

Sonja Gensler · Bernd Skiera · Martin Böhm

## Einsatzmöglichkeiten der Matching Methode zur Berücksichtigung von Selbstselektion

Eingegangen: 22. November 2004 / Angenommen: 15. Februar 2005  
© Wirtschaftsuniversität Wien, Austria 2005

**Zusammenfassung** Häufig ist es von Interesse, den Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu untersuchen. Um jedoch eine Kausalität adäquat evaluieren zu können, müssen Selbstselektionseffekte berücksichtigt werden. Hierfür wird die Matching Methode vorgeschlagen. Bei der Matching Methode besteht das Ziel darin, durch die Bildung von Paaren von Teilnehmern und Nicht-Teilnehmern den Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu bewerten. Dieser Beitrag stellt unterschiedliche Varianten der Matching Methode vor und vergleicht diese. Der Beitrag zeigt damit, wie bei betriebswirtschaftlichen Problemen Selbstselektionseffekte angemessen berücksichtigt werden können.

**Schlüsselwörter** Selbstselektion · Matching Methode · Kausalität

**Abstract** In many situations it is of interest to determine the impact of a specific treatment on an outcome variable. In order to evaluate the causality between treatment and outcome variable, self-selection effects have to be taken into consideration. An approach to account for these self-selection effects is the matching method. The aim of the matching method is to evaluate the impact of a specific treatment on the outcome variable by building pairs of participants and non-participants of the treatment. This article illustrates and compares different specifications of

---

Wir danken Prof. Dr. Bernd Fitzenberger, Dipl.-Vw, Marco Caliendo sowie den beiden Gutachtern für ihre wertvollen Hinweise zu einer früheren Fassung dieses Aufsatzes. Zudem danken wir dem E-Finance Lab e.V. Frankfurt, das dieses Projekt gefördert hat.

---

S. Gensler (✉) · B. Skiera · M. Böhm  
Lehrstuhl für Betriebswirtschaftslehre, insbesondere Electronic Commerce,  
Johann Wolfgang Goethe-Universität Frankfurt am Main, Mertonstr. 17, 60054 Frankfurt am Main, Germany  
E-mail: sgensler@wiwi.uni-frankfurt.de

the matching method and demonstrates how self-selection can be accounted for in managerial problems.

**Keywords** Selection bias · matching method · causality

**JEL classifications** C40 · C21 · M31

## 1 Einleitung

Häufig ist es von Interesse, den Effekt einer Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu untersuchen. So interessieren beispielsweise in der Volkswirtschaft und Politik folgende Fragestellungen: Verkürzt die Teilnahme an einer Arbeitsbeschaffungsmaßnahme die Dauer bis eine neue Beschäftigung erfolgt? Wirkt sich das Niveau der Schulbildung auf das spätere Einkommen aus? Aber auch in der Betriebswirtschaft finden sich zahlreiche Fragestellungen, die sich mit dem Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable befassen. Hierzu zählen zum Beispiel folgende Fragestellungen: Wirkt sich der Besitz einer Kundenkarte positiv auf den Umsatz des Kunden aus? Hat die Nutzung eines bestimmten Vertriebswegs einen Einfluss auf die Profitabilität eines Kunden? Wie wirkt sich die Innovativität von Unternehmen auf deren Erfolg aus? Inwieweit beeinflusst die Nominierung für einen Academy Award den Erfolg eines Films an den Kinokassen? Welche Wirkung hat die Einführung eines Category Managements auf den Absatz eines Händlers?

Bei all jenen Fragestellungen gilt es, den Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable zu untersuchen, um einen kausalen Zusammenhang zu bestätigen oder zu verwerfen. Wenn eine Kausalität festgestellt werden kann, ergeben sich daraus Implikationen für die Politik oder Management-Entscheidungen. Um jedoch eine Kausalität adäquat evaluieren zu können, ist es von Bedeutung zwischen der Ermittlung und der Interpretation des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu trennen. Denn häufig ist die Zuordnung eines Probanden zu einer Gruppe (Teilnehmer versus Nicht-Teilnehmer an der Maßnahme) nicht zufällig, sondern die Probanden ordnen sich einer Gruppe selbst zu (Selbstselektionseffekt). In solchen Fällen kann ein Vergleich der beiden Gruppen Unterschiede zwischen diesen identifizieren, jedoch kann der Effekt der Teilnahme an einer Maßnahme auf die Ergebnisvariable dann nicht in einfacher Weise untersucht werden. Denn der beobachtete Unterschied zwischen den beiden Gruppen kann sowohl durch den Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable als auch durch weitere nicht berücksichtigte Variablen (Störvariablen), die einen Selbstselektionseffekt abbilden, bedingt sein.

Häufig wird aber in der Unternehmenspraxis und Politik Kausalität unterstellt. So wurde beispielsweise anhand eines Mittelwertvergleichs festgestellt, dass Teilnehmer an einer Arbeitsbeschaffungsmaßnahme im Durchschnitt nach kürzerer Zeit eine neue Arbeitsstelle finden als Nicht-Teilnehmer (z.B. Hujer et al. 1997). Auf Basis dieses Ergebnisses wurde dann häufig die Schlussfolgerung gezogen, dass Arbeitsbeschaffungsmaßnahmen eine wirkungsvolle Maßnahme zur Senkung der Arbeitslosigkeit seien. Diese Schlussfolgerung kann jedoch so nicht erfolgen, da die Entscheidung über die Teilnahme an einer Arbeitsbeschaffungsmaßnahme nicht zufällig erfolgt, sondern maßgeblich von den Betroffenen mitgestaltet wird (Hujer et al. 1997). So belegt beispielsweise die Untersuchung von Hujer et al.

(2003), dass vor allem motivierte Arbeitssuchende an einer Arbeitsbeschaffungsmaßnahme teilnehmen. Diese (motivierten) Teilnehmer hätten aber auch ohne die Arbeitsbeschaffungsmaßnahme mit einer höheren Wahrscheinlichkeit schneller eine neue Arbeitsstelle gefunden als die (weniger motivierten) Nicht-Teilnehmer an der Arbeitsbeschaffungsmaßnahme. Der ermittelte Unterschied kann folglich nicht ausschließlich auf die Teilnahme an der Arbeitsbeschaffungsmaßnahme zurückgeführt werden, sondern wird auch durch die Motivation der Probanden bedingt. Es liegt dann ein Selbstselektionseffekt der Arbeitssuchenden bei der Einteilung der beiden Gruppen „Teilnehmer“ und „Nicht-Teilnehmer“ an der Arbeitsbeschaffungsmaßnahme vor. So erscheinen vor diesem Hintergrund beispielsweise die Investitionen in Höhe von 138 Milliarden Euro für aktive Arbeitsmarktpolitik in Ostdeutschland in einem anderen Licht (Bundesanstalt für Arbeit 2003).

Auch bei Untersuchungen, die den Einfluss der Vertriebswegenutzung auf die Kundenprofitabilität zu bestimmen versuchen, wurde ein kausaler Zusammenhang unterstellt. Ein Mittelwertvergleich hat hier gezeigt, dass Kunden, die das Internet nutzen, eine höhere Profitabilität aufweisen als Kunden, die das Internet nicht nutzen (Wehring 2002; Boston Consulting Group 2001). Daraus wurde die Schlussfolgerung gezogen, dass die Internetnutzung einen positiven Einfluss auf die Profitabilität der Kunden hat. Diese Studien vernachlässigen jedoch, dass sich Kunden, die das Internet nutzen, von Kunden, die das Internet nicht nutzen, signifikant unterscheiden. Es liegt demnach ein Selbstselektionseffekt vor (Degeratu et al. 2000) und es ist nicht möglich, die höhere Profitabilität von Kunden, die das Internet nutzen, gänzlich auf die Internetnutzung zurückzuführen. Dies gilt auch, wenn ein Unternehmen eine Kundenkarte einführt und auf Basis eines Vergleichs der beiden Gruppen „Besitzer der Kundenkarte“ und „Nicht-Besitzer der Kundenkarte“ zu dem Ergebnis kommt, dass die Kunden, die die Kundenkarte besitzen, einen höheren Umsatz tätigen und dass daher die Kundenkarte offensiv vertrieben werden soll. In diesem Beispiel könnte es sein, dass Kunden mit hohem Einkommen in der Regel mehr Umsatz generieren und gleichfalls eine höhere Wahrscheinlichkeit für den Besitz der Kundenkarte haben.

Zur Berücksichtigung des Selbstselektionseffekts ist zum Beispiel in der Medizin (z.B. D'Agostino 1998; Singer 1986) und der Volkswirtschaftslehre (z.B. Ashenfelter und Rouse 1998; Heckman et al. 1996; LaLonde 1986) die so genannte Matching Methode als ein nicht parametrisches Verfahren vorgeschlagen worden. Diese Methode hat jedoch in der Betriebswirtschaftslehre bislang kaum Beachtung gefunden. Ausnahmen stellen beispielsweise Christensen et al. (2004), Hitt und Frei (2002) und Degeratu et al. (2000) dar. Diese Studien zeigen, dass Selbstselektionseffekte existieren und dass die Berücksichtigung dieser Selbstselektionseffekte für eine adäquate Evaluierung der Effektivität von Maßnahmen entscheidend ist.

Aus diesem Grund ist es das Ziel dieses Beitrags, aufzuzeigen, wie Selbstselektionseffekte berücksichtigt werden können, um eine adäquate Evaluierung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu gewährleisten. Hierfür wird die Matching Methode dargestellt und es werden verschiedene Varianten der Durchführung der Matching Methode diskutiert. Bei der Darstellung der Matching Methode besteht das Ziel darin, den Einfluss unterschiedlicher Varianten bei der Anwendung der Matching Methode auf die Evaluierung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable aufzuzeigen. Um dieses Ziel zu erreichen, wird ein illustratives Beispiel aufgegriffen. Letztlich

soll dieser Aufsatz verdeutlichen, inwieweit die Matching Methode einen Beitrag leisten kann, bei betriebswirtschaftlichen Fragestellungen eine angemessene Berücksichtigung des Selbstselektionseffekts zu gewährleisten.

Der Beitrag ist daher wie folgt gegliedert. In Abschnitt 2 wird zunächst das Problem der Evaluierung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable beschrieben. In Abschnitt 3 wird dann die Matching Methode dargestellt und diskutiert. Darauf aufbauend werden dann in Abschnitt 4 Verfahren vorgestellt, die es erlauben, den Effekt der Teilnahme an einer Maßnahme auf die Ergebnisvariable bei Berücksichtigung des Selbstselektionseffekts zu evaluieren. Diese Ausführungen werden anhand eines illustrativen Beispiels verdeutlicht. Der Beitrag schließt in Abschnitt 5 mit einem Fazit.

