Centrum für Evaluation

Center for Evaluation

Arbeitspapier 19

Centrum für Evaluation – Universität des Saarlandes

Quasiexperimentelle Wirkungsevaluation mit Propensity Score Matching: Ein Leitfaden für die Umsetzung mit Stata

Christoph E. Müller

Müller, Christoph E.

Quasiexperimentelle Wirkungsevaluation mit Propensity Score Matching: Ein Leitfaden für die Umsetzung mit Stata

Saarbrücken, Centrum für Evaluation, 2012.

(CEval-Arbeitspapiere; 19)

Bezug: Centrum für Evaluation (CEval)
Universität des Saarlandes
Postfach 15 11 50
D-66041 Saarbrücken
info@ceval.de

Inhalt

1. Hintergrund

Die Durchführung sozialer Interventionsmaßnahmen ist meist mit der Erwartung verbunden, Wirkungen zu induzieren. Je nach Themenbereich, inhaltlicher Ausrichtung und praktischer Umsetzung bezieht sich der Wirkungsbegriff auf unterschiedlichste Konstrukte wie bspw. die Veränderung individuellen Verhaltens von Personen oder Organisationen, die Beeinflussung kognitiver Eigenschaften wie Einstellungen, Werthaltungen, Motivationen oder der gezielten Verbesserung objektiverer Tatbestände wie bspw. des Gesundheitszustands, der Wirtschaftskraft oder der Beschäftigungssituation.

Da mit der Umsetzung sozialer Interventionen angestrebt wird, Wirkungen hervorzurufen, können die Interventionen selbst als Treatment betrachtet werden, dessen Konsequenzen durch die Evaluation ermittelt werden müssen. Der Begriff ‚Treatment' stammt aus dem angelsächsischen Sprachgebrauch und bezeichnet eigentlich eine unabhängige Erklärungsvariable, die durch den Forscher/Evaluator[1] variiert werden kann. Im Fall der Wirkungsevaluation werden Interventionsmaßnahmen, ob sie nun durch den Forscher variiert werden können oder nicht, als Treatment verstanden, mit dessen Umsetzung bestimmte Konsequenzen verbunden sind. Diese Konsequenzen werden als Treatmenteffekte bezeichnet und können synonym zum Wirkungsbegriff aufgefasst werden.

Die Beantwortung der Frage, ob eine Intervention Wirkungen nach sich zieht oder nicht und von welcher Stärke mögliche Wirkungen sind, stellt eine der wichtigsten Aufgaben der Evaluation dar. Sie ist zugleich aber auch „eine der größten Herausforderungen einer Evaluation" (Stockmann, 2006, S. 104), da die Wirkungsevaluation vor allem dann von Nutzen ist, wenn Wirkungen kausal interpretierbar sind. Die Möglichkeit einer fehlerfreien Attribuierung von Wirkungen auf eine Intervention erfordert von Evaluatoren die Anwendung angemessener Untersuchungsdesigns und elaborierter Erhebungs- und Analysemethoden der empirischen Sozialforschung. Die Wahl der Vorgehensweise im Rahmen der Wirkungsevaluation muss dabei allerdings den Eigenheiten des Evaluationsgegenstands und den situativen Rahmenbedingungen der Evaluation Rechnung tragen (White, 2010).

Zur Problematik der kausalen Wirkungsmessung schreibt Stockmann (2006, S. 104) weiter: „Da das Ziel von Wirkungsevaluationen darin besteht, mit größtmöglicher Zuverlässigkeit festzustellen, ob eine Intervention die intendierten Wirkungen auslöst, sind die Einflüsse anderer Faktoren, die ebenfalls für die gemessenen Veränderungen verantwortlich sein könnten, auszuschließen." Aus forschungstheoretischer Perspektive bedeutet dies, dass alle Faktoren, die neben dem Treatment einen Einfluss auf die interessierenden Wirkungsvariablen haben, entweder kontrolliert werden müssen oder deren Einfluss eliminiert werden muss. Als Wirkung im kausalen Sinne kann daher nur eine Veränderung in einer Outcome-Variablen gelten, die ausschließlich auf das Treatment zurückzuführen ist.

Der vorliegende Leitfaden beschreibt im Folgenden in einer anwendungsorientierten Art und Weise die Grundlagen und die Anwendung eines ökonometrischen Verfahrens, das im Rahmen quasiexperimenteller Wirkungsevaluation zur Ausbalancierung beobachteter Drittvariab-

[1] Die in diesem Leitfaden verwendeten männlichen Bezeichnungen dienen ausschließlich der besseren Lesbarkeit und gelten grundsätzlich für beide Geschlechter.

len (sog. Kovariaten) zwischen Versuchs- und Vergleichsgruppen und damit zur Reduktion von Konfundierungsprozessen genutzt werden kann. Bei diesem Verfahren handelt es sich um das sog. Propensity Score Matching (PSM), welches maßgeblich auf Rosenbaum & Rubin (1983, 1985) zurückgeht und die Schätzung individueller und durchschnittlicher Treatmenteffekte erlaubt. Bevor jedoch näher auf das Verfahren und seine Anwendung mit dem Statistikprogramm Stata eingegangen wird, werden im nächsten Abschnitt seine methodischen Grundlagen beschrieben.

Es sei an dieser Stelle noch angemerkt, dass es sich beim PSM um ein relativ komplexes ökonometrisches Verfahren handelt, dessen Anwendung bereits einige Vorkenntnisse in den (quantitativen) Methoden der empirischen Sozialforschung und der Statistik erfordert. Dieser Leitfaden richtet sich daher vor allem an Evaluatoren, die bereits über Vorkenntnisse und Erfahrungen im Bereich der quantitativen Wirkungsanalyse verfügen.

2. Das fundamentale Evaluationsproblem und Lösungsansätze

Die Beantwortung der zentralen Frage einer jeden Wirkungsevaluation erfordert die Klärung, ob eine Intervention dazu in der Lage ist, Wirkungen zu entfalten. Wenn fortfolgend von Wirkungen die Rede ist, so sind damit Wirkungen im streng kausalen Sinne gemeint, also eindeutige Beziehungen zwischen Ursachen und durch sie hervorgerufene Konsequenzen (Rubin, 2004, 1974; Kim 1995). Im Rahmen der kausalen Wirkungsanalyse wird demnach untersucht, ob die Ursache ‚Interventionsdurchführung' zu bestimmten Wirkungen auf Seiten der Untersuchungsobjekte führt und wie stark diese Konsequenzen ausgeprägt sind.

Die Wirkungsanalyse folgt dabei dem Konzept des kontrafaktischen Zustands, welches im sog. Roy-Rubin-Modell formalisiert wurde (Rubin, 1974; Roy, 1951). Der kontrafaktische Zustand beschreibt den hypothetischen Zustand, in dem sich ein Untersuchungsobjekt, das einer Intervention ausgesetzt war, unter denselben Bedingungen befinden würde, wenn es diesem Einfluss nicht ausgesetzt gewesen wäre. Unter der hypothetischen Existenz des kontrafaktischen Zustands würde sich das Untersuchungsobjekt in den beiden Zuständen (faktisch vs. kontrafaktisch) lediglich hinsichtlich der Teilnahme bzw. Nichtteilnahme an einer Intervention unterscheiden und wäre ansonsten identisch.

Wenn im Falle eines dichotomen Treatmentfaktors (1 = Teilnahme; 0 = Nichtteilnahme) τ_i den individuellen Treatmenteffekt für jede Person i = 1, 2…, N bezeichnet und N die Population repräsentiert, dann kann die individuelle Differenz zwischen dem faktischen und kontrafaktischen Zustand folgendermaßen ausgedrückt werden[2]:

$$\tau_i = Y_i(1) - Y_i(0) \qquad (1)$$

[2] Die Notation ist an Caliendo & Kopeinig (2008) angelehnt.

Y(1) bezeichnet dabei den faktischen, Y(0) den kontrafaktischen Zustand in einer Outcome-Variable Y. Aus der in den Sozialwissenschaften (meist) gegebenen Unmöglichkeit der Messung des kontrafaktischen Zustands ergibt sich nun das sog. ‚fundamentale Evaluationsproblem' (vgl. Rubin, 2004, 1974; Heckman & Smith 1995). Dieses bezieht sich auf den Umstand, dass ein- und dasselbe Objekt zum selben Zeitpunkt nicht in zwei verschiedenen Zuständen beobachtet werden kann, womit die Schätzung individueller Treatmenteffekte verhindert wird. Stattdessen muss der kontrafaktische Zustand zur Abschätzung von Treatmenteffekten künstlich konstruiert werden, bspw. durch die Verwendung von Vergleichsgruppen. Übersetzt in die Sprache der Evaluation beschreiben Rossi et al. (2004, S. 234) die Vorgehensweise im Rahmen der kontrafaktischen Wirkungsevaluation wie folgt: „Evaluators assess the effects of social programs by comparing information about outcomes for program participants with estimates of what their outcomes would have been had they not participated."

Die Schätzung von Treatmenteffekten kann prinzipiell auf mehrere Arten erfolgen. In der Praxis beschränkt sich die Auswahl in der Regel aber auf den ‚Average Treatment Effect' (ATE) oder den ‚Average Treatment Effect on the Treated' (ATT). Unter dem ATE wird grundsätzlich der durchschnittliche kausale Treatmenteffekt in der Gesamtpopulation von Interesse verstanden (bspw. alle Langzeitarbeitslosen), der ATT beschreibt dagegen den durchschnittlichen kausalen Effekt eines Treatments auf diejenigen Personen, die tatsächlich an einer Intervention teilgenommen haben (bspw. alle Langzeitarbeitslosen, die an einer Weiterqualifizierungsmaßnahme teilgenommen haben). Gangl & DiPrete (2004, S. 8/9) beschreiben die beiden Parameter wie folgt:

> „Der ATE entspricht [...] konzeptionell dem erwarteten Effekt von T [Treatment] für Y [Outcome] einer zufällig aus der Gesamtpopulation [...] gezogenen Person, der ATT dagegen dem im Durchschnitt beobachteten Effekt von T für Y für eine zufällig aus der Experimentalstichprobe E gezogenen Person. Der ATE erfasst konzeptionell damit die typischen Folgen von T in der untersuchten Population, der ATT die typischen Folgen von T für die (möglicherweise selektive) Gruppe der tatsächlich von T Betroffenen."

Da der ATE auch Effekte auf Personen, für die eine Maßnahme möglicherweise nicht intendiert war, miteinbezieht (vgl. Heckman, 1997), kommt in der Praxis meist der ATT zur Anwendung. Aus diesem Grund beschränkt sich der vorliegende Leitfaden auf die Schätzung des ATT.

Bezeichnet D die dichotome Treatmentvariable mit den Ausprägungen 1 = Teilnehmer und 0 = Nichtteilnehmer und beschreibt $E[Y(1)|D = 1]$ den Erwartungswert im Outcome der Maßnahmenteilnehmer und $E[Y(0)|D = 1]$ den Erwartungswert des entsprechenden kontrafaktischen Zustands, dann lässt sich der wahre ‘Average Treatment Effect on the Treated' (τ_{ATT}) folgendermaßen berechnen:

$$\tau_{ATT} = E[Y(1)|D = 1] - E[Y(0)|D = 1] \qquad (2)$$

Da der kontrafaktische Zustand nicht messbar ist, muss er künstlich hergestellt werden. Eine erste Möglichkeit hierzu besteht in der Verwendung von Pretest-Posttest-Designs, wobei die Vorher-Messung als kontrafaktischer Zustand der Nachher-Messung aufgefasst wird (bspw. Reinowski, 2006). Die fehlerfreie Abbildung des kontrafaktischen Zustands in Form einer Vorhermessung ist allerdings nur dann möglich, wenn das entsprechende Untersuchungsobjekt zwischen den Messzeitpunkten in allen Merkmalen stabil bleibt, sodass sich das Objekt zwischen den beiden Messzeitpunkten ausschließlich darin unterscheidet, ob es einem Treatment ausgesetzt war oder nicht. Da sich das in diesem Leitfaden vorgestellte Verfahren des Propensity Score Matching vor allem auf die Analyse von Daten bezieht, die sowohl Informationen für eine Versuchs- als auch eine Vergleichsgruppe beinhalten, wird auf die Verwendung von Vorher-Messungen als Substitute des kontrafaktischen Zustands an dieser Stelle nicht näher eingegangen.

Eine zweite Möglichkeit zur künstlichen Abbildung des kontrafaktischen Zustands stellt also die Verwendung von Vergleichsgruppen dar, wobei die Schätzung von Treatmenteffekten dann anhand durchschnittlicher Gruppendifferenzen erfolgt. Der in Formel (2) dargestellte Ausdruck $E[Y(0)|D = 1]$ wird durch den Ausdruck $E[Y(0)|D = 0]$ ersetzt. Der einzige Unterschied zu Formel (2) besteht darin, dass D im Falle des kontrafaktischen Zustands gleich null wird. Dies bedeutet, dass der kontrafaktische Zustand durch die Verwendung von Untersuchungsobjekten, die nicht der entsprechenden Intervention ausgesetzt waren, geschätzt wird. Der ATT ist hier folglich als die Differenz des Erwartungswerts der Teilnehmer und des Erwartungswerts der Nichtteilnehmer definiert:

$$ATT = E[Y(1)|D = 1] - E[Y(0)|D = 0] \qquad (3)$$

Da sich Menschen aber prinzipiell in ihren Eigenschaften voneinander unterscheiden können, also $E[Y(0)|D = 0] \neq E[Y(0)|D = 1]$ sein kann, kann durch die Verwendung von Vergleichsgruppen der kontrafaktische Zustand nicht immer fehlerfrei abgebildet werden. Lediglich im Falle randomisierter Gruppen gilt $E[Y(0)|D = 0] = E[Y(0)|D = 1]$. Sofern bestehende Unterschiede in bestimmten Merkmalen zwischen den Gruppen auch einen Einfluss auf die Outcome-Variable aufweisen, liegt eine Konfundierung vor und die geschätzten Treatmenteffekte sind verzerrt. Der Schätzfehler, der sich aufgrund von Konfundierung durch Merkmalsunterschiede zwischen den Gruppen ergibt, wird auch als ‚Selection Bias' (SB) bezeichnet (Heckman et al., 1998a). Bei Verwendung von Vergleichsgruppen als künstliche Abbilder des kontrafaktischen Zustands kann der Selektionsfehler als Differenz zwischen dem wahren und dem geschätzten kontrafaktischen Zustand formalisiert werden:

$$SB = E[Y(0)|D = 1] - E[Y(0)|D = 0] \qquad (4)$$

Aufgrund der Existenz eines potentiellen Selektionsfehlers unbekannter Größe muss Formel (3) erweitert werden, sodass der Selektionsfehler bei der Schätzung von Treatmenteffekten berücksichtigt wird (siehe Formel (5)).

$$ATT = E[Y(1)|D = 1] - E[Y(0)|D = 0] + SB \qquad (5)$$

Generell gilt, dass Merkmalsunterschiede zwischen den Personen einer Versuchs- und einer Vergleichsgruppe nur dann ein Problem darstellen, wenn sie auch den Outcome beeinflussen. Dies wäre bspw. der Fall, wenn mit der Umsetzung einer Aufklärungsmaßnahme zum Klimaschutz angestrebt wird, die Teilnehmer zu einem klimafreundlicheren Verhalten zu bewegen, diese allerdings im Vorfeld schon ein höheres Umweltbewusstsein aufweisen und daher auch eher an der Maßnahme teilnehmen. Beobachtete Unterschiede im klimabezogenen Verhalten zwischen der Teilnehmergruppe und einer Vergleichsgruppe sind dann möglicherweise nicht auf die Maßnahmenteilnahme, sondern auf die Prädisposition Umweltbewusstsein zurückzuführen.

