirischen Forschung präzisiert werden

...zwischen **theoretischen Hypothesen** und **empirischen Hypothesen** zu unterscheiden

...die **Operationalisierung** von statistischen Variablen und deren **Skalenniveaus** kennen

...**manifeste Variablen** und **latente Variablen** zu unterscheiden

...wie **Tabellen bei strukturierten Daten** abgebildet werden

Bereits in der *Einführung in die Forschungsmethoden* und dem Verweis auf die forschungsmethodische Herangehensweise des *Sherlock Holmes* sind erste Begriffe aufgetaucht, die es für ein besseres Verständnis zu veranschaulichen und zu definieren gilt. Dabei wird zunächst auf den Unterschied zwischen Forschungsfragen und Hypothesen verwiesen, die beide für die Forschung unerlässlich sind. Bei den Hypothesen gilt es ferner zwischen theoretischen und empirischen Hypothesen zu unterscheiden. Die Überführung einer theoretischen Hypothese in eine empirische Hypothese wird als Operationalisierung dargestellt, bei der die bereits erwähnten Variablen in Anlehnung an ein zugrunde liegendes reales Phänomen konstruiert werden. Diese lassen sich primär in abhängige und unabhängige sowie sekundär in manifeste und latente Variablen unterteilen.

Forschungsfragen und Hypothesen

Forschung beginnt in der Regel mit einer zugrunde liegenden *Forschungsfrage*. Die Forschungsfrage präzisiert das Ziel der Forschung, indem sie beispielsweise erste Hinweise auf die *Zielgruppe* der Forschung, den *Anwendungskontext* sowie daran anschlussfähige Theorien aufweist. Mit der Forschungsfrage wird anderen Forscherinnen und Forschern immer auch eine Verortung der vorliegenden Forschung ermöglicht. Dabei setzt die Forschungsfrage häufig bereits das angestrebte Ziel sowie den *aktuellen Stand der Forschung* in Relation zueinander, um bisher noch offene oder unvollständig beantwortete Aspekte innerhalb einer Theorie zu erweitern oder erstmals zu beantworten – kurz: Die Forschungsfrage soll dazu dienen, neues Wissen zu erlangen!

Damit dies gelingt und für andere Forscherinnen und Forscher nachvollziehbar ist, scheinen gleich zu Beginn der eigenen Forschung ein paar allgemeine Hinweise hilfreich zu sein: Zunächst einmal sollte die Forschungsfrage *nicht zu allgemein gehalten* sein. Dies bedingt sich bereits aus der Anforderung eines Hinweises auf die Zielgruppe, den Anwendungskontext und die zugrunde liegenden Theorien. Dabei

dürfen durchaus bereits Teilaspekte fokussiert werden, um die Forschungsfrage möglichst konkret zu gestalten. Weiterhin wird es in der Forschungsfrage höchstwahrscheinlich um *Phänomene* gehen, die nicht allen anderen Forscherinnen und Forschern geläufig sind; schließlich sind Forscherinnen und Forscher vornehmlich Expertinnen und Experten auf ihrem Gebiet. Dies erfordert *klare Benennungen und eindeutige Definitionen*, um den wissenschaftlichen Diskurs um die eigene Forschungsfrage zu ermöglichen. Auch sollte es sich bei den mit der Forschungsfrage fokussierten Phänomenen um „erforschbare" Phänomene handeln. Nicht immer ist das menschliche Erleben, Verhalten und Handeln für Forscherinnen und Forscher *einsehbar, nachvollziehbar oder sogar aufschlüsselbar* – dieses Vorgehen wird im nächsten Abschnitt in Bezug auf die ebenfalls noch vorzustellenden Hypothesen als Operationalisierung bezeichnet. Beispielsweise ist die Religiosität einer Person aus forschungsmethodischer Perspektive nur sehr schwer zu erfassen. Zwar können die Häufigkeit der Gebete sowie des Besuchs eines religiösen Ortes Hinweise auf die Religiosität eines Subjekts liefern, die empfundene Intensität der Religiosität kann dabei aber oftmals nicht vollständig abgebildet werden und ist wahrscheinlich ohnehin intersubjektiv unterschiedlich ausgeprägt. Schlussendlich sollte sich Forschung, damit sie gelingen und die Forschungsfrage beantwortet werden kann, auch umsetzen lassen. Daher empfiehlt es sich, diese Kriterien gleich zu Beginn der eigenen Forschung und während der Ausformulierung der Forschungsfrage mitzudenken.

