HMD (2020) 57:106–116
https://doi.org/10.1365/s40702-019-00557-y

SCHWERPUNKT

Kausalanalyse mit maschinellem Lernen

Martin Huber

Eingegangen: 8. Mai 2019 / Angenommen: 22. Juli 2019 / Online publiziert: 8. August 2019

Zusammenfassung Die datenbasierte Kausalanalyse versucht, den kausalen Effekt einer Intervention auf ein interessierendes Ergebnis zu messen, häufig unter Kontrolle beobachtbarer Charakteristiken, die ebenfalls das Ergebnis beeinflussen. Beispiele für kausale Fragestellungen sind: Was ist der Effekt einer Marketingkampagne (Intervention) auf die Verkaufszahlen (Ergebnis) unter ansonsten identischen Marktbedingungen? Was ist der Effekt einer Zinsveränderung (Intervention) auf den Aktienkurs (Ergebnis) unter ansonsten identischen wirtschaftlichen Rahmenbedingungen? Die Kausalanalyse unterscheidet sich deshalb konzeptionell von der statistischen Vorhersage. Letztere versucht aus Kombinationen von Charakteristiken (zum Beispiel Zinssatz, Wirtschaftswachstum, Unternehmensgewinn) möglichst genau das Ergebnis (zum Beispiel Aktienkurs) vorherzusagen, ohne die kausalen Effekte der einzelnen Charakteristiken zu bestimmen. Im Zeitalter von „Big Data“ erfährt die Vorhersage in vielen Bereichen einen qualitativen Quantensprung aufgrund des Einsatzes von maschinellem Lernen. Letzteres vermag in großen Datensätzen jene Kombinationen von Charakteristiken zu lernen, die für die Vorhersage des Ergebnisses entscheidend sind. Dieser Beitrag diskutiert, wie die Vorzüge des maschinellen Lernens auch für die Kausalanalyse in großen Daten genutzt werden können. Die Messung eines kausalen Effektes ist möglich, wenn für Charakteristiken, welche die Intervention und das Ergebnis bedeutend beeinflussen, kontrolliert werden kann. Dies lässt sich durch sogenanntes „doppeltes maschinelles Lernen“ implementieren. Dabei werden sowohl die Intervention, als auch das Ergebnis als Funktion der anderen Charakteristiken vorhergesagt um letztendlich den Effekt der Intervention auf das Ergebnis zu schätzen. Der Beitrag diskutiert diesen Ansatz beispielhaft anhand eines bestimmten statistischen Modells und verweist auf mehrere Praxisbeispiele.

M. Huber (✉)
Universität Freiburg, Freiburg i. Ü., Schweiz
E-Mail: martin.huber@unifr.ch

Schlüsselwörter Kausalanalyse · Maschinelles Lernen · Kausales maschinelles Lernen · Doppeltes maschinelles Lernen · Lasso Regression

Combining Causal Analysis with Machine Learning

Abstract Data-based causal analysis aims at evaluating the causal effect of some intervention on an outcome of interest, frequently by controlling for observed characteristics also affecting the outcome. Examples for causal questions are: What is the effect of a marketing campaign (intervention) on sales (outcome) under otherwise identical market conditions? What is the effect of a change in interest rates (intervention) on stock prices (outcome) under otherwise identical economic conditions? Therefore, causal analysis conceptually differs from statistical prediction. The latter aims at predicting an outcome (e.g. stock prices) from a combination of characteristics (e.g. interest rates, economic growth, profits), however, without determining the causal effects of the various characteristics. In the age of "big data", the use of machine learning has entailed a boost in the quality of predictions in many domains. In sufficiently large data, machine learning is capable of learning those combinations of characteristics that are crucial for the prediction of the outcome. This article discusses how the benefits of machine learning can also be used for causal analysis in big data. Evaluating a causal effect is feasible if any characteristics that importantly affect both the intervention and the outcome can be controlled for. So-called "double machine learning" may achieve this goal. It consists of predicting both the intervention and the outcome as functions of the other characteristics to ultimately estimate the effect of the intervention on the outcome. The article discusses this approach based on a particular statistical model and refers the reader to several empirical examples.