## 2 Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable

Um die vorangegangenen Fragestellungen bezüglich des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu beantworten, ist der folgende Zusammenhang relevant:

$$\Delta_i = Y_i^1 - Y_i^0 \quad \forall i \in I \quad (1)$$

mit

- $\Delta_i$ : Veränderung des Werts der Ergebnisvariablen für den  $i$ -ten Teilnehmer der Maßnahme,
- $Y_i^1$ : Wert der Ergebnisvariablen für den  $i$ -ten Teilnehmer der Maßnahme,
- $Y_i^0$ : Wert der Ergebnisvariablen für den  $i$ -ten Teilnehmer, wenn dieser nicht an der Maßnahme teilgenommen hätte,
- $I$ : Indexmenge der Teilnehmer an der Maßnahme.

Um den Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu prüfen, wird somit der Wert der Ergebnisvariablen bei Teilnahme an der Maßnahme verglichen mit dem Wert der Ergebnisvariablen, wenn der Teilnehmer nicht an der Maßnahme teilgenommen hätte. Besteht ein signifikanter Unterschied zwischen den beiden Werten der Ergebnisvariablen, so drückt diese Differenz die Größe des Effekts der Teilnahme an der Maßnahme auf die Ergebnisvariable aus. Es wird dann entsprechend dem Kausalitätsbegriff nach Roy (1951) und Rubin (1974) ein kausaler Zusammenhang zwischen der Teilnahme an der Maßnahme und der Ergebnisvariablen postuliert. Dieses Verständnis von Kausalität unterscheidet sich von dem Kausalitätsbegriff nach Granger (1969). Denn Granger geht von Kausalität aus, wenn eine Variable  $y$  bei Berücksichtigung vergangener Werte der Variable  $x$  besser prognostiziert werden kann als wenn diese Variable nicht berücksichtigt wird. Beiden Kausalitätsbegriffen ist jedoch gemeinsam, dass diese eine eindeutige Richtung des Zusammenhangs zugrunde legen. So gehen die Kausalitätsbegriffe nach Roy und Rubin und Granger über einen rein assoziativen Zusammenhang hinaus.

Allerdings ist die Differenz der Werte der Ergebnisvariablen nicht beobachtbar (vgl. Tabelle 1). So ist beispielsweise für einen Kunden, der im Besitz einer Kundenkarte ist, der Umsatz bei Besitz der Kundenkarte beobachtbar, aber der Umsatz dieses Kunden ist nicht beobachtbar, wenn er keine Kundenkarte besitzen würde. Dieser fehlende Wert wird auch als „Counterfactual Outcome“ bezeichnet (Hujer

et al. 2003, S. 11). Daher handelt es sich bei der Evaluierung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable um ein Problem fehlender Daten. Denn für die Teilnehmer kann nicht beobachtet werden, welchen Wert die Ergebnisvariable aufweisen würde, wenn sie nicht an der Maßnahme teilgenommen hätten und für die Nicht-Teilnehmer kann nicht beobachtet werden, welchen Wert die Ergebnisvariable aufweisen würde, wenn sie an der Maßnahme teilgenommen hätten. Dieses Problem fehlender Daten wird auch als fundamentales Evaluierungsproblem bezeichnet (Imbens 2004, S. 5). Aufgrund dieses fundamentalen Evaluierungsproblems ist es im vorangegangenen Beispiel nicht möglich, den individuellen Effekt der Kundenkarte auf den Umsatz eines Kunden direkt zu messen.

**Tabelle 1** Beobachtbarkeit der zustandsabhängigen Werte der Ergebnisvariablen

	Zustandsabhängige Werte der Ergebnisvariablen	
	$Y_{i(j)}^1$ Ergebnis bei Teilnahme	$Y_{i(j)}^0$ Ergebnis bei Nicht-Teilnahme
Teilnehmer $i$	$Y_i^1$ Beobachtbar	$Y_i^0$ Nicht-beobachtbar
Nicht-Teilnehmer $j$	$Y_j^1$ Nicht-beobachtbar	$Y_j^0$ Beobachtbar

mit

- $Y_j^1$ : Wert der Ergebnisvariablen für den  $j$ -ten Nicht-Teilnehmer, wenn dieser an der Maßnahme teilgenommen hätte,  
 $Y_j^0$ : Wert der Ergebnisvariablen für den  $j$ -ten Nicht-Teilnehmer der Maßnahme.

Da das Counterfactual Outcome für die Teilnehmer nicht beobachtbar ist, wird meist der Wert der Ergebnisvariablen für die Nicht-Teilnehmer herangezogen, um den Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable zu evaluieren. Es wird dann ein Mittelwertvergleich für unabhängige Stichproben genutzt (Bortz 1999, S. 137f.). Im Beispiel würde demnach der Umsatz der Kunden, die eine Kundenkarte besitzen, mit dem Umsatz jener Kunden verglichen, die keine Kundenkarte besitzen. Es erfolgt somit kein individueller Vergleich, sondern ein aggregierter Vergleich der beiden Gruppen:

$$\Delta = E[Y_i^1] - E[Y_j^0] \tag{2}$$

mit

- $\Delta$ : durchschnittliche Veränderung des Werts der Ergebnisvariablen,  
 $E[Y_i^1]$ : durchschnittlicher Wert der Ergebnisvariablen für die Teilnehmer an der Maßnahme,  
 $E[Y_j^0]$ : durchschnittlicher Wert der Ergebnisvariablen für die Nicht-Teilnehmer an der Maßnahme.

Der dargestellte Vergleich der Gruppenmittelwerte basiert jedoch auf der Annahme, dass die Zuordnung der Probanden auf die beiden Gruppen zufällig erfolgt und somit Unabhängigkeit der beiden Gruppen vorliegt. In dem angeführten Beispiel der Kundenkarte könnte allerdings das Einkommen der Kunden die Vergabe der

Kundenkarte beeinflussen, wenn diese an die Bonität der Kunden gekoppelt ist. Um diese Effekte kontrollieren zu können, ist es notwendig, dass diese Störvariablen berücksichtigt und als Kontrollvariablen erfasst werden.

Basierend darauf, inwieweit diese Störvariablen kontrolliert werden können, wird zwischen einer experimentellen und einer quasi-experimentellen Untersuchung unterschieden. Bei einer experimentellen Untersuchung ist es möglich, die Probanden zufällig den Gruppen zuzuordnen, so dass die Störvariablen unter den Untersuchungsbedingungen annähernd gleich verteilt sind (Randomisierung der Stichprobe) und somit Unabhängigkeit der Gruppen unterstellt werden kann. Dies impliziert, dass die beiden Gruppen sich hinsichtlich der Störvariablen nicht systematisch unterscheiden. Der oben beschriebene Mittelwertvergleich kann dann weiterhin angewendet werden.

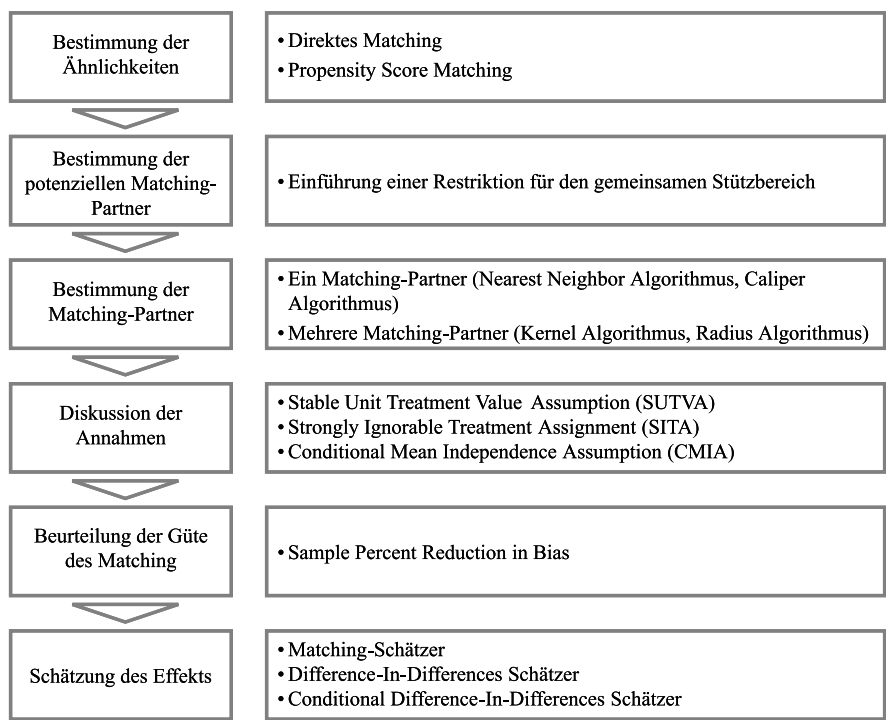
Häufig sind die Gruppen aber bereits gegeben und es ist nicht möglich, die Probanden diesen zufällig zuzuordnen. Da es dann möglich ist, dass das Ergebnis des Mittelwertvergleichs von Störvariablen überlagert wird, gilt es, im Nachhinein ein experimentelles Design nachzubilden, indem eine Berücksichtigung der Störvariablen erfolgt. In diesem Fall liegt eine quasi-experimentelle Untersuchung vor und eine Evaluierung des Effekts der Teilnahme an einer Maßnahme auf die Ergebnisvariable ist dann möglich. In dem angeführten Beispiel würde dies bedeuten, dass das Einkommen eines Kunden sowohl einen Effekt auf den Besitz der Kundenkarte als auch auf den Umsatz eines Kunden hat. Besitzen beispielsweise vor allem Kunden mit hohem Einkommen aufgrund der Vergabekriterien die Kundenkarte, so kann ein beobachteter höherer Umsatz der Kunden mit Kundenkarte im Vergleich zu jenen Kunden ohne Kundenkarte nicht ausschließlich dem Besitz der Kundenkarte zugerechnet werden. Vielmehr liegt dann ein Selbstselektionseffekt vor (Lee 2000, S. 383). Ist nur von Interesse, ob die Kunden, die eine Kundenkarte besitzen einen höheren Umsatz erzielen, so ist es nicht erforderlich dem Selbstselektionseffekt Rechnung zu tragen. Ist aber hingegen der Effekt des Besitzes der Kundenkarte auf den Umsatz eines Kunden von Interesse, so ist die Berücksichtigung der Störvariablen in Form von Kontrollvariablen bei der Evaluierung des Effekts von Bedeutung, um den Selbstselektionseffekt zu berücksichtigen.

So kann die Validität quasi-experimenteller Untersuchungen erhöht werden, indem die zu vergleichenden Gruppen nach allen relevanten Störvariablen parallelisiert werden (matched sample). Im Folgenden wird daher die Matching Methode beschrieben, die eine Parallelisierung der beiden Gruppen erlaubt und somit einem möglichen Selbstselektionseffekt Rechnung trägt.