Solange dies nicht der Fall ist, die Teilnehmer der Maßnahme also nicht deshalb eher teilnehmen, weil sie über ein höheres Umweltbewusstsein verfügen, liegt keine Konfundierung vor und die Treatmenteffekte können unverzerrt geschätzt werden. Sobald (selektionsbedingte) Unterschiede zwischen den Gruppen allerdings auch die Outcome-Variable Y beeinflussen, liegt der Tatbestand der Konfundierung vor und die Wirkungen werden über- oder unterschätzt. Um fehlerhafte Effektschätzungen zu verhindern, müssen alle konfundierenden Einflüsse eliminiert oder zumindest konstant gehalten werden.

Die Bedingung, welche vorgibt, dass Unterschiede in den Outcomes zwischen Treatment- und Vergleichsgruppe unabhängig von Selektionsprozessen sein müssen, also ausschließlich auf das Treatment zurückgeführt werden können, wird ‚Conditional Independence Assumption' (CIA) (Rosenbaum & Rubin, 1983; D'Orazio et al., 2006) oder auch ‚Unconfoundedness' (Lechner, 1999) bezeichnet. Diese ist eine starke Annahme, da bei ihrer Erfüllung alle Kovariaten kontrolliert oder konstant gehalten werden müssen, die sowohl die Interventionsteilnahme als auch den Outcome beeinflussen. Ist die CIA nicht erfüllt, so könnten geschätzte Effekte teilweise, oder im Falle einer sehr starken Konfundierung sogar vollständig, durch unbeobachtete Konfundierungsfaktoren hervorgerufen und Treatmenteffekte damit über- oder unterschätzt werden (Reinowski, 2006). In der Sozialwissenschaft sind meist Menschen die Untersuchungsobjekte. Da sich Menschen generell voneinander unterscheiden, ggf. auch in Merkmalen, die eine Outcome-Variable beeinflussen, ist es nicht ratsam, im Rahmen eines Kontrollgruppendesigns lediglich den Gruppenmittelwert der Outcome-Variablen willkürlich oder bewusst ausgewählter Nichtteilnehmergruppen als Abbildung des kontrafaktischen Zustands zu verwenden. Es ist dann nämlich möglich, dass Faktoren, die die Teilnahmeentscheidung der Teilnehmer beeinflussen, gleichzeitig auch den Outcome beeinflussen (Caliendo & Kopeinig, 2008, S. 34) und damit eine Konfundierung vorliegt.

Für den angemessenen Umgang mit dem ‚Selection Bias' eignen sich zur kausalen Wirkungsevaluation am besten experimentelle Designs (bspw. Shadish & Cook, 2009). Durch die randomisierte Zuweisung von Personen zu Teilnehmer- und Nichtteilnehmergruppen wird eine systematische Selektion in die Gruppen ausgeschlossen. Es sei an dieser Stelle allerdings darauf hingewiesen, dass auch die Durchführung eines Experiments nicht zwangsläufig dazu führt, dass Treatment- und Kontrollgruppen bzgl. aller Drittvariablen identisch sind. Bestehende Unterschiede sind dann jedoch auf zufällige Variation zurückzuführen und werden mit steigender Fallzahl mehr und mehr zwischen Experimental- und Kontrollgruppe ausbalanciert, weshalb sie sich bei der Berechnung von Treatmenteffekten gegenseitig aufheben (Heckman & Smith, 1995, S. 89). Dies bedeutet, dass sich mit steigender Fallzahl die (randomisierten) Gruppen in den Merkmalen der Teilnehmer immer ähnlicher werden.

Experimentelle Designs nutzen weiter häufig eine Kombination aus Vorher- und Nachhermessung, um Veränderungsprozesse und unterschiedliche Ausgangsniveaus im Outcome kontrollieren und die Präzision der Wirkungsschätzung erhöhen zu können. Anhand der Herstellung einer Laborsituation kann im Rahmen von Experimenten zudem die Erhebungssituation kontrolliert werden. Experimenten wird daher generell eine hohe interne Validität nachgesagt, die allerdings oftmals auf Kosten der Übertragbarkeit und damit der externen Validität erreicht wird. Dies liegt bspw. im sog. ‚Randomization Bias' begründet (Heckman & Smith, 1995), da im Experiment Personen zufällig dem Treatment zugewiesen werden, die in der Realität dem Treatment vielleicht nicht ausgesetzt gewesen wären. Dieser Umstand hat jedoch keinen Einfluss auf die Überprüfung der prinzipiellen Wirksamkeit einer Intervention (interne Validität), kann jedoch die Übertragung der Befunde auf Teilnehmer in realen Interventionssituationen beeinträchtigen (externe Validität).

Die Beeinträchtigung der externen Validität ist vor allem dann gegeben, wenn die Personen aus der Experimental- und der Kontrollgruppe nicht repräsentativ für die Gesamtpopulation sind. Dies ist bspw. bei Experimenten mit Studentenstichproben der Fall. Diese weisen zwar häufig eine hohe interne Validität auf, die Ergebnisse sind aufgrund der Tatsache, dass nur Studenten am Experiment teilnehmen, allerdings nicht ohne weiteres auf die Gesamtbevölkerung übertragbar.

Sofern die Durchführung eines Experiments nicht möglich ist, gibt es andere Ansätze zur Lösung des Selektionsproblems wie zum Beispiel den ‚Regression-Discontinuity-Ansatz' (Imbens & Lemieux, 2007) oder verschiedene Matching-Verfahren (Caliendo & Hujer, 2006; Heckman et al., 1998b), von denen in Gestalt des ‚Propensity Score Matching' ein inzwischen relativ weit verbreitetes ökonometrisches Adjustierungsverfahren im vorliegenden Leitfaden vorgestellt wird. Neben den beiden genannten existieren weitere Ansätze zur Lösung des Selektionsproblems und zur Schätzung kausaler Wirkungseffekte, zu denen an dieser Stelle allerdings auf die entsprechende Fachliteratur verwiesen wird (bspw. Shadish & Cook, 2009; Reinowski, 2006; Caliendo & Hujer, 2006; Rossi et al., 2004; Linden & Adams, 2010).

Zusammenfassung

- ✓ Eine vollständig kausale Zuschreibung von Wirkungen zu einer Maßnahme erfordert den Ausschluss aller möglichen Alternativerklärungen.
- ✓ Zur Beschreibung kausaler Wirkungen eignet sich das Konzept des kontrafaktischen Zustands. Dieser beschreibt den hypothetischen Zustand, in dem sich die Untersuchungsobjekte befinden würden, wenn sie nicht an einer Maßnahme teilgenommen hätten.
- ✓ Die Differenz zwischen faktischem und kontrafaktischem Zustand wird als kausale Wirkung interpretiert, da sich die Untersuchungsobjekte in beiden Zuständen lediglich darin unterscheiden, ob sie an einer Maßnahme teilgenommen haben oder nicht.
- ✓ Da der kontrafaktische Zustand (in aller Regel) nicht gemessen werden kann, muss er künstlich hergestellt werden. Dies geschieht sehr häufig durch die Verwendung von Kontroll- bzw. Vergleichsgruppen.
- ✓ Da sich Personen und damit auch Gruppen immer voneinander unterscheiden, besteht die Gefahr eines Selektionsfehlers bzw. einer Konfundierung. Konfundierung liegt dann vor, wenn (selektionsbedingte) Unterschiede zwischen Versuchs- und Vergleichsgruppe auch den Outcome von Interesse beeinflussen.
- ✓ Um das Selektionsproblem und damit auch die Gefahr der Konfundierung zu lösen, eignen sich vor allem experimentelle Untersuchungsdesigns, da systematische Selektionsprozesse durch die randomisierte Zuweisung von Personen zu Experimental- und Kontrollgruppe ausgeschlossen werden.
- ✓ Sofern die Durchführung von Experimenten nicht möglich ist, muss auf andere Strategien zur Lösung des Selektionsproblems wie bspw. Matching-Verfahren zurückgegriffen werden.

3. Propensity Score Matching

Während das Experiment dem beschriebenen Selektionsproblem mit der randomisierten Zuteilung von Personen zu Experimental- und Kontrollgruppe begegnet, muss für Situationen, in denen keine Randomisierung möglich ist, auf andere Strategien zur Lösung des Selektionsproblems zurückgegriffen werden. Ein Ausweichen auf den vielversprechenden Regression-Discontinuity-Ansatz, bei dem eine Zuordnung von Personen zu Teilnehmer- und Nichtteilnehmergruppe ebenfalls kontrolliert erfolgt, wird durch die Freiwilligkeit der Maßnahmenteilnahme in der Evaluationspraxis häufig unterminiert. In Studien, in denen der Forscher keinen Einfluss auf die Zuordnung von Untersuchungsobjekten zu Versuchs- und Vergleichsgruppe besitzt, vollzieht sich die Schätzung von Treatmenteffekten allerdings unter dem potentiellen Einfluss selektionsbedingter Verzerrungen. Zur Schätzung von Treatmenteffekten wird daher ein Analyseverfahren benötigt, welches nicht auf den Ausschluss von

(Selbst-) Selektionsprozessen angewiesen ist, sondern einen adäquaten Umgang mit diesen verspricht.

Ein Ansatz zur Lösung des Selektionsproblems ist der Matching-Ansatz. Die Grundannahme des Matchings liegt darin, dass für jede Person, die an einer Interventionsmaßnahme teilnimmt, der kontrafaktische Zustand mit Hilfe einer eigens generierten Unter-Kontrollgruppe aus Nichtteilnehmern konstruiert werden kann (bspw. Reinowski, 2006). Mit Hilfe des Matchings wird versucht, Interventionsteilnehmern möglichst ähnliche Vergleichsgruppenpersonen gegenüberzustellen, um dann durch die Bildung von meist gewichteten, durchschnittlichen Gruppendifferenzen Wirkungseffekte einer Intervention abschätzen zu können. Durch die Zuordnung möglichst ähnlicher Personen wird dabei eine Annäherung an den kontrafaktischen Zustand angestrebt; die aus Selektionsprozessen entstandenen Verzerrungen sollen so minimiert werden. Die Anzahl der für die Berechnung der individuellen Differenzen verwendeten Vergleichsgruppenpersonen kann dabei, ebenso wie die Art der Gewichtung, je nach methodischem Ansatz bzw. angewendetem Matching-Algorithmus stark variieren. Grundsätzlich gilt jedoch, dass die zum Matching verwendeten Personen der Vergleichsgruppe auf einem definierten Set von Variablen (den Kovariaten) den Personen aus der Treatmentgruppe so ähnlich wie möglich sein sollten.

Als präferiertes Matching-Verfahren zur Bildung einer Vergleichsgruppe wird in dieser Arbeit das inzwischen relativ weit verbreitete ‚Propensity Score Matching' (PSM) gewählt (für Beispiele siehe Luellen et al., 2005), welches auf Rosenbaum & Rubin (1983, 1985) zurückgeht. Der Propensity Score P(X) ist ein sog. ‚Balancing Score', ein eindimensionales Maß, welches als die bedingte Wahrscheinlichkeit der Teilnahme an einer Intervention auf Basis beobachteter Merkmale definiert ist. Der Vorteil des PSM gegenüber anderen Verfahren zur Kovariatenkontrolle ist einerseits in der Reduktion der Dimensionalität der zum Matching verwendeten Merkmale zu sehen (Dehejia & Wahba, 2002, S. 151), andererseits in der nichtparametrischen Anlage des Verfahrens (Jaenichen, 2002, S. 394).

Zwar kommen einige Studien zu dem Ergebnis, dass die Schätzung von Treatmenteffekten auf Basis des Propensity Score nicht dieselben Ergebnisse liefert wie die Effektberechnung anhand experimenteller Kontrolle (bspw. Wilde & Hollister, 2007; Agodini & Dynarski, 2004; LaLonde, 1986), andere Studien zeigen jedoch, dass sich die auf dem PSM basierenden Schätzer bei hoher Datenqualität und hochwertigen Designs durchaus zur Schätzung kausaler Treatmenteffekte eignen (bspw. Cook & Steiner, 2010; Cook et al., 2008; Smith & Todd, 2001). Eine hohe Datenqualität umfasst neben der Auswahl der relevanten Kovariaten, ausreichend großer Stichproben[3] und der reliablen Messung[4] des Outcomes und der Kovariaten auch die Vollständigkeit des Datensatzes[5]. Es ist zudem eine intensive Diskussion darüber im Gange, ob die Nutzung von Propensity Scores bessere Ergebnisse im Rahmen der

[3] Für die Verwendung der Probit-Regression zur Schätzung des Propensity Score wird bspw. empfohlen, eine Fallzahl von mindestens n = 100 zu verwenden (bspw. Long, 1997). Die Anwendung logistischer Regressionen ist dagegen auch mit geringeren Fallzahlen möglich.

[4] Bei Skalen sollte Cronbach's α bspw. mindestens den Wert 0.7 aufweisen.

