Potential Outcome Modell

Über die letzten vier Jahrzehnte entwickelte sich das *Potential Outcome Modell* zu dem Standardwerkzeug bei der Untersuchung von Kausalität in den Sozialwissenschaften. Dieses ist auch bekannt als Kontrafaktisches oder Rubin Modell. Im Zentrum dieses Modells steht die Frage, ob und wenn ja, inwiefern der Eintritt eines Ereignisses im Vergleich zum Ausbleiben desselben die Veränderung einer weiteren Größe einer Einheit hervorruft[1]. Damit dient eine Kausalanalyse immer der Beantwortung von Wenn-Dann-Fragen (Gangl 2010, S. 22-23).

Die zentralen Bausteine dieses Modells sind zum einen der *kausale Zustand* (D) welcher zwei sich ausschließende Ausprägungen beinhaltet[2] und zum anderen das *Outcome* (Y), welches die zu beeinflussende Größe darstellt. Der kausale Zustand kann entweder die Ausprägung *Treatment* (d=1) oder *Kontrolle* (d=0) annehmen (Morgan und Winship 2007, S. 31). Diese beiden Zustände können jeweils ein unterschiedliches Outcome verursachen[3], weshalb die beiden potentiellen Outcomes Y^1 $(Y|d=1)$ und Y^0 $(Y|d=0)$ unterschieden werden müssen (Holland 1986, S. 946).

Der einheitenspezifische kausale Effekt (δ_i) lässt sich aus der Differenz der realisierten Ausprägung der beiden potenziellen Outcomes, y^1 und y^0, je Einheit i ableiten, sodass im linearen Fall

$$\delta i = y^1_i - y^0_i \qquad (1)$$

gilt (Morgan und Winship 2007, S. 31 ff). Von den beiden benötigten Ausprägungen des Outcomes, y^1_i und y^0_i, ist jedoch in der Realität immer nur eine zu beobachten, da eine Einheit entweder den Treatment- oder den Kontroll-Zustand vor Messung des Outcomes annehmen kann. Dieses Problem wird als *Fundamentales Problem der Kausalanalyse* bezeichnet und bedingt, dass Gleichung (1) nicht lösbar ist und damit nie der einheitenspezifische kausale Effekt berechnet werden kann (Holland 1986, S. 947; Morgan und Winship 2007, S. 35).

Die Lösung dieses Problems beruht auf der Berechnung des durchschnittlichen kausalen Effekts (*average treatment effect [ATE]*) für eine Gruppe von Einheiten. Innerhalb dieser Gruppe können beide kausalen Zustände angenommen werden und damit (von unterschiedlichen Einheiten) auch beide potenziellen Outcomes beobachtet werden. Der ATE ist definiert als

$$\Delta_{ATE} \equiv E(\Delta i) = E[\,(y^1_i) - (y^0_i)\,], \qquad (2)$$

wobei E[.] der Erwartungs- bzw. in diesem Fall der Mittelwert der Gruppe ist. Somit entspricht die Differenz der Mittelwerte von Y^1 und Y^0 dem durchschnittlichen kausalen Effekt (Gangl 2010, S. 23). Dieser Zusammenhang gilt für Beobachtungen, bei welchen die potenziellen Outcomes unabhängig ($\coprod$) von der Zuweisung der kausalen Zustände sind und damit

$$(Y^0, Y^1) \coprod D, \qquad (3)$$

gilt[4]. Diese Ignorierbarkeit des Zuweisungsprozesses wird als *Unabhängigkeitsannahme* (*independence assumption [IA]*) bezeichnet und ist per Definition in randomisierten Experimenten gegeben. In

---

[1] In dieser Ausgangssituation wird ein kausaler Effekt bei genau einer Einheit, der einheitenspezifische Kausaleffekt, beschrieben.

2 Im Folgenden wird der vereinfachte Fall eines binären Treatments dargestellt. Die Logik des Potential Outcome Modells lässt sich aber auch auf Fälle eines nicht binären Treatments erweitern, ohne dass sich etwas an den grundlegenden dargestellten Zusammenhängen ändert (Morgan und Winship 2007, S. 31).