Während die zugrunde liegende Forschungsfrage den allgemeinen Rahmen der Forschung vorzugeben scheint, sind es in der Regel *Hypothesen*, die zur Beantwortung der Forschungsfrage herangezogen werden. Hypothesen haben den Vorteil, dass sie nicht nur *präzise* und *widerspruchsfrei* formuliert sind, sondern sie gelten auch als *prinzipiell widerlegbar*, sind gut *operationalisierbar* und *theoretisch fundiert*. Aufgrund der Komplexität vieler Phänomene ist es durchaus ratsam, mehrere Hypothesen zur Beantwortung einer Forschungsfrage heranzuziehen. Hypothesen können dabei gerichtet sein, indem sie die unabhängigen und die abhängigen Variablen in Anlehnung an die Theorie ausweisen; andernfalls handelt es sich um *ungerichtete Hypothesen*. Darüber hinaus gelten sie als spezifisch, wenn sie in Anlehnung an die Theorie nicht nur die unabhängigen und abhängigen Variablen ausweisen, sondern in Anlehnung an die Theorie ebenfalls Hinweise auf die erwartete Effektstärke beinhalten. Ohne eine Ausweisung der erwarteten Effektstärke spricht man von unspezifischen Hypothesen. Ferner wird zwischen theoretischen Hypothesen und empirischen Hypothesen unterschieden. *Theoretische Hypothesen* sind nicht nur anschlussfähig an den wissenschaftlichen Diskurs und leiten sich von den zugrunde liegenden Theorien ab, sondern zeichnen sich ebenfalls durch ihren konkreten Bezug zur Forschungsfrage aus. *Empirische Hypothesen* können als Erweiterung der theoretischen Hypothesen angesehen werden, indem sie die in den Hypothesen abgebildeten Variablen benennen und dabei deren Messgrößen sowie Vergleichbarkeit offenlegen. Die Überführung einer theoretischen Hypothese in eine empirische Hypothese wird als Operationalisierung bezeichnet. Erste wichtige Hinweise zur Operationalisierung folgen bereits im nächsten Abschnitt.

Operationalisierung zur Beantwortung von Hypothesen

Um den Prozess der *Operationalisierung* zu veranschaulichen, bedarf es einer beispielhaften theoretischen Fundierung. Eine solche findet sich in der sogenannten *Mincer-Einkommensgleichung*, benannt nach dem Ökonomen Jacob Mincer. Die Mincer-Einkommensgleichung setzt das Lohneinkommen einer Person in Abhängigkeit zu deren Schulbildungsniveau und der Berufserfahrung (vgl. Mincer 1958). Dabei stellen das Lohneinkommen eine abhängige und das Schulbildungsniveau sowie die Berufserfahrung zwei unabhängige Variablen dar. Von diesen bereits postulierten Zusammenhängen ausgehend lässt sich die bei Jacob Mincer zugrunde liegende *Forschungsfrage* rekonstruieren.

Forschungsfrage: Gibt es einen Zusammenhang zwischen dem Lohneinkommen einer Person und deren Schulbildungsniveau sowie Berufserfahrung?

Der Theorie entsprechend steigt das Lohneinkommen mit höheren Bildungsniveaus und mehr Berufserfahrung, sodass ausgehend von der Forschungsfrage und der zugrunde liegenden Theorie zwei einschlägige *theoretische Hypothesen* formuliert werden können:

Theoretische Hypothese$_1$: Es gibt einen positiven Zusammenhang zwischen dem Lohneinkommen einer Person und deren Schulbildungsniveau.

Theoretische Hypothese$_2$: Es gibt einen positiven Zusammenhang zwischen dem Lohneinkommen einer Person und deren Berufserfahrung.