[5] Bei Unvollständigkeit der Daten werden diejenigen Fälle, für die nicht alle Informationen zu Outcome und/oder Kovariaten verfügbar sind, aus den Analysen ausgeschlossen. Sofern die Anzahl fehlender Werte keine statistischen Analysen mehr ermöglicht, können Werte imputiert werden (bspw. Schafer & Graham, 2002).

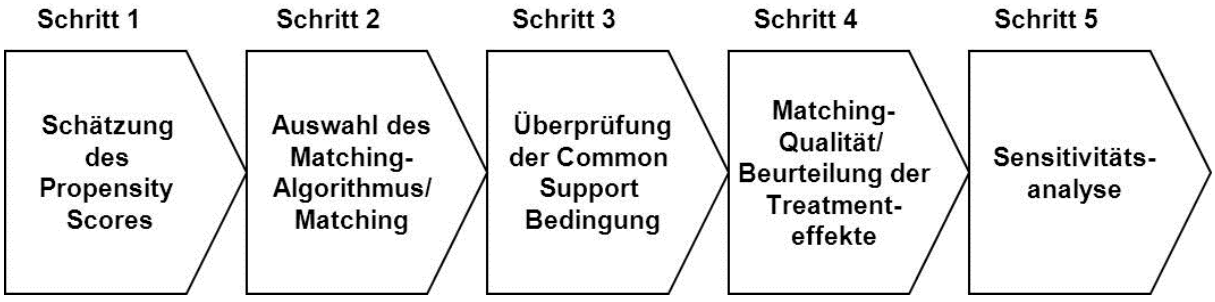
Schätzung von Treatmenteffekten liefert als gängigere statistische Verfahren wie bspw. OLS- oder logistische Regressionen (bspw. Cepeda et al., 2003). Die Ergebnisse einiger Simulations- und Vergleichsstudien deuten darauf hin, dass die Wahl des statistischen Verfahrens im Vergleich zur Kovariatenauswahl eine eher untergeordnete Rolle spielt (bspw. Cook & Steiner, 2010; Pohl et al., 2009; Shadish et al., 2008).

Bei der Schätzung von Propensity-Scores als bedingte Wahrscheinlichkeiten für die Teilnahme an einer Intervention wird angestrebt, alle Merkmale X, die simultan für die Interventionsteilnahme und für den interessierenden Outcome relevant sind, zu erfassen. Diese Basisannahme bezieht sich auf die bereits erwähnte CIA. Sie besagt, dass Unterschiede in den Outcomes zwischen Treatment- und Vergleichsgruppe unabhängig von Selektionsprozessen sein müssen, also ausschließlich auf das Treatment zurückgeführt werden können. Die Verletzung der CIA führt damit zu einer niedrigeren Robustheit geschätzter Treatmenteffekte. Aus diesem Grund ist die Durchführung einer Sensitivitätsanalyse zur Abschätzung der Robustheit der Treatmenteffekte gegenüber unbeobachteten Konfundierungsfaktoren erforderlich. Die Sensitivitätsanalyse stellt folglich ein Instrument dar, dessen Anwendung die Überprüfung der Robustheit geschätzter Treatmenteffekte in potentiellen Fehlerszenarien erlaubt.

Eine weitere Annahme, die bei der Durchführung des PSM berücksichtigt werden muss, ist die ‚Stable-Unit-Treatment-Value-Assumption' (SUTVA). Sie besagt vor allem, dass der Treatmenteffekt auf eine Person nicht von der Teilnahme anderer Personen an derselben Intervention beeinflusst werden darf, die Stabilität der Kausalwirkung also gegeben sein muss (bspw. Gangl & DiPrete, 2004).

Der gesamte Prozess des Propensity Score Matching vollzieht sich nun in mehreren Schritten. Die nachfolgende Abbildung 1 basiert (leicht abgeändert) auf Caliendo & Kopeinig (2008, S. 33) und beschreibt den Verlauf des Propensity Score Matching Prozesses.

Abbildung 1: Verlauf des Propensity Score Matching



Im folgenden Teilabschnitt wird zunächst dargelegt, wie der Propensity Score geschätzt werden kann und welche Bedingungen hierbei erfüllt sein müssen.

Zusammenfassung

- ✓ Sofern der Evaluator keinerlei Einfluss auf die Zuweisung der Untersuchungsobjekte zu Versuchs- und Vergleichsgruppe hat, können keine experimentellen oder Regression-Discontinuity-Designs zur Anwendung kommen.
- ✓ Eine weitere Strategie zur Lösung der Selektionsproblematik stellen Matching-Verfahren dar. Die Grundannahme des Matchings liegt darin, dass für jedes Untersuchungsobjekt, das einem Treatment ausgesetzt ist, der kontrafaktische Zustand mit Hilfe einer eigens generierten Unter-Kontrollgruppe aus Untersuchungsobjekten, die dem Treatment nicht ausgesetzt waren, konstruiert werden kann.
- ✓ Mit Hilfe des Matchings wird demnach angestrebt, Maßnahmenteilnehmern möglichst ähnliche Vergleichsgruppenpersonen gegenüberzustellen, um dann durch die Bildung von meist gewichteten, durchschnittlichen Gruppendifferenzen Wirkungseffekte einer Intervention abschätzen zu können.
- ✓ Ein mögliches Matching-Verfahren stellt das sog. Propensity Score Matching (PSM) dar. Der Propensity Score (PS) ist ein eindimensionales Maß, welches als die bedingte Wahrscheinlichkeit der Teilnahme an einer Intervention auf Basis beobachteter Merkmale definiert ist. Zwei Untersuchungsobjekte, die einen sehr ähnlichen PS aufweisen, sind sich auch in den Kovariaten sehr ähnlich und können daher miteinander verglichen werden.
- ✓ Von entscheidender Bedeutung für das PSM ist die Datenqualität: D.h. es sollten möglichst alle relevanten Kovariaten berücksichtigt werden, die Stichproben sollten groß genug sein, die Messungen sollten reliabel und der Datensatz vollständig sein.

3.1 Schätzung des Propensity Score

Die Schätzung des Propensity Score basiert prinzipiell auf zwei Entscheidungen des Forschers, welche die Auswahl des Schätzverfahrens und die Auswahl der für die Schätzung der zu verwendenden Kovariaten betreffen.

Die Art des zu wählenden Schätzverfahrens hängt maßgeblich davon ab, wie viele Stufen das Treatment besitzt bzw. wie viele Gruppen in eine Untersuchung einbezogen werden. Da sich in den empirischen Analysen dieser Arbeit ausschließlich der Zweigruppenfall mit Versuchs- und Vergleichsgruppe wiederfindet, werden zur Schätzung von Propensity Scores lediglich binäre diskrete Entscheidungsmodelle benötigt. In der Regel werden aufgrund ihrer Eigenschaften hierfür entweder die Probit- oder die Logit-Regression verwendet. Für binäre Treatmentvariablen spielt die Wahl zwischen Logit- und Probit-Verfahren allerdings nur eine untergeordnete Rolle, da beide Verfahren zu ähnlichen Resultaten führen (Caliendo & Kopeinig, 2008, S. 37) und sich lediglich in den Annahmen über die Verteilung der Residuen unterscheiden (vgl. Greene, 2009). In diesem Leitfaden wird das häufig verwendete Probit-

Verfahren zur Schätzung des Propensity Scores genutzt, welches auch im für die Durchführung des PSM verwendeten Stata-Modul *psmatch2* nach Leuven & Sianesi (2003) als Standardverfahren festgelegt ist.

Ungleich wichtiger als die Auswahl des Schätzverfahrens ist die Wahl der für die Schätzung von Propensity Scores benötigten Kovariaten (bspw. Cook & Steiner, 2010; Shadish et al., 2008; Brookhart et al., 2006). Wie eingangs erwähnt, ist der Propensity Score als bedingte Wahrscheinlichkeit der Interventionsteilnahme auf Basis beobachteter Merkmale definiert. Auf Basis dieser beobachteten Merkmale (Kovariaten) wird für jede Person aus Versuchs- und Vergleichsgruppe die Wahrscheinlichkeit geschätzt, Mitglied der Versuchsgruppe zu sein. Diese Wahrscheinlichkeit basiert ausschließlich auf den in der Schätzgleichung verwendeten Kovariaten. Bereits an dieser Stelle wird ersichtlich, dass die Auswahl der Kovariaten eine zentrale Rolle für das Matching spielt. Betrachtet man weiter die bereits dargelegte Annahme der ‚Unconfoundedness', so müssten vom Forscher theoretisch alle Kovariaten mit einbezogen werden, die einen Einfluss auf die Zuordnungswahrscheinlichkeit und den Outcome haben. In der empirischen Forschungspraxis ist dieses Vorgehen praktisch nicht möglich, weshalb in der Regel versucht wird, so viele Kovariaten wie möglich in die Schätzung mit einzubeziehen.

Zur Auswahl potentieller Kovariaten stehen statistische Methoden (Black & Smith, 2004; Heckmann et al., 1998b) wie bspw. die Prüfung statistischer Signifikanz einer Kovariaten im Probit-Modell zur Verfügung. Da es bei der Schätzung des Propensity Score allerdings nicht darauf ankommt, die Teilnahme an einer Intervention möglichst gut und signifikant vorauszusagen, sondern Kovariaten über die Gruppen hinweg auszubalancieren, spielt dieser Sachverhalt in der Praxis oft nur eine untergeordnete Rolle. Stattdessen werden die Kovariaten häufig nach Verfügbarkeit ausgewählt und alle, ob signifikant oder nicht, werden in den Schätzmodellen beibehalten. Damit wird der Auffassung von Rubin & Thomas (1996) gefolgt, welche sich für ein Belassen von Kovariaten im Modell aussprechen, sofern nicht eindeutig feststeht, dass die Kovariaten weder den Outcome beeinflussen noch überhaupt als Kovariaten geeignet sind. Dies wäre bspw. dann der Fall, wenn die Interventionsteilnahme die Ausprägungen der Kovariaten verändern würde.

Die eigentliche Schätzung des Propensity Score ist nach Festlegung des Modells und der Auswahl der Kovariaten nichts anderes als die Anwendung einer einfachen Probit- bzw. Logit-Regression[6]. Die Stärke der Koeffizienten sowie deren Signifikanz spielen für das PSM eine untergeordnete Rolle. Der Determinationskoeffizient Pseudo-R^2 lässt sich hingegen als Maßzahl für die Heterogenität zwischen Versuchs- und Vergleichsgruppe interpretieren. Ebenfalls von Relevanz ist der LR-χ^2-Test, der darüber Aufschluss gibt, ob überhaupt eine einzige Kovariate einen Einfluss auf die Zuordnung von Versuchs- und Vergleichsgruppe besitzt, der signifikant verschieden von null ist. Abgesehen von metrischen Kovariaten werden die Kategorien nominal und ordinal skalierter Kovariaten als Dummy-Variablen kodiert und als binäre Kovariaten in die Schätzfunktion einbezogen. Der Mittelwert einer solchen Kovariate entspricht dann dem prozentualen Anteil des Merkmals in den Stichproben. Diese

[6] Vgl. zur Berechnung sowie zur statistischen Evaluation der Ergebnisse *Greene (2009).*

Vorgehensweise hat auch den Vorteil, dass im Nachgang zum Matching bestimmte Tests zur Überprüfung der Matching-Qualität durchführbar sind (siehe Abschnitt 3.3).

Zusammenfassung

- ✓ In der Regel werden zur Schätzung des PS binäre diskrete Entscheidungsmodelle, also Logit- oder Probit-Regressionen, verwendet. Sofern die Fallzahl groß genug ist (bspw. n > 100), spielt die Wahl des Verfahrens nur eine untergeordnete Rolle. Bei kleineren Stichproben sollte eher auf das Logit-Verfahren zurückgegriffen werden.
- ✓ Zentral für die Güte des Matchings auf Basis des PS ist die Auswahl der Kovariaten! Es sollten so viele Kovariaten wie möglich, vor allem aber die „richtigen" Kovariaten berücksichtigt werden (also diejenigen, die sowohl mit der Zugehörigkeit zur Treatmentgruppe als auch mit dem Outcome hoch korrelieren).

3.2 Auswahl eines Matching Algorithmus und Common Support

Nachdem der Propensity Score für alle Personen der Versuchs- und Vergleichsgruppe geschätzt wurde, kann mit dem eigentlichen Matching-Prozess begonnen werden. Zur Durchführung stehen verschiedene Matching-Algorithmen zur Verfügung, welche sich in der Art der Zuordnung von Vergleichsgruppenpersonen zu einer Versuchsgruppenperson und in der Art der Gewichtung voneinander unterscheiden. Um potentielle Treatmenteffekte auf Sensitivität gegenüber einem bestimmten Algorithmus zu überprüfen, kommen in der Praxis häufig unterschiedliche Algorithmen für ein- und dieselbe Effektschätzung zum Einsatz. Im Folgenden werden drei ‚typische' Algorithmen vorgestellt. Für eine weiterführende Diskussion von Algorithmen sei an dieser Stelle auf die einschlägige Fachliteratur verwiesen (bspw. Guo & Fraser, 2010).

Nearest-Neighbour-Matching (NNM):

Der klassische Nearest-Neighbour-Algorithmus (‚One-to-One-Matching') ist ein weit verbreiteter und zugleich relativ simpler Zuordnungsalgorithmus. Im Falle des NNM wird einer Person der Versuchsgruppe diejenige Person der Vergleichsgruppe zugeordnet, welche ihr in der Ausprägung des Propensity Score am nächsten ist oder anders ausgedrückt, welche die ähnlichste Teilnahmewahrscheinlichkeit besitzt.

In der Regel wird das NNM ‚mit Zurücklegen' durchgeführt. Eine Person der Vergleichsgruppe, die bereits einer Person der Versuchsgruppe zugeordnet wurde, ist dann nicht ‚verbraucht', sondern kann weiteren Personen der Versuchsgruppe als Matching-Partner dienen. Auf diese Weise wird verhindert, dass Personen mit sehr unterschiedlichem Propensity Score einander zugeordnet werden (‚bad matches'). Es wird beim NNM vorgeschlagen, mehr als nur eine Person der Vergleichsgruppe einer Person der Versuchsgruppe zuzuordnen (Caliendo & Kopeinig, 2008, S. 42). Einer Person der Versuchsgruppe werden dann die k

Personen der Vergleichsgruppe zugeordnet, die ihr bzgl. des Propensity Scores am ähnlichsten sind. Dieses Vorgehen führt zu einer reduzierten Varianz des Effektschätzers, da mehr Informationen in die Berechnung individueller Treatmenteffekte einbezogen werden. Leider führt es aber auch zu einem erhöhten Bias, da auch schlechtere Matches verwendet werden müssen (bspw. Smith & Todd, 2005).