3 Ist dies der Fall wird von einem kausalen Effekt von D auf Y gesprochen

[4] Würde der Zuweisungsprozess neben dem kausalen Zustand ebenfalls die potentiellen Outcomes beeinflussen, ergäbe die Differenz der potenziellen Outcomes nicht den ATE. Ein Beispiel: Ein Wert des Zuweisungsprozesses (Z) von 1 verursacht, dass der kausale Zustand (D) und das potenzielle Outcome Y^1

Beobachtungsstudien ist diese Ignorierbarkeit nicht automatisch gegeben und zumeist auch nicht erfüllt. In diesen Fällen kann eine erweiterte Form der IA verwendet werden, die *konditionale Unabhängigkeitsannahme (conditional independence assumption [CIA]*). Diese entspannt IA mittels der Konditionierung auf den Faktor Z, welcher eine Menge von Variablen darstellt, sodass

$(Y^0, Y^1) \perp\!\!\!\perp D \mid Z, \qquad (4)$

gilt. Die Unabhängigkeit muss damit nur noch innerhalb der durch Z definierten Gruppen gelten (Imbens und Wooldridge 2009, S. 12–13). Sofern nicht auf Basis der CIA die Berechnung des kausalen Effekts vorgenommen wird, wird von einem *naiven Schätzer (naive estimator [NE]*) gesprochen.

Die Ergebnisse dieses Schätzers unterliegen einer Verzerrung, die aus zwei Komponenten besteht[5]. Der (1) *selection-bias* gibt die Verzerrung des ATE an, der darauf beruht, dass die Individuen der Treatment- und Kontrollgruppe unterschiedliche Werte der Outcome-Variable vor Zuweisung des kausalen Zustandes aufweisen können. Kann die Wahl des kausalen Zustands weiterhin durch die Einheiten, aufgrund der erwarteten Effekte, selbst beeinflusst werden, ergibt sich zusätzlich ein *self-selection bias*. Dieser beschreibt die potenziell unterschiedliche Auswirkung des Treatment-Zustands auf die Einheiten aus der Treatment-Gruppe gegenüber denen aus der Kontroll-Gruppe (Gangl 2010, S. 25).

Der Prozess der Auswahl geeigneter Variablen Z um die CIA zu erfüllen, wird als *Identifikation des kausalen Effekts* bezeichnet und stellt die Grundlage jeder konsistenten Kausalanalyse dar. Gangl (2010) fast die daraus abzuleitenden Schlussfolgerungen für die empirische Forschung folgendermaßen zusammen:

> *"Compared with this requirement, the usual setup in […] papers that list alternative causes of outcomes in their "theory" sections and then proceed to "test" the relative importance of 'competing hypotheses' by simultaneously including a series of observed variables in a regression specification is woefully inadequate and is eventually unlikely to identify any causal effect of interest." (ebd., S. 27).*

Mit Hilfe der Theorie der *directed acyclic graphs (DAG)* lassen sich die für die Identifikation von kausalen Effekten notwendigen Mechanismen und Regeln visuell darstellen (diese wird im Folgenden beschrieben).

Weitere grundlegende Voraussetzungen zur konsistenten Identifikation von kausalen Effekten sind in der *Stable Unit Treatment Assumption (SUTVA)* zusammengefasst. Diese beinhaltet zwei zentrale Elemente. Erstens müssen die kausalen Effekte jeder einzelnen Einheit unabhängig von dem Zuweisungsmuster der kausalen Zustände der anderen Einheiten sein und zweitens müssen die kausalen Zustände klar definiert sein, sodass die kausalen Zustände keine Untergruppen umfassen, die unterschiedliche Effekte hervorrufen (Rubin 1986, S. 961f).