### 3 Matching Methode

Die Matching Methode wurde bisher im ökonomischen Bereich vor allem in der volkswirtschaftlichen Literatur angewendet und nimmt eine Parallelisierung von zwei Gruppen vor (z.B. Heckman 1976; Heckman 1978; Ashenfelter 1978; Heckman 1979; Bjorklund und Moffitt 1987; Lechner 2002). Ziel der Matching Methode ist es, systematische Unterschiede in den Störvariablen zwischen der Gruppe der Teilnehmer und der Nicht-Teilnehmer an einer Maßnahme zu eliminieren (Heckman et al. 1996). So kann die nicht-zufällige Zuordnung der Probanden zu den Gruppen beseitigt und das Design einer experimentellen Untersuchung nachgebildet werden. Um dieses Ziel zu erreichen, werden „Zwillingspaare“ aus der Gruppe

der Teilnehmer und Nicht-Teilnehmer gebildet, die sich bezüglich der relevanten Störvariablen gleichen und sich nur bezüglich des Teilnahmestatus unterscheiden (Hujer et al. 2003, S. 18). Somit werden jedem Teilnehmer ein oder mehrere Nicht-Teilnehmer als Matching-Partner zugeordnet. Auf Basis der ermittelten Matching-Partner kann dann das Counterfactual Outcome bestimmt und so der Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable evaluiert werden. Die Bestimmung der Matching-Partner führt dazu, dass sich die Verteilungen der Störvariablen der Teilnehmer und Nicht-Teilnehmer an der Maßnahme angleichen. Mit anderen Worten, systematische Unterschiede in den Störvariablen der Teilnehmer und Nicht-Teilnehmer werden so eliminiert und damit die bei einem Experiment vorliegenden Bedingungen erreicht. Ein anschließender Mittelwertvergleich erlaubt dann eine unverzerrte Schätzung des Effekts der Teilnahme an der Maßnahme auf die Ergebnisvariable. Im Folgenden wird nun die Vorgehensweise der Matching Methode detailliert beschrieben (siehe Abb. 1).



**Abb. 1** Vorgehensweise der Matching Methode

### 3.1 Bestimmung der Ähnlichkeiten der Teilnehmer und Nicht-Teilnehmer

Um die Matching-Partner zu identifizieren, ist es erforderlich, die Ähnlichkeit der Teilnehmer und Nicht-Teilnehmer zu spezifizieren. Dabei kann unterschieden werden, ob ein Matching der Teilnehmer und Nicht-Teilnehmer auf der Basis einzelner Störvariablen erfolgt (direktes Matching) oder ob ein so genannter Propensity Score

für das Matching herangezogen wird. In beiden Fällen ist es erforderlich, zunächst jene Variablen zu identifizieren, die die Teilnahme an der Maßnahme und die Ergebnisvariable beeinflussen und daher den Selbstselektionseffekt determinieren. Bei einem direkten Matching werden dann diese Störvariablen herangezogen, um die Matching-Partner zu finden. Hierbei gilt es für jeden Teilnehmer einen Nicht-Teilnehmer zu identifizieren, dessen Ausprägungen der Störvariablen mit jenen des Teilnehmers übereinstimmen. Dieses Vorgehen eignet sich für Anwendungen, bei denen nur wenige Störvariable zu berücksichtigen sind. Bei einer Vielzahl von Störvariablen erweist sich das direkte Matching jedoch als nicht praktikabel, denn die Anwendung führt zu einem Dimensionalitätsproblem (Hujer et al. 2001, S. 178). Dieses Dimensionalitätsproblem besteht darin, dass es sich sehr schwierig gestaltet einen Matching-Partner zu identifizieren, der in allen Störvariablen dem Teilnehmer an der Maßnahme gleicht. Eine Berücksichtigung aller Störvariablen ist jedoch erforderlich, um den Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable adäquat evaluieren zu können. Es kommt dann zu dem folgenden Trade-Off: Je mehr beobachtete Störvariablen berücksichtigt werden, desto besser ist die Evaluierung des Selbstselektionseffekts, desto wahrscheinlicher ist es aber auch, dass für einen Teilnehmer kein geeigneter Matching-Partner gefunden wird.

Rosenbaum und Rubin (1983) schlagen daher zur Lösung des Dimensionalitätsproblems das Propensity Score Matching vor. Dabei werden die Matching-Partner in der Art bestimmt, dass sich deren Propensity Scores  $P(X)$  annähernd entsprechen. Der Propensity Score  $P(X)$  ist eine Funktion der Störvariablen  $X$  und ist definiert als die Wahrscheinlichkeit der Teilnahme an der Maßnahme. Bei der Bildung von Matching-Partnern werden also alle relevanten und beobachteten Störvariablen  $X$  indirekt durch deren Einfluss auf den Propensity Score  $P(X)$  berücksichtigt. Daher reduziert sich das Problem des Auffindens eines Matching-Partners auf eine Dimension und zwar auf den Wert des Propensity Scores  $P(X)$  (D'Agostino 1998, S. 2267). Der Propensity Score wird üblicherweise mittels Probit- oder Logit-Modellen (siehe Gleichung 3) geschätzt, bei denen die abhängige Variable die getroffene Teilnahmeentscheidung darstellt (Dehejia und Wahba 2002).

$$P_h(X) = \frac{1}{1 + \exp(\beta' \cdot X_h)} \quad \forall h \in H \quad (3)$$

mit

- $P_h(X)$ : Wahrscheinlichkeit, dass der  $h$ -te Proband an der Maßnahme teilnimmt,
- $\beta$ : Vektor der Parameter für die Störvariablen,
- $X_h$ : Vektor der Ausprägungen der Störvariablen für den  $h$ -ten Probanden,
- $H$ : Indexmenge der Probanden (entspricht der Menge aller Teilnehmer und Nicht-Teilnehmer:  $H = I \cup J$ ).

Da die Bildung des Propensity Scores die Grundlage für das Auffinden der Matching-Partner ist, ist es empfehlenswert, die Spezifikation des Modells zu testen (Lechner 1998; Cox und Wermuth 2004). So ist es zum Beispiel sinnvoll, die Signifikanz der berücksichtigten Störvariablen zu überprüfen.

In Anlehnung an das Beispiel der Kundenkarte würde die Variable „Besitz der Kundenkarte“ als abhängige Variable für den Propensity Score verwendet werden. Die unabhängigen Variablen werden durch die beobachteten Störvariablen repräsentiert. Hierfür könnten beispielsweise das Einkommen der Kunden oder weitere Variablen wie das Geschlecht oder die Wohnregion in Betracht kommen. Es kann



nun das Modell geschätzt werden, das jedem Teilnehmer und Nicht-Teilnehmer einen Propensity Score zuordnet und damit eine Wahrscheinlichkeit für den Besitz der Kundenkarte. Statt nun die einzelnen Störvariablen heranzuziehen, um die Matching-Partner zu identifizieren, wird lediglich der Propensity Score betrachtet. So wird durch die Bildung des Propensity Scores das mehrdimensionale Problem des direkten Matching auf eine Dimension reduziert, was häufig auch in einer Erhöhung der Anzahl der einem Teilnehmer ähnlichen Nicht-Teilnehmer resultiert. Dieser Effekt tritt ein, da nun mit dem Propensity Score eine einzige Variable und nicht mehrere Störvariablen zur Bestimmung der Ähnlichkeit der Teilnehmer und Nicht-Teilnehmer herangezogen wird.

Aufgrund der kompensatorischen Beziehungen zwischen den unabhängigen Variablen im Logit- oder Probit-Modell ist es jedoch möglich, dass Matching-Partner in ihren Propensity Scores übereinstimmen, jedoch nicht in den Ausprägungen ihrer Störvariablen. Aus diesem Grund wird in der neueren Literatur häufig ein hybrider Ansatz angewendet, bei dem neben dem Propensity Score auch einzelne Störvariablen zum Matching herangezogen werden (Lechner 1998). Dies soll dafür Sorge tragen, dass sich die Matching-Partner nicht nur im Propensity Score, sondern auch in einzelnen Störvariablen ähnlich sind.

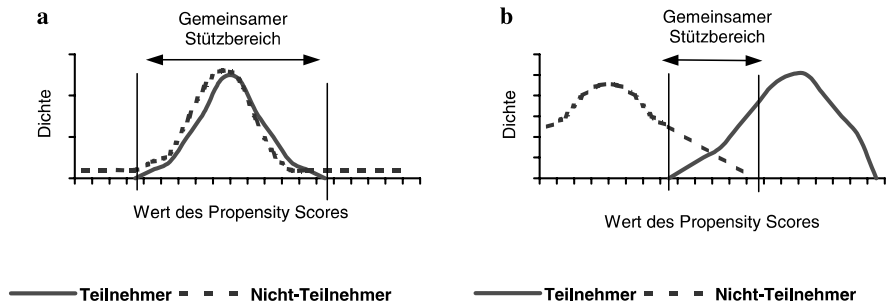
Da es sich bei dem Propensity Score um eine Wahrscheinlichkeit handelt, liegen im Intervall  $[0, 1]$  unendlich viele Ausprägungen des Propensity Scores vor. Aus diesem Grund ist es üblicherweise bei einem Matching auf Basis des Propensity Scores nicht möglich, eine exakte Übereinstimmung zwischen dem Propensity Score eines Teilnehmers und eines Nicht-Teilnehmers zu erreichen. Mehrere Algorithmen stehen nun zur Verfügung, um Matching-Partner zu identifizieren, die keine exakte Übereinstimmung zwischen den Propensity Scores voraussetzen. Hierbei gilt es die Menge der Nicht-Teilnehmer zu identifizieren, die der Menge der Teilnehmer möglichst ähnlich ist. Die Ähnlichkeit wird durch eine geringe Differenz zwischen den Propensity Scores der Teilnehmer und Nicht-Teilnehmer ausgedrückt. Im Folgenden wird daher zunächst beschrieben, wie die Menge an potenziellen Matching-Partnern für die Teilnehmer an der Maßnahme bestimmt werden kann und dann werden die am häufigsten eingesetzten Algorithmen zur Bestimmung der Matching-Partner vorgestellt.

### 3.2 Bestimmung der potenziellen Matching-Partner

Die Propensity Scores der Teilnehmer und Nicht-Teilnehmer bilden zwei Verteilungen. Um geeignete Matching-Partner zu identifizieren, ist es erforderlich, dass sich die Wertebereiche der Verteilungen der Propensity Scores der Teilnehmer und Nicht-Teilnehmer überschneiden. Denn nur dann ist sichergestellt, dass hinreichend ähnliche Matching-Partner identifiziert werden können. Dieser Überschneidungsbereich der beiden Verteilungen wird als gemeinsamer Stützbereich (Region of Common Support) bezeichnet (Heckman et al. 1998).

Um sicher zu stellen, dass sich die Matching-Partner möglichst ähnlich sind, kann eine Restriktion für den gemeinsamen Stützbereich eingeführt werden. In diesem Fall werden die Nicht-Teilnehmer und Teilnehmer eliminiert und damit nicht zum Matching herangezogen, deren Propensity Scores außerhalb des gemeinsamen Stützbereichs liegen. Dieser Zusammenhang wird in Abb. 2 a,b verdeutlicht. So kann die Einführung einer Restriktion für den gemeinsamen Stützbereich hel-

fen, ähnliche Matching-Partner zu identifizieren. Es muss jedoch beachtet werden, dass die Einführung einer Restriktion für den gemeinsamen Stützbereich dazu führen kann, dass lediglich der Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable für einen Teil der Stichprobe bestimmt wird.