Kernel-Matching:

Der zweite in der Studie verwendete Matching-Algorithmus ist das Kernel-Matching, welches einen Gegenpol zum k-NNM mit k = 1 darstellt. Beim Kernel-Matching handelt es sich um einen nicht-parametrischen Matching-Algorithmus, der gewichtete Durchschnittswerte von allen (oder fast allen)[7] Individuen in der Vergleichsgruppe benutzt, um den kontrafaktischen Zustand einer Versuchsgruppenperson abzubilden (Caliendo & Kopeinig, 2008, S. 43).

Der große Vorteil dieser Vorgehensweise besteht in einer deutlich reduzierten Varianz der Schätzer aufgrund der vielen einbezogenen Informationen und der damit verbundenen erhöhten Präzision der Schätzungen. Da allerdings auch sehr schlechte Matches, also Personen der Vergleichsgruppe, die einen sehr unterschiedlichen Propensity Score im Vergleich zur Person der Versuchsgruppe aufweisen, verwendet werden, steigt der ‚bias' bei Verwendung des Kernel-Algorithmus an.

Für die Durchführung des Kernel-Matchings muss der Forscher zwei Entscheidungen treffen: Zunächst muss er sich zur Gewichtung der Nichtteilnehmer für eine Dichte-Funktion entscheiden. Im vorliegenden Leitfaden wird der Gauß-Kern verwendet. Ebenfalls gängig wäre der Epanechnikov-Kern, dessen Ergebnisse sich häufig allerdings nicht nennenswert von denen des Gauß-Kerns unterscheiden. Einen größeren Unterschied für die Ergebnisse macht hingegen die gewählte Bandbreite der Dichte-Funktion, die vom Forscher festgesetzt werden muss (bspw. Pagan & Ullah, 1999). Genaue Regeln, welche Bandbreite verwendet werden sollte, existieren allerdings nicht.

Radius Matching:

Dehejia & Wahba (2002) schlagen das Radius-Matching als eine Art Zwischenstufe zwischen den beiden Polen des One-to-One-Matching und des Kernel-Matching vor. Beim Radius-Matching werden nur diejenigen Personen der Vergleichsgruppe zum Matching herangezogen, die sich in einem definierten Abstandsbereich (‚caliper') des Wertes des Propensity Scores der zu matchenden Versuchsperson befinden. Dehejia & Wahba (2002, S. 153/154) merken hierzu an: „A benefit of caliper matching is that it uses only as many comparison units as are available within the calipers, allowing for the use of extra (fewer) units when good matches are (not) available". Für die Durchführung des Radius Matchings lässt sich in den Analysen der Abstandsbereich frei wählen, wobei leider keine detaillierten Richtlinien zur Wahl des Calipers zur Verfügung stehen.

[7] Dies hängt von der verwendeten Dichte-Funktion ab.

Bevor mit der Durchführung des Matchings und der Schätzung von Treatmenteffekten begonnen werden kann, gilt es noch sicherzustellen, dass die Bedingung des ‚Common Support' erfüllt ist. Hierunter versteht man den Wertebereich, in dem der Propensity Score für die Untersuchungseinheiten aus Versuchs- und Vergleichsgruppe eine ähnliche Dichte hat. Ist dies nicht gegeben, sind das Matching und die darauf basierende Schätzung von Treatmenteffekten nicht möglich (Diaz & Handa, 2006, S. 325). Die für das PSM verwendete Stata-Applikation *psmatch2* bietet die Option an, alle Matching-Prozeduren und Schätzungen von Treatmenteffekten nur auf Basis derjenigen Fälle durchzuführen, die sich im Bereich des Common Support befinden.

Zusammenfassung

- ✓ Für die Durchführung des PSM stehen verschiedene Algorithmen zur Verfügung, die sich in der Art und Weise der Bildung des individuellen kontrafaktischen Zustands für eine Person der Versuchsgruppe voneinander unterscheiden.
- ✓ Es gibt keine Richtlinien dafür, welcher Algorithmus der „beste" ist. Alle Algorithmen weisen bestimmte Vor- und Nachteile auf. Um die Sensitivität der Effektschätzungen gegenüber dem verwendeten Verfahren zu reduzieren, kommen in der Praxis häufig mehrere Algorithmen zum Einsatz.
- ✓ Grundsätzlich gilt: Je mehr Personen zur Bildung des individuellen kontrafaktischen Zustands herangezogen werden, desto präziser wird die Schätzung aufgrund der reduzierten Varianz der Schätzer und desto größer wird der ‚bias', da auch schlechte Matches in die Schätzung mit einbezogen werden.
- ✓ Zur Erhöhung der Güte des PSM sollten nur diejenigen Fälle herangezogen werden, die die Bedingung des ‚Common Support' erfüllen. Diese sagt aus, dass der PS für die Untersuchungseinheiten aus Versuchs- und Vergleichsgruppe in einem Wertebereich eine ähnliche Dichte aufweist. Dies ist erforderlich, damit die Fälle aus Versuchs- und Vergleichsgruppe überhaupt miteinander verglichen werden können.

3.3 Matching-Qualität und Beurteilung der Treatmenteffekte

Ist der Prozess des Matchings abgeschlossen, so muss zunächst die Qualität des Matchings überprüft werden. Hierbei ist von Interesse, ob der geschätzte Propensity Score dazu in der Lage ist, die einzelnen Kovariaten zwischen den Gruppen auszubalancieren. Ist dies der Fall, dann ist die sog. ‚Balancing Property' erfüllt. Diese sagt aus, dass Untersuchungsobjekte mit dem gleichen Propensity Score die gleiche Verteilung von beobachteten (und unbeobachteten) Charakteristika unabhängig davon aufweisen, ob sie der Versuchs- oder Vergleichsgruppe angehören. Anders ausgedrückt bedeutet dies, dass die Zuordnung zu Versuchs- oder Vergleichsgruppe als zufällig aufgefasst werden kann und Interventionsteilneh-

mer und Nichtteilnehmer im Durchschnitt als näherungsweise identische Untersuchungsobjekte aufgefasst werden können (Becker & Ichino, 2002, S. 359).

Zur Überprüfung der Balancing Property kommen im vorliegenden Leitfaden zwei Methoden zum Einsatz. Zunächst wird überprüft, ob signifikante Unterschiede in einer Kovariate vor der Durchführung der Matching-Prozedur auf Basis des Propensity Score nach Beendigung des Prozesses immer noch signifikant sind. Damit die Kovariaten als näherungsweise ausbalanciert gelten können, darf sich nach dem Matching keine der Kovariaten mehr signifikant zwischen den Gruppen unterscheiden. In Anlehnung an Rosenbaum & Rubin (1985) können hierzu zweiseitige T-Tests durchgeführt werden.

Ein weiterer Ansatz geht zurück auf Sianesi (2004) und sieht vor, den Propensity Score nach dem Matching noch einmal neu zu berechnen und den Wert des Pseudo-R^2 mit demjenigen vor dem Matching zu vergleichen. Da das Pseudo-R^2 angibt, wie gut die Kovariaten die Teilnahmewahrscheinlichkeit erklären, sollte es nach dem Matching sehr niedrig und vor allem deutlich niedriger als vor dem Matching sein. Die gleiche Logik lässt sich auf den LR-χ^2-Test übertragen. Nach dem Matching sollte die Nullhypothese eines gemeinsamen Effekts von null nicht mehr verworfen werden können, es sollte also $p > .1$ gelten. Der Grad an Ausbalancierung von Kovariaten zwischen den Gruppen ist letztlich auch ein Kriterium dafür, welcher Matching-Algorithmus sich am besten zur Analyse eignet. Hierzu heißt es bei Morgan & Winship (2007, S. 114), dass generell anerkannt ist, dass die besten Matching-Algorithmen diejenigen sind, die die Balance in den analysierten Daten optimieren.

Zur Beurteilung der Treatmenteffekte stehen dem Forscher verschiedene Möglichkeiten zur Verfügung. So kann bspw. einfach abgelesen werden, ob das Vorzeichen des Treatmenteffekts in die postulierte Richtung zeigt. Zudem kann die Stärke des geschätzten Effekts beurteilt werden, bspw. indem man den Effekt in Relation zur maximalen Skalenbandbreite setzt. Ob ein Effekt als stark oder schwach zu betrachten ist, kann allerdings nur im Kontext der Untersuchung eingeschätzt werden. Schließlich ist es möglich, Standardfehler und darauf basierende T-Werte zu berechnen, anhand derer einseitige T-Tests zur Signifikanzprüfung von Treatmenteffekten durchgeführt werden können.

Im vorliegenden Leitfaden werden die T-Werte einmal auf Basis der von Leuven & Sianesi (2003) implementierten Berechnungsweise in *psmatch2* berechnet, welche auf Lechner (2001) basiert[8]. Eine zweite Möglichkeit zur Berechnung von Standardfehlern besteht in der Anwendung des Bootstrap-Resamplings (bspw. Lechner, 2002). Hierbei werden aus der bestehenden Stichprobe zufällig Stichproben mit Zurücklegen gezogen. Für jede der zufällig gezogenen Stichproben wird der gesamte Propensity Score Matching Prozess erneut durchgeführt (Berechnung der Propensity Scores, Bestimmung des Common Support, Schätzung der Treatmenteffekte etc.), was bei N gezogenen Stichproben zu N neu geschätzten Treatmenteffekten führt. Die Verteilung dieser geschätzten Effekte wird als Stichprobenverteilung

[8] Die Berechnung basiert auf den Annahmen unabhängiger Untersuchungsobjekte, fixierter Gewichte, Homoskedastizität der Outcome-Variablen sowohl in der Versuchs- als auch der Vergleichsgruppe sowie auf der Annahme, dass die Varianz in der Outcome-Variablen unabhängig vom Propensity Score ist.

eines Treatmenteffekts aufgefasst, aus welcher sich Standardfehler extrahieren lassen und welche als Approximation an die Populationsverteilung betrachtet werden kann (Caliendo & Kopeinig, 2008, S. 53). Da es sich beim nachfolgenden Beispiel um eine eher explorative Feldstudie handelt, wird ein Signifikanzniveau von α = 10% als akzeptabel betrachtet.

Zusammenfassung

- ✓ Die Qualität des PSM bemisst sich weitestgehend daran, inwiefern es einem verwendeten Algorithmus gelingt, die verwendeten Kovariaten zwischen den Gruppen auszubalancieren (‚Balancing Property').
- ✓ Generell gilt: Derjenige Algorithmus, der die Balance in den analysierten Daten optimiert, weist die höchste Qualität auf.
- ✓ Zur Überprüfung der Matching-Qualität existieren verschiedene statistische Tests. So kann anhand von zweiseitigen T-Tests bspw. überprüft werden, ob die Balance der Kovariaten zwischen den Gruppen nach dem Matching besser ist als vorher.
- ✓ Die Beurteilung der Treatmenteffekte erfolgt anhand von drei Kriterien:
  - ✓ Die Richtung des Effekts (negatives vs. positives Vorzeichen)
  - ✓ Die Stärke des Effekts (klare Kriterien, ab wann ein Effekt stark ist, gibt es allerdings nicht!)
  - ✓ Die Signifikanz des Effekts (zur Berechnung von Standardfehlern stehen verschiedene Verfahren zur Verfügung)

3.4 Sensitivitätsanalyse

Wie eingangs bereits erwähnt, ist in Selektionsprozessen, die die Zuordnung der Teilnehmer zu Versuchs- und Vergleichsgruppe und simultan auch die Outcome-Variable beeinflussen, ein großes Problem für die kausale Wirkungsabschätzung zu sehen. Selbst bei der Erhebung unzähliger Kovariaten kann nicht mit absoluter Sicherheit davon ausgegangen werden, dass nicht doch unbeobachtete Konfundierungsfaktoren die Treatmenteffekte verzerrt haben und damit die CIA verletzen. Um die Güte geschätzter Treatmenteffekte auch im Fall einer Verletzung der CIA beurteilen zu können, kann im Anschluss an die Schätzung der Treatmenteffekte eine Sensitivitätsanalyse auf Basis von ‚Rosenbaum Bounds' (Rosenbaum, 2002; DiPrete & Gangl, 2004) und des in diesem Zusammenhang ebenfalls von Rosenbaum (1993) vorgeschlagenen ‚Hodges-Lehmann-Punktschätzer' durchgeführt werden. Anhand der Sensitivitätsanalyse kann abgeschätzt werden, wie stark eine nicht erfasste Drittvariable, die auch die Outcome-Variable beeinflusst, den Selektionsprozess zu Teilnehmer-/ Nichtteilnehmergruppen beeinflussen müsste, um die Robustheit der Schätzung der Treatmenteffekte gegenüber einem Selektionsfehler zu gefährden.

Bei der Schätzung von Rosenbaum Bounds, welche obere (p_+) und untere (p_-) Signifikanzgrenzen für die Treatmenteffekte darstellen, geht man davon aus, dass mit zunehmender

Positiv- bzw. Negativselektion in die Treatment- oder Vergleichsgruppe die Treatmenteffekte über- bzw. unterschätzt werden, die Wahrscheinlichkeit eines positiven Treatmenteffekts im Fall von Positivselektion also abnimmt und bei Auftreten einer Negativselektion steigt. Die Berechnung von Rosenbaum Bounds, Hodges-Lehmann-Punktschätzern sowie von 95%-Konfidenzintervallen für die Treatmenteffekte kann mit Hilfe des Stata-Moduls *rbounds* nach Gangl (2004) durchgeführt werden.