Keywords Causal analysis · Machine learning · Causal machine learning · Double machine learning · Lasso regression

1 Einführung in die Kausalanalyse

Die statistische Kausalanalyse (siehe Pearl 2000 und Imbens und Rubin 2015 für Einführungen) befasst sich mit der empirischen, also datenbasierten Untersuchung der ursächlichen Wirkung eines bestimmten Phänomens, nachfolgend „Intervention" genannt, auf ein anderes Phänomen, nachfolgend „Ergebnis" genannt. Beispiele hierfür sind die Wirkung einer medizinischen Behandlung auf die Gesundheit, von Bildung auf Einkommen, von Innovation auf Wirtschaftswachstum und vieles mehr. Um den kausalen Effekt einer Intervention identifizieren zu können, müssen alle anderen Faktoren, die ebenfalls die Ergebnis beeinflussen, konstant gehalten werden. Man spricht deshalb auch vom „ceteris paribus"-Effekt, also der Wirkung der Intervention wenn andere Determinanten des Ergebnisses unverändert bleiben. So wird sichergestellt, dass der interessierende Effekt nicht durch Variation in anderen Faktoren kontaminiert wird.

Für die Formalisierung von kausalen Effekten eignet sich das statistische Konzept der sogenannten „potenziellen Ergebnisse“ (engl. „potential outcomes“), siehe zum Beispiel Neyman (1923) und Rubin (1974). Y bezeichnet zu diesem Zweck das in den Daten beobachtete Ergebnis, zum Beispiel den Gesundheitszustand gemessen anhand eines Indexes, und D die Intervention, zum Beispiel eine bestimmte medizinische Behandlung. Einfachheitshalber sei angenommen, dass D binär ist und den Wert 1 annimmt, falls die Behandlung durchgeführt wird, und den Wert 0, falls dies nicht der Fall ist. Ferner bezeichnet $Y(d)$ den Wert, den das Ergebnis annehmen würde, wenn Index D auf einen bestimmten Wert d gesetzt würde, wobei d eben entweder 1 oder 0 sein kann. Es handelt sich bei $Y(d)$ um ein potenzielles Ergebnis (im Gegensatz zum tatsächlich beobachteten Ergebnis Y), da D in der Realität nicht zwingenderweise diesen Wert annimmt. Der Effekt der medizinischen Intervention lässt sich nun als Unterschied der potenziellen Gesundheitszustände mit Behandlung ($d = 1$) vs. ohne Behandlung ($d = 0$) ausdrücken: $Y(1) - Y(0)$.

Für eine bestimmte Person sind derartige kausale Effekte allerdings nicht (ohne starke statistische Annahmen) identifizierbar, weil ein Individuum die Behandlung zu einem bestimmten Zeitpunkt entweder bekommt oder eben nicht. Das bedeutet, dass nur eines der beiden potenziellen Ergebnisse beobachtbar ist, aber niemals beide gleichzeitig. Formal besteht der folgende Zusammenhang des beobachteten und der potentiellen Ergebnisse für ein binäres D:

$$Y = D \cdot Y(1) + (1 - D) \cdot Y(0). \tag{1}$$

Der faktische Gesundheitszustand Y entspricht dem potenziellen Gesundheitszustand unter dem faktischen Behandlungszustand D, aber der kontrafaktische Gesundheitszustand, das heißt der potenzielle Gesundheitszustand, wenn D einen anderen Wert angenommen hätte, bleibt unbekannt. Dies ist als „fundamentales Problem der Kausalität“ bekannt.

In vielen kausalen Fragestellungen ist nicht unbedingt (nur) der kausale Effekt auf ein Individuum, sondern der Durchschnittseffekt auf alle Individuen, bezeichnet als Δ, von Interesse, zum Beispiel der mittleren Effekt der Behandlung in der Bevölkerung eines Landes:

$$\Delta = E(Y(1) - Y(0)). \tag{2}$$

„E“ bezeichnet den Erwartungswert. Mittlere Effekte erscheinen einfacher zu identifizieren als individuelle Effekte, wenn man sowohl Personen mit Behandlung ($D = 1$) und Personen ohne Behandlung ($D = 0$) beobachtet. Man könnte deshalb geneigt sein, den mittleren Effekt durch einen Vergleich der mittleren Gesundheitszustände in den Gruppen mit $D = 1$ und $D = 0$ zu berechnen: $E(Y|D = 1) - E(Y|D = 0)$, wobei „|“ für „gegeben“ oder „konditional auf“ steht.

In der Regel entspricht dieser Mittelwertvergleich allerdings nicht dem kausalen Effekt Δ, wenn Personen in Behandlung andere gesundheitsrelevante Charakteristiken aufweisen als jene ohne Behandlung. Die beiden Gruppen könnten sich etwa hinsichtlich des Alters, des Krankheitsverlaufs vor der Behandlung oder des Ein-

kommens systematisch unterscheiden, um nur ein paar Beispiele zu nennen. Die oben erwähnte „ceteris paribus"-Annahme ist in diesem Fall nicht erfüllt, sodass $E(Y|D=1) - E(Y|D=0)$ den Effekt der Intervention mit dem Effekt der Charakteristiken vermischt, weil gewissermaßen Äpfel mit Birnen verglichen werden, was die Charakteristiken betrifft.