**Abb. 2 a,b** Darstellung des gemeinsamen Stützbereichs. **a** Große Überschneidung der Verteilungen. **b** Geringe Überschneidung der Verteilungen

### 3.3 Bestimmung der Matching-Partner

Bei der Bestimmung der Matching-Partner kann grundsätzlich unterschieden werden, ob nur ein Nicht-Teilnehmer oder ob mehrere Nicht-Teilnehmer als Matching-Partner herangezogen werden. Zudem kann differenziert werden, ob eine maximale Distanz zwischen den Matching-Partnern berücksichtigt wird und ob ein Matching mit oder ohne Zurücklegen der Nicht-Teilnehmer erfolgt. So können die Alternativen zur Bestimmung der Matching-Partner, wie in Abb. 3 dargestellt, unterteilt werden.

#### 3.3.1 Berücksichtigung eines Matching-Partners

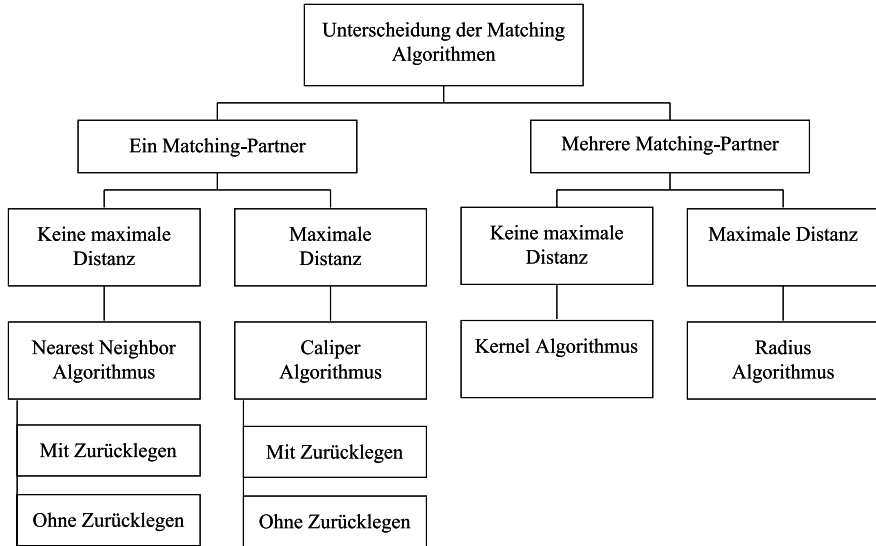
Bei der Berücksichtigung genau eines Nicht-Teilnehmers als Matching-Partner für einen Teilnehmer an der Maßnahme kann zwischen dem Nearest Neighbor Algorithmus und dem Caliper Algorithmus unterschieden werden.

Beim Nearest Neighbor Algorithmus wird jedem Teilnehmer  $i$  jener Nicht-Teilnehmer  $j$  als Matching-Partner zugeordnet, der die geringste Distanz zu dem Teilnehmer aufweist. So wird die Nachbarschaft eines Teilnehmers  $C_i$  durch den Nicht-Teilnehmer definiert, dessen Distanz zum Propensity Score des Teilnehmers am geringsten ist:

$$C_i = \left\{ j \mid \|P_i(X) - P_j(X)\| = \min_{j' \in J} \|P_i(X) - P_{j'}(X)\| \right\} \quad \forall i \in I. \quad (4)$$

Diesem Matching-Partner  $j$  wird das Gewicht  $w(i, j) = 1$  zugewiesen. Demnach ist  $w(i, j)$  gegeben durch (Heckman et al. 1998, S. 1024):

$$w_{i,j}^{NN} = \begin{cases} 1, & \text{falls } j \in C_i \\ 0, & \text{sonst} \end{cases} \quad \forall i \in I, \forall j \in J. \quad (5)$$



**Abb. 3** Algorithmen zur Bestimmung der Matching-Partner

Es ist möglich, dass mehr als ein Nicht-Teilnehmer in der Menge  $C_i$  enthalten ist. Dieser Fall wird jedoch in der Literatur bislang nicht thematisiert. In diesem besonderen Fall wird daher die Auswahl des Matching-Partners durch die spezifische Umsetzung des Algorithmus in der verwendeten Software getroffen. Die Auswahl des Matching-Partners kann zum Beispiel nach der Reihenfolge der gefundenen Nachbarn erfolgen. So wird dann bei mehreren gleich weit entfernten Nachbarn beispielsweise der zuerst identifizierte Nachbar herangezogen.

Im Gegensatz zum Nearest Neighbor Algorithmus wird beim Caliper Algorithmus eine maximale Distanz der Matching-Partner berücksichtigt (Dehejia und Wahba 2002, S. 153). Es wird ein Toleranzwert  $\varepsilon$  eingeführt:

$$C_i = \left\{ j \mid \left\| P_i(X) - P_j(X) \right\| = \min_{j' \in J} \left\| P_i(X) - P_{j'}(X) \right\| \wedge \left\| P_i(X) - P_j(X) \right\| < \varepsilon \right\} \quad \forall i \in I. \quad (6)$$

Diesen Toleranzwert darf die Differenz der Propensity Scores  $P_i(X)$  und  $P_j(X)$  nicht überschreiten. Ist dieses Kriterium für keinen Nicht-Teilnehmer erfüllt, weist also die in (6) dargestellte Menge keine Elemente auf, dann wird der Teilnehmer  $i$  ausgeschlossen. Weist die Menge  $C_i$  allerdings einen Nicht-Teilnehmer als Element auf, dann wird diesem ein Gewicht von eins zugewiesen, wohingegen allen anderen Nicht-Teilnehmern ein Gewicht von Null zugewiesen wird (Heckman et al. 1998, S. 1024). So gilt:

$$w_{i,j}^{CM} = \begin{cases} 1, & \text{falls } j \in C_i \\ 0, & \text{sonst} \end{cases} \quad \forall i \in I, j \in J. \quad (7)$$

So entspricht das Ergebnis des Caliper Algorithmus dann dem des Nearest Neighbor Algorithmus, wenn der Toleranzwert in der Art und Weise festgelegt wird, dass immer ein Nicht-Teilnehmer als Matching-Partner identifiziert wird.

Zudem besteht ein Zusammenhang zwischen der Berücksichtigung einer Restriktion für den gemeinsamen Stützbereich und der Höhe des Toleranzwerts. Bei einem geringen Wert für den Toleranzwert und keiner Berücksichtigung einer Restriktion für den gemeinsamen Stützbereich ist gewährleistet, dass die Matching-Partner sich relativ ähnlich sind, auch wenn diese außerhalb der Restriktion für den gemeinsamen Stützbereich liegen. Dennoch ist bei einem sehr geringen Wert für den Toleranzwert auch bei einem großen gemeinsamen Stützbereich nicht garantiert, dass ein Matching-Partner identifiziert werden kann.

Bei einem Matching ohne Zurücklegen wird jeder Nicht-Teilnehmer maximal einmal als Matching-Partner herangezogen. Weisen die Teilnehmer überwiegend hohe Werte, die Nicht-Teilnehmer hingegen überwiegend geringe Werte für den Propensity Score auf, dann werden zunächst die wenigen Nicht-Teilnehmer mit hohen Werten für den Propensity Score als Matching-Partner herangezogen. Im weiteren Verlauf stehen in der Gruppe der Nicht-Teilnehmer jedoch nur noch Teilnehmer mit einem geringen Wert für den Propensity Score zur Verfügung. Dies führt dazu, dass für die Teilnehmer zunehmend unähnlichere Nicht-Teilnehmer als Matching-Partner ausgewählt werden oder bei Anwendung des Caliper Algorithmus kein Nicht-Teilnehmer mehr als Matching-Partner identifiziert werden kann. Auch wird durch die Reihenfolge der Zuordnung der Matching-Partner die Auswahl der Matching-Partner beeinflusst. Um dem Problem des Auffindens immer unähnlicherer Matching-Partner zu begegnen, kann ein Matching mit Zurücklegen durchgeführt werden.

Bei einem Matching mit Zurücklegen kann ein Nicht-Teilnehmer mehrfach als Matching-Partner für verschiedene Teilnehmer herangezogen werden. Weisen zum Beispiel zwei Teilnehmer der Maßnahme den gleichen Propensity Score auf, können sie auch dem gleichen Nicht-Teilnehmer zugeordnet werden. Aus diesem Grund ist das Matching mit Zurücklegen generell mit einer größeren Ähnlichkeit der Matching-Partner verbunden (Smith und Todd 2000, S. 1). Andererseits birgt das Matching mit Zurücklegen, insbesondere bei extrem unterschiedlichen Verteilungen der Propensity Scores der Teilnehmer und Nicht-Teilnehmer, das Problem, dass wenige Nicht-Teilnehmer sehr häufig als Matching-Partner verwendet werden. Dies führt dazu, dass die Evaluierung des Effekts der Teilnahme an der Maßnahme auf die Ergebnisvariable auf Basis einer sehr kleinen Stichprobe von Nicht-Teilnehmern erfolgt. Dadurch sinkt jedoch die Stabilität der Schätzung des Effekts der Teilnahme an der Maßnahme auf die Ergebnisvariable. Wird zum Beispiel ein Nicht-Teilnehmer aus der Datenbasis entfernt, kann dies zu starken Änderungen des Werts des ermittelten Effekts der Teilnahme an der Maßnahme auf die Ergebnisvariable führen.

### *3.3.2 Berücksichtigung mehrerer Matching-Partner*

Bei der Berücksichtigung mehrerer Nicht-Teilnehmer als Matching-Partner für einen Teilnehmer an der Maßnahme wird zwischen dem Kernel Algorithmus und dem Radius Algorithmus unterschieden.

Der Kernel Algorithmus geht von der Annahme aus, dass jeder Nicht-Teilnehmer zumindest in gewissem Umfang als Matching-Partner geeignet ist. So werden für jeden Teilnehmer alle Nicht-Teilnehmer als Matching-Partner herangezogen (Hujer et al. 2001, S. 180).

$$C_i = \{J\} \quad \forall i \in I \quad (8)$$

Dabei erhalten Nicht-Teilnehmer, die eine geringe Distanz zu einem Teilnehmer aufweisen, ein hohes Gewicht und Nicht-Teilnehmer mit einer großen Distanz bekommen ein geringes Gewicht zugewiesen (Hujer et al. 2003, S. 68).

$$w_{i,j}^{KM} = \frac{K\left(\frac{P_i(X) - P_j(X)}{\tau}\right)}{\sum_{j' \in C_i} K\left(\frac{P_i(X) - P_{j'}(X)}{\tau}\right)} \quad \forall i \in I, j \in J \quad (9)$$

mit

$\tau$ : Bandbreitenparameter,  
 $K(\cdot)$ : Kernel-Funktion (z.B. Gauß'sche Normalverteilung).