---

ebenfalls den Wert 1 annehmen. Bei z = 0 nehmen D und Y^0 den Wert 0 an. Obwohl Z die Werte von Y^1 und Y^0 vollständig determiniert, würde fälschlicherweise ein Wert von 1 für den ATE nach Formel (2) berechnet werden.

[5] Durch die Dekomposition von des rechten Teils von Formel (2)

$\Delta ATE \equiv E(Y^1 \mid D = 1) - E(Y^0 \mid D = 0)$

$- [E(Y^0 \mid D = 1) - E(Y^0 \mid D = 0)]$

$- (1 - Pr(D = 1)) * [E(\Delta_i \mid D = 1)$

$- E(\Delta_i \mid D = 0)]$

können die beiden Komponenten isoliert werden. In Zeile 2 ist die Verzerrung aufgrund des selection-bias dargestellt und in Zeile 3 aufgrund des self-selection bias.

Directed Acyclic Graphs

Das dargestellte mathematische Potential Outcome Modell liegt dem nun präsentierten graphischen Modell der DAGs zugrunde, sodass beide Perspektiven zu äquivalenten Ergebnissen führen (Morgan und Winship 2007, S. 61f). Über eine Formalisierung der zugrundeliegenden Annahmen bieten DAGs die Möglichkeit mit Hilfe von Diagrammen kausale Effekte zu identifizieren und eventuelle Verzerrungen festzustellen. Dies geschieht über ein zweistufiges Verfahren. Zunächst wird (1) ein Kausal-Diagramm konstruiert, um daraus (2) eine Identifikationsstrategie des kausalen Effekts ableiten zu können, welche die CIA erfüllt (Pearl 1995, S. 669f).

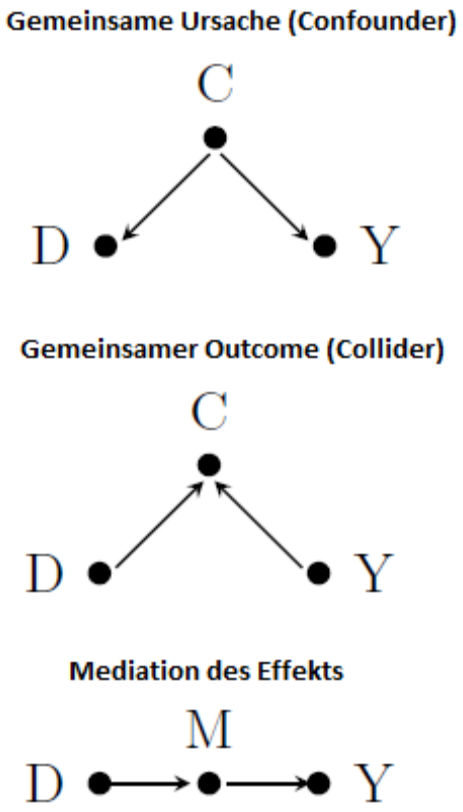

Abbildung 1: Die drei möglichen Zusammenhänge von drei Variablen

Ein DAG, wie in Abbildung 1 (Mediation des Effekts) dargestellt, besteht zunächst aus den beiden interessierenden Faktoren D und Y, welche mit einem ausgefüllten Punkt dargestellt werden, da diese beobachtet wurden. Variablen die nicht beobachtet sind werden mittels eines nicht ausgefüllten Punktes dargestellt. Ein Pfeil zwischen zwei Punkten repräsentiert eine kausale Beziehung beider Variablen und wird als kausaler Pfad bezeichnet. Mittels D -> Y wird also dargestellt, dass D eine Ursache von Y darstellt. Da kausale Effekt nur einseitig wirken können[6], sind in DAGs nur einseitige Pfade erlaubt. Desweitere dürfen durch die kausalen Pfade keine Kreise geschlossen werden.