Die Herausforderung der Kausalanalyse liegt also in der Isolierung der Wirkung einer Intervention von der Wirkung anderer Charakteristiken. Dies funktioniert per Definition in einem Experiment, in dem die Intervention zufällig (zum Beispiel durch Münzwurf) zugeteilt wird. Dann haben die Charakteristiken keinerlei Einfluss auf D, sodass $E(Y|D=1) - E(Y|D=0)$ einem Vergleich von Äpfeln mit Äpfeln entspricht, der Δ wiedergibt, also den Durchschnittseffekt identifiziert. Jedoch lassen sich aus verschiedenen Gründen (ethische Bedenken, Kosten, etc.) nicht für alle Fragestellungen und Situationen immer die passenden Experimente finden oder durchführen.

Eine weitere Strategie für die Identifikation von Δ basiert auf der Annahme, dass für alle Faktoren, die gleichzeitig die Intervention D und das Ergebnis Y beeinflussen, anhand beobachteter Charakteristiken (wie zum Beispiel Alter), bezeichnet mit X, kontrolliert werden kann. Es wird somit unterstellt, dass gegeben X, das heißt für Personen mit gleichen Werten in X, keine weiteren, unbeobachteten Charakteristiken gleichzeitig D und Y beeinflussen. Eine hinreichende Bedingung dafür ist die statistische Unabhängigkeit der potenziellen Ergebnisse von D gegeben X. Das bedeutet, dass Gruppen mit $D=1$ und $D=0$ mit identischen Werten in X – gleichsam einem Experiment – identische Verteilungen der potenziellen Ergebnisse aufweisen, wie die nachfolgende Annahme formalisiert:

Annahme 1: $Y(1), Y(0)$ sind unabhängig von $D|X$.

Annahme 1 ermöglicht die Identifikation des mittleren Effekts gegeben X, bezeichnet als Δ_X, anhand eines Vergleiches der mittleren Ergebnisse von Personen mit $D=1$ und $D=0$, die identische Werte in X aufweisen:

$$\Delta_X = E(Y|D=1,X) - E(Y|D=0,X). \tag{3}$$

Dies setzt voraus, dass für alle möglichen Werte in X (zum Beispiel jedes in der interessierenden Population vorkommende Alter) sowohl Personen mit $D=1$ als auch mit $D=0$ existieren. Diese Annahme eines gemeinsamen Stützbereichs bezüglich der Werte in X lässt sich wie folgt formalisieren:

Annahme 2:

$$0 < \Pr(D=1|X) < 1\,.$$

„Pr" bezeichnet eine Wahrscheinlichkeit. Annahme 2 besagt, dass keine Kombination von Charakteristiken X existiert, für die alle Personen entweder behandelt oder nicht behandelt werden. D darf zwar eine Funktion von X sein, aber keine deterministische.

Unter Annahmen 1 und 2 ist der mittlere Effekt Δ als Mittelwert des konditionalen Effekts Δ_X identifiziert:

$$\Delta = E(\Delta_X) = E(E(Y|D=1,X) - E(Y|D=0,X)). \tag{4}$$

Der konditionale Erwartungswert $E(Y|D,X)$ kann auf verschiedene Arten modelliert werden. Legt man ein einfaches lineares Modell zu Grunde, dann entspricht $E(Y|D,X) = \alpha + D\beta + X'\gamma$, wobei α, β und γ die Konstante sowie die Koeffizienten von D und X im linearen Modell bezeichnen. Flexibler ist die Annahme eines nichtparametrischen Modells $E(Y|D,X) = g(D,X)$, wobei g eine unbekannte Funktion ist, die möglicherweise Interaktionseffekte von D und X inkludiert, sodass Δ_X arbiträr mit unterschiedlichen Werten in X variieren kann. In diesem Fall kann der kausale Effekt zum Beispiel mithilfe von Kernregression oder sogenannten Matching-Verfahren geschätzt werden (siehe den Methodenüberblick in Imbens und Wooldridge 2009).