Der Bandbreitenparameter ist durch den Anwender festzulegen (Smith und Todd 2000). Dieser beeinflusst das Gewicht, das ein Nicht-Teilnehmer als Matching-Partner für einen Teilnehmer an der Maßnahme erhält. So führt eine Senkung des Bandbreitenparameters bei einer Normalverteilung als Kernel-Funktion dazu, dass Nicht-Teilnehmer mit einer größeren Distanz zu dem Teilnehmer ein geringeres Gewicht erhalten. Dennoch berücksichtigt der Kernel Algorithmus keine maximale Distanz zwischen den Matching-Partnern, sondern zieht alle Nicht-Teilnehmer heran. Dagegen wird beim Radius Algorithmus eine maximale Distanz der Matching-Partner berücksichtigt. So ergibt sich die Nachbarschaft eines Teilnehmers bei dem Radius Algorithmus aus:

$$C_i = \left\{ j \mid \|P_i(X) - P_j(X)\| < \varepsilon \right\} \quad \forall i \in I. \quad (10)$$

Es werden dann alle Nicht-Teilnehmer gleich gewichtet als Matching-Partner herangezogen, deren Distanz zu einem Teilnehmer den Toleranzwert nicht überschreiten:

$$w_{i,j}^{RM} = \begin{cases} \frac{1}{|C_i|}, & \text{falls } j \in C_i \\ 0, & \text{sonst} \end{cases} \quad \forall i \in I, j \in J \quad (11)$$

mit

$|C_i|$ : Anzahl der Elemente in der Menge  $C_i$ .

Eine abschließende Beurteilung der unterschiedlichen Algorithmen ist nicht möglich, da es von der zu untersuchenden Fragestellung abhängt, welcher der vorgestellten Algorithmen sich als vorteilhaft erweist. Häufig angewendet werden der Nearest Neighbor Algorithmus und der Kernel Algorithmus (z.B. Lechner 2002; Black und Smith 2004), da die Bestimmung des Toleranzwerts durch den Anwender beim Caliper und Radius Algorithmus kritisch ist.

### 3.4 Annahmen der Matching Methode

Der Matching Methode liegen die folgenden Annahmen zugrunde:

- Stable Unit Treatment Value Assumption (SUTVA)
- Strongly Ignorable Treatment Assignment (SITA)
- Conditional Mean Independence Assumption (CMIA).

Die als Stable Unit Treatment Value Assumption (SUTVA) bezeichnete Annahme unterstellt, dass die Teilnahme eines Probanden an der Maßnahme ausschließlich das Verhalten dieses Probanden beeinflusst und keinen Einfluss auf das Verhalten und die Ergebnisvariable anderer Probanden hat (Rubin 1990). So wird im Beispiel unterstellt, dass ein Kunde, der eine Kundenkarte besitzt, nicht das Teilnahmeverhalten und den Umsatz eines anderen Kunden beeinflusst. Diese Annahme ist dann kritisch, wenn beispielsweise Weiterempfehlungen oder auch Netzeffekte von Bedeutung sind. Empfiehlt ein Besitzer der Kundenkarte diese weiter und erwerben dann weitere Kunden die Kundenkarte, so kann nicht mehr davon ausgegangen werden, dass die Stable Unit Treatment Value Assumption erfüllt ist. Der Vergleich eines Teilnehmers mit einem gematchten Nicht-Teilnehmer würde dann vernachlässigen, dass sich die Teilnahme auch in einem anderen Verhalten von weiteren Teilnehmern niederschlägt.

Die Strongly Ignorable Treatment Assignment (SITA) Annahme besagt, dass nach dem Matching auf die beobachteten Störvariablen die Werte der Ergebnisvariablen nicht durch die Störvariablen beeinflusst werden:

$$(Y_h^1, Y_h^0) \perp D_h \mid P(X) \quad \forall h \in H \quad (12)$$

mit

- $D_h$ : Teilnahmestatus an einer Maßnahme des  $h$ -ten Probanden,
- $Y_h^1$ : Wert der Ergebnisvariablen für den  $h$ -ten Probanden, wenn er an der Maßnahme teilnimmt,
- $Y_h^0$ : Wert der Ergebnisvariablen für den  $h$ -ten Probanden, wenn er an der Maßnahme nicht teilnimmt.

Wenn die SITA Annahme erfüllt ist, kann der Wert der Ergebnisvariablen für die Nicht-Teilnehmer als Counterfactual Outcome für den Wert der Ergebnisvariablen der Gruppe der Teilnehmer verwendet werden. Ist die SITA Annahme hingegen nicht erfüllt, so kann der Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable nicht angemessen evaluiert werden, da eine Konfundierung vorliegt.

Diese Annahme ist jedoch eine sehr strenge Annahme und Heckman et al. (1999) zeigen, dass die so genannte Conditional Mean Independence Assumption (CMIA) ausreicht, um den Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable adäquat evaluieren zu können. Diese Annahme unterstellt lediglich:

$$E(Y^1 \mid D, P(X)) = E(Y^1 \mid P(X)) \text{ bzw. } E(Y^0 \mid D, P(X)) = E(Y^1 \mid P(X)). \quad (13)$$

Es werden folglich nur noch die Erwartungswerte der Ergebnisvariablen betrachtet. Ist diese Annahme nicht erfüllt, dann erfolgt lediglich eine verzerrte Schätzung des Effekts der Teilnahme an einer Maßnahme auf die Ergebnisvariable.

Wie im Beispiel der Kundenkarte bereits beschrieben, könnten sich die Besitzer der Kundenkarte systematisch von den Kunden ohne Kundenkarte hinsichtlich ihres Einkommens unterscheiden. Diese Störvariable hat jedoch nicht nur einen Effekt auf die Teilnahmeentscheidung, sondern auch auf die Ergebnisvariable Umsatz. Aus diesem Grund kann nicht eindeutig bestimmt werden, ob eine Änderung der Ergebnisvariablen der Teilnahme an der Maßnahme oder der Störvariablen Einkommen zugeschrieben werden muss. Durch eine Berücksichtigung des Einkommens als Störvariable bei dem Matching wird der Effekt dieser Störvariablen auf die Gruppenzugehörigkeit eliminiert. Somit besitzt das Einkommen keinen weiteren Einfluss auf den Besitz einer Kundenkarte und der Wert der Ergebnisvariablen ist nun unabhängig von dem Besitz einer Kundenkarte.

### 3.5 Beurteilung der Güte des Matching

Die Güte des Matching hängt davon ab, inwieweit eine Angleichung der Verteilungen der Störvariablen in den beiden Gruppen der Teilnehmer und Nicht-Teilnehmer erreicht werden kann. Das Sample Percent Reduction in Bias ist eine Messgröße, mit der geprüft werden kann, wie stark sich die Störvariablen von Teilnehmern und Nicht-Teilnehmern nach dem Matching angeglichen haben (Rosenbaum und Rubin 1985). Dazu wird für jede einzelne Störvariable der Mittelwert der Teilnehmer und Nicht-Teilnehmer vor und nach Matching verglichen.

$$SB_n = 1 - \frac{|\bar{x}_{i,n}^N - \bar{x}_{j,n}^N|}{|\bar{x}_{i,n}^V - \bar{x}_{j,n}^V|} \quad \forall n \in N \quad (14)$$

mit

$SB_n$ : Percent Reduction in Bias für die  $n$ -te Störvariable,

$\bar{x}_{i,n}^N$ : Mittelwert der  $n$ -ten Störvariablen für die Teilnehmer an der Maßnahme nach Matching,

$\bar{x}_{j,n}^N$ : Mittelwert der  $n$ -ten Störvariablen für die Nicht-Teilnehmer an der Maßnahme nach Matching,

$\bar{x}_{i,n}^V$ : Mittelwert der  $n$ -ten Störvariablen für die Teilnehmer an der Maßnahme vor Matching,

$\bar{x}_{j,n}^V$ : Mittelwert der  $n$ -ten Störvariablen für die Nicht-Teilnehmer an der Maßnahme vor Matching,

$N$ : Indexmenge aller Störvariablen.

Das Sample Percent Reduction in Bias liegt in der Regel im Intervall von 0 % bis 100 %, da durch das Matching die Differenz der Mittelwerte kleiner wird und somit die Ähnlichkeit der Teilnehmer und Nicht-Teilnehmer bezüglich einer Störvariablen erhöht wird.

## 4 Schätzung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable

Im Folgenden wird nun beschrieben, wie nach dem Auffinden der Matching-Partner der Effekt der Teilnahme an der Maßnahme auf den Wert der Ergebnisvariablen

geschätzt werden kann. Hierfür werden die folgenden Schätzer herangezogen: der Matching-Schätzer, der Difference-In-Differences Schätzer und der Conditional Difference-In-Differences Schätzer. Diese Schätzer haben das Ziel, den durchschnittlichen Effekt der Teilnahme an der Maßnahme auf die Ergebnisvariable – den so genannten durchschnittlichen Treatment-on-Treated Effekt (Average Treatment-on-Treated Effect, kurz: ATTE) – zu schätzen:

$$\Delta^{\text{ATTE}} = E[Y_i^1] - E[Y_i^0] \quad (15)$$

mit

$\Delta^{\text{ATTE}}$ : durchschnittlicher Treatment-on-Treated Effekt.

Der Effekt der Teilnahme bei den Nicht-Teilnehmern an der Maßnahme auf die Ergebnisvariable kann analog basierend auf  $E[Y_j^1] - E[Y_j^0]$  ermittelt werden. Jedoch wird in den empirischen Untersuchungen meist auf den durchschnittlichen Treatment-on-Treated Effekt fokussiert. Aus diesem Grund wird im Folgenden nur dieser Effekt – wie in Gleichung (15) dargestellt – betrachtet.

#### 4.1 Matching-Schätzer

Der Matching-Schätzer basiert auf einem Vergleich der durch das Matching identifizierten Partner von Teilnehmern und Nicht-Teilnehmern an der Maßnahme. Dabei setzt die Anwendung des Matching-Schätzers voraus, dass die bereits in Abschnitt 3.4 beschriebene Conditional Mean Independence Assumption (CMIA) erfüllt ist (Rässler 2002, S. 3). Es gilt dann:

$$E[Y_i^0 | P(X)] = E[Y_j^0 | P(X)] \quad (16)$$

mit

$E[Y_i^0 | P(X)]$ : durchschnittlicher Wert der Ergebnisvariablen für die Teilnehmer nach Matching, wenn diese nicht an der Maßnahme teilgenommen hätten (Counterfactual Outcome),

$E[Y_j^0 | P(X)]$ : durchschnittlicher Wert der Ergebnisvariablen für die Nicht-Teilnehmer nach Matching.