Dabei bezeichnet Gamma (Γ) die Odds-Ratio der Zuordnung von Personen zu Versuchs- und Vergleichsgruppe, die durch unbeobachtete Merkmale beeinflusst wird. Ein Γ von 1 bedeutet demnach ein Chancenverhältnis von 1:1, welches damit demjenigen entspricht, welches unter Randomisierung gegeben wäre. Dies würde bedeuten, dass keine unbeobachteten Faktoren Einfluss auf die Zuordnung zu Versuchs- und Vergleichsgruppe und simultan auf die Outcome-Variable nehmen, also keine Konfundierung vorliegt. In diesem Fall wären die Treatmenteffekte unverzerrt und die CIA wäre erfüllt. Anhand einer Steigerung des Parameters Γ und den sich entsprechend verändernden Signifikanzobergrenzen p_+ lässt sich bspw. abschätzen, wie hoch Γ maximal werden darf, bevor die Robustheit eines positiven Treatmenteffekts nicht mehr gegeben ist, der geschätzte Effekt also nicht mehr nur auf das Treatment, sondern maßgeblich auf Selektionsprozesse und die damit verbundene Konfundierung zurückzuführen ist.

Für jeden Matching-Algorithmus kann in den Analysen der maximal zulässige Wert von Γ berechnet werden, bei dem die Treatmenteffekte noch signifikant sind, also mit großer Wahrscheinlichkeit nicht stark über- oder unterschätzt werden. Die Nullhypothese, die unterstellt, dass der geschätzte Effekt ausschließlich auf Konfundierung zurückzuführen ist, kann bis zu dieser maximalen Höhe von Γ_{max} mit $p < .1$ abgelehnt werden. Hinsichtlich der Bewertung des Robustheitsgrads bezeichnet bspw. Aakvik (2001, S. 132) ein Γ von 2.0, unter dem die Treatmenteffekte noch signifikant sind, als „very large“. Der Hodges-Lehmann-Punktschätzer gibt unter der Annahme additiver Treatmenteffekte zudem an, wie stark der Treatmenteffekt im Γ_{max}-Szenario mindestens noch sein sollte.

Zusammenfassung

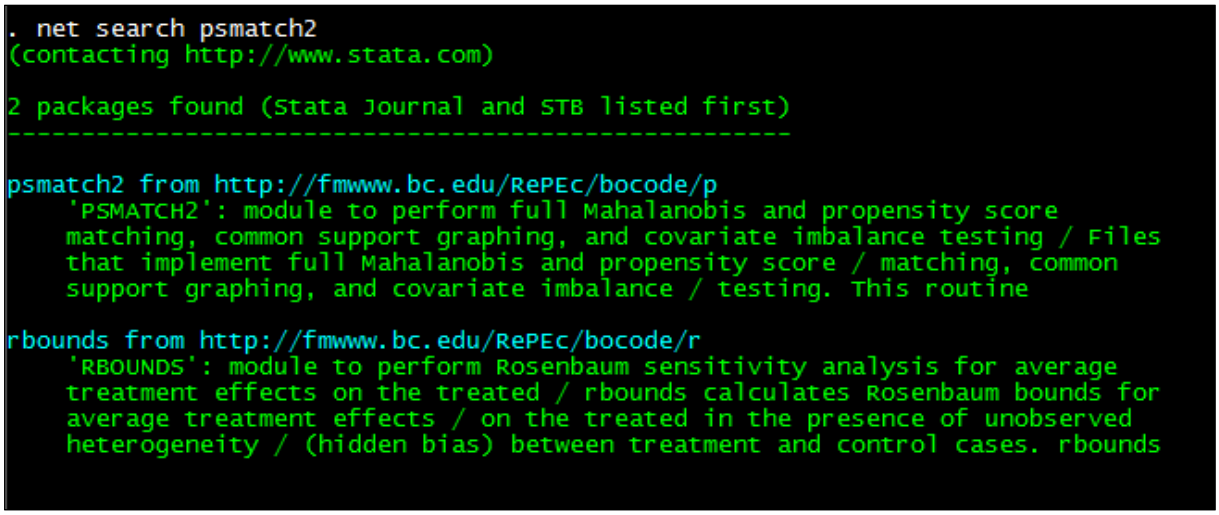
- ✓ Da bei Durchführung des PSM nicht bekannt ist, ob alle relevanten Kovariaten verwendet wurden, sollte eine Sensitivitätsanalyse durchgeführt werden. Grundsätzlich können durch Sensitivitätsanalysen potentielle Fehlerszenarien modelliert werden, wodurch die Robustheit der geschätzten Effekte gegenüber Verzerrungen durch unbeobachtete Kovariaten abgeschätzt werden kann.
- ✓ Im Fall des PSM können bspw. ‚Rosenbaum Bounds‘ berechnet werden. Anhand dieser kann abgeschätzt werden, wie stark eine unbeobachtete Kovariate die Zuteilung (Selektion) zur Treatmentgruppe verzerren müsste, sodass der durch das PSM geschätzte Effekt nicht mehr auf das Treatment, sondern vollständig auf Konfundierung zurückzuführen wäre.

4. Die Anwendung von Propensity Score Matching mit Stata

Obwohl Module zur Anwendung von Propensity Score Matching in der Basisversion von Stata nicht implementiert sind, bietet Stata die Möglichkeit, einzelne Module kostenfrei im Internet herunterzuladen. Für die Propensity Score-Analyse sind bspw. die Module *psmatch2* (Leuven & Sianesi, 2003) oder *pscore* (Becker & Ichino, 2002) frei im Netz erhältlich. Neben diesen beiden Modulen stehen eine Reihe weiterer, zum Teil ergänzender Module zur Verfügung. Ein differenzierter und sehr breiter Überblick über verschiedene Formen des Matchings sowie deren statistische Umsetzung findet sich bspw. bei Guo & Fraser (2010).

Mit dem Stata-Befehl `net search psmatch2` sucht Stata bspw. nach dem Modul *psmatch2* und stellt die entsprechenden Links zum Download automatisch bereit. Durch einfaches Anklicken des blau hinterlegten Links wird der Nutzer auf eine Seite geleitet, auf der das Modul problemlos durch Anklicken des Befehls (`click here to install`) installiert werden kann.

Abbildung 2: Suche eines Stata-Moduls im Internet



Sobald ein Programmmodul zur Analyse mit PSM installiert wurde, kann prinzipiell mit der Analyse begonnen werden. Um die Anwendung von PSM im Folgenden anschaulich darstellen zu können, werden im nächsten Teilabschnitt zunächst die zu evaluierende Intervention, die Outcome-Variablen, das Untersuchungsdesign sowie einige Vorarbeiten beschrieben, die vor allem mit der Aufbereitung der Daten in die für das PSM erforderliche Form zu tun haben.

4.1 Die Intervention

Als Beispiel wird eine Intervention verwendet, welche die Aufklärung von Verbrauchern zum Themenkomplex ‚Ernährungsverhalten und Klimaschutz' zum Ziel hatte. Bei der Intervention handelt es sich um ein Weiterbildungskonzept, mit welchem primär das Ziel verfolgt wurde,

den Teilnehmern die Zusammenhänge zwischen Ernährungsverhalten und den daraus resultierenden Konsequenzen für das Klima zu verdeutlichen. Intendierter Nebeneffekt der Intervention war es, durch die Teilnahme auf Seiten der Teilnehmer potentiell veränderte Muster bzgl. des Ernährungsverhaltens hervorzurufen. Das Konzept der Intervention versuchte die Ziele anhand von zwei Kernelementen zu erreichen: Mit Hilfe eines Vortrags sollte einerseits auf einer rein informativen Ebene über die Zusammenhänge zwischen Ernährung und Klimaschutz aufgeklärt werden. Anhand realer Essenssituationen sollten zudem die konkreten Auswirkungen des persönlichen Ernährungsverhaltens erfahrbar gemacht werden. Die Teilnehmer konnten hierbei zwischen klimafreundlichen, ökologisch hergestellten und nicht biologischen Lebensmitteln wählen. Beim gemeinsamen Essen wurde von den Mitarbeitern dann erläutert, welche Klimarelevanz die verzehrten Lebensmittel besitzen.

Im Rahmen der Wirkungsevaluation der Maßnahme soll nun geklärt werden, ob die Interventionsteilnahme Wirkungen erzielt oder nicht.[9] Wirkungen können sich beispielsweise in der Veränderung von Wissen, von Werthaltungen und Einstellungen, Verhaltensbereitschaften oder tatsächlichem Verhalten niederschlagen. Als Outcome-Variablen werden in der Untersuchung themenspezifische Verhaltensintentionen verwendet (siehe Abschnitt 4.3), welche ein dem tatsächlichen Verhalten vorgelagertes Erklärungskonstrukt darstellen und von denen angenommen wird, dass sie einen Einfluss auf das tatsächliche Verhalten der Teilnehmer ausüben.

4.2 Untersuchungsdesign und das Selektionsproblem

Für die Durchführung der Wirkungsüberprüfung wurde ein quasiexperimentelles Vergleichsgruppendesign gewählt. Hierbei wurden 103 Teilnehmer der Intervention unmittelbar nach Beendigung der Veranstaltung mittels standardisierter Fragebögen befragt. Die Fragebögen bestanden, abgesehen von soziodemografischen Variablen, aus Fünfpunkt-Ratingskalen mit den Polen „trifft überhaupt nicht zu“ bis „trifft voll und ganz zu“. Die Pole und die Zwischenschritte wurden entsprechend mit den Zahlenwerten von eins bis fünf kodiert, wobei der Wert eins für vollkommene Ablehnung steht und der Wert fünf für absolute Zustimmung. Die Daten der Vergleichsgruppe stammen aus einer deutschlandweiten Repräsentativbefragung von 1.002 Personen. Die Nichtteilnehmer wurden dabei anhand der gleichen Items befragt wie die Teilnehmer. Da die befragten Teilnehmer der Intervention alle aus Großstädten stammten (> 100.000 Einwohner), wurden für die Vergleichsgruppe ausschließlich diejenigen Personen aus der Repräsentativbefragung verwendet, welche zum Zeitpunkt der Befragung ebenfalls in einer Großstadt wohnhaft waren (Prinzip der Homogenisierung). Nach Bereinigung der Daten verblieben insgesamt 145 Personen in der Vergleichsgruppe.

Aufgrund der Abwesenheit eines randomisierten Zuteilungsverfahrens kann nicht generell ausgeschlossen werden, dass identifizierte Differenzen in den Outcome-Variablen zwischen den Gruppen auf Selektionsprozesse zurückzuführen sind oder die Abstinenz signifikanter Gruppenunterschiede nicht ebenfalls durch Selektionsprozesse hervorgerufen wurde. Da die

[9] Die Ergebnisse der quasiexperimentellen Wirkungsevaluation finden sich auch in Gaus & Mueller (im Druck).

Veranstaltungen meist am Vor- oder Nachmittag stattfanden, wäre es bspw. denkbar, dass die Interventionsteilnehmer mehrheitlich ohne Beschäftigung oder in Teilzeitbeschäftigung sind. Sofern die Teilnehmer aufgrund ihrer geringeren Arbeitsbelastung mehr Zeit dafür aufwenden, sich über klimabezogene Aspekte des Ernährungsverhaltens zu informieren und deshalb ein höheres Umweltbewusstsein aufweisen, so könnte sich ein Gruppenunterschied im Berufsstatus verzerrend auf die themenbezogenen Verhaltensabsichten auswirken.

Um derartige Konfundierungsprozesse ausschließen oder zumindest reduzieren zu können, sollten bei der Analyse daher Drittvariablen berücksichtigt werden. Für die vorliegende Untersuchung wurden ausschließlich verschiedene soziodemografische Drittvariablen berücksichtigt, auch wenn bekannt ist, dass die Kontrolle derartiger Kovariaten den Selektionsfehler nur bedingt reduzieren kann (bspw. Steiner et al., 2010; Pohl et al., 2009). Für die Darstellung der Anwendung des Verfahrens mit Stata ist dieser Umstand allerdings von untergeordneter Bedeutung, er sollte bei der Interpretation der Ergebnisse aber berücksichtigt werden.

4.3 Deskriptive Analyse und Vorarbeiten

Bevor im Nachfolgenden die zu kontrollierenden Kovariaten vorgestellt und für die Analyse mit PSM vorbereitet werden, wird zunächst einen einfacher T-Test zur Überprüfung von Mittelwertdifferenzen in den Outcomevariablen zwischen der Versuchs- und der Vergleichsgruppe durchgeführt. Hierzu muss der Stata-Datensatz eine binäre Treatment-Variable enthalten, die erfasst, ob eine Person Mitglied der Treatmentgruppe ist oder nicht. Der Mitgliedschaft in der Treatmentgruppe wird der Wert eins zugewiesen, der Mitgliedschaft in der Vergleichsgruppe der Wert null. Sobald diese Variable spezifiziert ist, kann der T-Test durchgeführt werden.

Im vorliegenden Datensatz trägt die binäre Treatmentvariable die Bezeichnung `treatment`, die Outcome-Variablen sind mit `OV1`, `OV2`, `OV3`, `OV4` und `OV5` gekennzeichnet, wobei die Outcome-Variablen folgende Aussagen repräsentieren:

- `OV1` = „Zukünftig werde ich beim Kauf von Nahrungsmitteln den Klimaaspekt stärker berücksichtigen.“
- `OV2` = „Zukünftig werde ich beim Kauf von Lebensmitteln darauf achten, Lebensmittel aus der Region zu kaufen.“
- `OV3` = „In Zukunft werde ich beim Kauf von Obst und Gemüse darauf achten, saisonale Produkte zu kaufen.“
- `OV4` = „Ich werde in Zukunft weniger Fleisch konsumieren.“
- `OV5` = „Ich werde mich zum Thema klimafreundliche Ernährung informieren.“

Für die Anwendung des PSM mit Stata wird der Übersichtlichkeit halber in allen Analysen nur die Outcomevariable OV4 betrachtet. Es wird also untersucht, ob die Teilnahme an der Intervention dazu führt, dass die Teilnehmer ihren Fleischkonsum reduzieren.

Durch die Nutzung des einfachen Stata-Befehls `ttest` können nun die Mittelwerte der beiden Gruppen miteinander verglichen werden. Um bspw. die Mittelwerte der Outcomevariable OV4 anhand einfacher T-Tests für unabhängige Stichproben miteinander vergleichen zu können, wird folgender Befehl in die Stata-Befehlsmaske eingegeben:

```
ttest OV4, by(treatment)
```

Zwei wichtige Ergebnisse zeigen sich bei näherer Betrachtung der Ergebnisse der Analyse (Abbildung 3). Zum einen existiert ein deutlicher Unterschied zwischen den Gruppenmittelwerten. So liegt der Mittelwert in der Treatmentgruppe mit 3.469 deutlich höher als in der Vergleichsgruppe (MW = 3.0). Da dieser Unterschied beim einseitigen T-Test auch noch hoch signifikant ist (p = .008 und damit p < .01), kann davon ausgegangen werden, dass die Teilnehmer der Intervention mit hoher Wahrscheinlichkeit ihren Fleischkonsum eher zu reduzieren beabsichtigen als die Personen aus der Vergleichsgruppe.