2 Chancen und Herausforderungen mit großen Daten

Ob die Anzahl und Qualität der beobachteten Charakteristiken X ausreichend ist, um Annahme 1 zu erfüllen, kann in der Regel nicht getestet werden. Allerdings kann die im Zeitalter der Digitalisierung und Datenvernetzung steigende Verfügbarkeit von Informationen die Plausibilität von Annahme 1 in vielen empirischen Fragestellungen erhöhen. Eine höhere Anzahl von gemessenen Variablen erhöht die Chance, mehr relevante Charakteristiken zu beobachten, anhand derer sich für Faktoren, die gleichzeitig die Intervention D und das Resultat Y beeinflussen, besser kontrollieren lässt. Auf der anderen Seite wächst aber auch die Wahrscheinlichkeit, mehr irrelevante Charakteristiken zu beobachten, die (zumindest über bereits berücksichtigte Kontrollvariablen hinaus) keinen Nutzen für die Erfüllung von Annahme 1 bringen.

Ein Beispiel hierfür könnte die Lieblingsfarbe der Patienten sein, wenn diese weder für die Behandlungswahrscheinlichkeit noch für den Gesundheitszustand eine Rolle spielt. In diesem Fall wäre die Lieblingsfarbe irrelevant für die (Nicht-)Erfüllung von Annahme 1 und somit auch für die sogenannte Verzerrung einer Schätzmethode, also die mittlere Abweichung des geschätzten Effekts vom wahren Δ über viele Stichproben mit konstanter Größe. Würde man diese Variable dennoch in X inkludieren, so hätte dies für die Identifikation von Δ in unendlich großen Stichproben weder Vor- noch Nachteile. Für die Schätzung von Δ in realen (und deshalb endlichen) Stichproben führt das Kontrollieren für irrelevante Charakteristiken aber generell zu einer Erhöhung der Varianz und somit des Schätzfehlers. Deshalb erscheint es nicht ratsam, grundsätzlich für alle im Datensatz beobachtbaren Charakteristiken zu kontrollieren, insbesondere wenn die Anzahl der Charakteristiken hoch ist relativ zur Stichprobengröße.

In der Kausalanalyse wird die Plausibilität von Annahme 1 sowie die Definition der Kontrollvariablen X häufig anhand theoretischer Überlegungen oder statistischer Zusammenhänge zwischen diversen Charakteristiken und der Intervention oder dem Ergebnis motiviert. Allerdings macht eine steigende Variablenzahl die manuelle Aus-

wahl der Kontrollvariablen (zum Beispiel durch Testen der partiellen Korrelation von potenziellen Kontrollvariablen mit D und Y) praktisch zusehends schwieriger. Dazu kommt das Problem, dass Unsicherheit in der (Modell-)Selektion hinsichtlich X in der statistischen Methode zur Schätzung von Δ typischerweise nicht berücksichtigt wird, was implizit unterstellt, dass die korrekte Definition von X von vornherein bekannt ist. Unter der Prämisse dass Annahme 1 im verfügbaren Datensatz erfüllt ist, bietet maschinelles Lernen (ML) die Chance, datengetrieben für relevante Charakteristiken zu kontrollieren und eine akkuratere Schätzung von Δ zu erhalten, als es durch die manuelle Wahl von X möglich ist, insbesondere wenn die Anzahl der beobachteten Charakteristiken groß ist.

3 Maschinelles Lernen für die Vorhersage

ML ist weitverbreitet zum Zweck der Vorhersage, welche nachfolgend kurz skizziert wird, um sie von den Anforderungen der Kausalanalyse unterscheiden zu können. Ziel ist es, ein Ergebnis Y (zum Beispiel Gesundheitszustand) anhand von beobachtbaren Charakteristiken X (Alter, Einkommen, etc.), die in diesem Zusammenhang auch Prädiktoren genannt werden, bestmöglich vorherzusagen. Dies entspricht einer möglichst akkuraten Schätzung des Zusammenhanges zwischen Y und X, welcher der Funktion $f(X)$ in der folgenden Gleichung entspricht:

$$Y = f(X) + W. \tag{5}$$

W repräsentiert die Abweichung zwischen $f(X)$ und Y aufgrund unbeobachteter Variablen, das sogenannte Residuum. Wenn $\widehat{f}(X)$ eine ML Methode zur Schätzung von $f(X)$ bezeichnet, so bedeutet „möglichst akkurat“, dass ein bestimmtes Fehlermaß für die Abweichung zwischen Vorhersage $\widehat{f}(X)$ und tatsächlichem Ergebnis Y minimiert wird. Ein gängiges Kriterium ist der mittlere quadrierte Fehler, $E((\widehat{f}(X) - Y)^2)$, der sich sowohl aus $E((\widehat{f}(X) - E(\widehat{f}(X)))^2)$, der Varianz von $\widehat{f}(X)$, als auch der quadrierten Verzerrung $(E(\widehat{f}(X) - Y))^2$ zusammensetzt.