Zentral ist die Eigenschaft von DAGs, dass nicht nur die Darstellung eines Zusammenhangs als Annahme beurteilt wird, sondern ebenfalls jede nicht eingezeichnete Verbindung zwischen zwei Variablen eine Annahme darstellt (Greenland et al. 1999, S. 38f). DAGs müssen das Wissen der

[6] Häufig wird auch von wechselseitigen Zusammenhängen zwischen zwei Variablen ausgegangen. Die Effekte von zwei Faktoren aufeinander werden jedoch in temporaler Betrachtung nie zeitgleich wirken, sondern zeitversetzt. Damit wäre ein vermeintlich wechselseitiger Einfluss X <-> Z in einem DAG in die zeitliche Abfolge der Wirkungen $X_{t=0}$ -> $Z_{t=1}$ -> $X_{t=2}$ zu zerlegen.

forschenden Person über das Arbeitsgebiet repräsentieren, damit eine Identifikation von kausalen Effekten möglich ist. Jeder Faktor, der die IA/CIA verletzt, verzerrt die kausalen Effekte unabhängig davon, ob dieser dargestellt worden ist oder nicht (Pearl 1995, S. 670).

Von grundlegender Bedeutung für die Identifikation von kausalen Effekten sind die sogenannten *back-door Pfad*e. Dies sind alle Reihen von kausalen Pfaden die den Faktor Y mit dem Faktor X ungeachtet der Pfeilrichtungen verbinden und mit einer Pfeilspitze auf X eingehen (Morgan und Winship 2007, S. 69). Ein Beispiel für einen back-door Pfad ist in Abbildung 1 (Gemeinsame Ursache) dargestellt. Variablen die sowohl einen (vermittelten) Einfluss auf X als auch auf Y ausüben werden als Confounder bezeichnet und rufen Verzerrungen des naiven Schätzers hervor.

Diese back-door Pfade können zwei Zustände aufweisen: *geblockt* oder *offen*, wobei jeder nicht geblockte Pfad offen ist. Ein offener back-door Pfad verletzt die IA/CIA, womit die geschätzten kausalen Effekte einer Verzerrung unterliegen.

Geschlossen ist ein Pfad automatisch, wenn (1) dieser einen Faktor beinhaltet, auf den von beiden Richtungen des Pfades Pfeile zulaufen (siehe Abbildung 1: Collider). Ein solcher Faktor wird als *Collider* bezeichnet. Weiterhin (2) muss ein Pfad geblockt werden, wenn für einen Faktor der auf diesem Pfad liegt und der kein Collider ist, der kausale Effekt konditioniert wird. Mit der Aufnahme eines solchen Faktors in die Menge von Variablen Z aus Gleichung 4 wird der Pfad geblockt.

Für eine konsistente Schätzung des kausalen Effektes muss das sogenannte *back-door Kriterium* erfüllt sein, welches äquivalent zur der IA/CIA ist (Pearl 1995, S. 671). Demnach erfüllt eine Menge von Variablen Z dieses Kriterium, wenn (1) keine der enthaltenden Variablen durch D beeinflusst wird (siehe Abbildung 1: Mediation) und (2) Z jeden back-door Pfad blockt (Pearl 1995, S. 674f).

Das back-door Kriterium beinhaltet die zentrale Einsicht von DAGs, dass die Konditionierung mittels eines Colliders einen zuvor automatisch geblockten Pfad öffnet und damit die CIA verletzt wird.

Praxis der Regressionsanalyse

Die Konditionierung der kausalen Effekte kann in der Praxis mittels der Aufnahme einer Variable in ein Regressionsmodell durchgeführt werden (dies wird als Kontrolle einer Variable bezeichnet). Bei der Aufnahme der Variablen und der Interpretation der Koeffizienten müssen einige Punkte berücksichtigt werden.

Um eine konsistente Schätzung des kausalen Effekts durchführen zu können, muss lediglich für solche Variablen kontrolliert werden, welche das back-door Kriterium und damit die CIA erfüllen (Gangl 2010, S. 27–28). Jede Variable, für die zusätzlich kontrolliert wird, ist eine potenzielle Gefahr für die Gültigkeit der CIA und kann damit eine Verzerrung der Schätzung bedingen (Greenland et al. 1999, S. 44).