Der in Gleichung (16) dargestellte Zusammenhang drückt aus, dass das Counterfactual Outcome für die Teilnehmer an der Maßnahme durch den beobachteten Wert für die Nicht-Teilnehmer ausgedrückt werden kann. Dieser Zusammenhang basiert auf der Annahme, dass der erwartete Wert der Ergebnisvariablen nach dem Matching unabhängig von der Teilnahme an der Maßnahme ist.

So kann der geschätzte durchschnittliche Treatment-on-Treated Effekt in der folgenden Weise ermittelt werden:

$$\begin{aligned} \Delta^{\text{MATTE}} &= E[Y_i^1 | P(X)] - E[Y_i^0 | P(X)] \\ &= E[Y_i^1 | P(X)] - E[Y_j^0 | P(X)] \end{aligned} \quad (17)$$

mit

$\Delta^{\text{MATTE}}$ : durchschnittlicher Treatment-on-Treated Effekt nach Matching.



Der durchschnittliche Treatment-on-Treated Effekt entspricht somit der Differenz der Mittelwerte der Ergebnisvariablen der Teilnehmer und der gematchten Nicht-Teilnehmer der Maßnahme nach Anwendung der Matching Methode.

Der durchschnittliche Wert der Ergebnisvariablen für die Nicht-Teilnehmer an der Maßnahme ergibt sich dabei in der folgenden Weise:

$$E[Y_j^0 | P(X)] = \frac{1}{|I|} \sum_{i \in I} \sum_{j \in J} w_{i,j} \cdot Y_j^0. \quad (18)$$

So ergibt sich der durchschnittliche Wert der Ergebnisvariablen für das Counterfactual Outcome der Teilnehmer als gewichteter Mittelwert der Werte der Ergebnisvariablen für die jeweiligen Matching-Partner. Hierbei wird die Gewichtung der Nicht-Teilnehmer durch den angewendeten Algorithmus determiniert.

#### 4.2 Difference-In-Differences Schätzer

Der Difference-In-Differences (DID) Schätzer setzt nicht die Bestimmung von Matching-Partnern voraus, sondern vergleicht die durchschnittliche Veränderung des Wertes der Ergebnisvariablen bei den Teilnehmern vor und nach der Teilnahme an der Maßnahme mit der durchschnittlichen Veränderung der Ergebnisvariablen bei allen Nicht-Teilnehmern (Hujer et al. 2003, S. 22).

$$\Delta^{\text{DID}} = E[Y_{i,t}^1 - Y_{i,t'}^0] - E[Y_{j,t}^0 - Y_{j,t'}^0] \quad t' < t_T < t \quad (19)$$

mit

$\Delta^{\text{DID}}$ : durchschnittlicher Treatment-on-Treated Effekt auf der Basis des Difference-In-Differences Schätzers,

$Y_{i,t'}^1$ : Wert der Ergebnisvariablen für den  $i$ -ten Teilnehmer an der Maßnahme zum  $t'$ -ten Zeitpunkt,

$Y_{j,t(t')}^0$ : Wert der Ergebnisvariablen für den  $j$ -ten Nicht-Teilnehmer an der Maßnahme zum  $t(t')$ -ten Zeitpunkt,

$t_T$ : Zeitpunkt der Teilnahme an der Maßnahme.

Aufgrund der beschriebenen Vorgehensweise berücksichtigt der Difference-In-Differences Schätzer einen möglichen Trendeffekt. Hierbei wird angenommen, dass dieser für die Teilnehmer und Nicht-Teilnehmer an der Maßnahme in gleichem Maße vorliegt. Die Vorgehensweise des Difference-In-Differences Schätzers impliziert jedoch auch, dass die durchschnittliche Verzerrung durch den Selbstselektionseffekt in den betrachteten Perioden das gleiche Ausmaß besitzt, so dass diese durch die doppelte Differenzbildung eliminiert werden kann (Heckman et al. 1998, S. 1029). Es liegt dem Difference-In-Differences Schätzer folglich die Annahme zugrunde, dass die Höhe des Selbstselektionseffekts über die Zeit hinweg konstant ist und linear additiv eingeht.

### 4.3 Conditional Difference-In-Differences Schätzer

Der Conditional Difference-In-Differences (CDID) Schätzer ist eine Kombination des Matching-Schätzers und des Difference-In-Differences Schätzers. So wird der durchschnittliche Treatment-on-Treated Effekt ermittelt, indem die doppelten Differenzen nun in Bezug auf die Matching-Partner gebildet werden. Es wird dabei entsprechend zu Gleichung (16) die folgende Annahme zugrunde gelegt:

$$E[Y_{i,t}^0 - Y_{i,t'}^0 | P(X)] = E[Y_{j,t}^0 - Y_{j,t'}^0 | P(X)] \quad t' < t_T < t. \quad (20)$$

Der durchschnittliche Treatment-on-Treated Effekt wird dann in der folgenden Weise ermittelt (Heckman et al. 1998):

$$\begin{aligned} \Delta^{\text{CDID}} &= E[Y_{i,t}^1 - Y_{i,t'}^0 | P(X)] - E[Y_{i,t}^0 - Y_{i,t'}^0 | P(X)] \\ &= E[Y_{i,t}^1 - Y_{i,t'}^0 | P(X)] - E[Y_{j,t}^0 - Y_{j,t'}^0 | P(X)] \quad t' < t_T < t \end{aligned} \quad (21)$$

mit

$\Delta^{\text{CDID}}$ : durchschnittlicher Treatment-on-Treated Effekt nach Matching auf Basis des Conditional Difference-In-Differences Schätzers.

So berücksichtigt der Conditional Difference-In-Differences Schätzer unbeobachtete, lineare und zeitinvariante Effekte. Es ist nun möglich, zwischen einem zeitlichen Trend und unbeobachteter Heterogenität zu differenzieren.

Der durchschnittliche Wert für den Trendeffekt bei den Nicht-Teilnehmern an der Maßnahme ergibt sich hierbei in der folgenden Weise:

$$E[Y_{j,t}^0 - Y_{j,t'}^0 | P(X)] = \frac{1}{|I|} \sum_{i \in I} \sum_{j \in J} w_{i,j} \cdot (Y_{j,t}^0 - Y_{j,t'}^0) \quad t' < t_T < t. \quad (22)$$

So entspricht das Counterfactual Outcome der Teilnehmer dem gewichteten Mittelwert der Differenzen der Werte der Ergebnisvariablen für die jeweiligen Matching-Partner.

### 4.4 Illustratives Beispiel

Aufbauend auf dem in Tabelle 2 dargestellten Zahlenbeispiel werden nun die unterschiedlichen Schätzer zur Bestimmung des durchschnittlichen Treatment-on-Treated Effekts anhand eines illustrativen Beispiels verdeutlicht. Das Zahlenbeispiel weist Beobachtungen von 9 Kunden auf, von denen 4 Kunden eine Kundenkarte besitzen. Als Störvariable wird das Haushaltsnettoeinkommen als kategoriale Variable berücksichtigt. Das Unternehmen möchte wissen, ob der Besitz der Kundenkarte dazu führt, dass Kunden mehr Umsatz bei dem Unternehmen tätigen. Es wird ein direktes Matching durchgeführt und es werden alle Nicht-Teilnehmer als Matching-Partner berücksichtigt. Zudem wird der Nearest Neighbor Algorithmus mit Zurücklegen angewendet.

Dem Unternehmen liegen die in Tabelle 2 dargestellten Informationen vor und nach der Einführung der Kundenkarte vor. Der Vergleich der Mittelwerte zwischen

**Tabelle 2** Zahlenbeispiel zur Schätzung des Effekts des Besitzes der Kundenkarte auf den getätigten monatlichen Umsatz

Kunde	Haushalts- nettoeinkommen	Besitz der Kundenkarte	Umsatz des Kunden je Monat vor Einführung der Kundenkarte	Umsatz des Kunden je Monat nach Einführung der Kundenkarte
1	2.000 € und mehr	Ja	140 €	177 €
2	2.000 € und mehr	Ja	130 €	167 €
3	2.000 € und mehr	Nein	140 €	152 €
4	1.500 €–1.999 €	Ja	30 €	65 €
5	1.000 €–1.499 €	Ja	40 €	75 €
6	1.500 €–1.999 €	Nein	30 €	40 €
7	1.000 €–1.499 €	Nein	40 €	50 €
8	500 €–999 €	Nein	30 €	40 €
9	500 €–999 €	Nein	40 €	50 €

den beiden Gruppen der Besitzer einer Kundenkarte und der Nicht-Besitzer nach Einführung der Kundenkarte ergibt einen Unterschied bezüglich des monatlichen Umsatzes von 54,60 € (siehe Tabelle 3).

**Tabelle 3** Ergebnis des Mittelwertvergleichs

	Mittelwert
Durchschnittlicher Umsatz der Gruppe der Besitzer einer Kundenkarte (1, 2, 4, 5)	121,00 €
Durchschnittlicher Umsatz der Gruppe der Nicht-Besitzer einer Kundenkarte (3, 6, 7, 8, 9)	66,40 €
Differenz der Mittelwerte der beiden Gruppen	54,60 €

Dieser Mittelwertunterschied kann durch den Besitz der Kundenkarte, aber auch durch Selbstselektion der Kunden bedingt sein. Aus diesem Grund wird im Folgenden die Matching Methode angewendet. Bei dieser erfolgt ein direktes Matching auf Basis der Variable Haushaltsnettoeinkommen, da vermutet wird, dass das Haushaltsnettoeinkommen (Störvariable) aufgrund der Konditionen der Kundenkarte deren Besitz beeinflusst. So werden dann die Kunden 1 und 2 mit 3 und die Kunden 4 und 5 mit 6 bzw. 7 verglichen. Die Berechnung des Matching-Schätzers führt zu dem in Tabelle 4 dargestellten Ergebnis.

**Tabelle 4** Ergebnis des Matching-Schätzers

	Umsatz je Monat bei Besitz der Kundenkarte	Umsatz je Monat des Matching-Partners	Differenz
Kunde 1	177,00 €	152,00 €	25,00 €
Kunde 2	167,00 €	152,00 €	15,00 €
Kunde 4	65,00 €	40,00 €	25,00 €
Kunde 5	75,00 €	50,00 €	25,00 €
Mittelwert	121,00 €	98,50 €	22,50 €

Tabelle 4 zeigt, dass der Besitz der Kundenkarte bei den Kunden eine durchschnittliche Steigerung des monatlichen Umsatzes und damit einen durchschnittlichen Treatment-on-Treated Effekt in Höhe von 22,50 € hervorruft.

Bei Anwendung des Difference-In-Differences Schätzers wird nun die Veränderung des monatlichen Umsatzes der Kunden, die eine Kundenkarte besitzen, mit der Veränderung des monatlichen Umsatzes der Kunden ohne Kundenkarte betrachtet. Es zeigt sich bei den 4 Kunden mit Kundenkarte eine durchschnittliche Differenz des monatlichen Umsatzes vor und nach Einführung der Kundenkarte von 36,00 €. Bei den Kunden ohne Kundenkarte ergibt sich hingegen eine durchschnittliche Differenz von 10,40 €, so dass ein Trendeffekt angenommen werden kann. Der durchschnittliche Effekt des Besitzes der Kundenkarte auf den monatlichen Umsatz beträgt somit 25,60 €.