Die Approximation $\widehat{f}(X)$ anhand von ML ist als blackbox zu verstehen. Sie hat nicht zum Ziel, die kausalen Effekte von Elementen in X auf Y möglichst korrekt wiederzugeben, sondern einzig Y bestmöglich vorherzusagen. Wenn $\widehat{f}(X)$ zum Beispiel durch eine sogenannte Lasso-Regression geschätzt wird, siehe Tibshirani (1996), so entspricht dies einer Regression, in der (im Unterschied zur Methode der kleinsten Quadrate) die Summe der Absolutwerte der Koeffizienten von X restringiert wird. Deshalb werden typischerweise manche Koeffizienten gegen oder genau auf Null verzerrt, um die Restriktion einzuhalten. Dies betrifft insbesondere die Koeffizienten von verhältnismäßig schlechten Prädiktoren und solchen, die stark mit anderen Prädiktoren korrelieren. Die an sich unerwünschte Verzerrung geht einher mit einer erwünschten Reduktion in der Varianz, weil durch die Restriktion – auch Regularisierung genannt – effektiv weniger Koeffizienten zu schätzen sind.

Es gilt somit in der Lasso-Regression, wie in anderen ML Methoden, die optimale Balance zwischen Verzerrung und Varianz zu finden, die den mittleren quadrierten Fehler minimiert, zum Beispiel mithilfe der sogenannten Kreuz-Validierung. Letz-

tere basiert auf der (zufälligen) Teilung eines Datensatzes in eine sogenannte Trainingsstichprobe, in der $\widehat{f}(X)$ (zum Beispiel basierend auf den Lasso-Koeffizienten) geschätzt wird und in eine Validierungsstichprobe, in der anhand der geschätzten Funktion der mittlere quadrierte Fehler für unterschiedliche Regularisierungen (zum Beispiel stärkere und schwächere Restriktionen auf die Lasso-Koeffizienten) ermittelt wird. Dabei werden die Rollen von Trainings- und Validierungsdaten getauscht und letztendlich jene Regularisierung gewählt, die den Durchschnitt des mittleren quadrierten Fehlers über alle Validierungsstichproben minimiert.

Die Schätzung des Modells und seine Evaluierung anhand des mittleren quadrierten Fehlers in getrennten, unabhänigen Stichproben ist wichtig, um eine Überanpassung von $\widehat{f}(X)$ (engl. „overfitting“) zu vermeiden: Würden beide Schritte in ein und derselben Stichprobe durchgeführt, ließe sich der mittlere quadrierte Fehler durch immer weniger restringierte Koeffizienten immer weiter reduzieren, wodurch die Verzerrung verringert wird. Dies birgt aber die Gefahr, dass $\widehat{f}(X)$ zu spezifisch auf zufällig auftretende (und durch das Residuum getriebene) Datenmuster in der gegebenen Stichprobe angepasst wird, die sich nicht systematisch in anderen Stichproben finden. Dadurch steigt die Varianz von $\widehat{f}(X)$ in solchen anderen Stichproben aufgrund der tendenziellen Wahl von zu vielen beziehungsweise zu (absolut) großen Koeffizienten. Aus diesem Grund wird zwischen Trainings- beziehungsweise Validierungsdaten unterschieden, um anhand der letzteren die zu erwartete Performanz von $\widehat{f}(X)$ in anderen, zum Beispiel zukünftig gesammelten Daten zu evaluieren.

Da Methoden der statistischen Vorhersage ausschließlich das Ziel haben, den Vorhersagefehler zu minimieren, sollten die der Approximation $\widehat{f}(X)$ zugrundeliegenden Parameter, wie zum Beispiel Lasso-Koeffizienten, nicht kausal interpretiert werden. Ohnehin können Lasso-Koeffizienten über verschiedene Stichproben substantiell variieren, sogar dann, wenn die Stichproben relativ ähnlich sind. Dies ist dann der Fall, wenn unterschiedliche Kombinationen von Prädiktoren jeweils ähnlich gute Vorhersagen liefern. ML Methoden zur Vorhersage sind im Allgemeinen ungeeignet für die Kausalanalyse, außer in sehr spezifischen Fällen, siehe Bühlmann und van de Geer (2011), die in empirischen Problemen mit größter Wahrscheinlichkeit nicht erfüllt sind.