Abbildung 3: Einfacher T-Test ohne Kovariatenkontrolle

```
. ttest OV4, by(treatment)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     144           3   .1202981    1.443578    2.762208    3.237792
       1 |      98    3.469388   .1563604    1.547889    3.159056     3.77972
---------+--------------------------------------------------------------------
combined |     242    3.190083   .0965131    1.501391    2.999966      3.3802
---------+--------------------------------------------------------------------
    diff |           -.4693878   .1946762                -.85288   -.0858956
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.4111
Ho: diff = 0                     degrees of freedom =      240

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0083         Pr(|T| > |t|) = 0.0167          Pr(T > t) = 0.9917

.
```

Ob dieser Gruppenunterschied allerdings kausal auf die Interventionsteilnahme zurückzuführen ist oder ob möglicherweise Selektionsprozesse für die Differenz verantwortlich sind, lässt sich anhand der vorgenommenen Analyse nicht überprüfen.

Um den potentiellen Selektionsfehler reduzieren zu können, werden im Rahmen des PSM daher Drittvariablen in der Analyse berücksichtigt. Diese Kovariaten müssen hierzu aber zunächst in die ‚richtige Form' gebracht werden, da lediglich Kovariaten mit metrischem Skalenniveau ohne vorherige Umkodierung in die Analyse aufgenommen werden können. Kategorial und ordinal skalierte Variablen müssen jedoch in binäre Variablen (sog. „Dummy-Variablen") überführt werden.[10] Das bedeutet, dass jede Kategorie einer nominal oder ordinal

[10] Zur Generierung neuer und Umkodierung bestehender Variablen siehe Kohler & Kreuter (2008).

skalierten Variable als eigene binäre Variable in die Analyse mit eingeht. Als Beispiel kann an dieser Stelle der höchste Bildungsabschluss der Untersuchungsteilnehmer herangezogen werden. Dieser weist ordinales Skalenniveau auf und die Variable umfasst die Ausprägungen „Hauptschulabschluss“, „Realschulabschluss“, „Abitur“ und „Hochschulabschluss“. Nach dem Generieren und Umkodieren der Dummy-Variablen erhält jede Person mit Hauptschulabschluss bei der Variable „Hauptschulabschluss“ den Wert eins, alle anderen erhalten den Wert null. Analog erhalten alle Personen mit Realschulabschluss bei der Variable „Realschulabschluss“ den Wert eins, während alle anderen Personen, die einen niedrigeren oder höheren Bildungsabschluss aufweisen, den Wert null aufweisen müssen. Für die verbleibenden Kategorien „Abitur“ und „Hochschulabschluss“ kommt dieselbe Vorgehensweise zum Einsatz.

Folgende Kovariaten werden für die weitergehende Analyse mit PSM verwendet, wobei die einzelnen Kategorien der nominal und ordinal skalierten Variablen „Bildungsabschluss“ und „Berufsstatus“ als binär kodierte Variablen in die Analyse mit eingehen.

Tabelle 1: Verwendete Kovariaten

Kovariate	Kategorie/Maßeinheit
Alter	Lebensjahr
Geschlecht	männlich / weiblich
Höchster Bildungsabschluss	Hauptschulabschluss / Realschulabschluss / Abitur / Hochschulabschluss
Berufsstatus	keine Arbeit / Rente / Ausbildung / Vollzeit / Teilzeit

4.4 Schätzung des Propensity Score und der Treatmenteffekte

Im nächsten Schritt der Analyse können nun der Propensity Score und die Treatmenteffekte geschätzt werden. Hierzu wird das Stata-Modul *psmatch2* genutzt, welches die Schätzung des Propensity Score und der Treatmenteffekte in einem einzigen Analyseschritt erlaubt. Bevor allerdings eine Schätzung möglich ist, muss ein Matching-Algorithmus ausgewählt werden.

Zunächst kommt die Variante des Nearest-Neighbour-Matching mit Zurücklegen (siehe Abschnitt 3.2) zur Anwendung. Die Option des Zurücklegens ist bereits standardgemäß festgesetzt und muss daher nicht manuell eingestellt werden. Anstelle einer einzelnen Person der Vergleichsgruppe, welche zur Bildung des individuellen kontrafaktischen Zustands herangezogen wird, sollen allerdings diejenigen beiden Personen herangezogen werden, deren Pro-

pensity Score einer Person der Treatmentgruppe am nächsten ist. Es wird also der Nearest-Neighbour-Algorithmus mit zwei nächsten Nachbarn verwendet. Für die Bildung des individuellen Treatmenteffekts wird durch das Programm automatisch das arithmetische Mittel der beiden nächsten Nachbarn in der Outcome-Variable OV4 gebildet und vom entsprechenden Wert der Person der Treatmentgruppe abgezogen.

In die Stata-Befehlszeile wird vor dem Komma der Befehl zur Schätzung des Propensity Score eingegeben, wobei direkt nach dem Befehl `psmatch2` die abhängige Variable des diskreten Entscheidungsmodells folgt, anhand dessen der Propensity Score geschätzt wird. Da die auf Basis der ausgewählten Kovariaten bedingte Wahrscheinlichkeit für jede Person geschätzt werden soll, zur Treatmentgruppe zu gehören, ist die abhängige Variable des Regressionsmodells die Variable „treatment“. Unmittelbar nach der abhängigen Variablen des Probit-Modells müssen dann die einzelnen Kovariaten aufgelistet werden. Als Standardverfahren zur Schätzung des Propensity Score ist die Probit-Regression festgelegt, durch das Hinzufügen der Option `logit` nach dem Komma wird anstelle der Probit-Regression eine logistische Regression zur Schätzung des Propensity Score verwendet.

Durch die Setzung eines Kommas in der Befehlszeile können nun die einzelnen Optionen näher spezifiziert werden. So wird die Variable OV4 bspw. durch den Befehl `out(OV4)` als Outcome-Variable festgelegt. Durch die Option `common` wird sichergestellt, dass nur diejenigen Fälle in die Analyse einbezogen werden, die auch die Bedingung des Common Support erfüllen (siehe Abschnitt 3.2). Schließlich wird der Matching-Algorithmus spezifiziert. Der Nearest-Neighbour-Algorithmus folgt dem Befehl `neighbor()`, wobei in die Klammer die Anzahl der nächsten Nachbarn einzugeben ist, die verwendet werden soll.

Der für die Durchführung des NN-Matchings mit zwei nächsten Nachbarn erforderliche Stata-Befehl wird folgendermaßen in die Befehlszeile eingetragen.[11]

```
psmatch2 treatment Geschlecht Alter Vollzeit Teilzeit Ausbildung
Rente keine_Arbeit Hauptschule Realschule Abitur Hochschule,
out(OV4) common neighbor(2)
```

Nachfolgende Abbildung 4 stellt die Ergebnisse der Analyse dar. Die obere Tabelle zeigt zunächst die Ergebnisse der Probit-Regression zur Schätzung des Propensity Score. Wie zu erkennen ist, hat sich die Fallzahl im Vergleich zur Analyse aus Abbildung 3 von 242 auf 217 Personen reduziert. Dies liegt darin begründet, dass aufgrund fehlender Werte in verschiedenen Kovariaten und des damit verbundenen listenweisen Fallausschlusses einige Personen in der Analyse nicht berücksichtigt werden konnten. Weiter zeigt sich, dass die beiden Gruppen relativ unähnlich hinsichtlich der verwendeten Kovariaten sind. Die mittlere Heterogenität lässt sich bspw. am Pseudo-R^2 von .246 festmachen. Zudem ist aus der Tabelle abzulesen, dass die Kovariaten Geschlecht, Alter und Hochschulabschluss einen signifikanten Beitrag zur Vorhersage der Gruppenzugehörigkeit von Personen aufweisen. Die beiden

[11] Die in der Befehlszeile angegebenen Variablenbezeichnungen entsprechen denjenigen aus dem Stata-Datensatz. Sie können beliebig verändert werden.

Kovariaten Rente und Hauptschulabschluss wurden von Stata aufgrund von Kollinearität automatisch aus der Schätzfunktion entfernt.

Abbildung 4: Nearest-Neighbour-Matching mit zwei nächsten Nachbarn

```
. psmatch2 treatment Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente keine_Arbeit Hauptschule Realschule
>  Abitur Hochschule, out(OV4) common neighbor(2)
note: Rente dropped because of collinearity
note: Hauptschule dropped because of collinearity

Probit regression                                 Number of obs   =        217
                                                  LR chi2(9)      =      68.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -104.54413                       Pseudo R2       =     0.2456

------------------------------------------------------------------------------
   treatment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Geschlecht |  -.6545116   .2266223    -2.89   0.004    -1.098683     -.21034
       Alter |  -.0401567   .0090497    -4.44   0.000    -.0578937    -.0224197
    Vollzeit |   .1950931   .3747103     0.52   0.603    -.5393256     .9295118
    Teilzeit |   .1919472    .410075     0.47   0.640     -.611785     .9956795
  Ausbildung |   .2813998   .5587986     0.50   0.615    -.8138253     1.376625
keine_Arbeit |   .6321679   .4463163     1.42   0.157    -.2425958     1.506932
  Realschule |  -.4122026   .3092369    -1.33   0.183    -1.018296     .1938906
       Abitur |  -.1889003    .33028    -0.57   0.567    -.8362372     .4584367
  Hochschule |  -.5205006   .2992334    -1.74   0.082    -1.106987     .0659861
       _cons |   1.813427   .7158779     2.53   0.011       .410332     3.216522
------------------------------------------------------------------------------
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
(31 missing values generated)

------------------------------------------------------------------------------
    Variable     Sample |    Treated     Controls   Difference          S.E.   T-stat
-------------------------+----------------------------------------------------------
         OV4  Unmatched |  3.26027397            3   .260273973   .211082677     1.23
                  ATT |  3.45901639   2.81147541   .647540984   .294795261     2.20
------------------------------------------------------------------------------
Note: S.E. for ATT does not take into account that the propensity score is estimated.

 psmatch2:  |  psmatch2: Common
  Treatment |        support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |         0         144 |       144
   Treated |        12          61 |        73
-----------+----------------------+----------
     Total |        12         205 |       217

.
```

Die mittlere Tabelle aus Abbildung 4 zeigt nun die Ergebnisse des eigentlichen Matchings. Wie in der Zeile „Unmatched“ zu erkennen ist, ist der Mittelwert in der Outcomevariable OV4 vor dem Matching etwas niedriger als noch beim einfachen T-Test aus Abbildung 3. Dies liegt daran, dass einige Fälle aufgrund fehlender Werte oder des Umstands, dass sie nicht im Bereich des Common Support liegen, aus der Analyse ausgeschlossen wurden. Die Mittelwertdifferenz im ungematchten Fall beträgt 0.260 und ist nicht signifikant. Im gematchten Fall, also in der Zeile „ATT“, erhöht sich die Mittelwertdifferenz deutlich auf 0.648 und verdreifacht sich damit nahezu. Am T-Wert von 2.20 ist zudem zu erkennen, dass der Treatmenteffekt signifikant auf dem 5%-Niveau ist (einseitiger Test).

Die Ergebnisse können nun folgendermaßen interpretiert werden: Während sich die Teilnehmer und Nichtteilnehmer der Intervention hinsichtlich ihrer Absicht, zukünftig weniger Fleisch zu konsumieren, zunächst nicht signifikant voneinander unterscheiden, erhöht sich der Treatmenteffekt unter Berücksichtigung der verwendeten Kovariaten deutlich. Das bedeutet, dass unter der Kontrolle der Kovariaten Geschlecht, Alter, Berufsstatus und Bil-

dungsabschluss die Interventionsteilnehmer eine signifikant stärkere Verhaltensabsicht aufweisen als die Nichtteilnehmer, ihren Fleischkonsum zukünftig zu reduzieren.

Die dritte Tabelle in Abbildung 4 zeigt schließlich an, wie viele Fälle aufgrund der Nichterfüllung der Common Support Bedingung aus der Analyse ausgeschlossen und wie viele Fälle tatsächlich zur Analyse herangezogen wurden.