**Tabelle 5** Ergebnis des Difference-In-Differences Schätzers

Durchschnittliche Differenz des Umsatzes je Monat der Kunden <i>mit</i> Kundenkarte	Durchschnittliche Differenz des Umsatzes je Monat der Kunden <i>ohne</i> Kundenkarte	Durchschnittlicher Effekt des Besitzes der Kundenkarte auf den Umsatz je Monat
36,00 €	10,40 €	25,60 €

Bei Anwendung des Conditional Difference-In-Differences Schätzers erfolgt nun eine Kombination des Matching-Schätzers und des Difference-In-Differences Schätzers (Tabelle 6).

**Tabelle 6** Ergebnis des Conditional Difference-In-Differences Schätzers

	Differenz des Umsatzes je Monat der Kunden <i>mit</i> Kundenkarte	Differenz des Umsatzes je Monat der Matching-Partner	Differenz
Kunde 1	37,00 €	12,00 €	25,00 €
Kunde 2	37,00 €	12,00 €	25,00 €
Kunde 4	35,00 €	10,00 €	25,00 €
Kunde 5	35,00 €	10,00 €	25,00 €
Mittelwert	36,00 €	11,00 €	25,00 €

Tabelle 6 zeigt, dass nun ein Effekt der Teilnahme an der Maßnahme von 25,00 € ermittelt wird. Zudem liegt ein Trendeffekt von 12,00 € für die Haushalte mit einem hohen Haushaltsnettoeinkommen und von 10,00 € für Haushalte mit einem geringen Haushaltsnettoeinkommen vor. So ist es bei Anwendung des Conditional Difference-In-Differences Schätzers im Gegensatz zum Difference-In-Differences Schätzer möglich, Heterogenität in den Trendeffekten aufzudecken.

Daher ergibt sich bei Betrachtung des ursprünglichen Mittelwertvergleichs, dass 54,21 % des beobachteten Effekts auf einen Selbstselektionseffekt und 45,79 % auf den durchschnittlichen Treatment-on-Treated Effekt zurückzuführen sind (siehe Tabelle 7).

**Tabelle 7** Aufspaltung des gesamten Effekts in den durchschnittlichen Treatment-on-Treated Effekt und Selbstselektionseffekt

Differenz der Mittelwerte vor Matching	Treatment-on-Treated Effekt bei Anwendung des CDID Schätzers	Selbstselektionseffekt bei Anwendung des CDID Schätzers
54,60 € (100 %)	25,00 € (45,79 %)	29,60 € (54,21 %)

#### 4.5 Vergleich der Schätzer zur Ermittlung des durchschnittlichen Treatment-on-Treated Effekts

Abschließend erfolgt nun ein Vergleich der unterschiedlichen Schätzer, wobei hierfür die folgenden Kriterien herangezogen werden:

- **Anforderungen an die Daten** Die vorgestellten Schätzer stellen unterschiedliche Anforderungen an die benötigten Daten. In der Praxis liegen häufig nur bestimmte Daten vor. Aus diesem Grund wird der Trade-Off zwischen der Menge an benötigten Daten und dem Erkenntnisgewinn bezüglich des durchschnittlichen Treatment-on-Treated Effekts diskutiert.
- **Berücksichtigte Effekte** Aufgrund des fundamentalen Evaluierungsproblems kann der wahre durchschnittliche Treatment-on-Treated Effekt nur näherungsweise bestimmt werden. Hierfür ist es erforderlich, dass der geschätzte durchschnittliche Treatment-on-Treated Effekt unverzerrt ermittelt wird. Aus diesem Grund werden die betrachteten Schätzer hinsichtlich ihrer Annahmen und deren Implikationen für eine Evaluierung des durchschnittlichen Treatment-on-Treated Effekts diskutiert. Hierfür werden die Effekte herangezogen, die durch die Schätzer berücksichtigt werden.

Um den **Matching-Schätzer** ermitteln zu können, sind lediglich Querschnittsdaten für einen Zeitpunkt erforderlich. Dementsprechend stellt der Matching-Schätzer geringe Anforderungen an die erforderlichen Daten. Diese geringen Datenanforderungen führen jedoch dazu, dass nur eine statische Betrachtung des durchschnittlichen Treatment-on-Treated Effekts erfolgt. Ist zu vermuten, dass ein Trend existiert, so kann dieser auf Basis des Matching-Schätzers nicht ermittelt werden. Auch erfasst der Matching-Schätzer lediglich die beobachtete Heterogenität, die durch die Störvariablen berücksichtigt wird. Für eine adäquate Ermittlung des durchschnittlichen Treatment-on-Treated Effekts ist es allerdings erforderlich, dass alle relevanten Störvariablen für die Bestimmung der Partner von Teilnehmern und Nicht-Teilnehmern der Maßnahme berücksichtigt werden. Ist dies nicht gewährleistet, so kann der Matching-Schätzer aufgrund unbeobachteter Heterogenität verzerrt sein, so dass der wahre durchschnittliche Treatment-on-Treated Effekt überschätzt oder unterschätzt wird. So bleibt festzuhalten, dass der Matching-Schätzer überhaupt nur dann den wahren durchschnittlichen Treatment-on-Treated Effekt abbilden kann, wenn Trends nicht von Bedeutung sind und keine unbeobachtete Heterogenität vorliegt.

Zur Berechnung des **Difference-In-Differences Schätzers** ist es erforderlich, dass gepoolte Daten, die zu zwei Zeitpunkten erhoben werden, vorliegen. Dabei

ist relevant, dass die Gruppe der Teilnehmer an der Maßnahme zum ersten Beobachtungszeitpunkt noch nicht an der Maßnahme teilgenommen hat, so dass  $Y_{i,t'}^0$  beobachtet werden kann. Aufgrund dieser Voraussetzung stellt die Ermittlung des Difference-In-Differences Schätzers höhere Anforderungen an die Daten als der Matching-Schätzer. Bezüglich der adäquaten Bestimmung des durchschnittlichen Treatment-on-Treated Effekts ist festzuhalten, dass der Difference-In-Differences Schätzer in der Lage ist, mögliche Trendeffekte zu berücksichtigen. Zu einer verzerrten Evaluierung des durchschnittlichen Treatment-on-Treated Effekts kann es jedoch kommen, wenn Heterogenität in den Trendeffekten existiert und wenn die zukünftigen Teilnehmer einer Maßnahme ihre Teilnahme antizipieren und daraufhin ihr Verhalten bereits vor der Teilnahme an der Maßnahme ändern (Fitzenberger und Speckesser 2001, S. 11).

Um den **Conditional Difference-In-Differences Schätzer** ermitteln zu können, sind ebenfalls gepoolte Daten, die zu zwei Zeitpunkten erhoben werden, erforderlich. Dabei ist ebenfalls relevant, dass die Gruppe der Teilnehmer an der Maßnahme zum ersten Beobachtungszeitpunkt noch nicht an der Maßnahme teilgenommen hat, so dass  $Y_{i,t'}^0$  beobachtet werden kann. Aufgrund dieser Voraussetzung stellt die Schätzung des Conditional Difference-In-Differences Schätzers höhere Anforderungen an die Daten als der Matching-Schätzer. Bei Anwendung des Conditional Difference-In-Differences Schätzers wird wie auch bei dem Difference-In-Differences Schätzer ein möglicher Trendeffekt berücksichtigt. Jedoch ist es bei Anwendung des Conditional Difference-In-Differences Schätzers zudem möglich, unterschiedliche Trendeffekte für die Teilnehmer und Nicht-Teilnehmer an der Maßnahme zu evaluieren. So kann der unbeobachteten Heterogenität in Abhängigkeit von den beobachteten Störvariablen Rechnung getragen werden.

In der Tabelle 8 wird der Vergleich der unterschiedlichen Schätzer zusammengefasst.

**Tabelle 8** Vergleich der unterschiedlichen Schätzer

	Anforderung an die Daten	Berücksichtigte Effekte
Matching-Schätzer	<ul style="list-style-type: none"> <li>• Erhebung der Ergebnisvariablen und jener Variablen, die für das Matching herangezogen werden sollen</li> <li>• Querschnittsdaten für einen Zeitpunkt</li> </ul>	<ul style="list-style-type: none"> <li>• Selbstselektionseffekt basiert auf beobachteter Heterogenität</li> <li>• Mögliche Trendeffekte werden nicht berücksichtigt</li> </ul>
Difference-In-Differences Schätzer	<ul style="list-style-type: none"> <li>• Erhebung der Ergebnisvariablen</li> <li>• Gepoolte Daten, die mindestens zu zwei Zeitpunkten erhoben wurden</li> </ul>	<ul style="list-style-type: none"> <li>• Selbstselektionseffekt basiert auf Bildung der Differenzen</li> <li>• Mögliche Trendeffekte können berücksichtigt werden</li> </ul>
Conditional Difference-In-Differences Schätzer	<ul style="list-style-type: none"> <li>• Erhebung der Ergebnisvariablen und jener Variablen, die für das Matching herangezogen werden sollen</li> <li>• Gepoolte Daten, die mindestens zu zwei Zeitpunkten erhoben wurden</li> </ul>	<ul style="list-style-type: none"> <li>• Selbstselektionseffekt basiert auf beobachteter Heterogenität</li> <li>• Mögliche Trendeffekte können berücksichtigt werden</li> <li>• Unbeobachtete Heterogenität in den Trendeffekten wird berücksichtigt</li> </ul>

## 5 Fazit

Dieser Beitrag zeigte auf, wie mit Hilfe der Matching Methode Selbstselektionseffekte berücksichtigt werden können. Die angemessene Berücksichtigung von Selbstselektionseffekten ist immer dann von Bedeutung, wenn nicht nur ein Unterschied zwischen zwei Gruppen festgestellt, sondern auch Rückschlüsse auf eine kausale Beziehung gezogen werden sollen. So gibt es in der Betriebswirtschaft zahlreiche Fragestellungen, bei denen der Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable von Interesse ist. Es sei aber darauf hingewiesen, dass für eine Prognose die Matching Methode nicht angewendet werden muss (Cox und Wermuth 2004).

Gezeigt wurde, wie der Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable angemessen ermittelt werden kann. Dabei wurden unterschiedliche Varianten der Matching Methode dargestellt. Da Barabas (2004) gezeigt hat, dass unterschiedliche Spezifikationen stabile Ergebnisse bei der Evaluierung des durchschnittlichen Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable liefern, ist die Matching Methode in einfacher Weise anwendbar. Voraussetzung für die angemessene Evaluierung des Effekts der Teilnahme an einer Maßnahme auf eine Ergebnisvariable ist jedoch, dass alle relevanten Störvariablen erfasst werden.