An dem oben genannten Beispiel wird deutlich, dass ausgehend von einer allgemeinen Forschungsfrage mehrere einschlägige Hypothesen formuliert werden können. Die theoretischen Hypothesen sollen in Übereinstimmung mit der Theorie dazu beitragen, die Forschungsfrage zu beantworten. Dabei halten die theoretischen Hypothesen die Behauptung der Forscherinnen und Forscher nicht nur präzise fest, sondern weisen gegebenenfalls zusätzlich eine positive oder negative Effektstärke aus (im vorherigen Abschnitt als spezifische Hypothesen und unspezifische Hypothesen definiert). Die bereits genannten theoretischen Hypothesen sind zugleich *Haupthypothesen*, da sie primär zur Beantwortung der Forschungsfrage beitragen. Sind weitere unabhängige Variablen zu berücksichtigen, so werden diese in der Regel als *Nebenhypothesen* festgehalten. Nebenhypothesen können im Zusammenspiel mit den Haupthypothesen dazu beitragen, ein möglichst umfassendes Abbild realer Phänomene nachzuzeichnen. So könnte man bei der Mincer-Einkommensgleichung beispielsweise das Verhandlungsgeschick einer Person beim Vorstellungsgespräch als weitere unabhängige Variable berücksichtigen, da dieses ebenfalls positiv auf das Lohneinkommen einer Person Einfluss nehmen kann.

In einem nächsten Schritt können die theoretischen Hypothesen operationalisiert und in *empirische Hypothesen* überführt werden. Dazu müssen die Forscherinnen und

Forscher festlegen, wie die realen Phänomene beobachtet und gemessen werden können. In den Ländern der Eurozone bietet sich für das Lohneinkommen einer Person beispielsweise das Bruttogehalt in Euro an. Zusätzlich muss bei der Operationalisierung konkretisiert werden, ob das monatliche oder das jährliche Lohneinkommen beobachtet und gemessen werden soll. Nur so können verschiedene Werte mehrerer Personen vergleichbar gemacht werden. Entsprechend lässt sich das Schulbildungsniveau über die *International Standard Classification of Education* (ISCED) beobachten und messen, welche in Bezug auf die Allgemeinbildung die unterschiedlichen Schulbildungsniveaus mit Werten von 0 bis 3 ausdifferenziert. Die Berufserfahrung lässt sich schließlich in Anzahl an Jahren beobachten und messen. Ausgehend von dieser Konkretisierung lassen sich die empirischen Hypothesen formulieren.

Empirische Hypothese$_1$: Es gibt einen positiven Zusammenhang zwischen dem monatlichen Brutto-Lohneinkommen einer Person in Euro und deren Schulbildungsniveau nach ISCED.

Empirische Hypothese$_2$: Es gibt einen positiven Zusammenhang zwischen dem monatlichen Brutto-Lohneinkommen einer Person in Euro und deren Berufserfahrung in Jahren.

Bei der Mincer-Einkommensgleichung handelt es sich übrigens um eine *bewährte* Theorie, da die empirischen Hypothesen bereits in mehreren voneinander unabhängigen Untersuchungen *verifiziert* werden konnten. Wenn sich eine empirische Hypothese hingegen nicht bestätigen lässt, so gilt sie als *falsifiziert*. Ob eine empirische Hypothese verifiziert oder falsifiziert werden kann, sagt bei einem gut begründeten forschungsmethodischen Vorgehen jedoch nichts über deren Qualität aus. Schließlich kann sowohl bei der Verifikation als auch bei der Falsifikation Wissen generiert und zur Diskussion gestellt werden.

Zusammenfassend lässt sich zur Operationalisierung sagen, dass Forscherinnen und Forscher mit ihren empirischen Hypothesen die erwartete Richtung der Effekte und die verwendete Einheit der Beobachtung und Messung offenlegen. Dadurch tragen empirische Hypothesen dazu bei, dass das Vorgehen von Forscherinnen und Forschern nachvollziehbar und vergleichbar ist. Mittels adäquater Ausweisung von Haupt- und Nebenhypothesen belegen die Forscherinnen und Forscher zusätzlich, dass sie die zugrunde liegenden realen Phänomene in angemessener Weise nachzuzeichnen vermochten, um zu ihren Ergebnissen zu gelangen. Nur auf diesem Wege können Hypothesen bewährt, verifiziert oder falsifiziert werden.

Variablen und Konstanten in Datensätzen

Während der Operationalisierung hat sich gezeigt, dass sich Hypothesen aus zwei oder mehr Variablen zusammensetzen können und dabei eine Unterteilung in abhängige oder unabhängige Variablen vorgenommen werden kann. In den quantitativen Forschungsmethoden spricht man immer dann von einer Variable, wenn verän-