Es sollte für keine Variablen kontrolliert werden, die Collidern darstellen. Dies öffnet zuvor verschlossene back-door Pfade und verletzt die IA bzw. CIA. Diese nun geöffneten Pfade müssten durch Kontrolle von geeigneten Variablen, die das back-door Kriterium erfüllen, erneut geschlossen werden.

Weiterhin entsteht durch die Kontrolle von Variablen, die als Mediator den Effekt zwischen D und Y vermitteln, eine Verzerrung des geschätzten kausalen Effekts, da dieser nun nicht mehr dem gesamten kausalen Effekt entspricht, sondern um den kontrollierten indirekten Effekt reduziert wurde.

Die Kontrolle von Variablen die nur D beeinflussen und über keinen Pfad in einer Verbindung mit Y stehen, öffnet zwar keinen back-door Pfad (verletzten damit nicht die IA bzw. CIA), reduziert aber die Varianz von D innerhalb der konditionalen Beziehung zu Y und verursachen damit die Gefahr einen kausalen Effekt zu verdecken (Gangl 2010, S. 28; Greenland et al. 1999, S. 43).

Die Interpretation der mittels Regression geschätzten Koeffizienten, muss sich auf diejenigen Koeffizienten beschränken, welche einen Zusammenhang zwischen zwei Variablen beschreiben, für die das back-door Kriterium erfüllt ist. In der Praxis sind dies zumeist nur die beiden Variablen D und Y und damit nur der Regressionskoeffizient von D. Alle weiteren Koeffizienten unterliegen, aufgrund der fehlenden Aufstellung eines DAGs für die zugrungeliegenden Variablen (z.B. Z_1 -> Y) und dem fehlenden Schließen der back-door Pfade, in der Regel Verzerrungen[7]. Da ohne einen DAG das Ausmaß der Verzerrungen nicht abschätzbar ist, können diese Koeffizienten nicht sinnvoll interpretiert werden (Gangl 2010, S. 28).

Literaturverzeichnis

Gangl, Markus (2010): Causal Inference in Sociological Research. In: *Annu. Rev. Sociol.* 36 (1), S. 21–47. DOI: 10.1146/annurev.soc.012809.102702.

Greenland, Sander; Pearl, Judea; Robins, James M. (1999): Causal Diagrams for Epidemiologic Research. In: *Epidemiology* 10 (1), S. 37–48. DOI: 10.1097/00001648-199901000-00008.

Holland, Paul W. (1986): Statistics and Causal Inference. In: *Journal of the American Statistical Association* 81 (396), S. 945–960. DOI: 10.1080/01621459.1986.10478354.

Imbens, Guido W.; Wooldridge, Jeffrey M. (2009): Recent Developments in the Econometrics of Program Evaluation. In: *Journal of Economic Literature* 47 (1), S. 5–86. DOI: 10.1257/jel.47.1.5.

Morgan, Stephen Lawrence.; Winship, Christopher (2007): Counterfactuals and causal inference. Methods and principles for social research. Cambridge: Cambridge Univ. Press (Analytical methods for social research). Online verfügbar unter https://doi.org/10.1017/CBO9780511804564.

Pearl, Judea (1995): Causal diagrams for empirical research. In: *Biometrika* 82 (4), S. 669–688. DOI: 10.1093/biomet/82.4.669.

Rubin, Donald B. (1986): Comment. Which Ifs Have Causal Answers. In: *Journal of the American Statistical Association* 81 (396), S. 961–962. DOI: 10.1080/01621459.1986.10478355.

[7] Die Folgerung der modernen Kausalanalyse daraus ist, dass je Regression im Normalfall nur ein Kausaleffekt geschätzt und interpretiert werden sollte. Für die konsistente Schätzung und Interpretation von eventuellen Interaktionen oder Effekt-Modifikationen des Kausaleffekts muss das DAG erweitert werden, sodass alle back-door Pfade zwischen D, Y und Interaktionsfaktor/ Effekt-Modifikator identifiziert und geschlossen werden können.