Es stellt sich aber die Frage, ob nicht eine einfache Regressionsanalyse ebenfalls in der Lage ist, den Effekt der Teilnahme an einer Maßnahme auf die Ergebnisvariable zu evaluieren. Das Berücksichtigen der Teilnahmeentscheidung, beispielsweise in Form einer Dummy-Variablen sowie der Störvariablen als unabhängige Variablen führt allerdings dazu, dass der Effekt der Teilnahme verzerrt geschätzt wird. Dies liegt darin begründet, dass die Störvariablen sowohl einen Einfluss auf die Teilnahmeentscheidung als auch auf die Ergebnisvariable haben. Es liegt dann eine Verzerrung aufgrund des Vorliegens endogener Variablen vor. In dem Beispiel ergibt sich bei einer Berücksichtigung des Besitzes der Kundenkarte sowie des Einkommens als unabhängige Variable und dem monatlichen Umsatz des Kunden als abhängige Variable ein Effekt des Besitzes der Kundenkarte von 23,00 €. Auch dieser geschätzte Effekt entspricht nicht dem wahren Effekt des Besitzes der Kundenkarte, da der Endogenität nicht angemessen Rechnung getragen wurde. So beeinflusst das Einkommen sowohl die Teilnahmeentscheidung als auch die Höhe des monatlichen Umsatzes eines Kunden.

Eine Annahme der Regressionsanalyse ist, dass keine Korrelation zwischen den unabhängigen Variablen und dem Fehlerterm existiert. Wenn jedoch eine Korrelation besteht, so sind die geschätzten Koeffizienten verzerrt und es liegen dann endogene Variablen vor. In der Literatur wird daher der Instrumental Variable Ansatz diskutiert, um dem Problem endogener Variablen Rechnung zu tragen (Wooldridge 2003). So wird im Instrumental Variable Ansatz ebenfalls der Effekt von unabhängigen Variablen auf die Ergebnisvariable anhand eines Regressionsmodells formuliert. Jedoch gilt es, für die endogenen Variablen so genannte Instrumente zu finden, die mit den entsprechenden endogenen Variablen korrelieren, nicht aber mit dem Fehlerterm der Regressionsgleichung, die den Wert der Ergebnisvariablen erklärt. Eine zweistufige Schätzung ermöglicht dann eine adäquate Evaluierung des Effekts der Teilnahme an einer Maßnahme, da zunächst in einem ersten Schritt der Einfluss des Instruments auf die endogene Variable – hier die Teilnahme an einer Maßnahme – spezifiziert wird. In einem zweiten Schritt wird dann der Effekt der Teilnahme an einer Maßnahme auf die Ergebnisvariable unverzerrt geschätzt. Der Instrumental

Variable Ansatz wurde bereits bei einigen betriebswirtschaftlichen Fragestellungen angewendet, um Selbstselektionseffekte zu berücksichtigen (Zettelmeyer et al. 2001; Erdem et al. 1999). In betriebswirtschaftlichen Anwendungen ist es jedoch häufig schwierig, Instrumente zu finden, die die oben beschriebene Eigenschaft besitzen. Die Matching Methode vermeidet die Identifizierung solcher Instrumente. Jedoch liegt ihr die zentrale Annahme zugrunde, dass alle relevanten Störvariablen als Kontrollvariablen erfasst wurden (Heckman und Navarro-Lozano 2004).

Weitere Ansätze, die sich mit kausalen Effekten beschäftigen sind beispielsweise Graphenmodelle, die eine Kombination aus Wahrscheinlichkeitstheorie und Graphentheorie repräsentieren (Lauritzen 1999; Edwards 2000). Hierzu zählen beispielsweise Bayesianische Netzwerke. Das Ziel dieser Modelle besteht jedoch in erster Linie nicht darin, den Effekt der Teilnahme an einer Maßnahme auf eine Ergebnisvariable zu evaluieren, sondern primär kausale Zusammenhänge abzubilden.

Abschließend ist anzumerken, dass die Berücksichtigung von Selbstselektionseffekten in der Zukunft eine bedeutendere Stellung in der betriebswirtschaftlichen Forschung einnehmen wird. Diese Einschätzung wird primär von der Entwicklung getragen, dass aufgrund der vermehrten Diskussion von Selbstselektionseffekten in der volkswirtschaftlichen Literatur auch in der betriebswirtschaftlichen Literatur das Problembewusstsein geschärft wird. Dies wird beispielsweise auch an der breiten Diskussion über Endogenität in betriebswirtschaftlichen Modellen deutlich. Somit ist davon auszugehen, dass die Diskussion über adäquate Methoden zur Berücksichtigung von Selbstselektion in betriebswirtschaftlichen Fragestellungen weiter an Bedeutung gewinnen wird.

## Literatur

- Ashenfelter O (1978) Estimating the effect of training programs on earnings. *Rev Econ Stat* 60:47–57
- Ashenfelter O, Rouse C (1998) Income, schooling, and ability: Evidence from a new sample of identical twins. *Q J Econ*: 253–284
- Barabas J (2004) How deliberation affects policy opinions. *Am Polit Sci Rev* 98:1–16
- Bjorklund A, Moffitt R (1987) The estimation of wage gains and welfare gains in self-selection models. *Rev Econ Stat* 69:42–49
- Black DA, Smith J (2004) How robust is the evidence on the effects of college quality? Evidence from matching. *J Econometrics* 121:99–124
- Boston Consulting Group (2001) The multichannel consumer – The need to integrate online and offline channels in Europe. Boston
- Bortz J (1999) Statistik für Sozialwissenschaftler. Springer, Berlin Heidelberg New York
- Bundesanstalt für Arbeit (2003) 138 Milliarden Euro für aktive Arbeitsmarktpolitik im Osten. Pressemitteilung vom 8. April 2003.  
[http://www.soliserv.de/presse-arbeitsamt\\_2quartal-03.htm](http://www.soliserv.de/presse-arbeitsamt_2quartal-03.htm), abgerufen am 1. November 2004
- Christensen B, Clement M, Albers S, Guldner S (2004) Zur Relevanz der Kontrollgruppenauswahl in der empirischen Forschung: Eine Analyse am Beispiel der Erfolgswirkung der Academy Awards im Filmgeschäft. Working paper, Christian-Albrechts-Universität zu Kiel
- Cox D, Wermuth N (2004) Causality: A statistical view. *Int Stat Rev* 72:285–305
- D'Agostino R (1998) Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 17:2265–2281
- Degeratu A, Rangaswamy A, Wu J (2000) Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price and other search attributes. *Int J Res Marketing* 17:55–78
- Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat* 84:151–161
- Edwards D (2000) Introduction to graphical modelling. Springer, Berlin Heidelberg New York



- Erdem T, Keane M, Sun B (1999) Missing price and coupon availability data in scanner panels: Correcting for the self-selection bias in the choice model parameters. *J Econometrics* 89:177–196
- Fitzenberger B, Speckesser S (2003) Weiterbildungsmaßnahmen in Ostdeutschland: Ein Misserfolg der Arbeitsmarktpolitik? In: Schmähl W (Hrsg) Wechselwirkungen zwischen Arbeitsmarkt und sozialer Sicherung (Soziale Sicherung und Arbeitsmarkt, Bd. 294). Duncker und Humblot, Berlin, S. 51–81
- Granger C (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37:424–438
- Heckman J (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Ann Econ Soc Meas* 5:475–492
- Heckman J (1978) Dummy endogenous variables in a simultaneous equations system. *Econometrica* 46:931–960
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica* 47:152–162
- Heckman J, Ichimura H, Smith J, Todd P (1996) Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. *Proc Natl Acad Sci* 93:13416–13420
- Heckman J, Ichimura H, Smith J, Todd P (1998) Characterizing selection bias using experimental data. *Econometrica* 66:1017–1098
- Heckman J, Ichimura H, Todd P (1998) Matching as an econometric evaluation estimator. *Rev Econ Stud* 65:261–294
- Heckman J, LaLonde R, Smith J (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (Hrsg) *Handbook of Labor Economics*, vol. III. Amsterdam, S. 1865–2097
- Heckman J, Navarro-Lozano S (2004) Using matching, instrumental variables, and control functions to estimate economic choice models. *Rev Econ Stat* 86:30–57
- Hitt LM, Frei FX (2002) Do better customers utilize electronic distribution channels? The case of PC banking. *Manag Sci* 48:732–748
- Hujer R, Caliendo M, Radic D (2001) Nobody knows...how do different evaluation estimators perform in a simulated labour market experiment? Working paper, Universität Frankfurt
- Hujer R, Caliendo M, Radic D (2003) Methods and limitations of evaluation and impact research. Report on VET Research in Europe, CEDEFOP Project 0730. Working paper, Universität Frankfurt
- Hujer R, Maurer K-O, Wellner M (1997) The impact of training on unemployment duration in West Germany – Combining a discrete hazard rate model with matching techniques. Working paper, Universität Frankfurt
- Imbens G (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev Econ Stat* 86:4–29
- LaLonde R (1986) Evaluating the econometric evaluations of training programs with experimental data. *Am Econ Rev* 76:604–620
- Lauritzen S (1999) Causal inference from graphical models. Research report R-99-2021, Aalborg University. <http://www.math.aau.dk/research/reports/R-99-2021.ps>, abgerufen am 7. Februar 2005
- Lechner M (1998) Mikroökonomische Evaluationsstudien: Anmerkungen zu Theorie und Praxis. In: Pfeiffer F, Pohlmeier W (Hrsg) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden, S. 13–38
- Lechner M (2002) Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Rev Econ Stat* 84:205–220
- Lee L-F (2000) Self-selection. In: Baltagi B (Hrsg) *A companion to theoretical econometrics*. Oxford, S. 384–409
- Rässler S (2002) *Statistical matching*. Springer, Berlin Heidelberg New York
- Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
- Rosenbaum P, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 39:33–38
- Roy A (1951) Some thoughts on the distribution of earnings. *Oxford Econ Pap* 3:135–145
- Rubin D (1974) Estimating causal effects to treatments in randomised and nonrandomised studies. *J Educational Psychol* 66:688–701

- 
- Rubin D (1990) Comment on “Neyman (1923) and causal inference in experiments and observational studies”. *Stat Sci* 5:472–480
- Singer B (1986) Self-selection and performance-based ratings: A case study in program evaluation. In: Wainer H (Hrsg) *Drawing inference from self-selected samples*. New York, S. 29–62
- Smith J, Todd P (2000) Does matching overcome lalondes critique of nonexperimental estimators? Working paper, Penn Institute for Economic Research, Philadelphia
- Wehring R (2002) Retail E-banking: Tinkering pays off. *ABA Banking J*: 11–23
- Wooldridge J (2003) *Introductory econometrics – A modern approach*. Mason
- Zettelmeyer F, Morton FS, Silvia-Risso J (2001) Cowboys or cowards: Why are internet car prices lower? NBER working paper 8667. National Bureau of Economic Research, Berkeley