Zur Veranschaulichung eines weiteren Matching-Algorithmus wird im Folgenden dieselbe Analyse anhand eines Kernel-Matchings durchgeführt (vgl. Abschnitt 3.2). Hierzu muss der oben verwendete Stata-Befehl lediglich in einem Punkt verändert werden: Statt der Option `neighbor(2)` wird die Option `kernel k(normal)` verwendet. Die Option `kernel` legt fest, dass ein Kernel-Matching durchgeführt werden soll, der Zusatz `k(normal)` bestimmt, dass als Dichtefunktion eine Normalverteilung verwendet wird. Bzgl. der Bandbreite wird die Standardeinstellung von 0.06 übernommen, welche in Stata bereits automatisch festgelegt ist. Der Befehl sieht damit folgendermaßen aus:

```
psmatch2 treatment Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente keine_Arbeit Hauptschule Realschule Abitur Hochschule, out(OV4) common kernel k(normal)
```

Abbildung 5: Kernel-Matching

```
. psmatch2 treatment Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente keine_Arbeit Hauptschule Realschule
>  Abitur Hochschule, out(OV4) common kernel k(normal)
note: Rente dropped because of collinearity
note: Hauptschule dropped because of collinearity

Probit regression                                 Number of obs     =        217
                                                  LR chi2(9)        =      68.07
                                                  Prob > chi2       =     0.0000
Log likelihood = -104.54413                       Pseudo R2         =     0.2456

------------------------------------------------------------------------------
   treatment |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Geschlecht |  -.6545116   .2266223    -2.89   0.004    -1.098683    -.21034
       Alter |  -.0401567   .0090497    -4.44   0.000    -.0578937   -.0224197
    Vollzeit |   .1950931   .3747103     0.52   0.603    -.5393256    .9295118
    Teilzeit |   .1919472    .410075     0.47   0.640     -.611785    .9956795
  Ausbildung |   .2813998   .5587986     0.50   0.615    -.8138253    1.376625
keine_Arbeit |   .6321679   .4463163     1.42   0.157    -.2425958    1.506932
  Realschule |  -.4122026   .3092369    -1.33   0.183    -1.018296    .1938906
      Abitur |  -.1889003     .33028    -0.57   0.567    -.8362372    .4584367
  Hochschule |  -.5205006   .2992334    -1.74   0.082    -1.106987    .0659861
       _cons |   1.813427   .7158779     2.53   0.011      .410332    3.216522
------------------------------------------------------------------------------
(31 missing values generated)

------------------------------------------------------------------------------
    Variable     Sample |    Treated     Controls   Difference          S.E.   T-stat
-------------------------+----------------------------------------------------------
         OV4  Unmatched |  3.26027397            3   .260273973   .211082677     1.23
                  ATT |  3.45901639   2.85343138   .605585017   .252701862     2.40
------------------------------------------------------------------------------
Note: S.E. for ATT does not take into account that the propensity score is estimated.

 psmatch2:  |  psmatch2: Common
  Treatment |        support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |         0        144 |       144
   Treated |        12         61 |        73
-----------+----------------------+----------
     Total |        12        205 |       217

.
```

Die Ergebnisse in Abbildung 5 zeigen zunächst, dass die Schätzung des Propensity Score die exakt selben Ergebnisse liefert. Unterschiede ergeben sich jedoch mit Blick auf den Treatmenteffekt (ATT). Während sich der Mittelwertunterschied im ungematchten Fall nicht von dem aus Abbildung 4 unterscheidet, erhöht sich der ATT beim Kernel-Matching zwar, jedoch nicht ganz so stark wie beim NN-Matching. Dadurch, dass zur Bildung des individuellen kontrafaktischen Zustands allerdings alle Fälle der Vergleichsgruppe herangezogen werden und somit mehr Informationen genutzt werden, verkleinert sich der Standardfehler und der T-Wert ist höher als beim NN-Matching. Unter Anwendung eines einseitigen T-Tests ist der ATT von 0.606 damit auf dem 1%-Niveau signifikant. Summa summarum gilt aber auch hier: Die Interventionsteilnehmer weisen unter Ausbalancierung der verwendeten Kovariaten deutlich stärkere Verhaltensabsichten bzgl. ihres zukünftigen Fleischkonsums auf als die Nichtteilnehmer.

4.5 Beurteilung der Matching-Qualität

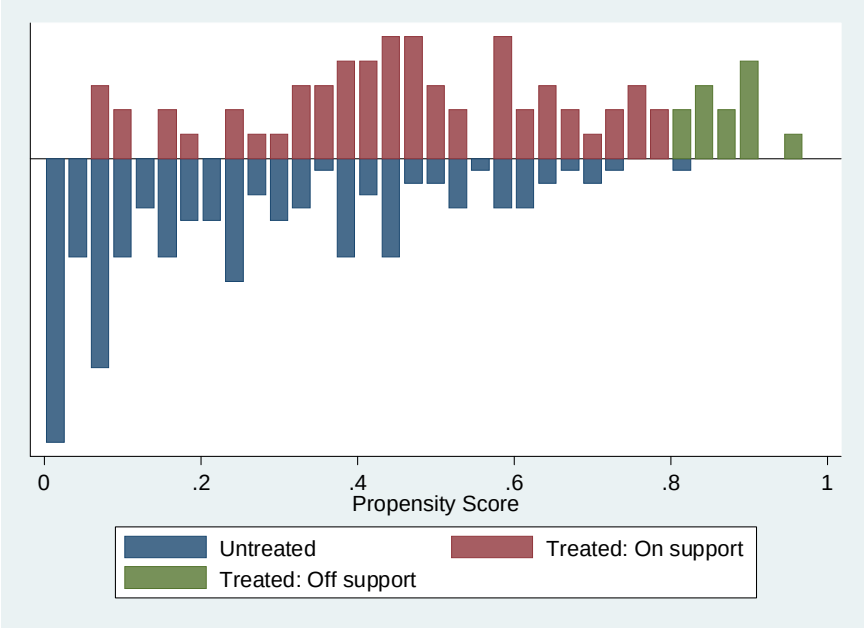
Bevor im Folgenden die Matching-Qualität anhand des Stata-Befehls `pstest` überprüft wird, soll kurz dargestellt werden, wie die Verteilung des Propensity Score über die beiden Gruppen hinweg grafisch veranschaulicht werden kann. Hierzu kommt der Stata-Befehl `psgraph` zur Anwendung. Mit diesem Befehl lassen sich automatisiert gruppierte Histogramme zur Darstellung der Verteilung des Propensity Score in den Gruppen erstellen. Zu beachten ist hierbei allerdings, dass der Befehl nur unmittelbar nach Durchführung des Matchings angewendet werden kann, da die zur Berechnung der Grafik erforderlichen Informationen von Stata dann noch gespeichert sind. Mit Hilfe der Option `bin()` lässt sich die Anzahl der Säulen im Histogramm festlegen. Im vorliegenden Fall wird eine Anzahl von 35 Säulen gewählt und der folgende Befehl in die Befehlszeile eingegeben:

```
psgraph, bin(35)
```

Abbildung 6 zeigt nun die grafische Verteilung des Propensity Score. Die roten Säulen repräsentieren die Häufigkeit der Fälle der Treatmentgruppe in einem bestimmten Intervall des Propensity Score, die grünen Säulen diejenigen Fälle, die aufgrund der Nichterfüllung der Common Support Bedingung von der Analyse ausgeschlossen wurden. Analog veranschaulichen die blauen Säulen die Fallhäufigkeit in einem bestimmten Intervall des Propensity Score in der Vergleichsgruppe.

Die anhand der statistischen Analyse bereits identifizierte Heterogenität zwischen den Gruppen wird in der Abbildung ebenfalls deutlich sichtbar. So konzentriert sich ein Großteil der Fälle der Vergleichsgruppe auf ein Intervall des Propensity Score von 0 bis 0.4, während der Großteil der Fälle in der Treatmentgruppe über dem Wert von 0.4 liegt. Konkret bedeutet dies, dass sich die Personen in den verwendeten Kovariaten relativ unähnlich sind.

Abbildung 6: Verteilung des Propensity Score



Die eigentliche Überprüfung der Matching-Qualität erfolgt nun anhand der in Abschnitt 3.3 beschriebenen Verfahren. So wird zunächst anhand zweiseitiger T-Tests überprüft, ob über die Gruppen hinweg ungleich verteilte Kovariaten durch den Propensity Score ausbalanciert werden können. Signifikante Gruppenunterschiede in den Kovariaten im ungematchten Zustand sollten nach dem Matching also nicht mehr auftreten. Zur Durchführung dieses Tests wird der Befehl `pstest` verwendet, der wie `psgraph` ebenfalls im Modul `psmatch2` enthalten ist. Auch `pstest` kann nur dann ohne weitere Spezifikationen angewendet werden, wenn zuvor bereits der Propensity Score geschätzt wurde. Zur Anwendung müssen hinter den Befehl `pstest` alle zu testenden Kovariaten aufgelistet werden:

```
pstest Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente kei-
ne_Arbeit Hauptschule Realschule Abitur Hochschule
```

In Abbildung 7 sind die Ergebnisse der Prüfung der Matching-Qualität dargestellt. Für jede Kovariate werden die Mittelwerte im gematchten und im ungematchten Fall ausgewiesen. Abgesehen von der Kovariate Alter repräsentieren die Mittelwerte aller anderen Kovariaten die relativen Häufigkeiten der Ausprägung eins. Mit anderen Worten: Der Anteil der Personen mit einem Hochschulabschluss in der ungematchten Treatmentgruppe beträgt bspw. 30.14 Prozent, in der Vergleichsgruppe 39.58 Prozent.

Abbildung 7: Prüfung der Matching-Qualität

```
. pstest Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente keine_Arbeit Hauptschule
> Realschule Abitur Hochschule
```

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
Geschlecht	Unmatched	.28767	.49306	-42.9		-2.94	0.004
	Matched	.34426	.3736	-6.1	85.7	-0.33	0.739
Alter	Unmatched	38.027	55	-115.8		-7.72	0.000
	Matched	40.639	41.444	-5.5	95.3	-0.32	0.748
Vollzeit	Unmatched	.32877	.36111	-6.8		-0.47	0.639
	Matched	.39344	.37271	4.3	35.9	0.23	0.816
Teilzeit	Unmatched	.28767	.16667	29.0		2.09	0.038
	Matched	.31148	.28763	5.7	80.3	0.28	0.777
Ausbildung	Unmatched	.16438	.04167	41.0		3.15	0.002
	Matched	.09836	.1351	-12.3	70.1	-0.63	0.533
Rente	Unmatched	.05479	.36111	-81.2		-5.14	0.000
	Matched	.06557	.09725	-8.4	89.7	-0.63	0.528
keine_Arbeit	Unmatched	.16438	.06944	29.7		2.20	0.029
	Matched	.13115	.10731	7.5	74.9	0.40	0.689
Hauptschule	Unmatched	.24658	.15972	21.6		1.55	0.124
	Matched	.14754	.14925	-0.4	98.0	-0.03	0.979
Realschule	Unmatched	.24658	.30556	-13.2		-0.91	0.366
	Matched	.29508	.26601	6.5	50.7	0.35	0.725
Abitur	Unmatched	.20548	.13889	17.6		1.26	0.209
	Matched	.21311	.24472	-8.4	52.5	-0.41	0.682
Hochschule	Unmatched	.30137	.39583	-19.8		-1.37	0.173
	Matched	.34426	.34002	0.9	95.5	0.05	0.961

Neben den Mittelwerten finden sich die Spalten `%bias` und `%reduct|bias|` in der Abbildung. Die erste dieser beiden Spalten beziffert die relative Differenz in einer Kovariate zwischen den Gruppen, die zweite die relative Reduzierung dieses Unterschieds durch das Matching. Bei näherer Betrachtung dieser Spalte ist zu erkennen, dass (abgesehen von einer Kovariate) der zuvor existente Unterschied in allen Kovariaten um mehr als 50 Prozent verringert werden kann.

Die beiden Spalten ganz rechts zeigen weiter die Ergebnisse der T-Tests. Der p-Wert bezieht sich hier auf den zweiseitigen Test. Es wird also überprüft, ob die Differenz zwischen den Gruppen signifikant von null verschieden ist. Während im ungematchten Zustand sechs signifikante Gruppenunterschiede zu verzeichnen sind, sind nach dem Matching keine signifikanten Gruppenunterschiede mehr erkennbar. Es kann also davon ausgegangen werden, dass das Matching auf Basis des Propensity Score zu einer hinreichend akzeptablen Matching-Qualität führt.

Ein zweiter Test, der ebenfalls unter Abschnitt 3.3 bereits angesprochen wurde, ist der Vergleich des Determinationskoeffizienten Pseudo-R^2 und des LR-χ^2-Tests vor und nach dem Matching. Hierzu wird derselbe Befehl verwendet wie zur Durchführung der zweiseitigen T-Tests, es wird jedoch nach dem Komma die Option `sum` eingefügt:

```
pstest Geschlecht Alter Vollzeit Teilzeit Ausbildung Rente kei-
ne_Arbeit Hauptschule Realschule Abitur Hochschule, sum
```

Zusätzlich zur Tabelle aus Abbildung 7 werden weitere Ergebnisse ausgegeben, unter anderem die nachfolgend dargestellte Tabelle aus Abbildung 8. Hier ist sehr deutlich zu erkennen, dass die Ausbalancierung der Kovariaten zwischen den Gruppen dazu führt, dass das Pseudo-R^2 nach dem Matching nahe null liegt und der LR-χ^2-Test nicht mehr signifikant ist, also keine der verwendeten Kovariaten mehr zur Vorhersage der Gruppenzugehörigkeit geeignet ist. Auch die hier durchgeführten Tests attestieren eine hohe Matching-Qualität. Die in Abschnitt 3.3 angesprochene ‚Balancing Property' kann damit als näherungsweise erfüllt betrachtet werden. Die Konsequenzen aus dieser (eher technischen) Feststellung sind eine höhere Güte der Schätzungen und letztendlich auch eine höhere Belastbarkeit der Ergebnisse.

Abbildung 8: Weitere Tests zur Überprüfung der Matching-Qualität

Sample	Pseudo R2	LR chi2	p>chi2
Unmatched	0.246	68.07	0.000
Matched	0.007	1.18	0.999

4.6 Sensitivitätsanalyse

Obwohl mit der Durchführung des PSM verfolgt wird, kausale Aussagen treffen zu können, ist dies hier mit einer nicht quantifizierbaren Unsicherheit verbunden. Zwar wurden einige potentielle Konfundierungsfaktoren im Rahmen des PSM zwischen den Gruppen ausbalanciert, jedoch wurden andere wie bspw. Themeninteresse, Themenrelevanz, Umweltbewusstsein oder gesundheitsbezogene Variablen etc. nicht berücksichtigt. Da diese Faktoren ebenfalls zu einer Konfundierung beitragen könnten, wurde der Selektionsfehler anhand der vorgenommenen Kovariatenauswahl vermutlich nur partiell reduziert.

Aufgrund des verbleibenden Risikos durch den potentiell verzerrenden Einfluss unbeobachteter Drittvariablen muss eine Sensitivitätsanalyse vorgenommen werden. Anhand dieser können Szenarien modelliert werden, wie stark unbeobachtete Kovariaten verzerrend wirken müssten, um die Robustheit der geschätzten Treatmenteffekte gegenüber selektionsbedingter Konfundierung zu gefährden.

Die Durchführung der Sensitivitätsanalyse erfolgt anhand der in Abschnitt 3.4 beschriebenen Rosenbaum Bounds und unter Verwendung des Stata-Moduls `rbounds`. Bevor die Sensitivitätsanalyse allerdings umgesetzt werden kann, muss zunächst eine neue Variable generiert werden, die die individuellen Treatmenteffekte beschreibt. Diese Variable, als delta_OV4 bezeichnet, wird im Anschluss an die Durchführung des Matchings folgendermaßen berechnet:[12]

```
gen delta_OV4 = OV4 - _OV4 if  _treated==1 & _support==1
```

Im nächsten Schritt kann nun mit folgendem Befehl die Sensitivitätsanalyse durchgeführt werden:

```
rbounds delta_OV4, gamma(1(0.1)2)
```

Unter der Option `gamma` lässt sich festlegen, in welchen Intervallen und bis zu welchem Wert von `gamma` die Rosenbaum Bounds berechnet werden. Im vorliegenden Fall wurden, ausgehend vom Wert 1, Intervallabstände von 0.1 bis zu einem `gamma` von 2.0 zur Berechnung der Rosenbaum Bounds gewählt.