4 Maschinelles Lernen für die Kausalanalyse

Um sich die Vorteile von ML im Kontext der Kausalanalyse für die datenbasierte Berücksichtigung von Kontrollvariablen zunutze machen, muss und kann ML an die Anforderungen der Kausalanalyse adaptiert werden. Zur besseren Illustration von einer naiven und angemessenen Anwendung von ML für die Schätzung des kausalen Effekts Δ sei in Anlehnung an die Diskussion in Chernozhukov et al. (2018) das folgende Modell für das Ergebnis Y und die Intervention D unterstellt:

$$Y = D\beta + g(X) + U, \tag{6}$$

$$D = m(X) + V. \tag{7}$$

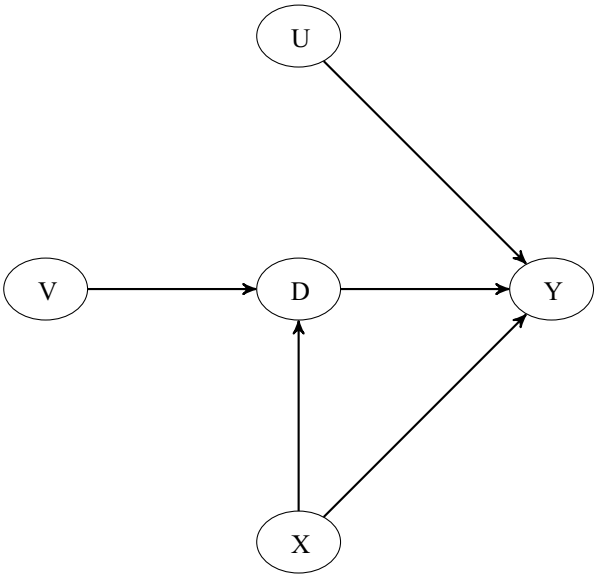

Abb. 1 Kausale Effekte zwischen Variablen in Gln. (6) und (7) unter Annahme 1

$g(X)$ und $m(X)$ bezeichnen unbekannte Funktionen, die den Zusammenhang zwischen den beobachteten Kovariaten X und Y beziehungsweise D abbilden. U und V sind Residuen mit einem Mittelwert von Null, welche unbeobachtbare Faktoren reflektieren, die ebenfalls Y beziehungsweise D beeinflussen. Dieses Modell impliziert, dass die potenziellen Ergebnisse durch $Y(1) = \beta + g(X) + U$ und $Y(0) = g(X) + U$ gegeben sind. Der Koeffizient β entspricht dem interessierenden kausalen Effekt:

$$\Delta = E(Y(1) - Y(0)) = E(\beta + g(X) + U - g(X) - U) = \beta. \tag{8}$$

Annahme 1 ist erfüllt, falls D statistisch unabhängig von U ist gegeben X. Abb. 1 veranschaulicht ein derartiges Szenario für die Variablen in Gln. (6) und (7), wobei jeder Pfeil einen kausalen Effekt darstellt. Da keine Pfeile zwischen V und U oder U und D existieren, ist D unabhängig von U gegeben X, wie von Annahme 1 unterstellt. Dies ist eine hinreichende Bedingung für die Identifikation des kausalen Effekts von D auf Y, weil in diesem Fall gilt (Fig. 1):

$$E(E(Y|D = 1, X) - E(Y|D = 0, X)) = E(\beta + g(X) - g(X)) = \beta. \tag{9}$$

Es wäre nun naiv, den ML Ansatz für die Vorhersage 1 : 1 zur Messung des kausalen Effekts von D zu übernehmen. Aus in Kap. 3 dargelegten Gründen führt das gleichzeitige Schätzen von β und $g(X)$ generell zu einem verzerrten Schätzer des kausalen Effektes von D auf Y, insbesondere wenn X und D stark korrelieren und/oder der Effekt von D auf Y verhältnismäßig klein ist. Während es für die optimale Vorhersage von Y in einer Lasso-Regression nützlich sein kann, den Koeffizienten von D zwecks Verringerung der Varianz gegen Null zu verzerren, wird in der Kausalanalyse eine möglichst geringe Verzerrung in der Schätzung von β an-

gestrebt. Ebenfalls naiv wäre es, Gl. (6) erst in den Trainingsdaten zu schätzen und das Residuum (das heißt die Differenz) zwischen Y und der geschätzten Funktion $\widehat{g}(X)$ in den Testdaten auf D zu regressieren, um letztendlich β zu schätzen. Nach wie vor besteht das Problem, dass es für die Vorhersage von Y nützlich sein kann, zwecks Varianzreduktion Lasso-Koeffizienten von X gegen Null zu verzerren. Die resultierende verzerrte Schätzung von $g(X)$ ist ungeeignet, um vollständig für den Zusammenhang zwischen D und X zu kontrollieren. Dies verzerrt wiederum den Schätzer von β in einer Regression des Residuums von Y auf D.

Ein annähernd erwartungstreuer (das heißt unverzerrter) Schätzer des kausalen Effekts, der über viele Stichproben gemittelt dem wahren β (und somit Δ) nahe kommt, kann durch sogenanntes doppeltes maschinelles Lernen (DML) konstruiert werden. Doppelt deshalb, weil in einem ersten Schritt sowohl $g(X)$ als auch auch $m(X)$, also D als Funktion von X, anhand von ML geschätzt werden, um die Vorhersagen $\widehat{g}(X)$ und $\widehat{m}(X)$ zu erhalten. In einem zweiten Schritt wird das Residuum $Y - \widehat{g}(X)$ linear auf das Residuum $D - \widehat{m}(X)$ regressiert, unter Verwendung der Standardmethode der kleinsten Quadrate. Dies führt zu verbesserten statistischen Eigenschaften, obwohl $\widehat{g}(X)$ als auch $\widehat{m}(X)$ für sich genommen aufgrund der Regularisierung wiederum verzerrt sind. Wie in Chernozhukov et al. (2018) ausgeführt ist für die durch Regularisierung verursachte Verzerrung des Schätzers von β aber nun das Produkt der Verzerrungen von $\widehat{g}(X)$ und $\widehat{m}(X)$ entscheidend, anstatt des problematischeren Produktes von D und der Verzerrung von $\widehat{g}(X)$ wie im naiven Ansatz.

Ein weiterer Faktor für die Konstruktion eines annähernd erwartungstreuen Schätzers des kausalen Effekts ist das Verwenden von unabhängigen Stichproben für die Schätzung von $g(X)$ und $m(X)$ auf der einen Seite und von β auf der anderen Seite. Dies wird ähnlich der Idee von Trainings- und Validierungsdaten in Kap. 3 durch eine zufällige Teilung des Datensatzes in zwei unabhängige Stichproben erreicht, wobei im einen Teil $g(X)$ und $m(X)$ und im anderen Teil β anhand der Residuen-Regression geschätzt werden. Das Schätzen von $g(X)$ in einer anderen Stichprobe als β vermeidet Überanpassung von $g(X)$, was zu einer Verzerrung des geschätzten kausalen Effekts führen würde. Allerdings hat die Teilung des Datensatzes einen Preis: Ein Teil der Beobachtungen wird nicht für die Schätzung von β verwendet, was die Varianz erhöht. Eine Möglichkeit, dem entgegenzuwirken ist das sogenannte Kreuz-Anpassen (engl. „cross-fitting“), bei dem die Rollen der Datenteile getauscht werden. Einmal dient der eine Datenteil für das Schätzen von $g(X)$ und $m(X)$ und der andere für das Schätzen von β und das andere Mal ist es umgekehrt. Der Mittelwert der beiden geschätzten kausalen Effekte wird schließlich als Schätzer für β herangezogen und besitzt eine geringere Varianz als jeder der beiden Schätzer für sich allein.

DML hat unter bestimmten Bedingungen vergleichbare Eigenschaften wie parametrische statistische Standardschätzer wie etwa die lineare Regression: Der Schätzer von β ist asymptotisch annähernd normalverteilt und Wurzel-N konsistent, das heißt er konvergiert gegen β mit der Rate $1/\sqrt{N}$, wobei N die Anzahl der Beobachtungen bezeichnet. Dies erlaubt es, anhand der asymptotischen Standardtheorie p-Werte und Konfidenz-Intervalle für den geschätzten kausalen Effekt zu berechnen. Eine Voraussetzung dafür ist, dass $g(X)$ und $m(X)$ hinreichend akkurat von den ML

Methoden approximiert werden. Im Fall der Lasso-Regression bedeutet dies, dass eine relativ zur Stichprobengröße begrenzte Anzahl an Variablen ausreichend ist, um bereits eine gute Approximation der Funktionen zu erreichen. Zum Beispiel konvergiert der Schätzer von β mit der Rate $1/\sqrt{N}$ gegen den wahren kausalen Effekt, wenn $\widehat{g}(X)$ und $\widehat{m}(X)$ nur mit der langsameren Rate $1/N^{1/4}$ gegen $g(X)$ und $m(X)$ konvergieren. Diese Rate wird von einigen ML Verfahren unter hier nicht weiter ausgeführten statistischen Annahmen erreicht.