Die Ergebnisse der Sensitivitätsanalyse sind in Abbildung 9 dargestellt. Von Interesse für die vorliegende Analyse sind lediglich die Signifikanzobergrenzen (sig+), da diese eine Positivselektion und damit eine potentielle Überschätzung der wahren Treatmenteffekte repräsentieren. Die Signifikanzobergrenzen sind demnach bei positiven Treatmenteffekten anzuwenden, da es einerseits eher unwahrscheinlich ist, dass ein bereits (starker) positiver Treatmenteffekt unterschätzt wird und andererseits überprüft werden soll, bei welcher Stärke einer selektionsbedingten Verzerrung der geschätzte Treatmenteffekt zu großen Teilen auf Selektion- bzw. Konfundierungsprozesse zurückzuführen wäre.

In der Tabelle ist zu sehen, dass im Szenario $\Gamma = 1$ der Treatmenteffekt signifikant mit $p < .05$ ist. Da die Analyse anhand von Rosenbaum Bounds im Gegensatz zum PSM medianbasiert ist und nicht das arithmetische Mittel verwendet, kann das Signifikanzniveau des Treatmenteffekts unter $\Gamma = 1$ leicht von demjenigen beim PSM abweichen. Das Ergebnis unter $\Gamma = 1$ bedeutet letztlich nichts anderes als dass im Falle der Abstinenz aller Selektions- und Konfundierungsprozesse der geschätzte Treatmenteffekt signifikant ist.

Die Resultate in Abbildung 9 zeigen weiter, dass auch im Szenario $\Gamma = 1.7$ der Treatmenteffekt noch signifikant auf dem 10%-Niveau ist. Selbst wenn ein bestimmter Konfundierungsfaktor also das Chancenverhältnis einer Person, der Treatmentgruppe zuzugehören, zu einem Verhältnis von 1.7:1 verzerren würde, wäre der geschätzte Effekt trotzdem nicht vollständig auf Konfundierung zurückzuführen. Die Nullhypothese, die unterstellt, dass der geschätzte Effekt ausschließlich auf Konfundierung zurückzuführen ist, kann bis zu dieser maximalen Höhe von $\Gamma = 1.7$ mit $p < .1$ abgelehnt werden. Mit ansteigendem Γ verringert sich

[12] Die im Befehl auftretenden Variablen _OV4, _treated und _support werden von Stata automatisch nach Beendigung des Matchings generiert und gespeichert.

demnach die Wahrscheinlichkeit des Eintretens eines positiven Treatmenteffekts. Bei zunehmendem Γ nehmen also der Treatmenteffekt und die Wahrscheinlichkeit, dass die Nullhypothese falsch ist, ab und sig+ steigt an.

Die Robustheit der Effektschätzungen hängt damit von der Höhe von Γ ab. Es gilt: Je höher Γ wird und der geschätzte Effekt dennoch signifikant ist, desto robuster sind die Effektschätzungen.

Abbildung 9: Sensitivitätsanalyse mit Rosenbaum Bounds

```
. rbounds delta_ov4, gamma(1(0.1)2)

Rosenbaum bounds for delta_ov4 (N = 61 matched pairs)

Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
-------------------------------------------------------------------------
    1         .001147   .001147   .657347   .657347   .308626   .969691
  1.1         .003178   .000367   .565896   .750711   .229533   1.09291
  1.2         .007306   .000116    .50026   .812435   .123477   1.16066
  1.3         .014557   .000036   .455257   .836225   .028547    1.2452
  1.4         .025922   .000011   .426684   .858324  -.001662   1.30534
  1.5         .042199   3.4e-06   .391319   .885542   -.06181   1.33491
  1.6          .06388   1.0e-06   .363078   .916663    -.1018   1.36003
  1.7           .0911   3.2e-07   .333178   .940712  -.132525   1.39986
  1.8         .123645   9.6e-08   .310717   .968883  -.162947   1.41915
  1.9         .161001   2.9e-08   .290963   1.00534  -.174535   1.43956
    2         .202432   8.7e-09   .237176   1.07012  -.189902   1.49308

* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)

.
```

Neben den Signifikanzobergrenzen und -untergrenzen finden sich in Abbildung 9 zudem die Hodges-Lehman Punktschätzer unter den gegebenen Γ-Szenarien. Unter der Annahme additiver Treatmenteffekte liefert der Hodges-Lehman Punktschätzer Informationen darüber, wie hoch der Treatmenteffekt noch wäre, wenn ein Konfundierungsfaktor das Chancenverhältnis, in der Treatmentgruppe zu sein, auf dem Niveau eines bestimmten Γ verzerrt. Im Fall von Γ = 1.7 heißt das, dass bei einer selektionsbedingten Verzerrung des Chancenverhältnisses von 1.7:1 der Treatmenteffekt immer noch geschätzte 0.333 Skalenpunkte betragen würde.

4.7 Zusammenfassung und Diskussion

Ziel des vorliegenden Leitfadens war es, dem Leser einen grundlegenden Einblick in die Methodik des Propensity Score Matching und dessen Anwendung im Programmpaket Stata zu

gewähren. Trotz des weitgehenden Verzichts auf die Verwendung von Formeln stellt der Leitfaden dennoch vergleichsweise hohe Anforderungen an den Leser. Dies hängt vor allem mit der Komplexität des Verfahrens zusammen, dessen Funktionsweise und Anwendung ohne fundierte Vorkenntnisse der Methoden der empirischen Sozialforschung und der Statistik nur schwerlich nachvollziehbar sein dürften. Zwar findet das PSM in der Evaluationsforschung eine immer weitere Verbreitung, in der Praxis kommt es aufgrund der Komplexität seiner Anwendung jedoch nach wie vor eher selten zum Einsatz.

Zur Veranschaulichung des Analyseprozesses wurde eine Beispielintervention gewählt, bei der die Teilnehmer durch das Hören eines Vortrags und den Konsum verschiedener ökologisch und nicht-ökologisch produzierter Lebensmittel die Konsequenzen des Ernährungsverhaltens auf den Klimawandel nachvollziehen sollten. Ziel der Veranstalter war es, zu einer Veränderung des Ernährungsverhaltens im Sinne des Klimaschutzes beizutragen. Anhand der quasiexperimentellen Wirkungsevaluation mit PSM wurde daher mithilfe eines exemplarischen Outcomes überprüft, ob die Interventionsteilnahme zu Wirkungen auf Seiten der Teilnehmer führt oder nicht. Es zeigte sich, dass die Interventionsteilnehmer unter Kontrolle beobachteter Kovariaten eine höhere Verhaltensintention aufwiesen, zukünftig auf den Konsum von Fleisch zu verzichten, als die Personen aus der Vergleichsgruppe. Im Rahmen der Umsetzung des PSM wurde dargestellt, welche Vorarbeiten zu treffen sind, wie der PS und die Treatmenteffekte geschätzt werden und wie eine Sensitivitätsanalyse zur Beurteilung der Robustheit der Schätzungen durchgeführt wird. Vor allem aufgrund des Umstands, dass zum Matching lediglich einige soziodemografische Merkmale herangezogen wurden, ist die Sensitivitätsanalyse im verwendeten Beispiel von besonderer Bedeutung.

Obwohl das PSM unter bestimmten Bedingungen einen fruchtbaren Ansatz zur quasiexperimentellen Wirkungsevaluation darstellt, sei abschließend noch einmal darauf hingewiesen, dass es sich dabei eher um eine Hilfsstrategie der Wirkungsschätzung handelt. Denn: Die Umsetzung methodisch hochwertiger Designs wie bspw. experimenteller Untersuchungsanordnungen macht die ökonometrische Schätzung von Treatmenteffekten überflüssig, da die angesprochene Selektionsproblematik hier bereits durch Randomisierung gelöst wird. Sofern eine Drittvariablenkontrolle anhand ökonometrischer Verfahren aufgrund situationsbedingter Gegebenheiten allerdings erforderlich ist, sollte der Fokus weniger auf die Wahl des statistischen Verfahrens, sondern eher auf die theoretisch begründete und sorgfältige Auswahl der Kovariaten gerichtet werden.

5. Literatur

Aakvik, A. (2001). Bounding a matching estimator: The case of a Norwegian training program. *Oxford Bulletin of Economics and Statistics*, *63* (1), 115-143.

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics*, *86* (1), 180-194.

Becker, S.O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, *2* (4), 358-377.

Black, D., & Smith, J. (2004). How robust is the evidence on the effects of the college quality? Evidence from matching. *Journal of Econometrics*, *121* (1), 99–124.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score matching. *American Journal of Epidemiology*, *163* (12), 1149-1156.

Caliendo, M., & Hujer, R. (2006). The microeconometric estimation of treatment effects - An overview. *Allgemeines Statistisches Archiv*, *90* (1), 199-215.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22* (1), 31-72.

Cepeda, M.S., Boston, R., Farrar, J.T., & Strom, B.L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, *158* (3), 280-287.

Cook, T.D., & Steiner, P.M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15* (1), 56-68.

Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27* (4), 724-750.

Dehejia, R., H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *84* (1), 151–161.

Diaz, J.J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. *Journal of Human Resources*, *41* (2), 319–345.

DiPrete, T.A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34* (1), 271-310.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching. Theory and practice*. Chichester [u.a.]: Wiley.

Gangl, M., 2004: RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated, Software, http://ideas.repec.org/c/boc/bocode/s438301.html (Stand 18.01.2012).

Gangl, M., & DiPrete, T. (2004). Kausalanalyse durch Matchingverfahren. S. 396-420 in: Diekman, A. (Hrsg.): *Methoden der Sozialforschung, Sonderheft 44/2004 der Kölner Zeitschrift für Soziologie und Sozialpsychologie*.

Gaus, H., Mueller, C.E. (im Druck): Evaluating free-choice climate education interventions applying propensity score matching. *Evaluation Review* (Manuskript angenommen zur Veröffentlichung im Januar 2012).

Greene, W.H. (2009). Discrete choice modelling. S. 473-556 in: *Palgrave handbook of econometrics Vol. 2: Applied econometrics*. Basingstoke: Palgrave Macmillan.

Guo, S., & Fraser, M.W. (2010). *Propensity score analysis: Statistical methods and applications.* Thousand Oaks: Sage.

Heckman, J.J. (1997). Instrumental variables - a study of the implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, *32* (3), 441–462.

Heckman, J.J., & Smith, J.A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9* (2), 85-110.

Heckman, J.J., Ichimura, H., Smith, J., & Todd, P. (1998a). Characterizing selection bias using experimental data. *Econometrica*, *66* (5), 1017-1098.

Heckman, J.J., Ichimura, H., & Smith, J. (1998b). Matching as an econometric evaluation estimator. *Review of Economic Studies*, *65* (2), 261–294.

Imbens, G.W., & Lemieux, T. (2007). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142* (2), 615-635.

Jaenichen, U. (2002). Mikroevaluationen: Bildung von Vergleichsgruppen zur Schätzung individueller Förderwirkungen. S. 387-397 in: Kleinhenz, G. (Hrsg.): *IAB-Kompendium Arbeitsmarkt- und Berufsforschung*. Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 250.

Kim, J. (1995). Causation. S. 125-127 in: Audi, R. (Hrsg.), 1999: *The Cambridge dictionary of philosophy*. Cambridge: Cambridge University Press (2.A.).

Kohler, U., & Kreuter, F. (2008). *Datenanalyse mit Stata.* München: Oldenbourg (2.A.).

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Software, http://ideas.repec.org/c/boc/bocode/s432001.html (Stand 18.01.2012).

Linden, A., & Adams, J.L. (2010). Using propensity sore-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice*, *16* (1), 175-179.

LaLonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76* (4), 604-620.

Lechner, M. (2002). Some practical issues in the evaluation of heterogenous labour market programmes by matching methods. *Journal of the Royal Statistical Society*, *165* (1), 59-82.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. S. 1-18 in: Lechner, M., Pfeiffer, F. (Hrsg.): *Econometric Evaluation of Labour Market Policies*. Heidelberg: Physica.

Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business Economic Statistics*, *17* (1), 74-90.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage.

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity scores. An introduction and experimental test. *Evaluation Review*, *29* (6), 530-558.

Morgan, S.L., & Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research*. Cambridge, N.Y.: Cambridge University Press.

Pagan, A., & Ullah, A. (1999). *Nonparametric econometrics*. Cambridge: Cambridge University Press.

Pohl, S., Steiner, P.M., Eisermann, J., Soellner, R., & Cook, T.D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31* (4), 463-479.

Reinowski, E. (2006). Mikroökonometrische Evaluation und das Selektionsproblem - Ein anwendungsorientierter Überblick über nichtparametrische Lösungsverfahren. *Zeitschrift für Evaluation*, *5* (2), 187-226.

Rosenbaum, P.R. (2002). *Observational studies*. New York: Springer (2. Aufl.).

Rosenbaum, P.R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, 88 (424), 1250-1253.

Rosenbaum, P.R., & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39* (1), 33-38.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70* (1), 41–55.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2004). *Evaluation. A systematic approach*. Thousand Oaks: Sage (7.A.).

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3* (2), 135–145.

Rubin, D.B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29* (3), 343-367.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66* (5), 688-701.

Rubin, D.B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52* (1), 249–264.

Schafer, J.L., Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Shadish, W.R., & Cook, T.D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*, 607-629.

Shadish, W.R., Clark, M. H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, *103* (484), 1334-1344.

Sianesi, B. (2004). An evaluation of the Swedish system of active labour market programmes in the 1990s. *The Review of Economics and Statistics*, *86* (1), 133–155.

Smith, J.A., & Todd, P.E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*, (1-2), 305–353.

Smith, J.A., & Todd, P.E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, *91* (2), 112-118.

Steiner, P.M., Cook, T.D., Shadish, W.R., & Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15* (3), 250-267.

Stockmann, R. (2006). *Evaluation und Qualitätsentwicklung: Eine Grundlage für wirkungsorientiertes Qualitätsmanagement.* Münster: Waxmann.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation*, *16* (2), 153-164.

Wilde, E.T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, *26* (3), 455–477.