DML zum Messen kausaler Effekte lässt sich auch auf allgemeinere kausale Modelle als jenes in Gl. (6) anwenden, welches aus illustrativen Zwecken gewählt wurde, da in diesem Fall lineare Residuen-Regression für die Schätzung von β verwendet werden kann. Wie in Chernozhukov et al. (2018) diskutiert könnte man jedoch unter Verwendung einer anderen, sogenannten doppelt-robusten Schätzmethode Gl. (6) durch das folgende nichtparametrische Modell ersetzen, welches Interaktionseffekte zwischen D und X zulässt:

$$Y = g(D, X) + U. \tag{10}$$

DML wird die empirische Kausalanalyse in den folgenden Jahren zweifelsohne revolutionieren. Während bisher gängige Verfahren implizit unterstellen, dass die zu berücksichtigenden Charakteristiken X gegeben und somit bekannt sind, kontrolliert DML datengetrieben für die wichtigsten Prädiktoren von D und Y. Diese Fähigkeit, die wichtigen von den (annähernd) irrelevanten Charakteristiken zu unterscheiden, kann die Schätzung kausaler Effekte in der Praxis enorm vereinfachen. Das ist insbesondere dann der Fall, wenn die Anzahl potenzieller Kontrollvariablen groß ist und (im Fall von Lasso-Regression) arbiträre Interaktionen zwischen den Charakteristiken sowie Nichtlinearitäten in der Schätzung von $g(X)$ und $m(X)$ erlaubt sein sollen. Es gilt aber nach wie vor, dass Annahme 1 (annähernd) für die im Datensatz beobachteten Charakteristiken erfüllt sein muss. Sind wichtige Kontrollvariablen nicht verfügbar, wird auch DML keine akkurate Schätzung liefern. Moderne statistische Verfahren können also niemals Annahmen ersetzen, die die Identifikation eines kausalen Effektes überhaupt erst ermöglichen.

5 Empirische Beispiele

Zur praktischen Illustration von DML werden als Abschluss dieses Beitrags drei empirische Anwendungen aus unterschiedlichen Forschungsfeldern andiskutiert. Knaus (2018) untersucht den Einfluss von musikalischer Betätigung auf die kognitiven und nicht-kognitiven Fähigkeiten von Jugendlichen, gemessen anhand schulischer Leistungstests und Persönlichkeitsmerkmalen. Die Analyse basiert auf knapp 6900 Beobachtungen und fast 380 potenziellen Kontrollvariablen aus dem Jahr 2010 des „Nationalen Bildungspanels“ für Deutschland und die Ergebnisse deuten im Allgemeinen auf positive Effekte von musikalischer Betätigung hin. Yang et al. (2019) evaluieren den Effekt von grossen, sogenannten „Big 4“ Wirtschaftsprüfergesellschaften vs. kleineren Wirtschaftsprüfungskanzleien auf die Qualität von Unternehmensprüfungen. Die Autoren analysieren gut 87.000 Beobachtungen von

Unternehmen zwischen den Jahren 1988 und 2006 aus der Datenbasis „Compustat“ und finden einen ökonomisch bedeutenden Qualitätseffekt. Semenova et al. (2018) untersuchen die Effekte von Produktpreisen auf die Produktnachfrage (also sogenannte Preiselastizitäten) für einen europäischen Lebensmittelgroßhändler. Sie analysieren dazu knapp 2 Millionen wöchentliche Beobachtungen zum Preis und Verkauf von fast 4700 unterschiedlichen Produkten über einen Zeitraum von ca. vier Jahren. Die Studie veranschaulicht, wie sich DML für die Analyse und Verbesserung von Prozessen in Unternehmen einsetzen lässt, etwa zur Optimierung der Preissetzungsstrategie. Die Kombination von Kausalanalyse mit maschinellem Lernen wird deshalb vor dem Hintergrund der steigenden Datenverfügbarkeit nicht nur in der öffentlichen Forschung, sondern auch in der Privatwirtschaft eine immer bedeutendere Rolle einnehmen.

Danksagung Der Autor bedankt sich bei Michael Knaus und Anthony Strittmatter für wertvolle Anregungen.

Literatur

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Heidelberg

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. Econom J 21:C1–C68

Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, Cambridge

Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47:5–86

Knaus MC (2018) A double machine learning approach to estimate the effects of musical practice on student's skills. arXiv 10300:1805

Neyman J (1923) On the application of probability theory to agricultural experiments. Essay on principles. Stat Sci 5:463–480 (Reprint)

Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701

Semenova V, Goldman M, Chernozhukov V, Taddy M (2018) Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. Arbeitspapier, MIT

Tibshirani R (1996) Regresson shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58:267–288

Yang J-C, Chuang H-C, Kuan C-M (2019) Double machine learning with gradient boosting and its application to the big n audit quality effect. Arbeitspapier 19-05, USC Dornsife Institute for New Economic Thinking